

Всероссийская Научная Конференция
**Электронные библиотеки: Перспективные Методы и
Технологии, Электронные коллекции**
RCDL 2008: Дубна

Труды RCDL 2008

Лекция: Preservation Planning with Plato. [слайды доклада]

Kulovits H., Rauber A.

RCDL'1999–RCDL'2008: DL, VDL, Semantic Web/GRID, GRID. [слайды доклада]

Шириков В.П.

Решение некоторых задач Text Mining при помощи концептуальных графов. [слайды доклада]

Богатырев М.Ю., Тяхтин В.В.

Проблемы технологий создания систем смысловой обработки данных. [слайды доклада]

Барахнин В.Б., Федотов А.М.

Концептуализация данных наукометрических исследований в научных электронных библиотеках. [слайды доклада]

Зацман И.М.

О некоторых особенностях восприятия электронного документа.

Шкилев В.Д., Желева Н.А.

Онтологии для тематического поиска данных в коллекциях электронной библиотеки. [слайды доклада]

Вдовицын В.Т., Лебедев В.А.

Метаданные в системе управления многоязычной лингвистической базой знаний. [слайды доклада]

Лунева Н.В.

Применение библиографических ссылок в автоматизированном учете книгообеспеченности и в представлении результатов поисковых запросов. [слайды доклада]

Тимонина Л.С., Шарова Т.С.

Автоматическое построение терминологической базы знаний. [слайды доклада]

Чаньшев О.Г.

О распределенных фактографических системах. [слайды доклада]

Марчук А.Г.

Формализм записи лексико-синтаксических шаблонов в задаче автоматизации процесса построения онтологий.

Рабчевский Е.А., Булатова Г.И., Шарафутдинов И.М.

Электронная система на основе ms wss для гармонизации терминологии аналитической химии. [слайды доклада]
Широкова В.И., Колотов В.П., Аленина М.В

Анализ и сравнение систем интеграции неоднородных информационных ресурсов. [слайды доклада]
Вовченко А.Е., Калиниченко Л.А.

Использование онтологии верхнего уровня для отображения информационных моделей. [слайды доклада]
Скворцов Н.А., Ступников С.А.

Открытые электронные коллекции с адаптивным визуальным интерфейсом фасетной навигации.
Обухова О.Л., Гершкович М.М., Бирюкова Т.К., Соловьев И.В., Чочиа А.П.

Автоматическое формирование рубрикатора полнотекстовых документов. [слайды доклада]
Пескова О.В.

Создание легко обновляемых текстовых индексов. [слайды доклада]
Веретенников А.Б.

Программный комплекс «Смалт». [слайды доклада]
Рогов А.А., Гурин Г.Б., Котов А.А., Сидоров Ю.В., Суровцова Т.Г.

Концепция объемного текста.
Тупик Н.В.

Автоматическое составление обзорных рефератов новостных сюжетов. [слайды доклада]
Тарасов С.Д.

О методе автоматического реферирования, основанном на результатах рубрицирования документов. [слайды доклада]
Абрамов В.Е., Абрамова Н.Н.

RESUMAGIC: система автоматической обработки резюме. [слайды доклада]
Сафронов А.В.

Обнаружение ссылочного спама. [слайды доклада]
Шарапов Р.В., Шарапова Е.В.

Подход к выявлению подмножеств похожих документов. [слайды доклада]
Антонов А.В., Баглей С.Г., Мешков В.С.

Изучение характеристик сообществ русскоязычной блогосферы. [слайды доклада]
Сычев А.В., Гадебский И.А.

Ранжирование источников информации в системе мониторинга новостей infostream. [слайды доклада]

Ландэ Д.В., Брайчевский С.М., Дармохвал А.Т., Морозов А.Ю.

Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров.

Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В.

Автоматическая классификация веб-сайтов.

Маслов М.Ю., Пяллинг А.А., Трифонов С.И.

Проект электронной библиотеки методик и результатов исследований текстовых коллекций для системы «Источник». [слайды доклада]

Каргинова Н.В., Кравцов И.В., Москин Н.Д., Варфоломеев

Система поиска в электронной коллекции изображений петроглифов Карелии. [слайды доклада]

Рогов А.А., Рогова К.А., Спиридонов К.Н., Быстров М.Ю.

Алгоритмы распределения потоков информации в графической поисковой системе в реальном масштабе времени. [слайды доклада]

Гостев И.М., Подгорбунский А.Г.

Электронные коллекции ядерных данных как средства фундаментальных и прикладных исследований в области физики атомных ядер и ядерных реакций. [слайды доклада]

Бобошин И.Н., Варламов В.В., Комаров С.Ю., Песков Н.Н., Степанов М.Е., Чесноков В.В.

Новая полная электронная карта данных о форме и размерах атомных ядер.

Бобошин И.Н., Варламов В.В., Гангтрский Ю.П., Комаров С.Ю., Маринова К.П., Султанов Р.И.

Научные и научно-организационные электронные ресурсы Объединенного Института Ядерных Исследований. [слайды доклада]

Борисовский В.Ф., Кореньков В.В., Куняев С.В., Ленская Н.А., Мусульманбеков Ж.Ж., Никонов Э.Г., Фило

Библиотека программ JINRLIB. [слайды доклада]

Попкова Л.В., Сапожников А.П., Сапожникова Т.Ф., Федорова Р.Н.

Унификация структур данных в области изучения, освоения и использования ресурсов мирового океана. [слайды доклада]

Вязилов Е.Д., Федорцов А.А., Кобелев А.Е.

Развитие инфраструктуры распределенных хранилищ спутниковых данных: интегрированная распределенная среда неоднородных информационных ресурсов исследования земли из космоса.

Кудашев Е.Б., Филонов А.Н.

Вариации магнитного поля земли как составной элемент баз данных космических экспериментов по физике магнитосферы. [слайды доклада]

Зайцев А., Петров В., Одинцов В., Шилимов В.

Информационные системы, ориентированные на представление данных о погоде и климате в сети Интернет. [слайды доклада]

Лаврентьев Н.А., Фазлиев А.З.

Эволюция стандарта ZOOCOD— концепции отражения зоологических иерархических классификаций в плоских таблицах реляционных баз данных. [слайды доклада]

Лобанов А.Л., Смирнов И.С., Дианов М.Б., Голиков А.А., Халиков Р.Г.

Детерминированное и случайное в отождествлении результатов астрометрических наблюдений пульсаров. [слайды доклада]

Авраменко А.Е.

Электронные документы научно-технической информации в системе ВНТИЦ.

Голосов Ю.И., Брагина Г.А., Пржиялковская М.Н.

О создании открытой научной электронной библиотеки периодических изданий НАНУ. [слайды доклада]

Резниченко В.А., Проскудина Г.Ю., Кудим К.А.

Проект создания и реализации городского библиотечного портала Dubna Library. [слайды доклада]

Кореньков В.В., Никонов Э.Г., Филозова И.А.

Роль традиционных библиотек в формировании проблемно-ориентированных информационно-коммуникационных социальных пространств.

Мешков Н.А., Сорокина Л.П., Цветкова Н.В.

Создание электронной коллекции личных дел студентов Казанского университета (1917–1925 гг.).

Амалиева Г.Г., Елизаров А.М., Сальникова А.А.

Научно-образовательная электронная библиотека ВУЗА. [слайды доклада]

Абросимов А.Г., Зуев Д.С.

Результаты мониторинга и основные рекомендации по вопросам развития электронных библиотек вузов.

Котляр Э.А., Гужеля Д.Ю., Полихина Н.А.

Стендовый доклад: Ассоциативно-семантическая неоднородность восприятия информации студентами рискоопасных профессий и методологические подходы к ее формализации.

Талалаева Г.В.

Стендовый доклад: Подход к построению сервис-ориентированной электронной библиотеки в среде Semantic Web на основе репланирования.

Новицкий А.В.

Стендовый доклад: Автоматический синтаксический анализ русских текстов.

Леонтьева Ан.Б., Кагиров И.А.

Стендовый доклад: Разработка методов автоматического индексирования текстов на естественном языке для информационно-поисковых систем.

Бевзов А.Н.

Стендовый доклад: Группировка пользователей интернета, основанная на истории их веб-сессий.

Киселева Ю.

Стендовый доклад: Сервис для обработки удаленных рядов данных на основе WAVELET-функций.

Корсаков Е.В., Молородов Ю.И.

**Copyright 1999 - 2010 © Электронные библиотеки: Перспективные
Методы и Технологии, Электронные коллекции**
Контактный адрес: [contacts\[AT\]rcdl.ru](mailto:contacts[AT]rcdl.ru)

Сервис для обработки удаленных рядов данных на основе Wavelet-функций*

© Е.В. Корсаков, Ю.И. Молородов

Институт вычислительных технологий СО РАН
yumo@ict.nsc.ru

Аннотация

Описан сервис для математической обработки временных рядов данных наблюдений, хранящихся на сервере-хранилище информации. Доступ к данным и программам обработки организован с помощью Веб-интерфейса. Сервис обеспечивает доступ к данным и их обработку с помощью вейвлетов Морле и Добеши. Цель работы – создание доступной по сети Интернет распределенной информационно-вычислительной среды для обработки результатов измерений.

1 Введение

В последние годы возрос интерес к изучению проблемы глобальных изменений окружающей среды. Особую роль при этом играют атмосферные аэрозоли (АА). Они – неотъемлемая компонента атмосферы и объяснение всех сложных процессов, происходящих в ней, невозможно без подробной информации об их химическом составе и поведении. Специалистам, занимающихся прогнозированием, моделированием и анализом процессов, происходящих в окружающей среде необходимо своевременно получать информацию о характеристиках, химическом составе и поведении атмосферного аэрозоля.

Для этого необходимо разрабатывать информационные системы, в которых информация будет структурирована и собрана в виде, доступном для понимания пользователю любой степени подготовки. Эту проблему, как и задачу выявления регулярных сезонных и суточных колебаний в поведении аэрозолей можно решить, если создать информационно-вычислительные системы на основе архитектуры клиент-сервер. Они решают задачи сбора, обработки и публикации сведений об атмосферных аэрозолях, в том числе:

- Добавление, хранение, редактирование исходных данных.

- Обработку временных рядов данных современными алгоритмами.
- Обеспечение авторизации пользователей и ограничение доступа (при необходимости) к информационным структурам и данным.

Данные могут храниться как на стороне клиента, так и на серверах владельцев информации.

Разработанные интерфейсы должны обеспечить представление исходных данных или результат их обработки в пригодном для анализа виде. Современные математические алгоритмы обработки временных рядов данных наблюдений позволяют проводить их частотно-временной анализ, в том числе и с помощью Вейвлет-анализа [1–4].

2 Алгоритмы обработки и анализ

2.1 Расчет функции спектр размеров

Ранее [5] мы описывали метод измерения характеристик аэрозолей и методик обработки на основе регуляризации по Тихонову. Эти алгоритмы позволяют решить обратную задачу восстановления спектра размеров из разнообразных данных, полученных при измерениях. Спектр размеров является предпочтительным описанием для включения АА в какие-либо модельные расчеты. Дифференциальные характеристики атмосферного аэрозоля, как правило, представляются в форме, имеющей аналитическую зависимость от приведенного размера частиц, называемого радиусом или диаметром. Соответственно, для получения интегральных характеристик атмосферного аэрозоля, имеющих дальнейшее прикладное значение, необходимо проводить интегрирование дифференциальных характеристик с учетом спектра размеров. Программный модуль позволял привести многочисленные и разрозненные данные по физическим свойствам атмосферного аэрозоля к единообразной форме. В качестве такой формы мы использовали счетное распределение частиц по размерам $dN / d(\log(r))$. Для полидисперсного аэрозоля, значение измеренной концентрации $P(m)$ ($m = 1, \dots, M$), M – число экспериментов) может быть выражено через нормированный спектр размеров аэрозоля:

$$f(x) = dN / (2.303 * N * d(\log(r))). \quad (1)$$

В нашем случае $P(1) = N$ – полная счетная концентрация. Зная $f(x)$, можно восстановить искомую функцию «спектр размеров». Тогда спектр размеров

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

по счетной концентрации представляется в виде числа частиц dN в интервале размеров $d(\log(r))$.

2.2 Вейвлет-анализ: основные определения

Вейвлет-анализ – ориентирован на локализацию и классификацию особых точек сигнала, вычисление его различных фрактальных характеристик, частотно-временной анализ нестационарных сигналов.

Следуя [1] опишем основные принципы анализа временных рядов. Вейвлет-преобразование $Wf(x, a)$ одномерного сигнала $f(t)$ заключается в его разложении по базису, сконструированному из некоторой специальной материнской базисной функции (вейвлета) – $\Psi(t, x, a)$ посредством масштабных изменений a и переносов x :

$$(W_{\Psi}f)(x, a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} \Psi\left(\frac{t-x}{a}\right) f(t) dt \quad (2)$$

При каждом масштабе a , согласно (2) базисный вейвлет растягивается по горизонтали и сжимается по вертикали.

Т.о., в отличие от преобразования Фурье, вейвлет-преобразование обеспечивает двухмерную развертку исследуемого одномерного сигнала, при этом масштаб и координата рассматриваются как независимые переменные. В результате этого появляется возможность анализировать свойства сигнала одновременно во временном и в частотном пространствах.

Для анализа вариаций мы будем использовать вейвлет Морле. Он хорошо приспособлен для анализа квазипериодических процессов. Он позволяет наглядно представить результаты анализа и легко интерпретировать их в терминах анализа Фурье.

$$\Psi(t) = \frac{1}{2\pi} \exp\left(i\omega_0 t - \frac{t^2}{2}\right) \quad (3)$$

В общем случае выбор анализирующего вейвлета является неоднозначным и зависит от конкретной задачи. Вейвлет Морле был выбран потому, что он хорошо приспособлен для анализа квазипериодических процессов (атмосферные вариации, являющиеся собственными колебаниями атмосферы [3], принято считать квазипериодическими), также он позволяет наглядно представить результаты анализа и легко интерпретировать их в терминах анализа Фурье.

2.3 Структура сервиса обработки данных

Сервис обработки, реализующий возможности Вейвлет-анализа, состоит из пяти функциональных модулей (рис. 1).

По запросу клиента выполняется авторизация и подключение к удалённому серверу по протоколу FTP. Затем клиент просматривает содержимое каталогов, выбирает интересующие его файлы и копирует их на сервер-обработчик. Модуль подготовки данных к анализу последовательно извлекает информацию из файлов, и готовит их к анализу. При этом удаляется служебная информация, из строк выделяются сведения о времени и амплитуде сигнала

ла в каждый момент времени, выделяется «сбойные» показания (сбой аппаратуры регистрации), сигнал, соответствующий «выпавшему» времени заменяется на «-1». После этого значения сигнала построчно записываются в файл temporary.txt. Далее модуль подготовки передает файл на кластер. Обращение к кластеру вызвано достаточно большими затратами времени на обработку. На кластере хранятся алгоритмы: быстрого преобразования Фурье, а также необходимые для работы вейвлет-функции: вейвлет Хаара, вейвлеты Добеши и вейвлеты Морле. По окончании обработки посылается сообщение клиенту. После этого клиент сможет посмотреть таблицы обработки и нарисовать графики.

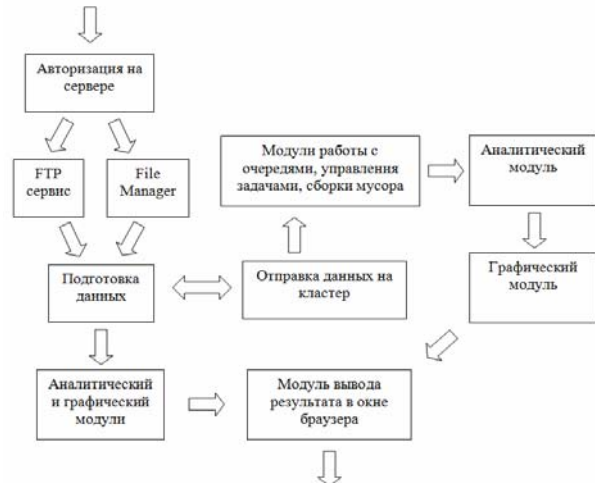


Рис. 1. Общая схема сервиса

2.4 Схема измерений

В пригородной зоне г. Новосибирска в п. Ключи (9 км от Академгородка) находится станция наблюдения за состоянием пограничного слоя атмосферы. Здесь проводятся измерения массовой концентрации субмикронной фракции атмосферного аэрозоля. Для этого используется нефелометр ФАН-А, связанный с компьютером IBM-PC класса Pentium IV. Измерения проводятся в течение одного месяца каждого сезона с частотой одно измерение в минуту (1440 замеров/сут.).

Файлы данных переносятся для постоянного хранения на сервере организации, хозяина данных.

2.5 Результаты обработки

Для работы с данными мы использовали ортогональный вейвлет-анализ на основе вейвлетов Добеши. Они позволяют нам исключить шум сигнала, сжать его, а также выделить его низкочастотную и высокочастотную компоненты. Анализ данных на основе вейвлетов Морле ориентирован на локализацию и классификацию особых точек сигнала, вычисление его различных фрактальных характеристик, частотно-временной анализ.

На рис. 2–3 представлены результаты обработки ряда данных измерений, полученных в течение суток 5 ноября 2006 г. в пункте наблюдений Ключи.

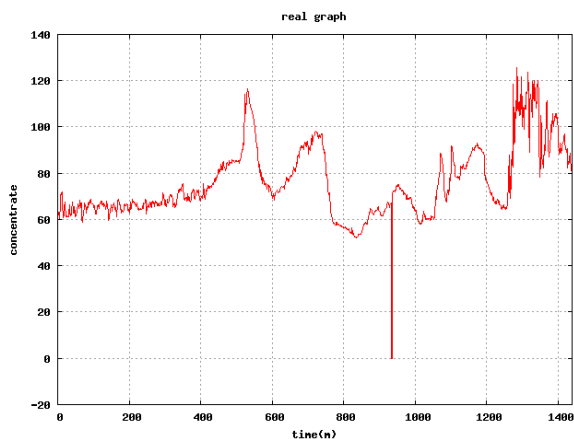


Рис. 2. Исходный сигнал

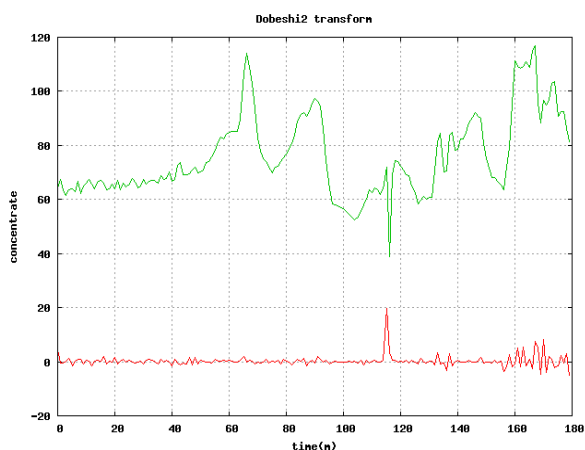


Рис. 3. Обработка Вейвлетом Добеши



Рис. 4

На рис. 3 показана низкочастотная составляющая сигнала (верхний график) и его высокочастотная компонента (нижний график). На рис. 4 показан результат обработки среднечасового сигнала полученного за 1 месяц в январе/феврале 2006 г. На верхнем графике показано его среднее значение, а на нижнем – среднее квадратичное отклонение от среднего. Поскольку на нем явно не видны какие-либо периодические структуры, то это сигнал был обработан с помощью вейвлетов Морле. Рис. 5 показывает результат обработки. Более явно показаны

периодические структуры на рис. 6 (увеличена нижний фрагмент рис. 5).

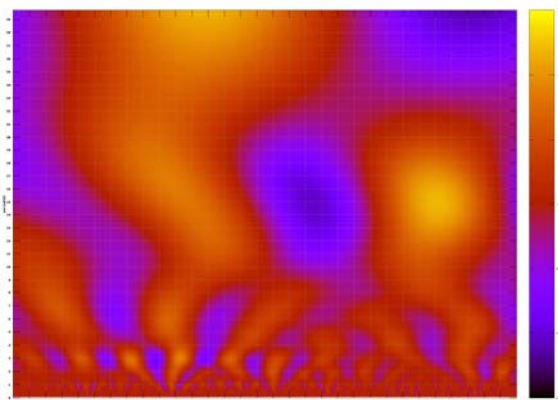


Рис. 5. Вейвлет-образ зима. Функции Морле

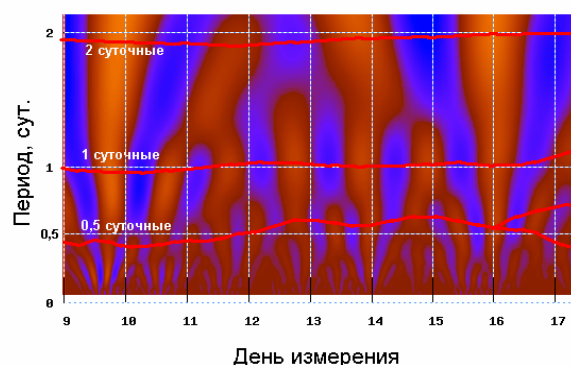


Рис. 6. Пространственно-временная изменчивость массовой концентрации субмикронной фракции атмосферного аэрозоля в пригородной зоне Новосибирска (зима 2006 г.)

Поскольку целью работы являлась только демонстрация возможностей метода обработки, то мы не показали обработку временных рядов данных полученных в весенний, летний и осенний периоды 2006 г.

3 Заключение

Ортогональное вейвлет преобразование Добеши используется для подавления шумов в сигнале. Как видно из полученных результатов быстрое вейвлет преобразование (БВП) Добеши с $M = 1$, не совсем подходит для подавления шумов в сигнале, что видно из рисунка раздела (3.2). Данное преобразование снизило среднее значение концентрации, хотя закономерности в поведении сигнала все же сохранились. Вейвлет преобразование с $M = 2$ (3.3), сохраняет среднее значение концентрации в сигнале в отличие от вейвлета Хаара.

Непрерывное вейвлет преобразование Морле может быть использовано для локализации и классификации особых точек, частотно-временного анализа нестационарных сигналов. Преобразование Морле мы также использовали для диапазона измерений полученных за один месяц.

Литература

- [1] Хуторова О.Г., Тептин Г.М. Исследование мезомасштабных вариаций в тропосфере по наблюдениям концентрации примесей // Известия РАН. Физика атмосферы и океана/ – 2001, т. 37, №6. – С. 853–856.
- [2] Torrence G. and Compo G.P. A practical guide to wavelet analysis. – Bull. Amer. Meteorol. Soc., 1998, vol.79, No 1, pp. 61–78.
- [3] Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения. – Т. 166, № 11. – С. 1145–1170.
- [4] Шитов А.Б. Разработка численных методов и программ, связанных с применением вейвлет-анализа для моделирования и обработки экспериментальных данных // Дис. на соискание ученой степени канд. физ.-мат. наук. – ИвГУ, 2001. – С. 125.
- [5] Молородов Ю.И., Куценогий П.К., Сударикова И.А. Содержательное наполнение Атласа «Атмосферные аэрозоли Сибири» // География и природные ресурсы. – 2004. Спец. выпуск. – С. 86–90.

Wavelet-function processing service for remote data rows

E. Korsakov, Yu. Molorodov

The technology and service was developed for remote processing of regular measurements experimental data files of submicronic fraction atmospheric an aerosol (AA), received by measurements of computerized nefelometer a FAN in a ground layer of the atmosphere, near the Novosibirsk (settlement Kluchi). Necessity of such data processing of measurements for the atlas is caused not only ordering of the saved up empirical numbers, but also carrying out of statistical processing together with other components AA, for example, organic and element carbon, elements etc. The purpose of offered processing is revealing of cyclic laws of a signal: daily, monthly, seasonal fluctuations and definition of local upper limit. Service provides connection to a remote data storage server, the primary analysis and preparation of the data on a server output agent and carrying out of the analysis on the basis of wavelet-functions of Haara, Dobeshi and Morlet. Results of the analysis are represented in the form convenient for the further analyze.

* Работа поддержана РФФИ (грант № 06-07-89060), президентской программой «Ведущие научные школы РФ» грант № НШ 931.2008.9 и Интеграционным проектом СО РАН № 34.

Группировка пользователей Интернета, основанная на истории их веб-сессий

© Юлия Киселева

Научный руководитель: Борис Асенович Новиков

Санкт-Петербургский государственный университет
julianakiseleva@gmail.com

Аннотация

В данной работе рассматривается вопрос персонализации пользователей Интернета на основе групп, отражающих интересы. В настоящее время является наиболее важным для исследований в области логического анализа данных Интернета. Существующие подходы группировки веб-пользователей основываются на снимках их веб-сессий. Данный подход описан в [1]. Группы пользователей Интернета образуются, исходя из истории их веб-сессий, широко используется в области веб-рекламы

1 Введение

Логический анализ данных Интернета – набор подходов для выявления шаблонов поведения пользователей. На данный момент является активной областью для исследований [1]. Существующие подходы и техники включают в себя статистический анализ [1], последовательные шаблоны [2], ассоциативные правила [3], классификацию [4] и другие методы. Важным аспектом логического анализа данных Интернета является выявление групп пользователей с близкими интересами.

Существующие подходы для группировки веб-пользователей состоят из трех этапов:

1. Подготовка данных - данная фаза представляет собой извлечение данных с сервера, затем проводится идентификация пользователей и их веб-сессий.
2. Выявление групп пользователей.
3. Анализ полученных групп.

Веб-данные по природе своей являются динамически развивающимися, и как следствие этого факта существуют два подхода для группировки пользователей:

1. возможность выявления похожих пользователей в процессе эволюции данных;

2. возможность выявления похожих пользователей за фиксированный промежуток времени. В этом случае результат группировки нуждается в постоянном обновлении, по мере поступления новых данных.

В данной статье мы сфокусировались на втором подходе выявления похожих пользователей. Проанализировали результаты – группы веб-пользователей, которые были полученные с использованием двух метрик, введенных в работе для измерения близости между пользователями. Целью представленной работы является получение единой методики оценки близости между пользователями Интернета.

2 Метрики похожести пользователей

2.1 Набор данных

Для исследования был использован лог, который содержит информацию о 1343 пользователях, общее количество запросов в нем – 66380. Будем называть запросы одного пользователя документом. Далее для построения групп схожих пользователей будем использовать поисковые их запросы.

2.2 «Очистка» данных

Перед началом эксперимента следует очистить данные, для этих целей можно использовать WordNet. WordNet – это большая лексическая база данных английского языка [6], при помощи него можно избавиться от опечаток, совершенных пользователями. Так же помощью отдельного фильтра убираем стоп-слова, такие как «how», «and» и другие. Проведенная очистка уменьшила рассматриваемое множество на 1.4%.

2.3 Обработка данных

Целью исследования является нахождение наиболее близких пользователей и объединение их в группы. Предполагаем, что пользователи, попавшие в одну и ту же группу, обладают схожими интересами. В данной работе используем две метрики для определения близости между запросами пользователей, они описаны ниже. Сначала создаем пространство всех слов, которые встретились в доку-

ментах. Затем для каждого пользователя получаем вектор весов слов, которые встречаются в его и только в его документе (в наборе его запросов). Вектор выглядит следующим образом:

$$d_j = w(t_1), w(t_2) \dots w(t_n), \quad (1)$$

где $w(t_j)$ – это вес $tf-idf$ слова t_j во всем множестве слов (term frequency–inverse document frequency) [5], где tf частота слова t_j :

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

в представленном выражении $n_{i,j}$ – это сколько раз встретилось рассматриваемое слово в документе d_j , а знаменатель представляет собой количество всех слов в документе d_j .

Обратная частота – это мера, показывающая общую важность слова:

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

- где $|D|$ общее количество документов в наборе.
- $|\{d_j : t_i \in d_j\}|$ количество документов, в которых встречается слово t_i .

Согласно второй метрике, каждый пользователь представлен набором векторов, координаты которых представляют собой веса слов, встретившихся в его запросе. Вектор-запрос представлен в следующем виде:

$$Q_i = w(t_1), w(t_2) \dots w(t_n), \quad (2)$$

где $w(t_j)$ – это вес слова j в запросе Q_i .

2.4 Метрика (1): средняя мера близости

Итак, выше определи метрики, которые помогут нам объединить пользователей с похожими интересами в группы.

Первая метрика, представленная в 2.2, для определения близости между пользователями использует скалярное произведение векторов их весов. Ясно, что наиболее близкие пользователи имеют наибольшее значение скалярного произведения их весовых векторов.

В результате проведения эксперимента получаем матрицу близости между всеми пользователями, каждый элемент матрицы $\{a_{ij}\}$ это мера близости между i -м и j -м пользователями. Соответственно, $a_{ij} \in [0,1]$.

2.5 Метрика (2): максимальная схожесть запросов пользователей

Вторая метрика, определенная в 2.2, определяет близость между пользователями как максимум скалярного произведения весовых векторов их запросов. Соответственно, если пользователи имеют два одинаковых запроса, то они имеют близость = 1. Такая метрика не чувствительна в случаях, когда пользователи U_1 , U_2 и U_3 имеют одинаковые запросы

(например, “vegas”), близость между этими пользователями равна 1, но мы можем встретить ситуацию, когда U_1 ввел запрос “vegas” 10 раз, U_2 – 8 раз, а U_3 – только 1 раз. В подобных ситуациях определенная выше метрика работает не слишком удачно.

2.6 Диаграммы распределения, полученных метрик

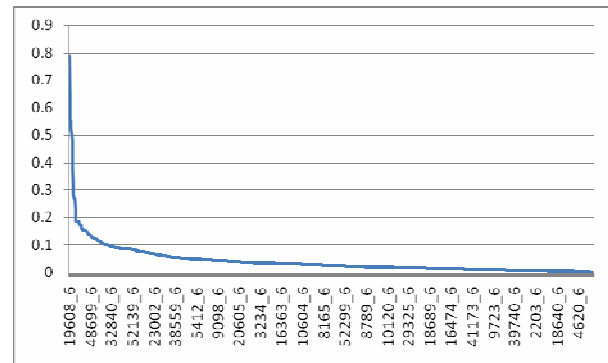


Диаграмма 1: распределение величин, полученное на основе первой метрики, ось x – user_id, ось y – величины первой метрики.

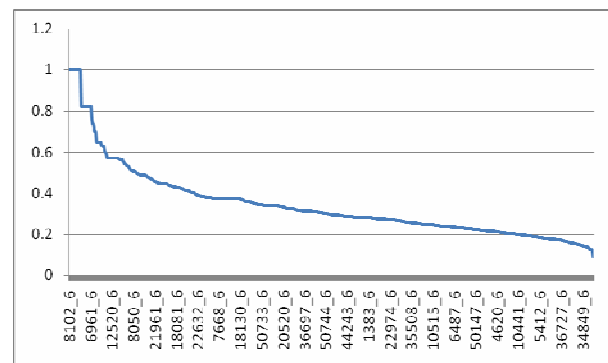


Диаграмма 2: распределение величин, полученное на основе второй метрики, ось x – user_id, ось y – величины второй метрики.

3 Полученные результаты и будущие эксперименты

Основной целью исследования является создание метрики для нахождения схожих пользователей, которая будет лишена недостатков обнаруженных, в процессе экспериментов, в представленных метриках. Для оценки полученных данных были использованы метод «Общего котла» [7] и анализ тематических срезов данных, т.е. рассматривался срез данных, который содержал одно или несколько тематических слов. Затем ассессорами оценивалось, насколько реально близки запросы, находящиеся в полученном срезе.

Итогом исследований представляется создание карты интересов пользователей Интернета, что является важным особенно при показе тематической рекламы и также упорядоченности информации, извлечение полезных знаний.

Литература

- [1] C. Buchwalter, M. Ryan, and D. Martin. The state of online advertising: data covering 4th Q 2000. In TR Adrelevance, 2001.
- [2] Q. Yang, H.H. Zhang, and T. Li. Mining web logs for prediction models in www caching and prefetching. In *Proc.of ICCNMC'01*, 2001.
- [3] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from the web usage data. In *Proc. Of WIDM*, 2001.
- [4] T. Li, Q. Yang, and K. Wang. Classification pruning for web-request prediction. In *Proc.of WWW*, 2001.
- [5] Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: *Current Trends in Database Technology – EDBT*, Springer-Verlag GmbH (2004) 588–596
- [6] Agirre, Eneko and David Martinez. “Integrating selectional preferences in WordNet.” In: *Proceedings of the first International WordNet Conference*, Mysore, India, 21-25 January 2002.
- [7] И. Некрестьянов, М. Некрестьянова, А. Нозик. К вопросу об эффективности метода «общего котла» // *Труды RCDL'2005*. – Ярославль, 2005.

Grouping Web-users based on Query Log

Julia Kiseleva

Web pages are personalized based on the interests of an individual. Personalization implies that the changes are based on implicit data, such as items purchased or pages viewed. In our research we don't approach to strongly user's personalization. Generally, typical web user grouping approach consists of three phases: data preparation, group discovery and group analysis. This is work in progress report. At this stage of our research we focus on user similarity metrics that later will be user to group users. In this report we present description of our approach, define several metrics and conduct experiments to evaluate their quality.

Разработка методов автоматического индексирования текстов на естественном языке для информационно-поисковых систем

© Алексей Николаевич Бевзов

Институт автоматики и электрометрии СО РАН
bvz@iae.nsk.su

Аннотация

Рассматривается разработка методов автоматического индексирования текстов на естественном языке (ЕЯ), которые могут использоваться при создании информационно-поисковых систем (ИПС). Основное внимание уделяется методологическому подходу, предназначенному для определения ключевых слов (терминов) в тексте на ЕЯ и определения веса этих терминов. Суть подхода состоит в том, что текст рассматривается не как простой набор слов ("bag of words"), а учитывается положение слов в предложении.

1 Введение

В настоящее время задача автоматической обработки и поиска смысловой информации в текстах на естественном языке является весьма актуальной. В этой области мы сталкиваемся с парадоксальной ситуацией [1]. Суть ее в том, что с одной стороны, в связи с развитием компьютерных технологий и средств связи доступ к информации упрощается. С другой стороны, возможность точного поиска необходимой информации уменьшается ввиду все возрастающего потока самой информации. Существующие на сегодняшний день методы обработки и поиска информации не вполне отвечают современным требованиям.

Среди наиболее известных методов, используемых для поиска информации, можно отметить: 1) библиотечные классификации (УДК, ББК, Классификация Дьюи, Классификация конгресса США) [2]; 2) поиск информации с помощью поисковых машин (ПМ) Интернет с указанием ключевых слов в поисковом запросе [3]; 3) использование лингвистических процессоров для автоматизированной обработки информации в текстах на ЕЯ [1], [4], [5]. Такие методы обработки и поиска информации наряду с положительными свойствами имеют ряд недостатков принципиального характера.

Ограничения библиотечных классификаций состоят в том, что, они требуют обязательного знакомства как со стороны администратора ИПС, так и со стороны пользователя, что существенно затрудняет ввод новых документов в поисковые системы, а также их поиск. Недостаток поиска информации по ключевым словам с помощью ПМ Интернет состоит в том, что ПМ часто выдают огромное количество ссылок, обрабатывать которые приходится вручную. Принципиальное ограничение лингвистических процессоров обусловлено их привязкой к тому языку, на котором ИПС может обрабатывать и производить поиск смысловой информации, т.к. в основе логики работы таких процессоров лежит знание о семантических и синтаксических конструкциях конкретных ЕЯ.

Данная работа описывает методику и результаты проведенных исследований, направленных на развитие и поиск альтернативных методов обработки и поиска информации в текстах на ЕЯ. В основе этих методов лежат идеи синергетического подхода к изучению сложных явлений [6]. В качестве исследуемых образцов текстов были использованы описания паттернов проектирования, приведенные в работе [7]. Данное исследование возникло как результат развития ИПС [8], первоначально предназначенной для более удобной работы с паттернами проектирования, которые используются при разработке программного обеспечения.

2 Текст как сложный синергетический объект

Суть предлагаемого подхода заключается в том, чтобы посмотреть на текст как на некоторый сложный синергетический объект, развернутый в пространстве и времени и получить различные временные ряды (ВР) из обычного текста на ЕЯ, а затем исследовать полученные данные с целью поиска в этих текстах наиболее значимой семантической информации.

Для исследования текста с помощью различных алгоритмов синергетики [9] была разработана специальная методика предварительной обработки текста с целью получения различных ВР. Ниже приведены основные этапы этой обработки.

1) Нормализация текста. На этом шаге производилась «оцифровка» текста. Сначала из текста выбрасывались предлоги и каждому слову текста присваивался уникальный идентификационный номер.

2) Выделение повторяющихся слов (ПС) выполнялось для того, чтобы отслеживать динамику эволюции текста на всем его протяжении.

3) Определение всех предложений в тексте. Критерием наличия предложения являлись такие знаки препинания, как ". ? ! ;".

4) Определение границ областей относительно позиции гармонического центра (ГЦ). Для каждого предложения определялись значения ГЦ, соответствующие пропорции золотого сечения. В этой части методики мы опирались на идеи, представленные в [10]. Позиция ГЦ равна 0.618 от начала предложения и измеряется в словах. Кроме того, выделялись границы трех областей относительно позиции ГЦ.

5) Расчет параметров ГЦ для каждого слова. На этом этапе для каждого слова проводился расчет параметров этого слова относительно позиции ГЦ предложения.

После выполнения указанных этапов предварительной обработки были предложены различные модели получения ВР, которые в дальнейшем исследовались синергетическими методами с целью выявления наиболее значимой смысловой информации.

3 Построение различных моделей текста

Для получения из текстов различных моделей ВР использовалась различная интерпретация ряда параметров, как, например, определение наблюдаемой переменной (НП), т.е. ключевого слова, способ получения значений НП (ЗНП) и т.д. Идентификация каждой модели задавалась с помощью: 1) определения того, как получаются ВР в результате компьютерной обработки текста; 2) определения параметра time (измеряется в словах) – момент проведения измерения ЗНП. 3) определения параметра deltaT (измеряется в словах) – длина временного интервала от одного времени измерения до другого. 4) определения параметра N – количество точек ВР и как они получаются. 5) определения того, что понимается под НП; 6) определения того, как считаются ЗНП;

В целом было рассмотрено несколько десятков моделей обработки текста с целью получения различных ВР. Ниже приводится описание некоторых моделей.

1) Идея получения ВР состоит в том, чтобы проанализировать динамику повторяемости ПС на протяжении всего текста через равные заданные «временные интервалы». Под «временными интервалами» подразумеваются фиксированные фрагменты текста, измеряемые в словах.

2) Идея получения ВР состоит в том, чтобы проанализировать динамику попадания выбранного ПС в разные области ГЦ на протяжении всего текста в каждом предложении.

3) Идея получения ВР состоит в том, что нас интересует динамика смещения позиции выбранного ПС относительно позиции ГЦ в предложении на протяжении всего текста. При этом, если ПС попадает в центр области ГЦ, то ЗНП равно некоторому максимальному значению, и это значение уменьшается до нуля по мере движения к левой или правой границе предложения.

4 Обработка ВР с целью получения некоторых идентификационных характеристик текста на ЕЯ

Обработка ВР для каждой из моделей велась аналогично тому, как это можно сделать при исследовании различных физических объектов. В работе использовался один из алгоритмов обработки ВР, описанный в [6], а именно – алгоритм расчета фрактальной размерности реконструированного аттрактора для идентификации интересующего нас объекта.

Предполагалось, что единственной информацией, которой мы обладаем относительно исследуемого объекта (в нашем случае – текста), является ВР (1), полученный согласно описанной выше методике обработки текста

$$X_0(t_0), \dots, X_0(t_i) \dots X_0(t_N), \quad (1)$$

где

$X_0(t_0)$ – экспериментально измеренное значение какой-либо переменной X_0 исследуемого объекта в момент времени t_0 .

$X_0(t_i)$ – экспериментально измеренное значение этой же переменной X_0 в момент времени $t_i = t_0 + \Delta t \cdot i$.

Из этого ВР, получались n -мерные вектора $X1, X2, \dots, XN$ для реконструкции аттрактора. Значения компонент этих векторов получаются из элементов исходного ВР так, как это показано в таблице 1.

Таблица 1. Получение векторов реконструкции из исходного ВР

Вектор $X1$	Вектор XN
$X_0(t_0)$	$X_0(t_N)$
$X_0(t_0 + \Delta t)$	$X_0(t_N + \Delta t)$
.....	
$X_0(t_0 + (n-1) \cdot \Delta t)$	$X_0(t_N + (n-1) \cdot \Delta t)$

Для этих векторов можно получить интегральную корреляционную функцию аттрактора по формуле:

$$C(r) = \frac{1}{N^2} \sum_{\substack{i,j=1 \\ i \neq j}}^N \theta(r - |X_i - X_j|), \quad (2)$$

где

N – количество точек ВР (точнее, количество векторов реконструкции)

i, j – номера векторов реконструкции ($1 - N$)

$\theta(x)$ – функция Хевисайда: $\theta(x) = 1$ если $x > 0$,

$\theta(x) = 0$ если $x < 0$

X_i – точка (вектор) фазового пространства с координатами $(X_0(t_i), \dots, X_0(t_i + (n-1) \cdot \Delta t))$

$|X_i - X_j|$ – расстояние между точками X_i и X_j

r – некоторая заданная величина.

Если с учетом (2) построить график зависимости $\ln(C(r))$ от $\ln(r)$, то размерность аттрактора (d) исследуемой динамической системы будет определяться наклоном зависимости $\ln(C(r))$ от $\ln(r)$.

С целью поиска наличия аттрактора исследуемого объекта использовался описанный в [6] алгоритм.

1) Получить несколько графиков для корреляционной функции (2), исходя из данного ВР (1), рассматривая последовательно возрастающие значения размерности фазового пространства n .

2) Посмотреть, как изменяется при возрастании n наклон этих графиков d вблизи начала координат (2).

3) Если величина d в зависимости от n выходит на плато выше некоторого относительно небольшого n , то представленная данным ВР система должна иметь аттрактор. В этом случае размерность d представляет собой минимальное число переменных, необходимых для моделирования поведения, соответствующего данному аттрактору.

Кроме построения графиков корреляционной функции, для исходных ВР строились также графики хаотических ВР (ХВР). Хаотичные временные ряды получались из исходных текстов, в которых хаотичным образом были перемешаны слова.

Критерием возможности использовать для индексирования тот или иной набор повторяющихся слов считалась ситуация, когда графики наклона корреляционной функции для исходного ВР и аналогичные графики для ХВР будут отличаться.

После получения различных ВР для указанных в п. 3 моделей, был проведен численный эксперимент, в котором эти ВР были обработаны согласно описанной выше методике.

5 Результаты экспериментов

Ряд проведенных экспериментов показал, что для некоторых моделей (модель 1) графики, полученные в результате обработки ВР и ХВР качественно не отличаются.

Для некоторых моделей (модификация модели 3) можно наблюдать качественное различие в поведе-

нии графиков обработки ВР и ХВР. Так, на рис. 1 и 2 представлены результаты обработки временных рядов, полученные для повторяющегося слова «адаптер» в тексте с описанием паттерна проектирования «Адаптер».

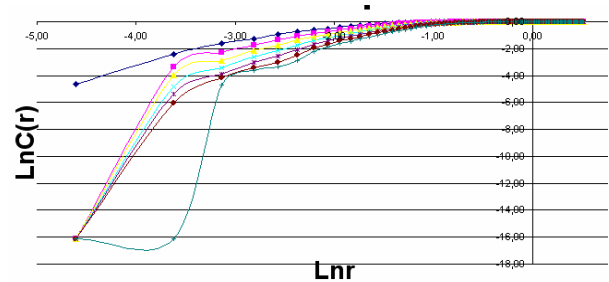


Рис. 1. Графики наклона корреляционной функции для исходного ВР.

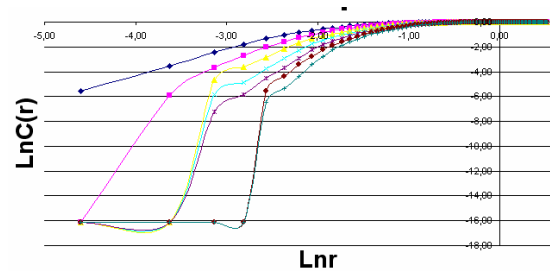


Рис. 2. Графики наклона корреляционной функции для хаотичного ВР.

Кроме указанного типа обработки ВР, опираясь на идею модели 3, был проведен расчет веса терминов (повторяющихся слов) документов, равных усредненной величине близости повторяющихся слов в тексте к положению ГЦ в предложении. После этого были проведены численные эксперименты для сравнения результатов поиска по предложенной методике расчета веса терминов с классическими методами расчета веса терминов.

Для выполнения процедуры поиска каждый документ был представлен в виде вектора в N -мерном евклидовом пространстве (N – количество терминов). Компоненты векторов соответствовали весу терминов, а коэффициент подобия между векторами документов и запросов считался по косинусной метрике.

Список терминов был образован из 70% наиболее частых слов с отсечением 15% наиболее частых и 15% наименее частых слов. Результаты проведенных экспериментов представлены в таблице 2. При этом коллекция текстовых документов состояла из описаний 23 паттернов проектирования, как они приведены в работе [7].

В качестве запросов использовались документы из коллекции документов в усеченном виде. Усеченный документ-запрос, которому необходимо было найти наиболее соответствующий ему документ в коллекции, формировался из 75, 50 и 25% исходного текста, начиная от начала. В таблице результаты поиска, соответствующие различным степеням усечения текста, разделены символом “/”.

Таблица 2. Результаты поиска документов

DP	tf (b)	tf	tf*idf	tf ГЦ
1	15/21/21	1/2/4	1/8/7	1/12/11
2	6/6/6	1/2/9	1/1/4	1/1/2
3	10/10/14	1/1/1	1/1/2	1/1/5
4	3/13/13	1/2/2	1/1/1	1/1/1
5	5/5/5	1/1/1	1/1/1	1/1/1
6	12/8/18	1/1/1	1/3/3	1/2/3
7	4/4/5	1/1/1	1/1/1	1/1/13
8	11/10/10	1/1/6	2/1/2	1/1/3
9	3/3/3	1/1/2	1/1/2	1/1/1
10	9/8/13	1/1/1	1/1/2	1/1/1
Среднее	8.8/ 9.8/ 10.9	1.0/ 1.3/ 2.8	1.0/ 1.9/ 2.5	1.0/ 2.2/ 5.1

При этом в столбцах таблицы использованы следующие обозначения:

DP – порядковый номер паттерна в работе [7];

tf (b) – бинарный расчет веса термина (1 – слово присутствует в документе, 0 – отсутствует);

tf – расчет веса термина только исходя из его частоты употребления в документе;

tf*idf – классический расчет веса термина с учетом его частоты как в одном документе, так и во всей коллекции документов;

tf_ГЦ – усредненная величина близости повторяющихся слов к положению ГЦ в предложении (1 – если слово находится в положении ГЦ и линейное убывание до 0 на границах предложения).

В столбцах tf(b), tf, tf*idf, tf_ГЦ приведен ранжированный порядковый номер «правильного» документа, полученного в результате поиска, т.е. чем ниже этот номер, тем точнее выполнен поиск.

6 Заключение

Предложенная методика выявления ключевых слов в тексте и проведенные по этой методике численные эксперименты, позволяют говорить об обнадеживающих результатах данного подхода, поскольку для некоторых моделей текста поведение текста, содержащего смысл, отличается от поведения хаотически перемешанного текста. Проведенные эксперименты по поиску документов в коллекции документов позволяют говорить о сопоставимости результатов поиска в случае расчета веса терминов по предложенной методике и с использованием классических методов поиска.

Литература

- [1] Краюшкин Д.В. Анализ технологий предварительной обработки документальной информации // Системы и средства информатики. – Вып. 15 / под ред. чл.-кор. РАН И.А. Соколова. – М., 2005.
- [2] Мидоу Ч. Анализ информационно-поисковых систем. – М., 1970.
- [3] Адамович, И.М. Поиск информации в WEB. Сравнительная оценка поисковых машин / И.М. Адамович, М.Ю. Заикин, Д.В. Земсков, А.Н. Пешков // Системы и средства информатики. – Вып. 13 / под ред. чл.-кор. РАН И.А. Соколова. – М., 2003.
- [4] Козеренко Е.Б. Методы категоризации в построении многоязычного лингвистического процессора // Системы и средства информатики. – Вып. 15 / под ред. чл.-кор. РАН И.А. Соколова. – М., 2005.
- [5] Кузнецов И.П. Особенности обработки текстов естественного языка на основе технологии баз знаний. // Системы и средства информатики. – Вып. 13 / отв. ред. чл.-кор. РАН И.А. Соколов. – М., 2003.
- [6] Пригожин И., Стенгерс И. Познание сложного. – М., 2003.
- [7] Гамма Э., Хелм Р., Джонсон Р., Влиссидес Дж. Приемы объектно-ориентированного проектирования. Паттерны проектирования. – СПб., 2001.
- [8] Бевзов А.Н. Использование экспертных систем в задачах проектирования программного обеспечения // Датчики и системы. – 2002, № 12.
- [9] Малинецкий Г.Г., Потапов А.Б. Современные проблемы нелинейной динамики. – М., 2002
- [10] Москальчук Г.Г. Структура текста как синергетический процесс. – М., 2003.

Development of automatic indexation methods for natural text processing for data retrieval systems

Alexey Bevzov

The development of NLP (natural language processing) methods that can be used for automatic indexation in DRS (data retrieval systems) is considered. Main attention is given to methodological approach of extracting key words (terms) and their weight definition from natural text. The essence of the proposed approach is that text is considered not as a “bag of words”, but words position in clause is taken into account.

Автоматический синтаксический анализ русских текстов

Анастасия Леонтьева, Ильдар Кагиров

Санкт-Петербургский институт информатики и автоматизации РАН
{an_leo, kagirov}@iias.spb.su

Аннотация

В данной статье описывается структура разработанного модуля автоматического синтаксического анализа (МСА). Данный модуль был разработан на основе модуля морфологического анализа русского языка SMART [1]. МСА служит для анализа структуры простых предложений и выделения в них синтаксических групп. Планируется внедрение МСА в систему распознавания русской слитной речи для повышения качества распознавания.

1 Введение

МСА – это программа или часть программы, выполняющая синтаксический анализ. На сегодня создание автоматического МСА является одной из самых актуальных задач в компьютерной лингвистике, решение которой позволило бы достичь высокого уровня формализации языковых структур в разнообразных прикладных целях: от создания систем автоматического распознавания речи до поисковых систем в Интернет.

Под целью синтаксического анализа в настоящей статье понимается вычленение базовых синтаксических структур и установление синтаксических связей между ними. На выход МСА поступает цепочка слов, разбитая на группы, причем каждая группа имеет связь с другими группами. Кроме того, МСА идентифицирует такие синтаксические категории, как подлежащее и сказуемое.

Однако создание МСА для русского языка упирается в большое количество сложностей, связанных с недостаточно разработанной теоретической базой в общем и прикладном языкознании; структуры человеческого языка отличаются разнообразием и часто высоким уровнем сложности, предусмотреть который чрезвычайно тяжело. В связи с этим в настоящей статье предлагается структура МСА, работающего с простыми синтаксическими структурами; создание МСА, справляющегося с текстом на русском языке любой сложности, представляется на настоящем этапе невозможным.

Предложенный нами модуль синтаксического анализа является частью анализатора текста русского языка SMART (на нынешний момент разрабатывается), ориентированного на грамматический разбор текстов на литературном русском языке. Наш МСА выгодно отличается от многих разработок в этой области своей ориентированностью на синтаксический уровень языка, а не исключительно на формальные (внешние) признаки словоформ. Иными словами, на выход МСА подается *синтаксическая структура*, а не цепочка словоформ с простейшей разметкой по морфологическому согласованию.

Например, в МСА, демонстрируемом на сайте <http://www.aot.ru>, реализован именно «согласовательный» подход, при котором МСА ищет морфологические характеристики, свидетельствующие о зависимости двух и более членов предложения, тогда как структуры типа «очень хорошо» или «в пальто» остаются неразобранными. К сожалению, такой подход на сегодняшний день является хоть и самым примитивным, но и довольно надежным – зачастую представить более сложные синтаксические зависимости в терминах грамматики составляющих оказывается проблематичным.

2 Синтаксический анализ предложения

Поскольку число предложений бесконечно, при синтаксическом разборе имеет смысл ориентироваться на более мелкие единицы – фразовые категории (ФК). ФК – это группа, в которой имеется одна вершина, а также может быть одно или несколько зависимых от этой вершины [2], [3]. Таким образом, алгоритм автоматического анализа сводится к вычленению ФК в составе предложения и поиску связей между ними.

Для разработки модуля автоматического синтаксического анализа был использован корпус текстов, состоящий из клауз с нераспространенной синтаксической структурой из [4]. Клаузы составлены в соответствии с нормами литературного русского языка. Этот корпус, безусловно, нуждается в расширении и усложнении, но на нынешнем этапе разработки модуля синтаксического анализа он отвечает основному поставленному требованию: идентификация отдельных ФК в структуре клаузы и определение связей между ними.

На основании анализа используемого корпуса были выделены пять основных синтаксических групп: именная группа (ИГ), глагольная группа (ГГ), группа прилагательного (ПГ), предложная группа (ПрГ), инфинитивная группа (ИнфГ). Для удобства за каждой группой был закреплен порядковый номер. Каждая синтаксическая группа имеет вершину, то есть слово, от которого зависят все остальные слова в группе. Вершиной ИГ является имя существительное или личное местоимение. Вершиной ГГ – личные формы глагола. Вершиной ПГ вы-

ступает краткое прилагательное. Вершиной ПрГ является предлог. Вершиной ИнфГ – инфинитив. Связи и соотношения слов внутри групп представлены в таблице 1.

Символ «*» означает, что элементы группы могут стоять также и в обратном порядке. При этом тип связи между ними сохраняется неизменным.

Следует отметить, что глагольные группы представлены двумя уровнями. Глагольная группа второго уровня ГГ'' включает в себя, помимо глагольной группы первого уровня ГГ', другие элементы.

Таблица 1

Типы ФК, используемых в модуле синтаксического анализа

ИГ {Сущ}/{М}	ГГ' {ГГ''}	ПГ {КрПрил}	ПрГ {Пр}	ГГ'' {Глаг}	ИнфГ {Инф}
—	ИГ* управление	/ИГ управление	/ИГ управление	—	ИГ* управление
/ {ИГ} управление	ИГ/ управление /ИГ управление	{Нар}/ примыкание		{AUX} → {КрПрил})	
{Прил}/ согласование	ПрГ* примыкание			{Нар}* примыкание	
/CON- {Сущ}	ИГ/ управление /ПрГ примыкание			{Нар}/ примыкание / {Нар} примыкание	
	/ИГ управление //ПрГ) примыкание				
	/ИнфГ примыкание				

ИГ – вершиной является Имя существительное Сущ или местоимение М; ГГ'' – вершиной является финитный глагол Глаг; ГГ' – вершиной является группа ГГ''; ПГ – вершиной является краткое прилагательное КрПрил или прилагательное П; ПрГ – вершиной является предлог Пр; ИнфГ – вершиной является инфинитив Инф; AUX – вспомогательный глагол.

Настоящая таблица представляет систематизацию ФК, встречающихся в 500 тестовых предложениях. В каждой ФК действуют подчинительные связи одного из трех типов; на уровне морфологии это находит отражение в том, что при согласовании зависимое слово (Прил) принимает те же показатели рода, числа и падежа, что и вершина (Сущ или М); при примыкании наблюдается простое синтаксическое соположение вершины и неизменяемого слова-зависимого без дополнительного маркирования на морфологическом уровне, а при управлении зависимое слово (Сущ или М) стоит в определенном косвенном падеже, причем выбор падежа определяется по словарю, в характеристиках слова-вершины. Для определения падежа, в котором стоит зависимое слово при подчинительной связи, используется словарь [5]. Предполагается со временем создать свой словарь, специально приспособленный для нужд автоматического синтаксического анализа.

3 Описание алгоритма автоматического анализа синтаксической структуры предложения

Работа анализатора основывается на базе данных по ФК. В тексте ищутся только такие ФК, которые внесены в базу. Предложением (в случае с нераспространенными предложениями это клауза) называется отрезок текста между двумя показателями конца предложения – точкой / восклицательным знаком / вопросительным знаком + пробел и точкой / восклицательным знаком / вопросительным знаком.

Анализ начинается с того, что модуль морфологического анализа определяет морфологические характеристики и частеречную принадлежность анализируемого слова. Далее начинается формирование гипотез о текущей ФК. Любая ФК может быть представлена в виде: $XГ = (x; Г)$, где $XГ$ –

название ФГ, x – вершина, $YГ$ – зависимое. $YГ$ может принимать значения $YГ = 0$, где 0 – это пустое множество:

$XГ = волк$; $x = волк$; $YГ = 0$.

Или $YГ = y$; $ZГ$ где $ZГ$ – ФК, идентичная по структуре $YГ$.

Теоретически, разложение $YГ$ на составляющие может быть бесконечным: $_1[кот, _2[который пугает и ловит синицу, _3[которая часто ворует пшеницу, _4[которая _5[в тёмном чулане]_5 хранится _6[в доме, _7[который построил (Джек)]_1]_2]_3]_4]_5]_7$

На следующем шаге по базе ФК определяются типы ФК, в которые может входить анализируемая словоформа. Это решается за счет простого перебора НС, которые возможны в текущем контексте при данных морфологических характеристиках слова.

После формирования гипотезы относительно ФК анализатор переходит к поиску и анализу следующей словоформы; далее аналогичным методом ищутся вершины и зависимые (см. раздел 5). По сути, все предложение разлагается на две группы – группу подлежащего (ИГ) и группу сказуемого (ГГ), построенных по модели $XГ = (x; YГ)$. Причем, если словоформа A – существительное в именительном падеже, у него нет вершины, оно объявляется подлежащим, если словоформа B – глагол в личной форме, у него нет вершины, он объявляется сказуемым и согласуется в числе и лице с подлежащим.

Анализ идет до тех пор, пока все словоформы в предложении не будут связаны друг с другом.

4 Программная реализация МСА

Структурная схема анализатора представлена на рис. 1.

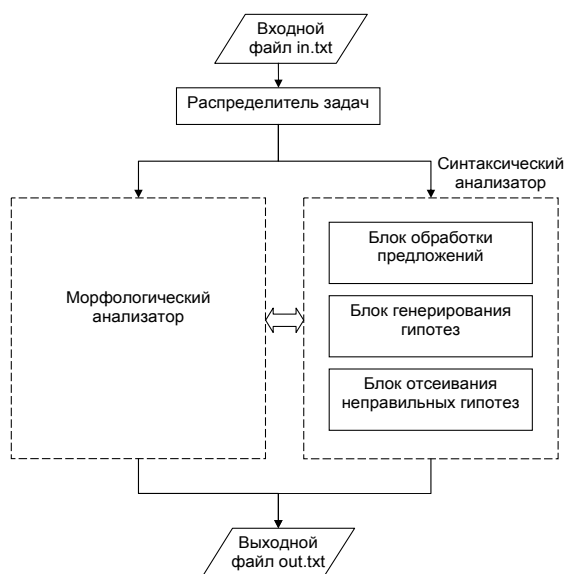


Рис. 1. Структурная схема модуля синтаксического анализа

Как было сказано выше, входные данные представляют собой список простых предложений. Рас-

пределитель задач передает эти данные в блок синтаксического анализа.

В блоке обработки предложения обрабатываются пословно. Исходная словоформа передается в блок морфологического анализа [6], в котором для нее подбираются все возможные варианты основ и соответствующие грамматические показатели. Каждое предложение считается пословно, после чего словоформа поступает на вход морфоанализатора. В результате определяются все возможные основы и соответствующие грамматические показатели. Если данной словоформе соответствует только одна основа, она поступает в процедуру построения гипотез. В зависимости от части речи и грамматических показателей выделяется соответствующая синтаксическая группа. После предварительной обработки словоформа поступает в блок генерирования гипотез. Этот блок является основным. На его вход поступает словоформа. Если это первое слово в предложении, то в соответствии с частью речи определяется синтаксическая группа.

В зависимости от частей речи группа может определяться однозначно, либо могут быть варианты. Например, если на вход поступило имя существительное, то первая группа в предложении будет именной. В этом случае запоминается порядковый номер группы и обрабатывается следующее слово. Если же первым словом в предложении является наречие, то оно может относиться как к глагольной группе, так и к группе прилагательного. В этом случае для однозначного определения группы требуется следующее слово.

Начиная со второго слова в предложении, важную роль играет не только часть речи данной словоформы, но и информация о группе или группах, которые выделены на данный момент. Поступившая на вход словоформа может, как принадлежать текущей синтаксической группе, так и выделяться в другую синтаксическую группу. В этом случае формируется дополнительная гипотеза, и рассматриваются оба варианта. В конечном счете, в предложении выделяются группа подлежащего и группа сказуемого. Важно отметить, что некоторые группы, например, глагольная, могут содержать в себе другие группы. История выделения слов в группы сохраняется в виде индекса.

Таким образом, из-за морфологической и синтаксической неоднозначности для одного предложения может быть сформировано несколько гипотез, ср. [7], [8]. Далее эти гипотезы поступают в блок отсеивания неправильных гипотез. Данный блок имеет два уровня проверки. На первом уровне проверяется согласование синтаксических групп в рамках одного предложения. Это согласование определяется, исходя из грамматических характеристик вершин групп.

Выходной файл представляет собой список предложений, каждое из которых разбито на синтаксические группы. Если предложение содержит слово, отсутствующее в словаре, то оно выводится без разбора.

5 Заключение

Для тестирования разработанного модуля были выбраны случайным образом 50 простых предложений из ГОСТ Р 50860-95. Они были представлены в виде списка и обработаны. Результаты выглядят следующим образом: 32% предложений разобрано не было. При анализе результатов были выявлены следующие ошибки:

1. Наличие в анализируемом тексте синтаксических конструкций, отличных от исходных синтаксических групп – 26%.

2. Наличие в тексте слов, отсутствующих в словаре – 6%.

3. Семантическая (морфологическая) неоднозначность слов, порождающая несколько вариантов разбора предложения. Данный вид ошибки привел к выводу неправильных гипотез, которые программа не смогла отсеять – 22%.

Для решения этих ошибок необходимо расширить базу синтаксических групп, за счет анализа большого количества текстов. Также планируется разработка модуля для автоматического пополнения словаря. И одной из важнейших задач является разработка частотного словаря, что позволит уменьшить количество гипотез для одного предложения и увеличить скорость и качество работы синтаксического анализатора.

В результате описанных в настоящей статье разработок был создан модуль синтаксического анализа текста русского языка. Синтаксический модуль позволяет производить разбор предложения. Следует заметить, что данная работа является только первым шагом при разработке полноценного синтаксического анализатора. В будущем планируется сделать более информативный вывод, представляя предложения в виде дерева. Таким образом, будут наглядно показаны синтаксические зависимости внутри предложений и указаны типы связи. Также планируется провести тестирование анализатора на произвольном тексте, взятом из художественной или научной литературы. На данном этапе разработки разбор производится только для простых предложений, так что планируется разработать и реализовать алгоритм разбора сложных предложений. Модуль синтаксического анализа может работать как самостоятельная программа, а может выступать в качестве синтаксической составляющей декодера для русской слитной речи. Поскольку словарь распознавателя содержит в себе отдельно основы и окончания, то в связи с вариативностью русского языка могут быть ошибки при подборе окончаний. Разрабатываемый синтаксический анализатор позволит устранить подобные ошибки, таким образом, повысив качество распознавания русской слитной речи.

Литература

- [1] An. Leontieva, "The Module of Morphophonetic Word Processing for Composing a Vocabulary for Russian Continuous Speech Recognizer". Scientific-theoretical journal «Artificial intelligence», Donetsk, Ukraine, Vol. 3, 2007, pp. 319–327.
- [2] Фитиалов С.Я. Об эквивалентности грамматик НС и грамматик зависимости // Проблемы структурной лингвистики, 1967. М.: Наука, 1967. 145 с. С. 16–43.
- [3] Тестелец Я.Г. Введение в общий синтаксис. М.: Российский государственный гуманитарный университет, 2001. 800 с.
- [4] Передача речи по трактам радиотелефонной связи: Требования к разборчивости речи и методы артикуляционных изменений. ГОСТ 16600 – 72. М.: Издательство стандартов, 1973. 90 с.
- [5] Большой толковый словарь русского языка / Под ред. Д.Н. Ушакова. М.: Альта-принт, 2005. 1239 с.
- [6] Kagirow I.A., Leontyeva An.B. Grammar-Based Speech- and Word-splitting // Proceedings of 3rd Language & Technology Conference. October 5–7, Poznac, Poland. Poznac: Fundacja Uniwersytetu im. A. Mickiewicza, 2007. 578 s. P. 413–417.
- [7] Попов Э.В. Общение с ЭВМ на естественном языке. М.: Едиториал УРСС, 2004. 260 с.
- [8] Mel'nik I. A. Dependency syntax: Theory and practice. Albany, NY: SUNY Press, 1988. 428 p.

Automatic syntactic analysis of russian texts

Anastasia Leontyeva, Ildar Kagirow

The paper presents a syntactic approach for analysis of Russian texts. The approach was used to develop the module of the syntactic analysis, which could be used as a single program or as the auxiliary module for speech recognition. At the first stage of development the module analyses only the sentences with a simple structure. During the analysis the syntactic groups are singled out and dependencies between the constituents are determined.

Подход к построению сервис-ориентированной электронной библиотеки в среде Semantic Web на основе репланирования

© А.В. Новицкий

Институт программных систем НАН Украины, г. Киев
ale-nov@yandex.ru

Аннотация

В работе рассмотрены вопросы построения сервис-ориентированной электронной библиотеки. Описан формализованный подход описания сервисов электронной библиотеки на основе дескриптивной логики. Особое внимание уделено решению проблемы планирования веб-сервисов в среде с нечеткими целями на основе знаний о предметной области.

1 Введение

Усовершенствование технологии дали новый толчок к развитию электронных библиотек. Основным требованием к будущим электронным библиотекам, есть постоянное расширение функциональности.

Использование сервис-ориентированной архитектуры является важным элементом достижения интероперабельности. В такой среде сложное программное обеспечение, такое как электронные библиотеки, может быть спроектировано на основе имеющихся в распоряжении сервисов, которые доступны не только локально, но и через Интернет. Причем доступ к сервисам будет происходить динамически.

Для того чтобы обеспечить свободный доступ служб к библиотеке она должна быть реализована на открытой архитектуре. Т.е. взаимодействие сервисов должно быть описано с помощью стандартизированных правил и сред. Что даст возможность свободно вводить новый сервисный элемент, не перестраивая систему заново.

Под *сервисом* мы понимаем некоторую услугу цель, которой удовлетворить запросы пользователя. Сервисы реализуются посредством веб-сервисов. *Веб-сервис* это программное приложение. Основное отличие между сервисом и веб-сервисом состоит в том, что сервис одновременно можно реализовать разными программными веб-сервисами. В свою

очередь веб-сервисы могут между собой объединяться, создавая новый веб-сервис, этот процесс называют композиция веб-сервисов.

Существует много способов композиции веб-сервисов [2]. Для того чтобы произвести композицию сервисов или обратиться к нему в рамках технологии Semantic Web сервис необходимо описать.

Консорциум W3C рекомендует использовать язык Консорциум W3C стандартизует язык OWL-S, который предназначен для описания онтологии сервисов, реализуемых через веб-сервисы. OWL-S включает три подонтологии, которые называются профилем, моделью процесса и основой. Онтология профиля сервиса описывает типы ввода и вывода, определяет предыдущие условия, которые сокращенно записывается как *IOPE* (Inputs, Outputs, Preconditions and Effects). Модель процесса описывает, как сервис работает. Онтология основания сервиса определяет информацию, необходимую для сервисного обслуживания [9].

2 Веб-сервисы и планирование их композиции

2.1 Обнаружение и композиция веб-сервисов

Взаимодействия сервисов – это динамический процесс, для проведения анализа которого необходима формализация. В основу языка OWL-S положен математический аппарат дескриптивной логики. Формализация сервисов отличается от обычного описания онтологий. Это связано с тем, что при описании сервисов мы имеем дело с динамической средой. Дескриптивная логика (DL) предназначена для того, чтобы описывать логику действий и поведение сервисов для достижения цели в зависимости от входных и выходных условий.

Дескриптивная логика играет значительную роль в Semantic Web, особенно с того времени, когда язык OWL был рекомендован W3C, которым мы можем воспользоваться для семантического описания страниц.

Под запросом r мы будем понимать, некоторые предоставленные в заранее оговоренном формате параметры, которые мы передаем веб-сервису. Обозначим веб-сервис через S_i^{in} где i - уникальный

номер сервиса, $in(out)$ – набор входных (выходных) параметров сервиса.

Когда с помощью одного веб-сервиса невозможно достичь исполнения запроса r необходимо создать композицию нескольких веб-сервисов $\{S_1, S_2, S_3, \dots, S_n\}$ при этом должно выполняться соотношение $(r_{in} \cup S_1^{out} \cup \dots \cup S_n^{out}) \supseteq r_{out}$. Это так называемая проблема композиции веб-сервисов (Web Service Composition - WSC).

Аннотации, описывающие семантику сервиса, могут упростить задачу композиции веб-сервисов.

Язык OWL-S развивает OWL для семантического описания веб-сервисов и включает в себя разные аспекты веб-сервисов, в том числе и функциональные. Чтобы описать функциональные возможности, сервисы рассматриваются в качестве процессов, которые имеют начальные условия и действия. Однако для представления динамического поведения таких процессов у статического языка OWL нет достаточных возможностей.

В работе мы будем рассматривать сервисы как действия, которые имеют начальные и конечные условия. Эти условия выражаются с помощью дескриптивной логики. Дескриптивная логика оперирует базой знаний, которая включает два компонента Tbox и Abox. Tbox определяет терминологию, т.е. словарь прикладной области, а Abox содержит утверждения в терминологии данного словаря [3]. Для композиции веб-сервисов окружающее состояние мира частично описывается в Abox.

Единичный законченный веб-сервис который может удовлетворять запрос назовем атомным сервисом.

Мы остановимся на двух основных проблемах композиции сервисов: выполнимость (executability) и проектирование (projection)[8].

Обе задачи уместны для исследования сервисов. Очевидно, что нужно выбирать сервис, который гарантировано, будет выполняться при данной, возможно точно неизвестной ситуации. Кроме того, выполнение сервиса подразумевает достижение какой-то цели, поэтому следует использовать именно те сервисы, которые эту цель могут достичь – следовательно, решать проблему обнаружения сервиса.

2.2 Дескриптивная логика

Основными понятиями дескриптивной логики являются атомные понятия (atomic concepts) и атомные роли (atomic roles). Сложные описания могут быть индуктивно построены из них с помощью конструкторов понятий (concept constructors). В дальнейшем мы будем использовать буквы **A**, **B** для обозначения атомных понятий и **R** для обозначения атомных ролей и буквы **C**, **D** для описания произвольных понятий. Более детальную информацию о дескриптивной логике можно найти в книге [3]. Известно, что в полной мере язык OWL-DL описывается дескриптивной логикой, которая имеет диалект SHOIN(D).

2.3 Семантическое описание сервисов

Профиль сервиса (ServiceProfile) предназначен для объявления и обнаружения сервиса. Модель процесса (ServiceModel) дает детализированное описание операций сервиса. Основание (ServiceGrounding) описывает процесс взаимодействия с сервисом через сообщения.

Вообще говоря, ServiceProfile необходим для обнаружения сервиса, а ServiceModel и ServiceGrounding вместе дают информацию о том, как использовать найденный сервис.

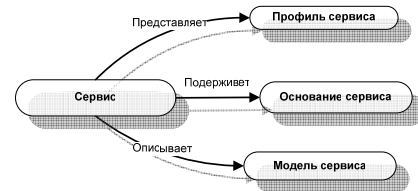


Рис. 1. Структура сервиса

Основание сервиса определяет детали доступа к сервису. Это главным образом протокол и формат сообщения, порядок поступления, транспортировка и адресация [9].

2.4 Формальное определение сервисов

Введем формальное описание веб-сервисов. Для упрощения будем рассматривать так называемые основные сервисы (ground services) т.е. сервисы в которых все параметры индивидуальные имена.

Определение. Пусть T – нециклический Tbox. Атомным сервисом для нециклического Tbox T будем называть выражение $S = (pre, sname, inputs, outputs, post, pref)$, где pre – ограниченное множество утверждений в Abox, которые образуют предыдущие статические (и динамические) условия, $sname$ – идентификационное имя автономного сервиса, $inputs$ – область определения класса входных параметров сервиса, которая формируется в виде $A(a)$ или $r(a, b)$, где A – примитивное понятие в T , r – роли и a, b атомные понятия.

$outputs$ – область определения класса исходных параметров сервиса, которая формируется в таком же виде, как и $inputs$,

$post$ – ограниченное множество пост-условий в форме P/E , где P – утверждение в Abox, а E – примитивное понятие со ссылкой на Tbox T т.е. $A(a)$, $\neg A(a)$, $r(a, b)$ или $\neg r(a, b)$ [7].

Формальная семантика сервисов может быть определена на основе перехода к интерпретации. Сервис S может трансформировать интерпретацию I к I' то есть $I \Rightarrow_S^{T, A} I'$ [8].

2.5 Отношение между веб-сервисами

При композиции веб-сервисов естественно стоит вопрос об установлении отношения между ними. Введем следующие виды отношений между сервисами.

Пусть T – нециклический Tbox, A – Abox, S_i и S_j – сервисы, которые являются подсервисами сложного сервиса S . Причем S_i предоставляет отличный от S_j тип сервиса. Отношение R между подсервисами S_i и S_j могут быть определены следующим образом:

Отношение независимости: $S_i + S_j = S_j + S_i$ – это означает, что каждый из сервисов независим один от другого и порядок выполнения этих сервисов не будет влиять на конечный сервис. Отношение идентичности: $S_i = S_j$ – это значит, что сервисы предоставляют один и тот же сервис тем не менее имеют разные атрибуты. Отношение условной идентичности: $S_i \cong S_j$ – это значит, что сервис S_i предоставляет те же самые функции что и S_j в некоторых ситуациях. Отношение замены: $S_i \text{ р } S_j$ – это значит, что сервис S_i можно заменить на сервис S_j в любом случае. Отношение условной замены: $S_i \% S_j$ – это значит, что сервис S_i можно заменить на сервис S_j в некоторых ситуациях. Отношение перекрытия: $S_i \otimes S_j$ – это значит, что существует общая часть в этих сервисных обслуживаниях. При композиции отношений перекрытия общая часть должна быть исключена. Отношение последовательности: $S_i \rightarrow S_j$ – это значит, что сервис S_i должен полностью закончить свою работу перед стартом сервиса S_j .

2.6 Базовые сервисы

Для построения электронной библиотеки на основе сервис-ориентированного подхода необходимо выделить сервисы, на основе которых будет выполняться композиция. Этот набор сервисов обусловлен анализом работы конечного пользователя (на примере научного сотрудника) а также на основе анализа требований к современным электронным библиотекам и принято во внимание ПО Eprints. При выделении базовых сервисов мы руководствовались тем, что каждый базовый сервис отвечает определенному уровню абстракции структуры библиотеки. Сервисы возможна разделить на две группы, внутренние сервисы которые не предназначены для конечных пользователей (имеют программные интерфейсы) и сервисы которые имеют веб-интерфейсы конечного использования. Приведем список последних:

Сервис преобразования форматов. Предназначенный для преобразования документов одного

формата в другие. Этот сервис является необходимым, поскольку он обеспечит работу библиотеки с разными форматами данных.

Сервис построения графа цитирования. Выполняет функции построения графа цитирования, где узлами графа будут объекты или документы, а связи между узлами являются собственно ссылками. При построении семантического графа узлами могут выступать, как ресурс или часть ресурса, а также элементы базы знаний, а связи между узлами выступают в качестве отношений между объектами.

Сервис конвертации метаданных. Отвечает за конвертацию одних схем метаданных в другие.

Сервис управления ресурсами. Обеспечит управления авторским правом, политиками и дубликатами объектов электронной библиотеки.

Сервис базового поиска. Обеспечивает поиск по базовым наборам метаданных.

Сервис расширенного поиска. Обеспечивает поиск по более широкому набору полей с возможностью логической комбинации.

Сервис навигации и классификации по определенному типу метаданных

Сервис управление доступом. Обеспечивает авторизацию пользователей, а также восстановления и установления паролей доступа.

Сервис сохранения результатов поиска. Предназначен для пользователей, которые прошли авторизацию, причем возможно выборочное сохранения результатов.

В последнее время в связи с развитием среды Semantic Web мы вводим *сервис аннотирования информации*. Сервис аннотирования представит еще один способ активного взаимодействия пользователя с электронной библиотекой. Объемы информационных ресурсов в электронной библиотеке очень значительны и использование аннотаций значительно упростит поиск информации.

Формальная модель, которая предложена в [12], выделяет два подхода к пониманию аннотаций: аннотации как метаданные и аннотации или как контент.

В первом случае мы имеем дело с разнообразными схемами метаданных (Dublin Core, MARC и др.) которые используются для описания информационных ресурсов. Эти аннотации в первую очередь направлены на пользователя.

Во втором случае аннотации как предоставления контента предназначены для автоматизированной машинной обработки. Эти аннотации предоставляют семантику документа. Семантическая аннотация – аннотация написана формальным языком с хорошо определенной семантикой и основана на онтологиях.

2.7 Обобщенная формальная модель аннотации

Цифровые объекты DO могут быть разными по содержанию: текст, аудио, картинки, также разнообразные форматы представление этих данных. Модель, которая будет представлять цифровой объект должна отображать фактическое содержание данно-

го объекта. Среди множества разработанных моделей, мы используем модель, которая предложена в [12, 13] с некоторыми изменениями которые касаются логики выделения потоков, исключили управления идентификации создателей аннотации и функции *hsm* что на наш взгляд делает модель эффективней.

В каждый момент времени k мы имеем некоторое множество цифровых объектов $DO(k)$ при аннотировании документа, мы получаем некоторое новое множество документов в момент времени $DO(k+1)$. Далее если переменная будет явно зависеть от k то она будет указываться, иначе она опускается.

Аннотация $a \in A(k)$ это кортеж:

$$a = \left(h_a \in H(k), \right. \\ \left. A_a \subseteq SN(k) \times LT \times ST(k) \times \right. \\ \left. \times SM(k-1) \times H(k-1) \right)$$

где h_a - уникальный собственный идентификатор аннотации a то есть $h(h_a) = a$;

A_a множество n -арных отношений аннотации a определяется как произведение множеств SN , LT , ST , SM и H ; SN - определяет множество типов аннотаций; ST - множество сегментов; SM - определяет множество потоков; H - множество уникальных идентификаторов

2.8 Описание сервиса аннотирования с помощью дескриптивной логики

Пусть S_0 сервис аннотирование веб-ресурсов. В рамках дескриптивной логики он примет следующий вид:

S_0 :

$pre_1 \leq 1 hasCount.link$

$sname_1 = AnnotationProcess$

$input_1 = \{sn, sm_i, st_{sm_i}, LT\}$

$outputs_1 = \{LT(a, do)\}$

Ациклический Tbox T_j для этих сервисов определяется следующим образом:

$link \equiv URI \sqcup DOI \sqcup OpenURL \sqcup PURL$

$sn \equiv text \sqcup image \sqcup audio \sqcup video$

$sm \equiv \exists number.T$

Используя дескриптивную логику возможно описать и другие сервисы.

3 Композиция веб-сервисов

Для крупномасштабных задач WSC, какими является электронная библиотека, алгоритмы полного перебора не являются эффективными, в данном случае более пригодными являются алгоритмы автоматического поиска подоптимальных решений

выборки сервисов на основе знаний о предметной области.

Это вызвано тем, что в процессе функционирования сервис-ориентированой электронной библиотеки среда в котором функционируют веб-сервисы постоянно изменяется. Изменение среды порождается двумя основными факторами:

во-первых, проблемы, которые порождают распределенные системы: временные задержки и ненадежность транспортного протокола; нехватка памяти общего использования между частями распределенной системы; проблемы отказа доступа и параллельных запросов, а также проблемы связаны с программной несовместимостью вследствие обновления части распределенной системы;

во-вторых, человеческий фактор, в процессе обслуживания пользователя, могут измениться требования пользователя, а, следовательно, это определенным образом будет влиять на результат. Рядовые пользователи целевой аудитории не имеют четкого представления об архитектуре библиотеки, следовательно, не могут заранее четко определиться с целями которые должны удовлетворять сервисы.

Мы предлагаем рассматривать композицию на основе целе-ориентированой парадигмы, то есть исходя из первоначальных условий и имеющегося множества сервисов, осуществить композицию. Причем, поскольку веб-сервисы размещены в семантической среде, то выбирать такие планы композиции, которые могут быть полезными для конечного пользователя. В отличие от классической постановки задачи, от спецификации цели к поиску сервисов, которые смогут эту цель достичь. Мы предлагаем выходить из того предположения, что существует множество сервисов, которые могут достигать некоторых, заранее неизвестных целей, которые может поставить перед собой пользователь.

В такой сложной системе цели, которые появляются перед композицией могут изменяться. Мы также выходим с того что знания об окружающем мир является не полными, а следовательно, и цели являются неполными. Для таких конфликтных целей Horst Rittel и Melvin Webber в [14], [15] ввели понятие викид (wicked) задачи. Поэтому классические методы планирования могут оказаться неэффективными. В среде, где цели не являются четко определенными и являются динамичными, процесс планирования принимает другое смысловое значение, а именно, план выполнения действий не является однозначным алгоритмом достижения цели,. Функцию планов можно сформулировать следующим образом [3]: проверка ресурсов; начинают координационные процессы или помогают упростить координационные процессы в начале; устанавливают ответственность и идентификацию; трековый прогресс и, что более важно, генерация намерений; узнают и управляют рисками; поддерживают импровизацию, и формализуют представление человека (пользователя) о проблеме, которую план пытаются решить.

Такая комбинация функциональных свойств приводит к тому, что планы необходимо дополнительно исследовать и проводить перепланирование. Т.е. изменять планы во время выполнения этих планов.

Рассмотрим самый простой случай в электронной библиотеке. Пользователю необходимо найти некоторую информацию, система генерирует возможные варианты поиска информации и пользователь решил, что лучше для этого воспользоваться базовым (простым) поиском, однако система не может однозначно определить цель (цель неточна) а следовательно и возможен план развития, поскольку в процессе выполнения плана пользователь может изменить цель (например, некоторая дополнительная обработка результатов поиска), а следовательно, нужно репланировать, то есть изменять первоначальный план. При этом пользователь появляется перед викид-проблемой, поскольку появляется дилемма выбора между важностью возможно найденной информации и времени затраченному на ее поиск.

Аннотирование информации значительно упрощает ее машинную обработку с целью систематизации и поиска.

Решение описанных проблем возможно только в среде Semantic Web. Что требует четкой формализации всех сервисов и построение модели процессов. В таком случае, возможно, получить гибкую, легко расширяемую систему с возможностью автоматически решать проблему WSC. Следующим этапом будет практическая реализация сервисов на основе их моделей.

Литература

- [1] Yves Petinot, C. Lee Giles, Vivek Bhatnagar, Pradeep B. Teregowda, Hui Han, Isaac Council. A Service-Oriented Architecture for Digital Libraries
- [2] J. Rao, X. Su. A Survey of Automated Web Service Composition Methods. In Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition, SWSWPC 2004, San Diego, California, USA, July 6th, 2004.
- [3] The description logic handbook: theory, implementation, and applications Edited by Franz Baader Deborah L. McGuinness Daniele Nardi Peter F. Patel-Schneider
- [4] Lirong Qiu, Fen Lin, Changlin Wan, Zhongzhi Shi. Semantic Web Services Composition Using AI planning of Description Logics
- [5] Description Logics: Foundations for Class-based Knowledge Representation Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini Dipartimento di Informatica e Sistemistica Universit'a di Roma "La Sapienza" Via Salaria 113, I-00198 Roma, Italy
- [6] Description Logics: Foundations for Class-based Knowledge Representation Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini Dipartimento di Informatica e Sistemistica Universit'a di Roma "La Sapienza" Via Salaria 113, I-00198 Roma, Italy

- [7] Yingjie Li1, Xueli Yu1, Lili Geng1, Li Wang. Research on Reasoning of the Dynamic Semantic Web Services Composition. Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on Volume , Issue , 18-22 Dec. 2006 Page(s):435 – 441
- [8] Franz Baader, Maja Milicic, Carsten Lutz, Ulrike Sattler, Frank Wolter. Integrating Description Logics and Action Formalisms for Reasoning about Web Services. LTCS-Report 05-02, Chair for Automata Theory, Institute for Theoretical Computer Science, Dresden University of Technology, Germany, 2005
- [9] OWL-S: Semantic Markup for Web Services <http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/>
- [10] Ragone, A., Di Noia, T., Di Sciascio, E., Donini, F.M., Colucci, S. Fully automated Web services orchestration in a resource retrieval scenario. Web Services, 2005. ICWS 2005. Proceedings. 2005 IEEE International Conference on Publication Date: 11-15 July 2005
- [11] WebServices Description Language (WSDL)1.1 <http://www.w3.org/TR/2001/NOTE-wsdl-20010315>
- [12] GONCALVES, M.A., FOX, E.A., WATSON, L.T., AND KIPP, N.A. 2004a. Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. ACM Trans. Inform. Syst. 22, 2 (Apr.), 270–312.
- [13] M. Agosti, N. Ferro. A Formal Model of Annotations of Digital Content. ACM Transactions on Information Systems (TOIS), 2008, 26(1):3:1–3:57
- [14] Gary Klein, "Flexexecution as a Paradigm for Replanning, Part 1," IEEE Intelligent Systems, vol. 22, no. 5, pp. 79–83, Sept/Oct, 2007
- [15] Gary Klein "Flexexecution, Part 2: Understanding and Supporting Flexible Execution," IEEE Intelligent Systems, vol. 22, no. 6 Nov./Dec., pp. 108–112.

The approach to construction service-oriented to digital library in the environment of Semantic Web on a basis replanning

A.V. Novytskyi

In work it is considered constructions service-oriented of electronic library. The formalized approach of the description of services of electronic library on the basis of descriptive logic is described. The special attention is given to the decision of a problem of planning of web services in the environment with the indistinct goals on the basis of knowledge of a subject domain.

Ассоциативно-семантическая неоднородность восприятия информации студентами рискованных профессий и методологические подходы к ее формализации*

© Г.В. Талалаева

Институт экологии растений и животных УрО РАН
gvthal@ipae.uran.ru

Аннотация

Статья посвящена анализу закономерностей передачи информации в системе «человек – компьютер – электронные библиотеки». Обращено внимание на роль эмоциональной и ассоциативной составляющей познавательного процесса в интерпретации поступающей информации. Проблема исследована на примере феномена радиофобии и реакции студентов-физиков на слово «здоровье». Делается вывод о целесообразности создания нового поколения интерфейсов, которые учитывают актуальное эмоциональное состояние пользователей.

1 Постановка проблемы

При создании электронных библиотек в любой области знаний (науки, образования, культуры, здравоохранения, социального менеджмента) проблема интерфейса «человек и машина» является актуальной. К этой проблеме было обращено внимание участников всех конференций «Электронные библиотеки», начиная с 1999 г. [6, 7, 9]. Для пользователей Интернет семантическая интерпретация базовых понятий радиоэкологии приобретает особое социальное значение, так как их разное прочтение является одним из механизмов формирования радиофобии. В настоящее время изучение феномена радиофобии ведется многопланово. От успешности этих исследований во многом зависит развитие ядерной энергетики в XXI веке [2]. Установлено, что восприятие радиоэкологической информации определяется не только семантикой материала, но и психологическими качествами пользователей этой информации. В доступной нам литературе мы нашли два варианта объяснения механизма возникно-

вления семантической неоднородности восприятия, казалось бы, инвариантных понятий. Один из них мы называем когнитивным (познавательным), второй – коммуникативным. Первый заключается в дефиците специальных знаний у обывателей; второй – в неумении профессионалов психологически грамотно преподнести научную информацию широкой аудитории.

На наш взгляд, есть еще третья причина возникновения радиофобий как примера сбоя передачи информации в многозвеновой системе коммуникации. Эта причина – ассоциативно-семантическая неоднородность восприятия информации людьми. В обычной жизни передача социально значимой информации, так же как и ассоциативный поиск нужной информации, сопровождается активацией мотивационно-эмоциональной компоненты коммуникации людей. Однако, в современных компьютерных программах мотивационный компонент поиска желаемой пользователю информации, как правило, не учитывается. В связи со сказанным мы полагаем необходимым формализовать алгоритмы восприятия информации в трехзвеновой системе: «эмоциональное состояние человека – его доминирующая актуальная мотивация (текущая потребность) – семантическая интерпретация заданного понятия».

2 Цель исследования

Цель исследования – разработать методологию оценки взаимосвязи текущего ассоциативно-эмоционального состояния респондента с вариантами интерпретации им одной и той же семантической информации. Было сформулировано две гипотезы: первая – отвергающая роль мотивационной сферы для эффективности повседневной работы пользователей; вторая – принимающая ее. Мы полагали, что если вторая гипотеза верна, то необходимо будет делать вывод о целесообразности усиления семантических и лингвистических принципов кодирования информации в электронных библиотеках дополнительным – ассоциативно-мотивационным. По сути, это предложение было бы ничем иным, как экстраполяцией систем с биологической обратной

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

связью (БОС) из сферы прикладной медицины [1] в сферу библиотечного дела (формирования поисковых и аналитических систем баз данных в масштабах электронных библиотек и глобальной сети Интернет).

3 Материал, метод и методология исследования

Исследование выполнено на примере студентов кафедры экспериментальной физики физикотехнического факультета ГОУ ВПО УГТУ-УПИ первого и пятого курса обучения. Для структурно-функционального анализа стилей восприятия информации был разработан специальный методологический прием, основой которого стала пирамида иерархии жизненных потребностей (мотиваций) А. Маслоу [5]. Тестирующим инструментом было выбрано слово «здоровье». Наши предыдущие исследования [3] показали, что архитектура актуальных потребностей уральцев существенно отличается от той, которая была обнаружена А. Маслоу полвека назад по результатам обследования представителей американского социума. Поэтому в авторскую методику тестирования ассоциативных связей слова «здоровье» с актуальной потребностью пользователя мы включили восемь потребностей, наиболее актуальных для современных россиян. Это потребность в удовлетворении базовых биологических нужд (в еде, одежде, жилище); безопасности; любви; самоуважении и уважении со стороны других людей; самодостаточности и осмыслении бытия; красоте и эстетическом наслаждении; добре; творческой самореализации. Тест осуществлялся в четырех повторах в режиме виртуального моделирования. Испытуемым было предложено последовательно представить себя в негативном, нейтральном, позитивном эмоциональном состоянии и в состоянии прослушивания радиопередачи. В каждом из названных состояний было предложено ассоциировать слово «здоровье» с актуальной эмоциональной потребностью (одной из восьми выше перечисленных).

4 Полученные результаты

Анализ результатов тестирования показал, что распределение ответов респондентов в выше перечисленных состояниях не одинаковы. Кроме того, внутри каждого из состояний спектры заявленных потребностей первокурсников пятикурсников не совпадали. Первокурсники в положительном состоянии ассоциировали слово «здоровье» с потребностью в любви, пятикурсники – с потребностью в самоактуализации и в удовлетворении этических потребностей. В нейтральном состоянии слово «здоровье» у пятикурсников ассоциировалось с темами эстетики (стремлением к искусству, красоте), у первокурсников – этики (потребностью в добре, моральном поведении). В негативном состоянии у обеих групп респондентов слово «здоровье» ассо-

циировалось с темой безопасности. При ситуации виртуального моделирования прослушивания радиопередачи первокурсники восприняли слово «здоровье» как аналог биологической проблематики пятикурсники – как эквивалент философских тем. Первые ассоциируют слово «здоровье» в данном состоянии с удовлетворением потребностей в безопасности, обеспечением физиологических нужд, вторые – с самоуважением и осмыслением бытия. Полипараметрический анализ показал, что внутри каждой возрастной группы в каждом из состояний помимо выше описанных моделей восприятия информации (доминирующих) присутствовали и дополнительные.

5 Обсуждение

Представленные факты свидетельствуют о ассоциативно-семантической неоднородности и вариативности процесса восприятия информации. Кроме того, проведенные исследования убедили нас в том, что алгоритм формирования ассоциативно-семантических взаимосвязей не случаен. Он не носит хаотического характера, не является абсолютно индивидуальной характеристикой, но, наоборот, подчиняется определенными закономерностями. Следовательно, подлежит формализации, может быть представлен в виде формализованной схемы и быть основой для разработки усовершенствованного интерфейса системы «компьютер – пользователь – электронные библиотеки». В случае принятия данной идеи к технической разработке, модифицированный интерфейс будет способен дифференцировать эмоциональные состояния пользователя и адекватно его текущему состоянию задавать направление информационного поиска.

Мы согласны, что сегодня психоинформационные феномены, представленные в данной статье, могут быть интерпретированы двояко: либо как новая фундаментальная проблема, находящаяся на прорывном направлении науки, либо как частный случай ранее известных науке закономерностей. Выбор позиции может варьировать в зависимости от методологической позиции исследователя, конкретных прикладных задач его деятельности, глобального вектора развития информационных технологий, приоритетного направления работ в конкретном научном учреждении и т.д. В любом случае признание важности психо-эмоциональных аспектов поиска, восприятия, кодирования и декодирования информации позволяет нам рассматривать изложенные в статье феномены как ключевые алгоритмы. Понимая при этом под термином «алгоритм» формальную систему, в которой «на каждом шаге процесса построения объектов в общем случае применимо не одно правило, а несколько, и, следовательно, возможен выбор следующего шага из нескольких возможных» [8]. Признание важности ассоциативно-семантического подхода к кодированию и декодированию информации является стимулом развития нового поколения современной компьютерной техники

и технологий искусственного интеллекта. В противном случае (в случае отрицания значимости психоэмоционального состояния человека в организации и эффективности информационного поиска) приведенные в статье факты следует рассматривать рангом ниже: не как алгоритмы, а как частный случай решения общих, уже известных задач. Например, задач создания интеллектуального человеко-машинного интерфейса и планирования целесообразного поведения автономных роботов на основе фреймообразных структур, успешно применяемых для обеспечения функционирования безэкипажных мобильных боевых комплексов, беспилотных летательных и космических аппаратов, автономных подводных и надводных плавсредств, бытовых робототехнических систем [4]. Необходимость создания модифицированного интерфейса, позволяющего более результативно осуществлять циркуляцию информации в системе «компьютер – пользователь – электронные библиотеки», абсолютно четко осознана научным сообществом, представители которого работают в области психофизики, социальной и экологической медицины. Доказательство тому – доклады участников Ежегодной Всероссийской научной школы-семинара «Методы компьютерной диагностики в биологии и медицине – 2008»¹, освещающие успешные попытки создания нового типа интерфейса «человек – машина», в котором, в том числе учитывается эмоциональное состояние пользователя. К перечню таких докладов, на наш взгляд, можно отнести презентации аппаратуры, выполненные сотрудниками кафедры медицинской физики СГУ (зав. каф., проф. А.В. Скрипаль), специалистами Института биофизики клетки РАН и Московского государственного технического университета им. Н.Э. Баумана (докладчик д.б.н., в.н.с. А.И. Федоров), коллегами из Дрезденского технологического университета и Московского государственного педагогического университета (докладчик д-р У. Кэмпф).

Но выше перечисленные технологии создания интерфейсов не адаптированы пока к решению задач социального менеджмента рискованных территорий, информационному поиску в базах данных электронных библиотек и глобальной сети Интернет. Мы надеемся, что такие работы будут выполнены, и осуществляться они в конструктивном диалоге специалистов разных областей научного знания (гуманитарного, технического и естественнонаучного).

Литература

- [1] Биоуправление-4: Теория и практика. – Новосибирск : ЦЭРИС, 2002. – 350 с.
- [2] Гагаринский А.Ю. Ядерная навигация в турбулентном мире // Ядерное общество. – 2003, №5–6. – С. 28–31.
- [3] Грачева А.В., Талалаева Г.В. Радиофобия: психологические предпосылки для искажения информации о радиации // Молодежь и будущее атомной промышленности России / Сб. докл. конф. студентов вузов Урала. – Снежинск, 2004. С. 84–85.

- [4] Макаров И.М. Искусственный интеллект и интеллектуальные системы управления / И.М. Макаров, В.М. Лохин, С.В. Манько, М.П. Романов. – М. : Наука, 2006. – С. 295–329.
- [5] Маслоу А. Мотивация и личность. – СПб. : Питер, 2003. – 352 с.
- [6] Одинцов И.О., Холчева А.В. Применение методов искусственного интеллекта в задаче поиска информации в Интернете // RCDL 1999. – <http://www.rcdl.ru/ru/1999/proceedings.html>.
- [7] Романова Е.В., Романов М.В., Некрестьянов И.С. Использование интеллектуальных сетевых роботов для построения тематических коллекций // RCDL 1999. – <http://www.rcdl.ru/ru/1999/proceedings.html>.
- [8] Фридланд А.Я. Информатика: процессы, системы, ресурсы. – М. : БИНОМ. Лаборатория знаний, 2003. – С. 163.
- [9] Maurita Peterson Holland, Kari R. Smith. Developmenta Digital Collective and Its Application to Cultural Heritage // RCDL 1999. P. 202–208. – <http://www.rcdl.ru/ru/1999/proceedings.html>.

Associative and semantic heterogeneity on student informational perception and methodological approaches to its formalization

G.V. Talalaeva

This article is dedicated to phenomenon of come dawn the information on system “people – computer – web-libraries”. The role of the emotional and associative components cognitive process was studied. The problem investigated on the case of radiophobia and perception the word “health” by the physicist students. The conclusion of necessary to create the new generation of computer that can differ different psychological status of its users was made.

* Исследования поддержаны грантом РГНФ № 08-06-00030а «Коммуникативная активность как фактор демографического поведения молодежи в условиях изменяющейся России».

¹ Саратовский госуниверситет им. Н.Г. Чернышевского, 3–5 июля 2008 г.; <http://medphys.sgu.ru/>

Результаты мониторинга и основные рекомендации по вопросам развития электронных библиотек вузов

© Э.А. Котляр

Д.Ю. Гужеля

Н.А. Полихина

«Центр инновационных технологий в образовании и науке»
polihinaNA@mail.ru

Аннотация

В докладе представлены результаты и основные выводы по комплексному исследованию электронных библиотек инновационных вузов РФ и анализу международного опыта в создании и функционировании университетских медиатек. Освещаются выявленные проблемы и потребности различных групп пользователей в организации структуры электронных библиотек и работе с ними. Сформулированы рекомендации по созданию и совершенствованию работы электронных библиотек вузов РФ.

В ноябре 2007 года «Центр инновационных технологий в образовании и науке» провел комплексное исследование электронных библиотек инновационных вузов являющихся победителями по результатам конкурсного отбора образовательных учреждений высшего профессионального образования, внедряющих инновационные образовательные программы в 2006–2007 годах. А также был проанализирован опыт в создании и функционировании электронных библиотек ведущих мировых вузов [1]. Исследование проводилось в рамках государственного контракта «Модернизация библиотек ВУЗов в целях комплексного информационного обеспечения учебного процесса».

Методология анализа международного опыта в создании и функционировании университетских медиатек: Метод - Desk research. Проанализированы данные, полученные с сайтов медиатек ведущих вузов мира (по данным Института высшего образования шанхайского университета Цзяо Тун) и других информационных источников. Выборка квотная, построенная при помощи метода экспертных оценок. Размер выборки: медиатеки 15 лучших вузов различных регионов мира.

Методология исследования электронных библиотек инновационных вузов РФ: Выборка – инновационные вузы РФ. Целевая совокупность – сту-

денты, профессорско-преподавательский состав, представители администрации библиотек. География – Москва, Санкт-Петербург, Челябинская область, Новосибирская область, Пермская область, Ставропольский край.

Ключевые этапы Мониторинга:

- Экспертный опрос заведующих библиотеками, представителей администрации вузов и профессорско-преподавательского состава. Метод – глубинное интервью.
- Опрос преподавателей, студентов и представителей администрации библиотек вузов методом личного стандартизованного интервью. Метод – формализованное интервью по анкете. Объем выборки: 424 интервью.
- Официальный запрос в инновационные вузы относительно функционирования электронных библиотек по специально разработанной форме.

Современное состояние электронных библиотек инновационных вузов

Проведенный в рамках проекта Мониторинг и анализ потребностей в создании и модернизации электронных библиотек (медиатек) для информационного обеспечения ВУЗов Российской Федерации позволил нам проанализировать текущую ситуацию в медиатеках инновационных вузов. Сочетание количественных и качественных методов исследования, а также анализ статистики вузовских медиатек позволили получить достаточно полную картину реальной ситуации в медиатеках инновационных вузов.

На сегодняшний день медиатеки инновационных вузов страны нельзя назвать полностью сформированными, очень многие ресурсы и сервисы недоступны на сегодняшний день.

- Почти половина медиатек не имеет полнотекстовых изданий, сайта, обеспечивающего удаленный доступ к ресурсам.
- Половина медиатек не содержит ссылок на ресурсы Интернет, не предоставляет программно-го обеспечения пользователям.
- В большинстве медиатек отсутствует электронная доставка документов.
- Особенно печально выглядит ситуация с видеотекой и фонотекой. У большинства электронных библиотек таких ресурсов просто нет.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Электронный каталог

Практически во всех инновационных вузах, участвовавших в Мониторинге, есть электронные каталоги. Электронные каталоги в библиотеках ведущих российских вузов начали появляться в начале 1990-х гг. Однако формирование фондов электронных библиотек и создание электронных каталогов шли самостоятельно и независимо друг от друга. Настоящие полнотекстовые электронные библиотеки стали формироваться лишь через спустя 5–10 лет после создания электронных каталогов, а в ряде вузов, включенных в выборку Мониторинга, этот процесс не был начат до сих пор.

Электронный каталог, как правило, отражает все новые поступления библиотеки. Ретро-издания в каталоге полностью или частично отсутствуют в связи с тем, что часть фонда постоянно находится на руках.

Полнотекстовые фонды

Основу полнотекстовых фондов на сегодняшний день, как правило, составляют электронные подписные издания и реферативные базы. Более половины медиатек содержат полнотекстовые документы – в виде отсканированных материалов или структурированных книг.

Аудио и видеофайлы

В настоящее время этот вид информации в электронных библиотеках представлен очень слабо. В вузах, в которых аудио- и видеоматериалов традиционно присутствуют как часть учебного процесса (например, языковые вузы), подразделения, обеспечивающие пользователей этими материалами, могут структурно не принадлежать к библиотеке и функционировать независимо от подразделений, обеспечивающих работу электронной библиотеки.

Галерея изображений

Создание галерей изображений, позволяющих описать атрибуты для каждого отдельного изображения и осуществлять по ним поиск, в настоящее время не входит в ближайшие планы библиотек. Это связано, с одной стороны, с отсутствием выраженного запроса со стороны пользователей библиотек, которые, как правило, имеют доступ к необходимым им иллюстративным материалам (в основном, через Интернет). Для специалистов-библиографов каталогизация изображений является новой и нестандартной задачей, не вполне ясной по своему техническому воплощению.

Внешние ссылки на Интернет сайты, описанные в каталоге как единицы хранения

В настоящее время ссылки на Интернет-сайты присутствуют в электронных библиотеках, однако,

в неоформленном с точки зрения каталогизации виде. Это направление деятельности в электронных библиотеках осложняется двумя основными проблемами:

- Значительная трудоемкость этой работы и недостаток квалифицированных кадров;
- Установка администрации библиотеки о целесообразности каталогизации ссылок на ресурсы в Интернет в связи с их широкой доступностью.

Дополнительные сервисы

Электронные библиотеки предоставляют своим пользователям возможность работать с полнотекстовыми электронными документами вне помещения библиотеки. Это достигается тремя способами:

- Удаленный доступ к электронному каталогу и работа с полнотекстовыми материалами через Интернет;
- Поиск необходимых изданий в электронном каталоге в библиотеке и пересылка полных текстов по электронной почте;
- Возможность распечатать необходимую информацию (как правило, платная услуга).

Перечень платных услуг библиотеки вуза зависит от таких факторов, как бюджет вузовской библиотеки (например, не все вузы могут позволить себе свободный доступ в Интернет из читального зала библиотеки), затраты на расходные материалы, а также от расхода рабочего времени персонала.

Консультативная помощь пользователям

По данным Мониторинга, вопрос консультативной помощи пользователям электронной библиотеки не особо актуален, не воспринимается специалистами-библиотекарями и пользователями как проблемный. Интуитивная понятность интерфейса программных оболочек электронных библиотек, сходство поисковых систем электронного каталога с привычными для пользователей поисковыми машинами Интернет, единая или сходная структура электронного и бумажного каталога библиотеки позволяют пользователям быстро осваивать принципы работы в электронной библиотеке. Обращение с компьютерным интерфейсом электронной библиотеки представляет сложность для тех читателей библиотек, кто плохо владеет компьютером, однако такие проблемы являются сравнительно редкими и характерны, в основном, для старшего поколения профессорско-преподавательского состава.

Вместе с тем практически во всех обследованных библиотеках организована консультативная помощь пользователям. Чаще всего – это специальные инструкции для пользователей, размещаемые в читальном зале, или присутствие специальных консультантов.

Несколько реже осуществляется учет обратной связи от пользователей, сбор их пожеланий, посто-

янный Мониторинг их потребностей, хотя данное направление имеет большое значение для эффективной работы электронных библиотек.

Совершенно неразвита в электронных библиотеках такое направление как проведение специальных тренингов для пользователей, хотя это имеет большое значение для формирования компетенций читателей электронных библиотек и пользователей электронных каталогов вузовских библиотек. Лишь в 3 библиотеках из 20 обследованных проводятся такие тренинги. Основу их составляют сведения по структуре электронного каталога и организации поиска, а также содержательном наполнении новых поступлений электронных библиотек. В некоторых библиотеках помимо групповых занятий для пользователей практикуются индивидуальные консультации. Однако, по данным Мониторинга, такие консультации на данный момент оказываются менее эффективными в плане приобщения читателей библиотек к электронным ресурсам, чем групповые тренинги, что связано с недостатком библиотечных кадров, достаточно компетентных, чтобы осуществлять грамотные консультации пользователей.

Ключевые проблемы медиатек вузов на сегодняшний день

Данные проведенного Мониторинга позволили выявить ключевые проблемы университетских медиатек:

- Согласно российскому законодательству (Государственным образовательным стандартам высшего профессионального образования) студенты не могут пользоваться исключительно электронными учебными пособиями.
- Значительная проблема касается ресурсов и сервисов электронных библиотек. Среди общедоступных ресурсов – только электронный каталог, видео и фонотеки практически отсутствуют, ссылки на ресурсы Интернет в каталог не включаются, электронная доставка документов и другие сервисы затруднены или отсутствуют.
- Недостаточная комплектация полнотекстовых фондов электронных библиотек не позволяет удовлетворять потребности читателей в необходимой им литературе.
- Недостаточная обеспеченность электронных библиотек необходимым оборудованием не позволяет с одной стороны, комфортно организовать работу библиотекарей и пользователей, а с другой – эффективно пополнять полнотекстовые фонды библиотеки.
- Недостаточное количество сотрудников библиотеки в целом, и особенно тех, кто отвечает за работу непосредственно электронной библиотеки. Нехватка специалистов негативным образом сказывается на эффективности работы электронной библиотеки и препятствует наполнению ее фондов необходимыми информационными ресурсами.
- Помимо нехватки специалистов значительную проблему также составляет недостаточная ква-

лификация сотрудников для работы в рамках инновационной модели библиотеки.

- Отсутствие в электронных библиотеках удаленного доступа к ресурсам через Интернет, что связано как с ограниченными техническими возможностями вузов, так и с ограничениями, накладываемыми на функционирование электронных библиотек российским законодательством. Отсутствие удаленного доступа к ресурсам электронной библиотеки создает дополнительные неудобства для пользователей.
- Одной из проблем на пути развития электронных библиотек является отсутствие согласованной архитектуры создаваемых электронных библиотек, тем более сервис-ориентированной архитектуры. В связи с этим возникает важнейшая задача «модуляризировать» электронную библиотеку с тем, чтобы иметь возможность управлять разнородным цифровым контентом путем обращения к тем или иным сервисами отдельных модулей внутри данной электронной библиотеки, а также между ней и другими информационными системами, которые с ней связаны.
- Существует психологический барьер в сознании сотрудников библиотек и преподавателей вузов. Дело в том, что в их сознании существует некое предубеждение, что удаленная работа с информационными ресурсами и возможность свободно оперировать информацией в электронном виде нежелательна для учебного процесса, поскольку в результате студенты могут проявить недостаточную познавательную активность. Кроме того, работа с бумажными носителями информации более привычна и удобна для пользователя, и не может быть полностью заменена электронными источниками информации.
- Одной из важнейших проблем на пути создания и модернизации электронных библиотек вузов является юридическая проблема. Действующее законодательство РФ не регулирует напрямую деятельность электронных библиотек, а именно не содержит механизмов, правил их создания и функционирования. Эта ситуация приводит к тому, что библиотеки вынуждены пользоваться принципом «что не запрещено законом, то разрешено», увеличивая тем самым вероятность предъявления им обвинений в нарушении авторских и смежных с ними прав.

Отношение потенциальных пользователей к созданию межвузовских электронных библиотек

Среди пользователей присутствует значительная заинтересованность в создании межвузовских электронных библиотек. Эта заинтересованность проявляется в трех основных направлениях:

- Фонды инновационных библиотек представляют значительный интерес для вузов «с периферии»;

- Инновационные вузы заинтересованы в доступе к фундаментальным библиотекам других вузов, в течение длительного времени функционирующих в качестве основы для развития научных школ;
- Интерес представляет доступ к информационным фондам вузов аналогичного профиля.

ВЫВОДЫ ИССЛЕДОВАНИЯ ЭЛЕКТРОННЫХ БИБЛИОТЕК ИННОВАЦИОННЫХ ВУЗОВ РФ

Контент

На сегодняшний день наиболее актуальным для пользователей библиотек инновационных вузов является решение проблемы недостаточного уровня наполнения научными журналами, переводной и зарубежной литературой, современными учебными пособиями. Недостаток в библиотеке вуза этих категорий литературы оказывает наиболее выраженное влияние на степень общей удовлетворенности читателей фондами библиотек. По данным исследования, потенциальные пользователи электронных библиотек вузов заинтересованы в пополнении фондов библиотек разнообразными категориями электронных ресурсов. Однако первоочередное значение имеют учебные издания, научные журналы, а также переводная и зарубежная литература.

- Научные журналы, переводная и зарубежная литература наиболее часто отмечались как пользователями, так и представителями администрации библиотек в числе категорий, недостаточно представленных в фондах библиотек в большинстве инновационных вузов, вошедших в выборку исследования.
- В вузах, читатели которых испытывают дефицит учебной и методической литературы, как администрацией библиотек, так и пользователями с большей вероятностью отмечается неудовлетворенность объемом фондов. Мониторинг выявил заинтересованность представителей администрации библиотек вузов в приобретении электронных версий учебных пособий, рекомендованных для студентов, поскольку это повышает доступность учебной литературы, пользующейся спросом и отсутствующей в библиотеках вузов в необходимом количестве в печатном виде
- Потребности в научной и учебной литературе существенно варьируют в зависимости от вуза. Степень удовлетворенности пользователей фондами библиотеки вуза, и, в частности, электронными ресурсами, определяется как объективными характеристиками фондов (объем, тематическое разнообразие), так и уровнем запросов пользователей, связанным со спецификой обучения в вузе, имеющим тенденцию к росту в вузах, использующих инновационные методы обучения, практикующих активное включение студентов научную работу. Библиотеки, распо-

лагающие недостаточными для удовлетворения потребностей пользователей печатными фондами, характеризуются большим разнообразием запросов пользователей в отношении разных категорий литературы.

Пополнение фондов

- Основным источником пополнения электронных библиотек для вузов, на сегодняшний день располагающих наиболее развитыми электронными ресурсами, является подписка на внешние электронные базы (реферативные, базы периодики), приобретение электронных книг у внешних издателей, а также расширение доступа пользователей к учебно-методическим работкам подразделений вуза.
- Расширение доступа к внешним электронным ресурсам требует, помимо привлечения дополнительного финансирования, решения юридических вопросов, связанных с авторским правом, технических вопросов, связанных с защитой информации от несанкционированного доступа, вопросов программной совместимости внешних электронных ресурсов и электронной библиотеки вуза.
- Проблема авторского права является основным препятствием для расширения доступа пользователей к имеющимся фондам электронных библиотек. Это касается как литературы, имеющейся в вузе в печатном и электронном виде, так и электронных изданий. Проблема авторских прав чаще всего называлась участниками исследования в качестве препятствия к развитию межвузовских электронных библиотек, расширению доступа к ресурсам библиотеки вуза для внешних пользователей и переводу в электронный вид необходимых библиотеке печатных материалов, имеющихся в недостаточном количестве.
- Для вузов естественнонаучного и технического профиля особенно большое значение в качестве источника пополнения фондов электронной библиотеки, по сравнению с вузами гуманитарного и экономического профиля, имеют публикации вуза. Это касается как учебных, методических изданий, так и научных публикаций. Это связано с тенденцией к узкопрофильной специализации в этих дисциплинах, спецификой учебных программ в разных технических вузах.

Популяризация концепции электронных библиотек вузов: коммуникативный аспект развития электронных библиотек

- На сегодняшний день, концепция электронных библиотек нуждается в популяризации. По данным исследования, более широкому использованию уже имеющихся электронных ресурсов библиотек вузов препятствует недостаточная осведомленность о них потенциальных пользо-

вателей, фрагментарность сведений об обновлениях фондов электронных библиотек, а также приверженность пользователей библиотек вузов привычной структуре удовлетворения потребностей в научной и учебной литературе.

- На данный момент в российских инновационных вузах отсутствует единая концепция интеграции различных электронных ресурсов вуза в электронную библиотеку. Термины «медiateка» и «электронная библиотека» на сегодняшний день имеют редкое употребление в среде персонала и читателей библиотек вузов. Термин «медiateка», как правило, не связывается с электронной библиотекой в широком смысле и понимается, в первую очередь, как ресурсы мультимедиа, аудио и видеофайлы. Пользователи электронных ресурсов библиотек, как правило, не соотносят эти ресурсы со специальным подразделением библиотеки. Представители администрации библиотек под «электронной библиотекой» понимают в первую очередь полнотекстовые библиотеки, а электронные каталоги, имеющиеся на сегодняшний день во всех вузах, включенных в выборку исследования, зачастую воспринимаются как отдельная категория ресурсов, не подразумевающая наличия электронной библиотеки.
- Перспективы расширения круга пользователей электронных библиотек вузов связаны, в частности, с высоким уровнем распространения электронных ресурсов – в первую очередь, ресурсов сети Интернет – в качестве дополнительного источника информации, помимо библиотек вузов, читателями библиотек вузов. Ресурсы сети Интернет являются источником, наиболее активно используемым как преподавателями, так и студентами для решения учебных задач и в научной работе. Электронный формат учебных и научных материалов широко используется как преподавателями, так и студентами, а возможность доступа к необходимой литературе через Интернет является преимуществом, повышающим привлекательность источника этой литературы для пользователей.
- Наиболее эффективным способом информирования потенциальных пользователей о ресурсах и сервисах электронных библиотек вузов является проведение очных обучающих занятий (курсов, тренингов). Повышение информированности профессорско-преподавательского состава о возможностях электронных библиотек является одновременно способом популяризации электронных ресурсов в среде студентов, поскольку рекомендации преподавателей играют ведущую роль в определении перечня литературы, активно используемой студентами вузов.
- На сегодняшний день персонал библиотек в полной мере не удовлетворяет потребности читателей в консультативной помощи, необходимой для максимально эффективного использования ресурсов электронных библиотек.

Характеристика читательской аудитории электронных библиотек

На сегодняшний день, наиболее активными пользователями электронных библиотек вузов являются преподаватели, студенты старших курсов и аспиранты. Это связано, с одной стороны, с минимальными ограничениями в доступе для этих категорий читателей и, с другой стороны, с тем, что на сегодняшний день электронные ресурсы, прежде всего, используются в работе над научными публикациями.

Исследование выявило существенные различия в структуре использования электронных ресурсов различными профессиональными категориями читателей. Наиболее активными пользователями электронных ресурсов, как по частоте обращения в электронную библиотеку, так и по степени разнообразия запросов, являются представители экономических и гуманитарных специальностей.

Проблемы доступа к ресурсам электронных библиотек

Наименьшее количество ограничений в доступе к ресурсам электронных библиотек испытывают на себе преподаватели вузов. Ограничения для студентов связаны, в первую очередь, с недостатком технических возможностей для свободного доступа большего числа желающих к электронным библиотекам – в силу отсутствия необходимого количества рабочих мест для пользователей. Решение данной проблемы связано, как правило, с развитием телекоммуникационной инфраструктуры электронной библиотеки: интеграции локальной сети вуза. Снятие ограничений в доступе к ресурсам электронных библиотек для внешних пользователей (не являющихся студентами и преподавателями вуза) лежит в плоскости решения проблем авторского права и развития системы удаленного доступа к электронным библиотекам через сеть Интернет.

Исследование выявило высокую степень заинтересованности читателей в расширении возможностей для удаленного доступа к ресурсам электронной библиотеки – через сеть Интернет. Особенно актуально это для студенческой аудитории, поскольку студенты, по сравнению с преподавателями, имеют на сегодняшний день меньше возможностей доступа к электронной библиотеке через локальную сеть вуза – в силу недостатка в вузе рабочих мест с доступом к локальной компьютерной сети вуза.

Характеристика электронных ресурсов

Исследование выявило основные требования к интерфейсу поисковых систем электронных библиотек и электронных каталогов библиотек вузов. Это возможность поиска как по ключевым словам, так и тематического поиска, возможность расширенного поиска, с учетом большого количества атрибутов единиц хранения, а также возможность полнотекстового поиска.

Среди полнотекстовых материалов, которыми располагают электронные библиотеки вузов, на сегодняшний день преобладают материалы в pdf формате, представляющие собой электронные копии печатных материалов, созданные путем сканирования последних или при электронном наборе книг, издаваемых в вузах. Исследование выявило потребность пользователей в пополнении фондов полнотекстовых электронных ресурсов, удобных для чтения и печати, использование которых создает преимущества, связанные с электронной формой этих материалов (удобство отбора, поиска информации).

Исследование выявило существенные региональные различия в структуре электронных ресурсов инновационных вузов, вошедших в выборку исследования. Для региональных вузов, вошедших в выборку, в отличие от московских, характерна более полная представленность разнообразных полнотекстовых материалов в электронных библиотеках вузов, а также более высокая степень интеграции электронных ресурсов в пределах вуза, создающая предпосылки для развития системы удаленного доступа к ресурсам электронной библиотеки (в частности, наличие единого портала доступа).

Материально-техническая база и проблемы финансирования

Основные бюджетные затраты в связи с созданием и модернизацией электронных библиотек касаются следующих аспектов: закупки электронных изданий и подписки на электронные базы, приобретения компьютерной техники для работы с электронными материалами в пределах вуза, телекоммуникационного оборудования и оргтехники (прежде всего, для перевода печатных фондов библиотеки в электронный вид), оплаты труда персонала, участвующего с созданием и модернизации электронной библиотеки.

Исследование выявило большую степень дифференциации между вузами в уровне развития материально-технической базы электронных библиотек (наличия необходимой компьютерной инфраструктуры, помещений). При этом, по оценкам представителей администрации библиотек всех вузов, участвующих в исследовании, дальнейшее проведение программы создания и модернизации электронных библиотек, требует привлечения дополнительных материально-технических ресурсов.

В ряде вузов исследование выявило тенденцию дисбаланса в финансировании развития материально-технической части библиотек и повышения фонда заработной платы сотрудников библиотек и иных категорий персонала, принимающих участие в развитии электронных библиотек в вузах: представителей профессорско-преподавательского состава, ИТ персонала. Характеристикой этого дисбаланса является недостаток средств на повышение заработной платы при удовлетворительном финансировании модернизации инфраструктуры электронных библиотек.

Одним из препятствий развитию в вузах электронных библиотек являются кадровые проблемы:

недостаток квалифицированных кадров и общая неуккомплектованность штата библиотек. Низкая заработная плата и связанная с этим низкая престижность труда библиотекарей на сегодняшний день не позволяет библиотекам вузов самостоятельно решить проблему кадрового дефицита. В наибольшей степени это касается привлечения квалифицированного ИТ персонала. Особенно остро проблема нехватки кадров стоит в связи с интенсификацией процессов создания и модернизации электронных библиотек, поскольку этот переходный этап повышает потребность библиотек в персонале.

ВЫВОДЫ И РЕКОМЕНДАЦИИ ПО СОЗДАНИЮ ЭЛЕКТРОННЫХ БИБЛИОТЕК ВУЗОВ ОСНОВАННЫЕ НА АНАЛИЗЕ МЕЖДУНАРОДНОГО ОПЫТА

Создание медиатек – важный шаг на пути модернизации образовательного процесса. Насколько можно судить по международному опыту создания, функционирования и использования медиатек, наиболее развитые регионы уже с начала 90-х годов XX века осознали важность создания единых информационных пространств в своих высших учебных заведениях.

Единые информационные пространства позволяют с одной стороны упростить, а с другой углубить и улучшить образовательный процесс. Если студентам не надо будет прикладывать особых усилий и проводить сложные манипуляции для поиска информации, то это только послужит доводом в пользу использования дополнительной информации во время обучения.

Важно отметить, что западные университеты очень серьезно подходят к вопросу развития медиатек. Практически во всех медиатеках создан наблюдательный совет, следящий за деятельностью и требующий отчета от библиотек за выполненные работы. Ежегодно медиатеки выпускают годовые отчеты, где описывают свою деятельность, ее результаты и планы на ближайший год. Также некоторые медиатеки выпускают утвержденные наблюдательным советом планы на среднесрочное (3–5 лет) развитие.

Исходя из международного опыта создания и функционирования медиатек, можно предложить следующие рекомендации по созданию единых информационных пространств в вузах Российской Федерации:

- Создание адекватной проекту технической базы для работы библиотекарей, что подразумевает хорошую оснащенность компьютерами (например, в одном из Университетов ЮАР для этих целей используется 495 компьютеров, приложение 16), сканерами для оцифровки материалов.
- Использование современных технологий в работе библиотекарей: доступ к сети Интернет со всех компьютеров, доступ к внутренней сети медиатеки, использование различных online сервисов для каталогизации материалов, отслеживания статуса и сроков сдачи выданных ма-

териалов, бронированию материалов, подписке на электронные материалы, общению с пользователями и т.д.

- Обучение работников медиатек работе с новыми техническими средствами и новыми технологиями.
- Создание достаточного количества удобных, оснащенных современным оборудованием рабочих мест для пользователей.
- Оснащение пользовательской зоны всеми видами оборудования для работы с различными форматами информации.
- Грамотное, продуманное построение структуры Web-сайта медиатеки.
- Создание удобного поиска с расширенными возможностями и интуитивно понятным интерфейсом.
- Использование современных технологий и сервисов для создания и функционирования пользовательских сервисов.
- Сотрудничество медиатек с различными библиотеками и медиатеками других вузов, в том числе для предоставления своим пользователям такой услуги как «Заем материалов в сторонних библиотеках».

Примером наиболее успешных и удобных в использовании университетских библиотек могут служить медиатеки западноевропейских вузов. Они также информационно и технологически насыщены, как и американские, но гораздо более удобны в использовании.

Литература

- [1] <http://rating.rbc.ru/article.shtml?2007/09/20/31640751>
- [2] Гилберт Черчилль. Маркетинговые исследования. СПб. : Питер, 1999.

Results of monitoring and the basic recommendations concerning development of digital libraries of high schools

E.A. Kotlyar, D.Yu. Guzhelya, N.A. Polikhina

The report results complex research concerning digital university libraries in Russia and international experience about it. The report illustrates problems and requirements of different groups of digital libraries users. There are recommendations of development and improvement of digital libraries in this report also.

Научно-образовательная электронная библиотека вуза*

© А.Г. Абросимов, Д.С. Зув

Казанский государственный университет
dzuev@ksu.ru

Аннотация

В статье рассматриваются критерии, которым должна удовлетворять научно-образовательная электронная библиотека вуза. Обсуждаются проблема и функциональные особенности электронной научно-образовательной библиотеки вуза.

Как известно (см., например, [1]), отличительной чертой сегодняшнего этапа развития человечества является представление информации и знаний не только в традиционной печатной, но и в электронной, цифровой форме, что позволяет принципиально иначе создавать, хранить, организовывать доступ и использовать информацию в любой форме. К тому же современные информационные технологии привели к тому, что большое число современных информационных ресурсов сразу создается в электронном виде. В условиях формирования информационного общества чрезвычайно важным инструментом устойчивого экономического и социального развития является обеспечение публичного (в том числе удаленного) доступа к социально значимой информации, в первую очередь, научного, образовательного и культурного характера.

К сожалению, степень доступности информационных ресурсов России, имеющих огромную ценность и столь же огромные объемы, до сих пор остается слишком низкой. Сегодня очевидна необходимость эффективной кооперации всех учреждений, занимающихся созданием и распространением информации и знаний – научных учреждений, информационных центров, библиотек, архивов, музеев и т. д., в области создания информационных систем, предназначенных для хранения и предоставления доступа к информации – электронных библиотек (ЭБ).

В создании информационных ресурсов весьма значительную роль играют вузы, основные информационные ресурсы которых – образовательный и научный – в настоящий момент уже создаются, как правило, в электронном виде. При этом многие вузы уже имеют большой опыт работы с электронными каталогами, создаваемыми с помощью автоматизи-

рованных библиотечно-информационных систем, которые в определенной степени можно считать предтечами электронных библиотек. Подтверждением этого является проект Федеральной целевой программы развития образования на 2006–2010 гг. «Модернизация инфраструктуры высших учебных заведений гуманитарного профиля», в рамках которого было разработано программное обеспечение и 15-ю университетами создан сводный электронный каталог информационных образовательных ресурсов гуманитарного профиля.

Для вузов предоставление открытого доступа к своим ресурсам имеет ряд принципиально важных позитивных моментов. Во-первых, это форма рекламы. Если вуз выставляет на своем сайте хорошие учебно-методические пособия, то это означает, что в нем работают хорошие преподаватели, у которых есть чему научиться, причем даже беглый просмотр вузовских сайтов убеждает в том, что выставляются действительно достойные материалы, как правило, в тех направлениях, в которых эти вузы наиболее сильны. Во-вторых, размещение в Интернете – это и форма общественной экспертизы работы. Сейчас мало кто решится выставить на всеобщее обозрение плагиат или откровенно слабую компилятивную работу.

Сегодня большинство ЭБ – тематические и содержат в основном электронные аналоги печатных изданий, ЭБ же вуза содержит более широкий спектр информационных ресурсов, т. е. ЭБ вуза политематическая. Это определяется хотя бы тем, что в вузе существует всегда не менее двух направлений деятельности – образовательная и научная, а в классических университетах, объединяющих множество научных направлений, ЭБ однозначно не может быть посвящена единственной тематике.

Все вышеизложенное, естественно, вызывает необходимость создания полноценной электронной библиотеки в вузе, аккумулирующей все собственные ресурсы, а также внешние ресурсы, доступ к которым вуз приобрел или получил бесплатно. Но специфика образовательного учреждения выставляет особые требования к создаваемой информационной системе: во-первых, как уже упоминалось выше, это необходимость интегрировать электронные коллекции, как собственные, так и приобретенные. Во-вторых, в каждом вузе ведутся разработки или внедрение информационных систем, предназначенных для автоматизации образовательной, административной и т. д. деятельности, и электронная биб-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

лиотека должна рассматриваться как естественная часть этого комплекса информационных систем, который можно назвать системой информатизации деятельности вуза.

Так, например, помимо всех традиционных функций, электронные библиотеки можно рассматривать как хранилище данных. При этом хранимая в них информация может быть востребована как другими информационными системами вуза, так и при работе различных подразделений вуза. Например, при конкурсном отборе на должности должен быть предоставлен список научных трудов. Подобный список может быть подготовлен самим претендентом и может содержать неточности и ошибки. В тоже время большинство трудов кандидата (уже работавшего) используется в учебном процессе и хранится в научной электронной библиотеке. Таким образом, информация, хранимая в одной из систем, востребована другими системами [2].

В-третьих, электронные коллекции вуза различаются по описываемым материалам, по области применения. Как следствие, каждая коллекция имеет свой профиль метаданных. Примером могут служить две центральные коллекции – образовательная и научная. Для образовательной коллекции естественно использовать форматы метаданных LOM, RUSLOM, а для научной можно использовать, например, Dublin Core

Разумеется, Казанский университет не является исключением, и попытки создания если не полноценной электронной библиотеки, то различных электронных коллекций предпринимаются уже давно. На данный момент можно утверждать, что в КГУ созданы электронные коллекции, разнородные по своему содержанию (например, коллекция периодической печати 19-го века [3, 4], коллекция учебно-методических материалов и т. п.), каждая из которых используется как для научных изысканий, так и для образовательной деятельности.

Наличие разнородных, разрозненных и плохо связанных между собой электронных коллекций не может удовлетворять всех потребностей пользователей, необходимо объединять все уже созданное и создаваемое в единую информационную систему. Подобное объединение позволило бы более четко систематизировать, учитывать созданные информационные ресурсы, расширило бы спектр предоставляемых сервисов: единовременный общий поиск по всем коллекциям, регистрация и движение электронных документов, ручное и автоматическое пополнение коллекций, предоставление единой точки входа для взаимодействия с другими системами, и т. д.

При объединении всех электронных коллекций в одно целое (т. е. при создании новой информационной системы) нельзя не учитывать специфику вуза и изначально планировать интеграцию создаваемой электронной библиотеки в единую информационную среду образовательного учреждения. Дополнительно создаваемое программное обеспечение для электронной библиотеки должно удовлетворять и другим критериям. Рассмотрим их подробнее.

Электронная библиотека представляет собой иерархическую систему – она не является монолитной, и основной ее структурной единицей являются коллекции, которые объединяют близкие по типу электронные документы. В свою очередь ЭК тоже могут разделяться и составлять иерархическую структуру. Коллекции могут различаться по типу хранимых электронных документов, по профилю метаданных, по программным средствам, реализующим основные функции коллекции, и другим признакам.

Следовательно:

- каждая коллекция может иметь свой профиль метаданных, свое лингвистическое обеспечение;
- соответственно, программное обеспечение каждой коллекции должно учитывать ее специфику;
- объединение коллекций должно происходить на основе коммуникативного формата метаданных, в качестве которого может выступать, например, Dublin Core; к базовым элементам Dublin Core могут добавляться дополнительные уточняющие квалификаторы, которые отражают специфику профиля каждой коллекции; при необходимости метаданные коллекции могут преобразовываться в коммуникативный формат, в данном случае все специфические элементы могут быть преобразованы, например, в ключевые слова;
- общий поиск по всей электронной библиотеке осуществляется по общим элементам метаданных, при этом специфичные элементы профиля метаданных коллекции могут игнорироваться;
- поиск внутри коллекций должен проводиться в соответствии с профилем метаданных этой коллекции, с использованием списков подстановок и авторитетных файлов, что влечет высокую релевантность поиска;
- основная форма поисковой системы предлагает поиск по всей электронной библиотеке или выбор конкретной коллекции, тогда предлагается переход к поисковой подсистеме коллекции, поскольку поисковая система ЭБ тоже имеет иерархическую структуру, как и сама библиотека; на разных уровнях могут возникать свои поисковые подсистемы, учитывающие специфику текущего раздела;
- создание программного обеспечения каждой электронной коллекции производится автоматически; для этого структура метаданных коллекций описывается на формальном машиночитаемом языке (например, с использованием XML Schema), причем первичный вариант структуры создается до формирования самой электронной коллекции; на основе созданного описания структуры метаданных генерируется вся программная система: экранные формы, таблицы баз данных, поисковая система и т. д.;
- при условии использования XML Schema [3, 4] как инструмента описания структуры метаданных может потребоваться разработка адаптированного редактора XML Schema, программного обеспечения для анализа и разбора XML Schema, программного обеспечения для ввода, редактирования и просмотра

метаданных, отдельных поисковых систем для коллекций;

- такой подход позволит сделать процесс формирования коллекций итерационный [4], поскольку качественный анализ всего материала для создания структуры метаданных перед созданием электронной коллекции провести не всегда представляется возможным, в процессе же формирования коллекции в структуру метаданных могут вноситься уточнения, поэтому необходимо программное обеспечение корректировки описания профиля и списков подстановок атрибутов;

- для облегчения поиска по электронной библиотеке также требуется механизм построения общих списков атрибутов или авторитетных файлов по всей электронной библиотеке, например, список авторов-сотрудников вуза, списки факультетов, кафедр, специальностей и т. д.;

- возможность интеграции с другими существующими электронными библиотеками и коллекциями может потребовать и разработки собственных веб-сервисов, и использования специальных протоколов для обмена информацией и удаленного поиска; например, проект «Электронное полнотекстовое объединенное собрание» АРБИКОН (<http://www.arbicon.ru/projects/epos/>) предоставляет возможность подключения своего каталога для поиска по протоколу Z39.50;

- возможность интеграции с внутренними информационными системами вуза, т. е. потребуется разработка интерфейса, позволяющего из любого модуля информационной системы формировать запрос к поисковой системе коллекции и передавать полученные результаты; для реализации подобного интерфейса дополнительно требуется разработка протокола передачи найденной информации между системами (естественным вариантом нам представляется сохранение результатов поиска в виде XML-документа, который передается или обрабатывается вызвавшим приложением); необходимо заметить, что вовсе не обязательно производить поиск по всей электронной библиотеке, поскольку на каждом этапе известно, в какой конкретно электронной коллекции отражена необходимая информация.

Заметим, что ЭБ вуза – научно-образовательная. Причем, помимо научных и образовательных коллекций, библиотека может содержать и коллекции административных документов, и смешанные коллекции имеющие особый профиль метаданных, которые нельзя напрямую отнести ни к научной составляющей ни к образовательной (например, ЭК периодической печати 19 - нач 20вв). Это не может позволить использовать без существенных доработок существующее ПО для библиотек. Во главу угла образовательной составляющей вузовской ЭБ конечно ставится обеспечение учебного процесса, что накладывает свои отпечатки на всю структуру библиотеки – на информационные ресурсы, на функции, на предоставляемые сервисы.

Основной особенностью информационных ресурсов ЭБ вуза является их неоднородность в раз-

личных аспектах – разнообразие сред представления (текст, числовые данные, статические изображения, видео, аудио, мультимедиа). При этом разнородные данные могут относиться к одному и тому же исследованию, однако фактически, учитывая их разнородность, не всегда могут храниться в одной и той же электронной коллекции. Таким образом, подобные информационные ресурсы должны иметь возможность объединения в коллекции по различным признакам, при этом один и тот же ресурс может состоять в нескольких коллекциях ЭБ.

Так, метаданные, описывающие электронные образовательные ресурсы (ЭОР), должны учитывать особенности предметной области ЭБ, т. е. включать элементы, специфичные для описания образовательных ресурсов. Таковой, например, является схема метаданных LOM, точнее, ее адаптация RUSLOM, учитывающая особенности российского образовательного процесса.

Любой ЭОР образовательной ЭБ может быть «привязан» к программе конкретного учебного курса (дисциплины), что должно отражаться в метаданных. Учебный курс также может быть каталогизирован и снабжен развернутым описанием.

Для каждого ЭОР может устанавливаться свой уровень доступа. На собственно электронный документ – никаких ограничений не накладывается, он может быть любого типа, это могут быть и аналоги печатных изданий, и ЭД более сложной структуры. Тем не менее, следует выделить тип электронных документов, наиболее специфичных для образовательной ЭБ: учебно-методические комплексы (УМК), включающие: развернутые учебные программы по дисциплинам; контрольно-измерительные материалы, средства текущего и итогового контроля знаний; ссылки на полнотекстовые электронные учебно-методические материалы, интерактивные электронные учебные курсы из вузовской ЭБ, других ЭБ и Интернета, а также ссылки на учебную литературу, имеющуюся в библиотеке вуза.

Такие УМК удобно использовать в дистанционном образовании, так как они представляют законченный продукт, позволяющий полноценно изучать дисциплину.

УМК это сложный составной электронный документ, для обеспечения работ с УМК необходимо обеспечить в ЭБ и возможность аналитического описания, и эффективные средства навигации в самом электронном документе с разделением прав доступа различным пользователям.

Если же рассматривать научную составляющую вузовской ЭБ, то здесь тоже необходимо выделять свои особенности. Область применения научных ЭБ нельзя ограничить только обеспечением научных исследований, хотя подобная область сама по себе очень широка – достаточно сказать лишь о разнообразии и предметов исследования, и технологий исследования. Научная составляющая вузовской ЭБ также представляет и серьезный интерес для образования, поскольку все передовые достижения нау-

ки должны отражать и на преподавании различных дисциплин. Серьезное влияние и на цели создания и на специфику применения научных ЭБ оказывает тот факт, что научное сообщество всегда стремилось к широкому сотрудничеству, осознавало необходимость обмена информацией, несмотря на разнообразие предметов исследования, разнообразие технологий научных исследований. Демократичность научного сообщества является одной из основных причин успешного развития научных ЭБ и, как следствие, при создании научной ЭБ необходимо учитывать многоязыковую среду научного сообщества, что влечет необходимость перевода метаданных на различные языки, неоднородность используемой информации, предоставлять возможность оперирования данными на разных уровнях абстракции.

Но практически во всех случаях основой информационных ресурсов ЭБ являются научные публикации в различных формах. Помимо этого ЭБ так же может содержать библиографическую информацию, результаты различного рода экспериментов, наблюдений, измерений, моделирования исследуемой реальности, модели исследуемых процессов, явлений, феноменов, представленные в разнообразных формах, новостную, событийную информацию (календарь конференций и т.п.). Таким образом, информационное наполнение научной составляющей вузовской ЭБ достаточно разнородно.

Другой особенностью информационных ресурсов научных ЭБ является возможность представления предметной области в форме, позволяющей непосредственно проводить исследование, то есть использование научной ЭБ не только для информационной поддержки, но и в качестве исследовательского полигона.[5]

Такой подход к формированию информационных ресурсов вузовской ЭБ требует учитывать возможность сверхбольших объемов содержимого ЭБ, что характерно для результатов различного рода экспериментов, наблюдений, измерений, моделирования реальных процессов.

Еще одной специфической особенностью информационных ресурсов научных ЭБ, имеющей существенное значение и определяющей особые требования к функциональным характеристикам ЭБ, является фактор старения информационных ресурсов, как результатов различного рода экспериментов, так и информации, содержащейся в статьях, монографиях и других изданиях. В завершение приведем некоторые выводы:

- научные ЭБ обладают специфической предметной областью, и с этим связаны особые требования к ним;
- специфичны по содержанию и неоднородны их коллекции информационных ресурсов;
- научные ЭБ должны быть оснащены специфическими сервисами, благодаря которым они могут использоваться не только как источник информационных ресурсов, но и как полигон для непосредственных научных исследований [5].

Это краткое перечисление только основных возможностей, которые составляют специфику вузовских электронных библиотек. Вдобавок к приведенным и, как нам кажется, необходимым функциональным возможностям программного обеспечения электронной библиотеки необходимо также учитывать и требования безопасности – т. е. необходимы авторизация и разграничение прав доступа пользователей, регистрация электронных документов, инструменты архивного копирования.

Прежде чем приступить к созданию собственной электронной библиотеки, нами был проведен анализ функциональных возможностей существующих электронных систем хранения данных. Рассматривалось пять существующих систем – Dspace, ePrints, Greenstone, ELSA, Fedora. Была сделана попытка сравнения по нескольким критериям: учитывались возможности поиска и навигации, возможности разграничения прав доступа и самокаталогизации, структура и метаданные электронных документов, возможность обмена по различным протоколам. Отметим, что все рассматриваемое программное обеспечение является свободно распространяемым.

Все рассматриваемые системы предоставляют полнотекстовый и атрибутный поиск различной полноты, профиль метаданных основан на Dublin Core.

Все зарубежные аналоги в основном предоставляют доступ по протоколу OAI-PMH (OAI metadata harvesting protocol), единственная российская разработка ELSA поддерживает обмен по протоколам Z39.50 и SRU/SRW.

Наиболее функционально полными системами можно считать DSpace и Fedora, часть предоставляемых ими возможностей вообще не присутствуют в других системах.

Greenstone является наиболее простой из рассмотренных систем. Очевидно, что простота использования является решающей при выборе данной системы, но простота реализации влечет за собой функциональную ограниченность. С другой стороны, это единственная система, позволяющая организовывать электронные коллекции на сменных носителях.

ELSA – это на текущий момент единственный российский аналог подобных систем, учитывающий российские тенденции развития электронных библиотек. Предлагаемые возможности импорта описаний из АБИС уникальны, однако существует возможность импорта только из АБИС «Руслан». Также отсутствует возможность выгрузки в xml, что ограничивает интероперабельность системы.

Также необходимо заметить, что все рассмотренные нами системы основываются на схеме метаданных Dublin Core, причем эта схема является первичной и создается всегда, часто автоматически. Конечно, в части систем существует возможность хранить и другой тип метаданных, но это обычно опционально и трактуется как своего рода «дополнительная глава» к основному ресурсу. Если рассматривать научную составляющую ЭБ, причем

именно хранилище электронных аналогов печатных изданий, то подобный подход не может вызвать каких-либо нареканий, схема DC здесь безупречна. Подобные системы вполне успешно используются для хранения трудов ученых, используются как источник информационных ресурсов. Однако, если подходить к научной ЭБ как к исследовательскому полигону, то встает вопрос об успешной интеграции различных специфических сервисов. Здесь, на наш взгляд, нельзя сделать однозначного вывода об использовании уже существующих систем или о создании собственной. С одной стороны, написание аналога уже работающих систем ЭБ с похожим функционалом не может представлять какого-либо интереса, с другой стороны интеграция различных сервисов гораздо проще в программном продукте, в котором изначально предполагалось наличие подобных сервисов. Здесь, на наш взгляд, окончательное решение остается за конкретным коллективом разработчиков и зависит напрямую от квалификации этого коллектива.

Таким образом, для научной составляющей вузовской ЭБ как хранилища электронных версий печатных изданий, существующие решения могут быть приемлемыми, тем более что большинство из них для этого и предназначались. В случае же с образовательными электронными ресурсами этого явно недостаточно. Метаданные образовательных ресурсов все же логичнее создавать по схеме, специально для этого предназначенной (LOM, RusLOM), Структура большого ЭОР как электронного документа гораздо более сложна, и требует, как уже отмечалось, продуманной навигации как внутри самого ЭОР, так и по ЭБ в целом. Если в научной ЭБ, текст отдельного информационного ресурса (напр., статьи или автореферата), как правило, интересен читателю полностью, то пользователю образовательной ЭБ обычно необходимо всего лишь несколько глав из большого электронного учебника, которые описывают интересные ему факты. Большие ЭОР используются как справочники, поскольку ни один студент при подготовке к экзамену не будет читать большой электронный учебник целиком, из которого ему интересны только несколько параграфов. Т. е. требования к навигации по образовательным ресурсам гораздо более жесткие.

Учитывая вышеприведенные требования к образовательной составляющей вузовской ЭБ можно однозначно сделать вывод о непригодности существующих систем ЭБ для вузовской научно-образовательной электронной библиотеки.

Примеры успешного развертывания подобных систем - электронных архивов конечно существуют в Интернете, но это, в основном, архивы трудов научных работников, т. е. научные ЭБ. В этом случае, использование свободно распространяемого ПО вполне оправдано.

Также на основе проведенного анализа можно утверждать, что ни одна из существующих разработок не может удовлетворить наши потребности из-

за невозможности интеграции с существующими программными системами КГУ. Более того, иностранные аналоги не учитывают российскую специфику развития ЭБ, что влияет на структуру электронного документа, предоставляемые сервисы, поддерживаемые протоколы обмена информацией и, разумеется, усложнит процесс внедрения подобных систем. В случае образовательной составляющей ЭБ адаптация существующих систем может сопоставимой с разработкой собственной системы.

На данный момент начата разработка собственного программного обеспечения для ЭБ, на основе выработанных критериев разрабатывается техническое задание системы, создаются прототипы электронных коллекций, программное обеспечение для электронных коллекций, пробные сервисы ЭБ.

На первом этапе принято решение хранить подобные коллекции средствами существующей АБИС, ведется разработка поискового интерфейса с возможностью наложения различных фильтров для поиска материалов по кафедрам, факультетам, дисциплинам и специальностям. Поскольку АБИС является Z39.50 сервером, то фактически разрабатывается сервис формирования поискового запроса type-1 к системе.

Литература

- [1] Армс В. Электронные библиотеки [текст] /пер. с англ. – М.: ПИК ВИНТИ, 2001. – 276 с.
- [2] Зуев Д.С. Программная оболочка для интеграции электронных коллекций Казанского университета [Электронный ресурс]// Информационные технологии, компьютерные системы и издательская продукция для библиотек: материалы конф. «Libcom-2007». – электрон. текстовые дан. – М.: ГПНТБ России, 2007, тез. докладов. – 2007 – 1 электрон. опт. диск (CD-ROM). – Загл. с этикетки диска. – ISBN 5-85638-100-9. – № гос. регистрации 0320501386. – Сведения доступны также по Интернет: <http://www.gpntb.ru/libcom7/disk/12.pdf>
- [3] Абросимов А.Г. Метаданные описания коллекции периодической печати [Электронный ресурс] // Электронные библиотеки: рос. науч. электронный журн. – 2005. – Т. 8, вып. 2. – режим доступа: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal2005/part2/Abrosimov>, свободный
- [4] Абросимов А.Г., Зуев Д.С., Принцип построения программного обеспечения электронной коллекции периодической печати //Актуальные проблемы современной науки: Труды 3-го Международного форума (8-й Международной конференции молодых учёных и студентов). Естественные науки. Ч 1, 2: Математика. Математическое моделирование, Самара: Изд-во СамГТУ, 2007. – с. 78–83
- [5] Коголовский М. Р. Особенности научных электронных библиотек . Тезисы докладов научной конференции, посвященной 10-летию РФФИ

«Электронные библиотеки и информационное обеспечение научной деятельности», Москва, 25–26 ноября 2002 г.

- [6] Когаловский М.Р. Научные коллекции информационных ресурсов в электронных библиотеках. труды 1-й Всерос. науч. конф., Санкт-Петербург, 19–21 октября 1999. – СПб., 1999. – С : б.н. http://ict.edu.ru/ft/002340/scc_coll.pdf.
- [7] Антопольский А.Б., Вигурский К.В. Концепция электронных библиотек. Электронные библиотеки: рос. науч. электронный журн. – 1999. – Т. 2, вып. 2. <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/1999/part2/antopol>.

Scientifical and Educational Digital Libraries in institutes of higher education

A.G. Abrosimov, D.S. Zuev

We look at functional properties of scientifical and educational digital libraries especially in institutes of higher education.

* Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (проект 07-01-12146).

Создание электронной коллекции личных дел студентов Казанского университета (1917–1925 гг.)^{*}

© Г.Г. Амалиева, А.М. Елизаров, А.А. Сальникова

Казанский государственный университет
amelizarov@gmail.com

Аннотация

Работа посвящена созданию электронной коллекции личных дел студентов Казанского университета (КГУ) за 1917–1925 гг. Разработана компьютерная модель личного дела студента КГУ как электронный аналог нарративного материала. Данная коллекция является составной частью Электронной библиотеки КГУ.

1 Введение

Как известно, одной из черт происходящего сейчас перехода от индустриального общества к информационному является то, что информация и знание становятся одним из основных факторов развития и структурирующим социальным фактором, под воздействием которого меняются формы экономической деятельности и социальные взаимоотношения. Сохранение, развитие и рациональное использование этого стратегического ресурса будущего имеют огромное значение для любого общества и государства. Отличительной чертой сегодняшнего этапа развития человечества является представление информации и знаний не только в традиционной печатной, но и в электронной, цифровой форме, что позволяет принципиально иначе создавать, хранить, организовывать доступ и использовать информацию в любой форме. К тому же современные информационно-телекоммуникационные технологии привели к тому, что большое число современных информационных ресурсов сразу создается в электронном виде, т. е. формируются электронные библиотеки (см., например, [1]).

В условиях формирования информационного общества чрезвычайно важным инструментом устойчивого экономического и социального развития является обеспечение публичного (в том числе удаленного) доступа к социально значимой информации, в первую очередь научного, образовательного и культурного характера. Нельзя не признать, что

степень доступности информационных ресурсов России, имеющих огромную ценность и столь же огромные объемы, до сих пор остается слишком низкой. Осознание этого привело к необходимости переосмысления роли, задач и методов работы традиционных центров хранения и распространения информации. Стала очевидной необходимость эффективной кооперации всех учреждений, занимающихся созданием и распространением информации и знания – научных учреждений, информационных центров, библиотек, архивов, музеев и т. д. Все это привело к появлению нового, в определенной степени синтетического направления – электронные библиотеки, объединяющего специалистов в области информационных технологий, библиотекарей, информационных работников, работников музеев и архивов, издателей, теле- и радиовещателей и многих других.

Сегодня в разработках электронных библиотек, обеспечении их поддержки и доступа к ним востребован весь спектр ключевых технологий управления информацией, которые используются в современных информационных системах. Поэтому тенденции развития информационных технологий, сформировавшиеся в последние годы, весьма существенно влияют на функциональные возможности электронных библиотек.

2 Развитие в КГУ технологий формирования и семантически ориентированной разметки данных коллекций электрон-ных документов

2.1 Интегрированный программный комплекс для формирования коллекций электронных документов, генерации метаданных и XML/RDF-разметки

В КГУ в течение последнего десятилетия совместными усилиями ряда структурных подразделений университета, прежде всего, Научной библиотеки им. Н. И. Лобачевского – НБЛ, ведутся интенсивные работы по созданию информационных ресурсов в самых различных областях знания. Важную роль в этом играет НБЛ, на базе которой создан ряд электронных коллекций, существенно отличающихся как по предметным областям, так и по технологиям формирования баз данных. В этом смысле имею-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

щиеся информационные ресурсы достаточно разобщены. Несмотря на то, что доступ к большинству из них осуществляется через сайт КГУ, не существовало единой информационной системы, позволяющей получить представление обо всем богатстве информационного пространства КГУ и обеспечить доступ к нему. Доступ к коллекциям осуществляется или через электронный каталог НБЛ (библиографическое описание документов содержит ссылку на их электронные версии), или с сайта НБЛ по веб-интерфейсу. Например, для коллекции периодической печати 19 – начала 20 веков разработано оригинальное программное обеспечение, учитывающее специфику данной коллекции и ориентированное на использование XML-технологий.

В последние годы в КГУ создается интегрированный программный комплекс, нацеленный на развитие технологий формирования и семантически ориентированной разметки данных коллекций электронных документов (ЭД); разработку единой системы навигации, поиск и обеспечение доступа к электронным коллекциям КГУ и других производителей, к которым университет имеет доступ – Корпоративная библиотечная сеть г. Казани, АРБИКОН, НЭБ e-library.ru и др.; на организацию с помощью этого комплекса широкого доступа к научным электронным коллекциям и библиотекам в области гуманитарного знания, создаваемым в КГУ, как читателей НБЛ, так и внешних пользователей, с целью проведения фундаментальных научно-исследовательских работ; на организацию с помощью этого комплекса совместного доступа к научным электронным коллекциям и библиотекам КГУ и других ведущих производителей электронных ресурсов через единую точку входа; на создание новых форм обмена научной информацией и формирование новых видов информационных ресурсов в области гуманитарного знания (см., например, [2 – 6], а также <http://ruslan.ksu.ru>, http://Z3950.ksu.ru/phil_ont/index.htm, <http://isl.ksu.ru/i83.htm>, <http://jenny.ksu.ru/cgi-bin/zgate>).

Названный программный комплекс осуществляет следующие функции:

- формирование коллекций и описаний ЭД;
- обеспечение доступа по http-интерфейсу и протоколу Z39.50;
- обеспечение работы с ЭД;
- атрибутивный поиск по метаданным с использованием индексных баз;
- формирование архивов ЭД;
- обеспечение доступа других производителей через единую точку входа.

Программный комплекс включает следующие основные информационные системы.

Система формирования метаданных коллекций и ЭД. Каждая коллекция ЭД должна иметь свой набор метаданных, содержащий описание:

- коллекции, включающее описание структуры и количественные характеристики (количество разделов, ЭД и т. д.);

- раздела коллекции, индивидуальное для каждой коллекции;

- электронного документа, основанное на стандартном библиографическом описании.

Система поиска, навигации и обеспечения доступа к электронным ресурсам КГУ и других организаций – обеспечивает:

- разработку web-сервисов для поиска и навигации, использующих несколько распределенных электронных ресурсов;

- навигацию и доступ к электронным ресурсам КГУ и других организаций с единой точкой входа.

Система обеспечения работы с ЭД. В настоящий момент наиболее актуальными являются три типа ЭД:

- полностью текстовый ЭД (наиболее распространенные форматы HTML, Word);

- ЭД, содержащий только образы страниц, полученных в результате сканирования;

- смешанные ЭД, содержащие и образы страниц, и распознанный текст.

Система атрибутивного поиска по метаданным с использованием индексных баз. Используются два уровня поиска по метаданным:

- поиск по ЭБ с использованием метаданных, общих для всех ЭД, таких, как автор, название;

- расширенный поиск по конкретной коллекции с использованием метаданных, отражающих специфику ЭД, включенных в коллекцию.

Система формирования архива документов – предоставляет возможности компактного размещения массивов ЭД с учетом будущего расширения и гарантирует их сохранность.

Существенный вклад в информационное наполнение программного комплекса дают исторические коллекции.

2.2 Историческая коллекция

Проводится работа по созданию и наполнению электронной библиотеки КГУ эго-документами, раскрывающими историю университетской культуры в Казани. Научная значимость такой работы обусловлена тем обстоятельством, что в России до сих пор нет единой программы изучения истории российских университетов. Архивы старейших классических университетов страны находятся зачастую в тяжелом состоянии: часть из них погибла в пожарах и войнах, часть утрачена, в советское время резко сузился объем комплектования. Казанский университет в этом отношении представляет уникальное явление. Его архив сохранился в целостности и насчитывает 17 тысяч единиц хранения, отложившихся в Национальном архиве Республики Татарстан. В отделе рукописей и редких книг НБЛ имеется рукописная коллекция мемуаров и писем. Часть этого наследия хранится в музее истории КГУ, Объединенном музее Республики Татарстан. Вся эта коллекция позволяет проводить исследования за императорский период истории университета, но ее со-

стояние и трудности доступа к ней внушают опасение и желание сохранить ее и сделать доступной для современных и последующих поколений исследователей. Что же касается истории университета за советский период его существования, то в имеющемся комплексе документов фактически отсутствуют материалы личного происхождения (письма, мемуары, дневники). Между тем, подготовка 200-летнего юбилея Казанского университета (2004) показала, что множество таких документов хранится в личных коллекциях ныне здравствующих свидетелей и участников событий (см. также [7–9]). Создание электронной версии позволит ввести их в научный оборот и тем самым значительно обогатить существующее представление о советской академической повседневности. Эти свидетельства и фотографии – бесценный источник для изучения противоречивой действительности университетского человека советского времени. В отличие от Московского и Ленинградского университетов, в Казани не была издана серия исторических документов и свидетельств. В результате у историков отсутствует необходимая документальная основа для изучения провинциальной университетской жизни.

2.3 Характеристика студенческих личных дел

Источниковый комплекс личных дел студентов Казанского университета 1917 – 1925 гг., хранящийся в НБЛ КГУ, включает 9435 единиц хранения в составе фонда 22. Личное дело (досье) представляет собой совокупность документов, содержащих сведения о работнике [10]. Личное дело начинало формироваться с момента подачи студентом заявления в приемную комиссию, а вновь поступающие документы подшивались уже в старое дело. Процесс формирования дела продолжался на протяжении всего срока студенческого обучения. Личные дела студентов хранятся в папках, все документы подшиты к перегибу левой стороны листа. Личные дела хранятся в алфавитном порядке по фамилиям студентов.

2.4 Выбор информационной среды

Учитывая специфику представления данных для реализации проекта, в качестве инструментария для создаваемой информационной системы используются следующие продукты:

- Microsoft Access 2003 – гибкий в использовании программный пакет, соединяющий широкие возможности обработки текста и легкую в использовании систему организации баз данных. К преимуществам программы относится открытый характер базы, т. е. возможность в будущем его пополнения новыми сведениями и возможность построения на основе главного банка данных иных форм и конструкций для различных видов обработки информации, проведения статистического анализа. Все данные хранятся в нескольких связанных между собой таблицах (реляционная база данных);

- MySQL – система управления реляционными базами данных, которая применяет язык SQL (Structured Query Language – язык структурированных запросов), используемый по всему миру стандартный язык запросов к базам данных.

Первоначально база данных ведется в формате Microsoft Access 2003, а затем при размещении на сайте КГУ она конвертируется в формат MySQL.

2.5 Текущее состояние проекта

Личные дела студентов Казанского университета – исторические документы, требующие специализированного описания, метаданных. Этот вид исторических источников обладает рядом особенностей: нарративный характер сведений, неоднородность состава личных дел по объему и видам документов, противоречивость сведений документов разных видов, официально – личностный характер. При видовом разнообразии материала личных дел есть, тем не менее, возможность выделить основную форму анкеты, которая и легла в основу сформированного массива данных. Поскольку личные дела не были полностью идентичными, понадобилась дополнительная унификация информации с помощью специально выработанного вопросника (унифицированная анкета). Вопросник заполнялся на каждого студента по его личному делу.

На данный момент времени коллекция студенческих личных дел содержит сведения 500 единиц хранения. Также в базе данных можно осуществить поиск и подсчет по таким параметрам, как пол, национальность, социальное происхождение и др. В будущем коллекция будет пополнена новой информацией.

Литература

- [1] Армс В. Электронные библиотеки : пер. с англ. – М. : ПИК ВИНТИ, 2001. – 276 с.
- [2] Абросимов А.Г. Выбор программного обеспечения электронных библиотек // 11-я Международная конференция и выставка Libcom-2007 «Информационные технологии, компьютерные системы и издательская продукция для библиотек»: докл. и тез. докл. – М. : ГПНТБ России, 2007, электронный ресурс. – <http://www.gpntb.ru/libcom7/disk/29.pdf>.
- [3] Абросимов А.Г., Елизаров А.М., Федоров А.О., Халиков Л.И. Электронные библиотеки как технология интеграции электронных коллекций и издания электронных журналов, раздел 1.4. коллективной монографии Информационная инфраструктура гуманитарного вуза: качество образования и интеграция в мировое информационное пространство. – М. : ВИНТИ, 2007. – С. 49–100.
- [4] Абросимов А.Г., Зуев Д.С. Принцип построения программного обеспечения электронной коллекции периодической печати // Актуальные проблемы современной науки : Труды 3-го Международного форума (8-й международной

конференции молодых ученых и студентов). Естественные науки. Ч. 1, 2: Математика. Математическое моделирование. – Самара : Изд-во СамГТУ, 2007. – С. 78–83.

- [5] Биряльцев Е.В., Гусенков А.М., Елизаров А.М. О доступе к электронным коллекциям в виде реляционных баз данных на основе онтологий // Труды 9-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 15–18 октября 2007 г. – Переславль-Залесский : Изд-во «Университет города Переславля», 2007. – С. 211–216.
- [6] Зуев Д.С. Программная оболочка для интеграции электронных коллекций Казанского университета // 11-я Международная Конференция и Выставка Libcom-2007 «Информационные технологии, компьютерные системы и издательская продукция для библиотек»: докл. и тез. докл. – М. : ГПНТБ России, 2007, электронный ресурс. – <http://www.gpntb.ru/libcom7/disk/12.pdf>.
- [7] Вишленкова Е.А., Малышева С.Ю., Сальникова А.А. Университет как память и как знание // Диалог со временем: Альманах интеллектуальной истории. Т. 18. – М. : ЛКИ, 2007. – С. 35–67.
- [8] Малышева С.Ю., Сальникова А.А. Старый мир при новой власти: Казанский университет в советском политическом пространстве // Мир Клио : сб. статей : в 2 т. – Т.2. – М. : ИВИ РАН, 2007. – С. 97–114.
- [9] Сальникова А.А. Символы и образы советского университетского пространства и их специфика // Линия судьбы : Сб. статей, очерков, эссе. – М. : Собрание, 2007. – С. 349–358.
- [10] Источниковедение новейшей истории России: теория, методология, практика / А.К. Соколов, Ю.П. Бокарев, Л.В. Борисова и др. ; под ред. А.К. Соколова. – М. : Высш. шк., 2004. – 687 с.

Construction of the digital person files collection of the students of Kazan University (1917–1925)

G.G. Amalieva, A.M. Elizarov, A.A. Salnikova

This work is devoted to problem, which connect with digital person files collection of the students of Kazan University (KSU) in 1917–1925. The computer model of KSU student person files was developed like an electronic copy of the narrative material. This collection is the part of the total KSU digital library.

* Работа поддержана грантом РГНФ (проект № 07–01–12146)

Роль традиционных библиотек в формировании проблемно-ориентированных информационно-коммуникационных социальных пространств

© Н.А. Мешков

Л.П. Сорокина

Н.В. Цветкова

Централизованная библиотечная система «Люблино» г. Москвы
biblion@list.ru

Аннотация

Раскрывается роль традиционных библиотек в формировании проблемно-ориентированных информационно-коммуникационных социальных пространств.

Возникновение библиотечного дела как отрасли информационной, культурно-просветительской и образовательной деятельности мировая история связывает с именами Аристотеля и Птолемея и с основанием Птолемеем в начале 3 в. до н.э. Александрийской библиотеки (весьма показательно, что уже тогда при Александрийской библиотеке работал штат копиистов, занимавшихся перепиской, т.е., фактически, тиражированием, книг). Современная публичная библиотека уже не является, как это было изначально, простым хранилищем книг, или, как сказано в Большой Советской Энциклопедии, «культурно-просветительным и научно-вспомогательным учреждением, организующим общественное пользование произведениями печати». Согласно Федеральному Закону РФ «О библиотечном деле», библиотека – это «информационное, культурное, образовательное учреждение, располагающее организованным фондом тиражированных документов и предоставляющее их во временное пользование физическим и юридическим лицам». В данном контексте под документом понимается материальный объект с зафиксированной на нем информацией в виде текста, звукозаписи или изображения, предназначенный для передачи во времени и пространстве в целях хранения и общественного использования. Таким образом, в состав библиотечных фондов могут сегодня входить не только произведения печати, но и другие носители информации, и следующим естественным шагом в развитии публичных библиотек должен стать переход от предоставления во временное пользование физическим и юридическим лицам документов как материальных объектов с зафиксированной на них информацией к непосред-

ственному предоставлению всем желающим свободного доступа к этой информации. Ведущей тенденцией мирового цивилизационного процесса является в наши дни формирование глобального информационного общества, характеризующегося «высоким уровнем развития информационных телекоммуникационных технологий и их интенсивным использованием гражданами, бизнесом и органами государственной власти» [11]. В глобальном информационном обществе обмен информацией не имеет ни временных, ни пространственных, ни политических границ. Главным условием благополучия каждого человека, каждой организации и каждого государства в глобальном информационном обществе становится знание, полученное благодаря беспрепятственному доступу к информации и умению работать с ней. «Богатство, власть, общественное благополучие и культурное творчество в России в XXI веке во многом будут зависеть от ее способности развить модель информационного общества, приспособленную к ее специфическим ценностям и целям» (М. Кастельс, 2004 г.) [4].

В преамбуле утвержденной Президентом России 7 февраля 2008 года «Стратегии развития информационного общества в российской федерации» сформулирована задача «уже в среднесрочной перспективе реализовать имеющийся культурный, образовательный и научно-технологический потенциал страны и обеспечить Российской Федерации достойное место среди лидеров глобального информационного общества». Решение этой задачи требует совместных целенаправленных усилий органов государственной власти, бизнеса и гражданского общества.

К числу основных путей достижения цели формирования и развития информационного общества в Российской Федерации – повышения качества жизни граждан, обеспечения конкурентоспособности России, развития экономической, социально-политической, культурной и духовной сфер жизни общества, совершенствования системы государственного управления на основе использования информационных и телекоммуникационных технологий, относится сохранение культуры многонационального народа Российской Федерации, укрепление нравственных и патриотических принципов в общественном сознании, развитие системы культурного и гуманитарного просвещения. На основе передовых

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

информационно-коммуникационных технологий должны быть созданы принципиально новые социальные технологии как совокупности методов и средств решения социальных проблем, направленных на формирование условий жизни и развития общества, общественных отношений, социальной структуры с целью обеспечения потребностей человека, создания условий для реализации его способностей и интересов с учетом одобряемой обществом системы ценностей и взаимозависимости между общественным прогрессом и экономическим развитием.

Основой глобальной компьютерной коммуникации является Интернет. Интернет предлагает человечеству новое качество жизни, основанное на интеграции и активной эксплуатации мировых информационных ресурсов. В отличие от средств массовой информации, воздействующих на человека, Интернет организует взаимодействие людей. Миссия Интернет заключается в создании благоприятных условий для открытого общения каждого с каждым и каждого со всеми. К ноябрю 2006 г. количество сайтов в Интернет превысило 100 млн, в Рунет – российском сегменте Интернет – 500 тыс. Общее количество пользователей Интернет в России, по некоторым оценкам, достигает сегодня 40 млн чел. (наиболее активно осваивают Рунет учащаяся молодежь и пенсионеры – основные посетители публичных библиотек).

Современный массовый читатель предпочитает чтению для удовольствия, целью которого является удовлетворение духовных потребностей, чтение для работы – ознакомление, изучение, использование документов с целью получения информации. И ему, в принципе, не важно, в каком виде (как бумажный документ, как файл, записанный на CD- или DVD-диске или «скачанный» из Интернет) он эту информацию получит. И также ему не важно, с какого именно сайта (своей районной библиотеки, Российской Государственной библиотеки им. В.И. Ленина, любой другой российской библиотеки, а равно и любой другой организации или частного лица) он в каждом конкретном случае будет «скачивать» нужные ему файлы. Главное – чтобы это можно было сделать быстро и, в идеале, бесплатно. Таким образом, становится возможным возложить обязанности по обеспечению основной массы россиян учебной, рабочей и справочной информацией на крупные специализированные федеральные библиотеки. Районные же библиотеки будут вынуждены искать тогда другие способы выполнения своих социальных функций, внедрять другие социальные технологии.

К числу наиболее перспективных социотехнических методов построения информационного общества относится метод «открытого контента» – генерации и распространения информации и знаний (результатов индивидуальной и коллективной интеллектуальной деятельности) как общественного достояния через Интернет [5]. «Открытый контент» создается в процессе реализации открытых Интер-

нет-проектов, связанных с созданием и поддержкой электронных библиотек, энциклопедий, веб-журналов, открытых книг и сетевой литературы, банков идей и научных публикаций, дискуссионных форумов и форумов личного опыта, открытых образовательных и коммуникативных сред. Степень эффективности метода «открытого контента» как формы социальной организации в информационном обществе определяется уровнем кооперации творчески мыслящих индивидов на основе общих внеэкономических ценностей.

Ускоренными темпами развиваются в последние годы в России информационно-коммуникационные социальные сети, объединяющие множество акторов (агентов), которые могут вступать во взаимодействие друг с другом и связи между которыми преимущественно являются социальными (дружественные отношения, совместная работа, обмен информацией) [10]. Самый показательный на сегодняшний день пример социальной сети в Рунет – сеть Одноклассники.РУ, число участников которой уже в первом квартале 2008 года превысило 10 млн чел. Информационно-коммуникационные социальные сети служат основой для формирования проблемно-ориентированных информационно-коммуникационных социальных пространств.

Участвуя в создании социальных сетей и формировании на базе Рунет проблемно-ориентированных информационно-коммуникационных социальных пространств, российские библиотеки должны внести весомый вклад в развитие в стране информационного общества. Этическая цель традиционных библиотек в информационном обществе заключается в увеличении объема информации, являющейся общественным достоянием. «Информационным сетям следует быть открытыми для контента из любых источников, поощряя все заинтересованные стороны создавать контент, вместо того, чтобы только потреблять его. Всем действующим лицам в информационном обществе следует стремиться поднять каждого участника на тот уровень, где он поймет, как работает система и как он может действовать коллективно со всеми, разделяя ответственность за успех системы в целом. Открытое, интегрированное и межкультурное образование, совмещенное с обучением навыкам информационного и коммуникационного управления, является решающим» [1].

Одна из основных возможностей, предоставляемых информационным обществом, – качественное улучшение системы охраны здоровья людей. Новые информационно-коммуникационные технологии повышают эффективность пропаганды здорового образа жизни, делают широко доступной профилактическую медицину, создают основу для получения любым пациентом, где бы он ни находился, регулярных врачебных консультаций, превращают в реальность телемедицину, опирающуюся на национальные и мировые медико-информационные ресурсы. Электронная коммерция помогает органам здравоохранения, всем субъектам медико-

производственной деятельности оптимизировать закупки медицинского оборудования, расходных материалов, медикаментов, а гражданам – улучшить условия приобретения лекарственных средств и других изделий медицинского назначения.

«Сохранение и укрепление здоровья народа достойно возведения в ранг национальной идеи» (Г.Н. Голухов, Н.А. Мешков, Ю.В. Шиленко, 1998 г.) [2]. В XXI веке именно здоровье народа во всех его проявлениях будет как целью, так и основной движущей силой общественного прогресса, обретет статус главной социальной ценности, ключевого фактора национальной безопасности, обобщающего критерия эффективности управления государством. Воспринятая общественным сознанием, идея здоровья должна стать в России реальной консолидирующей и мобилизующей силой. Только физически и нравственно здоровый народ в состоянии адекватно воспринимать и оценивать жизнеутверждающие принципы патриотизма, соборности и коллективизма, добра и справедливости, взаимоуважения и терпимости, мира и согласия, знания и веры, мечты и идеала. Только он может сделать их нормой повседневной жизни.

В условиях российской действительности особо важную роль играет сформулированный И.А. Гундаровым закон «духовно-демографической детерминации здоровья»: при прочих равных условиях улучшение (ухудшение) нравственно-эмоционального состояния общества сопровождается улучшением (ухудшением) демографической ситуации. Главным показателем эффективности проводимых в стране реформ становится качество жизни россиян, измеряемое с помощью комплексного индикатора, состоящего из следующих четырех блоков [3]:

- здоровье человека, его трудоспособность и долголетие, определяемые по продолжительности активной творческой жизни;
- способность иметь детей и продолжать род, оцениваемая по уровню естественного воспроизводства населения;
- радость жизни, удовлетворенность индивидуальными условиями бытия и положением дел в государстве, определяемые с помощью социологических опросов;
- духовное состояние, при анализе динамики которого в качестве единиц измерения используются нравственные аномалии – убийства, грабежи, разводы, брошенные дети.

Активизация процессов рыночного реформирования и развития российского здравоохранения в условиях демократизации общественных потребностей в свободном доступе к информации, имеющей отношение к здоровью людей, медицине, фармацевтике во всех аспектах медико-производственной деятельности обуславливает необходимость широкого и всестороннего обсуждения проблем развития отечественного медико-производственного комплекса (МПК) в новых, коренным образом изменивших-

ся социально-экономических условиях. Абстрагируясь от функционального и экономико-правового статуса участников здравоохранительной деятельности, под МПК мы понимаем множество взаимосвязанных и взаимодействующих между собой субъектов, реализующих в процессе своего согласованного функционирования одну общую цель – обеспечение валео-демографической безопасности Российской Федерации на основе повышения эффективности и качества медико-социальной помощи населению страны, сохранения и укрепления физического и психического здоровья россиян, максимального продления активной творческой жизни и гармоничного развития личности [6]. Единый медико-производственный комплекс страны призван обеспечивать единство всех воспроизводственных процессов охраны и укрепления здоровья россиян, предоставление гражданам доступной квалифицированной медико-социальной помощи, решение текущих и перспективных задач реализации медико-демографической политики государства.

Особый отпечаток на суть и характер задач, стоящих сегодня перед российским здравоохранением, накладывает изменение отношения общества к отдельным его членам, признание приоритетности прав личности. Если раньше отношения «врач – больной» строились на сугубо «патерналистской» основе (врач лечил, а больной беспрекословно выполнял все назначенные ему предписания, лишь иногда доводя до сведения врача свои жалобы), то теперь, в соответствии с принципом информированного согласия пациента на медицинское вмешательство, больной – это уже не просто пациент, запуганный непонятными ему действиями, указаниями и терминологией врача, а равноправный партнер врача, знающий все о своей болезни, об альтернативных способах ее лечения и их возможных результатах, о рисках, возникающих при применении каждого из них, о последствиях того или иного решения и имеющий право знакомиться со своей медицинской документацией, соглашаться или не соглашаться с предлагаемым ему лечением, менять по своему усмотрению лечащего врача, подавать иски по возмещению причиненного ему в процессе лечения физического, материального и морального вреда (нравственных страданий). Только признавая автономию воли пациента и культивируя уважение к его личности, можно в полном объеме реализовать творческий потенциал врача, привлечь больного к активному участию в профилактических, лечебно-диагностических и реабилитационных процессах.

Необходимым условием высокой эффективности информационной поддержки медико-производственной деятельности и пациентов является формирование на базе Интернет единого общероссийского информационно-коммуникационного медико-производственного пространства (ИКМПП) [12]. В ИКМПП должны быть представлены все виды лечебно-профилактических организаций, врачи всех специальностей, производители и продавцы

лекарственных средств и других изделий медицинского назначения, страховые медицинские организации (обязательное и добровольное медицинское страхование), адвокаты, специализирующиеся на защите прав пациентов и врачей, научно-исследовательские институты (разработчики лечебно-диагностических технологий, лекарственных средств, материалов, медицинского и другого оборудования) и учебные заведения (основное и дополнительное последипломное образование), профессиональные медицинские и фармацевтические ассоциации, общественные объединения пациентов, редакции и издательства (научная и научно-популярная медицинская, экономическая и юридическая литература) и др. Главным результатом выхода производителей медицинских услуг и товаров медицинского назначения в Интернет будет улучшение качества и увеличение объема медицинской помощи населению страны, повышение социально-экономической эффективности функционирования российского МПК.

С расширением присутствия субъектов МПК в Интернете ускоряется процесс социальной институционализации ИКМПП – организации и координации социального взаимодействия россиян в ИКМПП, появления его стандартизированных и регулярно воспроизводимых элементов. Социальная институционализация ИКМПП возникает и развивается на основе следующих четырех основных предпосылок:

- формирования в условиях глобализации и демократизации общественных потребностей в свободном доступе к информации, имеющей отношение к здоровью людей, медицине, фармацевтике, валео-демографической безопасности России во всех аспектах медико-производственной деятельности;
- формирования и развития необходимых организационных структур (каталогов медико-информационных ресурсов, поисковых систем, специализированных порталов и сайтов), а также связанных с ними ценностных стандартов и социальных норм, регуляторов поведения человека и социальных групп в ИКМПП;
- возникновения социальных условий и возможностей, соответствующих социализации человека, интернационализации новых ценностей и норм, способных формировать необходимую систему потребностей личности, ценностных ориентаций и ожиданий в области охраны здоровья и продолжения рода;
- интеграции ИКМПП в структуру российского МПК, нуждающегося в связи с ухудшением валео-демографической ситуации в стране в выполнении новых социально-значимых функций в качестве условия своего дальнейшего развития.

Активизируется процесс формирования валеологического Интернет-сообщества – объединения людей, имеющих общие интересы, общую цель – сохранение и укрепление здоровья россиян, обеспечение валео-демографической безопасности России.

Процесс социальной институционализации ИКМПП должен завершиться принятием набора формальных и неформальных санкций (правовых и моральных), с помощью которых общество будет осуществлять социальный контроль за различными типами поведения (действий) граждан и организаций в этой области Интернет.

Темпы роста ИКМПП и сила его влияния на ситуацию в российском здравоохранении определяются уровнем развития его инфраструктуры и количеством и качеством представленных в нем медико-информационных ресурсов, прежде всего – медико-социологической информации, которая должна помочь нашему обществу сконструировать «паттерн здорового образа жизни» (А.В. Решетников, 2003 г.) [9]. Дополнительное ускорение развитию ИКМПП придает повышение образовательного уровня россиян – через пропаганду здорового образа жизни и формирование научно-обоснованных потребностей в медицинских услугах и товарах медицинского назначения и расширение возможностей реализации этих потребностей.

Основными целями участия традиционных библиотек и созданных на их базе электронных библиотек в формировании и развитии ИКМПП являются:

- улучшение валео-демографической ситуации в соответствующих административно-территориальных образованиях и в стране в целом;
- повышение доступности и качества медицинской помощи, лекарственных средств и других изделий медицинского назначения;
- повышение эффективности управления государственным и муниципальным здравоохранением.

Для достижения этих целей традиционные и электронные библиотеки должны активно включиться в решение следующих основных задач:

- содействие формированию и развитию ИКМПП и интеграции его в структуру российского МПК;
- обеспечение прав граждан и организаций на свободный поиск, получение, передачу, производство и распространение информации, имеющей отношение к здоровью людей, медицине, фармацевтике, валео-демографической безопасности России во всех аспектах медико-производственной деятельности;
- формирование соответствующей стратегическим целям развития России системы потребностей личности, ценностных ориентаций и ожиданий в области охраны здоровья и продолжения рода, пропаганда здорового образа жизни и семейных ценностей;
- осуществление мониторинга качества жизни населения соответствующих административно-территориальных образований, измеряемого с помощью комплексного индикатора, предложенного И.А. Гундаровым;
- привлечение внимания политических элит, предпринимательских кругов и населения к проблеме обеспечения валео-демографической безопасно-

сти России, расширение возможностей для осуществления общественного контроля за ходом реализации приоритетного национального проекта «Здоровье»;

- пропаганда результатов научных исследований и практического опыта применения Интернет-технологий в здравоохранении.

Не секрет, что на первом этапе реализации приоритетного национального проекта «Здоровье» Министерства здравоохранения и социального развития России неожиданно для себя столкнулось с проблемой непонимания и, нередко, неприятия медицинским сообществом и населением проводимых им мероприятий. Одна из основных причин этого – низкий уровень информационной поддержки проекта. Признание этого факта содержится, в частности, в цитате из стенограммы выступления Министра здравоохранения М.Ю. Зурабова на VI Пироговском съезде 28 сентября 2006 года: «Мы осознаем, что информирование медицинского сообщества о сути тех мероприятий, которые проводятся в эти годы, и о плане, который намечен на более продолжительный период, у нас страдает».

При всем многообразии имеющихся сегодня в Рунете медицинских порталов по-прежнему актуальной остается задача создания единого общероссийского информационно-аналитического портала «Демографическая безопасность России» – основанной на последних достижениях Интернет-технологий глобальной саморазвивающейся проблемно-ориентированной информационно-коммуникационной системы, высокоэффективного средства массовой информации, агитации и пропаганды, инструмента общественного контроля за деятельностью органов государственной власти и местного самоуправления, федеральных и муниципальных учреждений здравоохранения, других субъектов МПК по реализации приоритетного национального проекта «Здоровье» [8]. Важнейшей общественно-политической задачей портала "Демографическая безопасность России" является содействие формированию у россиян нового валеологического мышления на принципах здорового образа жизни и семейных ценностей. Главная научно-практическая задача портала «Демографическая безопасность России» заключается в содействии формированию и развитию ИКМПП, накоплении, хранении и актуализации информации по вопросам личного и общественного здоровья, создании уникальной по содержанию и объему базы данных по проблемам валео-демографической безопасности России.

Портал «Демографическая безопасность России» представляет собой «виртуальную организацию», предоставляющую возможность рассредоточенной и децентрализованной работы. В качестве основной особенности технического решения портала «Демографическая безопасность России» следует отметить его модульную иерархическую структуру, обеспечивающую высокую степень унификации за счет использования ограниченного числа типовых

модулей и простейших программ-редакторов для их настройки и наполнения содержанием. Создание контента портала осуществляется непосредственно на местах – в органах государственной власти и местного самоуправления, организациях здравоохранения, региональных и местных отделениях политических партий, общественных организациях, библиотеках и др. Основой жизненной силы портала «Демографическая безопасность России» как виртуальной организации являются знания, касающиеся самых разных аспектов медико-производственной деятельности.

Портал «Демографическая безопасность России» должен обеспечивать:

- установление прямых информационных связей между производителями и потребителями продукции МПК;
- реализацию принципа информированного согласия пациента на медицинское вмешательство посредством использования ссылок на материалы по основной и смежным темам и предоставление доступа к специализированным базам данных;
- комплексную информационную поддержку всего спектра медицинских и сопутствующих им сервисных услуг и товаров медицинского назначения;
- индивидуализацию подхода к каждому пользователю;
- анонимность и конфиденциальность информации о пользователях;
- возможность проведения широкомасштабных медико-социологических и маркетинговых исследований;
- бессрочное хранение и обновление научной и коммерческой информации.

В структуре портала «Демографическая безопасность России» необходимо выделить следующие основные отделы:

- медицинские услуги (по специальностям);
- лекарственные средства;
- изделия медицинского назначения;
- гомеопатическое лечение;
- санаторно-курортное лечение;
- профилактика и оздоровление;
- лечебно-диагностические и химико-фармацевтические технологии, медицинское и медико-производственное оборудование, инструменты и материалы;
- медицинское право;
- медицинское страхование;
- общественные объединения пациентов;
- профессиональные медицинские и фармацевтические ассоциации;
- профессиональное образование и трудоустройство работников МПК;
- социология медицины;
- медицинская наука и пропаганда медицинских знаний;
- научная и научно-популярная медицинская, экономическая и юридическая литература.

На работников портала «Демографическая безопасность России» как средства массовой информации должны полностью проецироваться требования, предъявляемые российским законодательством к журналистам. Они обязаны уважать права, законные интересы, честь и достоинство граждан и организаций; проверять достоверность сообщаемой ими информации; сохранять конфиденциальность информации и ее источника; удовлетворять просьбы лиц, предоставивших информацию, об указании на ее источник, а также об авторизации цитируемого высказывания, если оно оглашается впервые; получать согласие (за исключением случаев, когда это необходимо для защиты общественных интересов) на распространение сведений о личной жизни граждан от самих граждан или их законных представителей; при получении информации от граждан и должностных лиц ставить их в известность о проведении аудио- и видеозаписи, фотосъемки и др. Сокрытие или фальсификация общественно значимых сведений, публикация сообщений о непроверенных или недоказанных методиках лечения без указания на их непроверенность или недоказанность, распространение слухов под видом достоверных сведений недопустимы. Обеспечение плюрализма мнений по спорным вопросам медицинской теории и практики должно достигаться за счет включения в каждое сообщение ссылок на другие материалы по обсуждаемой тематике. Рекламные сообщения следует выделять примечанием «Публикуется на правах рекламы».

В результате развития интеграционных процессов в информационно-коммуникационном медико-производственном пространстве возникает мощный синергетический эффект. Основными факторами синергизма в ИКМПП являются:

- концентрация прежде рассредоточенных информационных ресурсов МПК в одном месте – информационно-аналитическом портале «Демографическая безопасность России» (возникающий в результате этого кумулятивный эффект будет более весомым, чем в случае размещения тех же ресурсов на сайтах отдельных субъектов МПК);
- функциональная специализация субъектов МПК (разделение обязанностей при выполнении разделенных функций в здравоохранении способствует росту производительности труда, повышению качества оказываемых медицинских и сервисных услуг, выполняемых работ, изготавливаемых изделий медицинского назначения);
- расширение и углубление информационных связей между взаимодействующими группами участников рынка медицинских услуг и товаров медицинского назначения (за счет гармонизации отношений в ИКМПП улучшается координируемость индивидуальных и совместных действий субъектов МПК);
- активизация деятельности одних групп субъектов МПК, обусловленная присутствием в ИКМПП других групп субъектов МПК, выступающих в роли катализаторов (чем больше различных

субъектов МПК участвует в работе портала «Демографическая безопасность России», тем выше энергетический потенциал поля их взаимодействия);

- взаимодополняемость субъектов МПК (с ростом взаимодополняемости субъектов МПК расширяется пространство допустимых вариантов групповых решений и улучшаются возможности выбора из них наиболее приемлемых);
- взаимозаменяемость субъектов МПК (частичная взаимозаменяемость субъектов МПК необходима для обеспечения устойчивости работы портала «Демографическая безопасность России» в случае временного или постоянного (полного или частичного) прекращения функционирования некоторых из них);
- привлечение внимания Рунет-аудитории к порталу «Демографическая безопасность России», обусловленное концентрацией в одном месте всего спектра информации по вопросам здоровья и демографии (при условии, что пользователям будет предложена простая и эффективная система навигации по ИКМПП).

При проведении грамотной редакционной и организационно-экономической политики суммарная отдача от вложения средств в создание и обеспечение функционирования портала «Демографическая безопасность России» будет заведомо выше суммы отдачи по всей стратегической зоне хозяйствования субъектов МПК, участвующих в его работе, и в сфере деятельности органов государственной власти и местного самоуправления, политических партий, общественных организаций, библиотек и граждан, направляющих свои усилия на поддержку государственной медико-демографической политики, без учета преимуществ использования общих ресурсов – их взаимозаменяемости и взаимодополняемости.

Для анализа социально-экономических явлений и процессов, происходящих в ИКМПП, предлагается использовать методы и средства теории самоорганизации – синергетики. Теория самоорганизации исходит из того, что все происходящие в социально-экономических системах явления могут быть описаны простыми, а иногда и очень простыми, системами уравнений. Любое изучаемое явление предлагается идеализировать так, чтобы эта идеализация приводила к системе, состоящей из двух обыкновенных автономных дифференциальных уравнений, качественный анализ которых осуществляется на фазовой плоскости (если в дальнейшем будет выявлено расхождение теории с практикой, исходную модель можно будет усложнить, введя в систему еще одно уравнение, и т.д.). Адекватный математический аппарат для изучения проблем самоорганизации дает качественная теория дифференциальных уравнений. При использовании качественных методов основной акцент делается на получении качественного результата, определении наиболее характерных черт всего явления в целом, прогнозировании будущих событий. Задача качественного анализа обратна оптимизационной: требуется найти

простейшую модель, описывающую изучаемое явление, получить новые сведения об объекте исследования и выработать рекомендации для практики. Математическая часть качественного исследования состоит в поиске неприводимых топологических структур, на которые разбиваются фазовые портреты изучаемых систем уравнений, прикладная часть – в сопоставлении этих структур конкретным социально-экономическим явлениям и проведении бифуркационного анализа (идентификации быстрых изменений, скачков, разрывов непрерывности – катастроф). Аналогами целей оптимизационного подхода здесь служат аттракторы, на которые "сваливается" система, когда попадает в их область притяжения. Применяв целевой подход, можно изменить тип аттрактора, к которому изначально стремилась изучаемая система, искусственно вызвав бифуркацию ее фазового портрета.

В первую очередь методы и средства теории самоорганизации целесообразно использовать при изучении следующих социально-экономических явлений и процессов, происходящих в ИКМПП [7]:

- информационного развития ИКМПП;
- экономического развития ИКМПП;
- предпринимательской активности субъектов медико-производственной деятельности в ИКМПП;
- развития науки в ИКМПП;
- подготовки организаторов здравоохранения и врачей-менеджеров;
- выработки коллективных решений, направленных на повышение социально-экономической эффективности функционирования МПК. Особое внимание необходимо уделить анализу влияния процессов, происходящих в ИКМПП, на развитие общей валео-демографической ситуации в стране.

Обобщая сказанное, сформулируем следующие основные выводы:

1. Ведущей тенденцией мирового цивилизационного процесса является в наши дни формирование глобального информационного общества.

2. К числу основных путей достижения цели формирования и развития информационного общества в Российской Федерации относится сохранение культуры многонационального народа РФ, укрепление нравственных и патриотических принципов в общественном сознании, развитие системы культурного и гуманитарного просвещения.

3. В условиях информационного общества традиционные библиотеки должны искать новые способы выполнения своих социальных функций, внедрять новые социальные технологии.

4. Участвуя в создании социальных сетей и формировании на базе Рунет проблемно-ориентированных информационно-коммуникационных социальных пространств, российские традиционные библиотеки и созданные на их базе электронные библиотеки должны внести весомый вклад в развитие в стране информационного общества.

5. Одна из основных возможностей, предоставляемых информационным обществом, – качественное улучшение системы охраны здоровья людей.

6. Эффективным средством информационной поддержки государственной медико-демографической политики, инструментом общественного контроля за ходом ее реализации призван стать информационно-аналитический портал «Демографическая безопасность России».

7. Важнейшей общественно-политической задачей портала «Демографическая безопасность России» является содействие формированию у россиян нового валеологического мышления на принципах здорового образа жизни и семейных ценностей.

8. Главная научно-практическая задача портала «Демографическая безопасность России» заключается в содействии формированию информационно-коммуникационного медико-производственного пространства, накоплению, хранению и актуализации информации о здоровье и здравоохранении, создании уникальной по содержанию и объему базы данных по проблемам валео-демографической безопасности России.

9. Основной особенностью технического решения портала «Демографическая безопасность России» является его модульная иерархическая структура.

10. Основой жизненной силы портала «Демографическая безопасность России» являются знания, касающиеся самых разных аспектов медико-производственной деятельности.

11. В результате развития интеграционных процессов в ИКМПП возникает мощный синергетический эффект.

12. Для анализа социально-экономических явлений и процессов, происходящих в ИКМПП, предлагается использовать методы и средства теории самоорганизации – синергетики.

Литература

- [1] Альтовский Е. Этический кодекс для информационного общества. – МОО ВПП ЮНЕСКО «Информация для всех», 2007. – <http://www.ifap.ru>.
- [2] Голухов Г.Н., Мешков Н.А., Шиленко Ю.В. Медико-производственный комплекс: современное состояние и перспективы развития. – М.: Алтус, 1998.
- [3] Гундаров И.А. Демографическая катастрофа в России: причины и пути преодоления // Почему умирают русские: Последний шанс. – М.: Эксмо, 2004.
- [4] Кастельс М. Галактика Интернет. – Екатеринбург: У-Фактория, 2004.
- [5] Меськов В.С., Куликова И.В., Мамченко А.А. Открытый контент как феномен и модель обустройства обществ, базирующихся на знаниях // Открытое образование. – 2006. – № 5.

- [6] Мешков Н.А. Системное моделирование медико-производственного комплекса // Автоматизация и современные технологии. – 2003. – № 11.
- [7] Мешков Н.А. Качественный анализ явлений и процессов, происходящих в проблемно-ориентированном информационно-коммуникационном социально-экономическом пространстве // Вестник МГОУ. Серия «Экономика». – М. : МГОУ, 2006. – № 4.
- [8] Мешков Н.А. Информационно-аналитический портал "Демографическая безопасность России" // Демографическая безопасность России. Материалы форума. – М. : Издат. дом НП, 2007.
- [9] Решетников А.В. Здоровье нации: хотим получить, не желая добиться // Российское здоровье и общество. – 2003. – № 3.
- [10] Сазанов В.М. Социальные Сети – инструмент инновационного развития. – http://ntl-cbm.narod.ru/cbm-net/pr-08/sna_comm.doc
- [11] «Стратегия развития информационного общества в российской федерации» (утв. Приказом Президента РФ от 7 февраля 2008 г. № Пр-212).
- [12] Meshkov N.A. Principles of informational environment for Medical industry complex forming // Information and Telecommunication Technologies in Intelligent Systems: Proceeding of International Conference. – Barcelona, Spain, 2003.

The role of traditional libraries in formation of the problem-oriented information communication social spaces

N.A. Meshkov, L.P. Sorokina, N.V. Tsvetkova

The role of traditional libraries in formation of the problem-oriented information communication social spaces is uncovered.

Проект создания и реализации городского библиотечного портала Dubna Library

© В.В. Кореньков
korenkov@cv.jinr.ru

Э.Г. Никонов
e.nikonov@jinr.ru

И.А. Филозова
Irina.Filozova@jinr.ru

Объединенный институт ядерных исследований

Аннотация

Статья посвящена описанию проекта по созданию библиотечного портала г. Дубна. Приводится обзор технического и программного оснащения библиотек-участниц проекта, общее описание разрабатываемой информационной системы и ее базовых компонентов. Описывается текущее состояние и план дальнейших работ.

1 Введение

На современном этапе развития глобальной сети Интернет наблюдается тенденция к объединению информационных ресурсов по некоторому признаку – тематическому, профессиональному и др. – и создание виртуальных сообществ или групп в библиотечном Интернет-пространстве. Как правило, акцент делается на создание сводных каталогов. Для этого библиотеки объединяются в корпоративные системы, в которые могут входить библиотеки города, области, региона. Для повышения качества обслуживания пользователей в последнее время возросла популярность программных продуктов, называемых библиотечными порталами. Это класс систем, которые позволяют создать единый интерфейс доступа к разнородным библиотечным ресурсам и управлять доступом к ним. Уже накоплен существенный практический опыт создания и развития профессионально-ориентированных информационно-библиотечных порталов, ориентирующиеся на профессиональное библиотечное сообщество. Например, Российская Библиотечная Ассоциация (РБА, <http://www.rba.ru:8101/index.html>), информационный проект Научной библиотеки МГУ и компании "Библиотечная компьютерная сеть" СИГЛА (<http://www.sigla.ru/>) и др. [4].

Библиотеки ведут и развивают электронные каталоги своих фондов. Кроме традиционных каталогов в библиотеках накапливаются собственные электронные ресурсы, они являются подписчиками на различные электронные издания. В результате этих процессов библиотеки должны обеспечивать полно-

ценную информационную поддержку целого комплекса накопленных ими ресурсов, учитывая их специфические особенности [1]. Для своих читателей библиотеки традиционно представляют бесплатный доступ к своим электронным каталогам, подписным изданиям, электронным библиографическим базам данным и пр. По мере роста количества библиотек, выставляющих свои ресурсы в Интернет-пространстве, становится естественным иметь возможность поиска по нескольким каталогам на одном Web-сайте. Для этого нужны единые правила взаимодействия. Библиотечные порталы есть следствие развития этих тенденций. Они играют роль посредников в организации распределенного поиска. Порталы могут быть реализованы как отдельные Web-сайты, обеспечивающие всю необходимую функциональность, либо как самостоятельный программный продукт, который устанавливается и настраивается в зависимости от текущих потребностей библиотеки и категорий пользователей.

Портал – это сетевой узел или комплекс узлов, подключенных к Интернет по высокоскоростным каналам, обладающий развитым пользовательским интерфейсом и предоставляющий единый, с концептуальной и содержательной точки зрения, доступ к широкому спектру ресурсов и услуг, ориентированных на профессиональное сообщество. Главная задача портала – интеграция в единую управляемую систему служебных и профессиональных информационных потоков и ресурсов Интернет – пространства различного качества, обеспечение удобного доступа к информации различных категорий пользователей.

Формирование единого информационного пространства – это последовательность мер, практических шагов, позволяющих достичь интеграции информационных ресурсов.

При реализации проектов библиотечных порталов необходимо учитывать следующие виды деятельности:

- создание логической сети, объединяющей библиотеки;
- определение правил взаимодействия библиотек при создании общих разделяемых ресурсов и их использовании;
- использование открытых стандартов при создании информационных ресурсов и организации доступа к ним.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Цель настоящей работы – описание проекта портала Dubna Library, потенциальными участниками которого являются:

- ✓ Научно-техническая библиотека Объединенного института ядерных исследований (НТБ ОИЯИ);
- ✓ Универсальная библиотека ОИЯИ;
- ✓ Библиотечный комплекс университета «Дубна».

2 Цели создания портала Dubna Library

- o Обеспечение полноты и повышение оперативности библиотечно-информационного обслуживания всех категорий пользователей за счёт предоставления им объединенных информационных ресурсов и средств поиска с использованием современных информационных и телекоммуникационных технологий для решения научных, образовательных, деловых и других задач;
- o Обеспечение свободного доступа всех категорий пользователей к библиотечно-информационным ресурсам других библиотек в рамках действующего законодательства;
- o Сокращение финансовых, трудовых и других затрат на создание и поддержку библиотечно-информационных ресурсов, предоставление услуг пользователям;

3 Общее описание

Портал разрабатывается как frontend-решение к объединенному электронному библиотечному фонду. В качестве backend-решения используется единое распределенное электронное хранилище библиотечного фонда, основанное на межбиблиотечном взаимодействии на базе стандартного протокола ANSI/NISO Z39.50 [2][3]. Основными задачами портала являются – распределенный поиск ресурсов в гетерогенных источниках, отображение результатов поиска и дальнейшая работа с ними, идентификация пользователя и разграничение прав доступа, предоставление полных текстов документов при их наличии в информационных системах библиотек и соблюдении правовых ограничений. Функциональность портала не должна зависеть от различий в используемых форматах данных в каталогах библиотек. Внутри библиотечного хранилища все действия по наполнению, сопровождению, обмену и синхронизации будут производиться в оригинальных (родных) для данного каталога форматах: MARC-форматы (UniMARC, RusMARC), а также SUTRS. Для каждого из форматов записей разрабатываются формы вывода. Реализация портала выполняется на технологии AJAX.

Основные задачи создаваемого портала:

- ✓ Сформировать запрос и найти документы в электронных каталогах библиотек-участниц проекта.
- ✓ Оформить заказ на документ в библиотеках, которые технологически позволяют это сделать.
- ✓ Обеспечить доступ к полным текстам документов при их наличии в информационных системах библиотек и соблюдении правовых ограничений.

4 Обзор технического и программного оснащения участников проекта

4.1 Научно-техническая библиотека ОИЯИ

На текущий момент компьютерная сеть Научно-технической библиотеки ОИЯИ укомплектована 19 компьютерами. Из них 14 компьютеров – рабочие места персонала НТБ, 3 компьютера размещены в читальных залах (со сканерами и принтером они используются посетителями библиотеки), 2 узла эксплуатируются как серверы. Один из серверов – Web- и файловый сервер, второй – сервер Liber, сервер Z39.50, шлюз WWW-Z39.50, модуль Web-доступа Open Public Access Catalog (OPAC) к АИБС LiberMedia. Все рабочие места НТБ автоматизированы. Также функционируют 8 принтеров и 3 сканера (2 из них – в читальных залах).

Автоматизация работы библиотеки обеспечивается Автоматической Информационной Библиотечной Системой (АИБС) *LiberMedia*. Система *LiberMedia* – единый программный пакет управления библиотекой, автоматизирующий все библиотечные циклы: комплектование, обработку, работу с читателями, статистику, обмен библиографическими записями и т.д. Обеспечивается поддержка обмена информацией по протоколу Z39.50, возможность работы системы в сети библиотек с взаимным доступом к каталогам друг друга и возможностью обмена библиографическими описаниями и организации МБА. Имеется возможность ведения корпоративной каталогизации сети библиотек, доступа к данным и приложению через глобальные сети посредством World Wide Web.

4.2 Библиотечный комплекс университета «Дубна»

Сеть библиотеки университета «Дубна», объединяет более 50 компьютеров, большая часть которых предоставлена читателям. Автоматизация проводится на основе АИБС MAPK SQL, которая поддерживает полнотекстовые базы данных. АИБС MAPK SQL обеспечивает комплексную автоматизацию библиотечных процессов на базе новейших информационных технологий – создание электронного каталога, формирование и печать выходных форм, обслуживание читателей, ведение статистической и управленческой отчетности, анализ книгообеспеченности.

4.3 Универсальная библиотека ОИЯИ

В качестве программного обеспечения библиотечной работы используется АИБС “Моя библиотека”. Система предназначена для создания и ведения электронных каталогов. Поддерживает библиотечные процессы от комплектования литературы до регистрации перемещения фонда. Программа является клиент – серверным приложением, что позволяет вести единый электронный каталог. Система предоставляет современный поисковый аппарат, обеспечивает широкие возможности по созданию вы-

ходных отчетов, таких как: инвентарная книга, книга суммарного учета, акт списания литературы и др. Система реализована полностью через Web, что позволяет ограничить работы по установке и сопровождению системы только на самом сервере. В качестве сервера могут использоваться компьютеры под операционной системой Windows или Linux. Большое преимущество системы – возможность работы в многоязыковом окружении, поскольку данные хранятся и обрабатываются в кодировке UNICODE. Система адаптирована для работы в рамках корпораций, для чего в ее составе имеется модуль поиска и импорта записей с серверов Z39.50 и собственный сервер Z39.50 для предоставления соответствующих серверов по данному протоколу. Для управления системой используются стандартные средства работы WWW: Apache, PHP, My SQL.

5 Архитектура портала

Функционирование создаваемого портала предполагает взаимодействие трех основных уровней (см. рисунок 1):

- ✓ содержательный уровень, на котором решаются вопросы информационного наполнения портала, предоставляемых сервисов, механизмов доступа к информационным ресурсам в соответствии с разработанной политикой безопасности;
- ✓ организационно-технический уровень, предусматривающий технические аспекты разработки портала и техническую поддержку пользователей, определение категорий пользователей, каталогизацию и структуризацию ресурсов портала, обеспечение механизмов их обновления, изменения, предоставления сервисов различным категориям пользователей, интеграцию имеющихся информационных ресурсов в единое информационное пространство, сбор и анализ статистических данных по эксплуатации портала;
- ✓ уровень сопровождения, отвечающий за разработку методики использования информационных технологий в библиотечном процессе, оказание консультационных услуг пользователям и т.д.

Структура портала реализуется на принципе модульности. Модель должна поддерживать три основных компонента (см. рисунок 2):

1. общий модуль, отвечающий за управление порталом, – компонент, позволяющий осуществлять общее администрирование порталом, определяет структуру контента (масштабируемого, иерархического набора разделов и подразделов) и систему сервисов и служб, доступных пользователям;
2. административный модуль – компонент, отвечающий за управление категориями пользователей;
3. модуль управления контентом – компонент, позволяющий осуществлять управление блоками информации, представленной на портале.

Учитывая специфику портала, предполагается выделить следующие категории пользователей:

- ✓ администраторы портала – пользователи, обеспечивающие эффективную работу портала, занимающиеся мониторингом эксплуатации портала и т.д.;
- ✓ модераторы контента – пользователи, имеющие право обновлять и редактировать информационные ресурсы портала;
- ✓ читатели – пользователи Интернета, имеющие расширенные права (имеющие доступ к электронным ресурсам библиотек-участниц портала);
- ✓ гости – пользователи Интернета, обладающие минимальными правами на просмотр открытых разделов портала и использование общедоступных сервисов.

Все категории пользователей могут воспользоваться сервисами и службами портала.

Под сервисом портала понимаются доступные через внешний или внутренние интерфейсы портала, реализованные в его составе программные средства, основная задача которых – обеспечить выполнение затребованной пользователем операции над содержанием портала и эффективно использовать его ресурсы. Службой портала называется функциональная подсистема портала, доступная через внутренние интерфейсы и используемая несколькими функциональными подсистемами (сервисами) портала. Рассмотрим перечень базовых сервисов и служб создаваемого портала. В состав этих сервисов и служб входят:

• **Общепортальные сервисы:**

- ✓ Сервис интерактивного общения:

- Сервис электронной рассылки;
- Форум;
- On-line консультация;

В этом разделе предполагается организовать службу, где пользователи могут получить ответы на вопросы, касающиеся работы с порталом.

- Сервис проведения опросов и голосований;
- Электронная доска объявлений.

- ✓ Справочная система;

Справочная система должна содержать описательные сведения по каждой части портала, необходимые рекомендации для пользователей.

- ✓ Новости:

1. Новостная лента;
2. Архив новостей;

• **Поисковый сервис** обеспечивает обработку поисковых запросов пользователей и отображение результатов запроса.

• **Сервис управления**, осуществляющий аутентификацию пользователей, авторизацию доступа к ресурсам, сбор статистики и мониторинг;

• **Сервис управления профилем пользователя** позволяет настраивать персональное инструментальное окружение. Включает средства для настройки персональных данных пользователя, персонального календаря, настройки внешнего вида портала, средства организации системы закладок, личных библиотечных полок, службу заказа и доставки доступного ресурса и т.д.



Рис. 1. Схема функционирования портала

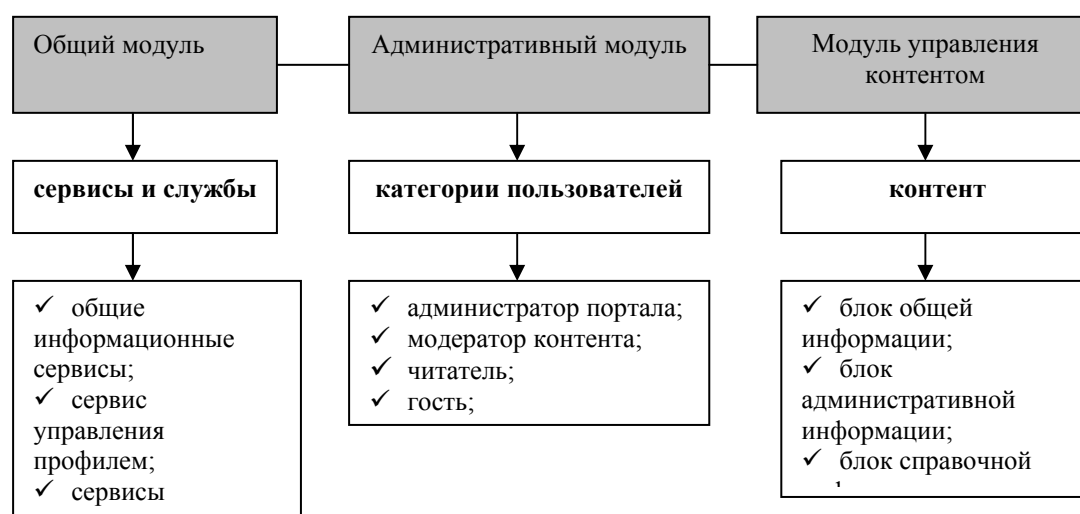


Рис. 2. Функциональная схема портала

Информационное наполнение портала будет представлено следующими блоками:

- О блок общей информации, доступный для всех категорий пользователей – модуль включает в себя новости, общую информацию о библиотеках и предоставляемых ими ресурсах и т.д.;
- О блок административной информации, доступный для администраторов портала, модераторов контента – модуль содержит информацию об организации работы портала, особенностях и ограничениях доступа к информационным ресурсам и т.д.

- О блок справочной информации – модуль содержит необходимые описания разделов портала, инструкции и рекомендации для пользователей и т.д.

5.1. Поисковый сервис

Задачей этого модуля является формирование запроса на выборку библиографической информации из выбранных пользователем (отмеченных в поисковой форме) электронных каталогов, слияние полученных данных в список библиографических записей и его отображение. Реализация поиска в удаленных каталогах базируется на протоколе Z39.50. На

текущий момент он является самым распространенным протоколом доступа к библиографической информации. Возможно, в дальнейшем для включения в портал других библиотек города понадобится реализация дополнительных решений. Это будет необходимо в случае отсутствия в них стандартных поисковых протоколов для внешнего доступа. Проблема может быть решена с помощью создания набора программ-адаптеров, преобразующих заданный API-интерфейс доступа к удаленному каталогу во внутренний формат каталога.

Для систематизации документов библиотеками используются различные тезаурусы, предметные каталоги и классификационные индексы. Эти данные планируется использовать для организации вспомогательных средств поиска для возможности производить поиск необходимой рубрики или термина.

Поисковые формы должны позволять пользователю выбирать интересующие его каталоги из списка имеющихся в наличии каталогов, предоставлять описание каталога. В режиме *простого поиска* поиск производится по вхождению указанных слов в поисковый элемент в любом месте и в любом порядке. Этот способ рекомендуется использовать в случаях, когда у пользователя нет уверенности в точной формулировке своего запроса. Расширенный поиск позволяет применять различные способы формализации – логические операторы И, ИЛИ, И-НЕ, поиск по точному совпадению и т.п. Стандартные поля для поиска: Заглавие, ISBN, ISSN, Издательство, Год издания, Автор, Серия, Рубрика.

5.2 Обработка поискового запроса и отображение результатов поиска

Поисковый запрос формируется в оригинальном для каталога формате. Это уменьшает риск потери и искажения информации. Процесс обработки запроса фиксирует количество найденных записей. Далее производится вывод записей в виде списка библиографических описаний порциями по 20 (или 10) записей на странице. Этот параметр является настраиваемым. Пользователь может отбирать записи из результатов поиска и работать с ними:

- Переход к заполнению бланка заказа при наличии в каталоге библиотеки функции on-line заказа документа;
- Сортировка отобранных записей по выбранным признакам (автору, заглавию, дате);
- Отсылка отобранных записей по e-mail;
- Просмотр и печать записей в виде списка библиографических описаний;
- Загрузка полного текста документа с учетом прав доступа.

5.3 Сервис управления

Часть интегрированных посредством портала ресурсов являются платными или доступными при соблюдении лицензионных соглашений. Реализация такой идентификации возможна на определении принадлежности текущего пользователя к зарегистрированным библиотекам по IP-адресам. Для каж-

дой библиотеки определены права доступа к ресурсам, на которые она подписана или имеет доступ на определенных условиях. Информация об идентификации пользователя используется также для разрешения ссылок и доступа к сервисам, разрешенным в рамках данного набора привилегий. Для этого для каждого сеанса пользователя необходимо сохранять информацию о характеристиках данного сеанса с помощью cookie – использованных каталогах, текущих параметрах поискового запроса, настройках портала и т.п.

6 Текущее состояние и план дальнейших работ

На данный момент проведен анализ технического и программного оснащения библиотек – потенциальных участниц проекта, поисковых требований пользователей, разработана архитектура портала, определен базовый состав его сервисов и служб, а также основных информационных блоков.

План дальнейших работ предполагает:

1. Создание прототипа портала на базе двух библиотек – НТБ ОИЯИ и Библиотечного комплекса университета «Дубна»;
2. Отладку системы на работающем прототипе;
3. Введение в эксплуатацию;
4. Определение направлений развития портала.

Литература

- [1] Борисовский, В.Ф. Электронные информационные ресурсы ОИЯИ: концепция и технологические решения / Борисовский В.Ф. [и др.] // Четвертая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2002 (Дубна, 15–17 октября 2002 г.): в 2 т. – Дубна : ОИЯИ, 2002. – Т.1. – С. 157–168.
- [2] ANSI/NISO Z39.50 – 2003 Information Retrieval: Application Service Definition and Protocol Specification http://www.niso.org/standards/standard_detail.cfm?std_id=465
- [3] ANSI/NISO Z39.89 – 2003 The U.S. National Z39.50 Profile for Library Applications http://www.niso.org/standards/standard_detail.cfm?std_id=734
- [4] Хохлов А.Ю. Сигла: портал доступа к библиографической информации // RCDL'2004, <http://www.impb.ru/~rcdl2004>

Creation and realisation project of the city library portal Dubna Library

V.V. Korenkov E.G. Nikonov I.A. Filozova

The paper is devoted to the the description of the project on creation of a library portal Dubna town. The hardware and software of the project libraries-participants project is given. The general description of the developed information system and its base components is considered. It is described current state and plan of the further works.

О создании открытой научной электронной библиотеки периодических изданий НАНУ

© В.А. Резниченко, Г.Ю. Проскудина, К.А. Кудим

Институт программных систем НАН Украины, г. Киев
reznich@isofts.kiev.ua, gupros@isofts.kiev.ua, kuzma@isofts.kiev.ua

Аннотация

В работе рассматриваются вопросы создания открытой научной электронной библиотеки периодических изданий НАН Украины. Сделан обзор некоторых известных проектов порталов периодических изданий. А также особенности функциональных возможностей программного обеспечения DSspace применительно к решаемой задаче.

1 Введение

Работы по данному направлению выполняются в рамках проекта «Разработка и реализация проектных решений автоматизированного библиотечного сервиса в интересах организаций НАН Украины». Проект предполагает выполнение нескольких заданий, одно из которых – «Разработка технологий, методов и средств создания научных электронных библиотек НАН Украины». Согласно этому заданию проводились работы:

- по созданию научных электронных библиотек в научно-исследовательских организациях НАНУ;
- а также выработка методов и средств создания интегрированных электронных библиотек в Академии наук в целом.

Доклад представляет собой обобщение нашего опыта по созданию интегрированной электронной библиотеки периодических изданий – одного из основных видов информационных ресурсов научно-исследовательских академических организаций на основе современных тенденций и решений.

Существующие в настоящее время и разрабатываемые новые системы электронных библиотек (ЭБ) характеризуются следующими тремя ключевыми особенностями:

- прежде всего это информационные системы, то есть системы сбора, хранения, обработки и выдачи информации;
- во-вторых, это системы, поддерживающие

именно электронные документы (ресурсы) различного типа;

- и, наконец, важным моментом в современном понятии ЭБ является то, что такие системы работают в глобальных информационных сетях передачи данных.

Следует отметить, что в мире на протяжении последних 20 лет интенсивно проводятся исследования и разработки по созданию ЭБ. Осознание того факта, что ЭБ являются одной из важнейших составляющих информатизации общества, привело к тому, что во многих ведущих странах мира были приняты национальные программы по созданию ЭБ, которые поддерживаются правительствами и которые имеют внушительную финансовую поддержку. Вот далеко не полный перечень таких программ:

- Национальная электронная библиотека США (Библиотека конгресса США).
- «Электронная библиотека XXI столетия» – Япония.
- Национальная программа Великобритании eLib по созданию ЭБ.
- Программа ЕС «Создание Европейского библиотечного пространства».
- Программа «Электронные библиотеки России».
- Программа «Национальная электронная библиотека Украины».
- Национальные НЭБ США, Франции, Великобритании, Канады, России.

В результате этих усилий в настоящее время в сети Интернет появилось множество ЭБ различной направленности, в том числе и научных электронных библиотек (НЭБ). Они в основном создаются в НИИ, учебных заведениях, международных ассоциациях. Среди них имеются и НЭБ открытого доступа. Наиболее известны два международных каталога НЭБ открытого доступа:

- OpenDOAR – каталог репозитория открытого доступа, насчитывающий 1172 НЭБ, <http://www.opendoar.org>;
- ROAR – реестр репозитория открытого доступа, насчитывающий 1101 НЭБ, <http://roar.eprints.org>.

Следовательно, в настоящее время в открытом доступе уже более 1100 НЭБ, тогда как до середины прошлого года не было ни одной НЭБ от Украины.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2008, Дубна, Россия, 2008.

Однако уже к концу года в этих реестрах появилось 7 украинских НЭБ, в том числе 2 от Академии наук, 4 от ВУЗов и 1 от общественной организации при поддержке фонда «Возрождение». Причем 4 из них были созданы при нашем непосредственном участии. Эти библиотеки создавались на базе научных и учебных заведений, специфика которых определила характер и состав содержащихся в них научных информационных ресурсов.

Основная тема нашего доклада – создание НЭБ периодических изданий открытого доступа (предполагается публикация приблизительно 100 журналов). Здесь также сделано достаточно много в мире. Международный каталог журналов открытого доступа DOAJ в настоящее время содержит 3245 таких журналов и среди них 10 из Украины.

Анализ специализированных журналов Украины и НАНУ, в частности [1], показал, что 80% периодических изданий физико-математических наук, 87% химических, 95% геологических, 95% географических, и 97% технических наук практически остаются невидимыми.

Следует отметить, в Национальной библиотеке Украины им. Вернадского (НБУВ) с 2006 года ведется Общеакадемический портал научной периодики, куда включено приблизительно 500 научных периодических изданий Украины. В нем предпринята первая в Украине успешная попытка объединить научные информационные ресурсы периодических изданий. Наши предложения и решения являются не альтернативой тому, что уже создано в НБУВ, а его дальнейшим развитием с учетом новых, современных концепций. В частности, предполагается основное внимание уделить следующим нерешенным вопросам, которые на наш взгляд являются узким местом:

- поиск по картотеке статей (электронный каталог), нет полнотекстового поиска;
- нет средств пополнения библиотеки с рабочих мест в институтах;
- нет средств интеграции в международные каталоги, реестры, репозитории.

Далее в работе рассматриваются: примеры известных проектов порталов научной периодики открытого доступа (раздел 2); формулировка целей и концептуальных положений открытой НЭБ периодических изданий НАНУ (раздел 3); обоснование выбора платформы ПО для решения этой задачи (раздел 4); краткая информация о ПО (раздел 5). Вопросы масштабируемости системы DSpace рассмотрены в разделе 6; некоторые важные функциональные возможности системы DSpace рассмотрены в разделах 7–10 и 12; процедура внесения статей в систему описана в разделе 11.

2 Обзор известных проектов порталов научной периодики открытого доступа

Рассмотрим более детально особенности построения и реализации подобных систем на примере

довольно крупных и функционально приемлемых решений.

Хорватская система периодических изданий <http://hrcak.srce.hr/?lang=en>, на момент написания статьи содержала 163 наименований журналов по различным направлениям научной деятельности (естественные, медицинские, технические и гуманитарные науки), 2087 выпусков, 24,609 статей, из которых 22,287 полнотекстовых. Здесь предоставляется как простой, так и расширенный поиск, который включает булеву (операторы И, ИЛИ, И-НЕ) комбинацию до четырех полей метаданных (автор, название, аннотация, издательство, ключевое слово, название журнала) и полного текста, а также запрос можно уточнить дополнительными значениями области исследования, года выпуска статьи, языка и словом в тексте статьи. Так, например, на запрос с указанием только лишь русского языка, система выдала 71 полнотекстовую статью.

Бразильская интегрированная система публикации периодических изданий SciELO (Scientific Electronic Library Online, (<http://www.scielo.org/php/index.php>)) содержит:

- 550 журналов из Аргентины, Бразилии, Чили, Кубы, Испании, Португалии, Колумбии и Венесуэлы;
- 11,941 выпусков;
- 180,187 полнотекстовых статей;
- 3,530,025 ссылок цитирования.

Помимо предоставления сервисов просмотра и поиска (в том числе полнотекстового), вычисляются интегрированные статистические показатели популярности (использования) и влияния (импакт-факторы) периодических изданий, научных публикаций.

Наиболее мощной, на наш взгляд, является система HighWire Press (<http://highwire.stanford.edu/>), подразделение библиотеки Стэнфордского университета, которая публикует онлайн версии журналов с высоким импакт фактором (high-impact) и уровнем рецензирования (peer-reviewed) и другой научный (учебный) контент. HighWire Press награждена премией ALPSP (Association of Learned and Professional Society Publishers) за предоставляемые «Сервисы для некоммерческой публикации». HighWire в партнерстве с влиятельными академическими сообществами, университетскими издательствами, создает прекрасные онлайн-коллекции полностью доступной для поиска исследовательской и клинической литературы. Совместно эти партнеры выпускают 71 из 200 наиболее часто цитируемых научных журнальных публикаций.

HighWire Press представляет собой наиболее крупный в мире репозиторий полнотекстовых статей наук о жизни (биология, медицина, антропология, социология и т. п.), где (на момент написания этого доклада) содержится 1,904,070 статей, которые находятся в открытом доступе. Всего же ЭБ содержит 1157 журналов и 4,806,738 полнотекстовых статей из более чем 140 научных издательств.

3 Цели и концептуальные положения

Создание открытой НЭБ периодических изданий НАН Украины преследует следующие цели:

- Достижение качественно нового уровня, полноты и оперативности удовлетворения информационных потребностей научных сотрудников НАНУ.
- Повышение эффективности использования имеющихся информационных ресурсов НАНУ.
- Оперативное и полное информирование международной научной общественности о научных достижениях и разработках институтов НАНУ.

Концептуальными моментами создания этой НЭБ являются следующие:

- Единый портал для описания журналов и хранения статей.
- Централизованное ведение репозитория, администрирование и предоставление услуг.
- Предоставление возможности по пополнению НЭБ электронными статьями с рабочих мест институтов или редакций журналов.
- Интеграция с другими международными каталогами, реестрами, репозиториями.

Схема функционирования НЭБ представлена на рис. 1.



Рис. 1. Общая схема функционирования НЭБ

При таком решении централизованно осуществляются следующие функции:

- администрирование;
- ведение библиотеки;
- предоставление услуг (сервисов);
- сопровождение программного обеспечения.

На местах, где редакции институтов, как правило, не имеют соответствующих специалистов для установки и поддержки библиотечной системы, осуществляются только лишь функции, связанные с наполнением контентом данной системы, а именно:

- оформление статей своих периодических изданий в электронном виде;
- загрузка или передача статей в систему, используя Интернет-браузер.

В итоге мы получаем продукт сотрудничества IT-специалистов и прикладных специалистов информационного профиля различных предметных областей.

4 Выбор платформы системы ПО

С начала 2005 года нами было изучено и опробовано несколько платформ ПО для построения такого рода систем. Результаты этой работы освещались в [2–5].

Для реализации НЭБ периодических изданий открытого доступа была выбрана ПО DSpace [6].

Из особенностей и преимуществ применения данной системы можно выделить следующие:

- В подавляющем большинстве приложений ПО DSpace используется для создания репозитория одной отдельной организации. В нашем приложении DSpace устанавливается и администрируется централизованно, но применяется многими организациями-партнерами, редакционными коллегиями периодических изданий институтов НАНУ. Каждая такая организация имеет свой раздел (разделы) и отвечает за его своевременное наполнение.
- Не осуществляется самоархивирования своих материалов самими авторами статей (важная функциональная особенность данной системы). В данном случае система выступает в качестве провайдера сервисов. Контент наполняется исключительно сотрудниками редакций периодических изданий.
- Возможность предоставления функций просмотра и поиска по основным атрибутам (и логическим комбинациям таких атрибутов) публикации (названию, автору, году издания, ключевым словам), полнотекстового поиска и синдикации RSS о новых поступлениях в библиотеку [7].
- Благодаря хранению описаний публикаций (метаданных) в соответствии с международными стандартами (в данном случае – расширенное Дублинское ядро) и поддержке протокола OAI-PMH [8] существует возможность интеграции с международными каталогами, например, OAIster, OpenDOAR, ROAR и др.
- Индексация Google.

5 DSpace – общие сведения

DSpace – программное обеспечение (ПО) системы управления цифровыми ресурсами или активами (DAMS – Digital Asset Management System), предоставляющее услуги по доступу, наполнению, управлению и повторному использованию цифровых активов, преимущественно образовательных и исследовательских материалов. В настоящее время система DSpace установлена в более чем 220

университетах мира, с суммарным объемом хранения более чем один миллион цифровых ресурсов, представляющих исследовательские статьи, технические отчеты, наборы данных, изображений и видео. Кроме того, некоторые другие организации используют DSpace, для хранения, организации и сбережения своих цифровых активов, например, HP Labs. В настоящее время, DSpace – система с открытым исходным кодом, размещаемая на SourceForge и распространяемая по лицензии BSD [9], авторские права принадлежат HP и MIT [6]. Для поддержки и защиты поставщиков информации на платформе DSpace в 2007 была создана некоммерческая организация, которая призвана вовлечь многие организации (являющиеся пользователями DSpace) в управление ПО DSpace.

Система DSpace развивается также в рамках программы Google Summer of Code™ (GSoC) под руководством Роберта Тенсли (программиста первоначальной версии системы), где в настоящее время на платформе DSpace выполняются проекты [10]: сервисы целостности контента, статистика, поддержка версий, визуализация.

В настоящее время интенсивно обсуждается совместный проект сотрудничества DSpace и Fedora [11].

Таким образом, разработка системы DSpace находится в постоянном развитии и при непосредственном участии многочисленных (нескольких сотен) пользователей, что гарантирует долговременное использование, своевременное обновление, а также необходимую помощь и поддержку в случае возникновения непредвиденных проблем (например, утечки собственных кадров, IT-специалистов).

6 Масштабируемость

Перспективы развития системы также освещены в [12] на совещании групп пользователей системы DSpace, где особое внимание было уделено вопросам масштабируемости, способности системы адаптироваться к расширению предъявляемых требований и возрастанию объемов расширяемых задач. Проблемы масштабируемости рассматриваются в 3-х аспектах [13]:

- емкость, или вместимость, т.е. как DSpace будет справляться с большими объемами данных?
- пропускная способность, как быстро DSpace будет поглощать (импортировать) контент?
- параллелизм, как DSpace работает под нагрузкой, при одновременном доступе многих пользователей к репозиторию?

Исходя из этого были сформулированы требования масштабируемости для DSpace:

- DSpace может содержать 10 млн. записей;
- DSpace может обрабатывать файлы произвольных размеров;
- Время занесения элемента (статьи) в 10 млн. репозиторий не превышает 1 сек;

- DSpace поддерживает параллельную работу 10 пользователей на внесение информации и 100 пользователей на чтение, при наличии достаточной полосы пропускания, приемлемого процессора и емкости памяти;
- стратегия реализации DSpace – сбалансированные по загрузке кластеры серверов, управляющие одним репозиторием.

Наряду с обнадеживающими перспективами масштабируемости системы DSpace, анализ функциональности ЭБ [14] позволил выделить три важнейшие функциональные возможности, благодаря которым был сделан выбор в пользу данной системы:

- поддержка иерархической модели коллекций документов;
- достаточно хорошо поддающийся настройке пользовательский интерфейс;
- гибкая настройка прав доступа к коллекциям, отлаженные механизмы безопасности (аутентификации, авторизации и т.п.).

Остановимся несколько подробнее на перечисленных особенностях и наших проектных решениях.

7 Архитектура

Способ организации данных в DSpace [4] отражает структуру организации, использующей данную систему. Разделы и подразделы образуют структурную иерархию, т.е. подразделы могут также содержать свои подразделы, внизу иерархии находятся коллекции. В данном случае раздел соответствует конкретному периодическому изданию, периодические издания в свою очередь структурированы в иерархию согласно направлениям деятельности и существующим отделениям (структурным подразделениям НАНУ, см. рис. 2).

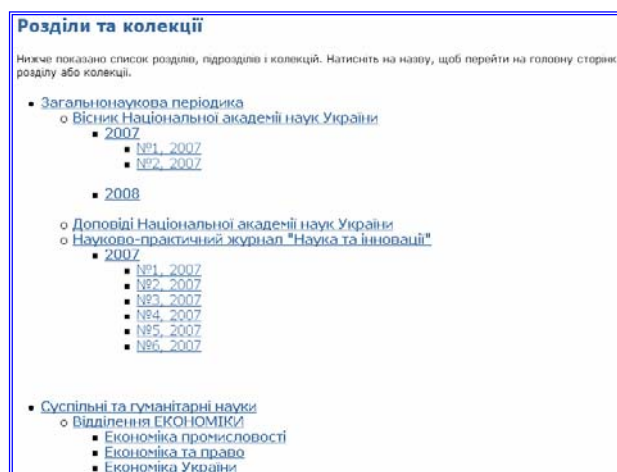


Рис. 2. Фрагмент иерархии разделов, подразделов и коллекций

КОЛЛЕКЦИЯ соответствует конкретному номеру журнала. Например, коллекция №1, 2007 журнала «Вестник Национальной академии наук Украины» (название журнала соответствует ПОДРАЗДЕЛУ, год выпуска журнала, например, 2007 – также соответствует ПОДРАЗДЕЛУ).

В результате имеем следующую структуру:

раздел (community) – направление деятельности НАНУ;

подраздел (sub-community) – отделение НАНУ и/или периодическое издание (журнал);

подраздел (sub-community) – год или том издания;

коллекция (collection) – выпуск или номер журнала.

Поскольку создаваемая библиотека не имеет предметной специализации, с большой вероятностью можно предположить, что каждому периодическому изданию или группе периодических изданий потребуется своя, специфическая процедура ввода и набор описательных атрибутов (обязательные, необязательные или факультативные, управляемые на момент внесения, свои тематические словари) при внесении своих ресурсов (в данном случае статей). Для реализации индивидуальных требований участников внесения контента в системе предусмотрена настройка как процедуры внесения, так и форм внесения, содержащих полный набор необходимых описательных атрибутов, форматов полных текстов статей, принятия лицензии по авторскому праву, дополнительные процедуры проверки своего контента и т.д. Установление разных параметров внесения для каждой коллекции (раздела) позволяет администратору системы устанавливать разный контент внесения, а также параметры проверки и одобрения (в случае необходимости) для каждой из коллекций и/или групп пользователей, которые определены в НЭБ.

Здесь же следует отметить важную возможность создания домашней странички для каждого раздела, подраздела, коллекции (рис. 3).



Рис. 3. Главная страница коллекции “№1, 2007” журнала “Вісник Національної академії наук України”

8 Настройка шагов внесения статей в коллекцию

Стандартная последовательность шагов внесения или процедура передачи документа в DSpace следующая:

Описание→Описание→Описание→Загрузка→Проверка→Лицензия→Завершение.

Однако при необходимости такая последовательность внесения может быть изменена, например, следующим образом:

Лицензия→Описание→Описание→Загрузка→Проверка→Завершение.

Существует возможность переупорядочивания шагов внесения, либо некоторые шаги могут быть вообще пропущены, причем каждой коллекции может быть назначена своя процедура внесения. Настройка и конфигурирование пользовательского интерфейса DSpace внесения контента подробно описана в [15].

9 Настройка пользовательских экранов внесения

В DSpace возможна настройка форм для разных коллекций. Это делается с помощью конфигурационного файла input-forms.xml. В [16] представлено подробное описание его формата. Там можно указать какие именно поля Дублинского ядра (ДЯ) нужно заполнять на этапе внесения ресурса (статьи) в данную коллекцию, их последовательность и текстовые обозначения (метки) этих полей, как можно разбить ввод метаописаний на несколько страниц. Также через веб-интерфейс для определенной коллекции можно ввести начальные значения для полей или необходимые списки выбора значений.

Описание того, как создавать свой собственный набор пользовательских форм также приведено в Приложении А.

10 Гибкая настройка прав доступа к коллекциям

DSpace обладает гибкой системой прав пользователя, которая тесно связана с применяемой моделью данных. Выделены следующие группы пользователей: администраторы, редакторы, участники процесса внесения, подписчики и пользователи с разрешением просматривать необщедоступные элементы (документы).

Гибкая система прав доступа в DSpace позволяет ограничивать доступ к различным частям репозитория. Каждому разделу можно назначить группу пользователей, которым разрешается доступ к данному разделу. Каждой коллекции назначается множество из отдельных пользователей и групп, которые могут вносить контент для этой коллекции, будут иметь доступ к содержимому, выполнять функции редакторов, или, наконец, администрировать коллекцию.

Пользователь может быть ассоциирован с несколькими группами одновременно. Каждый пользователь группы наделяется соответствующими правами.

Управление группами и отдельными пользователями, а также редактирование прав доступа к разделу или коллекции осуществляется через веб-интерфейс. Доступ к этим функциям имеет пользователь с правами администратора [4].

11 Процедура внесения статей в НЭБ

Для внесения статей в систему нами разработана «Инструкция пользователя. Внесение статей», согласно которой эти действия могут производить только зарегистрированные пользователи системы, наделенные правами вносить и изменять содержимое только строго определенного раздела (или разделов), а также, возможно, правами создания коллекции внутри этого раздела (если администратор системы сочтет это действие возможным). Такие пользователи могут вносить статьи только в существующие в НЭБ номера (выпуски) своего журнала. Каждый выпуск журнала имеет в НЭБ статус КОЛЛЕКЦИИ.

Когда новый пользователь регистрируется в системе, ему выделяется отдельная рабочая область, куда он может загружать свои документы. Существует две возможности инициирования внесения статей в журнал:

1. Выбрать опцию МОЯ БИБЛИОТЕКА и НАЧАТЬ НОВОЕ ВНЕСЕНИЕ, и затем выбрать ту коллекцию, куда будет произведено внесение.

2. Выбрать опцию просмотра РАЗДЕЛЫ И КОЛЛЕКЦИИ на главной странице системы (рис. 2), затем выбрать разрешенный для внесения раздел, коллекцию.

Далее нужно выполнить последовательность шагов в соответствии с конфигурацией внесения выбранной коллекции. Если такая конфигурация специально не определена, то следует процедура по умолчанию:

- выбрать опции, влияющие на множество полей, доступных для ввода на последующих шагах (рис. 4);

Рис. 4. Процедура внесения: Начало. Описание общих характеристик элемента (статьи)

- ввести значения основных полей метаданных;
- ввести ключевые слова, классифицирующие тему элемента, а также ввести дополнительные поля метаданных;
- загрузить файлы;
- проверить, все ли в порядке с загруженными файлами и их форматом и отредактировать в случае наличия ошибок;
- проверить, все ли в порядке во всех данных, введенных ранее; отредактировать в случае наличия ошибок;
- принять лицензионное соглашение.

Прерывание процесса внесения безопасно с точки зрения сохранности введенных данных.

12 Статистика в DSpace

Для обеспечения эффективного контроля функционирования НЭБ система DSpace включает модуль статистики, позволяющий собирать достаточно подробный статистический анализ использования публикуемой информации:

- количество уникальных посетителей;
- количество визитов в день/месяц;
- количество просмотренных страниц в день/месяц;
- количество визитов страницы, раздела, коллекции, документа;
- поисковые ключевые фразы и слова;
- наиболее популярные страницы и прочее.

По этой информации можно создать различные представления (разные статистические отчеты).

Также для статистического анализа работы системы может быть подключен внешний модуль статистики [10].

13 Заключение

ПО DSpace с открытым исходным кодом применено к специфичному контенту, а именно периодическим изданиям. В настоящее время осуществляется открытый доступ пока только к нескольким научным журналам, наполняемых редакционными коллекциями разных организаций НАНУ: <http://dspace.nbuv.gov.ua:8080/dspace>.

Архитектура системы выбрана таким образом, что каждое периодическое издание может иметь свою настройку как домашней странички издания, так и полей заполнения метаданных своих ресурсов (статей), а также, возможно, форм вывода (просмотра). Такой подход может быть использован для создания различных междисциплинарных репозиторий (а также тематических), например, научных отчетов многих организаций со 100% открытым доступом, а также 100% предоставлением полных текстов статей.

Что надо сделать?

- Выработка и поддержание политики функционирования НЭБ (регламент подачи статей, со-

- соблюдение авторских прав, положение о НЭБ, распоряжение по НАНУ, отслеживание соблюдения политики функционирования).
- Эксплуатация НЭБ (техническое оснащение, обеспечение бесперебойной круглосуточной работы, конфигурирование и настройка НЭБ, отслеживание соблюдения политики пополнения НЭБ).
- Программно-методическое сопровождение (сопровождение программного обеспечения, инструкции, методики, руководства и другие материалы, обучение и консультирование).
- Бесперебойное пополнение НЭБ (решение организационно-технических вопросов, обучение персонала, представление документов в электронном виде, пополнение ЭБ новыми выпусками журналов, соблюдение политики функционирования ЭБ).

Литература

- [1] Мриглод І., Мриглод О. Наука України у світовому інформаційному просторі // Вісник Національної Академії наук України. — 2007 — №10. — С. 3–18.
<http://dspace.nbuv.gov.ua:8080/dspace/bitstream/123456789/562/1/a1-10.pdf>
- [2] Резниченко В.А., Проскудина Г.Ю., Овдий О.М. Создание цифровой библиотеки коллекций периодических изданий на основе Greenstone // Электронные библиотеки. — 2005. — 8. Вып. 6.
<http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part6>
- [3] Новицкий А.В., Кудим К.А., Резниченко В.А., Проскудина Г.Ю. Создание научных архивов с помощью системы EPrints // Проблемы программирования. — 2007. — № 1. — С. 46–60.
<http://eprints.isofts.kiev.ua/157/>
- [4] Кудим К.А., Проскудина Г.Ю., Резниченко В.А. Создание научных электронных библиотек с помощью системы DSpace // Проблемы программирования. — 2007. — № 3. — С. 49–60.
<http://eprints.isofts.kiev.ua/233/>
- [5] Кудим К.А., Проскудина Г.Ю., Резниченко В.А. Сравнительный анализ функциональных возможностей систем электронных библиотек // Проблемы программирования. — 2007. — № 4. — С. 32–48. <http://eprints.isofts.kiev.ua/320/>
- [6] DSpace Federation Web site <http://dspace.org>
- [7] <http://ru.wikipedia.org/wiki/RSS>
- [8] The Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 of 2002-06-14.
<http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>
- [9] Open Source BSD License. Available at <http://www.opensource.org/licenses/bsd-license.php>
- [10] Robert Tansley Google Summer of Code // In. Proc. DSpace User Group Meeting. — Rome, Italy. — Oct 2007.
<http://www.aepic.it/conf/viewabstract.php?id=338&cf=11>
- [11] Fedora – Flexible Extensible Digital Object Repository Architecture. <http://www.fedora-commons.org/>
- [12] Robert Tansley DSpace Architecture: an update// In. Proc. DSpace User Group Meeting. — Rome, Italy. — Oct 2007. 2007
<http://www.aepic.it/conf/viewpaper.php?id=328&cf=11>
- [13] Toward the next generation: Recommendations for the. next DSpace Architecture. Written by the DSpace architecture review group John Mark Ockerbloom, chair. — Jan. 2007.
http://wiki.dspace.org/static_files/0/0e/DSpace-recs.pdf
- [14] Резниченко В.А., Проскудина Г.Ю., Овдий О.М. Функциональные возможности современных систем электронных библиотек // Проблемы программирования. Специальный выпуск №2–3. — Материалы конференции УкрПрог '2008. — С. 525–532. <http://eprints.isofts.kiev.ua/420/>
- [15] Customizing and Configuring Submission User Interface.
<http://researchspace.auckland.ac.nz/docs/submission.html>
- [16] Custom Metadata-entry Pages for Submission
http://www.dspace.org/index.php?option=com_content&task=view&id=155
- [17] Handle
http://www.dspace.org/index.php?option=com_content&task=view&id=149#handles
- [18] Configuring Controlled Vocabularies
http://www.dspace.org/index.php?option=com_content&task=view&id=147#controlledvocabulary

On creation the open access scientific digital library for periodical of NASU

V.A. Reznichenko, G.Yu. Proskudina, K.A. Kudim

The basic issues for creation of the open access scientific digital libraries for periodical of NAS Ukraine are considered in the work. The review of some known projects of portals of periodicals is presented. And also features of functionalities of software DSpace with respect to solved task are discussed.

ПРИЛОЖЕНИЕ А

Описание пользовательских форм метаданных

Описание набора страниц, на которых пользователь вводит свои метаданные, называются формой (или набором форм). Форма идентифицируется уникальным символьным именем. В XML-структуре форма разбивается на несколько страниц, каждая из которых представляет собой отдельную веб-страницу для сбора элементов метаданных. Чтобы настроить какую-либо из пользовательских коллекций в DSpace формами ввода, следует войти в карту формы (form-map). Фактически она представляет собой таблицу, связывающую коллекцию с соответствующей формой, соединяя идентификатор Handle коллекции с именем формы. Коллекции идентифицируются при помощи Handle [17], т.к. их имена могут изменяться и они не обязательно уникальны, тогда как идентификаторы Handle уникальны и постоянны. Специальная строка таблицы предназначена для набора форм по умолчанию, которые вызываются в случае, если специально не указан идентификатор Handle коллекции. Они применяются ко всем коллекциям, которые специально не заданы в этой таблице. В XML-примере набор форм назван как традиционной (traditional, для традиционного пользовательского интерфейса системы DSpace), но его можно назвать и иначе.

Структура input-forms.xml

Конфигурационный XML-файл имеет один высокоуровневый элемент <input-forms>, содержащий в определенном порядке три элемента.

```
<input-forms>
  <!-- Соответствие коллекций наборам форм -->
  <form-map>
    <name-map collection-handle="default" form-name="traditional" />
    ...
  </form-map>
  <!-- Определения набора форм -->
  <form-definitions>
    <form name="traditional">
      ...
    </form-definitions>
  <!-- Пары Имя/Значение, используемые в элементах множественного выбора -->
  <form-value-pairs>
    <value-pairs value-pairs-name="common_iso_languages" dc-term="language_iso">
      ...
    </form-value-pairs>
  </input-forms>
```

Добавление карты коллекции

Каждый элемент <name-map> в пределах тэга <form-map> связывает коллекцию с именем набора форм. Его атрибут collection-handle определяет Handle-адрес коллекции и атрибут form-name определяет имя набора форм. Например, следующий фрагмент показывает, как коллекция с Handle-идентификатором равным "12345.6789/42" присоединена к набору форм "TechRpt":

```
<form-map>
  <name-map collection-handle="12345.6789/42"
    form-name="TechRpt" />
  ...
</form-map>
<form-definitions>
  <form name="TechRpt">
    ...
  </form-definitions>
```

Здесь, в примере input-forms.xml, заложена хорошая идея – хранить определение name-map по умолчанию. В случае, когда для некоторых коллекций такие формы специально не определены, всегда можно получить набор форм по умолчанию.

Получение идентификатора Handle коллекции

Для того чтобы назначить пользовательский набор форм нужно знать идентификатор Handle коллекции. Его можно получить, выбрав просмотр «Разделы и коллекции» в меню домашней страницы системы, а затем найти ссылку нужной коллекции.

Например:

<http://myhost.edu/dspace/handle/12345.6789/42>

Подчеркнутая часть URL есть идентификатор Handle, известный каждому администратору системы. Именно он входит в атрибут collection-handle элемента <name-map>.

Добавление набора форм

Можно добавить новый набор форм путем создания нового элемента формы внутри элемента <form-definitions>. Он имеет атрибут name, значение которого как отмечалось выше, должно совпадать с атрибутом form-name элемента <name-map>, используемого для соответствующей коллекции.

Формы и страницы

Содержание формы есть последовательность элементов страницы. Каждый из которых соответствует определенной веб-странице или форме ввода элементов метаданных, последовательно представленных между начальной страницей «Описание» и финальной страницей «Проверка», представляющую сводный список всех собранных метаданных.

Форма может содержать от одной до шести страниц. Они представлены в том порядке, как они идут в XML-файле. Каждый страничный элемент

должен содержать атрибут `number`, равный порядковому номеру страницы, например:

```
<page number="1">
```

Элемент страницы, в свою очередь, содержит последовательность элементов полей. Каждое поле определяет интерактивный диалог, куда вводится одно из полей элементов метаданных Дублинского ядра.

Композиция полей

Каждое поле содержит следующие элементы, в порядке указания. Также обозначается обязательность данного элемента.

dc-schema (обязательный) – название применяемой схемы, например, для Дублинского ядра – `dc`. Это значение должно совпадать со значением элемента `schema`, определенном в `dublin-core-types.xml`.

dc-element (обязательный) – название элемента ДЯ, вводимого в это поле, например, соисполнитель (`contributor`).

dc-qualifier – квалификатор элемента ДЯ, вводимого в данное поле, например, `contributor.advisor`, значение этого элемента должно быть консультант. Незаполнения означает использование неквалифицируемого ДЯ.

repeatable – принимает значение истина, когда поле может содержать несколько значений, и ложь в противном случае. При указании этого поля как повторяемое (`repeatable`), серверлет пользовательского интерфейса добавит управляющий элемент, чтобы пользователь мог запросить несколько полей для ввода дополнительных значений. Предназначен для использования в произвольных повторяющихся полях, таких как ключевые слова, когда невозможно знать заранее сколько понадобится полей ввода для данного элемента.

label (обязательный) – текст, применяемый для обозначения данного поля, описывающий что нужно ввести, например, «Введите научного руководителя».

input-type (обязательный) – определяет вид интерактивного элемента, который вставляется в форму для сбора значений метаданных ДЯ. Должен принимать одно из следующих значений:

- **onebox** – одно текстовое поле.
- **twobox** – пара простых текстовых полей, используемых для повторяемых значений, например, тема или ключевые слова (`subject`) в ДЯ.
- **textarea** – большой блок для текста, состоящего из нескольких строк, например для резюме.
- **name** – имя с отдельными полями для фамилии и инициалов.
- **date** – календарная дата, требующая, чтобы был внесен по крайней мере год.
- **dropdown** – выбирает значения из ниспадающего списка. Для этого элемента следует также включить значение атрибута `value-pairs-name` для определения списка выбираемых значений меню. Используется для определения ограни-

ченного множества значений, например, для элемента ДЯ язык (`language`).

- **qualdrop_value** – вводит «подходящее значение», включающее как квалификатор из ниспадающего меню, так и произвольное текстовое значение. Используется для ввода полей типа альтернативные идентификаторы и коды для вносимых элементов, например, для элемента ДЯ идентификатор. Как и в случае ниспадающего списка нужно включить значение атрибута `value-pairs-name` для определения списка выбираемых значений меню.

hint (обязательный) – содержит текст, который будет появляться при «наводке», или инструкции, следующие за входными полями. Они могут оставаться пустыми, но все равно должны присутствовать.

required когда этот элемент включен с каким-либо контентом, это означает что это поле является обязательным для ввода. Если пользователь попытается оставить его пустым, появится предупреждающее сообщение.

Например, `<required>` Вы должны ввести название.`</required>`. Если оставить этот элемент пустым, это не отмечает это поле как обязательное для ввода. Например, `<required></required>`.

visibility – когда этот необязательный элемент включается вместе со значением, он ограничивает видимость поля в границах, определенных тем значением. Если элемент отсутствует или пустой, поле видимо при любых диапазонах. В настоящее время поддерживаются следующие границы.

- **workflow**: поле видимо только на стадии рабочего процесса внесения. Это используется в тех случаях, когда нужно скрыть от пользователя трудные поля, такие например, как предметная классификация, тем самым облегчая использование системы внесения.
 - **submit**: поле будет только видимо при первоначальном внесении, а не на стадии рабочего процесса.
- Например, `<visibility>workflow</visibility>`.

Пример описания одного из полей ДЯ:

```
<field>
  <dc-schema>dc</dc-schema>
  <dc-element>title</dc-element>
  <dc-qualifier />
  <repeatable>false</repeatable>
  <label>Название</label>
  <input-type>onebox</input-type>
  <hint>Введите основное название материала</hint>
  <required>Вы должны внести основное название!</required>
</field>
```

В системе предусмотрена возможность подключения предметного классификатора [18].

Автоматическое скрытие полей

Следует отметить, что некоторые поля могут быть автоматически скрыты или пропущены, при отображении пользовательской формы. Это зависит от разновидности вносимых документов и происходит потому, что в соответствии с первоначальным описанием документа программа пользовательского интерфейса DSpace пропускает ненужные поля ДЯ.

Например, если пользователь на первой странице «Описание» отмечает, что не будет альтернативных названий (первый из трех флажков), ввод элемента ДЯ `title.alternative` на пользовательской странице внесения автоматически опускается.

Когда пользователь инициирует внесение, DSpace вначале показывает так называемую «страницу начальных вопросов» (рис. 4), которая содержит три обозначенных флажками вопроса:

1. **Документ содержит больше одного названия, например, на разных языках.**
Управляет полем `title.alternative`.
2. **Документ был опубликован раньше.**
Управляет полями ДЯ:
 - `date.issued`
 - `publisher`
 - `identifier.citation`
3. **Документ состоит из больше чем одного файла.** Не влияет ни на одно входное поле метаданных.

Ответы на первые два вопроса управляют отображением некоторых полей метаданных ДЯ, даже если они определены в качестве полей пользовательской страницы.

И наоборот, если управляемые флажком поля метаданных не определены в пользовательской форме, то для избежания ошибок и недоразумений такой флажок не показан на начальной странице внесения.

Добавление групп значений

Наконец, описание пользовательской формы может потребовать определить «группы значений» для каких-либо полей с входными типами, которые на них ссылаются. Для этого нужно добавить элемент `value-pairs` к содержимому элемента `form-value-pairs`, содержащий атрибуты:

- **value-pairs-name** – имя по которому `input-type` ссылается на этот список.
- **dc-term** – поле квалифицированного ДЯ для которого выбирается значение из этого списка выбора.

Каждый элемент `value-pairs` содержит последовательность пар подэлементов, каждый из которых, в свою очередь, содержит два элемента:

- **displayed-value** – имя, показанное на веб-странице, для ввода меню;
- **stored-value** – запоминаемое значение в элементе ДЯ, когда выбирается данный вход.

В отличие от HTML-тэга `select`, здесь нет возможности обозначить один из входов значением по умолчанию. Поэтому всегда является выбором по умолчанию первое значение.

Пример. Здесь показано меню типов обычных идентификаторов.

```
<value-pairs value-pairs-name="common_identifiers"
dc-term="identifier">
  <pair>
    <displayed-value>№ Прав. Док.</displayed-
value>
    <stored-value>govdoc</stored-value>
  </pair>
  <pair>
    <displayed-value>URI</displayed-value>
    <stored-value>uri</stored-value>
  </pair>
  <pair>
    <displayed-value>ISBN</displayed-value>
    <stored-value>isbn</stored-value>
  </pair>
</value-pairs>
```

Установка пользовательских форм

Веб-приложение DSpace читает пользовательские формы только при запуске, поэтому важно помнить, что следует всегда перезапускать Tomcat (или любой другой используемый контейнер сервлета), чтобы вступили в силу сделанные изменения в файл `input-forms.xml`.

Любые ошибки в синтаксисе или семантике определения формы, типа плохо сформированного XML или ссылки на несуществующие названия полей, вызовет фатальную ошибку в интерфейсе пользователя DSpace.

Электронные документы научно-технической информации в системе ВНТИЦ

© Ю.И. Голосов, Г.А. Брагина, М.Н. Пржиялковская

Всероссийский научно-технический информационный центр (ВНТИЦ)

postmaster@vntic.org.ru

Аннотация

Данное сообщение посвящено системе приема непубликуемых документов научно-технической информации как составляющей системы документооборота Всероссийского научно-технического информационного центра (ВНТИЦ), предназначенного для создания и поддержки интегрального электронного автоматизированного хранилища ВНТИЦ, с предоставлением онлайнового доступа через сеть Интернет к его ресурсам.

Всероссийский научно-технический информационный центр (ВНТИЦ) формирует Федеральный фонд по непубликуемым источникам информации – отчетам о научно-исследовательских и опытно-конструкторских работах, кандидатским и докторским диссертациям, описаниям алгоритмов и программ, осуществляет ведение Единого реестра результатов научно-технической деятельности.

В настоящее время фонд ВНТИЦ насчитывает более 7 миллионов документов. Ежегодные поступления во ВНТИЦ составляют около 100 тысяч различных документов, отражающих контент научно-технической информации (НТИ). Обработка такого количества документов требует значительных финансовых и временных затрат.

Поставщиками информации в коллекцию являются организации науки и высшей школы, промышленные предприятия – исполнители НИОКР, диссертационные советы и авторы диссертаций, а также бюджетополучатели – министерства и ведомства, выступающие государственными заказчиками НИОКР.

Любой ресурс, в том числе и информационный, становится бесполезным и не может быть использован без инструментов систематизированного хранения, доступа и распространения.

Современное состояние ИТ-отрасли позволяет создавать информационные системы любой степени сложности, в том числе и основанные на принципах электронного документооборота. Однако, основная

масса разрабатываемых систем электронного документооборота, в том числе и ориентированных на использование и ведение депозитариев научно-исследовательских проектов и разработок, рассчитаны на использование в одной организации с определенной предметной областью и определенным типом информации, а также ограниченным количеством пользователей.

Система электронного документооборота научно-технической информации в федеральном информационном центре имеет ряд принципиальных особенностей. Прежде всего, это система, рассчитанная на прием, обработку, хранение и распространение больших объемов информации – несколько сотен тысяч документов в год. При этом объем документов колеблется от 3 Кб до 250 Мб. Информация, представленная в документах, не структурирована либо слабо структурирована, документы относятся к различным областям знаний, т.е. фонд политематичен.

Функционально система электронного документооборота научно-технической информации в Федеральном информационном центре должна решать следующие задачи:

1. Подготовка и прием электронных форм документов в интерактивном режиме;
2. Смысловая обработка документов (индексирование, рубрицирование, редактирование и т.д.);
3. Организация электронного хранилища данных с возможностью онлайнового доступа пользователей;
4. Выпуск электронных изданий;
5. Прием и выполнение заказов на копии первоисточников;
6. Аналитическая обработка и статистический анализ контента фонда;
7. Страхование копирования и хранение документов.

Традиционно, документы, представляемые во ВНТИЦ, поступали только на бумаге и, пройдя определенные стадии обработки, трансформировались в электронный вид. По мере развития компьютерных технологий и внедрения их в технологический процесс формирования федерального фонда по непубликуемым источникам информации, различные операции по обработке, хранению и распространению информации автоматизировались и модернизировались, создавая основу для системы электронного документооборота.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Было создано интегральное электронное автоматизированное хранилище ВНТИЦ, включающее банк данных государственных контрактов на НИОКР, политематические ретроспективные реферативно-библиографические базы данных по государственной регистрации и учёту НИОКР и диссертаций, а также хранилище полнотекстовых отчётов и диссертаций с реализацией организации онлайн-ового доступа пользователей через сеть Интернет к ресурсам электронного хранилища. Однако, операции по приему документов до настоящего времени оставались неизменны.

Прием электронных форм документов в сетевом режиме имеет ряд особенностей:

1. Прежде всего, в существующей практике юридическую силу по-прежнему имеют документы на бумажном носителе, поэтому наряду с электронной формой документа необходимо иметь его подтверждение на бумаге, оформленное по всем правилам, предусмотренным для такого рода документов.

2. В Российской Федерации существуют документы, регламентирующие бумажное представление документов научно-технической информации (ГОСТ 7.32-2001 от 1.07.2002 – Отчет о научно-исследовательской работе, структура и правила оформления), но отсутствуют стандарты или рекомендации, регламентирующие электронное представление как текстов документов научно-технической информации, так и их нетекстовых частей (формул, графических элементов, фотографий и т.д.).

3. Значительный объем полнотекстового документа (отчёта о НИОКР, диссертации) – до 250 Мб.

В настоящее время во ВНТИЦ разработана первая очередь системы электронного документооборота НТИ, включающая как технологию и средства приема, так и обработки, хранения и распространения реферативной информации.

Прием документов в электронном виде (регистрационных, информационных карт НИОКР, информационных карт диссертаций) реализован с использованием технологии ASP.NET. Доступ к системе осуществляется через сайт ВНТИЦ – www.vntic.org.ru. Далее принятые документы поступают в технологическую базу, обрабатываются и загружаются в электронное автоматизированное хранилище ВНТИЦ. Первоисточники – диссертации, тексты НИР и ОКР – сканируются и также поступают в электронное хранилище.

В 2009 году планируется проведение разработки второй очереди системы документооборота, ориентированной на работу с полнотекстовыми документами.

Литература

- [1] ГОСТ 7.32-2001 Межгосударственный стандарт. Отчет о научно-исследовательской работе (Структура и правила оформления).
- [2] Богданов Ю.М., Павлов Л.П. и др. Отчет НИОКР «Развитие информационной системы и

технологий формирования, ведения и распространения информации Единого реестра результатов научно-технической деятельности и государственного фонда неопубликованных документов», 2007 г., с 436, ВНТИЦ.

- [3] Л.Г. Титарев, Ю.М. Богданов и Л.П. Павлов “Digital Documents in Grey Literature : New Challenges ”, доклад, с. 7, Антверпен (Бельгия), 9-я Международная конференция по серой литературе GL 9.

Electronic Documents of Scientific and Technical Information in VNTIC System

Y.I. Golosov, G.A. Bragina, M.N. Przyjalkovskaya

The report describes the system of receiving of scientific and technical information documents, as the part of The Scientific and Technical Information Centre of Russia (the Russian acronym – VNTIC) documents circulation system, intended to provide the generation and support of VNTIC integral electronic documents repository, provided with Internet online information search.

Детерминированное и случайное в отождествлении результатов астрометрических наблюдений пульсаров*

© А.Е. Авраменко

Пушчинская радиоастрономическая обсерватория ФИАН

avr@prao.ru

Аннотация

Проведено сопоставление численных результатов отождествления измеряемых физических величин по двум моделям наблюдательных данных хронометрирования пульсаров. Показаны преимущества параметрической модели данных, на несколько порядков превосходящей по точности общепринятую разностную модель. Рассмотрены особенности отождествления по группам временных рядов, полученных в наблюдениях ансамбля пульсаров, в том числе на разных инструментах. Приведены обобщенные оценки достижимой точности для обеих моделей. Показаны примеры приложений коллекции пульсарных данных применительно к задачам рассматриваемой предметной области.

1 Введение

Высокостабильная периодичность импульсного излучения энергии пульсаров определила интерес к ним как к возможным альтернативным хранителям времени, не зависящим от земных условий и обладающим практически неограниченным энергетическим ресурсом. Измерение пульсарного времени основано на динамической модели движения небесных тел, которая определяет их пространственные координаты (интервалы) как функции независимой переменной – времени, которая в свою очередь определяется последовательностью интервалов наблюдаемых пульсарных событий [1].

Обсерватории, располагающие инструментами для наблюдения импульсного излучения пульсаров, проводят их регулярное хронометрирование и накапливают данные в течение многих лет. Наблюдательные данные хронометрирования формируются и поддерживаются, в соответствии с принятым форматом Принстонского университета [2], в виде унифицированных записей, содержащих метаданные, которые отражают условия наблюдений, характеристики инструмента, установочные параметры, и собственно численные результаты наблюде-

ний. Такой набор фиксируемых данных в многолетних коллекциях принят всеми обсерваториями для хранения и обмена. В численном представлении данные хронометрирования содержат:

- моменты прибытия импульсов (МПИ) на дату наблюдения в начало выбранной координатной системы, в качестве которого обычно принимают фазовый центр радиотелескопа или барицентр Солнечной системы;
- остаточные отклонения – разность между расчетными и наблюдаемыми барицентрическими МПИ;
- среднеквадратичные отклонения (СКО) остаточных отклонений МПИ.

По результатам хронометрирования определяются фундаментальные свойства материи, уточняются численные величины физических констант путем выявления влияния физических факторов на условия распространения импульсов излучения пульсара, проявляющихся в виде изменений МПИ. Вычисляются и регулярно уточняются пределы энергетической плотности стохастического фона гравитационного излучения на сверхнизких частотах [2], ведется мониторинг вариаций меры частотной дисперсии МПИ – величины концентрации электронов в ионизированной межзвездной среде в направлении на пульсар [3].

Вопрос точности определения МПИ становится центральным в подобных задачах, от него непосредственно зависит достоверность оценок физических величин. Вместе с тем, принятая оценка по остаточным отклонениям, определяемым как разность расчетных и наблюдаемых МПИ в пределах текущих суток на дату наблюдения, не учитывает абсолютных величин последовательности интервалов, отсчитываемых от общего начала в пределах всего промежутка наблюдений. Расчетные МПИ, по которым вычисляются остаточные отклонения, определяются по заранее принятым значениям параметров, которые напрямую не связаны с результатами текущих наблюдений и априори известны лишь с некоторой конечной точностью. Наблюдаемые МПИ, в свою очередь, подвержены вариациям инструментального и методического происхождения, величина которых сопоставима с остаточными отклонениями и представляет собой статистическую оценку погрешности хронометрирования каждого наблюдаемого события.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Указанное противоречие было разрешено путем преобразования выборочно наблюдаемых МПИ в последовательность интервалов, определяемых через параметры вращения пульсара, численные значения которых в свою очередь находятся приближением по критерию наименьших квадратов к последовательности наблюдаемых интервалов, которые формируются из наблюдаемых МПИ [4]. В результате выборочные данные трансформируются в дискретный континуум пульсарных событий, который представляет собой последовательность интервалов по всем излучаемым импульсам пульсара в пределах протяженности наблюдений. Интервалы пульсарного времени, детерминированные параметрами вращения пульсара, выделены в явном виде из наблюдаемых интервалов путем декомпозиции их на две последовательности: одна из них определяет собственно пульсарное время, а другая – непараметризуемый остаток, который представляет собой вариации наблюдаемых интервалов, вносимые хронометрированием, средой, другими внешними факторами нестабильности.

Для приведения формата наблюдательных данных в соответствие с их параметрическим представлением потребовалось учесть особенности представления интервалов пульсарного времени, детерминированных параметрами вращения пульсара в пределах протяженности наблюдений. Исходные наблюдательные данные, кроме содержащихся в них МПИ в пределах суток на дату наблюдения, которые определены форматом Принстонского университета, дополнены значениями интервалов между соседними наблюдаемыми событиями, по которым находятся величины суммарных интервалов от начального до любого события в пределах протяженности наблюдений. В коллекцию включены также установленные по наблюдаемым интервалам численные величины параметров вращения пульсара и случайные, не параметризуемые моделью хронометрирования, вариации наблюдаемых интервалов [5].

Благодаря трансформации наблюдательных данных к параметрическому виду, представляющему собой дискретный континуум пульсарных событий в пределах протяженности наблюдений, который не зависит критически от случайных вариаций наблюдаемых интервалов, стабильность интервалов пульсарного времени, синтезированного по параметрам вращения пульсара, оказалась несопоставимо, на несколько порядков выше, чем в последовательности исходных МПИ и их остаточных уклонений.

В связи с этим возникает задача рассмотреть, каким образом новые возможности коллекции пульсарных данных отразятся на результатах отождествления наблюдаемых физических величин применительно к приложениям предметной области. Во-первых, требуется определить численные отношения между результатами измерений физических величин и их вариациями, с одной стороны, и вычисленными значениями, изменениями наблюдаемых параметров вращения пульсара, по которым

определяются физические величины, с другой стороны. Во-вторых, требуется найти такие подходы и средства, которые реализуют достижимую точность определения наблюдаемых параметров вращения пульсаров, с учетом их уникальной долговременной стабильности и принципиально более высокой устойчивости к случайным вариациям наблюдаемых интервалов.

На решение этой двуединой задачи направлена настоящая работа. Будет проведено сопоставление результатов измерений наблюдаемых физических величин разностным методом – по остаточным уклонениям для каждого наблюдаемого события, а также по дискретному параметризованному континууму пульсарных событий в пределах протяженности наблюдений. Должны быть получены сравнительные оценки погрешностей результатов по обоим вариантам. Будут рассмотрены методы и средства формирования последовательностей параметризованных временных рядов коллекции, отвечающей требованиям минимизации влияния вариаций наблюдаемых интервалов.

Особое внимание уделено правилам отбора наблюдательных данных по критерию достоверности отождествления результатов наблюдений, а также выявлению признаков наблюдательных данных, не совместимых с параметрической моделью и ограничивающих их использование для измерений. Будут рассмотрены условия и признаки согласованности данных по группе (или ансамблю) наблюдаемых пульсаров, включая наблюдения одних и тех же пульсаров на различных инструментах.

В результате должны быть выявлены условия и количественные оценки наилучшего соответствия временных рядов пульсарного времени по данным хронометрирования для решения проблемно-ориентированных задач предметной области.

2 Отождествление физических величин по наблюдательным данным

Целью хронометрирования является получение количественных характеристик наблюдаемых и взаимодействующих с ними физических объектов по результатам измерений, содержащихся в численных рядах наблюдательных данных. От того, какие данные и в каком виде берутся в качестве исходных и какие применяются правила, отношения, связывающие свойства физических объектов и их признаки по исходным наблюдательным данным, зависят результаты отождествления, достоверность и надежность выявленных свойств.

Сопоставим с этой точки зрения две модели данных хронометрирования – разностную, в основу которой положены остаточные уклонения МПИ наблюдаемых пульсарных событий [2], и параметрическую, в которой наблюдательные данные представлены в виде последовательности интервалов, детерминированных наблюдаемыми параметрами вращения пульсара [5]. Выделим в рассматриваемых моделях те особенности, которые определяют

специфику выявления свойств наблюдаемых объектов, их отношения с наблюдаемыми признаками.

2.1 Две модели наблюдательных данных

Структура барицентрических наблюдательных данных в формате Принстонского университета на основе разностной модели хронометрирования приведена в табл. 2.1.

Таблица 2.1. Фрагмент архива пульсарных данных Принстонского университета [2]

PSR	TOA_bry (MJD)	resid (μs)	err (μs)	freq_bry (MHz)	Del_DM (cm-3 pc)
1937 + 21	47541.6518	0.686	0.34	2385.0196	0.000926
1937 + 21	47544.6458	1.126	0.32	2370.0103	0.000926
1937 + 21	47556.6093	1.207	0.16	2374.9734	0.000793
1937 + 21	47586.5358	0.925	0.14	2389.8897	0.000588
1937 + 21	47613.4625	0.549	0.14	2379.8407	0.000456
1937 + 21	47664.3236	0.764	0.22	2364.8436	0.000197
1937 + 21	47676.2990	0.456	0.20	2374.8611	0.000232
...	...	...	...	...	...

Наблюдательные данные представляет собой последовательность однотипных записей, каждая из которых включает набор численных величин, относящихся к текущей дате наблюдений MJD. Это прежде всего остаточные отклонения МПИ (resid) и их статистическая оценка СКО (err). Кроме того, приведены измеренные отклонения Del_DM меры дисперсии DM, сопоставленные с частотой излучения энергии пульсара freq_bry. Две последние величины учитываются при вычислении остаточных отклонений методом подгонки (фиттинга) параметров пульсара и среды распространения импульсов излучения.

Параметрическая модель данных [4] определяет связь интервалов наблюдаемых импульсов пульсара с периодом вращения пульсара и его производной:

$$\sum_i dTB_i = (1 + \alpha_i)(1 + 0,5\frac{\dot{P}}{P})P_o^*N + R_i \quad (1)$$

$$P_o = (1 + \alpha_i)P_o^*, \quad (2)$$

где $\sum_i dTB_i$ – наблюдаемые в барицентре Солнечной системы интервалы от начального до i -го наблюдаемого события, P_o , $\frac{\dot{P}}{P}$ – наблюдаемый в барицентре период вращения пульсара и его производная, N – общее число излученных импульсов пульсара от начального до i -го наблюдаемого, P_o^* – расчетное значение периода на начальную эпоху наблюдений, $(1 + \alpha_i)$ – коэффициент линейного приближения методом наименьших квадратов наблюдаемых интервалов к параметрическому виду, R_i – непараметризуемый остаток.

Параметрическая модель оперирует целостным представлением всей последовательности интервалов пульсарного времени, которые определяются численными параметрами вращения пульсара и отсчитываются от одного и того же начального собы-

тия. В явном виде выделен непараметризуемый остаток наблюдаемых интервалов R_i (1). Детерминированная параметрическая составляющая интервалов зависит от коэффициента $(1 + \alpha_i)$, который характеризует величину отклонений наблюдаемого периода вращения и таким образом представляет собой обобщенный количественный показатель неустойчивости наблюдаемого периода и соответственно пульсарного времени в пределах всей протяженности наблюдений. Он учитывает последствия влияния всех факторов неустойчивости, включая вариации меры дисперсии, и служит непосредственной мерой погрешности наблюдаемого периода и определяемых им интервалов пульсарного времени под влиянием этих факторов. Очевидно, в соответствии с (1), что условием идеальной стабильности является нулевая величина непараметризуемого остатка и, как следствие, постоянство коэффициента $(1 + \alpha_i)$ и соответственно периода P_o в пределах всей протяженности наблюдений. По степени отклонения этого коэффициента от единичного значения определяется реальная стабильность наблюдаемых интервалов и их детерминированной составляющей – пульсарного времени.

Благодаря введению интервалов пульсарного времени в параметрическую модель наблюдательных данных достигается, во-первых, переход от разностных характеристик МПИ отдельных наблюдаемых событий и их статистических характеристик к последовательности интервалов в виде дискретного континуума, детерминированного стабильными параметрами вращения пульсара. Во-вторых, параметрическая модель содержит в явном виде и определяет отклонение периода – времязадающего параметра модели пульсарных данных, что позволяет получить непосредственные величины погрешностей пульсарного времени и учитывать их последствия в контексте задач предметной области.

2.2 Отображение физических величин в моделях данных

В этом разделе рассмотрены особенности отображения меры частотной дисперсии DM, которая характеризует частотно-зависимое запаздывание моментов импульсов при распространении излучаемой энергии пульсара, зависящее от концентрации электронов в ионизированной межзвездной среде в направлении на пульсар.

В работе [2] было обнаружено монотонное уменьшение меры дисперсии в направлении на пульсар В1937+21 в пределах 8-летнего промежутка наблюдений (1985–1993 гг.) со скоростью порядка $0,001 \text{ пк-см}^{-3}/\text{год}$ (рис. 2а). С другой стороны, запаздывание моментов импульсов точно так же может быть обусловлено изменением наблюдаемого периода вращения пульсара. Оба эти параметра непосредственно влияют на расчетную величину вращательной фазы и соответственно на запаздывание наблюдаемого импульса пульсара. Поэтому измене-

ние во времени одной из этих величин можно сопоставить с поведением другой, и по результатам сделать вывод о влиянии той или другой величины на запаздывание наблюдаемых импульсов. С этой целью по разностной модели определялись отклонения наблюдаемых интервалов, вызванные изменением меры дисперсии, которые сопоставлялись с изменением рассчитанного по параметрической модели значения периода вращения пульсара и вызванного им запаздывания по наблюдаемым интервалам импульсов пульсара.

Дополнительное запаздывание импульса излучения на частоте f пропорционально изменению меры дисперсии и обратно пропорционально квадрату частоты [2]:

$$\Delta t = \frac{\Delta DM}{2,41 \cdot 10^{-16} f^2}. \quad (4)$$

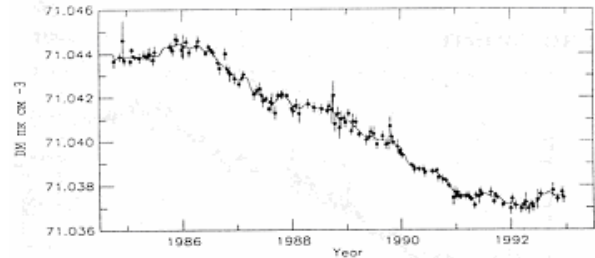
Отклонение интервалов пульсарного времени, рассчитанное по параметрической модели, в соответствии с (2), пропорционально относительному изменению наблюдаемого периода вращения, численно равное отклонению коэффициента линейного приближения в промежутке наблюдения, и величине этого промежутка t :

$$\Delta t = -\frac{\Delta P}{P} t = -\alpha t. \quad (5)$$

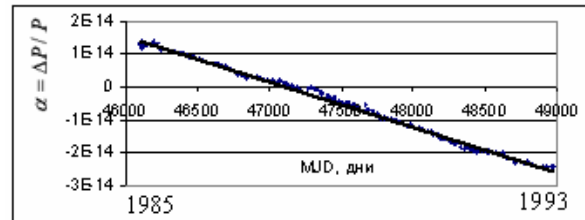
На рис. 2.1 приведены зависимости меры дисперсии, расчетного периода и отклонения интервалов в совпадающем промежутке наблюдений. На рис. 2.1а показана зависимость меры дисперсии В1937+21 во времени из работы [2], рассчитанная по разностной модели данных. На рис. 2.1б приведено изменение величины $\alpha = \Delta P / P$ на 8-летней протяженности наблюдений. На рис. 2.1в показано отклонение интервалов, определяемых по изменениям величины $\alpha = \Delta P / P$, которое рассчитано в соответствии с (5). Результаты, приведенные на рис. 2.1б, в, получены по тем же наблюдательным данным, что и на рис. 2.1, а [2], но с использованием параметрической модели данных [4].

На приведенных графиках видна выраженная корреляция отклонений периода вращения пульсара (рис. 2.1б) и изменений меры дисперсии (2.1а). Сопоставление полученных двумя независимыми методами отклонений интервалов также подтверждает их достаточно близкое совпадение. Величина отклонения интервалов за счет изменения периода вращения составляет около $8,7 \cdot 10^{-6}$ с за 7,4 года, а расчетная величина смещения МПИ на частоте наблюдения около 2,4 ГГц (табл. 2.1) за счет изменения меры дисперсии по приведенному на рис. 2.1а результату – около $6 \cdot 10^{-6}$ с в том же промежутке. Некоторое расхождение вычисленных величин можно объяснить тем, что определяемые двухчастотным методом значение меры дисперсии и ее изменения не исключает погрешностей, обусловленных многими факторами, в том числе значительной частотной зависимостью [6]. Из приведенного в табл. 2.1 фрагмента архива наблюдательных дан-

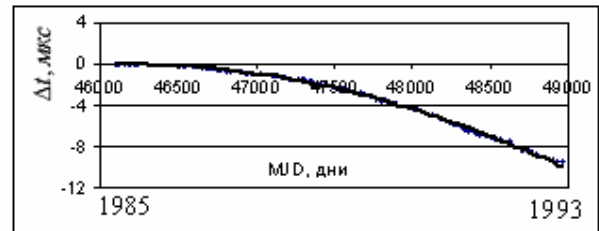
ных, по которым были получены рассматриваемые результаты, следует, что расчетные изменения меры дисперсии даже в небольшом, около месяца, промежутке имеют приблизительно пятикратный статистический разброс, сопоставимый по величине с обнаруженным в [2] систематическим линейным отклонением меры дисперсии на годовом промежутке.



а) зависимость меры дисперсии В1937+21 во времени (по данным работы [2])



б) изменение расчетной величины периода вращения В1937+21 и линейная аппроксимация (сплошная линия)



в) отклонение интервалов, обусловленное изменением расчетной величины периода вращения В1937+21, и его аппроксимация полиномом второй степени (сплошная линия)

Рис. 2.1. Мера дисперсии, расчетный период, отклонение интервалов

Сопоставление результатов, полученных по двум моделям данных, показывает, что выявленное в [2] разностным методом линейное изменение меры дисперсии не согласуется с поведением наблюдаемого периода вращения пульсара. Действительно, нарастающее отклонение вращательной фазы пульсара вследствие линейного изменения меры дисперсии компенсируется смещением на некоторую постоянную величину наблюдаемого периода вращения пульсара. Однако выявленное параметрической моделью линейное изменение наблюдаемого периода вращения пульсара, коррелированное с изменением меры дисперсии, может свидетельство-

вать о том, что закон изменения меры дисперсии не ограничивается только линейной зависимостью, но содержит также составляющие более высоких порядков, в первую очередь квадратичную, которая не принята во внимание и может стать причиной линейного изменения расчетного периода вращения пульсара, не совместимого с параметрической моделью наблюдаемых интервалов.

2.3 Корректировка параметрической модели по наблюдательным данным

Приведенные на рис. 2.1 результаты показывают высокую степень зависимости изменений расчетного периода вращения пульсара и меры дисперсии в пределах протяженности наблюдений. Их отношение, характеризующее отклик расчетного периода на изменение меры дисперсии, составляет около $0,5 \cdot 10^{-11} \text{ пк}^{-1} \cdot \text{см}^3$, что показывает чрезвычайно высокую чувствительность расчетного периода вращения к изменениям меры дисперсии. Отклонение расчетного периода при изменении меры дисперсии в пределах погрешности измерения разностным методом вызывает отклонение расчетного периода, которое на 3-4 и более порядка превышает его предельно различимую параметрической моделью величину.

Свойство параметрической модели данных уверенно выявлять даже небольшие, на несколько порядков меньше погрешностей разностной модели, отклонения наблюдаемых признаков позволяет с высокой точностью учитывать в параметрической модели данных устойчивые систематические отношения взаимосвязанных величин, в том числе и тех, которые не относятся непосредственно к рассматриваемой предметной области. В приведенном примере, когда внесенное изменение величины меры дисперсии приводит к отклонениям наблюдаемых интервалов, не согласующимся с параметрами модели (1), можно ввести коррекцию модели, включив в нее выявленную зависимость расчетного периода, коррелированного с изменениями меры дисперсии по времени.

С этой целью, выделив из (1) параметризованный компонент пульсарного времени, введем дополнительно линейную функцию, отражающую зависимость коэффициента α_i от величины интервалов пульсарного времени для компенсации выявленного изменения расчетного периода вращения пульсара. Пульсарное время из (1):

$$PT_i = (1 + \alpha_i)(1 + 0,5PN)P_o^* N. \quad (6)$$

Скорректированное пульсарное время:

$$PT_i = (1 + \alpha_i(1 + \delta)PT_i)(1 + 0,5PN)P_o^* N. \quad (7)$$

Изменение периода:

$$\frac{\Delta P}{P} = 1 + \alpha_i(1 + \delta)PT_i. \quad (8)$$

Вычисленное значение корректирующего коэффициента $\delta = 1,65 \cdot 10^{-22} \text{ с}^{-1}$.

Заметим попутно, что значение PT_i , вычисленное в соответствии с (7), формально от него же и зависит. Это, тем не менее, не вносит некорректность, поскольку величина корректирующей поправки, численно равная $\alpha_i \delta PT_i$, чрезвычайно мала по сравнению с α_i , и для расчета PT_i в (7) вполне можно было использовать наблюдаемые интервалы $\sum_i dTB_i$ (1), для численного результата (7) заметной разницы нет.

На рис. 2.2 показано изменение расчетной величины периода вращения после коррекции коэффициента α_i в параметрической модели данных. В результате корректировки коэффициента линейного приближения расчетная величина периода вращения пульсара B1937+21 не обнаруживает монотонного отклонения в пределах промежутка наблюдений, предположительно вызванного внесенной погрешностью измерения меры дисперсии. Уточнение модели данных позволило исключить влияние этой погрешности на последовательность интервалов пульсарного времени, детерминированных наблюдаемыми параметрами вращения пульсара.

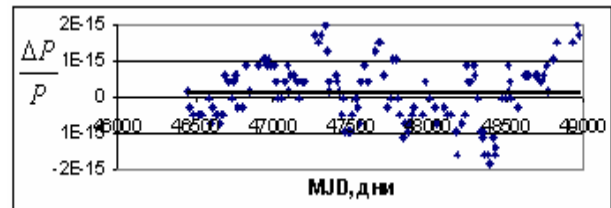


Рис. 2.2. Изменение расчетной величины периода вращения B1937+21 и линейная аппроксимация (сплошная линия) после коррекции модели данных

Для окончательного вывода о причинах изменения расчетного периода следовало бы переопределить барицентрические МПИ для более сложной, чем линейная, модели изменений меры дисперсии и сопоставить полученные по ним значения расчетного периода вращения пульсара.

3 Отождествление в связанных группах данных

Как было показано, вследствие статистической неопределенности наблюдательных данных значительных и вариаций наблюдаемых признаков, результаты отождествления могут допускать неоднозначную интерпретацию. Поэтому, с целью повышения точности и достоверности выявляемых свойств, обычно не ограничиваются наблюдением только одного пульсара, используют данные групповых наблюдений, с возможностью сопоставления и совместной оценки получаемых результатов. Рассмотрим кратко возможности наблюдений группы

(ансамбля) пульсаров применительно к параметрической модели данных.

3.1 Отождествление по ансамблю пульсаров

Обычно применяемая разностная модель данных в наблюдениях нескольких пульсаров в пределах общего промежутка позволяет снизить влияние случайных факторов и получить более стабильные результаты по усредненным с весовыми коэффициентами оценкам остаточных уклонений по группе наблюдаемых пульсаров [7]. Параметрическая модель данных, которая в явном виде определяет независимый временной ряд интервалов для каждого пульсара в группе, позволяет выполнить сопоставление наблюдаемых признаков и свойств временных рядов в ансамбле по абсолютным значениям интервалов.

На нескольких примерах покажем возможности параметрической модели данных для отождествления по группам наблюдательных рядов хронометрирования пульсаров.

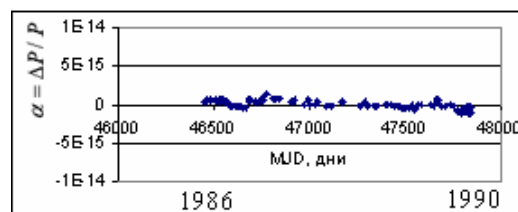
Одно из основных приложений хронометрирования пульсаров – выявить свойства пульсаров как высокостабильных хранителей времени и определить методику сопоставления пульсарного времени с атомными часами с целью их синхронизации по пульсарам. Представление наблюдательных данных в виде последовательности интервалов сближает пульсарное время по форме с атомным измерительным эталоном, который формирует последовательность единичных интервалов времени для хронометрирования пульсаров. Одной из вероятных причин выявленных изменений расчетного периода по наблюдениям пульсара B1937+21 (рис. 2.1б) могло стать изменение масштаба единицы величины (секунды) измерительного атомного эталона, по которому были определены интервалы пульсарного времени. Однако сделать с уверенностью такое заключение по наблюдениям только одного пульсара невозможно, тогда как сопоставлением результатов наблюдений другого пульсара в этом же промежутке и на том же радиотелескопе по одному и тому же измерительному атомному эталону оказывается возможным подтвердить или опровергнуть сделанные предположения.

На рис. 2.3 показаны изменения расчетного периода вращения и отклонения наблюдаемых интервалов пульсара B1855+09 в промежутке 1986–1990 гг. Тогда же и на том же инструменте наблюдался и пульсар B1937+21.

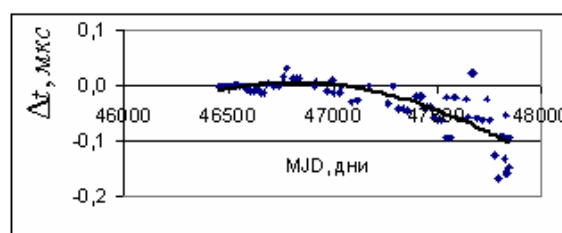
Сопоставление рис. 2.1 и рис. 2.3 показывает, что B1855+09 не обнаруживает изменений расчетного периода, подобных B1937+21. Следовательно, нет оснований считать причиной изменения расчетного периода B1937+21 отклонение единицы измерительного атомного эталона, по которому проводилось хронометрирование обоих пульсаров.

Наблюдения B1855+09 не противоречат сделанному предположению, что причиной выявленного изменения периода пульсара B1937+21 в виде линейного тренда стала погрешность в определении меры дисперсии, которая была учтена путем кор-

рекции модели наблюдательных данных, в результате чего вариации расчетного периода B1937+21 были приведены к стационарному виду (рис. 2.2). Причины изменения расчетного периода B1937+21, которые, очевидно, не связаны с погрешностями измерительного атомного эталона, должны быть выяснены дополнительно, принимая во внимание высказанные предположения относительно характера зависимости меры дисперсии.



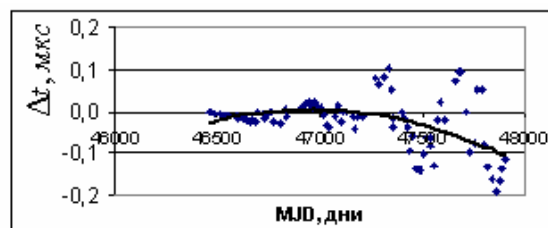
а) изменение расчетной величины периода вращения



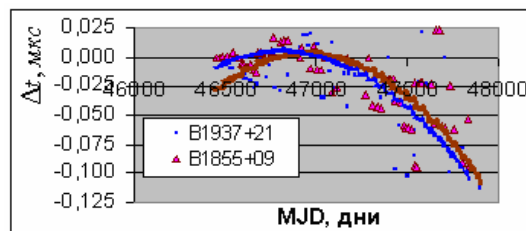
б) отклонение интервалов

Рис. 2.3. Расчетный период, отклонение интервалов пульсара B1855+09

На рис. 2.4 показаны сглаженные аппроксимирующим полиномом второго порядка отклонения интервалов пульсарного времени для B1937+21 и B1855+09 в пределах общего промежутка наблюдений около 4-х лет. Видно, что сглаженные отклонения практически совпадают для обоих пульсаров как по форме изменения во времени, так и по величине.



а) отклонение интервалов пульсара B1937+21



б) отклонение интервалов B1855+09 и B1937+21

Рис. 2.4. Отклонение интервалов пульсаров B1937+21 и B1855+09 в общем промежутке наблюдений

Принимая во внимание, что вариации расчетной величины относительного периода вращения пульсаров В1937+21 (после коррекции параметрической модели данных) и В1855+09 находятся около 10^{-15} (рис. 2.2, 2.3а), что близко к достижимому пределу стабильности лучших атомных стандартов, вполне допустимо, с учетом очень близкого совпадения отклонений интервалов для обоих пульсаров, интерпретировать эти отклонения как результат влияния нестабильности атомного измерительного эталона, используемого для хронометрирования обоих пульсаров. Средняя величина полученных таким образом по параметрической модели отклонений интервалов является расхождением атомных часов относительно пульсарного времени и служит для них численной синхронизирующей поправкой, погрешность которой находится в пределах 5-10 нс в 4-летнем промежутке наблюдений.

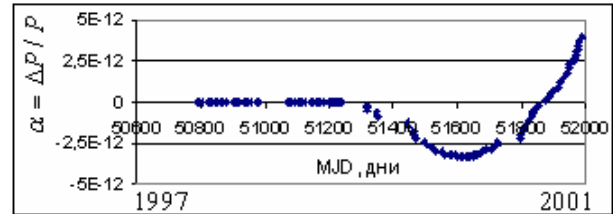
3.2 Отождествление по наблюдениям на разных инструментах

Не всегда рассчитанные по параметрической модели вариации наблюдаемого периода оказываются так близки по величине к своим предельно достижимым значениям, как в рассмотренном примере с пульсарами В1937+21 и В1855+09. Причины могут быть разные: инструментальные погрешности наблюдений, методические погрешности обработки данных, неточные априорные знания численных значений используемых при многопараметрической подгонке величин. Остается открытым вопрос о собственной стабильности вращения пульсаров. Известно, что у некоторых пульсаров период вращения может меняться во времени, и достаточно резко (глитчи). Характер этих изменений не укладывается в общепринятую модель хронометрирования пульсаров, и требуется дополнительное исследование таких особенностей поведения пульсаров. Ясно, что пульсары в промежутках аномальных отклонений периода не могут рассматриваться как полноценные хранители времени, и включение их в ансамбль для отождествления результатов наблюдений внесло бы дополнительные погрешности, снижающие общую эффективность использования группы пульсаров.

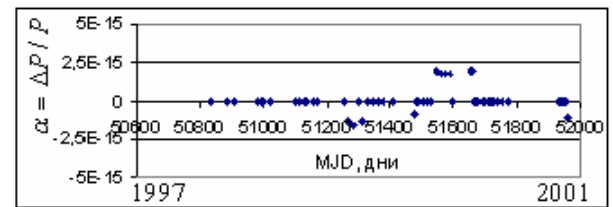
Как было показано на примере пульсара В1937+21, по наблюдениям только одного пульсара невозможно однозначно определить причину аномального отклонения расчетного периода вращения, можно лишь сделать предположение на этот счет. По наблюдениям группы из нескольких пульсаров можно обнаружить совпадающие по характеру и величине отклонения, которые имеют общую причину для группы, например, возможные вариации измерительного атомного эталона, выявляемые по нескольким пульсарам. Обнаруженное аномальное поведение пульсара может быть подтверждено по данным наблюдений этого пульсара на другом инструменте в том же промежутке времени. По результатам сопоставления изменений величины расчетного периода по наблюдениям на обоих (не-

скольких) инструментах делается заключение, подтверждающее или опровергающее предполагаемое аномальное поведение этого пульсара.

На рис. 2.5 в качестве примера показаны изменения расчетного периода вращения пульсара В1937+21 по наблюдениям, проведенным в течение 1997–2001 гг. на радиотелескопах РТ-64 в Калязине и РТ-32 в Кашиме (Япония). Результаты хронометрирования пульсара В1937+21 и их сравнительного анализа приведены в работе [8].



а) по наблюдениям В1937+21 в Кашиме



б) по наблюдениям В1937+21 в Калязине

Рис. 2.5. Изменение расчетной величины периода вращения В1937+21

Сопоставление приведенных на рис. 2.5 графиков показывает, что отклонения расчетной величины периода вращения пульсара В1937+21 по наблюдениям в Кашиме в промежутке MJD 51300–52000 приблизительно на три порядка превышают отклонения периода по наблюдениям в Калязине в том же промежутке. Расчетная величина периода вращения по наблюдениям в Калязине не обнаруживает заметных отклонений в промежутке MJD 51300–52000, наблюдаемые вариации периода хорошо согласуются с характеристиками предельно достижимой стабильности атомного эталона порядка 10^{-15} в общем промежутке наблюдений. Следовательно, нет достаточных оснований считать поведение пульсара В1937+21 аномальным в этом промежутке, а причины значительных отклонений расчетного периода вращения, полученных по наблюдениям в Кашиме, следует уточнить дополнительно.

Таким образом, связанные параметрической моделью группы наблюдательных данных предоставляют широкие возможности отождествления свойств пульсаров как высокостабильных хранителей времени. Из группы пульсаров, наблюдаемых на одном инструменте, отбираются пульсары по критерию стационарности вариаций расчетного периода в диапазоне, не выходящем за пределы погрешностей измерительного атомного эталона, и исключать из группы пульсары с аномальным отклонением.

ем наблюдаемых интервалов, численные значения которых не укладываются в диапазон, определяемый параметрической моделью хронометрирования.

По данным наблюдений одних и тех же пульсаров на разных инструментах выявляются характеристики долговременной стабильности параметров вращения пульсаров, проверяется достоверность аномальных отклонений интервалов и расчетного периода, которые обнаруживаются в наблюдениях.

4 Случайные вариации и достижимые пределы стабильности

На основе сопоставления двух моделей наблюдательных пульсарных данных на рассмотренных примерах приложений, приведем в качестве некоторого итога оценку порядка величин случайных вариаций наблюдаемых интервалов, которые определяют характерные значения погрешностей и соответственно достижимые пределы точности и долговременной стабильности по результатам их отождествления.

Затем кратко остановимся на особенностях реализации сценариев отождествления по наблюдениям связанных групп пульсаров с применением параметрической модели.

4.1 Обобщенная оценка неопределенности отождествления

В табл. 4.1 систематизированы оценки порядка величин погрешностей измеряемых интервалов пульсарного времени, которые определяют достижимые пределы точности отождествляемых величин для различных моделей данных и условий измерения.

Таблица 4.1. Случайные вариации и погрешность интервалов по наблюдениям пульсаров

Случайные вариации величин	Погрешность интервалов (1 год), секунд	Примечание
1. Вариации DM 0,001 пк·см-3/год DM = 71,040 пк·см-3	10 ⁻⁵	Хронометрирование разности фаз B1937+21 Частоты 1,4; 2,4 ГГц
2. СКО МПИ 10 ⁻⁶ с	10 ⁻⁶	Разностная модель МПИ
3. $\alpha = \Delta P / P$ = xx·10 ⁻¹⁵	10 ⁻⁷	Параметрич. модель: синтез периода P
4. $\alpha = \Delta P / P$ = xx·10 ⁻¹⁶	10 ⁻⁸	Пульсар в группе (ансамбле)
5. $\alpha = \Delta P / P$ = xx·10 ⁻¹⁷	10 ⁻⁹	С учетом вариаций атомного эталона
6. Предел точности параметрической модели	< 10 ⁻⁹	Уточнение параметрической модели

В левой колонке указаны вариации величин, которые определяются принятой моделью наблюдательных данных, а в средней колонке приведены обобщенные оценки погрешностей наблюдаемых интервалов. Вариации меры дисперсии, измеренные по разности фаз наблюдаемых МПИ пульсара B1937+21 [9], имеют неопределенность, сопоставимую с обнаруженным в работе [2] систематическим уменьшением меры дисперсии в промежутке 1 год. Среднеквадратичные отклонения разности измеренных и наблюдаемых МПИ, характерные значения которых находятся в диапазоне нескольких микросекунд, ограничивают этим порядком величины предел достижимой стабильности разностной модели данных.

Переход к параметрической модели пульсарных данных, основанной на синтезе наблюдаемого периода в пределах промежутка наблюдений, на несколько порядков увеличивает достижимый предел точности измерения интервалов. Особенно эффективна параметрическая модель применительно к группе пульсаров. В этом случае появляется возможность определить собственные вариации и систематические отклонения измерительного атомного эталона и произвести коррекцию обусловленных ими погрешностей хронометрирования. Определенный резерв снижения погрешностей существует за счет дальнейшего уточнения параметрической модели, предполагающего, например, расширение полинома (1) с включением производных периода более высокого порядка на больших промежутках.

Отметим, что параметрическая модель данных реализует принципиально более высокую стабильность уже на относительно коротких промежутках наблюдений, порядка одного года. Что же касается разностной модели, в которой ограничивающим фактором является величина немоделируемых вариаций остаточных уклонений, достигающая нескольких микросекунд в пределах рассмотренных здесь 8-летних наблюдений пульсара B1937+21, то эта модель не позволяет уменьшить предел случайных вариаций до значения ниже чем 10⁻¹⁴ [2]. Существенно, что это значение достигается только в отношении ко всему 8-летнему промежутку, тогда как при его уменьшении этот предел, определяемый по остаточным уклонениям, пропорционально возрастает.

Модифицированный формат наблюдательных данных [5] который дополнен значениями интервалов от начального до каждого наблюдаемого события и численными величинами параметров вращения пульсара в пределах протяженности наблюдений, согласуется с принятой параметрической моделью и отвечает требованиям существенного, в пределах нескольких порядков, снижения влияния случайных вариаций наблюдаемых интервалов на результаты отождествления по данным хронометрирования пульсаров.

4.2 Сценарии отождествления по группам пульсаров

Детерминированное и целостное представление с высокой точностью дискретного континуума наблюдаемых интервалов излучения импульсов пульсара, определяемое параметрической моделью данных, позволяет более эффективно реализовать возможности наблюдений нескольких пульсаров в составе ансамбля. Прежде всего, одновременными наблюдениями группы пульсаров на одном или нескольких инструментах достигается надежное тестирование наблюдательных данных, выявление устойчивых признаков, не подверженных критическому влиянию вариаций интервалов, которые обусловлены случайными факторами. Обнаруживаются имеющиеся аномальные отклонения интервалов, не совместимые с параметрической моделью хронометрирования пульсаров, которые стали бы причиной преобладающего вклада этих пульсаров в результирующие вариации, не дает оснований для включения таких пульсаров в группу.

Благодаря тому, что отождествляемые физические величины описываются параметрами модели, стабильные значения которых устанавливаются по наблюдаемым интервалам, а влияние всех факторов нестабильности проявляется в виде вариаций наблюдаемого периода вращения пульсара, можно определить последовательность действий с дискретными рядами наблюдательных данных как при обработке наблюдательных данных на стадии подготовки и отбора, так и в процессе выполнения задач приложения. Цель отбора данных заключается в выделении из коллекции таких последовательностей, в которых вариации наблюдаемых интервалов имеют стационарный характер и, следовательно, не вносят систематических погрешностей отождествляемых величин в пределах промежутка наблюдений. В противном случае должны быть приняты дополнительные меры по корректировке параметрической модели, если это возможно, как описано в разделе 2.3, либо введены ограничения на промежуток, в пределах которого выполняется требование стационарности вариаций, как необходимое условие для включения пульсара в группу.

По выделенным рядам наблюдаемых интервалов, стационарных в пределах общего для группы промежутка наблюдений, выявляются устойчивые признаки отождествляемых величин, характер которых подтверждается по наблюдениям нескольких входящих в группу (ансамбль) пульсаров. Например, близкое совпадение по форме и величине отклонений наблюдаемых интервалов в диапазоне значений около 10^{-15} и менее может интерпретироваться как отклонение измерительного атомного эталона времени, по которому ведется хронометрирование. Если у пульсара обнаружено заметное отклонение рассчитанного по наблюдательным данным периода вращения, то путем наблюдений этого же пульсара на других инструментах в группе с другими пульсарами и сопоставлением полученных

рядов наблюдательных данных подтверждается или опровергается предположение об аномальном поведении этого пульсара. Следует отметить также, что по характеру и величине непараметризуемого остатка наблюдаемых интервалов (1) может быть сделано заключение о необходимости уточнения параметрической модели данных по критерию минимизации остатка, в котором, наряду со случайными вариациями наблюдаемых интервалов, может содержаться параметризуемая компонента.

Таким образом, сопоставлением независимых рядов хронометрирования по наблюдениям группы пульсаров достигается эффективный отбор пульсаров по критерию стационарности случайных вариаций наблюдаемых интервалов, исключение из группы аномальных пульсаров, которые обнаруживают отклонения интервалов, не совместимые с параметрической моделью, а также дальнейшее уточнение модели данных по выявленным параметризуемым компонентам вариаций наблюдаемых интервалов.

5 Заключение

Сопоставление двух моделей наблюдательных данных хронометрирования пульсаров применительно к отождествлению измеряемых физических величин – параметрической, основанной на синтезе наблюдаемого периода вращения пульсара, и разностной, основанной на остаточных отклонениях моментов пульсарных событий, – показало существенно, на несколько порядков, более высокую точность и достоверность отождествления по параметрической модели. Эти свойства параметрической модели проявляются, в отличие от разностной модели, даже на сравнительно небольших, порядка года, промежутках наблюдений и сохраняются с увеличением продолжительности промежутка.

Параметрическая модель данных, благодаря представлению результатов наблюдений в виде дискретного континуума интервалов излучаемых импульсов, независимо для каждого пульсара, особенно эффективна в наблюдениях группы пульсаров. Достигается надежное тестирование и отбор наблюдательных данных по устойчивым признакам, не подверженным критическому влиянию случайных вариаций наблюдаемых интервалов в пределах всей протяженности наблюдений. Обнаруживаются аномальные отклонения наблюдаемых интервалов, не совместимые с параметрической моделью, при наличии которых пульсары не могут быть включены в группу.

Коллекция астрометрических пульсарных данных отвечает требованиям приложений применительно к задачам отождествления свойств предметной области по наблюдательным данным хронометрирования. Коллекция минимизирует влияние статистических погрешностей хронометрирования, поддерживая тем самым долговременную стабильность наблюдаемых характеристик и высокую точность отождествления, которые определяются свойствами параметрической модели данных, и не могут быть достигнуты с помощью разностной модели.

Литература

- [1] К.Одуан, Б.Гино. Измерение времени. Основы GPS. Пер. с англ. под ред. В.М.Татаренкова. М., ТЕХНОСФЕРА, 2002.
- [2] V.M.Kaspi, J.H.Taylor, and M.F.Ryba. High-precision Timing of Millisecond Pulsars. III. Long-term Monitoring of PSRs B1885+09 and B1937+21 // *The Astrophysical Journal*, 1994, 428, 713–728.
- [3] A.N.Lommen, D.C.Backer. New Limits on the Gravitational Wave Background and Other Results from the Pulsar Timing Array // *Bulletin of the American Astronomical Society*, 2001, Vol. 33, p. 1347.
- [4] А.Е.Авраменко. Параметрический синтез пульсарного времени // *Измерительная техника*, 2006, № 6, 39–44.
- [5] А.Е.Авраменко. К согласованному виртуальному и реальному времени в коллекции астрометрических пульсарных данных // *Труды RCDL* 2007. Переславль Залесский, 2007, 103–111.
- [6] J.M.Cordes, A.Wolszczan, R.J.Dewey, et al. Timing and Scintillations of the Millisecond Pulsar 1937+214 // *Astrophys.J.*, 1990, 349, 245–261.
- [7] G.Petit and P.Tavella. Pulsars and Time Scales. // *Astron.Astrophys.*, 1996, 308, 290–298.
- [8] Ю.П.Илясов, М.Имае, Ю.Ханадо, В.В.Орешко, В.А.Потапов, А.Е.Родин, М.Секидо. Результаты двухчастотного хронометрирования пульсара В1937+21 в Калязине и Кашиме в 1997–2000 гг. // *Письма в Астрономический Журнал* 31, № 1, 33, 2005.
- [9] D.C.Backer, S.Hama, S.Van Hook, and R.S.Foster. Temporal Variations of Pulsar Dispersion Measures // *Astrophys J.* 404, 636, 1993.

Deterministic and ambiguous in the identification of the results of astrometrical observations of pulsars

A.E. Avramenko

The comparison of numerical results of identification of the measured physical values based on the two models of the pulsar timing data, is performed. The advantages of parametric model, which accuracy on several orders exceeds the accuracy of accepted residual model, are shown. The features of identification based on groups of pulsar timing sets which were created by the observations of ensemble of pulsars, are considered. The estimations of the achievable accuracy and long time stability for both pulsar data models, are performed. The instances of the domain applications of the astrometrical pulsar data collection, are presented.

* Работа поддерживается грантом РФФИ № 06-07-89043.

Эволюция стандарта ZOOCOD – концепции отражения зоологических иерархических классификаций в плоских таблицах реляционных баз данных*

© А.Л. Лобанов, И.С. Смирнов, М.Б. Дианов, А.А. Голиков, Р.Г. Халиков

Зоологический институт РАН
smiris@zin.ru

Аннотация

Биологические базы данных опираются на таксономические таблицы, в которых должны полно отражаться сложные многоуровневые иерархические схемы классификации таксонов. Отражение иерархий в плоских таблицах реляционных баз данных является нетривиальной задачей. Несколько способов решения этой проблемы известны в теории информационных систем. В Зоологическом институте РАН (ЗИН) большинство этих способов было разработано независимо в ходе многолетних работ по созданию зоологических баз данных. Все эти разработки отражены в стандарте ZOOCOD, многократно описанном в ряде публикаций, успешно используемом уже более 20 лет и распространенном сейчас уже и за пределами ЗИН (БИН РАН, ИПЭЭ РАН, МГУ, МГПИ, Нижегородский госуниверситет и др.). На основе этого стандарта создана и развивается национальная информационная система по биоразнообразию России (BIODIV). Концепция представления многоуровневых иерархий в реляционных базах данных успешно перенесена с таксономии на географические данные, сведения по морфологии и экологии животных.

1 Появление концепции ZOOCOD

Крупнейшие в России и одни из крупнейших в мире фондовые коллекции Зоологического института РАН насчитывают более 60 млн единиц хранения, представляющих почти все биологическое разнообразие России, в значительной степени биоразнообразие Антарктики и биоразнообразие отдельных регионов мира.

Огромные масштабы коллекций и рутинные операции по их поддержанию потребовали разработки компьютерных методов ведения зоологических коллекций [16, 17].

В свою очередь, все зоологические и ботанические исследования опираются на строго иерархические системы (классификации) животных и растений. Самым простым и полным представлением таких классификаций являются их графические изображения в виде разветвленных схем (в теории графов именуемых деревьями) (рис. 1). Однако классификации, включающие сотни и тысячи таксонов, изображать таким способом неудобно. Поэтому биологами давно изобретен не имеющий количественных ограничений способ – текстовые систематические списки. В этих списках ранги и соподчинение таксонов отражаются сдвигом строк списка относительно друг друга (табл. 1). Каждому таксономическому рангу соответствует определенный сдвиг строки от левого края страницы (т.е. определенное число пробелов). Подчиненные таксоны идут в списке за таксоном, к которому они относятся (т.е. за родительским таксоном, если использовать терминологию логики). Для удобства чтения перед каждым таксоном иногда указывают его ранг.

С появлением компьютеров систематические списки стали составлять с помощью текстовых редакторов (а позднее – текстовых процессоров типа MS Word). И если такие типичные операции по ведению систематических списков, как перестановки, вставки и удаление таксонов, в текстовом редакторе выполняются просто, то соблюдение одинакового числа пробелов перед таксонами одного ранга в длинном списке выдержать уже трудно. Кроме того, текстовые редакторы практически исключают: получение из списка разнообразной статистики, автоматическую генерацию из большого списка подписков по каким-либо аспектам, логическую проверку содержания списка (в том числе логической целостности классификации и отсутствия пропусков обязательных рангов, диктуемых соответствующим кодексом; в зоологии это: тип, класс, отряд, семейство, род и вид) и использование словарей для единообразного написания повторяющихся слов (в первую очередь – это фамилии авторов таксонов).

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

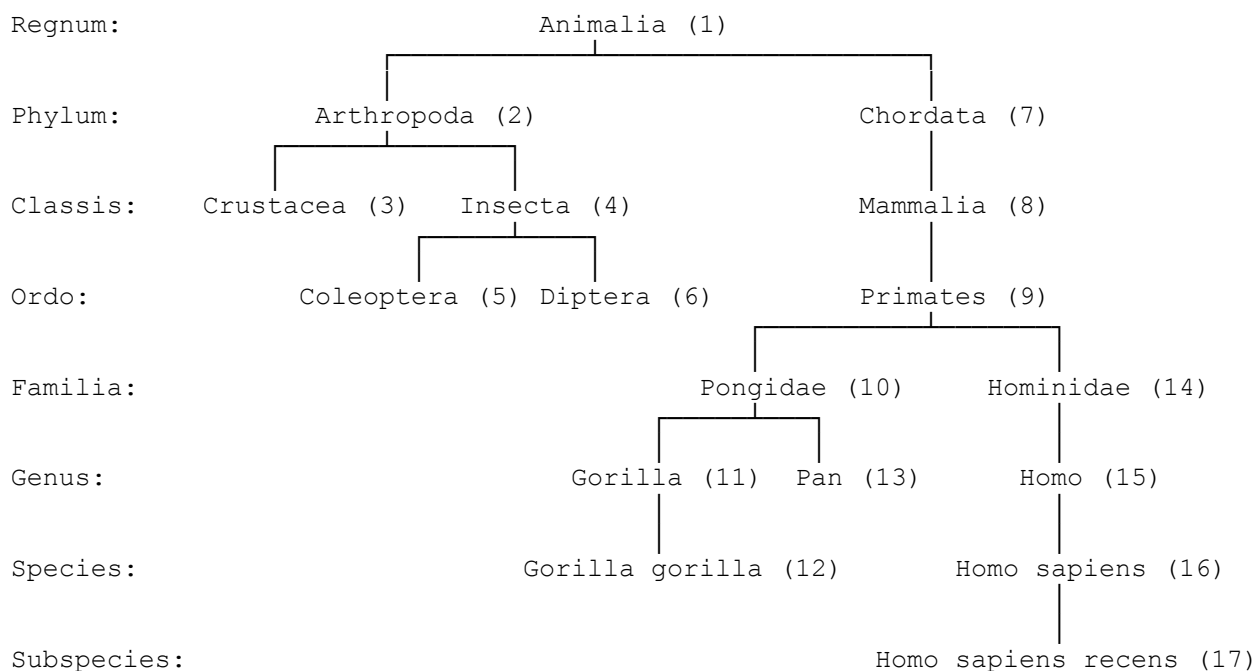


Рис. 1. Фрагмент древа зоологической классификации (номера в скобках показывают порядок обхода при левосписковом описании древа (см. в тексте далее) и при создании систематического списка)

Таблица 1
Фрагмент систематического списка, соответствующий древу классификации на рис. 1

Regnum: Animalia
Phylum: Arthropoda
Classis: Crustacea
Classis: Insecta
Ordo: Coleoptera
Ordo: Diptera
Phylum: Chordata
Classis: Mammalia
Ordo: Primates
Familia: Pongidae
Genus: Gorilla
Species: Gorilla gorilla
Genus: Pan
Familia: Hominidae
Genus: Homo
Species: Homo sapiens
Subspecies: Homo sapiens recens

Поэтому следующим этапом использования компьютеров для составления и дальнейшего ведения систематических списков стали попытки применения более совершенных программ - табличных процессоров (типа Excel) и систем управления базами данных (СУБД) самых разных моделей. Наиболее распространены три модели СУБД: иерархические, сетевые и реляционные.

Иерархическая модель очень подходит по своим особенностям для работы с иерархическими систематическими списками, но программы для ее поддержания являются экзотикой и отсутствуют на массовом компьютерном рынке. Сетевая модель

прошла небольшой период популярности, но из-за ряда присущих ей недостатков она была вытеснена с коммерческого рынка реляционной моделью [1]. В настоящее время практически все широко используемые и активно развиваемые производителями СУБД опираются на реляционную модель, которая помимо многих положительных качеств имеет под собой строгую математическую теорию и обоснованные ею правила нормализации баз данных, которые позволяют пользователям избегать ошибок, ввода излишней повторяющейся информации и дают возможность облегчить впоследствии работу по созданию программ, реализующих запросы к накопленной информации [3].

В итоге перед нами встала задача разработки способа, позволяющего легко и полно отражать в плоских реляционных базах данных сложные многоуровневые иерархические схемы классификации таксонов.

В основу первых решений этой задачи было положено моделирование в плоской таблице традиционного систематического списка [4, 7]. Такие списки существуют несколько сот лет, но не потеряли своего значения и сейчас. Они официально признаны и информатикой, где именуются «левосписковыми описаниями иерархического древа» [2]. Оказалось, что полную информацию об иерархии таксонов (их названия хранятся в поле LATNAM, куда заносится всегда только одно слово: для биноменов и тринименов – лишь последнее) можно передать всего двумя полями таблицы – цифровым кодом таксономического ранга (RANCOD) и систематическим кодом было (SYSCOD), который описывает порядок таксонов в списке (он может быть цифровым или символьным). Для возможности сравнения разных таблиц необходимо, чтобы RANCOD имел в

них одинаковый смысл. Первым шагом к стандарту было составление таблицы, включающей практически все применяющиеся в зоологии ранги (их оказалось более 40) (табл. 2).

Таблица 2

Стандартные зоологические таксономические категории (ранги) и их цифровые коды (значения для поля RANCOD)

Код ранга	Русское название	Латинское название
1	Царство	Regnum
2	Подцарство	Subregnum
4	Надотдел	Superdivisio
5	Отдел	Divisio
6	Подотдел	Subdivisio
8	Надтип	Superphylum
10	Тип	Phylum
12	Подтип	Subphylum
13	Аппендикс	Appendix
14	Инфратип	Infraphylum
18	Надкласс	Superclassis
19	Раздел	Part
20	Класс	Classis
24	Подкласс	Subclassis
26	Инфракласс	Infraclassis
30	Легион	Legio
32	Подлегион	Sublegio
36	Когорта	Cohors
37	Подкогорта	Subcohors
38	Надотряд	Superordo
40	Отряд	Ordo
42	Подотряд	Subordo
43	Инфраотряд	Infraordo
46	Серия семейств	Series
47	Кладус	Cladus
48	Надсемейство	Superfamilia
49	Группа семейств	Famil. grex
50	Семейство	Familia
52	Подсемейство	Subfamilia
58	Надтриба	Supertribus
60	Триба	Tribus
62	Подтриба	Subtribus
68	Надрод	Supergenus
70	Род	Genus
72	Подрод	Subgenus
80	Секция	Sectio
82	Подсекция	Subsectio
84	Группа видов	Spec. grex
88	Надвид	Superspecies
90	Вид	Species
92	Группа подвидов	Subsp. grex
93	Гибрид	Hybridus
94	Подвид	Subspecies
95	Племя	Natio
96	Вариетет	Varietas
97	Форма	Forma
98	Морфа	Morpha
99	Аберрация	Aberratio

Впервые новый подход был использован в работе над кадастром животных СССР [13]. Таксономические классификаторы – специализированные базы данных, получили широкое применение, и новый подход постоянно совершенствовался.

Принципиально важным было добавление поля SYN, в котором помечались синонимичные таксоны (они приобретали одинаковое значение SYSCOD со своими валидными таксонами), и поля ABBR, которое на первых порах играло роль уникального кода таксона (в нем программно генерировалась уникальная буквенная аббревиатура, имевшая максимальную мнемоничность). В основу концепции, получившей название «стандарт ZOOCOD», легли описание типичной структуры таксономической таблицы и правила присвоения кодов и порядка заполнения таблиц (табл. 3). Под этим названием было опубликовано первое подробное описание концепции и примеры ее применения в систематике млекопитающих [6].

2 Развитие концепции ZOOCOD

Использование концепции ZOOCOD и ее дальнейшее развитие проходило в дальнейшем чаще всего при исследованиях по систематике беспозвоночных – группы, отличающейся большим числом видов и надвидовых таксонов [5, 8, 10, 20, 21] и др. На основе концепции ZOOCOD были проведены важные обобщающие работы ЗИН РАН по биоразнообразию [14, 15] и созданы функционирующие много лет информационные системы ZOOINT [22] и OCEAN [23]. Следующим важным шагом в совершенствовании типовой структуры таблицы явилось введение поля GENUS, в котором специальной программой собирались все начальные части полного названия таксона видового (или еще более низкого) ранга. Соединение содержания полей GENUS+LATNAM образует полное название таксона. Для валидных таксонов это поле избыточно (сборку полного названия можно сделать поднимаясь в таблице вверх по полю LATNAM). Но для синонимичных комбинаций (отмеченных в поле SYN) только в поле GENUS можно было найти названия родов, в которых ранее фигурировал вид или подвид.

Естественным было появление поля ALTGEN, отметка в котором означала первичное описание вида в другом роде и необходимость заключения в круглые скобки фамилии автора таксона и года описания (в поля AUTHOR и YEAR скобки никогда не заносились). Со временем претерпело изменения поле SYSCOD, которое тоже стало частично отражать иерархию, но пока только для обязательных таксономических уровней. Усовершенствованный стандарт получил название ZOOCOD2.

Таблица 3

Фрагмент таблицы в стандарте ZOOCOD, соответствующий систематическому списку в таблице 1

GENUS	LATNAM	SYN	RANCOD	ABBR	SYSCOD
	Animalia		1	AN	100
	Arthropoda		10	AR	110
	Crustacea		20	CR	120
	Insecta		20	IN	130
	Coleoptera		40	INCO	13010
	Diptera		40	INDI	13013
	Chordata		10	CH	140
	Mammalia		20	MA	150
	Primates		40	MAPR	15010
	Pongidae		50	MAPRPO	15010100
	Hylobatidae	=	50	MAPRHY	15010100
	Gorilla		70	MAPRPOGOR	150101001000
Gorilla	gorilla		90	MAPRPOGORGOR	1501010010001000
	Pan		70	MAPRPOPAN	150101001010
	Hominidae		50	MAPRHO	15010105
	Homo		70	MAPRHOHOM	150101051000
Homo	sapiens		90	MAPRHOHOMSAP	1501010510001000
Homo sapiens	recens		94	MAPRHOHOMSAPRE	150101051000100010

В последние годы, в связи со спецификой разработок, использующих стандартные управляющие элементы и диалоговые панели Windows, концепция ZOOCOD претерпела значительные изменения и появилась 3-я версия – ZOOCOD3 (табл. 4). К описанному ранее набору полей теперь добавлены поля UNICOD (уникальный код таксона, делающий ненужным поле ABBR) и UPPCOD (значение UNICOD для ближайшего родительского таксона; для синонима таким является его валидный таксон). Эти поля позволяют пользоваться TreeView – специфическим

элементом Windows, изображающим иерархию таксонов в виде дерева. Хотя эти два поля полностью отражают многоуровневую иерархию, они неудобны для быстрого поиска всех потомков одного таксона. Для решения этой проблемы изменены правила заполнения поля SYSCOD (в новом виде оно называется HIERCOD). Его значениям приданы строгая и полная иерархичность, ширина поля делается равной максимальному числу уровней иерархии в таблице, а для заполнения используются не только цифры, но и все идущие за ними символы.

Таблица 4

Полный (максимальный) перечень полей таксономической таблицы стандарта ZOOCOD3

Имя поля	Тип	Длина	Назначение
GENUS	C	60	недостающие до полного названия вышестоящие термины (род – для вида, род и вид – для подвида, род, вид и подвид – для разновидности или формы)
LATNAM	C	25	латинское название (одно слово для данного ранга)
AUTHOR	C	45	автор таксона
YEAR	C	4	год описания таксона
RANCOD	N	2	код ранга таксона
SYN	C	1	код синонима
TAXON	C	85	полное название таксона (без автора и года описания)
ALTGEN	C	1	отметка описания в другом роде
ALTGENUS	C	25	название рода, в котором вид был впервые описан
SYSCOD	C	15	систематический код (необязателен при наличии HIERCOD)
HIERCOD	C	32	полностью иерархический систематический код (длина поля изменяется в зависимости от числа таксономических рангов в классификаторе)
UNICOD	C	10	уникальный код названия (генерируется автоматически)
UPPCOD	C	10	значение UNICOD у вышестоящего таксона
RANK	C	15	словесное обозначение ранга таксона (дублирование RANCOD для неподготовленного пользователя)
ABBR	C	14	оставлено исторически, ибо во многих старых таблицах играет роль UNICOD

Поле HIERCOD заполняется специальной утилитой, но зато поиск потомков заданного таксона выполняется молниеносно и предельно просто – наложением фильтра с полностью оптимизируемым условием. Интересно, что мы в своих разработках самостоятельно пришли ко всем способам представления иерархии в реляционных таблицах, известным в информатике [19, 24]. К полям GENUS и LATNAM добавлено поле TAXON, в котором хранится полное название таксона (т.е. биномен для видов и тринмен для подвидовых таксонов). Для возможности отказа от использования поля GENUS введено более короткое поле ALTGENUS, в которое заносится названия рода, в котором был впервые описан вид. По техническим причинам поле YEAR переименовано в YEARD.

Этот стандарт многократно описан в публикациях [9] и сейчас распространен и за пределами ЗИН (БИН, ИПЭЭ, МГУ, МГПИ, Нижегородский госуниверситет и др.). На базе этого стандарта была создана основа национальной информационной системы по биоразнообразию России [12].

Концепция представления многоуровневых иерархий в реляционных базах данных перенесена с таксономии на географические данные, сведения по морфологии насекомых и другие аспекты зоологической информации [11].

3 Правила унификации биологических баз данных

В итоге работ над концепцией ZOOCOD были сформулированы некоторые основные положения построения и унификации биологических (таксономических) баз данных, которые должны облегчить общение биологов и создание интегрированных систем:

1. Должна использоваться реляционная модель баз данных.

2. В соответствии с требованиями нормализации баз данных информационные системы должны включать отдельные таблицы по каждому аспекту сведений о животных или растениях.

3. Каждый элемент сведений должен храниться в конкретной системе только в одной записи одной таблицы. Это правило влечет за собой широкое использование словарей и классификаторов (или тезаурусов – словарей с иерархической структурой и родо-видовыми отношениями между понятиями).

4. В основе системы должны лежать классификаторы животных, растений, грибов, бактерий и других царств живых организмов.

5. Часто используемые биологами данные о географическом распространении организмов должны быть строго разделены на отдельные аспекты (административно-территориальный, физико-географический, биогеографический, ландшафтный, природно-зональный и т.п.). Термины каждого аспекта должны быть зафиксированы в классификаторах.

6. Добиться ведения разных баз данных в одной СУБД невозможно даже внутри одного института и

к этому не нужно стремиться. Важно только использовать профессиональные коммерческие СУБД, имеющие проверенные механизмы импорта и экспорта. В качестве критерия пригодности СУБД предлагается использовать формат DBF (dBaseIII). Если СУБД не способна импортировать или экспортировать этот формат, то ее не следует использовать.

7. Для обеспечения простоты обмена данными между разными СУБД, работающими в разных версиях операционных систем, следует ограничить длину имен файлов 8 символами, длин расширений имен 3 символами, а названия полей баз данных 10 символами; при этом использовать во всех этих именах только латинские буквы, цифры и символ подчеркивания (при этом не использовать разницу между заглавными и строчными буквами, если какая-то система такую возможность поддерживает). В системах, где длина символьного поля не ограничивается или может быть очень большой, не делать длину поля более 254 символов.

8. Если для представления иерархии не используется стандарт ZOOCOD, то все же нужно стремиться к соблюдению некоторых правил, облегчающих переход к этому стандарту:

а) для указания ранга таксона или понятия использовать только одно поле таблицы, приводя в нем коды или полные названия рангов;

б) иметь поле с уникальным кодом таксона и использовать эти коды для связи с таблицами фактов о таксонах;

в) в отдельном поле отмечать факт синонимии, указывая уникальный код валидного названия для каждого синонимичного;

г) иметь поле для указания систематического порядка записей, ибо по правилам нормализации толкование таблицы не должно зависеть от физического порядка записей (упорядочение по этому полю должно создавать аналог «систематического списка», который привычен и необходим большинству биологов);

д) наличие двух полей, отражающих иерархию способом, принятым теперь в приложениях для Windows (поля для отражения иерархии методом TreeView), ни в коей мере не отменяет необходимость пунктов а) и г), так как эти два поля не отражают истинные ранги таксонов и нужный порядок среди дочерних таксонов одного родительского.

4 Концепция ZOOCOD 3 и ее развитие

Стандарт ZOOCOD3 продолжает эволюционировать, приспосабливаясь к новым требованиям специалистов-систематиков и изменениям в компьютерных технологиях.

Легко заметить, что перечисленный набор полей (табл. 4) обладает значительной избыточностью и формирование окончательной структуры таблицы – дело вкуса составителя (поле GENUS лишнее при наличии поля TAXON; поля UNICOD и UPPCOD могут быть сгенерированы по значениям RANCOD

и HIERCOD; и т.д.). Такая избыточность оставлена намеренно, ибо окончательная структура должна выбираться с учетом предпочтений программиста, который будет писать программы работы с базой данных, обеспечивающие изменения классификации, дополнения, выполнение типовых запросов, генерацию вторичных списков и таблиц, страниц HTML и ASP, проведение различных проверок и анализа по разным аспектам.

Особенно важным нам кажется резюмировать решения, связанные с главной задачей - представлением иерархии. Обобщая сказанное выше можно выделить следующие способы представления сложной иерархии в плоских реляционных таблицах стандарта ZOOCOD3:

1. Использование совокупности полей: SYSCOD + RANCOD;

2. Использование совокупности полей: HIERCOD + RANCOD;

3. Использование совокупности полей: UNICOD + UPPCOD + RANCOD + SYSCOD.

Эти способы существенно отличаются при написании программ обработки таксономических таблиц.

Все изложенное выше относится к первым 20-ти годам эволюции концепции ZOOCOD, проходившей в основном с использованием СУБД FoxPro и Access. Но используемый FoxPro формат DBF все хуже и хуже воспринимается современными программами, работающими под управлением Windows. Кроме того, стандарт ZOOCOD3 выполняет не все требования нормализации. Поэтому в 2007–2008 гг. была предпринята еще одна серьезная ревизия стандарта: отказ от полей, избыточных в новой модели данных (ABBR, GENUS, HIERCOD); использование строгой реляционной модели.

Литература

- [1] Алимов А.Ф., Лобанов А.Л., Пугачев О.Н. Сравнительный анализ реляционного и сетевого подходов к созданию банков данных по систематике, экологии и географическому распространению животных // Журнал общей биологии. – 1993, т. 54, № 1. – С. 96–103.
- [2] Глушков В.М. Основы безбумажной информатики. – М.: Наука. 1982. – 552 с.
- [3] Дейт К. Дж. Введение в системы баз данных: пер. с англ. – 6-е издание. – Киев – М.: Диалектика, 1998. – 784 с.
- [4] Лобанов А.Л. Линейно-иерархическая структура баз данных о таксонах животных // Принципы и методы экоинформатики. – М., 1986. – С. 293–295.
- [5] Лобанов А.Л., Дианов М.Б., Смирнов И.С. Результаты разработок и использования зоологических информационно-поисковых систем // Информационно-поисковые системы в зоологии и ботанике: Труды Зоологического ин-та РАН. – Т. 278. – СПб., 1999. – С. 54–55.
- [6] Лобанов А.Л., Зайцев М.В. Создание компьютерных баз данных по систематике млекопитающих на основе классификатора названий

животных «ZOOCOD» // Вопросы систематики, фаунистики и палеонтологии мелких млекопитающих: Труды Зоологического ин-та РАН. – Т. 243. – СПб., 1991. – С. 180–198.

- [7] Лобанов А.Л., Сергеев Г.Е. Проект классификатора названий животных и принцип представления информации об распространении в структуре биологических баз данных // Принципы и методы экоинформатики. – М., 1986. – С. 214–215.
- [8] Лобанов А.Л., Смирнов И.С. Принципы построения и использования классификаторов животных в стандарте ZOOCOD // Базы данных и компьютерная графика в зоологических исследованиях: Труды Зоологического ин-та РАН. – Т. 269. – СПб., 1997. – С. 66–75.
- [9] Лобанов А.Л., Смирнов И.С. Место и роль информационных технологий в исследованиях Зоологического института РАН // Фундаментальные зоологические исследования. Теория и методы. – М. – СПб.: Изд-во КМК, 2004. – С. 283–318.
- [10] Лобанов А.Л., Смирнов И.С., Дианов М.Б. ZOOCOD – концепция представления зоологических иерархических классификаций в реляционных базах данных // Информационно-поисковые системы в зоологии и ботанике (Тезисы международного симпозиума, май 1999): Труды Зоологического ин-та РАН. – 1999, Vol. 278. – Р. 66.
- [11] Медведев С.Г., Лобанов А.Л. Информационно-аналитическая система по мировой фауне блох (Siphonaptera): результаты и перспективы развития // Энтомологическое обозрение. – 1999, 78(3). – С. 732–748.
- [12] Пугачев О.Н., Дианов М.Б., Лобанов А.Л., Смирнов И.С., Халиков Р.Г., Голиков А.А. Итоги разработки проекта «Информационная система по биоразнообразию животных России» (ZooDiv) // Отчетная научная сессия по итогам работ 2007 г., 8–10 апреля 2008: Тезисы докладов. – СПб.: Изд-во Зоологического ин-та, 2008. – С. 42–44.
- [13] Скарлато О.А., Алимов А.Ф., Лобанов А.Л., Умнов А.А. Машинные банки данных – подход к кадастру животного мира // Всесоюзное совещание по проблеме кадастра и учета животного мира: Научно-информац. материалы к совещанию. – Уфа, 1989. – С. 56–64.
- [14] Скарлато О.А., Старобогатов Я.И., Лобанов А.Л., Смирнов И.С. Базы данных по зоологической систематике и сведения о высших таксонах животных // Зоологический журн. – 1994а, Т. 73, Вып. 12. – С. 100–116.
- [15] Скарлато О.А., Старобогатов Я.И., Лобанов А.Л., Смирнов И.С. Биоразнообразие и возможности его анализа с применением компьютерных банков данных // Биоразнообразие. Степень таксономической изученности. – М.: Наука, 1994б. – С. 20–41.
- [16] Смирнов И.С., А.Л. Лобанов, А.Ф. Алимов, В.А. Кривохатский. Электронные коллекции

Зоологического института РАН // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Пятой Всероссийской научной конференции RCDL'2003 (Санкт-Петербург, 29–31 октября 2003 г.). – СПб. : НИИ Химии СПбГУ, 2003. – С. 275–278.

- [17] Смирнов И.С., А.Л. Лобанов, А.Ф. Алимов, А.Г. Кирейчук, А.Е. Вахитов. Электронные зоологические коллекции и технологии электронных библиотек // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Сб. тезисов постерных докладов Восьмой Всероссийской научной конференции (RCDL'2006). Суздаль, 17–19 октября 2006 г. – Ярославль : Ярославский государственный ун-т им. П.Г. Демидова, 2006. – С. 60–61.
- [18] Старобогатов Я.И. О проблемах номенклатуры высших таксономических категорий // Справ. по систематике ископаемых организмов. – М. : Наука, 1984. – С. 174–187.
- [19] Щеваев П.А. Способы хранения иерархических структур в реляционных базах данных // Новые информационные технологии и системы : Труды VI Междунар. научно-технической конф. Ч. 2. – Пенза : ПГУ, 2004. – С. 226–233.
- [20] Lobanov A.L., Smirnov I.S., Dianov M.B. ZOOCOD – conception of representation of zoological hierarchical classifications in relational databases // Information Retrieval Systems in Biodiversity Research (Abstracts of the International Symposium. May 1999). Proceedings of the Zoological Institute RAS. 1999. Vol. 278. P. 65.
- [21] Smirnov I. 1993. Working out a databank on marine invertebrates. – In: Abstracts of 8th International Echinoderm Conference. Dijon, September 6–10, 1993. P. 111.
- [22] Smirnov I.S., Lobanov A.L., Alimov A.F., Dianov M.B., Golikov A.A., Stepanjants S.D. Integrated Information System ZOOINT: the Present State and Prospects // Proceedings of the First East-European Symposium on Advances in Databases and Information Systems (ADBIS'97). St.-Petersburg, September 2–5, 1997. V. 2, 1997. P. 112–114.
- [23] Smirnov I., Lobanov A., Dianov M. 1994. Information system for Antarctic marine animals. – In: Abstracts of SCAR SIXTH BIOLOGY SYMPOSIUM: Antarctic Communities: Species, Structure and Survival. Venice, 30 May – 3 June, 1994. P. 247.
- [24] VanTulder Gijs, 2003. Storing hierarchical data in a DB: <http://www.sitepoint.com/print/hierarchical-data-database>

Evolution of the standard ZOOCOD – concept of reflection of zoological hierarchical classifications in the plane tables of relation databases

A.L. Lobanov, I.S. Smirnov, M.B. Dianov,
A.A. Golikov, R.G. Khalikov

Biological databases lean on taxonomic tables in which complex multilevel hierarchical schemes of taxa classification should be reflected full. Reflection of hierarchies in flat tables of relation databases is not trivial problem. Some ways of the decision of this problem are known in the theory of intelligence systems. In Zoological Institute of the Russian Academy of Science (ZIN) the majority of these ways has been developed independently during long-term works on creation of zoological databases. All these development are reflected in standard ZOOCOD repeatedly described in a number of publications. Already more 20 years it successfully is used and widespread outside of ZIN (BIN of the RAS, IPEE of the RAS, the Moscow State University, MSPI, the Nizhniy Novgorod state university, etc.). On the basis of this standard the national intelligence system on a biodiversity of Russia (BIODIV) was created and developed. The concept of representation of multilevel hierarchies in relation databases is successfully transferred from taxonomy on geographical data, data on morphology and ecologies of animals.

* Гранты РФФИ № 05-07-90179-в, 05-07-90354-в и 06-04-08020, проект «Антарктика» (ФЦП «Мировой океан»), программа «Биоразнообразие».

Информационные системы, ориентированные на представление данных о погоде и климате в сети Интернет

© Н.А. Лаврентьев, А.З. Фазлиев

Институт оптики атмосферы СО РАН, Томск
faz@iao.ru

Аннотация

Интерес пользователя к информационным ресурсам в сети Интернет, представляющих данные о погоде и климате, в первую очередь о температуре и давлении необычайно высок. В нашей работе представлен обзор ресурсов по погоде и климату, как информационных, так и вычислительных, определяющие спектр возможностей пользователей, который большинством из них не осознан. Рассмотрена организация центров хранения данных, форматы данных типовых моделей, информационные системы, представляющие данные о погоде. Дан краткий обзор порталов LEAD и ATMOS, предоставляющих возможность проведения пользователем индивидуальных расчетов с помощью мезомасштабных моделей.

1 Введение

Проблема предсказания климатических изменений, вызываемых человеческой деятельностью является одной из наиболее важных проблем, стоящих в настоящее время перед наукой. В качестве причин таких изменений в климатической системе можно рассматривать вырубку лесов, приводящую к опустыниванию и изменению альбедо; сжигание ископаемого топлива, приводящее к изменению концентрации углекислого газа в атмосфере; изменение концентрации малых газовых примесей, контролирующих концентрацию озона в атмосфере; освоение нефтяных и газовых месторождений в районах вечной мерзлоты, сопровождаемое нарушением растительного покрова, а также множество других воздействий. Решение этих проблем диктует необходимость создания национальных систем экспертной оценки неблагоприятных для страны последствий глобальных изменений климата.

Климатические модели дают возможность воспроизведения современного климата на основе численных экспериментов, исследования чувствительности "модельного" климата по отношению к малым изменениям параметров, характеризующих внешние воздействия, изучения основных климатообразующих процессов, оценки опасных региональных последствий глобальных изменений и многое другое. Они составляют также основу технологий четырехмерного усвоения многообразной информации о состоянии компонент климатической системы, к которым, в первую очередь, относятся атмосфера и океан [1].

В настоящее время при исследовании и прогнозе локальных атмосферных процессов, протекающих над ограниченной территорией, широко используются мезомасштабные метеорологические модели, которые опираются на нестационарные трехмерные уравнения гидротермодинамики атмосферы и параметризацию атмосферных процессов (поток коротковолновой и длинноволновой радиации, конвективных процессов, пограничного слоя, микрофизики влаги, турбулентности атмосферы, тепло- и влагообмена в подстилающей поверхности) [2]. Компьютерная реализация таких моделей основана на применении нетривиальных вычислительных алгоритмов и высокопроизводительных вычислительных ресурсов. В крупных мировых центрах (NCAR, EPA, ECMWF, NERC) для исследования атмосферных процессов созданы и свободно распространяются исходные коды программ моделей такого уровня. Использование этих моделей в исследовательской и прикладной работе в России ограничено. Это обусловлено, с одной стороны, недостаточным распространением необходимых вычислительных ресурсов, а с другой стороны, отсутствием необходимого числа специалистов.

На наш взгляд эта задача системно решается в ряде проектов (PRISM, LEAD, ...) [ссылка на мой обзор и ЛИД], реализуемых и осуществленных в США и ряде стран Европы. В докладе сделан краткий обзор центров хранения данных в США, описаны форматы данных, применяемые при решении погодных задач. Описываются информационные ресурсы о состоянии атмосферы предоставляемые в Европе, Америки и Азии. Отмечено, что большая

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

часть этих ресурсов порождена мезомасштабными моделями MM5 и WRF. Дан краткий обзор порталов LEAD и ATMOS, предоставляющих возможность проведения пользователем индивидуальных расчетов с помощью мезомасштабных моделей.

2 Центры хранения данных в США

Обращаясь к описанию информационных ресурсов в области климата и погоды в первую очередь стоит рассмотреть распределенные архивы США (центры хранения данных) в которых сконцентрирована основная информация и данные в науках о Земле. Эти центры были созданы как часть программы EOS (система наблюдений за Землей) и формировались двумя организациями: NASA и NOAA. Стоит отметить, что в одном из архивов (SEDAC) проводится синтез данных и информации о погоде и климате с социометрическими данными в форме полезной для тех кто принимает решения на основе результирующей информации.

В центрах данных большая часть информации получена в результате измерений, проведенных со спутников, самолетов, наземных платформ и компьютерных расчетов. Информационные системы, связанные с архивами, содержат необходимый инструментарий для анализа данных, работы с временными рядами и визуализации. Как правило, это прикладные программы, которые пользователь может устанавливать на своем рабочем месте. В последние пять лет начато формирование информационных систем с доступом по сети Интернет (см. параграф 5).

Стоит отметить, что большая часть этих данных является доступной только сотрудникам NASA. В каждом центре данных существуют данные доступные всем пользователям. Например, в центре на Аляске к их числу относятся данные о бореальных лесах, морских льдах и тропических лесах.

Распределенные архивы данных и информации в области наук о Земле, находящиеся в США, перечислены ниже:

- Alaska Synthetic Aperture Radar (SAR) Facility (ASF)
Полярные процессы
- EROS Data Center (EDC) Land Processes
Процессы взаимодействия с поверхностью
- Goddard Space Flight Center (GSFC)
Верхняя атмосфера, глобальная геосфера, динамика атмосферы и геофизика
- Jet Propulsion Laboratory (JPL) Physical Океанография
- Langley Research Center (LaRC) DAAC
Radiation budget, tropospheric chemistry, clouds, and aerosols
- National Snow and Ice Data Center (NSIDC)
Снег и лед, криосфера (не-SAR) и климат
- Oak Ridge National Laboratory (ORNL)
Биогеохимическая динамика

- Socioeconomic Data and Applications Center (SEDAC) at the Consortium for International Earth Science Information Network (CIESIN)
Человек и окружающая среда
- Global Hydrology Resource Center (GHRC)
Гидрологические циклы
Данные о климате и информации можно найти на сайтах организаций NOAA:
- National Climatic Data Center (NCDC) of National Oceanic and Atmospheric Administration (NOAA) Climate and weather
- National Geophysical Data Center (NGDC) of NOAA Solid Earth geophysics, marine geology and geophysics, solar terrestrial physics, and paleoclimatology
- National Oceanographic Data Center (NODC) of NOAA
Океанография
- Satellite Active Archive (SAA) of National Oceanic and Atmospheric Administration (NOAA)
Спутниковое дистанционное зондирование

3 Форматы данных для погодных и климатических моделей

Климатические и погодные модели, в основном, работают со следующими четырьмя форматами данных: GRIB [6], BUFR [7, 8], NetCDF [9–11] и HDF [12].

Самым распространенным форматом для представления метеорологической информации является GRIB (**G**ridded **B**inary). Он предназначен для хранения и обмена информацией о погоде в бинарной форме. Формат GRIB был разработан Всемирной метеорологической организацией (WMO)

Предложенный в 1985, GRIB был усовершенствован в 1988, 1990 (стандарт Edition 1) и 1994. К сожалению, стандартизован он не полностью, часть организаций использует свои собственные версии GRIB.

Каждая запись GRIB предназначена для передачи или хранения единственного параметра вместе со значениями, которые представляют собой массив точек. Закодированные GRIB данные реанализа или прогноза представляют собой непрерывный битовый поток, состоящий из последовательности октетов (1 октет = 8 бит). Запись логически разбивается на состоящие из октетов секции, каждая из которых содержит контрольную информацию или данные. Запись GRIB состоит из 6 секций, две из которых необязательны:

0. Индикаторная секция – метка «GRIB», длина сообщения, версия формата

1. Секция описания продукта (Product definition section/PDS) – длина секции, идентификатор закодированного анализа или прогноза

2. Секция описания сетки (Grid Description Section/GDS) – длина секции, геометрия сетки, данная секция необязательна

3. Секция битовой карты (Bit Map Section/BMS) – длина секции, последовательность битов для каждого из узлов сетки, 0 при отсутствии данных в узле, 1 при присутствии, секция необязательна

4. Секция бинарных данных (Binary Data Section/BDS) – длина секции, значения данных

5. Конечная секция – метка «7777», признак конца секции.

Хотя секция описания продукта и необязательна, рекомендовано её наличие во всех GRIB-сообщениях.

Представление данных в виде последовательности битов делает формат независимым от какого-либо одного машинного представления. Данные в формате GRIB предназначены исключительно для автоматической обработки.

Описания данных, находящихся в GRIB-файле, находится во внешних таблицах параметров. Наиболее распространены таблицы параметров NCAR, NCEP, ECMWF и NOAA FSL [6].

Формат данных BUFR (Binary Universal Form for the Representation of Meteorological Data) был также разработан WMO в 1988 году. BUFR представляет собою бинарный формат для представления в виде непрерывного битового потока метеорологических данных в обособленных точках (в отличие от GRIB, предназначенного для представления пространственно распределённых данных). Тем не менее, сама структура формата не предполагает чего-либо уникального только для метеорологии, и он может применяться к любым численным типам данных.

Файлы BUFR состоят из некоторого числа последовательных записей, называемых потоками. Запись BUFR, содержащая некоторые данные наблюдения, также содержит и полное описание этих данных, включая определение наблюдаемого параметра (высота, температура, давление, дата и т.д.), единицы измерения, точность округления, формат сжатия, применённого к данным, а также число бит в значениях содержащихся данных. Описание данных также содержится во внешних таблицах, представляющих собою основную часть BUFR-описания. Таблицы, определённые в BUFR, включают в себя:

Таблица A – категория данных

Таблица B – классификация и описание отдельных элементов данных

Таблица C – операторы описания данных

Таблица D – список общих последовательностей

Таблицы кодов и флагов — описание кодов и флагов, связанных с таблицей B [7,8].

Общий сетевой формат данных NetCDF (Network Common Data Format) был предложен Межуниверситетской кооперацией атмосферных исследований UCAR (University Cooperation for Atmospheric Research) в качестве формат самоописываемых расширяемых файлов с научными данными.

NetCDF представляет собой абстракцию, которая описывает данные как коллекцию самоописываемых, прозрачных с сетевой точки зрения объектов, доступных через простой интерфейс. Коллекции поименованных многомерных переменных могут быть доступны в произвольном порядке, без знания деталей хранения данных. Дополнительная информация (единицы измерения и пр.) также хранится вместе с данными. Для работы с данными в формате NetCDF существует специальное программное обеспечение [9].

Программное обеспечение, работающее с данными формата NetCDF обращается лишь к абстрактным типам данных, из чего следует, что все операции доступа и манипулирования данными в NetCDF-файле используют лишь набор функций, предоставленный интерфейсом. Внутренне представление данных, таким образом, скрыто от приложения, использующего интерфейс, так что формат данных может быть изменён без того, чтобы затрагивать уже существующие приложения. Формат NetCDF был разработан таким образом, чтобы его физическое представление не зависело от архитектуры какой-либо определённой машины.

Для достижения прозрачности с сетевой точки зрения и независимости от конкретной архитектуры NetCDF определён в терминах XDR (eXternal Data Representation) [10], протокола кодирования и описания данных, предоставляющего средства для представления данных в виде машинно-независимых битовых последовательностей.

Одна из целей формата NetCDF – предоставление эффективного доступа как к небольшим, так и крупным наборам данных, для чего используется прямой, а не последовательный, как в ранее описанных форматах, доступ к данным. Это может быть намного более эффективным в случае, когда данные считываются не в том порядке, в каком были записаны. Дополнительные затраты на представление в виде XDR зависят от множества факторов, но обычно они незначительны по сравнению с общим объёмом данных.

Простое использование NetCDF не достаточно для того, чтобы сделать данные самоописываемыми и имеющими какой-либо смысл для человека или машины. Названия переменных и единиц измерений должны поддаваться интерпретации и соответствовать установленным соглашениям [10].

Формат иерархических данных HDF (Hierarchical Data Format) создан для обмена научными базами данных американским Национальным центром суперкомпьютерных приложений NCSA (National Center for Supercomputing Applications).

HDF – самоописываемый формат файлов для переноса различных типов данных между различными компьютерами. Библиотека HDF содержит интерфейсы для хранения и поиска, сжатых и несжатых растровых изображений с палитрами цвета, интерфейс для хранения и поиска n-размерных наборов научных данных вместе с информацией о данных, таких как метки, единицы, форматы и масштабы по

всем измерениям. Версии 3.3 и выше предусматривает работу с форматом JPEG и поддерживает использование NetCDF.

В HDF [12] используются метаданные о наборах научных данных SDS (Scientific Data Sets), которые содержат:

- систему координат, используемой для интерпретации или отображения данных;
- масштабы по каждой оси;
- метки по каждой оси и для всего набора данных в целом;
- единицы измерения по каждой оси и для данных;
- допустимые максимальное и минимальное значения данных;
- информацию о калибровке данных;
- информацию о дополнительных или отсутствующих значениях.

4 Информационные ресурсы

Информационно-вычислительные ресурсы о состоянии атмосферы, в основном, представлены веб-сайтами, предоставляющими пользователю прогноз различных метеорологических величин на ближайшее время в каком-либо определённом регионе. Так, в США – это ресурсы государственных организаций: университетов Мэриленда [13], Висконсина [14, 15], Пурду [16], Иллинойса [17], Миллерсвилля [18]; Военно-морской академии [19], Национальной службы прогноза погоды (NWSF) [20], Национального управления по авиации и исследованию космического пространства (NASA) [21], различных подразделений Национальной администрации атмосферы и океана (NOAA) [22–24] и Национального центра атмосферных исследований (NCAR) [25–28]. Европейские сайты в большинстве своём посвящены исследованию средиземно-морского региона – это различные научные учреждения Италии [29–33], Словении [34] и Греции [35]. Кроме того, Европа представлена такими коммерческими ресурсами как швейцарский Meteoblue [36], предоставляющим прогноз на всю европейскую территорию, а также испанский Meteosim Truewind [37], основной задачей которого является использование климатических моделей для прогноза поведения ветра. Ресурсы Юго-Восточной Азии включают в себя сайты, предоставляющие прогнозы на территорию Вьетнама [38], Гонконга [39], Южной Кореи [40] и Тайваня [41], Центральная Америка представлена Мексикой [42], а Ближний Восток – Израилем [43].

Большинство описанных выше информационно-вычислительных ресурсов в качестве своей вычислительной основы используют мезомасштабные климатические модели MM5 и, в последнее время всё чаще, WRF. Исключениями являются лишь ресурсы, осуществляющие прогноз на средиземно-морский регион, которые используют специфическую для данного региона модель BOLAM, а также сайты предоставляющие глобальную климатиче-

скую информацию – RAP Real-Time Weather [27] и Meteosim Truewind [37], использующие глобальные климатические модели, такие как GFS. Основное множество ресурсов не даёт возможности сравнивать данные, полученные от разных климатических моделей, так как предоставляют лишь результаты расчётов одной конкретной модели. Исключениями являются сайты Национальной обсерватории Афин [35], предоставляющей доступ к результатам расчётов 3 различных моделей – MM5, WRF и BOLAM; Национального центра атмосферных исследований США [28], включающего модели MM5 и WRF; и уже упомянутые выше RAP Real-Time Weather (WRF, RUC, ETA, GFS) и Meteosim Truewinds (GFS и MASS). Самым распространённым видом представления результатов расчётов пользователю является графический – в виде статичного изображения, но большинство ресурсов также позволяет получить анимированную различными способами (анимированный GIF, JavaScript, Java-апплет) графику. Табличное отображение результатов встречается очень редко.

Предназначение описанных выше информационно-вычислительных ресурсов – предоставлять краткосрочный (от 6 часов до 2 суток) климатический прогноз на какую-либо область. Отсюда вытекают такие недостатки их для исследовательской работы как:

1. Отсутствие настроек используемой в ресурсе климатической модели – модель настраивается заранее на конкретную область, и пользователь не имеет никакого контроля над её поведением.
2. Невозможность самому инициировать запуск модели – модель запускается в строго определённое время, выполняет расчёт прогноза, а пользователю лишь предоставляются уже рассчитанные данные.
3. Небольшой временной промежуток, для которого доступны данные – у большинства ресурсов срок хранения архивов не превышает двух недель.

Примеры графического представления результатов вычислений и космического снимка сопровождающего прогноз погоды показаны на рис. 1 и 2.

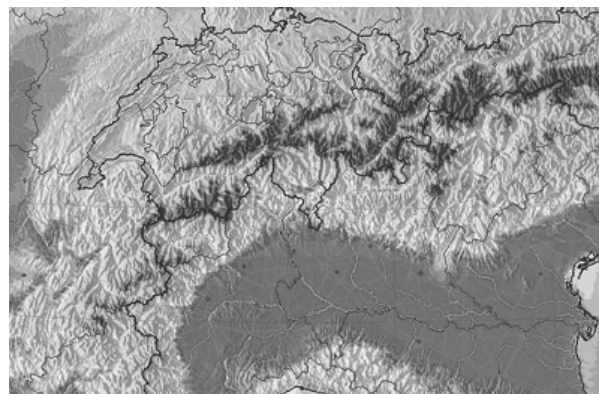


Рис. 1. Представление данных на карте с сайта. Meteoblue, [://www.meteoblue.ch/](http://www.meteoblue.ch/)

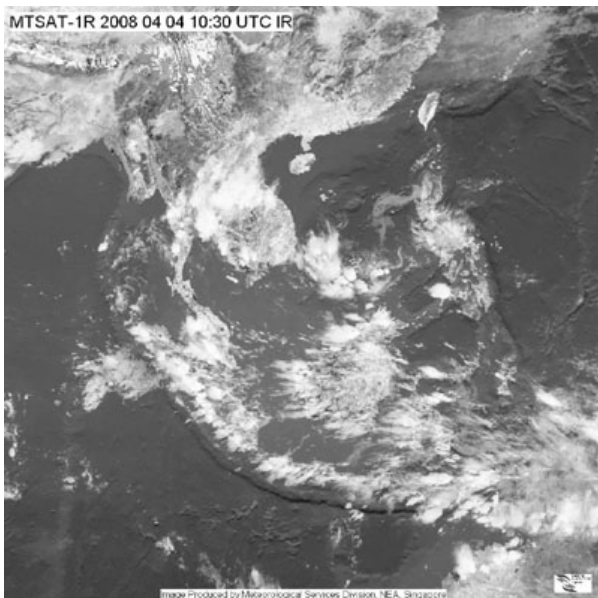


Рис. 2. Космический снимок с сайта AMI Environmental. South Asia, http://www.vnbaolut.com/index_en.html

5 LEAD портал

Каждый год только на территории США потери от наводнений, града, сильных ветров, торнадо, молний и прочих мезомасштабных погодных явлений составляют сотни человеческих жизней и десятки миллиардов долларов. И, хотя даже частичное предотвращение таких событий обещает огромные экономические и социальные преимущества, прогресс в информационной сфере затрудняется большим размером наборов метеорологических данных, а также огромными объемами вычислений и логической сложностью используемых численных моделей и систем сбора данных.

Портал LEAD (Linked Environments for Atmospheric Discovery) [4] предназначен для решения этой проблемы – это междисциплинарный проект, в котором участвуют 9 научных институтов и более 100 учёных. Для достижения результата проект фокусируется на двух основных целях.

Первая – это понижение входного барьера для использования специализированного метеорологического ПО. Используемый в этой области инструментарий включает в себя множество модулей для работы с данными, которые слишком сложны для индивидуального использования. Мезомасштабные модели представляют собой сложное ПО, развиваемое десятилетиями, работающее с огромным числом входных параметров и требующее обращения к сложным форматам и типам данных, причём при интеграции различных данных сложность увеличивается лавинообразно. Поэтому портал LEAD ориентирован на демократизацию доступа к технологиям исследования погоды.

Второй целью проекта LEAD является интеграция задач, возникающих при обнаружении, анализе и предсказании мезомасштабных атмосферных яв-

лений. Большинство технологий, использующихся для наблюдения за атмосферой, предсказания её эволюции, вычисления, передачи и хранения информации никак не связано между собой и функционирует раздельно. Радары сканируют области штормов не адаптивно, численные модели запускаются на фиксированный промежуток времени и с фиксированным набором параметров, киберинфраструктура не позволяет метеорологическому инструментарию работать в реальном времени, изменяя настройку своих режимов в ответ на изменение погоды. Как результат – современные информационные технологии в области исследования погоды далеки от оптимальных в каждом конкретном случае. Для преодоления этих ограничений LEAD развивает возможности, позволяющие вычислительным моделям и вспомогательному инструментарию откликаться динамически на изменение выходных данных предыдущих задач, данных наблюдений и пользовательских данных. Также развиваются средства, позволяющие моделям и иным инструментам динамически озадачивать адаптивные системы наблюдения, в первую очередь доплеровские лидары, для получения данных когда и где это необходимо.

Средства портала LEAD позволяют выполнять следующие действия:

- Запрос и получение информации, такой как: наборы данных о наблюдениях (включая потоки в реальном времени), выходные данные моделей, определения и связи между метеорологическими величинами, статус информационных ресурсов или потока работ.
- Выполнение расчётов с помощью численных моделей атмосферы (в частности, WRF, которая может использоваться как в базисном варианте, так и в полном, с расчётом полной физики, прилежащих территорий и неоднородными начальными условиями).
- Усвоение данных для объединения наблюдений при определённых физических ограничениях, с учётом фоновой информации для анализа 3D-моделей.
- Накопление и последующий анализ данных наблюдений и результатов вычислений для получения качественной информации о пространственно-временных отношениях между различными полями и процессами.
- Визуализация одно-, двух- и трёхмерных данных.

LEAD основан на сервисно-ориентированной архитектуре SOA и подразделяется на пять уровней:

1. Нижний уровень, представляет ресурсы, в число которых включены приложения и данные, распределённые по LEAD Grid.

2. Следующий уровень содержит веб-сервисы, обеспечивающие доступ к базовым средствам, а также к инструментарию, осуществляющему доступ к данным.

3. Множество конфигурирующих и исполняющих сервисов образуют следующий слой, отвечающий за управление потоком работ LEAD.

4. Четвёртый уровень содержит сервисы для работы с данными. Сервис онтологий разрешает концепты верхнего уровня по специфическим схемам, используемым в различных сервисах данных, а сервисы декодирования и обмена преобразуют данные из одной формы в другую.

5. Наконец, на верхнем уровне архитектуры находится LEAD-портал, являющийся контейнером для интерфейсов пользователя, называемых портлетами, обеспечивающих доступ к индивидуальным сервисам нижних уровней.

Таким образом, LEAD представляет собою сложный проект, включающий в себя:

- Специализированный веб-портал.
- ADAS(ARPS Data Assimilation System) – инструмент для получения данных и контроля их качества, включающий в себя подготовку начальных условий для вычислительной модели.
- MyLEAD – сервис для работы с персонализированными данными пользователя, в котором происходит сбор метаданных на всех этапах исследования.
- WRF – мезомасштабная климатическая модель, о которой уже говорилось выше.
- ADaM (Algorithm Development and Mining) – инструмент для поиска и извлечения закономерностей из данных наблюдений, множества усваиваемых данных и выходных данных моделей.
- IDV (Integrated Data Viewer) – приложение для визуализации многомерных геофизических данных.

Сила LEAD не в возможностях разных инструментов, а в организации их совместной работы для широкого круга проблем.

6 Портал АТМОС

Информационно-вычислительная система «Климат» (<http://climate.atmos.iao.ru>) [43–45], с одной стороны, демонстрирует возможности математического моделирования для решения отдельных климатических задач, а с другой стороны – на основе технологий глобальной сети и средств графической визуализации осуществляет взаимодействие с "потребителями" информации о результатах решения таких задач. В области региональных климатических моделей одна из вычислительных компонент ИВС «Климат» представляет собой мезомасштабную модель MM5, разработанную в Национальном центре атмосферных исследований США. ИВС также имеет и информационную компоненту, связанную с последующим преобразованием результатов расчетов в информацию, необходимую конкретным пользователям (студентам, аспирантам и преподавателям; специалистам по проблемам окружающей среды и т.д.).

- Пользовательский интерфейс для работы с моделью MM5 разделен на три части:

- Инициализация рельефа, категорий землепользования, растительности и других стационарных географических параметров выбранных областей.
- Подготовка данных для задания граничных условий, интерполяция начальных приближений на расчетную сетку модели.
- Подготовка к запуску модели MM5.

В разделе интерфейса, посвященном заданию стационарных географических параметров, пользователем вводятся следующие величины: долгота и широта центра расчетной области, максимальная для всех областей горизонтальная сетка, тип картографической проекции, типы сглаживания и растительности, количество вложенных областей и число вертикальных слоев. Затем для каждой области задается шаг сетки в километрах и номера индексов нижнего угла в координатах модели. Данные этого шага используются, в основном, для генерации входных параметров программы TERRAIN.

Следующий раздел интерфейса позволяет задать период моделирования, давление на верхней границе расчетной области и набор новых уровней давления в расчетной сетке, а также метод используемой интерполяции. Здесь вводятся конфигурационные параметры для пакетов REGRID (связки pregrid/regidder) и INTERPF.

На последнем шаге задаются параметры самой модели MM5, такие как: описание микрофизики влаги, параметризация облачности, параметризация планетарного пограничного слоя и параметризация излучения в атмосфере. На основе данных этого и предыдущих шагов создается набор конфигурационных файлов модели, после чего параллельная версия MM5 ставится в очередь на выполнение на вычислительном кластере Института оптики атмосферы. Постановка задач в очередь и управление очередью осуществляется с помощью менеджера Torque [47].

Так как вычислительные модули программы MM5 зависят от количества вложенных областей, выбранных пользователем для исследования, перед решением каждой задачи модель MM5 приходится перекомпилировать. Если на этапе компиляции или исполнения параллельной версии модели возникают ошибки при вычислениях (это может происходить как в случае неправильного задания входных условий, так и при некорректной работе модели), то они предоставляются конечному пользователю в виде, получаемом на выходе программы MM5. В зависимости от настройки, проводимой администратором кластера, в вычислениях могут использоваться от 2 до 20 процессоров (на текущий момент – 10, пять двухпроцессорных узлов).

Выходные данные, получаемые в результате расчетов, имеют достаточно большой объем. Так, например, для расчета изменения локальной метеорологии за сутки с шагом в 30 секунд для трех вложенных областей суммарный размер массива результатов достигает трех-четырех сотен мегабайт. Эти данные хранятся в файловой системе веб-

сервера, в директории, имя которой совпадает с уникальным идентификатором, что позволяет связать их с пользователем и решенными им задачами. Передавать по сети Интернет настолько значительные объемы числовой информации на текущий момент экономически необоснованно, поэтому пользователю предоставляется возможность строить отображение результатов в виде карт с изолиниями, цветной заливкой или отображением векторных полей. Пример такой графики представлен на рис. 3 для температуры на 30 апреля 1999 г. в области размером 50×50 км с центром в г. Томске (85 градусов восточной долготы и 56,5 градусов северной широты).

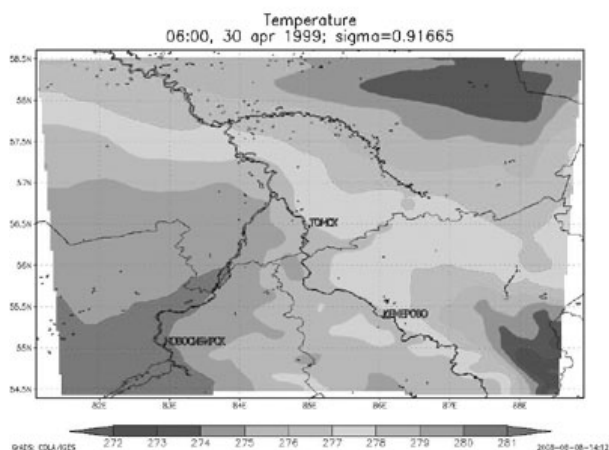


Рис. 3. Представление данных на карте в ИБС «Климат»

После выполнения задачи и получения выходных файлов модели, пользователю предоставляется возможность просмотра полученных результатов по каждой вычисленной величине и в любой момент расчетного интервала в графическом виде, используя для этого систему анализа и отображения данных GrADS [48].

Современная система моделирования WRF развивалась в течение последних нескольких лет. Это мезомасштабная система численного предсказания погоды следующего поколения, разработанная как для нужд оперативного прогнозирования погоды, так и более глубокого исследования атмосферы. Модель WRF разработана, чтобы быть гибкой, современной атмосферной системой моделирования, она является переносимой и эффективной для использования на имеющихся в распоряжении параллельных вычислительных платформах. WRF может использоваться при решении широкого диапазона задач в масштабах от метров до тысяч километров, включая идеализированные течения (например, конвекция, бароклинные волны), исследование параметризации, исследование сравнения данных, численное предсказание погоды в реальном времени, исследование прогноза и обучение.

Входные данные для модели WRF, также как и MM5, представлены в виде файлов в формате GRIB. В настоящий момент в ИБС «Климат» доступны

данные ECMWF для временного промежутка с 1991 по 2002 год.

Блок-схема системы моделирования WRF (для WRFV2)

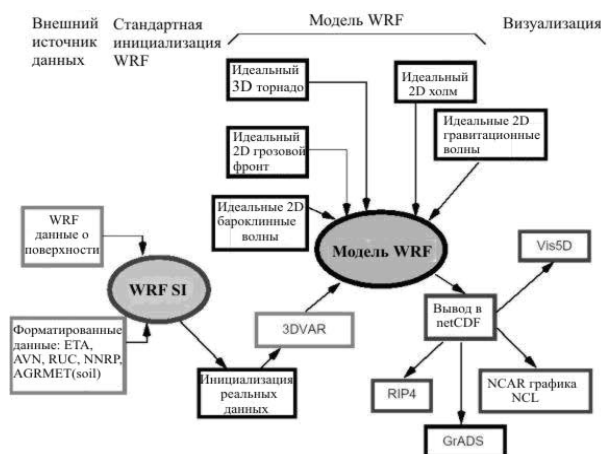


Рис. 4. Схема моделирующей системы WRF

Используемый в ИБС «Климат» расчетный блок модели WRF относится к версии 2.0. Он использует схему моделирования, представленную на рисунке 4. Как показано в диаграмме, система моделирования WRF состоит из следующих основных программ: стандартная инициализация WRF (WRF SI), WRF 3DVAR, модель WRF и графические инструменты для постобработки. **WRF SI** используется, прежде всего, для моделирований на основе реальных данных. Ее функции включают определение областей моделирования; интерполяцию данных земной поверхности (таких как ландшафт, землепользование, и типы почвы) для выбранной области моделирования; и дегриббинг и интерполяция метеорологических данных другой модели к этой области моделирования и к вертикальной координате модели. **WRF 3DVAR** является дополнительной программой, но может использоваться, чтобы включить наблюдения в интерполированные исследования, созданные WRF SI. Она также может использоваться, чтобы обновить начальное условие модели WRF, когда модель WRF находится в циклическом режиме. **Модель WRF** – ключевой компонент системы моделирования, который состоит из нескольких программ инициализации для идеализированного моделирования, и моделирований на основе реальных данных, и числовой интеграции программы. Он также включает программу, чтобы производить одностороннее вложение.

Для модели WRF ввод пользователем данных происходит уже в четыре шага:

- Инициализация рельефа, категорий землепользования и растительности.
- Задание стационарных географических параметров выбранных областей.
- Подготовка данных по временному периоду и уровням давления.
- Инициализация физических параметров модели.

На первом шаге интерфейса вводятся долгота и широта центра расчетной области, максимальные размерности вычислительной стеки, расстояние между узлами сетки, тип картографической проекции, число вертикальных слоев и количество вложенных областей.

На следующем этапе для каждой из областей моделирования пользователем заносятся отношения размеров областей, а также индексов, как нижнего левого, так и верхнего правого углов положения вложенной области в расчетной сетке материнской области.

На третьем шаге проходит задание временных характеристик периода моделирования и уровней давления. Результатом первых трех этапов является составление конфигурационных файлов программы WRFSI.

Затем, на четвертом и последнем этапе для каждой из областей инициализируются опции модели для микрофизики, длинноволновой и коротковолновой радиации, расчета параметров поверхности земли, пограничного слоя и облаков. Пользователь решает также учитывать ли при моделировании потоки влажности и тепла от поверхности, влияние облаков на параметризацию радиации и влияние снежного покрова, после чего, на основе полученных данных генерируются входные файлы конфигурации непосредственно для модели WRF.

7 Заключение

Авторы благодарны РФФИ и СО РАН (интеграционный проект №34) за финансирование работы.

Литература

- [1] Pielke R.A. Mesoscale Meteorological Modeling. Orlando: Academic Press, 1984. 622p.
- [2] Grell G.A. et al. A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). NCAR Tech. Note, NCAR/TN-398+IA. 1993. 122 p.
- [3] Фазлиев А.З., Информационные ресурсы и Интернет-технологии для наук об окружающей среде, Вычислительные технологии, Специальный выпуск, т.9, ч.1, с.11-21, 2000.
- [4] LEAD Portal
<https://portal.leadproject.org/>
- [5] ATMOS Portal
<http://atmos.iao.ru>
- [6] GRIB Edition 1,
<http://dss.ucar.edu/docs/formats/grib/gribdoc/>
- [7] WMO BUFR,
<http://www.esdswg.org/spg/docindexfolder/heritage/wmo-bufr/>
- [8] Thorpe, W., "Guide to the WMO Code Form FM 94-IX EXT. BUFR," Fleet Numerical Oceanography Center, Monterey, California
- [9] NetCDF Guide,
http://www.unidata.ucar.edu/software/netcdf/guide.txn_toc.html
- [10] eXternal Data Representation,
<ftp://ds.internic.net/rfc/rfc1832.txt>
- [11] NetCDF Conventions,
<http://www.unidata.ucar.edu/software/netcdf/conventions.html>
- [12] Hierarchical Data Format,
<http://www.hdfgroup.org/products/hdf5>
- [13] Department of Atmospheric and Oceanic Science, University of Maryland
<http://www.atmos.umd.edu/~mm5/>
- [14] University of Wisconsin – Madison
<http://helios.aos.wisc.edu/>
- [15] University of Wisconsin – Milwaukee
<http://sanders.math.uwm.edu>
- [16] Purdue University
<http://wpx.eas.purdue.edu/>
- [17] University of Illinois
<http://blizzard.atmos.uiuc.edu/rt/>
- [18] Millersville University, PA
<http://snowball.millersville.edu>
- [19] Naval Postgraduate School
<http://met.nps.edu>
- [20] National Weather Service Forecast Office
<http://www.srh.noaa.gov/>
- [21] NASA GHCC Real-Time WRF Weather Model System
<http://www.ghcc.msfc.nasa.gov/>
- [22] NOAA/NSSL
<http://www.nssl.noaa.gov/>
- [23] NCEP/EMC
<http://www.nco.ncep.noaa.gov/>
- [24] NOAA/GSD
<http://ruc.fsl.noaa.gov/taq.html>
- [25] NCAR WRF-ARW
<http://www.wrf-model.org/>
- [26] The SNU-NCAR Joint Center for High-Impact Weather and Climate Research
<http://jhwc.snu.ac.kr/weather/>
- [27] RAP Real-Time Weather (4 Models – WRF, RUC, ETA, GFS)
<http://www.rap.ucar.edu/weather/model/>
- [28] NCAR/MMM (2 Models – WRF, MM5)
<http://www.mmm.ucar.edu/>
- [29] Centro Funzionale Meteorologico di Protezione Civile della Regione Liguria
<http://www.meteoliguria.it/map/bolam/bolamin.htm>
- [30] Servizio Agrometeorologico Regionale – Sardegna
<http://www.sar.sardegna.it/servizi/meteo/>
- [31] Institute of Atmospheric Sciences and Climate (ISAC) of the Italian National Research Council (CNR)
<http://oracolo.le.isac.cnr.it/mpcf/METEO/wrf.php>
- [32] LaMMA - CNR IBIMET (Laboratory for Meteorology and Environmental Modeling of Tuscany Region) – WRF NMM
<http://www.lamma.rete.toscana.it/>
- [33] Meteo Riccione, Italy
<http://www.meteoriccione.it/>
- [34] Slovenian Meteorological Amateur Research Team
<http://www.pro-vreme.net/>

- [35] National Observatory of Athens (3 Models – WRF, MM5, BOLAM)
<http://cirrus.meteo.noa.gr/>
- [36] Meteoblue
<http://www.meteoblue.ch/>
- [37] Meteosim Truewind (2 Models – MASS, GFS)
<http://www.meteosimtruewind.com/>
- [38] AMI Environmental. South Asia
<http://www.vnbaolut.com/>
- [39] Institute for the Environment, the Hong Kong University of Science and Technology
<http://envf.ust.hk/dataview/mm5/>
- [40] DTC WRF-NMM
<http://www.dtcenter.org/wrf-nmm/>
- [41] Advanced Operational Aviation Weather System, Taiwan
<http://aoaws.caa.gov.tw/>
- [42] Institute of Astronomy and Meteorology of the University of Guadalajara, Mexico
<http://www.udg.mx/wrf/>
- [43] Weather-It-Is, Israel
<http://www.weather-it-is-israel.com/>
- [44] Беликов Д.А., Лаврентьев Н.А., Старченко А.В., Фазлиев А.З. Информационно-вычислительная система для коллективного исследования проблем атмосферного пограничного слоя // Вычислит. технологии, Специальный выпуск. – 2006, т.11. – С. 73–79.
- [45] Беликов Д.А. Вражнов Д.А. Лаврентьев Н.А. Старченко А.В., Фазлиев А.З. Использование моделирующих систем MM5 и WRF в веб-портале ATMOS // Измерения, моделирование и информационные ресурсы для изучения окружающей среды. – Томск : Изд-во ЦНТИ, 2006. – С. 36–41.
- [46] Лаврентьев Н.А. Старченко А.В., Фазлиев А.З. Организация удаленной работы с мезомасштабной моделью MM5 // Труды 11-й Байкальской Междунар. конф. «Информационные и математические технологии». Ч. 2. – Иркутск : ИСЭМ СО РАН, 2006. – С. 175–179.
- [47] Torque,
<http://www.clusterresources.com/pages/products/torque-resource-manager.php>
- [48] GRADS,
<http://grads.iges.org/grads/grads.html>

Informational system oriented on weather and climatic data representation in the Internet

N.A. Lavrent'ev, A.Z. Fazliev

Internet informational resources representing the data on weather and climate, especially on temperature and pressure are of great interest to the users. In our research work we present an overview of informational resources on weather and climate, both informational ones and computational, that defines the spectrum of possibilities of users about which they are often unaware. The topics of data storage centers organization, data formats of common models and informational systems representing weather data are discussed in this article. There is also a short overview of LEAD and ATMOS portals which grant a user a possibility of individual calculations conduction with the help of mesoscale models.

Вариации магнитного поля Земли как составной элемент баз данных космических экспериментов по физике магнитосферы

© А. Зайцев, В. Петров, В. Одинцов, В. Шилимов

Институт земного магнетизма, ионосферы и распространения радиоволн им. Н.В. Пушкова
Российской академии наук (ИЗМИРАН)
zaitsev@izmiran.ru

Аннотация

В данной работе рассмотрена структура БД некоторых современных космических экспериментов (Интербол, Кластер и Фемида) и оценена роль данных о вариациях магнитного поля Земли при формировании этих баз данных. Для исследований физических явлений в магнитосфере традиционным является использование координированных как спутниковых, так и наземных данных. Предложено рассматривать сеть наземных обсерваторий как комплексную систему датчиков, дающих сведения о взаимодействии планеты Земля с космической средой. На этой основе сделана оценка современного развития ИС и БД по космическим экспериментам по физике магнитосферы, оформившегося сегодня в виде самостоятельного направления по созданию и ведению электронных библиотек (ЭБ) и данных по комплексным космическим экспериментам в виде открытых баз данных (БД) в поддержку программ по исследованиям магнитосферы. При этом все более широко используются элементы виртуальных обсерваторий и доступ к распределенным базам данных. Авторы данной работы выполняют грант РФФИ 07-07-00377 «Информационная система диагностики магнитосферы методом геомагнитных меридианов».

1 Постановка задачи

По мере того, как совершенствуются методики и техника координированных космических экспериментов, направленных на исследования магнитосферы, все более возрастает роль наземных наблюдений как составного элемента баз данных по таким экспериментам. Без детальной привязки к естественной системе координат – земному магнитному

полю – не представляется возможным провести качественную интерпретацию и анализ сложных явлений в космическом пространстве. На создание и ведение баз данных затрачиваются средства, сравнимые с теми, которые затрачены на подготовку и проведение космических экспериментов. Если данные первых космических экспериментов были авторскими, т.е. доступ к ним был возможен только в рамках сотрудничества с его создателями, то в настоящее время доступ к таким данным реализуется в режиме открытых источников. При этом реализовано два режима – режим реального времени (как пример – данные спутника АСЕ, данные по проектам Кластер и Фемида) и режим отложенного доступа, особенно в том случае, если требуется дополнительная проверка, калибровка или верификация данных (как например база данных по проекту Интербол). Но при всех случаях данные в полном объеме становятся доступными по сети Интернет в формате веб-сервисов [7–14]. Поэтому актуальным является создание ИС и БД в открытом доступе с развитым сервисным и программным обслуживанием.

2 Источники данных

Для исследований физических явлений в магнитосфере традиционным является использование координированных баз как спутниковых, так и наземных данных. В известном смысле сеть наземных обсерваторий представляет собой комплексную систему датчиков, дающих сведения о взаимодействии планеты Земля с космической средой. Здесь основными элементами выступают данные магнитных и ионосферных обсерваторий, а также специальные сети, включающие радиотехнические системы типа радаров некогерентного рассеяния, лидаров, сеть риометров в высоких широтах, сеть камер всего неба для наблюдений полярных сияний, сеть станций космических лучей, и т.д. [8–11]. По всем этим инструментам накапливаются базы данных, идет их развитие, создаются новые средства первичной обработки и отображения данных. Примером таких сложных систем могут служить программа Интермагнит, программа AMIE (Assimilative

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Mapping of Ionosphere Electrodynamics), радары для диагностики ионосферы SuperDARN, см. <http://superdarn.jhuapl.edu/index.html> и т.д. По всем этим программам создаются БД и открытые справочно-информационные системы, а сами данные входят составным элементом в интегрированные базы данных по космическим экспериментам.

При этом число таких исходных баз данных с одинаковым типом данных может быть довольно велико. Для наземных геомагнитных данных в данной статье приведены ссылки на 5 источников собранных в базы геомагнитных данных, а если учесть отдельные обсерватории, данные которых доступны через Интернет, то общее количество источников данных достигает нескольких десятков. Если учесть, что, как правило, все эти источники имеют разную структуру самих баз и используют разные форматы хранения данных, то становится необходимой разработка отдельной управляющей СУБД, которая может формировать запросы к исходным базам и предоставлять пользователю выборку по всем оригинальным источникам данных.

3 Направление работы

В данной работе сделана попытка представить описание современного развития ИС и БД по космическим экспериментам по физике магнитосферы, оформившегося сегодня в виде самостоятельного направления по созданию и ведению электронных библиотек (ЭБ) и данных по комплексным космическим экспериментам в виде открытых баз данных (БД) в поддержку программ по исследованиям магнитосферы. Примером современных БД можно считать результаты работ по проектам Интербол, Кластер, THEMIS. Во всех этих экспериментах данные по вариациям магнитного поля служат отправной точкой при анализе отдельных уникальных событий или основным признаком при типизации или классификации наблюдаемых явлений. Геомагнитные индексы K_p , A_E и D_{st} являются в этих БД основой для оценки состояния магнитосферы Земли и базовым признаком (репером) при статистических исследованиях. Нами рассмотрена структура БД современных космических экспериментов и оценена роль данных о вариациях магнитного поля Земли при формировании этих баз данных. Кроме того, проведена оценка состояния известных баз данных (БД) по вариациям магнитного поля и намечена перспектива использования этих БД в конкретных исследованиях, в том числе по грантам РФФИ. Авторы данной работы выполняют грант РФФИ 07-07-00377 «Информационная система диагностики магнитосферы методом геомагнитных меридианов».

4 Инструменты ведения и пользования БД

Важным аспектом создания и развития ЭБ и БД является разработка стандартных инструментов и пакетов программ для работы с этими базами дан-

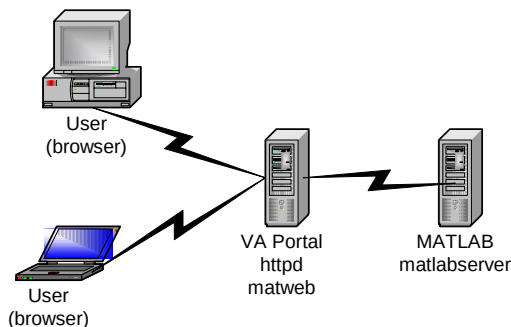
ных. Все инструменты делятся на две группы: первая группа – специализированные для первичной обработки данных конкретных приборов, будь то приборы на спутниках или сложные радиосистемы на Земле. Вторая группа использует пакет стандартных программ, обработки и анализа данных, из которых наиболее известны и распространены MATLAB и IDL. Пакеты хорошо разработаны, свободно распространяются в нашей стране на коммерческой основе, и активно используются специалистами по физике магнитосферы. Такой пример имеется в ИЗМИРАН, где используется MATLAB веб-сервер для работы с наземными геомагнитными данными, см. <http://matlab.izmiran.ru/magdata/>. Эта работа была выполнена в рамках гранта РФФИ 03-07-90066 "Интерактивный WEB-ресурс для адаптивной обработки экспериментальных данных на основе инструментального средства MATLAB Web Server". В рамках этого проекта был открыт веб-сервис, позволяющий вести обработку данных виртуального архива в интерактивном режиме с использованием всех возможностей вычислительной среды MATLAB. Эти же методики и инструменты могут быть использованы для работы и с другими типами наземных и космических данных.

Виртуальные архивы, в их нынешнем понимании, рассматриваются лишь как средства поиска и загрузки данных для обработки пользователем. В лучшем случае к ним могут быть добавлены программы начальной обработки и графического представления запрошенных данных. Для более сложной обработки данные должны быть загружены в вычислительную среду со всеми вытекающими проблемами совместимости форматов хранения данных и их представления в вычислительной среде.

Наиболее простой и эффективный способ избежать проблем согласования форматов заключается в создании универсального интерфейса между виртуальными архивами и развитыми системами математической обработки, такими, например, как MATLAB и IDL. В качестве примера можно привести интерактивный WEB-ресурс для обработки геофизических, данных, разработанный в ИЗМИРАН на основе стандартного пакета MATLAB Web Server (MWS). MWS работает в сети Интернет по сетевому протоколу TCP/IP и обеспечивает, с одной стороны, обмен данными между базами данных и вычислительной средой для их обработки и, с другой стороны, передачу пользователю результатов обработки данных и их отображение с помощью стандартного Web-браузера. Причем, пользователь, база данных и вычислительная среда могут находиться в совершенно различных местах, связанных между собой лишь посредством Интернет.

Установка необходимого оборудования и программных средств, обеспечивающих функционирование MWS, требуется только на специально выделенном сервере. В общем случае Web-браузеры пользователей работают на клиентских машинах (их может быть любое количество), в то время как

MATLAB, MWS и сетевой сервер (httpd) запущены на одном серверном компьютере. В более сложном случае сетевой сервер и базы данных могут располагаться на разных машинах.



В более сложном случае сетевой сервер и базы данных могут располагаться на разных машинах.

В качестве следующего шага предполагается полная интеграция виртуальных архивов с системами обработки данных. Эта интеграция должна быть прозрачной для конечного пользователя, и как можно более простой в использовании. В то же самое время она должна обеспечивать математическую обработку данных любой сложности, ограничиваемую лишь возможностями конкретной вычислительной среды, используемой для обработки.

В настоящее время существует большое количество разработанных специализированных пакетов MATLAB для обработки массивов данных в различных областях знаний. Однако их использование ограничивается кругом их разработчиков. Эти пакеты можно было бы преобразовать таким образом, чтобы их можно было находить и получить посредством поисковых систем и включить при необходимости в стандартные среды обработки. Это один из путей, ведущих к превращению виртуальных архивов (ВА) из простых средств хранения и поиска данных в эффективные средства виртуальной обработки данных – виртуальные лаборатории (ВЛ).

5 Авторские БД и опыт работы

Начиная с момента появления первых цифровых магнитно-вариационных станций, обеспечивших поступление регулярных данных с магнитных обсерваторий, в ИЗМИРАН были развернуты работы по созданию БД по вариациям магнитного поля Земли [1-2]. Первый этап этой работы был закончен в середине 90-х годов и первый CD-ROM был создан в 1995 году и размещен в сети Интернет, см. www.iki.rssi.ru/magbase. Размещение на сайте ИКИ РАН было вызвано необходимостью использования этой БД в проекте Интербол. В последующий период времени эти работы были продолжены, несмотря на существенное сокращение числа сотрудников. Примером может служить работа ЦМБС-6 в обсерватории Москва и ЦМБС-2 на станции Мыс Шмидта. Эти данные входят в текущие БД и привлекаются для работ по другим космическим проектам. В

качестве основного инструмента для работы с БД нами был разработан многопользовательский интерфейс в среде MATLAB, который обеспечивает передачу данных по запросам пользователей из виртуального архива в вычислительную среду MATLAB и обратную отправку пользователям результатов обработки заданных временных рядов в виде графических образов или файлов, содержащих результаты обработки и данные для последующего анализа средствами пользователя, см <http://matlab.izmiran.ru/magdata/>. Таким образом, MATLAB Web Server обеспечивает возможность поиска данных в распределенных базах данных в соответствии с заданными критериями поиска, передачу отобранных данных в вычислительную среду MATLAB для обработки в соответствии с заданными параметрами обработки и представление результатов обработки в соответствии с заданными формами отображения результатов. Сравнительная простота подготовки интерфейсных HTML-форм и MATLAB приложений (m-файлов) для создания собственных математических конструкций обработки данных с поддержкой сложной графики, работающих в любом стандартном Web-браузере, обеспечивает доступ к данным и вычислительным ресурсам виртуального архива широкому кругу пользователей.

Другим примером создания и ведения баз данных в ИЗМИРАН может служить электронный каталог знаков секторной структуры межпланетного магнитного поля (СС ММП) за достаточно продолжительный период времени с 1957 г. и по настоящее время.

Секторная структура ММП (СС ММП) была открыта в 60-х гг. прошлого века путем анализа данных прямых космических измерений. Выдающуюся роль сыграл в этом открытии известный американский физик и инженер Н. Несс (N. Ness). В настоящее время определения СС ММП базируются на трех основных источниках данных: наблюдениях магнитного поля Солнца в солнечных обсерваториях (Стенфорд, Маунт Вилсон и Гавайи), измерениях на спутниках в солнечном ветре в точке L1 (ACE, WIND), и магнитных измерениях на Земле в полярной шапке (пара обсерваторий Восток – Резольют-Бей).

Основная физическая идея косвенного метода нахождения знака ММП была предложена независимо и одновременно российским магнитологом С.М. Мансуровым и датским исследователем Л. Свалгаардом. Она основана на том, что перестройка магнитосферы при переходе из сектора одного знака в сектор другого знака сопровождается перестройкой токовых систем магнитосферы – ионосферы. Оба названных исследователя решили задачу нахождения знака ММП путем анализа геомагнитных вариаций в приполюсных (полярных) регионах. При этом удалось составить каталоги знаков ММП за длительный промежуток времени практически без серьезных пропусков. Л. Свалгаард опубликовал каталог секторной структуры ММП, восстановлен-

ной для эпохи, предшествующей началу космических измерений. Большая заслуга в деле создания каталогов принадлежит С.М. Мансурову, который заложил основу метода регулярного определения знака секторов ММП по наземным измерениям, и, начиная с 1975 г. такая работа ведется в ИЗМИРАН постоянно. Регулярно пополняемый каталог знаков ММП доступен пользователям на сайте ИЗМИРАН: <http://www.izmiran.ru/stp/polar/SSIMF/>.

Принимая во внимание тот факт, что по данным спутников в точке либрации L1 ведется регулярное определение секторной структуры ММП (см. <http://omniweb.gsfc.nasa.gov/html/polarity/>), мы имеем возможность сравнивать ее с таблицами, получаемыми на основе наземных данных в полярных широтах и оценивать долговременные изменения в динамике солнечного ветра на орбите Земли. Методы определения знака ММП по вариациям магнитного поля Земли дают возможность, в принципе, охватить исследованиями интервал времени порядка столетия.

Влияние СС ММП на общее состояние активности геомагнитного поля наблюдается во всех обсерваториях, включая средне- и низкоширотные. Этот факт служит основанием для развития методов определения СС ММП по данным среднеширотной обсерватории Москва с целью уточнения таблиц СС ММП, определяемых по станциям в полярной шапке.

При развитии информационных систем (ИС) и баз данных (БД) напрашивается построение виртуальных обсерваторий. В качестве пробного шага усилиями сотрудников ИЗМИРАН и Массачусетского технологического университета США разработан проект виртуальной магнитной обсерватории для работы с базами геомагнитных данных http://serv.izmiran.ru/webff/magdb_all.html [17, 19]. В основу VGMO заложена технология промежуточного хранения данных во временных файлах, так называемых «flat file», создаваемых на сайте, где развернута виртуальная обсерватория.

Однако создание такой управляющей базы данных как WEB сервиса сталкивается с несколькими проблемами. Первой проблемой являются авторские права. Многие открытые источники данных позволяют воспользоваться ими любому пользователю по запросу и регистрации, но не разрешают их дальнейшее распространение. Второй проблемой является сетевая безопасность. Создать эффективную СУБД основанную только на http протоколе довольно сложно, а использование апплетов и создание прямых каналов связи (сокетов) между клиентом и сервером часто ограничивается администраторами сетей. В качестве способа разрешения этих проблем была создана так называемая stand along версия VGMO. Она должна устанавливаться на компьютере конечного пользователя как самостоятельное приложение. В этом случае обе указанные проблемы снимаются.

При работе с космическими данными важный момент – создание ресурсов не только для науки, но и

для образования. В этом направлении авторы предприняли ряд шагов, в частности была издана книга «Космическая среда вокруг нас» с приложением избранных ресурсов по солнечно-земной физике в виде CD-ROM «Космос вокруг нас» [6]. Эффективность использования геомагнитных данных при анализе результатов космических измерений подтверждена в многочисленных публикациях, ссылки на которые размещены на сайтах проектов Интербол [16], Кластер [18] и других ресурсах сети Интернет.

6 Заключение

Представленные сведения по развитию баз данных (БД) и информационных систем (ИС) в поддержку координированных космических экспериментов, направленных на исследования магнитосферы, подтверждают значимость наземных геомагнитных наблюдений как составного элемента баз данных по таким экспериментам. При этом все более широко используются элементы виртуальных обсерваторий и доступ к распределенным базам данных. Дальнейшее развитие баз данных идет по двум направлениям. Первое – все более широкое и универсальное строительство БД и ИС, где организован доступ к максимально возможным ресурсам. Как пример можно представить российский проект «Электронная Земля» или Единую информационную систему по фундаментальным космическим исследованиям (ЕИС – ФКИ), создаваемую в Академии наук. Второе направление – совершенствование инструментов и содержания конкретных БД данных по отдельным обсерваториям. По мере совершенствования линий коммуникаций, удастся организовать высокоскоростные соединения и доступ к большим объемам данных почти в реальном времени. Для успеха научных исследований наша работа развивает практику и перспективы развития ИС и БД как основы системы и разработки средств для диагностики магнитосферы методом геомагнитных меридианов. Настоящая работа представляет собой попытку дать введение к современным космическим проектам (описание состояния этих дел на русском языке) с учетом нашего опыта поддержки космических экспериментов наземными (геомагнитными) данными.

Литература

- [1] Зайцев, А. Разработка Интернет ресурсов по солнечно-земной физике для науки и просвещения / А. Зайцев [и др.]. // Труды Четвертой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции RDCL-2002 (Дубна, 15–17 окт. 2002 г.) : в 2 т. – Дубна : ОИЯИ, 2002. – Т. 1 – 335 с. – С. 299–303.
- [2] А. Зайцев, В. Ишков, А. Козлов, В. Обридко, В. Одинцов, В. Петров, В. Назаров, М. Тихонова, Развитие Российских Интернет-ресурсов по

- солнечно-земной физике // Труды 5-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2003, Санкт-Петербург, Россия, 2003.
- [3] Длужневская, О.Б. Российская виртуальная обсерватория. Информационная система Центра астрономических данных / Длужневская О.Б. [и др.]. // Труды Четвертой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции RDCL-2002 (Дубна, 15–17 окт. 2002 г.) : в 2 т. – Дубна : ОИЯИ, 2002. – Т. 1 – 335 с. – С. 202–212.
- [4] Вольфенгаген, В.Э. Создание и развитие российских информационных, вычислительных и телекоммуникационных ресурсов для поддержки фундаментальных научных исследований (анализ результатов работ, поддержанных грантами РФФИ) / В.Э. Вольфенгаген [и др.]. // Труды Четвертой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции RDCL-2002 (Дубна, 15–17 окт. 2002 г.) : в 2 т. – Дубна : ОИЯИ, 2002. – Т. 1 – 335 с. – С. 12–16.
- [5] Ахлестин, А.Ю. Сайт SOLTER – «Атмосфера и гелио-геофизическая обстановка в Сибири» / А.Ю. Ахлестин [и др.]. // Труды Четвертой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции RDCL-2002 (Дубна, 15–17 окт. 2002 г.) : в 2 т. – Дубна : ОИЯИ, 2002. – Т. 1 – 335 с. – С. 281–289.
- [6] Будько Н., Зайцев А., Карпачев А., Козлов А., Филиппов Б. Космическая среда вокруг нас – введение в исследования околоземного космического пространства, книга и приложение на CD-ROM, издательство ТРОВАНТ, Троицк, с. 245, 2006.
- [7] Матлаб веб-сервер для работы с магнитными данными: <http://matlab.izmiran.ru/magdata/>.
- [8] Ресурс Интернет по геомагнитным данным СуперМаг: <http://supermag.jhuapl.edu>.
- [9] Геомагнитный меридиан 210 : <http://stdb2.stelab.nagoya-u.ac.jp/mm210/>.
- [10] Международный проект Интермагнит : <http://www.intermagnet.org/>.
- [11] Международный проект IMAGE – меридиан 105 : <http://www.ava.fmi.fi/image/>.
- [12] Сеть магнитометров на Аляске: <http://magnet.gi.alaska.edu/>.
- [13] Международный проект THEMIS (Фемида) : <http://themis.ssl.berkeley.edu/index.shtml>.
- [14] Международный проект Кластер, ЦУП : <http://www.jsoc.rl.ac.uk/jsocLinks.php>.
- [15] Наземная поддержка проекта Кластер : <http://www.wdc.rl.ac.uk/gbdc/gbdc.html>.
- [16] Международный проект Интербол : <http://www.iki.rssi.ru/interball/>.
- [17] Проект виртуальной магнитной обсерватории: http://serv.izmiran.ru/webff/magdb_all.html.
- [18] Международный проект Кластер, публикации в реферируемых журналах: <http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=39766>.
- [19] Papitashvili, V.O., Saxena A.B., Petrov V.G., C.R. Clauer, and Papitashvili N.E. A Virtual Global Magnetic Observatory: VGMO.NET, Earth, Planets and Space, 58, No. 6, 765–774, 2006.

Geomagnetic Field Variations as the Part of Data Base in the Space Experiments for Magnetosphere Physics

A. Zaitsev, V. Petrov, V. Odintsov, V. Shilimov

In the paper we consider of some characters and demands in the structure of contemporary data bases of space experiments as INTERBALL, CLUSTER and THEMIS. We search of demands for geomagnetic field variations as the part included in the data bases for such space experiments. The network of ground-based instruments was served as the reference level for space experiments. The collected data of magnetic observatories also used for virtual archives. This work partially supported by RFBR grant 07-07-00377.

Развитие инфраструктуры распределенных хранилищ спутниковых данных: интегрированная распределенная среда неоднородных информационных ресурсов исследования Земли из Космоса*

© Е.Б. Кудашев, А.Н. Филонов

Институт космических исследований РАН, Институт проблем информатики РАН
kudashev@iki.rssi.ru, filonov_a@inbox.ru

Аннотация

Статья представляет собой обзор текущего состояния технологий в области распределенных информационных систем, предназначенных для хранения и предоставления доступа к данным дистанционного зондирования Земли из космоса, а также возможных путей развития подобных систем в ближайшие годы. Рассматриваются отдельные технические аспекты реализации информационных систем трех поколений – от проекта первого поколения INTAS IRIS, до инициативы GMES HMA, которая сейчас находится на стадии разработки и тестирования.

1. От виртуальных центров научных данных к информационной инфраструктуре Spatial Data Infrastructure (SDI, NSDI, GSDI)

Широкое признание научного сообщества в последние годы получает концепция «Виртуального управления информацией». Эта парадигма привела к идее построения «Виртуального Научного Центра данных» [1, 2]. В связи с бурным ростом количества наблюдательных данных в астрономии стремительно завоевывают популярность и интенсивно разрабатываются концептуальные основы «Виртуальной Обсерватории» как интегрированной системы распределенных гигантских астрономических архивов и баз данных.

В США с 2000 года разрабатывается проект National Virtual Observatory [3], немного позднее начался подобный проект в Европе [4]. В июне 2002 года создан международный консорциум IVOA (The International Virtual Observatory Alliance) [5].

Основной целью этой Программы является разработка «Виртуального Научного Центра астрономических данных» [6]. Исследуются проблемы управления и распространения астрономическими данными. Астрономические каталоги становятся взаимосвязанными: создается сеть интегрированных, интероперабельных баз данных в объемах, измеряемых в петабайтах [6, 7].

Концепция Виртуальной Обсерватории обеспечивает исследователю возможность использовать астрономические данные от многих различных инструментов, используя передовые ИТ и развивающуюся международную кооперацию научного сообщества. Общий уровень финансирования проектов за последние 3-5 лет приблизился к 20 миллионам долларов США.

Тенденцию лавинообразного увеличения данных можно отметить в области наук о Земле, в геофизике и в метеорологии. Интенсивные исследования глобального потепления и изменения климата приводят, например, к тому, что коллекции данных в наши дни увеличиваются в объеме на величины порядка нескольких петабайт в день.

Наряду с «Виртуальной обсерваторией» развивается программа «Виртуальный Научный Центр данных геофизики плазмы и солнечно-земной физики». Международный консорциум SPASE – Space Physics Archive Search and Extract [8, 9] создан научным сообществом Европы и США. Проект в настоящее время получил поддержку National Science Foundation USA и начал реализовываться; уровень финансирования достигает нескольких миллионов долларов США. Технологические достижения, объединяя распределенные по всему миру гигантские цифровые архивы и базы данных, фундаментально изменяют характер исследований космической плазмы, позволяют решить проблему доступа к данным и обмена информацией в Солнечно-Земной физике.

«Виртуальный Научный Центр» в области изучения планет создается NASA (National Aeronautics and Space Administration) – проект PDS (The Planetary Data System). Образован международный консорциум и сформирована система on-line архивов, состоя-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

щая из нескольких географически распределенных центров, связанных Сетью Интернет [10, 11].

Естественным шагом в развитии концепции виртуальных обсерваторий является развитие геоинформационной инфраструктуры распределенных спутниковых данных. Проект выполняется в Институте космических исследований РАН и известен как Распределенная среда интегрированных информационных ресурсов в области космического мониторинга природной среды (Виртуальный Научный Центр спутниковых данных) [12, 28].

В США стратегия и концепция National Spatial Data Infrastructure (NSDI) начала разрабатываться с 1994 г. [13].

Основными целями программы National Spatial Data Infrastructure являются:

- Организация и проведение национальных дискуссионных форумов.
- Совершенствование механизма доступа к данным путем организации центров информационного обмена и создания баз метаданных.
- Создание баз пространственных данных (БПД).
- Создание тематических данных, критически важных для государства.
- Подготовка образовательных программ для обучения персонала.
- Развитие всестороннего партнерства и координация сбора и использования пространственных данных.

Концепция Spatial Data Infrastructure стремительно завоевала международную популярность. В 2003 г. Европейская Комиссия приняла решение о создании INSPIRE (Infrastructure for Spatial Information in Europe) – Глобальной геоинформационной инфраструктуры данных в Европе [14]. Аналогичный проект создается в Канаде – Canadian Geospatial Data Infrastructure (CGDI).

Основные задачи программы развития геоинформационной инфраструктуры Spatial Data Infrastructure (SDI) можно сформулировать следующим образом:

- Построение глобальной инфраструктуры ИТ и геоданных.
- Гармонизация спутниковой информации; использование согласованного набора стандартов, «понятных» всем участникам данной системы.
- Интероперабельность между независимо созданными приложениями и базами данных; Совместить между собой их интерфейсы, протоколы и форматы данных.
- Поддержка единой политики данных: а) доступ к данным и б) создание и поддержка спутниковой информации – для задач экологии, но открытой и для сельскохозяйственных и транспортных проблем.

Для того, чтобы глобальная геоинформационная инфраструктура стала реальностью в России, возникла необходимость определить те проблемы и особенно-

сти, с которыми сталкиваются национальные научные сообщества исследователей Земли из космоса при поиске оптимального доступа к спутниковым данным. Эти исследования на начальном этапе были выполнены авторами в проекте INTAS IRIS Project: Integration of Russian Satellite Data Information Resources with the Global Network of Earth Observation Information Systems [15, 16]. Проблема интеграции спутниковых архивов является ярким примером актуальности инфраструктурных решений при организации удаленного доступа и распределенной информационной среды. В проекте исследуются новые задачи аэрокосмического дистанционного зондирования, связанные с проблемой виртуального управления информацией в географически удаленных спутниковых цифровых архивах. Важнейшие научные задачи, охватывающие область исследования Земли из космоса, глобальный экологический мониторинг и аэрокосмическое дистанционное зондирование природной среды, актуальны также для процессов интеграции данных солнечно-земной физики, для программ интеграции Цифровых Архивов данных планетарных исследований. Проект IRIS представил концепцию Виртуального Научного Центра: было показано, что условием успешной интеграции служит использование технологий открытых систем и единых стандартов; решение проблемы гармонизации спутниковых данных для создания распределенных ресурсов. Основное внимание при решении задач поиска ресурсов, взаимодействия с ресурсами, извлечения ресурсов, совместного использования ресурсов необходимо уделять проблемам разработки масштабной информационной инфраструктуры интеграции ресурсов, удаленного доступа и управления гетерогенными ресурсами спутниковых архивов.

Главной целью проекта было улучшить доступ к российским спутниковым данным. Проект поддерживает объединение данных регионального спутникового экологического мониторинга. Обеспечивается доступ российских пользователей к европейской информационной системе INFEO (метаданные / on-line каталоги). При запросе в единой точке системы поиск данных осуществляется одновременно в ресурсах распределенных мировых каталогов. Запрос на метаданные оформляется единым образом для всех спутниковых центров, и информация ищется сразу по всем центрам, доступным из глобальных поисковых систем (в первую очередь, это система EOSDIS/ USA и система INFEO). Запрос на получение собственно спутниковых данных также выполняется автоматически и единым образом, а данные приходят в удобном формате для обработки в рамках GIS-систем.

Формирование геоинформационной инфраструктуры спутниковых данных определяется решением следующих задач: использование GIS и Web технологий для разработки новых средств поиска и обнаружения данных аэрокосмического зондирования; включение российских спутниковых каталогов в международную систему INFEO; создание описания российских спутниковых коллекций и регистрации их в IDN (International Directory Network).

Основной задачей проекта является интеграция географически удаленных российских спутниковых архивов. Требуется обеспечить работу пользователей с данными, которые получаются из различных источников и имеют разную структуру хранения. Создается единая информационная система, основанная на концепции интероперабельных распределенных архивов. Необходимо обеспечить требования действующих международных стандартов: ISO/TC 211, FGDC USA (Federal Geographic Data Committee), CEOS (Committee on Earth Observation Satellites).

2. Информационная система первого поколения: проект INTAS IRIS

Международная Информационная система INFEO (Information about Earth Observation) – инициатива Европейского Космического Агентства (ESA), призванная решить проблему доступа к распределенным данным дистанционного зондирования Земли (ДЗЗ). Основопологающим принципом системы является обеспечение интероперабельности – возможности взаимодействия гетерогенных сервисов и каталогов данных в рамках единой информационной системы.

В проекте IRIS были исследованы информационные задачи в области аэрокосмического дистанционного зондирования Земли, связанные с проблемой распределенного доступа и управлением информацией в географически удаленных спутниковых цифровых архивах. Разработка методов и программных средств сбора, тематической обработки, систематизации больших потоков аэрокосмической информации и хранения экстремальных объемов изображений особенно актуальна для космического мониторинга земной поверхности, водной среды, атмосферы, источников антропогенных воздействий на них, опасных природных процессов и объектов техносферы. Основные задачи Активного Архива данных спутниковых исследований, следуя работе [24], представим в такой последовательности: *Публикация* – процесс представления ресурса одним пользователем системы, в результате которого другие пользователи могут

найти этот ресурс и обратиться к нему. *Резюмирование* – процесс извлечения представительного множества характеристических сведений (метаданных) из ресурса или его окружения. *Индексирование* – процесс перевода информации, полученной после резюмирования, в форму, обеспечивающую эффективный распределенный поиск. *Поиск* – процесс обработки «индекса», имеющий целью сформировать ответ на запрос пользователя по распределенным метаданным. *Просмотр* – контролируемая пользователем деятельность, направленная на обследование распределенного информационного пространства, формируемого системой на основе ресурсов и их связей. *Извлечение* – деятельность пользователей, вызванная желанием получить (извлечь из распределенного информационного пространства) обнаруженную информацию. *Взаимодействие* – процесс, ведомый пользователем и находящийся в рамках, ведомых ресурсом, в течение которого ресурс предоставляет пользователю информационный или функциональный сервис. *Распределение* – деятельность системы, имеющая целью обеспечение работы с данными, расположенными на разных физических серверах, различных аппаратно-программных платформах. *Преобразование* – деятельность системы, направленная на перевод метаданных других систем, хранящихся в различных внутренних форматах, во внутренний формат системы для обеспечения интеграции информации и эффективного поиска. *Сотрудничество* – деятельность системы, ориентированная на совместную деятельность с другими информационными системами по обслуживанию поисковых запросов пользователей, в основном заключающихся в преобразовании поисковых запросов для этих систем и интеграции результатов их ответов. *Контроль* – деятельность системы по представлению различных уровней привилегий доступа к информации, её защите. *Управление* – деятельность по сопровождению системы, её ресурсов, соблюдению актуальности, целостности и сохранности информации в распределенном информационном пространстве. Схему функционирования Архива спутниковых данных иллюстрирует рис. 1 [25].

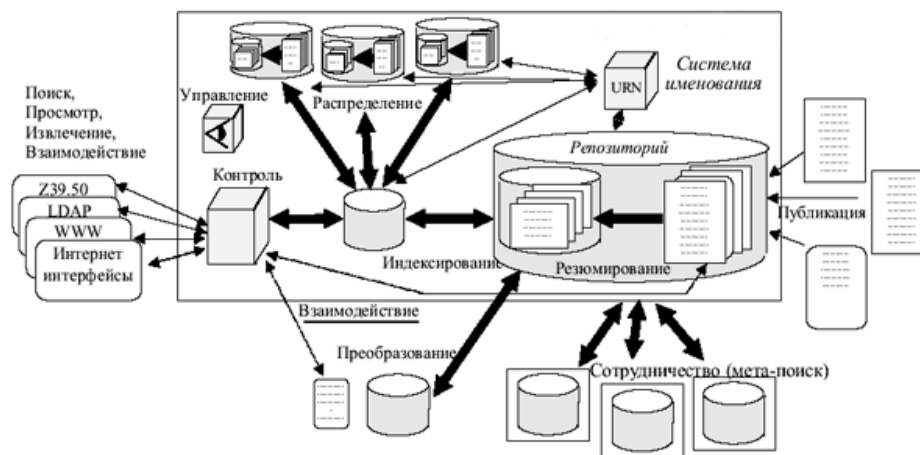


Рис. 1. Структура Активного Архива спутниковых данных

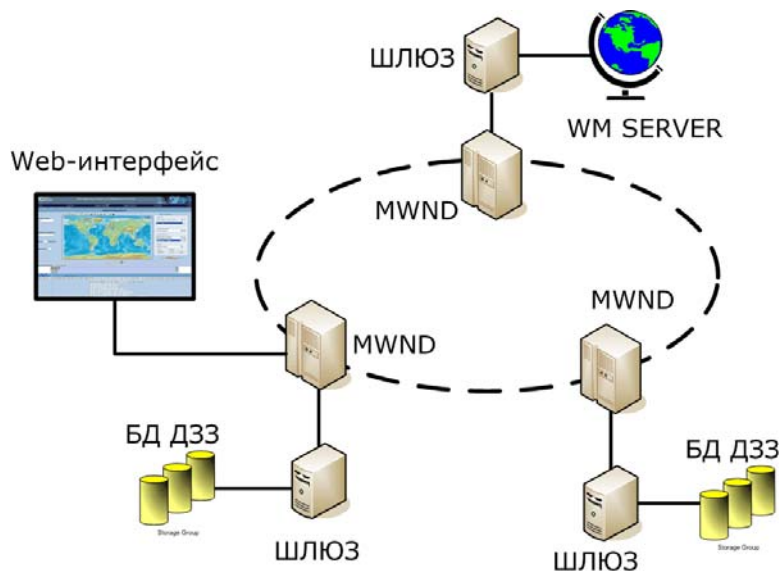


Рис. 2. Структурная схема распределенной системы первого поколения

Основная функция такой системы – создание единой информационной среды, предоставление единой точки доступа для поиска и получения информации по всем архивам и каталогам, зарегистрированным в системе, независимо от их географического расположения и внутреннего формата данных. Распределенная информационная система первого поколения представляет собой сеть узлов, так называемых Middleware Nodes (MWNDs) (рис. 2).

Каждый из узлов содержит метаданные коллекций, описывающие данные, доступ к которым предоставляется через Шлюзы (Gateways). Пользователь для поиска информации обращается к системе через WWW-сервер на одном из узлов MWND, где формируется распределенный запрос, рассылаемый релевантным шлюзам. Результаты выполнения запроса объединяются и возвращаются пользователю.

Шлюз представляет собой промежуточное звено между узлом MWND и каталогом или архивом поставщика данных. Шлюз обменивается сообщениями по протоколу CIP (Catalogue Interoperability Protocol) с узлом MWND, получая от него поисковые запросы и возвращая ему результаты поиска. Запросы передаются транслятору RDBMS, который преобразует их в запросы для непосредственного поиска в базе данных, либо формирует запрос для обращения к предоставляемому сервису (Web Map Server и т.п.).

В основе распределенной информационной системы лежит протокол CIP [17], регламентирующий правила взаимодействия пользователей и каталогов данных ДЗЗ. Для поддержки одновременного доступа пользователя к множеству каталогов используется трехуровневый принцип распределения запросов. Пользователь через web-интерфейс задает поисковый запрос и посылает его к узлу MWND, который, в свою очередь, перенаправляет запрос множеству серверов каталогов данных. Сервера, имеющие данные, удовлетворяющие критериям поиска, возвращают ответы на узел MWND, через

который пользователь получает объединенный результат поиска.

3 Технология интеграции ресурсов и сервисов второго поколения: среда SSE

В проекте INTAS IRIS авторы выполнили исследования инфраструктуры системы международного обмена космической информацией на основе технологий INFEO. Система INFEO, являясь активным многоцелевым архивом космических данных, предоставляет пользователям специализированные интерфейсы доступа к данным, обеспечивает долговременное хранение данных и позволяет производить поиск спутниковых снимков и коллекций данных.

Логическим развитием системы INFEO стала разработка и внедрение технологий интеграции второго поколения. Это было вызвано стремлением расширить участие ESA в глобальном экологическом мониторинге, улучшить в задачах охраны окружающей среды взаимодействие с мировыми информационными системами ДЗЗ [13] и необходимостью поддержать требования Директивы ЕС – INSPIRE [14] о совместимости с другими информационными системами. Развитие технологий интеграции второго поколения направлено на миграцию от подхода, основанного на технологии Z39.50/Gateways, к использованию веб-сервисов (XML над SOAP/HTTP) для связи внешних каталогов с каталогами INFEO. Технологии интеграции второго поколения, получившие название EOLI-XML, разрабатываются ESA в сотрудничестве с национальными космическими Агентствами: Франции (Centre National d'Etudes Spatiales – CNES), Италии (Italian Space Agency – ASI), Германии (German Space Agency – DLR), European Union Satellite Centre (EUCS), Joint Research Centre (JRC) при участии ряда партнеров в промышленности: SPOT Image, Siemens, Spacebel и др. EOLI-XML

представляет собой внешний интерфейс для обмена сообщениями между сервером и клиентом на основе протокола SOAP [18].

Технология EOLI-XML лежит в основе системы SSE (Service Support Environment) [19], разрабатываемой для решения указанных задач с помощью создания открытой, сервис-ориентированной, распределенной среды. Система SSE является инфраструктурой, единой средой для потребителей и провайдеров данных и сервисов. SSE объединяет своих пользователей на основе стандарта XML с использованием протоколов обмена сообщениями SOAP и языка описания веб-сервисов WSDL (Web Service Definition Language) [20]. Таким образом, любой сервис, интегрируемый в среду SSE, должен предоставить SOAP-интерфейс, в соответствии с SSE ICD (Interface Control Document) [21].

Итак, основные цели создания инфраструктуры SSE:

- Предоставить среду, максимально упрощающую взаимодействие типа «сервис-провайдер – пользователь» и «сервис-провайдер – сервис-провайдер».
- Упростить интеграцию уже существующих сервисов, предоставляя универсальный интерфейс на основе XML, позволяющий не изменять их структуры.
- Предоставить пользователям единую точку входа – Интернет-портал системы.

Система SSE состоит из двух основных компонент: SSE Portal Server и AOI Server, образующих вместе Интернет-портал, с которым взаимодействует конечный пользователь.

SSE Portal Server предоставляет web-интерфейс для доступа пользователей к portalу. Сервер построен с использованием промышленной технологии J2EE, которая удовлетворяет требованиям SSE: хорошая масштабируемость, интеграция с существующими информационными системами, гибкая политика безопасности, поддержка стандартных протоколов и языков, переносимость компонент без необходимости перекомпиляции. Выполнение этих требований является обязательным для успешного построения информационной инфраструктуры. В качестве сервера приложений выбран сервер JBoss с открытым исходным кодом.

AOI Server работает в связке с SSE Portal Server и служит для предоставления для сервис-провайдеров функций визуального выделения области на карте (area of interest, AOI) при задании критериев поиска, а также визуализации результатов поиска.

Так как большинство сервисов в качестве одного из входных параметров требуют указание географической области, SSE Portal предоставляет специальную поддержку этой возможности. При обращении к сервису пользователь должен указать географическую область и сделать это он может разными способами: выбрать из списка, указать область на карте, загрузить файл, описывающий AOI, задать AOI с помощью указания координат.

При обращении пользователя к сервису, SSE Portal использует апплет, загружаемый с сервера AOI для визуального отображения выделенной области на карте. SSE позволяет использовать для визуализации как локально хранимую информацию, так и подключаться к OGC WMS серверам для загрузки дополнительных слоев карты [22]. Выделенная таким образом на карте область затем преобразуется в описание в формате GML (Geography Markup Language) [23] описание вместе с запросом отправляется сервис-провайдеру в виде SOAP сообщения.

Типичный сценарий работы выглядит следующим образом. Пользователь через интерфейс SSE портала формирует запрос к нужному сервису, выделив на карте область AOI и указав характеристики искомой информации. После этого сформированный в формате SOAP-сообщения запрос пересылается на URL сервиса, указанный провайдером при регистрации. Здесь сообщение поступает на вход сервера веб-приложений, где в зависимости от поступившего запроса запускается соответствующий сервлет, который занимается дальнейшей обработкой запроса и передачей его на внутренний интерфейс сервиса. Для выполнения сервлетов сервис-провайдеру вместе с Toolbox необходимо установить web-сервер с поддержкой сервлетов (рекомендуется к установке веб-сервер Tomcat).

Обработка входного запроса и формирование ответа на запрос выполняется на основе WSDL-документа в соответствии с XML-схемой сервиса. При установке Toolbox уже имеется предопределенная схема, тем не менее, некоторые сервисы могут иметь специфические параметры входных и выходных сообщений, что можно легко исправить, отредактировав соответствующие файлы XML-схем.

Для поиска необходимой информации пользователь может воспользоваться поисковым интерфейсом по адресу <http://catalogues.eoportal.org/eoli.html>. Интерфейс выполнен в виде Java-апплета (рис. 3).

Существует возможность ограничивать поиск либо по поставщикам данных (панель Find Products), либо по коллекциям данных (Find Collections). В раздел Find Products входят крупные поставщики, имеющие несколько собственных обширных коллекций. Например, ESA (коллекции Radar Imagery, Radar Altimetry, Optical/Multispectral Imagery), DLR (коллекции Atmospheric Sensors, Thematic Maps, Optical Sensors, Digital Elevation Model, Synthetic Aperture Radar Data) и др. После выбора типа данных, по которым будет ограничен поиск, следует выбрать площадь земной поверхности. Для более точного поиска информации можно воспользоваться режимом расширенного поиска Advanced Query Mode. Этот режим позволяет дополнительно задавать ключевые слова для поиска, а также выбирать спутник и сенсор, с помощью которого были получены данные. Ключевые слова служат для отбора данных, поле описания которых содержит строку, заданную в поле Free Text.

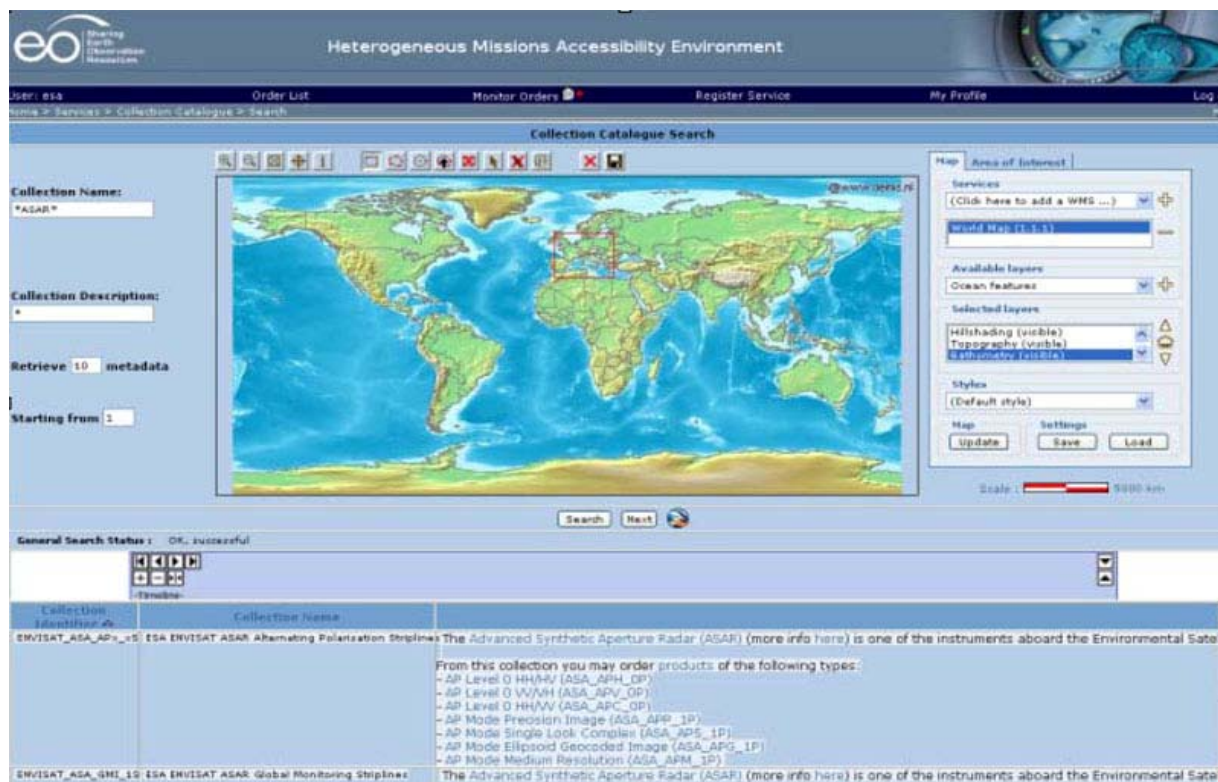


Рис. 3. Поисковый интерфейс пользователя

4 Интеграция неоднородных распределенных информационных ресурсов спутниковых данных: инициатива GMES и проект НМА

Основным предназначением инициативы Global Monitoring for Environment Security (GMES) является поддержка задач Европейского Союза, касающихся устойчивого развития и глобального управления с помощью обеспечения доступа к высококачественным данным дистанционного зондирования Земли в режиме реального времени. Доступ к такой информации имеет стратегическое значение как с точки зрения современного развития отдельных регионов Европы, так и с точки зрения Европейского Союза в целом, как члена мирового сообщества.

Информация, получаемая с помощью проекта GMES, используется в трех основных направлениях деятельности: экология и защита окружающей среды, поддержка развития инфраструктуры ЕС, обеспечение безопасности граждан, как в повседневной жизни, так и в режиме чрезвычайных ситуаций.

Европейское космическое агентство (ESA) – главный партнер европейской комиссии по разработке и внедрению проекта GMES, работающий в тесном сотрудничестве с национальными космическими агентствами Европейских стран, а так же Канады и России.

В настоящий момент доступ к данным дистанционного зондирования Земли каждого спутникового оператора осуществляется через отдельную точку доступа («портал») этого поставщика данных. Поль-

зователь вынужден выполнять поиск нужных ему данных отдельно в каждой точке доступа. Особенно актуальной проблема своевременного получения данных становится в условиях режима чрезвычайных ситуаций, когда счет идет на часы и минуты.

Ключевой особенностью GMES в таком аспекте является гармонизация доступа к гетерогенным данным разных поставщиков данных, включая национальные космические агентства. С этой целью в рамках инициативы GMES европейское космическое агентство инициировало проект Heterogeneous Missions Accessibility (HMA). Это совместный проект с участием большого количества корпораций, поставщиков данных, научных и коммерческих организаций из Европы и Канады.

Проект НМА обеспечит гармонизацию сервисов и данных наземных сегментов поставщиков данных, с помощью сервис-ориентированной архитектуры, предоставляя единообразный доступ к данным и сервисам каждого из участников проекта. В данный момент проект находится на стадии активного внедрения и тестовой эксплуатации. Основные этапы разработки проекта НМА показаны на рис. 4 [26].

Не только базовые спецификации, лежащие в основе проекта НМА, но и большая часть архитектурных решений проекта находятся в свободном доступе, то есть принадлежат области всеобщего достояния (public domain), что значительно увеличивает возможности распространения технологий и программных продуктов проекта. Широко используются стандарты и протоколы ISO (International Organization for Standardization) и OGC (Open

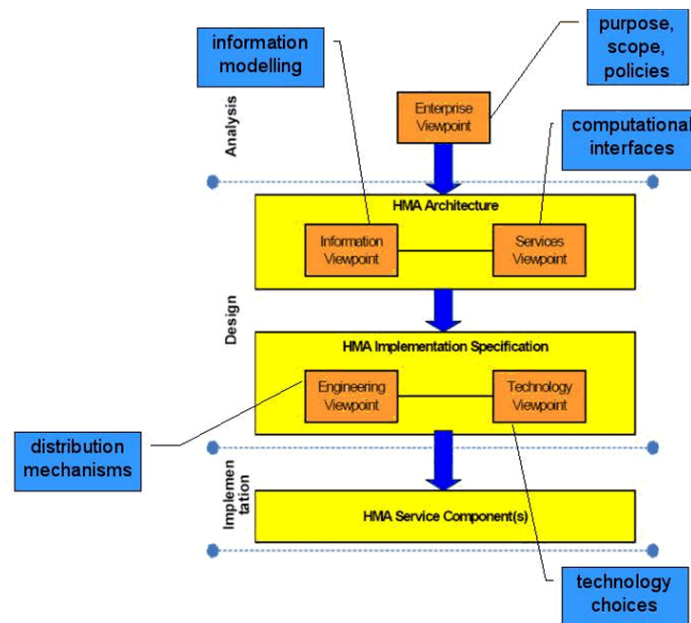


Рис. 4. Разработка проекта Heterogeneous Missions Accessibility (HMA)

Geospatial Consortium). Основополагающим стандартом системы является международный стандарт для построения открытых распределенных систем (RM-ODP ISO/IEC 10746-1:1998).

Модель построения распределенных систем RM-ODP была выбрана в качестве базовой для проекта HMA GMES, так как основные задачи проекта практически совпадают с задачами, для которых разрабатывалась модель RM-ODP:

- Поддержка распределенной обработки.
- Поддержка интероперабельности между гетерогенными системами.
- Скрытие механизмов распределенного взаимодействия от конечных пользователей и системных разработчиков.

Но есть и важное отличие от базовой модели – GMES-HMA будет действовать на основе слабо связанной сети систем и сервисов вместо распределенной вычислительной системы, базирующейся на взаимодействующих объектах.

Определены пять основных направлений развития проекта, определяющих минимальный набор требований и объектную модель, гарантирующую целостность и логическую связность системы:

- Организационное направление: определяет цели, область применения, ограничения и правила использования и развития системы.
- Информационное направление: определяет семантику информационных процессов.
- Вычислительное направление: декомпозиция всей системы на отдельные объекты, взаимодействующие между собой и с внешними интерфейсами с целью оптимизации вычислительных затрат.
- Направление инжиниринга: разработка механизмов и функций, необходимых для под-

держки распределенного взаимодействия между отдельными объектами системы.

- Технологическое направление: определяет выбор конкретных технологических решений, необходимых для выполнения поставленных задач.

Предложенная архитектура является сервис-ориентированной (Service Oriented Architecture, SOA), в которой поставщики данных получают возможность предоставлять доступ к своим сервисам через стандартизованные интерфейсы системы.

Ключевым компонентом системы HMA является интеграционный слой доступа к данным (Data Access and Interoperability Layer – DAIL), предоставляющий единый интерфейс доступа к каталогам и сервисам всех участников проекта GMES с помощью использования установленных стандартов интероперабельного доступа.

В структуре DAIL можно выделить основные компоненты:

- Среда выполнения сервисов – среда для выполнения основных компонентов DAIL, предоставляет средства выполнения сервисов и управления их взаимодействием с другими компонентами DAIL.
- База данных – служит для хранения данных, необходимых для функционирования служебных сервисов HMA.
- Рабочий процесс – основное приложение ядра архитектуры DAIL, служит для запуска и управления сервисами DAIL.
- Репозиторий профилей пользователей – база данных для хранения профилей пользователей системы и информации, необходимой для аутентификации пользователей в системе.
- Сервер коллекций – служит для хранения списка доступных в системе коллекций и отслеживания их статуса.

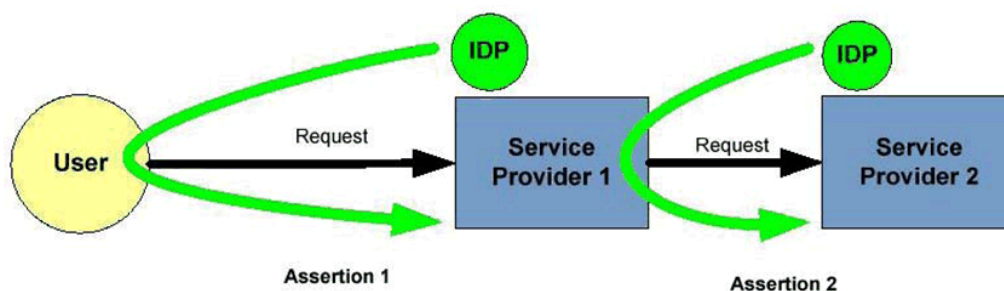


Рис. 5. Стратегия управления доступом к данным и сервисам в НМА

- Сервер сервисов – аналогично серверу коллекций, служит для хранения списка доступных в системе сервисов и отслеживания их статуса.
- Виртуальный FTP сервер – компонента для поддержки обмена данными между поставщиками и пользователями через систему НМА.

В процессе проектирования системы большое внимание было уделено вопросам безопасности, а при разработке учитывались такие стандарты, регламентирующие информационную безопасность, как ISO/IEC17799, ISO27001 и т.п. В качестве базовой для проекта была принята модель управления доступом «Distributor» (рис. 5). В данной модели пользователь авторизуется только у одного поставщика сервисов (SP1), который затем, в случае необходимости воспользоваться сервисами другого поставщика, обращается к нему под своими собственными правами; таким образом пользователю нет необходимости выполнять аутентификацию у каждого из поставщиков данных и сервисов, участвующих в цепочке обработки информации.

5 Заключение

Технологии и принципы построения распределенных систем, рассмотренные в данной статье имеют большой потенциал для дальнейшего развития в будущем и уже сегодня предоставляют широкие возможности для организации распределенных гетерогенных данных и сервисов в единую глобальную систему. Вектор развития каждого следующего поколения распределенных информационных систем состоит из трех основных слагаемых: уход от недостатков существующей системы, улучшение технических и информационных характеристик, решение новых задач, которых просто не существовало на момент разработки системы предыдущего поколения.

Коллектив авторов принимал активное участие в проектах всех трех поколений на этапе как разработки системы, так и практического ее внедрения. Так, на основе проектов INFEO IRIS и INSPIRE (технологии EOLI-XML и SSE) был построен и введен в действие российский сегмент распределенной информационной системы.

Российский сегмент распределенной информационной системы спутниковых архивов был реали-

зован при сотрудничестве ИКИ РАН и ИАПУ ДВО РАН. В качестве поставщика данных дистанционного зондирования Земли выступает Лаборатория спутникового мониторинга ИАПУ Дальневосточного Отделения РАН (<http://satellite.dvo.ru>). Программное обеспечение и серверное оборудование для функционирования шлюза сегмента установлено в Москве, в Институте Космических Исследований РАН (<http://iris.iki.rssi.ru>) [16]. Таким образом, мировое научное сообщество получило возможность доступа к данным одного из российских поставщиков данных ДЗЗ с помощью глобальной распределенной системы.

В заключении нельзя не отметить немаловажный аспект, выходящий за рамки данного обзора. К сожалению, в России до недавнего времени практически отсутствовала поддержка и заинтересованность государства в развитии аналогичных национальных информационных систем. Однако, в последнее время наблюдаются положительные сдвиги в этом направлении. Так, в 2006 году правительством РФ была одобрена и принята к исполнению «Концепция создания и развития инфраструктуры пространственных данных Российской Федерации» [27].

В 2004 году стартовал проект информатизации наук о Земле (геология полезных ископаемых, физика Земли, экология) «Электронная Земля: научные информационные ресурсы и информационно-коммуникационные технологии». В этом проекте исследован подход к решению задачи интеграции геопространственных (ГИС) данных на основе стандартов OpenGIS Consortium и ISO и специализированных методов их обработки с помощью технологий Grid и Интегрированной Системы Информационных Ресурсов (ИСИР). Этот проект является одним из пунктов Программы фундаментальных исследований Президиума РАН «Разработка фундаментальных основ создания научной распределенной информационно-вычислительной среды на основе технологий GRID». Проект имеет своей целью создание на основе технологий IT-инфраструктуры сетевой распределенной информационно-аналитической системы по наукам о Земле, обеспечивающей единообразный доступ к объединяемым в рамках проекта информационно-аналитическим и другим ресурсам. Создаваемая инфраструктура будет предоставлять возможность использования этих ресурсов для решения фундаментальных и прикладных задач, компьютерного моделирования и параллельных вычислений.

Участниками проекта являются свыше двадцати институтов РАН, как геофизической, химической, картографической направленности, так и решающих задачи в сфере распределенных вычислений и телекоммуникаций [29].

Литература

- [1] Кудашев Е.Б., Филонов А.Н. Технология и стандарты интеграции сервисов, каталогов и баз данных дистанционного исследования Земли из космоса // Труды Девятой Всероссийской научной конференции «Электронные Библиотеки»-RCDL'2007, с. 273–279. – Переславль-Залесский: Изд-во «Университет г. Переславля», 2007.
- [2] Antikidis J.P. Virtual Access to Information: an emerging concept // In Proceedings Workshop PV-2004. Ensuring the Long Term Preservation and Adding Value to the Scientific and Technical Data, pages 63–67, 2004.
- [3] Szalay A.S. The National Virtual Observatory // In Astronomical Data Analysis Software and System X. ASP Conf. Ser. Vol. 238, p. 3–12, 2001.
- [4] Virtual Observatory in Europe home page: <http://esavo.esac.esa.int/registry>
- [5] IVOA (International Virtual Observatory Alliance) home page: <http://www.ivoa.net/>
- [6] Arviset C, Dowson J, Hernandez J, Ortiz I., Osuna P., Salgado J, San Miguel G, Venet A. ESA RSSD Science Archives User Interfaces and Inter-Operability Systems. In Proceedings Workshop PV-2004 Ensuring the Long Term Preservation and Adding Value to the Scientific and Technical Data, pages 159–164, 2004.
- [7] Arviset C. ESA Scientific Archives and inter-operable Virtual Observatory Systems // In Proceedings PV-2007 Conference. Ensuring the Long Term Preservation and Adding Value to the Scientific and Technical Data – Munich, DLR, 2007.
- [8] Thieman J. Et al. The SPASE Standard for Heliophysics Data Description, Location, and Acquisition // In Proceedings PV-2007 Conference. Ensuring the Long Term Preservation and Adding Value to the Scientific and Technical Data – DLR, Munich, 2007.
- [9] Harvey C. et al. SPASE –Space Physics Archive Search and Extract – <http://www.igpp.ucla.edu/spase/>
- [10] Hughes S. et al. A Planetary Data system for the 2006 Mars Reconnaissance Orbiter Era // In Proceedings PV-2004 Workshop. Ensuring the long-term Preservation and Value Adding, p. - ESA/ESRIN, Frascati, 2004.
- [11] Hughes S. et al. Defining the Core Archive Data Standards of the International Planetary Data Alliance (IPDA) // In Proc. PV-2007 Conference. Ensuring the Long Term Preservation and Adding Value to the Scientific and Technical Data – DLR, Munich, 2007.
- [12] Kudashev E.B. Digital Library: Improving the accessibility to Russian Satellite Data for mitigating Natural Disaster // In Proc. PV-2004 Workshop. Ensuring the long-term Preservation and Adding Value to the Scientific and Technical data, p. 233–240 – ESA/ESRIN, Frascati, Italy. 2004
- [13] National Spatial Data Infrastructure (NSDI) home page, 2007. <http://www.fgdc.gov/nsdi>
- [14] INSPIRE home page, 2007. <http://www.ec-gis.org/inspire>
- [15] Е.Б. Кудашев, А.Н. Филонов. Организация информационной распределенной среды и интеграция спутниковых архивов // Труды Седьмой Всероссийской научной конференции «Электронные Библиотеки» – RCDL'2005. – Ярославль: Ярославский ун-т, 2005.
- [16] Кудашев Е.Б., Филонов А.Н. Интегрированная распределенная информационная система спутниковых данных в программах исследования Земли из космоса // Труды Восьмой Всероссийской научной конференции «Электронные библиотеки» – RCDL'2006, Суздаль, с. 131–139. – Ярославль: Ярославский ун-т, 2006.
- [17] Catalog Interoperability Protocol (CIP) Specification – Release B-24, 2005. http://wgiss.ceos.org/ics/documents/cip2.4/cipspec-2_4_75_6.pdf
- [18] Earthnet On-line XML Front-End Interface Control Document, EOLI-XML-006-ICD, Issue 1.7, 2003. <http://earth.esa.int/rtd/Documents>
- [19] Service Support Environment Architecture, Model and Standards, 2007. http://earth.esa.int/services/esa_doc/doc_sse.html
- [20] Web Services Description Language (WSDL) 1.1, W3C Note, 2001. <http://www.w3.org/TR/wSDL>
- [21] Interface Control Document (ICD) for the Service Support Environment (SSE), 2007. <http://services.eoportal.org/massRef/documentation/icd.pdf>
- [22] The Open Geospatial Consortium, 2007. <http://www.opengeospatial.org/>
- [23] ISO/TC 211/WG 4/PT 19136 Geographic information – Geography Markup Language (GML), 2004. http://portal.opengeospatial.org/files/?artifact_id=4700
- [24] Бездушный А.Н., Жижченко А.Б., Кулагин М.В., Серебряков В.А. Интегрированная система информационных ресурсов РАН и технология разработки электронных библиотек // Программирование. 2000. Т. 26. № 4. С. 177–185.
- [25] Балашов А.Д., Каленова Н.И., Кудашев Е.Б. Интеграция электронных архивов спутниковых данных // VIII Международная конференция по электронным публикациям EL-Pub2003. <http://www.ict.nsc.ru/ws/elpub2003/rep6254.zip>
- [26] Lawrence B., Pritchard M., Woolf A. Review of the Heterogeneous Mission Accessibility Project // BNSC EO International Co-operation Programme. Year 4 (2006/07): Documents and Presentations. <http://www.pinkbox.space.qinetiq.com/icp2/docs>

- [27] Концепция создания и развития инфраструктуры пространственных данных Российской Федерации // Распоряжение Правительства РФ №1157-п от 21 авг. 2006 г.
- [28] Kudashev E. Spatial Data Infrastructure development of Russian Satellite Data for Global Monitoring and Disaster Management support // 2008 IEEE International Geoscience & Remote Sensing Symposium (IGARSS 2008). July 6-11, 2008 | Boston, Massachusetts, U.S.A. Abstracts of Symposium Presentations. Abstract 1754.
- [29] Вершинин А.В., Бездушный А.Н., Серебряков В.А. «Электронная Земля»: интеграция ГИС-данных с использованием технологий Grid // Труды Восьмой Всероссийской научной конференции «Электронные библиотеки» – RCDL'2006, Суздаль. – Т. 1. Статья 94. – Ярославль : Ярославский ун-т, 2006.

Distributed Infrastructure Development of Satellite Data: Integrated Distributed Environment of Heterogeneous Information Resources for Earth Observation

E.B. Kudashev, A.N. Filonov

Our presentation will focus on distributed infrastructure development for the integration of Russian satellite data information resources with EO Portal/INFEO. Access to geo-data and services are organised via the geo-Portal. Geo-Portal provides an efficient means of distributing geo-data, and for users to search for and obtain information and gain access to data. Distributed Infrastructure of Spatial Data is based on standards, technology and architecture that will lead towards distributed GIS wherein updated data is accessed over the internet directly from data owners in the form of WMS and WFS. Distributed infrastructure for Spatial Information supports the easier access to geo-data and organize better products and services, which in turn will support increased use of geo-information, necessary to acquire process, store, distribute and improve utilisation of geospatial data. Direct integration with other datasets may be organised in an increasingly flexible way. It is intended to extend the current routine collection of digital data and create new information databases, to generate catalogues of Remote Sensing Data and filling in of the Digital Archive, browse images and ordering capability via the web, which are tailored to support environmental monitoring and natural disaster mitigation. Important here is interoperability of data and services. We are facing a transition: from data interoperability to service interoperability. Efficient organization of information resources and open access to spatially distributed experiment data are founded on the Web technology (Access to data including data search and request; Access distributed metadata catalogues, which may be: distributed catalogues themselves based on different application profiles or protocols). Appropriate metadata

management systems are built to provide for the collection and distribution of experiment data and thematic processing results; while the archive is linked to the regional centers of environmental monitoring via Internet. An important element is the elaboration of interface, archiving and network data exchange structures. This calls for the development of search engines and a remote interactive access regime for external users via Internet to catalogues of experiment data and processing results and the realization of the on-line access mode. The principle of distributed data processing gains an ever-growing importance for satellite monitoring. At the same time, remote users get access through Internet to satellite data obtained by the regional monitoring centers. Further development of Remote Sensing satellite data electronic archives implies introduction of automatic user-oriented information selection as a result of enhanced information models of electronic collections coupled to upgraded techniques of user interaction with the archive's information system.

* Работа проводится при поддержке РФФИ (гранты №№ 08-07-00002, 08-07-90401-Укр_а).

Унификация структур данных в области изучения, освоения и использования ресурсов Мирового океана*

© Е.Д. Вязилов, А.А. Федорцов, А.Е. Кобелев

Всероссийский научно-исследовательский институт гидрометеорологической информации
– Мировой центр данных
vjaz@meteo.ru

Аннотация

Представлены методы выделения унифицированных элементов данных при интеграции распределенных гетерогенных информационных ресурсов, применяемые в Единой государственной системе информации об обстановке в Мировом океане. Показаны проблемы идентификации унифицированных элементов. Предложены подходы по созданию многомерных структур с учетом жизненного цикла информации об объектах БД. Приведены примеры многомерных структур данных и реализации БД с элементами многомерности.

1 Введение

В настоящее время создаваемые в разных организациях базы данных (БД) не имеют общей концептуальной основы и по большому счету не совместимы друг с другом. Одновременное использование заложенной в них информации становится трудоемким процессом.

Вопросам стандартизации структур и схем данных в настоящее время уделяется большое внимание, так разработан стандарт ISO 19115, протоколы описания данных – SensorML, TML, MODBUS [11–14] которые упорядочивают представление метаданных и данных на уровне измерительных систем.

Ведутся работы по интеграции данных во многих предметных областях и странах [2, 4, 7–9]. Под эгидой Всемирной метеорологической организации (<ftp://ftp.wmo.int/Documents/wis/WIS-TechnicalSpecification-v1-0.doc>), Межправительственной океанографической комиссии ЮНЕСКО (<http://data.meteo.ru:81/odp/>), в рамках Европейского сообщества ведутся большие работы интеграции данных и созданию баз метаданных. Это проекты Sea Search (<http://www.sea-search.net>), Sea Data Net (<http://www.seadatanet.org>), др. [15]). При этом основное внимание в европейских проектах было уде-

лено созданию схем метаданных для таких объектов, как сведения о массивах и БД, проектах, организациях, рейсах научно-исследовательских судов и созданию общего индекса данных.

В Единой государственной системе информации об обстановке в Мировом океане (ЕСИМО, <http://data.oceaninfo.ru>) реализована технология интеграции [1] разнородных, распределенных информационных ресурсов (ИР, <http://data.meteo.ru:81/srbd/>). Поддержка этой технологии требует приведения имен локальных атрибутов данных к унифицированным элементам данных. Для этого необходимо развитие типовых моделей данных, стандартизации представления и обмена данными, метаданными.

Унифицированный элемент данных – это поименованный объект данных (атомарная порция), находящийся под управлением информационной системы. Унифицированные элементы данных определяют метаданные (сведениями о данных – страна, организация, платформа, пространственно-временные характеристики) и значения атрибутов морской среды и морской деятельности.

Для отображения атрибутов данных и их представления необходимо назначение соответствий имен атрибутов локальной БД их унифицированным элементам. Обмен и использование атрибутов данных в различных приложениях также осуществляется через унифицированные элементы, которые имеют уникальные имена (идентификаторы). Локальные атрибуты данных представлены в массивах и БД, структурированных (форматированных) или объектных файлах.

Имена унифицированных элементов данных отличаются от атрибутов в локальных системах данных при одном и том же содержании и представлении (единицы измерений, формат и др.). Унифицированные элементы хранятся в словаре (<http://data.oceaninfo.ru/udopweb/>), используются для унификации локальных атрибутов данных с различным представлением и имеют уникальные имена и соответствующие им атрибуты описания. С помощью унифицированных элементов можно получить информацию, как пользователям, так и программным средствам (идентификатор, краткое имя, пределы изменений, формат хранения поля).

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

2 Именованние унифицированных элементов данных

При регистрации ИР требуется установить однозначное соответствие между атрибутами массивов и БД, предоставляемых в качестве ресурсов ЕСИМО, и унифицированными элементами данных. Программный комплекс «Поставщик данных» технологии интеграции E2EDM [10] осуществляет преобразование (маппинг) локальных имен атрибутов данных в унифицированные имена элементов данных. Для этого преобразования каждому локальному имени атрибута требуется найти аналог по содержанию и убедиться, что свойства выбранных пар – единицы измерений и/или системы кодирования значений атрибута, пространственно-временное разрешение измерения или обобщения – совпадают.

Критерием выделения унифицированных элементов является возможность интеграции разнородной информации в рамках одной модели описания ИР [5] с минимизацией числа используемых локальных имен атрибутов.

В ЕСИМО для преобразования структурированных данных, хранящихся в БД и структурированных файлах данных, используются разделы словаря «Метаданные» – **М** и «Данные» – **Р**.

Для объектных файлов (карты, графики, документы) используется унифицированный элемент с префиксом **Р**, показывающий отношение к рубрике. Унифицированные элементы данных для объектных файлов используются в том случае, если есть возможность задания элемента на весь объект (например, jpeg-файл с унифицированным элементом «Давление воздуха», представленное в виде изолиний на карте).

Атрибуты метаданных специфицируют характеристики производства (получения, описания) данных. В метаданных определены унифицированные элементы, связанные с описанием ИР. Это идентификаторы структурной единицы данных, платформы наблюдений, инструмента измерений, метода обработки; проекта, географического объекта.

В разделе «Метаданные» выделены подразделы:

- общей идентификации данных;
- принадлежности стране, организации, автору, проекту и др.;
- спецификации производства (получения) данных – платформа, метод, прибор и др.;
- географические характеристики – районы, высота над уровнем моря, глубина, толщина слоя, расстояние, местоположение платформы (широта, долгота), направление движения платформы;
- временные характеристики (год, месяц, день, время начала и окончания события).

В разделе «Данные» выделены подразделы:

- морская среда;
- морская деятельность;
- социально-экономическая информация (СЭИ).

Атрибуты данных, специфицирующие отдельное свойство природного процесса (явления):

- могут быть измеренными значениями, или вычисленными статистическими, или новыми величинами, полученными на основе теоретических или эмпирических закономерностей – каждый атрибут (измеренный или вычисленный), включается в состав унифицированных элементов;

- указывают интенсивность явления – этот показатель является общим унифицированными элементом для многих унифицированных элементов;

- показывают аномалию (отклонение от многолетнего среднего значения), экстремальные значения, тенденцию, характеристику тенденции, количество зарегистрированных явлений (ситуаций) – включается в состав унифицированных элементов.

Значения атрибутов могут визуально наблюдаться, определяться путем химического анализа в пробе, измеряться, регистрироваться в виде порядкового номера. Эта информация указывается в описании метода определения атрибута и представляется в отдельном объекте метаданных [3, 4] – «Методы».

Для указания этих свойств атрибутов используется следующая структура имени унифицированного элемента:

ANNNN_SS_I (M, Q),

где **A** – раздел словаря (**P**, **M** или **R**),

NNNN – числовой код, определяющий атрибут наблюдений (измерений, вычислений);

SS – числовой код, отражающий название статистической характеристики значения атрибута (измеренное значение, минимум, максимум, среднее, повторяемость, вероятность, аномалия, продолжительность явления, число случаев, сумма, тенденция);

I – интенсивность явления, выраженная в значении атрибута;

M – метод получения атрибута;

Q – признак качества значения атрибута.

Примеры унифицированных элементов представлены ниже:

R0295_00 – температура воды, измеренная;

R0295_01_Q – признак качества температуры воды, минимальной.

Для упорядочения и улучшения поиска унифицированных элементов в словаре для поля «Название» применяется следующий принцип записи. В названии атрибута первым записывается основное слово, характеризующее элемент, а затем уточняющее, например, давление воды, измеренное; температура воды максимальная.

3. Проблемы определения унифицированных элементов

3.1 Множество атрибутов, имеющих одинаковое назначение

В распределенных ИР бывают ошибки нормализации данных. В таблицах одного объекта данных или метаданных может быть несколько значений атрибутов для даты, страны, организации, персоны, проекта и других атрибутов. Или, например, одни и

те же атрибуты (погода между сроками, наблюдения в один срок на разных площадках), имеющие несколько значений, хранятся в виде отдельных полей таблиц с фиксированными позициями. Это объясняется тем, что разработчики БД в локальных источниках данных не всегда выходят на необходимый уровень нормализации данных. Проще ввести два или несколько новых атрибутов, характеризующих разные процессы жизненного цикла (ЖЦ) объекта, чем создавать более универсальные структуры данных. Примерами нескольких значений атрибутов в ИР являются:

- страна автора проекта, производитель, владелец судна;
- даты проектирования, постройки, передачи в другую организацию, списания судна;
- даты рассмотрения, принятия, утверждения, подписания, ратификации документа.

Для решения этой задачи атрибуты, имеющие несколько значений, должны записываться вместе с дополнительным атрибутом «ЖЦ». Чтобы определить стадию ЖЦ таких атрибутов, как страна, организация, персона, проект необходимо описать стадии ЖЦ объекта и выделить необходимые метаданные.

Ниже представлены стадии ЖЦ, которые относятся к атрибутам:

- страна – проектирование, производство, владение, администрирование;
- организация – производство, владение, разработка, эксплуатация;
- персона – владение, разработка;
- проект – выполнение, участие.

Структура хранения таких данных в многомерной модели выглядит следующим образом:

Объект: идентификатор (прибор, судно, организация, др.)

Объект: ЖЦ (проектирование, разработка, эксплуатация, др.)

Дата: значение

Атрибут: название (страна, организация, персона, проект, др.)

Атрибут: значение (идентификатор страны, организации, персоны, проекта)

Такая структура применима для хранения значений атрибутов в различных каталогах объектов (судов, товаров и т.п.).

Для ЖЦ документа используются такие стадии как создание, модификация (редактирование), обсуждение (рассмотрение), согласование, принятие, ратификация, утверждение с одной и другой стороны, подписание, вступление в силу документа, передача в печать, издание, доступность, передача на хранение, прекращение срока действия, доставка, уничтожение. Хранение множества дат для документа должно производиться в отдельной таблице:

Объект: идентификатор (документа)

Объект: ЖЦ

Дата: значение

Если требуется уточнить условия применения, указать признак качества даты, метод записи време-

ни, то для этого можно ввести дополнительные атрибуты метаданных.

3.2 Большое количество однородных атрибутов в локальных БД

Иногда одному атрибуту дают разные названия и хранят в одной или разных таблицах, например, для оценки производства различных товаров в СЭИ представлено более ста названий атрибутов. Для работы с такими данными необходимо выделить атрибут «Класс». Классом может быть товар, промышленный объект, объект вылова, отходы, данные, экономические показатели и др. Чтобы правильно описать объект, необходимо выделить стадии ЖЦ производства и определить состояние объекта. Примеры стадий ЖЦ для различных объектов приведены ниже.

Товар (продукция): проектирование, создание (производство), ввод в действие – пуск, сдача, заказ, поставка, продажа, хранение, монтаж, постановка на учет, обслуживание, вылов, сдача, продажа, списание, уничтожение (утилизация).

Состояние транспортного объекта: отправление, переход, заход, прибытие, стоянка у причала, на рейде, ремонт, дрейф, производство наблюдений, погрузка, разгрузка, списание.

Выбросы (отходы): создание, хранение, обезвреживание, утилизация.

Проект: инициация; планирование; выполнение; контроль и мониторинг; завершение.

Данные: агрегация – расчет статистических характеристик, прогноз, классификация, результаты сравнения – аномалии.

Еще одним вариантом использования множественных имен атрибута, который встречается в БД, является тип географического объекта. Например, географический объект можно представить в виде географический названий материков, стран, федеральных округов, субъектов РФ, населенных пунктов, морей и океанов, др. Для этого атрибута необходимо хранить в таблице БД два атрибута – роль географического объекта (страна, район, др.) и значение географической области (Индийский океан).

4 Модель информационного пространства

Основной идеей типизации структур данных является использование минимальной атомарной единицы хранения. Сейчас в большинстве ИР атомарной единицей является группа атрибутов, измеренных или вычисленных для одной точки в определенный момент времени или за какой-то период. Если за атомарную единицу хранения принять отдельное значение атрибута, то многие атрибуты метаданных могут быть вынесены в отдельную таблицу «Каталог объекта», а изменяющиеся значения в пространстве и во времени – в таблицу «Факты».

Модель информационного пространства (ИП), объединяющего разнообразную информацию, представлена на рис. 1.

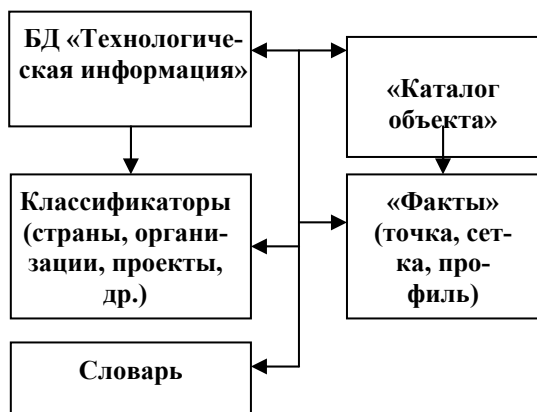


Рис. 1. Модель информационного пространства

Состав основных компонент такой модели включает:

- каталог объекта;
- таблица фактов;
- словарь унифицированных элементов;
- классификаторы;
- технологическая информация, определяющая состояние и связи объектов.

Каталоги включают справочные сведения о различных объектах по классам (метаданные, товары, отходы, документы, др.). Так как каталоги для конкретных объектов очень сильно отличаются по составу атрибутов, то для хранения каталогов эффективнее применять многомерную модель хранения данных. Таблица «Каталог объектов» имеет следующую структуру данных:

Идентификатор класса объекта (каталога)

Идентификатор экземпляра объекта

Имя атрибута (унифицированный элемент каталога)

Значение атрибута (каталога).

Факты отражают пространственно-временные координаты объектов, находящихся на различных стадиях ЖЦ. Данные могут быть представлены в виде точки, сетки, профиля. Эти типы данных имеют свою структуру данных [6]. Более подробно об этих структурах сказано в разделе «Примеры реализации». Таблица фактов должна включать такие атрибуты как класс объекта, ЖЦ, дата и время регистрации значения атрибута, имя атрибута, его значение. Кроме того, в зависимости от ситуации могут использоваться дополнительные атрибуты – страна, организация, персона, проект (если значения этих атрибутов изменяются в различных событиях).

Сведения об унифицированных элементах включают имя, временной и пространственный масштабы, полное и краткое наименование атрибута на русском и английском языках, точность наблюдений или расчетов, единицы измерения, диапазон изменчивости (мин, макс), вертикальное разрешение (мин и макс высота/глубина), используемый классификатор для значений атрибутов (международный, национальный), метод определения атрибута, международный код атрибута, описание.

Технологическая информация включает дату ввода, все даты редактирования, кто вводил и редактировал, показатели использования данных, другую информацию и предназначена для мониторинга состояния таблиц «Каталог объектов» и «Факты». Эта информация может также храниться в двух таблицах «Каталог» и «Факты».

Многомерные структуры данных для таблиц «Каталог объектов» и «Факты» представлены ниже.

Таблица «Каталог объектов»

Объект: Класс

Объект: тип

Объект: процесс

Страна: ID

Организация: ID

Персона: ID

Дата: значение

Атрибут: ID

Атрибут: значение

Таблица «Факты» для класса «Отходы»

Объект: Класс

Отходы: Тип

Отходы: Процесс

Страна: ID

Организация: ID

Персона: ID

Проект: ID

Геообъект: Тип

Геообъект: Значение

Дата: Значение

Атрибут: ID

Атрибут: Значение

Отходы: Класс опасности

Для каждого объекта составляется своя пара таблиц «Каталог объекта» и «Факты». То есть физически это будут разные таблицы, а по структуре они не будут отличаться.

5 Структура словарной базы данных

Создание и ведение ИП требует управления семантическим обеспечением (классификаторами, словарями). Для этого необходимо иметь в словарной базе (рис. 2):

- список объектов, включенных в ИП;
- список технологий, которые используют словарную базу (технология интеграции E2EDM, централизованная база метаанных, web-портал, другие компоненты);
- список задач, для которых необходимо использование словарной базы (ввод, визуализация, конвертирование данных);
- список программных средств, в которых используются классификаторы;
- список объектов метаанных, для которых необходима словарная база;
- список таблиц, где используются классификаторы;
- список необходимых классификаторов для каждой технологии;

- сведения о классификаторах (кодовых таблицах);
- кодовые таблицы;
- сведения об атрибутах (унифицированных элементах).

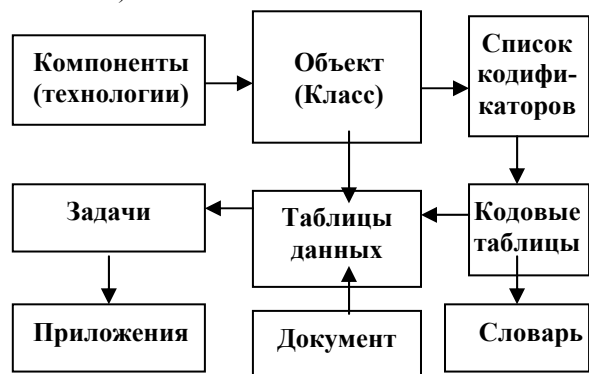


Рис. 2. Структура словарной БД

Число используемых классификаторов в ЕСИМО составляет около 300 единиц. Если для каждого классификатора создавать свою таблицу, то число таблиц резко увеличится, поэтому используется структура хранения классификаторов, включающая идентификатор классификатора, код объекта и значение кода. Для повышения эффективности и удобства использования классификаторов для каждой компоненты создаются рабочие словари. Состав классификаторов и кодов в них назначается в зависимости от потребностей технологий.

6 Примеры реализации

В рамках разработки интегрированной БД ЕСИМО созданы четыре типа хранения структур данных (данные, измеренные на профиле в случайных точках, данные в виде временных рядов, сеточные данные и каталоги объектных файлов). Эти структуры активно используют различные классификаторы.

Данные о временных рядах оформляются в виде двух таблиц – сведения о временном ряде и матрицы значений. Сведения о временных рядах включают идентификатор ряда, название станции, широту, долготу, дату начала и окончания наблюдений, горизонт, масштаб временного и пространственного осреднения, кто создал, когда, дату последней редакции, пополнения данных, название исходного массива. Матрица временного ряда значений включает идентификатор ряда, год, месяц, день, время, значение.

Сеточные данные хранятся в виде трех таблиц: сведения о сетке, сведения о поле, значения атрибута в узлах сетки. Сведения о сетке включают уникальный код, тип сетки, ссылку на источник данных, идентификатор модели, размерность сетки и шаг сетки в град по X, Y, широту и долготу точки, временной масштаб осреднения (сутки, месяц, год, другой), пространственный масштаб осреднения (квадрат, трапеция, точка), шаг при отрисовке изолиний.

Сведения о поле включают идентификатор поля, ссылку на сетку, значение года, месяца, дня, часа, минут наблюдений, имя атрибута, тип уровня, значение уровня. Данные в узлах сетки включают идентификатор точки, широту, долготу, значение атрибута.

Данные измерений по профилю состоят из двух таблиц: сведения о профиле и данные по профилю. Сведения о профиле включают идентификатор профиля и пространственно-временные координаты профиля. Данные по профилю включают идентификатор профиля, уровень, имя атрибута и значение.

Сведения об объектных файлах хранятся в виде каталога, имеющего ссылки на файлы с документами, картами и рисунками.

7 Заключение

Унификация процессов управления данными в гетерогенной среде, выработка наборов правил, применение типового инструментария и тем самым снижение расходов на администрирование данных возможна только через рассмотрение стадий ЖЦ объектов, сведения о которых отражаются в БД.

Необходимо создавать каталоги для организации хранения и поиска объектных файлов (документов, графических файлов, др.), а также для любой информации, представленной в виде сведений о каких-либо объектах.

В сущностях, для которых создаются БД, необходимо выделять общие поля типа идентификатор, дата, время создания, редактирования, хранения, название георайона, другие, в т.ч. поля для организации связи между объектами. При работе с атрибутами необходимо создавать обобщенные форматы полей (целое число, число с плавающей запятой, строка).

Предложенные структуры данных могут быть использованы в других предметных областях, при этом не надо добавлять новые сущности и поля.

Литература

- [1] Белов С.В., Бритков В.Б. Интеграция информационных ресурсов в задачах исследования морской среды // Информационные технологии и вычислительные системы. 2008, Вып. 1. С. 73–82.
- [2] Вавилов К., Щербина С. Web – Интеграция // Издательство «Открытые системы». Электронный журнал «Открытые системы». 2001. Вып. 1. 12 с. <http://www.osp.ru/os/2001/01/043.htm>
- [3] Вязилов Е.Д. Консолидация метаданных в области наук об окружающей среде // Журнал «Вычислительные технологии» Т. 10, Спецвыпуск. СВ-Томск, 2005. С. 30–38.
- [4] Вязилов Е.Д., Михайлов Н.Н., Карпенко Г.А., Кобелев А.Е. Широкий комплекс метаданных, как основа мониторинга и управления данными // 6-я Российская научно-техническая конференция «Современное состояние и проблемы

- навигации и океанографии («НО-2007»), 23–25 мая 2007, С-Петербург, ГНИНГИ. С. 493–496.
- [5] Вязилов Е.Д. О стандартизации структур данных в области морской среды // Электронный журнал «Новости ЕСИМО». 2007. Вып. 30. <http://meteo.ru/resource/magazine/news30.mht>
- [6] Вязилов Е.Д., Михайлов Н.Н. Интеграция гетерогенных информационных ресурсов в области морской деятельности // Журнал «Вычислительные технологии» Т. 10, Спецвыпуск. СВ-Томск, 2005. С. 21–29.
- [7] Ершова Г., Сафонов В. Интеграция в книжном мире. 21.10.2003. Журнал «Директор ИС», № 10. 2003 // Издательство «Открытые системы». <http://www.osp.ru/cio/2003/10/080.htm>
- [8] Журавлева О.В., Лисовский К.Ю., Томусяк Г.С., Томусяк Э.С. Функциональные методы обработки слабоструктурированных данных и их применение для построения электронных библиотек // Сборник трудов Третьей Всероссийской конференции по электронным библиотекам. КНЦ РАН, Петрозаводск, сентябрь 2001. <http://www.pair.com/lisovsky/misc/biblio/index.html>
- [9] Игнатович Н Брокер интеграции приложений // Журнал «Открытые системы», 2003. № 09. Издательство «Открытые системы». <http://www.osp.ru/os/2003/09/008.htm>
- [10] Belov S., Mikhaylov N., Vyazilov E. A model of the distributed marine information resources – approaches and decisions // ICES-IOC Study Group on the Development of Marine Data Exchange Systems using XML, Gothenburg, Sweden 26–27 May 2003. С. 79–100. <http://www.ices.dk/reports/OCC/2003/SGXML03.pdf>
- [11] Botts Mike. OpenGIS® Sensor Model Language (SensorML), Implementation Specification, Version 1.0 // Open Geospatial Consortium Inc., 2006. – 118 с.
- [12] Havens Steve. OpenGIS® Transducer Markup Language, Implementation Specification, Version 1.0.0 // Open Geospatial Consortium Inc., 2006. 258 pp. <http://xml.coverpages.org/OGC-06-010r6-TransducerMarkupLanguage-TML.pdf>
- [13] ISO 19115. – 2003. 146 pp.
- [14] MODBUS application protocol specification. Version 1.1b // Modbus-IDA, December 2006, http://www.modbus.org/docs/Modbus_Application_Protocol_V1_1b.pdf.
- [15] Ocean Biogeographic Information System // The Census of Marine Life. 2003. <http://www.iobis.org/>.

Unification of data structure for field research, exploration and resources using of World Ocean

E. Vyazilov, A. Fedortsov, A. Kobelev

The coding methods of system elements are presented in report for integration of distributed heterogenic information resources in Unified state system of information for World Ocean. The selecting problems of system element are showed. The approaches for development multi dimensional structures with using of life cycles of information about data base objects are proposed. The examples of multi dimensional data structures and creating of such data base are cited.

* Работа выполнена в рамках проекта РФФИ № 07-01-00662-а.

Библиотека программ JINRLIB*

© Л.В. Попкова,

А.П. Сапожников,

Т.Ф. Сапожникова,

Р.Н. Федорова

Объединённый институт ядерных исследований
tsap@jinr.ru

Аннотация

В статье содержится описание библиотеки компьютерных программ JINRLIB. В зависимости от способа сопровождения и распространения библиотека делится на две части: одна часть распространяется в виде объектных модулей, другая - в виде автономных пакетов прикладных программ. Описываются принципы организации и технология сопровождения программ и программных пакетов библиотеки.

БИБЛИОТЕКА ПРОГРАММ – это организованная на машинном носителе совокупность компьютерных программ, доступ к которым осуществляется по их именам или индексам. Библиотеки программ могут состоять из объектных модулей (двоичных файлов, полученных в результате трансляции), или текстов программ на исходном языке программирования [1].

JINRLIB – библиотека программ, предназначенная для решения широкого круга математических и физических проблем, возникающих в ходе научной деятельности сотрудников ОИЯИ [2].

1 Цели проекта:

1. обеспечить долговременную и стабильную поддержку пользователей компьютеров всех типов, существующих в ОИЯИ;
2. обеспечить сохранение результатов труда старшего поколения специалистов по численным методам;
3. создать у сотрудников ОИЯИ стимул для разработки собственных программ.

2 Состав библиотеки

Основу библиотеки составляют программы из библиотеки ДУБНА [3,4] – популярной в прошлом библиотеки программ ОИЯИ, и MATHLIB и KERNLIB – ядра программного обеспечения CERN [5]. Не все программы из этих библиотек сейчас

представляют интерес. Некоторые входят в состав библиотек компилятора, другие не используются в силу своей специфической области применения, потеряли актуальность или заменены более новыми. При создании новой библиотеки по мере возможности мы учитывали мнение пользователей, в частности, какими программами из библиотеки ДУБНА желательно заняться в первую очередь. Пополнение библиотеки происходит также новыми программами, создаваемыми сотрудниками ОИЯИ и их коллораборантами.

Каждая программа идентифицируется уникальным индексом или именем. Для индексирования программ, включаемых в библиотеки объектных модулей, используется привычная нашим пользователям классификация, принятая в свое время в библиотеке ДУБНА. Для идентификации остальных программ используется ее имя и классификация, принятая в издательском отделе ОИЯИ. Программы объединяются в библиотеки объектных модулей или существуют в виде самостоятельных пакетов прикладных программ. Поэтому в зависимости от способа сопровождения и распространения библиотека естественным образом делится на две части: одна часть распространяется в виде объектных модулей, другая — в виде автономных пакетов прикладных программ.

3 Принципы организации и технология сопровождения библиотек объектных модулей

1. Программа, включаемая в библиотеку объектных модулей, предоставляемую пользователям, это совокупность 3-х компонентов:

- Исходный текст. Это файл, содержащий тексты всех подпрограмм библиотечной программы, пригодных для компиляции на компьютерах различных типов. Текст программы должен быть написан на языках Фортран и/или С.
- Описание программы (по возможности – на русском и английском).
- Тестирующая программа. Сюда же относятся файлы, содержащие исходные данные для теста, если они требуются, и результаты эталонного прогона.

2. Все программы работают только с 64-разрядным представлением чисел с плавающей за-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

пятой. Использование в программе конкретных особенностей машинного представления чисел допускается только в исключительных случаях. Для Фортрана: каждая подпрограмма должна содержать строку

IMPLICIT REAL*8 (A-H, O-Z)

Для языков C и C++ должны использоваться типы Double и Long.

3. Имена файлов, содержащих компоненты библиотечной программы, получаются из ее библиотечного индекса.

Расширения имен: .for .f .c .C .txt .dat .res

Каждая программа располагается в отдельной директории, имя которой совпадает с библиотечным индексом. Например, программа A101 хранится в директории A101 в виде 6 файлов:

A101.for - содержит фортранный текст программы A101;
T_A101.for - файл с тестом программы A101;
A101.dat - набор входных данных для теста;
A101.res - результат прогона теста;
A101.txt - описание программы A101 на русском языке;
A101E.txt - описание программы A101 на английском языке.

Предлагаемая организация материалов библиотеки пригодна для использования современных технологий сопровождения больших систем (например, CVS [6]).

4. Программа включается в библиотеку только после того, как будет успешно откомпилирована, а по возможности и оттестирована на всех компьютерных платформах ОИЯИ.

5. Форма описания программ:

Название программы. Индекс
Автор(ы)
Язык
Назначение
Структура:
 Тип: Подпрограмма,
 подпрограмма-функция
 Имена входа для пользователя
 Используемые внешние
 подпрограммы
Обращение
Метод. Точность. Ограничения.
Ошибки исполнения
Литература
Пример

6. Головная директория библиотеки называется JINRLIB. Здесь размещаются:

- программы для полной пересборки библиотеки из ее компонентов;
- файл библиотеки, получаемый в процессе сборки;
- поддиректории A101-Z999 компонентов библиотеки;

- временные файлы, создаваемые в ходе сборки библиотеки.

7. Программы библиотеки проверяются на компьютерных платформах Unix (Linux) и Windows.

Программа создания (сборки) библиотеки индивидуальна для каждой платформы и для каждого компилятора. Как правило, это текстовый файл, содержащий установку переменных для создания нужного окружения, команды вызова компилятора и библиотекаря – командный файл для ОС Windows или shell-script для ОС Unix.

В работе с библиотеками программ объектных модулей можно выделить два этапа. Сначала с каждой программой идет «индивидуальная» работа – проверяются исходные тексты, проводится полное тестирование программы, анализируются результаты тестов, описание программы приводится к стандартному виду. После выполнения всех этих операций считается, что программа готова для включения ее в библиотеку объектных модулей – все необходимые файлы программы заносятся в директорию, совпадающую с ее библиотечным индексом, объектный модуль программы добавляется к соответствующей библиотеке. Программа готова для использования.

Второй этап есть собственно сборка библиотек объектных модулей из исходных текстов программ, входящих в состав библиотеки. Этот этап выполняется для обновления библиотек объектных модулей, а также при изменениях в аппаратном и/или программном обеспечении Центрального информационно - вычислительного комплекса (ЦИВК) ОИЯИ [7]. Учитывая большое количество программ библиотеки, наличие нескольких поддерживаемых в ОИЯИ компьютерных платформ и несколько систем программирования на каждой, основные усилия были направлены на то, чтобы сборка библиотек происходила максимально просто, «от одной кнопки». Для этого для каждой системы программирования были подготовлены соответствующие командные файлы (ОС Windows) или shell-скрипты (Unix). В качестве примера можно привести один из скриптов сборки библиотеки для ОС Linux:

```
#!/bin/sh
dir=`ls`
for name in $dir
do
#####
# Special compilation mode: z008
#####
if test -d $name
then
case $name
in
##### Fortran + C:
z008)
cd $name
g77 -c $name.f
```

```

gcc -c "$name".c.c
ar rv ../libjinr.a $name.o "$name".c.o
ranlib ../libjinr.a
rm *.o
cd ..
;;
##### Work directories:
copy|obsolete|tests|tmp)
;;
##### In general:
*)
cd $name
g77 -c $name.f
ar rv ../libjinr.a $name.o
ranlib ../libjinr.a
rm $name.o
cd ..
;;
esac
fi
done

```

4 Реализация

При изменениях в аппаратном и/или программном обеспечении ЦИВК ОИЯИ, по мере появления новых программ или модернизации старых готовится новая версия библиотек объектных модулей JINRLIB. Библиотеки готовятся на компьютерах с операционной системой (ОС) Linux для работы на ЦИВК ОИЯИ, а также на компьютерах с ОС Windows для всех доступных нам фортранных трансляторов. В настоящий момент на основных машинах ЦИВК работает ОС Scientific Linux версии 4 (SL4), установленная в 64-битном варианте, на некоторых машинах с 32-битной архитектурой стоит ОС Scientific Linux 3 (SL3).

Для вычислительной фермы ЦИВК ОИЯИ библиотеки объектных модулей математических программ общего назначения подготовлены для ОС Scientific Linux 4 с архитектурой CPU x86_64 для трансляторов:

- GNU Fortran 77 compiler (g77) – libjinr.a;
- GNU Fortran 95 compiler (gfortran) – libjinr95.a;
- Intel Fortran Compiler (ifort) – libjinri.a.

Также поддерживается и пополняется библиотека для ОС Scientific Linux 3 с архитектурой CPU i386 для GNU Fortran 77 compiler (g77) – libjinr.a.

Библиотеки размещаются в директории /usr/libexec/CERNLIB/jinr/.

Библиотеки объектных модулей для ОС Windows 9X/NT/2000/XP подготовлены для трансляторов GNU Fortran 77, Compaq Visual Fortran 6.6, Fortran PowerStation 4.0, Microsoft Fortran 5.00.

Для использования на компьютерах с ОС Windows 9X/NT/2000/XP библиотеку JINRLIB нужно записать в одну из директорий и указать путь к этой директории при вызове компилятора. Если

библиотека записана, например, в директории d:\fortran\lib, тогда ссылка на библиотеку с использованием переменной окружения PLACE=d:\fortran\lib будет выглядеть так: %PLACE%\jinr\lib

В таблице 1 приведены примеры вызова компиляторов с использованием библиотек объектных модулей JINRLIB.

5 Технология сопровождения автономных пакетов прикладных программ

Программы, которые по разным причинам не могут распространяться в виде библиотек объектных модулей (программа представляет собой независимый пакет, не является мобильной, недоступен исходный текст и т.п.), при желании автора (владельца) также могут регистрироваться в JINRLIB под уникальным именем со ссылкой на первоисточник. Вся информация, предоставленная автором (владельцем) программы, помещается на WWW-сервере и является доступной. По желанию автора это может быть аннотация программы, полное описание или руководство, исполняемый модуль или исходные тексты программы. Дополнительная информация может быть получена только у автора. Для предоставления таких программ используется форма, приведенная ниже:

1. Автор (владелец), e-mail
2. Имя программы
3. Область применения
4. Описание возможностей программы, способы представления результатов, алгоритмы, математические методы, физические модели и т.д.
5. Ссылки на публикации, руководства, инструкции, где описана программа
6. Компьютерная платформа, на которой работает программа. Среда программирования.
7. Дополнительные библиотеки и пакеты программ, необходимые для работы программы.

Информация, размещаемая в этом разделе, регулярно обновляется и пополняется. Обновляются разрабатываемые сотрудниками ОИЯИ пакеты прикладных программ, появляются новые программы. Для идентификации таких программ используется классификация, принятая в издательском отделе ОИЯИ. В настоящий момент насчитывается около 40 таких программных продуктов из разделов:

- Теоретическая физика высоких энергий
- Теоретическая физика низких энергий
- Физика тяжелых ионов
- Автоматизация обработки экспериментальных данных
- Вычислительная математика и техника

Таблица 1. Примеры вызова компиляторов с использованием библиотек объектных модулей JINRLIB

Scientific Linux g77 (lxpub01-lxpub07)	Вызов компилятора: g77 myprog.f -L/usr/libexec/CERNLIB/jinr -ljinr или: g77 myprog.f -L\$CERN/jinr -ljinr Просмотр содержимого библиотеки: ar t /usr/libexec/CERNLIB/jinr/libjinr.a
Scientific Linux gfortran (lxpub01-lxpub03)	Вызов компилятора: gfortran myprog.f -L/usr/libexec/CERNLIB/jinr -ljinr95
Scientific Linux ifort (lxpub01-lxpub02)	Вызов компилятора: ifort myprog.f -L/usr/libexec/CERNLIB/jinr -ljinri
Windows 9X/NT/2000/XP GNU Fortran 77	Вызов компилятора в командной строке: g77 myprog.for %PLACE%\libjinr.a Просмотр содержимого библиотеки: ar t %PLACE%\libjinr.a
Windows 9X/NT/2000/XP Compaq Visual Fortran 6.6	Вызов компилятора в командной строке: fl32 myprog.for %PLACE%\jinrlib.lib В оконном режиме: в Project/Settings/Link в Object/Library modules добавить jinrlib.lib, указав полный путь. Просмотр содержимого библиотеки: lib /list %PLACE%\jinrlib.lib
Windows 9X/NT/2000/XP Fortran PowerStation 4.0	Вызов компилятора в командной строке: fl32 /MD myprog.for %PLACE%\jinrlib.lib В оконном режиме: в Build/Settings.../Link в Object/library modules добавить jinrlib.lib, указав полный путь. Просмотр содержимого библиотеки - так же.
Windows 9X/NT/2000/XP Microsoft Fortran 5.00	Вызов компилятора в командной строке: fl myprog.for %PLACE%\jinrlib.lib Просмотр содержимого библиотеки: lib %PLACE%\jinrlib.lib , con или: lib %PLACE%\jinrlib.lib , listlib.txt

Решение не ограничивать библиотеку JINRLIB только программами, написанными на языках Фортран и С, позволило существенно пополнить библиотеку новыми программами, разработанными сотрудниками ОИЯИ и максимально отражающими широкий спектр научных задач института.

6 Поддержка программирования в среде MPI

В последнее десятилетие происходит бурное развитие технологий программирования параллельных вычислений, в частности, MPI [8]. Эта тенденция нашла свое отражение и в библиотеке JINRLIB.

Классическим примером больших вычислительных программ является MINUIT [9], программа минимизации функций многих переменных, созданная в начале 70-х годов прошлого века Ф.Джеймсом (ЦЕРН), и чрезвычайно популярная до сих пор. Популярность этой программы побудила выбрать именно ее как объект распараллеливания.

Была сформулирована следующая стратегия распараллеливания: библиотечная программа, подготовленная для работы в среде MPI, должна успешно работать при любом количестве NP параллельных процессов, вовлекаемых в решение прикладной задачи, в частности и при NP=1. Таким образом, возникает единый исходный текст библиотечной программы, равно пригодный к эксплуатации как на традиционных последовательных вычислительных системах, так и на современных кластерах, состоящих из большого числа процессоров. Необходимым условием для обеспечения такой унификации является наличие заранее подготовленных программ-заглушек (stubs) пакета MPI, что позволяет использовать библиотечные программы даже на тех машинах, где никакого MPI нет вообще. К счастью, базовые операции MPI сконструированы таким образом, что в качестве заглушек достаточно иметь пустые подпрограммы, лишь бы системный загрузчик смог реализовать все внешние ссылки. Эта идея была успешно опробована на упомянутой выше программе MINUIT, которая при работе на кластерах использует «настоящий»

MPI, а при работе в однопроцессорной конфигурации – его заглушки [10, 11].

7 WWW-сопровождение библиотеки

Специализированный WWW-сайт <http://www.jinr.ru/programs/jinrlib> обеспечивает электронный доступ к библиотеке JINRLIB, где можно найти каталог, тексты и описания программ и программных пакетов, библиотеки объектных модулей. Каждый программный пакет находится в отдельной директории со своим именем, что позволяет легко найти нужную программу и удобно сослаться на нужный документ. Ведется каталог вновь поступивших программ и программных пакетов. На сайте также размещены «Положение о библиотеке JINRLIB», классификация математических и прикладных программ, форма для предоставления программ в библиотеку, правила работы с библиотеками объектных модулей (примеры вызова транслятора, параметры компиляции, заказ нужной библиотеки и т.п.). Готовые библиотеки объектных модулей для конкретных компиляторов находятся там же.

Для некоторых программных пакетов на странице с их описанием по просьбе авторов заведены счетчики обращений, в определенной мере отражающие степень интереса к программе. Этого количества счетчиков явно недостаточно для получения полной картины. Но даже из этого небольшого количества информации можно сделать вывод, что наибольшую популярность у пользователей сайта имеют математические пакеты классической тематики, а также программные пакеты, ориентированные на решение конкретной прикладной задачи и имеющие свой круг пользователей.

Примером может служить математический пакет GaussDLL – комплекс программ матрично-векторных операций и решения систем линейных алгебраических уравнений (автор И.Н.Киян) [12], который появился в JINRLIB только в середине июня, а через месяц на его счетчике обращений уже было значение 90. Примером популярных программ другого типа являются регулярно обновляемые пакеты FITTER – программа для фитирования экспериментальных данных [13], и SAS – программа для первичной обработки спектров малоуглового рассеяния [14] (автор А.Г. Соловьев), к которым было более 1500 обращений за время их существования в библиотеке.

Возможно, следует завести такого рода счетчики на все программные пакеты и библиотеки программ для получения более подробной статистики и определения своеобразного рейтинга программ.

8 Заключение

JINRLIB находится в ряду таких библиотек, как CPCLIB [15], CERNLIB [5], библиотека численного анализа НИВЦ МГУ [16], но традиционно отличается от двух последних шириной охвата потенциальных предметных областей, что определяется широ-

той круга авторов программ, представленных в библиотеке.

Планы развития:

- освоение новых систем программирования, в первую очередь – свободно распространяемых;
- улучшение качества статистики использования программ библиотеки;
- разработка для некоторых программ динамических WEB-приложений;
- развитие WEB-сайта библиотеки JINRLIB: обратная связь с пользователем, получение отчетов об ошибках при использовании программ библиотеки.

Литература

- [1] Сапожников А.П., Сапожникова Т.Ф. Библиотека программ // Большая рос. энцикл. – Т. 3. – М. : Изд-во «Большая рос. энцикл.», 2005. – С. 460–461.
- [2] Библиотека программ JINRLIB: <http://www.jinr.ru/programs/jinrlib>
- [3] Федорова Р.Н., Широкова А.И. Библиотека программ на фортране. – P11-90-465. – Дубна : ОИЯИ, 1990.
- [4] Мазный Г.Л. Программирование на БЭСМ-6 в системе «ДУБНА». – М., 1978.
- [5] Библиотека программ CERNLIB. <http://cern.ch/cernlib/>
- [6] CVS – Concurrent Versions System: <http://www.nongnu.org/cvs/>
- [7] <http://lit.jinr.ru/>, раздел «Центральный информационно - вычислительный комплекс ОИЯИ».
- [8] MPI: The complete Reference. MIT Press, Cambridge, Massachusetts, 1997.
- [9] CERN Program Library Long Writeup D506. F.James. MINUIT. CERN, Geneva, Switzerland.
- [10] Сапожников А.П. Опыт распараллеливания больших вычислительных программ. Параллельная версия программы MINUIT. – P11-2003-216. – Дубна : ОИЯИ, 2003.
- [11] Сапожников А.П. Параллельная версия программы MINUIT – минимизации функций многих переменных. – <http://www.jinr.ru/programs/jinrlib/minuit/>
- [12] Киян И.Н. Комплекс программ матрично-векторных операций и решения СЛАУ с вещественными коэффициентами. – <http://www.jinr.ru/programs/jinrlib/gaussdll/>
- [13] Соловьев А.Г. Программа для фитирования экспериментальных данных. – <http://www.jinr.ru/programs/jinrlib/fitter/>
- [14] Соловьев А.Г. Программа для первичной обработки спектров малоуглового рассеяния. – <http://www.jinr.ru/programs/jinrlib/sas/>
- [15] Библиотека программ CPC. – <http://www.cpc.cs.qub.ac.uk/>
- [16] Библиотека численного анализа НИВЦ МГУ. – http://www.sccc.msu.su/num_anal/lib_na/libnal.htm

Program Library JINRLIB

L.V. Popkova, A.P. Sapozhnikov, T.F. Sapozhnikova,
R.N. Fedorova

The article is devoted to JINRLIB – library of computer programs, created in JINR. The foundation of this library is a kernel of very popular in the past DUBNA. Depending on the way of maintenance and expansion this library is divided onto two parts: the first part expands as a collection of object modules, second part lives as a set of standalone packages in their source form. The principles of organization, the technology of maintenance for the whole library are described. The examples for compiling and linking with usage JINRLIB are given for different computer platforms. The specialized WWW-site, where the library resides, is mentioned. This site provides electronic access to JINRLIB for all needed computer platforms.

* Авторы выражают благодарность В.П. Ширикову за настоятельные рекомендации не ограничивать библиотеку JINRLIB только программами, написанными на языках Фортран и С, что позволило существенно пополнить библиотеку новыми программами, а также А.Г. Соловьеву за плодотворные обсуждения по организации структуры сайта.

Научные и научно-организационные электронные ресурсы Объединенного института ядерных исследований

© В.Ф. Борисовский
vborisov@jinr.ru
Н.А. Ленская
lenskaya@jinr.ru
И.А. Филозова
Irina.Filozova@jinr.ru

В.В. Кореньков
korenkov@cv.jinr.ru
Ж.Ж. Мусульманбеков
genis@jinr.ru

С.В. Куняев
kouniaev@jinr.ru
Э.Г. Никонов
e.nikonov@jinr.ru

Объединенный институт ядерных исследований

Аннотация

Статья посвящена задачам поддержки электронных информационных ресурсов Объединенного института ядерных исследований. Рассматриваются вопросы обеспечения унифицированного и прозрачного доступа к разнообразным источникам данных и интеграции используемых приложений.

1 Введение

Информационные технологии сегодня широко используются в организации и поддержке фундаментальных и прикладных исследований в предметных областях науки. Их роль в этом процессе достаточно активна и определяется насущными задачами. Роль электронных ресурсов заметно усиливается. Они становятся равноправной частью традиционных информационных ресурсов, на которую распространяется опыт справочно-библиографического описания, классификации и поиска. Эти процессы характерны для большинства научных организаций, имеющих развитые вычислительные ресурсы и опыт работы в области информатизации отдельных областей знаний и располагающих большими информационными массивами, накопленными в процессе научной деятельности.

Электронные ресурсы (ЭР) с одной стороны, расширяют область использования традиционного фонда научно-технической информации (книги, журналы, препринты, информационные подборки и бюллетени), а с другой дополняют ее по содержанию и качеству.

Электронные каталоги стали основным средством библиографического доступа к фондам и кол-

лекциям библиотек. Развитие ЭР создает условия для оперативного доступа к различной информации (как локальной, так и удаленной). Доступ к этим информационным ресурсам осуществляется посредством специализированных поисковых систем, Web-интерфейсов или CD-DVD дисков.

В этой связи наблюдается бурное развитие информационных систем, направленных на хранение и обработку электронных данных различной природы. Также важна роль систем, обеспечивающих информационную поддержку фундаментальных и прикладных исследований в различных областях человеческих знаний.

2 Структура научных электронных ресурсов (НЭР) ОИЯИ

Объединенный институт ядерных исследований (ОИЯИ) является межправительственной международной организацией, ведущей исследования в области ядерной физики и смежных областях. С момента основания ОИЯИ был накоплен огромный объем разнородной информации, хранящейся как в структурированном, так и в неструктурированном или слабоструктурированном виде [1]. Эффективность использования этой электронной информации предполагает интеграцию и обеспечение прозрачного доступа к разнообразным информационным ресурсам ОИЯИ и к мировым электронным библиотечно-информационным ресурсам, в том числе, физических научных центров, университетов России, стран участниц ОИЯИ, научно-технических библиотек [2]. Актуальность задачи определяется тенденцией информационного хаоса в глобальной сети, когда доступ к информационным ресурсам становится затруднительным, а иногда невозможным. Особенно остро эта проблема проявляется в процессе поиска и обработки неструктурированных данных.

Источники научных электронных ресурсов института достаточно разнообразны и специфичны:

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

О Проблемно-тематический план научно-исследовательских работ и международного сотрудничества Объединенного института ядерных исследований за 1999–2009 гг. (ПТП НИР и МС);
О информационные ресурсы научно-технической библиотеки ОИЯИ:

- еженедельные экспресс-бюллетени "КНИГИ", "ПРЕПРИНТЫ", "СТАТЬИ", "ЖУРНАЛЫ", "ДИССЕРТАЦИИ И АВТОРЕФЕРАТЫ",

- архивы бюллетеней с 1998 по 2007 гг.,
- базы данных библиографических описаний с 1988 по 2007 гг.

- база данных ИНИС (Международная система по ядерной информации, используемая МАГАТЭ);

О научные издания Издательского отдела ОИЯИ:

- журнал "Физика Элементарных частиц и атомного ядра" (ЭЧАЯ) / Physics of Elementary Particles and Atomic Nuclei (PEPAN),

- журнал "Письма в ЭЧАЯ" / PHYSICS of PARTICLES and NUCLEI, LETTERS,

- бюллетень "Новости ОИЯИ" / "News of the Joint Institute for Nuclear Research,

- препринты,

- ежегодные отчеты о научной деятельности ОИЯИ,

- книги,

- труды научных конференций, совещаний, симпозиумов.

О научно-методические издания, регламентирующие научную и административно-организационную деятельность института:

- решения, протоколы сессий Комитетов полномочных представителей (КПП);

- протоколы сессий и резолюции Ученых советов (УС), Программно-консультативных комитетов (ПКК) и финансовых комитетов (ФК);

О научно-публицистические издания: газета "Дубна: наука, сотрудничество, прогресс";

О персональные электронные ресурсы (ПЭИ)

О архивные материалы (документы, фото, видео и т.д.);

О прочие издания (доклады, презентации, рекламные материалы).

На рис. 1 приведена функциональная схема структуры информационных ресурсов в контексте сетевой инфраструктуры.

2.1 Электронные ресурсы научно-технической библиотеки (НТБ) ОИЯИ

Одним из крупнейших поставщиков электронных информационных ресурсов ОИЯИ является научно-техническая библиотека ОИЯИ (НТБ ОИЯИ). Стандартной схемой доступа к библиографическим записям являются электронные каталоги. Исторически можно выделить два этапа в формировании электронных ресурсов НТБ ОИЯИ. Первый этап – 1988–1990 гг. – начало автоматизации библиотечного процесса на основе реляционной СУБД DBASE III.

На этом этапе были сформированы базы данных, содержащие к данному моменту порядка 650 тыс. записей. Эта база данных активно используется читателями и сегодня. Второй этап – 1999–2000 гг. – переход НТБ к использованию единого программного пакета управления библиотекой, автоматизирующего все библиотечные циклы – комплектование, обработку, работу с читателями, статистику, обмен библиографическими записями и т.д. (АИБС *LiberMedia*, Франция).

С момента перехода на АИБС *LiberMedia* электронные каталоги и базы данных создаются и поддерживаются на ее основе. В частности, перечень работ включает:

- ✓ текущий и ретроспективный ввод библиотечного фонда;

- ✓ формирование еженедельных информационных бюллетеней новых поступлений;

- ✓ формирование ежегодных Библиографических указателей работ сотрудников ОИЯИ;

- ✓ формирование ежеквартальных списков материалов конференций, симпозиумов, съездов, опубликованных в журналах, полученных НТБ ОИЯИ;

- ✓ формирование библиографических списков по запросам читателей.

На 01.01.2008 количество библиографических описаний в АИБС "LiberMedia" составило 89638, а общее количество экземпляров – 99403. С 2007 года база данных Liber пополняется библиографическими записями статей из журналов, поступающих в НТБ, со ссылками на полный текст (URL).

2.2 Отражение электронных ресурсов НТБ в сети

Web-сайт Научно-технической библиотеки функционирует с 1998 года. Он постоянно обновляется и развивается (<http://lib.jinr.ru>), отражая все направления деятельности НТБ. Далее приводится краткий обзор разделов, пользующихся наибольшей популярностью у пользователей.

Списки новых поступлений

В разделе "**Новые поступления**" еженедельно пополняется коллекция списков новых поступлений НТБ "КНИГИ", "ПРЕПРИНТЫ", "СТАТЬИ", "ЖУРНАЛЫ", "АВТОРЕФЕРАТЫ И ДИССЕРТАЦИИ", "КОНФЕРЕНЦИИ". Это электронные версии традиционных еженедельных информационных бюллетеней НТБ. С 2002 года их выпуск на бумажных носителях не производится. Поддерживаются архивы списков с 1998 года (см. <http://lib.jinr.ru/archru.htm> и <http://lib.jinr.ru/arch.htm>).

Доступ к электронным каталогам

Электронные каталоги доступны с главной страницы сайта НТБ в разделе "**Электронный каталог**" или по ссылке <http://lib.jinr.ru/cat.htm>. Они в совокупности содержат библиографические описания публикаций, поступивших в фонд НТБ ОИЯИ:

- практически весь книжный фонд
- публикации сотрудников ОИЯИ с 1988 г.

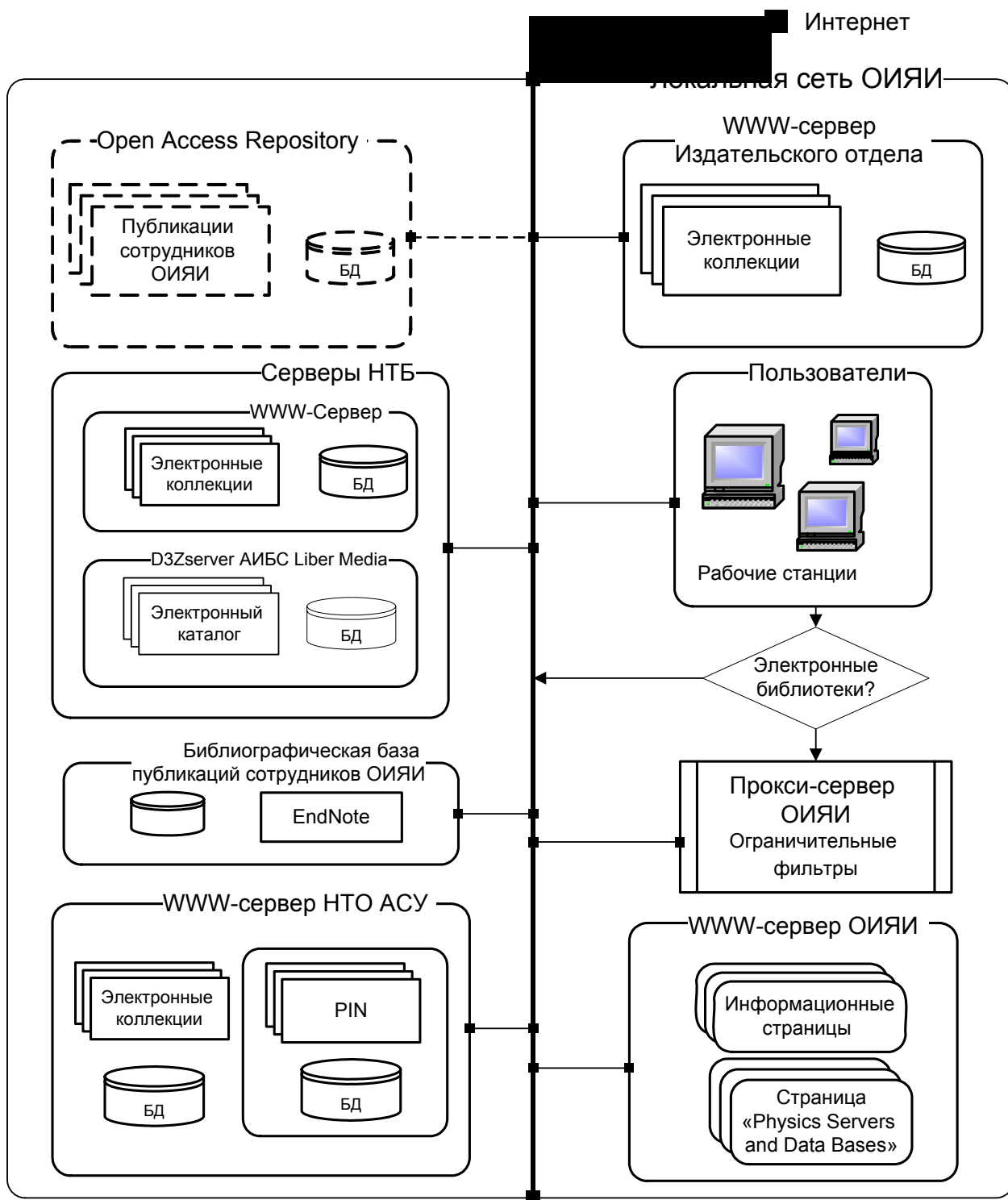


Рис. 1. Функциональная схема электронных ресурсов ОИЯИ

- журналы с 1999 г.
- все диссертации
- авторефераты диссертаций с 2002 г.
- препринты с 1990 г.
- материалы конференций, опубликованные в журналах с 2001 г.
- журнальные статьи с 1992 г.

Пользователям предоставляется возможность удобного информационного поиска, в обоих каталогах (Liber и DBASE). Имеется справочная система, содержащая рекомендации для пользователей. Для осуществления наиболее точного поиска реализован Рубрикатор, позволяющий автоматически вводить в поисковые поля значения тематических индексов. Доступ к электронному каталогу Liber через ИНТЕРНЕТ реализован с помощью шлюза WWW -

Z39.50 и программного модуля Open Public Access Catalog (OPAC). Поиск возможен по следующим полям: Заглавие, Издательство, Автор, Серия, Рубрика. Для более сложного поиска можно применять логические операторы И, ИЛИ, И-НЕ. В поле "Рубрика" можно указать слова или фразу из тематических индексов *Тезауруса* (<http://lib.jinr.ru/opac/libopac/opac.html>).

С 2005 г. можно заказывать литературу в режиме on-line через модуль OPAC. Читатели со своих рабочих мест, не посещая библиотеку, могут воспользоваться услугой *on-line заказ*, чтобы заказать имеющуюся в фондах литературу через электронный каталог (<http://lib.jinr.ru/opac/elcat.php>). Пользоваться этой услугой имеют право сотрудники ОИЯИ, зарегистрированные в библиотеке. Для заказа литературы в электронном каталоге необходимо кликнуть на ссылку *заказ*, отметить отдел, куда направляется заказ, и ввести запрашиваемые системой данные: номер читательского билета и имя/отчество. Заказ будет выполнен при наличии свободного экземпляра в течение одного дня. При отсутствии литературы на полке на момент заказа, читатель ставится на очередь. Одновременно можно заказать не более 5 единиц. В 2007 г. общее количество обращений к электронным каталогам составило около 18 тысяч.

Организация авторизованного доступа к электронным публикациям

Эффективные научные исследования немыслимы без информационного обеспечения. Это предполагает широкий доступ специалистов к научной информации и, прежде всего, к ведущим российским и зарубежным научным журналам и базам данных, что помогает поддержанию высокого творческого потенциала ученых, ознакомлению с новыми идеями и тенденциями в мировой науке, развитию взаимовыгодного международного научного сотрудничества.

Через раздел *"Электронные журналы"* читатели могут выйти на сайт "Publishing Offices over the World", где собраны ссылки на полнотекстовые электронные версии журналов, книг, энциклопедий и справочников. Здесь также представлены ссылки на поисковые системы и сервера, сайты крупных издательств, научных центров, институтов.

В НТБ ежегодно поступает около 170 названий периодических изданий. Благодаря подписке (которая стоит ежегодно немалых денег) сотрудники ОИЯИ имеют доступ к полнотекстовым электронным версиям большинства журналов по тематике ОИЯИ. Традиционно НТБ является постоянным подписчиком издательств AIP&APS, Elsevier, Springer, Institute of Physics Publishing на книги, журналы и энциклопедии по физике, математике, химии и информационным технологиям. В настоящее время ОИЯИ имеет доступ более чем к 3000 электронным книгам издательства Springer. Другие издания, не охватываемые подпиской, но востребованные отдельными читателями, стали доступными благодаря активному участию НТБ в различных библиотечных объединениях и консорциумах. Так,

участие в консорциуме НЭИКОН позволяет сотрудникам ОИЯИ читать полные тексты блока научно-популярных журналов NATURE, SCIENCE, журналов ACS (American Chemical Society), IEEE (Institute of Electronic and Electrical Engineers). Благодаря сотрудничеству с Математическим институтом им. Стеклова ОИЯИ является членом объединения институтов физико-математической направленности, имеющего доступ к информационно-поисковой системе "MathSciNet" Американского математического общества (AMS). Членство в консорциуме библиотек, поддерживаемого Роснаукой, дает возможность сотрудникам ОИЯИ пользоваться библиографической базой данных "Web of Knowledge" Института научной информации (ISI). Активно используется читателями "Научная электронная библиотека РФФИ", которая предоставляет доступ к полным текстам более 3000 названий журналов. На бонус, полученный от фирмы "EBSCO Information Services", продлена подписка на электронную энциклопедию "Physical Science and Technology"; на электронные версии архивных номеров журналов "Nuclear Physics A&B" на 2008 г. Общее количество обращений к электронным версиям журналов через "Электронную библиотеку РФФИ" и через сайты зарубежных издательств составило 90 тысяч. При загрузке полных текстов пользователи обязаны неукоснительно выполнять условия Лицензионных соглашений, подписываемых ОИЯИ с одной стороны и издательством – с другой. Для исключения случаев нарушения этих условий выход пользователей на сайты издательств осуществляется через прокси-сервер. При этом каждый пользователь должен зарегистрироваться и получить авторизацию. На прокси-сервере установлены фильтры, которые выявляют недопустимые по интенсивности загрузки соединения и автоматически блокируют сеанс связи и учетную запись нарушителя. С правилами загрузки статей из электронных журналов пользователи знакомятся на страничке "Registration".

Полезные сервисы

Раздел *"Сервисы"* содержит:

- ✓ Электронные версии Библиографических указателей работ сотрудников ОИЯИ с 2002 года, сформированные по тематическим разделам. Указатели содержат ссылки на полные тексты публикаций и именные указатели (author indexes).
- ✓ Статистику о распределении публикаций сотрудников ОИЯИ по различным журналам.
- ✓ On-line заказ литературы через модуль OPAC.
- ✓ Подписку на еженедельную рассылку Бюллетеней новых поступлений в НТБ.

Желающие получать электронные версии еженедельных информационных бюллетеней НТБ по электронной почте должны заполнить предлагаемую форму. На данный момент электронные версии информационных бюллетеней рассылаются примерно по 100 адресам электронной почты. Рассылка может производиться в двух форматах: html и doc. По мере

необходимости организуется внутренняя рассылка сообщений для извещения подписавшихся пользователей о планируемых выставках, семинарах, презентациях и прочих мероприятиях.

2.3 Электронные ресурсы Издательского отдела ОИЯИ

В конце 90-х годов в Издательском отделе ОИЯИ были начаты работы по созданию электронных коллекций публикаций ОИЯИ. Начиная с 1998 г. все выпуски журнала ЭЧАЯ сохранялись в электронном виде и размещались на сервере издательского отдела. В 2000 г. аналогичная коллекция начала создаваться для журнала "Письма в ЭЧАЯ". Параллельно была организована работа по созданию архива журналов. Для этого были отсканированы все выпуски журнала ЭЧАЯ за период с 1971 по 1997 гг., проведена трудоемкая работа по набору всех аннотаций (около 1200 аннотаций на русском и столько же на английском языке), созданию удобной для пользователя единой структуры электронного варианта журнала. Электронные варианты журналов доступны для всех желающих на сервере отдела. Доступ к электронным ресурсам организован на сайте Издательского отдела ОИЯИ (<http://www1.jinr.ru>).

✓ ЭЧАЯ

Обзорный журнал **"Физика элементарных частиц и атомного ядра" (ЭЧАЯ)** издается Объединенным институтом ядерных исследований с 1970 г. Архив журнала 1970 – 2007 гг. В публикуемых обзорах отражаются актуальные проблемы теоретической и экспериментальной физики элементарных частиц и атомного ядра, физики конденсированных сред, рассматриваются вопросы создания новых ускорителей и экспериментальных установок, автоматизации обработки экспериментальных данных, экологии и радиобиологии. В журнале печатаются обзоры как на русском, так и на английском языках.

✓ Письма в ЭЧАЯ

«Письма в журнал "Физика элементарных частиц и атомного ядра"», кратко – **«Письма в ЭЧАЯ»**, издаются с 1984 г. (до 2000 г. – «Краткие сообщения ОИЯИ»). В журнале публикуются статьи, содержащие результаты оригинальных теоретических, экспериментальных, научно-технических, методических и прикладных исследований. Содержание публикуемых статей определяется тематикой научных исследований в ОИЯИ. Архив журнала 1984–2007 гг.

✓ Новости ОИЯИ

Информационный бюллетень **"Новости ОИЯИ"** издается с 1988 г. "Новости ОИЯИ" выходят 4 раза в год и сообщают о новых научных результатах, полученных в лабораториях ОИЯИ, о ходе крупных экспериментов, создании новых установок, открытиях, изобретениях и т. д. Широко представлена хроника ОИЯИ. Информационный бюллетень "Новости ОИЯИ" рассылается как внутри Института, так и в страны-участницы ОИЯИ, а также в другие научные центры, сотрудничающие с Дубной. Архив 1998–2006 гг.

✓ Годовые отчеты ОИЯИ (Архив 1998–2007 гг.)

✓ Препринты и сообщения ОИЯИ,

а также авторефераты диссертаций, начиная с 1999 г. Архив 1999–2007 гг. Начиная с 1999 г. публикации сканируются и в соответствии с достигнутой с информационным отделом ЦЕРН договоренностью по согласованному протоколу размещаются одновременно в двух базах данных: издательского отдела и ЦЕРН. Это сделано для более оперативного информирования зарубежных ученых с последними научными результатами, полученными в ОИЯИ. Начиная с 2004 г. на этих серверах размещается полнотекстовая база данных.

✓ Книги (Монографии, юбилейные издания).

✓ Научные труды (Материалы конференций, симпозиумов, совещаний).

2.4 Персональные электронные ресурсы (ПЭИ)

Персональные электронные ресурсы посвящены выдающимся ученым, деятельность которых связана с ОИЯИ (<http://www.info.jinr.ru/Scientists.htm>):

Д. И. Блохинцев, Н. Н. Боголюбов, Бруно Понтекоров, В.П. Джелепов, Г.Н. Флеров, М.Г. Мещеряков, И.М. Франк, В.И. Векслер, А.М. Балдин, В.П. Саранцев, Н.Н. Говорун, Ф.Л. Шапиро.

Структура ПЭИ:

- Биография
- Научная деятельность
- Библиография научных работ
- Фото/видео-архив

3 Научные электронные ресурсы и оценка научной деятельности

В последнее время становится актуальной проблема оценки эффективности деятельности научных организаций и их сотрудников на основе показателей результативности научной деятельности (ПРНД). Один из аспектов этой оценки – публикации научных сотрудников. С начала 2008 года в ОИЯИ введена в эксплуатацию разработанная в отделе АСУ ОИЯИ информационная система "PIN: Персональная информация", предназначенная для распределенного сбора и анализа информации о результатах научной деятельности сотрудников. Программное обеспечение создано с целью выполнения Постановления Правительства РФ от 22 апреля 2006 г. № 236 «О реализации в 2006–2008 гг. пилотного проекта совершенствования системы оплаты труда научных работников и руководителей научных учреждений и научных работников научных центров Российской академии наук» и приказа от 23 мая 2007 г. № 144/352/33 Минобрнауки России, Минздравсоцразвития и РАН «Об утверждении Положения о порядке аттестации научных работников организаций, подведомственных Российской академии наук».

Информационная система PIN основана на Java Servlet технологии, позволяющей создавать сложные интернет-приложения. Все компоненты, использованные при разработке ПИН (Linux, Java, Tomcat,

СУБД MySQL), являются бесплатными. Без каких-либо изменений ПИН может работать и на платформе Microsoft Windows.

Основные функции системы PIN:

- ✓ Ввод данных о результатах научной деятельности сотрудников включая публикации, научно-организационную работу, педагогическую и инновационную деятельность в соответствии с приказами РАН и Минобрнауки;
- ✓ Верификация этой информации уполномоченными лицами (администраторами подразделений / отделов);

✓ Расчет формальных показателей результативности научной деятельности (ПРНД), подготовка заявок для аттестации и для начисления стимулирующих надбавок в соответствии с приказом Минобрнауки;

✓ Ввод научными сотрудниками дополнительной информации о своей работе;

✓ Визуализация информации с учетом прав доступа.

Интерфейс пользователя системы PIN показан на рис. 2.

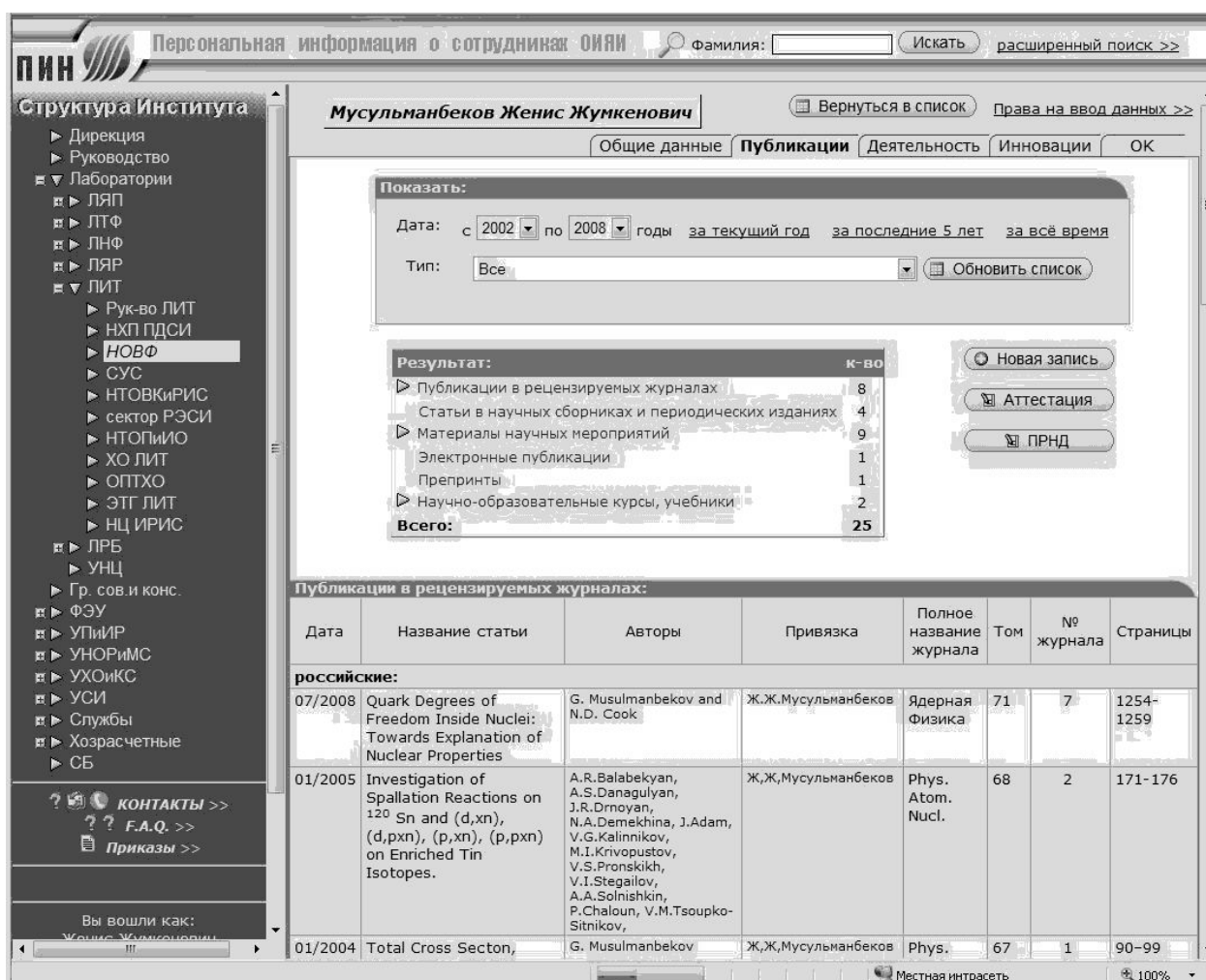


Рис. 2. Интерфейс пользователя в системе PIN

Одним из основных показателей результативности научной деятельности являются публикации сотрудников. Для учета публикаций научных сотрудников в ОИЯИ создается база библиографических данных, пополняемых в первую очередь самими авторами. С недавнего времени в ОИЯИ эксплуатируется коммерческий многофункциональный пакет EndNote, позволяющий создавать персональные библиографические базы данных, вести поиск библиографической информации в удаленных базах данных [3]. Помимо других средств пакет EndNote планируется использовать для более полного учета публикаций научных сотрудников ОИЯИ. Являясь

поисковой системой, EndNote позволяет быстро осуществлять поиск в электронных библиографических базах в Интернете, загружать ссылки во вновь созданную или существующую библиотеку EndNote, присоединять к библиографическим ссылкам полнотекстовые статьи, рисунки и таблицы. Поиск обеспечивается посредством протокола запросов "Z39.50". В EndNote имеется предустановленный список библиографических баз, по которым можно вести поиск и скачивать библиографические документы. В этот список входят всемирно известная база Web of Science Института Научной Информации (ISI), библиотеки различных университетов, а

также электронные каталоги крупных российских библиотек. Наиболее полной по охвату работ сотрудников ОИЯИ являются библиографическая база INIS и библио- и фактографическая база SPIRES. Однако, формат документов этих баз данных не позволяет производить поиск и загрузку документов в среде EndNote. Для решения таких проблем в пакете EndNote предусмотрена возможность импортирования нестандартных документов с помощью различных фильтров и шаблонов. Библиографическая информация скачивается из этих баз в текстовом виде, затем этот текстовый файл пропускается через созданный в соответствующем формате фильтр, и преобразованный таким образом файл загружается уже в локальную библиографическую базу. Библиографические записи из этой базы могут быть импортированы в соответствующие разделы системы "PIN: Персональная информация".

Таким образом, приложение PIN и пакет EndNote позволяют осуществлять пополнение и создание максимально полных списков библиографий научных сотрудников ОИЯИ, что способствует повышению качества сопровождаемых электронных коллекций.

В настоящее время начаты работы по созданию электронного полнотекстового архива (Open Access Repository) публикаций сотрудников ОИЯИ с открытым доступом в рамках мировой системы Open Access Information (OAI) [4]. Актуальность этих работ обусловлена ежегодным ростом цен на электронные подписные издания. Создание собственного сервера открытого архива и включение его в мировую систему OAI позволит иметь единый полнотекстовый архив публикаций сотрудников ОИЯИ, повысить доступность работ, увеличить их цитируемость, облегчить процедуру депонирования работ, сократить средства, расходуемые на традиционную подписку.

Литература

- [1] Борисовский, В.Ф. Электронные информационные ресурсы ОИЯИ: концепция и технологические решения / В.Ф. Борисовский [и др.] // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Четвертой Всерос. науч. конф. RCDL'2002 (Дубна, 15–17 октября 2002 г.) : в 2 т. – Дубна : ОИЯИ, 2002. – Т. 1. – С. 157–168.
- [2] Громов В.О. Электронная библиотека Объединенного института ядерных исследований / В.О. Громов [и др.] // Электронные библиотеки. – 1999. – Т. 2, вып.2.
- [3] Документация пакета EndNote <http://www.endnote.com/>
- [4] Проект Budapest Open Access Initiative <http://www.soros.org/openaccess/read.shtml>

Scientific and organizational digital resources of the Joint Institute for Nuclear Research

© V.F. Borisovsky V.V. Korenkov S.V. Kuniaev
N.A. Lenskaja G.G. Musulmanbekov E.G. Nikonov
I.A. Filozova

The paper is devoted to problems of informational support of digital resources in the Joint Institute for Nuclear Researches (JINR). The provision of the unified and transparent access to various data sources and the problem of applications integration are considered.

Новая полная электронная карта данных о форме и размерах атомных ядер*

© И.Н. Бобошин¹, В.В. Варламов¹, Ю. П. Гангровский², С.Ю. Комаров¹, К.П. Маринова²,
Р.И. Султанов¹

¹ Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына
Московского государственного университета им. М.В. Ломоносова, Москва,

² Объединенный институт ядерных исследований, Дубна Московской области
Komarov@depni.sinp.msu.ru

Аннотация

В Центре данных фотоядерных экспериментов (ЦДФЭ) НИИЯФ МГУ в рамках программы обеспечения российских пользователей точной и надежной ядерно-физической информации в течение ряда лет созданы несколько полных реляционных баз ядерных данных (БД) [1]. Источниками информации для них служат большие фонды числовых данных, создаваемые и поддерживаемые международной сетью Центров ядерных данных МАГАТЭ [2], участником которой является ЦДФЭ, а также и фонды данных, подготовленных в ЦДФЭ.

Доступ удаленных пользователей ко всем БД осуществляется через специально разработанный Web-интерфейс. Мощные и гибкие поисковые системы этих БД позволяют быстро находить данные, соответствующие огромному количеству признаков и их сочетаний и обрабатывать очень сложные, а иногда уникальные запросы. Вместе с тем они дают возможность решать задачи физики атомного ядра и ядерных реакций, многие из которых без них не могли быть не только решены, но и сформулированы. Это делает их и не только инструментами информационного обеспечения исследований, но и средствами самих исследований. С использованием возможностей расположенных на Web-сервере ЦДФЭ (<http://cdfe.sinp.msu.ru>) БД выполнен целый ряд оригинальных исследований и получена [1, 3] новая интересная информация о свойствах ядер и характеристиках ядерных реакций.

Одной из последних разработкой ЦДФЭ подобного типа является созданная ранее

электронная «Карта квадрупольных ядерных деформаций» [4], впоследствии дополненная современными данными о средне-квадратичных зарядовых радиусах (СКЗР) большого числа ядер и таким образом преобразованная в «Карту параметров формы и размеров ядер»

(<http://cdfe.sinp.msu.ru/services/radchart/radmain.html>) – полную коллекцию обсуждаемых данных.

Доклад посвящен краткому описанию новой БД и примеров ее использования для получения новой информации об атомных ядрах.

1 Введение

Как отмечалось в докладе на Конференции по электронным библиотекам в 2006 году [4] широкое применение как в фундаментальных, так и прикладных ядерно-физических исследованиях находит полная реляционная «База ядерно-спектроскопических данных ЦДФЭ» (<http://cdfe.sinp.msu.ru/services/ensdfr.html>). Она основана на информационном фонде Национального центра ядерных данных США, который создается и поддерживается в рамках широкого международного сотрудничества МАГАТЭ и предоставляется всем сотрудничающим центрам. Массив экспериментальных и оцененных данных по структуре ядер ENSDF включает в себя всю опубликованную к настоящему времени информацию (~500 Мб) об уровнях и переходах между ними для всех (~3500) известных в настоящее время атомных ядер – энергии, спины, четности, времена жизни, мультипольности, коэффициенты ветвления и смешивания, вероятности переходов, значения параметра $\log ft$ распадов и многие другие). В результате специально проведенных в ЦДФЭ работ созданная на основе этого массива реляционная БД “Relational ENSDF” была дополнена абсолютно уникальными данными об изо-спинах ядерных состояний, квадрупольных моментах Q и параметрах β_2 квадрупольной деформации ядер. Их уникальность заключается в том, что хотя эти данные и содержатся в исходном массиве

ENSDF, однако лишь в его неформализованных разделах, вследствие чего другими информационными системами аналогичного назначения пользователям эта информация не предоставляется.

Особый интерес в последнее время наряду с систематиками характеристик ядерных уровней и оболочек приобрели полные систематики данных о квадрупольных моментах Q ядер и параметрах β_2 их квадрупольной деформации, описывающих форму ядра. Повышенный интерес к общим сведениям о форме атомного ядра, а также к информации о том, как эта форма изменяется при переходе ядра из основного состояния в различные возбужденные состояния, прежде всего, связан с развитием исследований новых видов взаимодействий между нуклонами (протонами и нейтронами) ядра, находящимися на оболочках с различными квантовыми характеристиками. Обнаружено, что некоторые виды такого взаимодействия оказывают влияние не только на форму ядра, но и на формирование целого ряда его свойств и признаков, которые ранее считались принадлежащими лишь некоторым особо выделенным ядрам. В силу этой особенности такие выделенные ядра, у которых число протонов и/или число нейтронов равняется одному из 7 следующих – 2, 8, 20, 28, 50, 82, 126 – были названы магическими. Новые виды протон-нейтронного взаимодействия приводят, в конечном счете, к тому, что вполне определенные свойства магичности [5, 6] обнаруживаются у существенно большего, чем считалось ранее, числа ядер.

В связи с повышенным интересом к информации такого рода, данные о квадрупольных моментах Q атомных ядер и параметрах их квадрупольной деформации β_2 были систематизированы в особую электронную коллекцию. На ее основе была создана принципиально новая база данных – «Карта квадрупольных ядерных деформаций» (<http://cdfc.sinp.msu.ru/services/nsr/defchart/defmain.html>) [4], в которую вошли систематизированные данные обсуждаемого типа из трех авторитетных международных источников [7–9] (~2000 наборов данных для ~1500 ядер). В последнее время эта БД была дополнена данными о зарядовых радиусах ядер и фактически реорганизована в Карту параметров формы и размеров ядер (<http://cdfc.sinp.msu.ru/services/radchart/radmain.html>), которая и представляется в настоящем докладе.

2 Основные принципы представления данных в Карте квадрупольных ядерных деформаций, сохраненные в Карте параметров формы и размеров ядер

Поскольку новые данные о среднеквадратичных радиусах ядер были встроены в Карту квадрупольных ядерных деформаций [4], в настоящем докладе кратко представлены принципы представления данных:

- данные представлены в удобной графической форме, аналогичной форме представления ин-

формации на хорошо известных специалистам так называемых Картах нуклидов – блоки информации, описывающие многообразные спектроскопические свойства (параметры уровней, гамма-линии их распада, моды распада, времена жизни и др.) отдельных ядер, располагаются в пространстве координат «число протонов Z » – «число нейтронов N », при этом каждому ядру соответствует специальным образом окрашенный квадрат – блок информации для конкретного ядра (Z, N);

- используются 4 уровня масштаба, существенно облегчающие выбор интересующей пользователя области значений Z и N . Кроме того, предусмотрены панели быстрого поиска “Fast Search” и (при других масштабах) “Nucleus Search” для задания значений заряда Z и числа нейтронов N (или массового числа $A = Z + N$) конкретного, искомого пользователем, ядра;
- применена специальная расцветка элементов системы по типу традиционной расцветки географических карт. Так, коричневым цветом («горы») обозначены ядра с положительной деформацией ($Q > 0$), то есть вытянутые, синим («моря») – ядра с отрицательной деформацией ($Q < 0$), то есть сплюснутые, зеленым («равнины») – ядра с деформацией нулевой ($Q = 0$) или неизвестного знака;

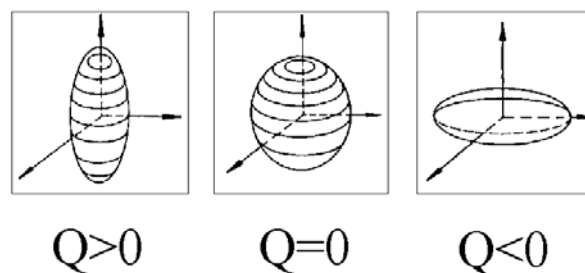


Рис. 1. Возможные варианты формы атомных ядер

- использованные 5 градаций каждого цвета идентифицируют параметры деформации по абсолютной величине – наибольшему значению соответствует более интенсивный цвет;
- предусмотрена панель “Deformation value search”. В данной панели возможно задание знака и диапазона значений параметра деформации β_2 . При нажатии кнопки “Go!” на Карте остаются окрашенными только ядра, имеющие деформацию, параметры которой находятся в заданном диапазоне;
- по умолчанию представление данных из трех указанных выше источников информации с возможностью отключения любого из них;
- по приводимым библиографическим ссылкам в виде уникальных идентификаторов БД ЦДФЭ по публикациям – NSR [10] возможен прямой выход на документы этой справочно-библиографической БД.

3 Данные о зарядовых радиусах в новой Карте

3.1 Источники информации

Среднеквадратичный зарядовый радиус (СКЗР) $R = \langle r^2 \rangle^{1/2}$ является фундаментальной характеристикой атомного ядра, изменение которой в зависимости от числа нуклонов (протонов и нейтронов) ядра часто является индикатором изменения фундаментальных свойств ядерной материи. Для его определения с достаточной точностью применяются различные методы, использующие особенности электромагнитных взаимодействий между ядром и его электронными оболочками. Основными из них являются эксперименты по определению энергий переходов в мюонных (μ^-) атомах, упруго рассеянных электронов (e^-), в которых возможно определение значений именно радиусов $R = \langle r^2 \rangle^{1/2}$, а также эксперименты по измерению энергий рентгеновских КХ-лучей и оптических изотопических сдвигов, в которых определяются разности радиусов $\delta \langle r^2 \rangle^{1/2}$ некоторых сравниваемых ядер.

В связи с этим в электронную карту были включены данные о СКЗР ядер, охватывающие более 700 изотопов ядер 60 элементов ($Z = 2 - 95$, $N = 3 - 148$). Представлены значения СКЗР, полученные тремя разными методами [11–13].

3.2 Организация доступа к данным

Для представления данных Карты оставлена та же [4] удобная графическая форма – расположение наборов данных в координатах «число протонов Z » – «число нейтронов N ».

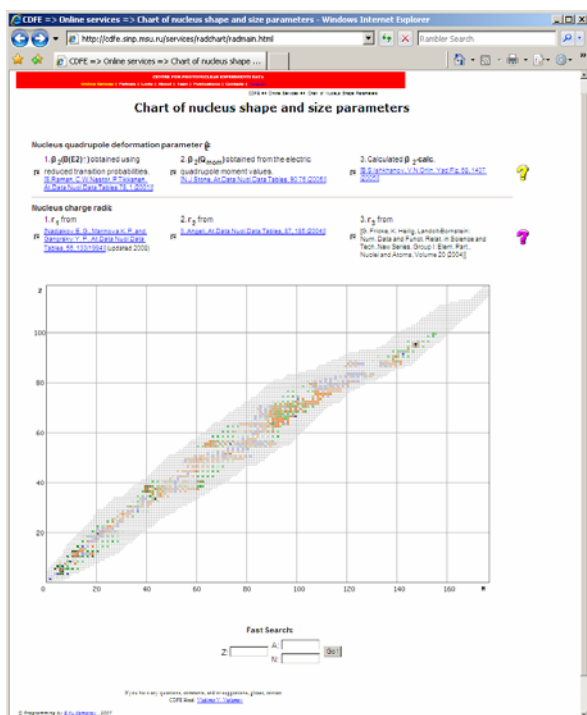


Рис. 2. Главная страница Карты данных о форме и размерах атомных ядер

Была использована та же специальная расцветка элементов Карты, выполненная по аналогии с той, что используется на традиционных географических картах.

Предусмотрены 4 уровня масштаба карты и панель быстрого поиска “Fast Search” для задания значений заряда Z и числа нейтронов N , конкретного ядра.

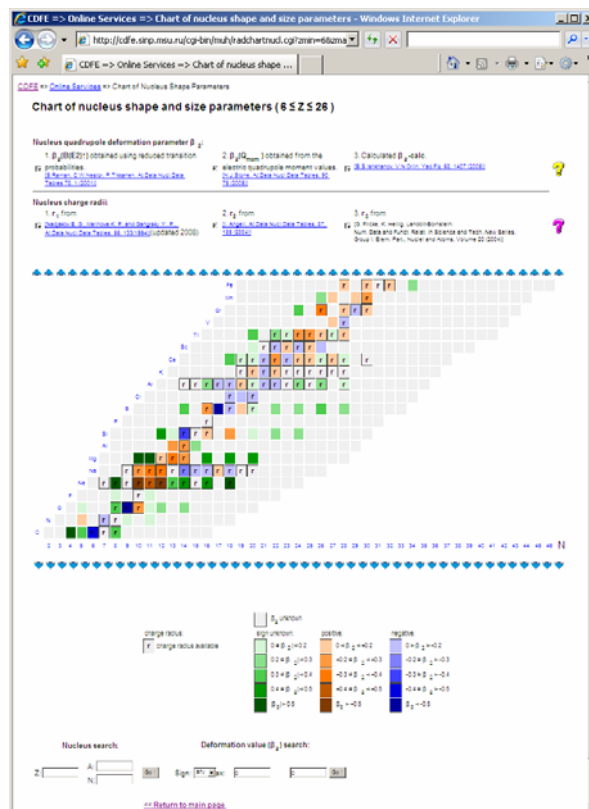


Рис. 3. Третий уровень масштаба Карты

Наличие данных о зарядовых радиусах обозначается следующим способом: в мелком масштабе дополнительной наклонной чертой (диагональ), в промежуточном масштабе – буквой «г» (рис. 3), в крупном масштабе – подчеркнутым курсивом для названия химического элемента.

3.3 Принципы организации программного обеспечения

Программные средства

Настоящая поисковая система представляет собой глубокую модернизацию Карты [4], выполненную с использованием программных языков Perl, JavaScript, каскадных таблиц стилей CSS и СУБД MySQL.

На главной странице новой Карты (1-й уровень масштаба) сверху представлены 3 источника данных по деформациям и 3 источника по зарядовым радиусам. Пользователь может выбрать один или несколько источников для отображения на карте нуклидов.

При нажатии на ветвь нуклидов, запрос пересылается на сервер ЦДФЭ, в программный модуль

“radchartpol.cgi”, где происходит генерация нужного png-изображения Карты с помощью вспомогательного модуля “radchartpng.cgi” на основе физических данных выбранных источников, и вставка этого изображения в html-страницу (2-й уровень масштаба).

Далее, при выборе пользователем более крупных масштабов карты – 3-го (рис. 3) и 4-го уровней (рис. 4), запрос предварительно обрабатывается модулем JavaScript “tmdef.js”, а затем пересылается на сервер, в программное ядро системы (“radchartnucl.cgi”). В файле “radchartnucl.cgi”, с помощью функций языка Perl происходит разбиение входных параметров запроса, обращение посредством библиотеки DBI в таблицу данных “deformations_new” СУБД MySQL, получение результатов из базы данных, построение html-таблицы, изображающей фрагмент ветви нуклидов, используя файл стилей “chartnucl.css”, и вывод этой таблицы на экран. В таблице стилей “chartnucl.css” располагаются параметры цветовой гаммы ядер. Подсветка выделенных ядер выполняется модулем “tmdef.js”.

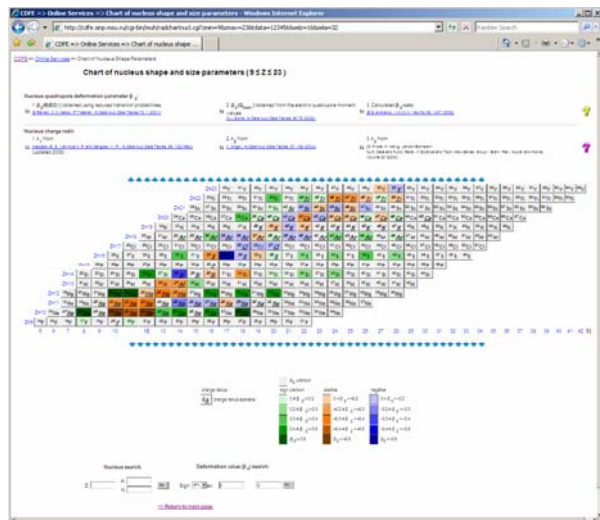


Рис. 4. Четвертый уровень масштаба Карты (область ядер с диапазоном атомных номеров $Z = 9-23$)

Обработка запроса с панелей “Fast Search”, “Nucleus Search”, “Deformation value search” и “Radius value search” (в процессе разработки) расположенных ниже таблицы нуклидов также выполняется сначала программой “tmdef.js”, а потом “radchartnucl.cgi”.

Построение таблиц физических параметров атомных ядер происходит с помощью модуля “radcard.cgi”. При этом снова выполняется запрос в базу данных MySQL, и на его основе происходит создание html-страницы с данными по деформациям и зарядовым радиусам искомого нуклида (рис. 5) и с гиперссылками в информационную систему NSR (“nsrsearch.cgi”).

Специальные кнопки (вопросительные знаки желтого и красного цветов) в правом верхнем углу главной страницы Карты (рис. 2) являются ссылкой

на документы справки (“radhelp.html” и “radhelp.html#rad”), представленные в русском и английском вариантах (русскоязычные варианты – “radhelp.html” и “radhelp_ru.html#rad”, соответственно).

45Sc (Z=21)

Chart of nucleus shape and size parameters

$\beta_2(Q_{\text{mom}})$

Parameter	Data 1	Data 2	Data 3	Data 4
Q-moment (barn) for 7/2 ⁺ state at E = 0.000 MeV	-0.216 ± 0.009	-0.22 ± 0.01	-0.22 ± 0.002	-0.236 ± 0.002
$\beta_2(Q_{\text{mom}})$	-0.136 ± 0.012	-0.139 ± 0.013	-0.139 ± 0.008	-0.149 ± 0.009
NSR Reference:	1971Ch25	1976E01	2000Ke12	2002M37
Journal Reference:	PR A4 1767 (71)	ZP A276 9 (76)	CPL 329 112 (00)	N.Naturforsch. 57a 595 (02)

β_2^{calc}

Parameter	Data
β_2^{calc}	-0.096 ± 0

Nucleus radii

Parameter	r_1	r_2	r_3
radius (fm)	3.646 ± 0.0035	3.6469 ± 0.0025	
Reference:	Yu. P. Gangrsky et al., HFI 171, 209 (2006); Yu. P. Gangrsky et al., private communication (2008)		
NSRCode:	2006GA47		

Рис. 5. Таблица физических данных Карты для ядра ^{45}Sc

Библиотечный модуль “parcheck.pm” выполняет функции безопасности, отсеивая неприемлемые варианты запросов к программным модулям системы.

4 Новая Карта в ядерно-физических исследованиях

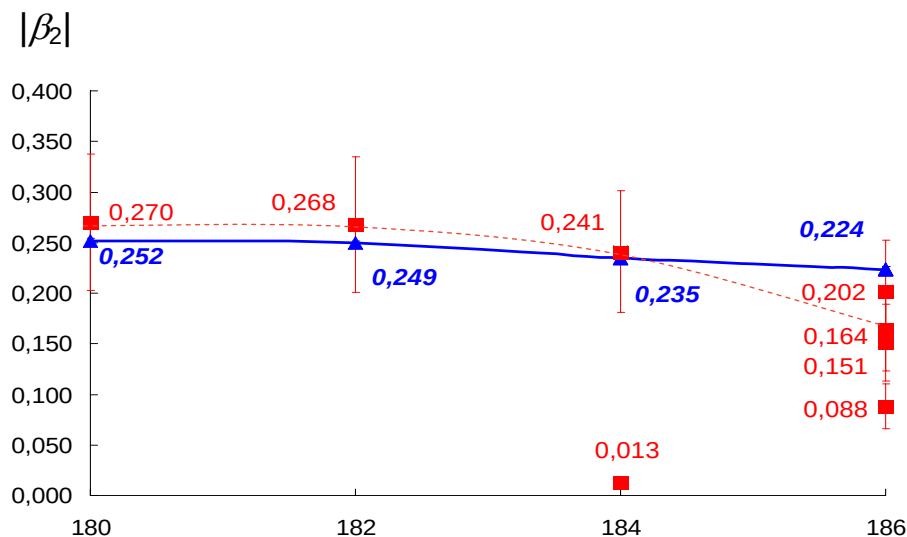
Как отмечалось выше, современные научные БД позволяют быстро находить данные, соответствующие некоторым уникальным признакам, а также и большому количеству относительно простых признаков и их сочетаний. Это позволяет обрабатывать запросы, которые могут быть, с одной стороны, простыми по физическому смыслу, но очень сложными, а иногда и уникальными по составу соотносимых признаков, а с другой стороны, относительно простые по используемым признакам, физический смысл ответов на которые заранее может быть не вполне ясен (ожидаем). Таким образом, развитые научные БД дают возможность решать задачи физики атомного ядра и ядерных реакций, многие из которых без них не могли быть не только решены, но и сформулированы. Это делает их и не только инструментами информационного обеспечения исследований, но и средствами самих исследований [14, 15].

4.1 Систематические расхождения данных о параметрах квадрупольной деформации ядер, полученных различными методами

С помощью новой реляционной БД – электронной Карты ядерных деформаций – были выявлены неизвестные ранее систематические расхождения данных о параметрах квадрупольной деформации ядер, полученных двумя традиционными разными способами, описанными выше – из приведенной вероятности переходов из основного состояния ядра в первое состояние с $J^\pi = 2^+_1$ и из квадрупольных моментов ядер.

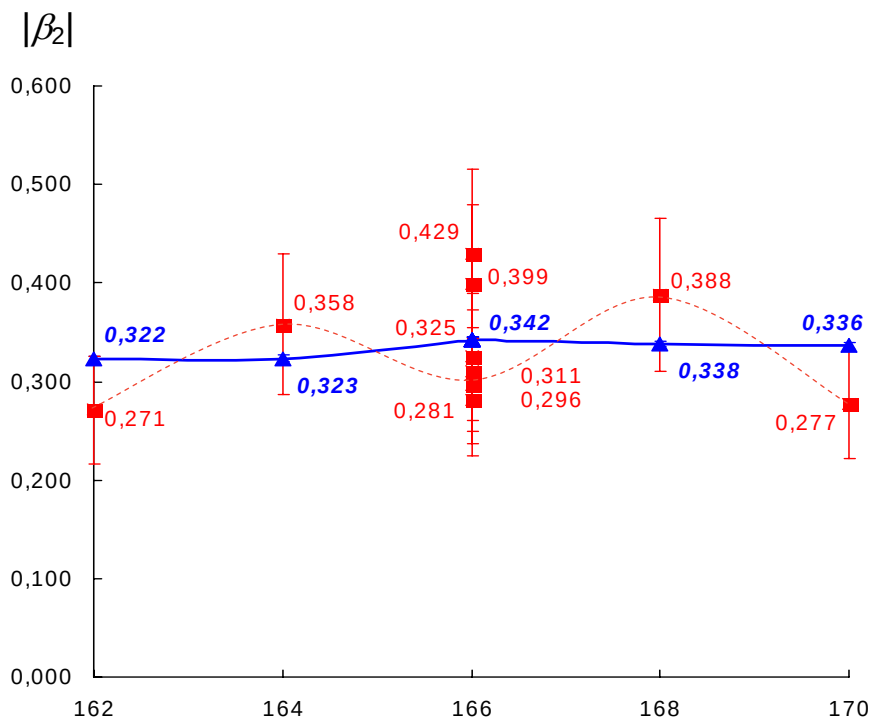
Уже при заполнении данными Карты квадрупольных ядерных деформаций было обнаружено, что все ядра, для которых имеется информация данного типа, четко разделяются (имеется лишь 4 случая некоторого пересечения данных) на две группы по способу ее получения.

На рис. 6 и 7 иллюстрируется практическое совпадение значений параметра квадрупольной деформации для ядер одной группы (представлены данные для изотопов W и Er), а на рис. 8 и 9 – значительные систематические расхождения этих значений для ядер другой группы (представлены данные для изотопов Ge и Sn).



А

Рис. 6. Практическое совпадение данных о параметре квадрупольной деформации, полученных разными методами для ядра ядер W (группа - Ti, Cr, Zr, Nd, Sm, Gd, Dy, Er, W)



А

Рис. 7. То же (что и на Рис. 6) для ядер Er

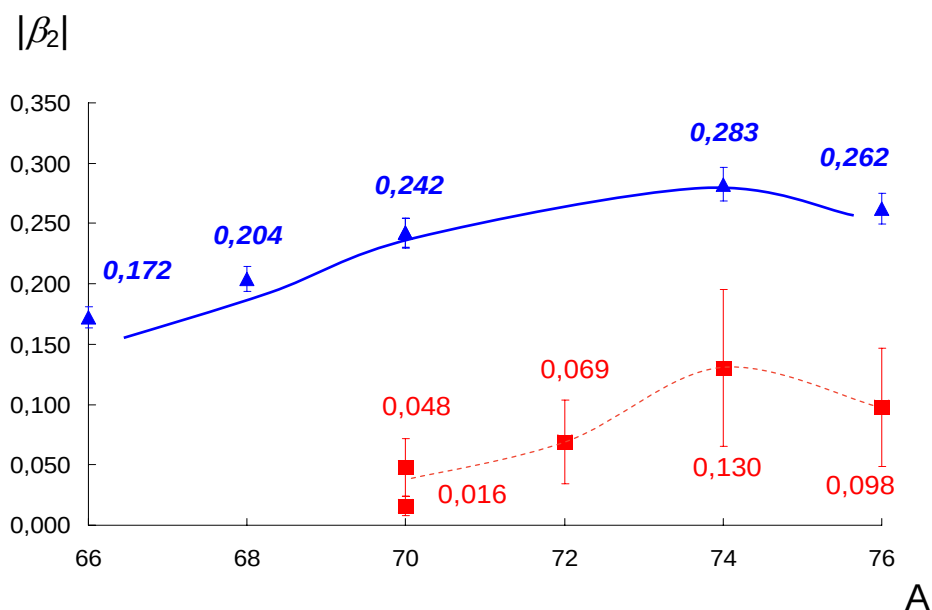


Рис. 8. Систематические расхождения данных о параметре квадрупольной деформации, полученных разными методами для ядер Ge - (группа - C, Si, Ar, Ca, Fe, Ni, Zn, Ge, Se, Kr, Sr, Mo, Ru, Pd, Cd, Sn, Te, Ba, Yb, Hf, Pt, Pb)

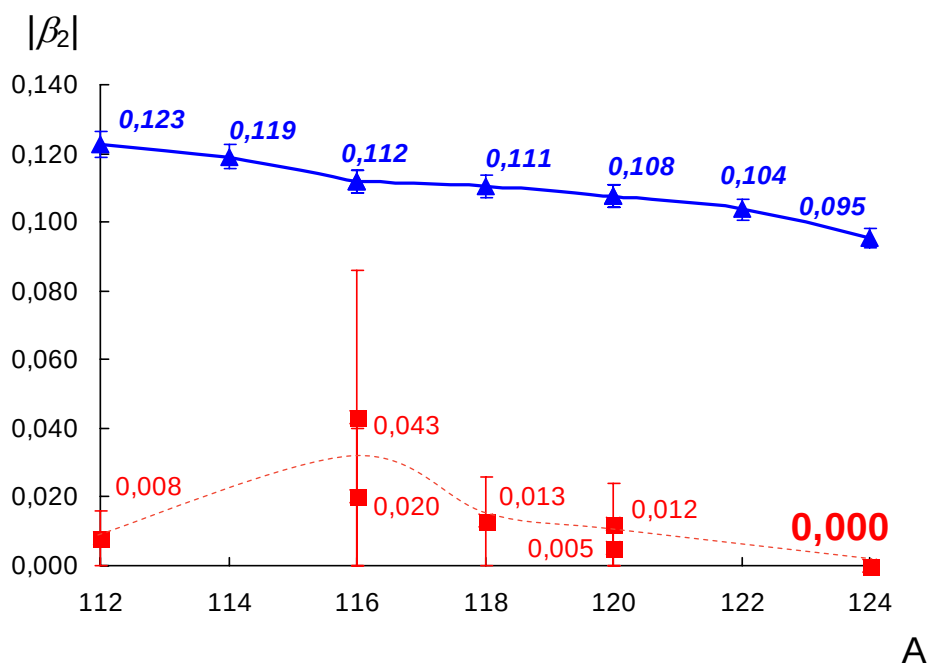


Рис. 9. То же (что и на рис. 8) для ядер Sn

Четкое разделение всех известных ядер на две группы, в одной из которых полученные разными способами значения близки, а во второй - значительно различаются (в пользу первого), может свидетельствовать [16] о проявлении в ядрах второй группы такого не изученного пока явления, как нулевые колебания поверхности ядра в основном состоянии.

Такие данные дают основания также и для обоснования зависимости этого явления от таких также не изученных характеристик ядерного вещества, как «жесткость» или «рыхлость».

4.2 Особенности зарядовых радиусов новых магических ядер

Хорошо известно наличие так называемых магических ядер, имеющих магические числа нуклонов (или протонов, или нейтронов, или тех и других – дважды магические ядра) – 2, 8, 20, 28, 50, 82, 126. Они имеют ряд специфических свойств, отличающих их от соседних изотопов и изотонов – меньшие размеры (радиусы), большая сферичность (меньшая деформация), большая устойчивость, особенные значения энергий отделения нуклонов и первых (самых нижних) уровней со спином и четностью $J^\pi = 2^+$.

На полной систематике данных по основному параметру магичности – энергии первого уровня с $J^\pi = 2^+$, полученной из ядерно-спектроскопической БД было установлено [17], что во-первых, таким классическим магическим ядрам отвечают специфические области вдоль прямых линий при соответствующих магических значениях числа протонов Z и числа нейтронов N . Однако кроме таких областей вдоль прямых линий имеются отчетливые замкнутые области – «острова магичности» при определенных парах магических чисел Z и N , некоторые из которых отличаются от перечисленных выше 7 магических чисел. Было показано, что такие области соответствуют новым дважды магическим ядрам то есть таким, которые имеют многие типичные свойства классических магических ядер, но не предусмотрены традиционной оболочечной моделью. С

помощью ядерно-спектроскопической БД такие ядра могут быть обнаружены, например, по характерным особенностям расположения как низколежащих состояний (например, уровней с выделенными значениями спина и четности – $J^\pi = 2^+$), так и самых верхних (вблизи границы Ферми) одночастичных состояний (подоболочек и оболочек). К настоящему времени таких ядер установлено уже довольно много – $^{92,94,96,98}\text{Zr}$, $^{90,92,94,96}\text{Sr}$, $^{26,28,30}\text{Si}$, ^{30}Si , $^{14,16,24,28,40,48}\text{O}$, ^{14}C и др.

Описываемая БД параметров формы и размеров ядра также может быть использована в таких исследованиях. Рис. 10 дает наглядное представление о том, как, например, одно из самых заметных ядер этого ряда новых магических – ^{96}Zr – выделяется среди соседних с ним изотопов и изотонов меньшими размерами.

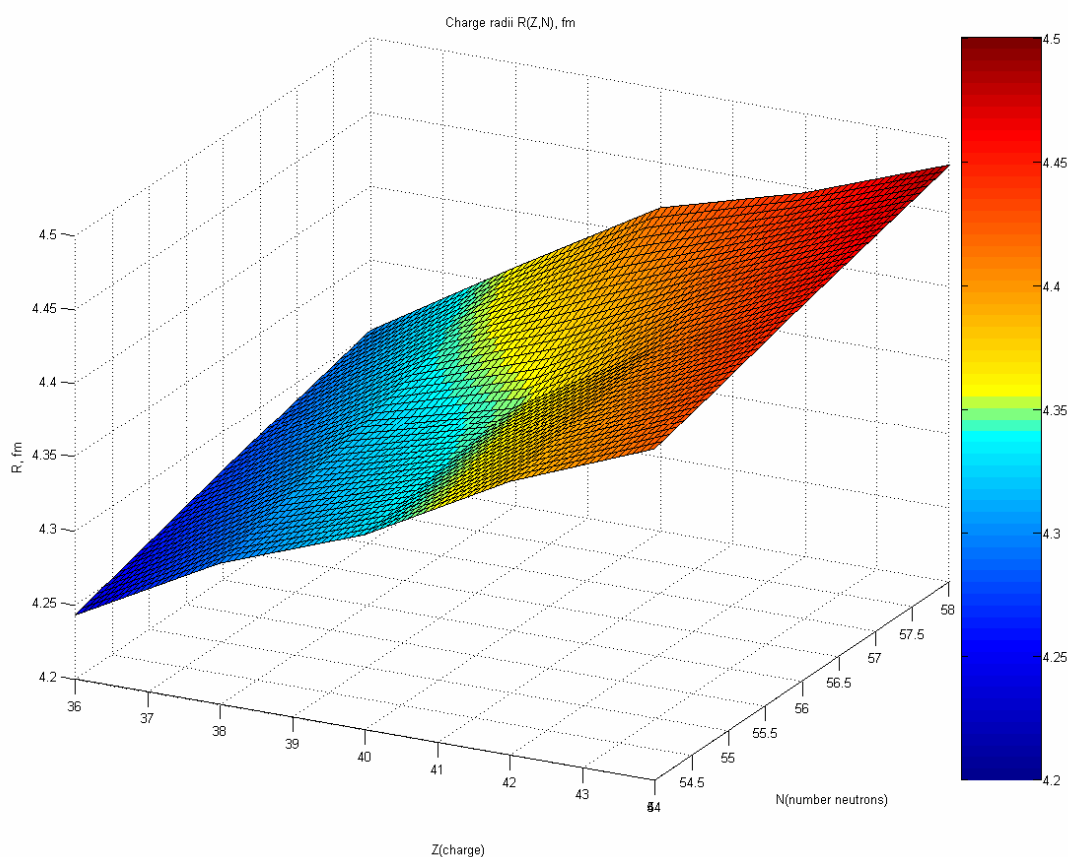


Рис. 10. Систематика значений зарядового радиуса для ядер в окрестности нового магического ядра ^{96}Zr . Хорошо видно, что это ядро достаточно отчетливо отличается меньшими размерами от своих соседей и по числу протонов Z (изотонов) и по числу нейтронов N (изотопов)

4.3 Систематики зарядовых радиусов ядер различных областей массовых чисел A

Характерная особенность магических ядер, заключающаяся в том, что они имеют меньшие по сравнению с соседями размеры также могут быть весьма наглядно иллюстрированы информацией из БД.

На рис. 9 представлен пример систематики зарядовых радиусов в области ядер с $Z = 3\text{--}30$. Ясно видны особенности (уменьшение среднеквадратичных

радиусов) именно в областях классических магических чисел (в данном случае – чисел нейтронов) 8, 20, 28.

Весьма интересно отметить, что вполне отчетливо выраженные особенности обсуждаемого типа наблюдаются и для числа нейтронов $N = 14$. Это число активно обсуждается в последнее время в качестве кандидата в новое магическое. В списке новых магических ядер, приведенном в предыдущем разделе, этому новому магическому числу ней-

тронов ядро ^{30}S , имеющее также новое магическое число протонов $Z = 16$.

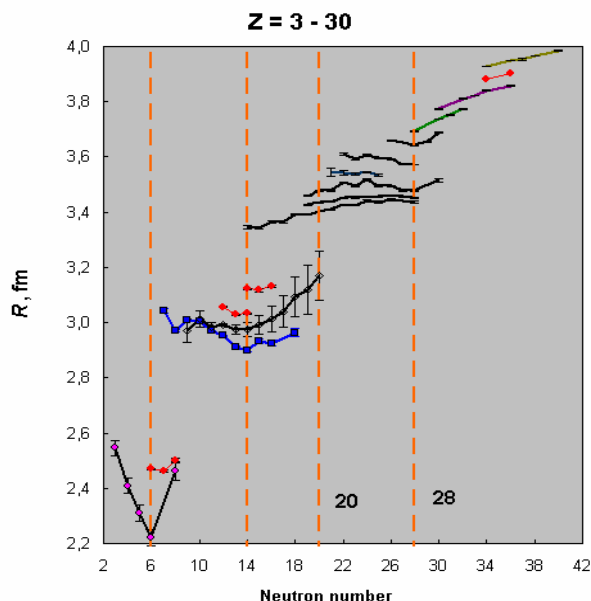


Рис. 9. Систематика зарядовых радиусов изотопов ядер, имеющих число протонов в области 3–30

5 Заключение

С использованием большого числа данных из авторитетных международных компиляций, систематизирующих информацию о форме (квадрупольные моменты и параметры квадрупольной деформации) и размерах (зарядовые радиусы) в Центре данных фотоядерных экспериментов НИИЯФ МГУ создана новая реляционная БД, постоянно доступная пользователям в среде Интернет.

Показано, что даже БД относительно скромного содержания – параметры формы ядер – предоставляют пользователям новые, достаточно нетрадиционные, в чем-то уникальные возможности проведения научных исследований, которые без таких БД были бы невозможны или, по крайней мере, существенно затруднены.

Безусловно, что своеобразные аналитические (или предсказательные) возможности БД, в первую очередь, определяются разнообразием и полнотой содержащейся в ней информации – наибольшее число интересных и оригинальных проблем было решено с помощью возможностей таких универсальных полных современных научных БД, какой является полная реляционная «База ядерно-спектроскопических данных ЦДФЭ “Relational ENSDF” (<http://cdfe.sinp.msu.ru/services/ensdfr.html>)», кратко представленная выше.

К настоящему времени перечень конкретных случаев получения новых данных, новой информации, нового знания с помощью запросов различной сложности к достаточно полным научным БД достаточно велик, однако и скромные случаи, описанные в настоящем докладе, позволяют характеризовать развитые научные базы как средства не только информационного обеспечения научных исследований, но и самих исследований.

Литература

- [1] Труды Всероссийской научной конференции «Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет РФФИ». И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров и др., 2007, с. 318.
- [2] Ed. by V.G. Pronyaev, The Nuclear Data Centres Network. IAEA Nuclear Data Section, INDC(NDS)-401, IAEA, Vienna, Austria, 1999.
- [3] В.В. Варламов, И.Н. Бобошин. Базы данных как метод исследования. *Природа*, № 12 (2005) 29.
- [4] Труды Восьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров и др., 2006, с. 145.
- [5] I.N. Boboshin, B.S. Ishkhanov, V.V. Varlamov. *Ядерная физика*, 67 (2004) 1872.
- [6] О.В. Беспалова, И.Н. Бобошин, В.В. Варламов, Т.А. Ермакова, Б.С. Ишханов, Е.А. Романовский, Т.И. Спасская. *Известия РАН, серия физическая*, 69 (2005) 123.
- [7] S. Raman, C.W. Nestor, P. Tikkanen. *Atomic Data and Nuclear Data Tables*, 78 (2001) 1.
- [8] N.J. Stone. *Atomic Data and Nuclear Data Tables*, 90 (2005) 75.
- [9] Б.С. Ишханов, В.Н. Орлин. *Ядерная физика*, 68 (2005) 1407.
- [10] Всероссийская научная конференция «Научный сервис в сети Интернет». В.В. Варламов, С.Ю. Комаров, С.Б. Сёмин и др. 2003, с. 52.
- [11] E.G. Nadjakov, K.P. Marinova, Yu.P. Gangrsky. *Atomic Data and Nuclear Data Tables*, 56 (1994) 133; K. P. Marinova, Yu. P. Gangrsky, data updated, private communication, 2008.
- [12] I. Angeli. *Atomic Data and Nuclear Data Tables*, 87 (2004) 185.
- [13] G. Fricke, K. Heilig. *Landolt-Bornstein: Numerical Data and Functional Relations in Science and Technology. New Series, Group I: Elementary Particles, Nuclei and Atoms, Volume 20* (2004).
- [14] В.В. Варламов, И.Н. Бобошин. Базы данных как метод исследования // *Природа*. – 2005, № 12. – С. 29–38.
- [15] Математическая морфология. Электронный математический и медико-биологический журнал – Смоленск: СГМА, 2000, том. 3, вып. 3. И.Н. Бобошин, А.В. Варламов, В.В. Варламов, Е.М. Иванов, М.Е. Степанов, В.В. Чесноков. Ядерно-физические исследования и электронные информационные ресурсы в Интернет. URL: <http://www.smolensk.ru/user/sgma/MMORPH/N-6.html/VARLAMOV/varlamov.htm>.
- [16] Proceedings of the International Conference on Nuclear Data for Science and Technology, April 22–27, 2007, Nice, France. I. Boboshin, B. Ishkhanov, S. Komarov, V. Orlin, N. Peskov and V. Varlamov. DOI: 10.1051/ndata:07103, p. 65.
- [17] И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров и др. *Известия РАН, серия физическая*, 71 (2007) 334.

New complete electronic Chart of Data on Shape and Values of Atomic Nuclei

I.N. Boboshin, V.V. Varlamov, Yu.P. Gangrsky,
S.Yu. Komarov, K.P. Marinova, R.I. Sultanov

Several complete relational nuclear databases (DB) were produced at the Centre for Photonuclear Experiments Data (Centr Danykh Fotoyadernykh Eksperimentov) of Lomonosov Moscow State University Skobeltsyn Institute of Nuclear Physics in the frame of program of giving exact and reliable nuclear physics information to the Russian users. Sources of data are both the large numerical data funds produced and managed by the international IAEA Nuclear Data Centers Network, participants of which is the CDFE and the CDFE's data.

All DB are available to users through the specially developed Web-interface. Powerful and flexible Search Engines give to one possibility to find rapidly the data in accordance with large amount of data features and their combinations and process very complicated and sometimes unique queries. At the same time they give to one possibility to solve atomic nuclei and nuclear physics tasks that could not only solved but formulated without those DB. That means that such DB are the tools not only of research informational support but of research themselves. Several original scientific research [1–3] have been carried out and new interesting information on properties of atomic nuclei and features of nuclear reactions has been obtained using the possibilities of CDFE DB (<http://cdfe.sinp.msu.ru>).

One of the last CDFE product of type under discussion is the “Chart of Quadrupole Nuclear Deformations” [4] later have been added by modern mean-root-square charge radii data for many nuclei and therefore transformed into the “Chart of Nucleus Shape and Size Parameters” – (<http://cdfe.sinp.msu.ru/services/radchart/radmain.html>) – complete collection of data under discussion.

The talk is devoted to the short description of new DB and examples of its using for obtaining new information on atomic nuclei.

* Работы выполнялись в Лаборатории анализа ядерных данных (Центр данных фотоядерных экспериментов) Отдела электромагнитных процессов и взаимодействий атомных ядер НИИЯФ МГУ и были частично поддержаны несколькими грантами РФФИ (№№ 99-07-90015, 03-07-90431, 04-02-16275, 04-07-99601), а также грантами Президента РФ для научных школ №№ НШ-1619.2003.2, НШ-5365.2006.2 и НШ-485.2008.2.

Электронные коллекции ядерных данных как средства фундаментальных и прикладных исследований в области физики атомных ядер и ядерных реакций*

© И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров, Н.Н. Песков, М.Е. Степанов,
В.В. Чесноков

Центр данных фотоядерных экспериментов
Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына
Московский государственный университет им. М.В. Ломоносова
Varlamov@depni.sinp.msu.ru

Аннотация

Описывается система реляционных баз данных (БД) по физике атомных ядер и ядерных реакций, созданная в рамках широкого международного сотрудничества – Сети Центров ядерных данных МАГАТЭ, участником которой много лет является ЦДФЭ НИИЯФ МГУ. Несколько больших (полных, репрезентативных) электронных коллекций ядерно-физических данных в сочетании с мощными и гибкими поисковыми системами позволяют решать широкий круг задач в традиционном для БД направлении информационного обеспечения научных и прикладных исследований. Вместе с тем, они позволяют на основе полученных ранее данных получать данные новые, что делает их новым инструментом (средством) собственно исследований.

Доклад включает краткие описания содержания основных БД ЦДФЭ и примеры использования БД для получения новых данных, новой информации и, в конечном счете, нового знания.

1 Введение

Любая электронная коллекция по существу представляет собой всего лишь «склад готовой продукции» – сведения об окружающей нас действительности, соответствующим образом собранные и оформленные. Если такие сведения являются отрывочными (фрагментарными), коллекция так и остается «складом». Однако, если эти сведения достаточно полно характеризуют описываемый ими объект, в полном соответствии с известным философ-

ским законом количество переходит в качество. Если говорить более правильно – количество, конечно же, так и остается количеством, но, начиная с некоторого уровня, начинает эффективно способствовать рождению и росту нового качества, становится своеобразной средой (почвой) на которой и вырастает новое качество. Это становится очевидным, когда достаточно полной коллекции данных определенного типа придается мощная и гибкая поисковая система, преобразующая коллекцию в реляционную базу данных (БД). Электронные средства поиска точно определенного объекта среди большого (огромного) количества «готовой продукции», «разложенной» по многим «полочкам», снабженной точными реквизитами («этикетками»), открывают многие новые возможности для получения новых научных данных и знаний, среди которых в первую очередь могут быть отмечены следующие:

- выявление на основе совместного анализа результатов многих экспериментов неизвестных ранее систематических закономерностей;
- определение на основе совместной оценки результатов различных экспериментов их систематических погрешностей;
- получение точных и надежных новых данных на основе объединенной оценки результатов различных экспериментов с учетом их систематических погрешностей;
- оценка результатов экспериментов, которые по тем или иным причинам не были (не могут быть) проведены.

Возможность исследовать подобные закономерности означает, что развитые БД обладают определенными аналитическими и предсказательными возможностями.

Иными словами – удобный одновременный и эффективный доступ к огромному количеству данных не только определенного типа, но и разных типов, позволяет задавать вопросы (формулировать запросы к БД), которые без такого доступа просто не могли «приходить в голову» исследователю по простой причине – для них не было оснований. Об-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

работка таких запросов во многих случаях приводит и к ответам, которые тоже не «приходили в голову». А такие ответы представляют собой, не что иное, как данные, неизвестные ранее, то есть новые. В конечном счете, такие ответы – новое знание об объекте исследования [1].

2 Реляционные базы ядерных данных ЦДФЭ

2.1 Ядерные данные в международной сети Центров ядерных данных МАГАТЭ

Атомные ядра представляют собой сложные физические системы, которые описываются очень большим количеством параметров, получаемых из различных экспериментов, а ядерные реакции и радиоактивные распады являются сложными процессами, в которых ядра превращаются друг в друга. В этой связи, БД по свойствам ядер и характеристикам реакций имеют значительный объем и весьма сложное строение. В силу ряда особых исторических причин (одна из основных – разработка ядерных реакторов сначала для военных, а затем мирных целей) создание сначала коллекций ядерных данных началось под эгидой Международного агентства по атомной энергии (МАГАТЭ) очень давно, носило планомерный, долговременный и глобальный характер и проводилось. К 90-м годам, когда во всем мире начался информационный бум, ядерная физика оказалась едва ли не самой передовой и подготовленной в этом отношении областью знаний. В настоящее время свыше 20 центров и групп из разных стран – Австрии, Китая, Кореи, России, Словакии, США, Украины, Франции, Швеции, Японии и др. объединены в различные Сети Центров ядерных данных МАГАТЭ [2]. Россию в этом сотрудничестве представляют 4 организации – Центр ядерных данных (ФЭИ, Обнинск), Центр атомных и ядерных данных (ИАЭ, Москва), Центр ядерно-физических данных (ВНИИЭФ, Саров), Центр данных фотоядерных экспериментов (НИИЯФ МГУ, Москва). Каждый из Центров отвечает в этом сотрудничестве за определенный тип данных (например, ЦДФЭ – за фотоядерные данные, то есть за данные по ядерным реакциям под действием γ -квантов) и обеспечение ими определенного географического региона. При этом каждый Центр имеет доступ ко всем накапливаемым массивам самых различных ядерных данных и возможность использовать их наиболее оптимальными для себя и своих пользователей способами, призванными обеспечивать эффективное информационное обеспечение как фундаментальных, так и прикладных ядерно-физических и смежных с ними исследований.

2.2 Основные БД ЦДФЭ

Одна из главных задач информационного обеспечения научных исследований - создание условий

для эффективного использования исследователями данных и знаний, накопленных ранее, с целью получения новых данных и знаний. В течение нескольких последних лет ЦДФЭ идет в этом направлении по пути создания системы полных баз ядерных данных – тех самых «складов готовой продукции» при наличии развитых систем поиска информации превращающихся в новое эффективное средство научных исследований. При использовании в качестве источников информации огромных фондов числовых данных, создаваемых и поддерживаемых международной Сетью, а также собственных больших и полных фондов данных, в ЦДФЭ НИИЯФ МГУ созданы, поддерживаются и предоставляются пользователям на Web-сайте ЦДФЭ 10 реляционных БД, содержащих огромное количество данных по ядрам и реакциям.



Рис. 1. Главная страница Web-сайта ЦДФЭ НИИЯФ МГУ

Основные БД, созданные с 1999 года, были достаточно полно представлены на соответствующих конференциях [3–9]:

- «База данных по ядерным реакциям» (<http://cdfe.sinp.msu.ru/exfor/index.php>) – выходы, сечения, дифференциальные сечения, функции возбуждения полных и парциальных реакций, энергетические, угловые, зарядовые, массовые и другие распределения, я частиц продуктов и др.; имеет целый ряд преимуществ по сравнению с некоторыми аналогами, в частности, обеспечивает поиск в режиме «обратная геометрия»; БД включает в себя материалы из

боле 250000 статей, содержит свыше 2 миллионов наборов данных, общий объем исходного массива превышает 500 Мб;

- «База ядерно-спектроскопических данных» (<http://cdfе.sinp.msu.ru/services/ensdfr.html>); – огромное количество характеристик уровней и переходов между ними во всех известных в настоящее время (~3200) ядрах; включает ряд данных, отсутствующих в некоторых аналогичных системах, и не имеет аналогов по мощности и гибкости поисковой системы; объем исходного массива данных составляет около 500 Мб;
- «Параметры основных состояний ядер» (<http://cdfе.sinp.msu.ru/services/gsp.en.html>); не имеет аналогов по полноте и разнообразию представленных данных о ядре в целом (распространенности стабильных изотопов или периоды полураспада радиоактивных, массы, дефекты массы, энергии связи и др.);
- «Параметры гигантского дипольного резонанса, сечения фотоядерных реакций» (<http://cdfе.sinp.msu.ru/services/gdrsearch.html>); уникальная БД – разработка ЦДФЭ, содержит все основные характеристики (энергетическое положение, амплитуда, ширина, интегральные сечения) гигантских дипольных резонансов, наблюдаемых в сечениях фотоядерных реакций;
- «Карта данных о форме и размерах атомных ядер» (<http://cdfе.sinp.msu.ru/services/radchart/radmain.html>); – уникальная БД – разработка ЦДФЭ; содержащая большое количество данных из авторитетных международных источников об основных параметрах, описывающих форму ядра (квадрупольный момент и параметр квадрупольной деформации), а также о среднеквадратичных зарядовых радиусах большого количества ядер;
- «Калькулятор порогов и энергий ядерных реакций и Графическая система построения данных об энергиях отделения нуклонов» (http://cdfе.sinp.msu.ru/muh/calc_thr.shtml); БД – разработка ЦДФЭ, позволяющая удобно и быстро определять энергии и пороги любых ядерных реакций, а также наглядно представлять изменения энергий отделения одного и двух нуклонов (нейтронов и протонов) с изменением массового числа ядра; имеет многие преимущества по сравнению с многочисленными аналогами;
- «Публикации по ядерной физике» (http://cdfе.sinp.msu.ru/services/nsr/Search_form.shtml); является уникальной реляционной БД такого содержания – рефераты в ключевых словах огромного количества публикаций, посвященных как экспериментальным, так и теоретическим исследованиям по ядерной физике; БД включает в себя свыше 500000, с огромной скоростью (около 10000 в год) пополняется и охватывает работы по тематике, начиная с пионер-

ских работ Э.Резерфорда, опубликованных в 1910–1912 годах;

- «Индекс фотоядерных данных с 1955 г.» (<http://cdfе.sinp.msu.ru/services/pnisearch.html>); уникальная БД – разработка ЦДФЭ, содержащая большое количество рефератов экспериментальных работ по электромагнитным взаимодействиям, во многом дополняющих содержание БД по публикациям.

Доступ удаленных пользователей ко всем БД осуществляется через специально разработанные Web-интерфейсы. При создании БД и их поисковых систем был использован единый подход (операционная система LINUX, система управления базами данных MySQL, Web-сервер “Apache”, технологии обработки запросов CGI, языки программирования C++, Perl и php). Это дало возможность на основе нескольких изначально независимых БД впервые построить интегрированную «Универсальную электронную систему информации по атомным ядрам и ядерным реакциям» (<http://cdfе.sinp.msu.ru/services/unifsys/index.html>) [6, 7] с объединенным интерфейсом, позволяющим существенно повысить эффективность использования, как фондов данных, так и поисковых возможностей интегрированных БД.

3 Возможности баз ядерных данных ЦДФЭ для ядерно-физических исследований

Безусловно, что упомянутые выше аналитические и предсказательные возможности созданных в ЦДФЭ и аналогичных БД связаны, в первую очередь, с результатами поисков по огромному количеству разнообразных признаков информации и их сочетаниям. Самые разнообразные поиски такого типа, очевидно, несколько различные для разных БД, достаточно проиллюстрировать одним, ориентированным на ядерно-спектроскопические данные. Так, например, может быть сформулирован запрос, обработка которого невозможна ни в одной другой из известных систем аналогичного назначения: найти такие атомные ядра, в определенной области энергий возбуждения которых имеются уровни с определенными характеристиками (спин, четность, время жизни, изоспин и др.), которые распадаются с испусканием γ -квантов, также имеющих точно определенные характеристики (энергия, мультипольность, коэффициенты ветвления и смешивания и др.), приводящих к образованию конечного ядра в состояниях также с точно определенными характеристиками. К этому запросу могут быть добавлены и другие требования, относящиеся, например, к типам реакций и/или распадов, в которых описанные выше процессы происходят и т.д. и т.п. Очевидно, что может быть получен результат (список определенных ядер с необходимыми свойствами), который никакими иными способами не может быть получен.

Однако и вполне тривиальные для развитых научных БД запросы, сформулированные в соответст-

вии с некоторыми впервые пришедшими в голову исследователя физическими вопросами, позволяют решать задачи физики атомного ядра и ядерных реакций, многие из которых без них не могли быть не только решены, но и сформулированы.

3.1 Поиск данных по ядерным реакциям по конечному ядру

Некоторые примеры возможностей первого типа были представлены, например, в докладе [10]. В контексте настоящего доклада может быть упомянут один из них – совершенно тривиальный с точки зрения возможностей БД по ядерным реакциям запрос на данные о сечениях различных ядерных реакций, объединенных одним и тем же конкретным конечным ядром. Обработка такого запроса иными средствами может оказаться по продолжительности «несовместимой с человеческой жизнью» поскольку ни в одном библиотечном каталоге, рубрикаторе, индексе нет сортировки информации по конечному ядру реакции (везде имеются сортировки по ядру-мишени и налетающей частице, кое-где – по небольшому числу вылетающих частиц). Приведенный пример относился к достаточно жесткому излучению изотопа ^{60}Co ($T_{1/2} = 1925.1$ дня), которое может оказывать отрицательное воздействие, например, на прецизионную электронику на борту космического аппарата. При полете космического аппарата все элементы его конструкции облучаются самыми различными частицами и ионами. При этом в самых различных материалах этих элементов про-

исходят самые разные реакции и образуются разнообразные ядра-продукты. Упомянутый тривиальный запрос к БД показывает, какие материалы и каких именно элементов конструкции должны быть заменены для того, чтобы исключить (существенно снизить) воздействие излучения изотопа ^{60}Co . Из результатов поиска отчетливо видно, что если поиск данных по относительно простым реакциям типа $^{59}\text{Co}(n,\gamma)$, $^{60}\text{Ni}(n,p)$, $^{59}\text{Co}(d,p)$, $^{63}\text{Cu}(n,\alpha)$ или $^{60}\text{Ni}(t,\text{He-3})$ без БД хотя и с большими трудностями, но может быть проведен, поиск по остальным сложным многочастичным реакциям (например, $^{59}\text{Co}(t,np)$, $^{59}\text{Co}(\alpha,n2p)$, $^{65}\text{Cu}(p,3n3p)$) практически безнадежен («в голову исследователя идея о нем не приходит»).

3.2 Решение проблем систематических расхождений результатов исследований фотоядерных реакций различными методами

Специалистам по фотоядерным реакциям давно и хорошо известен факт существенных расхождений, как по форме, так и по величине энергетических зависимостей сечений реакций, полученных в экспериментах двух типов – на пучках тормозного γ -излучения и квазимоноэнергетических аннигиляционных фотонов. Суть расхождений заключается в том, что, несмотря на близкие условия, сечения в экспериментах первого типа имеют, как правило, практически всегда, существенно более отчетливо выраженную промежуточную структуру и заметно большие амплитуды.

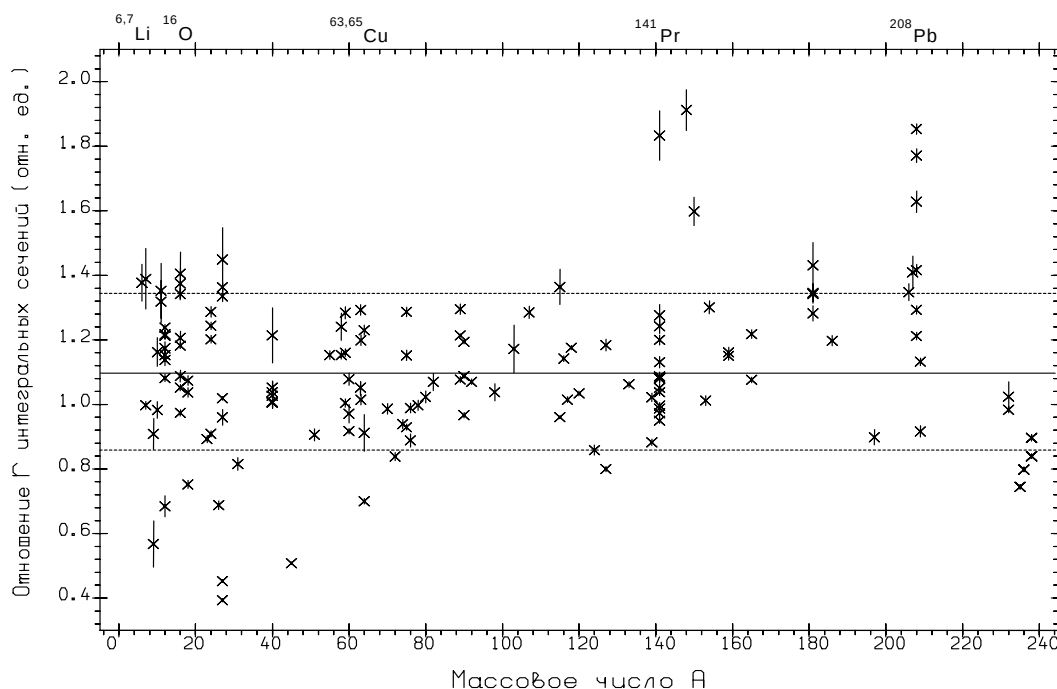


Рис. 2. Систематика данных, полученных в различных лабораториях, к данным одной из них – хорошо видны систематические расхождения величиной 12 %

Ранние попытки выяснить причины таких расхождений и разработать методы для их устранения, предпринимаемые исследователями, выполнявшими

эксперименты разного типа, приводили к диаметрально противоположным рекомендациям. Использование результатов поиска данных в БД по ядер-

ным реакциям в соответствии с простым запросом относительно интегральных сечений полной фотонейтронной реакции $(\gamma, xn) = (\gamma, n) + 2(\gamma, 2n)$ позволило выявить основные причины отмеченных расхождений. Из рис. 2 хорошо видно, что сечения реакции, полученные в одной из лабораторий, систематически отличаются по абсолютной величине от сечений других лабораторий, а способ устранения этих расхождений относительно прост – перенормировка данных с использованием коэффициента

1.12. Аналогичная выборка данных об интегральных сечениях парциальных реакций с образованием одного и двух нейтронов (γ, n) и $(\gamma, 2n)$ выявило (рис. 3) наличие существенно более серьезных расхождений, причина которых выявляется однозначно – неточности процедуры разделения процессов с образованием одного и двух нейтронов – но для устранения которых необходим серьезный пересчет результатов обеих лабораторий.

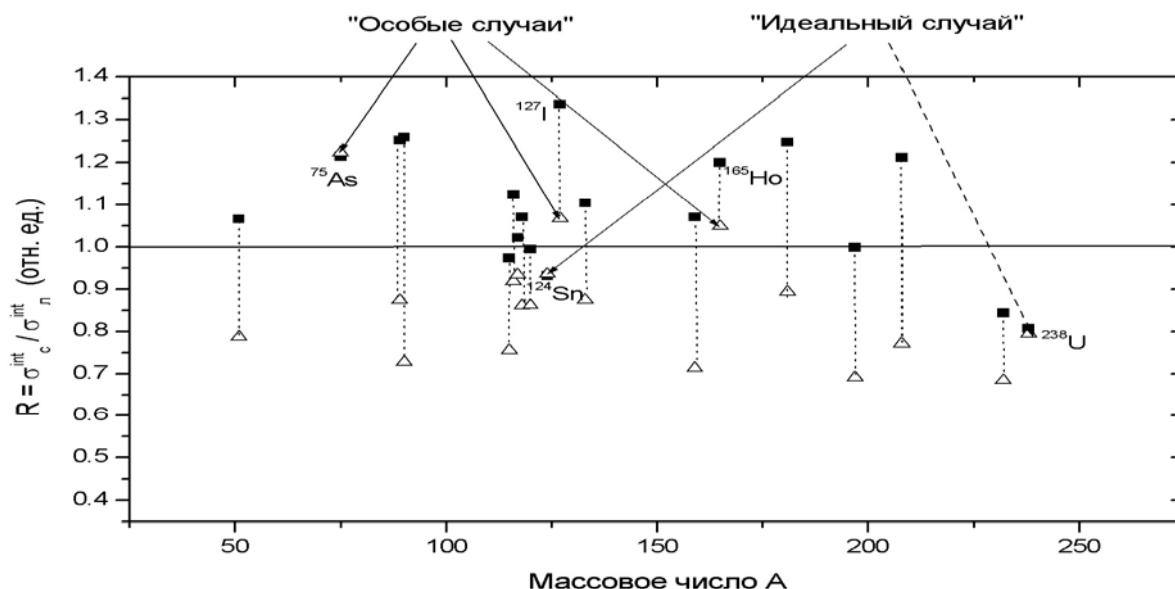


Рис. 3. Систематика данных о сечениях реакций с образованием одного и двух нейтронов, полученных в двух разных лабораториях. В то время, как сечения реакций с одним нейтроном имеют большие амплитуды в одной из них (прямоугольники имеют среднее значение ~ 1.2), сечения реакций с двумя нейтронами (треугольники имеют среднее значение ~ 0.8) – в другой

С учетом этого обстоятельства были предложены методы приведения результатов различных экспериментов к единой интерпретации и, тем самым, получения новых точных и надежных согласованных между собой данных. Системный анализ данных по сечениям фотонейтронных реакций с вылетом различного числа нейтронов также позволил выявить причины наблюдаемых значительных расхождений по абсолютной величине сечений парциальных реакций, полученных в различных экспериментах, и разработать метод их учета, по существу – снять проблему известных расхождений [11, 12].

3.3 Новая особенность атомных ядер – динамическая деформация в основном состоянии

С помощью новой реляционной БД – электронной Карты ядерных деформаций – были выявлены неизвестные ранее систематические расхождения данных о параметрах квадрупольной деформации ядер, полученных традиционными, но разными способами – из приведенной вероятности переходов из основного состояния ядра в первое состояние с $J^\pi = 2^+$ и из квадрупольных моментов ядер.

Четкое разделение [13, 14] всех известных ядер на две группы, в одной из которых полученные разными

способами значения близки, а во второй – значительно различаются (в пользу первого), может свидетельствовать о проявлении в ядрах второй группы такого не изученного пока явления, как нулевые колебания поверхности ядра в основном состоянии.

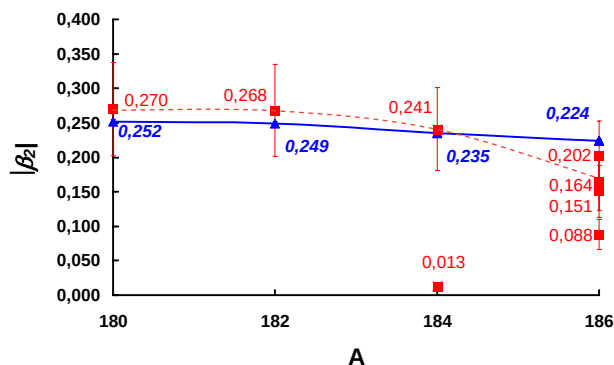


Рис. 4. Практическое совпадение данных о параметре квадрупольной деформации, полученных из данных о квадрупольных моментах ядер для ядер W (группа – Ti, Cr, Zr, Nd, Sm, Gd, Dy, Er, W, Os, Ra).

Такие нулевые колебания свидетельствуют о том, что наряду с традиционной и достаточно хорошо

исследованной «статической» деформацией часть ядер обладает деформацией «динамической», которая к настоящему времени практически не изучена.

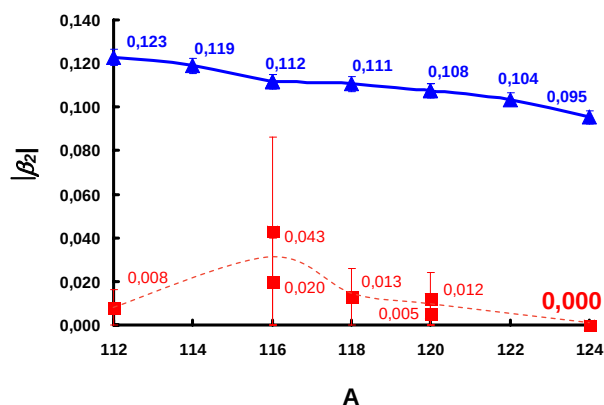


Рис. 5. Систематические расхождения данных для ядер Sn (группа – C, Si, Ar, Ca, Fe, Ni, Zn, Ge, Se, Kr, Sr, Mo, Ru, Pd, Cd, Sn, Te, Ba, Yb, Hf, Pt, Pb)

3.4 Энергии первых состояний ядер с $J^\pi = 2^+$ и новые магические ядра

Хорошо известно наличие так называемых магических ядер, имеющих магические числа нуклонов (или протонов, или нейтронов, или тех и других –

дважды магические ядра) – 2, 8, 20, 28, 50, 82, 126. Такие ядра имеют ряд специфических свойств, отличающих их от соседних изотопов и изотонов – большая сферичность (меньшая деформация), большая устойчивость, особенные значения энергий отделения нуклонов и первых (самых нижних) уровней со спином и четностью $J^\pi = 2^+$ и др. На полной систематике данных из ядерно-спектроскопической БД по энергиям первых уровней с $J^\pi = 2^+$ таким ядрам отвечают прямые линии – «горные хребты» – при соответствующих значениях Z и N.

Согласно традиционной классической оболочечной модели заполнение некоторых протонных или нейтронных оболочек приводит к формированию магических чисел нуклонов и выделению ядер, которые обладают таким числом нуклонов, по целому ряду свойств на фоне соседей. При магическом числе протонов магические свойства сохраняются для всех изотопов, при магическом числе нейтронов – для всех изотонов (такие ядра и располагаются вдоль отчетливых прямых линий систематики).

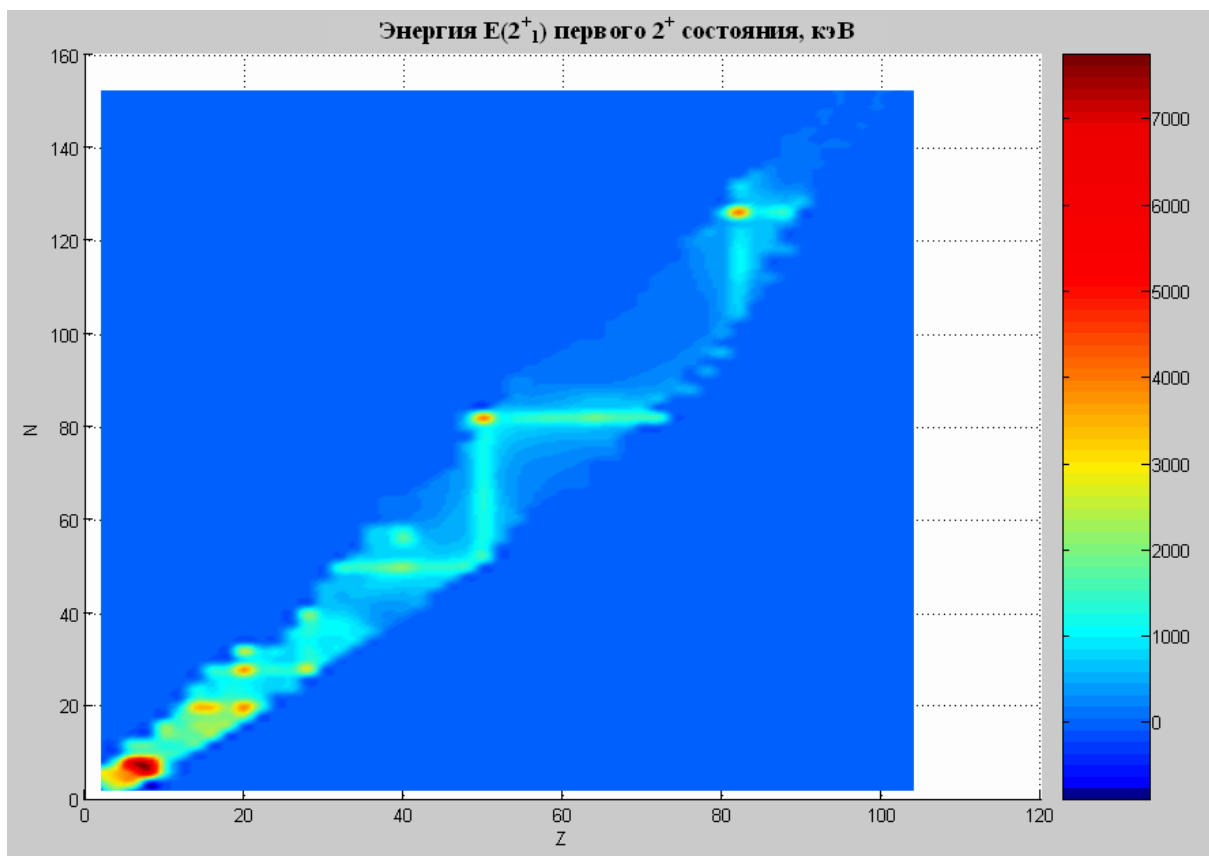


Рис. 6. Простая выборка из базы данных по структуре атомных ядер значений энергий первых уровней с $J^\pi = 2^+$. Хорошо видно, что кроме прямых линий, отвечающих классическим магическим ядрам, имеются отчетливо выраженные островки ($Z = 40, N = 56$), ($Z = 20, N = 32$), ($Z = 14-16, N = 14-16$) и др. – новые дважды магические ядра

Вместе с тем приведенная систематика с очевидностью показывает, что наряду с «линиями магичности» существуют некоторые замкнутые области – «острова магичности» – вблизи некоторых пар чисел протонов и нейтронов, из которых либо одно либо оба не предусматриваются классической моделью оболочек в качестве магических чисел. Наиболее отчетливо «острова магичности» наблюдаются для комбинаций ($Z = 40, N = 56$), ($Z = 20, N = 32$), ($Z = 14 - 16, N = 14 - 16$).

Таким образом, простейшая выборка из БД определенной характеристики ядер (энергия некоторого специфического уровня) довольно ясно свидетельствует о наличии нового класса новых магических ядер. Они не являются по определению магическими (традиционная магичность связана именно с заполнением определенных подоболочек), но имеют многие свойства (прежде всего, главное – большую энергию первого уровня с $J^\pi = 2^+$), типичные для дважды магических ядер.

3.5 Предсказательные возможности простого графического представления энергий отделения нуклонов от атомных ядер

Некоторые другие свойства новых магических ядер, типичные для классических магических ядер, могут быть легко исследованы также с помощью достаточно тривиальных запросов к другим БД.

Так, например, даже относительно простая «Графическая система построения данных об энергиях отделения нуклонов», дополняющая «Калькулятор порогов и энергий ядерных реакций» (обе системы – по существу, также реляционные БД – основаны на хорошо известной специалистам коллекции массовых характеристик ядер [15]), позволила внести огромный вклад в эти исследования. Она очень эффективно позволила наблюдать у новых магических ядер особенности массовых и зарядовых зависимостей энергий отделения нуклонов, которые характерны для классических магических ядер.

По существу, данные, приведенные на рис. 7, наглядно иллюстрируют то, каким образом научная БД может не только отвечать на некоторые специальным образом сформулированные вопросы к ней, но и сама задавать исследователям вопросы, становящиеся стимулом для проведения новых исследований. В данном случае выявления специфических для магических ядер особенностей (изломов в зависимостях от массового числа A энергий отделения нейтронов для случая нескольких четно-четных изотопов ядра Zr) этот вопрос формулируется следующим образом: Если наблюдаемые на рис. 7 изломы для $A = 90$ соответствуют хорошо известному магическому по числу нейтронов ($N = 50$) ядру ^{90}Zr , то, что собой представляет ядро ^{96}Zr , имеющее число протонов (40) и нейтронов (56), которые классической оболочечной моделью не рассматриваются как магические, потому что не соответствуют необходимым оболочкам.

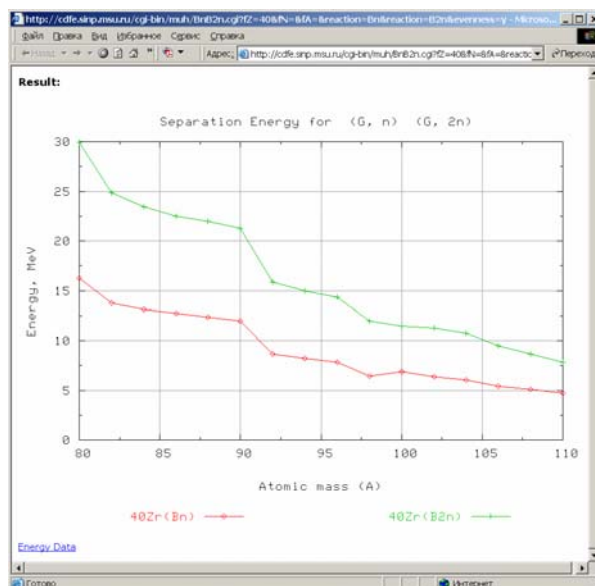


Рис. 7. Энергии отделения одного и двух нейтронов от изотопов циркония

Еще одна простая выборка из ядерно-спектроскопической БД информации об энергиях и числах нуклонов на подоболочках новых магических ядер устанавливает общие закономерности в структуре верхних (вблизи границы Ферми) оболочек таких ядер, свидетельствующие в пользу наличия некоторого дополнительного взаимодействия между заполненными протонными и нейтронными оболочками, возможно связанного с эффектами протон-нейтронного спаривания.

Совместное использование нескольких БД дает возможность провести системное изучение соотношения в ядрах, соседних с исследуемыми, таких специфических параметров «магичности», как значения энергии некоторых уровней, параметры формы ядра и энергии отделения одного и двух нуклонов, особенности взаимного расположения верхних протонных и нейтронных подоболочек ядер. Оказывается, что практически идентичное поведение всех перечисленных параметров «магичности», а также и некоторых других позволяет выявить достаточно большое количество новых неизвестных ранее магических ядер – $^{92,94,96,98}\text{Zr}$, $^{90,92,94,96}\text{Sr}$, $^{26,28,30}\text{Si}$, ^{30}Si , $^{14,16,24,28,40,48}\text{O}$, ^{14}C и других [16].

Таким образом, поиск ответа на вопрос, заданный одной БД, о том, что, собственно, представляет собой ядро ^{96}Zr , привел к обнаружению значительного количества новых дважды магических ядер, имеющих свойства, близкие к свойствам классических магических ядер и структуру верхних оболочек, такую же, как у ядра ^{96}Zr . Результаты тривиальных поисков в других БД открыли возможности объяснения факта их существования с использованием нового типа притягивающего взаимодействия между протонами и нейтронами типа спаривания [17].

4 Заключение

Перечень конкретных случаев получения новых данных, новой информации, нового знания с помощью относительно просто сформулированных запросов к достаточно полным научным БД может быть продолжен. Безусловно, что для того, чтобы получить решение уже совсем нетривиальной научной проблемы, в некоторых случаях результат тривиального поиска в БД, должен быть дополнен некоторым (иногда также простейшим, а иногда и очень нетривиальным анализом) или далеко не простой взаимной обработкой результатов разных экспериментов.

К научным проблемам, которые были таким способом не только сформулированы, но и решены, могут быть добавлены, например, и следующие:

- предсказания (без проведения сложнейших длительных и дорогих экспериментов на пучках поляризованных частиц) значений спинов и четностей большого числа уровней ядер на основании простейшего анализа (по существу – простого сопоставления систематик) данных о возбужденных состояниях определенного типа в различных ядрах;
- подтверждение на основе экспериментальных данных сделанного теоретически в НИИЯФ МГУ открытия (1987 г., № 342, авторы – Б.С. Ишханов, И.М. Капитонов, В.Г. Неудачин, В.Г. Шевченко, Н.П. Юдин) явления так называемого конфигурационного расщепления гигантского дипольного резонанса; оно было достигнуто на основе простейшей выборки из БД по ядерным реакциям данных по сечениям различных фотоядерных реакций для двух изотопов ${}^6,7\text{Li}$; они были по некоторому специальному признаку разделены на две группы, что и дало возможность получить параметры двух компонент расщепления;
- решение принципиальной проблемы существенных систематических расхождений по форме сечений фотоядерных реакций, полученных в экспериментах на различных фотонных пучках; оно было основано на совместном анализе систематик некоторых параметров этих сечений, полученных в результате простых запросов к БД по ядерным реакциям; разработанные с помощью результатов таких запросов методы взаимной корректировки результатов разных экспериментов, учитывали особенности условий определения реально достигаемого энергетического разрешения эксперимента и позволили решить проблему;
- объяснение непонятной ранее аномалии в значениях спинов основных состояний некоторых ядер, например, изотопов ядра калия – ${}^{39,40,41,43,45,47}\text{K}$; оно было основано на установленном факте, что некоторые ядерные орбиты реально оказываются расположены друг относительно друга не совсем так, как это предсказывается классической моделью оболочек; объ-

яснение было достигнуто в результате применения весьма нетривиального нового метода совместного анализа результатов двух систематик – простейших выборок из БД результатов исследования на одного и того же ядра в ядерных реакциях нуклонного срыва и подхвата;

- и многие другие.

Таким образом, все сказанное выше свидетельствует о том, что развитые научные БД действительно довольно простыми способами не только повышают эффективность исследований в областях, ставших источниками информации для этих самых БД, но и открывают совершенно новые возможности самих исследований. При этом, зачастую они позволяют решать довольно сложные, иногда принципиально новые задачи физики атомного ядра и ядерных реакций, многие из которых без них не могли быть не только решены, но и вообще сформулированы. Именно такие качества и позволяют говорить о таких БД как о новых инструментах (средствах) не только информационного обеспечения, информационной поддержки исследований, но и собственно самих исследований.

Литература

- [1] В.В. Варламов, И.Н. Бобошин. Базы данных как метод исследования // Природа. – 2005, №12. – С. 29–38.
- [2] Ed. by V.G. Pronyaev, The Nuclear Data Centres Network. IAEA Nuclear Data Section, INDC(NDS)-401, IAEA, Vienna, Austria, 1999.
- [3] Сборник докладов Второй Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». И.Н. Бобошин, А.В. Варламов, В.В. Варламов и др., 2000, с. 39.
- [4] Труды Всероссийской научной конференции «Научный сервис в сети Интернет». И.Н. Бобошин, В.В. Варламов, Е.М. Иванов и др., 2001, с. 19.
- [5] Бобошин, И.Н. Электронная коллекция научных данных по физике атомных ядер и ядерных реакций ЦДФЭ НИИЯФ МГУ / И.Н. Бобошин [и др.] // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Четвертой Всероссийской научной конференции RCDL'2002 (Дубна, 15–17 окт. 2002 г.) : в 2 т. – Дубна, 2002. – Т. 1. – 335 с. – С. 290–298.
- [6] Труды Всероссийской научной конференции «Научный сервис в сети Интернет». И.Н. Бобошин, Д.Д. Бранец, В.В. Варламов и др., 2004, с. 22.
- [7] Труды Седьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров и др., 2005, с. 56.
- [8] Труды Всероссийской научной конференции «Научный сервис в сети Интернет: технологии

- параллельного программирования». И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров и др., 2006, с. 198.
- [9] Труды Всероссийской научной конференции «Научный сервис в сети Интернет: многоядерный компьютерный мир. 15 лет РФФИ». И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров и др., 2007, с. 318.
- [10] Математическая морфология. Электронный математический и медико-биологический журнал – Смоленск: СГМА, 2000, т. 3, вып. 3. И.Н. Бобошин, А.В. Варламов, В.В. Варламов, Е.М. Иванов, М.Е. Степанов, В.В. Чесноков. Ядерно-физические исследования и электронные информационные ресурсы в Интернет. URL: <http://www.smolensk.ru/user/sgma/MMORPH/N-6-html/VARLAMOV/varlamov.htm>.
- [11] В.В. Варламов, Б.С. Ишханов. *ЭЧАЯ*, том 35, вып. 4, 2004, с. 858.
- [12] В.В. Варламов, Н.Н. Песков, Д.С. Руденко и др. Вопросы атомной науки и техники. Серия: Ядерные константы, № 1–2 (2003) 48.
- [13] Proceedings of the International Conference on Nuclear Data for Science and Technology, April 22–27, 2007, Nice, France. I. Boboshin, B. Ishkhanov, S. Komarov, V. Orlin, N. Peskov and V. Varlamov. DOI: 10.1051/ndata:07103, p. 65.
- [14] И.Н. Бобошин, В.В. Варламов, С.Ю. Комаров и др. Известия РАН, серия физическая, 71 (2007) 334.
- [15] G. Audi, A.H. Wapstra, C. Thibault. *Nucl. Phys.*, A729 (2003) 337.
- [16] И.Н. Бобошин, В.В. Варламов, Б.С. Ишханов и др. Известия РАН, серия физическая, 72 (2008) 308.
- [17] I.N. Boboshin New Magic Nuclei and Neutron-Proton Pairing (in Russian), MSU SINP Preprint – 2007 – 4/825, URL: <http://dbserv.sinp.msu.ru:8080/sinp/files/pp-825.pdf>.

Nuclear Data Electronic Collections as Tools for Basic and Applied Research in the Field of Atomic Nuclei and Nuclear Reactions

I.N. Boboshin, V.V. Varlamov, S.Yu. Komarov,
N.N. Peskov, M.E. Stepanov, V.V. Chesnokov

The system of atomic nuclei and nuclear reactions physics relational databases (DB) produced in the frame of wide international cooperation – the IAEA Nuclear Data Centers Network, participants of which is the Centre for Photonuclear Experiments Data (Centr Danykh Fotoyadernykh Eksperimentov) of Lomonosov Moscow State University Skobeltsyn Institute of Nuclear Physics is described. Several large (complete) electronic collections of nuclear data in combination with powerful and flexible Search Engines give to one possibility to solve wide range tasks of informational support of basic and applied research. At the same time they allow using previously obtained data to obtain new data that do them new tool of research namely.

The talk includes the short descriptions of the main CDFE's DB and an examples of DB using for new data, new information and new knowledge at last obtaining.

* Работы выполнялись в Лаборатории анализа ядерных данных (Центр данных фотоядерных экспериментов) Отдела электромагнитных процессов и взаимодействий атомных ядер НИИЯФ МГУ и были частично поддержаны несколькими грантами РФФИ (№№ 99-07-90015, 03-07-90431, 04-02-16275, 04-07-99601), а также - грантами Президента РФ для научных школ №№ НШ-1619.2003.2, НШ-5365.2006.2 и НШ-485.2008.2

Алгоритмы распределения потоков информации в графической поисковой системе в реальном масштабе времени

© И.М. Гостев, А.Г. Подгорбунский

Московский институт электроники и математики

igostev@list.ru, shurikp@gmail.com

Аннотация

В работе рассматриваются преимущества и недостатки динамического и статического распределений видеопотока в графической поисковой системе в реальном масштабе времени в распределенной вычислительной среде. Приводятся модели на основе ориентированных графов. Рассматриваются алгоритмы построения распределенной вычислительной сети. Проводится их сравнение, и обсуждаются особенности реализации.

1 Введение

С развитием вычислительной техники в человеческую жизнь неизбежно вторгаются новые понятия и технологии. Появление новых технологий предоставляет новые возможности, а на основании новых возможностей и новых технологий у людей формируются новые потребности.

Развитие библиотечного дела приводит к появлению гигантских банков данных, которые содержат разнородную информацию. К ней относятся: различные формы представления текстов (форматы .rft, .pdf, .tex, djvu и т. п.); графическая информация в виде рисунков, скетчей и др. изображений в самых различных форматах; видеофильмы; аудио книги и многие другие составляющие мультимедийной информации, включая различные комбинации. Обилие и разнородность информации приводит к необходимости использования современных технологий их хранения и поиска.

Если в настоящее время поиск текстовой информации еще как-то удовлетворительно решен, то с навигацией по мультимедийной информации остается практически в начальном состоянии уже длительное время. Известные системы поиска графической информации по содержанию графического документа, практически не изменились за послед-

ние 10 лет и не вышли за пределы исследовательских прототипов.

В [1] было приведено описание графической поисковой системы (Graphic Search System – GSS) для поиска графического содержания в глобальных банках данных, а в [2] введена математическая модель поисковой системы, которая затем была реализована. Однако хотя эта система и была позиционирована как поисковая система реального времени, она не позволяет производить поиск в видео документах. Это объясняется тем, что, как и в любой другой поисковой системе, в ней создается индексируемый список по некоторым заранее выбранным критериям поиска. Однако это удовлетворительно (по скорости) работает с графическими изображениями, но при переходе к поиску фрагментов в видеопотоках требует существенного повышения скорости обработки.

Оценка таких требований, выполненная в [3] показывает, что скорость вычислительной системы для обработки видео потока должна быть пропорциональна размеру изображения, и для телевизионного формата превышать существующие мощности одиночного компьютера примерно в 25-30 раз. Для решения этой проблемы были разработаны параллельные алгоритмы, позволяющие реализовать поисковую систему в распределенной среде.

При обработке непрерывного видеопотока графической информации необходимо решать несколько проблем. Во-первых, необходимо оценивать быстродействие вычислительной системы, на которой производится такая обработка. Во-вторых, – разработать специальные методы, которые обеспечат работу в реальном времени.

Интуитивно понятно, что при высоких скоростях видео потока включающего несколько этапов его обработки, невозможно обеспечить выполнение всех операций на одном компьютере. Необходимо строить процесс на основе распределенной вычислительной системы. В такой среде каждый из ее элементов будет выполнять свою операцию по обработке изображения и распознаванию образов. Управление всем процессом по обработке и распознаванию должно осуществляться одним контроллером, который будет распределять задания по вы-

числительным узлам системы, осуществлять контроль и синхронизацию их выполнения.

В такой системе каждый метод по обработке и распознаванию необходимо строить так, чтобы обеспечить выполнение операции при однократном проходе по изображению. Кроме того методы (по возможности) не должны иметь параметров настройки процесса на изображение. В работах [3] и [4] была предложена и развита идея построения распределенной системы обработки и распознавания изображений в реальном времени, основанной на принципах конвейера и нарезки потока на участки. При проведении последующих исследований возникла проблема распределения отдельных фрагментов потока по процессорам, на которых он должен обрабатываться. В теории параллельных вычислений известны несколько способов моделирования параллельных вычислений. Одним из них является представление процесса при помощи сетей Петри [5]. В настоящей работе рассматривается еще один метод, основанный на применении ориентированных графов [6].

2 Математическая модель процессов

Методы распределения потока можно разделить на статические и динамические. В первом случае разбиение потока определяется на этапе начальной инициализации системы и далее остается неизменным. Для любого фрагмента потока можно заранее сказать, по какому маршруту он пройдет, и какими экземплярами каждого метода будет обработан. При динамическом способе расчета, это определяется на стадии обработки. Статический подход имеет два основных преимущества и позволяет:

- снять с вычислительных узлов нагрузку по расчету маршрутизации фрагментов изображения на стадии выполнения;
- построить гораздо более эффективную и простую структуру для хранения фрагментов изображения в памяти каждого вычислительного узла.

Рассмотрим односвязный, ориентированный, ациклический граф, имеющий один исток и один сток, вершинами которого являются *экземпляры метода* или *узлы вычислительной сети методов*, а дугами – *каналы передачи потока*.

По каналам вычислительной сети методов передаются данные, состоящий из фрагментов потока. Методы, имеют несколько входящих каналов и несколько исходящих, осуществляют «нарезку» потока – разделение потока на несколько потоков, содержащих чередующиеся фрагменты, и «сборку» потока – слияние нескольких потоков в один.

Всем процессом обработки потока управляет компонент, называемый «контроллером потока». Контроллер потока в качестве входной информации принимает набор методов обработки и распознавания изображений и параметры этих методов.

Метод – процедура обработки изображения, реализованная в качестве библиотечной функции. В качестве входных данных метод принимает изобра-

жение (фрагмент изображения) и параметры для процедуры обработки. В результате работы метода получается выходное изображение и/или некоторые другие данные, полученные в результате применения процедуры, формат и смысл которых зависит от конкретного метода.

Под методом также понимается программа, функционирующая как узел вычислительной сети методов.

Далее рассматривается случай, когда:

- графический поток представляет собой одно очень длинное непрерывное изображение.
- *фрагментом* такого потока будет непрерывный участок изображения, состоящий из целого числа строк.
- Методы обработки изображения либо точечные либо пространственные.

Вычислительная сеть методов будет удовлетворять следующим ограничениям:

- Каждому элементу *последовательности методов* соответствует один или несколько узлов, являющихся экземплярами этого метода. Совокупность этих узлов образует *ярус* сети, соответствующий данному методу цепочки.
- Все входные каналы данного узла, соединяют его только с узлами яруса, соответствующего предыдущему методу цепочки. Все выходные каналы узла соединяют его только с узлами яруса, соответствующего следующему методу в цепочке.
- Первый ярус содержит единственный узел, соответствующий первому методу цепочки методов – методу загрузки (считывания) изображения и не имеет входных каналов.
- Последний ярус содержит единственный узел, соответствующий последнему методу цепочки – методу сохранения (выдачи) результирующего изображения и не имеет исходящих каналов.

Каждый метод имеет набор характеристик:

- относительная вычислительная сложность процедуры;
- смещение по высоте (y^-, y^+) ;
- параметры метода.

Относительная вычислительная сложность рассчитывается эмпирически на конкретной ЭВМ относительно эталонного метода, как соотношение средних времен обработки эталонного изображения.

Смещения по высоте представляют собой пару чисел, определяющих сколько строк до и после фрагмента требуется для обработки соответствующего фрагмента результирующего изображения.

Другими словами, смещение (y^-, y^+) означает, что для вычисления $I[s_1, s_2]$ необходим фрагмент $I[s_1 - y^-, s_2 + y^+]$. Эти смещения играют большую роль в процессе расчета схемы потока через вычислительную сеть и определяют некоторые требова-

ния к организации памяти для фрагментов изображения внутри вычислительных узлов.

3 Алгоритм построения вычислительной среды

3.1 Статическое распределения потока

Рассмотрим алгоритм статического распределения потока. Пусть исходный поток представляет собой одно длинное изображение – «ленту». Фрагмент потока – непрерывный участок изображения, содержащий целое число строк. Промежуточные потоки состоят из последовательности фрагментов. Если поток состоит из фрагментов одинаковой длины, следующих через равные промежутки, то он может быть описан схемой потока.

Схема потока представляет собой тройку $\langle S, L, P \rangle$ где S – начальная строка первого фрагмента. L – длина каждого фрагмента в строках, P – промежуток от конца предыдущего фрагмента до начала следующего. Т. о. $\langle S, L, P \rangle$ задает последовательность фрагментов $[S + (L + P) * i, S + (L + P) * i + L]$, $i \in N_0$

Две схемы называются *смежными*, если поток, полученный слиянием двух потоков (состоящий из всех фрагментов первого и второго потока) так же может быть задан схемой потока (такая схема называется *объединенной*).

Символом $\aleph_i(\langle S, L, P \rangle)$ обозначается схема потока, скорректированная в соответствии со смещениями по y для i -го метода (необходимого методу на входе, чтобы производить поток $\langle S, L, P \rangle$ на выходе):

$$\aleph_i(\langle S, L, P \rangle) = \langle S - y_i^-, L + y_i^- + y_i^+, P - y_i^- - y_i^+ \rangle$$

Алгоритм состоит из двух шагов – прямого хода и обратного. На первом шаге происходит определение количества вершин в каждом ярусе и построение связей между ярусами исходя из равномерного последовательного разделения потока в каждом ярусе. На втором шаге происходит коррекция соединений и разметки с целью исключения каналов, по которым передаются очень малые относительно среднего потока на ярусе фрагменты.

3.2 Алгоритм построения вычислительной сети с равномерным статическим распределением потока

Вход: цепочка методов $\langle M_1, K, M_N \rangle$, относительные вычислительные сложности c_1, K, c_N , пропускная способность эталонного метода q_e , требуемая пропускная способность цепочки q .

Выход: Построенный граф вычислительной сети $G = \{V, E\}$, ребра и вершины которого размечены схемами потоков $M : E \rightarrow \{\langle S, L, P \rangle\}$, $M' : V \rightarrow \{\langle S, L, P \rangle\}$.

Шаг 1. Для каждого $i \in [1, N]$ определить ярус

$$G_i = \{V_i, \emptyset\}, \text{ где } V_i = \left\{ v_i^k \mid k = 1.. \left\lceil \frac{q_e}{q} c \right\rceil \right\} - \text{набор упорядоченных узлов методов, являющихся экземплярами метода } M_i.$$

Положить $V = \bigcup_{i \in [1, n]} V_i$; $E = \emptyset$. Приписать истоку схему

потока $\langle 0, L, 0 \rangle$, где L – исходная длина блока, задаваемая контроллером потока на этапе конфигурации. Положить разметку

$$M = \emptyset, M' = \left\{ \left(v_1^1, \langle 0, L, 0 \rangle \right) \right\}.$$

Для каждого узла v_i^k разметить его схемой

$$\left\langle \left\lfloor \frac{k-1}{|V_i|} \right\rfloor, \left\lfloor \frac{k}{|V_i|} \right\rfloor - \left\lfloor \frac{k+1}{|V_i|} \right\rfloor, L - \left\lfloor \frac{k}{|V_i|} \right\rfloor + \left\lfloor \frac{k+1}{|V_i|} \right\rfloor \right\rangle$$

Для каждой пары узлов v_i^k и $v_{i+1}^{k'}$, $i \in [1, N-1]$, $k \in [1, |V_i|]$, $k' \in [1, |V_{i+1}|]$ в случае если $M'(v_i^k) \cap \aleph_i(M'(v_{i+1}^{k'})) \neq \emptyset$ добавить дугу между узлами: $E = E \cup (v_i^k, v_{i+1}^{k'})$ и пометить ее схемой пересечения:

$$M = M \cup \left\{ \left((v_i^k, v_{i+1}^{k'}), M'(v_i^k) \cap \aleph_i(M'(v_{i+1}^{k'})) \right) \right\}.$$

Шаг 2. Для каждого яруса V_i , $i = N, \dots, 2$ для каждого узла $v_i^k \in V_i$, $k \in [1, |V_i|]$:

а) Проверить, что объединенная схема выходов совпадает со схемой узла:

$$\bigcup_{(v_i^k, v) \in E} \{M(v_i^k, v)\} = M'(v_i^k).$$

В случае, если это не так и объединенная схема включает схему узла, то вычислить множество недостающих схем

$$\bigcup_{(v_i^k, v) \in E} \{M(v_i^k, v)\} \setminus \aleph_i(M'(v_i^k))$$

и каждый недостающий элемент этого множества прибавить к одной из схем входящих ребер узла, имеющую схему смежную с недостающей схемой и наибольшую длину.

б) Удалить входные дуги, длина схемы которых меньше определенного порога и схемы этих дуг приписать дугам наибольшей длины, имеющим смежные схемы.

3.3 Динамическое распределения потока

При динамическом распределении потока между обрабатываемыми узлами, маршрут по которому пройдет конкретный фрагмент изображения определяется на стадии выполнения. При этом из-за смещений по высоте, для обработки данного фрагмента могут потребоваться несколько строк предыдущего или последующего фрагмента, непосредственно прилегающих к данному. А так как, заранее неизвестно какие именно фрагменты пройдут через данный экземпляр метода, невозможно сказать требуется ли сохранять данный фрагмент. Кроме того, встает проблема перемещения фрагментов в памяти – для того чтобы обеспечить обработку данного фрагмента с учетом смещений по высоте, требуется разместить предыдущий и последующий фрагменты непосредственно до и после него. Эти два фактора в совокупности усложняют процедуру управления памятью и требуют реализации дополнительной логики маршрутизации на этапе обработки потока, включающую коммуникацию с контроллером потока, координирующем в этом случае маршрутизацию. Это противоречит одной из основных концепций, положенных в основу проектирования системы, согласно которой взаимодействие контроллера потока с методами на этапе обработки должно быть минимальным. Однако обе проблемы могут быть решены с помощью избыточной передачи фрагментов, так, чтобы каждый фрагмент на каждом ярусе сети мог быть обработан независимо. В этом случае освободить память, занимаемую фрагментом, можно сразу после его обработки. Единственным и очевидным недостатком такого метода является повышение общего объема пересылаемых по сети фрагментов вследствие избыточной передачи строк, однако при больших значениях длины пересылаемых фрагментов и небольшом количестве ярусов эти издержки сравнительно малы.

Далее приводится алгоритм для построения сети с динамическим избыточным распределением потока. При этом в отличие от статического метода, где схема потока, приписанная конкретному узлу или дуге сети, определяет какие в точности фрагменты изображения и в каком порядке пройдут через данную дугу или узел, схема потока при динамическом распределении определяет какие фрагменты могут пройти через узел или дугу. Все узлы одного яруса являются равноправными в том смысле, что любой узел предыдущего яруса может передать обработанный фрагмент любому из узлов следующего яруса.

3.4 Алгоритм построения вычислительной сети с равномерным динамическим распределением потока с избыточными строками

Вход: цепочка методов $\langle M_1, K, M_N \rangle$, относительные вычислительные сложности c_1, K, c_N ,

пропускная способность эталонного метода q_e , требуемая пропускная способность цепочки q .

Выход: Построенный граф вычислительной сети $G = \{V, E\}$, ребра и вершины которого размечены схемами потоков

$$M : E \rightarrow \{\langle S, L, P \rangle\}, \quad M' : V \rightarrow \{\langle S, L, P \rangle\}$$

Для каждого $i \in [1, N]$ определить ярус

$$G_i = \{V_i, \emptyset\}, \text{ где } V_i = \left\{ v_i^k \mid k = 1.. \left\lceil \frac{q_e}{q} c \right\rceil \right\} - \text{набор упорядоченных узлов методов, являющихся экземплярами метода } M_i.$$

Положить $V = \bigcup_{i \in [1, n]} V_i$; $E = \emptyset$. Приписать стоку схему потока $\langle 0, L, 0 \rangle$, где L – исходная длина блока, задаваемая контроллером потока на этапе конфигурации. Положить разметку $M = \emptyset, M' = \{\langle v_N^1, \langle 0, L, 0 \rangle \rangle\}$,

Для каждого яруса $i \in [N-1, 1]$ разметить все его вершины схемами

$$\left\langle -\sum_{j=i}^N y_j^-, L + \sum_{j=i}^N (y_j^- + y_j^+), -\sum_{j=i}^N (y_j^- + y_j^+) \right\rangle$$

Для каждой пары узлов v_i^k и $v_{i+1}^{k'}$, $i \in [1, N-1]$, $k \in [1, |V_i|]$, $k' \in [1, |V_{i+1}|]$ добавить дугу между ними: $E = E \cup (v_i^k, v_{i+1}^{k'})$ и пометить ее схемой

$$\left\langle -\sum_{j=i+1}^N y_j^-, L + \sum_{j=i+1}^N (y_j^- + y_j^+), -\sum_{j=i+1}^N (y_j^- + y_j^+) \right\rangle$$

В построенной таким образом сети каждый узел каждого яруса соединен дугой с каждым узлом последующего яруса. При большом количестве узлов в соседних ярусах, это может привести к значительной дополнительной нагрузке на вычислительные и сетевые ресурсы кластера (каждая дуга соответствует двум дополнительным потокам в процессах отправляющего и принимающего узлов). Для снижения количества дуг можно использовать для вычисления множества дуг первый алгоритм, а для разметки дуг и узлов – второй.

Каждый узел может послать обработанный фрагмент любому из узлов следующего яруса. Простейшим способом маршрутизации обработанных фрагментов из данного узла является чередующаяся рассылка фрагментов последовательно каждому из узлов следующего яруса. При этом будет достигаться равномерное распределение загрузки между всеми узлами каждого яруса, в предположении, что каждый узел обладает равным количеством ресурсов.

4 Обсуждение и выводы

Подводя итог, можно выделить следующие преимущества и недостатки статического и динамического способа распределения потока:

Распределенная система обладает высокой отказоустойчивостью – при выходе из строя одного из узлов сети, его работу перераспределяется на остальные узлы того же яруса. Поэтому при динамическом распределении потоков неисправный узел будет заменен. При статическом же методе распределения, отказ узла приведет к появлению в выходном изображении необработанных участков.

Динамическое распределение дает возможность более гибкой балансировки загрузки узлов вычислительной сети. Если какой-то узел перегружен, его работу можно распределить на другие узлы того же яруса. Статическое распределение не позволяет этого делать.

При динамическом распределении суммарный объем пересылаемых данных возрастает, вследствие избыточной передачи строк. Появляется дополнительная нагрузка на узлы, связанная с маршрутизацией и балансировкой исходящих фрагментов.

По сравнению со статическим методом возрастает число потоков приема и передачи фрагментов и число сетевых соединений.

В целом преимущества динамического метода превосходят его недостатки, что делает его предпочтительным в дальнейшей разработке системы.

Литература

- [1] Гостев И.М. Графическая поисковая система GSS2001 // Сборник «Информационные технологии в системах вычислительной техники». Вып. 2. – М. : МГИЭМ, 2002.
- [2] Гостев И.М., Мирошкин А.В. Математическая модель одного класса поисковых систем // Вестн. РУДН. Сер. Прикладная и компьютерная математика. – 2004, Т. 3, №1. – С. 93–98.
- [3] Gostev I. M., Sevastianov L. A. About construction distributed calculation in real time system of image processing and pattern recognition // Proc Int. conf. «Distributed computation and GRID in science and education » Dubna JINR June 2004. – Pp.79–85.
- [4] Gostev I. M., Sevastianov L. A. Internal Architecture of Distributed Real-time System of Image Processing and Pattern Recognition // Proc. XX Int. Symp. on Nuclear Electronics & Computing (NEC 2005), Varna, Bulgaria, Sept. 2005.
- [5] Питерсон Дж. Теория сетей Петри и моделирование систем. – М. : Мир, 1984. – 264 с.
- [6] Мелихов А.Н. Ориентированные графы и конечные автоматы. – М. : Наука, 1971. – 416 с.

Information distribution algorithms for the real-time image search system

I.M. Gostev, A.G. Podgorbunsky

This paper deals with issues of distribution of video-flow used in the real-time distributed image search system. It introduces architecture based on graphic-flow controller and distributed network of nodes, applying image recognition and analysis methods to flow parts, static and dynamic methods of distribution of flow among these nodes, their benefits and drawbacks. The paper describes models of distribution and algorithms of distributed computational network building, compares them and discusses details of implementation.

Система поиска в электронной коллекции изображений петроглифов Карелии*

© А.А. Рогов,

К.А. Рогова,

К.Н. Спиридонов,

М.Ю. Быстров

Петрозаводский государственный университет

rogov@psu.karelia.ru, ksushar@mail.ru, spiridonov@psu.karelia.ru, maksimkab@yandex.ru

Аннотация

При создании любой коллекции электронных изображений возникает три задачи:

1. Приведение графических документов к единому стандарту, удобному для формирования электронной коллекции.
2. Формирование электронной коллекции графических документов.
3. Разработка комплекса программ по поиску в данной электронной коллекции графических документов.

Задача поиска по электронным коллекциям графических документов в общем случае является достаточно сложной. На практике, ее удается решить только для определенных классов изображений.

В данной статье описаны способы решения трех вышеперечисленных задач на примере создания коллекции электронных изображений петроглифов Карелии, реализованные в информационно-поисковой системе «PIRS». Эта система, обеспечив наглядность и доступность исторического материала, позволяет, наряду с традиционными, применять математические методы и компьютерные технологии при исследовании петроглифов.

Разработанные информационные методы анализа петроглифов и полученные с их помощью результаты позволяют использовать их при анализе петроглифов других регионов России и во всем мире. В настоящее время ведутся разработки по присоединению в систему «PIRS» информации о петроглифах Северной Фенноскандии, в первую очередь петроглифы Мурманской области (Канозеро) и Норвегии (Альта). Данная система рассчитана в первую очередь на исследователей наскальных изображений, а также будет интересна любому человеку, интересующемуся петроглифами. В

связи с тем, что в других публикациях авторами подробно рассмотрены разработанные ими методы, в данной статье сделана попытка общего описания системы со ссылкой на предыдущие работы.

1 Введение

Петроглифы Карелии были выбиты более 3 тысяч лет назад на скалах восточного побережья Онежского озера и берегах Белого моря [9]. Наскальные рисунки постоянно подвергаются таким природными воздействиями как: осадки, ветер, температурные колебания, зарастание мхом и т.д. В результате, некоторые петроглифы могут найти только специалисты. К перечисленным воздействиям добавились и случаи вандализма людей, стремящихся увековечить свое пребывание рядом с историческими памятниками. С целью сохранения информации об уникальном наследии и создания возможностей для его изучения была разработана информационно-поисковая система «PIRS». Информационно-поисковая система «PIRS» состоит из четырех блоков: электронной коллекции петроглифов, модуля подготовки изображений для данной коллекции, модуля онлайн-доступа и модуля локального доступа к электронной коллекции [8].

В качестве средства создания электронной коллекции была применена технология Microsoft Access (mdb), а сами изображения представляют собой файлы графических форматов BMP и JPEG.

Для разработки модулей системы использовалась среда Borland C++ Builder 6 (за исключением модуля онлайн-доступа).

2 Электронная коллекция петроглифов

Электронная коллекция петроглифов является основой системы. В нее входят графитные копии, фотографии, карты, черно-белые схемы, характеристики и текстовые описания петроглифов. Выделены: группы петроглифов по местонахождению, сюжетные группы и отдельные петроглифы. Описание сюжетных групп содержит: название, кодовый номер, занимаемая площадь, карта местности, текстовая информация, фотографии и графитные копии (не менее трех для каждой). Каждый петроглиф в группе имеет собственное описание, включающее следующую информацию: кодовый номер, назва-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

ние, высота над уровнем моря, глубина вырезки, обрастание мхом, сохранность, угол поворота, фотографии и графитные копии каждого петроглифа (более 10 для каждого).

Электронная коллекция графических документов зарегистрирована в государственном регистре баз данных (ФГУП Научно-технический центр «Информрегистр», №0220611248).

3 Модуль подготовки изображений для электронной коллекции петроглифов

Модуль подготовки изображений для электронной коллекции петроглифов состоит из подпрограмм, реализующих решение следующих задач: выделение отдельных изображений из группы, сегментация изображений, приведение изображений к единому стандарту, нанесение защитных надписей на фотографии.

3.1 Сегментация

Сегментация изображений [3] производится с использованием двух методов: порогового [5] и мультифрактальной параметризации структур. Следует отметить, что эти методы не являются взаимоисключающими, а выгодно дополняют друг друга. На входе и выходе – графические файлы. Пороговый метод сегментации работает следующим образом: по рисунку определяется набор эталонных цветов (векторов RGB) фона и петроглифа. Для каждого пикселя, определяется к какому из эталонных векторов входной вектор ближе всего. Если он ближе к одному из цветов фона, то он перекрашивается в белый цвет, иначе – в черный цвет. После этого производится фильтрация получившегося черно-белого изображения.

Точность алгоритмов при сегментации текстурной мозаики, состоящей из двух текстур Бродаца [1] размером 512×512 , составляет порядка 95%.

Применение данных методов при работе с электронной коллекцией графических документов петроглифов Карелии показало следующие результаты: точность сегментации – в среднем 90% (при удачном выборе эталонов), время работы алгоритма – 1 с для изображения размерностью 100 килопикселей и глубиной цвета 24 бита.

3.2 Мультифрактальная параметризация структур

В методе мультифрактальной параметризации структур одной из основных характеристик изображения является спектр фрактальных размерностей Реньи D_q , где $q \in R$ [11]. Для численного расчета таких спектров была использована методология мультифрактальной параметризации структур, первый вариант которой был разработан в лаборатории прочности металлических материалов ИМЕН РАН в 1993 [2]. В качестве меры отличия μ двух спектров D_q и используется следующее выражение:

$$M = \sum_q |D_q - D'_q| \quad (1)$$

Функционирование данной подпрограммы включает следующую последовательность шагов:

- подача изображения на вход программы;
- предложение пользователю выделить область петроглифа S_{petr} и область скалы S_{rock} ;
- автоматическое разбиение каждой из областей S_{petr} и S_{rock} на квадратные фрагменты размером $n \times n$ пикселей;
- расчет спектра Реньи D_q для каждого квадратного фрагмента и $q = 0, 1, 2, 3, 4, 5, 10, 20, 30, 40$;
- отделение пикселей, относящихся к изображению петроглифа от остальных, относящихся к фону т.е. скале.

При настройке и анализе работы данной подпрограммы была выявлена зависимость качества сегментации от размеров квадратных фрагментов. В результате серии экспериментов выявлены следующие свойства спектров D_q :

- для положительных значений q , среднее значение D_q слабо зависит от размеров квадратных фрагментов, по которым идет анализ;
- при увеличении размера квадратных фрагментов, величина стандартного отклонения значений D_q для $q \geq 0$ убывает;
- минимальный размер фрагментов, при котором наблюдается приемлемая точность расчетов спектров D_q , равен 12×12 пикселей;
- оптимальный размер фрагментов равен 30×30 пикселей.

3.3 Выделение отдельных изображений

Для подпрограммы выделения отдельных изображений на вход подается черно-белая копия группы, а на выходе – графические файлы с отдельными изображениями.

При этом возникает следующая проблема: если разные петроглифы соединены линией, то они воспринимаются как одно изображение. В этом случае необходимо разделение вручную. Заметим, что таких изображений не больше 5%. Программа показала свою работоспособность на изображениях размером до 30 мегапикселей.

3.4 Нанесение защитных надписей

Нанесение защитных надписей на фотографиях, представленных на сайте, необходимо для защиты авторских прав. В качестве защиты выбрана надпись адреса сайта по центру фотографии.

4 Модуль онлайн-доступа

Модуль онлайн-доступа реализован в виде раздела «Каталог» сайта «Петроглифы Карелии». Поиск происходит по признаку «местонахождение петроглифа» путем перемещения с помощью гиперссылок по картам.

Кроме того, сайт «Петроглифы Карелии» содержит следующие разделы:

- «Информация» включает в себя общую информацию о петроглифах Карелии. Отдельно описано каждое скопление изображений в каждой стране (это описание включает место расположения петроглифов, познавательную информацию, петроглифы как объект международного сотрудничества, состояние, бережение и использование петроглифов в научных исследованиях, туризме и образовательных целях).
- «Исследователи» предназначен для ознакомления посетителей сайта с теми, кто причастен к открытию и изучению петроглифов. Кратко представлена биография К.И. Гревингга, Г. Халлстрема, А.М. Линевского, А.Я. Брюсова, В.И. Равдоникаса, К.Д. Лаушкина, А.Д. Столяра, Р.Б. Климова, Ю.А. Савватеева, Н.В. Лобановой.
- «Публикации» содержат статьи и ссылки о петроглифах.
- «Новости» отражают свежую информацию о конференциях, встречах, проектах, открытиях исследователями новых петроглифов.
- «Фото и видео» – материалы, которые пополняются периодически и являются визуальным дополнением информации, представленной в других разделах.
- «Математические методы» – характеризует применение данных методов и технологий в истории (на примере изучения петроглифов).
- «Форум» необходим для поддержания обратной связи пользователей и разработчиков, а также для обеспечения возможности интерактивного общения между заинтересованными лицами.

В настоящий момент на сайте находится более 500 фотографий групп петроглифов с их описаниями. Общее количество представленных петроглифов превышает 2000 фигур. Адрес сайта: <http://smalt.karelia.ru/~petroglyphs>. При желании пользователь может получить доступ к ресурсу через мобильные устройства. Для этого необходимо всего лишь иметь возможность выхода в интернет и установленный на мобильном устройстве HTTP браузер.

5 Модуль локального доступа к электронной коллекции петроглифов

Модуль локального доступа к электронной коллекции петроглифов имеет больше функциональных возможностей по сравнению с онлайн версией и

состоит из следующих подпрограмм: поиск по признакам, поиск по изображениям, классификация по признакам, классификация по схожести текстуры изображений [6, 7]. Данный инструментарий позволяет упростить работу исследователя при анализе изображений петроглифов.

5.1 Поиск по признакам изображений

Рассмотрим на примере петроглифов лосей и оленей Беломорья. Было выделено 16 признаков: тип головы, тип корпуса, наличие/отсутствие холки, изгиб передней и задней пар ног и т.д. Все признаки проанализированы на статистическую зависимость с помощью критерия Пирсона. Были найдены зависимые признаки, например, изгиб спины и наличие/отсутствие холки, тип головы и серьга на шее и т.д. Общая схема работы с данным поиском следующая: пользователю предлагается выбрать значения признаков и точность поиска (количество совпадений признаков). После выбора условий поиска будут найдены петроглифы, соответствующие выделенным критериям.

Пользователю доступны изображения найденных петроглифов и полная информация о них.

5.2 Классификация по признакам

Из базы данных пользователю предлагается выбрать петроглифы, которые требуется разбить на определенные группы, а также набор признаков. При помощи иерархического кластерного анализа (метод ближайшего соседа), все выделенные петроглифы разбиваются на группы.

5.3 Классификация по схожести текстуры графического изображения

Классификация основана на гипотезе о том, что разные люди использовали разную технику выбивания петроглифов. Это позволило поставить задачу о проведении классификации петроглифов по степени близости (различию их текстур). Для решения этой задачи необходимо было найти характеристику поверхности скалы являющуюся инвариантом петроглифа, определенного по технике его выбивки. В качестве такой характеристики был взят описанный выше спектр фрактальных размерностей Реньи [10].

Заметим, что инвариант графического изображения должен сохранять свои значения:

- при изменении яркости одного и того же изображения;
- при повороте изображения;
- при масштабировании изображения.

Была проведена серия экспериментов на предмет наличия данных свойств у спектров Реньи D_q . В качестве исходных изображений были взяты текстуры из альбома Бродаца [1] и фрагменты изображений петроглифов. Опишем эксперименты более подробно.

1. Свойства спектров D_q при изменении яркости одного и того же изображения.

Для проверки изменения значений спектра D_q при изменении яркости изображения необходимо иметь набор одного и того же изображения с разными значениями яркости.

Уровень яркости исходных изображений брался за нулевой. Было установлено, что уменьшению разброса значений спектров способствует предварительная обработка изображений, в частности – γ -коррекция изображения [4].

Для каждого изображения и для различных его вариаций уровней яркости после γ -коррекции рассчитывался спектр фрактальных размерностей Реньи D_q .

Было установлено, что величина вариации значений D_q анализируемых изображений при данном преобразовании изменяется в пределах от 0.029 до 0.2 для $q = -40$ и от 0.008 до 0.027 для $q = 40$. Проведенные исследования показали, что яркостные преобразования несущественно влияют на значения спектров Реньи D_q .

2. Свойства спектров D_q при повороте одного и того же изображения.

Было установлено, что величина вариации значений D_q анализируемых изображений при данном преобразовании изменяется в пределах от 0.005 до 0.018 для $q = -40$ и от 0.0025 до 0.022 для $q = 40$. Проведенные исследования показали, что преобразование поворота несущественно влияет на значения спектров Реньи D_q .

3. Свойства спектров D_q при масштабировании одного и того же изображения.

Масштаб исходных изображений приняли за 100%. Далее в графическом редакторе размер исходных изображений уменьшили, получая копии текстуры масштабам 80%, 40%, 20%. Величина вариации значений D_q анализируемых изображений при данном преобразовании изменяется в пределах от 0.006 до 0.2 для $q = -40$ и от 0.0003 до 0.033 для $q = 40$. Проведенные исследования показали, что преобразование масштаба несущественно влияет на значения спектров Реньи D_q .

Кроме того, были выявлены дополнительные свойства спектров фрактальных размерностей Реньи:

1. Различие значений D_q для фрагментов между областями петроглифа и скалы на одном фотоснимке.
2. Близость значений D_q для фрагментов из области петроглифа на различных фотоснимках одного и того же петроглифа.
3. Различие значений D_q для фрагментов из областей петроглифа различных петроглифов.

4. Слабая зависимость средней величины D_q для определенных q от размеров фрагментов.

Проведенные эксперименты показали, что спектр фрактальных размерностей Реньи можно использовать для классификации изображений по схожести текстур. Программу, реализованную на основе фрактальной размерности Реньи, можно описать следующей последовательностью шагов:

- пользователем из базы данных выбирается набор изображений петроглифов, классификацию которых необходимо провести;
- для каждого петроглифа рассчитывается спектр;
- на основе введенной меры различия (1) методом иерархического кластерного анализа производится классификация изображений петроглифов.

В результате работы программы было выявлено, что при классификации изображений с точностью до уровня яркости ошибка классификации составляет порядка 7-11% от общего количества изображений, с точностью до поворота – 6-9%, с точностью до масштабирования – 7-12%.

5.4 Поиск по изображениям

Поиск по изображениям предназначен для поиска изображений, похожих на данное или на его фрагмент. На вход подается исследуемое изображение, а на выходе должны появиться изображения из базы данных, наиболее похожие на исходное.

На сегодняшний день все новейшие материалы по петроглифам Карелии представляют собой набор цветных фотографий. Определенную сложность поиска создает фактическое отсутствие некоторых частей изображения. Поиск также осложняется тем, что часто невозможно определить, где верх, а где низ изображения. При этом, требование, что при поиске необходимо только совпадение контура изображения, позволяет упростить поиск, а значит, изображение петроглифа можно рассматривать, как бинарное (скале соответствует белый цвет, а петроглифу – черный). В зависимости от выбранных параметров поиска (точность поиска, процент совпадений элементов изображений) будет найдено одно или несколько изображений. Для поиска используются сеть адаптивного резонанса и структурный метод поиска. В настоящее время заканчивается разработка двухуровневого метода поиска.

В результате поиска, пользователю предоставляется доступ к информации о кодовом номере, месторасположении, характерных признаках найденного петроглифа и петроглифах, близких к нему по ранее описанным признакам.

5.4.1 Структурный метод поиска

Структурный метод поиска основан на выделении скелетов изображений и их сравнении. Скелетом изображения является ломаная линия, аппроксимирующая множество всех центров вписанных в

фигуру окружностей максимального радиуса. Использование структурного метода поиска петроглифа осуществляется при помощи следующего алгоритма:

1. выделение скелета изображения;
2. кусочно-линейная аппроксимация скелета с заданным параметром близости;
3. перевод кусочно-линейной аппроксимации в последовательность символов – цепочку примитивов, представляющих собой последовательность цифр, разделенных пробелом, обозначающих длины отрезков скелетона, углов между ними. Начиная с крайней левой точки, происходит обход всего скелетона по контуру против часовой стрелки;
4. анализ кодов рисунков на совпадение кодов или их фрагментов с учетом правил их преобразования.

Две цепочки примитивов считаются равными, если существует такая циклическая перестановка, когда с определенной точностью совпадут углы, а длины будут пропорциональны. Сравнение цепочек примитивов происходит при помощи стандартного алгоритма сравнения строк. Сложность сравнения заключается в том, что цепочки являются замкнутыми.

5.4.2 Двухуровневый метод поиска

Ускорение поиска при увеличении объема базы данных потребовало разработку новых методов. В настоящее время разрабатывается двухуровневый поиск, который осуществляется следующим образом. Вначале применяется быстрый, но неточный способ отбора похожих изображений, сужающих зону поиска. Затем используется точный, но медленный метод для окончательного отбора изображений. Для быстрого первичного отбора используется следующий алгоритм. Вокруг петроглифа описывается окружность минимального радиуса, затем с шагом в 10 градусов проводятся радиусы, на которых определяются отрезки, соединяющие дугу окружности и границу петроглифа. Длина каждого отрезка делится на длину радиуса. Последовательность полученных чисел называется кодом петроглифа. Изображения считаются близкими, если существует циклическая перестановка кода петроглифа, такая что расстояния между кодами петроглифов меньше заданного числа. Для окончательного отбора была применена нейронная сеть Хемминга и сеть адаптивного резонанса.

5.5 Определение зон видимости изображений

Изучая области видимости наскального рисунка можно ответить на вопрос о целевой аудитории автора, не только распознать изображение петроглифа, но и делать выводы о символическом значении данного образа для древних людей, понимать особенности быта и образа жизни. Эта задача решалась за несколько этапов:

1. Анализ видимости объекта: выделение параметров, влияющих на видимость, построение математической модели, функции от известных параметров для определения видимости.

2. Оцифровка топографических карт: ввод карты, получение матрицы высот, анализ точности результатов.

3. Применение функции видимости для нахождения на карте областей видимости петроглифов, зная размеры и местоположение последних.

Для статических объектов параметры яркости и контрастности являются постоянными и определяются экспериментальным путем на местности. Определяющим видом параметром становится угол наблюдения, который прямо зависит от размеров и расстояния до объекта. При расчетах дальности требуется, чтобы угловые размеры наблюдаемого объекта превышали остроту зрения нормального человека. Поэтому, задав минимальный угол зрения α по формуле (2) можно вычислить максимальное расстояние x до объекта пока он будет видимым.

$$\alpha = \arctg\left(\frac{H-h}{x}\right) - \arctg\left(\frac{H-v}{x+r}\right), \quad (2)$$

где H – высота точки наблюдения (рост человека с учетом высоты точки стояния),

h – высота нижнего края петроглифа,

v – высота верхнего края петроглифа,

r – расстояние между точками петроглифа.

На данный момент создана функциональная модель видимости, создана библиотека функций для определения угла зрения при заданных параметрах изображения, позиции наблюдателя и расстояния от него до объекта, и определения максимально возможного расстояния, на котором сохраняется видимость объекта. Реализовано проверочное приложение для определения угла наблюдения.

В ходе экспедиции к беломорским петроглифам был произведен ряд наблюдений для определения эмпирических значений угла видимости не закрашенных петроглифов. Приняв для расчета среднее значение расстояния видимости, ландшафт местности и минимальный линейный размер деталей петроглифа был определен необходимый угол зрения для наблюдения петроглифа. Для восприятия изображения необходимо чтобы расстояние до объекта не превышало размеры объекта более чем в 150 раз. Расчеты показывают, что при этом угловые размеры объекта составляют 23 угловых минуты. Однако эксперименты показывают, что при солнечном освещении для наблюдения петроглифа достаточно угла в 4-5 угловых минуты.

Оцифровка топографических карт проводится в автоматизированном режиме. Ввиду небольшого количества карт, требуемых оцифровки, не ставилась задача полностью автоматизировать данный процесс.

Непосредственно к этой задаче примыкает задача оценки численности группы людей, одновременно видящих данный петроглиф или сюжетную группу. Она решается методом имитационного моделирования. В качестве параметров задается средний рост взрослых и детей, соотношение их количества и т.д.

6 Заключение

Разработанные информационные методы анализа петроглифов и полученные с их помощью результаты позволяют использовать их при анализе петроглифов других регионов России и во всем мире. В настоящее время ведутся разработки по присоединению в систему «PIRS» информации о петроглифах Северной Фенноскандии, в первую очередь петроглифы Мурманской области (Канозеро, более 1000 изображений) и Норвегии (Альта, более 5000 изображений). Данная система рассчитана в первую очередь на исследователей наскальных изображений, а также будет интересна любому человеку, интересующемуся петроглифами. Число пользователей не ограничено.

7 Литература

- [1] USC-SIPI Image Database – <http://sipi.usc.edu/database/>
- [2] Встовский Г.В., Колмаков А.Г. и др., Введение в мультифрактальную параметризацию структур материалов. – М. : R&C Dynamics, 2001. – 115 с.
- [3] Верденская Н. В. Сегментация изображений – статистические модели и методы // Зарубежная радиоэлектроника. Успехи современной радиоэлектроники, 2002. – №12.
- [4] Гонсалес Р., Вудс Р. Цифровая обработка изображений. – М. : Техносфера, 2005. – 1072 с.
- [5] Грузман И.С., Киричук В.С. и др. Цифровая обработка изображений в информационных системах : Учеб. пособие. – Новосибирск : Изд-во НГТУ, 2000. – 168 с.
- [6] Рогов А.А. Применение методов теории распознавания для анализа петроглифов Карелии / Материалы IX международной конференции «Интеллектуальные системы и компьютерные науки», Т. 2, ч. 2. – М. : Изд-во ЦПИ при механико-математическом факультете МГУ, 2006. – С. 260–262.
- [7] Рогова К.А., Быстров М.Ю. Задачи анализа изображений в информационно-поисковой системе PIRS // Материалы XIII Всероссийской конференции «Математические методы в распознавании образов». – М. : МАКС Пресс, 2007. – С. 528–530.
- [8] Рогова К.А. Информационно-поисковая система «Петроглифы Карелии» //Интеллектуальные системы и компьютерные науки. – М. : Изд-во ЦПИ при мех.-мат. факультете МГУ, 2006. – С. 262–264.
- [9] Саватеев Ю.А. Залавруга. – Л. : Наука, 1970. – 250 с.
- [10] Спиридонов К.Н. К вопросу об инварианте графического изображения // Материалы XIII Всероссийской конференции «Математические методы в распознавании образов». – М. : МАКС Пресс, 2007. – С. 393–396.
- [11] Спиридонов К.Н. Применение мультифрактального анализа при изучении петроглифов Карелии // Материалы IX международной конфе-

ренции «Интеллектуальные системы и компьютерные науки». – М. : Изд-во ЦПИ при механико-математическом факультете МГУ, 2006. – Т. 2, ч. 2. –С. 278–280.

The Search Systems in Electronic Collection of Karelian Petroglyphs Images

A.A. Rogov, K.A. Rogova, K.N. Spiridonov,
M.Yu. Bystrov

While creating any electronic collection of images, three tasks should be solved:

1. Graphic images reduction to common standard. This standard should be convenient to use for forming a database.
2. Creating a database of graphic images.
3. Developing a complex of programs for searching in electronic document collection.

The article describes the ways of solving the above-listed tasks by the example of the creation of electronic images collection of Karelian petroglyphs, that are realized in the information retrieval system “PIRS”. This system provides obviousness and availability of historical materials and allows, along with traditional methods, using mathematical methods and computer technologies for the investigation of petroglyphs.

The developed informational methods of petroglyphs’ analysis and results achieved by these methods can be applied for the analysis of petroglyphs from other regions of Russia and all over the world. Nowadays information about Northern Fennoscandia petroglyphs is being added to the system “PIRS”, first of all, petroglyphs from Murmansk region (Kanozero) and Norway (Alta).

The system was created mostly for researchers of rock art, but would be also interesting for any person who is interested in petroglyphs.

♦ Данная научная разработка поддержана грантами РГНФ №05-01-12118в, № 08-01-1211в (руководитель Н.В. Лобанова).

Проект электронной библиотеки методик и результатов исследований текстовых коллекций для системы «Источник»*

© Н.В. Каргинова, И.В. Кравцов, Н.Д. Москин, А.Г. Варфоломеев

Петрозаводский государственный университет
avarf@psu.karelia.ru

Аннотация

Данная статья посвящена применению формата записи произвольных правил RuleML для сохранения методик и результатов исследований в цифровых библиотеках исторических документов. Предлагается формализация процесса исследования текстовых коллекций: сначала в текстах выделяются логические или семантические фрагменты, затем из них образуются структуры текстов в виде векторов или графов, эти структуры разбиваются на типы, и делаются выводы о зависимости типов структур от других характеристик текстов. Эти выводы записываются в виде правил, пригодных для машинной обработки, и сохраняются в библиотеке.

Рассмотренный подход используется авторами на практике в разработке информационной системы «Источник», предназначенной для организации работы сетевых сообществ исследователей текстовых исторических источников.

1 Введение

С развитием компьютерных технологий все чаще исследования по истории и лингвистике опираются на большие коллекции текстовых документов. Такие коллекции, представленные в Интернете, составляют основу для формирования сетевых сообществ исследователей, разделяющих между собой форматы представления данных (TEI [20], MEF [15], CEI [13]), а иногда и сами тексты для совместного изучения и редактирования («Манускрипт» [8], Monasterium [16], TextGrid [21]). Естественным выглядит следующий шаг – предоставить в распоряжение сообщества не только данные и инструменты, но и описания проведенных ранее исследований. Казалось бы, для этого достаточно научных публикаций в их традиционной форме. Однако традици-

онные публикации не являются машиночитаемыми, с их помощью невозможно строить базу знаний сообщества. Хотелось бы дополнить такие публикации некими формализованными описаниями проведенных исследований, на основе которых компьютер не только мог бы производить поиск подходящих методик и результатов, но и выдвигать собственные гипотезы.

Технологии для достижения поставленной цели существуют. В рамках современного направления Semantic Web разрабатываются стандарты представления бизнес-правил [19], в виде которых могут быть записаны и научные выводы, гипотезы, формулы или алгоритмы. Примерами таких стандартов являются язык PMML [18], служащий для записи регрессионных и других предиктивных моделей анализа данных, и форматы группы MKM [22] для обмена математическими результатами. Однако в историко-филологических исследованиях текстовых источников формализация методик и результатов в виде бизнес-правил также может быть осуществлена. К сожалению, пока нельзя привести ни одного примера электронных коллекций текстов, предоставляющих подобную информацию. На наш взгляд, причиной является то, что пока подавляющее большинство текстовых коллекций ориентированы на предоставление данных для локальной работы индивидуальных исследователей, и не предусматривают обмен формализованными методиками и результатами в рамках сетевых сообществ.

Задача разработки формата представления методик и результатов исследований текстовых коллекций была поставлена в [11] в рамках проекта информационной системы «Источник» [7]. Разработка этой системы, ориентированной на поддержку сетевого научного сообщества исследователей текстов, отражена в публикациях [4, 9, 11]. В качестве особенностей данной системы следует указать ее многофункциональность, а также последовательное применение технологий XML на разных этапах работы с текстом.

В данной статье мы предлагаем проект библиотеки методик и результатов исследований текстовых коллекций для системы «Источник» с возможностью публикации результатов в машиночитаемой форме с помощью языка RuleML [23].

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

2 Информационная система «Источник»

2.1 История развития системы

Первоначально информационная система «Источник» возникла как сетевой сервис для подготовки печатных изданий средневековых исторических документов, прежде всего комплекса «Moscowitica-Ruthenica» [5]. Для реализации была выбрана технология XML. Однако текст, один раз размеченный, мог быть повторно использован, причем не только для публикации, но и для анализа его структуры и содержания. Так возникла идея разрабатывать многофункциональный сетевой сервис, включающий в себя различные инструменты анализа текста [11]. Одним из важных способов анализа является визуализация текстовой информации, для этого в системе «Источник» было решено широко применять SVG-графику [4].

Другим важным направлением развития системы «Источник» было превращение ее в систему поддержки работы сетевого сообщества исследователей произвольных текстов с возможностью накопления размеченных текстов и отчетов о проведенных исследованиях в виде своеобразной «базы знаний» сообщества [9,11]. Но описания методов и результатов исследований, хоть и были структурированы, не имели вид правил «если-то» и не могли быть использованы для автоматического порождения гипотез и получения новых знаний. Настоящая статья как раз направлена на решение проблемы записи методик и результатов в виде правил вывода.

2.2 Анализ текстовых коллекций

Работая в информационной системе «Источник», исследователь размечает тексты коллекции с помощью одной или нескольких XML-разметок. Затем, изучая размеченные тексты, исследователь может делать выводы об обнаруженных закономерностях.

Например, в ходе формулярного анализа средневековых грамот [6] исследователь вручную выделяет в текстах устойчивые смысловые фрагменты, такие как «приветствие», «описание ситуации», «ссылка на предыдущий документ» и сравнивает последовательности таких фрагментов для разных текстов, делая выводы о наличии норм и классифицируя документы по их «формулярам». В контент-анализе массовых источников [1] исследователь тоже выделяет фрагменты текста – смысловые единицы на основании присутствия в них словосочетаний-«индикаторов», указывающих на некие «категории анализа», затем подсчитывает количество или объем смысловых единиц, классифицируя документы по степени отражения в них тех или иных категорий. При структурном анализе фольклорных песен в текстах выделяются устойчивые фрагменты (мотивы), упоминания объектов (например, основных персонажей, животных, предметов обихода и т.д.), и глаголы (или отглагольные формы), указывающие на отношения между объектами [3].

Для организации работы исследователей с методиками и результатами информационная система, на наш взгляд, должна обладать следующими функциями:

- предоставлять пользователю интерфейс для разметки текстов, записи правил разметки и правил вывода (методики исследования), а также самих выводов (результатов исследования);
- предоставлять сетевому сообществу возможность накапливания не только коллекций размеченных текстов, но также библиотеки правил и выводов как основы базы знаний сообщества;
- обеспечивать возможность поиска, просмотра и сравнения методик и результатов других исследователей;
- предлагать пользователю ознакомиться с гипотезами, выведенными системой автоматически на основе имеющихся данных и правил;
- предоставлять возможность публикации методик и результатов исследований в машиночитаемой форме для того, чтобы они также становились доступными для произвольных семантических сервисов в сети.

Возможны различные варианты автоматической генерации гипотез в зависимости от полученных результатов исследователя и уже имеющихся результатов других исследователей. Например, если исследователь получает какой-либо вывод на основании определенного правила, то система может попробовать применить это правило к другим похожим коллекциям и предложить сравнить результаты. Если есть несколько исследований с похожими целями или методиками, проведенных на одной коллекции, система может самостоятельно сравнить полученные результаты и сделать выводы об их сходстве или различии. Если проводится повторное исследование коллекции или текста, система может сообщить все полученные ранее результаты, чтобы исследователь мог их подтвердить или опровергнуть.

3 Формализация процесса исследования

Для успешного решения задачи формализованной записи методик и результатов исследований необходимо предложить некоторую обобщенную схему, формализующую сам процесс исследования текстовых коллекций. На наш взгляд, процесс исследования можно представить в виде следующей последовательности этапов.

Исследователь разбивает текст на логические или семантические фрагменты. При этом он может выбрать разметку из уже имеющихся в системе вариантов или предложить свою. В последнем случае исследователь должен описать принципы, положенные в основу разметки, для того, чтобы ее в дальнейшем могли использовать другие исследователи.

Для представления разбиения текста в системе «Источник» используется язык XML. Для хранения синтаксических правил разметки используются форматы DTD или XSD. На данном этапе также могут фиксироваться некоторые исследовательские

правила, с помощью которых производится разбиение, чтобы в дальнейшем можно было частично автоматизировать этот процесс. Например, если в тексте уже присутствует какая-то разметка, в частности, на абзацы, то наличие в абзаце слова-индикатора может служить основанием для преобразования данного абзаца в логический фрагмент определенного типа. Такое преобразование может быть записано с помощью языка XSLT.

После того, как текст разбит на логические фрагменты (блоки), его структуру необходимо представить в виде некоторого объекта, с которым будет в дальнейшем работать система. В самом простом варианте это может быть список (вектор), в котором каждый элемент будет являться идентификатором соответствующего блока в разметке. Можно представить структуру теста в виде следующей таблицы, где номер – это порядковый номер блока в структуре текста, тип блока – значение из заданного набора типов блоков выбранной разметки, характеристики блока могут использоваться для хранения дополнительной информации о том фрагменте текста, который относится к данному блоку. В качестве характеристик могут выступать объем фрагмента текста, вычисленный в словах или символах, или процентное соотношение объема блока к общему объему текста.

№	Тип блока	Характеристика блока
1	A	20
2	B	25
...	...	...

Приведенная таблица адекватно передает линейную структуру текста, используемую, например, в формулярном анализе или в контент-анализе. В случае использования более сложной иерархической или графовой структуры, как, например, в анализе фольклорных песен, таблица представления результатов тоже должна быть усложнена.

После представления структуры текста в выбранном стандартном виде можно проводить анализ при помощи правил вывода.

Правила в системе делятся на заданные изначально и формируемые в ходе работы исследователей. Заданные изначально правила обеспечивают анализ исследований, получение новых выводов и применение уже существующих закономерностей к текущему исследованию. Также изначально может быть задан набор фактов, устанавливающих, например, что какая-то структура является общей для какого-либо набора текстов или что две структуры похожи друг на друга:

*структура 1 – общая для структур
(структура 2, ..., структура n)*

структура 1 похожа на структуру 2 на n %

Некоторые факты могут быть представлены в виде функций (алгоритмов). Например, функция *compare_structure* с двумя аргументами – структурами возвращает их сходство в процентах.

compare_structure(структура 1, структура 2) = n %

К заданным изначально правилам также относятся шаблоны, с помощью которых могут записываться и новые правила, полученные в ходе исследования. Ниже приведены примеры шаблонов.

Шаблон 1

ЕСЛИ <i>текст 1 — структура 1</i> <i>текст 2 — структура 2</i> <i>текст n — структура n</i> <i>(текст 1, ..., текст n) – тип 1</i> <i>структура k – общая для структур (структура 1, ..., структура n)</i>
ТО <i>структура k – общая структура для типа 1</i>

В соответствии с данным шаблоном, если для каждого текста из набора была получена определенная структура, и известно, что все тексты относятся к одному типу, то можно сделать предположение об общей структуре для текстов данного типа.

Шаблон 2

ЕСЛИ <i>текст 1 — структура 1,</i> <i>текст 2 — структура 2,</i> <i>текст n — структура n,</i> <i>(текст 2, ..., текст n) имеют тип 1,</i> <i>(структура 1, структура 2..., структура n)</i> <i>похожи на m%,</i> <i>m больше, чем пороговое значение</i>
ТО <i>текст 1 имеет тип 1</i>

В соответствии с этим шаблоном, если имеется несколько текстов достаточно близкой структуры, и про часть из них известно, что они относятся к определенному типу, можно предположить, что оставшиеся тексты также относятся к этому типу. Следует отметить, что шаблоны могут отражать вовсе не только дедуктивные правила вывода, но, как в указанных примерах, «правдоподобные» рассуждения, основанные на индукции и аналогии [2].

Правила, формирующиеся в ходе работы исследователей, представляют собой формализованную запись выводов исследователей. Они могут формироваться как конкретизация имеющихся в системе шаблонов. В таком случае на место свойств, типов и структур подставляются конкретные значения. Эти правила могут быть сохранены в библиотеке системы и использованы в дальнейших исследованиях.

Для каждого правила должны задаваться область видимости и коэффициент доверия. По умолчанию исследователь видит только свои правила, но при желании может ознакомиться с правилами других исследователей. Если в ходе исследования какое-то

уже сохраненное в системе правило подтверждается или опровергается, это может вызвать изменение его коэффициента доверия. Таким образом, в дальнейшем исследователи могут выбирать наиболее подтвержденные правила.

4 Хранение правил

Рассмотренные выше правила могут быть представлены в системе в каком-либо внутреннем формате для ускорения процессов поиска, сравнения и вывода. Но при публикации результатов необходимо, чтобы правила были понятны не только данной системе, но и другим системам, ориентированным на работу с текстами. Для этого необходимо представить тексты, разметки и правила с помощью стандартных форматов, позволяющих сохранить семантику. Также, так как система «Источник» является web-ориентированной, чтобы обеспечить работу исследователей из разных мест, необходимо использовать стандарты, позволяющие представить информацию в Интернете.

Для записи правил в библиотеке методик и результатов системы «Источник» предлагается использовать активно разрабатываемый в настоящее время стандарт RuleML.

RuleML (Rule Markup Language) [23] представляет собой язык разметки для описания правил. С помощью данного языка можно публиковать и обмениваться правилами, созданными в рамках разных систем и задач. Он включает в себя такие языки, как PMML (для предиктивных моделей), MathML (для записи функциональных зависимостей), XSLT (для правил преобразований из XML в XML). Но ядром RuleML является XML-вариант языка Datalog. Он позволяет записывать факты и правила в форме продукций.

Пример описания шаблона 2 на языке RuleML.

```
<Implies>
  <head>
    <Atom>
      <Var>text 2</Var>
      <Rel>type of text</Rel>
      <Var>type 1</Var>
    </Atom>
  </head>
  <body>
    <Atom>
      <Var>text 1</Var>
      <Rel>type of text</Rel>
      <Var>type 1</Var>
    </Atom>
    <Atom>
      <Var>text 1</Var>
      <Rel>structure of text</Rel>
      <Var>structure 1</Var>
    </Atom>
    <Atom>
      <Var>text 2</Var>
      <Rel>structure of text</Rel>
      <Var>structure 2</Var>
    </Atom>
  </body>
</Implies>
```

```
</Atom>
<Atom>
  <Var>structure 1</Var>
  <Rel>is -like</Rel>
  <Var>structure 2</Var>
  <Ind>m %</Ind>
</Atom>
</body>
</Implies>
```

5 Примеры записи результатов анализа фольклорных текстов

Теперь рассмотрим, как можно применить правила, записанные на языке RuleML, для описания закономерностей, полученных в результате анализа фольклорных песен на основе их теоретико-графовых моделей [3]. Одним из способов анализа таких моделей является сравнение их агрегирующих графов с небольшим числом вершин и ребер. Эти графы позволяют выделить структуру основных отношений в тексте, отбросив несущественные объекты и связи. Поскольку в общем виде задача агрегации трудно решается, данные графы были построены при помощи метода, предложенного в работе [10]. В нем накладывается следующее ограничение: разбиение объектов осуществляется на непересекающиеся группы, объединение которых дает исходное множество объектов. Число групп должно быть установлено заранее – в нашем случае оно, как правило, определялось по числу мотивов песни.

На рис. 1 приведен пример агрегирующего графа для бесёдной песни «Широкая борода» (текст песни и ее теоретико-графовая модель приводятся в [3]). Здесь группа объектов №2 – это *сад, плечи, головушка, отец-мать, род-племя, дом, блюдечко, серебряный поднос, высок терем*, т.е. объекты, связанные с девушкой.

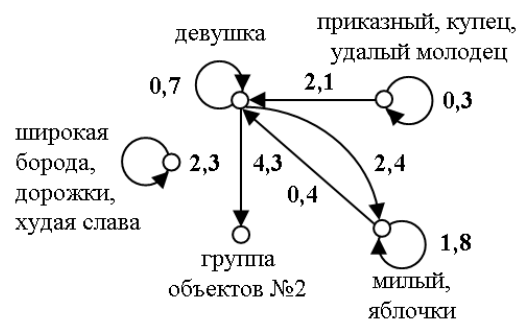


Рис. 1

На основе коллекции бесёдных песен была составлена выборка из 50 текстов. Если не учитывать изолированные вершины и петли, у большинства песен агрегирующий граф имеет вид дерева (см. рис. 2). При этом вершина с максимальной степенью (с номером 1), как правило, соответствует главному действующему лицу – парню или девушке.

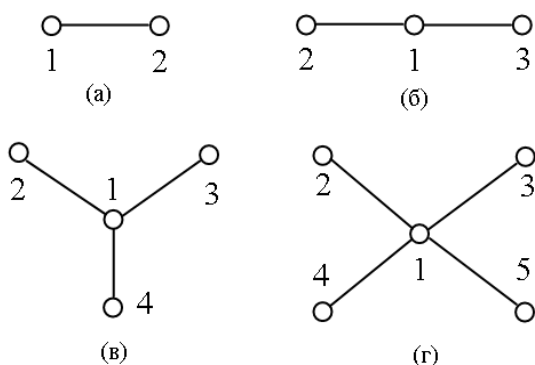


Рис. 2

Как показали эксперименты, данные структуры можно связать с основными характеристиками песен, например: «если песня имеет тип «утушенная песня», то агрегирующий граф с большой вероятностью имеет структуру 1». На языке RuleML данная закономерность может быть записана следующим образом:

```
<Implies>
  <head>
    <Atom>
      <Rel>имеет структуру</Rel>
      <Var>агрегирующий граф</Var>
      <Var>1</Var>
      <Var>с большой вероятностью</Var>
    </Atom>
  </head>
  <body>
    <Atom>
      <Rel>имеет тип</Rel>
      <Var>песня</Var>
      <Var>утушенная песня</Var>
    </Atom>
  </body>
</Implies>
```

Также можно выделить закономерности с обратным порядком, например: «если агрегирующий граф имеет структуру 1, то эта песня с большой вероятностью имеет частый темп исполнения». На языке RuleML данная закономерность может быть записана следующим образом:

```
<Implies>
  <head>
    <Atom>
      <Rel>имеет темп</Rel>
      <Var>песня</Var>
      <Var>частый</Var>
      <Var>с большой вероятностью</Var>
    </Atom>
  </head>
  <body>
    <Atom>
      <Rel>имеет структуру</Rel>
      <Var>агрегирующий граф</Var>
      <Var>1</Var>
    </Atom>
  </body>
</Implies>
```

Описав таким образом полученные закономерности, исследователь получает возможности для автоматизированного анализа правил и получения новых знаний

6 Реализация логического вывода

Для успешной работы системы необходимо обеспечить логический вывод на основе имеющихся фактов и правил. Для этого нужна машина логического вывода.

Возможны два варианта реализации. Первый заключается в использовании в системе какого-либо внутреннего формата для хранения фактов и правил. В таком случае в качестве такого формата можно, например, взять форматы представления знаний таких сред как CLIPS или SWI-Prolog. Эти среды предназначены для систем, основанных на знаниях, и являются свободно-распространяемыми продуктами. Программы, написанные в обеих этих средах, могут быть интегрированы с программами на других языках, в том числе с Java, что позволит создать web-интерфейс для системы. Сама программа, осуществляющая логический вывод может выполняться на сервере, а передача входных и выходных данных может осуществляться через web-интерфейс.

Стандарты XML, RDF и RuleML будут играть роль форматов для представления результатов и методики исследования другим сообществам исследователей и другим системам. Для преобразования из внутренних форматов во внешние необходимо будет использовать уже имеющиеся программы (например, SWI-Prolog содержит библиотеку для разбора данных в формате RDF) или реализовать собственную.

Второй вариант реализации заключается в использовании RuleML и в качестве внутреннего формата представления знаний. В таком случае, необходимо использовать машины логического вывода, работающие с форматом RuleML. Примерами таких машин является Bossam [12], которая позволяет строить приложения в рамках концепции Semantic Web, а также OO jDREW [17] – объектно-ориентированная дедуктивная машина вывода для RuleML. OO jDREW представляет собой библиотеку, написанную на языке Java. Таким образом можно написать набор приложений на Java, которые будут реализовывать функции системы, и осуществлять логический вывод, используя функции библиотеки OO jDREW.

Также существует система DR-DEVICE [14], позволяющая осуществлять рассуждения в условиях неполной и противоречивой информации. Это актуально в условиях Semantic Web, например, конфликты могут возникнуть при объединении онтологий. Система поддерживает использование фактов, правил и приоритетов правил.

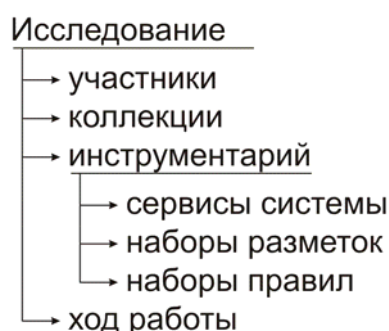
DR-DEVICE позволяет загружать RDF документы из Интернета и использовать их в качестве входной информации – фактов – для программы. Прави-

ла могут быть записаны либо на языке, подобном языку CLIPS, либо с использованием синтаксиса RuleML. Полученные результаты можно выгружать в Интернет в виде RDF документов.

7 Проект библиотеки

В рамках системы «Источник» создается библиотека методик и результатов исследований коллекций текстов. Цель такой библиотеки – организация хранилища аналитических публикаций, которые в своем составе содержат не только исходные источники материалы и описание полученного результата в виде научной статьи, но и сам инструментарий исследования с промежуточными выкладками.

Рассмотрим структуру информации в библиотеке об одном «типовом» исследовании.



Каждое исследование является фактом связывания набора составляющих его объектов и имеет собственный уникальный идентификатор и также является отдельным объектом системы.

В рамках одного исследования определяется состав участников, работающих над ним. Как правило, это руководитель исследования, который определяет состав остальных объектов исследования, и исполнители, которые получают возможность работать с определенными в исследовании объектами.

Исследование проводится на основе фиксированной коллекции документальных источников. Как правило, это полная обособленная коллекция документов.

Исследование заключается в применении к источниковому материалу специализированного инструментария: программных сервисов и алгоритмических модулей, фиксированных наборов структурных и семантических разметок, определенных на данных разметках наборов правил получения результатов.

Инструментарий применяется последовательно, согласно заранее описанному ходу работы. Например, сначала производится физическая разметка текстов (выделение словоформ), затем – структурно-семантическая разметка, после этого производится обработка полученной структуры с помощью заранее определенных правил (функций, преобразований) с целью получения новой разметки или новых правил, содержащих выявленные закономерности информации.

Результат всего исследования либо отдельной его стадии записывается в виде XML-документа, состоящего из трех частей: результата, правил, посылок. В качестве посылок (аргументов правил) выступают первично размеченные тексты, результатом является новая разметка или новые правила (закономерности).

В случае отсутствия заранее разработанного инструментария для анализа проводится «экспериментальное» исследование, результатом которого являются новые, выработанные исследователем, схемы разметки текстов или правила получения результатов.

Накопленные результаты могут быть рассмотрены как исходный материал для выдвижения и проверки новых гипотез и проведения новых исследований.

Кроме рассмотренного разреза «исследования», навигацию по библиотеке можно будет осуществлять на базе остальных составляющих ее объектов. Например, при просмотре с точки зрения участников можно будет видеть, в каких исследованиях они участвовали. Для коллекций текстов можно будет просмотреть примененные к ним разметки и правила, и, наоборот, для той или иной разметки получить примеры её использования.

Литература

- [1] Бородин Л.И. Контент-анализ и проблемы изучения исторических источников // Математика в изучении средневековых повествовательных источников. – М., 1986. – С. 8–30.
- [2] Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. Достоверный и правдоподобный вывод в интеллектуальных системах. – М., 2004.
- [3] Варфоломеев А.Г., Кравцов И.В., Москин Н.Д. Проект специализированного Интернет-ресурса для представления и анализа фольклорных песен // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Пятой Всероссийской научной конференции RCDL'2003 (Санкт-Петербург, 29–31 октября 2003 г.). – СПб., 2003. – С. 339–343.
- [4] Варфоломеев А.Г., Кравцов И.В., Филатов В.О. SVG-визуализация в цифровых библиотеках рукописных документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Девятой Всероссийской научной конференции RCDL'2007 (Переславль-Залесский, Россия, 14–18 октября 2007 г.). – Переславль-Залесский, 2007. С. 230–235.
- [5] Иванов А. С., Варфоломеев А.Г. Использование технологии XML для введения в научный оборот комплекса документов «Moscovitica-Ruthenica» // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Шестой Всероссийской научной конференции RCDL'2004 (Пущино, 29

- сентября – 1 октября 2004 г.). – Пущино, 2004. – С. 285–289.
- [6] Иванов А. С., Варфоломеев А.Г. Технология XML как инструмент компьютерного источниковедения (на примере формулярного анализа документов приказного делопроизводства) // Круг идей: Алгоритмы и технологии исторической информатики : Труды IX конференции Ассоциации «История и компьютер» / ред. Л.И. Бородин, В.Н. Владимиров. – М. ; Барнаул, 2005. – С. 241–281.
 - [7] Источник. Информационная система для работы сообществ исследователей текстов. (<http://istochnik.karelia.ru>)
 - [8] Манускрипт. Древние славянские памятники. (<http://manuscripts.ru>)
 - [9] Кравцов И.В., Варфоломеев А.Г. Принципы организации информационного пространства сетевого сообщества исследователей рукописных текстов // Информационное общество. Интеллектуальная обработка информации. Информационные технологии : Материалы 7-й международной конференции НТИ-2007 (Москва, 24–26 октября 2007 г.). – М., 2007. – С. 383–386.
 - [10] Куперштох В. Л., Трофимов В. А. Алгоритм анализа структуры матрицы связи // Автоматика и телемеханика. – 1975, №11. – С. 170–180.
 - [11] Филатов В.О., Кравцов И.В., Варфоломеев А.Г. Информационная система для работы с полнотекстовыми базами данных исторических документов на основе технологии XML // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Восьмой Всероссийской научной конференции (RCDL'2006), Суздаль, 17–19 октября 2006 г. – Ярославль, 2006. – С. 337–344.
 - [12] Bossam Rule/OWL Reasoner (<http://bossam.wordpress.com>)
 - [13] Charters Encoding Initiative (CEI) (<http://www.cei.lmu.de>)
 - [14] DR-DEVICE. A Defeasible Logic Reasoner for the Semantic Web. (<http://lpis.csd.auth.gr/systems/dr-device.html>)
 - [15] Model Editions Partnership (MEP) (<http://adh.sc.edu>)
 - [16] Monasterium Project (<http://monasterium.net>)
 - [17] Object Oriented jDREW (<http://www.jdrew.org/ooidrew>)
 - [18] Predictive Model Markup Language (PMML) (<http://www.dmg.org/pmml-v3-2.html>)
 - [19] Rule Interchange Format Working Group (http://www.w3.org/2005/rules/wiki/RIF_Working_Group)
 - [20] Text Encoding Initiative (<http://www.tei-c.org>)
 - [21] TextGrid Project (<http://www.textgrid.de>)
 - [22] The MKM Interest Group (Mathematical Knowledge Management) (<http://www.mkm-ig.org>)
 - [23] The Rule Markup Initiative (<http://www.ruleml.org>)

Information system “Istochnik” e-library project for storing sets of rules and results in the area of texts collections analyzing

N. Karginova, I. Kravtsov, N. Moskin,
A. Varfolomeyev

This paper represents one new way for special e-library organizing. The main idea is for simultaneously storing of linguistic text sources, analyzing methods, research services and final results. Also library is designed as instrument for remote collaborative work.

This project is the part of network information system “Istochnik” realization. There are many different XML-technologies using in project for text structure and semantic markup. Results of research are representing by the rules. Rules can be use for logical deduction and new formalized knowledge forming by machine. For rules we use RuleML technology.

* Статья написана в рамках проекта, поддержанного грантом Российского гуманитарного научного фонда (проект № 08-01-12136в).

Автоматическая классификация веб-сайтов

© М.Ю. Маслов, А.А. Пяллинг, С.И. Трифонов

Компания «Яндекс»
{maslov, pyal, trifon}@yandex-team.ru

Аннотация

Предлагается алгоритм автоматической классификации веб-сайтов на основе анализа их текстового содержания. Алгоритм сравнивается с алгоритмом SVM-light на данных веб-каталога Яндекса.

1 Введение

Целью настоящей работы было построение алгоритма автоматической классификации, способного эффективно классифицировать Рунет (сегмент интернета в доменных зонах .ru и .su, и страницы на русском языке вне этих доменов). Результат классификации предполагается использовать при ранжировании ссылок в ответе поисковой системы. Поэтому алгоритм должен быть:

а) Достаточно незатратным по вычислительным ресурсам, чтобы была возможность классифицировать все сайты Рунета по 500-1000 тематическим рубрикам веб-каталога.

б) В то же время, достаточно качественным, сопоставимым наиболее эффективными известными алгоритмами: чем точнее результат классификации, тем это должно лучше сказаться на ранжировании результатов поиска.

Основной особенностью задачи классификации Веба является большой объем информации. Оценить объем Рунета можно с помощью данных поисковой системы Яндекс. К настоящему моменту в поисковом индексе Яндекса содержится порядка $3 \cdot 10^9$ уникальных веб-страниц, с $4 \cdot 10^6$ «живых» сайтов Рунета (под «живыми» понимаются сайты, у которых есть хотя бы одна уникальная веб-страница из индексной базы Яндекса).

Эталонным набором данных, или учителем, при классификации веба обычно служит один из веб-каталогов, таких как Yahoo!, ODP, Каталог@Mail.ru или Яндекс.Каталог [1-2, 12]. Характерный объем «всемирного» веб-каталога – порядка 10^6 – 10^7 записей (Yahoo!, ODP), а веб-каталога Рунета – порядка 10^5 записей (Mail.ru, Яндекс).

Сопоставление объемов учителя и «генеральной совокупности» веб-страниц в задаче классификации Веба может привести к мысли, что учитель очень

мал по сравнению с «генеральной совокупностью». Так, в [3] задача классификации Веба характеризуется как semi-supervised learning, т.е. как промежуточный случай между задачей классификации с учителем и задачей классификации без учителя (кластеризации). Логика такова: в веб-каталоге [2] описано порядка 10^6 веб-страниц. Во всем Вебе их гораздо больше – порядка 10^{10} . Т.е. задача классификации Веба – классификация с очень маленьким учителем.

Однако, веб-каталог как учитель можно трактовать иначе. Можно считать, что если запись веб-каталога ссылается на входную страницу сайта, то описание в веб-каталоге относится не только к этой входной странице, но и к сайту в целом. При таком подходе объем и представительность учителя радикально возрастает. Например, в [1] на настоящий момент содержится порядка 10^5 описаний сайтов (т.е. столько записей, ссылающихся на корневую страницу сайта), что составляет порядка 2% от общего количества сайтов Рунета. При чем эти 10^5 сайтов содержат суммарно порядка 10^9 веб-страниц, что составляет порядка 30% от общего количества уникальных веб-страниц в индексе Яндекса. В этом контексте уже нет речи о малости учителя. И проблему репрезентативности учителя по отношению к «генеральной совокупности» веб-страниц решить гораздо легче.

Таким образом, особенности постановки задачи, решаемой в данной работе, следующие:

а) Рассматривается задача классификации **веб-сайтов**. В отличие от классификации **веб-страниц** – задачи, подразумеваемой в большинстве работ на эту тему.

б) Сайт в обучающей выборке представлен не только входной страницей, но и репрезентативной выборкой своих страниц.

Отметим, что у такого «посайтового» подхода есть следующие проблемы:

а) **Зашумленность**. Сайты нередко содержат заметное число страниц, не относящихся к теме. Характерным примером являются интернет-форумы. Типичный тематический форум состоит из нескольких разделов по теме, и раздела, предназначенного для общения без ограничения темы. Такой раздел часто бывает довольно объемным. Кроме того, форумы часто подвергаются спамовым атакам, что приводит к большой зашумленности плохо модулируемых и немодулируемых форумов.

б) **Политематичность сайтов.** Нередко сайт посвящен широкой предметной области, или сразу нескольким предметным областям. Редакторы каталогов в подобных случаях стараются выделить и описать отдельно тематические подразделы, сайтов. Но часто эти подразделы выделить не удается из-за особенностей структуры сайтов. В таких случаях к нескольким тематическим рубрикам относится весь сайт, и тогда каждая его страница получает несколько тем. Но каждая конкретная страница сайта обычно посвящена одной конкретной теме. И в случае сайта с N темами точность учителя оказывается равной $1/N$.

2 Related work

В работе [3] констатируется, что методы автоматической классификации, основанные на анализе текстов, в случае веб-страниц работают плохо. И показано, что при использовании текстов входящих и исходящих ссылок страницы, и учета ссылочной структуры веба в целом, совместно с лексикой текстов веб-страниц, качество классификации можно повысить.

Идея об использовании ссылочной информации совместно с текстом страниц при классификации веба используется также в ряде других работ, в частности [5–8].

В [9] показано, что использование аннотаций сайтов, сделанных редакторами веб-каталога [10], позволяет существенно улучшить качество классификации. Далее, предложен алгоритм автоматического построения аннотаций веб-страниц. И показано, что с помощью таких автоматических аннотаций качество классификации тоже можно повысить.

В семинаре РОМИП [11] задачи классификации веб-сайтов и веб-страниц явно разделены (возможно, впервые). Участникам «дорожки» классификации веб-сайтов предоставляется учитель, построенный на основе русскоязычной части веб-каталога ODP [12]. В нем 2100 сайтов, 300 тыс. страниц, с сайта бралось не более 500 страниц, в порядке обхода «в ширину». Проверка качества алгоритмов проводилась на другом множестве: на сайтах с хостинга Narod.ru (в дорожках 2003–2006 гг.) или сайтах из белорусского интернета (в дорожке 2007 г). Таким образом, обучающее и проверочное множество сайтов в РОМИП могут отличаться по своим характеристикам.

В [13] была предложена постановка задачи, аналогичная РОМИП. Набор данных, предоставляемый стипендиатам, был построен на основе каталога Яндекса [1]. В выборке 40 тыс. сайтов, 1.5 млн страниц, с сайта берется тем больше страниц, чем больше сайт, от десятков до тысяч страниц. В выборку страниц с сайта попадают а) страницы, близкие к корню сайта ($1/3$ страниц) б) случайная выборка страниц из дополнительной части сайта ($2/3$ страниц) [14]. Этот набор данных был использован, в частности, в работе [15], в которой анализировалась эффективность ряда известных методов при реше-

нии задачи классификации сайтов по текстовому содержанию их страниц.

3 Построение проверочной и обучающей выборки. Тестирование классификаторов

Был создан набор данных из 4 млн веб-страниц в соответствии с процедурой [14]. Полученный набор данных делился на две выборки – обучающую выборку ($2/3$ сайтов) и проверочную (оставшаяся $1/3$ сайтов).

Алгоритмы классификации обучались на всех страницах из обучающей выборки. Далее, проводилась классификация страниц на проверочной выборке. Возможным результатом классификации страницы было три варианта:

1. Страница относится к заданной теме
2. Страница не относится к заданной теме
3. Страницу не удалось определить (мало данных)

Если для сайта число страниц по теме было больше, чем число страниц, не относящихся к теме, то считалось, что сайт относится к теме. В противном случае считалось, что сайт к теме не относится. Качество классификаторов определялось F1-мерой полученного результата классификации сайтов.

«Среднее» качество классификаторов оценивалось на 12 случайно выбранных рубриках разных уровней иерархии (см. табл. 1). Кроме того, был проведен расширенный набор экспериментов и проанализированы ошибки классификатора для темы Порно (см. ниже разделы 4.3, 5 и 6). Качественная классификация сайтов по теме Порно востребована в связи с задачей семейной фильтрации при поиске [19]. Тема Порно была синтезирована посредством объединения пяти подтем рубрики Яндекс.Каталог/<http://yacatalog.narod.ru/admin/sprav/themes.xhtml?level=1&parent=0>Развлечения/ Личная жизнь. Три из них публичные: это Эротика и секс, Эротические галереи, Эротические игры. И две не публикуются в Яндекс.Каталоге: это Интим-услуги и Порнография.

4 Классификатор сайтов по ключевым словам (КС-классификатор)

Одним из наиболее мощных методов решения задачи классификации текстовых документов считается метод опорных векторов, или SVM (см. напр. обзор [16]). Для сравнения с описываемым алгоритмом мы выбрали реализацию SVM-Light [18], поскольку она является одной из наиболее широко используемых реализаций метода SVM, и одной из наиболее эффективных по вычислительным ресурсам во время обучения, с возможностью применения к очень большим наборам данных [17].

К сожалению, использование данного алгоритма для классификации всех сайтов, известных поисковой машине Яндекс, по 500 темам, нам представляется затруднительным.

Во-первых, алгоритм хранит все входные данные в памяти, в результате возникают ограничения

по числу страниц, на которых строятся правила (в данной работе 3GB оперативной памяти хватило для хранения 600 тыс. страниц, при средней длине страницы 320 слов).

Во-вторых, при малых объемах обучающей выборки алгоритм в наших экспериментах оказался неустойчивым. При использовании 200 тыс. страниц в качестве обучающего множества, F1-мера для темы Порно была ~30%, а F1-мера на выборке из 600 тыс. страниц ~83%.

В связи с этим возникла потребность в создании достаточно простого и устойчивого алгоритма, способного классифицировать большие объемы данных по большому числу тем. В результате был построен классификатор байесовского типа, который и описан в этой статье.

При построении классификатора использовались следующие идеи:

1. Классификация строится по каждой теме независимо друг от друга.
2. Тему документа можно определить по списку встреченных слов (bag of words).
3. Для каждой темы есть слова, характерные для темы, и слова, которые в теме не встречаются.
4. Множество характерных для темы слов можно разбить на 2 части:
 - a. Характеристические слова – множество слов, без которых невозможно раскрыть тему. Если в документе нет ни одного слова из этого множества, то документ к теме не относится.
 - b. Остальные слова, характерные для темы. Мы не можем определить только по наличию или отсутствию этих слов в документе, в теме документ или нет.
5. Кроме характерных слов, классификатор существенно использует слова, нехарактерные для темы.

Ниже описаны процедуры обучения и собственно классификации сайтов.

4.1 Процедура обучения

1. Из исходного множества документов (4 млн документов) собирается множество словоформ F которые встречаются в исходном корпусе слов больше чем N раз (бралось $N = 500$, получено 135 тыс. словоформ).
2. Для каждой темы собирается статистика: сколько раз слово встретилось в теме N_w и не в теме $\overline{N_w}$, а также сколько всего слов во всех документах в теме N_{tot} и не в теме $\overline{N_{tot}}$. Далее: $P_w = N_w / N_{tot}$, где P_w – вероятность встретить данное слово. $P_w^L = 1 - (1 - P_w)^L$, где P_w^L – вероятность того, что в тесте длиной L есть данное слово для документов, которые от-

носятся к заданной теме. Точно так же считается $\overline{P_w^L}$ – вероятность того, что в тесте длиной L есть данное слово для документов, которые не относятся к заданной теме.

$W_w^L = \ln(P_w^L / \overline{P_w^L})$, где W_w^L – вес слова, логарифм вероятности того, что документ относится к заданной теме, при условии, что в документе длиной L встретилось данное слово.

3. Слова сортируются в порядке убывания веса. Учитываются положительные веса для N_{top} первых слов (характеристические слова). N_{top} – подбирается из условия максимальной F1-меры на обучающем множестве. Для разных тем N_{top} разное: чем шире тема, тем N_{top} больше. Для остальных слов с положительным весом вес обнуляется. Для остальных слов с отрицательным весом вес учитывается.

В формуле выше существенным образом участвует длина документа L , различная для различных документов. Однако, для организации эффективно-счета полезно заменить изменяющееся значение L на постоянную усредненную величину. Эксперименты показали, что F1-мера результата от такой замены изменяется незначительно.

Средняя длина документов в обучающей выборке была 320 слов на множестве словоформ F . Было протестировано несколько вариантов с разной длиной среднего документа L при вычислении веса W_w^L . Оптимальный вариант выбран равным 50.

4.2 Процедура классификации

1. Классификация документа. Выделяются все слова из множества F . Вычисляется среднее значение логарифма вероятности того, что документ соответствует теме для слова в документе:

$$P = \frac{\sum_w W_w^L * N_w}{\sum_w N_w}$$

Если $P > 0$ – документ в теме, если $P < P_{min}$ – документ не в теме. Максимальная F1-мера результатов классификации была получена при $P_{min} = -0.5$.

2. Классификация сайта. Подсчитывается число документов в теме и вне темы. Если документов в теме больше, чем вне темы – сайт в теме. Если нет – вне темы.

Преимущество построенного классификатора – высокая скорость работы, хорошее качество классификации. При работе создается множество характеристических слов, которые могут в дальнейшем использоваться для классификации запросов, построения тезауруса, иерархического наследования ключевых слов при классификации тем-родителей. Можно использовать ключевые слова, полученные из других источников, в частности, взять список внешних тематических слов и объединить его со списком автоматически определенных слов с максимальным весом.

Таблица 1. Результаты тестирования КС-Классификатора на 12-и случайно выбранных темах

Уровень	Тема	Точность	Полнота	F1-мера
1	Дом	65.3	71.9	68.5
1	Учеба	51.4	79.0	62.3
2	Дом/Все для праздника	48.5	76.6	59.4
2	Общество/Религия	82.3	72.7	77.2
2	Hi-Tech/Мобильная связь	47.2	69.6	56.2
2	Бизнес/Строительство	37.8	62.2	47.1
3	Общество/Власть/Силовые структуры	41.4	59.0	48.6
3	Hi-Tech/Мобильная связь/Сотовые телефоны	46.0	72.7	56.3
3	Спорт/Экстремальный спорт/Пейнтбол	71.4	88.2	78.9
3	Дом/Все для праздника/Фейерверки	57.9	100	73.3
3	Дом/Кулинария/Рецепты	49.4	89.4	63.6
3	Порно	72.7	87.3	79.3

Таблица 2. Анализ ошибок КС-классификатора для темы Порно

Класс ошибки	Описание ошибки	Процент ошибок	Ошибочные сайты
Сайт ошибочно отнесен к теме	Сайт имеет сходную тематику – женское белье, медицина	22%	www.echo-h.ru, marmeladova.com, shop.flirt.ru, www.night-dress.ru, www.chocolatelily.ru
	На сайте порно есть, но с точки зрения Яндекс.Каталога это «развлекательный портал» (типичные разделы: эротика, юмор, обои и т.п.), который к теме Порно редакторы не относят	59%	Anekdots.smeha.net, www.strasti.ru, makecool.ru, www.s-info.ru, www.adamast.ru, www.klassno.com, www.loveorgasm.ru, www.prikol.net, zateynik.ru, bigpenis.com1.ru, www.megaportal.ru, www.erotofun.ru, death.com.ru
	Сайт имеет другую тему, автомат его отнес в Порно, возможно, из-за заспамленных форумов	9%	www.khabensky.net, www.raskolbass.ru
	Неактуальность записи в Яндекс.Каталоге	5%	www.gayvideo.ru
	Сайт недоступен	5%	www.forum.prikola.net
Сайт ошибочно не отнесен к теме	На сайте используется специфический язык (научный, медицинский, торговый)	27%	www.ograna.com, oganezov.boom.ru, www.bondage.ru, www.shorox.ru, www.sexrecord.ru, www.donnasummer.ru, vein.ru, www.fetishshop.ru, www.lesbi.ru
	Много страниц не по теме (форумы с обсуждением других тематик, ошибки типа soft-404 и т.п.)	27%	www.erochat.ru, www.x-time.ru, www.kiss-chat.ru, www.pornovids.ru, ecret-video.ru, photoart.prm.ru, vipxxx-party.ru, www.pickupforum.ru, www.xoxma.ru
	Политематические сайты, большинство страниц относится к другим темам	25%	www.rulz.ru, 1001.vdv.ru, kobelissimo.narod.ru, www.comics.ru, www.babes.ru, www.komi.com, excluz.com, www.krasgay.ru
	Явная ошибка автомата	6%	www.nudecelebs.ru, www.blevota.com
	Ошибочная запись в Яндекс.Каталоге	12%	letstest.info, sexxx6.sbn.bz, www.sexxx.sbn.bz, www.mpeg74.ru
	Сайт недоступен	3%	britva.gay.ru

Работа алгоритма была протестирована на 12 случайно выбранных темах. Результаты представлены в табл. 1.

Сравнение результатов при использовании порно-словника, составленного вручную, представлено в табл. 3.

Таблица 3. Влияние составленного вручную порно-словника

	Точность	Полнота	F1
Авто-список (500 слов)	74.4	88.5	80.8
Авто-список (200 слов) порно-словник (300 слов)	71.4	91.6	80.3

Добавление порно-словника происходило следующим образом. Для каждого слова из порно-словника вычислялся вес слова. Далее, выбиралось 300 слов из порно-словника с максимальными весами. К полученному списку слов добавлялось 200 слов с максимальными весами из списка, полученного автоматически.

4.3 Анализ полученных результатов

Список ключевых слов достаточно сильно изменялся при изменении обучающего множества сайтов, при использовании списка ключевых слов, собранных вручную. Также делались попытки очистить полученные списки слов, удалить случайные слова. Однако алгоритм оказался достаточно устойчивым по отношению к изменению списка ключевых слов. Качество работы изменялось незначительно.

Качество классификатора зависит от количества используемых в классификаторе словоформ. При использовании 200 тыс. словоформ F1-мера результатов классификации упала на 2%. При использовании 50 тыс. словоформ F1-мера результата оказалась также на 1–2% хуже.

Анализ ошибок КС-классификатора для темы Порно приведен в табл. 3.

5 Двухстадийный классификатор

При использовании классификатора по ключевым словам вариация параметров N_{top} , P_{min} позволяет гибко менять соотношение полноты и точности, т.е. получать полноту, близкую к 100%, за счет малой точности.

Сделаем две стадии обучения и классификации сайтов. На первой стадии построим классификатор с максимальной полнотой. На этом этапе отделим все тематические сайты от далеких по теме. Для темы Порно это приведет к тому, что в тему подмешают сайты с женской одеждой, развлечениями.

Множество сайтов, отнесенных на первой стадии в тему Порно, используем для обучения классификатора на второй стадии. Так как множество не-порносайтов на второй стадии существенно отличается от множества не-порносайтов на первой стадии, то статистика для слов существенно изменится. На второй стадии порносайты будут эффективно отделяться от сайтов с женской одеждой, развлечениями, а не от всего множества не-порносайтов. Это позволяет поднять полноту и точность конечного классификатора.

Результат работы двухстадийного КС классификатора представлены в табл. 4.

Таблица 4 Результаты тестирования двухстадийного классификатора

	Точность	Полнота	F1
Первая стадия	29.5	96.2	45.2
Вторая стадия	77.8	90.9	84.1

Для проверки качества построенного классификатора, была проведена классификация сайтов с помощью классификатора SVM-light. Получена точность 84.4%, полнота 82.4%, F1-мера 83.4%. Таким образом F1-мера двухстадийного классификатора оказалась немного выше, чем при использовании классификатора SVM-light [18].

6 Классификация темы Порно по полным данным

С помощью КС-классификатора был произведен обход всех сайтов из индекса Яндекса.

При обучении классификатора использовалась треть сайтов веб-каталога Яндекса. С помощью построенного классификатора был проведен обход всех страниц всех сайтов находящихся в поисковом индексе Яндекса, тогда как при обучении и предварительном тестировании классификатора использовалась только малая репрезентативная выборка страниц сайта.

Алгоритм классифицировал 4 млн сайтов. Из них 34500 были отнесены к теме Порно. При сравнении результатов классификации с данными веб-каталога Яндекса была получена точность 75.3%, полнота 87%, F1-мера 80.7%.

7 Выводы

В данной работе предложен классификатор сайтов по ключевым словам. Основными преимуществами его является большая скорость работы, простота построения правил, устойчивость, что позволяет использовать его для эффективной классификации больших объемов данных и для большого количества тем. Качество работы классификатора сопоставимо с качеством работы алгоритма SVM-light. Классификатор при своей работе строит множество ключевых слов для темы, которые могут быть использованы отдельно от классификатора.

Литература

- [1] Яндекс.Каталог. <http://yaca.yandex.ru>
- [2] Yahoo! Directory, <http://dir.yahoo.com>
- [3] Каталог@MAIL.RU <http://list.mail.ru/index.html>
- [4] S. Chakrabarti. Mining the Web : Discovering Knowledge from Hypertext Data, Morgan-Kaufmann Publishers, 2003.

- [5] А.В. Сычев, М.М. Баженов. «Автоматическое пополнение веб-каталога на основе идентификации веб-сообществ с последующей фильтрацией документов по контенту» // Интернет-математика 2007 : Сборник работ участников конкурса. – Екатеринбург : Изд-во Урал. ун-та, 2007.
<http://download.yandex.ru/IMAT2007/sychev.pdf>
- [6] G. Attardi, A. Gulli, F. Sebastiani. Automatic Web Page Categorization by Link and Context Analysis // In Chris Hutchison and Gaetano Lanzarone (eds.), Proc. of THAI'99, 1999, 105–119
- [7] E. Glover et. al. Using web structure for classifying and describing web pages, Proc. of the 11th international conference on World Wide Web, May 07–11, 2002, Honolulu, Hawaii, USA
- [8] P. Calado, M. Cristo, E. Moura et al. Combining link-based and content-based methods for web document classification. – CIKM'03, Nov. 3–8. – [Electronic resource]
<http://homepages.dcc.ufmg.br/~nivio/papers/cikm03.ps>
- [9] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma, Web-page Classification through Summarization
- [10] LookSmart Web directory.
<http://search.looksmart.com> (ссылка устарела)
- [11] Сайт ПОМИП www.romip.ru
- [12] Open Directory Project <http://www.dmoz.org/>
- [13] Интернет-математика 2005: Автоматическая обработка веб-данных Москва, 2005.
<http://company.yandex.ru/grant/list.xml>
- [14] Описание наборов данных для участников программы научных стипендий Яндекса 2005 г.
http://company.yandex.ru/grant/datasets_description.xml
- [15] М.В. Киселев. Оптимизация процедуры автоматического пополнения веб-каталога. // Интернет-математика 2005: Автоматическая обработка веб-данных. – М., 2005.
http://company.yandex.ru/grant/2005/08_Kiselev_102710.pdf
- [16] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.
- [17] O. Ivanciuc. Applications of Support Vector Machines in Chemistry, Rev. Comput. Chem. 2007, 23, 291–400.
- [18] T. Joachims, Training Linear SVMs in Linear Time, KDD, 2006. <http://svmlight.joachims.org/>
- [19] Семейный поиск Яндекса.
<http://help.yandex.ru/search/?id=481933>

Web sites automatic categorization

Mikhail Maslov, Aleksei Pyalling, Sergey Trifonov

An automatic Web page categorization method is proposed. It's efficiency is evaluated and compared with SVM-light method using Yandex web directory data.

Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров

© А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков

МГТУ им. Н.Э. Баумана

arka@inteltec.ru, dmitryb@inteltec.ru, mvvbox@yandex.ru, simakov99@mail.ru

Аннотация

В статье изложен нейросетевой метод кластеризации коллекций текстовых документов на основе карт Кохонена. Также предложен метод синтеза аннотаций для формируемых кластеров, в основе которого лежит идея поиска устойчивых словосочетаний.

1 Введение

Отличительной особенностью методов кластеризации является способность автоматически выделять группы в потоке входных данных. В контексте обработки текстов на естественном языке это свойство является особенно привлекательным, когда возникает необходимость оперативно выделить группы в большом массиве текстовых документов. Эта задача, например, актуальна для информационно-поисковых систем, когда пользователю необходимо дать поверхностное представление обо всем списке найденных документов, не принуждая его просматривать большую часть этих документов.

В такой постановке задачи дополнительно возникает необходимость синтеза аннотаций кластеров, кратко отражающих тематику их документов.

2 Основные методы кластеризации

Объекты, с которыми мы оперируем — это тексты или части документов, т.е. некоторые фрагменты текстов. Каждый из документов представляется набором слов (или словосочетаний, или других единиц языка). В этом наборе могут встречаться конструкции, которые не должны влиять на результаты поиска: некоторые слова общей лексики, предлоги и другие строки. Имеются также слова, которые непосредственно влияют на отнесение документа в какую-либо категорию, — это термины. Каждый термин является элементарным признаком, множество терминов составляет пространство. Множество документов — это множество точек или векторов этого пространства. Координатами точки являются вели-

чины значимости каждого термина для данного документа. Величина значимости может оцениваться различными моделями. Часто используются следующие модели оценки значимости:

- Хэммингово расстояние, 0 означает, что термин в данном документе не встречается, 1 — встречается,
- относительная частота встречаемости слова,
- модель TF-IDF.

При кластеризации документов следует стремиться к тому, чтобы схожесть документов, попадающих в кластер, была смысловой. При этом численно схожесть задается следующими способами:

1. множество документов, значения близости между любыми двумя элементами которого не меньше определённого порога

$$d^2(X_i, X_j) \geq D_{min}, \text{ где } X_i, X_j \in X - \text{множество всех векторов признаков};$$

2. значение близости между любым документом множества и центроидом кластера не меньше определённого порога, центроид кластера рассчитывается как среднее арифметическое между всеми документами кластера:

$$C_S = \frac{1}{|S_l|} \sum_{X_i \in S_l} X_i, \text{ где } S_l \in S - \text{множеству кластеров};$$

3. В качестве критерия качества разбиения для алгоритмов с заданным числом кластеров используются критерии:

$$Q_D(S) = \sum_{j=1}^k \sum_{X_i \in S_l} d^2(X_i, \bar{X}_j) - \text{сумма внутриклассовых дисперсий},$$

$$Q_M(S) = \sum_{j=1}^k \sum_{X_i \in S_l} d^2(X_i, X_j) - \text{сумма попарных внутриклассовых расстояний между элементами}.$$

Для разбиения с заранее неизвестным числом классов критерии качества задают в виде комбинации двух критериев, один из которых характеризует внутриклассовый разброс наблюдений, второй — возрастающей (или неубывающей) функцией числа классов k . Например, в виде формулы:

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

$$D_U(S) = \left(\sum_{j=1}^k \sum_{X_i \in S_i} d^2(X_i, X_j) \right) \cdot (ck(S)),$$

где c – некоторая положительная константа, а $k(s)$ – число классов, при разбиении S .

2.1 Suffix Tree Clustering

Метод Suffix Tree Clustering предложил Орен Замир (Oren Zamir) в диссертационной работе «Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results» [7]. Метод Suffix Tree Clustering кластеризует тексты в виде суффиксного дерева. Суффиксное дерево – дерево, содержащее все суффиксы данной строки. Кластеры образуются в узлах специального вида дерева – суффиксного дерева, которое строится из слов и фраз входных документов [1][6][7].

Достоинства метода:

- высокая скорость работы. По времени и занимаемой памяти дерево строится пропорционально количеству документов $O(n)$,
- наглядность представления результатов (извлекаются ключевые слова для обозначения кластеров),
- алгоритм не нуждается в обучении и задании порога срабатывания;

Недостатки метода:

- сравнительно низкая точность (по данным источника в среднем 30%),
- важен порядок слов в документе для определения названий кластеров, состоящих из нескольких слов,
- не выявляется скрытая семантика среди документов, которая может присутствовать не только на текстовом уровне; проблемы синонимии и омонимии.

Данный метод применяется в поисковой машине Vivisimo (www.vivisimo.com), которая была разработана в университете Карнеги, первоначально как экспериментальный некоммерческий проект.

2.2 Single Link, Complete Link, Group Average

Иерархические методы делят на агломеративные и дивизимные, первые из которых объединяют объекты в множества, а вторые наоборот разделяют единое множество объектов на подмножества [2].

Методы Single Link, Complete Link, Group Average относятся к агломеративным иерархическим методам, которые получили широкое распространение. Пусть задано множество объектов $O_1, O_1, \dots, O_n$, для которых определены признаки в виде матрицы:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(p)} & x_2^{(p)} & \dots & x_n^{(p)} \end{bmatrix}$$

где $x_i^{(j)}$ – значение j -го признака на i -м объекте, соответственно, каждый столбец матрицы $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})'$ характеризует объект O_i , т.е. является результатом статистического об следования объекта по p признакам.

Достоинства методов:

- иерархическая кластеризация.

Недостатки методов

- необходимо задавать порог – максимальное количество документов в кластере,
- высокая вычислительная сложность для алгоритмов Single Link и Group Average – $O(n^2)$, а в Complete Link – $O(n^3)$, где n – количество документов.

2.3 K-means

В основе K-means (K-средних) лежит итеративный процесс стабилизации центроидов кластеров. Основной характеристикой кластера является его центроид и вся работа алгоритма направлена на стабилизирование или, в лучшем случае, полное прекращение изменения центроида кластера [2].

Достоинства метода:

- низкая вычислительная сложность – $O(knT)$, где n – число документов, k – число кластеров, T – количество итераций,
- метод не нуждается в обучении и при необходимости может накапливать сведения для дальнейшего увеличения точности работы – использование Байесовских оценок параметров кластеризации.

Недостатки метода:

- требуется задание количества кластеров, как минимум на начальных этапах – до использования априорной информации,
- сравнительно низкая точность,
- в том случае, когда центроиды кластеров выбираются случайным образом, результаты, получаемые над одной и той же выборкой документов, будут отличаться. Это может происходить по причине неудовлетворительной работы генератора случайных чисел и вследствие равномерного распределения документов в пространстве – без явных областей сгущения;

2.4 Concept Indexing

Этот метод используется для уменьшения размерности пространства признаков. В пространстве признаков, размерность которого уменьшена, выполняется стандартное отображение множества документов [8].

Основное отличие метода от других заключается в том, что алгоритм может быть обучаемым или необучаемым.

Необучаемый алгоритм поиска кластеров заключается в непосредственном разбиении на k кластеров (k -way) или методом рекурсивной бисекции. В начале процесса непосредственного разбиения про-

извольно выбираются k документов, которые являются центрами для кластеров, затем к этим центрам относят документы, которые имеют наивысшее значение коэффициента близости и производят пересчёт центроидов возникших кластеров. Процесс продолжается до исчерпания всех документов или ограничиваются определённым числом итераций.

Метод рекурсивной бисекции подразумевает разбиение всего объема документов на 2 части, затем эти части тоже подвергаются разбиению, по необходимости. Процесс продолжается рекурсивно до получения k кластеров. Алгоритм рекурсивной бисекции использует значение «суммарного несходства» (aggregate dissimilarity) между документами в кластере, для того чтобы решить какой кластер разбивать дальше. Это значение равно:

$$\text{AggregateDissimilarity} = |S_i|^2 (1 - \|C_i\|_2^2)$$

где $|S_i|$ - объем кластера. $\|C_i\|_2$ - норма центроида i -го кластера.

Обучаемый вариант метода производит классификацию приходящих документов. Документы распределяются по предопределённым кластерам – классам. Кластеры определяются своими центроидами, присвоенными кластерам на этапе обучения. Практически обучение заключается в фиксировании предопределённого количества центроидов кластеров, а классификация заключается в отнесении документа к кластеру, расстояние до центроида которого, наименьшее.

Достоинства метода:

- алгоритм рекурсивной бисекции обладает высокой скоростью работы – выше линейной ($O(N \times \log k)$), где N – число документов, k – число кластеров,
- для определения точности алгоритма используется величина RI (retrieval improvement):

$$RI = \frac{\sum_{d \in D} n_d^r}{\sum_{d \in D} n_d^o}$$

где n_d^r – количество документов, принадлежащих классу d в «уменьшенном» пространстве признаков размерности r , n_d^o – количество документов, принадлежащих классу d в оригинальном пространстве признаков размерности,

- использует значения матрицы $tf \cdot idf$;
- Недостатки метода:
- один из вариантов метода – обучаемый, что для автоматических систем это, безусловно, недостаток, однако обучение заметно улучшает качество работы,
 - требуется задание количества кластеров, на которые будет разбиваться множество документов.

2.5 SOM

Самоорганизующиеся карты (Self-Organizing Maps, SOM) Кохонена выполняют обобщение предъявляемой информации. Наиболее полное изложение теории SOM можно найти в фундамен-

тальной работе Kohonen T. Self-Organizing Maps. В результате работы нейросети Кохонена получается образ, представляющий собой карту (как правило, двумерную) распределения векторов из обучающей выборки [3][4].

Данная сеть обучается без учителя на основе самоорганизации. Известно значительное число модификаций данной сети, в самом общем виде алгоритм самоорганизации следующий: По мере обучения вектора весов нейронов стремятся к центрам кластеров – групп векторов обучающей выборки. На этапе решения информационных задач сеть относит новый предъявленный образ к одному из сформированных кластеров, указывая тем самым категорию, к которой он принадлежит.

Сеть Кохонена состоит из одного слоя нейронов. Число входов каждого нейрона равно размерности входного образа. Количество нейронов определяется той степенью подробности, с которой требуется выполнить кластеризацию набора библиотечных образов. При достаточном количестве нейронов и удачных параметрах обучения сеть Кохонена может не только выделить основные группы образов, но и установить структуру полученных кластеров. При этом близким входным образам будет соответствовать близкие карты нейронной активности.

Обучение начинается с инициализации весов нейронов путем задания случайных значений или на основе обучающей выборки, векторы $\{X_i\} \in \mathbb{R}^n$. В дальнейшем происходит процесс самоорганизации, состоящий в модификации весов при предъявлении на вход векторов обучающей выборки. Для каждого нейрона можно определить его расстояние до вектора входа $\{W_j\} \in \mathbb{R}^n$:

$$\|X_i - W_c\| = \min_j \{X_i - W_j\},$$

где W_c – нейрон-победитель W_c . На текущем шаге обучения t будут модифицироваться веса нейронов из окрестности нейрона W_c :

$$W_j^{t+1} = W_j^t + \alpha(t) \cdot h_{ci}(t) \cdot (X_i - W_j^t).$$

Первоначально в окрестности $h_{ci}(t)$ любого из нейронов находятся все нейроны сети, в последствии эта окрестность сужается. В конце этапа обучения подстраиваются только веса самого ближайшего нейрона. Темп обучения $\alpha(t) < 1$ с течением времени также уменьшается. Образы обучающей выборки предъявляются последовательно, и каждый раз происходит подстройка весов. Нейронная сеть Кохонена может обучаться и на искаженных версиях входных векторов, в процессе обучения искажения, если они не носят систематический характер, сглаживаются.

Для наглядности представления карты нейроны Кохонена могут быть упорядочены в двумерную матрицу, при этом под окрестностью нейрона-победителя принимаются соседние (по строкам и

столбцам) элементы матрицы. Результирующую карту можно представить в виде двумерного изображения, на котором различные степени возбуждения всех нейронов отображаются квадратами различной площади.

Вычислительная сложность сети SOM $O(dn^2)$, где n – число векторов в обучающей выборке, d – размерность входных векторов.

В 1996 предложена модификация алгоритма SOM, названная WEBSOM, нашедшая практическое применение в кластеризации больших массивов текстовой информации.

2.6 ART

Сети ART позволяют проверить, соответствует ли «новый» образ «старому», что невозможно выполнить другими типами нейронных сетей [5, 9]. Так, например, многослойный персептрон, обучающийся по методу обратного распространения, запоминает весь пакет обучающей информации, при этом образы обучающей выборки предъявляются в процессе обучения многократно. Попытки затем обучить персептрон новому образу приведут к модификации синаптических связей с разрушением структуры памяти о предыдущих образах. Особенностью нейронных сетей с адаптивным резонансом является то, что они сохраняют «пластичность» при запоминании новых образов, и, в то же время, предотвращают модификацию старой памяти. Нейросеть имеет внутренний детектор новизны – тест на сравнение предъявленного образа с содержимым памяти. При удачном поиске в памяти предъявленный образ классифицируется с одновременной уточняющей модификацией синаптических весов нейрона, выполнившего классификацию. О такой ситуации говорят, как о возникновении адаптивного резонанса в сети в ответ на предъявление образа. Если резонанс не возникает в пределах некоторого заданного порогового уровня, то успешным считается тест новизны, и образ воспринимается сетью, как новый. Модификация весов нейронов, не испытавших резонанса, при этом не производится.

Сеть ART-1 является классификатором входных двоичных образов по нескольким сформированным сетью категориям. Решение принимается в виде возбуждения одного из нейронов распознающего слоя, в зависимости от степени похожести образа на шаблон критических черт данной категории. Если эта степень похожести невелика, т.е. образ не соответствует ни одной из имеющихся категорий, то для него формируется новый класс, который в дальнейшем будет модифицироваться и уточняться другими образами, формируя свой шаблон критических признаков. Для описания новой категории отводится новый, ранее не задействованный нейрон в слое распознавания. Эта особенность сетей ART позволяет использовать их для автоматической классификации текстов.

Используемые в сети ART компоненты вектора признаков являются дихотомическими, что не позволяет задавать значимость терминов. Кроме

того, сеть ART имеет высокую чувствительность к шуму, что делает ее непригодной для задачи кластеризации текстов.

3 Предложенная модель нейросети

Предложенная модель многослойной SOM позволяет сохранить достоинства обычной SOM, такие как простое обучение и быстрая работа для обученной сети. Основной идеей было создание такой сети, которая имеет гибкую и динамическую архитектуру, способную адаптировать свой размер и согласно данному набору данных.

После обучения сети, Таким образом, тексты представляются не единым объектом, а разделяются на несколько отдельных сегментов, что в принципе также позволяет множественное размещение одного и того же документа в различных тематических ветвях, используя алгоритмы размытых множеств.

Многослойная SOM представляет собой структуру из множества уровней, каждый из которых состоит из нескольких независимых растущих SOM. Начиная с самого верхнего уровня, каждая карта растет в размерах, чтобы представить детальное представление на каждом уровне. После улучшений глубины детализации, кластеры анализируются, чтобы определить какие из них представляют данные с минимальным уровнем детализации. Эти кластеры, которые являются слишком разнотипными входными данными, расширяются на новой растущей SOM на нижележащем уровне, который представляет данные более детально.

Такая карта снова растет в размерах, пока определенные улучшения качества представления данных не достигнуты. Кластеры представленные на уже однородном наборе данных не требуют дальнейшего расширения на более низкие уровни.

Т.е. многослойная SOM представляет собой дерево из нескольких уровней, каждый узел которого содержит SOM. SOM на первом уровне имеет размерность 1×1 и приближенно организует входящие данные. Каждому кластеру из карты первого уровня соответствует SOM второго уровня, представляющих собой более детальное представление данных и т.д. Входные данные для одной карты представляет собой набор, которые попал в связанный кластер на верхнем уровне. Кластеры второго уровня аналогично связаны с SOM третьего уровня, которая обеспечивает достаточно разряженное представление для входных данных. Карты имеют различные размеры, согласно структуре данных. Нулевой уровень обеспечивает представление всего набора данных и контроль за процессом роста дерева.

Каждый слой иерархии представляет собой обычную карту Кохонена, состоящую из одного слоя нейронов. Число входов каждого нейрона равно размерности входного образа. Обозначим $X = (x_1 \dots x_N)$ множество входных векторов, каждый из которых представляет семантический образ документа, каждый вектор нормализован $\|x_i\| = 1$, W

– матрица связей размерности $m \times n$, элементами которой являются весовые векторы.

Обучение начинается с задания случайных значений матрице связей W . В дальнейшем происходит процесс самоорганизации, состоящий в модификации весов при предъявлении на вход векторов обучающей выборки. Для каждого нейрона можно определить его расстояние до вектора входа.

$$d_m = \|x_k - w_{ij}\| \quad (1)$$

Далее выбирается нейрон $w = w^*$, для которого это расстояние минимально.

$$d_{min} = \min_{\substack{i=1..m \\ j=1..n}} \|x_k - w_{ij}\| \quad (2)$$

На текущем шаге обучения t будут модифицироваться только веса нейронов из окрестности нейрона w^* :

$$w_i^j(t+1) = w_i^j(t) + \alpha(t) \cdot h_{ci} \cdot [x(t) - w_i^j(t)],$$

$$\text{где } h_{ci} = \exp\left(-\frac{\|r_c - r_i\|}{2 \cdot \delta(t)^2}\right)$$

Первоначально в окрестности любого из нейронов находятся все нейроны сети, в последствии эта окрестность сужается. В конце этапа обучения подстраиваются только веса самого ближайшего нейрона. Скорость обучения $\alpha(t) < 1$ с течением времени также уменьшается. Образы обучающей выборки предъявляются последовательно, и каждый раз происходит подстройка весов.

Для наглядности представления карты нейроны Кохонена могут быть упорядочены в двумерную матрицу, при этом под окрестностью нейрона-победителя принимаются соседние (по строкам и столбцам) элементы матрицы. Результирующую карту можно представить в виде двумерного изображения.

При обучении карты ее точный размер не известен, для определения необходимости увеличения размера карты предложен следующий критерий:

Для каждого кластера вычисляется коэффициент «различности» векторов в кластере m_{ij} .

$$m_{ij} = \frac{1}{n_C} \cdot \sum_{x_i \in C_{ij}} \|w_{ij} - x_i\| \quad (3)$$

где C_{ij} – множество векторов, отнесенных к нейрону матрицы ij , n_C – мощность множества C_{ij} .

Для 0-го слоя данных критерий можно сформулировать следующим образом:

$$m_0 = \frac{1}{N} \cdot \sum_{x_i \in X} \|w_0 - x_i\| \quad (4)$$

где N – общее количество входных векторов.

В процессе обучения возможно изменение размера карты, как предложено. Определяется нейрон w' , для которого максимальный m_{ij} из всех.

$$m_{max} = \max \left(\frac{1}{n_C} \cdot \sum_{x_i \in C_{ij}} \|w_{ij} - x_i\| \right) \quad (5)$$

Далее, из соседних нейронов определяется тот, для которого расстояние наибольшее, и между этим нейронами и w' вставляется новый ряд или строка в матрицу W в зависимости, от того какая из метрик меньше.

$$d_{max} = \max_{w \in C_N} \|w - w'\| \quad (6)$$

где C_N – множество соседних нейронов. При этом для веса добавленных нейронов инициализируется как среднее от весов их соседей.

В качестве критерия остановки роста размеров карты предложен следующий критерий:

$$M = \frac{1}{m \cdot n} \sum_{\substack{i=1..n \\ j=1..m}} m_{ij}, M < \tau_1 \cdot m_u \quad (7)$$

где τ_1 – числовой коэффициент из $0 < \tau_1 < 1$, m_u – коэффициент «различности» для кластеров вышестоящего уровня. При этом уменьшение τ_1 приводит к увеличению результирующей карты, и, соответственно, к большей детализации представления, и наоборот.

Учитывая, что в данном случае количество кластеров ограничено сверху и снизу значениями M_{min}, M_{max} , то возможно оценить разброс значений m_u и выбрать фиксированный размер карты.

После завершения обучения карты, каждый кластер должен быть проверен – нужно ли создавать карту следующего уровня или нет. Т.е. должен быть критерий оценки, насколько содержащиеся в кластере тексты «различны». Для этого предложен следующий критерий:

$$m_{ij} < \tau_2 \cdot m_0, \text{card } C_{ij} < \mu \quad (8)$$

где τ_2 – числовой коэффициент, $0 < \tau_2 < 1$, выбираемый эмпирически, C_{ij} – множество входных векторов, отнесенных к данному кластеру, μ – ограничение минимального числа текстов в кластере, т.к. с точки зрения уточнения результатов поиска кластеры с малым числом текстов бесполезны, данный коэффициент лежит в пределах 10..25.

Если критерий не выполняется, т.е. $m_{ij} \geq \tau_2 \cdot m_0$ или $\text{card } C_{ij} > \mu$, то создается следующий уровень кластера, для которого повторяется процесс обучения. Коэффициент τ_2 позволяет управлять детальностью кластеризации, уменьшение этого коэффициента приводит к увеличению числа уровней иерархии и, соответственно, к повышению детализации представления, и наоборот.

Для формирования кластеров в заданных пределах M_{min}, M_{max} можно использовать алгоритм k-means (в т.ч. и модификацию алгоритма для размытых множеств), для определения оптимального количества кластеров k_{opt} с точки зрения минимизации внутриклассовых дисперсий.

4 Экспериментальная проверка модели

Для проведения эксперимента была выбрана коллекция текстов, представляющих собой блоки новостной информации. Общее количество текстов – 542.

Эксперимент проводился на ПЭВМ со следующими основными параметрами: один процессор Pentium 4, 3 Гц, объем ОЗУ 1 Гб.

В проведенных экспериментах искусственно использовалась плоская кластеризация вместо иерархической, что обусловлено целями эксперимента.

Время кластеризации коллекции текстов составило 18 минут.

В результате кластеризации, программа сформировала 19 кластеров. Размерность входных векторов составила 1577. Распределение количества текстов по кластерам с ранжированием показано на рис. 1, среднее количество текстов в кластере 28,52, медиана 22, дисперсия 399.26.

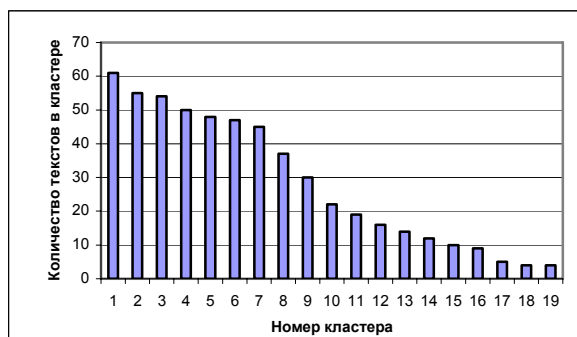


Рис. 1. Размеры кластеров

Для оценки качества кластеризации использовалась экспертная оценка. Для каждого кластера определялась преобладающая тема текстов и, исходя из этого, оценивалось количество правильно кластеризованных документов в каждом кластере. Результаты оценки представлены на рис. 2. Процент правильно кластеризованных текстов по всей коллекции составил 59,06%. Среднее количество неправильно кластеризованных текстов в кластере составило 11,68, медиана 7, дисперсия 139,67.

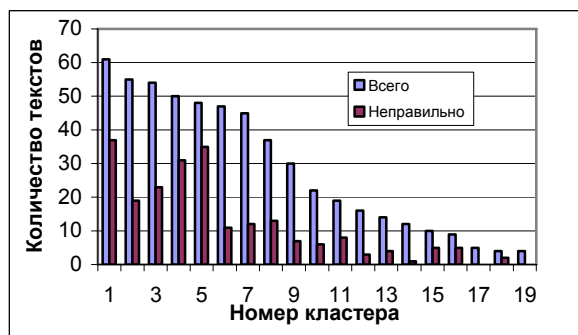


Рис. 2. Оценка качества кластеризации

Проведенный эксперимент показал пригодность предложенной модели для задачи кластеризации текстовых коллекций. Использование SOM с дина-

мической архитектурой позволяет кластеризовать большие коллекции текстов, вместе с тем, для уменьшения времени кластеризации предпочтительно проводить обучений нейросети на выборке из всей коллекции текстов.

5 Метод синтеза аннотаций кластеров

Предложенный метод синтеза аннотаций кластеров реализует подход, в соответствии с которым в основу аннотации закладываются фрагменты текстов, подлежащих кластеризации. Таким образом, реальный синтез связанного текста не выполняется, а задача сводится к поиску текстового фрагмента, наиболее релевантного для всего кластера. В таком случае синтаксическая связанность аннотации гарантируется, поскольку полагается, что обрабатываемые тексты содержат связанные тексты. Исходя из соображений практической применимости, к формируемым аннотациям предъявляется два дополнительных требования.

Лаконичность: фрагменты текста, включаемые в аннотацию, должны быть короткими фразами, содержащими 2-3 слова. В противном, случае при употреблении в аннотациях развернутых предложений, восприятие результата кластеризации будет затруднено. Напомним, что кластеризация в нашем случае необходима для более лаконичного представления результата поиска документов по большой текстовой коллекции (более 1 млн. текстов). Таким образом, пользователь, получая набор кластеров, каждый из которых охарактеризован краткими фразами, имеет возможность быстро сориентироваться в списке документов, возвращенных в результате поиска.

Контекстная независимость: форма, в которой записан фрагмент, включаемый в аннотацию, не должна предполагать наличия контекста. Предположим, что релевантной аннотацией для текстов кластера является фраза «начало массового производства». Если данный фрагмент изъять из его контекста «Почти сразу же поступил ответ и от Samsung, сообщившей о начале массового производства на втором своем заводе седьмого поколения», получим словосочетание «начало массового производства», использование которого в роли аннотации кластера без контекста неприемлемо.

Предложенный метод синтеза аннотаций предполагает, что для каждого кластера сформированы списки ключевых многословных терминов, имеющих наибольшую информационную значимость для кластера. Каждое слово термина представлено канонической формой, например, приведенному выше фрагменту будет соответствовать термин «начало, массовый, производство». Процедура синтеза разделена на следующие три этапа:

1. поиск релевантных текстовых фрагментов для каждого термина,
 2. выбор нормализованного фрагмента,
 3. выбор наиболее предпочтительного фрагмента.
- Рассмотрим указанные этапы подробнее.

5.1 Поиск релевантных текстовых фрагментов

Поскольку термин образован каноническими формами слов, для поиска всех вхождений термина в текст кластера необходимо также привести все слова текста к их каноническим формам. Для этих целей использовался морфологический анализатор, функционирование которого описано в [10]. Преобразованный таким образом в сплошную цепочку канонических форм текст кластера подлежит сканированию, в процессе которого по критерию точного совпадения отыскиваются все вхождения термина в обрабатываемый текст. Найденные вхождения термина являются кандидатами на роль аннотации кластера.

5.2 Выбор нормализованного фрагмента

На данном этапе из всех найденных кандидатов необходимо отобрать один единственный, соответствующий двум сформулированным выше критериям о лаконичности и контекстной независимости.

Для осуществления такого отбора предложен эвристический критерий, сформулированный согласно следующим соображениям.

1. Поскольку для комфортного восприятия аннотации кластеров должны быть предельно краткими (2-3 слова), то целесообразно включать в аннотации фразы, представленные именными группами (глагольные группы представленные глаголом, согласованным с наречием, не являются информативными, а при включении в группу дополнения теряется свойство лаконичности аннотации).

2. Для обеспечения контекстной независимости формы, в которой записан фрагмент, необходимо отбирать кандидатов, в которых употребляются слова в именительном падеже.

Все фрагменты-вхождения термина фильтруются согласно этим двум критериям следующим образом. Сначала из всех вхождений термина остаются только те, среди слов которых присутствует существительное в именительном падеже. Затем, если среди оставшихся вхождений присутствуют те, в составе которых есть прилагательное в именительном падеже, то все прочие вхождения удаляются. В качестве примера рассмотрим термин «ценный, бумага» (напомним, что каждое слово термина представлено своей канонической формой). В кластере, содержащем тексты с проспектами эмиссии ценных бумаг коммерческих банков, найдены следующие вхождения данного термина (см. рис. 3).

ценных бумаг ценной бумаги ценным бумагам ценных бумагах ценные бумаги ценными бумагами ценной бумагой <i>ценными бумага</i> ценной бумаге ценная бумага

Рис. 3. Пример вхождений термина

После фильтрации данного перечня первым критериями данный список сокращается до следующего (см. рис. 4).

ценной бумаги ценные бумаги <i>ценными бумага</i> ценная бумага

Рис. 4. Первый шаг фильтрация

Как видно из рис. 4 в результате кроме всего прочего попали фрагменты, содержащие грамматические ошибки (выделено курсивом), однако применение второго критерия оставляет в данном списке только два корректных фрагмента (см. рис. 5).

ценные бумаги ценная бумага

Рис. 5. Второй шаг фильтрации

Тем не менее, после двух шагов такой фильтрации может остаться большое количество вариантов (например, если группа представлена двумя существительными, первое из которых – главное слово – употребляется в различных формах, а второе – подчиненное слово – употребляется в единственной форме, например, в форме родительного падежа, совпадающей с формой именительного падежа омонима). Пример результирующего списка для того же кластера приведен на рис. 6.

проспекта эмиссии проспект эмиссии проспекте эмиссии проспекты эмиссии проспекту эмиссии проспектом эмиссии

Рис. 6. Влияние омонимии

Если морфологический анализатор обладает способностью снятия омонимии, то описанная проблема не является актуальной и список фрагментов на рис. 6 был бы сокращен до двух, выделенных жирным шрифтом. Однако в нашем случае морфологический анализатор не был в состоянии оставить у слова «эмиссии» единственную форму родительного падежа, поэтому потребовался дополнительный критерий, позволяющий фильтровать подобные списки.

Суть данного критерия заключается в следующем. Все слова, составляющие фрагмент, разделяются на две следующие группы.

- Не изменяющиеся слова. К этой группе относятся слова, употребляющиеся в одних и тех же формах в большинстве анализируемых фрагментов.
- Изменяющиеся слова. К этой группе относятся слова, форма которых изменяется от одного фрагмента употребления термина к другому.

Для примера на рис. 6 к первой группе относится слово «эмиссии», а ко второй – все приведенные формы слова «проспект». Слова первой группы фиксируются, так что предполагается, что их нормализованная форма найдена. Для каждого изменяющегося слова отыскивается форма именительного падежа.

Итоговым результирующим нормализованным фрагментом текста для исходного термина объявляется фрагмент с наибольшей группой неизменяемых слов и с наибольшим количеством изменяемых слов, употребленных в форме именительного падежа. Для примера на рис. 6 этому критерию удовлетворяют именно выделенные жирным шрифтом фрагменты, из которых произвольным образом выбирается один единственный.

5.3 Выбор предпочтительного фрагмента

К данному этапу для обрабатываемого кластера имеется список терминов, с каждым из которых связан нормализованный фрагмент. Задача данного этапа выбрать из всех терминов наиболее предпочтительный так, что связанный с ним фрагмент образует аннотацию кластера. Для этих целей в данной работе взят за основу критерий TF-IDF при условии, что кластер документов рассматривается как один единый текст. Таким образом, частота термина рассматривается не в контексте одного документа, а в контексте всего кластера.

Поскольку желательно избегать употребления терминов из одних кластеров в аннотациях к другим, необходимо применять более строгие санкции к терминам, которые встречаются более чем в одном кластере. Более того, предпочтение следует отдавать терминам, имеющим равномерное распределение в текстах кластера, т.е. следует избегать включения в аннотацию терминов, встречающихся с одной стороны часто в одном документе кластера и редко – в другом. Также важно поощрять термины, покрывающие наибольшее число документов кластера. В связи с этими замечаниями, предложена следующая модификация функции TF-IDF, оценивающая предпочтительность термина для включения его в аннотацию.

$$p_t^c = \frac{f_t^c}{\sum_i f_i^c} \cdot \frac{C}{C_t} \cdot \frac{D_t^c}{D^c} \cdot \frac{1}{1 + \log(l + \sigma_t^c)}, \quad (9)$$

где f_t^c – частота употребления термина t в текстах кластера c , f_i^c – частота употребления i -го термина в текстах кластера c , C – количество кластеров, C_t – количество кластеров, в которых встречается термин t , D_t^c – количество документов в кластере c , в которых встречается термин t , D^c – количество документов в кластере c , σ_t^c – среднее отклонение частоты употребления термина t в текстах кластера c .

Таким образом, из всех терминов кластера отбирается единственный, с максимальным значением

предпочтительности p_t^c . Аннотацией кластера объявляется нормализованный фрагмент, связанный с данным термином.

6 Экспериментальная проверка метода

Эксперименты с методом синтеза аннотаций были выполнены над искусственной коллекцией, содержащей 4 группы документов из разных предметных областей:

bank: проспекты эмиссии ценных бумаг коммерческих банков (170 текстов),

crime: новости с криминальными сводками УВД некоторых регионов РФ (154 текста),

itnews: новости с анонсами ИТ-продуктов (315 текстов),

law: распоряжения кабинета министров Республики Татарстан (97 текстов).

В рамках каждого кластера были синтезированы списки из двух и трех словных именных групп, составленных из канонических форм слов исходных текстов. Далее в каждом списке было оставлено по 20 наиболее часто встречающихся словосочетаний, которые в итоге были объявлены терминами соответствующего кластера. Над каждым списком терминов выполнялись описанные выше процедуры синтеза аннотаций, отдельно для двух и трех словных терминов.

Поставленный таким образом эксперимент, с изначально явно непересекающимися кластерами, гарантирует корректную трактовку семантики кластеров экспертом и, как следствие, минимизацию ошибки при оценке результата аннотирования. Дополнительно такой эксперимент позволяет управлять ошибкой результата кластеризации, начиная с 0%, когда кластеры полностью соответствуют описанным четырем группам, и, заканчивая 100%, когда документы из всех групп равномерно распределены по четырем группам одинакового размера.

В таблице 1 приведены результаты синтеза аннотаций на основе двух словных терминов. Для каждого кластера отбиралось три термина с наибольшим весом p_t^c .

Каждая строка таблицы содержит результаты синтеза аннотаций для соответствующего кластера. В колонках таблицы отражены результаты для трех серий экспериментов, где приведен процент смешивания исходных эталонных групп. Первой колонке соответствует эксперимент без смешивания, что соответствует идеальному результату кластеризации, последней колонке соответствует результат кластеризации, когда 40% каждой группы равномерно заполнено содержимым из других групп. В каждой ячейке таблицы приведены первые три аннотации с максимальным весом p_t^c , полученные для соответствующего кластера в соответствующем эксперименте. Каждой аннотации поставлено в соответствие значение величины p_t^c .

Таблица 1. Синтез аннотаций на основе двух словных терминов

	0%	30%	40%
bank	уставной фонд: 0.22 собрание акционеров: 0.14 ценные бумаги: 0.12	ценные бумаги: 0.06 уставной фонд: 0.05 собрание акционеров: 0.03	уставной фонд: 0.043 ценные бумаги: 0.04 собрание акционеров: 0.02
crime	мера пресечения: 0.12 розыскные мероприятия: 0.11 обнаружение трупа: 0.09	мера пресечения: 0.027 розыскные мероприятия: 0.027 обнаружение трупа: 0.02	мера пресечения: 0.0209 розыскные мероприятия: 0.0204 <i>изъятие наркотиков</i> : 0.015
itnews	технические характеристики: 0.05 жесткий диск: 0.038 карта памяти: 0.02	жесткий диск: 0.044 <i>мобильный телефон</i> : 0.016 технические характеристики: 0.011	жесткий диск: 0.018 технические характеристики: 0.01 карта памяти: 0.006
law	республика Татарстан: 0.176 министр республики: 0.15 кабинет министров: 0.071	республика Татарстан: 0.076 министр республики: 0.03 кабинет министров: 0.029	республика Татарстан: 0.045 кабинет министров: 0.02 министр республики: 0.019

Для всех серий экспериментов видно, что фактически первые три наилучшие аннотации сохраняются независимо от качества кластеризации, что позволяет судить о применимости предложенного метода (во всем эксперименте появилось только две новые аннотации, они в таблице выделены жирным курсивом). Вместе с тем с ухудшением качества кластеризации значения всех весов уменьшаются, что является закономерным, поскольку с ростом степени смещения кластеров уменьшаются все четыре множителя выражения (9).

Акцентируя внимание на формулировках самих аннотаций, можно заключить о практической применимости предложенных критериев о лаконичности и контекстной независимости, а также о методе, синтезирующем аннотации, удовлетворяющие этим критериям.

В экспериментах с трех словными аннотациями имеют место закономерности, аналогичные приведенным в таблице 1. Отличие заключается лишь в самих формулировках:

для **bank** выделены аннотации: «выпуск ценных бумаг» и «уставной фонд банка»,

для **crime** выделены аннотации «труп неизвестного мужчины» и «тяжкие телесные повреждения»,

для **itnews** выделены аннотации «технология прямой печати» и «время автономной работы»,

для **law** выделены аннотации «министр республики Татарстан» и «постановления кабинета министров».

7 Заключение

Предложенная модель кластеризации текстов в ходе экспериментов продемонстрировала свою работоспособность. Использование семантического образа документа в качестве входного вектора для нейросети позволяет адекватно представлять содержание документа и получать в результате кластеризации результаты, пригодные для практического применения.

Предложенный метод синтеза аннотаций гарантирует синтаксическую связанность формируемых фраз, поскольку берет за основу связанные фразы текстов, подлежащих кластеризации. С точки зрения человека, аннотации, удовлетворяющие предложенным критериям лаконичности и независимости от контекста, воспринимаются комфортно и в полной мере могут дать поверхностное представление о предметной области документов, заключенных в кластере.

Свойства предложенных методов позволяют рекомендовать их совместное применение в виде единого подхода при построении информационно-поисковых систем нового поколения, обеспечивающих представление больших списков текстовых документов в виде лаконичного набора поименованных кластеров, дающих представление о структуре и составе найденного массива.

Литература

- [1] Cao G., Song D., Bruza P. Suffix Tree Clustering on Post-retrieval Documents Information. The University of Queensland, 2003.
- [2] Jain A. Dubs R., Clustering methods and algorithms, 1988 // Prentice-Hall Inc.
- [3] Kohonen T., Self-Organizing Maps, Springer, Berlin, 1995, 3rd extended edition, 2001.
- [4] Kohonen T., et al. Self-Organization of a Massive Document Collection. Kohonen Maps. Elsevier, 1999.
- [5] Mucoz Alberto, "Compound Key Word Generation from Document Databases Using A Hierarchical Clustering ART Model", Intelligent Data Analysis, Elsevier Science Inc., 1997.
- [6] Zamir O. and Etzioni, O. Web Document Clustering: A Feasibility Demonstration. In Proceedings of ACM/SIGIR'1998.
- [7] Zamir O. Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results // A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy. University of Washington, 1999.

- [8] Прикладная статистика: Классификации и снижение размерности : справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин ; под ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989. – 607 с. : ил.
- [9] Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа / А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Пятой всероссийской научной конференции (RCDL'2003). – СПб. : НИИ Химии СПбГУ, 2003. – С. 140–149.
- [10] Обучение морфологического анализатора на большой электронной коллекции текстовых документов / А.М. Андреев, Д.В. Березкин, К.В. Симаков // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Седьмой всероссийской научной конференции (RCDL'2005). – Ярославль : Ярославский государственный ун-т, 2005. – С. 173–181.

The method of clustering texts collections and clusters annotating

A. Andreev, D. Berezkin, V. Morozov, K. Simakov

The article contains the method of neural clustering collections of text documents on the basis of Self-Organization Map (SOM). We also propose a method for synthesizing annotations formed clusters, based on a method of finding collocations and stable phrases.

Ранжирование источников информации в системе мониторинга новостей InfoStream

© Д.В. Ландэ, С.М. Брайчевский, А.Т. Дармохвал, А.Ю. Морозов

Информационный центр «ЭЛВИСТИ», Киев, Украина
dwl@visti.net smb@visti.net hval@visti.net alex@visti.net

Аннотация

Описано несколько подходов к ранжированию и отбору информационных источников при построении служб мониторинга новостей. Эти подходы базируются на принципе обеспечения максимальной полноты информации при минимальном количестве источников, выборе наиболее оригинальных, тематически стабильных, максимально цитируемых источников. Представлены результаты соответствующих статистических исследований. Рассматриваемые в работе принципы используются при отборе источников для корпоративных применений на базе системы мониторинга новостей InfoStream.

1 Введение

Эффективное использование возможностей современного информационного пространства невозможно без общих представлений о природе соответствующих информационных процессов, устойчивых закономерностях.

Системы интеграции и мониторинга новостей из открытых веб-сайтов сети Интернет сегодня все чаще становятся основными компонентами информационных служб различного уровня. Ни для кого уже не секрет, что даже самая закрытая новостная информация, передаваемая информационными агентствами, с минимальной временной задержкой становится доступной в Сети. Можно отметить разнообразный диапазон параметров информационных источников как по объемам публикуемой информации, так и по содержанию – от сообщений информационных агентств – до «живых журналов».

Вместе с тем мощные возможности Интернет порождает проблему оптимизации состава и количества источников, используемых корпоративной информационной системой с целью обеспечения

приемлемого качества, удовлетворяющего потребностям пользователей.

В этой связи актуальными оказываются вопросы ранжирования и выбора источников новостной информации – веб-сайтов, к которым требуется обеспечить доступ через один интерфейс как в поисковом режиме, так и в режимах аналитического обобщения.

Принципам ранжирования как отдельных веб-документов, так и документальных массивов посвящено большое количество научных работ и практических разработок [1–5]. Ссылочное ранжирование веб-сайтов сегодня является отдельным направлением интернет-бизнеса – SEO (search engine optimization) [6]. Вместе с тем, вопросам ранжирования и отбора информационных ресурсов с учетом их новостного контента, объемов и стабильности тематики публикаций уделяется значительно меньшее внимание [7, 8].

Безусловно, основным критерием при отборе источников для таких систем мониторинга новостей является их содержание. Как было показано в [8], распределение источников по контенту, соответствующему тематическим потребностям корпоративного пользователя удовлетворяет закону Бредфорда [9], соответственно, при отборе источников обязательно должно учитываться их ранжирование по степени соответствия тематике. Однако, реализация такого отбора приводит к известным сложностям. На практике такое ранжирование осуществляется экспертами [5] путем оценивая количества документов, релевантных некоторому отлаженному пакету тематических запросов, адресуемых к фрагменту базы данных, составленной из документов анализируемого источника. А это неизбежно приводит к элементу субъективизма со всеми вытекающими последствиями.

Поэтому нам представляется перспективным дополнить традиционный подход более объективными и строгими методами, позволяющими оптимизировать процесс формирования информационной базы систем интеграции контента.

Именно такой подход как основной, дополненный некоторыми представленными ниже, применяется в службе ИЦ «ЭЛВИСТИ» при отборе источников для корпоративных применений системы интеграции и мониторинга новостей InfoStream [10].

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

2 Распределение источников по количеству генерируемых документов

В качестве экспериментального корпуса в наших исследованиях использовался информационный массив, охватываемый системой интеграции и мониторинга новостей InfoStream. В частности, ниже приведены распределения, относящиеся к массиву документов за март 2008 года объемом свыше 1.2 млн. документов из более чем 2500 источников – открытых веб-сайтов.

На рис. 1 приведен график распределения (в полулогарифмическом масштабе) количества документов, опубликованных источниками, ранжированными по этому параметру. Центральная часть графика хорошо аппроксимируется прямой, что свидетельствует о близости представленной зависимости к гиперболической (т.е. о действии обобщенного закона Ципфа). На рис. 2 приведено общее количество документов, охватываемых системой мониторинга в зависимости от учитываемых в ней источников, также ранжированных по количеству опубликованных документов. Поскольку закон Ципфа предполагает аппроксимацию плотности распределения гиперболической зависимостью вида a/x , то функция распределения количества документов:

$$f(x) \sim \int \frac{a}{x} dx = a \ln x + C$$

в разумном приближении описывается логарифмическим законом.

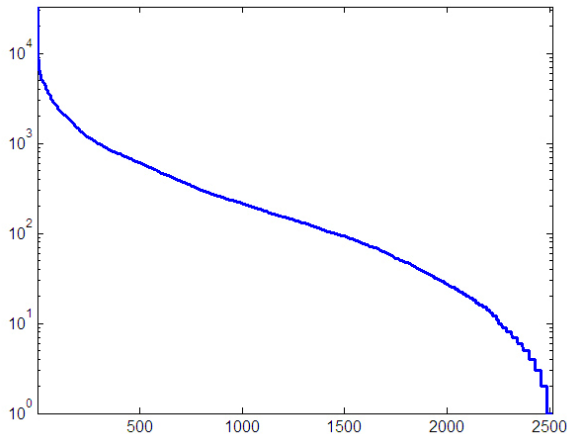


Рис. 1. Ранжированный список источников (ось OX) по количеству публикаций (ось OY)

Приведенная зависимость позволила построить критерий отбора необходимой части источников для различных корпоративных применений из общего списка охватываемых системой мониторинга InfoStream, удовлетворительно решающих задачи пользователей.

Если предположить, что все источники давали бы одинаковый вклад по количеству опубликованных документов, то рассматриваемая зависимость была бы линейной и выражалась бы формулой:

$$f_{lin}(n) = n \frac{f_{max}}{N},$$

где f_{max} – максимальный объем охватываемых документов, N – количество источников.

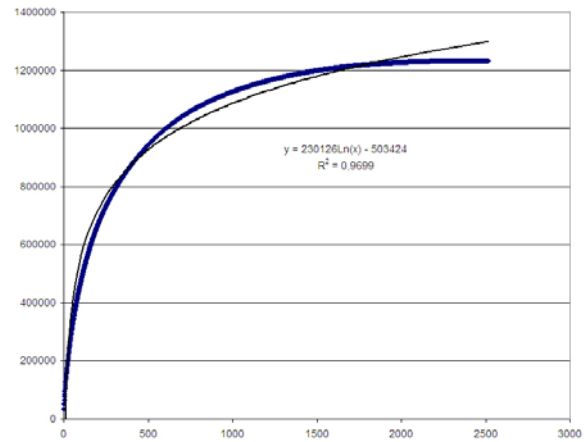


Рис. 2. Количество публикаций в системе мониторинга (ось OY) в зависимости от источников (ось OX), ранжированных по количеству документов

Очевидно, что отклонение реальной зависимости от линейной сначала возрастает, а затем уменьшается до нуля. Будем называть количество источников пороговым n_p , когда значение реальной зависимости $f(n)$ максимально отклоняется от приведенной линейной:

$$n_p = \arg \max \{f(n) - f_{lin}(n)\}.$$

На рис. 3 приведена иллюстрация значений n_p для различных значений N , т.е. когда выбирается N наиболее продуктивных источников. Что интересно (и вполне соответствует характеру функции $f(n)$), значения n_p практически линейно зависят от N (рис. 4): $n_p \sim 0.24N$, при этом количество охватываемых документов, соответствующих n_p при максимальном количестве источников (2514, рис. 5) достигает 80 процентов от f_{max} .

При этом можно заметить, что построенная зависимость удовлетворяет принципу Парето: приблизительно 20% наиболее продуктивных источников публикуют 80% документов.

3 Наиболее цитируемые источники

Как уже было отмечено, цитируемость отдельных документов и веб-сайтов сегодня является одним из основных критериев оценки рангов документов в сетевых поисковых системах (PageRank, HITS, Salsa, TrustRank, h-индекс и др.) Идея оценки уровня цитируемости позволила построить одну из первых моделей веб-пространства [11]. Главное ее достоинство состоит в том, что она естественным образом сочетает в себе содержательный аспект с воз-

возможностью использования количественных параметров, значения которых определяются вполне объективно.

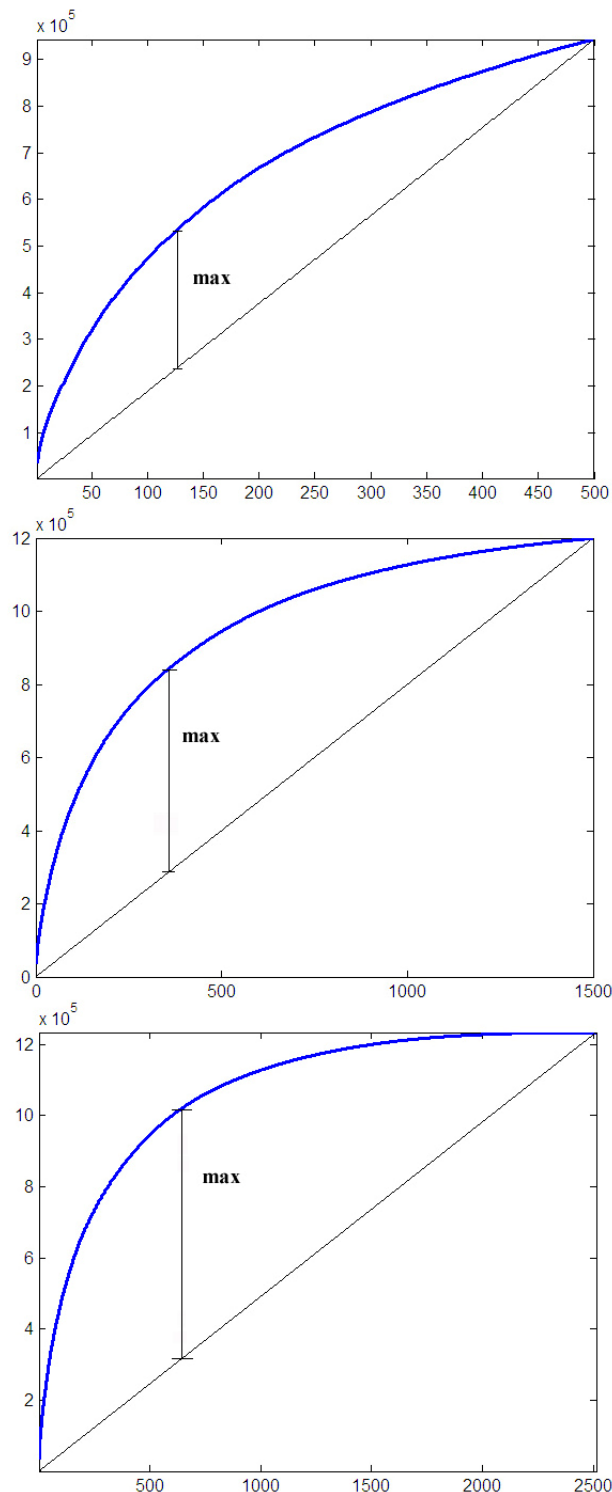


Рис. 3. Количество публикаций в системе мониторинга при подключении новых наиболее интенсивных источников (500, 1500, 2500)

Однако при изучении цитируемости новостных веб-ресурсов как информационных источников

необходимо учитывать ряд условий, актуальных для таких источников.

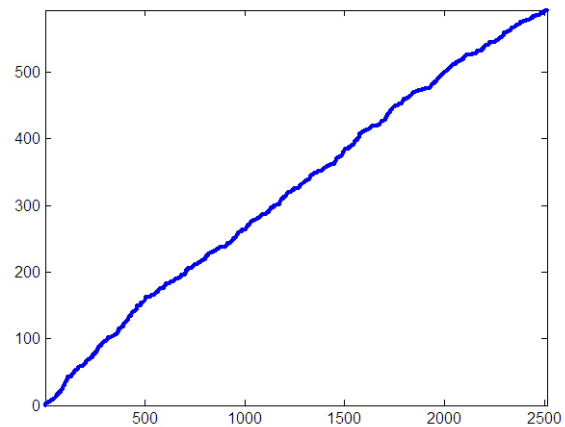


Рис. 4. Изменение порогового значения (ось OY) при изменении исходного количества источников (ось OX)

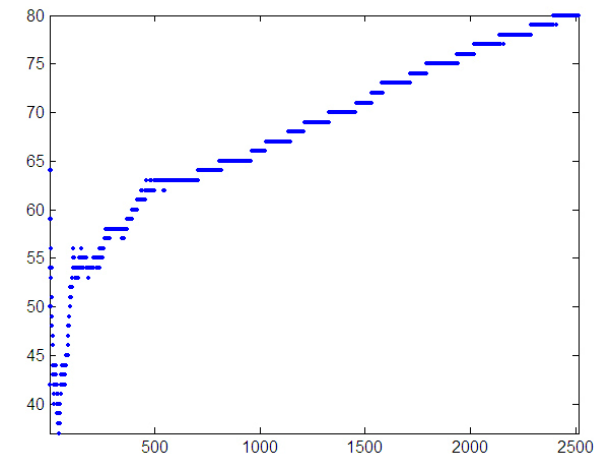


Рис. 5. Удельное количество документов, охватываемых системой (ось OY) при изменении исходного количества источников (ось OX)

Можно было бы попросту применить модель А. Бредера [11] и критерий типа PageRank [1] к новостной составляющей Web-пространства, однако такой подход нельзя считать корректным по ряду причин:

- новостные потоки характеризуются повышенной динамикой [12], что сильно влияет на природу гиперссылок. Например, на наиболее актуальные сообщения в течение определенного времени ссылок может вообще не существовать;
- модель Бредера слабо учитывает особенности «скрытого» Web, т.е. тех информационных Web-ресурсов, на которые не существует прямых гиперссылок (в свое время им рассматривались лишь ресурсы, уже охваченные поисковой системой AltaVista);
- в новостных потоках необходимо учитывать не только гиперссылки, но и ссылки контекстные, причем не только на объекты из открытой части Web-пространства (это могут быть зачастую ссылки на ресурсы, доступные только по паро-

- лю, или даже оффлайновые публикации изданий, возможно и присутствующих в Интернет);
- кроме того, модель Бредера не включает такого понятия, как содержательное дублирование информации.

Для построения модели для каждого из 2500 активных источников за март 2008 года был составлен запрос в следующем виде:

`<код источника>#<шаблон для поиска> [#<шаблон для поиска>...#<шаблон для поиска>].`

Совокупность подобных запросов была объединена в пакет, в результате специальной обработки которого для каждого сообщения, относящегося к определенному источнику – веб-сайту, были выявлены исходящие ссылки на другие источники (ссылки на собственный источник исключались). Было выявлено, что исходящие контекстные ссылки присутствовали на 484945 сообщениях с 2323 веб-сайтов.

Также было получено распределение новостных источников по количеству веб-сайтов, имеющих на них ссылки. Всего за месяц ссылки указывали на 1459 источников (без самоцитирования). Оказалось, что на первые 100 источников ведут ссылки с более 80% веб-сайтов. На рис. 6. представлен график ранжированного распределения новостных источников веб-сайтов по количеству сайтов, имеющих на них ссылки. Следует обратить внимание на то, что приведенный график позволяет достаточно четко выделить две зоны, имеющие различные статистические характеристики (углы наклона на графике): первая зона (левая на рис. 6), включающая информационные агентства и крупнейшие издания и вторая, которая соответствует сайтам госорганов, специальных изданий, компаний, публикующим пресс-релизы.

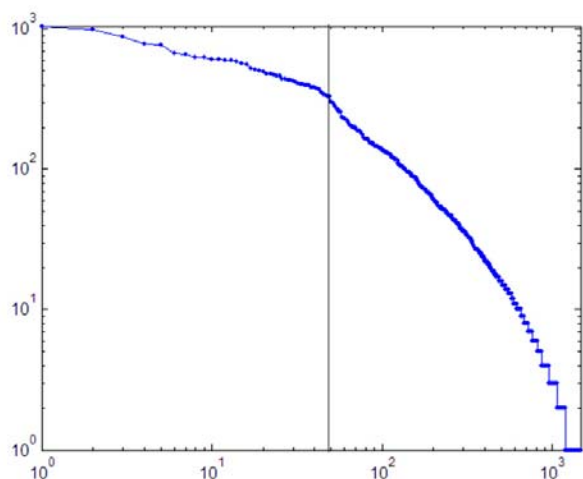


Рис. 6. Зависимость количества ссылающихся веб-сайтов (ось OY) от ранга новостного источника (ось OX) в логарифмической шкале

Ниже приведен начальный фрагмент ранжированного списка новостных источников, на которые

ведут ссылки с максимального количества веб-сайтов:

Web-сайт	Количество ссылающихся веб-сайтов
ИА «Интерфакс»	1051
«РосБизнесКонсалтинг»	983
"Reuters"	882
ИТАР-ТАСС	787
РИА «Новости»	773
УНИАН	675
Радио «Свобода»	662
НТВ	631
«Коммерсантъ»	623
BBC	598
«Комсомольская правда»	595

Следует отметить, что оценка уровня источника информации как «автора» преимущественно по количеству веб-сайтов, с которых на него ведут гиперссылки, вполне согласуется с предложенным Мораном и Лемпелем алгоритмом Salsa [13].

4 Выбор наиболее оригинальных источников

Специальное место в исследовании занимало изучение смыслового дублирования информации [14]. При этом следует отметить, что процент дублирующихся по смыслу документов в системе мониторинга InfoStream значительно меньше, чем во всем новостном Web-пространстве. Это объясняется подбором источников для сканирования, в число которых не входят многие новостные интеграторы.

Как уже отмечалось нами ранее, одной из главных особенностей новостной информации является наличие большого количества сообщений, дублирующих друг друга. Так, о событии мирового значения напишут все СМИ, причем, скорее всего, на одной из первых страниц. Потребитель же (за исключением некоторых специфических направлений аналитических исследований информационного пространства) желает получать по каждому событию одно сообщение. Поэтому исследование характера и свойств дублирования информации приобретает в современных технологиях исключительно важное значение. В том числе, крайне актуальной становится задача отбора наиболее оригинальных источников, позволяющих (по крайней мере статистически) исключить не только формальное, но и содержательное дублирование информации.

Технология InfoStream, позволяющая с высокой достоверностью выявлять содержательные дубликаты сообщений [10]. Дублирование сообщений на веб-сайтах зависит от различных причин, поэтому проведенные измерения для ранжированного по количеству публикаций списка источников показывают разный уровень, при этом информация не носит наглядного характера. Вместе с тем, сглаживание с

помощью метода скользящей средней (с окном наблюдения, равным 20), позволил получить график (рис. 7), наглядно свидетельствующий об устойчивой тенденции: чем более продуктивен источник информации, тем больше он содержит заимствований из других источников.

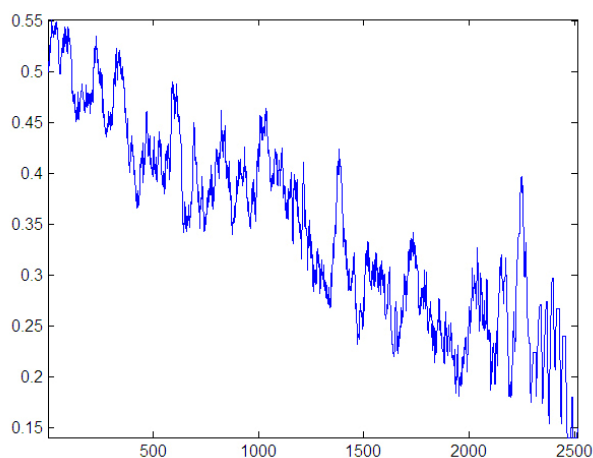


Рис. 7. Усредненное удельное количество дублирующихся документов (ось OY) по ранжированному по количеству публикаций списку источников (ось OX)

5 Тематическая стабильность

Одной из важных характеристик информационных источников в новостном сегменте Сети является их стабильность, понимаемая как генерация постоянного числа документов в единицу времени (естественно, с учетом периодичности изданий) [15]. Примером стабильных источников могут служить крупные информационные агентства, регулярно поставляющие потребителям примерно одинаковые объемы информации на протяжении длительного времени, а примером нестабильных – блоги, многие из которых активно действуют в течение нескольких дней, а затем угасают.

Естественно, источник, регулярно выпускающий свою продукцию, с большей вероятностью отразит в своих публикациях важные события, чем источник, выходящий нерегулярно, от случая к случаю (он может попросту «проскочить мимо события»).

С другой стороны, крупные издания, обеспечивающие полноценное освещение нашей жизни, как правило на первое место выводят масштабные, значительные в общественном понимании события, о которых мы все равно так или иначе узнаем, если не из телевизора, так из разговоров в метро. События же меньшего, так сказать, общественного веса, но при этом, возможно, интересные и важные для отдельных групп потребителей, либо вообще отсутствуют, либо теряются «на последних страницах». Поэтому задача оптимального учета стабильности источников отнюдь не тривиальна и требует, на наш взгляд, серьезных исследований.

Ниже мы обратимся к одному из важных ее аспектов.

Как было показано, ежедневное общее количество документов, публикуемых на основных информационных веб-сайтах приблизительно постоянно, и колеблется в основном в зависимости от дня недели [8]. Вместе с тем тематика публикаций подвержена существенным колебаниям.

Один из возможных подходов к решению проблемы ранжирования источников информации основывается на подходе, заключающемся в изучении динамики порождаемых ими тематических информационных потоков.

На практике среди множества проблем подбора и анализа источников контента большое значение, в частности, имеет учет параметров их тематической стабильности. При этом тематическая стабильность и стабильность публикации информации источниками зачастую играют решающую роль при проведении аналитических исследований. Например, такие важные свойства информационных источников, как тематическую корреляцию и полноту, имеет смысл учитывать только для источников, публикующих документы относительно стабильной тематической направленности.

Тематическую стабильность источника можно определить как корреляцию наборов тематических рубрик, которым соответствуют документы из этого источника в различные периоды времени. Нам представляется, что конкретный набор рубрик мало влияет на предлагаемый ниже метод расчета стабильности источников (под тематической рубрикой в данном случае понимается тематика, семантика которой, в частности, находит свое отражение в виде запроса на информационно-поисковом языке). Предполагается, что документу присваивается та или иная рубрика, если он соответствует определенному запросу. Перечень рубрик и соответствующих им запросов был выбран авторами на основании опыта работы с политематическими новостными ресурсами сети Интернет. Эти рубрики и запросы установлены и апробированы в течение длительного времени в системе контент-мониторинга InfoStream. В настоящее время система включает 35 основных тематических рубрик.

При исследовании тематической направленности некоторых источников информации были обнаружены документы, отклоняющиеся от основной направленности этих источников. Такие документы, если их количество относительно невелико, не должны влиять на рассчитываемый ниже уровень стабильности источников. Конечно, автоматическая рубрикация во многом зависит от качества запросов, однако некоторыми погрешностями в рубрикации при статистическом исследовании можно пренебречь.

Для вычисления уровня стабильности источника информации использовалась формула, основанная на так называемом R/S -анализе [16]. Следует отметить, что этот подход имеет непосредственное отношение к фрактальному анализу, применение которого, однако, выходит за рамки настоящей ра-

боты. R/S -анализ позволяет исследовать «изрезанность» кривой, образуемой временным рядом на основе отношения разброса значений к среднеквадратичному отклонению.

Авторами был предложен параметр тематической стабильности K временного ряда интенсивности публикаций на веб-сайтах (источниках), который выглядит следующим образом:

$$K = \frac{1}{N} \sum_{i=1}^N \frac{S_i}{R_i},$$

где N – количество тем (рубрик) источника;

S_i – среднеквадратичное отклонение по рубрике i ;

R_i – размах значений по рубрике i .

Значение S_i вычисляется по формуле:

$$S_i = \sqrt{\frac{1}{M} \sum_{j=1}^M \left\{ r_{ij} - \frac{1}{M} \sum_{k=1}^M r_{ik} \right\}^2},$$

где r_{ij} – количество вхождения рубрики i за день j , M – количество значений ряда измерения (недель, например).

Значение R_i вычисляется следующим образом:

$$R_i = \max_{1 \leq k < M} X_{ik} - \min_{1 \leq k < M} X_{ik},$$

где X_{ik} – накопленное к моменту k отклонение по рубрике i , вычисляемое по формуле:

$$X_{ik} = \sum_{j=1}^k (r_{ij} - \frac{1}{M} \sum_{l=1}^M r_{il}).$$

На рис. 8. представлена кривая значений коэффициентов стабильности для источников (было измерено поведение свыше 2500 источников за 2007 год), ранжированных по этим значениям.

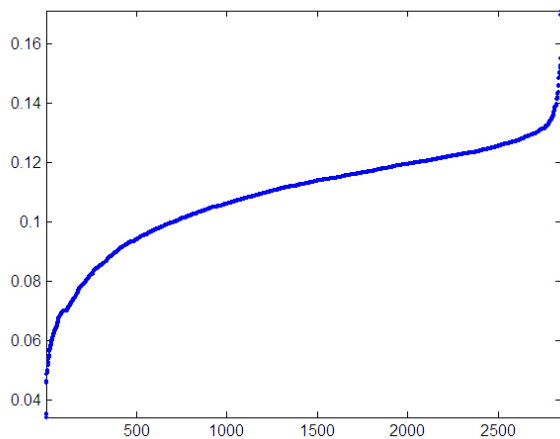


Рис. 8. Ранжированный список источников (ось OX) по параметру тематической стабильности (ось OY)

В частности, самыми тематически стабильными документами (значения правой верхней части диаграммы), оказались периодические профессиональные издания, такие как «Континент Сибирь», «Зеркало недели», «Русский Вестник», «Политический журнал», «Власть денег» и т.п., с определенной периодичностью печатающие постоянное количество сообщений по тематикам, распределенным в

приблизительно в одинаковых пропорциях. Подтвердилась гипотеза о том, что именно профессионализм информационного источника коррелирует с тематической стабильностью. В частности, практически все ведущие информационные агентства, выпускающие политематическую информацию, тем не менее, вошли в состав наиболее тематически стабильных.

Кроме приведенной тематической, исследовалась более простая, диаграмма внетематического распределения источников, ранжированная по коэффициентам стабильности. Полученные данные еще раз подтвердили тот факт, что электронные издания более склонны изменять тематику публикаций, чем свои объемы, выраженные общим количеством публикаций.

6 Заключение

Сегодня становится ясно, что разработка качественно новых средств работы с сетевыми ресурсами переходит в разряд приоритетных задач. В частности, без развитых средств наблюдения за сетевыми информационными источниками невозможно обеспечить соответствующих репрезентативность выборки, а эта задача сегодня является одной из самых актуальных при отборе источников для корпоративных применений систем контент-мониторинга.

Отметим лишь несколько практических применений ранжирования информационных источников. Во-первых, это даст возможность выявления первоисточников информации, например, для размещения в них рекламных материалов, материалов информационного влияния и т.п. Во-вторых, можно сократить затраты времени и средств путем игнорирования, исключения из поиска и анализа заведомо слабых, «шумовых» источников. Кроме того, для оперативного нахождения актуальной информации корректное ранжирование может способствовать нахождению действительно полезных первоисточников и служб интеграции информации.

Результаты данных исследований источников информации могут использоваться при ранжировании выдачи информационно-поисковых систем, подсчете медиа-рейтингов, позволяют рекомендовать пользователям наиболее тематически стабильные и оригинальные источники информации, например, для включения их в список «персональных» в интерфейсах систем контент-мониторинга информационных ресурсов.

Следует отметить, что, несмотря на то, что в данной работе приведено четыре критерия ранжирования источников информации, окончательный «универсальный» критерий не приводится. Теоретически его можно было бы записать, например, как линейную комбинацию приведенных критериев с некоторыми экспертно определяемыми коэффициентами. Однако практика, диктуемая информационными потребностями корпоративных пользователей, показывает, что при выборе источников информации останавливаются на одном из приведен-

ных критериев, дополняя его некоторыми неформальными соображениями (вкусовые предпочтения, учет региональных факторов и т.п.).

В заключение выражаем благодарность А. Григорьеву за организационную поддержку данного исследования, С. Бороденкову и Р. Мазуркевичу и А. Стукаленко за участие в обсуждении затронутых проблем и кропотливую работу с контентом информационных источников.

Литература

- [1] Brin S. and Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7, - 1998.
- [2] Kleinberg J.M. Authoritative sources in a hyperlink environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.
- [3] Kleinberg J.M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, 1999.
- [4] Luo G., Tang Ch., Yu P.S. Resource-adaptive real-time new event detection // In Processing of SIGMOD'07, June 11-14, 2007, Beijing, China.
- [5] Kent, P. Search engine optimization for dummies . Hoboken, Wiley, 2004, 354 p.
- [6] Gianna M. Del Corso, Antonio Gullн, Francesco Romani. Ranking a stream of news. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. – pp. 97–106.
- [7] Брайчевский С.М., Ландэ Д.В. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1, 2005. – № 11. – С. 21–33.
- [8] Bradford S.C. Sources of Information on Specific Subjects // Engineering: An Illustrated Weekly Journal (London), 137, 1934 (26 January), pp. 85–86.
- [9] <http://infostream.ua>
- [10] Broder A., Kumar R., Maghoul F., Prabhakar R., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the web. URL: <http://www.almaden.ibm.com/cs/k53/www9.final/>
- [11] Ландэ Д.В. Структура новостного Web-пространства // Научно-техническая информация. Сер. 2. – М., 2006. – № 8. – С. 17–20.
- [12] Lempel R. and Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKE effect // In Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, 2000. – pp. 387–401.
- [13] Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю. Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды 8-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006. – С. 115–119.
- [14] Ландэ Д.В., Григорьев А.Н., Брайчевский С.М. Стабильность источников информации как один из параметров информационных потоков //

Компьютерная лингвистика и интеллектуальные технологии : труды международной конференции Диалог'2006 – М. : Наука, 2006. – С. 332–334.

[15] Федер Э. Фракталы. – М. : Мир, 1991. – 254 с.

Ranking of information sources in the news monitoring system InfoStream

D.V. Lande, S.M. Brajchevskiy, A.T. Darmokhval,
A.Y. Morozov

A few going is described near ranking and selection of informative sources at the construction of services of news monitoring. These approaches are based on principle of providing of maximal plenitude of information at the least of sources, choice of the most original, thematically stable, maximally quoted sources.

The results of the proper statistical researches are presented.

The principles examined in-process are used for the selection of sources for corporate applications of the news monitoring system InfoStream.

Изучение характеристик сообществ русскоязычной блогосферы

© А.В. Сычев, И.А. Гадебский

Воронежский государственный университет
sav@cs.vsu.ru

Аннотация

В статье представлены результаты расчета характеристик для сообществ блог-хостингов LiveJournal и LiveInternet, полученных на основе атрибутов профилей сообществ. Также приведены графики и таблицы, полученные в результате кластеризации сообществ и их интересов.

Результаты могут быть использованы для решения ряда прикладных задач связанных с поиском сообществ и интересов.

1 Введение

Социальные сети в WWW — интересный феномен, который сформировался за буквально за последние 5 лет. Они объединяют в себе блоги (сетевые дневники), сети медиа-ресурсов, сети персональной информации (MySpace, LinkedIn, Facebook, Мой круг, Одноклассники и др.), системы закладок (del.icio.us), wiki-энциклопедии и другие. Количество пользователей в данных сетях увеличивается с беспрецедентной скоростью, и это вызывает вполне заслуженный интерес к данным сетям со стороны исследователей. В частности, во многих исследованиях рассматриваются вопросы о структуре и свойствах больших длительно развивающихся социальных групп и сообществ в этой электронной среде.

За рубежом проводятся исследования данного феномена по довольно широкому спектру направлений и имеются многочисленные научные публикации, например [1–8].

В [5] представлены результаты изучения изменений, происходящих в блогах в течение всего жизненного цикла, выделены отдельные фазы жизненного цикла сетевого журнала (блога) и соответствующие им индикторы. Применение методов кластеризации для решения таких задач как поиск «горячих» тем, аннотирование и классификация блогов рассматривается в работах [2, 3, 6]. Эффективный алгоритм поиска кластеров терминов из сообщений в блогах, существующих в рамках опре-

деленных временных интервалов предложен в [1]. Там же представлены результаты исследования временных характеристик таких кластеров. В работе [4] блог-сообщества рассматриваются в их структурной и временной динамике. Авторами предложен метод факторизации сообществ (на основе графа сообществ), позволяющий выявлять сообщества, не обнаруживаемые традиционными методами.

Задача поиска «скрытых» друзей по подобию в распределении тематик в их блогах рассмотрена в [7]. Исследованию он-лайн профилей пользователей посвящена работа [8]. В ней на основании экспериментальных данных показано, что пользователи реально используют лишь «тонкий срез» информации из профиля для того, чтобы составить представление о других пользователях. Там же рассмотрены такие характеристики атрибутов профилей как полезность для восприятия, предсказуемость и диагностируемость.

В России исследования по данной тематике носят эпизодический характер и являются чаще всего коммерчески ориентированными. Из числа наиболее известных исследований в России можно упомянуть исследования, регулярно проводимые компанией Яндекс [9] и проект «Пульс блогосферы» (<http://blogs.yandex.ru/pulse/>).

Целью исследования, результаты которого представлены в данной статье, является анализ структуры и связи между атрибутами профилей сообществ в наиболее популярных в России блог-хостингах LiveJournal и LiveInternet, а также поиск эффективных методик обработки данных из профилей сообществ, позволяющих получить дополнительную информацию о сообществах и их интересах.

2 Исходные данные

Для проведения исследования были использованы профили сообществ из блог-хостингов LiveJournal (<http://www.livejournal.com>) и в LiveInternet (<http://www.liveinternet.ru>). В качестве исходных данных был получен список русскоязычных сообществ LiveJournal. Список был извлечен путем скачивания и анализа веб-страниц, на которые указывают гиперссылки со страницы «Реестр русскоязычных сообществ “Живого журнала”» (<http://lj.com.ru>) и «Топ сообществ» LiveInternet (<http://www.liveinternet.ru/top/community/>). Для проведения исследования всего было скачано 2905 профи-

лей сообществ LiveJournal и 35984 профилей сообществ LiveInternet.

3 Хронология создания сообществ

На рис. 1 показана гистограмма, отражающая хронологию развития русскоязычной части блогостинга LiveJournal с момента создания по настоящее время. На гистограмме хорошо прослеживаются сезонные циклы и основные этапы становления и развития данной среды.

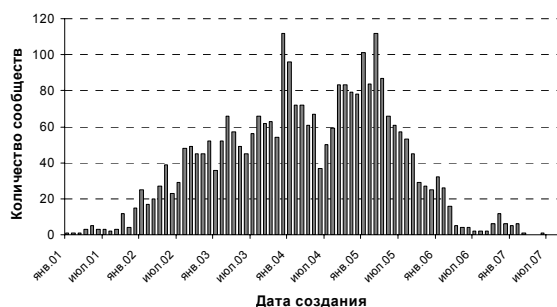


Рис. 1. Распределение сообществ LiveJournal по дате создания

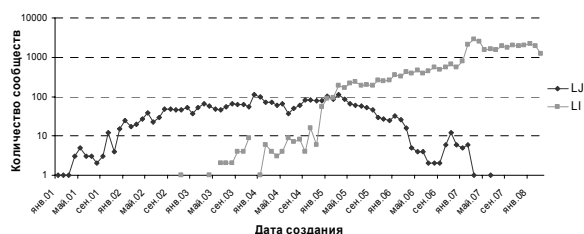


Рис. 2. Распределение сообществ LiveJournal и LiveInternet по дате создания (в логарифмическом масштабе)

Для сравнения на рис. 2 приведены в виде графика данные по хронологии создания сообществ одновременно для LiveJournal и LiveInternet.

На первый взгляд, исходя из этого графика можно сделать вывод об упадке блогостинга LiveJournal (LJ) и его вытеснении LiveInternet (LI) и другими блогостингами. Однако исследование, проведенное компанией Яндекс [1], показывает, несмотря на то, что LJ уступает LI по количеству блогов, преимущество по количеству активных блогов все же на стороне LJ. Сравнение медианных значений для количества записей в таблицах 1 и 3 позволяет сделать вывод о том, что блогосфера в указанных блогостингах находится просто на разных этапах развития.

4 Атрибуты профилей

4.1 LiveJournal

В таблице 1 приведены сводные данные по атрибутам профилей русскоязычных сообществ LiveJournal.

Для каждой пары атрибутов был рассчитан коэффициент корреляции, соответствующие значения корреляции между атрибутами профилей приведены в таблице 2

Из таблицы 2 следует, что наиболее тесно связанными между собой параметрами профиля являются количество членов (4), количество читателей (5), количество записей (11), количество пользователей с правом записи (14).

Таблица 1. Усредненные количественные показатели профилей русскоязычных сообществ в LiveJournal

№	Название поля профиля	Максимум	Среднее	Медиана
1	Количество интересов	150	25	11
2	Количество смотрителей	27	1,8	1,0
3	Количество модераторов	12	0,3	0,0
4	Количество членов	17653	599	195
5	Количество читателей	16191	551	186
6	Входит в сообщества (количество)	29	0,1	0,0
7	Тип аккаунта (0-беспл., 1-улучш., 2-платн.)	2	0,03	0,00
8	Дата создания	22.06.2007	01.04.2004	13.04.2004
9	Дата обновления	10.03.2008	13.10.2007	24.02.2008
10	Количество подарков	5	0,01	0,00
11	Количество записей	71271	1089	193
12	Написано	965	5	0
13	Получено	1335748	10302	627
14	Количество пользователей с правом записи	17650	578	171

Таблица 2. Оценка корреляции между атрибутами профилей русскоязычных сообществ в *LiveJournal*

№	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0,11	0,09	0,10	0,11	0,01	0,05	0,16	0,19	0,03	0,04	-0,05	0,05	0,10
2		0,23	0,31	0,30	0,02	0,16	-0,02	0,13	0,07	0,30	0,01	0,30	0,30
3			0,26	0,29	-0,01	0,11	0,04	0,06	0,04	0,12	0,01	0,16	0,25
4				0,91	0,06	0,19	-0,21	0,17	0,10	0,73	0,08	0,62	0,99
5					0,05	0,15	-0,18	0,18	0,07	0,58	0,06	0,52	0,90
6						0,00	-0,11	0,02	0,00	0,05	0,05	0,03	0,04
7							0,07	0,02	0,00	0,05	0,05	0,03	0,04
8								0,12	-0,01	-0,20	-0,21	-0,15	-0,21
9									0,03	0,12	-0,05	0,08	0,18
10										0,13	0,09	0,14	0,10
11											0,15	0,74	0,73
12												0,13	0,08
13													0,62

Таблица 3. Усредненные количественные показатели профилей русскоязычных сообществ в *LiveInternet*

№	Название поля	Максимум	Среднее	Медиана
1	Дата регистрации	23.03.2008	22.04.2007	04.06.2007
2	Записей в дневнике	20570	60	2
3	Комментариев в дневнике	1002402	256	3
4	Написано сообщений	990725	246	4

4.2 LiveInternet

В таблице 3 приведены сводные данные по атрибутам профилей сообществ LiveInternet.

Для каждой пары атрибутов был рассчитан коэффициент корреляции, результаты сведены в таблице 4.

Таблица 4. Оценка корреляции между атрибутами профилей сообществ в *LiveInternet*

	2	3	4
1	-0,16	-0,05	-0,03
2		0,44	0,44
3			0,99

5 Распределение интересов в сообществах

Изучение интересов сообществ может предоставить исследователю вполне компактную и достаточно точную информацию, позволяющую судить о тематической направленности этих сообществ и предпочтениях пользователей – членов сообществ.

5.1 LiveJournal

В результате обработке профилей сообществ LiveJournal был сформирован список интересов, характеристики которого приведены в таблицах 5–6. Хотя бы 1 интерес был указан в профиле 2260 сообществ. Общее количество интересов получилось равным 43247.

В таблице 5 приведен список 30 наиболее упоминаемых интересов в профилях русскоязычных сообществ LiveJournal.

Таблица 5. Список наиболее распространенных интересов в профилях русскоязычных сообществ *LiveJournal*

№	Интерес	Сообществ
1	музыка	167
2	любовь	120
3		115
4	литература	108
5	фотография	97
6	искусство	86
7	секс	86
8	книги	84
9	история	83
10	кино	83
11	фото	79
12	психология	77
13	красота	76
14	Россия	73
15	люди	68
16	женщины	67
17	философия	67
18	свобода	66
19	творчество	65
20	дети	63
21	путешествия	62
22	поэзия	62
23	music	61
24	дизайн	61
25	природа	60
26	юмор	59
27	жизнь	59
28	общение	59
29	реклама	55
30	интернет	52

В таблице 6 показаны характеристики распределения интересов в русскоязычных сообществах *LiveJournal*, усреднение проводилось по сообществам. Величина **ICW** рассчитывалась как сумма частот интересов, указанных в профиле сообщества. Вес интереса был равен частоте его встречаемости в профилях всех сообществ. Величина **CIC** рассчитывалась как количество интересов из профиля сообщества, указанных также в профиле хотя бы одного другого сообщества.

Таблица 6. Характеристики распределения интересов в русскоязычных сообществах *LiveJournal*

	Взвешенное количество интересов (ICW)	Простое количество интересов (IC)	ICW / IC
Макс.	2486	150	167
Средн.	230	32	9,2
Мед.	120	18	5,7

Корреляция (ICW, IC) 0,63

Корреляция (IC, ICW / IC) -0,16

	Совпадений по интересам (CIC)	Всего интересов в сообществе (IC)	CIC / IC, %
Макс.	137	150	100
Сред.	18	32	59
Мед.	10	18	59

Корреляция (CIC, IC) 0,81

Корреляция (CIC, CIC/IC) -0,21

5.2 LiveInternet

Характеристики распределения интересов в сообществах *LiveInternet* приведены в таблице 7.

Таблица 7. Характеристики распределения интересов в русскоязычных сообществах *LiveInternet*

	Взвешенное количество интересов (ICW)	Простое количество интересов (IC)	ICW / IC
Макс.	59760	402	1598
Средн.	1778	18	131
Мед.	583	7	78

Корреляция (ICW, IC) 0,77

Корреляция (IC, ICW / IC) -0,08

	Совпадений по интересам (CIC)	Всего интересов в сообществе (IC)	CIC / IC, %
Макс.	402	402	100
Сред.	15	18	84
Мед.	6	7	94

Корреляция (CIC, IC) 0.985
Корреляция (CIC, CIC/IC) 0.029

Как видно из сравнения таблиц 6 и 7, в LiveJournal существенно больше доля уникальных интересов, неповторяющихся в других сообществах.

6 Кластеризация сообществ по интересам, указанным в профиле

Для получения более содержательных результатов был проведен ряд экспериментов по кластеризации сообществ и интересов, указанных в профилях сообществ.

С этой целью был использован метод аггломеративной кластеризации Ланса-Уильямса. Первичное расстояние между сообществами рассчитывалось по формуле:

$$\rho(c_1, c_2) = \frac{|c_1 \cap c_2|}{\sqrt{|c_1|} \cdot \sqrt{|c_2|}}$$

Сообщество c_i рассматривалось как множество интересов, указанных в его профиле. При проведении процедуры кластеризации расстояние между кластерами рассчитывалось по формуле среднего расстояния.

При проведении кластеризации интересов расчет расстояния между интересами выполнялся по аналогичной формуле, при этом вместо размера сообщества подставлялся размер множества сообществ, в которых указан данный интерес.

В качестве исходных данных для процедуры кластеризации сообществ (интересов) была использована матрица «сообщество–интерес», на основе которой строилась матрица «сообщество–сообщество» («интерес–интерес»).

Характеристики матриц приведены в таблицах 8–10.

При проведении кластеризации по интересам в связи с ограничениями вычислительного характера учитывались только интересы, указывавшиеся в двух и более сообществах, т.е. фактически кластеризация выполнялась на прореженных матрицах.

Таблица 8. Исходные данные для построения матрицы «Сообщество–Интерес»

Блог-хостинг	Сообществ всего	Сообществ, в которых указаны интересы	Макс. кол-во интересов в профиле сообщества	Интересов (во всех сообществах)	Кол-во интересов, указанных более чем в 1 сообществе
LiveJournal	3083	2260	150	43245	9416
LiveInternet	35983	11200	402	43398	18542

Таблица 9. Характеристики матриц «Сообщество–Интерес» и «Сообщество–Сообщество» (без прореживания)

Блог-хостинг	Параметр	Матрица «Сообщество–Интерес»		Матрица «Сообщество–Сообщество»	
		Агрегирование по сообществам	Агрегирование по интересам	Число остальных сообществ, с которыми есть пересечение	Средняя степень пересечения сообщества с остальными сообществами, %
LiveJournal	Максимум	150 (интересов)	167 (сообществ)	638	80
	Среднее	31,7	1,7	66,0	6
	Медиана	18	1	33	5
	Сообществ (Интересов)	2260	43245	2065	
	Разреженность матрицы, %	99.93		96.81	
LiveInternet	Максимум	402 (интересов)	1598 (сообществ)	7602	100
	Среднее	17,6	4,2	465,5	12
	Медиана	7	1	188	10
	Сообществ (Интересов)	11200	47398	10488	
	Разреженность матрицы, %	99.96		95.56	

Таблица 10. Характеристики матриц «Сообщество-Интерес» и «Интерес-Интерес» (с прореживанием)

Блог-хостинг	Параметр	Матрица «Сообщество-Интерес» (прореженная)		Матрица «Интерес-Интерес» (прореженная)	
		Агрегирование по сообществам	Агрегирование по интересам	Число остальных интересов, с которыми есть пересечение	Средняя степень пересечения интереса с остальными интересами, %
LiveJournal	Максимум	139 (интересов)	167 (сообществ)	2657	100
	Среднее	17,7	4,02	68,08	49
	Медиана	10	2	24	50
	Сообществ (Интересов)	2142	9416	8867	
	Разреженность матрицы, %	99.81		99.23	
LiveInternet	Максимум	402 (интересов)	1598 (сообществ)	10374	100
	Среднее	15,6	9,09	228,07	41
	Медиана	6	3	38	40
	Сообществ (Интересов)	10829	18542	16600	
	Разреженность матрицы, %	99.92		98.63	

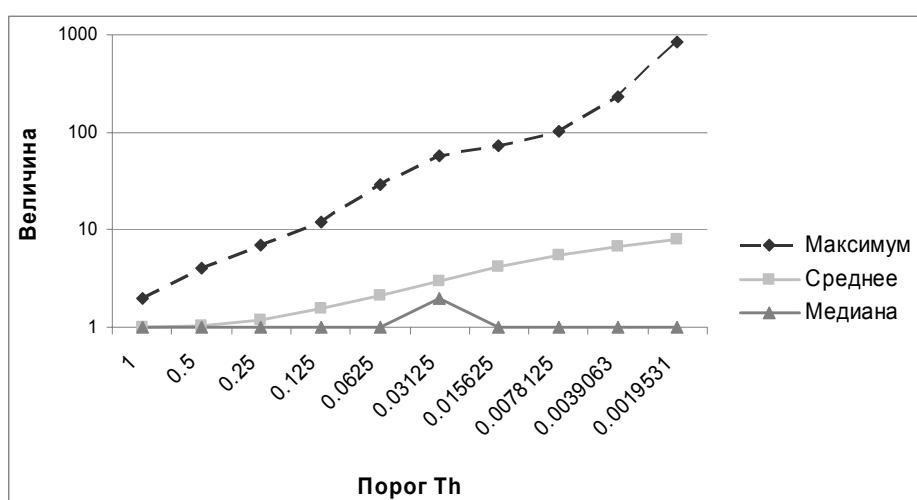


Рис. 3. Характеристики кластеров сообществ *LiveJournal*

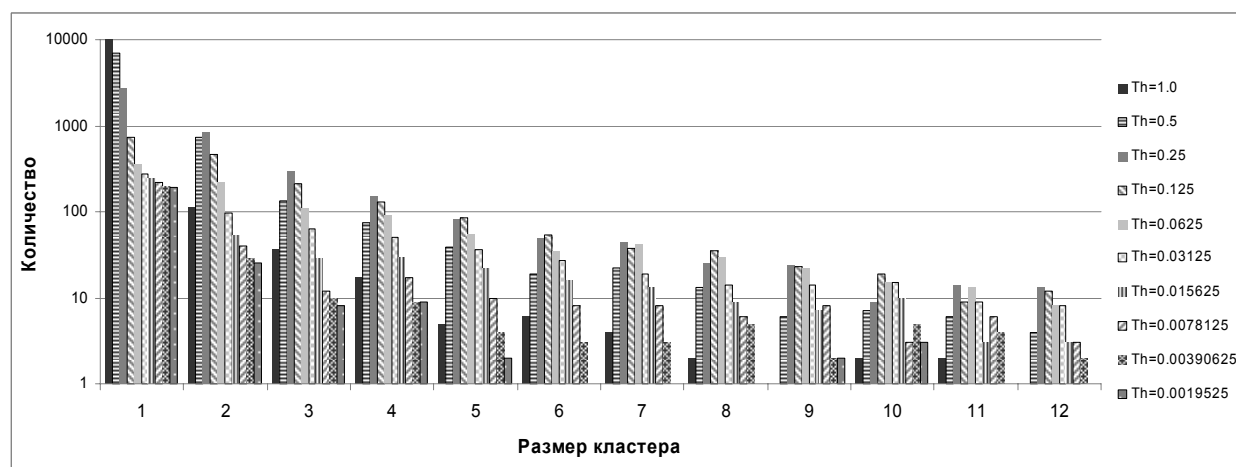


Рис. 4. Распределение кластеров сообществ *LiveInternet* при различных значениях порога кластеризации Th

6.1 Кластеризация сообществ

На основе матриц «Сообщество–Сообщество» были получены кластеры сообществ при различных значениях порога кластеризации Th . Характер зависимости размера кластера (максимум, среднее и медиана) от порога Th для LiveJournal представлен на рис. 3. Для LiveInternet была получена зависимость похожей формы, причем, в обоих случаях наиболь-

шее значение медианы достигается для значения порога в районе от 0.02 до 0.05. Распределение кластеров сообществ в зависимости от порога кластеризации для LiveInternet приведено на рис. 4. На гистограмме размер кластера ограничен 12 сообществами, поскольку самая существенная часть сообществ оказалась сконцентрированной в кластерах с размером в диапазоне от 1 до 12.

Таблица 11. Сообщества *LiveInternet* из наибольшего кластера (для порога $Th = 0.5$)

№	Название	№	Название	№	Название
1	___ДИЗАЙН___	21	HappyTreeFriends_4ever	41	вСе_ЛуЧшЕе_ЗдЕсЬ
2	-_Картинки_-	22	JannLissssy_Design	42	Всё_самое_лучшее_для_днева
3	_-Pictures-only-for-you-_	23	Lilac_style	43	Всевозможные_аватары
4	_ЗаКаЖи_АвАтАр_	24	-Making_avatars-	44	выполняем_заказики
5	_МеГо_КаРтИнКи_	25	mega_art_design	45	Делаем_аватары_на_заказ
6	_ТВОЁ_	26	MP_STUDIO	46	ДиЗайн
7	A_F_Y_D	27	Pics_and_Diz	...	
8	Anything_For_Your_Diary	28	PiCtUrEs_sTyLe	67	Просто_така
9	Avatarki_special_for_your	29	Pretty_Blogs	68	РисОваННые_ДиЗайны
10	AvatarLand_BG_	30	THE-BEST-PICTURES	69	Смени_ДиЗ
11	-Avatars-	31	-Voncouver-	70	собрание_картинок
12	Avatars_for_you_diary	32	we_make_designs	71	сообществО_х
13	Beauty_Diary_4_u	33	Your_orders	72	Фон_для_тебя
14	design_for_diarys	34	Аватары_для_Жизни	73	Фоны_для_дневников
15	Designes_By_-Rapsody-	35	авафонэпик	74	Фоны_и_эпиграфы_Ап
16	Dizzziki	36	Беру_на_Ли-ру	75	Фоны_Эпиги_Авики
17	Ducks	37	Всё_Лучшее_Для_Дневников	76	фсё_для_днеффничка
18	FASHOIN_PICTURES	38	Все_для_Вашего_днева-	77	хХСмотри_сукаХх
19	fine_arts_group	39	Все_для_ДневничкОв	78	Цветной_Клан
20	ForYourDiary	40	Все_картинки_здесь	79	Юзер_пиг_и_Ко

Таблица 12. Интересы, указанные в профилях сообществ из таблицы 11

№	Интерес	CF	ICF	PF	CF-ICF	№	Интерес	CF	ICF	PF	CF-ICF
1	фоны	73	218	364	0,821	20	lilac	1	1	1	0,031
2	эпиграфы	70	196	315	0,815	21	purple	1	1	1	0,031
3	аватары	70	393	699	0,633	22	violet	1	1	1	0,031
4	картинки	51	342	831	0,488	23	сиреневый	1	1	2	0,031
5	дизайны	21	116	186	0,285	24	фотошоп ...	1	1	1	0,031
6	заказы	9	30	59	0,168	25	схемы оформле- ния	1	1	2	0,031
7	дизайн	11	205	463	0,126	26	велкомы	1	1	1	0,031
8	фотошоп	6	144	307	0,077	27	заголовки	1	1	1	0,031
9	готовые дизайны	3	7	11	0,072	28	фотографии.	1	1	3	0,031
10	фотографии	4	105	369	0,056	29	глиттер	1	1	1	0,031
11	подписи	2	4	6	0,052	...					
12	поиск картинок	2	4	10	0,052	61	эмо	1	89	341	0,015
13	схемы	2	8	22	0,047	62	рисунки	1	91	224	0,014

14	оформление	2	8	27	0,047	63	гламур	1	97	288	0,014
15	анимации	2	12	29	0,044	64	интернет	1	99	335	0,014
16	дневники	3	138	288	0,039	65	критика	1	129	222	0,013
17	сообщества	2	31	95	0,037	66	стихи	1	147	507	0,013
18	создание аватарок	1	1	1	0,031	67	аниме	1	250	560	0,011
19	happy tree friends	1	1	9	0,031	68	любовь	1	267	1048	0,010

В таблице 11 приведен в качестве примера приведен наибольший по размеру кластер сообществ *LiveInternet*, сформированный при значении порога $Th = 0.5$. Интересы, содержащиеся в профилях этих сообществ, представлены в таблице 12. Для каждого интереса также указаны: **CF** – как часто встречается интерес в профилях сообществ, образующих кластер, **ICF** – количество других кластеров, содержащих сообщества с этим интересом, **PF** – частота встречаемости интереса в профилях всех сообществ, **CF-ICF** – метрика, аналогичная TF-IDF, и показывающая специфичность интереса для данного кластера.

Расчет показал, что коэффициент корреляции между **CF-ICF** и **PF** равен 0.28, что указывает на возможность применения метрики **CF-ICF** для ранжирования кластерообразующих интересов. Процедура кластеризации может быть использована как

для поиска латентных «суперсообществ», так и для выявления множества интересов, определяющих общую тематику таких «суперсообществ». Варьируя величину порога кластеризации Th , можно задавать соотношение между полнотой и точностью данного множества.

6.2 Кластеризация интересов

На основе матриц «Интерес-Интерес» были получены кластеры сообществ при различных значениях порога кластеризации Th .

Характеристики кластеров представлены на рис. 5–6. В таблицах 13–14 приведены интересы из наибольшего кластера и соответствующие им сообщества.

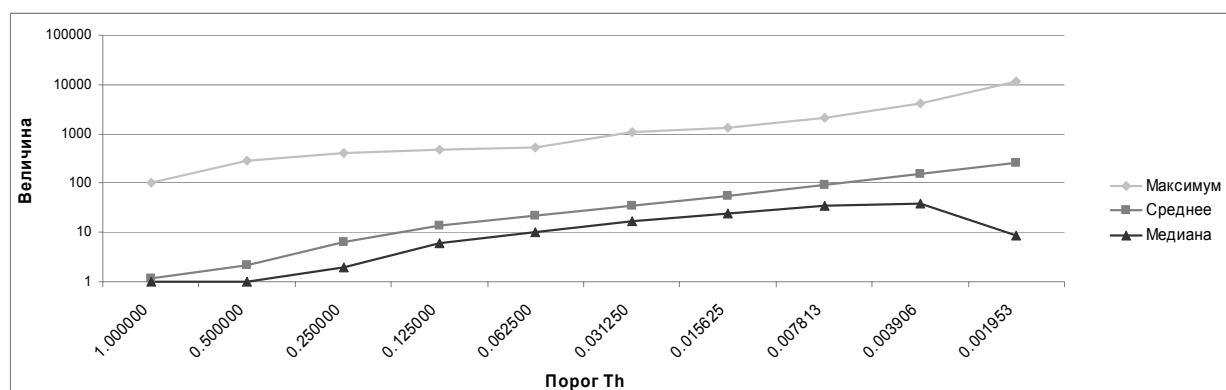


Рис. 5. Характеристики кластеров интересов *LiveInternet*

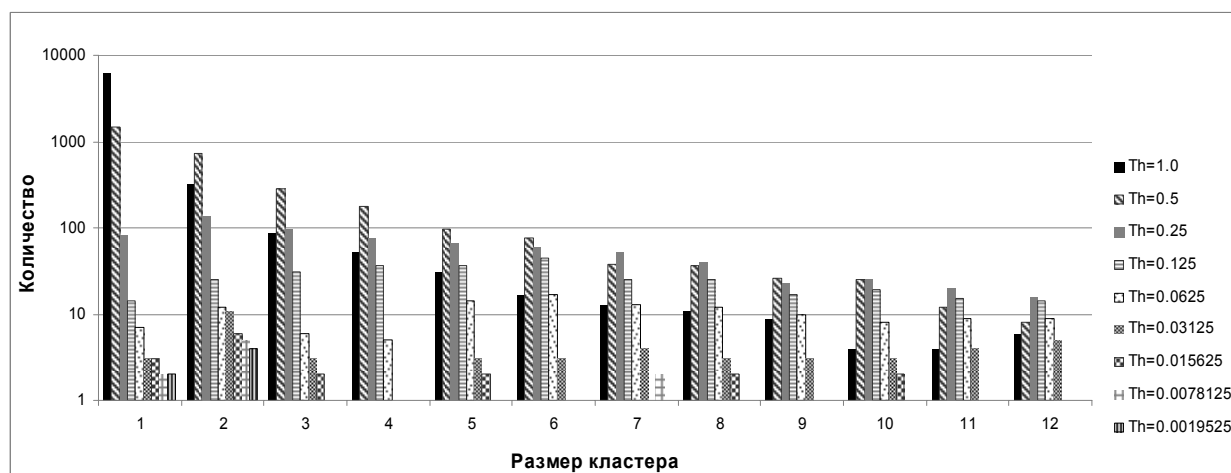


Рис. 6. Распределение кластеров интересов *LiveJournal* при различных значениях порога кластеризации Th

Таблица 13. Интересы *LiveJournal* из наибольшего кластера (для порога $Th = 0.5$)

№	title	№	title	№	title	№	title
1	alfa romeo	21	hyundai	41	renault	61	ВАЗ
2	audi	22	isuzu	42	rover	62	вождение
3	auto	23	jaguar	43	saab	63	ГАЗ
4	automobile	24	jeep	44	seat	64	ГАИ
5	bmw	25	kia	45	skoda	65	гаишники
6	buick	26	lancia	46	subaru	66	ГИБДД
7	cadillac	27	land rover	47	suzuki	67	давить на газ
8	car	28	lexus	48	toyota	68	дворники
9	cars	29	lincoln	49	volkswagen	69	джип
10	chevrolet	30	mazda	50	volvo	70	дизель
11	chrysler	31	mercedes	51	vw	...	
12	citroen	32	mitsubishi	52	авария	110	схождение
13	daewoo	33	mustang	53	авто	111	сцепление
14	daihatsu	34	nissan	54	автозапчасти	112	теория
15	dodge	35	opel	55	автомобиль	113	топливо
16	fiat	36	peugeot	56	автопром	114	тормоз
17	ford	37	pontiac	57	автосервис	115	тормоза
18	geo	38	porsche	58	АЗС	116	тормозная жидкость
19	headlights	39	porsche	59	аккумуляторы	117	УАЗ
20	honda	40	range rover	60	бензин	118	фаркоп

Таблица 14. Сообщества, указанные в профилях сообществ из таблицы 13

№	nick	CF	№	nick	CF
1	avto_piter	120	11	ru_fl	3
2	spb_auto	112	12	ru_luxury	3
3	ru_auto	111	13	italian_cars	2
4	ru_auto_shop	43	14	mountains	2
5	ru_rightcars	9	15	ru_ford	2
6	avto_ru	8	16	ru_hyundai	2
7	_megafon_	4	...		...
8	megafondv	4	44	strana_sovetov	1
9	moto_ru	3	45	su_tormoz	1
10	ru_4x4	3	46	surgut	1

7 Заключение

Изучение сообществ русскоязычной блогосферы – обширная и многогранная задача. В данной статье представлены некоторые результаты, касающиеся одного из аспектов общей задачи. На основе информации, представленной в профилях сообществ наиболее популярных блог-хостингов LiveJournal и LiveInternet в рамках проведенного исследования решались такие задачи как реконструкция хронологии создания сообществ, вычисление корреляции между атрибутами профилей. Наибольшее внимание в работе было уделено одному из самых информативных и значимых атрибутов – интересам сообще-

ства. Был изучен характер распределения интересов в профилях как путем расчета простейших числовых характеристик, так и с помощью процедуры кластеризации. Последняя может быть эффективно использована для решения таких прикладных задач как поиск латентных «суперсообществ» и определения их тематики, автоматического структурирования пространства интересов, автоматической оценки тематической принадлежности и специфичности интересов, а также для ряда других интересных задач.

Литература

- [1] N.Bansal, F.Chiang, N.Koudas, F W.Tompa. Seeking Stable Clusters in the Blogosphere. VLDB 2007, Vienna. pp. 806–817.
- [2] C.H.Brooks, N.Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proc. of WWW, 2006, pp. 625–632.
- [3] A.Qamra, B.L.Tseng, E.Y.Chang. Mining blog stories using community-based and temporal clustering. In Proc. of the 15th CIKM Conference, 2006. pp. 58–67.
- [4] Y.Chi, S.Zhu, X.Song, J.Tatemura, B. L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In: Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007. pp. 163–172.
- [5] D.Gurzik and W.G.Lutters. From the personal to the profound: understanding the blog life cycle. CHI 2006, April 22–27, 2006, Montreal, Quebec, Canada. ACM Press, New York: 2006. pp. 827–832.
- [6] B.Li, S.Xu, J.Zhang. Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments. In Proc. ACMSE'07, March 23-24, ACM.
- [7] D.Shen, J.Sun, Q.Yang, Z.Chen. Latent Friend Mining from Blog Data. In: 6th International Conference on Data Mining, Hong Kong, China, 2006. pp. 552–561.
- [8] K.Stecher, S.Counts. Thin Slices of Online Profile Attributes. In Proc. of International Conference on Weblogs and Social Media – 2008. [Electronic resource]. – Mode of access: research.microsoft.com/~scottlt/pubs/ICWSM_Thin_Slices.pdf.
- [9] Состояние блогосферы российского интернета. По данным поиска по блогам Яндекса. Весна 2008 г. [Электрон. ресурс] – 2008. – Режим доступа: http://download.yandex.ru/company/yandex_on_blogosphere_spring_2008.pdf

Russian Blogspace Community Features Research

A.V. Sychev, I.A. Gadebskij

The paper is dedicated to the Russian blogspace research. Some numerical features related to community profiles (hosted in the most popular in Russia blogspaces www.livejournal.com and www.liveinternet.ru) are presented. The diagram of communities creation chronology is provided. The correlation factor between pairs of community profile attributes is estimated. The clustering of communities and its interests is illustrated by tables, histograms and graphics. This procedure could have important practical applications like detecting latent supercommunities and its topic distillation, interest space automatic structuring and others.

Подход к выявлению подмножеств похожих документов

© А.В. Антонов, С.Г. Баглей, В.С. Мешков

Корпорация «Галактика»
{alexa, baglei, meshkov}@galaktika.ru

Аннотация

В статье описывается подход к определению подмножеств похожих документов, реализованный в поисково-аналитической системе Галактика-Zoom. Метод основан на работе со статистико-лингвистическими данными, составляющими Информационный портрет (ИнфоПортрет), формируемый в системе Галактика-Zoom. ИнфоПортрет формируется в соответствии каждому документу в заданном подмножестве документов. Элементами ИнфоПортрета являются языковые инварианты (слова и пары слов), статистически отличающие данную выборку документов. Основная задача, которая решается с помощью описываемого подхода – разбиение заданного множества документов на подмножества похожих документов и подмножества уникальных документов. В качестве критерия близости в подмножествах выбрано расстояние Дженсена–Шеннона.

1 Введение

Общей проблемой, снижающей эффективность работы пользователя с поисковой системой, является избыточность информации при выдаче результатов по запросу. Причинами возникновения избыточности могут являться, например, нечетко сформулированные запросы к поисковой системе, омонимичность элементов поисковых запросов, а также другие.

Задача уменьшения избыточности, может решаться различными способами. Один из наиболее действенных способов снижения избыточности – диалог пользователя с системой, то есть, режим, при котором пользователю предоставляется возможность уточнения своих информационных предпочтений.

Как часть диалога пользователя с системой может рассматриваться формирование подмножеств похожих документов, которое позволяет пользователю рассматривать некоторое подмножество в качестве некоторой отдельной сущности. Работая с полученными подмножествами похожих документов, пользователь может более быстро восприни-

мать данные, с которыми он работает.

Исследования проводились в рамках моделей, использующих расстояния *Kullback-Leibler* [6] или – *Jensen-Shannon* [5] между двумя распределениями признаков. Описание применяемой в системе модели приведено в работе [4]. Мы использовали этап формирования ИП – множества языковых инвариантов (слов и словосочетаний) для выявления признаков, характеризующих каждый документ из рассматриваемого множества. Выбор наиболее тематически значимых слов для работы метода являлся интуитивным, и основывался на наших предыдущих работах в этой области. Кроме того, мы учитывали необходимость сохранения быстродействия в работе системы. Для этой цели требовался этап предварительной фильтрации признаков-слов и словарных пар, характеризующих отдельный документ. Естественным шагом было использование в качестве смыслового фильтра существующий в системе механизм построения Информационного Портрета документа.

В качестве меры близости документов рассматривается расстояние между распределениями весов элементов (слов и/или словосочетаний) их ИП. Опорным считается ИП наибольшего размера, то есть, такой ИП, в котором количество составляющих его элементов максимально. Для формирования подмножеств документов был выбран порог. Похожими считаются документы, у которых расстояние не меньше заданного.

Построение подмножеств документов осуществляется с помощью утилиты сравнения документов выборки и измерения расстояния между ними.

Галактика-Zoom представляет собой поисково-аналитическую систему обработки больших объемов неструктурированных данных. Подробно архитектура, принципы работы, характеристики системы описаны в работах [1, 2].

Основным объектом в системе Галактика-Zoom является понятие Информационного портрета выборки документов (ИнфоПортрета). ИнфоПортрет представляет собой список языковых инвариантов (слов и словосочетаний), отличающих данную выборку от прочих. Технология построения информационного портрета, детально описанная в работах [2, 3, 4], основана на статистических методах обработки текстовой информации. Используя характеристики элементов сформированного ИнфоПортрета и собственной статистики документа, возможно формирование информационного портрета отдель-

ных документов. То есть, для каждого документа система формирует список слов и словосочетаний, статистически отличающих данный документ от прочих в выборке. ИнфоПортрет представляет собой информацию, описывающую содержание документа в целом.

Система позволяет не прибегать к условию равнозначности слов, то есть существует возможность обработки весовых значений элементов, составляющих ИП. Мы использовали следующие преимущества предоставляемые системой с точки зрения задачи поиска похожих документов:

- возможность получения величины относительной значимости слов и словосочетаний для документа;
- возможность упорядочивания значимых слов и словосочетаний в документе исходя из величины их относительной значимости в выборке.

2 Постановка задачи

2.1 Входные данные

$D = \{d_i\}, i = \overline{0, n-1}$ – множество документов, загруженных в базу.

$P = \{p_i\}, i = \overline{0, n'-1}$ – множество ИП, соответствующих документам из D .

S – вектор номеров документов из множества D , упорядоченный по убыванию размера ИП.

$C = \{C_i\}, i = \overline{0, k}$ – подмножества похожих документов базы.

U_0 – список документов базы, у которых отсутствует ИП.

U – список уникальных документов.

2.2 Промежуточные данные

S^1 – вектор номеров документов из множества D , упорядоченный по убыванию размера соответствующего им ИП.

S^w – рабочий вектор номеров документов, упорядоченный по убыванию размера соответствующего им ИП, к которому применяется алгоритм построения подмножеств похожих документов.

2.3 Выходные данные

$P^1 = \{p_i\}, i = \overline{n, n+m'-1}$ – множество ИП документов.

S – вектор номеров документов упорядоченный по убыванию размера соответствующего им ИП.

$C^1 = \{C_i^1\}, i = \overline{0, k_1}$ – подмножества похожих документов.

U_0^1 – список документов, у которых отсутствует ИП.

U^1 – список уникальных документов базы.

2.4 Алгоритм построения подмножеств похожих документов (ППД) на векторе S^w

Первый документ из вектора порождает первое подмножество похожих документов и сравнивается со всеми последующими, основываясь на расстояниях Kullback-Leibler или Jensen-Shannon, задающих меру подобия ИП (оценку близости документов). Если данная мера подобия для какого-нибудь ИП оказывается выше заданного порогового значения, то текущий документ приписывается первому подмножеству. Сравнение продолжается, пока не исчерпывается вектор документов.

После этого происходит обработка следующего документа, не вошедшего в первое подмножество, с которым последовательно сравниваются все последующие актуальные документы и т.д.

В результате формируется некоторое неизвестное заранее количество подмножеств похожих документов.

1. Инициализация. Построение и последующее сохранение ИП каждого документа из множества D^1 . Формирование и сохранение вектора S^1 . Документы, у которых отсутствуют ИП, добавляются в список U^0 .

2. Множество D^1 разбивается на два непересекающихся подмножества: D_+^1 – документы, размер ИП которых не меньше наибольшего размера ИП в векторе S^0 и D_-^1 – документы с меньшими размерами.

3. Если подмножество D_+^1 не пустое, то формирование вектора S^w из документов множества D_+^1 , центроидов из подмножеств C похожих документов и из списка U уникальных документов. Применение алгоритма ППД к документам из множества D_+^1 как возможным новым центроидам. Если центроид из C приписывается к новому подмножеству, то для всех документов похожих на него пересчитывается мера близости относительно нового центроида.

4. Если остались уникальные документы из подмножества D_+^1 или подмножество D_-^1 не пустое, то формирование вектора S^w из уникальных документов подмножеств D_+^1 , D_-^1 и центроидов подмножеств похожих документов. Применение алгоритма ППД к уникальным документам из подмножества D_+^1 и центроидам с возможным пополнением их новыми документами из D_-^1 .

5. Если после выполнения шагов 3 и 4 остались уникальные документы из множества D^1 , то формирование вектора S^w из всех оставшихся уникальных документов. Применение алгоритма ППД к каждому уникальному документу из множества D^1 .

Примечание. В процессе выполнения шагов 3, 4 и 5 формируются подмножества C^1 и список U^1 .

3 Результаты экспериментов

Для оценки качества работы метода нами было проведено исследование на экспериментальной базе документов. В качестве наполнения базы использовался массив, состоящий из нормативно-справочных документов. Моделировалась ситуация проведения разбиения определенной выборки на множества похожих и множество уникальных документов.

Таблица 1. Основные характеристики базы документов

Параметр	Количество
Документов в базе	5000
Слов в базе	63100
Словомест в базе	2008235
Словосочетаний в базе	13621
Мест словосочетаний в базе	172383

Рассматривая работу алгоритма поиска похожих документов как часть функциональности поисковой системы Галактика-Zoom, в качестве элементов исходного набора данных были определены документы тестовой выборки из базы, заданной случайным образом.

В качестве модели документа мы также использовали ИнфоПортрет с соответствующими ему характеристиками, формируемый в системе Галактика-Zoom. Далее приведены основные параметры выборки и полученных подмножеств похожих документов.

Таблица 2. Основные параметры полученных подмножеств

Количество подмножеств	31
Количество документов в выборке	523
Количество документов, включенных в подмножества	426
Количество документов, общих для подмножеств	0
Минимальное число документов в подмножестве	3
Максимальное число документов в подмножестве	21
Минимальное число объектов в подмножестве	3
Максимальное число объектов в подмножестве	10

Анализ полученных результатов показал, что в результате работы алгоритма получено достаточно плотное покрытие исходного массива данных непересекающимися подмножествами. Параметры полноты и точности формирования подмножеств вполне приемлемо для эффективного практического использования алгоритма.

4 Заключение

Мы применили алгоритм построения подмножеств похожих документов, основанный на измерении меры близости между отдельными документами выборки. Основная цель применения алгоритма – представление однородной или близкой к однородной части исходных данных в удобном виде для восприятия пользователем системы.

Экспериментальная проверка показала, что получены приемлемые результаты в части полноты и точности формируемых подмножеств. Таким образом, использование метода оказалось оправданным.

Литература

- [1] Антонов А. Методы классификации и технология Галактика-Zoom // Международный форум по информации, Москва, ВИНТИ, 2003. т. 28.
- [2] Антонов А, Курзинер Е. Автоматическое выделение предметной области большого необработанного текстового массива // Компьютерная лингвистика и интеллектуальные технологии : Труды Международного семинара Диалог-2002.
- [3] Антонов А. Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации // Сб. ВИНТИ. – 2001, №8.
- [4] Антонов А., Мешков В. Современные проблемы поисковых систем и некоторые пути их преодоления (Сер. «Аналитика-Капитал»). – М., 2000.
- [5] Fuglede, B., Topsøe, F. Jensen-Shannon Divergence and Hilbert space embedding. // University of Copenhagen, Department of Mathematics, <http://www.math.ku.dk/~topsøe/ISIT2004JSD.pdf>.
- [6] Kullback, S., Leibler, R. // On information and sufficiency, Annals of Mathematical Statistics 22:79–86, 1951.
- [7] Антонов А., Баглей С., Мешков В., Суханов А. Кластеризация документов с использованием метаинформации // Компьютерная лингвистика и интеллектуальные технологии : Труды Международного семинара Диалог-2006.

An Approach to Detection of Similar Documents Subsets

Alexander Antonov, Stanislav Bagley,
Valentin Meshkov

In this paper, we propose an approach to similar documents detection. We consider automatic decomposition of an initial document set into a collection of non-overlapping subsets of similar documents and a subset of independent documents as the result of an approach work. The model underlying the approach uses statistical information about words and pairs of neighboring words in texts. To measure the proximity between documents we use Jensen-Shannon divergence applied to that statistical information. We

give an example of test run of an approach in the paper performed on a collection of legal documents.

Обнаружение ссылочного спама

© Шарапов Руслан Владимирович

© Шаропова Екатерина Викторовна

Муромский институт (филиал) Владимирского государственного университета
info@vanta.ru

Аннотация

В статье рассматриваются подходы к выявлению ссылочного спама на основе анализа содержания страницы. Основное место в работе посвящено выявлению рекламных (платных) ссылок. Анализируются признаки, характерные для рекламных ссылок. Дается алгоритм выявления спам-ссылок и приводятся результаты его работы.

1 Введение

Количество информации, доступной пользователям глобальной сети Интернет, с каждым годом становится все больше и больше. Постоянно растет число сайтов, увеличивается число их страниц. Например, поисковая система Яндекс, на момент написания статьи, осуществляла поиск по 17 миллионам сайтов и 4 миллиардам веб-страниц [13]. Интернет становится не только средством получения информации и общения, но и средством ведения бизнеса. Естественно, нахождение сайта по наиболее популярным поисковым запросам и положение сайта в списке результатов поиска является актуальной проблемой для большинства владельцев сайтов. В связи с тем, что сайтов одной тематики иногда бывает слишком много, и каждый хочет быть на вершине списка результатов поиска по ключевым запросам, владельцы начинают прибегать к различным ухищрениям, чтобы поднять свои сайты как можно выше к началу списка. Простым выходом кажется применение различных технологий манипулирования поисковыми системами, например с помощью поискового спама. Существует большое количество методик, используемых для того, чтобы ввести в заблуждение поисковые системы [8]. Рассмотрим одну из них – ссылочный спам.

Увеличение числа ссылок на сайты стало одним из основных методов манипулирования поисковыми системами в последнее время. Масштабы манипуляции постоянно растут. Если несколько лет назад основным способом являлся так называемый обмен ссылками, который проводился вручную, то теперь ему на смену пришли различные способы автоматического размещения ссылок. Можно выделить не-

сколько вариантов такого размещения:

1. Использование специализированных программ для автоматического добавления ссылок в каталоги, гостевые книги, форумы и т.д.
2. Покупки ссылок у рекламных брокеров.

С первым вариантом поисковые системы научились бороться, выявляя ресурсы, где есть возможность простого, немодерируемого добавления ссылок. Вес ссылок с таких ресурсов сильно снижается. Размещение же ссылок с использованием рекламных брокеров представляет для поисковых систем много большую проблему.

В настоящее время в русском сегменте интернет действует около десятка крупных рекламных брокеров, занимающихся продажей текстовых ссылок. Только один из них, Sape.ru, имеет возможность размещать ссылки на более чем 35 миллионах страниц. Несмотря на то, что ссылки в таких системах называют «рекламными», их основная цель – не реклама с целью привлечения посетителей (ссылки часто размещаются в самых неприметных местах страницы и пользователь их просто не замечает), а улучшение своего положения в поисковых системах. Стоимость такой «рекламы» также часто бывает номинальной, иногда всего 0.01\$ за месяц размещения. Ссылки, размещенные с помощью рекламных брокеров, будем называть «платными», подчеркивая что ссылки имеют искусственное происхождение (т.е. не имеют никакого отношения к содержанию страницы). Они размещаются владельцем страницы за деньги, а не из «уважения» к сайту, на который ссылаются.

В чем же основная опасность крупномасштабного ссылочного спама, наблюдаемого последние несколько лет? Опасность заключается в том, что ссылки активно используются современными поисковыми системами для ранжирования результатов поиска. Со ссылками связано и понятие Индекса цитируемости в Яндекс и определение PageRank в Google. Массовое увеличение ссылок неестественного происхождения (ссылочного спама) может сильно «испортить» эффективность их работы. Ситуация осложняется тем, что «платные» ссылки могут размещаться на любых сайтах, в том числе и на очень уважаемых и популярных ресурсах. Таким образом, становится невозможным простое деление страниц на «хорошие» и страницы для ссылочного спама.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Агентство <http://www.tretyakoff-realty.ru/> дает ипотечный кредит под 9 процентов.

Агентство [даст ипотечный кредит под 9 процентов.](#)

Рис. 4. Пример ссылки с текстом по краям

Часто ссылки указывают на ресурс с достаточно общей тематикой (например, при ссылках на источник новостей или сайт автора какой-либо статьи).

Нужно заметить, что, согласно статистике цифровых фотокамер. Так, например, в фотоаппаратов, что эквивалентно 19% от общ конкурент Canon, в 2007 году распродала око.

Аналитики IDC также отмечают, что доля составила 42,7%. Компания Canon в пр ближайший конкурент, фирма Nikon - около 11

Источник: www.astera.ru

Рис. 5. Пример ссылки на источник

Для правильного определения тематики ссылки может помочь глубокий анализ тематика сайта, на который ведет ссылка. Понятно, что задача эта трудоемка и требует большого количества времени.

6. Тематическая близость соседних ссылок. Для этого необходимо проанализировать тематику группы ссылок, размещенных на странице. Если ссылки не являются тематическими и имеют явно выраженный разброс тематики, то они – рекламные.

7. Место расположения ссылок. Для этого необходимо проанализировать расположение ссылок на странице. Чем дальше ссылки от основного содержания страницы, тем более вероятно, что они являются рекламными. Например, часто такие ссылки размещаются внизу страницы или в правом столбце, когда основной текст располагается посередине.

8. Код ссылок. Многие автоматизированные системы установки ссылок (биржи, обменники, брокеры) устанавливают код автоматически по шаблону. Наличие блока идентичных по коду ссылок может указывать на их спамерское происхождение.

9. Динамичность/Время жизни ссылок. Частое изменение ссылок на страницах без изменения остального содержания может свидетельствовать о неестественном их происхождении. Ссылки могут либо просто на время исчезать со страниц (в случае неполадок систем по автоматическому размещению ссылок), либо часть их может заменяться на новые.

10. Сообщение о платных ссылках. О платных ссылках могут сообщить конкуренты, бывшие покупатели ссылок, бывшие сотрудники и т.д.

11. Просмотр страницы человеком. Просмотр страниц модератором и выявление ссылочного спама вручную.

4 Алгоритм обнаружения ссылочного спама

Теперь рассмотрим алгоритм, способный выявить спамерские ссылки. Он состоит из нескольких этапов.

Этап 1. Формирование предварительного набора спам-ссылок S . Набор формируется из следующих ссылок:

- выбранных вручную;
- определенных алгоритмом ранее, как спам;
- определенных на основе анализа кода рекламных брокеров для автоматического размещения ссылок.

Наибольший интерес представляет именно последний способ. У некоторых рекламных брокеров можно обнаружить отличительные особенности в размещении кодов, которые могли бы помочь их идентифицировать. Например, при автоматическом размещении кода Prospero.ru можно заметить следующее:

```
<a class=prospero
href="http://www.logipark.ru">таможное оформ-
ление Янония</a>
<a class=prospero
href="http://www.svadbaexclusive.com/">ЗАГСы Мо-
сквы, организация свадьбы в Москве</a>
```

При кэшировании рекламных ссылок иногда можно увидеть такой код:

```
<!--from cache 14:18:25 13.04.2008-->
<a href="http://www.clinicsex.ru/"
target=_blank>цитомегаловирус затем гарднерел-
лез анализы мочи</a>
<a href="http://zemnovosti.ru"
target=_blank>Статьи земельная тематика</a>
<!--/from cache-->
```

Оригинальный способ предлагается на [12]. Суть его заключается в том, что ссылки от рекламных брокеров устанавливаются для определенных страниц, и чаще всего с помощью одного и того же кода. Соответственно, рекламный брокер узнает о том, какой код разместить на странице, анализируя строку адреса страницы, например, <http://www.site.ru/index.php?cat=1&page=11>. Тогда, передав дополнительный параметр (например, <http://www.site.ru/index.php?cat=1&page=11&aa=bb>) можно ввести в заблуждение рекламного брокера, и он не установит рекламные ссылки на страницу. Сравнив содержание страницы в первом и втором случае, появляется возможность выявить платные ссылки.

Еще один метод заключается в отслеживании динамики изменения содержания страницы. Если в течение времени на странице изменяется только группа ссылок, то эта группа может являться платными ссылками. Аналогичные выводы можно сде-

лать, увидев на странице в какой-то момент времени следующее сообщение:

```
<b>Warning</b>: mysql_connect(): Too many
connections in <b>/home/clx/inc/conf.inc</b> on line
<b>56</b><br />
```

Надо заметить, что не все ссылки, определенные алгоритмом как спам стоит заносить в набор S , а только те, чьи признаки спама носят явно выраженный характер (чтобы исключить случайного попадания ссылок в разряд спама).

На этапе 1 можно использовать различные алгоритмы классификация и машинного обучения.

Этап 2. Выявление спам-ссылок на основе содержания страницы. Основная идея состоит в анализе содержания страницы и выявлении признаков спама. За каждый признак спама на ссылку налагается penalty q_i . Если суммарное penalty превышает определенный порог, ссылка признается спамом.

Шаг 1. Страница сканируется на наличие ссылок S_b , занесенных в список S , сформированный на Этапе 1. При обнаружении таких ссылок сканируется область вокруг них. Если ссылки обнаружены, то им назначается penalty q_1 , величина которого снижается по мере удаления от ссылки S_b .

Шаг 2. Страница сканируется на наличие признаков рекламного блока. Признаком могут служить слова «Реклама», «Спонсоры», «Наши Партнеры» и т.д. При обнаружении признаков рекламного блока, ссылкам в его окрестностях назначается penalty q_2 .

Шаг 3. Страница сканируется на наличие ссылок на рекламного брокера. При обнаружении таких признаков рекламного блока, ссылкам в его окрестностях назначается penalty q_3 .

Шаг 4. Страница сканируется на наличие информации о продаже ссылок (и о том, каких можно купить). При обнаружении таких признаков, ссылкам в их окрестностях назначается penalty q_4 .

Шаг 5. Страница сканируется на наличие большого блока ссылок. Если количество ссылок в блоке больше определенного порога, им назначается penalty q_5 .

Шаг 6. Ссылки сканируются на признаки кода рекламного брокера, в случае обнаружения которого ссылкам назначается penalty q_6 .

Шаг 7. Проверяется соответствие тематики ссылки и общей тематики страницы. В случае несоответствия, ссылке назначается penalty q_7 . Для проверки тематики часто бывает достаточно просто просканировать текст страницы на совпадение слов с текстом ссылки.

Шаг 8. Проверяется соответствие тематики ссылки и тематики ссылок в ее окрестностях. В случае несоответствия, ссылке назначается penalty q_8 .

Шаг 9. Проверяется место размещения ссылки. Если ссылка находится в самом конце страницы, ей назначается penalty q_9 .

Этап 3. Анализ структуры сайта с целью выявления спама. Этот этап является самым сложным. Его

цель – выявить особенности структуры сайта и места на страницах, где встречаются «платные» ссылки.

Для этого из страниц сайта удаляется весь изменяющийся контент (кроме ссылок). Далее производится объединение страниц с одинаковым шаблоном в кластеры. Следующий этап: для каждого кластера удаляются повторяющиеся ссылки и идентифицируются области, где ссылки меняются на каждой странице кластера. Для ссылок, входящих в такие области назначается penalty q_r .

Этап 4. Для каждой ссылки все начисленные penalty суммируются. Если сумма превышает определенный порог, делается вывод, что ссылка – спам. В этом случае ссылка заносится в список S .

5 Практическая реализация

Алгоритм был реализован с использованием CodeGear RAD Studio 2007. Программа состоит из двух независимых модулей:

- модуль формирования набора спам-ссылок S ;
- модуль обнаружения ссылочного спама.

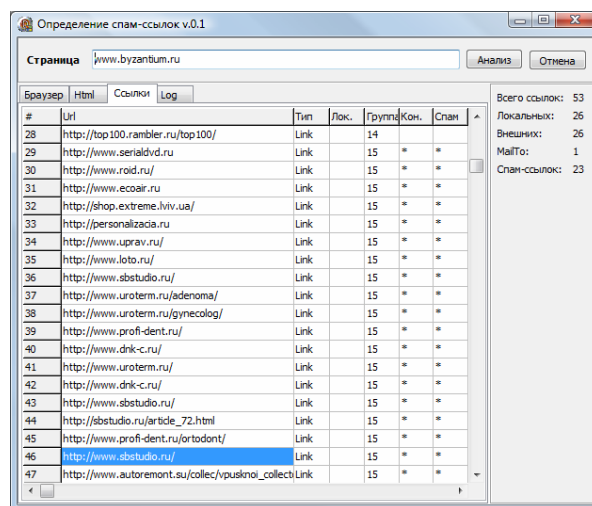


Рис. 6. Окно анализа страницы на спам-ссылки

Для удобства анализа в программе есть возможность просмотра анализируемой страницы как в браузере, так и в виде html-текста.

В программе есть возможность варьировать величину penalty и коэффициенты затухания (снижения величины penalty по мере удаления от признака спама). Коэффициент затухания 0 показывает, что величина penalty не снижается в блоке ссылок. Коэффициент затухания 1 показывает, что величина penalty накладывается только на одну ссылку и не передается на соседние.

Так как формат представления ссылок может отличаться, то из ссылок были удалены все лишние теги и информация, т.е. ссылка вида

```
<a href="http://gmstar.ru/uslugi/dir/mdr/ohrana/org-ohr/" target="_blank">охрана: пост, нулевая, ох-
ранные услуги, собакой</a>
```

преобразуется в вид

<http://gmstar.ru/uslugi/dir/mdr/ohrana/org-ohr/> охр-
на: пост, пультавая, охранные услуги, собакой

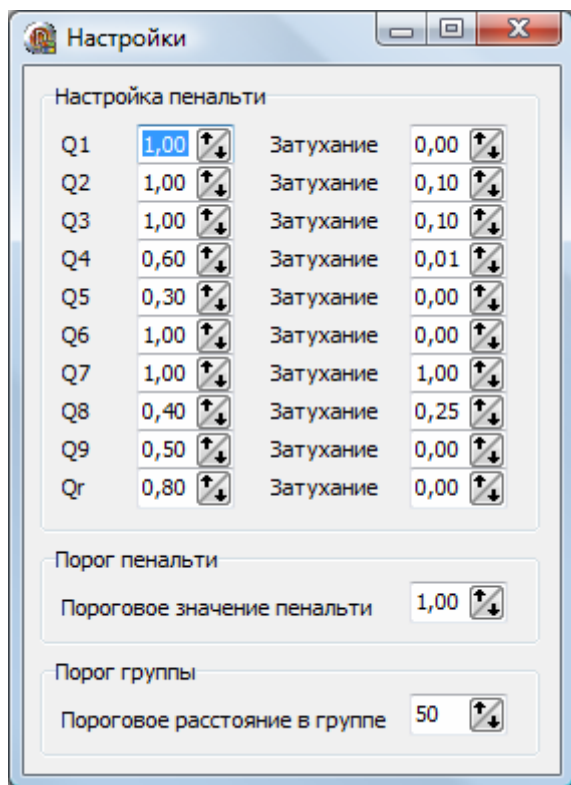


Рис. 7. Окно настройки параметров программы

Это позволяет исключить влияние второстепенных факторов, таких как *target="blank"*, наличие кавычек ' или " и т.д. Указанное преобразование проводится как для набора спам-ссылок *S*, так и для ссылок на странице.

Каждая ссылка, найденная на странице, представляется в программе в виде отдельного объекта с набором характеристик. Все ссылки объединяются в группы, в зависимости от их расположения. Если вблизи ссылки нет других ссылок (одиночная ссылка), то группа создается только из одной ссылки. Далее происходит выявление локальных и внешних ссылок, ссылок *mailto* и т.д. Для внешних ссылок происходит вычисление всех необходимых характеристик (нахождение признаков спама).

Для осуществления возможности анализа времени жизни ссылок и динамики их изменений, все обрабатываемые страницы сохраняются в централизованном архиве. Для сокращения объема хранимой информации страницы при сохранении сжимаются zip-архиватором.

6 Результаты исследований

Для формирования начального набора спам-ссылок *S* были просканированы 20 сайтов, размещающих платные ссылки (информация о местах размещения платных ссылок были предоставлены

нам владельцами сайтов). Число страниц на каждом сайте – от 100 до 5000. После удаления дубликатов был получен набор из 15000 платных ссылок.

В алгоритме были реализованы этапы 1, 2 и 4. Реализация этапа 3 только начата и планируется к завершению в дальнейшем.

При слабо выраженных иных факторах, тематическая близость становится наиболее значимым мериллом спама. Особо это актуально для одиночных ссылок. В наших исследованиях мы использовали «наивный» подход для определения тематической близости, основанный на совпадении набора слов. Конечно, такой подход является достаточно приближенным и не точным. Кроме того, в рамках наших исследований мы остановились лишь на анализе страницы, на которую ведет ссылка, а не всего сайта. Это позволило ускорить задачу анализа страниц. Недостатком такого подхода явилось возникновение неточностей в определении тематической близости (так как часто ссылки ведут на главную страницу, а не на подраздел сайта). В частности такой подход вызвал ошибочное отнесение ряда ссылок в разряд спама.

Для сведения к минимуму подобных ошибок в дальнейшем мы планируем использовать более совершенные методы определения тематического подобия, подобные описанным в работе [14].

Для оценки качества работы алгоритма использовалась методика, описанная в [2].

$$\text{Precision} = \frac{\text{Число спам - ссылок, отмеченных как спам}}{\text{Число ссылок, отмеченных как спам}}$$

$$\text{Recall} = \frac{\text{Число спам - ссылок, отмеченных как спам}}{\text{Общее число спам - ссылок}}$$

$$\text{FalseSpam} = \frac{\text{Число обычных ссылок, отмеченных как спам}}{\text{Общее число обычных ссылок}}$$

$$\text{FalseNotSpam} = \frac{\text{Число спам - ссылок, отмеченных как не спам}}{\text{Общее число спам - ссылок}}$$

Для тестирования были вручную отобраны 100 страниц с числом внешних ссылок от 1 до 30 на каждой. Общее количество ссылок составило 783. Для каждой страницы были вручную отмечены спам-ссылки, которых оказалось 519. В результате работы алгоритма 490 ссылок были отмечены как спам, их которых 449 действительно были спам-ссылками (совпали с отобранными вручную). Результаты оценки качества работы алгоритма приведены в табл. 1.

Таблица 1. Результаты тестирования алгоритма

Precision	0.92
Recall	0.86
FalseSpam	0.155
FalseNotSpam	0.135

Большинство ошибок в выявлении спам-ссылок возникло из-за неглубокого анализа тематической близости. Ряд ссылок были отнесены к разряду спама (групповые ссылки) из-за того, что часть ссылок в группе были признаны как не соответствующие тематике страницы.

Одиночные рекламные ссылки (в основном, размещенные вручную) близкой к странице тематики не были выявлены как спам. Это связано с тем, что у таких ссылок признаки спама часто отсутствуют. Решением может служить только анализ структуры страниц сайта и выявление мест размещения рекламы (планируется осуществить в будущем), а также анализ времени жизни ссылок, для чего необходим длительный мониторинг страниц.

Таким образом, предложенный алгоритм демонстрирует достаточно неплохие результаты в определении спам-ссылок. Ошибки в определении ссылок можно сократить путем совершенствования методов анализа структуры страниц и оценки тематического подобию.

Литература

- [1] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The connectivity sonar: Detecting site functionally by structural patterns. In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, pages 38–47, Aug. 2003.
- [2] L. Becchetti, C. Castilio, D. Donato, S. Leonardi, R. Baeza-Yates. Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection. Technical report, DELIS – Dynamically Evolving, Large-Scale Information Systems, 2006.
- [3] B. D. Davison. Recognizing nepotistic links on the web. In *AAAI-2000 Workshop on Artificial Intelligence for Web Search*, Austin, TX, pages 23–28, July 30 2000.
- [4] N. Eiron, K.S. McCurley, and J.A. Tomlin. Ranking the web frontier. In *Proceedings of the 13th International World Wide Web Conference (WWW)*, pages 309–318, New York, NY, USA, 2004. ACM Press.
- [5] Eric Enge. 15 Methods for Paid Link Detection <http://www.stonetemple.com/blog/?p=167>
- [6] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics – Using statistical analysis to locate spam web pages. In *Proceedings of the 7th International Workshop on the Web and Databases (WebDB)*, Paris, France, 2004.
- [7] Z. Gyungyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, Toronto, Canada, 2004.
- [8] Z. Gyungyi and H. Garcia-Molina. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*, 2005.
- [9] Tim Nash. How to find a paid link? <http://paymentblogger.com/2007/10/07/how-to-find-a-paid-link/>
- [10] B. Wu and B. D. Davison. Identifying link farm pages. In *Proceedings of the 14th International World Wide Web Conference (WWW)*, 2005.
- [11] Алексей Кравцов. Ссылочный спам: найти и обезвредить <http://www.kravcov.ru/2007/03/11/nnueiue-niai-e-eae-n-ie-i-aidhiouny/>
- [12] Детектор продажных ссылок, 2008. <http://venality.name/>
- [13] Компания Яндекс <http://company.yandex.ru/>
- [14] Некрестьянов И.С. Тематико-ориентированные методы информационного поиска: Диссертационная работа к.т.н.: 05.13.11 / Санкт-Петербургский государственный университет – СПб., 2000. – 80 с.

Link spam Detection

R.V. Sharapov, E.V. Sharapova

In article approaches to detecting of a reference spam on the basis of the analysis of the maintenance of page are considered. The basic place in work is devoted detecting of advertising (paid) links. Signs of paid links are analyzed. The algorithm of detecting a spam-links is given and results of its work are resulted.

Resumagic: система автоматической обработки резюме

© А.В. Сафронов

ООО «Хэдхантер»
safronov@hh.ru

Аннотация

В статье описывается система Resumagic, предназначенная для автоматической обработки резюме. Рассмотрены задачи фильтрации и классификации резюме. Приведены результаты применяемых методов.

1 Введение

В своей повседневной работе специалисты по подбору персонала часто сталкиваются с задачей обработки потока входящих резюме. Информация о соискателях попадает к рекрутеру из нескольких разных источников, в том числе и по электронной почте. В больших кадровых агентствах количество резюме, присылаемых соискателями в течение дня, может достигать нескольких сотен; к этому числу на практике следует также прибавить некоторое количество незапрашиваемой корреспонденции (спама). Каждое входящее письмо нуждается в трудоемкой обработке.

Для автоматизации процесса обработки резюме была разработана система Resumagic. Ее основные функции:

- Импорт резюме из внешних источников;
- Отделение резюме от прочих документов;
- Извлечение фактов из текста резюме;
- Классификация резюме;
- Нормализация основных полей;
- Выявление резюме, принадлежащих одному человеку.

Далее кратко рассматриваются наиболее интересные элементы системы.

2 Извлечение фактов

Главным элементом рассматриваемой системы является модуль извлечения фактов. Система обеспечивает выделение следующих полей:

- Личная информация (ФИО, пол, дата рождения/возраст, семейное положение, гражданство);
- Контактная информация (телефоны, электронная почта, ICQ);

- Желаемая должность и пожелания по уровню дохода;
- Опыт работы (период работы, название компании, сфера деятельности компании, должность, отдел, обязанности, стаж);
- Образование (год выпуска, название учебного учреждения, факультет, кафедра, специальность, специализация, квалификация);
- Владение иностранными языками;
- Знание программного обеспечения и технологий;
- Личные качества, хобби;
- Рекомендации (ФИО, должность, название компании, контакты).

Процесс извлечения фактов в описываемой системе состоит из следующих этапов:

1. Графематический анализ. Выполняется разбиение текста на слова и предложения. При этом учитываются характерные для процесса пересылки текста по электронной почте случаи потери форматирования и как следствие – склеивание слов и строк.
2. Морфологический анализ. Морфологический анализ выполняется с помощью набора правил словообразования и словаря лемм-исключений, которые не могут быть описаны набором правил. Важной особенностью данного этапа состоит в том, что язык резюме в значительной мере отличается от языка классических корпусов русского языка (имеет место широкое использование профессиональной терминологии).
3. Выделение лингвистических конструкций вокруг ключевых слов. Иначе говоря, выделяются цепочки слов, обозначающие некоторые объекты предметной области. Например, на основе ключевого слова «университет» будет, в частности, выделено словосочетание «московский государственный университет».
4. В зависимости от контекста, выделенная цепочка может быть интерпретирована как факт. Например, упоминание московского государственного университета в контексте подраздела «образование» означает, что человек учился в данном образовательном учреждении. В контексте подраздела «опыт работы» данное словосочетание интерпретируется как название места работы.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

3 Классификация резюме

Под классификацией (рубрикацией) резюме понимается его отнесение к одной или нескольким категориям, описывающим профессиональную область или специализацию соискателя. Другими словами, для резюме определяется некая общая характеристика, описывающая соискателя в целом (например, «продажи», «IT-специалисты», «маркетинг»).

Единого стандарта классификации резюме не существует. На практике разные кадровые агентства и job-сайты используют свои собственные схемы классификации. Встречаются как простые одноуровневые классификаторы, так и разветвленные таксономии. Как правило, количество классов верхнего уровня находится в пределах от 10 до 30. Иногда могут применяться схемы, основанные на тегах.

В настоящее время классификатор сайта hh.ru имеет 2 уровня. Каждое резюме может относиться только к одной профессиональной области (верхний уровень) и к нескольким специализациям.

В системе Resumagis была реализована функция автоматической классификации резюме в соответствии со стандартом HeadHunter. Для обучения и настройки автоматического классификатора использовалась выборка из 5661 резюме. Для оценки качества классификации использовалась тестовая выборка из 1887 резюме. Нами были произведены эксперименты классификацией резюме с помощью метода k ближайших соседей и метода PrTFIDF.

3.1 Классификация резюме с помощью метода k ближайших соседей

Метод k ближайших соседей (kNN, k nearest neighbors) заключается в поиске в обучающей выборке документов, наиболее похожих на анализируемый документ, после чего этому документу присваивается самая распространенная категория среди найденных.

Для поиска ближайших соседей вводится метрика близости документов. Каждый документ рассматривается как вектор в n-мерном пространстве, где n – общее количество признаков (термов). В качестве элементов вектора используется мера TF-IDF. В таком представлении метрика схожести 2-х документов определяется как значение косинуса угла между их векторами:

$$\text{Similarity} = \frac{\sum_{i=1}^n a_i \cdot b_i}{\sqrt{\sum_{i=1}^n a_i^2} \cdot \sqrt{\sum_{i=1}^n b_i^2}}$$

где a_i и b_i – элементы векторов сравниваемых документов, т.е. TF-IDF.

$$\text{Tfidf} = \text{TermFreq} \cdot \log\left(\frac{N_{doc}}{DocFreq}\right)$$

где

TermFreq – частота встречаемости термина в документе

Ndoc – количество документов в коллекции

DocFreq – количество документов, в которых встречается терм.

Для оценки эффективности нашей реализации метода kNN был произведен ряд экспериментов на тестовом наборе данных. В качестве основного критерия оценки использовалась метрика F1.

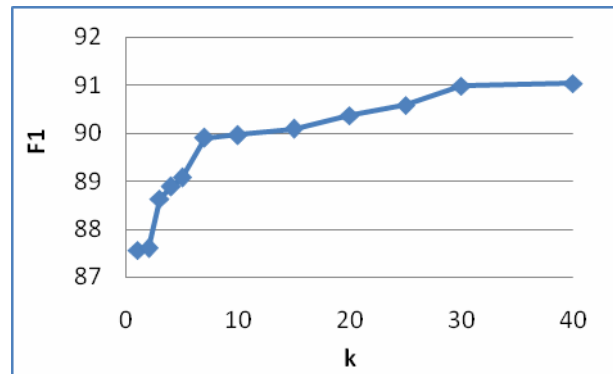


График 1. Зависимость качества классификации резюме (F1) от количества ближайших соседей, принимающих участие в классификации (k).

Таблица 1. Точность, полнота и F1 в зависимости от k

k	Precision	Recall	F1
1	0,778	1	0,876
2	0,778	1	0,876
3	0,796	1	0,886
4	0,800	1	0,889
5	0,803	1	0,890
7	0,816	1	0,899
10	0,817	1	0,899
15	0,819	1	0,901
20	0,824	1	0,904
25	0,828	1	0,906
30	0,834	1	0,909
40	0,835	1	0,910

Как видим из таблицы 1, с помощью нашей реализации метода kNN достигается точность классификации резюме 0,835 при полноте 1, что можно считать удовлетворительным уровнем при автоматической обработке резюме.

3.2 Классификация резюме с помощью метода PrTFIDF

Метод PrTFIDF описан в работе [2]. Формула вычисления наиболее вероятной рубрики выглядит следующим образом:

$$H(d') = \arg\max_{C_j \in C} \sum_{\omega \in \Omega} \frac{\Pr(\omega|C_j) \cdot \Pr(C_j)}{\sum_{C' \in C} \Pr(\omega|C') \cdot \Pr(C')} \cdot \Pr(\omega|d')$$

где

$\Pr(\omega|C_j)$ – средняя частота термина ω в документах категории C_j ;

$\Pr(C_j)$ – соотношение количества документов в категории C_j к количеству документов в обучающей выборке;

$\Pr(\omega|d')$ – частота термина ω в классифицируемом документе.

Метод PrTFIDF также был реализован в рамках платформы Resumagic. Было произведено сравнение результатов работы метода PrTFIDF с результатами kNN. Анализ прогонов на тестовой выборке резюме показал, что при оптимальном количестве используемых для классификации термов (около 1000) методы показывают близкие значения полноты и точности, при небольшом преимуществе метода PrTFIDF.

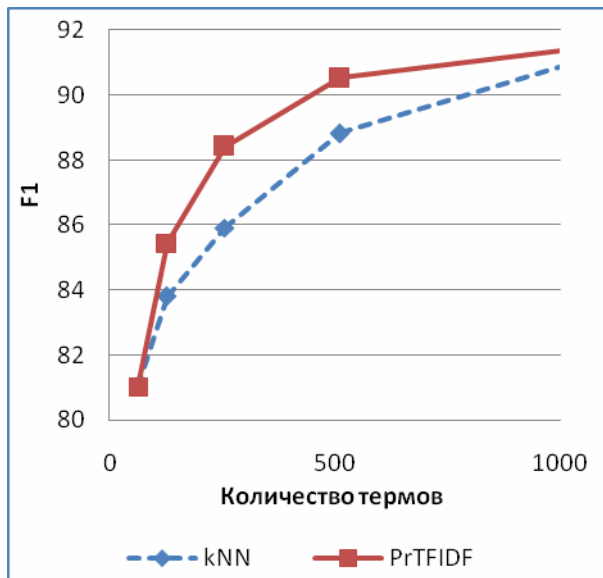


График 2. Сравнение kNN и PrTFIDF

В связи с этим в качестве основного метода автоматического определения профессиональной области в платформе Resumagic был выбран PrTFIDF.

3.3 Выбор термов

Важным фактором, влияющим на качество классификации резюме, является алгоритм выбора термов. Для уменьшения размерности пространства признаков обычно используются два подхода:

- отбрасывание неважных термов;
- кластеризация термов.

При отбрасывании термов для каждого из них рассчитывается некий вес, соответствующий их «важности». Отбрасываются все термы, вес которых ниже определенного уровня. Веса термов могут рассчитываться по-разному. Например, в качестве веса можно брать частоту терма. В Resumagic для вычисления веса используется следующая эвристическая формула:

$$W(\omega) = \text{Freq}(\omega|D) * \sum_{C_j \in C} \left(\frac{\text{Freq}(\omega|C_j) * N}{\text{Freq}(\omega|D)} - 1 \right)^2$$

где

$\text{Freq}(\omega|D)$ – частота терма ω в обучающей выборке;

$\text{Freq}(\omega|C_j)$ – частота терма ω в категории C_j ;

N – количество категорий.

Эксперименты на тестовой выборке резюме показали, что предложенная формула позволяет более

точно оценивать важность терма, чем простая оценка частоты. Особенно это заметно при небольшом количестве термов.

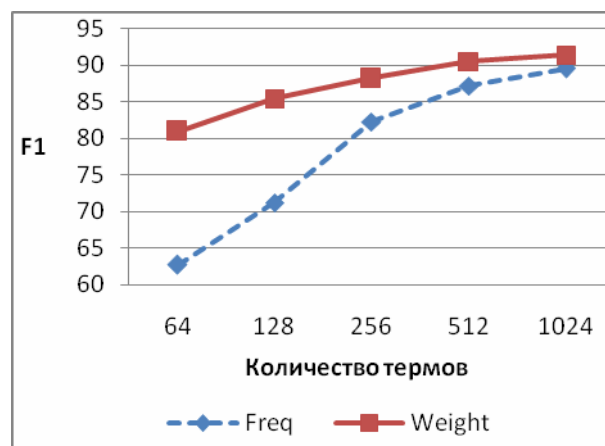


График 3. Сравнение методов выбора термов

Также для уменьшения размерности пространства признаков в Resumagic используется лемматизация. Как показали наши эксперименты, использование морфологии позволяет несколько увеличить качество классификации резюме.

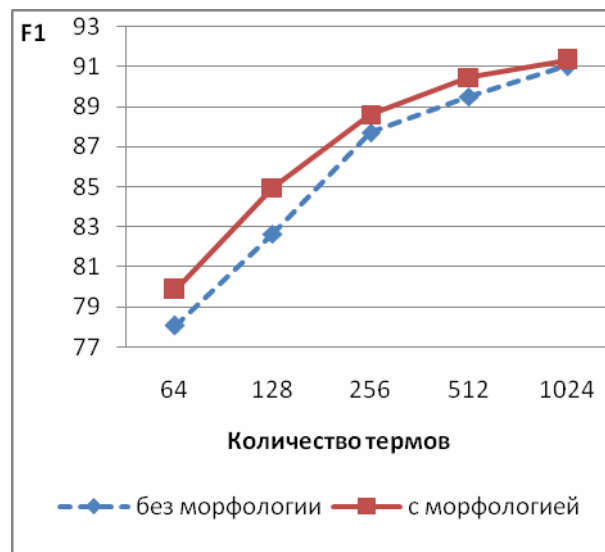


График 4. Сравнение качества классификации с учетом морфологии и без нее.

Использование в качестве термов не только отдельных слов, но и словосочетаний приводит к дальнейшему улучшению качества классификации резюме. Очевидно, что иногда роль слова с точки зрения классификации может сильно изменяться в зависимости от того, в каком контексте это слово употребляется. Например, словосочетания «продвинутый пользователь» и «поддержка пользователей» характерны для совсем разных категорий резюме. Поэтому в платформе Resumagic в качестве термов используются не только отдельные слова, но и словосочетания, что позволяет поднять качество классификации.

3.4 Классификация на основе составленных вручную правил

Помимо метода PrTFIDF для классификации резюме в Resumagic применяется классификатор на основе инженерного подхода. Его работа основана на множестве составленных вручную правил. Инженерный классификатор работает уже не с исходным текстом резюме, а с его структурированным представлением. Можно сказать, что этот классификатор в известном смысле воспроизводит логику рекрутера, занимающегося ручной классификацией резюме. Инженерный классификатор обладает меньшей полнотой (recall), чем PrTFIDF, но при этом обеспечивает почти 100% точность (precision).

Сочетание классификатора на основе правил и метода PrTFIDF обеспечивает распознавания профессиональной области с точностью порядка 0,92 при полноте 0,99.

4 Отделение резюме от прочих документов

Чтобы отличать резюме от других документов, Resumagic рассчитывает специальный числовой индекс под названием ResumagicRank. Этот индекс представляет собой число в диапазоне от 0 до 100. Чем выше ResumagicRank, тем больше текст похож на резюме.

ResumagicRank представляет собой интегральную эвристическую характеристику, которая формируется на основе множества различных факторов. Логика ее расчета можно схематически описать следующей формулой:

$$\text{ResumagicRank} = \log \left(\sum_{i=1}^{NP} WP_i + \sum_{i=1}^{NB} WB_i \right)$$

где

NP — количество фактов личного характера, извлеченных из документа. Под «личными фактами» подразумевается непосредственная информация о человеке (ФИО, возраст, семейное положение и т.д.).

WP_i — вес (значимость) i -того личного факта.

NB — количество фактов делового характера, извлеченных из документа. Под деловыми фактами подразумевается информация, связанная с профессиональной деятельностью человека, его умениями и знаниями (опыт работы, образование, желаемая должность, желаемая зарплата и т.д.).

WB_i — вес (значимость) i -того «делового» факта.

Как видно из формулы, важным требованием к документу является одновременное присутствие как деловой, так и личной информации.

Наши эксперименты показали, что ResumagicRank позволяет с высокой точностью отличать резюме не только от обычного спама, но и от «резюмеподобных» писем. Например, в обычной рабочей переписке может содержаться достаточно много информации, характерной для резюме (хотя бы подписи к письму). Также некоторые информационные рассылки могут включать контактную информацию, сведения о людях, компаниях и т.п.

ResumagicRank используется для сортировки входящих документов по 3 разделам: «резюме», «не резюме» и «нераспознанные». Документы, которым присвоен ResumagicRank > 66, относятся к категории «резюме». Если ResumagicRank < 33, то документ помечается как «не резюме». Остальные документы относятся к классу нераспознанных и нуждаются в дальнейшей ручной проверке. Как правило, в «нераспознанные» попадают резюме, составленные в нетрадиционной манере или содержащие слишком мало сведений. Также к категории «нераспознанных» могут быть отнесены некоторые письма, содержащие большое количество фактов о людях (например, новостные рассылки с информацией о назначениях и отставках в правительстве).

5 Определение региона

Город проживания не всегда указывается в резюме в явном виде. Кроме того, иногда рекрутеры могут интересоваться кандидатами не из конкретного города, а из достаточно большого списка разных городов (например, ближнего Подмосковья).

Для распознавания региона в Resumagic имеется специальный механизм, работающий по следующему алгоритму:

1. Поиск в тексте резюме непосредственного упоминания региона. Иногда область, в которой проживает соискатель, указывается явным образом. Например, в некоторых резюме соискатели пишут что-то вроде этого: «регион проживания — Краснодарский край».
2. Если регион не указан явно, Resumagic пытается определить его на основании города, в котором проживает соискатель. Для этого имеется внутренний справочник соответствий между городами и регионами. С помощью этого справочника, например, можно определить, что город Борисоглебск относится к Воронежской области. На этом этапе проблема может состоять в том, что и город не всегда указывается кандидатом в своем резюме. Если город не указывается в явном виде, то для его определения используются дополнительные проверки:
 - a. Код города в телефоне. В Resumagic имеется справочник телефонных кодов для основных городов России.
 - b. Станция метро. Иногда в резюме отсутствует прямая информация о городе, но упоминается ближайшая станция метро. Особенно это характерно для тех резюме, которые соискатели отправляют конкретному работодателю или в конкретное кадровое агентство, предполагая, что город и так очевиден. Многие названия станций метро являются «эндемичными», т.е. встречаются только в одном городе.

3. В разных регионах встречаются города с одинаковыми названиями. Однако при указании почтового адреса в резюме иногда указывается индекс, что позволяет надежно определить регион.

6 Защита от дублирования

При поступлении новых резюме необходимо следить за тем, чтобы в базе данных не возникало дублирования соискателей. Ручная проверка на дублирование отнимает некоторое время. При этом, как правило, в базе все равно постепенно появляются «двойники». Если одному соискателю соответствует несколько разных записей, то это может привести к путанице и снижению эффективности работы рекрутера.

Для решения этой проблемы в описываемой системе реализована автоматическая защита от дублирования.

Защита от дублирования включает в себя следующие уровни:

1. Поиск по полному совпадению текстов резюме на основе хеш сумм. После очистки от элементов форматирования для входящего резюме вычисляется CRC32, а затем в базе находятся все резюме с совпадающим значением CRC32. Найденные резюме проходят дополнительное проверочное сравнение с входящим документом, в результате чего происходит обнаружение полных дубликатов. Эта проверка позволяет защититься в случаях, когда соискатель присылает одно и то же резюме по нескольку раз.
2. Поиск по совпадению ФИО, даты рождения или контактной информации. Если из текста входящего резюме были успешно выделены фамилия, имя, отчество, дата рождения или контактная информация, то Resumagic производит поиск соискателя по этим полям. При этом принимаются во внимание также распространенность фамилии и имени соискателя.
3. Поиск частичных совпадений резюме. Если предыдущие шаги не дали результатов, то система пытается принять решение с помощью «индекса сходства резюме». Индекс сходства – это число от 0 до 100, которое показывает, насколько два резюме похожи между собой. Эта методика позволяет найти дублирующихся соискателей, даже если они внесли изменения в текст своего резюме. Для этого сначала в базе данных производится поиск соискателей только по имени и фамилии или по контактной информации. Потом для всех найденных резюме определяется степень их схожести на входящее резюме. Если система находит в базе текст с высоким индексом сходства, то это означает, что кандидат прислал обновленное резюме с некоторыми изменениями. В этом

случае можно уверенно обновлять соискателя с максимальным индексом схожести. Если же индекс схожести низок, то это говорит о том, что резюме принадлежат разным лицам. Тогда можно спокойно создавать нового соискателя, не опасаясь дублирования.

Литература

- [1] Resumagic. <http://resumagic.ru>
- [2] Joahims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF. Proceedings of ICML-97, 14th International Conference on Machine Learning. http://www.cs.cornell.edu/people/tj/publications/joachims_97a.pdf
- [3] Гершензон. Технология извлечения структурированной информации из неструктурированного текстового массива. 2006. Jug.ru. http://www.jug.ru/servlets/images/meeting_2006_12_23/FactExTech.ppt
- [4] Качаева, Южиков. Автоматизированная система распознавания и классификации резюме.RCDL-2007. http://rcdl2007.pereslavl.ru/papers/paper_45_v1.pdf
- [5] Петров, Кузнецов. Особенности сетевого англоязычного лингвистического процессора для формализации текстовой информации на естественном языке. Диалог-2006. <http://www.dialog-21.ru/dialog2006/materials/html/Petrov.htm>
- [6] Resume Mirror. <http://www.resumemirror.com/products/resume-processing.html>

Resumagic: automatic CV processing

A.V. Safronov

This paper describes Resumagic, a system intended for automated parsing of curriculum vitae texts [CVs]. Problems of filtering and classification of resumes were reviewed. Also results of applied methods are given.

О методе автоматического реферирования, основанном на результатах рубрицирования документов

© В.Е. Абрамов

ЗАО СКБ «ТЭЛКА»
abramval@yandex.ru

Н.Н. Абрамова

НИЦИ при МИД России
nabramova@mid.ru

Аннотация

Работа посвящена методу автоматического реферирования, в котором используются результаты предварительно проведенного рубрицирования текстовой информации.

Обычно реферирование и рубрицирование рассматривают как самостоятельные задачи, хотя они взаимосвязаны: опираются на одни и те же процедуры автоматической обработки текстов (морфологический и синтаксический анализ). Идея метода заключается в том, чтобы максимально использовать результаты предшествующих этапов обработки и составлять реферат после определения основных тем документа. Полученные результаты показывают, что предложенный метод дает приемлемое качество реферата.

1 Введение

В настоящее время рубрицирование и реферирование информации широко используется в автоматизированных системах при формировании баз данных. Обычно автоматическое рубрицирование и реферирование рассматривают как самостоятельные задачи, хотя некоторые исследователи [10] указывают на взаимозависимость этих задач.

Методы автоматического рубрицирования имеют много общего с методами реферирования, хотя присутствует специфика, характерная для каждого из этих двух классов задач.

Основной целью рубрицирования является определение тем документа. Предложения, в которых выражены темы документа, можно рассматривать как краткое изложение смысла документа, то есть его реферат. Составление реферата можно начинать, используя темы, определенные при рубрицировании.

Одна из целей исследования – показать, что использование результатов автоматического рубрици-

рования дает приемлемое качество реферата независимо от языка реферируемого текста. Преимуществом метода реферирования, использующего результаты автоматического рубрицирования, является сокращение времени обработки документов.

2 Существующие подходы

За прошедшие годы появилось много публикаций в области автоматического реферирования. Расширилось даже само понятие «автоматическое реферирование», которое относится не только к текстовой информации, но и к мультимедийной. Наряду с развивающейся традиционной задачей составления реферата отдельного документа появились новые сферы применения, к которым можно отнести мультязычное реферирование, составление обзорных рефератов по набору документов, реферирование гибридных источников, содержащих текстовую и фактографическую информацию, создание видеорефератов на основе анализа видеообъектов, относящихся к какому-либо видеофайлу, а также реферирование интернет-сайтов. В обзорной статье [14] дан подробный анализ положения дел в области автоматического реферирования. На сайте постоянно действующей конференции DUC можно найти последние публикации в области автоматического реферирования, сравнительные оценки методов реферирования и результаты тестирования различных систем [13].

Так как целью нашей работы является решение традиционной задачи реферирования, то остановимся на подходах, которые в настоящее время наиболее распространены в этой сфере. Все методы, применяющиеся в реферировании текстовой информации, можно разделить на два направления – квазиреферирование и генерирование рефератов. Методы, относящиеся к направлению квазиреферирования, основаны на выделении из текстов наиболее информативных фрагментов (предложений), передающих основной смысл текста документа. Методы второго направления основаны также на выделении из текстов документов наиболее информативной информации и генерирования с помощью ее новых текстов. Практически все современные системы реферирования относятся к направлению квазиреферирования, хотя в последние годы появи-

лись интересные работы, развивающие второе направление [2, 6, 11, 17].

Многие известные методы реферирования базируются на подходе, предложенном Г. Луном [18] в 50-х годах прошлого века, который заключается в выделении в тексте частотных слов, вычислении весов предложений с помощью суммирования частот (весов) входящих в их состав слов и включения в реферат предложений с наибольшими весами.

Для современных методов характерно сочетание традиционного подхода с некоторыми модификациями. Например, в качестве значимых элементов выбираются не слова, а словосочетания [3], вводятся дополнительные критерии выбора значимых слов: вес слова увеличивается в зависимости от его нахождения в заголовке, в первом и последнем предложениях или выделения шрифтами в тексте или в запросе пользователя [5].

В работе [4] предлагается эффективный метод реферирования на основе машинного обучения.

Яцко В.А. предложил метод симметричного реферирования [12], в котором вес предложения определяется количеством связей между данным предложением и предложениями, находящимися слева и справа от него. Для этого в каждом предложении определяется список ключевых слов, входящих в предварительно составленный тематический словарь, а затем в предложениях, расположенных слева и справа, подсчитывается количество найденных в них ключевых слов (связей) из определенного ранее списка. Сумма лево- и правосторонних связей определяет вес предложения.

Известны подходы к реферированию на основе предварительно проведенной тематической кластеризации документа с дальнейшим выделением ключевых предложений из каждого кластера [21] и с предварительной разбивкой документа на части (с учетом его структуры), построением реферата для каждой части и отбором наиболее важных фрагментов [8].

Важнейшей частью задачи реферирования является оценка качества полученного реферата. Этой проблеме посвящено много работ [15, 19, 20, 22]. Помимо традиционных методик, связанных с экспертными оценками качества рефератов, в последние годы развиваются автоматические методы оценки. Например, в работе [22] предлагается оценивать реферат по наличию в нем частотных словосочетаний из оригинала и близости распределения относительных частот появления этих словосочетаний в документе и в реферате.

3 Идея исследования

Большинство исследователей рассматривают автоматическое рубрицирование и реферирование как самостоятельные задачи без связи между ними. Однако во многих информационных системах при формировании баз данных проводится не только рубрицирование информации, но и ее реферирование, поэтому наряду с полным текстом документов

дается их сжатое содержание в виде реферата или аннотации. В разработанной авторами системе автоматического рубрицирования текстов (САРТ), которая работает в промышленном режиме полтора года, имеется функция реферирования [7].

Как известно, методы рубрицирования и реферирования опираются на одни и те же процедуры автоматической обработки текстов: морфологический и концептуальный анализ, т.е. обработку документа можно проводить однократно, увеличивая производительность системы.

Преимущества метода заметны при обработке больших потоков информации в системах, работающих в режиме мониторинга.

Для оценки качества реферирования необходимо разработать методику экспертной оценки и провести анализ ее результатов, что позволит найти пути улучшения работы алгоритма реферирования.

4 Составление реферата отдельного текста документа на основе результатов автоматического рубрицирования

4.1 Описание метода реферирования

После рубрицирования массива из n документов $D = \{d_i, i = 1, \dots, n\}$ каждому документу d_i приписывается кортеж

$$(T, P) = \{t_1, p_1, t_2, p_2, \dots, t_j, p_j, \dots, t_m, p_m\},$$

где $j = 1, \dots, m$ – номер темы,

T – набор тем в документе d_i ,

P – набор весов тем в документе d_i (суммарных частот появления в тексте словосочетаний, имеющих в описании тем из набора T).

В свою очередь каждая из тем $t_j \in T$ определяет множество ключевых слов и словосочетаний, имеющихся в тексте документа d_i и в описании темы t_j , и частотами появления этих словосочетаний. Каждая тема t_j описывается с помощью кортежа

$$(W, F) = \{w_{j_1}^{(1)}, f_{j_1}^{(1)}, w_{j_2}^{(2)}, f_{j_2}^{(2)}, \dots, w_{j_j}^{(j)}, f_{j_j}^{(j)}\}, \text{ где}$$

$w_{j_j}^{(j)}$ – слово или словосочетание из документа

d_i , определяющее тему t_j ;

$f_{j_j}^{(j)}$ – частота появления в документе d_i слова

или словосочетания $w_{j_j}^{(j)}$;

l_j – количество слов или словосочетаний, описывающих тему t_j .

Выбор наиболее информативных предложений из текста документа d_i проводился по следующему алгоритму:

1. Для каждой темы формировался список предложений, которые ее характеризуют.
2. Далее все предложения подвергались обработке с помощью словаря стоп-слов, включающего

служебные части речи, а также неинформативные слова и словосочетания.

3. Определялся вес каждого предложения путем суммирования частот появления в нем слов и словосочетаний, определяющих темы.
4. В каждом из отобранных предложений удалялись примечания (слова в круглых, косых или угловых скобках) и некоторые обороты с помощью специального словаря. Отличие между словарем оборотов и словарем стоп-слов в том, что неинформативные слова целиком задаются списком, а в словаре оборотов имеется только начальная часть фразы, например, «Как сообщается», «Как стало известно». После распознавания в тексте удаляется из предложения не только эта часть, но и весь оборот до знака препинания вместе с ним.
5. Все предложения с удаленными стоп-словами, примечаниями и оборотами, являвшиеся кандидатами для включения в реферат, последовательно проверялись на тождественность. Предложения считались тождественными при совпадении 80 % слов. Из всех близких предложений в реферат включалось предложение с максимальным весом.
6. Вычислялся коэффициент сжатия реферата как отношение объема реферата к объему документа (в байтах). Коэффициент сжатия обычно задается параметрически. Кроме этого определялось количество предложений в реферате в начале работы и после каждого удаления предложений.
7. Если был получен реферат с коэффициентом сжатия более заданной величины, то удалялось предложение с самым маленьким весом. Затем опять оценивалась длина реферата (п. 6) и в случае неуспеха выбиралось следующее предложение с минимальным весом и т.д. В том случае, если имелось несколько предложений с одинаковым минимальным весом, оставлялось предложение с наименьшей длиной (длина равнялась количеству информативных слов в предложении).
8. Процесс останавливался, если был получен реферат, удовлетворяющий критерию сжатия, или осталось только одно предложение (такая ситуация не встречалась).

4.2 Примеры работы системы

Рассмотрим работу алгоритма на примере русского текста. Исходный текст приводится на рис. 1. Он сопровождается результатами автоматического рубрицирования: дается перечень тем и объектов.

После этапа рубрицирования были переданы следующие данные: темы, их веса и ключевые слова и словосочетания, определяющие темы, с частотой их встречаемости в тексте. Эти данные представлены в табл. 1.

1. В Южной Осетии может пролиться кровь из-за спорных фруктовых садов.
2. Командование Смешанных сил по поддержанию мира (ССПМ) в зоне грузино-осетинского конфликта выступает инициатором встречи представителей администрации Знаурского района Южной Осетии с населением приграничных сел Грузии, в связи с нерешенностью территориальных споров вокруг фруктовых садов, сообщил корреспонденту ИА REGNUM помощник командующего ССПМ по работе со СМИ подполковник Юрий Верещак.
3. 14 октября группой военных наблюдателей от трех сторон совместно с представителем Миссии ОБСЕ был проведен мониторинг в районе населенного пункта Нули (территория Грузии) и населенного пункта Гвертев (Южная Осетия) по факту обострения ситуации в данном районе.
4. Для предотвращения возможных инцидентов в районе садов выставлен временный наблюдательный пост миротворческих сил от России с наблюдателями от трех сторон.
5. 15 октября для разрешения проблемы была проведена встреча представителей сторон, однако они к взаимоприемлемому решению не пришли.
6. До настоящего времени вопрос остается открытым" - сказал Верещак.
7. Как сообщил глава администрации Знаурского района Южной Осетии Заур Цховребов, суть конфликтной ситуации заключается в необоснованных претензиях жителей Нули на яблоневые сады обрабатываемых осетинским населением Гвертев.
8. "Сады находятся на территории Южной Осетии и эти претензии мы не понимаем", – сказал Цховребов,
9. добавив, что "на предложенную 17 октября командованием ССПМ и Миссией ОБСЕ повторную встречу, грузинская сторона не явилась".
10. "Чтобы ситуация окончательно не вышла из под контроля, мы предложили провести встречу завтра.
11. Надеемся, что представители от грузинского села все-таки на нее явятся", – сказал глава администрации.

Темы:

Политические представители сторон и посредников урегулирования конфликта

Миротворцы России в зоне конфликтов на территории стран СНГ

Угроза применения силы, санкций и блокады, выдвижение ультиматумов

Территориальные споры в населенных пунктах зоны конфликта

Объекты:

Грузия

Южная Осетия

Организация по безопасности и сотрудничеству в Европе

Рис. 1. Исходный текст документа (рус. яз.)

Таблица 1
Исходные данные для реферирования

Тема	Вес темы	Определяющие ключевые слова с частотой встречаемости
Политические представители сторон.....	15	наблюдатели 2, наблюдательный пост 1, миссия ОБСЕ 2, представители сторон 1, грузинская сторона 1, Россия 1, Южная Осетия 5, Грузия 2
Угроза применения силы.....	12	инцидент 1, конфликт 1, конфликтная ситуация 1, кровь 1, обострение ситуации 1, Южная Осетия 5, Грузия 2
Территориальные споры.....	11	спорные фруктовые сады 1, спор вокруг фруктовых садов 1, претензии 1, территориальный спор 1, Грузия 2, Южная Осетия 5
Миротворцы России.....	5	Командование Смешанных сил по поддержанию мира 1, командование ССПМ 1, конфликт 1, миротворческие силы 1, Россия 1

Таблица 2
Веса предложений

Номер пред-я	Вес пред-я	Определяющие ключевые слова с частотой встречаемости
1	7	Южная Осетия 5, кровь 1, спорные фруктовые сады 1
2	11	Командование Смешанных сил по поддержанию мира 1, конфликт 1, Южная Осетия 5, Грузия 2, территориальный спор 1, спор вокруг фруктовых садов 1
3	12	наблюдатели 2, миссия ОБСЕ 2, Южная Осетия 5, Грузия 2, обострение ситуации 1
4	6	инцидент 1, наблюдатели 2, наблюдательный пост 1, миротворческие силы 1, Россия 1
5	1	представители сторон 1
6	0	
7	7	Южная Осетия 5, конфликтная ситуация 1, претензии 1
8	5	Южная Осетия 5
9	4	командование ССПМ 1, миссия ОБСЕ 2, грузинская сторона 1
10	0	
11	0	

Далее подсчитывались веса предложений, равные суммарной частоте встречаемости в них определяющих темы ключевых слов и словосочетаний из табл. 1.

Веса предложений вместе с найденными в этих предложениях ключевыми словами и словосочетаниями, определяющими темы, приведены в табл. 2.

Из всех предложений удалялись стоп-слова, примечания и обороты, которые определялись с помощью соответствующих словарей.

В соответствии с выбранным коэффициентом сжатия реферата 0,3 было отобрано три предложения со следующими номерами: 1, 2, 3.

Окончательный текст реферата приведен на рис.2.

Приведем также пример работы алгоритма для английского языка. На рис. 3 дается исходный текст, а составленный по нему реферат приводится на рис. 4.

В Южной Осетии может пролиться кровь из-за спорных фруктовых садов.
Командование Смешанных сил по поддержанию мира в зоне грузино-осетинского конфликта выступает инициатором встречи представителей администрации Знаурского района Южной Осетии с населением приграничных сел Грузии, в связи с нерешенностью территориальных споров вокруг фруктовых садов.
14 октября группой военных наблюдателей от трех сторон совместно с представителем Миссии ОБСЕ был проведен мониторинг в районе населенного пункта Нули и населенного пункта Гвертев по факту обострения ситуации в данном районе.

Рис. 2. Текст реферата (рус. яз.)

A ceasefire between Georgia and breakaway South Ossetia signed late on Friday night was ended when shooting broke out in the South Ossetian village of Sarabuk and the Georgian village of Tamarasheni, killing 17 people and injuring 30, MosNews writes. Of the 17 casualties, 15 were on the South Ossetian side, the Russian Information Agency Novosti reported, citing Georgian regional police.

The shootings, carried out by both sides the very day when Georgian and Ossetian military representatives met in Tskhinvali to establish peacekeeping posts with Russian troops, involved mercenaries from the neighboring North Caucasus region, where Chechnya is fighting its own separatist war with Moscow, Interfax reported, citing regional police.

The Joint Control Commission involving Georgia, Russia and South Ossetia, meeting Friday to regulate the conflict there, signed a preliminary ceasefire at midnight, with pledges from Russia to add peacekeepers to the breakaway region, which has said it wants to join Russia.

Fighting is continuing in the area, the Itar-Tass news agency reports.

Georgian Interior Minister Irakli Okruashvili said there would be no more talks following the recent attacks.

Flying into Eredvi by helicopter to evacuate the wounded, he accused South Ossetia of double standards, saying the authorities in the region's main city Tskhinvali were clearly incapable of sticking to the ceasefire.

As part of the deal, the two sides had agreed to create additional buffer zones between their positions.

These would be patrolled by Russian peacekeepers and monitored by the Organisation for Security and Co-operation in Europe (OSCE).

The commander of Georgia's peacekeeping battalion in South Ossetia, Alexander Kiknadze, said Georgian villages in the region came under heavy artillery fire in the latest attack.

He told Georgia's Rustavi 2 television that his battalion returned fire and had inflicted casualties on the South Ossetian side.

But Irina Gagloyeva, a spokeswoman for the South Ossetian authorities, said Georgian forces were the first to fire.

Темы:

Политические представители сторон и посредников урегулирования конфликта
Миротворцы России в зоне конфликтов на территории стран СНГ

Объекты:

Грузия
Южная Осетия
Россия

Рис. 3. Исходный текст документа (анг. яз.)

A ceasefire between Georgia and breakaway South Ossetia signed late on Friday night was ended when shooting broke out in the South Ossetian village of Sarabuk and the Georgian village of Tamarasheni, killing 17 people and injuring 30.

The Joint Control Commission involving Georgia, Russia and South Ossetia, meeting Friday to regulate the conflict there, signed a preliminary ceasefire at midnight, with pledges from Russia to add peacekeepers to the breakaway region, which has said it wants to join Russia.

Flying into Eredvi by helicopter to evacuate the wounded, he accused South Ossetia of double standards, saying the authorities in the region's main city Tskhinvali were clearly incapable of sticking to the ceasefire.

These would be patrolled by Russian peacekeepers and monitored by the Organisation for Security and Co-operation in Europe.

Рис. 4. Текст реферата (анг. яз.)

4.3 Оценка работы метода

Экспертная оценка проводилась только для текстов на русском языке. Методика оценки заключалась в следующем.

Трем экспертам были предложены 10 текстов документов и их рефераты. Экспертам предстояло ответить на следующие вопросы, выбрав ответ из шкалы оценки:

1. Насколько полно реферат отражает содержание документа? (0 – не отражает, 1 – не достаточно полно, 2 – удовлетворительно).

2. Присутствует ли избыточность в реферате? (0 – да, много, 1 – да, не слишком много, 2 – нет).

3. Удовлетворяет ли реферат представлению о связности текста?

(0 – нет, 1 – встречаются не связанные предложения, 2 – да).

4. Оцените длину реферата

(0 – слишком длинный, 1 – очень короткий, 2 – оптимальный).

Результаты экспертных оценок приводятся в табл. 3. По каждому тексту подсчитывались суммарные оценки экспертов, а для каждого эксперта вычислялась его суммарная оценка по всем текстам. Наиболее часто эксперты снижали оценки из-за длины реферата (слишком длинный) и из-за встречаемости предложений, нарушающих картину связности текста. От длины текстов, по которым составлялись рефераты, экспертные оценки практически не зависели.

Сравнительная оценка данного метода с другими широко известными методами реферирования не проводилась на большой выборке тестовых заданий. Однако мы провели эксперимент по реферированию на ряде русскоязычных текстов на основе дан-

ного метода и метода, предложенного Г.Г. Белоноговым [3].

Таблица 3
Экспертные оценки качества рефератов

N текста	Оценки экспертов			
	1	2	3	Σ
1	7	8	6	21
2	6	5	6	17
3	6	7	5	18
4	5	8	6	19
5	8	6	4	18
6	5	7	6	18
7	5	6	6	17
8	7	7	8	22
9	6	5	5	16
10	6	7	7	20
Итого	61	65	59	186

Этот метод заключается в выделении из текста каждого предложения частотных словосочетаний, вычислении весов словосочетаний, равных произведению количества слов в словосочетании на частоту их встречаемости в тексте, подсчете весов предложений с помощью суммирования весов словосочетаний, входящих в их состав, и включения в реферат предложений с наибольшими весами.

Рефераты, полученные по двум методам, для большинства текстов были похожи.

Например, для текста, рассмотренного в п. 3.2, с помощью метода Белоногова Г.Г. был составлен реферат, приведенный ниже:

Командование Смешанных сил по поддержанию мира (ССПМ) в зоне грузино-осетинского конфликта выступает инициатором встречи представителей администрации Знаурского района Южной Осетии с населением приграничных сел Грузии, в связи с нерешенностью территориальных споров вокруг фруктовых садов, сообщил корреспонденту ИА REGNUM помощник командующего ССПМ по работе со СМИ подполковник Юрий Верещак. 14 октября группой военных наблюдателей от трех сторон совместно с представителем Миссии ОБСЕ был проведен мониторинг в районе населенного пункта Нули (территория Грузии) и населенного пункта Гвертев (Южная Осетия) по факту обострения ситуации в данном районе.

Видно, что в реферате, составленном по методу Белоногова Г.Г., отсутствует первое предложение из исходного текста, которое имеется во втором реферате. Однако во втором реферате текст предложений сокращен за счет исключения пояснений, а в первом – предложения включаются в реферат в том виде, в каком они извлекаются из текста (фрагменты, не вошедшие во второй реферат выделены курсивом).

5 Выводы и обсуждение результатов

Предложенный метод составления рефератов, рассматриваемый в данной работе, может быть с успехом применен в современных информационных системах при обработке больших информационных потоков, когда проводится автоматическое рубрицирование документов. По сравнению с системами реферирования, в которых проводится полный цикл обработки документов, данный метод позволяет значительно сократить временные затраты на составление реферата.

Проведенная независимыми экспертами оценка качества реферирования показала, что метод, в целом, дает удовлетворительные результаты. Отмеченная экспертами в ряде рефератов слишком большая их длина является недостатком, который можно преодолеть за счет укорачивания длинных предложений.

В работе [5] описывается метод выделения по одному фрагменту из каждого предложения реферата вплоть до исчерпывания заданного лимита (в символах). Однако на наш взгляд, при таком подходе невозможно сформировать гладкий связный текст. Здесь требуется другой подход, близкий к предложенному в [9], основанный на синтаксическом анализе текста предложения. Отталкиваясь от результатов синтаксического анализа предложения, можно отсекавать распространенные дополнения, причастные и деепричастные обороты. В дальнейших исследованиях мы собираемся опробовать алгоритм синтаксического анализа для реферирования.

В ряде случаев эксперты отметили несвязность текста реферата. Для решения проблемы связности можно использовать методы распознавания анафорических связей. Эта сложная задача может быть предметом отдельного исследования, однако с целью улучшения вида реферата на основе эмпирических наблюдений можно разработать алгоритм для распознавания анафоры. Авторы располагают подобным алгоритмом и программой с точностью распознавания ~60% [1].

Литература

- [1] Абрамова Н.Н., Абрамов В.Е. Автоматическое составление обзорных рефератов новостных сюжетов // Интернет-математика 2007. – Екатеринбург, 2007. – С. 1–11.
- [2] Альгулиев Р.М. Автоматическое реферирование документов с извлечением информативных предложений // Вычислительные технологии. – 2007. – Т. 12, № 5. – С. 5–15.

- [3] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. – М. : Русский мир, 2004. – 246 с.
- [4] Браславский П.И., Густелев В. Система автоматического реферирования новостных сообщений на основе машинного обучения // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Девятой Всероссийской научной конференции RCDL '2007 (Переславль-Залесский, Россия, 15–18 октября 2007 г.). – Переславль-Залесский : Изд-во «Университет гор. Переславля», 2007. – С. 142–147.
- [5] Браславский П.И., Колычев И.С. Автоматическое реферирование веб-документов с учетом запроса // Интернет-математика 2005. – М. : Ян-декс, 2005. – С. 485–501.
- [6] Гладун В.П., Святогор Л.А. Конспектирование естественно-языковых текстов с сохранением тематически-связной информации // Горизонты прикладной лингвистики и лингвистических технологий : Труды международной конференции MegaLing'2006 (Украина, Крым, Партенит, 20–27 сентября 2006 г.). – <http://www.megaling.crimea.edu/archieve/2006/rep/orts.en.htm>.
- [7] Глобус Е.И., Абрамов В.Е., Абрамова Н.Н. Автоматическое рубрицирование текстовой информации (на русском, английском, немецком и французском языках). Свидетельство об официальной регистрации программы для ЭВМ № 2006613783, 31 октября 2006 г.
- [8] Губин М.В., Меркулов А.И. Эффективный алгоритм формирования контекстно-зависимых аннотаций // Компьютерная лингвистика и интеллектуальные технологии : Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня 2005 г.). – М. : Наука, 2005. – С. 116–120.
- [9] Емашова О.А., Мальковский М.Г. Функциональные стили русского языка и их влияние на задачу автоматического реферирования текстов // Компьютерная лингвистика и интеллектуальные технологии : Труды международной конференции «Диалог'2007» (Бекасово, 30 мая – 3 июня 2007 г.). – <http://www.dialog-21.ru/dialog2007/materials/html/25.htm>
- [10] Журавлев С.В., Добров Б.В. УИС "РОССИЯ". Автоматическое тематическое индексирование полнотекстовых документов // Научно-практическая конференция «Проблемы обработки больших массивов неструктурированных текстовых документов» (21-22 мая 2001 года, г. Москва). – <http://www.fep.ru/text/dataarrays03.html>.
- [11] Ефименко И.В. Лингвистические аспекты кросс-языкового реферирования: синтез текстов под управлением предметных онтологий // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006 (25–28 сентября 2006 г., Обнинск) : Труды конференции : в 3 т. – М. : Физматлит, 2006.
- [12] Яцко В.А. Симметричное реферирование: теоретические основы и методика // НТИ. Сер. 2, № 5, 2002. – С. 18–28.
- [13] Document Understanding Conferences (DUC) Web site, 2008. <http://duc.nist.gov>
- [14] Hahn U., Mani I. The Challenges of Automatic Summarization // Computer, vol.33, no.11, pp. 29–36, Nov., 2000. <http://www.osp.ru/os/2000/12/178370/>.
- [15] Harman D., Over P. The effects of human variation in DUC summarization evaluation. In Text summarization branches out workshop at ACL'2004.
- [16] Lee C.B., Kim M.S., Park H.R. Automatic summarization based on principal component analysis // EPIA 2003 : Portuguese conference on artificial intelligence N°11, Beja, 2003, vol. 2902, pp. 409–413.
- [17] Leontyeva N. Semantic dictionary for Text Understanding and Summarization // International Journal of Translation. Vol. 15. No. 1. Ed. M. Blekhan. New Delhi, 2003.
- [18] Luhn H. The automatic creation of literature abstracts. In IBM Journal of Research and Development, Vol. 2(2), pp. 159–165, 1958.
- [19] Mani I. et al. The Tipster Summac Text Summarization Evaluation // Proc. 9th Conf. European Chapter of the November 2000.
- [20] Nenkova A., Passonneau R. Evaluating content selection in summarization: The pyramid method. In Proceedings of HLT/NAACL 2004, Boston, MA, USA, 2004.
- [21] Nomoto T., Matsumoto Y. The diversity-based approach to open-domain text summarization. In Information Processing & Management, 2003, 39, pp. 363–389.
- [22] Tait J. Making Better Summary Evaluations. Proceedings of the International Workshop: "Crossing Barriers in Text Summarisation Research" // Horacio Saggion and Jean-Luc Minnel (eds.), Borovets, Bulgaria, September 2005, pp. 26–31. http://www.dcs.shef.ac.uk/~saggion/CBTS_Papers/MS08.pdf.

On the Automatic Summarization Method based on the Results of Automatic Text Classification

V.E. Abramov, N.N. Abramova

This paper describes an automatic summarization method using the results of previously conducted classification textual information. Usually, summarization and classification are considered as separate tasks, although they are interrelated: based on the same procedure automatic word processing (morphological and parse). The idea of the method consists in maximum use of the results of previous

phases of processing and compile of summaries after determining basic order paper. Based on the received results it is possible to conclude that the proposed method gives acceptable quality of the summarization.

Автоматическое составление обзорных рефератов новостных сюжетов*

© С.Д. Тарасов

Балтийский Государственный Технический Университет им. Д.Ф.Устинова «ВОЕНМЕХ»
tarasov_sd@mail.ru

Аннотация

Работа посвящена одной из актуальных проблем автоматического реферирования – составлению обзорных рефератов по набору документов. Рассмотрен новый на сегодняшний день алгоритм ранжирования связных структур (Manifold Ranking Algorithm) применительно к автоматическому реферированию новостных сюжетов. Алгоритм позволяет учитывать как зависимости между предложениями внутри одного документа, так и зависимости между всеми предложениями коллекции. Проведен анализ возможности использования алгоритма для русского языка. Построена пробная система автоматического реферирования. Приведены результаты работы системы. Сформулированы основные проблемы реализации системы и возможные методы их решения. Оценка качества работы системы произведена при помощи критерия ROUGE. Произведено сравнение результатов работы построенной системы с результатами в DUC 2003, DUC 2005.

1 Введение

Задача автоматизации реферирования текстовой информации на сегодняшний день остается очень актуальной, несмотря на огромное количество появившихся в последние годы публикаций. Это вызвано, в первую очередь, необходимостью в условиях постоянного роста информации знакомить специалистов и других заинтересованных людей с необходимыми им документами, представленными в сжатом виде, но с сохранением их смысла. В обзорной статье [1] описывается современное состояние в области автоматического реферирования, а также основные направления и пути развития. В традиционных методах реферирования чаще всего используются различные модификации подхода Г. Луна

[9], известного с конца 50-х годов XX века, который заключается в отборе предложений с наибольшим весом для включения их в реферат, а также подходы, сочетающие традиционный подход с некоторыми новыми элементами. Вес предложения определяется как сумма частот, входящих в него значимых слов. В работе [2] описан метод, в котором в качестве значимых элементов выбираются не слова, а словосочетания. Развитие этого подхода есть в работе [3].

При формировании и показе сообщений новостных сюжетов приобретает актуальность задача составления обзорных рефератов (по некоторому набору документов), в которых были бы представлены все основные вопросы, затрагиваемые в каждом документе, но в обобщенном виде без повторений информации.

Составление обзорных рефератов относится к новым сферам применения автоматического реферирования, также как и получение одноязычных рефератов, охватывающих источники на разных языках, использование гибридных источников (например, статистической информации и сведений из баз данных), составление мультимедийных рефератов.

За рубежом в рамках конференций по проблемам автоматического аннотирования DUC (Document Understanding Conference) и текстового реферирования TSC (Text Summarization Challenge) данному направлению исследований придается очень большое значение. Автоматическое реферирование набора новостных сюжетов реализовано в таких крупных новостных ресурсах, как Google News (<http://news.google.com/>), NewsBlaster (<http://www.newsblaster.com/>), Yandex News (<http://news.yandex.ru/>).

Недавно разработанный алгоритм ранжирования связных структур (Manifold Ranking Algorithm) [4] может быть использован для ранжирования любых информационных примитивов: текстов, предложений, изображений, звуков. В этом случае любой вид информации должен быть представлен в векторном пространстве. В задачах автоматического аннотирования данный алгоритм может быть применен для ранжирования предложений набора документов и отбора наиболее значимых из них для включения в обзорный реферат. В задачах ранжирования результатов информационного поиска «отправной точкой»

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

алгоритма является запрос, и ранг предложений определяется как мера их «информационной близости» запросу. В задаче автоматического реферирования «отправной точкой» алгоритма можно считать «тему» кластера. В этом случае, правда, тема должна быть сформулирована максимально четко и подробно, т.к. алгоритм основан на анализе лексики и не может учитывать возможную семантическую близость лексически «удаленных» предложений. Принцип автоматического реферирования набора документов на основе алгоритма пространственного ранжирования подробно описан в [5]. Там же приведена довольно подробная оценка работы метода. Мне представилось интересным опробовать данный алгоритм для автоматического реферирования новостных сюжетов на русском языке, а также провести свои оценки качества реферирования. В процессе работы были выявлены некоторые недостатки метода, а также намечены основные возможные пути улучшения качества реферирования.

2 Алгоритм ранжирования связанных структур

2.1 Обзор

Алгоритм Manifold Ranking позволяет описать связную структуру текста при помощи матриц. Изначально алгоритм предполагает выделение элементов (предложений) наиболее близких заданному (теме). Такая интерпретация характерна задаче информационного поиска. Для автоматического реферирования также выделяется набор предложений, наиболее близких заданной теме кластера, однако обязательным является применение алгоритма отсеечения «похожих» предложений, что особенно актуально для многодокументного аннотирования.

Автоматическое реферирование набора документов с использованием алгоритма ранжирования связанных структур состоит из двух этапов:

Вычисление ранга каждого предложения. Этим решается задача ранжирования всех предложений в соответствии с их «близостью» заданной теме кластера.

Применение алгоритма отсеечения предложений, наиболее похожих на те, что уже попали в обзорный реферат. Этим решается задача исключения из обзорного реферата одинаковых или близких предложений.

В результате некоторое количество предложений с наибольшим рангом выбирается для результирующего реферата. Порядок следования предложений в общем случае никак не специфицируется подходом. Мною был реализован самый простейший алгоритм выборки предложений в порядке их относительного следования с приоритетом для более коротких предложений, что является естественным для русского языка. Строго говоря, вопрос связности полученного реферата является отдельной темой исследования. Некоторые методы решения представлены в [6]. Исходя из спецификации, алго-

ритм ранжирования связанных структур оперирует двумя понятиями:

Информационная значимость: По заданному набору предложений $\bar{X} = \{x_i | 1 \leq i \leq n\}$ и заданной теме T вычисляется вектор $\bar{f} = \{f_0 \dots f_n\}$ инфор-

мационной значимости каждого предложения $\bar{x}_i$. Информационная значимость предложения определяется как степень близости к заданной теме T . Предполагается, что тема кластера T наиболее полно отражает содержание набора документов и содержит наиболее полный набор лексики.

Информационная новизна: Для каждого предложения определяется его близость с другими предложениями набора. В итоге суммарный рейтинг, который определяет попадание предложения в обзорный реферат, рассчитывается с учетом, как информационной значимости предложения, так и его «информационной новизны».

2.2 Алгоритм

Алгоритм ранжирования связанных структур является универсальным алгоритмом ранжирования объектов с учетом их внутренней связанной структуры. Объекты должны быть представлены векторами в Евклидовом пространстве. В этом случае полагается, что «близость» двух объектов представленных векторами может быть вычислена, как Евклидова мера или скалярное произведение векторов. Целью алгоритма является упорядочить объекты, с учетом внутренних связей объектов между собой. Формально, связанная структура объектов представляется как некий взвешенный граф, вершинами которого являются сами объекты, а в качестве весов дуг задаются евклидовы расстояния между объектами. В случае ранжирования предложений с целью отбора наиболее значимых из них для построения обзорного реферата алгоритм можно формализовать следующим образом:

- [1] Задается набор структур (предложений) $\bar{X} = \{\bar{x}_i | 1 \leq i \leq n\} \subset R^m$, где $\bar{x}_0$ – описание темы кластера T . Мы полагаем, что тема формулируется одним предложением.
- [2] Вводится $f: \bar{X} \rightarrow R$ – отображение, которое ставит в соответствие каждой точке $\bar{x}_i (0 \leq i \leq n)$ значение ранга f_i . Мы можем рассматривать f как вектор $f = [f_0, f_1, \dots, f_n]^T$
- [3] Задается вектор $y = [y_0, y_1, \dots, y_n]^T$. Согласно алгоритму $y_0 = 1$, т.к. $\bar{x}_0$ – тема кластера (в задачах информационного поиска $\bar{x}_0$ соответствует фразе поискового запроса), и $y_i = 1, i \in (1, n)$ для всех остальных предложений.

- [4] Каждое предложение (объект) представляется в векторном пространстве следующим образом:
 $x_i = [tf_0, tf_1, \dots, tf_n]^T$, где tf_k – стандартная TF-ISF мера относительной важности термина t_k .
- [5] Набор предложений представляет собой взвешенный граф с матрицей весов W . Для каждой пары $\overline{x_i}$ и $\overline{x_j}$ предложений вычисляется вес их «лексической близости» при помощи стандартной евклидовой меры. Таким образом:

$$W_{i,j} = Sim(\overline{x_i}, \overline{x_j}) \quad (1)$$

где

$$Sim(\overline{x_i}, \overline{x_j}) = \frac{\overline{x_i} \cdot \overline{x_j}}{\|\overline{x_i}\| \cdot \|\overline{x_j}\|} \quad (2)$$

Причем $W_{i,i} = 0$ для того, чтобы полученный граф не содержал циклов. Следует отметить, что полученная матрица весов является симметричной относительно своей главной диагонали.

Рассмотрим пример для случая с тремя предложениями-кандидатами, которые необходимо ранжировать.

Тема кластера: «Мама мыла раму»:

Таблица 1. Предложения кластера

№	Обозн.	Текст
1	X0	Первого октября мама мыла раму
2	X1	Мама мыла раму
3	X2	Мама мыла раму тряпкой

В результате морфологического разбора могут быть вычислены TF-ISF метрики для всех термов предложений (табл. 2). Предложение с номером 0 представляет собой тему кластера (x_0).

Таким образом, исходные данные представления предложений в евклидовом пространстве будут следующие:

$$\begin{aligned} x_0 &= (0.33; 0.33; 0.00; 0.00; 0.33; 0.00) \\ x_1 &= (0.20; 0.20; 0.48; 0.48; 0.20; 0.00) \\ x_0 &= (0.33; 0.33; 0.00; 0.00; 0.33; 0.00) \\ x_0 &= (0.25; 0.25; 0.00; 0.00; 0.25; 0.60) \end{aligned}$$

Матрица весов в этом случае будет выглядеть следующим образом:

$$W = \begin{pmatrix} 0,000 & 0,457 & 1,000 & 0,587 \\ 0,457 & 0,000 & 0,457 & 0,268 \\ 1,000 & 0,457 & 0,000 & 0,587 \\ 0,587 & 0,268 & 0,587 & 0,000 \end{pmatrix}$$

Построенный граф может быть представлен как:

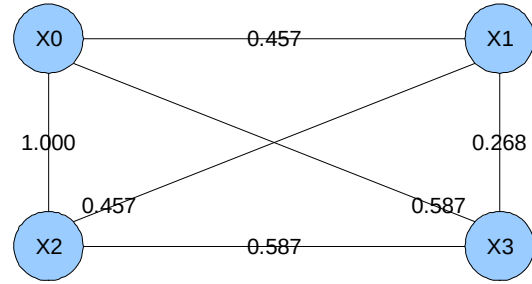


Рис. 1. Граф связности текста

Матрица весов подвергается симметричной нормализации

$$S = D^{-1/2} \cdot W \cdot D^{-1/2}, \quad (3)$$

где D – диагональная матрица, где $D_{i,i}$ равен сумме элементов i -ой строки матрицы W . Нормализация матрицы необходима, для того, чтобы итеративный алгоритм сходиллся.

$\overline{f}$ вычисляется как результат итеративного процесса:

$$\overline{f}(t+1) = \alpha \cdot S \cdot \overline{f}(t) + (1-\alpha) \cdot \overline{y} \quad (4)$$

Согласно теореме в [4] итеративный процесс сходится к $\overline{f}^*$. Далее полагается, что f_i^* – полученный ранг предложения с номером i . Интуитивно, алгоритм заключается в постепенном распространении объектами своего ранга на смежные объекты-вершины. Таким образом, ранг f_i каждого предложения $\overline{x_i}$ вычисляется не только с учетом «близости» его к эталонному объекту (теме кластера T), но и с учетом связной структуры текста, т.е. ранг «распространяется» по графу с учетом весов связей структур.

Для приведенного выше примера итеративный процесс для вычисления ранга выглядит следующим образом:

$$\begin{aligned} f^0 &= (0.00; 0.00; 0.00; 0.00) \\ f^1 &= (0.40; 0.00; 0.00; 0.00) \\ f^2 &= (0.40; 0.09; 0.12; 0.09) \\ f^3 &= (0.46; 0.13; 0.14; 0.13) \\ f^4 &= (0.48; 0.16; 0.17; 0.16) \\ f^5 &= (0.50; 0.17; 0.19; 0.18) \end{aligned}$$

Таблица 2. Представление набора предложений в евклидовом пространстве

№	МАМА (С)	МЫТЬ (Г)	ОКТЯБРЬ (С)	ПЕРВОЕ (Ч)	РАМА (С)	ТРЯПКА (С)
0	0,33	0,33	0,00	0,00	0,33	0,00
1	0,20	0,20	0,48	0,48	0,20	0,00
2	0,33	0,33	0,00	0,00	0,33	0,00
3	0,25	0,25	0,00	0,00	0,25	0,60

Можно также предположить, что связи между предложениями одного документа, а также связи между предложениями различных документов набора должны быть дифференцированы. В этом случае полагается, что:

$$W = W_{inner} + W_{int ra}, \quad (5)$$

где W_{inner} – матрица весов связей предложений внутри одного документа, а $W_{int ra}$ – матрица весов связей предложений разных документов.

В этом случае целесообразно ввести коэффициенты:

$$W = \lambda_1 \cdot W_{inner} + \lambda_2 \cdot W_{int ra}. \quad (6)$$

2.3 Алгоритм усечения сходных предложений

Для формирования итогового обзорного реферата по набору документов необходимо выполнить следующее:

исключить из рассмотрения предложения, повторяющиеся по своей структуре те, что уже попали в обзорный реферат.

Выполнить итоговую сортировку предложений с целью получения более-менее связного текста.

Для задачи внесения в итоговый ранг фактора «информационной новизны» используется следующий алгоритм:

- [1] Инициализируются два множества $A = \emptyset$ и $B = \{x_i \mid i = 1, 2, \dots, n\}$. Для каждого предложения В текущий ранг принимается равным f_i^* .

$$RankScore(x_i) = f_i^*, i = 1, 2, \dots, n \quad (7)$$

- [2] Предложения множества В сортируются в соответствии с их текущим рангом в порядке убывания.

- [3] Полагая, что предложение $\overline{x_i}$ имеет наивысший ранг, оно перемещается из В в А. Ранг оставшихся в В предложений $\overline{x_j}$ рассчитывается как

$$RankScore(x_j) = RankScore(x_j) - \omega \cdot \overline{S_{j,i}} \cdot f_i^* \quad (8)$$

Где $\omega > 0$ – фактор усечения сходных предложений, а

$$\overline{S} = D^{-1} \cdot W \quad (9)$$

- [4] Процесс повторяется, пока В не станет пустым.

Вопрос окончательной сортировки предложений в обзорном реферате является темой отдель-

ного исследования. Следует отметить, что в данном случае возможны как лексико-ориентированные алгоритмы, основанные на более детальном анализе предложений, так и семантические алгоритмы, основанные на сравнении поверхностно-семантических графов. Последние также позволяют исключать лексически разные, но семантически подобные предложения, однако не обладают такой простотой, быстротой и универсальностью, как первые. Строго говоря, темой отдельного исследования может быть вектор параметров, влияющих на окончательный порядок следования предложений в итоговом реферате, таких как количество слов в предложении, количество значимых слов, принадлежность предложения к определенному документу, дата опубликования документа, относительная сложность восприятия предложения и т.д.

3 Реализация

Система автоматического реферирования новостных сюжетов реализована как набор скриптов на языке PHP с Web-интерфейсом. Для матричных вычислений было разработано специальное расширение для языка PHP *php_math* на основе MTL[8], позволяющее осуществлять быстрые матричные вычисления. Для морфологического анализа была задействована библиотека *phpmorphology*, основанная на словарях проекта АОР[7]. Система позволяет подбирать параметры $\alpha, \lambda_1, \lambda_2, \omega$ и количество предложений, попадающих в обзорный реферат. По исходному корпусу строится обзорный реферат с выводом всех промежуточных расчетов. Система позволяет получать обзорные рефераты с заданными параметрами на лету. Система доступна по адресу <http://openthesaurus.ru/manifold/>.

4 Исходные данные

В качестве исходных данных для оценки работы алгоритма был взят набор кластеров новостной тематики, любезно предоставленный НИВЦ МГУ.

Для кластера «На севере Омской области выпал разноцветный снег» содержащего 8 документов (всего 61 предложение) был получен обзорный реферат из 4 предложений при значении параметров $\alpha = 0.6, \lambda_1 = 0.3, \lambda_2 = 1, \omega = 15$.

«Представители властей заявили, что если вдруг выяснится, что разноцветный снег в Сиби-

ри выпал из-за промышленных выбросов, нарушителей привлекут к уголовной ответственности. Пока специалисты только говорят, что аномальные осадки не опасны для здоровья. Кроме того, необычный снег выпал в Томской и Тюменской областях. Вчера были обнаружены первые лабораторные исследования выпавшего 31 января в Омской области желто-оранжевого снега».

5 Оценка

Предварительная оценка работы алгоритма позволяет утверждать о возможности применения его в модифицированном виде для корпусов новостных сюжетов на русском языке. На сегодняшний день на основе ручных аннотаций, любезно предоставленных НИВЦ МГУ (Б.В. Добров) проведена оценка качества системы реферирования при помощи меры ROUGE.

$$ROUGE-N = \frac{\sum_{S \in Re} \sum_{f \in Sum} \sum_{n-gram \in S} Count_{match}(n-gram)}{\sum_{S \in Re} \sum_{f \in Sum} \sum_{n-gram \in S} Count(n-gram)} \quad (10)$$

Таблица 3. Результаты оценки

№	Тема кластера	ROUGE-1	ROUGE-2	ROUGE-3
1	В Гальском районе Абхазии неизвестные похитили главу районной избирательной комиссии	0.4286	0.2545	0.2037
2	Китай успешно запустил на орбиту навигационный спутник "Бэйдоу"	0.2581	0.0656	0.0333
3	Секретаршу из Coca-Cola признали виновной в краже секретов компании	0.5600	0.4286	0.3542
4	На севере Омской области выпал разноцветный снег	0.4655	0.3157	0.2500

Таблица 4. Сравнение с DUC

	DUC 2003	DUC 2005	Построенная система
ROUGE-1	0.37332	0.38434	0.42805
ROUGE-2	0.07677	0.07317	0.26610

Приведенные оценки являются предварительными и требуют дальнейшей корректировки с учетом большого количества обзорных рефератов.

6 Будущая работа

Для улучшения качества работы системы необходимо выполнить следующие действия:

- [1] Улучшить алгоритм распознавания и разрешения анафор.
- [2] Добавить в систему синонимию. Например, лексемы «км» и «километр», «15» и «пятнадцать» должны рассматриваться как совершенно идентичные. Для других вариантов возможно введения коэффициента синонимии.

Мера ROUGE-N представляет собой обобщенную статистическую меру, выражающую какой процент лексических единиц (N-gram, - последовательностей из N лексем), входящих в состав ручной, построенной независимым экспертом, аннотации, попадает в обзорный реферат.

5.1 Результаты оценки

Для предварительной оценки качества работы системы были вычислены меры ROUGE-1, ROUGE-2, ROUGE-3 нескольких обзорных рефератов, полученных при помощи системы, для которых есть ручная аннотация. Для получения рефератов были использованы следующие значения параметров: $\alpha = 0.6$, $\lambda_1 = 0.3$, $\lambda_2 = 1$, $\omega = 15$.

Результаты оценки представлены в таблице 3. Полученные результаты можно сравнить с результатами, полученными в DUC [5] (таблица 4).

- [3] Провести более основательную оценку качества работы системы на основе большего количества ручных рефератов.

7 Заключение

Задача автоматизации реферирования текстовой информации на сегодняшний день остается очень актуальной. Алгоритм ранжирования связанных структур зарекомендовал себя с положительной стороны как довольно эффективный для задач автоматического аннотирования, и в то же время, относительно легко реализуемый, и может применяться для автоматического реферирования корпусов новостных сюжетов на русском языке. Однако алгоритм требует дополнительных доработок с учетом особенностей как русского языка,

так и естественного языка вообще. Так, необходимо более детально рассматривать предложения, содержащие прямую речь, разработать более совершенный (чем в [3]) алгоритм разрешения анафор, а также обеспечить связность текста полученного обзорного реферата.

Литература

- [1] Hahn U., Mani I. "The Challenges of Automatic Summarization," Computer, vol.33, no.11, pp. 29-36, Nov., 2000.
<http://doi.ieeecomputersociety.org/10.1109/2.881692>
- [2] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. – М. : Русский мир, 2004. – 246 с.
- [3] Н.Н. Абрамова, В.Е. Абрамов. Автоматическое составление обзорных рефератов новостных сюжетов // Труды 9-й Всероссийской научной конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 2007.
- [4] Zhou et al., 2003b D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. Schölkopf. Ranking on data manifolds. In Proceedings of NIPS'2003.
- [5] "Manifold-Ranking Based Topic-Focused Multi-Document Summarization" DUC 2003
<http://www.ijcai.org/papers07/Papers/IJCAI07-467.pdf>
- [6] Barzilay R. Sentence Ordering in Multidocument Summarization. Computer Science at Columbia University, Web seit, 2007,
http://www.cs.columbia.edu/nlp/papers/2001/barzilay_al_01.pdf
- [7] Автоматическая обработка текста.
<http://aot.ru/>
- [8] The Matrix Template Library.
<http://www.osl.iu.edu/research/mtl/>
- [9] Luhn The Automatic Creation of Literature Abstracts (context) – 1958.

Automatic compilation of news stories reviews

S.D. Tarasov

This work deal with one of the topical problems of automatic summarization – multi-document summarization in respect to news stories. This paper presents a novel extractive approach based on manifold-ranking of sentences to this summarization task. The manifold-ranking algorithm differentiates the intra-document and inter-document links between sentences with different weights. The possibility of the use the algorithm for Russian language is analyzed. A sample system for automatic summarization is build. This paper represents the sample summaries and describes experiments of summarization evaluation. The main problems of implementation of the system and possible methods of their solutions are formulated. The ROUGE criteria was used for evaluation. The results of work of built system are compared with the results of DUC 2003, DUC 2005.

* Автор выражает благодарность Б.В. Доброву (НИВЦ МГУ) за предоставленные материалы ручных аннотаций кластеров, а также за помощь в написании данной статьи.

Концепция объемного текста

© Н.В. Тупик

Филиал Дагестанского государственного университета (ДГУ), г. Каспийск
tupik_nv@mail.ru

Аннотация

Предлагается концепция объемного текста, состоящего из набора текстовых слоёв, где каждый следующий слой сжат относительно предыдущего, т.е. имеет меньше символов. Конечными вершинами служат вводимые в тексте положения. Каждое положение выражено коротким утвердительным предложением. Многовершинность такого текста позволяет рассматривать его как объемное тело с фрактальной структурой и предоставляет, как возможность читателю индивидуально и динамически выбирать траектории ознакомления с текстом, так и право выбора начальных точек этих траекторий. Таким образом, читатель смотрит на текст с «открытыми глазами».

1 Введение

Одним из условий развития общества является получение, накопление, сохранение и передача накопленной информации. «Для совершения всех ошибок не хватит никаких жизней никакого общества и далеко от печки таким образом не уйдёшь». Что и происходило на заре развития человечества, когда плотность населения была низкой, и это затрудняло передачу информации (накопленного общественного опыта) между соседними общинами, а гибель самих общин или их наиболее опытных членов заставляло следующее поколение начинать совершать все ошибки с самого начала (начинать опять от печки). Именно поэтому общественное развитие и шло медленно.

Вначале информационную функцию несли метки на местности (поломанные ветви, отмечающие тропы и места пропитания, зарубки, камни (менгиры) и т.д.). Затем эта функция перешла к изображениям местности, потом подключилась письменность, сначала в виде последовательных стилизованных изображений и в итоге в виде слогового и буквенного письма [1] на плоских носителях, в качестве которых служила поверхность камня (строительных конструкций), папирус, пергамент, береста, бумага. Плоский характер носителя текста и сформировал,

преимущественно, те особенности его формы представления, которые дошли до наших времён.

2 Плоский текст

2.1 Цели и задачи

Наличие текстов привело к тому, что процесс формализации информации с её закреплением, и её передача начали разделяться. Передача информации осуществлялась по принципу «делай как я», затем в виде заученных связанных фрагментов формализованных данных, в рифмованной или песенной форме, чтобы не исказить их при передаче. Таким способом тибетские монахи хранили, использовали и передавали свод добытых ими медицинских знаний, собранных затем в трактате «Чжуд-ши» [5]. Оба указанных принципа передачи информации действовали только во временном лаге жизни ближайших поколений связанных между собой членов общества. Наличие письменности позволило как оперативно и массово передавать формализованную информацию между членами обществ, проживающими в радиусе взаимодействия друг с другом, так и передавать её будущим поколениям без непосредственной (живой) связи между ними, в том числе и тем, с кем непосредственной связи быть не могло («через века»), например археологам.

Письменность, позволяя передавать формализованную информацию, требует при этом единства алфавита (представлений), скрываемого за письменными знаками или их наборами, в отличие от рисунков, для которых такой алфавит требуется в меньшей степени или не требуется вовсе, если в качестве него может выступать окружающая среда. Если нет уверенности в единстве используемого семантического алфавита, то тогда устная речь (обмен мнениями, диалог) более предпочтительна для оперативной передачи информации, чем текстовое сообщение. Диалог позволяет в процессе передаче информации выявить семантические различия и на ходу скорректировать сообщение таким образом, чтобы его семантическая часть была донесена до слушателя правильно. Это правило действует и для оперативной передачи информации в тех областях деятельности (особенно в развивающихся научных направлениях), где отсутствует ещё общепринятый семантический алфавит и специалисты говорят «на разных языках» при этом пользуясь одинаковыми наборами письменных знаков (терминов), т.е. в одинаковые термины вкладывают различный (часто

даже не пересекающийся) смысл. Для передачи информации будущим поколениям (без возможности прямого диалога) любой способ её передачи (устный, текстовый, видеоряд) будет эквивалентен текстовому сообщению.

2.2 Взрывной рост объемов

С появлением текстов их объёмы, генерируемые в мире и доступные для читателя, от века к веку росли в геометрической прогрессии, и сегодня объем генерируемой за год информации составляет петабайты (10^{15} байт) и продолжают расти. Если вычесть из этой информации медиа-, аудио- и графическую форму представления, то и тогда объем генерируемой за год текстовой информации будет очень и очень значительным. Ещё раз отметим, что в текстовый вид обычно превращают формализованную информацию, которая не может быть адекватно выражена в других формах, таких как рисунки, графики, картинки и т.д. так как представленные в таком виде они всё равно требуют пояснения.

2.3 Внесение структуры в текстуру текста

Рост объемов текстовой информации потребовал разработки механизмов его представления таким образом, чтобы облегчить читателю её восприятие. Например, многие древние тексты содержали только согласные в словах и в них были опущены гласные. Это значительно затрудняет расшифровку таких текстов сегодня, по прошествии нескольких тысячелетий, когда утерян (существенно изменился) контекст, при котором эти тексты создавались и в котором они однозначно интерпретировались. В качестве механизмов, облегчающих восприятия текстов было введение пробелов между словами. Роль этого новшества очень легко оценить, если попытаться прочитать заголовок газетной статьи, который не вмещался в полосу и пробелы в нём были сокращены до минимально возможного размера. Таким подходом к делу особенно часто страдают («грешат») местные газетные издательства. В этом случае наглядно проступает роль пробелов в тексте, для облегчения его восприятия, а ведь в русских текстах количество пробелов порядка 10% от всех печатных знаков материала. В данном материале из 28696 знаков 3689 пробелов, и это практически равно количеству слов в тексте. Следующим шагом было разбивка текста на предложения, абзацы (попробуйте почитать на каком-нибудь сайте текст, сформированный без абзацных отступов строк и пробелов между абзацами), главы, разделы (часто маркируя разделы буквицей). Все эти механизмы направлены на то, что бы однородную текстуру текста разнообразить, тем самым внося в него хоть какую-то структуру, за которую может зацепиться глаз читающего, и тем самым дать ему возможность осуществлять навигацию по тексту.

Следующим уровнем внесения структуры в текст были: оглавление, краткое содержание глав, подзаголовки, различного типа предметные указате-

ли (упоминаемых авторов, используемых терминов и т.д.) со ссылками на те или иные страницы текста, где эти термины упомянуты и т.д. [3]. Они давали возможность фрагментировать исходный текст и обеспечивали быстрый переход (получение доступа) к нужному фрагменту. С переводом плоского текста в электронный формат, этот уровень внесения рельефа в плоскостную структуру текста получил наибольшее развитие за счёт механизма гиперссылок, который легко реализуется в HTML-документах. К этому же уровню внесения рельефа в однородную текстуру текста относятся таблицы, схемы, графики, рисунки, графические элементы оформления в теле текста и т.д., т.е. графические образы, которые существенно отличаются по текстуре от самого текста и всегда на его фоне резко выделяются, тем самым концентрируя на себе внимание. К графическим элементам оформления плоского текста относятся и фрагменты текста (в том числе и отдельные слова), выполненные отличным от основного текста шрифтом (кеглем, размером, типом шрифта, цветом и т.д.). Пролистывая текст «по диагонали», читатель наталкивается на выделяющиеся в нём по текстуре графические образы, «на лету» оценивает их и, если они представляют для него интерес, начинает знакомиться с фрагментом текста, относящегося к данному графическому образу. Подборки графических образов, сгруппированные на вклейках или в конце текста без указания на фрагменты текста, к которым они относятся, роль элементов фрагментации исходного текста не выполняют, а служат лишь иллюстрациями, при последовательном знакомстве читателя с самим плоским текстом. Избыток графических элементов текста, особенно графических элементов оформления текста, приводит к потере их маркирующей функции в тексте, т.к. за ними перестаёт быть виден сам текст или текст читающим воспринимается как сложно оформленный в графическом плане. В этом случае, маркирующие элементы из выделяющихся (указывающих) переходят в разряд фоновых, т.е. составляющих основную текстуру текста. По этой причине не рекомендуется в основном теле текста использовать более трёх типов шрифтов или их начертаний. Заголовки, подзаголовки, оглавления, подрисовочные надписи и пр. в этот список не входят и обычно выполняются шрифтом и начертанием, отличным от тела основного текста с целью внесения структуры в однородный текст. Но механизм разбивки на главы, подзаголовки и т.д. задачу в конечном итоге не решает, т.к. следующий за заголовком текст опять-таки представляет собой последовательный плоский текст, из которого ещё предстоит выбрать те фрагменты (в процессе чтения), которые и дадут представление о вынесенном в заголовок положении.

2.4 Попытки ввести альтернативу тексту

Существующие сегодня попытки замены текста его рисованным аналогом («комиксом») отражают историческую тенденцию, направленную на повышение рельефности представления текста, которая

должна обеспечить увеличение скорости восприятия больших объемов информации за счёт «схватывания на лету», одним взглядом, сути передаваемого автором смысла. Но комиксы хороши для представления информации общебытового характера, которая хорошо интерпретируется через типовые, хорошо узнаваемые образы, и совершенно непригодны для передачи формализованной (свёрнутой и логической по своей сути) информации, такой как научная. Эту информацию представить в виде комиксов практически невозможно, т.к. для неё очень трудно подобрать образы, а полученные образы носят очень частный и неоднозначный характер и требуют развёрнутых текстовых пояснений.

Введение терминологии, особенно в новых направлениях исследований, выполняет роль сворачивания информации (создания нового образа), но выполненную текстовыми же средствами, и это не влияет на текстуру (рельефность) самого текста, но значительно сокращает его объем.

Таким образом исключить текст из восприятия нельзя, но и его плоский характер не позволяет разом схватить содержание всего материала. Не помогают, а скорее запутывают дело, и перекрёстные ссылки внутри материала (технология гипертекста), которые получили распространение при представлении текстов в электронном виде. Именно поэтому технология гипертекста не получает пока широкого распространения, т.к. она не позволяет познакомиться с материалом сверху вниз — от главных выдвигаемых положений в сторону их конкретизации и обоснования. "Плоский текст" не даёт возможности одним взглядом окинуть и оценить его содержательную составляющую, а гиперссылки обеспечивают только навигацию по тексту без сворачивания информации.

2.5 Межтекстовая навигация

Современные способы, обеспечивающие уже не выбор внутри текста, а выбор самих текстов из их совокупности, основаны на использовании введения и заключения к тексту, его аннотаций, рефератов, резюме, ключевых слов, индексов УДК и ББК, индексаторов и каталогов тематических подборок, поисковых машин и т.д. К сожалению, сегодня большинство надтекстовых элементов составляется самим автором текста, и они могут не учитывать надтекстовых корреляционных связей. При составлении этих элементов специалистами, возникает опасность субъективной расстановки акцентов, в зависимости от квалификации, научной специализации и пристрастий составителей. К тому же все эти элементы у специалиста меняются со временем и один и тот же текст, реферированный в разное время, может получить довольно различающиеся по своей сути результаты. Дело несколько может исправить система автоматического составления данных элементов текста, которая будет составлять эти элементы единым образом для всех текстов, независимо от времени их составления.

На сегодняшнем уровне не решают вопрос и поисковые машины, которые по заданному информа-

ционному запросу подбирают тексты и их фрагменты, релевантные запросу. Причина в том, что поисковые машины ищут входящие в текст термины и связки терминов, не ранжируя их по значимости для выбранных текстов или их фрагментов, т.е. не учитывают иерархию вложенного в текст смысла, которую хотел бы получить автор запроса. Хотя в этом направлении работы сейчас и ведутся. В результате использование указанных механизмов хоть и сокращает объем и расширяет зону поиска, но тем не менее не освобождает от необходимости последующего поэтапного (последовательного) знакомства с содержимым выбранных текстов или их фрагментов. И только после их прочтения можно сделать апостериорный вывод о том, следовало знакомиться с этой информацией или это оказалось пустой тратой времени.

2.6 Текст как оперативный обмен информацией

Кроме функции сохранения и передачи накопленного общественного опыта, текст выполняет и функцию оперативного адресного или перекрёстного (безадресного) обмена информацией в пределах горизонтального временного среза (текущей общественной практики) между членами общества. При адресной, т.е. от специалиста к специалисту, передаче текстовой информации её коэффициент новизны не велик, расхождение семантических полей пишущего и читающего, относительно затрагиваемых в тексте вопросов, так же минимален или заранее известен с большой долей вероятности. Таким образом, при адресной передаче информации она фактически для читающего хорошо сегментирована и ранжирована по уровням значимости. Читающий, в этом случае, знакомится с текстом не в виде сплошного его прочтения, а в виде выборочного знакомства с некоторыми его фрагментами.

При безадресной передаче информации такого единства и определённости семантических полей пишущего и читающего, как и точек зрения (моделей мира, из которых они берутся при написании текста или к которым они возводятся при его чтении) на излагаемую в тексте информацию, нет. При этом существует большая неопределённость для читающего в потребности подробного ознакомления с излагаемой информацией. Так как это может оказаться пустой тратой времени и ничего нового для читающего в этом тексте не будет, или в нём будет отсутствовать та информация, ради получения которой читающий и обратился к данному тексту. Использование общедоступных средств, таких как интернет, электронные архивы и тематические подборки и т.д. облегчает оперативный безадресный доступ к информации, создавая единое унифицированное (по способам размещения и поиска, формам представления и т.д.) информационное пространство для текстов, но имеет обратную сторону — «проклятие размерности». Это проклятие заключается в том, что объём релевантной запросу текстовой информации, который поисковая машина или иные средства тематического поиска и указания выбрала,

превышает мощностные характеристики восприятия индивидуума, если пользоваться стандартными способами извлечения смысловой информации из текста, а сужение запроса приводит к потере значимой для пользователя информации.

2.7 Выводы

Вот здесь и требуются новые способы представления текстовой информации, позволяющие быстрее («на лету») схватить её смысл и принять решение о нужности или ненужности более глубокого ознакомления с данным текстом или некоторыми его фрагментами. Например, программа ACDSec [2] снабжает каталоги с подборками изображений индивидуальным для каждой подборки рельефом, представляющим собой значительно уменьшенные копии нескольких изображений из подборки, наложенные на изображение каталога. В результате каждый каталог становится зрительно уникальным, если не содержит, конечно, одинаковые подборки рисунков. Окинув взглядом набор таких «рельефных» каталогов, значительно проще выбрать нужный, чем перебирать последовательно все, совершенно однотипные по внешнему виду каталоги, заглядывая внутрь каждого и после этого принимая решения – нужен ли этот каталог или следует перейти к следующему. Аналогичным механизмом придание рельефности каталогам снабжена и системная программная оболочка Windows Vista.

Все приведённые выше механизмы используются в современных текстах и направлены на одну цель – придать однородному потоку текста хоть какой-нибудь рельеф. Если рассмотреть введение указанных механизмов в динамику (во времени), то видно, что чем дальше идёт развитие общества и чем большие потоки информации в нём задействуются, тем больше количество и сложнее суть тех механизмов, которые используются для придания тексту рельефности. Современный огромный поток даже тематически подобранной текстовой информации настоятельно требует изменения формы представления текста и обеспечения такой его формы, которая предоставляла бы возможности взглянуть на текст как на объемное тело и быстро составить о нём общее представление. И в то же время иметь возможность тут же осуществить иерархическое погружение во фрагменты текста, которые заинтересовали читающего. Очевидно, подошла очередь и для перехода к полностью объемному механизму представления текста, опираясь на возможности электронных проигрывателей, каковыми являются компьютеры. Этому благоприятствует и повсеместное распространение компьютеров.

3 Объемный текст

3.1 Концепция

Концепция «объемного» (рельефного) текста заключается в том, чтобы представить текст в виде некоего рельефа, восприятие которого позволяло бы

сразу («на вскидку»), без углубления в сам текст, оценить его смысловое содержание, степень новизны, выбрать необходимые для более подробного ознакомления фрагменты, иметь возможность знакомиться с ними на разных уровнях детальности и по различным схемам. В этом случае появляется возможность динамически менять логику и регулировать детальность (глубину) ознакомления с текстом и т.д. Т.е. иметь возможность работать с текстом как с телом сложной формы, имеющим фрактальное строение, что выражается в самоподобии формы и иерархии смыслов. Такой способ построения объемного текста позволит начинать знакомство с текстом с любого места; осуществлять движение по тексту согласно логике заложенных в него смыслов; дать возможность оценить объем прочитанных фрагментов и соотношение прочитанных объемов; хранить траектории чтения; осуществлять навигацию, редактирование, реструктурирование, формирование собственных выборок из текста и т.д.

3.2 Логика перехода от плоского к объемному тексту

При объемном представлении исходный плоский текст подвергается многократному сворачиванию. На самом верхнем уровне расположены вводимые положения, которые имеют максимальную степень свёрнутости относительно исходного текста. Они служат вершинами объемного текста, и представляют собой короткие утвердительные предложения, отражающие вводимые в материале новые положения. Таких положений немного. Все подробности, детали и обоснования опускаются, как опускается и «информационное» обоснование вводимых положений. Краткие разъяснения по вводимым положениям даются на первом уровне детализации текста. При этом это должны быть два-три небольших утвердительных предложения. И далее по уровням продвижения вниз (углубления в текст) разъяснения по поводу вводимого положения расширяются, вплоть до подошвы – самого плоского текста, который является самым нижним (максимально расширенным) уровнем рельефа.

В плоском тексте могут быть значительные по величине фрагменты, которые в иерархии основных вводимых положений не отражены никак (появляются только на более нижних уровнях детализации текста), и являются лишь обоснованием такого представления, приводимые автором текста. Это могут быть примеры для такого обоснования, и в этом случае они занимают практически последние, самые нижние уровни рельефа, и как отдельные положения совершенно не выходят наверх, хотя в самом плоском тексте могут занимать львиную долю места. Но для вводимых положений это проходные моменты, лишь подтверждающие их правильность. Для американских стандартов, когда одна мысль – одна статья, вершина для плоского текста может оказаться единственной, а все другие появляются только на более низких уровнях детализации этой единственной вершины.

3.3 Особенности

Особенностью объемного текста является многовершинность на любом уровне, когда множество вершин на некотором уровне далее (при подъеме вверх) покрывается одной вершиной на следующем уровне. Таким образом, все вводимые положения, в конечном счёте, могут быть покрыты одной вершиной. Одна вершина на некотором уровне может превратиться во множество вершин на более низком уровне и при этом иметь продолжение исходной вершины на этом уровне. Чем глубже уровень, тем больше количество и размеры фрагментов текста и меньше степень их сжатия относительно плоского текста.

3.4 Динамическая структура

Технология представления плоского текста перед загрузкой его в 3D-интерфейс следующая. На самом верхнем уровне даны вводимые положения. Их немного и они даны краткими утвердительными предложениями. Далее эти положения всё более и более раскрываются по мере продвижения вниз по уровням (1, 2, 3 и т.д.) вплоть до самого нижнего уровня – фрагмента плоского текста, соответствующего данному положению (данной вершине). При этом в исходном плоском тексте этот фрагмент не обязательно должен быть слитным. Он может представлять собой набор из фрагментов плоского текста, исходно расположенных в разных его местах. Количество уровней между вершиной и основанием для различных вершин может быть разным. Для только намечаемых в материале положений количество уровней будет малое, для подробно рассматриваемых – большое.

Исходный набор утвердительных предложений играет роль вершин над плоским текстом, а постепенное разворачивание этих предложений по мере углубления в текст представляет собой переход от точечной вершины к более широкому её сечению, т.е. конкретизации (обоснованию) вплоть до подошвы. Исходный текст так же не остаётся неизменным. В нём динамически могут быть сформированы фрагменты как соответствующие вводимым положениям (покрываемых данной вершиной), так и адекватных промежуточным вершинам, возникающим по мере разворачивания содержимого исходной вершины (вводимого положения).

3.5 Роль читателя

Для читающего концепция «объемного» текста сводится к иерархическому смысловому его построению и наличию широкого набора механизмов навигации по такой структуре с возможностью редакторского оперирования фрагментами текста. Степенью и траекторией углубления в материал динамически (адаптивно) руководит читающий текст индивидуум, в соответствии со своими исходными знаниями, информационными потребностями и поставленными при знакомстве с данным мате-

риалом целями, в том числе и формируемыми по мере знакомства с самим текстом.

При переходе с уровня на уровень объем текста под вводимым положением постоянно расширяется. Читатель имеет возможность выбирать для ознакомления с каждым вводимым положением тот или иной уровень детальности представления текста. Для разных положений этот уровень представления текста читателем может быть выбран произвольно и независимо от других положений.

Если читатель постепенно погружается в уровень (работа в режиме приращения уровня), то у него есть возможность на каждом уровне любой вершины продолжить углубление в текст, не повторяя уже сказанного на предыдущем уровне. Поскольку на каждом уровне текст может представлять собой многовершинное продолжение, то и погружение в текст «в приращениях» возможно по каждой вершине независимо, т.к. текст на любом уровне имеет многовершинное продолжение своих фрагментов, независимо от других фрагментов, находящихся на этом уровне. При этом если читатель перейдёт полностью на более низкий уровень, то приращённые фрагменты, соответствующие этому новому уровню, автоматически станут достоянием этого уровня: войдут в состав уровня сразу по всем наметившимся вершинам предыдущего уровня. Но при этом, правда, будет осуществлён повтор всех фрагментов всех вышестоящих уровней данной вершины.

3.6 Спектр возможностей

Можно выбирать различные схемы знакомства с текстом. Например, такие стандартные как снизу (от плоского текста) вверх (к основным вводимым положениям) или сверху вниз, или начинать знакомство с любого уровня детальности и затем динамически, по мере знакомства с текстом, изменять степень углубления в то или иное вводимое в тексте положение или их группу, или распространять эту степень углубления на весь текст и т.д. Такие способы чтения не являются прерогативой только объемного текста. Они уже хорошо наработаны и ими читающие широко пользуются при знакомстве с плоским текстом. Например, начинают знакомиться с текстом с оглавления, выбирая только заинтересовавшие их главы или подразделы, или пролистывая текст «по диагонали» знакомятся только с локальными фрагментами, так или иначе выделенными графически (схемы, таблицы, рисунки и т.д.) в тексте, и обратившими на себя внимание. Задача объемного текста – предоставить более удобные технические и технологические механизмы поддержки таких методов чтения, и способствовать их развитию.

В результате чтение текста (знакомство с текстом) становится активным процессом, в котором читающий сам выбирает путь ознакомления с материалом, точку начала этого пути и уровень исходной и последующей детализации первоначально небольшого набора исходных вводимых положений.

Работа с текстом для читателя может закончиться и на этапе знакомства с набором вводимых положений, если таковые ему известны или не представляют для него интереса. Это значительно экономит время читающего, позволяет обрабатывать очень большие объёмы исходной текстовой информации, помогает быстро выявлять в них необходимую (новую и актуальную) информацию.

3.7 Связь с моделью мира индивидуума

Предлагаемая концепция объемного текста есть, по сути фрагмент модели мира [4], представляемой автором на суд читателя, а «гуляние» по такой модели читателя – есть попытка читателя соотнести свою модель мира с авторской, и здесь траектория «гуляния» очень много может сказать как о модели мира читающего, так и о механизмах взаимодействия и взаимовлияния таких моделей между собой

3.8 Преимущества

Использование концепции «объемного текста» позволяет осуществить качественный переход в процессе переноса текстов с бумажных на электронные носители. Существующее сегодня механическое перенесение бумажных текстов на электронные носители «один к одному», в лучшем случае с добавлением электронного оглавления и некоторых перекрёстных ссылок между отдельными фрагментами текста, не позволяет в полной мере использовать возможности электронных проигрывателей, каковыми являются компьютеры. Применение объемного текста позволит значительно поднять качество работы индивидуума с текстовой информацией.

Развитием технологии «объемного» текста является возможность динамического (автоматического) формирования такого текста, исходя из информационных потребностей читающего, его модели мира и актуальной поисковой активности на данный момент. Одни будут искать в тексте вводимые положения, другие – отличие от общепризнанного или особенности материала, не встречавшиеся в других текстах и т.д. Создание механизма автоматического построения рельефа по исходному плоскому тексту учитывающего модель мира индивидуума и его актуальные информационные потребности, позволит быстро, без потерь и информационного шума находить адекватную задачам индивидуума информацию в море исходных текстов, доступных в первую очередь по электронным каналам связи (интернет, электронные архивы и книги, сайты, блоги, чаты и т.д.). Появится возможность создания персональных электронных агентов, подбирающих в сети и электронных архивах и представляющих в объемном виде информацию для оперативного удовлетворения индивидуальных информационных потребностей различных индивидуумов с постоянно меняющимися информационными запросами. Причём делать это можно будет независимо от деятельности самого индивидуума (параллельно с ней) и кругло-

суточно, а информацию представлять в очень сильно сжатом виде (с учётом динамически меняющейся модели мира индивидуума) и исходя из текущей информационной потребности. После разработки полностью автоматических механизмов формирования объемного текста на одном и том же плоском тексте можно будет динамически формировать множество объемных текстов, отвечающих текущим информационным потребностям каждого конкретного читателя или их группам.

3.9 Аналогия

Ситуация с плоским и объемным текстом постараюсь объяснить на следующем примере-анalogии. Вы приходите на восточный базар и хотите выбрать себе ковёр или ковры. Но оказывается, что на базаре теперь продавцов много, а места мало и все ковры свёрнуты в трубочки и стоят стройными совершенно внешне одинаковыми рядами. Это аналогия текста – сплошная однородная текстура. В каждый момент времени можно развернуть и посмотреть только один ковёр. Это аналогия области нашего внимания. Вы стоите перед перспективой: брать и разворачивать каждый ковёр и смотреть – подходит или нет. Примерно так мы и поступаем при чтении плоского текста. Идём по нему в подряд, пока не выловим интересующие нас моменты и не опустим, с сожалением о потраченном времени и силах, то, что нам известно или не очень то и нужно было.

Возвращаемся на ковровый базар. Понятно, что торговцев такая ситуация не устраивает и они расставляют ковры группами по типам рисунков, качеству изготовления и т.д. В плоском тексте этому соответствует разбивка на главы, разделы и т.д. Но далее-то опять начинается процедура разворачивания каждого ковра, хотя уже в перспективе не всех, а только некоторых.

Поэтому предлагается следующая вещь. От каждого ковра отрезается характерный кусочек и из них составляется мозаика, которая размером с один ковёр и в развёрнутом виде лежит перед глазами. В концепции объемного текста это и есть вершины, т.е. вводимые в текст положения. Если какой-то фрагмент понравился, то переходят к группе ковров, содержащих этот фрагмент и достают мозаику следующего уровня детальности только для этой группы, там кусочки от ковров уже побольше, и представление о ковре можно составить лучше. Ну и так далее. Поскольку ковры в нашем примере виртуальные, то мы от них можем отрезать многократно любые кусочки, они от этого цельными быть не перестанут. В конечном итоге перед вами разворачивают выбранный вами таким способом ковёр, или вы, быстро просмотрев лишь некоторые из мозаик, ушли с базара без ковра, т.к. сегодня нужных вам ковров в наличии не оказалось. Но при этом вы значительно сэкономили время и использовали его, например, для посещения коньячного завода с дегустацией продукции. После этого конечно можно опять вернуться на восточный базар и купить пер-

вый попавшийся ковёр. Но это дело вкуса, ... того коньяка, который вы дегустировали.

Если развернуть все ковры и сложить их в один слой, для чего может не хватить не только площади всего восточного базара, но и площади всего города или даже государства, то мы получим плоский текст. Наложим на все эти развёрнутые ковры одним слоем все развёрнутые мозаики самого нижнего уровня. Затем на них наложим тоже одним слоем все развёрнутые мозаики следующего уровня и так далее. В результате на вершине получившейся пирамиды окажется только одна единственная мозаика самого верхнего уровня, которая размером всего с один ковёр. Получившееся объемное тело и является аналогом объемного текста, но выполненного из другого (текстового) материала.

Т.е. концепция объемного текста предполагает составление из плоского текста мозаик разного уровня детальности и их упорядоченное расположение. Понятно, что мозаика каждого нижележащего уровня становится всё больше и больше, а мозаикой самого нижнего уровня детальности будет сам плоский текст. В данном случае пример расположения космоснимков разного масштаба на Google [6] является более близкой аналогией, чем пример с коврами. Но с коврами пример более понятен с технологической точки зрения, а в Google – с точки зрения технической реализации и потребных для этого ресурсов, в том числе и мощностей проигрывателей. Понятно, что без компьютера текст представить в объемном виде на бумажных носителях будет совершенно нетехнологично. С такой формой представления текста будет очень неудобно работать, т.к. основная работа идёт по вертикали, а не в горизонтальной (вдоль слоя текста) плоскости.

4 Заключение

В приложении приведён образец текста, подготовленного для загрузки в 3D-интерфейс и представленный в технологии HTML. Это текст статьи «Интеллектуальная продукция и инновации в общественной практике», которая в плоском варианте занимает 56 страниц. Для объемного текста 3D-интерфейс является органической (функциональной) необходимостью, а не иллюстрацией, и этот интерфейс ещё предстоит разработать. Естественно, что объемный текст нельзя представить в отрыве от такого проигрывателя, как компьютер. Органическое сочетание информационных возможностей компьютера и концепции объемного текста и позволяет осуществить переход на новый этап представления формализованной (логической) информации, более адекватный современной быстротекущей общественной практике, чем выработанный тысячелетия назад способ передачи (обмена) информацией в виде плоского текста.

Литература

- [1] Ф. Иоганнес. История письма. – М. : Наука, 1979. – 464 с.
- [2] Компания ACD Systems, 2008.
<http://www.ACDSYSTEMS.COM>
- [3] Памятная книга редактора. – М. : Книга, 1988. – 416 с.
- [4] Тупик Н.В. Повышение эффективности обучения // Материалы электронной научно-практической конференции «Эффективность использования новых информационных технологий в учебном процессе» (ЭНИТ-2000). – Ульяновск, 2000. – С. 58–69.
- [5] «Чжуд-ши» – памятник средневековой тибетской культуры. – Новосибирск : Наука, 1988ю – 349 с.
- [6] Картографический ресурс Google Maps, 2008.
<http://maps.google.com>

Concept of volumetric text

N.V. Tupik

The concept of volumetric text is proposed in this article. The volumetric text consists of the text tier set, where the every next tier is more compressed than the previous (have less amount of words). The most important ideas are the top layer of the text. Every thought is expressed by one short affirmative sentence. The text, built in such case, can be regarded as a three-dimensional solid with the fractal structure. The reader can individually and dynamic choose the trajectory how to familiarize himself with the text as well as chose independently the initial point of this trajectory. In that way the reader look at the text so to speak with “open eyes”.

Программный комплекс «СМАЛТ»*

© А.А. Рогов, Г.Б. Гурин, А.А. Котов, Ю.В. Сидоров, Т.Г. Суровцова

Петрозаводский государственный университет
rogov@psu.karelia.ru

Аннотация

В современном языкознании на смену таким традиционным методам получения языковых данных, как интроспекция, сбор текстового материала, эксперимент, опрос, приходит **корпусный метод**. Создание лингвистических корпусов текстов осознается как одна из актуальных задач современного языкознания. Корпусы активно используются в практике составления словарей, в проведении разнообразных исследований языка.

В Петрозаводском государственном университете работы по компьютерной обработке текстов ведутся с 1995 года. Их результатом явилась разработка программного комплекса «Статистические методы анализа литературных текстов» (ПК «СМАЛТ»), имеющего в своей основе базу данных текстов, состоящую из публицистических статей разной тематической направленности из петербургских журналов XIX века в оригинальной орфографии.

1 Введение

В настоящее время на смену таким традиционным методам получения языковых данных, как интроспекция, сбор текстового материала, эксперимент, опрос, приходит **корпусный метод**, создание лингвистических корпусов текстов осознается как одна из актуальных задач современного языкознания [2, 4]. Корпусы активно используются в практике составления словарей, в проведении разнообразных исследований языка. Наиболее известны общепризнанные образцы корпусов: Британский национальный корпус, Мангеймский корпус немецкого языка, Чешский национальный корпус и др. Отечественная лингвистика несколько отстает в этом отношении. Достаточно вспомнить, что до недавнего времени существовал единственный русскоязычный корпус, Упсальский корпус русских текстов, который был создан в 60-е гг. прошлого века вне России и, по оценке многих, во многом устарел. Однако

в последнее время появилось немало интересных проектов такого рода, самый масштабный из них – Национальный корпус русского языка (www.ruscorpora.ru).

В Петрозаводском государственном университете работы по компьютерной обработке текстов ведутся с 1995 года. Их результатом явилась разработка программного комплекса «Статистические методы анализа литературных текстов» (ПК «СМАЛТ»), имеющего в своей основе базу данных текстов, состоящую из публицистических статей разной тематической направленности из петербургских журналов «Время», «Эпоха», «Современник», «Гражданин» «Светоч», «Молва», «Библиотека для чтения», «Заря» XIX века [1–3] в оригинальной орфографии. Проект был поддержан грантами РГНФ № 02-04-12015в, 05-04-12418в, 08-04-12105в (руководитель А. А. Рогов). Адрес в Интернете: <http://smalt.karelia.ru>.

2 Программный комплекс «СМАЛТ»

Программный комплекс «СМАЛТ» предоставляет несколько систем доступа к единой базе данных, хранящей синтаксические и морфологические разборы литературных произведений. Общая схема ПК «СМАЛТ» представлена на рис. 1. Он состоит из базы данных, системы подготовки данных, системы контроля знаний учащихся, системы доступа к БД и информационно-аналитической системы для анализа текстов. Для хранения базы данных используется СУБД Interbase 6.0. В качестве исходного источника данных для клиентского приложения используется текстовый файл в кодировке Unicode, что позволяет избежать проблем, связанных с использованием в отдельных текстах символов, специфичных как для отдельных языков, так и для орфографии разных периодов одного языка.

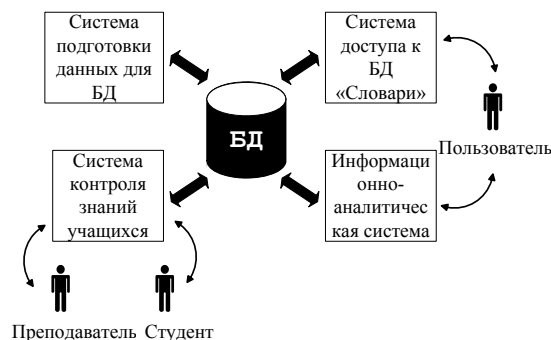


Рис. 1. Схема программного комплекса

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

С помощью системы подготовки данных ПК «СМАЛТ» была проведена лемматизация и морфологическая разметка текстов, запланирована синтаксическая аннотация предикативных клауз. Обработка каждого текста в БД предполагает три стадии: преформатирование, грамматический анализ, синтаксический анализ (рис. 2). На этапе преформатирования выполняется автоматизированное разбиение исходного текста на единицы, среди которых выделяются часть (или раздел), абзац, предложение, слово. Полученное разбиение может быть откорректировано вручную. Важнейшим модулем ПК «СМАЛТ» является морфологически размеченный корпус текстов русской публицистики второй половины XIX века, который может рассматриваться как самостоятельный продукт.



Рис. 2. Система подготовки размеченных текстов

3 Состав и особенности корпуса, его адресаты

Большинство современных русскоязычных корпусов ориентированы на язык XX–XXI веков, тексты предшествующих периодов, в силу трудности их автоматической обработки, включаются в корпус реже. Данный корпус является историческим, поскольку сформирован из оригинальных текстов русской публицистики 60–70-х годов XIX века.

Публицистические тексты обязательно включаются в состав современных лингвистических корпусов по понятным причинам: именно в публицистике в силу ее определенной жанровой свободы и тесной связи с социально-культурной, политической и экономической жизнью социума полнее и ярче отражаются разнообразные языковые изменения, прослеживаются формирующиеся тенденции развития языка. Насколько нам известно, публицистические тексты эпохи второй половины XIX века, представляющие богатейший языковой материал, до сих пор специально не привлекались в качестве особого объекта корпусной презентации.

В корпусе принципиально сохранены исконная графика текстов, а также все особенности дореформенной орфографии. Как известно, орфография той эпохи не была устойчивой, широко была распространена орфографическая и фомоморфологическая вариативность, например: *очень-многіе, само-по-себе, до-сихъ-поръ, на-дняхъ, какъ-будто, ничемъ другъ-къ-другу необязанныхъ, студентскій миръ,*

взмахнутый, въ самомъ-достойнѣйшемъ, самоновѣйшій, низачто, само-малѣйшей, истинно-умные расположонъ, предстоитъ современем, состарѣлась, мущина, выростеть, коммунистѣ, колосальный и проч.

Сохранение этих особенностей в диахроническом корпусе может быть полезно исследователю, описывающему динамику норм правописания [см. 4, 6]. При этом корпус удобен для пользователя, незнакомого с особенностями дореволюционной графики и орфографии: в корпусе реализован поиск слов по современной орфографии, позволяющий отыскивать, например, по лемме *мужчина* все орфографические варианты (*мужчина, мущина, мужщина*).

Потенциально корпус адресован самому широкому кругу пользователей – всем, кто интересуется русским языком и русской литературой 19 столетия: профессиональным лингвистам, в том числе историкам языка, филологам-литературоведам, студентам, преподавателям средней школы и школьникам. Это обусловило некоторые особенности морфологической разметки, подачи материала и системы поиска.

4 Морфологическая разметка корпуса

Общеизвестно, что «представление в корпусе информации о морфологических формах и значениях (часть речи, род, падеж, вид...) является самостоятельной научной проблемой» [6], которая может решаться по-разному [4]. Данный корпус опирается в основном на морфологическую модель, представленную в «Грамматическом словаре русского языка» А. А. Зализняка (М., 1977; 4-е изд. М., 2003).

Однако специфика корпуса и тем более корпуса языка XIX века, ориентированного на широкого пользователя, такова, что в некоторых случаях требовались особые решения. Для сохранения упорядоченности и единообразия разметки, особенно при принятии решений в области частеречной разметки, последовательно применялись рекомендации «Словаря русского языка» (2-е изд. М., 1981–1984) и «Толкового словаря русского языка» С. И. Ожегова и Н. Ю. Шведовой. Несомненные минусы этого решения очевидны, однако принципиальной установкой разработчиков было обеспечение доступности и простоты в использовании корпуса. Все перечисленные критерии учитывались при формировании системы грамматических параметров, с помощью которой описывались словоформы.

В корпусе использована система морфологической разметки, использующая систему традиционных морфологических понятий, причем в двух вариантах – в виде формально-грамматической и морфолого-семантической разметки

Разметка 1 опирается на следующий инвентарь частей речи: *существительное, прилагательное, числительное, местоимение, глагол, причастие,*

деепричастие, наречие, предикатив, союз, предлог, модально-дискурсивное слово или частица, междометие, компонент идиомы, антропоним – и предоставляет возможность поиска по значениям базовых морфологических категорий соответствующих частей речи.

Разметка 2 ориентирована на школьную традицию и включает дополнительные грамматические параметры: лексико-грамматические разряды существительных, прилагательных, числительных, местоимений, типы склонения и спряжения. Она предназначена для использования в образовательных целях и может рассматриваться как параллельный обучающий корпус, подобный тому, что реализован в рамках Национального корпуса русского языка.

Словарь, наполнение которого происходит в процессе разбора, существенно ускоряет проведение морфологического анализа, а также позволяет рассматривать разные виды омонимии, возникающие в тексте. Формирование собственного словаря позволяет в перспективе работать с текстами на разных языках

Отличие «лексико-грамматического» и «формально-грамматического разбора» можно пояснить на примере. В словах «первый ученик», при лексико-грамматическом анализе слово «первый» будет кодировано как прилагательное (в смысле лучший), а при формально-грамматическом разборе как числительное. Формально-грамматический разбор обладает меньшей вариативностью, но зато меньшей субъективностью. Заметим, что взаимнооднозначное соответствие между разборами удалось установить только в 90% случаев. Остальные 10% не удалось формализовать.

В настоящий момент в базе данных словаря с морфолого-семантическим разбором находится более 40 000 лемм из текстов общим объемом более 140 000 словоформ. В словаре с формально-грамматическим разбором находится более 26 000 лемм из текстов общим объемом около 100 000 словоформ.

5 Подача материала и системы поиска

Реализация модулей доступа к БД системы производится с использованием языка PHP 4. Общая схема доступа представлена на рис. 3. Пунктирными линиями обозначены модули, разработка которых незакончена. Для обеспечения поддержки символов дореволюционного русского алфавита все тексты произведений, словоформы хранятся в кодировке Unicode. Для отображения используется шрифт Palatino Linotype.

Для удобства работы и полноты информации корпус реализован в виде словаря с алфавитной системой построения. Реализовано несколько систем поиска по различным критериям: 1) по словам в старой орфографии; 2) по словам в современной орфографии; 3) по грамматическим признакам (с возможностью сохранения заданных параметров). Кроме того, возможен поиск через *Сводный список*

текстоформ: 1) Алфавитный, 2) Алфавитно-частотный (с указанием частоты встречаемости, отсортированные по убыванию).



Рис. 3. Функции доступа к базе данных размеченных текстов «СМАЛТ»

При использовании любого поиска пользователь получает информацию в следующей последовательности: 1) морфологический разбор (или множество морфологических разборов); 2) сведения об авторе и произведении, сведения о контексте с точностью до номера главы, параграфа и предложения; 3) контекст в пределах предложения; 4) расширенный контекст – полный оригинальный текст.

6 Дополнительные возможности ПК «СМАЛТ»

Предусмотрена возможность расширения корпуса электронных документов за счет данных, подготовленных в других информационных системах. Проведена работа по синхронизации данных ПК «СМАЛТ», и конкордансов В. И. Даля (http://elibrary.karelia.ru/dahl/user_new/index.php). Одна база данных реализована на платформе СУБД Interbase 6.0, другая – на платформе СУБД Oracle. Создано универсальное приложение, которое проводит синхронизацию данных.

Для пользователей, которые активно работают с ПК «СМАЛТ», в целях экономии денежных средств на Интернет-трафик, а также для пользователей, испытывающих проблемы с доступом к сети Интернет, была разработана локальная версия ПК «СМАЛТ», которую можно установить на локальные (персональные) рабочие станции пользователей. В локальную версию включена возможность работы с корпусом публицистических произведений. Для этого достаточно около 500 мегабайт свободного дискового пространства на компьютере и устройство для чтения компакт-дисков. Сам процесс установки достаточно прост даже для неспециалистов - инструкция по установке загружается автоматически и достаточно просто следовать пошаговым инструкциям, указанным в ней. Во время инсталляции на рабочую станцию устанавливается web-сервер Apache, серверный язык обработки скриптов PHP и система управления базами данных InterBase. Все вышеперечисленное программное обеспечение является свободно распространяемым и не требует покупки дополнительных лицензий.

Таким образом, на рабочие станции пользователей устанавливается программное обеспечение ПК «СМАЛТ», которое позволяет работать с корпусом произведений. В дальнейшем, для увеличения количества разобранных произведений пользователю достаточно будет обновить базу данных. Кроме этого в локальной версии ПК «СМАЛТ» предусмотрена функция выбора произведений, по которым строится словарь, что позволяет сделать его более сбалансированным. Адрес в Интернете: <http://smalt.karelia.ru>.

7 Описание информационно-аналитической системы для анализа текстов

Модуль статистической обработки, входящий в ПК «СМАЛТ», представляет собой информационно-аналитическую систему, которая включает в себя ряд общеизвестных методов, позволяющих проводить классификацию и кластеризацию текстов и групп текстов на основании определяемых лингвостатистических параметров, вычисляемых по тексту, например, распределению длины предложения или употреблению частей речи на заданных позициях предложения. Могут быть рассчитаны и количественные характеристики для произведения или групп произведений, например, количество употребления части речи, средняя длина предложения и т.п. Лингвостатистические параметры определяются на основании морфологических и синтаксических разборов литературных произведений, получаемых с использованием корпуса произведений, поддерживаемого ПК «СМАЛТ». Модульная структура и использование архитектуры «клиент-сервер», позволяет обеспечить расширяемость, гибкое изменение конфигурации, отсутствие необходимости в установке.

Для проведения кластеризации и классификации текстов в информационно-аналитической системе предлагаются следующие группы методов:

- разбиение анализируемых текстов на группы с использованием методов кластерного анализа;
- разбиение анализируемых текстов на группы с использованием методов дискриминантного анализа;
- метод «сильного графа» для оценки парной связи грамматических и синтаксических классов.

Кроме этого, есть группа методов, осуществляющая проверку статистических гипотез об однородности распределения частотных характеристик текстов.

Для облегчения работы пользователя выполняемое им исследование хранится в виде проекта, включающего тексты, которые он хочет проанализировать, и использованные им методы. Для анализа доступны разобранные тексты, входящие в корпус произведений ПК «СМАЛТ», в основном это публицистические статьи XIX века. Могут быть использованы и другие тексты, но для этого они должны быть предварительно подготовлены с ис-

пользованием ПК «СМАЛТ», а именно получены их морфологический и синтаксический разборы. База знаний информационно-аналитической системы подскажет пользователю методы анализа данных, которыми он может воспользоваться. Предложенные методы классифицируются по категориям, как наиболее предпочтительные для выбранных текстов, информативные или часто используемые. Если необходимо, то для выбранных методов указываются начальные условия.

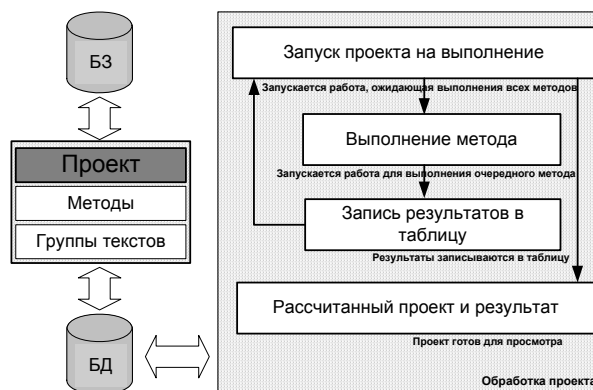


Рис. 4. Интерфейс пользователя

Число рассматриваемых методов можно расширять за счет модульной структуры информационно-аналитической системы. В процессе работы накапливается информация, позволяющая определить методы, которые обеспечивают более эффективную работу, то есть провести некоторое ранжирование методов. На основании этого проводится комплексное рассмотрение результатов тестов, с помощью введения оценки, учитывающей каждый из полученных результатов. Так каждый из тестов может давать положительное или отрицательное заключение о близости групп произведений, а комплексная оценка может учитывать все результаты и выдавать единый результат, причем вклад каждого теста берется в соответствии с введенным ранее ранжированием методов.

После выбора методов пользователь запускает проект на выполнение. Так как часть методов предполагает достаточно большой объем вычислений, то запускается работа, которая обеспечивает полную обработку проекта и не зависит от действий, выполняемых пользователем. Пользователь может закрыть браузер или наблюдать за тем, как происходят вычисления, см. рисунок 4.

После окончания расчетов можно проводить анализ полученных данных. Если обнаружена близость групп текстов по рассматриваемым признакам, то выводится соответствующее сообщение. На основании результатов всех методов, с учетом их ранжирования и используя «метод комитетов», может быть сделан вывод о наличии или отсутствии близости групп текстов.

В процессе работы пользователь может изменить проект, а именно добавить новые произведения и методы, в этом случае потребуется повторное проведение вычислений. Если получен интересный результат, например, обнаружено, что употребление глаголов определенным автором сильно отличается от остальных, то проект может быть продублирован исследователем, чтобы сохранить текущее состояние и начать работу с того же места, но в новом проекте, добавив для рассмотрения новые лингвостатистические параметры. Доступ к проекту может быть разделен с другими зарегистрированными участниками информационной системы.

8 Результаты апробации

С использованием ПК «СМАЛТ» был проведен ряд исследований по выявлению групп текстов, принадлежащих перу одного автора, а также их количественных характеристик [3, 5, 6, 8].

Проверялась методика Гейра Хетсо [9] для установления авторства. Было показано, что результаты исследования не зависят от того, какой текст рассматривается, в авторской орфографии и пунктуации или отредактированный для полного собрания сочинений. Проверена гипотеза о нормальности исходных выборок, тем самым правомерность использования статистических критериев в исследовании Г. Хетсо. Выявлена неустойчивость параметров «Средняя длина слова в буквах» и «Индекс разнообразия лексики» на разных объемах выборок.

Рассмотрен параметр «Распределение частей речи на первых трех и последних трех позициях в предложении». Показано, что увеличение числа формально-грамматических признаков не приводит к существенному изменению результатов классификации.

Метод оценки парной связи для грамматических классов (метод «сильного графа» [1]) не дал возможности провести атрибуцию авторства обработанных статей, но была выявлена возможность анализа жанров.

Был проведен анализ публицистических произведений, с использованием синтаксического разбора, на основе методов многомерного статистического анализа, а также метода «сильного графа». Группы объектов, которые получились в результате, не дали четкого разделения по авторам произведений, причем группировка менялась в зависимости от выбранной меры и метода построения кластеров. При этом происходило разделение текстов маленького и большого размера, предположительно из-за того, что маленькие тексты имеют более бедный набор рассматриваемых параметров.

9 Заключение

С помощью программного комплекса создан и поддерживается размеченный корпус, основанный на публицистических статьях XIX века. Для текстов, включаемых в корпус, проводятся в автоматизированном режиме морфологический и синтаксические разборы. Доступ к корпусу поддерживается с помощью словарей, содержащих контексты. Испол-

зованные технологии могут быть применены для создания, хранения и доступа к мультязычному корпусу, состоящего из произвольных текстов.

Входящая в ПК «СМАЛТ» информационно-аналитическая система для анализа текстов включает методы, позволяющие выделять группы текстов близкие по рассматриваемому набору лингвостатистических параметров. С одной стороны корпус используется как единый тестовый материал, дающий возможность исследовать сами методы на надежность и устойчивость, сравнивать их, создавать рекомендации по применению.

А с другой стороны он сам является объектом для проведения исследований. Так с помощью ПК «СМАЛТ», совместно со специалистами-филологами, проводился поиск лингвостатистических параметров, определяющих стиль автора.

Литература

- [1] Бородин Л. И., Милов Л. В., Морозова Л. Е. К вопросу о формальном анализе авторских особенностей стиля в произведениях // Математические методы в историко-экономических и историко-культурных исследованиях. – М., 1977, – С. 298–326.
- [2] Волков С. Св. Корпус текстов и исторический словарь / С. Св. Волков, В. П. Захаров // Русский язык конца XIX века: Проблемы изучения и лексикографического описания. – СПб., 2004. – С. 38–43.
- [3] Захаров В. Н. Программная система поддержки атрибуции текстов статей Ф. М. Достоевского / В. Н. Захаров, А. А. Леонтьев, А. А. Рогов, Ю. В. Сидоров // Труды Петрозаводского государственного университета: Сер. Прикладная математика и информатика. Вып. 9. – Петрозаводск, 2000.
- [4] Корпус текстов как особый тип лингвистической электронной библиотеки / С. Св. Волков, А. С. Герд, О. Н. Гринбаум и др. // Словарь русского языка XIX века. Проблемы. Исследования. Перспективы. – СПб., 2003. – С. 92–108.
- [5] Рогов А.А., Сидоров Ю.В., Солопова А.И., Суворцова Т.Г. Информационно-аналитическая система «СМАЛТ» // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог 2007" (Беласово, 30 мая – 3 июня 2007 г.) / под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М. : Издательский центр РГГУ, 2007. – С. 470–474. – на рус. яз.
- [6] Рогов А.А., Сидоров Ю.В., Суворцова Т.Г. Математические методы атрибуции литературных текстов небольшого объема // Материалы XIII Всероссийской конференции «Математические методы в распознавании образов». – М. : МАКС Пресс, 2007. – С. 525–528. – на рус. яз.

- [7] Стенограмма обсуждения «Проекта Словаря русского языка XIX века» // Словарь русского языка XIX века. Проблемы. Исследования. Перспективы. – СПб., 2003. – С. 109–154.
- [8] Суровцова Т. Г. Экспертная система для выявления скрытых количественных характеристик литературных произведений // Информационные технологии моделирования и управления. – 2007, №6(40). – С. 650–655. – на рус. яз.
- [9] Хетсо Г. Принадлежность Достоевскому: К вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах *Время* и *Эпоха* / Г. Хетсо – Oslo: Solum Forlag A.S., 1986. – 82 с.

Program Complex “SMALT”

A.A. Rogov, G.B. Gurin, A.A. Kotov,
Yu.V. Sidorov, T.G. Surovcova

In modern linguistics traditional methods of language data collection such as introspection, text data collection, experiment, and questioning are being replaced now by the corpus method. The creation of linguistic corpora of texts is one of the most important tasks of modern linguistics. Corpora are actively used in lexicography and in different linguistic research projects.

Researches of Petrozavodsk state university have been working with electronic text processing since 1995. The result of this work is the program complex “Statistical Methods of Literary Texts Analysis” (PC “SMALT”). It is based on a text database, which is composed of publicistic articles on different topics from St.Petersburg magazines of the XIX century in original spelling.

* Проект был поддержан грантами РГНФ № 02-04-12015в, 05-04-12418в, 08-04-12105в (руководитель А. А. Рогов). Адрес в Интернете: <http://smalt.karelia.ru>

Создание легко обновляемых текстовых индексов

© А.Б. Веретенников

Уральский государственный университет им. А.М. Горького
abvmt2@mail.ru, abvmt@e1.ru

Аннотация

Этот документ описывает структуру данных, предназначенную для создания текстовых индексов, которые легко обновлять. Также дано описание системы, в которой реализованы алгоритмы работы с данной структурой данных.

1 Введение

Для создания текстовых индексов для электронных коллекций, состоящих из естественных текстов, например художественная литература, сайты в сети Internet, в основном используются инвертированные файлы [6] и их аналоги. К сожалению, инвертированные файлы сложны в обновлении. Чтобы добавить новые данные в инвертированный файл часто требуется переписать весь индекс.

В текстовые коллекции часто добавляются новые данные, и требуется быстро обновлять индекс. Автором была разработана новая структура данных – CLB индекс, для создания индекса, в который можно легко добавлять новые данные.

Начальная идея CLB индекса описана в [1]. В этой работе была описана структура данных CLB дерево, разработанная автором данной статьи.

CLB дерево обладает рядом недостатков, которые были устранены с помощью новых алгоритмов, которые описаны в данном тексте.

2 Поддержка морфологии

CLB индекс предполагает возможность поиска в текстах с поддержкой морфологии языка. Более того, поддержка морфологии языка позволяет сделать ряд оптимизаций.

3 Идея индекса

CLB дерево представляет собой В-дерево [3, 4, 5], содержащее в себе слова проиндексированных текстов. Вместе с каждым словом в В-дереве сохраняется ссылка на список связанных блоков, в кото-

ром сохраняется информация обо всех вхождениях данного слова в текстах.

Например, для каждого вхождения каждого слова сохраняется номер или идентификатор соответствующего документа и позиция слова в этом документе.

Т. е. у нас есть файл, состоящий из блоков заданного размера. Несколько блоков объединяется в список: в каждом блоке из данного списка сохраняется ссылка на следующий блок.

При создании индекса, когда слово встречается первый раз, мы помещаем его в В-дерево и создаем список, состоящий из одного блока, в который записываем информацию о вхождении данного слова в тексте.

Если слово встречается повторно, информация о следующем вхождении добавляется в данный блок. Если свободное место в блоке заканчивается, мы создаем новый пустой блок (увеличивая файл в размере) и прописываем на него ссылку в текущем блоке, и далее информация добавляется в новый блок. Таким образом, мы достигаем возможности легкого обновления индекса.

В [1] были даны теоретические оценки для количества дисковых операций при создании индекса и поиске с его помощью.

Однако, данный подход обладает рядом недостатков. В частности, блоки располагаются не последовательно, что замедляет скорость их чтения при поиске. Данная проблема была решена за счет разработки более эффективных алгоритмов, описание которых дано в разделе 4.1.

Кроме того, некоторые слова редко встречаются в текстах. Соответственно, размер информации о вхождениях для таких слов меньше чем размер одного блока. Следовательно, при описанной организации индекса у нас возникают частично заполненные блоки.

Например, пусть некоторое слово встречается всего один раз. Информация об одном вхождении занимает несколько байт, например, номер документа 4 байта и позиция слова в документе – еще 4 байта, итого 8. При определенных оптимизациях информация об одном вхождении занимает 1–3 байта. Для хранения информации о вхождении данного слова выделяется блок в индексе, размер блока обычно 4–16 кб. Т. е. место на диске расходуется неэффективно.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Данная проблема также была решена, описание дано в разделе 4.3.

4 Построение индекса

4.1 Организация данных для эффективного чтения

Зафиксируем некоторое число c и соответствующее ему число $m = 2^c$. Цель алгоритма заключается в том, чтобы организовать хранение данных таким образом, чтобы список блоков был разбит на группы, и все блоки, входящие в заданную группу были расположены в файле последовательно.

Например, если у нас есть 25 блоков и $m = 8$. Разбиваем 25 блоков на группы следующих размеров 8, 8, 8, 1.

Т. е. весь список блоков разделяется на группы по m блоков, за исключением последней группы, в которой может быть меньше чем m блоков, т. к. длина списка может не делиться нацело на m .

При проведении экспериментов размеры блоков брались 4–16 килобайт. Число m было выбрано таким образом, чтобы суммарная длина группы последовательно располагающихся блоков была не меньше 8 мегабайт.

В результате скорость считывания данных из списка блоков практически совпадает со скоростью последовательного чтения данных.

4.2 Алгоритм

Далее кратко описан алгоритм записи списка блоков подобным образом.

Вначале у нас есть только один блок. Далее, пусть у нас есть k заполненных подряд расположенных блоков $B_1, \dots, B_k$, в частности последний блок также заполнен, и нам требуется взять где-то новый блок.

Рассмотрим случай, когда $k = 2^x$, где x – некоторое целое число, меньшее c (число c было введено в начале раздела 4.1).

Мы ищем $2k$ подряд располагающихся блоков $N_1, \dots, N_{2k}$. Затем копируем информацию из старых k блоков в первую половину новых блоков, т. е. в блоки $N_1, \dots, N_k$ соответственно. Старые блоки $B_1, \dots, B_k$ помечаются как свободные. Запись далее осуществляется в блок N_{k+1} . Остальные блоки $N_{k+2}, \dots, N_{2k}$, помечаются как зарезервированные.

Как мы ищем $2k$ подряд располагающихся блоков? Поиск осуществляется либо среди блоков, обозначенных как свободные, или мы добавляем блоки в конец файла, увеличивая его размер.

Если $x = c$, то это означает, что у нас есть группа подряд располагающихся блоков длины m . В этом случае мы начинаем формировать новую группу блоков.

В остальных случаях, т. е. если k не является степенью 2, мы осуществляем запись в блоки, которые были ранее зарезервированы описанным выше способом.

4.3 Эффективное заполнение блоков

Для эффективного заполнения блоков введем новый тип блоков. Пусть некоторые блоки разделяются на части равного размера. При этом, некоторым словам может соответствовать не список блоков, в крайнем случае состоящий из одного блока, а только часть блока. Если блок делится на части, то в нем же сохраняется информация о том, какие из его частей заполнены, а какие свободны. Разработан эффективный алгоритм управления такими блоками.

Как только суммарный объем информации для заданного слова становится достаточно большим, информация перемещается в новый пустой блок обычного типа.

5 Кэширование

5.1 Кэширование для слов, входящий в словарь морфологического анализатора

Очевидно, что если мы будем каждый раз при добавлении информации о вхождении слова записывать данные в файл, мы будем создавать индекс очень долго.

Основная оптимизация осуществляется за счет использования морфологии языка. Автором использовался морфологический анализатор [2].

Слова, входящие в состав данного анализатора составляют 80-90% от общего количества слов в обычных тестах (например, художественная литература, новости). В словарь анализатора входит 3,5 млн словоформ русского языка, образованных от 205 тыс. базовых форм.

Словарь анализатора позволяет по словоформе получить набор ее базовых форм. Для большинства словоформ существует только одна базовая форма, однако для многих словоформ существует несколько базовых форм. В данной работе автор отождествляет термины «слово» и «словоформа». Где это необходимо используется термин «базовая форма слова».

При сохранении данных о вхождении слова в тексте, слово сначала приводится в своей базовой форме. Если базовых форм несколько, то информация добавляется для всех базовых форм.

Соответственно для каждой базовой формы слова существует список блоков, для хранения информации о вхождении данной базовой формы в текстах.

При определенном размере блока мы можем хранить в оперативной памяти последний блок для каждого из этих списков блоков. Этот блок сохраняется на диск, только когда он становится полностью заполненным.

Данный подход можно применять только для слов, входящих в словарь морфологического анализатора, т. к. мы знаем количество базовых форм и можем заранее создать кэш требуемого размера.

Слов, не входящих в словарь анализатора может быть гораздо больше, например, в одной из используемых автором текстовых тестовых коллекций бы-

ло 65 млн различных слов, не входящих в словарь морфологического анализатора. О том, как обрабатывать такие слова, описано в разделах 5.3, 5.4.

5.2 Оптимизация кэширования

Метод, описанный в разделе 5.1, задает ограничения на размеры блоков.

Например, если у нас есть 200 000 базовых форм слов и размер блока 4 кб., то общий размер кэша равен 800 мб. Соответственно мы не можем значительно увеличивать размер блока.

Чтобы снять эти ограничения все базовые формы слов разделяются на группы. Для каждой группы базовых форм слов создается временный файл.

Вначале при чтении файлов запись информации о вхождении слов осуществляется во временные файлы. Затем независимо обрабатывается каждый временный файл. При этом требуется организовывать кэш только для базовых форм слов, входящих в соответствующую группу.

Например, пусть у нас есть 200 000 базовых форм слов. Разделим этот набор на группы из 5000 базовых форм слов в каждой.

В результате, при обработке конкретной группы требуется сохранять в памяти только 5000 блоков, что требует значительно меньше памяти и позволяет увеличить размер блока и соответственно уменьшить количество операций записи.

Автор обычно использует блоки, размер которых равен 16 кб. В результате проведенных испытаний, не выявилось значительного повышения производительности при увеличении размера блока до 32–64 кб. Следует отметить, что имеет смысл выбирать размер блока кратным размеру кластера файловой системы.

5.3 Кэширование для слов, не входящих в словарь морфологического анализатора

Подход, описанный в разделе 5.1, мы не можем использовать для слов, не входящих в словарь анализатора. В данном случае можно реализовать некие стандартные стратегии кэширования, например, при необходимости добавления нового блока в кэш выгружать из кэша блок, запись в который наиболее давно не производилась.

5.4 Оптимизированная обработка слов, не входящих в словарь морфологического анализатора

Автором также был реализован подход описанный далее. Слова, не входящие в состав морфологического анализатора, как правило, встречаются очень редко. Соответственно, суммарный объем информации об их вхождениях достаточно мал.

Предлагается сохранять информацию о вхождениях для нескольких слов в одном списке блоков. При этом, при записи информации о вхождении, вместе с ней сохраняется тег, определяющий, к какому слову она относится.

Данный метод значительно сокращает количество операций записи при создании индекса, при этом эффективность поиска снижается незначительно.

Например, пусть в одном списке блоков сохраняется информация о вхождениях слов: *aaa*, *bbb*, *www*. Присвоим этим словам номера 1, 2, 3. Если мы ищем слово *aaa*, то мы считываем весь этот список блоков. При чтении из списка блоков информации о каждом вхождении считываем также и тег, соответствующий данному вхождению. Рассматриваем только те записи, у которых тег равен 1.

При создании индекса контролируется, сколько раз уже встретилось слово. Если оказывается, что слово встречается слишком часто, то для него создается новый список блоков, в который переносится информация о вхождениях данного слова, далее в данном списке блоков будет сохранена информация только об этом слове.

Автором был проведен эксперимент. Было зафиксировано число 4096, и для слов, число вхождений которых превышало данное число, создавались отдельные списки блоков.

Для тех слов, число вхождений которых меньше или равно 4096, в один список блоков помещалась информация сразу о 32 различных словах.

Обрабатывалось 35 гб. текстов, содержащих 69 млн. различных слов.

В результате оказалось, что максимальное количество записей в списке блоков, для списков, содержащих несколько слов, было 19 431.

Данное число показывает, что эффективность не пострадала, т. к. объем указанных списков блоков не превышает нескольких десятков килобайт (т. к. одна запись занимает примерно 2–4 байта).

Слов, которые встречаются больше чем 4096 раз, оказалось всего 19 593.

7 Обработка стоп слов

Как правило, часто встречающиеся слова, например предлоги, не добавляются в индекс, т. к. считается, что они не несут в себе нового смысла. Подобные слова называются стоп словами и составляют до 30–40% текста. Кроме того, отказ от индексирования таких слов позволяет значительно увеличить производительность.

Примеры стоп слов

"и", "в", "он", "не", "я", "что", "на"

Частота встречаемости данных слов в типичных текстах

3.025%, 2.298%, 1.767%, 1.609%, 1.432%, 1.284%

Если все-таки потребуется искать фразу, содержащую слово "и", то мы столкнемся с большими трудностями. Слово "и" встречается в 3% случаях, соответственно объем данных, соответствующих слову "и" очень большой.

Для ускорения подобного поиска реализован следующий опциональный вариант: вместе с вхождением каждого слова сохранять также информацию о местонахождении рядом с ним стоп слов.

Это позволяет осуществлять эффективный поиск точный фраз, содержащих стоп слова.

8 Поиск

За счет алгоритма создания индекса, описанного в разделе 4, по своим возможностям и эффективности CLB индекс не уступает инвертированным файлам.

Реализованы различные виды поиска, в частности поиск слов с учетом расстояния между ними и точный поиск фраз. Поиск может быть осуществлен с учетом или без учета порядка слов.

Были проведены эксперименты, показывающие, что скорость чтения информации о вхождениях одного слова из индекса, совпадает со скоростью линейного чтения с диска.

9 Эксперименты по созданию индекса

9.1 Эксперимент 1

Обрабатывались файлы, сжатые в архивы формата RAR. В основном файлы имели формат txt, html.

Суммарный размер файлов 16,3 гб (сжатый размер). Суммарный размер файлов 42,5 гб (несжатый размер).

Суммарный размер текста 36,5 гб.

Всего файлов 300 тысяч.

Доля словоформ, входящих в словарь морфологического анализатора 87%.

Различных словоформ, не входящих в словарь морфологического анализатора – 65 млн.

Время создания индекса 7 часов, из них 4 часа – чтение файлов, 3 часа – создание индекса.

Объем индекса 34 гб.

Эксперимент проводился на следующей конфигурации:

Процессор: Intel Core 2 Duo E6700, 2.66 GHz, кэш: L1 Data – 2 x 32 кб, L1 inst. 2 x 32 кб, L2 – 4096 кб.

Оперативная память: 4 гб, DDR2 800.

Жесткий диск: Seagate Barracuda 7200.10, 7200 RPM, кэш 16 мб., объем 500 гб.

FSB 1066 MHz.

9.2 Эксперимент 2

Обрабатывались не сжатые файлы форматов txt, html.

Суммарный размер файлов 75 гб.

Суммарный размер текста 70 гб.

Всего файлов 600 тысяч.

Доля словоформ, входящих в словарь морфологического анализатора 93%.

Время создания индекса 10 часов, из них 4 часа – чтение файлов, 6 часов – создание индекса. Объем индекса 50 гб.

Эксперимент проводился на следующей конфигурации:

Процессор: Intel Pentium 4, 3.0 GHz, кэш: L1 Data – 16 кб, L1 trace – 12 Kuops, L2 – 2048 кб.

Оперативная память: 2048 гб, DDR2 533.

Жесткий диск: Seagate Barracuda 7200.8, 7200 RPM, кэш 8 мб., объем 200 гб.

FSB: 800 MHz.

Последний эксперимент был повторен, при этом запись индекса была осуществлена в 2 приема. Вначале был создан индекс на основании 40 гб текста, затем в него были добавлены остальные данные. Суммарное время в этом случае составило 12 часов.

Оба эксперимента были осуществлены без отбрасывания стоп слов.

10 Сравнение с инвертированными файлами

10.1 Описание тестовых данных

Было произведено сравнение по скорости создания индекса. Обрабатывались текстовые документы, общий объем 35,2 гб, 191 074 файла. Все файлы были в кодировке Windows-1251 (CP1251). Язык документов – русский. Все файлы представляли собой обычный текст.

10.2 Описание конфигурации оборудования

Эксперименты проводились на следующей конфигурации:

Процессор: Intel Core 2 Duo E6700, 2.66 GHz, кэш: L1 Data – 2 x 32 кб, L1 inst. 2 x 32 кб, L2 – 4096 кб.

Оперативная память: 4 гб, DDR2 800.

Жесткий диск: Seagate Barracuda 7200.10, 7200 RPM, кэш 16 мб., объем 750 гб.

FSB 1066 MHz.

10.3 Создание индекса

Создание инвертированного файла: время 9 часов, размер 40 гб.

Создание CLB индекса: время 3 часа, 32 мин., размер 24 гб.

10.4 Добавление в индекс одного файла среднего размера

Был проведен замер скорости добавления в индекс одного документа, размер документа 1,2 мб.

Время добавления одного документа 1,2 мб. для CLB индекса: 9 мин.

Время добавления одного документа 1,2 мб. в инвертированный файл: 57 мин.

10.5 Добавление в индекс одного файла малого размера

Был проведен замер скорости добавления в индекс одного документа, размер документа 534 байта.

Время добавления одного документа размером 534 байта для CLB индекса: 22 с.

Время добавления одного документа размером 534 байта в инвертированный файл: 57 мин (т. е. такое же, как при размере файла 1,2 мб).

10.6 Поиск

Время поиска в инвертированном файле и CLB-индексе практически совпадают.

10.7 Выводы

Проведенные эксперименты показывают высокую эффективность CLB индекса при добавлении в него данных небольшого размера.

11 Сравнение с существующими разработками

11.1 Описание тестовых данных

Было произведено сравнение по скорости создания индекса и размеру индекса. Использовались те же документы, которые указаны в разделе 10.

11.2 Описание конфигурации оборудования

Эксперименты проводились на следующей конфигурации:

Процессор: Intel Pentium 4, 3.0 GHz, кэш: L1 Data – 16 кб, L1 trace – 12 Kuops, L2 - 2048 кб.
Оперативная память: 4 гб, DDR2 533.
Жесткий диск: Seagate Barracuda 7200.8, 7200 RPM, кэш 8 мб., объем 200 гб.
FSB: 800 MHz.

11.3 SearchInform Desktop

<http://www.searchinform.com>

Размер индекса 16,15 гб.
Время создания 9 часов.

11.4 Архивариус 3000

<http://www.likasoft.com/>

Размер индекса 24,83 гб.
Время создания 6 часов 46 мин.

11.5 Создание CLB индекса

Размер индекса 26,2 гб.
Время создания 5 часов 49 мин.

11.6 Выводы

Эксперименты показывают высокую скорость создания CLB индекса.

12 Библиотека для создания индекса

Автором разработана библиотека для создания индексов и поиска в текстах, в которой реализована описанная структура данных и алгоритмы.

Библиотека может индексировать файлы в различных форматах, например RTF, PDF, CHM, HTML, DJVU и кодировках, например UNICODE, UTF8, CP1251, ASCII, KOI8.

Поддерживается обработка архивов форматов ZIP, CAB, RAR, 7Z, ARJ, TAR, и др.

Библиотека реализована в виде COM сервера для операционных систем Windows.

Состав библиотеки

1. Ядро, осуществляет создание индекса и поиск.
2. Модуль поддержки морфологии.
3. Модуль распознавания кодировки. При распознавании кодировки также учитывается морфология.
4. Модуль поддержки форматов файлов. Поддержка форматов файлов и архивов реализована с помощью подключаемых дополнительных модулей, которые могут быть реализованы в виде динамических библиотек или написаны на Java. Модуль поддержки форматов файлов реализован в виде отдельного процесса для повышения надежности системы.
5. Модуль атрибутов документов, для сохранения описания документов.
6. Модуль репозитория, для сохранения текстов документов. Создается для того, чтобы при поиске можно было быстро получать фрагмент текста, содержащий найденную фразу. Тексты в репозитории могут сохраняться с использованием различных алгоритмов сжатия.
7. Модуль COM осуществляет доступ к остальным модулям извне с помощью COM, что позволяет использовать библиотеку в различных языках программирования.

Реализованные алгоритмы достаточно нетребовательные к ресурсам компьютера. Для создания индекса достаточно иметь 300–400 мегабайт свободной оперативной памяти.

Автором проводились эксперименты по созданию индексов на машине с оперативной памятью размером 512 мб.

Следует отметить, что эффективность описанных в данном документе алгоритмов значительно возрастет с применением дисков SSD (Solid-state drive), за счет более быстрого чтения блоков малого размера. При этом эффективность таких структур данных как инвертированные файлы возрастет менее, т. к. для добавления в инвертированный файл информации его все равно придется практически переписать целиком.

Литература

- [1] Веретенников А. Б., Лукач Ю. С. Еще один способ индексации больших массивов текстов // Известия Уральского государственного университета. Сер. «Компьютерные науки». – 2006, №43. – С. 103–122.
- [2] Лукач Ю. С. Быстрый морфологический анализ флективных языков // Междунар. алгебраическая конф. : К 100-летию со дня рождения П. Г. Конторовича и 70-летию Л. Н. Шеврина : Тез. докл. – Екатеринбург : Изд-во Урал. ун-та, 2005. – С. 182–183.
- [3] Bayer, R., McCreight, E. Organization and maintenance of large ordered indexes. Acta Informatica 1, 3 (1972), 173-189.
- [4] Ferragina, P., Grossi, R. The string B-tree: a new data structure for string search in external memory and its applications. Journal of the ACM, 46, 2 (1999), 236-280.
- [5] Ferragina P., Grossi R. An experimental study of SB-trees. 7th ACM-SIAM symposium on Discrete Algorithms, 1996.
- [6] Prywes, N. S., Gray, H. J. The organization of a Multilist-type associative memory. IEEE Trans. on Communication and Electronics, 68 (1963), 488-492.

The Effective Creating of Easy Updatable Text Indexes

A.B. Veretennikov

Author introduce a new data structure for fast indexing of large volumes of texts and a software system based on this structure. The latter is more effective for inserting data into indexes than other known data structures, and is comparable with them in the data retrieval.

Автоматическое формирование рубрикатора полнотекстовых документов

© Ольга Вадимовна Пескова

МГТУ им. Н. Э. Баумана
opeskova@mail.ru

Аннотация

Рассмотрены вопросы построения информационно-поисковой системы на основе механизма автоматической классификации полнотекстовых документов. Изложены подходы и алгоритмы формирования образов документов, их кластеризации и оценки полученного разбиения данных. Предложен метод автоматического формирования рубрикатора коллекции документов в виде, унаследованном от традиционного библиотечного предметного рубрикатора. Проверена работоспособность метода на небольших коллекциях, содержащих русскоязычные документы различного размера и содержания.

1 Введение

На протяжении последних десятилетий во всём мире наблюдается стремительный рост количества и объёмов коллекций полнотекстовых документов, т. е. множеств документов, содержащих тексты на естественном языке, и доступных через средства телекоммуникации для поиска и доставки пользователю. Примерами могут служить фонды электронных библиотек, электронные архивы журнальных статей, различные собрания научно-технических материалов в локальных или глобальных сетях и другие. В связи с этим одной из важнейших задач является создание программных средств, предоставляющих пользователю эффективные механизмы поиска документов.

Традиционными механизмами поиска в коллекциях полнотекстовых документов, являются: *поиск по ключевым словам* (а) и *классификационный поиск* (б). Поиск по ключевым словам из-за кажущейся простоты использования применяется в большинстве поисковых систем. Однако его эффективность существенно зависит от удачного описания пользователем своей информационной потребности в

форме запроса на естественном языке. Следовательно, пользователь, малоознакомый с искомой предметной областью или малоопытный в вопросах использования поисковых машин, не находит требуемых документов. К такому же результату часто приводит необходимость отбора пользователем документов среди огромных списков, найденных поисковой машиной документов.

Классификационный поиск самостоятельно или в его сочетании с поиском по ключевым словам благодаря интуитивно понятному интерфейсу навигации позволяет легко формулировать и уточнять информационные потребности, что повышает эффективность и удобство поиска документов. Во-первых, в результате классификации всей коллекции документов пользователю становится доступным средство тематической навигации по коллекции, что позволит малоопытному пользователю легко углубляться в искомую предметную область. Во-вторых, механизмы поиска по запросу могут использовать информацию о классификации документов для уменьшения ширины поисковой области, таким образом сокращая число нерелевантных документов в результатах поиска. В-третьих, трудоёмкость выбора среди результата поиска может быть снижена, если документы поисковой выборки предоставлять пользователю не в виде огромных списков, а в виде набора тематических групп, на которые автоматически разбиваются релевантные запросу документы. В таком случае пользователь легко отбросит документы неинтересующих его тематик.

Настоящая работа посвящена классификационному поиску. Однако, традиционные механизмы классификационного поиска – универсальные библиотечные классификаторы (УДК, ГРНТИ, ББК) и специализированные предметные рубрикаторы, имеющие фиксированную структуру, не успевают изменяться вслед за темпом развития науки и техники или требуют высоких затрат как на адаптацию классификаторов, так и на классификацию по ним документов.

Современные методы классификационного поиска основаны на механизме автоматической классификации текстов. Выделяют два вида методов автоматической классификации полных текстов: (а) *методы категоризации* и (б) *методы кластеризации*. В обоих случаях входными данными являются информационно-поисковые образы документов,

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

представляемые в виде множества признаков, характеризующих содержание текста документа.

Методы категоризации распределяют документы по предопределенному набору рубрик на основе знания, полученного из обучающего множества. Разработке и тестированию алгоритмов данного вида, а также связанным с ними алгоритмам представления текстов посвящены труды таких авторов как T. Joachims, D. D. Lewis, R. E. Schapire, H. Schutze, F. Sebastiani, Y. Yang, I. Dagan, S. T. Dumais, M. C. Ageev, И. Е. Кураленок, И. С. Некрестьянов, В. И. Шабанов и ряда других. Однако данный подход решает не все проблемы традиционного классификационного поиска, поскольку регулярная актуализация рубрикатора связана с высокими экспертными затратами на анализ ситуации и подготовку новых обучающих данных. Широко известные методы категоризации подробно изложены в работе [4] и далее рассмотрены не будут.

Методы кластеризации классифицируют документы в условиях отсутствия предопределенного набора рубрик и множества документов-образцов, разбивая документы на группы (кластеры) на основе анализа тематической близости между ними. Разработке алгоритмов данного вида и способов оценки качества получаемого разбиения данных, а также связанным с ними алгоритмам представления текстов посвящены труды таких авторов как C. J. van Rijsbergen, G. Salton, D. Manning и H. Schutze, T. Kohonen, O. Eli Zamir, J. C. Bezdek, M. Halkidi, Д. В. Ландэ, М. В. Киселев, К. М. Кириченко и ряда других. Решить представленную проблему потенциально способны только алгоритмы кластеризации. Их применение позволит получить средство навигации как по всей коллекции документов, так и по её подмножествам, динамически формируя для каждого случая наиболее подходящий предметный рубрикатор, например, для результата поиска по ключевым словам.

Поскольку данная работа зародилась в процессе автоматизации университетской библиотеки, то особое значение имеет вид автоматического рубрикатора, который должен быть унаследован от традиционного библиотечного предметного рубрикатора, основанного на многолетнем опыте обслуживания читателей. Кроме того, метод построения рубрикатора должен быть применим к коллекциям русскоязычных документов, которые могут сильно различаться по размеру (статьи, учебники, книги, технические руководства и др.) и по тематике (физика, металлургия, транспорт, информатика, радиотехника, авиация, энергетика, приборостроение и др.). Таким образом, поставлена задача разработать метод автоматического построения рубрикатора в ориентированном на читателя виде для политематической коллекции документов без ограничения на их объём и оценить применимость для её решения метода кластерного анализа.

2 Обзор известных методов

2.1 Способы формирования образов документов

Формирование образов документов заключается в преобразовании исходного текста на естественном языке в некоторое формализованное представление его смысла. В задаче кластеризации образы документов представляют собой многомерные векторы в пространстве их признаков, например, слов документов. Тогда метод кластерного анализа определяет сходство документов путём вычисления геометрической близости их векторов. В работе выполнено обобщение известных способов формирования образов документов (рис. 1), направленных на решение проблемы автоматической обработки текстов, обусловленной неоднозначностями естественного языка [22, 2, 8, 25, 23 и др.].



Рис. 1. Этапы формирования образа полнотекстового документа для задачи кластеризации и способы к их реализации (статистический подход)

Выполненное обобщение показало, что наиболее приемлемым является подход, который использует в качестве смысловых признаков одиночные слова, прошедшие морфологический анализ и оценку их значимости по схеме TFIDF. Что касается многословных признаков, теоретически они представляются более подходящими для описания смысла текста. Однако известные попытки их применения показывают, что качество автоматической классификации, если и увеличивается, то незначительно. Причём это достижимо только в том случае, когда уделяется особое внимание тщательному отбору «качественных» фраз, что связано с большими вычислительными затратами, иначе может наблюдаться даже снижение качества поиска или классификации [2, 8, 5]. Что касается расширения понятия признака документа информацией из тезаурусов и других внешних словарей, то требование применимости метода классификации к политематическим коллекциям исключает возможность их использования.

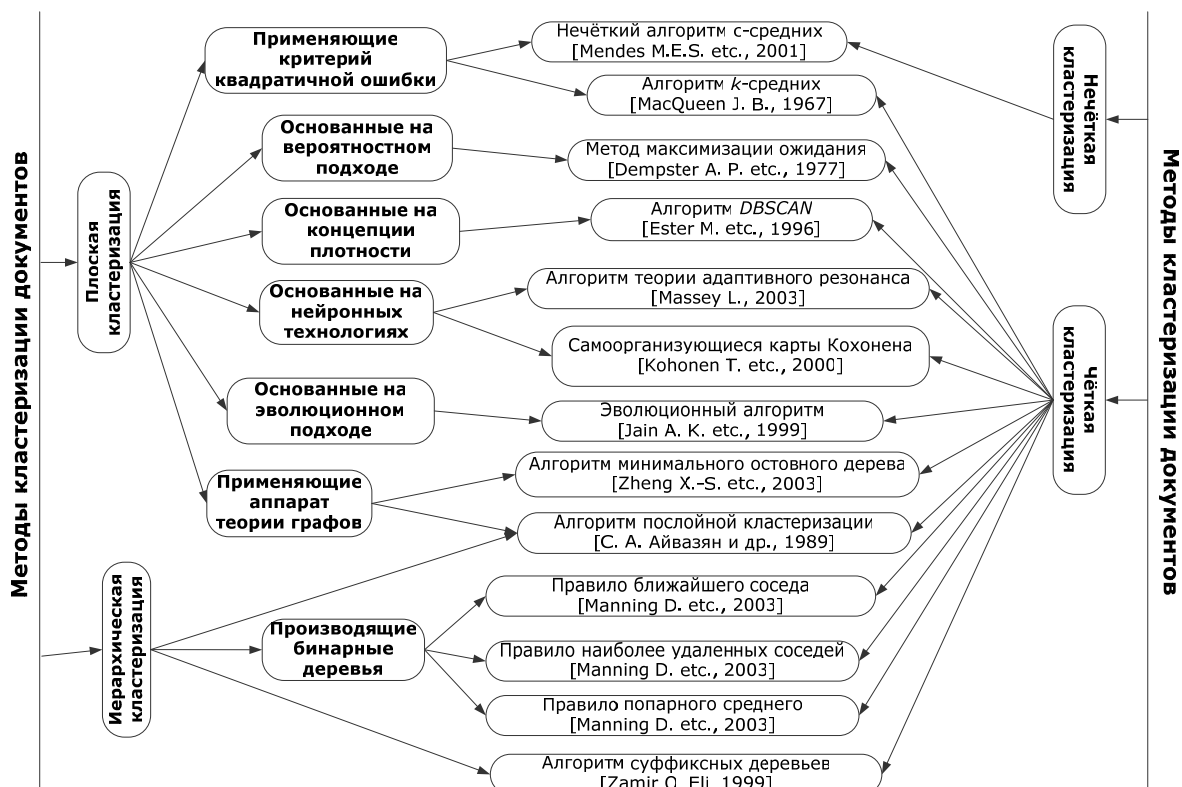


Рис. 2. Алгоритмы автоматической кластеризации документов

Заметим, что количество однословных признаков может достигать десятков или сотен тысяч даже для коллекций небольшого размера. Процесс кластеризации в таких условиях связан не только с высокими вычислительными затратами, но и с низким качеством разбиения на кластеры. Поэтому остро стоит проблема отбора информативных признаков из всего первоначального множества слов. В работе сделана попытка решить данную проблему в условиях отсутствия априорной информации о массиве документов. Соответствующий алгоритм будет изложен ниже.

2.2 Алгоритмы кластеризации документов

Наибольшее распространение получили следующие алгоритмы кластеризации полнотекстовых документов (рис. 2): иерархические (основанные на правилах ближайшего соседа, наиболее удалённых соседей и попарного среднего, суффиксные деревья) [18, 26], квадратичной ошибки (алгоритм k -средних, нечёткий алгоритм s -средних) [17, 21], алгоритмов теории графов (алгоритм минимального остовного дерева) [27], вероятностные (алгоритм максимального ожидания) [11], основанные на концепции плотности (алгоритм DBSCAN) [12], нейросетевые (самоорганизующиеся карты Кохонена, алгоритмы теории адаптивного резонанса) [19, 15] и эволюционные (генетические) [14].

Для кластеризации документов в настоящей работе потребовался алгоритм, способный разбить документы на такие кластеры, которые пригодны для формирования рубрикатора в унаследованном

от библиотечного предметного рубрикатора виде. Например, Предметный Рубрикатор в библиотеке МГТУ является иерархическим с глубиной иерархии не более 3 уровней, содержит очень «плотные» рубрики, это приводит к тому, что в рубрикаторе 25% всех рубрик (4520) являются корневыми, и в добавок между рубриками имеется родственная связь типа «смотри также». Следовательно, необходим алгоритм, позволяющий разбивать документы на иерархические кластеры с возможностью простого управления как количеством уровней, так и степенью детализации кластеров-рубрик на каждом.

Обобщение известных методов кластеризации показало, что традиционные иерархические и плоские алгоритмы для этого не пригодны. Поэтому предложен подход, называемый послойной кластеризацией, который взят из теоретического труда Айвазяна С. А [6]. Этот подход позволит получать необходимую структуру кластеров и для него характерны: (а) более низкие вычислительные затраты по сравнению с традиционными иерархическими алгоритмами, (б) отсутствие необходимости сложной настройки входных параметров и (в) независимость результата кластеризации от выбора начальных точек поиска разбиения данных. Однако, его применение к текстовым документам неизвестно. Поэтому в работе была поставлена и решена задача на основе выбранного подхода создать алгоритм, обеспечивающий ещё и высокую достоверность автоматической классификации текстов.

2.3 Способы оценки кластерного решения

Оценка качества разбиения документов на кластеры различными алгоритмами выполняется путём вычисления значений мер качества полученного результата кластеризации. Выделяют следующие два вида мер качества кластеризации документов: *внешние* (а) и *внутренние меры* (б) [3, 9, 13 и др.]. Внешние меры основаны на сравнении автоматического разбиения с полученным от экспертов «эталонным» разбиением этих же данных. Идея, положенная в основу этих мер, заключается в том, чтобы для каждой пары документов автоматически сопоставить два решения о сходстве их тематиках. Одно решение получено от экспертов, второе – в процессе автоматической классификации. Примерами внешних мер являются традиционные для оценки систем поиска такие характеристики, как: полнота, точность, ошибка классификации, F_1 -мера и другие. В данном случае их оценка производится на основе следующей таблицы:

Таблица 1
Основные категории результата кластеризации документов и количество документов, принадлежащих каждой категории

Для каждой пары документов d_j и d_i	d_j и d_i принадлежат одному кластеру в «эталонном» разбиении	d_j и d_i принадлежат разным кластерам в «эталонном» разбиении
d_j и d_i принадлежат одному кластеру в автоматическом разбиении	a	c
d_j и d_i принадлежат разным кластерам в автоматическом разбиении	b	d

Далее внешние меры вычисляются по известным формулам:

$$Recall = \frac{a}{a+b},$$

$$Precision = \frac{a}{a+c},$$

$$Error = \frac{b+c}{(a+b+c+d)},$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall},$$

где *Recall* – это мера полноты, *Precision* – это мера точности, *Error* – ошибка автоматической классификации, F_1 – F_1 -мера.

Внутренние меры основаны на анализе полученного разбиения без привлечения экспертов, они

анализируют такие свойства как отделимость и компактность кластеров документов. т. е. оптимальным кластерным решением считается то, расстояние между кластерами которого максимально, и в то же время расстояние между документами внутри каждого кластера минимально. Приведём несколько примеров. Для краткости не будем описывать все формулы, по которым вычисляются перечисленные индексы, их можно найти в указанной литературе.

Мерой качества результата иерархической кластеризации является *кофенетический коэффициент корреляции* (CPCC). Для его вычисления необходимо определить кофенетическую матрицу S_C , каждый элемент которой представляет собой уровень близости, на котором документы d_i и d_j впервые встретились в одном кластере, (номер уровня в иерархии кластеров). CPCC – это статистический индекс, показывающий степень сходства кофенетической матрицы S_C и действительной матрицы близости документов коллекции S , вычисляемый следующим образом:

$$CPCC(C) = \frac{(1/M) \sum_{i=1}^{N_D-1} \sum_{j=i+1}^{N_D} (s_{ij} s_{C_{ij}} - \mu_s \mu_{s_c})}{\sqrt{\left((1/M) \sum_{i=1}^{N_D-1} \sum_{j=i+1}^{N_D} (s_{ij}^2 - \mu_s^2) \right) \left((1/M) \sum_{i=1}^{N_D-1} \sum_{j=i+1}^{N_D} (s_{C_{ij}}^2 - \mu_{s_c}^2) \right)}}$$

где $-1 \leq CPCC \leq 1$, $M = \frac{N_D(N_D-1)}{2}$, N_D – количество

документов, s_{ij} и $s_{C_{ij}} - (i, j)$ -е значения матриц S и S_C соответственно, $\mu_s = \frac{1}{M} \sum_{i=1}^{N_D} \sum_{j=i+1}^{N_D} s_{ij}$, $\mu_{s_c} = \frac{1}{M} \sum_{i=1}^{N_D} \sum_{j=i+1}^{N_D} s_{C_{ij}}$

– здесь средние значения матриц S и S_C соответственно. Чем ближе значение CPCC к нулю, тем ниже сходство между матрицами.

Мерами качества плоского чёткого разбиения документов являются: *индекс Дана* (DI), модифицированный Беждеком для снижения чувствительности к шуму; *мера Дейвиса-Булдина* (DB), являющаяся функцией отношения суммы внутрикластерного разброса к межкластерному расстоянию; индекс Калинского и Гарабача (Calinski и Harabasz) (CH), вычисляющий отношение следа матрицы межкластерного разброса к следу матрицы внутрикластерного разброса, I -индекс и другие [10, 20, 24 и др.]. Оптимальное разбиение данных достигается при минимизации значения второго индекса и максимизации значений остальных индексов.

Мерами качества плоского нечёткого разбиения документов являются, например: *модифицированный индекс коэффициента разделения* (MPC), учитывающий только нечёткую принадлежность, и *индекс Хие и Бени* (XB), учитывающий не только нечёткую принадлежность, но и структуру данных [16]. Оптимальное разбиение данных достигается при максимизации значений этих индексов.

Заметим, что какими бы ни были надёжными меры обоснованности на одних данных, на других их поведение может существенно измениться. Дело в том, что данный подход к оценке обоснованности разбиения данных хорошо работает при условии,

что достоверные кластеры являются компактными, плотными и хорошо разделимыми. В задаче кластеризации текстов документы в общем случае группируются на кластеры произвольной формы. Тогда традиционные основы критериев обоснованности (дисперсия, плотность, отделимость) не являются достаточными. В высокоразмерном пространстве признаков документов сходство между документом и центроидом кластера в общем случае является низким и, поэтому трудно ожидать компактных кластеров.

3 Предложенный метод автоматического формирования рубрикатора коллекции документов

Рассмотрим предложенные алгоритмы для каждого этапа формирования рубрикатора.

3.1 Алгоритм формирования образов

Алгоритм формирования образов на первом шаге строит множество признаков документов всей коллекции $P = \{p_i\}, i = \overline{1, N_P}$. Для этого из текстов документов извлекаются все слова, из них удаляются общеупотребительные, входящие в список стоп-слов, оставшиеся слова проходят морфологический анализ по алгоритму М. Портера [1], выделяющего псевдоосновы слов. В результате получается исходное множество признаков документов, которое задаёт многомерное пространство $\Pi^{N_P}(\vec{D})$. Исходные образы документов строятся в виде векторов в данном пространстве $(\vec{D} = \{\vec{d}_j\}, j = \overline{1, N_D})$. Координатами вектора документа являются веса соответствующих признаков в данном документе, вычисленные по схеме TFIDF.

Следующий шаг – редукция признаков, целью которой является повышение качества представления содержания документов их образами. Предложенный алгоритм редукции заключается в последовательной обработке исходного множества признаков с использованием, во-первых, общеизвестного способа отсекаания признаков (*принудительной редукции*), во-вторых, с применением нового подхода к выявлению малоинформативных признаков документов (*избирательной редукции*).

Принудительная редукция состоит из следующих операций:

1) Удалить из множества признаков P все признаки, частота встречаемости которых в коллекции менее $\tau_{\min}^{DF}$ или более $\tau_{\max}^{DF}$ документов, где $\tau_{\min}^{DF}$ и $\tau_{\max}^{DF}$ – заданные пороговые значения.

2) Удалить из образа каждого документа все признаки, имеющие наиболее низкие веса, доля низковесовых признаков задаётся общим параметром τ^{WP} .

Избирательная редукция в отличие от принудительной признаёт малоинформативными не одни и те же слова для всех документов сразу. Она мотивирована тем, что один и тот же признак может являться значимым для описания одной предметной области и незначимым для другой, и при этом иметь достаточно высокую частоту встречаемости в документах обеих областей. Тогда анализировать его информативность следует по отдельности для каждой тематики, представленной в коллекции. Таким образом, избирательная редукция состоит из следующих операций:

1) Выполнить первоначальную кластеризацию документов с высоким значением меры близости документов τ^{sim} , где $0 \leq \tau^{sim} \leq 1$ – входной параметр алгоритма (алгоритм кластеризации как в п. 3.2). В результате выделяется множество кластеров «очевидно родственных» документов $C^0 = \{C_1^0, \dots, C_{|C^0|}^0\}$, т. е. документов, принадлежность которых к одному тематическому классу ярко выражена в использовании большого количества одинаковых слов. Полагается, что все признаки, встретившиеся в «очевидно родственных» документах, являются представительными для этого тематического класса.

2) Среди всех признаков «очевидно родственных» документов отобрать признаки, обладающие наиболее высокой различительной способностью. Благодаря таким признакам в группу данного тематического класса не попали документы, являющиеся менее близкими по смыслу с остальными документами этой группы. Для выполнения данного отбора следует:

а) для каждой группы документов

$$C_i^0 = (\vec{d}_1^{(i)}, \dots, \vec{d}_{N_i}^{(i)}) = \begin{pmatrix} w_{1(i)}^{(1)} & \Lambda & w_{N_i(i)}^{(1)} \\ & \Lambda & \\ w_{1(i)}^{(N_P)} & \Lambda & w_{N_i(i)}^{(N_P)} \end{pmatrix},$$

где N_i – количество документов в C_i^0 ; $w_{j(i)}^k$ – вес k -го признака в j -м документе, принадлежащем i -му кластеру C_i^0 , $k = \overline{1, N_P}$, $j = \overline{1, N_i}$,

перейти в пространство документов данной группы $\Pi^{C_i^0}(P)$, где объектами наблюдения являются признаки

$$\begin{pmatrix} \vec{p}_1 \\ \dots \\ \vec{p}_{N_P} \end{pmatrix} = \begin{pmatrix} w_{1(i)}^{(1)} & \Lambda & w_{N_i(i)}^{(1)} \\ & \Lambda & \\ w_{1(i)}^{(N_P)} & \Lambda & w_{N_i(i)}^{(N_P)} \end{pmatrix};$$

б) в каждом полученном подпространстве $\Pi^{C_i^0}(P)$ выполнить кластеризацию признаков со значением $\tau^{sim}=1$ (алгоритм кластеризации как в п. 3.2). Этот шаг основан на предположении, что в подпространстве родственных документов информативные признаки окажутся слабо схожими между

собой, а малоинформативные признаки будут в векторном виде лежать на одной прямой;

б) считать информативными только те признаки, которые образовали одноточечные кластеры, остальные удалить из образов документов, принадлежащих группе C_i^0 .

3.2 Модифицированный метод послойной кластеризации

Итоговые образы документов разбиваются на кластеры с применением послойного подхода, суть которого заключается в представлении информации о документах в виде графа, вершины которого соответствуют документам, а рёбра помечены значениями меры близости этих документов. Мера близости вычисляется как косинус угла между векторами документов. Тогда для заданной последовательности порогов меры близости $\{\tau_1^{sim}, \tau_m^{sim}\}$

определяются подграфы $G^1 \subseteq \dots \subseteq G^m$ путём удаления из исходного графа рёбер, помеченных значением меры близости, меньшим порога τ_t^{sim} . Затем для каждого подграфа выделяются компоненты связности. Множество выделенных компонент t -го подграфа считается разбиением коллекции документов на уровне близости τ_t^{sim} , иначе называемое t -м слоем классификации документов. На выходе алгоритма получается m вложенных разбиений, которые отражают иерархические связи между кластерами документов.

Первые испытания данного подхода показали необходимость его модификации, вызванной склонностью некоторых документов являться «узкими» переключками между кластерами, что приводило к ошибочному объединению кластеров на более высоких уровнях. Для уменьшения влияния таких документов на результат кластеризации, в работе предложено заменять все документы, образовавшие кластеры, центроидами данных кластеров, т. е. их средними элементами, для вычисления кластеров на последующих уровнях.

На заключительном этапе иерархический набор кластеров преобразуется в рубрикатор следующим образом:

- выявляются родственные связи между кластерами одного и того же уровня путём вычисления меры близости между их центроидами;

- формируется описание рубрик, состоящее из краткого названия и списка ключевых слов, детально раскрывающих тематику рубрики. Список ключевых слов – это наиболее весомые слова центроида кластера. А название – это самое весомое слово, встречающееся и в заглавиях документов кластера, и в его центроиде.

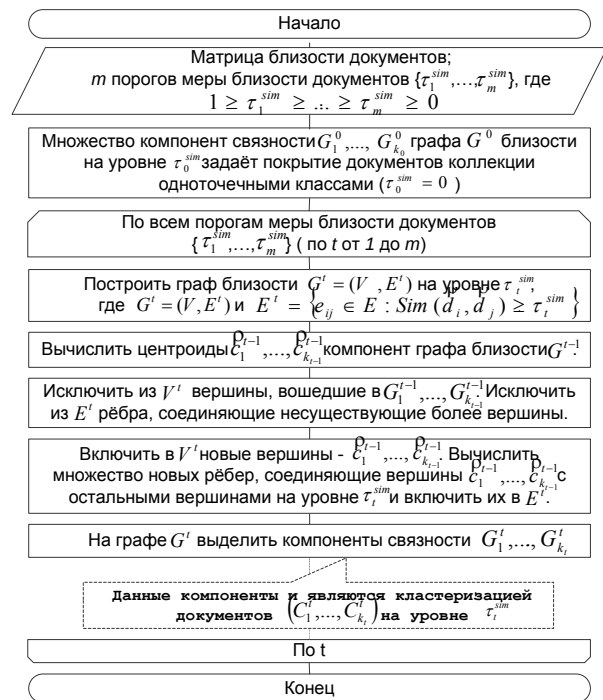


Рис. 3. Модифицированный алгоритм послойной кластеризации

4 Испытания

Разработанные алгоритмы реализованы в программном комплексе, с которого помощью исследован предложенный подход к кластеризации полнотекстовых документов. Тестовыми данными являлись две коллекции русскоязычных полнотекстовых документов:

- 1) документы on-line библиотеки по информационным технологиям CITFORUM [7] (1572 документа, тематика различных областей информационных технологий, документы различного объема (от обзорных статей до полных учебников по программированию), общий объём – 81,35 МБ текста) – коллекция CL1572;

- 2) электронные ресурсы библиотеки МГТУ им. Н. Э. Баумана – авторефераты диссертаций (234 документа, тематика различных научных специальностей, общий объём – 18,14 МБ текста) – коллекция TAL234.

Пример экранной копии интерфейса навигации по подмножеству коллекции CL1572 (200 случайных документов) с помощью автоматически построенных рубрикаторов представлен на рис. 4. У каждой рубрики имеется однословное название, подписанное на графе, при выборе рубрики отображается список её ключевых слов, список родственных рубрик и список документов, ей принадлежащих. Родственные связи также отражены на самом графе рубрикатора. В зависимости от выбора порогов близости документов для алгоритма кластеризации может быть получен рубрикатор с разной степенью укрупнения рубрик.

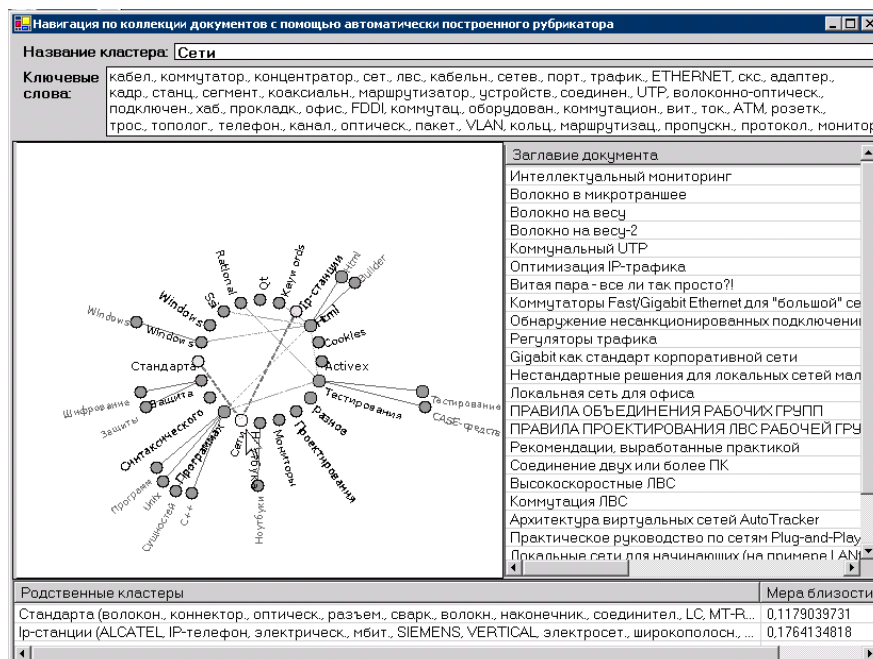


Рис. 4. Пример интерфейса навигации по выборке из коллекции документов с помощью автоматически построенного рубрикатора (выбрана рубрика "Сети")

4.1 Испытание предложенного алгоритма формирования образов документов

Испытания алгоритма формирования образов документов проведены на коллекции CITFORUM, содержащей разнородные по объёму и содержанию документы. Сравнивался результат кластеризации с применением предложенного подхода к отбору информативных признаков и без него.

Таблица 2

Оценка способа формирования образов.

(1) – без редукции, (2) – с принудительной редукцией, (3) – с принудительной и избирательной редукцией

	(1)	(2)	(3)
Внешние меры качества кластеризации			
MicroF ₁ -мера	0,190	0,466	0,505
Error	0,506	0,101	0,084
Внутренние меры качества кластеризации			
CPCC	-0,188	-0,508	-0,580
DI	0,539	0,598	0,577
DB	0,196	0,258	0,180
CH	3,086	6,3687	7,226
I-Index	0,0014	0,0016756	0,0018
Скорость кластеризации			
Время (с)	1783	584	85

Анализ показал, что применение разработанного алгоритма редукции, сократило количество призна-

ков в 3,5 раза и связей типа «признак-документ» в 5,7 раз, что, с одной стороны, снизило вычислительные затраты алгоритма кластеризации, а с другой стороны, повысило качество разбиения данных.

Значения параметров алгоритма редукции подобраны на той же тестовой коллекции путём оценки качества кластеризации для различных значений. В результате выработаны следующие рекомендации: $\tau_{\min}^{DF} = 1$ и $\tau_{\max}^{DF} = (75\% \text{ от количества документов})$; $\tau^{WP} = 0,60$; $\tau^{sim} = 0,40$.

4.2 Испытание модифицированного алгоритма послойной кластеризации

Испытание алгоритма кластеризации проведено как на коллекции CITFORUM, так и на коллекции авторефератов.

На коллекции CITFORUM кластеризация выполнена тремя алгоритмами:

1) иерархическим агломеративным алгоритмом по правилу попарного среднего; усечение полученного дерева выполнено при минимальном пороге меры близости документов внутри кластеров, равном $\tau^{sim} = 0,20$;

2) исходным алгоритмом послойной кластеризации при $\tau_1^{sim} = 0,40$ и $\tau_2^{sim} = 0,20$.

3) модифицированным алгоритмом послойной кластеризации при $\tau_1^{sim} = 0,40$ и $\tau_2^{sim} = 0,20$.

Рубрикатор, полученный для полной коллекции CITFORUM, представлен на рис. 5.

Посмотреть примеры автоматически построенных рубрикаторов для некоторых подмножеств коллекции CITFORUM и осуществить навигацию по ним возможно по адресу <http://demo.pesкова.ru>.

5 Заключение

Проведённые эксперименты показали эффективность предложенного в работе метода автоматического формирования рубрикатора документов и положенного в его основу алгоритма кластеризации полных текстов. Программная реализация метода предназначена для электронных библиотек как элемент их поисковых систем. Такой элемент способен являться как самостоятельным поисковым механизмом, так и служить средством повышения качества работы других поисковых механизмов, например, поиска по ключевым словам. Также предложенный метод может быть использован при разработке средств анализа динамики развития научно-технического знания в фондах электронных документов.

К сожалению, на момент написания статьи не удалось провести испытание метода на тестовых коллекциях, содержащих десятки или сотни тысяч документов, это планируется сделать в дальнейшем.

Литература

- [1] [Алгоритм выделения псевдооснов М. Портера] [Электронный ресурс]. – Режим доступа: <http://snowball.sourceforge.net>, свободный.
- [2] Губин М. В. Модели и методы представления текстового документа в системах информационного поиска / М. В. Губин // Научно-техническая информация. Сер. 1. – 2004. – №12. – С. 12-24.
- [3] Гусарова Л. Проверка обоснованности кластерного решения / Л. Гусарова, И. Яцкив // Reliability and statistics in transportation and communication (RelStat'03). – Рига, 2004. – Т. 5, №2. – С.49-56.
- [4] Пескова О. В. Методы автоматической классификации текстовых электронных документов // Научно-техническая информация. Сер. 2. – 2006. – №3. – С. 13-20.
- [5] Поляков П. Ю., Плешко В. В. RCO на РОМИП 2006 // Труды четвертого российского семинара РОМИП'2006 (Суздаль, 19 октября 2006 г.). – СПб. : НУ ЦСИ, 2006.
- [6] Прикладная статистика: Классификация и снижение размерности : Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин ; под. ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989. – 607с. : ил.
- [7] [Электронная библиотека по информационным технологиям CITFORUM] [Электронный ресурс]. – Режим доступа: <http://www.citforum.ru>, свободный.
- [8] Bekkerman R., Allan J. Using Bigrams in Text Categorization [Electronic resource]. – 2003. –

Access mode:

www.cs.umass.edu/~ronb/papers/bigrams.pdf.

- [9] Bezdek J. C., Pal N. R. Some New Indexes of Cluster Validity // IEEE Transactions On Systems, Man And Cybernetics. – 1998. – Vol. 28, No. 3. – P. 301-315.
- [10] Boutin F., Hascolt M. Cluster Validity Indices for Graph Partitioning // Proceedings of the Eight International Conference on Information Visualization (IV'04). IEEE – 2004.
- [11] Dempster A. P. Maximum likelihood from incomplete data via the EM algorithm / A. P. Dempster, N. M. Laird, D. B. Rubin // Journal of the Royal Statistical Society. Series B. – 1977. – Vol. 39, No. 1. – P. 1-38.
- [12] Ester M. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). – Portland, 1996. – P. 226-231.
- [13] Halkidi M. On Clustering Validation Techniques / M. Halkidi, V. Batistakis, M. Vazirgiannis // Journal of Intelligent Information Systems, Kluwer Academic Publishers. Manufactured in The Netherlands. – 2001. – 17:2/3. – P. 107-145.
- [14] Jain A. K. Data Clustering: A Review / A. K. Jain, M. N. Murty, P. J. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, No. 3. – P. 264-323.
- [15] Kohonen T. Self organization of a massive document collection / T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela // IEEE Transactions on neural networks. – 2000. – Vol. 11, No. 3. – P. 574-585.
- [16] Kuo-Lung W., Miin-Shen Y. A cluster validity index for fuzzy clustering // Pattern Recognition Letters. – 2005. – Vol. 26. – P. 1275-1291.
- [17] MacQueen J. B. Some Methods for classification and Analysis of Multivariate Observations // Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. – Berkeley, 1967. – Vol. 1. – P. 281-297.
- [18] Manning C. D., Schütze H. Foundations of statistical natural language processing. – Cambridge: MIT Press, 1999. – 620 p.
- [19] Massey L. Evaluating quality of text clustering with ART1 // Proceedings of the International Joint Conference on Neural Networks. – Portland, 2003. – Vol. 2. – P. 1402-1407.
- [20] Maulik U., Bandyopadhyay S. Performance Evaluation of Some Clustering Algorithms and Validity Indices // IEEE Transactions On Pattern Analysis And Machine Intelligence. – 2002. – Vol. 24, No. 12. – P. 1650-1654.
- [21] Mendes M.E.S., Sacks L. Dynamic Knowledge Representation for e-Learning Applications // Proc. of the 2001 BISC International Workshop on Fuzzy Logic and the Internet, FLINT'2001. – Berkeley, 2001. – P. 176-181.

- [22] Salton G., Buckley C. Weighting approaches in automatic text retrieval // Information Processing and Management. – 1988. – Vol. 24(5). – P. 513-523.
- [23] Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34, No. 1. – 47 p.
- [24] Stein B. On Cluster Validity and the Information Need of Users / B. Stein, S. M. zu Eissen, F. Wißbrock // 3rd IASTED Int. Conference on Artificial Intelligence and Applications: Proceedings of AIA 03. – Benalmadena, 2003. – P. 216-221.
- [25] Yang Y., Pedersen J. O. A Comparative Study on Feature Selection in Text Categorization // The Fourteenth International Conference on Machine Learning: Proceedings of ICML'97. – San Francisco, 1997. – P. 412-420.
- [26] Zamir O. E. Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results [Electronic resource]. – 1999. – Access mode: http://turing.cs.washington.edu/papers/zamir_thesis.pdf.
- [27] Zheng Xiao-Shen Algorithm of documents clustering based on minimum spanning tree / Zheng Xiao-Shen, He Pi-Lian, Tian Mei, Yuan Fu-Yong // International Conference on Machine Learning and Cybernetics. – Xi-an, 2003. – Vol. 1. – P. 199-203.

Automatic Full-text Documents Classifier Building

Olga V. Peskova

Developing information retrieval systems based on full-text document classification mechanism is discussed. Text representation approaches, text clustering methods and clustering validation techniques are described. Automatic full-text documents subjects heading building method is proposed. Method effectiveness is evaluated on two Russian document collections that contain different size and different subject documents.

Открытые электронные коллекции с адаптивным визуальным интерфейсом фасетной навигации

© О.Л. Обухова, М.М. Гершкович, Т.К. Бирюкова, И.В. Соловьев, А.П. Чочиа

Институт проблем информатики РАН
support@intergallery.ru

Аннотация

В статье рассматривается специальный вид электронных коллекций, состоящих из независимых объектов, полностью определяемых своими атрибутами. Для данного вида коллекций авторами разработана формальная модель, в которой каждый элемент коллекции представлен фасетной формулой объекта, структура коллекции хранится в виде фасетной таблицы, процесс навигации состоит из последовательности шагов по построению фасетной формулы запроса, при этом состояние фасетной таблицы динамически отображается на пользовательский интерфейс. Актуальное подмножество фасетной таблицы, отвечающее интересам конкретного пользователя, формируется на основе обработки фасетной формулы запроса пользователя. Разработанная формальная модель положена в основу реализационной модели, в которой визуальный интерфейс адаптируется к актуальному состоянию коллекции и выбранному пользователем сценарию доступа к объектам коллекции.

1 Введение

Электронные коллекции, состоящие из независимых объектов, полностью определяемых набором атрибутов, составляют немалую долю активно эксплуатируемых сетевых ресурсов. Такого рода коллекциями являются коллективно ведущиеся порталы, справочники и энциклопедии, библиотечные фонды, виртуальные музеи, Интернет – магазины. Методы доступа к объектам таких коллекций используют готовые технические решения – иерархическую навигацию и атрибутивный поиск, что часто неэффективно для потребителя. Навигация гарантирует выборку объекта, однако предоставляет жестко

заданный путь, соответствующий одному варианту классификации, а для различных пользователей могут быть удобными разные пути. Атрибутивный поиск позволяет одновременно выбирать несколько признаков классификации, однако, результат его не гарантирован.

Для коллекций, объем которых составляет тысячи объектов, а пользователи существенно различаются по интересам, большинство из применяемых в настоящее время средств доступа неэффективно. Пользователь не всегда точно знает, по каким признакам ему надо выбирать искомый объект. Субъективное мнение поставщиков информации о том, какими признаками наделены объекты, может не совпадать с точкой зрения пользователей об этих же объектах.

Поэтому авторы поставили задачу на уровне визуального интерфейса навигации динамически отображать полную информацию о составе коллекции, при этом каждому пользователю предоставлять с помощью специальных интерфейсных элементов только ту информацию, которая соответствует его кругу интересов, определенному в процессе работы с коллекцией.

Для достижений этой цели был разработан метод динамического построения визуального интерфейса для открытых электронных коллекций

2 Метод динамического построения визуального интерфейса фасетной навигации

2.1 Принцип преемственности

В методе динамического построения визуального интерфейса фасетной навигации на этапе разработки были использованы те же правила, которые лежат в основе как иерархической навигации, так и атрибутивного поиска. Построение иерархической навигационной схемы связано с классификацией объектов коллекции. Дерево навигации представляет собой рубрикатор соответствующей классификации. При удачном разбиении на рубрики минимизируется число шагов для доступа к объектам. Но в реальной жизни количество вариантов, интересных для разных пользователей, весьма велико. Поиск, в том виде, в котором он представлен в большинстве функционирующих коллекций, по сути, реализует

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

принцип фасетной классификации. Фасетная система классификации [1], в отличие от иерархической, позволяет выбирать признаки классификации независимо друг от друга. Признаки классификации называются фасетными признаками (они соответствуют «атрибутам» в терминах информационно-поисковых языков). Фасетный признак вместе с множеством допустимых значений образует фасет.

2.2 Технология адаптивного визуального интерфейса фасетной навигации

Ранее авторами был предложен метод динамической фасетной классификации объектов в открытых электронных коллекциях [2], определяющий структуру хранения данных в виде фасетной таблицы. Каждый объект определяется фасетной формулой объекта, представленной в виде пар: {фасетный признак – его значение для данного объекта}.

Фасетная таблица, содержащая фасетные формулы для всех объектов коллекции, представляет данные для всех возможных вариантов их классификации. Авторами была разработана модель фасетной навигации [3], определяющая навигацию как пошаговый интерактивный процесс построения фасетной формулы запроса. На каждом шаге пользователь уточняет фасетную формулу, выбирая фасет и задавая подмножество его искомым значений. Задача навигационного инструментария – обеспечить, чтобы формируемому запросу соответствовала непустая выборка. Один и тот же пользователь в разных случаях, а, тем более, различные пользователи предпочтут различные траектории выбора фасетов, что соответствует всевозможным вариантам классификационных схем.

В настоящей работе внимание сконцентрировано на технологии адаптивного визуального интерфейса фасетной навигации в открытых электронных коллекциях. Пользователю предоставляется возможность влиять на структуру и состав визуального интерфейса в зависимости от поступающей информации. Динамически появляющаяся информация на каждом шаге доступа к объекту предоставляет пользователю данные, анализ которых позволяет ему сделать свой следующий выбор. Текущее состояние навигационного инструментария отражает актуальное подмножество фасетной таблицы, которое фиксирует набор фасетов и их значений на каждом шаге алгоритма фасетной навигации.

В общем случае, фасетная таблица представляет собой многомерную матрицу, структура которой и значения ячеек являются переменными величинами для открытых электронных коллекций. Задача технологии адаптивного визуального интерфейса заключается в динамическом отображении содержимого фасетной таблицы на пользовательский интерфейс. Актуальное подмножество фасетной таблицы, отвечающее интересам конкретного пользователя, формируется на основе обработки фасетной формулы запроса пользователя.

Для получения ненулевого результата поиска объекта в коллекции визуальный интерфейс должен

предоставлять только актуальные элементы, которые определены двумя факторами:

- состоянием фасетной таблицы;
- предысторией формирования фасетной формулы запроса.

2.3 Формальная модель адаптивной фасетной навигации

Введем определения понятий «Фасетная формула объекта» и «Фасетная формула запроса».

Пусть

$G = \{g_i \mid i = 1, \dots, k\}$ – коллекция информационных объектов, где g_i – информационный объект;

$\Phi(G) = \{\Phi_j(G) \mid j = 1, \dots, l\}$ – множество фасетов на G .

При рассмотрении конкретной коллекции G будем обозначать: $\Phi_j(G) = \Phi_j$.

$\varphi^{\Phi_j} = \{\varphi_{j1}, \dots, \varphi_{jm}\}$ – множество допустимых значений фасета Φ_j .

Каждому объекту g_i в коллекции G ставится в соответствие множество:

$\varphi(g_i) = \{\varphi_j(g_i) \mid j = 1, \dots, l\}$, где $\varphi_j(g_i)$ – значение фасета Φ_j для объекта g_i : $\varphi_j(g_i) \in \{\varphi^{\Phi_j}, \varepsilon\}$ (если значение $\varphi_j(g_i)$ не задано, то $\varphi_j(g_i) = \varepsilon$).

Фасетная формула $FF(g_i)$ объекта g_i представляет собой множество пар: $\{ \langle \text{фасетный признак; значение фасета для объекта } g_i \rangle \}$ для тех фасетов, значения которых заданы для объекта g_i :

$$FF(g_i) = \left\{ \langle \Phi_j, \varphi_j(g_i) \rangle \mid \Phi_j \in \Phi(G), \right. \\ \left. \varphi_j(g_i) \in \varphi^{\Phi_j} \forall j \mid j \in \{1, \dots, l\}, \varphi_j(g_i) \neq \varepsilon \right\}.$$

Фасетная таблица содержит фасетные формулы для всех объектов коллекции:

$FT(G) = \{FF(g_i) \mid g_i \in G, i = 1, \dots, k\} = \{ \langle \Phi_j, \varphi^{\Phi_j} \rangle \mid j = 1, \dots, l \}$ – множество фасетов с их допустимыми значениями на объектах коллекции G .

Для формирования результирующего запроса с целью получения искомой выборки объектов $RG \subset G$ необходимо в процессе пошагового алгоритма задать значения некоторого подмножества фасетов, указывая в каждом требуемые значения фасетных признаков.

Множество допустимых фасетов и множество допустимых значений для каждого фасета на очередном шаге алгоритма определяется актуальным подмножеством объектов коллекции (состоящим из объектов, имеющих выбранные на предыдущих шагах значения фасетов).

Определим фасетную формулу запроса как множество пар:

{<фасетный признак; значение фасета, формирующее запрос>};

$$RFF = \{ \langle \Phi_j, \varphi_j^R \rangle \mid j \in J^R \},$$

где J^R – множество индексов фасетов, значения которых были выбраны (отмечены) в процессе пошагового уточнения фасетной формулы запроса;

φ_j^R – значение фасета Φ_j , выбранное для формирования запроса в процессе пошагового уточнения фасетной формулы запроса ($\varphi_j^R \in \Phi_j$).

Операция выборки (*retrieve*) формирует подмножество объектов $RG \subset G$ в соответствии с фасетной формулой запроса:

$$RG = \text{retrieve}(G, RFF)$$

В предельном случае, если фасетная формула запроса совпадает с фасетной формулой объекта, то объект однозначно определен и попадает в выборку.

Но на практике фасетная формула объекта, которая задается владельцем объекта при занесении его в коллекцию, может не совпадать с представлениями пользователя о том, с каким набором фасетов и их значений объект был занесен в коллекцию.

Разработанная авторами технология адаптивной фасетной навигации предоставляет пользователю возможность строить фасетную формулу запроса в форме пошагового интерактивного процесса с той целью, чтобы на каждом шаге получать информацию, анализ которой позволит сделать очередной выбор.

На каждом шаге пользователь уточняет фасетную формулу, выбирая значение для очередного (одного) фасета из множества допустимых значений фасета на данном шаге. Такой способ навигации обеспечивает получение на каждом шаге непустого подмножества объектов.

Обозначим: J^A – множество индексов допустимых («allowable») фасетов на очередном шаге, $J^A \subset \{1, \dots, l\}$.

При выполнении указанного алгоритма на каждом шаге $J^R \subset J^A$, где J^R – множество индексов фасетов, значения которых были отмечены на данном и предыдущих шагах.

На 1-м шаге алгоритма выборка соответствует всей коллекции G , допустимые фасеты – всем фасетам коллекции, а фасетная таблица содержит фасетные формулы для всех объектов коллекции, то есть, все фасеты и все их значения для всех объектов коллекции:

$$RG^1 = G;$$

$$\Phi^1 = \Phi(G) = \{ \Phi_j \mid j \in J^A \}, \quad \text{где}$$

$J^A = \{1, \dots, l\}$ – множество индексов допустимых фасетов на 1-м шаге;

$$FT^1 = FT(G).$$

На 1-м шаге алгоритма пользователь формирует запрос в виде следующей фасетной формулы запроса:

$$RFF^1 = \{ \langle \Phi_j, \varphi_j^R \rangle \mid j \in J^{R^1} \} - \text{множество,}$$

состоящее на 1-м шаге из одной пары: <фасетный признак, отмеченное значение фасета>, при этом множество J^{R^1} состоит из одного индекса ($J^{R^1} = \{j\}$) – индекса фасета, значение которого $\varphi_j^R \in \Phi_j$ было выбрано для формирования запроса на 1-м шаге.

В особых случаях (например, редкие свойства объектов), уже первый шаг может привести к получению искомого объекта.

На очередном (n-ом) шаге алгоритма в формировании фасетной формулы запроса участвуют только те фасеты (образующие множество допустимых фасетов на n-ом шаге), которые имеют непустые значения на объектах, принадлежащих выборке предыдущего шага, и только эти значения (допустимые значения) будут использоваться для построения запроса.

Таким образом, на n-ом ($n > 1$) шаге n-1 фасетов (принадлежащих множеству допустимых фасетов на n-ом шаге) имеют заданные значения, выбранные на предыдущих (1, ..., n-1) шагах, и пользователю предоставляется выбор значения одного из оставшихся фасетов, принадлежащих множеству допустимых фасетов на n-ом шаге.

На n-ом шаге алгоритма получаем следующее состояние навигационной системы (после выполнения предыдущего запроса):

$$\Phi^n = \{ \Phi_j \in \Phi^{n-1} \mid \exists g \in RG^{n-1} : \varphi_j(g) \neq \varepsilon \} - \text{множество допустимых фасетов на n-ом шаге:}$$

$$\Phi^n = \{ \Phi_j^n \mid j \in J^{A^n} \},$$

где $J^{A^n} = \{j \mid j \in \{1, \dots, l\}, \Phi_j \in \Phi^n\}$ – множество индексов допустимых фасетов на n-ом шаге алгоритма;

$$FT^n = \{ \langle \Phi_j^n, \varphi_j^{\Phi_j^n} \rangle \mid j \in J^{A^n} \} \quad (1)$$

– фасетная таблица на n-ом шаге,

где $\varphi_j^{\Phi_j^n}$ – множество допустимых значений фасета Φ_j на n-ом шаге;

$$RG^n = \text{retrieve}(RG^{n-1}, RFF^{n-1}) = \{ g_i \mid g_i \in RG^{n-1}, \varphi_j(g_i) = \varphi_j^{R^{n-1}} \forall j \in J^{R^{n-1}} \} \quad (2)$$

– подмножество выбранных объектов на n-ом шаге. Новый запрос (на n-ом шаге) имеет фасетную формулу запроса:

$$RFF^n = \{ \langle \Phi_j, \varphi_j^R \rangle \mid j \in J^{R^n} \}, \quad (3)$$

представляющую собой совокупность пар: <фасетный признак, отмеченное значение фасета>, где J^{R^n} – множество индексов фасетов, значения кото-

рых были выбраны (отмечены) на шагах $1, \dots, n$; φ_j^R - выбранное значение фасета Φ_j .

Данное теоретико-множественное описание информационных структур и процессов задает четкую основу для спецификации метода динамического построения визуального интерфейса. В общем случае, фасетная таблица представляет собой многомерную матрицу, структура которой и значения ячеек являются переменными величинами для открытых электронных коллекций. Задача метода динамического формирования визуального интерфейса заключается в динамическом отображении содержимого фасетной таблицы на пользовательский интерфейс. Актуальное подмножество фасетной таблицы, отвечающее интересам конкретного пользователя, формируется на основе обработки фасетной формулы запроса пользователя.

Элементы визуального интерфейса формируются программным образом и определяются текущим состоянием фасетной таблицы (формула (1)) и предысторией формирования фасетной формулы запроса.

Текущее состояние визуального интерфейса позволяет пользователю получить полное представление о составе коллекции. Дополнительно к этой информации пользователь может проанализировать текущую выборку, описываемую формулой (2), и сделать очередной шаг при выборе объектов, сформировав фасетную формулу запроса в соответствии с формулой (3).

Динамически появляющаяся информация на каждом шаге доступа к объекту предоставляет пользователю данные, анализ которых позволяет ему сделать свой следующий выбор.

Метод динамического формирования визуального интерфейса лежит в основе построенной авторами модели адаптивной фасетной навигации. Модель построена на примере открытой коллекции произведений искусства.

Данная предметная область выбрана авторами для достижения достаточной иллюстративности и образности при демонстрации и анализе разрабатываемых идей и их воплощения в набор методов и технологий, реализующих адаптивную фасетную навигацию в открытых коллекциях. Набор фасетов и их возможных значений являются общепринятым подходом специалистов в данной области.

3 Презентационные решения метода динамического построения визуального интерфейса фасетной навигации

Метод динамического построения визуального интерфейса гарантирует то, что при выборе любой из траекторий движения по фасетам пользователь за конечное число шагов получит доступ к искомому объекту.

Авторы предложили следующую форму визуального представления навигационного инструментария на примере открытой электронной коллекции произведений искусства.

Фасеты представлены в виде «облаков», для просмотра всех возможных значений фасета используются механизмы «прокрутки», что позволяет эффективно использовать место на экране (рис. 1). Количество фасетов выбирается в соответствии с принципом, сформулированным американским психологом Миллером [4]: для того, чтобы выбор был эффективным, количество элементов в нем не должно быть больше семи-девяти.

Задание любого из значений произвольно выбранного фасета приводит к перезагрузке подмножества доступных значений в остальных фасетах в соответствии с состоянием фасетной таблицы.

Сочетание фасетов и их значений на определенном этапе представляет легко обозримую картину, что позволяет пользователю сделать очередной выбор для достижения цели (рис. 2).

Рисунок 3 иллюстрирует тот факт, что пользователь мог указать любую траекторию выбора фасетов для достижения той же цели.

На первом шаге пользователю предоставляется полный обзор состояния фасетной таблицы (рис. 1).

Проиллюстрируем на примере, приведенном на рис. 1, введенные выше понятия фасетов и области их значений. Объекту g , представляющему собой работу Казимира Малевича «Весенний пейзаж», ставятся в соответствие следующие значения фасетов:

$\Phi_1 = \text{«Вид искусства»} : \varphi_1(g) = \text{«Живопись»};$

$\Phi_2 = \text{«Жанр»} : \varphi_2(g) = \text{«Пейзаж»};$

$\Phi_3 = \text{«Стиль»} : \varphi_3(g) = e \text{ (значение не задано)};$

$\Phi_4 = \text{«Техника»} : \varphi_4(g) = \text{«Масло»};$

$\Phi_5 = \text{«Материал»} : \varphi_5(g) = \text{«Холст»};$

$\Phi_6 = \text{«Автор»} : \varphi_6(g) = \text{«Малевич»}.$

В том случае, если пользователь выбрал в фасете «Вид искусства» значение «живопись», в остальных фасетах произошла перезагрузка значений: представлены только те значения фасетов, которые определены для множества объектов с фасетным признаком {вид искусства: живопись}.

Приведем примеры различных сценариев действий пользователя при выборе пути доступа к искомому объекту. Сценарий действия пользователя формируется в процессе анализа поступающей информации и относится к адаптивным стратегиям выбора пути [5].

На втором шаге пользователь выбирает в фасете «Техника» значение «масло». После перезагрузки

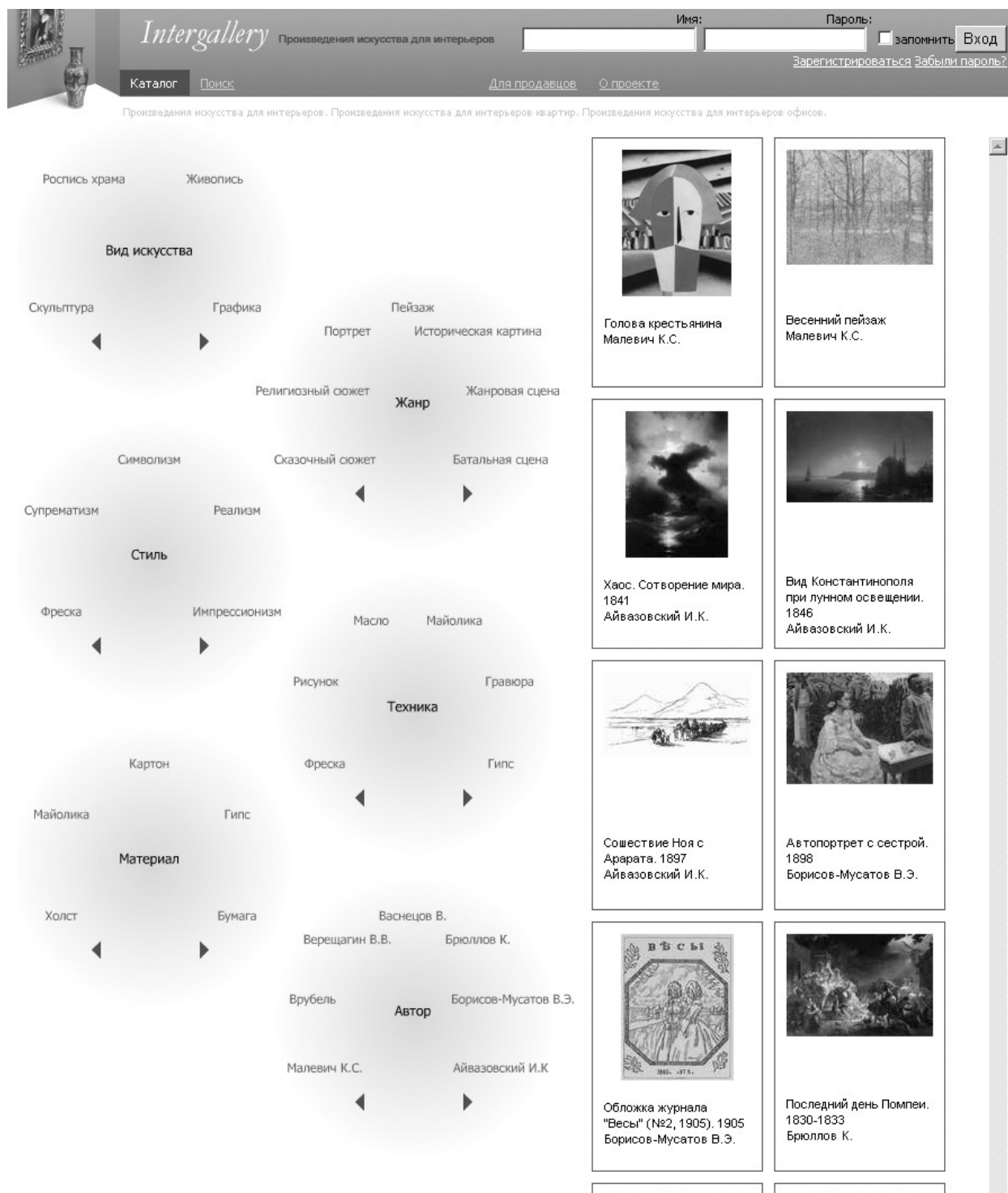


Рисунок 1

значений в остальных фасетах пользователю предоставляется легко обозримое отображение актуального для него подмножества фасетной таблицы (рис. 2). Анализ полученной информации позволит пользователю определить дальнейшую стратегию пути доступа к искомому объекту.

На рис. 3 проиллюстрирован тот факт, что, пользователь мог выбрать иную последователь-

ность фасетов. Тем не менее, он бы получил тот же искомый результат.

Метод динамического формирования визуального интерфейса в объединении с оригинальными решениями дизайна пользовательского интерфейса позволяют реализовать основные принципы адаптивной фасетной навигации в открытых электронных коллекциях.

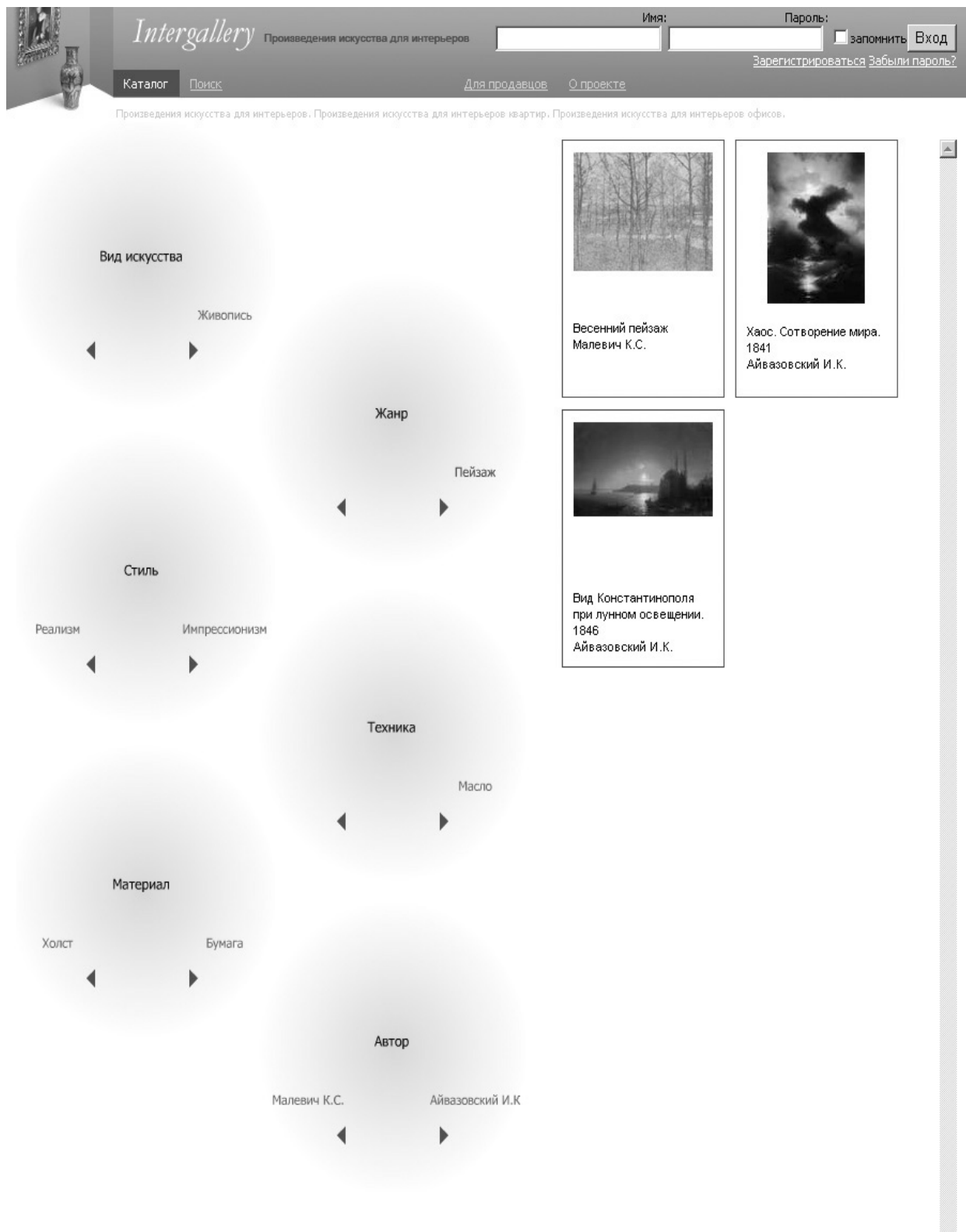


Рисунок 2

4 Реализационная модель адаптивного визуального интерфейса фасетной навигации

Макет реализационной модели размещен в Интернет по адресу www.intergallery.ru. Программная реализация выполнена в форме структурированного

web-сервиса (рис. 4). При разработке использовались программные средства:

- реализация осуществляется в среде ОС: Microsoft Windows 2000/2003;
- веб-сервер: Microsoft Internet Information Server 5.0/6.0;



Рисунок 3

- контейнер серверного приложения: Microsoft ASP.NET 2.0;
 - СУБД: Microsoft SQL Server 2005;
 - в качестве средств разработки используется Microsoft Visual Studio 2005, ASP.NET 2.0;
 - для построения пользовательского интерфейса используется Adobe Flash со встроенным языком ActionScript.
- ASP.NET web-сервисы используют индустриальные стандарты:



Рисунок 4

- XML. Обмен данными между пользовательским интерфейсом и сервером (а именно с "подсистемой задания фасетной формулы и представления результатов запроса") производится в формате XML [5], [9]
- SOAP. Протокол обмена сообщениями между Web службой и клиентом, основанный на XML [8].
- Web Services Description Language (WSDL). Описывает параметры сообщений Web-службы для взаимодействия с клиентами [6], [7].

5 Заключение

В предлагаемой модели динамического построения визуального интерфейса достигается выполнение следующих задач:

- задачи многовариантного выбора путей доступа к объектам в открытых электронных коллекциях;
- задачи поддержки естественных иерархий в данной предметной области;
- задачи динамического отображения актуального подмножества фасетной таблицы;
- задачи динамического формирования визуального интерфейса процесса навигации;
- задачи адаптации пользовательского интерфейса к выбранному пользователем сценарию поиска элементов коллекции.

Литература

- [1] Ранганатан Ш.Р. Классификация двоеточием. Основная классификация : пер. с англ. / под ред. Т.С. Гомолицкой и др. – М., 1970. – 422 с.

- [2] Н.А. Маркова, О.Л. Обухова, И.В. Соловьев, А.П. Чочиа. Web технология динамической классификации квази-однородной электронной коллекции // Труды 9-й Всероссийской научной конференции «Электронный библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 2007. – Переславль-Залесский, 2007. – Том 2. – С. 29–32.
- [3] Н.А. Маркова, О.Л. Обухова, И.В. Соловьев, А.П. Чочиа. Эффективная фасетная навигация в электронных коллекциях // Системы и средства информатики. – М. : Наука, 2007. – № 17. – С. 214–222
- [4] Миллер Дж. Магическое число семь, плюс или минус два. - В кн.: Инженерная психология. М. 1964
- [5] Коновалов И.Н., Коновалов М.Г. Модель и алгоритмы адаптивного управления для системы распределенных вычислительных ресурсов // Международный симпозиум «Информационные технологии и общество», 24 апреля – 01 мая 2007, Москва. – М., 2007. – С. 76–78.
- [6] XML Web Services Basics// <http://msdn.microsoft.com/en-us/library/ms996507.aspx>
- [7] Web Services Description Language (WSDL) 1.1 // <http://www.w3.org/TR/wsdl>
- [8] Web Services Description Language (WSDL) Version 2.0 // <http://www.w3.org/TR/wsdl20/>
- [9] SOAP Version 1.2 // <http://www.w3.org/TR/soap/>
- [10] XML // <http://www.w3.org/XML/>

Open digital collections with adaptive visual interface of facet navigation

Olga Obuhova, Maxim Gershkovich,
Tatiana Biryukova, Ivan Soloviev, Anton Chochia

We propose a formal model to describe a special type of digital collections consisting of independent objects defined by their attributes. Each element of collection is described by object's facet formula. Structure of collection is represented by a facet table. Navigation suggests a stepwise building of facet formula of the request, while the status of the facet table is represented in the user interface dynamically. Relevant subset of facet table, corresponding to particular user's demands, is build by processing the request's facet formula. Proposed formal model is used to produce an executive model with visual interface adapting to the current status of collection and scenario of access to the collection's objects defined by user.

Использование онтологии верхнего уровня для отображения информационных моделей*

© Н.А. Скворцов, С.А. Ступников

Институт проблем информатики РАН
nsvk@ipi.ac.ru, ssa@ipi.ac.ru

Аннотация

Отображение информационных моделей является необходимым шагом при решении задач над неоднородными информационными ресурсами в распределённых информационных системах, электронных библиотеках, предметных посредниках. При отображении моделей производится поиск близких конструкций моделей и их выражение друг через друга. Многообразие конструкций, используемых в информационных моделях, может быть сведено к ограниченному набору элементарных составляющих. Онтологии верхнего уровня описывают такие элементы для широких классов моделей (структурных, объектных, поведенческих и других). Описание конструкций конкретных моделей производится с помощью композиции понятий онтологии верхнего уровня и конкретизации значений их свойств. Таким образом, в эталонной схеме информационной модели, описывающей набор её конструкций и связей между ними, элементы могут быть аннотированы описаниями в терминах онтологии верхнего уровня. В статье представлен подход к семантической интеграции информационных моделей, основанный на обнаружении общности между описаниями их конструкций в терминах онтологии верхнего уровня. Такая интеграция является существенным шагом при построении отображения моделей.

1 Введение

Известные работы по отображению информационных моделей в основном связаны с конкретными моделями или их классами. Работ по методикам отображения произвольных информационных моделей немного. Обычно они базируются на подхо-

дах к конструированию функций отображения элементов моделей (например, [3]).

Следует отметить исследования [6], проводимые в лаборатории композиционных методов проектирования информационных систем ИПИ РАН, посвящённые коммутативному отображению моделей на основе уточнения спецификаций. В настоящее время также реализован Конструктор унифицирующих информационных моделей (Унификатор моделей) [9] – инструмент, поддерживающий процесс построения верифицируемых отображений исходных информационных моделей в каноническую информационную модель (т.е. процесс унификации моделей).

В моделях часто используется весьма ограниченное количество разных видов отношений, это позволяет утверждать, что методы семантической интеграции могут быть эффективными при решении задачи отображения моделей. Перспективным представляется онтологический подход, позволяющий учитывать формальные описания моделей в терминах онтологии. Толчком к разработке подхода, представленного в данной статье, послужила работа [2], в которой для трансляции схем и данных сложные элементы моделей описываются композициями ограниченного числа элементарных.

2 Отображение информационных моделей

Отображение информационных моделей является необходимым шагом при решении задач над неоднородными информационными ресурсами. Модели информационных ресурсов приводятся к однородному представлению, называемому канонической информационной моделью, в котором будут производиться дальнейшие манипуляции информацией из разных ресурсов. Модели информационных ресурсов при этом называются исходными, а каноническая модель – целевой. Задача унификации множества исходных информационных моделей становится особенно актуальной при необходимости масштабирования по количеству неоднородных информационных ресурсов.

И в случае отображения отдельных моделей, и в случае унификации множества моделей с помощью канонической модели, возникает задача поиска для конструкций исходной информационной модели

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

релевантных (имеющих сходную семантику) конструкций целевой. Ниже мы рассмотрим интеграцию конструкций фрагмента модели OWL (языка семантической разметки для публикации и совместного использования онтологий в Веб) [12] и канонической модели, основанной на языке СИНТЕЗ [7]. Интеграция конструкций необходима для последующего построения отображения OWL в каноническую модель. (Замечание: Спецификации эталонных схем, онтологий и аннотаций также представлены на языке СИНТЕЗ. Это связано не с выбранным для статьи примером модели, а с тем, что в проектах язык СИНТЕЗ является универсальным языком спецификаций.)

3 Эталонные схемы моделей

Спецификация информационной модели включает описание её конструкций. Мы используем описание моделей в виде *эталонных схем*, содержащих все значащие конструкции спецификаций в данной модели и отношения между этими конструкциями. Эталонная схема представляет собой абстрактное описание модели, независимое от конкретного синтаксиса (эталонную схему можно рассматривать как модель уровня M2 по классификации MOF [11]). Эталонные схемы могут включать элементарные типы (например, string), тип перечисления, абстрактные типы данных, ассоциации (представляемые атрибутами абстрактных типов).

В качестве примера рассмотрим фрагменты эталонных схем моделей OWL и СИНТЕЗ.

Эталонная схема OWL содержит тип *Ontology*. Онтология может состоять из нескольких директив (тип *Directive*) – это отражается атрибутом *directive* типа *Ontology*. Директивами являются аксиомы, в частности аксиомы классов (тип *ClassAxiom*) и объектных свойств (тип *ObjectPropertyAxiom*). Аксиомы классов содержат их описания (тип *Description*), в частности, определение иерархии классов *SubclassOf*. Объектные свойства могут быть нескольких видов (задаваемых значениями перечислимого типа *ObjectPropertyKind*) и содержат ограничения на область определения (атрибут *_domain*) и область значений (атрибут *_range*) и ссылку на инверсное объектное свойство (атрибут *inverseOf*).

```
{ Ontology;
  in: type;
  name: string;
  directives: {set; type_of_element: Directive}
},
{ Directive;
  in: type
},
{ Axiom;
  in: type;
  supertype: Directive
},
{ ClassAxiom;
```

```
  in: type;
  supertype: Axiom;
  name: string;
  descriptions: {set; type_of_element:
    Description}
},
{ Description;
  in: type
},
{ SubclassOf;
  in: type;
  supertype: Description;
  reference: Description
},
{ ObjectPropertyAxiom;
  in: type;
  supertype: Axiom;
  name: string;
  inverseOf: ObjectPropertyAxiom;
  kind: {set; type_of_element:
    ObjectPropertyKind};
  super: {set; type_of_element:
    SuperProperty};
  _domain: {set; type_of_element:
    ObjectPropertyDomain};
  _range: {set; type_of_element:
    ObjectPropertyRange}
},
{ ObjectPropertyKind;
  in: enum;
  enum_list: {Functional; InverseFunctional;
    Simmetric; Transitive}
}
```

Основной единицей спецификации модели СИНТЕЗ является модуль, задаваемый в эталонной схеме типом *Module*. Модуль может содержать типы (атрибут *_type*) и классы (атрибут *_class_specification*). Типы (*Type-Specification*) составляют иерархию тип/подтип (атрибут *_supertype*), абстрактные типы (*Abstract-Type*) содержат атрибуты (*attributes*). Классы (*Class-Declaration*) могут составлять иерархию класс/подкласс (атрибут *_superclass*). Отношения между классами задаются метаклассами ассоциаций (тип *Association-Metaclass*), для которых могут быть заданы область определения (*_domain*), область значений (*_range*) и ссылка на инверсный метакласс (со знака ‘_’ начинаются имена тех атрибутов, которые совпадают с ключевыми словами языка СИНТЕЗ).

```
{ Module;
  in: type;
  name: string;
  _class_specification: {set; type_of_element:
    Class-Declaration};
  _type: {set; type_of_element:
```



```

    Type-Specification}
},
{ Type-Specification;
  in: type;
  _supertype: {set; type_of_element:
    Type-Specification}
},
{ Abstract-Type;
  in: type;
  supertype: Type-Specification;
  name: string;
  attributes: {set; type_of_element:
    Abstract-Type}
},
{ Attribute-Specification;
  in: type;
  name: string;
  attribute-type: Type-Specification;
  metaslot: Attribute-Metaslot
},
{ Class-Declaration; in: type;
  name: string;
  _superclass: {set; type_of_element:
    Class-Declaration}
},
{ Association-Metaclass;
  in: type;
  supertype: Class-Declaration;
  _inverse: Association-Metaclass;
  _domain: Class-Declaration;
  _range: Class-Declaration;
  _association_type: Association-Type
}

```

4 Онтология верхнего уровня элементов информационных моделей

Для унификации представленных конструкций модели OWL с канонической моделью будет использоваться онтология верхнего уровня, включающая в себя понятия таких простых элементов спецификаций, как типы, атрибуты, множества, перечисления. Поэтому в данной статье рассмотрены лишь структурные элементы информационных моделей. Подобная онтология может быть выстроена на основе теории, изложенной в работе [5]. Ниже приведен фрагмент спецификации данной онтологии на языке СИНТЕЗ.

Большинство конструкций, присущих классу структурных моделей, могут быть составлены из трёх разновидностей элементарных конструкций: тип, атрибут и иерархическое отношение.

Типы (*_Type*) могут быть абстрактными (*AbstractType*), представимыми (*PrintableType* – числа, строки, изображения) либо составными (*ConstructedType* – агрегация неоднородных конструкций *Aggregation* или группирование однородных конструкций *Grouping*).

Атрибуты (*Attribute*) имеют множество разновидностей с различными свойствами: кол-во аргументов (*argumentNumber*), область определения (*_domain*), область значений (*_range*), множественность (*isMultivalued*), инверсный атрибут (*_inverse*).

Иерархические отношения могут выстраиваться над типами или над атрибутами и означать либо отношение подмножества, либо отношение обобщения/специализации.

```

{ _Type;
  in: concept;
  supertype: Construct
},
{ AtomicType;
  in: concept;
  supertype: _Type
},
{ PrintableType;
  in: concept;
  supertype: AtomicType;
  kind: {enum; enum_list: {string, number,
    image}}
},
{ AbstractType;
  in: concept;
  supertype: AtomicType
},
{ ConstructedType;
  in: concept;
  supertype: _Type
},
{ Aggregation;
  in: concept;
  supertype: ConstructedType;
  components: {set; type_of_element: _Type};
  arity: integer;
  inv: {in: predicate, invariant;
    { predicate:
      {all p/Aggregation
        (card(p.components) = p.arity)}}};
  };
},
{ Grouping;
  in: concept;
  supertype: ConstructedType;
  activeDomain: _Type;
  element: _Type
},
{ Attribute;
  in: concept;
  supertype: Construct;
  argumentNumber: integer;
  _domain: {set; type_of_element: _Type};
  _range: {set; type_of_element: _Type};
  _inverse: Attribute;
  minCardinality: integer;
  maxCardinality: integer;
}

```

```

minInverseCardinality: integer;
maxInverseCardinality: integer;
isOptional: boolean;
isMultivalued: boolean;
isKey: boolean
},
{ OneArgumentAttribute;
in: concept;
supertype: Attribute;
oneArgInv: {in: predicate, invariant;
{ predicative:
{all p/OneArgumentAttribute
(p.argumentNumber = 1)}};
};
}

```

Онтологии верхнего уровня для других классов моделей можно почерпнуть из работ, подобных [5]. Например, для процессных моделей создание онтологии может основываться на работе [1], в которой выделены элементарные шаблоны, использующиеся в большинстве современных систем управления потоками работ. Лингвистические, тезаурусные, более развитые по разнообразию отношений онтологические модели могут аннотироваться в терминах онтологии DOLCE [8], комбинация понятий которой может описывать тонкости множества видов отношений.

5 Аннотация эталонных схем моделей онтологическими описаниями

Следующим шагом к интеграции конструкций исходной и целевой моделей является аннотация спецификаций эталонных схем моделей в терминах онтологии верхнего уровня. Это значит, что мы должны определить, чем являются описываемые конструкции в эталонных схемах модели с точки зрения онтологии. На языке СИНТЕЗ аннотация типов в эталонной схеме представляется утверждением типов экземплярами новых онтологических понятий, определяемых как подтипы понятий онтологии. Таким образом, с одной стороны, спецификации схем независимы, с другой, каждый элемент может являться экземпляром подтипа понятия онтологии. Новые онтологические понятия не должны противоречить самой онтологии.

Аннотация в терминах онтологии типа модуля (*Module*) языка СИНТЕЗ в эталонной схеме представляет собой утверждение его экземпляром нового понятия *AggregationOfTwoGroups*. Это новое понятие есть подтип понятия агрегации (*Aggregation*) арностью 2 (для определения разделов классов и типов) и компонентами типа группирований (*Grouping*), соответствующими множествам классов и типов.

```

{ Module;
in: type, AggregationOfTwoGroups;
name: string;

```

```

_type: {set; type_of_element: Type-
Specification};
_class_specification: {set; type_of_element:
Class-Declaration}
}

{ AggregationOfTwoGroups;
in: concept;
supertype: Aggregation;
arityInvariant: {in: predicate, invariant;
{ predicative: {all
p/AggregationOfTwoGroups
(p.arity = 2 & all r/_Type
(in(r, p.components)->r/Grouping))}};
};
}

```

Так типы эталонной схемы могут аннотироваться как экземпляры произвольных композиций понятий онтологии.

6 Подход к интеграции эталонных схем информационных моделей

Рассмотрим онтологию *O* верхнего уровня, описывающую набор конструкций, используемых в достаточно широком классе информационных моделей; эталонные схемы исходной модели *V* и целевой модели *U*; и множества аннотаций A_V и A_U соответственно в терминах онтологии *O*, связанных с определениями эталонных схем. Назовем *аннотацией* пару $\langle u, c \rangle$, где *c* – понятие из *O* либо композиция понятий из *O*, *u* – элемент эталонной схемы, являющийся экземпляром понятия *c*. Таким образом, типы и атрибуты в спецификациях эталонных схем моделей *V* и *U* являются экземплярами понятий, которые в свою очередь есть подтипы понятий онтологии *O* (рис. 1).

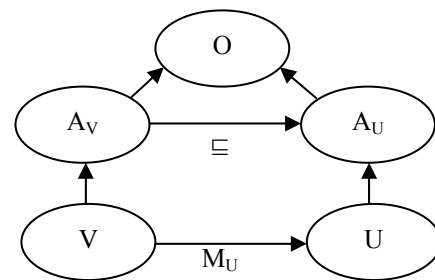


Рис. 1. Интеграция эталонных схем моделей по аннотациям в терминах онтологии

Движущей целью является построение отображения $M_U(V)$ модели *V* в модель *U*. Однако модели изначально никак не связаны, а следовательно, необходимо найти семантически близкие конструкции моделей. Онтологически релевантными принимаются пары конструкций $\langle u, v \rangle$, $u \in U$, $v \in V$, для которых существуют аннотации $\langle v, c \rangle \in A_V$ и $\langle u, d \rangle \in A_U$ такие, что *c* является подпонятием *d*.

Таким образом, абстрагируясь от определений эталонных схем, возможно рассуждать только в терминах онтологии, единственного однородного представления, имеющегося для исходной и целевой моделей. В расчёт берутся только определения самой онтологии O и аннотации A_V и A_U в терминах онтологии.

Работая в терминах одной онтологии, можно поставить следующую задачу. Найти понятия из аннотаций A_U , для которых понятия из A_V являются подтипами. Понятие c является подпонятием d , если супертипы c в онтологии O являются подтипами супертипов d , типы атрибутов c являются подтипами типов атрибутов d и из инвариантов c следуют инварианты d .

Например, понятие, которым аннотируется метакласс ассоциаций языка СИНТЕЗ (*MetaclassAssociationConcept*), совпадает с понятием, которым аннотируется объектное свойство OWL (*ObjectPropertyConcept*), так как оба являются атрибутами с одним аргументом, имеют одинаковые понятия в качестве описания области определения и значений и инверсных ассоциаций.

```
{ MetaclassAssociationConcept;
  in: concept;
  supertype: oneArgumentAttribute;
  _domain: AbstractType;
  _range: AbstractType;
  _inverse: OneArgumentAttribute;
}
```

Таким образом, полагается, что конструкции метакласса ассоциаций языка СИНТЕЗ и объектных свойств OWL, онтологически релевантны и метаклассы ассоциаций могут использоваться для моделирования объектных свойств языка OWL. В отличие от метакласса ассоциаций, атрибуты в языке СИНТЕЗ не имеют инверсных связей, а область значений у них шире.

```
{ AttributeConcept;
  in: concept;
  supertype: oneArgumentAttribute;
  _domain: AbstractType;
  _range: _Type;
  _inverse: none;
}
```

Атрибут *_inverse* понятия *AttributeConcept* (аннотирующего тип атрибута языка СИНТЕЗ) имеет пустой тип, и тип *OneArgumentAttribute* атрибута *_inverse* понятия *ObjectPropertyConcept* (аннотирующего объектное свойство языка OWL) не является его подтипом. Поэтому онтологически релевантными эти конструкции не будут.

7 Применение подхода

Представленный онтологический подход к семантической интеграции информационных моделей применим к любым разновидностям моделей при наличии соответствующих онтологий верхнего уровня. Подход может использоваться при решении любых задач, требующих отображения моделей: при интеграции информационных систем, создании библиотек над неоднородными ресурсами информации, предметных посредников.

В частности, предполагается реализация подхода в рамках Унификатора Моделей [9]. Целью Унификатора является доказательное приведение (унификация) множества разнотипных информационных моделей ресурсов к каноническому, унифицированному представлению для достижения интероперабельности ресурсов.

Унификация исходной модели V есть приведение ее к канонической информационной модели U , т.е. создание такого расширения E (оно может быть пустым) канонической модели U и такого отображения M_U исходной модели V в расширенную каноническую, что исходная модель V покрывалась бы расширенной канонической моделью U . К необходимости создания расширения эксперт приходит в процессе интеграции исходной и канонической моделей, когда обнаруживается, что средств ядра (или ядра с зарегистрированными расширениями) канонической модели недостаточно для выражения некоторых элементов исходной модели. Расширение канонической модели может включать введение новых образцов параметризованных замкнутых логических формул, выражающих в целевой модели зависимости данных, характерные для исходной модели, параметризованных родовых типов данных, представляющих новые типы данных, отсутствующие в ядре канонической модели, метафреймов, аннотирующих дополнительные свойства конструкций ядра в расширенной модели.

Представленный подход семантической интеграции эталонных схем моделей позволит автоматизировать поиск релевантных конструкций для регистрируемых в Унификаторе моделей среди конструкций ядра канонической модели и уже сформированных расширений. Расширения конструкции, онтологически релевантные конструкциям регистрируемой модели, проверяются на возможность повторного использования. Если для некоторых конструкций регистрируемой модели онтологически релевантные конструкции не найдены ни среди средств ядра, ни среди конструкций имеющихся расширений, то предполагается, что для них необходимо создать новые расширения ядра канонической модели.

Реализация подхода может быть выполнена для онтологической модели, основанной на дескриптивной логике, с использованием автоматических средств вывода в ней. Вместо отношения супертипа/подтипа между абстрактными типами данных в этом случае будет использоваться отношение по-

гlossenция [4], в данной логике доказываемое автоматически. В частности, определения онтологии верхнего уровня и аннотаций могут задаваться на языке OWL DL [12], а в качестве системы вывода использоваться средство доказательства Pellet [13].

8 Заключение

В статье представлен подход к семантической интеграции информационных моделей для их отображения, применяющий вывод над аннотациями конструкций модели в терминах онтологии верхнего уровня. Он применим к широкому кругу моделей и не сложен в реализации.

Литература

- [1] W. M. P. van der Aalst, et al. Workflow Patterns. – Distributed and Parallel Databases, 14(3):5-51, 2003.
- [2] P. Atzeni, P. Cappellari, and P. A. Bernstein. Model independent schema and data translation. EDBT 2006.
- [3] G. Caplat, J. L. Sourrouille. Model Mapping in MDA. Workshop in Software Model Engineering, 2002.
- [4] I. Horrocks, U. Sattler, S. Tobies. Practical reasoning for very expressive description logics. Logic Journal of IGPL, 8(3), 2000
- [5] R. Hull, R. King. Semantic Database Modeling: Survey, Applications, and Research Issues. ACM Computing Surveys, Vol. 19, 1987
- [6] L. A. Kalinichenko. Methods and tools for equivalent data model mapping construction. Advances in Database Technology: Proc. of the International Conference on Extending Database Technology EDBT'90. LNCS 416. – Berlin-Heidelberg: Springer-Verlag, 1990. – P. 92-119.
- [7] L. A. Kalinichenko, S. A. Stupnikov, D. O. Martynov. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007. – 171 p.
- [8] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, L. Schneider. The WonderWeb Library of Foundational Ontologies Preliminary Report. ISTC-CNR, Italy, 2003
<http://www.loa-cnr.it/Papers/DOLCE2.1-FOL.pdf>
- [9] В. Н. Захаров, Л. А. Калиниченко, И. А. Соколов, С. А. Ступников. Проектирование канонических информационных моделей для интегрированных информационных систем. Информатика и приложения. – 2007. – Т. 1, вып. 2. – С. 15–38.
- [10] С. А. Ступников. Моделирование композиционных уточняющих спецификаций. Диссертация канд. техн. наук. – Москва: ИПИ РАН, 2006. – 195 с.

- [11] Meta Object Facility (MOF) Core Specification Version 2.0.
<http://www.omg.org/docs/formal/06-01-01.pdf>
- [12] OWL Web Ontology Language Guide. W3C.
<http://www.w3.org/TR/owl-guide/>
- [13] Pellet: The Open Source OWL DL Reasoner.
<http://pellet.owldl.com/>

Application of upper level ontology for mapping of information models

N.A. Skvortsov, S.A. Stupnikov

Information model mapping is a necessary step in task solving over heterogeneous information resources in distributed information systems, digital libraries, subject mediators. For model mapping, close constructions of models are found and express one through another. Variety of constructions used in information models, may be reduced to a set of simple constituents. Upper level ontologies describe such constituents for wide classes of models (structural, object, behavioral, and others). Description of particular model constructions performed by combining concepts of upper level ontology and concretizing values of their properties. So in information model's reference schema, which describes a set of its constructions and their relationships may be annotated by descriptions in terms of upper level ontology. The paper represents an approach to semantic integration of information models, based on discovering commonalities between descriptions of their constructions in terms of upper level ontology. Such integration is an essential step for model mapping.

* Работа проведена при поддержке грантов, 06-07-89188-а, 08-07-00157-а и программы фундаментальных исследований отделения нанотехнологий и информационных технологий РАН «Фундаментальные основы информационных технологий и систем», проект 1-10.

Анализ и сравнение систем интеграции неоднородных информационных ресурсов*

© А.Е. Вовченко, Л.А. Калиниченко

Институт проблем информатики РАН
itsnein@gmail.com, leonidk@synth.ipi.ac.ru

Аннотация

Существует два подхода к решению задач в информационных системах над множеством интегрированных информационных ресурсов: движимый ресурсами и ориентированный на проблему. В статье предпочтение отдается второму. Одной из проблем при таком подходе остается та, что в различных инфраструктурах (веб-сервисы, грид-архитектуры, семантический веб, и др.) до сих пор открытым является вопрос интегрированного представления множественных источников информации для исследователя, решающего задачу. В статье рассматриваются основные принципы интеграции неоднородных информационных ресурсов, на основании чего выбирается набор критериев, по которым сравнивается ряд известных прототипов систем интеграции.

1 Введение

В различных областях науки наблюдается экспоненциальный рост объема получаемых экспериментальных (наблюдательных) данных. Например, в астрономии текущий и ожидаемый темп роста данных от наземных и космических телескопов удваивается в течение периода от шести месяцев до одного года. Сложность использования таких данных увеличивается еще и вследствие их естественной разнородности. Число организаций, получающих данные наблюдений в отдельных областях науки в мире, велико. Разнообразие (информационная несогласованность) получаемой информации вызывается, в частности, не только большим числом организаций, производящих наблюдения, и их независимостью, но и разнообразием объектов наблюдения, непрерывным и быстрым совершенствованием техники наблюдений, вызывающим адекватные изменения структуры и содержания накапливаемой информации. Это приводит к необходимости исполь-

зования неоднородной, распределенной информации, накопленной в течение значительного периода наблюдений технологически различными инструментами.

Чрезвычайно быстро развивается также программный инструментарий, включающий многообразные сервисы для поддержки различных видов обработки информации при решении научных задач и проведении исследований. Такие сервисы производятся различными научными организациями, их описания неоднородны и неполны. Увеличивающийся разрыв между исследователями и источниками данных и сервисов приводит к необходимости поиска новых путей организации решения задач над множественными распределенными коллекциями данных и программ, которые концентрируются в специализированных центрах данных и вычислительных ресурсов. Разработан (разрабатывается) ряд инфраструктур, которые технически позволяют реализовать решение задач над множественными информационными источниками. Среди них Веб сервисы, Грид-архитектуры, Семантический Веб, технологии распределенных баз данных, интероперабельные технологии промежуточного слоя, и др. Они составляют техническую среду для организации решения задач.

Традиционно при решении задач специалисты используют привычные для них источники информации, и формулируют задачи, учитывая лишь такие источники. Подобные способы формулирования и решения задач называются далее *движимыми* конкретными *ресурсами* информации, отобранными до или, в лучшем случае, в процессе формулирования задачи. Очевидна неполнота информации, которую удается охватить при таком подходе. Множество источников данных и сервисов, существующих в Интернете, их разнообразие, вызывают потребность в радикальном изменении такого традиционного подхода. Существо этого изменения заключается в том, что задачи должны формулироваться независимо от существующих источников информации, и лишь после такой формулировки, должна осуществляться идентификация релевантных задаче источников, приведение их к виду, требуемому в задаче, их интеграция, идентификация сервисов, которые позволяют реализовать отдельные части абстрактного процесса решения задачи. Только после этого должно осуществляться конструирование конкрет-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

ного процесса решения на основе отобранных источников данных и сервисов. Этот подход, называемый *ориентированным на проблему*, должен значительно больше опираться на использование техники представления знаний, нежели традиционный. Так, формулирование задачи должно быть основано на определении ее проблемной области, включающем ее терминологию и систему понятий, абстрактное описание соответствующей материальной системы, определение адекватных моделей и теорий, абстрактное описание требуемых в задаче характеристик объектов реального мира, определение методов, алгоритмов и процессов решения задачи.

Одной из проблем при таком подходе остается та, что во всех названных выше инфраструктурах до сих пор открытым является вопрос интегрированного представления множественных источников информации для исследователя, решающего задачу.

На сегодняшний день существует множество средств интеграции неоднородных информационных ресурсов, созданных для тех или иных нужд, достаточно лишь взглянуть на их далеко неполный список [1]. Предварительная классификация подходов к интеграции представлена в статье [2]. В данной статье рассматриваются подходы к интеграции ресурсов, основанные на взглядах [3].

Различаются два принципиально различных подхода к проблеме интеграции: 1) двигаясь от ресурсов к задачам (глобальная схема образуется как интегрированная схема множества ресурсов независимо от приложения) и 2) двигаясь от приложения к ресурсам (описание предметной области приложения образуется независимо от ресурсов (в терминах понятий, структур данных, функций, процессов), а затем релевантные приложению ресурсы отображаются в это описание).

Первый подход, движимый информационными ресурсами, является не масштабируемым по отношению к числу ресурсов, не дает возможности достижения семантической интеграции ресурсов в контексте конкретного приложения, не ведет к доказательной идентификации релевантных приложению ресурсов, не способствует повышению стабильности спецификации посредника в процессе эволюции ресурсов, релевантных приложению. Эти недостатки являются характерными для подхода, при котором глобальная схема является взглядом (Global as View (GAV)) [3, 4]. GAV может служить в качестве базовой техники подхода, движимого информационными ресурсами.

Другой подход (движимый приложениями) предполагает создание предметного посредника, который поддерживает взаимодействие между приложением и ресурсами на основе определения прикладной области (определения посредника). Второй подход имеет очевидные преимущества по отношению к подходу, движимому информационными ресурсами. Процесс регистрации неоднородных информационных ресурсов в предметном посреднике в подходе, движимом приложениями, основан на

технике GLAV, комбинирующей два подхода – LAV (Local As View) – когда локальная схема ресурса является взглядом над схемой посредника, и GAV. Согласно LAV [4], схемы регистрируемых ресурсов рассматриваются как материализованные взгляды над виртуальными классами посредника. В этом случае GAV взгляды служат для разрешения различных конфликтов между спецификациями ресурсов и посредника и обеспечивают правила трансформации результатов запроса от ресурса в представление в посреднике. Подобная техника регистрации обеспечивает стабильность спецификации приложения при изменении конкретных информационных ресурсов и их фактического присутствия (удаление ресурса, добавление новых ресурсов, и пр.), а также масштабируемость посредников по отношению к числу ресурсов, регистрируемых в них. В GLAV [5], голова определения взгляда LAV может содержать произвольный запрос над схемой ресурса для устранения конфликтов на основе техники GAV. Кроме перечисленных, существует подход именуемый BAV (Both As View) [6], суть которого заключается в построении реверсивных взглядов. В статье [6] продемонстрировано как из BAV взглядов можно получить взгляды LAV и GAV.

Целью настоящей работы является анализ соответствия ряда известных исследовательских прототивов средств интеграции неоднородных ресурсов некоторому набору критериев полноты средств интеграции. В следующем разделе дано описание выбранного набора критериев для сравнения. Затем представлен ряд известных исследовательских прототипов средств интеграции и осуществлено формирование сводной таблицы анализа этих прототипов.

2 Выбор критериев для сравнения средств интеграции

Выбор критериев является важным при реализации какого-либо сравнения. Набор критериев, с одной стороны, должен быть полным, отражающим по возможности все основные аспекты средства интеграции, а с другой стороны, он должен быть операбельным без необходимости включения в него всего множества различных параметров средств интеграции. Весьма ограниченный набор критериев сравнения предложен в [7], где для оценки подходов интеграции предлагается критерий ASME (Abstraction, Selection, Modeling, Explicit Semantics). Abstraction заключается в абстрагировании пользователя, использующего посредник, от ресурсов, иначе говоря все ресурсы и их детали при интеграции скрываются от пользователя. Selection означает возможность выбора пользователем тех ресурсов, которые необходимы для решения его задачи. Modeling означает возможность создания пользователем своего собственного представления предметной области для решения конкретной задачи. Наконец, Explicit Semantics означает поддержку возможности связывания схемы с семантическим описанием. Кроме

того, данный критерий означает возможность задания запросов к метаданным.

Несмотря на ограниченность критерия ASME, в настоящей работе представляется возможным в качестве базы использовать эти четыре критерия, существенно дополнив их. Критерии “А” и “М” являются характеристикой интеграционного подхода (LAV, GAV, ...), поэтому они заменены указанием используемого подхода. Критерии “S” и “E” остаются без изменения.

Для обоснования критериев далее будут рассмотрены основные принципы построения средства интеграции.

2.1 Интеграционный подход (LAV, GAV и другие)

Как было отмечено во введении, существуют два подхода: движимого ресурсами (GAV) и движимого приложением (LAV). LAV представляет интерес в контексте решения задач над неоднородными информационными ресурсами. Используя подход, ориентированный на проблему, специалист формулирует задачу, описывая базовые сущности и понятия предметной области, в рамках которой предполагается решать задачу. Среди прочего, специфицируются структуры данных, специфицируются понятия предметной области, выражаемые соответствующими онтологическими спецификациями, описываются функции, процессы и пр. Спецификация предметной области, представляет собой спецификацию предметного посредника, для решения класса задач. Сущности и понятия предметной области, определенные таким образом, не зависят от существующих информационных ресурсов. Также в пользу данного подхода могут служить доводы, высказанные в [7], при описании критериев Abstraction, Modeling.

Поэтому в качестве критерия был выделен критерий используемого подхода (Approach, “А”).

2.2 Создание расширяемой канонической информационной модели

При интеграции неоднородных ресурсов в посреднике нужно уметь семантически отождествлять объекты, представленные в различных информационных моделях, и семантически правильно отображать схемы интегрируемых ресурсов в схему посредника. Поскольку в общем случае ресурсы неоднородны (представлены в различных моделях), при интеграции неоднородных ресурсов для однородного представления их семантики требуется приведение различных информационных моделей к унифицированному виду в рамках некоторой унифицирующей информационной модели, которая называется *канонической*.

Для унификации разнородных спецификаций прежде всего требуется умение сопоставлять спецификации различных ресурсов друг с другом так, чтобы можно было отвечать на вопрос, можно ли при реализации посредника использовать спецификацию

существующего ресурса вместо фрагмента спецификации посредника. Для этого достаточно доказать, что рассматриваемые спецификации находятся в отношении уточнения. Говорят, что спецификация А уточняет спецификацию D, если А можно использовать вместо D так, что пользователь D не будет замечать этой замены. Средства доказательства факта уточнения спецификации некоторого компонента спецификацией другого компонента (реализуемые на основе теоретико-модельных нотаций и соответствующего инструментария [10–12]) составляют фундамент применяемых методов конструирования унифицирующих (канонических) моделей представления информации в посредниках. Каноническая информационная модель служит в качестве общего языка, эсперанто, для адекватного выражения семантики разнородных моделей представления информации, используемых в разнообразных информационных ресурсах, релевантных посреднику. Методы отображения информационных моделей и синтеза расширяемых канонических информационных моделей для слоя предметных посредников подробно рассмотрены в [13]. Там же предложена архитектура Унификатора информационных моделей.

Задачей Унификатора моделей является унификация множества информационных моделей (называемых *исходными*), использующихся в различных информационных ресурсах некоторой информационной системы. Унификация исходной модели R есть приведение ее к канонической информационной модели C, т.е. создание такого расширения E ядра канонической модели (которое может быть и пустым) и такого отображения M исходной модели в расширенную каноническую, что исходная модель *уточняет* расширенную каноническую модель. Уточнение моделей означает, что для любой допустимой спецификации г модели R ее образ M(r) при отображении M уточняется спецификацией г. Синтез канонической модели заключается в построении указанных расширений ядра канонической модели для всех исходных моделей и объединении таких расширений. В качестве ядра канонической информационной модели применяется язык СИНТЕЗ [14], который ориентирован на спецификацию предметных посредников и синтез канонических моделей. Во многих проектах подобные вопросы остаются без должного внимания, прототип разрабатывается только для нескольких моделей данных (чаще всего реляционной, XML). Поэтому возможность Унификации (Unification “U”) произвольных информационных моделей представляется важным критерием.

2.3 Семантический поиск соответствий спецификаций схемы посредника и схем информационных ресурсов

Регистрация релевантных посреднику ресурсов рассматривается как задача композиционного проектирования систем [8, 9]. Регистрация ресурсов есть процесс целенаправленной трансформации спецификаций, включающий декомпозицию специ-

фикаций посредника на непротиворечивые фрагменты, поиск среди спецификаций релевантных ресурсов подходящих типов данных – кандидатов для уточнения ими спецификаций типов посредника, построение выражений, определяющих классы ресурсов в виде композиции классов посредника.

Принципиальным моментом в этой схеме является реализация доказательства уточнения фрагментов спецификаций посредника спецификациями ресурсов в процессе построения отображений таких спецификаций. Идентификация релевантных ресурсов (предшествующая регистрации) основана на использовании трех моделей – *модели метаданных*, характеризующих свойства ресурсов, собранных в некотором реестре, *онтологической модели*, позволяющей формально определять понятия предметной области посредника или ресурса, и *канонической модели*, позволяющей формально определять структуру и поведение объектов предметной области посредника и информационных ресурсов. Рассуждения в канонической и в онтологической модели основаны на семантике канонической модели и средствах доказательства уточнения. При этом в онтологической модели необходимо достичь согласования понятийной семантики спецификаций посредника и регистрируемых в нем ресурсов. Рассуждения в модели метаданных являются эвристическими на основе нефункциональных требований к нужным в классе задач ресурсам. Необходимые модели метаданных и алгоритмы поиска составляют часть метода. В целях проектирования спецификации посредника и ресурсов задаются в однородном их представлении в канонической модели, хотя для этого может потребоваться преобразование в такую модель из некоторого другого языка спецификаций, например из UML.

Выше описан подход, используемый в проекте SYNTHESIS, в других проектах семантика может использоваться иначе, но несомненна важность наличия возможности использования семантических описаний, а также возможности поиска по метаданным спецификаций релевантных ресурсов, или для других задач. В [7] такой критерий был предложен, и назван Explicit Semantics (“E”).

2.4 Ориентация средства интеграции на решение задач

Важным аспектом является ориентация средства интеграции на решения научных задач. В связи с этим целесообразно введение двух дополнительных критерия. Как было сказано выше при описании предметной области, специалист может задавать спецификации понятий, онтологий, структур данных, а также спецификации функций (методов) и процессов. Последнее особенно важно в контексте решения задач, т.к. обработка данных с помощью различных средств является неотъемлемой частью процесса решения задачи. Таким образом, возникает критерий достаточной функциональности (Functionality, “F”).

Вторым важным аспектом при решении задач, является расширяемость архитектуры. В настоящий период существует множество инфраструктур решения задач: веб-сервисы, семантический-веб, виртуальные обсерватории в астрономии, грид-инфраструктуры. Последние представляют особую важность, т.к. различные грид-инфраструктуры обладают огромными ресурсами. Поэтому важным аспектом является возможность создания гибридных инфраструктур (Hybrid, “H”), иначе говоря, возможность встраивания средств интеграции в существующие инфраструктуры решения задач.

Последним аспектом при решении задач, является наличие удобных в использовании инструментов для работы со средством интеграции. В контексте решения задач, основной предполагаемой аудиторией пользователей средством интеграции являются специалисты в некоторой прикладной области. В астрономии – астрономы, в биологии – биологи, и т.д. Надеяться на то что, специалисты в своей области окажутся еще и хорошими ИТ специалистами не представляется возможным, поэтому наличие инструментов (Tools, “T”) с хорошим пользовательским интерфейсом является важным критерием.

Таким образом, для сравнения были выделены 7 критериев: Approach, Unification, Selection, Explicit Semantic, Functionality, Hybrid, Tools – AUSEFHT.

3 Выбор средств интеграции для сравнения

Выбор проектов для сравнения, учитывая их число и разнообразие, является непростой задачей. Прежде всего, в данной работе интересны те проекты, которые можно было бы достаточно содержательно сопоставлять с проектом SYNTHESIS [15-19] разрабатываемым в ИПИ РАН. С другой стороны, интерес представляют и отличные от проекта SYNTHESIS работы, отличающиеся оригинальными решениями. В работе были рассмотрены следующие проекты: LAV – Agora [20, 21], Infomaster [22, 23], SIRUP [24–26], PICSEL [27–29], Information Manifold [30]; GAV – MedMaker [31], MOMIS [32, 33], TSIMMIS [34, 35]; BAV – AutoMed [36–39]; GLAV – SYNTHESIS. Основное внимание уделено LAV сводимым проектам (LAV, BAV, GLAV), но также рассмотрены и некоторые известные GAV проекты.

4 Результаты сравнения

На таблице 1 видно, какой прототип соответствует описанным критериям. Помимо проекта SYNTHESIS, разрабатываемого в ИПИ РАН, выделены пять LAV – ориентированных проектов: AutoMed, Infomaster, Information Manifold, PICSEL, SIRUP. AutoMed основан на подходе, известном как both as view (BAV). BAV обеспечивает отображение между схемами в обоих направлениях, что немало важно для P2P приложений или веб-сервисов. В качестве языка запросов используется Automed

Intermediate Query Language (AIQL). Благодаря отображению этих моделей в модель, основанную на гиперграфе, названную HDM [9], AutoMed позволяет осуществлять трансляцию схем между различными моделями. К настоящему времени описано отображение следующих моделей: реляционной, XML, ER, UML, ORM. Несоответствие критерию S не столько существенно, а вот отсутствие поддержки функций, является существенным недостатком.

PICSEL – подход, ориентированный на интеграцию сервисов. Все информационные источники и сервисы обработки информации представляются в виде сервисов. В качестве языка запросов используется расширенный язык дескриптивной логики CARIN (разработанный в проекте Information Manifold). В проекте реализованы возможности автоматической генерации описания сервисов по их спецификациям на языке WSDL. Главными недос-

татком проекта является отсутствие возможности расширения. Инструменты будут работать лишь с теми сервисами, для которых заранее были написаны адаптеры. Вопрос о возможности интеграции данного прототипа в различные инфраструктуры авторами рассмотрен не был.

Проект SIRUP, соответствует всем критериям предложенным в [7], но имеет два существенных недостатка: отсутствие поддержки функций и возможности расширения при желании интегрировать ресурсы с новыми информационными моделями. Основными же достоинствами проекта является ориентация на онтологии. Язык описания, названный IConcept, позволяет связывать каждое описываемое понятие с онтологическими понятиями. Кроме того разработан АПИ, позволяющий осуществлять запросы к онтологиям на языке AIQL (SQL-like).

Таблица 1. Результат соответствия рассмотренных прототипов выбранным критериям

Название проекта	A	U	S	E	F	H	T
Agora	LAV	No	No	No	No	No	Yes
AutoMed	BAV	Yes	No	Partially	No	Yes	Yes
Infomaster	LAV	No	No	No	No	No	No
SYNTHESIS	GLAV	Yes	Yes	Yes	Yes	Yes	Yes
PICSEL	LAV	No	Yes	Yes	Yes	No	Yes
SIRUP	LAV	No	Yes	Yes	No	Yes	Yes
Information Manifold	LAV	No	No	No	No	No	Yes
MedMaker	GAV	No	Yes	Partially	No	No	Yes

5 Заключение

В работе рассмотрены основные принципы для построения системы интеграции неоднородных ресурсов в контексте решения научных задач. На основе этих принципов были сформулированы 7 критериев для сравнения различных прототипов средств интеграции.

Хотя единственным проектом, удовлетворяющим всем выбранным критериям, является проект SYNTHESIS, другие аспекты, такие как, например, поддержка интеграции сервисов, генерация адаптеров, декларативная спецификация посредников и т.д., в других проектах проработаны лучше. Поэтому планируется дальнейшее расширение системы критериев.

Литература

- [1] <http://www.ifi.unizh.ch/~piegler/IntegrationProjects.html>.
- [2] Patrick Ziegler, Klaus R. Dittrich. Data Integration – Problems, Approaches, and Perspectives. In John Krogstie, Andreas L. Opdahl, and Sjaak Brinkkemper, editors, Conceptual Modelling in Information Systems Engineering, pages 39–58. Springer, Berlin, 2007.

- [3] Jeffrey D. Ullman. Information Integration Using Logical Views. In Proc. of the 6th Int. Conf. on Database Theory (ICDT'97), 1997.
- [4] Alon Y. Halevy. Answering Queries Using Views: A Survey. VLDB Journal, 10(4), 2001.
- [5] Marc Friedman, Alon Levy, and Todd Millstein. Navigational Plans for Data Integration. In Proceedings of the National Conference on Artificial Intelligence (AAAI), 1999.
- [6] P.J. McBrien, A. Poulouvasilis. Data Integration by Bi-Directional Schema Transformation Rules. In Proceedings of ICDE03, IEEE, Pages 227–238, 2003.
- [7] Patrick Ziegler. User-Specific Semantic Integration of Heterogeneous Data: What Remains to be Done? Technical Report ifi-2004.01, Department of Informatics, University of Zurich, 2004.
- [8] Briukhov D. O., Kalinichenko L. A. Component-based information systems development tool supporting the SYNTHESIS design method // 2nd East-European Conference “Advances in Databases and Information Systems” Proceedings. – Berlin-Heidelberg:Springer-Verlag, 1998. P. 305–327.
- [9] Briukhov D.O., Kalinichenko L. A., Skvortsov N. A. Information sources registration at a subject mediator as compositional development // East-European Conference “Advances in Databases and

- Information Systems” Proceedings. Lithuania, Vilnius, Springer, LNCS No. 2151, 2001.
- [10] Abrial J.-R. B-technology: Technical overview. – B-Core (UK) Ltd., 1993.
 - [11] Abrial J.-R. The B-book: Assigning programs to meanings. — Cambridge: Cambridge University Press, 1996.
 - [12] The B-toolkit online documentation. <http://www.bcore.com/ONLINEDOC/BToolkit.html>.
 - [13] Zakharov V.N., Kalinichenko L.A., Sokolov I.A., Stupnikov S.A. Development of Canonical Information Models for Integrated Information Systems. Informatics and Applications. – 2007. – V. 1, Issue 2. – P. 15–38. (In Russian)
 - [14] Kalinichenko L. A., Stupnikov S. A., Martynov D. O. SYNTHESIS: A language for canonical information modeling and mediator definition for problem solving in heterogeneous information resource environments. – M.: IPI RAS, 2007. 171 p.
 - [15] Briukhov D.O., Kalinichenko L.A., Martynov D.O. Source Registration and Query Rewriting Applying LAV/GLAV Techniques in a Typed Subject Mediator. Proc. of the Ninth Russian Conference on Digital Libraries RCDL'2007.
 - [16] Briukhov D.O., Vovchenko A. E., Zakharov V.N., Zhelenkova O. P., Kalinichenko L.A., Martynov D.O., Skvortsov N.A., Stupnikov S.A. The Middleware Architecture of the Subject Mediators for Problem Solving over a Set of Integrated Heterogeneous Distributed Information Resources in the Hybrid Grid-Infrastructure of Virtual Observatories Informatics and Applications. – 2008. – V. 2, Issue 1. – P. 2–34.
 - [17] Kalinichenko L.A., Briukhov D.O., Martynov D.O., Skvortsov N.A., Stupnikov S.A. Mediation Framework for Enterprise Information System Infrastructures. Proc. of the 9th International Conference on Enterprise Information Systems ICEIS 2007. – Funchal, 2007. – Volume Databases and Information Systems Integration. – P. 246–251.
 - [18] Kalinichenko L.A., Stupnikov S.A., Martynov D.O. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007.
 - [19] Zakharov V.N., Kalinichenko L.A., Sokolov I.A., Stupnikov S.A. Development of Canonical Information Models for Integrated Information Systems. Informatics and Applications. – 2007. – V. 1, Issue 2. – P. 15–38.
 - [20] Ioana Manolescu, Daniela Florescu, Donald Kossmann. Answering XML Queries over Heterogeneous Data Sources. Proc. of the Int'l. Conf. on Very Large Databases (VLDB) 2001, Roma, Italy,.
 - [21] Ioana Manolescu, Daniela Florescu, Donald Kossmann, Florian Xhumari, Dan Olteanu. Agora: Living with XML and Relational. Proc. of the Int'l. Conf. on Very Large Databases (VLDB) 2000, Cairo, Egypt.
 - [22] Michael R. Genesereth, Arthur M. Keller, Oliver Duschka. Infomaster: An Information Integration System. in proceedings of 1997 ACM SIGMOD Conference, May 1997.
 - [23] Oliver M. Duschka, Michael R. Genesereth. Infomaster – An Information Integration Tool. In proceedings of the International Workshop "Intelligent Information Integration" during the 21st German Annual Conference on Artificial Intelligence, KI-97. Freiburg, Germany, September 1997.
 - [24] Patrick Ziegler. Evaluation of SIRUP with the SIRUP Classification of Data Integration Conflicts. Technical Report ifi-2007.07, Department of Informatics, University of Zurich, 2007.
 - [25] Patrick Ziegler, Christoph Sturm, Klaus R. Dittrich. Unified Querying of Ontology Languages with the SIRUP Ontology Query API. In 11. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2005), volume P-65 of Lecture Notes in Informatics, pages 325–344, Karlsruhe, Germany, March 2–4, 2005.
 - [26] Patrick Ziegler, Klaus R. Dittrich. User-Specific Semantic Integration of Heterogeneous Data: The SIRUP Approach. In Mokrane Bouzeghoub, Carole Goble, Vipul Kashyap, and Stefano Spaccapietra, editors, First International IFIP Conference on Semantics of a Networked World (ICSNW 2004). volume 3226 of Lecture Notes in Computer Science, pages 44–64, Paris, France, June 17–19, 2004. Springer.
 - [27] Chantal Reynaud, Gloria Giraldo. An application of the mediator approach to services over the Web. Concurrent Engineering, 2003.
 - [28] Chantal Reynaud. Building scalable mediator systems. Topical Day in Semantic Integration of Heterogeneous Data, IFIP World Computer Congress, 2004.
 - [29] Marie-Christine Rousset, Chantal Reynaud. Pícel and Xyleme: two illustrative information integration agents. Book chapter in Intelligent Information Agents Research and Development in Europe, Springer-Verlag, 2003.
 - [30] Alon Levy, Anand Rajaraman, Joann Ordille. Querying Heterogeneous Information Sources Using Source Descriptions. Proceedings of the Twenty-second International Conference on VLDB 1996.
 - [31] Yannis Papakonstantinou, Hector Garcia-Molina, Jeffrey Ullman. MedMaker: A Mediation System Based on Declarative Specifications. Proceedings of the 12th International Conference on Data Engineering, 1995.
 - [32] D. Beneventano, S. Bergamaschi. The MOMIS Methodology for Integrating Heterogeneous Data Sources. IFIP World Computer Congress. Toulouse France, 22–27 August 2004.
 - [33] D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini. Building a Tourism Information Provider

with the MOMIS System. Information Technology & Tourism Journal, 2005.

- [34] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, J. Widom. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In Proceedings of IPSJ Conference, pp. 7–18, Tokyo, Japan, October 1994.
- [35] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, V. Vassalos, J. Widom. The TSIMMIS approach to mediation: Data models and Languages. In Journal of Intelligent Information Systems, 1997.
- [36] M. Boyd, P.J. McBrien. Comparing and Transforming Between Data Models via an Intermediate Hypergraph Data Model. Journal on Data Semantics IV, Pages 69–109, Springer-Verlag, 2005.
- [37] M. Boyd, S. Kittivoravitkul, C. Lazanitis, P.J. McBrien, N. Rizopoulos. AutoMed: A BAV Data Integration System for Heterogeneous Data Sources. In Proceedings of CAiSE04, Springer Verlag LNCS Vol 3084, Pages 82–97, 2004.
- [38] L. Zamboulis, A. Poullovassilis. Using AutoMed for XML Data Transformation and Integration. In Proceedings of DIWeb'04, CAiSE Workshop Proceedings Volume 3, Pages 58–69.
- [39] L. Zamboulis, N. Martin, A. Poullovassilis. A Uniform Approach to Workflow and Data Integration. In Proceedings of U.K. e-Science All Hands Conference, September, 2007.

Analysis and comparison of systems for heterogeneous information resources integration

A.E. Vovchenko, L.A. Kalinichenko

Two approaches for problem solving in information systems over multiple heterogeneous integrated information resources are distinguished: resource driven and problem driven. In the paper the second approach is emphasized. One of the obstacles for such approach remaining to be unresolved consists in the integrated representation of multiple heterogeneous information resources for a researcher solving the problem in various infrastructures (such as grid architectures, Web services, Semantic Web, etc.). In the paper the basic principles of heterogeneous information resources integration are considered. These principles are used to make a choice of criteria for comparison of several well known integration system prototypes.

* Работа выполнена при частичной финансовой поддержке РФФИ (гранты 06-07-08072-офи-а, 06-07-89188-а, 08-07-00157-а) и программы ОНТИ РАН «Фундаментальные основы информационных технологий и систем» (проект 1-10).

Электронная система на основе MS WSS для гармонизации терминологии аналитической химии*

© В.И. Широкова, В.П. Колотов, М.В. Аленина

Институт геохимии и аналитической химии им. В.И.Вернадского РАН,
119991, ГСП-1, ул. Косыгина, 19
shirokova@geokhi.ru

Аннотация

Для сообщества российских химиков-аналитиков создана электронная система на основе MS WSS (Windows SharePoint Services), объединяющая географически удаленных пользователей, для совместной работы. Разработанная система интерактивных сайтов предназначена для обеспечения задач как постоянных подразделений Научного совета по аналитической химии (комиссий, отделений), так и временных рабочих групп, создаваемых для решения различных общих задач (например, таких как создание глоссариев, предметных баз данных, организации конференций и др.). Начата работа по созданию онтологии аналитической химии. В качестве первого шага поставлена задача разработки электронного глоссария терминологии. Имея в виду множество существующих источников терминов (ГОСТы, рекомендации Международного союза по теоретической и прикладной химии, системы и др.) и существующие разночтения в их трактовке, ведется работа по созданию всеобъемлющей базы данных терминов по аналитической химии из разных источников для их сопоставления, выявление противоречий или разночтений. Очевидно, что обойтись без публичного обсуждения обнаруженных разночтений и выработки консенсуса профессиональным сообществом химиков-аналитиков невозможно. Функциональность развернутых WSS сайтов попадает точно в цель поставленной задачи. С целью систематизации имеющихся рекомендаций проведен критический анализ структур дескрипторов XML-данных, разработанных различными международными организациями (IUPAC, ASTM и др.) для представления химической и аналитической информации.

Введение

Опыт эксплуатации российского интернет-портала по аналитической химии [1], показал определенные ограничения, свойственные традиционным сайтам, администрируемым централизованно. Так например, НСАХ РАН состоит из многочисленных комиссий по направлениям аналитической химии, научных ассоциаций и семинаров, региональных бюро и др. подразделений, имеющих свои задачи. Их эффективная работа подразумевает наличие способов прямой публикации/распространения информации в Интернете, обмена сведениями по актуальным проблемам и др. Таким образом, в профессиональной среде химиков-аналитиков назрела необходимость рациональной интеграции центрального портала с большим количеством сайтов, обеспечивающих децентрализованную поставку и интерактивное управление узкоспециальной информацией с рабочего места пользователя. Для этого была развернута электронная интернет-среда, основанная на использовании службы MS WSS v.2 (Server 2003). Задачей электронной системы является поддержка работы Научного Совета РАН по аналитической химии (НСАХ РАН) и организация эффективной совместной работы членов совета в рамках различных проектов по аналитической химии (подготовка книг, предметных баз данных, проведение конференций, глоссария по аналитической химии для обеспечения «семантического» Интернета, выполнение планов комиссиями и отделениями совета и др.).

Развернутые интерактивные WSS-сайты НСАХ РАН

Понятие «интерактивный интернет-сайт» подразумевает, что пользователи такого сайта, в соответствии с наделенными правами, могут самостоятельно размещать на них разнообразную информацию (новости, объявления, документы, организовывать и поддерживать форумы и др.) пользуясь привычным интернет браузером. Вместе с тем, вся размещаемая информация на таких сайтах и центральном портале интегрирована, например, для того, чтобы обеспечить работу поисковой системы или обеспечить ав-

томатизированную поставку новостей на центральный портал. Задачи поддержания целостности такого централизованного хранилища информации возлагаются на разработанное программное обеспечение.

MS WSS (v.2) отличается гибкостью настройки, многоязычным интерфейсом, тесной интеграцией с MS SQL-сервером, широкой поддержкой возможностей языка XML. Сайты на основе WSS используют развитую технологию активных серверных страниц (ASP.NET) и допускают удаленное управление и программирование с использованием XML-протокола для обмена информацией в распределенных системах. Последняя версия службы (v.3) отличается заметно более развитыми особенностями (наличие master-страниц, поддержкой ведения сетевых дневников – блогов и др.).

Для выполнения этой задачи использован двух-процессорный сервер (Xeon 1.7 МГц, отказоустойчивый дисковый массив RAID 1 на основе SCSI дисков), операционная система Windows Server2003 (Standard Edition), включая Интернет-сервер (IIS v.6.0) и сервер баз данных (MS SQL Server2000 Standard Edition SP3). Установлена служба WSS v.2 (SP2) в конфигурации Server Farm, обеспечивающая масштабируемость приложения. Так как новый сервер вошел в состав уже существующего домена, то конфигурация Server Farm, позволяет использовать один и тот же контент (базу данных) для нескольких виртуальных серверов. Такой сценарий может быть применен, например, для параллельной работы в интернет/интранет сети. После установки WSS, на IIS был создан виртуальный сервер верхнего уровня Научного Совета по аналитической химии РАН, который был расширен для использования WSS. Этому виртуальному серверу была назначена определенная база данных.

Реализованная архитектура системы подразумевает, что созданный корневой сайт является родительским для всех остальных сайтов системы. Далее с использованием администрирования корневого сайта созданы дочерние сайты для всех подразделений НСАХ РАН и временных групп, а также «входной» wss-сайт, который предоставляет список всех дочерних сайтов, включая соответствующие ссылки [2]. Развернутая система включает более 40 WSS-сайтов различного назначения.

Для автоматизации развертывания типовых сайтов с предопределенной функциональностью и структурой были разработаны различные шаблоны для сайтов (например, комиссий, отделений, конференций, работы над книгами и глоссариями). С использованием шаблонов были созданы сайты для всех подразделений совета (см. входной wss-сайт совета [2]). Генерация одного сайта по шаблону занимает менее минуты, а сам сайт обладает достаточно продвинутой функциональностью: размещение документов с сортировкой их по различным темам, размещение новостей, объявлений, интернет-линков, организация тематических интернет-форумов, поддержка списка контактов группы, ор-

ганизация совместной работы (например, задания, контроль их исполнения) и др.

Очевидно, что интерактивный сайт обеспечивает большую открытость в плане размещения информации, что автоматически означает и большую уязвимость для несанкционированного использования. Для этого разработана стратегия децентрализованного администрирования системы с целью обеспечения устойчивого и надежного ее функционирования. Корневой WSS сайт администрируется исключительно разработчиками портала, а каждый дочерний сайт имеет двух администраторов: один является администратором сервера, а второй представляет соответствующую удаленную группу. Функционирование этой системы основано на ролевой системе безопасности, когда каждый сайт (независимо от соподчиненности) имеет свою независимую систему учетных записей, назначающую посетителям определенные права. При попытке изменить информацию сайта (базы данных) система запрашивает учетное имя и пароль и далее проверяет возможность выполнения планируемого изменения в рамках назначенной роли. Анонимный доступ разрешает лишь просматривать информацию и документы. Администратор группы назначает роли посетителям дочернего сайта (Web-мастер, автор или читатель), а также, в случае необходимости (и наличия навыков) позволяет доступ к общей настройке представления информации на дочернем сайте. Аутентификация посетителей для доступа к функциям управления сайтом проводится средствами безопасности ОС сервера, для чего в системе учетных записей создано соответствующее организационное подразделение (ou). Для аутентификации администраторов всех уровней был назначен SSL (Secure Socket Layer) протокол, обеспечивающий шифрование и безопасную передачу данных через брандмауэр, что особенно важно для удаленного администрирования WSS. Для обычных пользователей (анонимных или ролевых) установлена базовая аутентификация, при использовании которой никакого шифрования информационных потоков не проводится. Кроме того, для каждого виртуального сервера назначен изолированный пул приложения (отдельный процесс операционной системы), что потенциально обеспечивает большую отказоустойчивость. В автоматическом режиме организовано резервное копирование содержания всех сайтов (синхронно с резервным копированием базы данных).

В соответствии с архитектурой WSS, информация всех сайтов хранится в единой базе данных, содержащей десятки реляционно-связанных таблиц и большое число хранимых процедур, обеспечивающих целостность данных. Таким образом, интеграция информации в пределах виртуального сервера существует как неотъемлемое свойство службы WSS.

Одной из целей проекта является интеграция разнородных данных или, другими словами, разработки хранилища информации (Data Warehouse)

интернет-портала и WSS-сайтов. При разработке концепции хранилища руководствовались следующими принципами: данные должны быть предметно ориентированными, должны быть маркированы во времени, должны быть потенциально интегрируемы (поддержка отношений). Процесс построения хранилища включал ряд этапов: доступ к гетерогенным источникам данных (включая ресурсы Интернета), их преобразование и размещение в хранилище; создание метабазы, описывающей смысл информации, помещенной в хранилище. Практика объединения разнородной информации показала перспективность использования объектно-ориентированного подхода. Для практической реализации упрощенного хранилища данных в контексте проводимой работы, мы опирались на имеющиеся аппаратные и программные возможности. Использовали язык описания данных XML, обеспечивающий межплатформенную независимость. Удобство использования XML состоит также в том, что SQL-сервер обеспечивает удобное взаимодействие с информацией такого рода. Данные интернет-портала, размещенные на другом SQL-сервере, были представлены в виде XML-структур, проиндексированы и предоставлены SQL-серверу, обслуживающему сайты WSS. Для синхронизации связанных данных, размещаемых на разных серверах, использовали триггеры (программы SQL-сервера, вызываемые при наступлении определенных событий, таких как загрузка, обновление или удаление данных). Таким образом, ключевая информация, размещаемая на многочисленных WSS-сайтах, мгновенно интегрируется и отображается на головном портале, использующем обычную ASP.NET архитектуру.

Результаты эксплуатации развернутых электронных систем показали высокую эффективность разработанных решений, положенных в их основу.

Гармонизация терминологии по аналитической химии

Одной из главных задач, на решение которой нацелена развернутая система, является обеспечение работ по различным общим проектам совета, которые подразумевают вовлечение ученых разных специальностей. Актуальным проектом является гармонизация терминологии аналитической химии, в том числе и в связи с активно обсуждаемой концепцией семантического Интернета, подразумевающей размещение информационных ресурсов, содержащих структурированную и формализованную информацию, «понятную» компьютерам [3]. Концепция предполагает наличие в сети предметных онтологий (по сути словарей дефиниций, понятий и терминов той или иной области знания) и увязывание размещаемых в Интернете материалов с этими онтологиями с помощью содержательных дескрипторов XML (вместо HTML). Онтологии, в том числе и по аналитической химии, должны создаваться сообществом специалистов. В этой связи использование возможностей WSS-сайтов является, пожалуй,

оптимальнейшим способом организации такой работы.

Специфика аналитической химии состоит в том, что многие ее методы возникли на стыке с другими науками, что привело к определенному взаимовлиянию терминов, появлению терминов-синонимов, трактуемых по-разному в различных областях аналитической химии. Это относится как к англоязычной, так и к русскоязычной терминологии.

В рамках международного сообщества под руководством Международного союза по теоретической и прикладной химии (ИЮПАК) ведется работа как по систематизации терминологии в различных областях химии, так и по ее гармонизации (единству интерпретации синонимов). ИЮПАКом издаются рекомендации по терминологии различных химических дисциплин, в том числе и аналитической химии [4]. В 2004 г опубликован «Русско-английский и англо-русский словарь терминов по аналитической химии» [5], обобщивший все предыдущие публикации. С января 2006 г. введен ГОСТ Р 52361-2005 «Контроль объекта аналитический. Термины и определения», содержащий определения основных терминов аналитической химии [6]. Периодически публикуются переработки и дополнения терминологии. Тем не менее, противоречия в терминологии остаются.

Например, до сих пор перевод такого ключевого термина аналитической химии как «ассигасу» не однозначен. Его трактуют и как «точность» результатов анализа (близость к опорному значению) или как «правильность». А «точность» часто приравнивают к термину «precision». В свою очередь, «precision» в ряде рекомендаций переводится как «воспроизводимость», которая также рекомендуется как «reproducibility»! И подобных примеров можно привести множество. Очевидно, что такая ситуация неприемлема. Заметным пробелом в решении этой проблемы является отсутствие единого электронного источника терминологии по аналитической химии, включающего и ее трактовку на русском языке и английском языках. Этот пробел несомненно, ощущается как научным, так и образовательным сообществами.

Разрешение неточностей и противоречий формулировок терминов, устранение различного рода невязок является лишь начальным шагом гармонизации терминологии, к сожалению, обычно и единственным. Упускается важнейшая особенность терминологии как языка, представляющего сущности той или иной науки и взаимосвязи этих сущностей. Как наука развивается от простого к сложному, так и терминология отражает этот путь, определяя более сложные понятия через существующие более простые. Глоссарий есть список терминов и их определений (дефиниций). Роль последних состоит в толковании терминов. Однако, очень часто толкование терминов делается относительно произвольно, не используя непосредственно термины более общего типа, а применяя другие определения, в лучшем случае близкие синонимы. Иерархию (систему)

терминов в той или иной области, конечно, должен определять эксперт-профессионал, а соответствие дефиниций предложенной системе может оценить специалист по информационным технологиям, оценивая присутствие требуемых базовых терминов в тексте определений.

Таким образом, другая сторона гармонизации терминологии заключается в формировании компьютерно-ориентированной иерархии терминов, устанавливая между ними семантически прослеживаемые материнско-дочерние отношения, свободные от противоречий.

Решение этой задачи является заметным шагом к построению онтологии, которая отличается более высокой степенью абстракции, «склеивая» терминологию в систему знаний (используя понятия объектно-ориентированно программирования, когда каждый объект наделяется наследуемыми свойствами и методами функционирования). Результирующая онтология, размещенная в качестве электронного ресурса в сети, является ключевым узлом семантической сети.

База данных глоссария

Работа по созданию онтологии аналитической химии начата сообществом химиков-аналитиков в рамках НСАХ РАН.

В качестве первого шага решается задача разработки электронного глоссария терминологии. Структура реляционной базы данных электронного глоссария представлена на рис. 1.

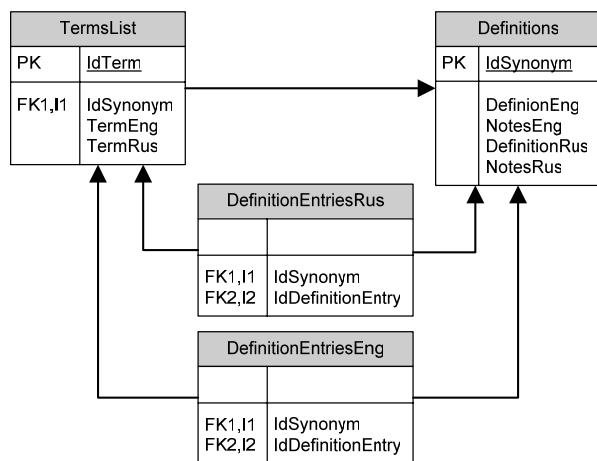


Рис. 1. Структура реляционной базы данных глоссария (PK и FK – первичные и внешние ключи для обеспечения целостности данных и их каскадного изменения/удаления)

Имеется всего две основные таблицы: общий список терминов (TermsList) и список определений терминов (Definitions). Таблица TermsList содержит единственное ключевое поле идентификатора записи (целое число), идентификатор семантической идентичности терминов и собственно термин (на двух языках). Следует отметить важность второго поля. Например, в Международном глоссарии [10]

следующие термины представлены в единой семантической группе с общим определением: «measured quantity value»≡ «measured value of a quantity»≡ «measured value»= quantity value representing a measurement result. В этом случае фрагмент таблицы TermsList выглядит так:

IdTerm	IdSynonym	TermEng	...
...	...	...	...
123	123	measured value	...
124	123	measured value of a quantity	...
125	123	measured quantity value	...

Значение поля IdSynonym будет идентично для всех семантических форм термина и равно идентификатору вхождения первой (обычно наиболее распространенной) формы. Общее определение для всего семантической группы представлено в таблице Definitions, где ключевым полем является идентификатор семантической группы. Такая структура базы данных обеспечивает гибкие возможности для расширения семантических форм и эффективна для оценки семантической идентичности при поиске информации. Две другие таблицы “DefinitionEntries”Eng/Rus” (для англо- и русскоязычной терминологии) предназначены для кодировки вхождения базовых (родительских) терминов в определения более общих (дочерних) терминов. Эти таблицы позволяют построить иерархию и отношения терминов, выявить «висячие» (вне иерархии) термины высокого порядка (заведомо не базовые), определения которых не включают термины более низкой иерархии. Наличие таких случаев может свидетельствовать о неудачно составленном определении термина. Компиляция упомянутых двух таблиц может проводиться автоматически путем последовательного поиска вхождений терминов соответствующего языка в определения терминов (для этой задачи используется хранимая процедура, которая может запускаться триггером при загрузке/изменении термина или его определении). С использованием данных таблиц “DefinitionEntries”Eng/Rus” автоматически строится древовидная структура терминологии, которая должна в основном соответствовать иерархии терминов, профессионального сообщества. Гармонизация терминологии достигается при разрешении выявленных противоречий.

База данных глоссария и WSS-списки

Таблицы на WSS сайтах называются списками. Существуют средства экспорта/импорта таких списков из баз данных (или таблиц Excel). Особенно тесное взаимодействие между списками WSS разработано для MS Access (MS Office2007), где возможен как экспорт/импорт данных, так и динамическая связь между таблицами MS Access и WSS-списками. Однако, обеспечить полноценное пред-

ставление системы реляционно связанных таблиц с помощью списков WSS невозможно (данные всех списков хранятся в единой таблице базы данных WSS). Поэтому вопросы обеспечения целостности данных, включая каскадное обновление/удаление данных, должны решаться особыми средствами.

В разработанной нами системе в качестве основной базы данных используется стандартный MS SQL2000, включая все необходимые средства поддержания целостности (для удобства работы шлюзом к базе данных является проект MS Access). Далее, на сайте WSS созданы два списка, отвечающие основным таблицам глоссария (TermsList и Definitions). Список Definitions включает поле подстановки из TermsList (IdSynonym) для обеспечения ограниченной целостности данных, а TermsList (как любой список) имеет уникальный идентификатор записи.

Как уже упоминалось данные всех списков хранятся в общей таблице WSS, которая называется UserData (или AllUserData в случае WSS v.3), а также ряде вспомогательных таблиц, содержащих идентификаторы списков. С помощью запросов данные из базы данных глоссария экспортируются в базу данных WSS. В результате наполненные списки становятся доступными в Интернете. Специалисты НСАХ РАН могут ознакомиться с системой терминологии и отредактировать для каждой проблемной записи стандартную вложенную форму (файл Excel), где сформулировать проблему и предложения. Непосредственное редактирование полей списка запрещено. На основании предложений администратор модифицирует записи в списке, сопровождая это комментариями в той же форме. По окончании разработки глоссария проводится обновление таблиц исходной базы данных с использованием запроса к WSS-базе данных.

Заключение

В настоящее время на сайте размещен электронный глоссарий терминов ГОСТ [6]. Ведется работа по обработке терминов ИЮПАК, а также данных из других источников. Определены запросы базы данных и программы по сопоставлению терминов и выявлению противоречий или разночтений с автоматическим размещением на WSS-сайтах для публичного обсуждения, сбора и последующего учета откликов научной общественности и принятия окончательного решения.

Литература

- [1] Российский Интернет-портал «Аналитическая химия в России»: <http://www.rusanalytchem.org>
- [2] WSS-сайты НСАХ РАН: <http://www.rusanalytchem.org/wss>
- [3] Berners Lee T. Semantic Web Road Map: <http://www.w3.org/DesignIssues/Semantic.html>
- [4] Compendium of Analytical Nomenclature: Definitive rules 1997. Orange Book. 3rd edition Inczedy, J.; Lengyel, T. and Ure, A.M. Blackwell Science, 1998 [ISBN 0-86542-6155]. On-line version: http://www.iupac.org/publications/analytical_comp_endium
- [5] Русско-английский и англо-русский словарь терминов по аналитической химии / сост. и ред. проф. В.М. Иванов, акад. Ю.А. Золотов. – М. : Изд-во Лаб-пресс, 2004. – 192 с. [ISBN 5-902658-01-2].
- [6] <http://194.67.119.6:83/glossary/>
- [7] What is AnIML?: <http://animl.sourceforge.net/>
- [8] Standard XML data dictionaries for chemistry.: <http://www.iupac.org/projects/2002/2002-022-1-024.html>
- [9] About the IUPAC Compendium of Chemical Terminology (Gold Book): <http://goldbook.iupac.org/about.html>
- [10] International Vocabulary of Metrology – Basic and General Concepts and Associated Terms, VIM 3rd edition, JCGM/WG 2, Document N341. 2007.

Electronic System based on MS WSS for Harmonization of Terminology in Analytical Chemistry

V.I. Shirokova, V.P. Kolotov and M.V. Alenina

An electronic media for Russian analysts has been developed on the base of MS WSS (Windows SharePoint Services). The media joins geographically remote users for making of common projects. The developed system consists of interactive Internet sites intended both for supporting of needs of the permanent divisions of the Scientific Council on Analytical Chemistry (commissions, divisions) and temporary working groups, which are creating for solution of various common tasks (compilation of glossaries, subject data bases, organization of conferences, etc.). One of the most urgent tasks consists in development of ontology in analytical chemistry. This work has been started recently. Development of electronic glossary of terminology (Russian/English) has been selected as the first step. Taking into account that now various official sources of terms are available (Russian standards, IUPAC recommendations, etc.) which often contradict each other, it was decided to create a comprehensive data base covering all sources. On the next step of the work the terms and their definitions should be compared to reveal conflicting variants of reading. The final phase implies elaboration of consensus of the whole professional community of the Russian analysts concerning the terms and definitions. It is clear that just public discussions via common electronic media will allow resolve such great task. It has been proved that functionality of the deployed WSS sites exactly suites the formulated tasks.

For systematization of the recommendations issued by different International organizations (IUPAC, ASTM, others) a critical analysis of the XML-data descriptors for representation of analytical chemistry information has been done too.

* Работа проводится при поддержке Российского фонда
фундаментальных исследований (грант № 08-03-00893).

Формализм записи лексико-синтаксических шаблонов в задаче автоматизации процесса построения онтологий *

© Е.А. Рабчевский, Г.И. Булатова, И.М. Шарафутдинов

Пермский государственный университет
evgeny@ranat.ru

Аннотация

Рассматривается использование лексико-синтаксических шаблонов для автоматизации построения семантических моделей, соответствующих тексту. Обсуждаются шаблоны, результаты их применения, а также методы выявления новых шаблонов, опубликованные в работах других авторов. Формулируется оригинальный подход к формализации лексико-синтаксических шаблонов. Разработана XML Схема языка для формализации шаблонов. Разрабатывается программа для выявления синтаксических групп на основе синтаксического анализатора Dictum. На основе шаблонов решается задача построения таксономии понятий.

1 Введение

Онтологии активно используются научными сообществами для описания терминологии, в электронной коммерции – для описания товаров и услуг, в других приложениях в качестве баз знаний интеллектуальных систем, в Интернет – для организации данных. Поисковой системой SWOOGLE на сегодня проиндексировано свыше 10 тысяч онтологий и словарей, доступных в Веб. Все больше Интернет ресурсов внедряют семантические модели в свою архитектуру. Появляется все больше промышленных средств для работы с онтологиями.

Существует достаточное количество технических средств для создания и редактирования онтологий, однако создание онтологий, помимо знаний в инженерии знаний, требует от автора экспертных знаний в исследуемой предметной области. Поэтому задача автоматизации процесса построения онтологий является весьма актуальной.

В Интернет можно найти описательную часть большинства предметных областей человеческой деятельности. В основном этот контент представлен

текстами на естественном языке. Поэтому задача формирования онтологий из текстового содержания Веб ресурсов представляется наиболее актуальной.

2 Лексико-синтаксические шаблоны

Лексико-синтаксические шаблоны позволяют построить семантическую конструкцию, которая соответствует концептуальному содержанию единицы текста. Для этого используются особенности языка, на котором представлен текст.

Марти Хэрст выявила существенное количество шаблонов для идентификации отношения гипонимии [1]. Её исследования показали, что, используя шаблоны на большом корпусе текстов определенной тематики, можно построить «достаточно адекватную» таксономию понятий соответствующей предметной области. В её шаблонах в качестве элементов используются, например, понятие именной группы, знаки препинания, конкретные слова и другое.

Так, например шаблон «NP {, NP}* {,} and other NP», где NP – условное обозначение именной группы, идентифицирует отношение гипонимии, которое продемонстрировано на части предложения «...temples, treasures, and other important civic buildings...». С помощью указанного шаблона из данного фрагмента текста могут быть выявлены следующие отношения hyponym("temple", "civic building"), hyponym("treasure", "civic building").

Большакова и др. сформулировали язык для записи лексико-синтаксических шаблонов [2]. По их мнению, элементами шаблонов могут быть:

- литералы, т.е. конкретные лексемы;
- определенные части речи;
- определенные грамматические конструкции;
- условия, уточняющие грамматические характеристики рассмотренных элементов.

Важно отметить, что исследования Большаковой направлены на выявление определений «новых» терминов в научной литературе (до сих пор неизвестных интеллектуальной системе), то есть соответствующие шаблоны ориентированы на ограниченный тип отношений.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

3 Подходы к автоматизации выявления шаблонов

В работе Хэрст [3] предложен способ выявления новых шаблонов. Метод состоит из пяти этапов:

- выбирается некое отношение R, например гипонимии;
- подбирается список терминов, для которых заранее известно, что они участвуют в выбранном отношении с другими терминами (в качестве обучающей выборки был использован тезаурус WordNet);
- в исследуемом корпусе находятся места, в которых появляются выбранные термины, участвующие в заданном отношении, «их окружение» фиксируется;
- найденные «окружения» анализируются, и формулируется новый шаблон для заданного отношения.

Недостатком данного метода является недостаточная формализация понятия окружения, а также выполнение 4-го этапа без использования средств автоматизации.

В работе сотрудников Шеффилдского Университета [4] сделана попытка автоматизации формулировки шаблонов из «окружения» терминов, состоящих в заданном отношении. В основе их работы лежит метод машинного обучения Lazy-NLP [5]. Алгоритм оперирует с морфологией текста и лексическими категориями (например, частями речи) и также использует обучение с учителем, в ходе которого пользователь указывает верные и неверные лексикализации выбранного отношения.

4 Формализация шаблонов

В контексте задачи автоматизации построения онтологий, обозначенной во введении, интересно использование шаблонов для выявления онтологических конструкций.

Предыдущие исследования [6] показали, что использование шаблонов может быть эффективным для построения отношений, использующихся в онтологиях.

Для записи лексико-синтаксических шаблонов авторы предлагают использовать унифицированный язык, который оперирует с понятиями графематики, морфологии и синтаксиса.

Для унификации записи шаблонов предлагается описывать шаблон с помощью входной и выходной схем. Входная схема позволяет идентифицировать шаблон в тексте, выходная — описывает семантическую конструкцию, которая соответствует концептуальному содержанию исследуемого текста.

В рамках данного исследования был произведен анализ шаблонов других авторов, а также сформулированы оригинальные лексико-синтаксические шаблоны. В результате анализа общего объема лексико-синтаксических шаблонов авторы пришли к выводу, что в задаче выявления семантических мо-

делей онтологического характера, элементами входной схемы шаблона должны быть:

- литерал
- синтаксическая единица, такая как:
 - слово, являющееся определенной словоформой определенной леммы
 - слово определенной части речи
 - синтаксическая группа (грамматика непосредственно составляющих [7])
- синтаксические элементы могут накладываться грамматические условия

Элементы шаблона могут участвовать между собой в отношениях грамматики зависимостей [7], то есть один элемент может быть подчинен другому.

5 Синтаксическое дерево предложения и синтаксические группы

Синтаксические группы состоят из последовательности слов, входящих в часть текста. Группа характеризуется типом, который определяется типом ее главной подгруппы. Тип подгруппы определяется типом (частью речи) главного слова. Примером синтаксической группы является именная группа, группа в которой главная подгруппа — имя существительное.

Рассмотрим предложение, как дерево, в котором вершинам соответствуют слова предложения, а ребрам — подчинительные связи. И каждому ребру дерева соответствует вопрос от главного слова к зависимому. Тогда синтаксическую группу можно понимать как определенную сосредоточенную часть дерева или подграф. Тип группы определяется типом главного (то есть верхнего) слова в группе. В этом контексте, предложение можно рассматривать, как две главные группы, главными словами в которых являются подлежащее и сказуемое.

Таким образом, можно ввести понятие группы определенного слова. Это группа, главным словом в которой является заданное слово, само слово соответственно входит в группу. А группа, подчиненная определенному слову — это группа этого слова без самого этого слова, или подграф, подчиненный узлу заданного слова. Главное слово группы само по себе может интерпретироваться, как член предложения.

Очевидно, что в графе предложения несколько узлов могут иметь одного общего родителя.

6 LSPL — язык для записи шаблонов

Предлагается XML-язык для формализации шаблонов. Входная схема шаблона записывается на XML-основанном языке LSPL (Lexical-Syntactic Pattern Language — язык разметки лексико-синтаксических шаблонов).

В элементе `<inputShema>` записывается входная схема шаблона, как последовательность элементов шаблона, записываемых в тегах `<element>`. В атрибутах данного тега записываются тип элемента (literal, wordForm, partOfSpeech или syntacticGroup), id — идентификатор элемента в шаблоне по порядку,

connectedId - идентификатор элемента, которому подчиняется данный элемент, и флаг обязательности присутствия presence.

У элемента <element> также выделяется атрибут mainWord. Атрибут используется приложением только в случае, если тип элемента шаблона – синтаксическая группа. Наличие данного атрибута необходимо для установления того, в каком объеме синтаксическая группа участвует в шаблоне. Возможны случаи, когда элементом шаблона является только главное слово, вся группа вместе с главным словом, или группа без главного слова. Соответственно значения атрибута будут only (или без указания данного атрибута), included, not_included.

В элементе <content>, дочернем для <element>, указывается содержание элемента, которое для литерала означает слово, которое должно быть элементом шаблона, для словоформы – лемму, от которой берется словоформа. Либо часть речи или синтаксическая группа, которыми можно охарактеризовать элемент шаблона.

Для литерала значение тега <content> «plainText» зарезервировано под обозначение части текста, которая также участвует в шаблоне, но на нее не накладывается каких-либо лингвистических условий. Например, это может быть использовано для обозначения элемента в шаблоне, к которому должны быть применены другие шаблоны.

В элементе <grammaticalValue>, дочернем для <element>, в виде атрибутов указываются дополнительные грамматические условия, которые накладываются на элемент, такие как род, число и падеж. Как одно из возможных значений указанных атрибутов может быть указано значение переменной, которое может быть использовано прикладной программой. Тело тега может использоваться, для обозначения всех грамматических характеристик с помощью одного идентификатора.

У одного элемента <element> может быть несколько элементов <grammaticalValue>. Это позволяет формулировать шаблоны в более общей форме, например, шаблон может позволять элементу находиться в нескольких падежах.

Рассмотрим пример записи шаблона следующего вида:

[NP1] – это [NP2], [который (грамматически согласованный с NP2)] [V] [NP3 (подчиненная V)].

Где в квадратных скобках обозначены элементы шаблона, NP – именная группа, V – группа глагола.

Представим входную схему данного шаблона

```
<pattern>
<inputSchema>
  <element type="syntaxGroup" id="1">
    <content>nounPhrase</content>
  </element>
  <element type="literal" id="2">
    <content>-</content>
  </element>
  <element type="literal" id="3">
    <content>это</content>
  </element>
```

```
<element type="syntaxGroup" id="4">
  <content>nounPhrase</content>
  <grammaticalValue>var1<grammaticalValue>
</element>
<element type="literal" id="5">
  <content>,</content>
</element>
<element type="wordForm" id="6">
  <content>который</content>
  <grammaticalValue>var1<grammaticalValue>
</element>
<element type="syntaxGroup" id="7">
  <content>verbPhrase</content>
</element>
<element type="syntaxGroup" id="8">
  <content>nounPhrase</content>
</element>
</inputSchema>
</pattern>
```

В атрибуте path элемента <outputSchema> указывается ссылка на OWL файл выходной схемы шаблона, в котором указывается соответствующая семантическая модель. Либо сам OWL-код записывается в теле тега <outputSchema>. Приведем выходную схему данного шаблона.

```
...
<owl:Ontology rdf:about="" />
<owl:Class rdf:ID="element4" />
<owl:Class rdf:ID="element8" />
<owl:Class rdf:ID="element1" />
<rdfs:subClassOf rdf:resource="#element4" />
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource="#element7" />
    <owl:hasValue rdf:resource="#element8" />
  </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
<owl:ObjectProperty rdf:about="#element7">
  <rdfs:domain rdf:resource="#element4" />
</owl:ObjectProperty>
...
```

7 Сложные шаблоны и группы шаблонов

Рассмотрим несколько шаблонов:

[NP1] – это [NP2], [который (грамматически согласованный с NP2)] [V] [Vinf] [NP3 (подчиненная Vinf)].

[NP1] – это [NP2], [который (грамматически согласованный с NP2)] [V] [P].

Где Vinf – глагол в инфинитиве, P – предикатив.

Эти шаблоны и шаблон из примера в шестой главы частично совпадают по входной и выходной схемам. Так в выходной схеме всех упомянутых шаблонов будет присутствовать конструкция, приведенная в примере главы 6. Различия будут заключаться в том, каким образом будут называться свойство, на которое накладывается ограничение owl:hasValue и соответствующие классы.

В этом смысле приведенные шаблоны целесообразно записывать в виде одного (сложного шаблона), как конструкцию из элементов, указанных в главе 4, и литерала со значением «plainText». А при применении данного шаблона к тексту – помимо самого шаблона, к литералу со значением «plainText» и части сложного шаблона применять другой или другие шаблоны.

Шаблоны, которые могут применяться к части сложного шаблона и к литералу со значением «plainText» в рамках одного шаблона могут быть объединены в группы, принадлежность к которым указывается в атрибуте group тега <pattern>.

Понятие группы шаблонов позволяет упростить запись шаблонов.

8 Применение LSPL

В рамках данного исследования была проведена попытка использования лингвистических процессоров [8] и синтаксического анализатора Dictum [9] для идентификации шаблонов в тексте. Однако оба средства не обеспечили полной функциональности, заложенной в LSPL.

На данный момент разрабатывается программа, которая позволила бы выявлять синтаксические группы на основе синтаксического анализатора Dictum.

Последний предоставляет на выходе текст входного предложения в виде дерева, в узлах которого находятся слова с обозначенными частями речи и грамматическими условиями.

В дальнейшем планируется использовать полученное средство и несколько шаблонов, подобных тому, что указан в примере, для построения таксономии понятий для предметной области «Семейство продуктов компьютерной телефонии Rander».

9 Заключение

Создан формализм LSPL. Использование созданного языка LSPL позволит формировать систематизированную, открытую для развития среду лексико-синтаксических шаблонов.

Предполагается, что данный формализм послужит базой для последующего исследования методов автоматического формирования шаблонов.

Авторы считают, что создание данного формализма является базовой задачей в контексте задачи автоматизации процесса построения семантических моделей на основе применения лексико-синтаксических шаблонов.

Литература

- [1] Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes France, July 1992.
- [2] Большакова Е.И., Баяева Н.В., Васильева Н.Э. Структурирование и извлечение знаний, представленных в научных текстах // Девятая нацио-

нальная конференция по искусственному интеллекту с международным участием КИИ-2004 : Труды конференции : в 3 т. Т. 2. – М. : Физматлит, 2004.

- [3] Hearst, M.A. Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998.
- [4] Christopher Brewster, Fabio Ciravegna and Yorick Wilks. User Centred Ontology Learning for Knowledge Management. Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems, Stockholm, June 27–28, 2002, Lecture Notes in Computer Sciences, Springer Verlag.
- [5] Fabio Ciravegna: Adaptive Information Extraction from Text by Rule Induction and Generalisation. in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, August 2001.
- [6] Рабчевский Е.А., Булатова Г.И., Автоматическое построение онтологий // Научно-технические ведомости СПбГПУ. – 2007. – №4. – Санкт-Петербург : Изд-во Политехнического Университета, 2007.
- [7] Марчук Ю.Н. Компьютерная лингвистика: учебное пособие. – М. : АСТ: Восток–Запад, 2007.
- [8] Автоматическая обработка текста <http://www.aot.ru>
- [9] Синтаксический анализатор Dictum <http://www.dictum.ru>

Application of lexical-syntactic patterns to the automation of ontology building process

E.A. Rabchevsky, G.I. Bulatova, I.M. Sharafutdinov

The paper is devoted to the usage of lexical-syntactic patterns for the automation of semantic model extraction from text process. Some patterns, results of their usage and extraction of new patterns that were published in paper of other authors are discussed. We offer the original way for lexical-syntax pattern formalization. We name it LSPL - Lexical-Syntactic Pattern Language and we create proper XML Schema. The program for extraction of syntactic groups based on syntactic analyzer Dictum is developed. We plan to use some patterns and program for extraction of syntactic groups to build taxonomy some domain.

* Автор благодарит профессора, заведующего кафедрой КСиТ Пермского государственного университета, М.А. Марценюка за обсуждение работы.

О распределенных фактографических системах

© А.Г. Марчук

Институт систем информатики им. А.П. Ершова СО РАН, НГУ
mag@iis.nsk.su

Аннотация

В работе рассматриваются вопросы, связанные с фиксацией и структуризацией фактов. Речь идет о модели, создаваемой в виде набора сущностей вида: персоны, организационные системы, географические системы, документы и некоторые другие, записи существенных атрибутов этих систем и установления отношений между сущностями. Подход ориентирован на процессы, развивающиеся во времени и пространстве и позволяет создавать базы данных не устаревающие со временем и не зависящие от места восприятия и позиции воспринимающего.

Фактографический подход применим к ряду прикладных задач таких, как создание электронных архивов, музеев, исторических энциклопедий и справочников. Также была показана его эффективность к созданию простых систем делопроизводства. Изучается возможность использования структурированного поля фактов для создания аналитических систем и систем извлечения знаний.

Особенностями подхода являются: использование концепции Semantic Web [1], создание онтологии неспецифических сущностей, ориентация на распределенный характер создаваемого информационного поля, применение новой концепции информационного пространства документов [2].

Предлагаемые подходы и решения были частично реализованы в ряде прикладных проектов, выполненных и выполняемых в Институте систем информатики СО РАН.

1 Введение

Законченные полные версии принятых статей необходимо подать в оргкомитет до 14 июля 2008 года, используя форму подачи финальной.

Одно из самых дорогостоящих в современной информационной индустрии – это данные, собираемые и накапливаемые для решения конкретных за-

дач. Проблема заключается в том, что подавляющее большинство источников данных плохо сочетается между собой и существует огромное дублирование информационных единиц, описывающих одно и то же. Например, все мы, как абоненты разных сервисов и порталов Интернета зарегистрированы в разных базах данных своей персональной информацией, тем не менее, очередная заинтересовавшая нас информационная система снова предлагает ввести данные очень похожего профиля. Нам говорят, что персональные данные являются приватными и существует законодательство об ограничении распространения персональной информации. Все правильно, но та же ситуация сохраняется и с теми данными, которые не подпадают под ограничения, являются по существу публичными и, на самом деле, предназначены для распространения. Это относится к данным об организациях, о географических объектах, об открытых документах. Легко сообразить насколько интереснее, актуальнее и, соответственно дороже были бы данные, размещенные в едином источнике, заслуживающем доверия. Назовем эту проблему проблемой формирования единого информационного поля (фактов).

Некоторые достижения в направлении формирования единого информационного поля имеются. Для регистрации информационных сервисов имеется система UDDI [3], для унификации новостного потока используется RSS, для описания информационных ресурсов широкого профиля – Dublin Core [4], для частных случаев таких ресурсов, есть стандарты типа MARC, CIMI [5] и др. Проблема видится в отсутствии целостного подхода к структуризации разносортных данных, который позволил бы человеку или машинному агенту надежно находить требуемую ему информацию и использовать для решения тех или иных задач.

Другим аспектом группы проблем, находящихся в рассмотрении в данной работе является правильность устройства данных. Изучение применяемых схем данных показывает, что во многих случаях их разработка велась без опоры на какие-либо принципы кроме принципов, связанных с нормализацией и других очевидных аспектов. Наиболее распространенной ошибкой разработчиков базы данных является видение ситуации как мгновенного среза. Например, в базе данных сотрудников организации зафиксированы только нынешние сотрудники с указанием даты приема в организацию. А как же бывшие сотрудники? База данных сразу бы приобрела

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

черты исторической. Вроде такую модификацию проделать довольно просто, достаточно ввести поле завершения пребывания в организации. Однако, все не так просто. Довольно быстро выявляется группа новых проблем, препятствующих введению в базу данных историчности. Оказывается, работник не пребывал в одной должности и она как-то менялась, некоторые работники (девушки), меняли свои фамилии выходя замуж, оказывается нестабильно все – со временем менялись паспортные данные, телефоны, места проживания, научные степени, наконец, – структура предприятия, название предприятия.

Предлагаемый подход мы называем фактографическим, т.е. подходом, основанным на записи фактов. Заметим – на записи, а не стирании или редактировании. Мы ставим перед собой задачу сформировать принципы структуризации разнотипных данных, при которых все ранее записанные факты сохраняют свою справедливость вне зависимости от времени, географии или других факторов воспроизведения этих фактов. А изменение базы фактов осуществляется лишь добавлением новых информационных единиц (фактов). Простой пример: традиционное указание в поле записи о персоне возраста как числа лет – неправильный способ фиксации данных, правильный – указание даты рождения. К этому полю, в дальнейшем, будет добавлено другое поле: даты завершения жизненного пути. Поставленная цель идеальна. Реально в реализации имеются средства и изменения информационных полей и связей и их уничтожения. Однако, как будет показано далее, эти традиционные редактирующие действия приобретают новый смысл.

Наиболее естественным данный подход является в задачах исторического профиля, когда фиксируются факты, относящиеся к большому временному периоду и часто – к большой географической разбросанности. Принципы структуризации и основные методические моменты были разработаны при формировании и ведении базы данных кафедры программирования НГУ, где динамика изменений – очень большая и переход студентов из одной категории в другую исчисляется единицами лет. Данная задача относится к классу делопроизводственных, но ее особенности высветили ключевые проблемы фактографического подхода. В дальнейшем, подход получил реализацию и развитие в ряде проектов исторической направленности, но и тут, сопряжение исторического материала с фактами сегодняшнего дня (делопроизводством) является актуальной проблемой, если исходить из простого соображения, что сегодняшние события завтра станут историей.

Возвращаясь к проблемам приватности информации, в предлагаемом подходе эта проблема только обостряется. Действительно, при достаточно полном наборе информации о разных аспектах персон какого-то социума, сведение воедино разных информационных единиц выглядит пугающим при ее попадании в руки злоумышленников. Например, о персоне могут быть изложены и дата рождения и

места работы, проживания, коммуникационные адреса, отношение родства с другими персонами, посещаемые мероприятия, планы этих мероприятий, имена коллег и их телефоны у которых можно получить дополнительную информацию и т.д. и т.п. Причем уверяю вас, по публичным людям практически всю такую информацию можно получить в открытых источниках, потенциальной опасностью обладает именно информация, сведенная воедино. Вот почему наш коллектив предпочитает заниматься делами уже ушедших дней и увы, ушедших людей.

Еще одной особенностью представленного в данной работе подхода является ориентация на создание распределенных систем. Дело в том, что множество информационных ресурсов создается и поддерживается самостоятельными командами, имеющими независимые цели и предпочтения. Как правило, они хотят распоряжаться результатами своего труда в соответствии с собственным разумением, не ориентируясь на внешние процессы. Самый простой вариант для этого случая – полностью автономная разработка, что имеет существенные минусы: во-первых, не всегда используются профессиональные реализационные решения, во-вторых, созданный информационный массив не покрывает общего информационного поля, в-третьих, судьба информационного ресурса по истечению проекта часто плачевна. Наш подход позволяет решить все три проблемы не меняя автономного характера работы. Подход предлагает ряд методических и технологических решений, которые могут быть выполнены независимо и профессионально, а в дальнейшем развиваться в соответствии с изменениями базовых систем программирования. В подходе предусмотрено разделение информационного поля на приватное и общедоступное, имеются механизмы расширения схемы данных до уровня учета специфики локальной задачи, имеются средства настройки базовых решений на предметную область и предпочтения пользователей. В случае завершения проекта, созданная база данных может быть «втянута» в централизованное хранилище и продолжать участвовать в формировании общего информационного пространства.

2 Единое информационное пространство

В настоящее время существует множество подходов к фиксации фактов исторической направленности. К таким подходам относятся: базы данных, всемирная паутина World Wide Web, системы описания информационных ресурсов, основанные на метаинформации, электронные библиотеки. В последние годы активно развивается новое направление в информационных технологиях: Semantic Web (SW) [1].

Недостатками использования реляционных баз данных для решения задач фиксации фактов являются централизованный характер хранения и обработки, трудности в сопровождении и модернизации,

неудобства работы со спецификациями и словарями. Другая технология – WWW вообще мало ориентирована на формальные подходы к структуризации данных и, в основном, предназначена для использования информации, выполнения поиска и навигации человеком. Существенное развитие произведено в системах структуризации, основанных на метаинформации и в электронных библиотеках [6]. Проблемой этого подхода является «плоский» характер структуризации, больше предназначенный для больших объемов однородной информации и, как правило, не порождающий базы данных разнородных элементов. В отличие от упомянутых подходов, методология SW дает возможность формализовать существенные знания о смысле данных и использовать эти знания для создания информационных систем нового поколения, позволяет создавать распределенные системы, основанные на полномасштабной базе данных. Поле данных, при этом, легко использовать как для формирования WWW-интерфейсов, так и для выполнения запросов от машинных агентов.

В качестве базового подхода для широкого класса исторических фактографических систем был определен подход Semantic Web. Это обусловлено тем, что в подходе комплексно могут быть решены проблемы использования несвязанных или слабосвязанных между собой источников информации. Кроме того, формализмы, имеющиеся в SW и прилагающих стандартах, адекватны задаче работы с распределенным полем разнородной информации. Работа с данными, специфицированными средствами OWL [7], отличается также возможностью выполнения логического вывода для ряда практически важных случаев, а также применения других технологий, относящихся к искусственному интеллекту. В настоящее время ведутся исследования по использованию программирования в ограничениях, нечеткой логики, семантическому выводу и кластеризации данных.

Предлагаемый подход подразумевает формирование модели, в которой сущности внешнего мира представляются атрибутированными информационными единицами, отношения между сущностями реализуются либо в виде прямых ссылок, либо в виде составных конструкций определенного вида. Спецификация такой модели, воплощенная в виде онтологии позволяет создавать более универсальные программы обработки, использовать спецификации для удобного представления данных, многоязыковой реализации информационных систем, для анализа данных на предмет полноты и корректности. Важным элементом технологии RDF, которая находится «в центре» методологии SW, является концепция формирования семантической сети из отдельных высказываний и групп высказываний. Объединение высказываний в единый граф выполняется на основе слияния одинаково идентифицированных информационных единиц (айтемов). Это дает легкую возможность объединять воедино RDF-модели, хранимые в разных местах, то есть, поро-

ждает естественную распределенность информационного поля.

Рассмотрим особенности построения распределенного информационного поля. С точностью до деталей, информационное поле можно представлять как набор RDF-документов [8], логически связанных между собой единой идентификацией одинаково понимаемых объектов и отношений. Будем использовать термин «публикация» для характеристики того, что документ расположен по фиксированному для него URL и «виден» из Интернета. Возможны детали, связанные с возможным ограничением видимости для разных категорий пользователей, а также возможным косвенным характером доступа к документу. Последнее означает, что допустима ситуация, когда например реляционная база данных, погруженная в СУБД, имеет специальный интерфейс, преобразующий базу данных или ее часть в запрашиваемый RDF-документ. Опубликованные документы противопоставляются неопубликованным, т.е. таким, доступ к которым возможен только для «своих».

Таким образом, любой интернетовский агент имеет равный доступ к общим опубликованным документам и локальный доступ к своим документам. Из этого множества регулярно построенных документов агент может сформировать базу данных в виде (семантической) сети или, в некоторых случаях, в виде системы реляционных таблиц. Использование такой собранной базы данных может происходить в соответствии с функциями того агента, который собрал базу данных. В итоге, построена общая схема организации сосуществования информационных систем, обладающих собственным информационным контентом, при этом опирающихся на общее информационное пространство.

Общее пространство фактов будет полезным только если будет решена проблема «понимания», т.е. если каждый из документов с общими данными будет использовать структуризацию данных, сводимую к той, которая используется конкретным агентом (конкретной информационной системой). Но это еще не все. Нужно также, чтобы объекты разных баз данных, означающие одинаковые сущности, смогли бы объединиться при слиянии баз данных. Первую задачу назовем задачей совместимости онтологий, вторую задачу – проблемой отождествления.

В самом «жестком» варианте, единое информационное пространство выстраивается из RDF-документов, структурированных в соответствии с единой онтологией, а отождествление выполняется базовым для RDF способом – через одинаковые идентификаторы сущностей. В «мягком» случае, данные из подгружаемых документов подвергаются переструктуризации в соответствии с различиями в онтологиях, а проблема отождествления решается каким-то алгоритмическим способом. Заметим, что при недостаточности информации, алгоритм отождествления вообще может не дать положительного решения для слишком большого количества случа-

ев. Действительно, мировая практика идентификации персон предполагает указание для конкретной персоны полного имени, места рождения и даты рождения, последние две позиции, относительно конкретной персоны, вообще могут отсутствовать в базе данных. Не забудем также, что имя может писаться на разных языках, что имя может измениться и т.д. и т.п.

Можно сделать вывод, что «мягкий» подход к обеспечению совместимости данных может оказаться непрактичным. Наш подход в решении задачи совместимости онтологий заключается в том, что для разных документов общего информационного пространства допускается использование разных онтологий, но лишь в виде расширения общего онтологического построения, которое мы называем базовой онтологией. Здесь уместно привести аналогию из области использования библиотек компонентов для создания специализированных программных комплексов. Если этих библиотек много и они не согласованы между собой, то их использование довольно затруднительно. Если же библиотеки выстраиваются в иерархию по принципу послойного расширения, то при независимости (ортогональности) таких расширений, множество специальных программных комплексов могут эффективно их использовать для достижения своих целей.

Проблему отождествления мы решаем следующим образом. Предпринимаются меры для того, чтобы отождествление производилось на наиболее ранних этапах формирования конкретных доменов данных, например при первичном вводе данных, особенно если ввод осуществляется вручную. Это могут помочь соответствующие программные средства и «тотальные» базы данных типовых сущностей, таких как персоны, организации, города и др. Тем не менее, для реальных ситуаций, когда данные вводятся независимыми процессами, всегда остаются случаи, требующие алгоритмического решения задачи отождествления, соответствующие инструменты должны присутствовать в арсенале систем редактирования и манипуляций.

Возникает вопрос: а правильно ли формировать целостную модель (базу данных) в каждом из мест обработки? Нет ли возможности распределить не только данные, но и обработку этих данных? Представляется, что некоторые подходы к этой задаче возможны, однако в целом, задача является математически и методически открытой и, в данной работе, мы ее обсуждать не будем.

3 Базовая онтология

Ключевым вопросом в реализации подхода, явился вопрос создания онтологии неспецифических (базовых) сущностей. Термин «неспецифических» используется в смысле противопоставления «специфическим» сущностям. Дело в том, что практически каждая информационная система обладает своим предметом, своей спецификой. При этом,

наряду со специфическими сущностями, например какими-нибудь номерами счетов, транзакциями и остатками средств, очень часто в онтологии необходимо также иметь такие классы сущностей, как персоны, организации, места нахождения, коммуникационные координаты, т.е. сущностей, являющихся продуктом общекультурного взгляда на мир и современную цивилизацию, лишенные конкретной специфики и одинаково понимаемыми создателями систем и пользователями.

К сожалению, на данный момент, подобной онтологии неспецифических сущностей как мирового стандарта или рекомендации – не существует, хотя в ряде стандартов (DC, FOAF [9], CIMI) прослеживаются ее элементы. Исходя из сформулированных далее требований и принципов, нами была порождена и обоснована базовая онтология, которая в настоящее время используется в ряде проектов.

Базовая онтология формировалась в течение довольно длительного времени при работе над реальными проектами. И можно точно утверждать, что ее построение еще не закончено. Накапливавшиеся в процессе эволюции онтологии изменения требовали периодической трансформации накопленных данных, но методы абстрактного программирования, основанные на использовании явных спецификаций данных, позволяют сохранить при таких изменениях большую часть кода программ и интерфейсов.

Был выработан набор принципов построения онтологий, в частности, базовой онтологии.

1. Отражение фактов реального мира. При создании общего информационного поля лучше придерживаться фактографического способа формирования модели мира. Это означает, что предмет фиксации в базе данных должны быть реально существующие сущности, причем сведения об этих сущностях должны иметь объективный характер, не зависящий от особенностей предметных областей, для которых эти факты используются. Например, если мы создаем историческую базу данных, то при отражении конкретной персоны, нас интересуют такие факты, как дата рождения, полученное образование, места работы и проживания. Но мы воздержимся от фиксации таких субъективных моментов как оценка достижений человека, его популярности или влиятельности. Эти моменты могут возникнуть при оформлении конкретного взгляда на исторические процессы, фиксируемого конкретной информационной системой, которая, как уже отмечалось, может активно использовать общее информационное поле фактов.

2. ER-модель как основа онтологического построения. Традиционная модель Entity – Relationship является и естественной и эффективной при данном подходе. Действительно, отражая сущности и атрибуты сущностей, мы порождаем поле независимых информационных единиц. Отражая отношения между сущностями, мы формируем связанную базу данных, в которой от одного объекта можно «добраться» до другого, что соответствует устройству мира. Если база данных распадается на

независимые части, это свидетельствует либо о неполноте данных, либо о неправильной структуризации данных.

3. Балансировка по детальности описаний, минимизация национальных и культурных различий. Описать мир во всех подробностях невозможно и нецелесообразно. Поэтому надо ограничить детальность описаний некоторой глубиной. Это относится и к набору сущностей и к набору отношений. Детальность описаний также лимитируется возможностями одинакового понимания вводимых понятий и градаций разными группами разработчиков и пользователей. Например, виды организационных структур могут весьма различаться в разных странах и даже в одной стране в исторической ретроспективе. Также, руководящие должности, территориальное административное деление, виды документов, родственные отношения и т.д., могут называться очень по-разному и не будут поняты или обработаны при анализе данных разных временных периодов и стран.

4. Специальная реализация атрибутированных отношений. Одной из распространенных ошибок, совершаемых создателями онтологий или схем баз данных является игнорирование возможных атрибутов для конкретного отношения. Для онтологий, создаваемых средствами OWL, такое ошибочное решение обычно воплощается в использовании в качестве отношения прямой ссылки через определение объектного свойства. Например, часто отношение «работа» между персоной и организацией рассматривается как объектное свойство. Это неправильно. Человек вступает в отношение с организацией в конкретный момент времени, может расторгнуть эти отношения в другой, в отношении «работа» могут быть и другие атрибуты, например должность.

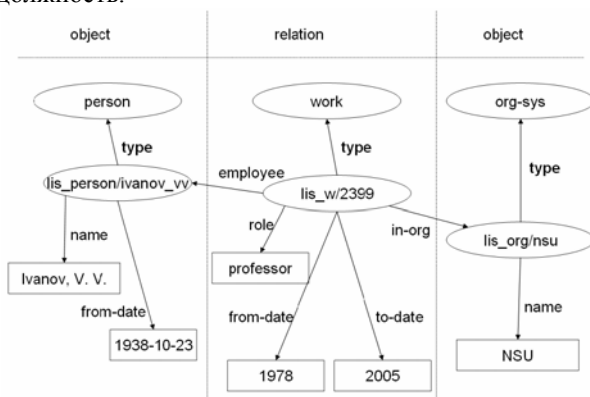


Рис. 1. Пример связывания объектов атрибутированным отношением

5. Прямая фиксация контекстной информации. Люди привыкли излагать свою мысль в определенном контексте. Например, «Путин сказал...». Совершенно очевидно, что имеется в виду вполне конкретный человек из десятков тысяч Путиных, имеется в виду недавнее время и, скорее всего, вне этого предложения обрисовано мероприятие, место его проведения и другие контекстные особенности.

В слишком большом количестве случаев некоторый контекст остается и в используемых формальных спецификациях данных. Это резко сужает возможность объединения данных, созданных в разных неявно присутствовавших контекстах. Часто делается попытка использовать контекстную информацию для группирования информационных единиц, например, объединение в списки персон, работающих или проживающих в одном месте, группирование по ролям (преподаватели-студенты, врачи-больные, работники, сгруппированные по иерархии организационной структуры предприятия и др.). Анализ показывает, что использование подобных группирующих или иерархических отношений в качестве прямой основы структуризации данных никакой полезности не несет, но может существенно усложнять доступ к данным. Подобные группирования являются вторичными и должны выводиться из записанных в базе данных свойств данных и отношений между ними.

6. Сводимость базовой онтологии к системе простых реляционных таблиц. Этот принцип означает, что онтология формируется так, чтобы средой структуризации и хранения данных могли бы быть не только RDF-документы, но и реляционные таблицы простого вида, причем соответствие между ними было бы эффективным в смысле реализации. Этот принцип видится важным для обеспечения совместимости формируемых RDF-данных с существующими реляционными данными и для использования реляционных СУБД как среды для поддержания модели семантической сети. Такое соответствие видится достаточно естественным: каждому классу сущностей сопоставляется таблица, первичным ключом которой является идентификатор сущности, используемые поля данных (атрибуты, DatatypeProperty) превращаются в типизированные колонки, прямые ссылки (ObjectProperty) превращаются в колонки внешних ключей. Такое соответствие существенно ограничивает набор принимаемых решений при формировании базовой онтологии. Во-первых, приходится отказаться от множественности для указания типа сущности, кроме того, онтология требует обязательной типизации сущностей. Во-вторых, не используется наследование для свойств DatatypeProperty и ObjectProperty. В-третьих, на уровне онтологии не используются «сложные» конструкции RDF такие как blank nodes, контейнеры (containers), списки (list), reification. В-четвертых, желательно не допускать множественности значений в колонках реляционных таблиц. Это, как правило, можно сделать, оставляя только «правильные» отношения и управляя направлением стрелок в свойствах, связывающих разные сущности.

В основе созданной базовой онтологии определены четыре класса сущностей: персоны, организационные системы, документы и географические системы. Между базовыми сущностями допустим ряд отношений, общая схема основных отношений изображена на рисунке.

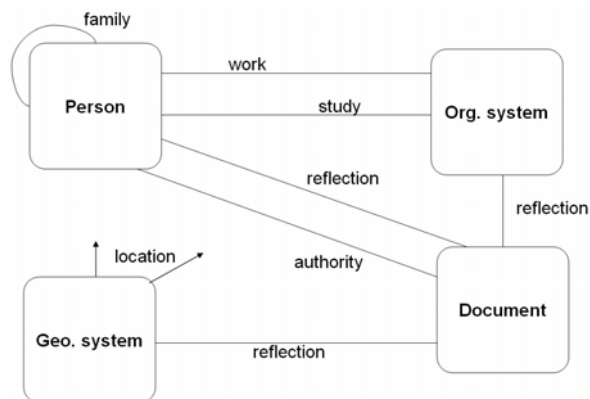


Рис. 2. Основная часть базовой онтологии

Под организационными системами мы понимаем (временное) сообщество людей, ориентированных на достижение конкретных целей. В эту категорию попадают и организации и мероприятия и команды и ассоциации людей и организаций. Под понятие географических систем попадают города, страны и прочие места проявления активности людей. Документы – широкий набор целостных образований (книги, статьи, файлы, фотографии и др.), сущностью которых является зафиксированное информационное наполнение. Между введенными сущностями онтологией определены следующие отношения. Между людьми учитываются только семейные отношения, между персонками и организационными системами возможно отношение «участник», причем в силу важности, выделено специальное отношение «обучение». Отметим, что уже стало традицией, в *Curticula Vitae* помещать информацию о полученном образовании. Между документами и сущностями других классов, основным отношением является отношение отражения. Под этим мы понимаем, что в содержимом (content) документа изображена или упомянута данная сущность. Кроме того, персоны могут выступать в качестве авторов документа, что фиксируется отношением «авторство». Географические системы выступают в качестве места для какого-то зафиксированного в базе данных момента в виде отношения «размещение» (location).

Из других классов сущностей, имеющих в базовой онтологии, следует упомянуть коллекции и архивы. Кроме того, на рисунке на изображен ряд унарных и бинарных (сложных) отношений, таких как: родственные организации, титулы, степени и награды, коммуникационные адреса и др. В качестве примера, покажем как для персоны указать адрес электронной почты.

```

mail rdf:about="em134983">
  <user rdf:resource="mag" />
  <e-code>mag@iis.nsk.su</e-code>
  <from-date>1990</from-date>
</email>

```

Легко видеть, что введенный способ «прикрепления» адреса электронной почты к персоне, не ограничивает ни количества адресов, ни возможности «закрытия» устаревшего адреса. Последнее легко

выполняется добавлением к записи поля <to-date>YYYY-MM-DD</to-date>, указывающего время завершения действия отношения email.

В базовой онтологии широко используется введение подклассов для указания частных случаев и даже некоторые классы помечаются как абстрактные (что делается средствами, дополнительными к формализму OWL). Важно отметить, что в силу использования механизма классов для сложных отношений, показалось, что определенный для корневого класса в OWL термин Thing терминологически не соответствует понятию отношения. Например, отношения «участник», «размещение» и др., являющиеся сложными отношениями, вряд ли логично называть «вещью» (thing). В нашей онтологии, корневой класс является абстрактным и назван entity. Для классов основных сущностей, которые уже соответствуют понятию «вещь», мы используем класс sys-obj – системный объект. От этого абстрактного класса наследуются персоны, организационные системы и др.

Существенным нововведением в базовой онтологии является понятия датирования и именования. Сначала о датировании. Вообще, указание временного диапазона определено прямо для конечного класса, как основное свойство системных объектов и отношений через два атрибута (DatatypeProperty) from-date и to-date, означающие временные отметки начала и конца сущности или отношения. Однако оказалось, что этого недостаточно. Дело в том, что во многих случаях практической работы с данными, точные значения этих атрибутов указать затруднительно. Зато можно сделать высказывания типа: «эта сущность (отношение) началась до такой-то даты», «эта сущность еще не закончилась (или уже закончилась)» и др. Также иногда можно указать приблизительную дату, но это не то же самое, что указать дату точно.

Некоторое решение указанных проблем можно найти во введенном унарном отношении датирования (dating). Отдельным датированием для указанной сущности определяется некоторая дата, которая имеет квалификатор и точность задания этой даты. Квалификатор может быть: before-date, from-date, in-date, to-date, after-date, т.е. один из пяти вариантов отметки точки на временной оси, на которой расположен интервал существования данного явления. Например,

```

<dating rdf:about="da298345">
  <referred rdf:resource="em134983"/>
  <date>2006</date>
  <d-specificator>to</d-specificator>
  <d-accuracy>1-0-0</d-accuracy>
</dating>

```

Означает, что мы отметили временную точку 2006, известную с точностью до одного года, как дату окончания сущности, имеющей идентификатор da298345.

Теперь рассмотрим именование. Для системных объектов имеется стандартный атрибут name, означающий имя объекта. Однако, как и в предыдущем случае, этого оказывается недостаточно. Соответст-

венно, введено унарное отношение naming, означающее сопоставление системному объекту (еще одного) имени или синонима. Кроме того, что мы получили возможность множественного обозначения объекта, именование можно атрибутировать, например можно указывать с какого и по какое время существовал такой вариант имени.

4 Ввод и редактирование данных, визуализация информации

Системы, базирующиеся на формальных спецификациях, обладают существенным достоинством высокого уровня абстракции проектирования. Программист рассуждает не в терминах персон, городов, организаций, а в терминах математической модели, отношение которой с реальным миром задается онтологией, которую он также рассматривает как математический объект. Не всегда удастся достаточно глубоко выдержать такой уровень абстракции, но в базовых действиях: визуализации, редактировании, поиске, навигации – это возможно и желательно. Полученные программные средства становятся универсальными, для их использования достаточно задать онтологию и некоторые другие спецификации. А полученная система способна «разговаривать» с пользователем в терминах заданной спецификациями предметной области и на языке, удобном пользователю.

Рассмотрим RDF-построение как математический объект и некоторые из тех задач, которые пользователь желает решать.

RDF-модель, в своей основе, представляет собою ориентированный граф, узлами которого являются сущности (в терминологии стандарта – subject) и строковые константы. Дугами являются связи «сущность – сущность» и «сущность – константа». Другие свойства модели – типизацию, идентификацию, соответствие формальным спецификациям (онтологии), мы будем в рассуждениях привлекать по необходимости. Назовем айтемом типизированную сущность нашей модели. Если сосредоточить взгляд на конкретном айтеме, то легко увидеть его каноническое представление:

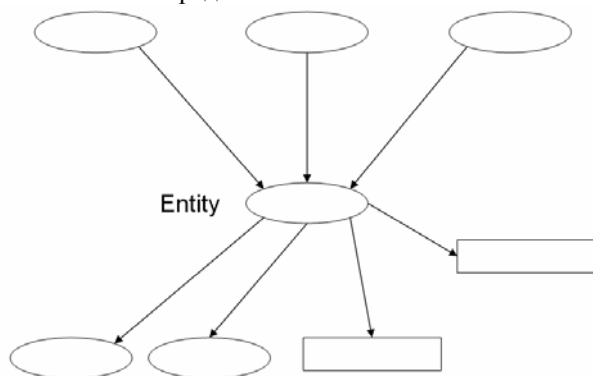


Рис. 3. Каноническое представление айтема

Заметим, что мы ограничиваемся рассмотрением только типизированных моделей, т.е. таких моде-

лей, в которых определены классы сущностей и определены свойства (DatatypeProperty, ObjectProperty) и их возможные использования в модели. Часто мы также ограничиваемся рассмотрением онтологий, в которых множественность (cardinality) исходящих дуг ограничена единицей. Рассмотренная в предыдущем разделе базовая онтология обладает указанными свойствами.

Как мы видим, фокусирование на конкретном айтеме выделяет также подграф, являющийся ближайшим окружением данного объекта. Легко видеть, что в совокупности, атрибуты айтема, вместе с окружением айтема, представляют существенную информацию для его характеристики. Назовем такой подграф информационным портретом айтема (сущности). Например, сфокусировав внимание на некоторой персоне, в соответствии с базовой онтологией, мы получим информацию о возрасте, родственниках, образовании, работе, проживании, участии в мероприятиях, коммуникационных адресах и т.д. Иногда, при этом полезно будет «заглядывать» на более, чем одну дугу в окружении изучаемого объекта.

Если ввести в модели метрику, например как взвешенную сумму числа дуг между узлами графа, то задачу построения информационного портрета можно свести к задаче выделения из графа шара определенного диаметра, с центром в айтеме объекта изучения. Такая геометрическая подзадача является не единственной интересной постановкой. Если выделить два узла (айтема), то можно выявлять некоторую окрестность точек, сумма расстояний от которых до фокусов не превышает заданной величины, т.е. строить эллипсоид. Полезная интерпретация данного построения заключается в формировании информационного портрета айтема с точки зрения другого айтема типа: что это за объект и что связывает данный объект со мной. Возможны и другие способы определения подграфа, основанные не только на метрике и назначении весовых коэффициентов для дуг, но и учитывающие специально назначенные веса для узлов. Несложно сконструировать оболочку для N назначенных точек и интерпретировать ее как раскрытие заданной темы. Подобные построения будут эффективны как при вычленении содержательной информации при построении справок (досье), так и для организации поисковых действий по множеству ключевых слов или объектов.

Работа с сетевыми структурами RDF имеет свои особенности. Это касается и визуализации информации и поиска и редактирования. Причем все три процесса могут быть перемешаны в рамках единой работы. Начнем работу с поиска, сформулируем поисковый запрос, вид которого нам пока не существует. По запросу могут быть найдены айтемы-кандидаты, соответствующие критериям запроса. При слишком полном запросе или при недостаточности информационного покрытия, таких кандидатов может быть несколько. Следующим этапом является визуализация множества айтемов-

кандидатов в таком виде, чтобы пользователь смог найти нужный. Визуализация списка айтемов как правило базируется на визуализации отдельного айтема, вопрос лишь в детальности создаваемого портрета. После выделения найденного айтема, пользователя может интересовать его детальный информационный портрет. Легко видеть из особенностей построения RDF-моделей, что информационный портрет состоит из множества значений атрибутов и множества айтемов ближайшего окружения. Множество значений атрибутов легко реализуется в виде набора пар: имя атрибута – значение атрибута и представимо в традиционной форме «в столбик» или «в строчку (таблицы)». А множество ассоциированных айтемов мы можем визуализировать уже упоминавшимися средствами. К этой композиции достаточно добавить навигационные переходы на айтемы окружения и базовые принципы визуализации семантических сетей RDF, в общем, изложены.

Трудности появляются в процессе редактирования в связи с необходимостью постоянно, на уровне оператора решать две задачи, обе из которых, как минимум, непривычны этому оператору в контексте редакторских действий. Первая – поиск и отождествление информационных единиц, вторая – связывание одних объектов с другими через простые или сложные отношения. В чем состоит непривычность или сложность данных действий для оператора? Дело в том, что на уровне ввода и редактирования данных, наиболее стойкими и привычными конструкциями для ввода и редактирования, являются запись, в случае одиночного объекта редактирования и таблица, для множественности объектов редактирования. В работе с визуальным интерфейсом это означает или редактирование полей формы или редактирование ячеек таблицы. Работа со ссылками – обычно не предусматривается. Значит, проблема заключается в том, как устанавливать именованную связь данного айтема с другим айтемом базы данных. Рассмотрим эту задачу для случая «прямых» ссылок, в графовой модели это соответствует исходящим из айтема дугам, ведущим к другим айтемам. Пусть в ячейке, имеющей ссылочный тип (ObjectProperty), надо средствами действий с визуальным интерфейсом установить конкретную ссылку. Каким должен быть интерфейс ввода ссылки? Причем желательно, чтобы этот интерфейс был как бы «внутри» ячейки и не портил общий портрет редактируемого айтема. Имеется несколько вариантов решения, применяемых в различных системах.

Самый простой способ: превратить набор айтем-кандидатов на связывание в визуальный элемент интерфейса типа ListBox или ComboBox. Несмотря на то, что этот вариант является самым понятным пользователю, имеется ряд недостатков, сильно ограничивающих его применимость. Во-первых, без введения дополнительной информации со стороны пользователя, список кандидатов на ассоциирование может оказаться непомерно большим, что лишает этот способ практичности. В этом плане,

вариант ComboBox предпочтительнее, поскольку в нем встроена возможность ввода поисковой информации. Во-вторых, айтема, с которым мы хотим установить связь может не оказаться в базе данных. Тогда в списке будет фигурировать вариант «другое» и для этого варианта нужна специальная форма инициации ссылаемого айтема. В-третьих, для определения того, какой из айтемов списка является искомым, часто надо посмотреть детали кандидата, что также является нетипичным действием для редактирования.

Другим вариантом является внедрение поискового интерфейса в данное конкретное место редактирования, возможно, прямо в ячейку редактирования. В этом случае, сначала дается форма для формулирования поискового запроса, потом появляется, упоминавшийся ранее, список айтемов из которого требуется сделать выбор и имеется «кнопка» формирования нового айтема, если поиск не выявил его существования. Новый айтем строится исходя из имеющегося контекста и введенной в поисковом запросе информации. Довольно громоздко может выглядеть такое универсальное решение в случае, когда тип прикрепляемого айтема может различаться, тогда, как правило, сначала пользователь устанавливает тип, потом может генерироваться форма поискового запроса, только потом мы можем добираться до списка кандидатов. Достоинством данного подхода является его универсальность и относительная сочетаемость со стандартными реализационными схемами Web-интерфейсов и оконных приложений.

Третьим вариантом является использование технологии drag and drop для установления прямых ссылок между айтемами. Смысл такого установления заключается в том, что мы ассоциированный айтем можем найти или породить независимым интерфейсом, потом «взять» его и «положить» в требуемую ячейку или прямо на текущий айтем. Положительным моментом такого подхода является использование базовых представлений о модели структуризации: то, что есть сущности (айтемы) и есть отношения и процесс редактирования заключается в поиске или вводе сущностей, редактировании их атрибутов и установлению отношений между различными сущностями. Тем не менее, модель редактирования данных, пользователю непривычна и иногда от пользователя требуются знания о деталях используемой онтологии.

Возможно, наиболее естественной для пользователя, была бы модель текстового ввода сущности, на которую устанавливается ссылка. Например, в поле «место проживания», можно было бы позволить указать некоторый текст с минимальными элементами формального синтаксиса, скажем: г. Новосибирск. Система смогла бы разобрать такую запись с помощью словарей понятий и сокращений и найти соответствующую сущность в базе данных для использования в качестве ссылки, а в случае отсутствия – ввести новый элемент и установить на него ссылку. Понятно, что при отождествлении возмож-

на альтернативность, но об этом, можно предупредить и дать варианты для выбора. Такая схема требует развитой базы данных по основным категориям сущностей, указаний текущего контекста редактирования и системы соглашений о применяемых служебных словах и сокращениях или качественной системы анализа естественно-языкового текста, основанной на алгоритмах искусственного интеллекта. Кроме того, эта схема не является универсальной. Например, если речь идет о документах (фото-документах и др.), то поисковое действие вряд ли будет базироваться на имени документа. Также сложно идентифицировать на словесном уровне мероприятия, а иногда и организации.

Вышеупомянутые особенности редактирования семантических сетей, позволяют сделать вывод об отсутствии привычных пользовательских моделей для выполнения поисковых и редактирующих действий. Возможно, нужно предусматривать наличие альтернативных способов для получения того же результата.

Довольно часто проблемой становится не отсутствие информации, а ее избыток. Например, относительно конкретной персоны в базу данных могут быть занесены тысячи фото и других документов, в которых он отражен, сотни документов, в которых он автор, десятки конференций, в работе которых он участвовал и т.д. Целостное изображение такого информационного портрета приводит к трудностям в нахождении пользователем нужной информации и связей между информационными единицами. Для приведенного примера, целью пользователя может являться нахождение номера телефона данной персоны, что может оказаться непростым занятием при неудачном представлении упоминавшегося ранее списка айтемов, ссылающихся на данный.

Таким образом, речь идет об ограничении видимости, задаваемой статически, при настройке конкретного интерфейса или динамически, при работе с конкретным объектом. Мы опробовали два подхода для решения данной проблемы. Первый заключается в определении множеств классов и свойств, которые в данный момент перестают рассматриваться. Эти множества задаются группами, так называемыми темами. Например, тема «документы» включает в себя несколько классов документов, а также ряд сложных отношений и их свойств, связанных с документами. Это – авторство, отражение, публикация. В интерфейсе работы с данными есть настройочный список тем, в котором можно отметить те темы, которые будут использоваться при просмотре или редактировании данных и те, которые использоваться не будут.

Другой способ связан с введением информационной значимости айтемов. Этот субъективно вводимый фактор, дает возможность списки айтемов сортировать по некоторому критерию важности и оставлять в формируемом информационном портрете только самую существенную информацию.

5 Исторические информационные системы

Был сформулирован и обоснован подход к созданию электронных архивов. Суть этого подхода заключается в том, что порождаются электронные образы единиц архивирования, создается или используется база данных традиционных сущностей (персоны, организации, географические точки, события), осуществляется «привязка» единиц архивирования к элементам базы данных, создается информационная система, позволяющая осуществлять навигацию и поиск по базе данных.

В Институте систем информатики СО РАН, в течение 10 лет ведутся исследования и выполняются разработки по созданию информационных систем исторической направленности. В частности, разработан и уже несколько лет эксплуатируется электронный архив документов А.П. Ершова. В результате выполненных работ, были сформулированы основные принципы создания фактографических систем исторической направленности: система должна отражать факты реального мира; для этого используются формальные системы, позволяющие через спецификации определять номенклатуру объектов отражения и основные свойства данных; необходимо создавать информационные системы распределенного типа; базы данных должны быть контекстно независимыми; должны активно использоваться международные стандарты; интерфейсами к системам должны быть для людей Web-клиенты (браузеры), а для программных агентов – Web-сервисы.

Наиболее крупным проектом последнего периода времени явился проект фотоархива Сибирского отделения <http://soran1957.ru>, запуск в эксплуатацию первой очереди которого был приурочен к 50-летию СО РАН. Архив в настоящее время содержит около 15 тыс. фотодокументов, база данных охватывает период с 1957 года и содержит информацию о 9 тыс. персон, 2 тыс. организаций и событий, также в базу данных помещена информация о членах Российской академии наук за весь период ее функционирования, взятая из базы данных РАН.

Архив представляет собой программно-аппаратный комплекс, реализующий сканирование и другую обработку фотоматериалов, обеспечивающий сохранность и защиту информации от несанкционированного доступа, предоставляющий средства редактирования базы данных, рабочие и публичные интерфейсы. Публичный интерфейс архива был запущен в опытную эксплуатацию в 2007 году, за первые три месяца эксплуатации, его посетило около 20 тысяч посетителей, пользователями просмотрено более 600 тысяч страниц.

Институт систем информатики СО РАН будет продолжать работу над фотоархивом как в части технологии, так и в части пополнения новыми данными, подробного описания фотодокументов, исправления неточностей. На этой же платформе предполагается создавать и другие корпоративные и

личные архивы содержащие базу фактов, а также документы, фото, видео, аудио документы, описание людей, организаций, событий и проектов.

Литература

- [1] Berners-Lee Tim, Hendler James, Lassila Ora, The Semantic Web. In Scientific American, volume 284(5), pages 34–43, 2001.
- [2] Марчук А.Г. Распределенные электронные архивы, библиотеки и базы данных. Препринт 122, Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, 25 с., 2004.
- [3] Online community for the Universal Description, Discovery, and Integration OASIS Standard — <http://uddi.xml.org>
- [4] The Dublin Core Metadata Initiative — <http://dublincore.org/>
- [5] The CIMI Profile. Release 1.0H. A Z39.50 Profile for Cultural Heritage Information — http://www.cimi.org/public_docs/HarmonizedProfile/HarmonProfile1.htm
- [6] Когаловский М.Р. Перспективные технологии информационных систем, М.: ИТ-Экономика, 2003. 288 с.
- [7] Web Ontology Language (OWL). — <http://www.w3.org/2004/OWL>
- [8] Resource Description Framework (RDF). — <http://www.w3.org/RDF>
- [9] FOAF — <http://www.foaf-project.org/>

About Distributed Factographic Systems

Alexander Marchuk

In this research work special model is proposed for fixation and structuring of historical facts. The model is based on factographic principles of collection of formally specified data and information. The structuring model includes several classes of instances such as: persons, organizing systems, geographic systems, documents and some other. Instances are connected by simple or complex relations. Proposed approach is oriented to processes, upcoming in time and space, it allows to create databases, which does not become obsolete and does not depend upon time, place and position of perceiving person. Factographic approach can be used in a set of such applications as: building of digital archives, museums, historical encyclopedia and directories. Same ideas can be efficiently implemented in small systems for office work organizing. New possibilities in use of structured field of facts for analytics and knowledge extraction are under investigation. Specific features of the proposed approach are: main concept is from Semantic Web, new ontology of non-specific entities is proposed, strong orientation on distributed databases, use of the idea of information space of documents. This approach was partially implemented in applied projects of A.P. Ershov Institute of Informatics Systems.

Автоматическое построение терминологической базы знаний*

© О.Г. Чанышев

ОФИМ СО РАН
fedorov22@yandex.ru

Аннотация

В докладе представлены метод автоматического построения и структура терминологической базы знаний предметной области, заданной множеством естественно-языковых текстов. Автоматически строится предметный указатель, путем кластеризации терминоподобных словосочетаний по принципу вхождения в них главного («кардинального») слова. Вводится мера ассоциативной близости терминоподобных словосочетаний. Кратко описан процесс выделения терминоподобных словосочетаний. Основным фактором, определяющим их адекватность предметным областям, является доминантность. Приведен результат оценки адекватности.

1 Введение

Несомненный интерес представляют работы по автоматизации построения баз знаний [2, 4, 5] путем экстракции знаний из естественно-языковых текстов, представляющих некоторую предметную область. Практически полезной и не требующей детального семантико-синтаксического анализа предложений для ее построения может быть терминологическая база знаний.

Настоящая работа выполнена в рамках комплекса исследований, посвященных автоматическому анализу естественно-языковых текстов и автоматизации построения семантических сетей отдельных текстов и предметных областей. В ходе исследований были разработаны: Ассоциативная модель реального текста [6, 7], позволившая, в частности, выделять наиболее тематически значимые слова (доминанты), оригинальные алгоритмы распознавания различных фрагментов текстов (рубрик и их наименований, списков литературы, ФИО, границ предложений) [9], простой алгоритм классификации текстов на основе анализа пересечений множеств до-

минант [5].

Естественным продолжением стала разработка метода автоматического выделения терминоподобных словосочетаний (ТС) [9], для практической реализации которого использовался весь наработанный ранее инструментарий. Главное дополнение к известным из литературы условиям, которым должны отвечать ТС (контактность, устойчивость, объектность (обязательное наличие существительного), смысловая завершенность и ограниченная распространенность), – условие доминантности: каждое словосочетание должно включать слово, принадлежащее множеству доминант как минимум в одном из текстов анализируемого множества. Это условие обеспечивает адекватность выделенных словосочетаний предметной области и повышает эффективность фильтрации.

В докладе акцент смещен в сторону организации информации о терминоподобных словосочетаниях в базе знаний интеллектуального прототипа терминологической ИПС (в дальнейшем – просто ИПС). Структура фактов определяется требованиями к ИПС.

2 Требования к терминологической ИПС

1. Прежде всего, показалась привлекательной идея «педагогического» характера – дать возможность пользователю получить представление о содержании ТС в пределах множества текстов предметной области путем определения и вывода всех предложений вхождения элементов его грамматической парадигмы.

2. Эту возможность следовало дополнить быстрым доступом к энциклопедическим и толковым словарям, в частности, для построения конспектов по «локальным темам», определяемым терминоподобным сочетанием.

3. Представляется, что наиболее удобным для пользователя вариантом выбора ТС является выбор по предметному указателю, в котором ТС кластеризованы в группы по принципу общего главного или «кардинального» слова.

4. Очевидно, следовало добавить сервисы, создание которых не представляет сложностей:

предоставление списка текстов вхождения выбранного ТС для доступа к ним;

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

поливариантный выбор ТС – по норме, по элементам грамматической парадигмы, по заданной подстроке.

5. Чтобы удовлетворить первому требованию, необходимо автоматически определять вхождения ТС в тексты и предложения текстов, тем самым, создавая дерево вхождений, в котором узлы – тексты, предложения и ТС связаны отношениями включения. Это дерево можно сделать терминологической семантической сетью, если ввести два вида отношений между парами ТС – отношение тематической важности и отношение ассоциативной близости. В таком случае представляется возможность для интерпретации запросов пользователя, сформулированных на естественном языке как запросов на поиск путей между ТС в терминологической семантической сети.

Конечно, на данном этапе исследований интерпретация будет основываться на анализе ключевых слов, в качестве которых выступают ТС. «Анализ ключевых слов есть метод анализа предложений на предмет наличия ключевых слов, которые может распознавать компьютер. Ключевые слова становятся значениями объектов предикатов. При этом не так важна грамматическая структура предложения, так как программа не анализирует связей между словами. Компьютер реагирует одинаково на различные варианты входного текста; роль играет лишь наличие ключевых слов» [3].

Результат может быть представлен пользователю в виде множества предложений вхождения терминов, выделенных из запроса и промежуточных терминов. Учитывая, что каждому терминоподобному сочетанию приписывается тематический вес, появляется возможность сравнить различные пути по двум параметрам: по значению ассоциативной близости и их тематической важности.

Таким образом, в список первоочередных задач входили две основных: задача определения кардинального слова и задача выбора меры ассоциативной близости ТС.

3 Необходимые определения

3.1 Ассоциативная мощность

Несмотря на то, что Ассоциативная модель была неоднократно изложена в ранних публикациях, для лучшего восприятия настоящего текста приведем основные положения, касающиеся понятия «ассоциативная мощность» – основы меры тематической важности ТС.

За установление ассоциативных связей между предложениями текста отвечает множество независимых лексем связи (НЛС), обладающее тем свойством, что для каждой произвольной пары из этого множества существуют минимум два предложения, в которые они входят по отдельности (элементы НЛС не принадлежат стоп-словарю и их частота встречаемости в различных предложениях больше 1).

Мера тематической важности («тематический вес») лексемы из НЛС (*Ш* – **ассоциативная мощность**) равна числу других лексем из НЛС, встречающейся с данной в предложениях текста, с поправкой, учитывающей «личный» вклад лексемы в установлении связей – лексемы первого предложения вхождения не учитываются, поскольку в нем присутствуют только «родители» и «братья» данной.

Из НЛС отбираются **доминантные лексемы** (доминанты) по критерию $Ш > 0.5R$, где R – ранг текста (максимальный номер группы лексем с равными *Ш* в частично упорядоченной по убыванию *Ш* последовательности НЛС).

3.2 Нормы и парадигмы ТС

Норма (лемма) слова. Норма для каждого слова словосочетания определяется по строке морфологической информации, выдаваемой морфопарсером *mystem*.

Норма ТС получается заменой слов на леммы.

Парадигма ТС (слова) – множество грамматических вариантов ТС (слова) в пределах заданного множества текстов. Размер парадигмы – число вариантов.

3.3 Тематические веса

Для работы с множествами текстов **вес доминанты** в фиксированном тексте полагается равным ее обратному рангу в убывающей по значению ассоциативной мощности последовательности доминант.

Вес нормы доминанты во множестве файлов равен сумме весов ее доминантных грамматических форм.

Вес словосочетания равен сумме весов входящих доминант.

Вес нормы сочетания равен сумме весов элементов его парадигмы.

3.4 Кардинальные слова и предметный указатель

Кардинальные слова

Кардинальным словом нормы словосочетания назовем норму доминанты с наибольшим весом. По нашей (вероятно, наивной) оценке кардинальные слова – это однословные термины, не обязательно принадлежащие рассматриваемой предметной области. В Таблице 1 для шести предметных областей приводятся данные авторской оценки «термин – не термин».

Метод оценки: не вдаваясь в анализ принадлежности слова к специализированной или общеупотребительной лексике, в качестве терминов рассматривались существительные и аббревиатуры. Редко – прилагательные, явно выступающие в роли существительных в конкретном тексте. Например, *большая, испытываемый* в рамках предметной области «Психология».

Ниже приводится полный список кардинальных слов предметной области «Компьютерная лингвистика». Кардинальные слова упорядочены по убыванию весов. Не термины набраны жирным курсивом.

слово, документ, текст, знание, данный, модель, цепочка, правило, система, обработка, отношение, термин, проектирование, информация, рубрикация, запрос, статья, анализ, группа, словарь, выборка, предложение, **строить**, ооо, wordnet, словосочетание, **позволять**, ошибка, вид, **объектовый**, построение, поиск, **контекстовый**, язык, объект, понятие, качество, сеть, работа, вывод, связь, **значимый**, **называть**, грамматика, время, связанность, смысл, **семантический**, medsearch, capr, управление, процесс, длина, number, технология, число, база, значимость, **иметь**, рубрика, **ассоциативный**, встречаемость, класс, рассуждение, савс, **лекарственный**, tip, представление, **логический**, эксперт, логика, **конкретный**, рубрицирование, произведение, множество, проверка, **являть**, структура, тип, **макаревич**, функция, описание, решение, образ, создание, разработка, единица, день, **максимальный**, **языковой**, эталон, бд, text, метод, память, **and**, шаблон, лексема, **электронный**, the, средство, **именной**, **второй**, предприятие, **нетрадиционный**, диалог, продукция, сотрудник, **начальный**, интернет, **лингвистический**, значение, пример, человек, **следующий**, алгоритм, **научно-технический**, oracle, **комбинаторный**, **реальный**, **an**, окрестность, **описывать**, моделирование, **один**, запись, изделие, пользователь, **стоить**, **курейчик**, **московский**, обеспечение, сервер, **деловой**, доступ, производительность, компания, задание.

Таблица 1. Результаты оценки термин – не термин для кардинальных слов

Предметн. обл.	Число кард. слов	Число не терминов	Отношение
Иск. интел.	172	33	0.19
Сетевые операц. системы	162	32	0.20
СУБД	168	33	0.20
Психология	234	62	0.26
Философия	329	105	0.32

Построение предметного указателя

Кардинальные слова упорядочиваются по убыванию их тематических весов. Все нормы ТС подразделяются на группы по признаку вхождения в них кардинальных слов. Нормы ТС в группах упорядочиваются по возрастанию их длины (длина – число слов). В группах организуются подгруппы по признаку включения более коротких ТС.

Пример (для лучшей читабельности ТС приведены вручную к номинативной форме):

система
система ии
современная система ии
построение системы ии

история развития систем ии
система искусственного интеллекта
совершенствование системы искусственного интеллекта
современная система искусственного интеллекта

система понимания естественного языка

В результате группирования часть кардинальных слов может остаться без своих включающих словосочетаний. В таком случае для них организуются ссылки на соответствующие группы. Пример:

понимание->система->система понимания естественного языка

По сути – это алгоритмически простой вариант иерархического группирования терминов. Более сложные варианты рассмотрены в работе [1].

3.5 Контекстная мера ассоциативной близости

Для групп ТС, идентифицированных кардинальными словами K_i, K_j – это функция от трех параметров: N – числа общих текстов (содержащих хотя бы по одному элементу парадигмы из состава терминов различных групп), L_{min} – минимального расстояния между предложениями, включающими элементы парадигм различных групп и L – среднего расстояния между предложениями во всем множестве общих текстов.

Например:

$A(K_i, K_j) = aN / (1 + L \cdot L_{min})$, а – нормировочный коэффициент.

Очевидно, эту же меру можно использовать и для определения близости пары норм терминоподобных словосочетаний. В таком случае L и L_{min} – это среднее и минимальное расстояния между предложениями, включающими элементы парадигм норм ТС.

Тематическая важность группы ТС определяется как сумма весов норм ТС, входящих в группу.

4 Формирование терминологической базы знаний

4.1 О выборе языков программирования

Для реализации алгоритма формирования базы знаний (автоматического анализа текстов, определения словосочетаний и итогового формирования фактов в синтаксисе Пролога) выбран свободно распространяемый язык программирования Unicon (расширенный Icon), а прототип ИПС реализован на PDC Prolog v.5.2. Personal Edition.

Выбор основан на следующих соображениях: реализация процесса автоматического анализа текста с целью выделения ТС требует большого объема исходного кода. Язык Unicon обладает очень простым синтаксисом и чрезвычайно удобными и функционально мощными встроенными типами данных (таблицы, множества, списки, записи) и полным набором стандартных математических операций и встроенных функций. Компилятор Unicon строит программы, по скорости работы не усту-

пающими программам, реализованными на Си. Unicon мог бы быть идеальным языком быстрого прототипирования, если бы не проблемы с кириллицей в интерактивном режиме работы.

Основываясь на собственном многолетнем опыте работы с Prolog'ом, можем утверждать, что система программирования PDC Prolog v.5.2., удовлетворяя определению Объектно-ориентированных баз данных (СУБД+язык программирования) очень удобна для экспериментов с ИПС в силу логической простоты и высокой технологичности. В частности, организация всякого рода многоуровневых меню, управления процессами не представляет никаких сложностей из-за наличия множества специализированных встроенных предикатов. Основные усилия должны быть сосредоточены на организации баз данных. Данные могут быть представлены в виде простой и удобной системы внутренних баз данных, которые содержат факты Пролога. Базы могут модифицироваться, сохраняться в файлы и загружаться в оперативную память из файлов. Для коммерческих систем с очень большим количеством информации следует использовать систему внешних (бинарных) баз данных.

Таким образом, комбинация этих двух языков программирования полностью удовлетворяет потребностям исследователя при разработке прототипов интеллектуальных систем («экспериментальных стендов»), избавляя от необходимости обращаться к профессиональным программистам.

4.2 Основные этапы выделения ТС

На вход программной системы поступает список полных имен тестовых файлов, представляющих фиксированную предметную область (или некоторый объемный текст).

1. В каждом текстовом файле программа выделяет и классифицирует лексемы и строки. Затем, на основании полученной признаковой информации, распознает границы предложений.

2. Выделяются НЛС и доминанты.

3. Параллельно (последовательность не важна) определяются все возможные контактные устойчивые словосочетания. Если частота контактного словосочетания больше единицы, то в справочниках соответствующих предложений вхождения эта пара лексем заменяется на их сочетание и новая «лексема» добавляется в словарь текста. Процесс повторяется для лексемы-сочетания.

4. Используя морфоанализатор mystem, для каждого слова сочетания получаем морфологическую информацию.

5. Словосочетания отдельных файлов объединяются. Определяются веса словосочетаний, затем они группируются в парадигмы. Каждому элементу парадигмы сопоставляется список включений, содержащий наименования текстов, полные имена файлов и списки предложений вхождения для каждого файла.

6. Удаляются словосочетания, не содержащие доминант.

7. Оставшиеся словосочетания пропускаются через лексико-морфологический фильтр.

8. Пересчитываются веса словосочетаний и их норм.

9. Формируются файлы фактов предметной области в синтаксисе Prolog'a.

4.3 Коэффициенты фильтрации

В настоящее время база знаний содержит описания следующих предметных областей (и одной монографии, представленной десятью файлами):

«СУБД», «Искусственный интеллект», «Философия», «Психология», «Сетевые операционные системы» (монография).

Количества текстов предметных областей (в порядке перечисления): 13,18,54,98,10,15.

Коэффициенты фильтрации (значение отношения числа всех отвергнутых словосочетаний к общему числу выделенных):

Психология – 0.39

Искусственный интеллект – 0.37,

Сетевые операционные системы – 0.34,

СУБД – 0.30,

Философия – 0.28.

Процент малоинформативных словосочетаний в итоговых множествах (по визуальной субъективной оценке) не превышает 2%.

Отметим, что в приведенном макроалгоритме принципиальной новизны обладает только п.6, выполнение которого обеспечивает адекватность выделенных словосочетаний предметным областям.

4.4 Оценка адекватности ТС предметным областям

Адекватность выделенных ТС предметным областям оценивалась путем классификации множеств норм словосочетаний, выделенных нами (контрольные множества), по предметным областям, представленным наименованиями статей словарей (эталонные множества словосочетаний) [9]. Принадлежность контрольных множеств к ПО определялась по максимальному суммарному весу норм словосочетаний, принадлежащих пересечению с эталонными множествами.

В качестве эталонных множеств словосочетаний были выбраны:

- а) нормированные наименования статей “Новейшего философского словаря под редакцией Грицанова А.А.” (<http://linguists.narod.ru/downloads5.html#spec>, 1390 наименований, “Философия-эталон”);

- б) нормированные термины “Психологического словаря” (<http://psi.webzone.ru/>, 2172 наименования, “Психология-эталон”).

- в) в качестве эталонного множества сочетаний, представляющих информатику, были выбраны нормированные наименования статей “Словаря компьютерной лексики” (<http://slovar.boom.ru/Head.html>, 1213 наименований, “КомпЛекс-эталон”).

В контрольные множества словосочетаний были включены нормы словосочетаний “СУБД”, “СетОпСист”, “Иск. Инт.” “Философия”, “Психология”. А также для контроля качества подборки мы проанализировали “Психологические теории и концепции личности. Краткий справочник.” (http://www.gumer.info/bibliotek_Buks/Psihol/Psi_Teo/index.php)

и нормированные двухсловные словосочетания включили в контрольную подборку (“ПсихТеор”).

Результат представлен в Таблице 2. Первая строка в каждой ячейке – результат до фильтрации, вторая – после фильтрации. Первое число в строке – суммарный вес словосочетаний контрольной подборки, принадлежащих множеству пересечения с эталонной подборкой, второе – размер пересечения.

Таблица 2. Результат классификации множеств словосочетаний.

Предм. обл. Эталон	КомпЛекс эталон	Философ. Эталон	Психология Эталон
СУБД	359.05 46 359.05 39		0.00 1
СетОпСист	82.05 69 82.05 56		1.79 2 1.79 1
Иск.Инт.	49.61 13 49.61 10		5.73 5 5.73 4
Философия	0.18 1 0.18 1	184.24 33 184.09 28	7.65 7 7.16 6
Психология	2.53 12 2.27 10	2.47 3 2.47 2	47.83 53 47.83 44
ПсихТеор		2.35 4 2.35 3	28.57 13 28.57 13

4.5 Некоторые характеристики программного обеспечения

Максимальный размер исполняемого модуля – 796 Кб.

Вычислительная система:

Intel(R) Core(TM)2 Duo CPU E6750 @ 2.66GHz
Память - 2 Гб.

Максимальное время – 30 сек для подборки текстов из предметной области «Философия», 54 текста общим объемом – 2 Мб.

5 База знаний

5.1 Структура

Структурно терминологическая база знаний организована в виде семантической сети, основу которой составляет дерево вхождений ТС в тексты и предложения, а группы ТС, идентифицированные кардинальными словами, связаны отношениями ассоциативной близости и тематической важности.

5.2 Файлы

Базу знаний системы составляют текстовые файлы, содержащие тексты предметных областей, тек-

сты энциклопедических и толковых словарей и файлы фактов (метаинформация).

В систему включены следующие словари:

«Словарь компьютерной лексики», «Современный энциклопедический словарь», «Современный философский словарь», «Все монархи мира. Др.Греция, Рим», «Все монархи мира. Западная Европа», «Биографический справочник», «Толковый словарь Ефремовой».

5.3 Файлы, содержащие тексты

Структура текстов в базе знаний отличается от структуры исходных текстов наличием номеров предложений (в угловых скобках, «тэгах») перед каждым предложением. Они добавляются автоматически в процессе распознавания границ предложений.

5.5 Файлы фактов

Множество файлов фактов составляют:

файл состава базы знаний, файл словарного состава, файлы – описатели вхождений ТС, файлы предметных указателей, файлы весов норм ТС.

Для сокращения описания доменного состава фактов введем следующие обозначения:

ПО – предметная область

ННТС – номер нормы терминоподобного словосочетания,

НТС – норма терминоподобного словосочетания,

ЭП – элемент парадигмы,

НТ – наименование текста,

ПИТФ – полное имя файла, содержащего текст,

СПВ – список номеров предложений вхождения,

НСл – наименование словаря.

Файл состава базы знаний содержит множество однотипных фактов

имена_баз(<имя_ПО>, <имя_файла_ПО>),

сопоставляющих наименованию предметной области OS-имя файла, содержащего факты, описывающие связи между нормами словосочетаний, множествами парадигм и множествами включающих предложений различных текстов.

Файл словарного состава содержит множество однотипных фактов, сопоставляющих наименованию словаря имя содержащего файла:

имя_словаря(<НСл>, <ПИТФ>).

Файлы – описатели вхождений ТС содержат факты следующих типов:

норма(<ННТС>, <НТС>)

элемент_парадигмы(<ННТС>, <ЭП>)

файл(<ННТС>, <НТ>, <ПИТФ>, <СПВ>)

Пример:

норма(12, "объект база данных")

элемент_парадигмы(12, "объект базы данных")

элемент_парадигмы(12, "объектом базы данных")

элемент_парадигмы(12, "объекта базы данных")

элемент_парадигмы(12, "объектам базы данных")

файл(12, "С.Д. Кузнецов. Введение в СУБД. Часть 7.", "D:\TEXTS\SUBD\bd_kuz7.ntf", [76, 99])

Файлы предметных указателей содержат факты следующих типов

группа(<КС><НТС>)

ссылка_на_группу(<КС1><КС2><НТС>),

где КС – кардинальное слово.

Пример:

группа("искусственный", "искусственный жизнь")

группа("искусственный", "база искусственный жизнь")

группа("искусственный", "изучение искусственный жизнь")

ссыл-

ка_на_группу("жизнь", "искусственный", "искусственный жизнь")

ссыл-

ка_на_группу("жизнь", "искусственный", "база искусственный жизнь")

ссыл-

ка_на_группу("жизнь", "искусственный", "изучение искусственный жизнь")

Файлы **весов норм ТС** содержат факты, сопоставляющие нормам ТС их веса:

вес_нормы(<НТС>, <вес_НТС>)

6 Терминологическая ИПС

Файлы фактов загружаются во внутренние базы данных прототипа терминологической информационно-поисковой системы (ИПС).

В настоящее время ИПС позволяет:

- выбирать предметные области и загружать ассоциированные с ней факты;
- вводить или выбирать интересующие пользователя ТС из групп или из списков ТС, отсортированных по алфавиту, либо по убыванию их весов;
- предоставлять пользователю все предложения вхождения элементов парадигмы ТС или группы ТС из всего множества текстов, представляющих выбранную предметную область;
- производить поиск непосредственно в текстах;
- искать необходимую информацию в словарях;
- строить отношения ассоциативной близости и тематической важности между группами идентифицированными кардинальными словами.

При запуске программы прежде всего загружаются файл состава базы знаний и файл словарного состава. Для дальнейшей работы пользователь должен выбрать предметную область. После этого загружаются факты, связанные с данной предметной областью. Имена файлов фактов образуются по правилу

< имя_файла_ПО_без_расширения>.<расширение>

Конструируются факты, задающие отношения ассоциативной близости и тематической важности между группами, идентифицированными кардинальными словами:

а_близость(<КС><значение_а_близости>),

т_важность(<КС><значение_т_важности>)

6.1 Пример использования фактов для вывода всех предложений вхождения элементов парадигмы ТС

1. Задается норма ТС (*НТС*).
2. Используя факт *норма(<ННТС>, <НТС>)*, определяется значение *<ННТС>*.
3. Отбираются факты *файл(<ННТС>, <НТ>, <ПИТФ>, <СПВ>)* с заданным значением *<ННТС>*.
4. Из каждого файла с ОС-именем *<ПИТФ>* выбираются предложения с номерами, содержащимися в списке *<СПВ>*.
5. Предложения с предшествующими наименованиями текстов (*<НТ>*) объединяются в «строку» и выводятся на экран.

6.2 Использование других фактов

Факт *элемент_парадигмы(<ННТС>, <ЭП>)* используется для поиска значения *<ННТС>* по заданному элементу парадигмы или при поиске *<ННТС>* по заданной подстроке.

Факты

группа(<КС><НТС>),

ссылка_на_группу(<КС1><КС2><НТС>)

используются для выбора *НТС* из групп после предварительного выбора кардинального слова, а также для построения отношений близости и тематической важности между группами, идентифицированными кардинальными словами.

6.3 Примеры вывода предложений вхождения.

Предметная область – СУБД, норма термина – «тип данный».

ТЕКСТ >> С. Д. КУЗНЕЦОВ. ВВЕДЕНИЕ В СУБД. ЧАСТЬ 9.

<119>Недаром в стандарте SQL появились специальные **типы данных** *date* и *time*.

<155> Соответствующий базис обеспечивают как предыдущие работы в области БД, так и давно развивающиеся направления языков программирования с абстрактными **типами данных** и объектно-ориентированных языков программирования.

<192>Как видно, при таком наборе базовых понятий, если не принимать во внимание возможности наследования классов и соответствующие проблемы, объектно-ориентированный подход очень близок к подходу языков программирования с абстрактными (или произвольными) **типами данных**.

<440>Такая информация в случае реляционной базы данных сохраняется в системных отношениях-каталогах и содержит главным образом имена базовых отношений и имена и **типы данных** их атрибутов.

ТЕКСТ >> С. Д. КУЗНЕЦОВ. ВВЕДЕНИЕ В СУБД: ЧАСТЬ 3

<22>Основными понятиями реляционных баз данных являются: **тип данных**, домен, атрибут, кортеж, первичный ключ и отношение.

<26>4.1.1 **Тип данных.**

<27>Понятие **тип данных** в реляционной модели данных полностью адекватно понятию **типа данных** в языках программирования.

Предметная область – «Философия», норма термина – «право человек».

ТЕКСТ >> ВВЕДЕНИЕ В ФИЛОСОФИЮ. ДУХОВНАЯ ЖИЗНЬ ОБЩЕСТВА.

<216>В праве сформулированы **права человека**, например, право на сохранение и продолжение жизни, в то же время в праве отражены нравственные обязательства индивида.

<230>Во Всеобщей декларации **прав человека** записано: "Каждый человек имеет право на жизнь, на свободу и на личную неприкосновенность".

<235>Если же естественные **права человека** ущемлены, то он чувствует себя неуверенно, теряет способность к творчеству, тем самым сдерживая собственное развитие и развитие общества.

<237>Незыблемость **прав человека**, его чести и достоинства, его интересов, их охрана и гарантированность являются принципами правового государства.

Предметная область – «Психология», норма термина – «активация асфс».

ТЕКСТ >> И.СМИРНОВ, Е.БЕЗНОСЮК, А.ЖУРАВЛЁВ. ПСИХОТЕХНОЛОГИИ.

<2535>Л.В.Бережкова и В.М.Смирнов считают, что АСФС, как феномен долгосрочной памяти, характеризуется специфичным для нее паттерном сверхмедленных колебаний потенциала, который проявляется при **активации АСФС** и сопровождается эффектами уменьшения проявлений основного заболевания и повышением психической активности.

<2541>В этой работе указано, что АСФС-II можно формировать двумя путями: простые матрицы, когда на фоне однократного введения этимизола предъявляют сенсорную стимуляцию и последующие сеансы **активации АСФС-II** проводят только с помощью сенсорной стимуляции разной модальности, но равной частоты; и сложные матрицы, когда в структуру АСФС-II включают фармакологические препараты, например, для невротиков используют транквилизаторы, а для больных с фантомно-болевыми синдромами - анальгетики.

<2543>У здоровых напряженно работающих лиц через 4-7 сеансов **активации АСФС-II** объем непосредственной памяти возрос на 42%, оперативной - на 77%, индекс кратковременной памяти - на 92% (двойной тест), устойчивость внимания повысилась на 50%, пропускная способность зрительного анализатора возросла до 42% (корректурный тест), оптимизировалось эмоциональное состояние, улучшилась способность к обучению, повысилась психическая активность.

<2551>После однократной **активации АСФС-I** лечебный эффект превышает два года (Бородин Ю.С., Зайцев Ю.В., 1985).

6.4 О применимости меры ассоциативной близости

В настоящее время проводятся эксперименты по интерпретации запросов пользователя на семантической терминологической сети. Пока в качестве примера, иллюстрирующего применимость введенной меры ассоциативной близости (А) приведем по 10 ассоциаций с кардинальным словом «память», расположенных в порядке убывания ненормированного значения А, из трех предметных областей.

«СУБД»:

память-журнал 5
память-функция 5
память-файл 4
память-управление 2.86
память-число 1.54
память-страница 0.41
память-объект 0.21
память-организация 0.21
память-значение 0.21
память-кортеж 0.18

«Сетевые операционные системы»:

память-адрес 2.22
память-страница 1.58
память-сетевой 1.46
память-использование 0.68
память-область 0.57
память-пространство 0.50
память-управление 0.35
память-таблица 0.23
память-сервер 0.21
память-сообщение 0.21

«Психология»:

память-асфс 0.59
память-психика 0.16
память-семантический 0.07
память-уровень 0.05
память-информация 0.04
память-мозг 0.04
память-состояние 0.03
память-расстройство 0.03
память-исследование 0.02
память-реакция 0.02

6.5 О временных характеристиках

Из всех операций в данной реализации ИПС самой «времяемкой» оказалась загрузка словаря объемом 13.6 Мб – 3 сек.

Максимальная продолжительность построения отношений близости для всех групп – около 1 сек.

7 Выводы

Построен прототип интеллектуальной терминологической ИПС. Структура его автоматически сгенерированной базы знаний полностью отвечает сформулированным требованиям.

Адекватность выделенных терминоподобных словосочетаний предметным областям обеспечива-

ется, в основном, исключением словосочетаний, не содержащих доминантных лексем.

Реализован относительно простой и эффективный метод автоматического построения предметного указателя на основе выделения «кардинальных слов» словосочетаний.

Возможность представления пользователю множества предложений вхождения выбранного терминологического сочетания является хорошей основой для построения конспектов по «локальным темам» и может использоваться в автоматизированных обучающих системах.

Результаты первых экспериментов позволяют говорить о приемлемости введенной меры ассоциативной близости для интерпретации запросов пользователя как поиска путей между терминологическими словосочетаниями в семантической терминологической сети.

Литература

- [1] Виноградова Н.В., Митрофанова О.А., Паничева П.В. Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике // Труды 9-й всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 2007.
- [2] Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды 5-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2003, Санкт-Петербург, Россия, 2003.
- [3] Ин Ц., Соломон Д. Использование Турбо-Пролога : пер. с англ. – М. : Мир, 1993.
- [4] Загоруйко Н.Г., Налетов А.М., Гребенкин И.М. На пути к автоматическому построению онтологии (<http://www.dialog-21.ru/Archive/2003/Zagorujko.htm>)
- [5] Чанышев О.Г. О возможности построения онтологий на основе доминантных лексем: результаты автоклассификации текстов // Вестник Омского государственного университета. – 2004. – Вып. 3. – С.45–47.
- [6] Чанышев О.Г. Диссертация «Ассоциативная модель реального текста и ее применение для автогенерации баз знаний о связях» на соискание степени кандидата технических наук по специальности 05.13.16. – применение вычислительной техники, математического моделирования и математических методов в научных исследованиях // Омск, ОФИН СО РАН, 1998 – 120 с.
- [7] Чанышев О.Г. Ассоциативная модель реального текста и ее применение в процессах автоиндексирования // Труды Седьмой национальной конференции по искусственному интеллекту с

международным участием КИИ'2000. – М. :

Физ.-мат. лит., 2000. – С. 430–438.

- [8] Чанышев О.Г. Автоматическое извлечение кандидатов в термины предметной области из представляющих ее текстов // Информационные технологии. – 2008, №2. – С. 2–7.
- [9] О.Г. Чанышев. О распознавании фрагментов естественно-языкового текста // Вестник Омского государственного университета. – 2002. – Вып. 4. – С. 14–16.

Automatic construction of terminological knowledge base

O.G. Chanyshev

In the report the method of automatic construction and structure of the terminological knowledge base of the subject domain set by set of naturally-language texts are stated. The index, by clusterization term-like word-combinations by a principle of entering in them of the main thing ("cardinal") word automatically construction. The measure of associative affinity term-like word-combinations is entered. Process of extraction term-like word-combinations is briefly described. A major factor defining their adequacy to subject domains is presence most thematically important words - dominants. They also are defined automatically. The result of an estimation of adequacy is given. Also the result of definition of associative affinity of a cardinal word "память"(memory) for three subject domains is given.

* Работы выполнены по проекту № 1.4.2. ОМН РАН за 2007 г.

Применение библиографических ссылок в автоматизированном учете книгообеспеченности и в представлении результатов поисковых запросов

© Л.С. Тимонина, Т.С. Шарова

Библиотечный комплекс университета «Дубна»
biblioteka@uni-dubna.ru

Аннотация

В статье приведена методика расчета коэффициентов книгообеспеченности учебного процесса в ВУЗе с учетом библиографических ссылок на переиздания и требований к устареваемости литературы, рекомендуемой в качестве основной в различных циклах дисциплин. Рассматривается возможность использования технологии библиографических ссылок для представления результатов поисковых запросов. В работе описана реализация и обоснование предложенных технических решений в АИБС «МАРК-SQL».

Электронная каталогизация в библиотеках открывает все новые возможности предоставления информации о содержании фондов, расширяет сервисы, доступные читателю. Пользователю библиотечными базами данных уже не надо задумываться о том, какой каталог ему необходим: алфавитный, предметный, систематический, ему достаточно определиться с поисковым термом или их комплексом, задать этот вопрос информационной библиотечной системе и получить ответ в удобном формате.

Механизмы электронной каталогизации отлажены, регламентированы и структурированы. С помощью форматов семейства MARC можно составить библиографическое описание на любую единицу хранения фонда библиотеки. Тем не менее, есть в этой области библиотечного дела задачи, решение которых сделает фонд более доступным читателю.

Библиотечный комплекс университета «Дубна» более десяти лет занимается проблемами электронной каталогизации и автоматизации процессов библиотечного дела [1, 2].

Неотъемлемой частью работы любой вузовской

библиотеки является расчет коэффициентов книгообеспеченности на одного обучающегося, по направлениям и специальностям, по циклам дисциплин, по конкретным дисциплинам, формам обучения, уровню компонента (федеральный, региональный), видам учебной литературы и т.д. Эта информация необходима не только для комплектования фонда, но и предоставляется при лицензировании и аккредитации учебной деятельности ВУЗов [3].

Суть этой части библиотечной деятельности сводится к мониторингу баланса между количеством учебной, научной, справочной литературы в фонде по отраслям знаний, ее актуальности по дате издания, количеству обучающихся по специальностям, направлениям, ступеням обучения. Существуют нормативные требования к показателям книгообеспеченности учебных заведений, стандартизированные статистические формы отчетов [4]. Кроме того, достоверные и точные сведения о содержании фонда и его соответствии потребностям учебного процесса существенно влияют на качество библиотечного обслуживания.

При наличии электронного каталога, в котором отражен весь фонд библиотеки, полностью автоматизированной книговыдачи, задача автоматизации учета книгообеспеченности является весьма актуальной проблемой. На рынке программных продуктов автоматизации расчета показателей книгообеспеченности предлагаются различные решения: специализированное программное обеспечение, направленное только на решение конкретной задачи, либо дополнительные модули комплексных АИБС [5, 6]. В нашей библиотеке используется программный продукт НПО «Информсистема» АИБС «МАРК-SQL», которая включает в себя АРМ «Книгообеспеченность».

При внедрении выше указанного АРМа мы столкнулись с проблемой взаимодействия сведений о группах (перечень и состав), учебных планах и рекомендуемой литературе к дисциплинам, с фондом библиотеки, также находящимся в постоянной динамике изменений. Кроме того, при анализе фонда на книгообеспеченность учебного процесса необходимо учитывать сведения о количестве книг разных годов изданий, а в некоторых случаях, переизданий под разными заглавиями. Например:

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Асмус, В.Ф. История античной философии. — М.: Высшая школа, 1965

Асмус В.Ф. Античная философия (история философии). — М.: Высшая школа, 2003.

Кроме того, существует понятие аналогов изданий, представляющих собой издания, которые равноценны в учебном процессе при изучении какой-либо дисциплины для определенной специальности [7]. В этом случае каждый студент должен получить в библиотеке хотя бы один экземпляр из аналогичных изданий. В настоящее время все аналоги изданий непосредственно указываются в списках основной и дополнительной литературы к дисциплинам и учитываются как отдельные наименования учебной литературы.

Тогда коэффициенты книгообеспеченности рассчитываются исходя из количества экземпляров всех переизданий учебной и учебно-методической литературы из фонда ВУЗа с учетом степени устареваемости литературы по циклам дисциплин. Согласно требованиям Министерства образования РФ для дисциплин общегуманитарного и социально-экономического цикла в качестве рекомендованной литературы могут использоваться издания за последние 5 лет, по естественнонаучным и математическим дисциплинам — за последние 10 лет, по общепрофессиональным дисциплинам — за последние 10 лет, по специальным — за последние 5 лет [4].

Таким образом, при формировании списка литературы, обеспечивающей ту или иную дисциплину, необходимо учесть не только наименования необходимых изданий, но их переизданий. При этом список рекомендуемой литературы должен постоянно актуализироваться, так как происходит движение фонда — пополнение, списание и устаревание рекомендуемой литературы.

Конечно, автоматизированный учет книгообеспеченности по сравнению с традиционным бумажным подходом — огромный прорыв в библиотечном деле. Цель любой автоматизации — исключение из процесса человеческого труда там, где это возможно, поэтому, в рамках освоения АРМа «Книгообеспеченность» была поставлена задача разработать и внедрить технологию автоматизированного учета всех переизданий одной и той же книги с учетом степени устареваемости литературы по циклам дисциплин, а также без него.

Признак устареваемости может носить формальный характер, так как в учебном процессе используются учебные пособия, написанные ведущими учеными и специалистами более двадцати лет назад, неоднократно переизданные. Таким образом, задавшись целью автоматизировать учет книгообеспеченности вуза, было решено модернизировать электронный каталог, а именно, неким образом связать библиографические описания одной книги разных изданий между собой.

Очевидно, такая связь должна носить уникальный и однозначный характер, т.е. электронные документы в БД (библиографические описания) долж-

ны ссылаться друг на друга уникальным, неизменяемым в базе данных параметром или параметрами.

Как определяет ГОСТ 7.76–96, ссылка — это запись, связывающая между собой части документа, библиографические записи, классификационные деления, предметные рубрики [8].

В конкретной программной реализации базы данных АИБС «МАРК-SQL» каждый электронный документ, представляющий собой библиографическое описание, имеет уникальный номер DOC_ID, которой можно использовать для связи документов между собой в базе данных. Кроме того, в формате Marc21, используемого в МАРК-SQL, версии 1.10, есть специальное выделенное поле 0010 «Контрольный номер», который формируется с применением генератора случайных чисел и, таким образом, уникален для каждого электронного документа.

Можно отметить, что идея связанности электронных документов уже реализована разработчиками в АИБС «МАРК-SQL», когда речь идет об иерархической связи вышестоящего и нижестоящего документа, например, библиографическое описание статьи из журнала — это нижестоящий документ, а библиографическое описание самого журнала, источника статьи, — вышестоящий документ. Технология связывания таких электронных документов основана на поле 0010 «Контрольный номер»: данные этого поля источника статьи (вышестоящего документа) вносятся в поле 773w «Контрольный номер источника», в электронное библиографическое описание статьи (нижестоящего документа). Таким образом, можно сформировать подмножество статей одного источника, или получить информацию, просматривая статью, об источнике этой статьи в целом и его местонахождении [9].

В ходе работы для ссылок библиографических описаний друг на друга сначала использовался уникальный номер документа DOC_ID, однако выяснилось, что номер документа не сохраняется при передаче документа между двумя базами, что является существенным недостатком для использования этого поля как уникального идентификатора документа, так как обычно в библиотеках имеется несколько баз. Поэтому для формирования ссылок было выбрано поле 0010 «Контрольный номер», хотя программная реализация связей документов через это поле значительно сложнее, чем по номеру документа DOC_ID.

В качестве поля, используемого для хранения данных о ссылках на другие издания было выбрано поле 787 «Поисковый признак на единицу, связанную с описываемой прочими отношениями» формата MARC21, а именно повторяемое подполе 787w «Контрольный номер связанной записи».

В случае решения задачи связи аналогов изданий можно было бы использовать другие подполя 7XXw. Технологически задача обработки хранимых в таких полях данных решается аналогично рассматриваемой.

Автор	Название	Издательство	Дат...	Объем	Ссылки на издания
Вентцель Е.С., Овчаров Л.А.	Теория вероятностей и ее инженерные приложения	Наука	1988	480с.	RU/IS/BASE/232717317
Вентцель Е.С., Овчаров Л.А.	Прикладные задачи теории вероятностей	Радио и связь	1983	416с.	
Вентцель Е.С.	Теория вероятностей	Физматгиз	1962	564с.	RU/IS/BASE/236869650, RU/IS/BASE/242909885, RU/IS/BASE/2501761
Вентцель Е.С., Овчаров Л.А.	Теория вероятностей	Наука	1969	368с.	RU/IS/BASE/242909322
Вентцель Е.С.	Исследование операций. Задачи, принципы, методол...	Наука	1988	208с.	RU/IS/BASE/222189353
Вентцель Е.С.	Теория вероятностей	Высшая школа	1999	576с.	RU/IS/BASE/264534548, RU/IS/BASE/242909885, RU/IS/BASE/2501761
Вентцель Е.С., Овчаров Л.А.	Теория случайных процессов и ее инженерные прило...	Высшая школа	2000	383с.	
Вентцель Е.С.	Теория вероятностей и ее инженерные приложения	Высшая школа	2000	480с.	RU/IS/BASE/263563679
Вентцель Е.С.	Познание продолжается	Просвещение	1969	526с.	
Вентцель Е.С.	Исследование операций. Задачи, принципы, методол...	Дрофа	2006	208с.	RU/IS/BASE/264537645
Вентцель Е.С., Овчаров Л.А.	Теория случайных процессов и ее инженерные прило...	Наука	1991	383с.	
Вентцель Е.С.	Элементы теории игр	Физматлит	1999	67с.	

Рис. 1. Ссылки на другие издания в модуле «Поиск» книги Вентцель Е. С. Теория вероятностей

Допустим, в электронном каталоге есть описание четырех изданий учебника Вентцель Е.С. «Теория вероятностей». Тогда, в электронном библиографическом описании каждого издания в поле 787w «Контрольный номер связанной записи» будут содержаться контрольные номера трех других изданий этой книги (рис. 1).

Понятно, что процесс занесения ссылок электронных документов друг на друга необходимо автоматизировать, во избежание ошибок каталогизаторов и из-за монотонности и однозначности самого процесса заполнения поля 787w кодированными значениями. Поэтому для связки документов между собой были написаны скрипты для заполнения поля 787w в автоматизированном режиме.

Функционально процесс объединения ссылками документов состоит из следующих шагов: во-первых, отбираются в подмножество документы, которые должны быть связаны, во-вторых, последовательно выделяются каждый из документов подмножества и убираются все связи, существовавшие ранее, в-третьих, заполняется поле 787w «Контрольный номер связанной записи» контрольными номерами документов поля 0010 из отобранного подмножества.

Приведенный пример с учебником по теории вероятностей Вентцель Е.С. стал тестовым для отработки технологии связывания ссылками библиографических записей, описывающих разные издания одного и того же материала. О том, что есть несколько изданий этого учебника просто было известно. В принципе, выявить по памяти несколько десятков наименований учебников, базовых для высшего учебного заведения, не составляет для библиотекаря особого труда. Но как быть с другими, менее выдаваемыми и спрашиваемыми читателями изданиями, тогда как фонд библиотечного комплекса содержит около 100 тыс. библиографических описаний. Тем более что для задач книгообеспеченности необходимо заполнить ссылками все множественные издания.

При решении задачи тотального выявления всех библиографических документов, требующих дополнения ссылками друг на друга была сделана выборка из электронного каталога всех изданий, у которых при полном совпадении заглавия совпадали бы и авторы. При этом рассматривались только те документы, которые бы имели одинаковый тип документа, например, книги (поле 901a «Тип документа»). Для этого был разработан скрипт

FindByZagl_Autor.scpl, который для активного документа находит все документы, имеющие то же заглавие и автора.

В результате использования частичной автоматизации процесса поиска и объединения разных изданий ссылками был обработан библиотечный фонд порядка 40 тысяч наименований, что заняло у 3-х каталогизаторов, помимо исполнения основных обязанностей, около трех недель дополнительного рабочего времени.

В процессе формирования ссылочного аппарата, используемого при расчете коэффициентов книгообеспеченности, выяснилось, что также целесообразно его применение в представлении результатов поисковых запросов в качестве дополнительной информации читателю о других изданиях найденного наименования.

Обычно результат поиска информации зависит от сформированного поискового запроса, от полноты и правильности библиографического описания, аналитическая часть которого может быть составлена по-разному. Поэтому зачастую не все варианты изданий одного наименования бывают результатом библиографического поиска. Указание в библиографическом описании ссылок на другие издания позволит пользователю получить более полные результаты поиска. Например, результатом тематического поиска по ключевому слову «метод суперпозиции» является список документов, состоящий из двух позиций:

1. Гмурман В. Е. Теория вероятностей и математическая статистика — 6-е изд., стер. — М.: Высшая школа, 1998. — 480с.

2. Гурский Е.И. Теория вероятностей с элементами математической статистики. — М.: Высшая школа, 1971. — 328с.

При открытии первого документа с контрольным номером RU/IS/BASE/161266507, в результате доработки поисковой формы читателя, в модуле «Поиск» отобразится результат, показанный на рис. 2.

Рассмотрим теперь на тестовом примере расчет коэффициентов книгообеспеченности для типовой реализации и модернизированного алгоритма подсчета с учетом ссылок наименований.

Пусть к дисциплине «Теория вероятностей и математическая статистика», читаемой кафедрой высшей математики, в качестве основной рекомендуемой литературы указаны:

22.17я73
Г 559
Книга
Текст. материал

Гмурман, В.Е.
Теория вероятностей и математическая статистика : Учебное пособие для вузов (гриф) / Гмурман Владимир Ефимович. - 5-е изд., перераб. и доп. - М. : Высшая школа, 1977. - 480с. - Доп.: с. 451.- Прил.: с. 461.- Предм. указ.: с. 474.

Ключевые слова: Бернулли формула; вероятность; выборки; выборочный метод; геометрические вероятности; дискретная случайная величина; дисперсия; закон больших чисел; испытания; классическое определение; комбинаторика; Лапласа теоремы; марковские цепи; математика; математическая статистика; математическое ожидание; метод суперпозиции; моделирование; Монте-Карло метод; нормальное распределение; однофакторный дисперсионный анализ; параметры распределения; плотность распределения; показательное распределение; сводные характеристики; система двух случайных величин; случайные величины; случайные события; случайные функции; спектральная теория; статистическая проверка гипотез; статистические гипотезы; статистические оценки; статистическое определение; стационарные случайные функции; теорема сложения вероятностей; теорема умножения вероятностей; теория вероятностей; теория корреляции; функции распределения

Пункт книговыдачи	Сигла хранения	Всего	В наличии	Заказано	Операция
аб.	аб.	1	1	0	 Читатель не зарегистрирован

Кол-во экземпляров: всего - 1

Другие издания этой книги

1. **22.17я73**
Г 559 Гмурман, В.Е.
Теория вероятностей и математическая статистика : Учебное пособие для вузов (гриф) / Гмурман Владимир Ефимович. - 6-е изд., стер. - М. : Высшая школа, 1998. - 480с. - ISBN 5-06-003464-X.
аб. - 198 в наличии 137
ч/з(е) - 2 в наличии 2
2. **22.17я73**
Г 559 Гмурман В.Е.
Теория вероятностей и математическая статистика : Учебное пособие для вузов (гриф) / Гмурман Владимир Ефимович. - 7-е изд., стер. - М. : Высшая школа, 2001. - 480с. - ISBN 5-06-003464-X.
аб. - 221 в наличии 147

Рис. 2. Результат тематического поиска и открытия одного из связанных документов

1. Вентцель, Е.С. *Теория вероятностей*. — 6-е изд., стер. — М.: Высшая школа, 1999. — 576с.

2. Вентцель, Е.С. *Теория вероятностей и ее инженерные приложения*. — 2-е изд., стер. — М.: Высшая школа, 2000. — 480с.

В нашем ВУЗе эта дисциплина читается в течение года у групп второго курса 2011, 2012 и во втором семестре у группы 2013, численность каждой из групп составляет десять человек. Тогда стандартный отчет по книгообеспеченности дисциплины, предлагаемый в модуле «Книгообеспеченность» АИБС «МАРК-SQL 1.10» будет выглядеть как показано на рис.3.

В случае использования ссылок на переиздания учебной литературы этот отчет будет иметь вид, представленный в Приложении 1. Причем, так как в АИБС «МАРК-SQL 1.10» существует возможность указать блок дисциплин, к которому принадлежит каждая дисциплина, то для получения итоговых значений коэффициентов книгообеспеченности была разработана система маркировки красным цветом изданий, не удовлетворяющих требованиям к новизне литературы для соответствующих блоков. При этом считается коэффициент книгообеспеченности и с учетом требований к новизне, и без их учета, во втором случае получаемое значение также маркируется красным цветом.

Можно заметить, что рассмотренный выше пример достаточно прост, так как учебные издания из списка основной литературы не приписаны к ника-

кой более дисциплине. Обычно при расчете книгообеспеченности эта ситуация не учитывается, однако для отражения реальной картины спроса учебной литературы нами были разработаны отчеты, которые учитывают эту ситуацию и в перечне номеров групп указываются все уникальные номера, у которых в читаемых дисциплинах в списке основной литературы есть это учебное издание. Получаемые таким образом значения коэффициентов книгообеспеченности, с учетом переизданий в фонде позволяет получить наглядную картину обеспеченности студентов учебной литературой.

В заключение можно отметить, что внедренный механизм ссылок на другие издания еще недостаточно проработан и универсален. В данный момент ведутся работы по разработке системы поддержки актуальности ссылочного аппарата при динамически изменяемом фонде. В частности, продумывается технология оповещения каталогизатора при удалении библиографической записи, имеющей ссылки на другие издания, а также последующая автоматизированная корректировка связанных библиографических записей.

Технология применения библиографических ссылок в представлении результатов поисковых запросов работает в библиотечной сети АИБС «МАРК-SQL». В дальнейшем планируется использовать эти разработки для представления поисковых запросов работает в библиотечной сети АИБС «МАРК-SQL».

Книгообеспеченность дисциплины основной литературой

Кафедра Высшей и прикладной математики

Дисциплина Теория вероятностей и математическая статистика (Высшей и прикладной математики)

№	Литература	Курс	Номера групп	Семестр	Кол-во студентов	Кол-во книг	Кол-во книг/студента
1	Вентцель, Е.С. Теория вероятностей.	2курс; 2курс; 2курс	2011; 2012; 2013	Год; Год; Год	10; 10; 10 Всего:30	10	0.33
2	Вентцель, Е.С. Теория вероятностей и ее инженерные приложения.	2курс; 2курс; 2курс	2011; 2012; 2013	Год; Год; Год	10; 10; 10 Всего:30	10	0.33

Рис. 3. Книгообеспеченность дисциплины

В дальнейшем планируется использовать эти разработки для представления поисковых запросов на сайте Библиотечного комплекса, а также при реализации корпоративного каталога Университета «Дубна» и его филиалов.

В заключение следует сказать, что разработанный аппарат ссылок является мощным инструментом отображения структуры библиотечного фонда как для профессионалов, так и для читателей. Эта технология позволяет более точно анализировать фонд библиотеки с точки зрения текущего спроса на учебную и научную литературу, а также получать и прогнозировать реальные значения коэффициентов книгообеспеченности в ВУЗе с учетом переизданий и устареваемости рекомендуемой литературы.

Литература

- [1] Шарова Т.С. Автоматизированная книговыдача: мечты и реальность // Материалы 9-й Международной конференции «Крым — 2002 «Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества». — Судак; Ялта, 2002.
- [2] Шарова Т.С. Электронная коллекция в вузовской библиотеке / Т.С. Шарова, Н.А. Гольцева, Ю.А. Крюков // Университетские библиотеки: прошлое, настоящее, будущее: материалы Междунар. науч.-практ. конф. — СПб., 2003. — С. 105-111.
- [3] Приказ Министерства Российской Федерации от 10.11.2002 № 3906 «Об утверждении перечня документов и материалов, представляемых к лицензионной экспертизе при проведении комплексной оценки деятельности высшего учебного заведения».
- [4] Приказ Минобразования России от 11.04.2001 № 1623 «Об утверждении минимальных нормативов обеспеченности высших учебных заведений учебной базой в части, касающейся библиотечно-информационных ресурсов».
- [5] Доброва М. В., Белугина Е. А. Книгообеспеченность учебного процесса. Как ее подсчитать? // Шестая Межрегиональная научно-практическая конференция «Университетская библиотека в образовательном пространстве XXI века»

(28.11.2005 - 30.11.2005). — Электронная версия статьи на сайте <http://dspace.lib.usu.ru>

- [6] Сайты производителей АИБС и ПО расчета книгообеспеченности. Система автоматизации библиотек ИРБИС <http://www.elnit.ru>; АИБС МАРК-SQL <http://www.informsystema.ru>; АБИС «Руслан» <http://obs.ruslan.ru>; АБИС 1С «Библиотека ВУЗа» <http://www.vuz.mart.ru>; АБИС «Буки» <http://www.buki.yar.ru>
- [7] Палей Д. Э. Учет книгообеспеченности в библиотеке ВУЗа: проблемы и пути их решения / Палей Д. Э., Курчинский Д. Н., Смирнов В. Н. // Восьмая международная конференция «LIBCOM-2004», November 15–19, 2004.
- [8] ГОСТ 7.76–96 Комплектование фонда документов. Библиографирование. Каталогизация. Термины и определения.
- [9] Руководство пользователя АИБС «МАРК-SQL»/НПО «Информсистема».

The use of references in automated accounting of the availability of literature and the presentation of the results of search queries

© L.S. Timonina, T.S. Sharova

The article described method of calculating the ratio of teaching literature in university, taking into account the references to reissue and the requirements for novelty literature for the various cycles of disciplines. This work illustrated the possibility of use of technology bibliographic references to present the results of search queries and described the rationale and implementation of the proposed technical solutions in the automated library system “МАРК-SQL”.

Книгообеспеченность дисциплины
"Теория вероятностей и математическая статистика (Высшей и прикладной математики)"
основной литературой с учетом возможного использования этой литературы в других дисциплинах
Блок дисциплины: Математические и общие естественнонаучные дисциплины
Требования к устареваемости: 10 лет

Контингент студентов, прослушивающих эту дисциплину по семестрам:

Номер группы	Количество студентов
Осенний семестр	
2011	10
2012	10
Итого: 20*	
Весенний семестр	
2011	10
2012	10
2013	10
Итого: 30*	

* Итоговое значение количества студентов, прослушивающих эту дисциплину в семестре, равное нулю, означает, что данная дисциплина в этом семестре не читается. Тем не менее, в таблице будет показано распределение основной рекомендованной литературы для этой и других дисциплин по каждому из семестров.

Читает дисциплину кафедра "Высшей и прикладной математики"

Заведующий кафедрой Л.А.Муравей

Серым цветом выделены книги, не удовлетворяющие требованиям по устареваемости

№	Литература	Общее кол-во книг	Номера групп в которых эти книги используются как основная литература	Общее количество студентов использующих эти книги в качестве основной	Кол-во книг/ студента
Осенний семестр					
1	22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебное пособие для вузов (гриф) / Вентцель Елена Сергеевна. - 6-е изд.,стер. - М. : Высшая школа, 1999. - 576с. : ил. - Прил.:с.561.-Лит.:с.573.-Предм.указ.:с.574. - ISBN 5-06-003650-2. Экз.: 10 22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебник для втузов / Вентцель Елена Сергеевна. - 2-е изд.,перераб.и доп. - М. : Физматгиз, 1962. - 564с. Экз.: 1 22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебник для втузов (гриф) / Вентцель Елена Сергеевна. - 4-е изд.,стер. - М. : Наука, 1969. - 576с. - Прил.Табл.:с.561.-Лит.:с.573.-Предм.указ.:с.574. Экз.: 1 22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебник для втузов (гриф) / Вентцель Елена Сергеевна. - 3-е изд.,испр. - М. : Наука, 1964. - 576с. - Табл.:с.561.-Лит.:с.573.-Предм.указ.:с.574. Экз.: 1	Общее количество 13 Новые 10 Устаревшие 3	2011 2012	20	Книгообеспеченность с учетом степени устареваемости 0.50 без учета степени устареваемости 0.65

2	22.17я73 В 296 Вентцель, Е.С. Теория вероятностей и ее инженерные приложения : Учебное пособие для вузов (гриф) / Вентцель Елена Сергеевна, Овчаров Лев Александрович. - 2-е изд., стер. - М. : Высшая школа, 2000. - 480с. : ил. - (Высшая математика для вузов). - ISBN 5-06-003830-0. Экз.: 10 22.17 В 296 Вентцель, Е.С. Теория вероятностей и ее инженерные приложения / Вентцель Елена Сергеевна, Овчаров Лев Александрович. - М. : Наука, 1988. - 480с. - (Физико-математическая библиотека инженера). - ISBN 5-02-013748-0. Экз.: 5	Общее количество 15 Новые 10 Устаревшие 5	2011 2012	20	Книгообеспеченность с учетом степени устареваемости 0.50 без учета степени устареваемости 0.75
Весенний семестр					
1	22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебное пособие для вузов (гриф) / Вентцель Елена Сергеевна. - 6-е изд., стер. - М. : Высшая школа, 1999. - 576с. : ил. - Прил.:с.561.-Лит.:с.573.-Предм.указ.:с.574. - ISBN 5-06-003650-2. Экз.: 10 22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебник для вузов / Вентцель Елена Сергеевна. - 2-е изд., перераб. и доп. - М. : Физматгиз, 1962. - 564с. Экз.: 1 22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебник для вузов (гриф) / Вентцель Елена Сергеевна. - 4-е изд., стер. - М. : Наука, 1969. - 576с. - Прил.Табл.:с.561.-Лит.:с.573.-Предм.указ.:с.574. Экз.: 1 22.171я73 В 296 Вентцель, Е.С. Теория вероятностей : Учебник для вузов (гриф) / Вентцель Елена Сергеевна. - 3-е изд., испр. - М. : Наука, 1964. - 576с. - Табл.:с.561.-Лит.:с.573.-Предм.указ.:с.574. Экз.: 1	Общее количество 13 Новые 10 Устаревшие 3	2011 2012 2013	30	Книгообеспеченность с учетом степени устареваемости 0.33 без учета степени устареваемости 0.43
2	22.17я73 В 296 Вентцель, Е.С. Теория вероятностей и ее инженерные приложения : Учебное пособие для вузов (гриф) / Вентцель Елена Сергеевна, Овчаров Лев Александрович. - 2-е изд., стер. - М. : Высшая школа, 2000. - 480с. : ил. - (Высшая математика для вузов). - ISBN 5-06-003830-0. Экз.: 10 22.17 В 296 Вентцель, Е.С. Теория вероятностей и ее инженерные приложения / Вентцель Елена Сергеевна, Овчаров Лев Александрович. - М. : Наука, 1988. - 480с. - (Физико-математическая библиотека инженера). - ISBN 5-02-013748-0. Экз.: 5	Общее количество 15 Новые 10 Устаревшие 5	2011 2012 2013	30	Книгообеспеченность с учетом степени устареваемости 0.33 без учета степени устареваемости 0.50

Средняя книгообеспеченность дисциплины

Серым цветом выделены значения книгообеспеченности, рассчитанные без учета требований к устареваемости

Осенний семестр	0.50 0.70
Весенний семестр	0.33 0.47
За год	0.41 0.58

Метаданные в системе управления многоязычной лингвистической базой знаний

© Н.В. Лунева

Институт проблем информатики РАН
nl2@mail.ru

Аннотация

В работе рассматривается использование метаданных в системе управления функционированием многоязычной лингвистической базы знаний, их роль в процессе взаимодействия функциональных подсистем и блоков программного комплекса базы знаний. Метаданные применяются для описания обрабатываемых и служебных данных системы, их текущего состояния и состояния самой системы, а также для передачи информации между функциональными подсистемами и блоками многоязычной лингвистической базы знаний и управления их работой.

1 Введение

В настоящее время метаданные являются широко распространенным инструментом, используемым для:

- обмена сведениями между различными инструментами и подсистемами информационной системы,
- хранения информации о внутренней структуре, форматах и содержании информационных ресурсов системы,
- интеграции разнородных данных системы в единое информационное пространство,
- описания самой системы.

Многоязычная лингвистическая база знаний создается в рамках проекта «Интертекст» на основе параллельных научных и патентных текстов и мета-языковых лингвистических представлений и предназначена для отладки семантико-синтаксических представлений в лингвистических процессорах систем машинного перевода и обработки текстовых знаний. Концептуальные и лингвистические основы данного проекта изложены в работах Е.Б. Козеренко [1-6, 9-11], общее описание архитектуры и метаданных, интерфейса и функциональных компонент

многоязычной лингвистической базы знаний рассмотрены в работе [7].

В основе многоязычной лингвистической базы знаний лежит разработка системы правил фразовых структур, функциональные значения языковых единиц закодированы как метки фразовых структур, и типы атрибутов-значений определяют функционально-категориальную семантику. Множество языковых структур, представленных в виде синтактико-семантических комплексов, выстраиваются в иерархию правил. Отношения зависимости реализуются через механизм головных вершин фразовых структур, а сами фразовые структуры задают линейные последовательности языковых объектов [1, 10].

Программный комплекс лингвистической базы знаний обеспечивает ввод обрабатываемого текста и его разметку, распознавание функционально-семантических структур введенного текста, построение на их основе набора соответствующих структур для результирующих текстов и формирование самих текстов, а также подбор примеров из архивов базы знаний. Кроме того, обеспечивается управление функционированием системы и ее настройками, служебными и пользовательскими архивами, использованием словарей [7].

Данные, используемые лингвистической базой знаний, представляют собой набор разнородных массивов информации, хранящейся в различных форматах – коллекции параллельных текстов научных и патентных документов, коллекции «переводческой памяти», комплексы фразовых структур, словари, служебные файлы и пользовательские архивы. Метаданные столь разнородной информации образуют комплекс, каждая часть которого описывает свой специфический класс данных системы.

Метаданные в многоязычной лингвистической базе знаний используются как для описания массивов данных и коллекций системы, так и в сфере управления функционированием блоков и подсистем базы знаний, а также для организации связи между ними в процессе обработки текста (распознавания и трансфера) и взаимодействия с пользователем. Метаданные, описывающие текущее состояние программного комплекса в процессе выполнения задания (метаописание), формируются в самом начале сеанса работы пользователя на основе шаблона либо загружаются из вызванного при входе в систе-

му предыдущего сеанса работы. В процессе работы метаописание модифицируется, чтобы отображать текущее состояние системы и проекта, что позволяет при необходимости приостановить заданный процесс обработки текста и впоследствии продолжить его с того же самого места, а также изменять текущие настройки и параметры среды.

2 Метаописание в многоязычной лингвистической базе знаний

Наряду с использованием метаданных для описания архивов и коллекций многоязычной лингвистической базы знаний и их содержимого, метаданные применяются для организации взаимодействия функциональных блоков и подсистем между собой. Метаданные используются для управления работой программного комплекса многоязычной лингвистической базы знаний, процессами распознавания введенного текста, построения функционально-семантических структур, трансфера, подбора примеров и формирования целевого текста. Метаданные обеспечивают обмен информацией о состоянии и выполняемых действиях, требуемых для работы внутренних данных между программными элементами базы знаний, и управление коллекциями данных лингвистической базы знаний. Кроме того, метаданные хранят настройки программной среды на конкретную задачу и пользователя, обеспечивая взаимодействие с ним.

Набор метаданных, хранящий описание текущего состояния многоязычной лингвистической базы знаний в процессе выполнения конкретной задачи, естественно назвать метаописанием. Каждое метаописание уникально для пользователя, решаемой им задачи и стадии выполняемого процесса. В процессе работы программного комплекса метаописание постоянно модифицируется, чтобы отображать текущее состояние системы и задачи, что позволяет при необходимости приостановить заданный процесс обработки текста и впоследствии продолжить его с того же самого места, а также изменять текущие настройки и параметры среды. Принципиальная схема взаимодействия функциональных блоков и подсистем многоязычной лингвистической базы знаний посредством метаописания приведена на рис. 1.

При самом первом входе пользователя в среду базы знаний формируется исходное метаописание на основе стандартного шаблона и конфигурации. Настройки, выбираемые пользователем для удобства работы в системе и в целях обработки конкретного текста либо другого задания, отражаются в метаописании, как и задаваемые пользователем задачи и их состояние.

Когда пользователь приостанавливает свою работу и выходит из системы, метаописание сохраняется, и при следующем входе пользователь имеет возможность продолжить приостановленный процесс с той самой точки, где работа была прервана. С другой стороны, настройки среды «под себя», сде-

ланные пользователем, сохраняются в файле конфигурации, и при его следующем входе может быть сформировано исходное метаописание уже на их основе.

Функционирование многоязычной лингвистической базы знаний основывается на взаимодействии между собой ряда подсистем. При передаче обрабатываемого текста от одной подсистемы или блока



Рис. 1. Схема взаимодействия функциональных блоков и подсистем многоязычной лингвистической базы знаний через метаописание текущего состояния. Цифрами обозначены процессы:

- 1 – формирования метаописания задачи,
- 2 – управления функционированием,
- 3 – модификации текущего метаописания,
- 4 – настройки и управления системой, а также модификации метаописания.

другим вместе с ним необходимо передавать информацию о текущем состоянии обрабатываемых данных, среде обработки, требуемых или ожидаемых далее результатах и т.п. Передача такой информации от одной подсистемы другим, как и информации о том, какие именно подсистемы и службы базы знаний должны далее отработать над текущей задачей, выполняется через метаописание.

Подсистема обработки входного текста обеспечивает ввод исходного текста и его характеристик в пользовательскую среду базы. Далее выполняется его подготовка для подачи в подсистему распознавания функционально-семантических структур – распознавание внутренней структуры текста и его фрагментация на логико-структурные блоки – фразы и фрагменты фраз.

Подсистема распознавания структуры входного фрагмента обеспечивает предварительное распознавание словарного состава заданного фрагмента текста и его морфологическую разметку для дальнейшего семантико-синтаксического анализа по правилам комплекса фразовых структур входного языка, построенных на основе новой когнитивной трансферной грамматики, впервые представленной в работах [3,9], и специализированных словарей входного языка, и построение его формальной семантической структуры. Результаты анализа отображаются в заданной пользователем нотации для оценки соответствия построенных структур исходному тек-

сту. Предусматривается возможность коррекции получаемых результатов и выбор структуры для дальнейшей работы системы, наиболее релевантной тексту.

Подсистема построения набора релевантных структур целевого языка на основе формальной семантической структуры обрабатываемого фрагмента текста, комплекса фразовых структур целевого языка (в общем случае – целевых языков) и словарей строит формальную семантическую структуру для построения выходного фрагмента текста. В связи с многовариантностью трансфера возможно построение набора выходных структур, характеризующихся как разной частотой использования в целевом языке, так и контекстом использования.

Подсистема построения выходного текста (текстов) на основе полученных выходных структур строит фрагмент выходного текста или его варианты. Также пользователю могут быть предложены подходящие примеры текста для построенных структур на основе существующих в системе коллекций «переводческой памяти». В зависимости от режима задаваемых целевых языков («многоязычности») и накопленных коллекций пользователю могут быть предоставлены разнообразные наборы текстов на разных языках.

Описания структур входного и целевых языков представляют собой независимые комплексы иерархически выстроенных правил фразовых структур соответствующих языков, выстроенные в единой нотации на основе когнитивной трансферной грамматики. Между собой комплексы фразовых структур разных языков связаны системой межязыковых отсылок, устанавливающих возможные направления трансфера от одного языка к другому. Подсистема фразовых структур языков обеспечивает ведение соответствующих архивов структур и вызов в подсистемы распознавания и трансфера необходимых пользователю комплексов структур языка.

Подсистема отображения, кроме обеспечения поэтапного вывода на экран результатов работы системы, предусматривает возможность коррекции пользователем построенных фразовых структур и выбор наиболее релевантной структуры на разных этапах работы.

На каждом этапе обработки метаописание модифицируется – структуры метаописания дополняется информацией, характеризующей текущий этап обработки, его состояние и возможные дальнейшие действия. После загрузки входного текста из файла или ввода его вручную в соответствующем окне ввода подсистема обработки входного текста хранит в исходном метаописании характеристики входного текста. Кроме того, отображаются выполненные над текстом операции, при необходимости дополняется сделанное ранее служебное описание текста описанием его фрагментации, а также сохраняется описание текущего уровня обработки введенного текста.

Подсистема распознавания структуры входного фрагмента дополняет метаописание характеристи-

ками текущего этапа анализа и требуемых для анализа инструментов многоязычной лингвистической базы знаний. Там же сохраняются данные о полученных при анализе фрагмента текста результатах и о действиях пользователя над результирующими структурами. Аналогичным образом действуют все остальные подсистемы, получая в метаописании команды, запускающие их работу, и описания обрабатываемых данных, и в свою очередь модифицируя метаописание результатами своей работы и формируя очередные команды для продолжения работы над текущей задачей.

3 Метаданные управления и связи

Метаданные, используемые для управления работой многоязычной лингвистической базой знаний, образующие метаописание, можно разделить на следующие категории:

- метаданные состояния системы;
- метаданные процесса обработки;
- информационные метаданные.

Метаданные состояния системы описывают текущую конфигурацию многоязычной лингвистической базы знаний, работающего с системой пользователя, а также запрошенные им ресурсы. Описание текущей конфигурации включает в себя открытые пользователем окна и панели инструментария, их настройку и заданные объем и формат отображения исходных данных и результатов. Принадлежность пользователя к той или иной группе пользователей определяет объем его прав и возможностей при работе в системе, в первую очередь объем доступа к коллекциям базы знаний. В этой же секции метаописания хранится информация о текущих исходном и целевом языках, тематических областях, запрошенных словарях, коллекциях «переводческой памяти» и параллельных текстах. Одновременное использование нескольких словарей, а также различных коллекций, предполагает указание последовательности их применения (иерархии). Основные атрибуты описания состояния системы и их значения приведены в таблице 1. Атрибуты описания открытого окна приведены в таблице 2, характеристики открытых панелей в общем случае им аналогичны и характеризуют заданные параметры отображения панелей инструментов.

Табл. 1. Атрибуты описания состояния многоязычной базы знаний

Атрибут	Описание
Тип пользователя	Указывает, к какой группе пользователей принадлежит данный пользователь – администраторы, опытные или обычные пользователи, лингвисты и т.п.
Личные коллекции пользователя	Указывает место хранения личных коллекций и данных пользователя, своего рода «личное пространство» пользователя в системе

Число открытых окон	Указывает длину списка открытых (затребованных) пользователем окон
Список открытых окон	Содержит список открытых пользователем окон и их характеристики
Активное окно	Указывает последнее активное окно, с которым работал пользователь
Число открытых панелей	Указывает длину списка открытых (затребованных) пользователем панелей
Список открытых панелей	Содержит список открытых пользователем панелей и их характеристики
Активная панель	Указывает последнюю активную панель, с которой работал пользователь
Формат отображения текста	Указывает формат отображения исходного или целевого текста
Нотация фразовых структур	Указывает форму отображения распознанных фразовых структур для исходного фрагмента текста
Исходный язык	Указывает язык исходного текста (по умолчанию – русский)
Целевой язык	Указывает целевой язык, на который предполагается трансфер исходного текста. В режиме работы с фразовыми структурами может не использоваться
Научная область	Указывает терминологическую сферу языка научного документа или патента
Число открытых словарей	Указывает длину списка открытых (затребованных) пользователем словарей
Список открытых словарей	Содержит список открытых пользователем словарей и их характеристики
Порядок просмотра словарей	Указывает порядок просмотра словарей в процессе поиска требуемого слова

Табл. 2. Атрибуты описания открытого окна

Атрибут	Описание
Тип окна	Указывает, какого типа информация выводится в данное окно
Размещение	Указывает расположение окна на экране, заданное пользователем
Размер шрифта	Указывает размер шрифта, заданный пользователем

Метаданные процесса обработки текста хранят информацию о состоянии текущего этапа обработки и его фазе, результатах и параметрах дальнейшего процесса. Процесс обработки текста естественным образом разбивается на ряд этапов, описываемых метаданными:

- ввод исходного текста и его характеристик в пользовательскую среду,
- фрагментация текста на логико-структурные блоки (например, фразы),
- распознавание-разметку текущего фрагмента текста по морфологическим характеристикам составляющих его слов,
- построение его формальной семантической структуры или набора подходящих для данного фрагмента структур,
- поиск подходящих примеров в коллекциях базы знаний,
- построение набора фразовых структур целевого языка.

Особняком стоят метаданные процесса, сигнализирующие о состоянии паузы в работе программного комплекса базы в связи с ожиданием команды пользователя. Основными ожидаемыми командами могут быть ввод параметров работы системы, ввод текста на обработку и запуск процесса, указание предпочтительной фразовой структуры в построенном наборе и т.п. Результатом подачи той или иной команды является модификация выполняемой задачи и, соответственно, модификация текущего метаописания либо смена выполняемой задачи, сопровождаемая созданием нового метаописания.

Каждый этап обработки исходного текста состоит из ряда фаз. Например, разметка текущего фрагмента текста для последующего построения фразовой структуры фрагмента разбивается на ряд фаз:

- выделение слова,
- определение типа слова,
- поиск базовой словоформы по словарям и определение атрибутов, характеризующих текущее слово,
- сохранение списка возможных словоформ и соответствующих им наборов атрибутов для данного слова.

Ряд атрибутов описания состояния процесса обработки текста программным комплексом многоязычной лингвистической базы знаний приведен в таблице 3.

Табл. 3. Атрибуты текущего процесса обработки

Атрибут	Описание
Команда	Указывает введенную пользователем команду и заданные им параметры
Этап обработки	Указывает, на каком этапе обработки текста находится текущая задача
Фаза этапа	Указывает текущую фазу этапа обработки
Инструмент	Указывает вызванный на данных фазе или этапе блок или инструмент для выполнения задачи

Сформированный массив списков словоформ с атрибутами подается на вход подсистеме построения фразовых структур для дальнейшей многофазовой

вой обработки. Построение набора формальных семантических структур планируется выполнять на базе интегрированных в многоязычную лингвистическую базу знаний системы автоматического синтаксического анализа и модуля унификации, описанных в [8]. Набор построенных фразовых структур может быть предъявлен пользователю для принятия решения о дальнейшей работе программного комплекса (например, выбора варианта, наиболее релевантного исходному фрагменту текста).

В многоязычной лингвистической базе знаний используется метод моделирования языковых структур, который вырабатывался в процессе создания целого ряда проектов: интеллектуальных сред обработки знаний (ДИЕС, ИКС, ЛОГОС-Д) [5,6,11], а также экспериментальной системы машинного перевода Cognitive Translator [9]. Основной принцип построения логико-лингвистической модели - это декларативное представление правил синтаксического разбора на основе гибридной унификационно-порождающей грамматики. В системах ДИЕС, ИКС, ЛОГОС-Д значения языковых конструкций представлены на глубинно-семантическом уровне в виде расширенных семантических сетей (РСС) и продукционных правил. При этом используется подход «синтаксической нормализации» [5,6]. Суть этого подхода состоит в том, что структуры с неличными формами глаголов приводятся к форме сложноподчиненных предложений с придаточными условия и причины. Это позволяет отобразить их в глубинно-семантические пропозициональные структуры.

В системе машинного перевода Cognitive Translator был разработан специализированный внутренний язык представления лингвистических знаний, поддерживающий механизм отображения, и отношений зависимости, и составляющих [9]. Лингвистическая база знаний ИНТЕРТЕКСТ предназначена для поддержки как поверхностно-синтаксических отображений, так и глубинных представлений языковых структур. Например, синтаксические правила для языковых конструкций со значением причины и условия представлены следующим образом [4]:

SYNRULE S/if_cl: S{fin+, sub_conj+} -> S{fin+,sub_conj+} PUNCT+ S{fin+,cl+}

Since there is a continual loss of micrometeoritic space because of the radiation effects, there must be a continual replenishment: otherwise micrometeorites would have disappeared from interplanetary space.

Поскольку имеет место постоянное сокращение микрометеоритного пространства по причине радиационных эффектов, должно быть постоянное пополнение: иначе микрометеориты исчезли бы из межпланетного пространства. (В связи с тем, что постоянно происходит сокращение микрометеоритного пространства, вызванное радиационными эффектами, по всей вероятности, происходит его постоянное пополнение, в противном случае, микрометеориты уже давно исчезли бы из межпланетного пространства).

SYNRULE S/vppt_if: S{fin-,sub_conj-} -> S{fin-,sub_conj-} PUNCT+ S{fin+,cl+}

Debugged, the program ran

Исправленная, программа заработала (Поскольку она была исправлена, программа заработала).

Информационные метаданные служат для передачи информационных сигналов от одной подсистемы или функционального блока другим подсистемам и функциональным блокам программного комплекса базы знаний, а также для организации информационных сообщений пользователю. Информационные метаданные хранят информацию о текущем состоянии процесса, обнаруженных конфликтных ситуациях (ошибках) и их разрешении, передаче задачи следующей подсистеме или возврат назад для повторной обработки, а также сообщения внутри подсистемы от одного функционального блока другому. Информационные метаданные могут обладать сложной структурой и включать в себя ряд параметров, характеризующих возникшую ситуацию и точку ее генерации. Основные атрибуты информационных сигналов системы приведены в таблице 4.

Табл. 4 Атрибуты информационных сигналов процесса обработки

Атрибут	Описание
Завершение	Указывает текущее состояние выполняемого/завершенного этапа
Завершение фазы	Указывает текущее состояние выполняемой/завершенной фазы
Следующий инструмент	Указывает требуемый далее инструмент
Конфликт	Указывает параметры ситуации, неразрешенной системой

4 Заключение

Применяемые в многоязычной лингвистической базе знаний метаданные, кроме обычного описания данных и коллекций, хранят текущее описание состояния базы знаний и выполняемых в ней действий, что позволяет организовать связи между подсистемами программного комплекса в зависимости от выполняемых заданий и управление их функционированием.

Наборы метаданных, описывающие текущее состояние многоязычной лингвистической базы знаний в процессе выполнения задания, организуются в виде метаописания, формирующегося в начале сеанса работы. В процессе работы базы метаописание модифицируется, чтобы отображать текущее состояние системы, что позволяет при необходимости приостановить заданный процесс и впоследствии продолжить его с того же самого места.

Каждая функциональная подсистема или блок базы знаний, получая доступ к метаописанию, считывает из него информацию о текущем состоянии работы над указанными данными и своих действи-

ях. В свою очередь, подсистема в процессе своего функционирования модифицирует метаописание, заносая в него результаты выполненного этапа или его фазы и входные параметры для дальнейших действий программного комплекса базы, которые служат командами для других подсистем. За счет организации через метаописание управления многоязычной лингвистической базой знаний достигается гибкая самонастройка всего комплекса на особенности каждой решаемой задачи.

Литература

- [1] Козеренко Е.Б. Лингвистические аспекты информатики // Системы и средства информатики. Специальный выпуск Научно-методологические проблемы информатики. Москва, ИПИРАН, 2006, 88-111.
- [2] Козеренко Е.Б. Логико-статистические методы представления языковых структур в машинном переводе. // Труды Международной конференции Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии", М.: "Наука", 2005.
- [3] Козеренко Е.Б. Моделирование переноса функциональных значений для англо-русского машинного перевода. // Труды Международной конференции Диалог'2004 "Компьютерная лингвистика и интеллектуальные технологии", М.: "Наука", 2004.
- [4] Козеренко Е.Б. Моделирование языковых структур со значением условия и причины для англо-русского лингвистического процессора. // Труды Международной конференции Диалог'2003 "Компьютерная лингвистика и интеллектуальные технологии", М.: "Наука", 2003, стр. 312-316.
- [5] Козеренко Е.Б. Унифицированные категориально-функциональные представления для синтаксической разметки полнотекстового документа // Системы и средства информатики. М.: Наука, 2002 Вып.11.
- [6] Козеренко Е.Б. Функциональная семантика в компьютерных решениях // Труды Международного семинара «Диалог'2002» по компьютерной лингвистике и интеллектуальным технологиям. г. Протвино Московской обл. 2002 т.1. Стр. 218-226.
- [7] Лунева Н.В. Архитектура и метаданные многоязычной лингвистической базы знаний // Системы и средства информатики. Вып.17. Москва, ИПИРАН, 2007, стр. 317-336.
- [8] Перекрестенко А.А. Разработка и программная реализация системы автоматического выделения синтаксических групп для естественных языков знаний // Системы и средства информатики. Вып.17. Москва, ИПИРАН, 2007, стр. 273-291.
- [9] Kozerenko E.B. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International

Conference on Machine Learning, Models, Technologies and Applications, June, 23-26, 2003, Las Vegas, USA.// CSREA Press, 2003. P. 49-55.

- [10] Kozerenko E.B. INTERTEXT: A Multilingual Knowledge Base for Machine Translation // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 25-28, 2007, Las Vegas, USA.// CSREA Press, pp. 238 - 243, 2007.
- [11] Kozerenko E.B. Portable Language Engineering Solutions for Multilingual Processors // Proceedings of the International Conference on Artificial Intelligence IC-AI'02// CSREA Press, 2002, pp. 447-453.

Metadata in the management system of multilingual linguistic knowledge base

N.V. Luneva

The use of metadata in the management system of multilingual linguistic knowledge base, their role in the interaction of functional subsystems and units of software of knowledge base are considered. Metadata are used to describe external and internal data of the system, their current status as well as the status of the system and to exchange information between functional subsystems and units of multilingual linguistic knowledge base.

Онтологии для тематического поиска данных в коллекциях электронной библиотеки*

© В.Т. Вдовицын, В.А. Лебедев

Институт прикладных математических исследований Карельского научного центра РАН
vdov@krc.karelia.ru

Аннотация

В работе рассматриваются вопросы практического построения и применения онтологий для организации тематического поиска данных в коллекциях электронной библиотеки. Представлен исследовательский прототип системы сервисов, обеспечивающих построение и редактирование онтологий, контроль целостности коллекций, а также формирование запросов и поиск релевантных документов.

1 Введение

Электронная библиотека Карельского научного центра РАН (ЭБ КарНЦ РАН) – <http://dl.krc.karelia.ru> предназначена для формирования, хранения, сопровождения и поиска электронных коллекций научных информационных ресурсов, отображающих результаты исследований природы и культурно-исторического наследия Карелии. В настоящее время ЭБ содержит коллекции по сосудистым растениям, афиллофороидным грибам, биотопам, млекопитающим, минеральным ресурсам и др. Кроме того создана и пополняется коллекция научных публикаций сотрудников, изданных редакционно-издательским отделом КарНЦ РАН. Предполагается дальнейшее развитие ЭБ как в плане расширения состава и содержания коллекций, так и совершенствования программных сервисов поддержки процессов формирования, хранения, сопровождения и поиска данных.

В предшествующих публикациях [4, 6, 7] изложены технологические основы построения ЭБ с применением XML. В данной работе основной акцент сделан на развитии сервиса поиска данных с применением онтологий.

Первоначально доступ к данным коллекций ЭБ осуществлялся выбором из списков названий документов в коллекции или набором на клавиатуре названий некоторых рубрик документов. Впоследствии выяснилась актуальность тематического подбора документов из разных коллекций по некоторому поисковому запросу, формирование которого не

должно представлять затруднений для пользователя. Таким образом, мы подошли к проблеме применения онтологий для поиска. Различным аспектам применения онтологий посвящен ряд наших работ [5, 11, 12], а также работы других авторов [2, 8–10, 13, 15]. Например, в обзорной работе [2] представлена, в частности, классификация моделей и языков для конструирования онтологий, а в обзоре [13] рассматривается использование онтологий для поддержки задач интеграции в семантически гетерогенных информационных системах и приводится сравнительный анализ инструментов.

Данная работа посвящена анализу состава, структуры и применения онтологий в ЭБ КарНЦ РАН.

2 Состав онтологий

Как известно, понятие «онтология» ввел в научный оборот информатики Т. Gruber [1], определив его как «точная спецификация концептуализации» предметной области. В дальнейшем в связи со сложностью реальных предметных областей и соответственно затруднительностью построения универсальных онтологий М. Uschold и М. Grundiger [3] предложили добавить в это определение уточнение – «с учетом задач информационной системы».

Назначением ЭБ КарНЦ РАН является представление в сети Интернет материалов по изученности свойств, состояния ресурсов региона в виде набора коллекций, содержащих описания классов однородных объектов, например таких, как минеральные ресурсы, водные объекты, виды животных и растений. При этом каждый такой класс определяется соответствующим понятием.

Понятие обозначается термином и является представлением класса реальных или абстрактных объектов и обладает смыслом и значением. В компьютерной информатике используется только значение, которым является совокупность (агрегат) свойств, атрибутов, признаков, характеристик класса объектов. Свойство (атрибут) – абстрактный объект, значение которого выбирается из некоторого именованного множества (домена). При этом домены могут быть различного типа: числовые, символьные, графические (рисунки, графики), составные и др.

Свойства и характеристики реальных объектов могут изучаться различными научными дисциплинами и играть различные роли в их описании или использовании. В соответствии с этим свойства мо-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

гут различными способами группироваться (агрегироваться), что порождает иерархическую тематическую структуру над свойствами. Тем самым вводятся дополнительные понятия. Помимо этого, как правило, однородность объектов в составе класса относительна, что часто вызывает необходимость выделения подклассов, то есть введения дополнительных понятий.

Таким образом, описание предметной области (области знаний) содержит некоторый набор понятий, которые вступают между собой в различные отношения: подчинения, включения, пересечения, ассоциации и др.

Отношения между понятиями обычно сводятся к небольшому числу классов отношений между терминами: классификации, агрегации, ассоциации. Классификации различаются по основаниям классифицирования (по их названиям), агрегации – по названиям целого. Среди ассоциаций важную роль играет отношение равнозначности, выражающее полную или частичную синонимию терминов. Обычно применяются две схемы выражения бинарных отношений между терминами: ARB или $R(A,B)$, где: R – имя отношения; A, B – термины.

Отношения между объектами предметной области обычно конкретны и представляются в виде некоторой характеристики класса объектов, в которой присутствуют выражения, содержащие термины, обозначающие другой класс объектов (реальных или абстрактных) или его характеристики. Такие характеристики задаются функциями, в частности, ролевыми функциями. Последние выражают роли некоторого класса объектов по отношению к объектам другого класса. Например, в экосистемах одни животные по отношению к другим могут быть врагами, конкурентами и т.п. Отметим, что совокупность конкретных отношений определяется назначением создаваемой информационной системы для решения заданного класса задач.

Для пояснения сказанного выше рассмотрим пример. Пусть необходимо создать систему «Определитель видов растений». Систематическая классификация растений основана на классификациях их морфологических признаков: подземных органов (корней, клубней и др.), вегетативных органов (стеблей, листьев), генеративных органов (цветков, соцветий, плодов и др.), то есть, на основе выделения таксонов по форме и размерам соответствующих органов. Каждый таксон именуется многословным термином, составляющими которого являются термины классификаций признаков данного органа по различным основаниям.

Существует несколько алгоритмов идентификации видов по выделенным признакам. В основном они схожи: выделяются таксоны ряда органов для определения класса, затем рассматриваются дополнительные таксоны для определения семейства внутри класса и, наконец, наиболее характерные таксоны используются для идентификации вида. Таким образом, процесс определения вида выполняется в два этапа – производится сравнение форм идентифици-

руемого растения с имеющимися в справочнике описаниями растений, затем формируется ответ – название вида растения (см. например [14]). «Не подготовленный» пользователь, как правило, не владеет терминологией, идентифицирующей, например, формы морфологических органов растений, поэтому целесообразно дополнить словарь терминов базой данных, в которой каждому термину ставится в соответствие изображение органов растений. Термины и изображения группируются в соответствии с иерархией, а также по стадиям вегетации растений. База данных должна содержать признаки каждого вида растений, сгруппированные по частям растений и стадиям вегетации. Таким образом, для решения задачи определения вида при помощи некоторой информационной системы на наш взгляд необходимо создать соответствующую прикладную онтологию, включающую систему формирования и исполнения запросов и базу данных, содержащую совокупность признаков каждого вида растения.

В данной работе предлагается система тематического поиска данных в коллекциях ЭБ, которая включает следующие основные компоненты:

- словарь связей терминов (таксономию) и их определений,
- сервис доступа к словарю с отбором релевантных терминов для включения в запрос,
- сервис формирования запроса в виде логического выражения (с визуальным контролем его корректности),
- сервис формирования по этому выражению оператора "Select",
- сервис поиска релевантных данных и выдачи их пользователю.

Таким образом, в состав онтологии в нашем случае включается словарь, отображающий иерархию терминов с альбомом изображений, соответствующих их названиям, и перечисленные выше программные сервисы, определяющие правила и ограничения использования этих терминов для построения пользователем тематических запросов к коллекциям ЭБ.

Предполагается, что система должна работать следующим образом. Пользователь сначала выбирает группу органов, например, подземные, вегетативные, генеративные органы. Далее для уточнения запроса ему выдается список, например, вегетативных органов: стебель, листья. А при выборе, например, листьев на экран выдается группа изображений листьев, в которой он должен сделать выбор и т.д. до формирования списка терминов, описывающих морфологию растения. На основе этого списка формируется запрос, который и подается системе исполнения запросов для поиска релевантных данных в коллекциях ЭБ и выдачи ответа.

3 Структура словаря

Коллекции ЭБ предназначены для описания состава, состояния и взаимодействий конкретных объектов (например, рек и озер) или классов (например, видов растений). Определение понятий, отношений и

механизмов управления, необходимых для описания процессов решения задач в избранной предметной области, выполняется специалистами предметниками в процессе разработки структур документов и баз данных. При этом выделение рубрик, атрибутов и их значений опирается на традиционную систему понятий данной предметной области. В итоге рубрикация документов и выражение содержания рубрик определяют структуру и состав словаря для данной коллекции. Как было указано выше, все понятия концептуальной схемы попарно связаны отношениями. Следовательно, словарь онтологии представляет собой совокупность триад вида $R(A,B)$ (или ARB), связи между которыми устанавливаются по тождеству написания названий отношений, первых или вторых терминов. В результате получается ациклический граф, в котором вершины помечены терминами, а дуги – названиями отношений.

Иерархию в этом графе определяют классы отношений классификации и агрегации. На нижних уровнях находятся термины, обозначающие характеристики и их значения. Выше располагаются агрегаты и классы в соответствии с их подчинением. Класс отношений равнозначности (синонимии) следует выделить из структуры основного графа, так как они могут проявляться на любом уровне иерархии и из-за нерегулярности излишне усложняют структуру словаря.

Среди терминов часто присутствуют многословные, например, «сухие еловые леса на сравнительно богатых почвах», «листья очередные простые яйцевидно-эллиптические тупые сидячие 0,4–2,2 см длиной, 0,2–1,1 см шириной». Наличие многословных терминов свидетельствует о многократной классификации данного класса объектов по различным основаниям. Включение многословных терминов непосредственно в состав словаря значительно усложнило бы процесс определения релевантности документов. Поэтому, на наш взгляд целесообразно включать в состав словаря каждую независимую классификацию отдельно.

Расположение независимых классификаций на одном уровне иерархии приводит к разрастанию соответствующего фрагмента словаря, делая его слабо обозримым, и затрудняет выбор нужных значений. Вследствие этого целесообразно на наш взгляд разносить их по уровням иерархии.

Различение отношений по названиям не всегда несет полезную информацию для пользователя. Связи между терминами часто настолько привычны, что для пользователя не требуется уточняющих сведений. В этом случае возникает возможность не помечать дуги графа, что значительно упрощает реализацию словаря.

Как известно, в общем виде граф G есть пара $\langle N, E \rangle$, где N – множество вершин, E – множество дуг. Если каждую дугу графа представить в виде пары смежных вершин, обозначенных терминами, и добавить к полученному множеству список висячих вершин, то иерархический граф словаря может быть представлен в виде «двухколонок» таблицы, в

которой в первой колонке представлен список «вершин-родителей», а во второй – список «вершин-потомков». Такой словарь может быть реализован в виде реляционной или объектной базы данных.

Существует две возможности реализации словаря: в виде множества нормализованных таблиц или в виде единой таблицы. В первом случае обеспечивается доступ к словарю по уровням иерархии, начиная с корня (поиск «в глубину»). Во втором случае становится возможным обращение к любому слову на любом уровне иерархии, если обеспечить предварительный поиск слова по набору его начальных букв. Однако, если структура словаря не древовидна, то есть содержит «полуциклы» (а это неизбежно при наличии многословных терминов), то возникают затруднения при построении сложных запросов. Это будет видно ниже при рассмотрении структуры запросов.

Сервис доступа к словарю должен обеспечивать последовательный отбор терминов от корня к листьям. Получающийся при этом список терминов сохраняет их иерархическую соподчиненность, что позволяет строить запросы, учитывающие иерархию терминов. Тем самым отсекается большая часть «информационного шума» при поиске.

Словари рассмотренного вида определяют объем и содержание каждого термина, обозначающего понятие, то есть его значение в коллекциях электронной библиотеки. Значения характеристик, за исключением тех терминов, семантика которых не требует специальных определений ввиду ее прозрачности для пользователей, могут определяться заданием соответствующих функций, рисунков или словесным определением (дефиницией) в зависимости от типа значений. Это позволяет использовать такие словари для решения задач поиска данных. Кроме того, такого рода словари могут использоваться и для контроля целостности коллекций ЭБ, то есть для выявления опечаток, неучтенных синонимов и др.

4 Структура запроса

Задача состоит в том, чтобы построить на основе словаря логическое выражение, семантически и синтаксически соответствующее аргументу оператора "Select". При этом логическое выражение должно удовлетворять следующим требованиям:

- Должна сохраняться иерархическая соподчиненность терминов. Это обеспечивает отсеивание «информационного шума» в тех случаях, когда термины, во-первых, могут появляться в разных документах в различных рубриках и на разных уровнях (например, название одного вида или рода животных может относиться к рубрике «систематика» в одном документе и к рубрике «враги, конкуренты, паразиты» в другом), во-вторых, компоненты многословных терминов могут присутствовать в текстах в другом словесном окружении.
- Должно четко проводиться различие между списками терминов, присутствие которых обязательно или желательно. Например, все компо-

ненты многословных терминов являются обязательными, что должно соответственно оформляться синтаксически.

- Связь между характеристиками и их значениями должна указываться явно.

В соответствии с этими требованиями принята следующая структура запроса:

(Вид объекта) [класс ... Л] [тема ... Л] [характеристика [= <список значений>]] <Л / V> ... <Л / V> [класс ... Л] [тема ... Л] [характеристика [= <список значений>]],

где: <список значений> ::= <значение> | <значение> Л <список значений> | <значение> V <список значений>

Нетерминалы, заключенные в квадратные скобки, представляют термины из их множества на соответствующем уровне иерархии. Они могут присутствовать или отсутствовать в запросе в соответствии с иерархией скобок. Многоточие после нетерминалов «класс» и «тема» обозначает, что в данном списке может быть один и более терминов, образующих иерархию. Нетерминал «Вид объекта» по существу представляет множество коллекций ЭБ и определяется в момент выбора коллекции, поэтому в списке он заключен в круглые скобки. Нетерминал «класс» означает наличие некой классификации, возможно иерархической, видов объектов (например, озер, видов животных), описания которых представлены в коллекции. Нетерминал «тема» указывает наличие агрегации свойств объектов в группы, например, по отношению к научным дисциплинам, изучающим их. Объекты в коллекции представлены описаниями их свойств и значений. В зависимости от интереса пользователя в данный момент запрос может содержать те или иные термины от одного до десятка и более.

Построение запроса по указанной выше схеме представляет определенные трудности для пользователя, поэтому в системе предусмотрен сервис построения запросов.

Реализация словаря в виде совокупности взаимосвязанных таблиц, содержащих дуги графа словаря, относящиеся к одному уровню иерархии, позволяет пользователю последовательно отбирать термины в процессе «поиска в глубину». В результате этого получается список терминов, релевантный запросу пользователя. Так как в процессе поиска нужных терминов в словаре, как правило, отбираются и избыточные термины, то пользователю предоставляется возможность не включать некоторые из них в запрос. В процессе формирования «суженного» списка терминов автоматически выполняется его пополнение синонимами. Далее работает сервис формирования запроса в соответствии с указанной выше схемой и с учетом иерархической соподчиненности терминов. По этому запросу выполняется поиск, в результате которого на экран выводится список релевантных документов (возможно, пустой) и вид логического выражения запроса, что обеспечивает возможность его корректировки. Процесс формирования запроса и вид отклика показаны на рис. 1.

The figure shows three screenshots of the 'электронная библиотека КарНИ РАН' (Electronic Library of the Karelian Research Center of the Russian Academy of Sciences) interface.

Top Screenshot: The main search page. It includes a logo, navigation links (о проекте, поиск, публикация, форум, ссылки), and a search section titled 'Поиск по онтологии'. The search section has buttons for 'выделить все' and 'убрать выделение'. Below is a 'Выберите коллекцию:' section with checkboxes for 'Афиллофориодные грибы Карелии', 'Виртуальная флора Карелии', and 'Млекопитающие Карелии'. There are also buttons for 'выбрать уровни онтологии' and 'очистить'. At the bottom, it says 'Последние изменения 30 апреля 2008' and 'webmaster@www.krc.karelia.ru'.

Middle Screenshot: A detailed view of the 'Поиск по онтологии' section. It shows a tree structure for 'Афиллофориодные грибы Карелии' and 'Виртуальная флора Карелии'. Under 'Афиллофориодные грибы Карелии', there is a sub-section 'Экология грибов' with checkboxes for 'Встречаемость и охрана', 'Местообитания', 'Особенности экологии', 'Распространение в Карелии', 'Субстрат', and 'Тип питания'. Under 'Виртуальная флора Карелии', there is a sub-section 'Экология растений' with checkboxes for 'Жизненная форма', 'Роль в растительном покрове', 'Хозяйственное значение', 'Экологические характеристики', 'Экотип', 'Распространение в Карелии', 'Встречаемость и охрана', and 'Местообитания'. There are buttons for 'выбрать уровни онтологии' and 'очистить'.

Bottom Screenshot: A window titled 'Поиск по онтологии' showing a list of selected terms on the left and a list of found documents on the right. The left list includes 'Лекарственное', 'Фанерофит', 'Ядовитое', and 'Подстилка'. The right list includes 'Подстилка'. Below the lists are buttons for 'вывести список найденных документов'. Below the window, there is a text block explaining the search process: 'Запрос формируется с использованием ключевых слов, показанных в левом окне. Как правило, используются термины нижних уровней онтологии.' and 'Конъюнкцией (И) соединяются ключевые слова, присутствие которых в документе обязательно. Дизъюнкцией (ИЛИ) соединяются слова, из которых в документе может присутствовать хотя бы одно.' It also includes instructions on how to use the interface: '■ Отметить курсором одно или несколько ключевых слов (при нажатой клавише Ctrl) и нажать кнопку [X]. Набор ключевых слов, образующих дизъюнкцию (ИЛИ) будет занесен в правое окно. Далее отмечаются следующие ключевые слова и переносятся в запрос. Они будут соединены в конъюнкцию (И) с предыдущими. (дизъюнкция (ИЛИ) от дизъюнкции отделяется в списке пунктиром, обозначающим связь И) и т.д.' and '■ Нажать на кнопку [X] вывести список найденных документов'. At the bottom, there is a text block: 'В случае, когда в коллекции не находится релевантных документов, следует вернуться назад и сформировать из того же списка ключевых слов более простой запрос. При неудаче нескольких попыток вернуться в начало и сформировать новый список ключевых слов для нового запроса.'

Рис. 1. Интерфейсы для формирования запроса и представления результатов поиска

5 Применение онтологии

Итак, целью применения онтологии в составе нашей ЭБ является обеспечение «дружественного» интерфейса с коллекциями ЭБ для организации тематического поиска объектов, чьи свойства соответствуют запросам пользователя в данный момент. При этом под термином «онтология» в нашем случае понимается словарь, отображающий иерархию терминов моделируемой предметной области с альбомом изображений, соответствующих их названиям, и перечисленные выше программные сервисы, определяющие правила и ограничения использования этих терминов для построения пользователем тематических запросов к коллекциям ЭБ.

Словарь формируется на базе ГРНТИ словарей предметных областей, описания которых представлены в коллекциях ЭБ. При этом в ряде случаев имеет место пересечение множеств терминов тех словарей, объекты коллекций которых взаимодействуют между собой (например, отражают взаимодействие организмов в экосистемах). Каждый предметный словарь является отображением концептуальной схемы описания объектов данной коллекции, принятой в соответствующей научной дисциплине.

Система сервисов формирования и исполнения запросов разрабатывается с использованием СУБД MySQL и обеспечивает не только поиск данных в коллекции, но и загрузку, редактирование словарей и некоторые другие функции, связанные с применением онтологий [5, 7, 12].

Как было показано выше, структура словаря представляет собой иерархический граф с полуциклами, в котором дуги обозначены парами смежных вершин (то есть, парами терминов). В этом случае граф словаря может быть представлен таблицей, число граф (столбцов) которой соответствует числу уровней иерархии, и которая естественным путем поддается декомпозиции.

Загрузка и декомпозиция каждого предметного словаря ЭБ, так же как и контроль его полноты и целостности (то есть, проверка связности графа, проверка отсутствия циклов и аутентичность написания терминов), выполняется автоматизированно. Эта операция реализуется посредством «поиска в глубину» экспертом-специалистом в данной предметной области, который имеет возможность самостоятельно редактировать словарь. На рис. 2 представлен интерфейс пополнения и редактирования словаря.

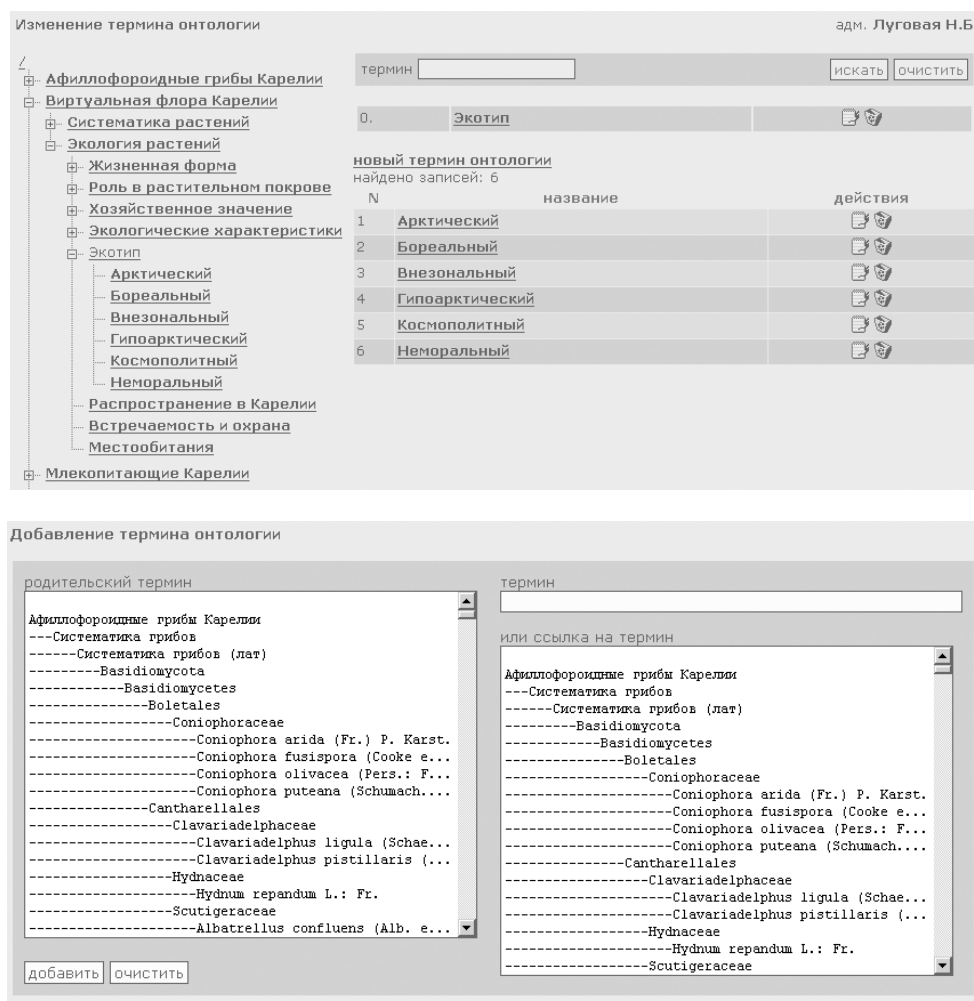


Рис. 2. Интерфейс для пополнения и редактирования словаря

Таким образом, совокупность тестируемых предметных словарей объединяется в единую базу данных на основе рубрикатора ГРНТИ с целью обеспечить пользователю возможность выбора интересующей предметной области. При этом использование разработанной таким образом онтологии предоставляет пользователю ряд следующих преимуществ:

- Возможность обозревать концептуальную схему предметных областей, представленных своими описаниями в ЭБ.
- Обеспечивает отбор терминов для запроса со всеми синонимами в аутентичном написании.
- Способствует построению «правильных» запросов.
- Обеспечивает переформулирование запроса в случае получения «пустого» отклика.

Таким образом, одной из основных функций применения онтологии в ЭБ является функция построения тематических запросов на поиск данных в коллекциях ЭБ, реализуемая по схеме, указанной в предыдущем разделе.

Документы коллекций ЭБ сформированы в виде: реляционных баз данных, слабо структурированных текстов в формате XML-DTD, а также в форматах HTML и RTF. В соответствии с этим организован и поиск данных, релевантных запросу – отдельно для реляционных баз данных, для XML-документов, для текстов HTML и RTF. Поскольку полнотекстовый поиск в документах сравнительно дорогая операция, он выполняется один раз после загрузки документов коллекции в процессе индексации документов, которая обеспечивает нахождение в текстах терминов соответствующей онтологии и формирование индекса.

Специальный сервис предусмотрен для выявления взаимодействий между объектами, которые фиксируются путем автоматического преобразования перекрестных ссылок в гиперссылки между документами. Для этого выделяется раздел онтологии, включающий, например, список названий «экологических отношений» между животными в экосистемах, то есть ролей (враги, конкуренты, паразиты, болезни и т.п.) и их значений (списки врагов, конкурентов, паразитов и т.п.). В процессе просмотра индексов выявляются соответствующие отношения и их значения, по ним определяется коллекция, формируется запрос к индексу и осуществляется поиск соответствующего вида животного или растения, а его логический адрес присоединяется к значению в исходной записи и т.д. до окончания просмотра всего индекса. После этого выполняется просмотр документов коллекций и по установленным логическим адресам формируются гиперссылки. Например, «враги зайца» – волк, лиса, сова и т.д., для формирования гиперссылок сначала в индексе находятся соответствующие им записи и их логические адреса присоединяются к названиям «врагов». Затем в коллекции находится документ «Заяц», в нем – рубрика «Враги» и для каждого «врага» на основе логического адреса формируется гиперссылка. Реализация данного сервиса позволит

автоматически связывать в единый комплекс набор коллекций, описывающих флору и фауну региона.

6 Заключение

В данной статье представлены результаты исследований и разработок, связанные с применением онтологического подхода для организации тематического поиска данных в коллекциях электронной библиотеки. Необходимость тематического поиска обусловлена желанием предоставить пользователю возможности построения «правильных» (т.е. адекватно отражающих на данный момент информационные потребности пользователя по интересующей его тематике) запросов и автоматического формирования из коллекций ЭБ релевантных запросу ответов в виде комплекса тематически связанных документов. Для решения поставленной задачи мы применили онтологический подход, суть которого в нашем случае состоит в построении словаря связей терминов (таксономии) и их определений, а также программных сервисов, определяющих правила и ограничения использования этих терминов для построения пользователем тематических запросов к коллекциям ЭБ. В перспективе планируется с помощью специалистов-предметников, как расширять словарь связей терминов и их определений для всех коллекций ЭБ, так и уточнять правила и ограничения использования этих терминов при построении тематических запросов с целью совершенствования программных сервисов формирования и исполнения таких запросов к коллекциям ЭБ. Исследовательский прототип системы основных сервисов, реализующий вышеописанный онтологический подход к организации тематического поиска данных в коллекциях ЭБ, представлен на сайте – <http://dl.krc.karelia.ru/ontology/>.

Литература

- [1] Gruber T. A translation approach to portable ontology specification. In. Knowledge Acquisition. Volume 5. (1993) 199-220.
- [2] Leonid Kalinichenko, Michele Missikoff, Federika Schiappelli, Nikolay Skvortsov Ontological Modeling // Proceedings of the Fifth National Russian Research Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" Saint-Petersburg, Russia, October 19-31, 2003. pp. 7-13.
- [3] Uschold M., Gruninger M. Ontologies: Principles, Methods and Applications // Knowledge Engineering Review. Volume 11. (1996).
- [4] В.Т. Вдовицын, А.Д. Сорокин, Н.Б. Луговая. Электронная библиотека научных информационных ресурсов КарНЦ РАН: состояние и перспективы развития // Труды Шестой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Пушкино, 29 сентября–1 октября 2004 г. с. 41-46.

- [5] В.Т. Вдовицын, В.А. Лебедев, С.В. Брагин, В.Г. Старкова, Н.Б. Луговая. Развитие сервисов электронной библиотеки. // Труды Всероссийской научной конференции «Научный сервис в сети Интернет: технологии параллельного программирования» (Новороссийск, 24–29 сентября 2007 г.). М., 2007. с. 305–310.
- [6] Вдовицын В.Т., Сорокин А.Д., Луговая Н.Б. Развитие программных сервисов и контента ЭБ КарНЦ РАН // Труды Седьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 4–6 октября 2005 года). – Ярославль: Ярославский государственный университет им. П.Г. Демидова, 2005. – с. 92–97.
- [7] Вдовицын В.Т., Лебедев В.А., Луговая Н.Б., Сорокин А.Д., Старкова В.Г. Разработка и развитие технологии публикации и поиска документов в электронных коллекциях // Труды Восьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Суздаль, Россия 17–19 октября 2006 г. – Ярославль: Ярославский государственный университет им. П.Г. Демидова, 2006. с. 162–167.
- [8] Гаврилова Т.А. Использование онтологий в системах управления знаниями // Технология менеджмента знаний
http://www.kmtec.ru/publications/library/.../use_ontology_in_suz.shtml?cmd=prin
- [9] Добров Б.В., Лукашевич Н.В. и др. Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам. // Труды седьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Ярославль, 4–6 октября 2005 г. с. 70–79.
- [10] Жыжырий Е.А. Щербак С.С. Математическое обеспечение систем поиска, основанных на онтологиях // Личный сайт, посвященный Semantic Web. Shcherbak.net/mat_ober/
- [11] Лебедев В.А., Старкова В.Г., Брагин С.В. Представление онтологии научной коллекции «Водные ресурсы региона» // Труды Шестой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Пущино, 29 сентября – 1 октября 2004 г. с. 86–92.
- [12] Лебедев В.А., Старкова В.Г., Брагин С.В. Применение онтологии для ведения и доступа к данным коллекций «Природные ресурсы региона» // Труды Седьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 4–6 октября 2005 г.). – Ярославль: Ярославский государственный университет им. П. Г. Демидова, 2005.
- [13] Овдей О.М., Проскудина Г.Ю. Обзор инструментов инженерии онтологий // Труды Шестой

Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Пущино, 29 сентября – 1 октября 2004 г.

- [14] Раменская М.Л. Определитель высших растений Карелии. Петрозаводск: Гос. изд-во Карельской АССР, 1960. С. 359.

- [15] Россеева О.И., Загорулько Ю.А. Организация эффективного поиска на основе онтологий // Международная конференция ДИАЛОГ. Компьютерная лингвистика и интеллектуальные технологии. <http://www.dialog-21.ru/>

Ontology for thematic search in digital library collections

V. Vdovitsyn, V. Lebedev

The paper considers problems of constructing and applying ontology for thematic search in digital library collections. The prototype of the system of services to facilitate ontology assembly and editing, control over integrity of the collections, as well as formation of and query search for relevant documents is represented.

* Данная работа выполняется при поддержке РФФИ (грант № 08-06-00085а).

О некоторых особенностях восприятия электронного документа

© В.Д. Шкилев, Н.А. Желева

Министерство информационного развития, Республика Молдова
schilov@registru.md

Аннотация

С позиций монадного символизма рассмотрен жизненный цикл архивов, бумажных и электронных документов.

Электронный документ – это относительно новый феномен виртуального мира. Этот феномен родился чуть позже рождения самого виртуального мира, а фактически стал замечен после внедрения Интранет и Интернет. Многие страны и научные направления дают разные определения этому понятию. Искать в этом множестве определений единственно верное было бы неправильным. Как утверждают философы, истина не может быть полностью познана при ее изучении с одной точки зрения. Это в полной мере относится и к изучению такого понятия как «электронный документ». Восприятие этого понятия также не может быть однозначным и сравнительный анализ множества определений понятия «электронный документ» может многое дать исследователю. Перечислять все известные определения в рамках небольшой работы просто неуместно, поэтому мы остановимся на определении этого понятия, принятого в Федеративном Законе Российской Федерации:

«Электронный документ – сведения, представленные в форме набора состояний элементов электронно-вычислительной техники (ЭВТ), иных электронных средств обработки, хранения и передачи информации, могущей быть преобразованной в форму, пригодную для однозначного восприятия человеком, и имеющий атрибуты для идентификации документа» [1].

Считать это определение единственно правильным было бы, как мы уже сказали, неверным. Какое принципиальное отличие электронного документа от привычного всем нам бумажного документа? Прежде всего – это документы из различных миров. Бумажный документ представляет наш обычный материальный мир, он существует в виде физической единицы. Электронный документ – это «бесплотный образ» или виртуально реальный документ [2]. Конечно мир материальный и мир виртуальный

не изолированы друг от друга. Часть электронных документов может быть распечатана на бумажном носителе, так же как и любая информация на бумажном носителе может быть оцифрована и представлена в виртуальном мире. Говоря о бумажном документе мы подразумеваем более широкое толкование – аналоговый документ.



Рис. 1. Жизненный цикл бумажных и электронных документов

Исходя из символа «ин-янь», обсудим жизненный цикл бумажных и электронных документов. Хотелось отметить, что распространенное восточное название символа «инь-ян» не совсем правомочно, поскольку практически все древние цивилизации, в том числе и древние славянские предки обращались к этому символу. Лучшим названием для этого символа является – символ мироздания. Именно о нем великий Пифагор говорил «монада – это все», в том числе и жизненный цикл электронных и бумажных документов. Монада легко объединяет символы восточной и западной философии [4], показывает единство всех религиозных символов [5], раскрывает единство кластерной и волной теории инновации [6], по-новому рассматривает экономическую безопасность государства [7–9].

Информация с символа считывается просто. Вертикальная ось символа – это стрела времени. На начальном периоде жизненного цикла (нижняя часть символа) преобладают бумажные документы, электронные документы существуют в зачаточном виде. Значимость электронного документа проявля-

ется в виде «островка» среди бумажных документов, причем положение островка соотносится с максимумом бумажных документов. Не случайно «островок» электронного документа соответствует пониманию информационного кризиса при использовании бумажных документов. После рождения «островка» значимость бумажных документов постоянно падает, жизненный цикл бумажных документов, как впрочем и бумажных архивов, постоянно сокращается. Но символ говорит и о том, что дальнейшее бурное развитие электронных документов породит новый «островок» бумажных документов. Так, например, брачное свидетельство, записанное на диске, возможно будет пользоваться меньшим уважением в обиходе, чем бумажный документ с этим же содержанием. Символ подтверждает, что в течении всего жизненного цикла документов будут всегда сосуществовать и бумажные и электронные документы, но значимость бумажных будет плавно сокращаться, а значимость электронных документов – возрастать. Символ также свидетельствует не только о противопоставлении бумажных и электронных документов (*архивов*), но и о их неразделимом единстве. Символ отражает философский закон единства и борьбы противоположностей. Символ, в отличие от знака, допускает индивидуальную трактовку информационного раскрытия.

Фактически до 2000 года электронные документы и бумажные документы в информационном смысле не отличались друг от друга. Можно ли считать электронный документ эквивалентным бумажному? Может при переходе на электронную форму документа меняется только форма записи? Такое восприятие электронного документа существует и будет существовать до того момента, когда пользователь будет видеть в вычислительной технике только удобную печатную машинку.

Но рано или поздно технологические преимущества электронного документа должны проявиться в полной мере. Что же это за преимущества, которых нет у бумажного документа? Может быть электронный документ дешевле? Нет, переход на электронную форму документа дорог. Проводить экономическое сравнение только сравнивая стоимость изготовления документов путь неверный. Так можно доказать преимущества телеги перед современным автомобилем. Сравнивать нужно не только стоимость изготовления документов, но и всю систему документооборота, включая поиск, хранение, тиражирование, транспортировку и т. д. С такой позиции эффективность использования информации в форме электронного документа уже не вызывает сомнений. Статистика это доказывает. Количество электронных документов в мире удваивается за год, в то время как бумажных документов увеличивается только на 7%. Наступит время, когда рост числа бумажных документов сначала приостановится, а затем начнет стремительно сдавать свои позиции перед электронными документами (*вся эта информация содержится в символе*). По оценке экспертов, через десять лет соотношение бумажных и

электронных документов будет выглядеть через пропорцию 30–70%. Что сегодня привлекает пользователя в электронном документе? Это, прежде всего, резкое увеличение производительности труда, быстрый и многоаспектный поиск информации. Да, средства для обработки бумажных документов имеются, но они не идут ни в какое сравнение с аналогичными средствами по обработке электронных документов.

И бумажный, и электронный варианты документов нуждаются в информационной защите. Электронные документы защищаются с помощью электронно-цифровой подписи, бумажные документы в последние годы получили возможность защиты такого же уровня [3]. Такая технология защищает не саму информацию в аналоговом документе, а только цифровой код аналогового документа. А есть ли недостатки у электронного документа? Возможно к недостаткам можно отнести то, что в электронной среде теряются такие понятия как «смысл», «знания» и т. д., а остается только детерминированная последовательность битов. Нас (*людей*) это пока не беспокоит по той причине, что при переводе информации из электронной версии на бумажную «смысл» и «знания» никуда не исчезают.

Анализ показывает, что технологические возможности электронного документа несомненно шире и это накладывает дополнительные требования к обеспечению государственных стандартов по документообороту, требует существенного изменения архивной практики. В связи с этим следует ожидать появление как минимум двух версий архивов – одна версия будет ориентирована на ведомственное потребление и будет основана на хранении CD-дисков, другая будет ориентироваться на внешнее потребление и основываться на использовании Интернет или Интранет. Вторая версия архива более перспективна, архивы на CD-дисках временное явление, ее можно воспринимать как первый этап построения электронных архивов. Конечно в зависимости от того закрытая или открытая информация хранится в архиве разделение архивов пройдет по той же схеме, что и бумажные архивы, то есть они разделятся на электронные архивы для Интернет и электронные архивы для Интранет. Это не случайность. Виртуальный мир будет развиваться по тем же базовым законам по которым развивается и наш материальный мир.

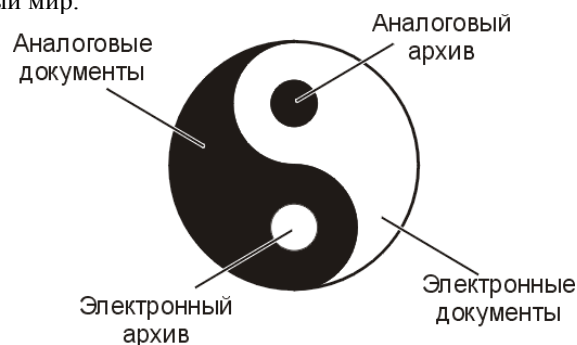


Рис. 2. Жизненный цикл архивов

Электронный архив – это качественно новый уровень использования информации, в том числе и документальной информации, он кардинально повышает эффективность ее использования. Это достигается за счет встроенных информационных поисковых систем и ряда других преимуществ (*возможность многопользовательского режима, обеспечение удаленного доступа к информации и т.д.*).

Существуют ли качественные изменения между восприятием бумажного и электронного документа? Они уже зарождаются, при переходе на электронную форму человек все меньше участвует в создании документа. В мире бумажных документов некоторое отчуждение документа от личности происходит при широком внедрении шаблонов. Но при переходе на электронный документ такое отчуждение будет стремительно нарастать, влияние человека будет заключаться в выработке правил формирования и преобразования электронного документа. Особенно заметен этот процесс уже сегодня в кредитно-финансовой сфере, где большинство финансовых электронных документов человек вообще никогда не увидит. В сферу деятельности человека входит изучение сводного документа, составленного компьютером на основе сотен электронных документов. На философском уровне обобщений это означает, что переход на электронную форму документа сопровождается качественно новым, структурным скачком, переходом на более высокий уровень абстрактного обобщения. Человек при этом освобождается от рутинной работы с документацией, ряд простых электронных документов уже сегодня создаются без участия человека.

Зададимся простым вопросом. Можно ли считать набранный на клавиатуре текст и выведенный на экран дисплея электронным документом? Большинство людей считает, что можно. Но это не так. На экране дисплея мы видим конечно бумажный (*аналоговый*) документ. Граница между электронным и бумажным документом проходит не здесь. Если набранный текст воспринимается человеком как осмысленная информация – то это бумажный документ и абсолютно неважно нацарапан текст на пергаменте или бересте, написан он на бумаге или набран на экране дисплея. Но экран дисплея в некотором смысле можно признать фильтром или границей, разделяющей мир бумажного и мир электронного документа. То, что мы видим на экране – это бумажный документ, а то что существует за ним – в виде строгой последовательности битов – это электронный документ. Где еще можно найти нечто эквивалентное фильтру? Это нечто существует в нашей голове, точнее между левым и правым полушарием. Не будем искать этому нечто медицинского эквивалента. Просто вспомним, что существует принципиально два качественно разных способа мышления: эмоциональное (*иногда называют образное*) и аналитическое (*абстрактное*), характерное для математиков. Эти два способа мышления присутствуют в каждом человеке, но, как правило, человек пользуется преимущественно одним типом

мышления. В случае использования образного мышления, которое характерно для людей искусства, идентификация объекта представляет собой практически мгновенный процесс. Образное мышление базируется на «похожести», а не на тщательном сравнении деталей образа. Каждый из нас помнит эпизоды из своей жизни, когда образное мышление давало сбой, когда убежденность в том, что ты узнал издавна знакомого человека продолжается вплоть до сближения. И только приглядевшись, то есть перейдя на другой режим мышления и начав анализировать и сравнивать, ты понимаешь, что перед тобой незнакомец.

Аналитическое мышление абстрактно и основано на логическом восприятии. Именно поэтому для этого типа мышления характерна однозначность результата. Это именно тот тип мышления, которым пользуется компьютер. Карпов побеждает специализированную на шахматах компьютерную программу не потому, что он может быстрее и логичнее просчитать множество комбинаций. Компьютер эти функции выполняет гораздо быстрее. Победа приходит к шахматисту за счет привлечения другого типа мышления – образного. Принципиальное различие этих двух типов мышления объясняет неэквивалентность электронного и бумажного документов. Способен ли человек непосредственно воспринимать электронный документ, то есть рисовать образные картины на основе просмотра машинного кода? Большинство людей такой способностью не обладают, но профессионалы высокого класса на это способны. А способен ли самый мощный компьютер воспринять бумажный документ, содержащий неоднозначные символы? Пока нет, поскольку компьютер не обладает эмоциональным мышлением.

Ответить на вопрос чем отличается бумажный документ от электронного – это фактически ответить на вопрос – чем отличается человек от компьютера. Электронные и бумажные документы функционируют в принципиально различных средах существования. Бумажный документ функционирует среди мыслящих людей, электронный – в среде программных и технических средств. Человек воспринимает Слово совсем иначе, чем компьютер. Для человека Слово – это символ, для компьютера – это просто слово – строгая последовательность битов. Если знаки – это изображения, несущие конкретное значение, то символы – изображения несущие более широкое толкование, чем знаки. Достоинством символов является возможность множественного трактования, зачастую с использованием противоречивых терминов и значений. Если выражаться более точно, то символы представляют не саму информацию, они являются носителями определенной концепции об информации. Говоря о символе информации, мы рассуждаем не об информации как таковой, а о представлениях о ней, которые могут быть индивидуальны. **Если говорить языком математики, то бумажный документ отображается в виде множества бесконечной мощности, в то время**

как электронный документ, построенный на последовательной цепочке детерминированных заключений, всегда будет множеством конечной мощности.

Именно поэтому, в силу конечности множества всегда можно создать абсолютно неразличимую «копию» электронного документа. Поэтому и принято утверждать, что применительно к электронному документу нельзя применить такое понятие, как «экземпляр». В мире строгой логики – это или один и тот же электронный документ, либо два различных. Именно отсюда возникают юридические осложнения при пользовании электронным документом. В среде бумажных документов не существует понятия «идентичности», здесь существует понятие «похожести». Осмысление бумажного документа принципиально является приближенной и связано это с простыми понятиями: число – это точность и однозначность, а образ (*символ*) – нечеткость и индивидуальность восприятия. Именно поэтому в бумажном документообороте и существует понятие «копия».

Философское обобщение позволяет сделать следующий вывод: в каком-то смысле среды обитания бумажных и электронных документов дуальны (*Рис. 1*). Но анализ понятия «документ» как бумажного, так и электронного, позволяет говорить о дуальности документа вообще. Дуальность или двойственность документа проявляется в том, что его можно воспринимать одновременно как вещь (*предмет*) и как информацию. Для бумажного документа это очевидно, информационный и вещественный аспекты настолько тесно связаны между собой, что не нуждается в каком-то обосновании. Другое дело электронный документ, где вещественный аспект не так очевиден. В электронной среде нет ничего кроме двоичных сигналов, поэтому нет качественного отличия между информационным и вещественным аспектом. И тем не менее электронный документ, в зависимости от условий можно считать или вещью или информацией. При этом наблюдается полная аналогия с понятием квантово-волнового дуализма в мире элементарных частиц. Это наверное можно сформулировать как понятие квантово-волнового дуализма электронного документа. Если перед пользователем стоит задача передачи документа, то электронный документ проявляет свои волновые качества, но если рассматривать правовые аспекты электронного документа, то это скорее квант (*вещь*). Человеку есть о чем задуматься, квантовые эффекты, которые проявлялись ранее в микромире все чаще становятся заметнее и в макромире.

Человек обладает знаниями, способностью обрабатывать информацию посредством мышления и более тонких процессов, которые в рамках данной статьи можно назвать интуицией. Компьютер этого лишен. Человек способен в нескольких мазках кисти гениального Пикассо увидеть быка, компьютер на это не способен. Конечно, существует быстро прогрессирующая теория распознавания образов, но «научить» распознать образ в той же мере, что и

человека, практически невозможно. Для этого нужно, прежде всего, познать самого себя, а это еще сложнее. Да, компьютер «умеет» различать достаточно сложные геометрические образы, но при встрече со сложными неоднозначными символами он «пасует».

Переход с бумажного документа на электронный можно сравнить по своей важности с изобретением письменности. Но это не последний перелом в восприятии информации. Можно допустить, что электронному документу суждена не такая уж продолжительная жизнь. Возможен ли при дальнейшем развитии полный отказ от документа в любой форме, от электронного архива? И да и нет. Вопрос терминологии. Можно ли признать за архив планетарную ноосферу? Если ноосфера – это архив, а ее пользователь – человек, имеющий неограниченные возможности при подключении и передачи бесконечной информации, если достигнутое соглашение между конкретными людьми становится достоянием всех и всегда может быть подтверждено (*вернее не нуждается в подтверждении*), то нужен ли документ вообще? Ответить утвердительно или отрицательно на эти вопросы пока нельзя, можно, как минимум, предположить, что необходимость в большей части документов естественным образом отпадет. Мы пока слишком мало знаем о информационных возможностях ноосферы. Слишком мало пока пользователей такой системы – людей, обладающих мышлением ноосферного масштаба. Исторически это связано с тем, что мы находимся на промышленном этапе эволюционного развития, который потребовал от нас строгой специализации (*деления на пекарей, слесарей, дипломатов и т.д.*). Но эволюционные процессы остановить невозможно и новый информационный перелом рано или поздно наступит для всех.

Литература

1. Закон Российской Федерации «Об электронном документе».
2. Закон Республики Беларусь «Об электронном документе» (от 10.01. 2000 г., №357) // Управление защиты информации. – 2000, Т. 4, №1. – С. 60–65.
3. Шкилев В.Д., Адамчук А.Н., Недиогло В.В. Электроразрядная технология защиты документов особой важности. Электронная обработка материалов. – 2008, №2. – С. 4–10.
4. Шкилев В.Д. О философском символе, объединяющем символы западной и восточной философии // Материалы XIII международного симпозиума по эниологии. – Алушта, 2004. – С. 226–229.
5. Шкилев В.Д. Этика построения единства религиозных символов. Всемирный Этический Форум. Международный Конгресс «Этика и гуманизм». – Алушта, 2005. – С. 272–279.
6. Шкилев В.Д. О единстве кластерной и волновой теории инноваций. Монадное понимание инно-

вационного процесса. Сборник докладов Международной конференции «Экономика и интеллектуальная собственность». – Кишинев, 2007. – С. 224–231.

7. Шкилев В.Д. О необходимости формирования в рамках национальной информационной системы единых принципов в идентификации объектов учета и контроля // Сборник «Информационные технологии» ВІТ+ 2003.
8. Шкилев В.Д. Информационные технологии и экономическая безопасность // Мысль. – 2006, сентябрь.
9. Шкилев В.Д. О монадном анализе инновационной экономики // Probleme teoretice ei practice ale economiiei proprienrtii intelectuale. – Chieinru, 2006. – S. 125–128.

About some features of perception the electronic document

V.D. Schiliov, N.A. Jeleva

The life cycle of archives, paper and electronic documents is reviewed from the point of monade symbolism.

Концептуализация данных наукометрических исследований в научных электронных библиотеках

© И.М. Зацман

Институт проблем информатики РАН
im@a170.ipi.ac.ru

Аннотация

Доклад посвящен проблеме концептуализации индикаторов сферы науки. Концептуализация индикаторов рассматривается как частный случай более общей проблемы обеспечения когнитивной интероперабельности в процессе совместной деятельности группы пользователей электронной библиотеки, то есть, обеспечение согласованного понимания ими информационных, программных и других объектов электронной библиотеки. Процессы концептуализации индикаторов анализируются в контексте совместной деятельности экспертов и ряда других групп пользователей, использующих электронные библиотеки для вычисления значений индикаторов и проведения наукометрических исследований. В докладе рассматриваются термины и семиотические основания постановки проблемы концептуализации наукометрических данных электронных библиотек и индикаторов сферы науки, а также предлагаемого подхода к ее решению.

1 Введение

Учитывая многозначность слова *концептуализация*, определим в явном виде значения двух словосочетаний: *концептуализация данных* и *концептуализация индикаторов*, которые являются ключевыми терминами для описания постановки *проблемы концептуализации данных электронных библиотек и индикаторов, вычисляемых на основе этих данных* (далее по тексту - проблема концептуализации).

Под *концептуализацией данных* понимается процесс формирования концептов, являющихся смысловым содержанием наукометрических данных и результатов их обработки, которые (концепты, данные и результаты обработки) используются экспертами в процессе анализа и оценивания результа-

тов научно-технической деятельности.

Под *концептуализацией индикаторов* понимается процесс формирования концептов, являющихся смысловым содержанием **новых видов** индикаторов и других количественных оценок.

Предполагается, что в процессе проведения анализа научно-технической деятельности возникает потребность в новых видах индикаторов, которые создаются и означиваются в рамках итеративного процесса их построения и вычисления значений на основе данных электронных библиотек.

Процесс концептуализации каждого нового индикатора сопровождается, как правило, разработкой алгоритма вычисления его значений и выбором названия, которое является вербальным обозначением формируемого концепта, а также может сопровождаться выбором табличных, образных или вербально-образных форм представления значений этого индикатора (например, в виде графика или диаграммы).

Таким образом, в докладе слово *концептуализация* обозначает не результат концептуального моделирования, а итерационный процесс формирования и изменения во времени смысловых значений индикаторов сферы науки. Описание процессов формирования и изменения смыслового содержания индикаторов влечет за собой необходимость рассмотрения ряда вопросов, связанных с семиотикой и лингвистикой.

Проблема концептуализации данных электронных библиотек, в том числе, частные задачи концептуализации данных наукометрических исследований и результатов их обработки являются относительно новыми. В частности, если искать на сайте журнала D-Lib Magazine (www.dlib.org) работы, в которых хотя бы один раз упоминаются любые вопросы концептуализации, то в результате выполнения такого запроса можно найти только около десяти публикаций или аннотированных библиографических ссылок на публикации в других журналах.

В найденных работах говорится о разных объектах (предметах) и средствах концептуализации, в том числе, о концептуализации пространства электронной библиотеки [1], концептуализации функциональности электронной библиотеки (ЭБ), включающей описание предметной области и абстрактной модели данных и процессов ЭБ [2], онтологии как средства концептуализации компьютерных дан-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

ных [3], концептуализации цифровых ресурсов [4] и концептуализации как процесса формирования нового взгляда на данные ЭБ [5].

Проблема формирования смыслового содержания индикаторов имеет отношение ко всем перечисленным вопросам концептуализации. Если же дополнительно необходимо учитывать еще и эволюцию семантики индикаторов, то появляются новые задачи концептуализации, касающиеся экспликации стадий эволюции семантики индикаторов во времени.

Кроме индикаторов, которым в докладе уделено основное внимание, существуют и другие категории показателей в сфере науки. В процессе описания проблемы концептуализации в общем случае различаются следующие пять категорий: *индикаторы*, для которых далее будет дано определение; *характеристики* ресурсов, используемых в процессе научно-технической деятельности (НТД), кадрового потенциала или инфраструктуры сферы науки; *нормативно заданные финансовые* и другие *параметры* в сфере науки (например, бюджет РАН); *критерии* принятия решений в процессе организации НТД; количественные и качественные *экспертные оценки* [6].

Основной целью доклада является описание семиотических оснований постановки проблемы концептуализации и предлагаемого подхода к ее решению. Возникает естественный вопрос: «Какое отношение имеют проблема в целом и отдельные задачи концептуализации индикаторов сферы науки к ЭБ?»

Речь идет не только о том, что в ЭБ могут храниться данные наукометрических исследований. Отличительной чертой предлагаемого подхода к решению проблемы в целом и отдельных ее задач является то, что этот подход основан на использо-

вании средств лингвистического обеспечения ЭБ, в том числе, классификационных систем, семантических словарей и тезаурусов.

В докладе предполагается, что значения индикаторов и других количественных оценок могут быть вычислены на основе наукометрических данных, полученных в результате обработки собранных информационных ресурсов ЭБ, многоаспектно характеризующих сферу науки.

Для иллюстрации различий между традиционными и новыми индикаторами, рассмотрим ретроспективные сведения о числе научных статей ученых России и ряда других стран, а также об общем числе статей, опубликованных в ведущих мировых журналах в 1994-2005 годах (см. табл. 1, сформированную на основе работ [7] и [8]).

Данные этой таблицы получены в результате обработки информационных ресурсов ЭБ Института научной информации США, которые представляют собой библиографические описания научных статей (включая пристатейные списки цитируемых работ), опубликованных в ведущих мировых журналах.

Числовые значения в ячейках 8 строки таблицы представляют собой удельные веса (в процентах) научных статей ученых России в общем числе статей, опубликованных в ведущих мировых журналах в период 1994-2005 годы. Иначе говоря, 8 строка таблицы содержит значения **традиционного** индикатора сферы науки «Удельный вес научных статей ученых России».

Если же поставить цель оценить распределение научных статей в зависимости от возраста их авторов, то такой индикатор будет **новым**. В настоящее время значения этого индикатора вычисляются только для отдельных научных коллективов, организаций или научных фондов, финансирующих научную деятельность.

Таблица 1

Country	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Worldwide	567 204	564 645	581 751	588 488	602 381	610 203	630 452	629 353	638378	661708	688615	709541
USA	199 769	193 337	193 153	189 752	190 395	188 004	192 743	190 592	190495	196427	202075	205320
Germany	36 901	37 645	39 213	41 415	42 953	42 963	43 509	42 674	42436	42228	43009	44145
China	7 821	9 061	10 526	12 172	13 781	15 715	18 479	21 134	23269	28768	34846	41596
India	9 928	9 370	9 753	9 618	9 945	10 190	10 276	10 801	11665	12461	13367	14608
Russia	21 612	18 603	18 553	18 133	17 166	17 145	17 180	15 658	15847	15146	14921	14412
Russia/ Worldwide	3,81%	3,29%	3,19%	3,08%	2,85%	2,81%	2,73%	2,49%	2,48%	2,29%	2,17%	2,03%

Прежде чем описать суть проблемы концептуализации, сформулируем три дефиниции.

Индикатором сферы науки будем называть один из трех следующих видов показателей в виде количественных оценок, вычисляемых на основе данных и информационных ресурсов ЭБ, собранных в процессе информационного мониторинга:

- указатель эффективности НТД, включая, как частный случай, эффективность научных и научно-технических программ;
- указатель степени достижения целевых (запланированных) результатов НТД;

- указатель инновационного потенциала направлений научных исследований и влияния результатов НТД на развитие сферы науки и других сфер или взаимосвязей с ними (здравоохранение, образование, инновационная система страны, экономика и т.д.).

Индикаторы применяются экспертами и другими пользователями ЭБ систем мониторинга в процессе анализа и оценки деятельности юридических и физических субъектов сферы науки, а также для оценки научно-технических программ и инновационного потенциала направлений научных исследований [9]. Предлагаемое определение термина «индикатор»

намеренно ограничено только тремя перечисленными видами показателей, так как для других их видов в работе [6] было предложено использовать перечисленные ранее названия других категорий показателей: характеристики, параметры, критерии и экспертные оценки.

Важно отметить, что основной целью исследования являются не все возможные виды индикаторов, а только те из них, которые являются новыми и находятся в процессе разработки и «кристаллизации» их семантики. Процесс концептуализации касается, в первую очередь, именно новых индикаторов, смысловое содержание которых эволюционирует во времени намного быстрее по сравнению с традиционными индикаторами. При этом понимание смыслового содержания новых индикаторов иногда может быть несогласованным даже между экспертами. Иначе говоря, новый индикатор, значения которого вычисляются по некоторому итерационно изменяемому алгоритму, в процессе «кристаллизации» его семантики разными экспертами часто интерпретируется по-разному.

Для более четкого описания проблемы в целом и отдельных задач концептуализации определим еще два термина.

Личностную концептуализацию индикатора определим как процесс формирования концепта одним из пользователей ЭБ данных наукометрических исследований, являющегося его пониманием индикатора в рамках его системы знаний о предметной области мониторинга и наукометрии. Предполагается, что результаты процесса формирования концепта находят свое отражение в лингвистическом обеспечении ЭБ в соответствии с некоторым регламентом.

Процесс личностной концептуализации конкретного индикатора может сопровождаться выбором для него названия, которое будет авторским вербальным обозначением соответствующего сформированного личностного концепта, а также выбором табличных, образных или вербально-образных форм представления значений этого индикатора (например, в виде графика или диаграммы), что также отражается в лингвистическом обеспечении ЭБ.

Коллективную концептуализацию индикатора определим как процесс формирования концепта в процессе совместной деятельности группы пользователей ЭБ (минимум, два пользователя), являющегося согласованным между ними пониманием смысла индикатора. Иначе говоря, смысловое содержание индикатора должно быть согласовано в пределах группы пользователей ЭБ в процессе их совместной деятельности.

Предполагается, что результаты этого процесса согласования находят свое отражение в лингвистическом обеспечении ЭБ в соответствии с некоторым регламентом.

Процесс согласованной концептуализации конкретного индикатора, как правило, сопровождается выбором для него названия, которое является вер-

бальным обозначением соответствующего сформированного концепта, а также выбором табличных, образных или вербально-образных форм представления значений этого индикатора.

Приведенные дефиниции позволяют описать суть проблемы концептуализации, которая заключается в том, что эксперты и другие группы пользователей, работающие с одной и той же ЭБ, иногда по-разному (несогласованно) интерпретируют одни и те же данные наукометрических исследований, а также результаты их обработки с использованием одинаковых процедур. В этом случае, одной из актуальных задач проектирования средств лингвистического обеспечения ЭБ данных наукометрических исследований, в том числе, классификационных систем, семантических словарей и тезаурусов, является обеспечение согласованного понимания пользователями идентичных объектов ЭБ, которые используются при обработке наукометрических данных.

При таком понимании сути проблемы концептуализации она может рассматриваться как частный случай более общей проблемы обеспечения когнитивной интероперабельности пользователей ЭБ в процессе их совместной работы [10]. В соответствии с определением Р. Будденберга реализация когнитивной интероперабельности в процессе совместной работы предполагает обеспечение согласованного понимания пользователями идентичных информационных объектов (идентичной информации) [11].

Важно отметить, что различия в интерпретации информационных объектов, наукометрических данных и индикаторов могут служить причиной конфликтов среди экспертов, целью совместной деятельности которых является получение согласованных оценок научных и научно-технических программ. В работе [12] был приведен краткий обзор методики экспертной оценки научно-исследовательских программ США и тех нерешенных проблем, с которыми столкнулись эксперты в процессе применения этой методики.

Анализу и обсуждению этих проблем был посвящен семинар по оценке федеральных научно-исследовательских программ США, состоявшийся 4-5 декабря 2003 года. В частности, в отчете по итогам этого семинара отмечалось, что между экспертами Административно-бюджетного управления при Президенте США и экспертами федеральных агентств США, финансирующих программы, достаточно часто возникают конфликты из-за различного понимания индикаторов эффективности реализации научно-исследовательских программ. Основной причиной упомянутых конфликтов нередко являются различия в понимании смысла индикаторов разными экспертами. В качестве превентивной меры (в целях предотвращения подобных конфликтов), в материалах семинара предлагается проводить семинары для экспертов [13].

Из краткого описания сути проблемы концептуализации уже видна та ключевая роль, которую

ЭБ играют в решении этой проблемы, так как результаты процессов концептуализации, в том числе, фиксация семантики индикаторов и ее эволюции во времени, предлагается отражать в лингвистическом обеспечении ЭБ данных наукометрических исследований, их классификационных системах, семантических словарях и тезаурусах.

2 Термины и семиотические основания постановки проблемы

В предыдущем разделе использовались термины «знания», «концепты», «информация (информационные ресурсы)» и «данные». Для более четкого описания проблемы концептуализации необходимо явно зафиксировать тот смысл каждого из этих терминов, который и будет использоваться при ее описании.

Начнем со слова *знания*, которое в докладе будем трактовать как результаты познавательной и креативной деятельности человека, носителем которых может быть только человек, и в которых могут быть выделены отдельные «кванты» знаний¹⁾.

Информацию определим как формы эксплицитного и отчужденного от человека представления его знаний, предназначенные для передачи, непосредственного сенсорного восприятия и понимания их другими людьми.

В процессе описания проблемы концептуализации основное внимание будет уделяться тем «квантам» знаний, называемых концептами, которые являются элементарными единицами или сочетаниями элементарных единиц плана содержания, выражаемого в рамках некоторого естественного языка (в общем случае, в рамках той или иной знаковой системы [14]).

Приведенные определения терминов «знания» и «концепты» предполагают, что они являются результатом процесса членения знаний человека на «кванты», которые могут быть выражены (обозначены) в рамках некоторой знаковой системы. Процесс членения неразрывно связан с процессом выражения знаний человека в сенсорно воспринимаемой и отчужденной от человека форме, например, в виде текста на естественном языке, диаграммы или в виде геоизображения на языке карты.

Отметим, что в системе знаний человека могут быть выделены несколько планов содержания в зависимости от того языка или той знаковой системы, которыми он пользуется для представления своих знаний в отчужденной форме. Элементарные единицы плана содержания, имеющие значение в рамках некоторой знаковой системы, будем называть элементарными концептами. В докладе предполагается, что все эксперты и другие пользователи ЭБ данных наукометрических исследований используют один и тот же естественный язык, то есть ис-

пользуют одинаковые планы содержания и выражения в процессе формирования смысла индикаторов.

Здесь необходимо отметить разницу между элементарным концептом и значением слова, когда речь идет о естественном языке. В процессе описания проблемы концептуализации рассматривается только сигнификативный аспект значения. Экспрессивно-эмоциональные оценки и коннотации не рассматриваются, так как речь идет о представлении знаний в виде научно-технических текстов, графиков и диаграмм, а не в виде художественных произведений. Следовательно, термин «элементарный концепт» используется для обозначения только сигнификативной составляющей значения слова [15].

Термин *знаковая информация* определим как результаты процесса представления концептов человеком-генератором этих результатов в плане выражения сферы социальных коммуникаций пользователей ЭБ в любой отчужденной форме, которая является сенсорно воспринимаемой другими участниками коммуникаций. Отметим, что при таком определении термин «знаковая информация» имеет отношение только к формам представления концептов, а введенный ранее термин «информация» - к формам представления любых «квантов» знаний, включая концепты.

Выделение в системе знаний человека нескольких планов содержания позволяет учитывать различия в членении знаний экспертов в разных естественных языках и других знаковых системах [16]. В дальнейшем учет одновременного использования разных планов содержания может понадобиться при решении проблемы концептуализации в процессе совместной работы пользователей ЭБ, говорящих на разных языках.

Термин «коды» определим как компьютерные эквиваленты двоичных цифр (или их последовательностей), которые могут представлять собой намагниченность или ее отсутствие, наличие электрического тока или его отсутствие, способность к отражению света или ее отсутствие в цифровой среде [17]. Двоичные цифры «0» и «1», о которых говорится в определении термина «коды» принадлежат сфере социальных коммуникаций ЭБ, а их компьютерные эквиваленты - цифровой среде.

При описании проблемы концептуализации будем выделять среди всех возможных кодов цифровой среды ЭБ три следующие категории:

- коды, соотнесенные с концептами знаний пользователей ЭБ (*первая категория*),
- коды, соотнесенные с эксплицитными и отчужденными от пользователей ЭБ формами представления концептов в плане выражения сферы социальных коммуникаций ЭБ (*вторая категория*),
- коды, соотнесенные с информационными объектами, алгоритмами, программами и другими категориями денотатов ЭБ (*третья категория*).

¹⁾ «Кванты» знаний являются переводом английского словосочетания «knowledge parts».

Например, для кодирования смысла индикаторов будут использоваться коды первой категории, для кодирования названий индикаторов - коды второй категории, а для кодирования программ вычисления значений индикаторов - коды третьей категории.

Также определим еще одну (нулевую) категорию кодов, которую будем называть «цифровыми данными» и к которой будем относить все коды цифровой среды, не относящиеся к трем выше определенным категориям. Таким образом, к цифровым данным будем относить результаты любых измерений, полученных с помощью цифровых компьютерных и других технических систем (устройств), и результаты вычислений, не являющихся формой представления знаний человека.

Формы представления цифровых данных в сфере социальных коммуникаций ЭБ в сенсорно воспринимаемом виде будем называть данными. Отметим, что данные как отображение в сфере социальных коммуникаций результатов цифровых измерений и вычислений могут сопровождаться метаданными, являющиеся одним из видов информации.

Перечислим рассмотренные термины, разделив их на три части в зависимости от их природы (ментальная, социальная и цифровая):

- знания, концепты и элементарные концепты (*ментальная сфера пользователей ЭБ*),
- информация, данные и знаковая информация (*сфера социальных коммуникаций пользователей ЭБ*),
- цифровые данные и три категории кодов (*цифровая среда ЭБ*).

Отметим, что определение термина «концепт» существенно зависит от дефиниции семиотического термина «знаковая система», так как любой концепт определяется в рамках той или иной знаковой системы. Однако по определению знака две его стороны (форма как материально выраженная его составляющая и значение как его идеальная составляющая), будучи поставлены в отношение постоянной связи, опосредованной сознанием, составляют устойчивое единство, которое посредством сенсорно воспринимаемой формы знака репрезентирует конвенционально приданное ему значение [18]. При этом, знаковая система определяется как совокупность знаков, объединенных между собой различными видами связей [14].

Таким образом, из определения знака следует, что концепты являются и конвенциональными, и стабильными, то есть, чтобы определить личностные и коллективные концепты пользователей ЭБ, необходимо предварительно определить понятия авторского и коллективного знаков.

Авторский знак отличается от традиционного семиотического знака тем, что две стороны авторского знака – форма и значение знака – могут находиться в отношении временной связи, опосредованной сознанием одного пользователя ЭБ, составлять нестабильное единство, которое посредством сенсорно воспринимаемой формы знака этим пользователем может репрезентировать персонально при-

данное ему значение в течение некоторого периода времени.

Коллективный знак отличается от авторского тем, что две стороны коллективного знака могут находиться в отношении временной связи, опосредованной сознанием каждого из пользователей-участников совместной деятельности, то есть используют и согласованно понимают коллективные знаки, как минимум, два пользователя ЭБ.

Авторские и коллективные знаки не возникают и тем более не функционируют раздельно. В своей совокупности с традиционными знаками они образуют единую знаковую систему, на основе которой проектируется лингвистическое обеспечение ЭБ данных наукометрических исследований. Иначе говоря, авторские и коллективные знаки означиваются в совокупности с традиционными знаками, являющимися составными элементами естественного языка или другой знаковой системы²⁾.

Используя определения авторского и коллективного знаков, определим элементарный личностный концепт как значение авторского знака, а личностный концепт как значение выражения на естественном языке, содержащего хотя бы один авторский знак, либо как новое значение выражения на естественном языке (например, дефиниция нового индикатора) без авторских знаков, новый смысл которого определен автором в явном виде.

Определим элементарный коллективный концепт как значение коллективного знака, а коллективный концепт как значение выражения на естественном языке, содержащего хотя бы один коллективный знак, либо как новое значение выражения на естественном языке без коллективных знаков, новый смысл которого определен в явном виде и согласованно понимается как минимум, двумя пользователями ЭБ.

Сопоставляя приведенные определения с дефинициями из предыдущего раздела, можно сказать, что в результате личностной концептуализации индикаторов получаются личностные концепты и авторские знаки, а в результате согласованной концептуализации индикаторов – коллективные концепты и знаки.

В заключение этого раздела отметим важное положение: из приведенных определений терминов следует, что соотносительность знаковой информации ЭБ, например, библиографических описаний научных статей, и обозначаемых этой информацией концептов, позволяющих определить источник публикации этих статей, по определению является конвенциональной для всех пользователей ЭБ, понимающих язык (языки) представления этой информации.

Однако данные и значения новых индикаторов, полученные в результате обработки этой информации, не всегда и не всеми пользователями ЭБ могут интерпретироваться согласованно. Поэтому, в на-

²⁾ Приведенные определения авторского и коллективного знаков являются развитием положений из работы [18].

звании доклада говорится именно о *концептуализации данных*, которые, в отличие от информации, в рамках рассмотренной системы терминов по определению не являются результатом процесса конвенционального представления знаний пользователей ЭБ.

3 Проблема концептуализации и виды обеспечения электронной библиотеки

В процессе разработки, вычисления значений и применения индикаторов сферы науки принимают участие разные группы пользователей ЭБ данных наукометрических исследований [19]:

- лица, принимающие решения,
- организационные и программные менеджеры,
- эксперты по областям знаний,
- специалисты по мониторингу,
- лингвисты и IT-специалисты, участвующие в разработке ЭБ.

Как отмечалось ранее, проблема концептуализации заключается в том, что пользователи, работающие с одной и той же ЭБ, иногда по-разному интерпретируют значения индикаторов, полученные в результате обработки одних и тех же ее цифровых данных и кодов ЭБ с использованием одинаковых процедур. Естественно, что различия в понимании индикаторов являются серьезным препятствием в проведении анализа и получении согласованных экспертных оценок результатов НТД.

Поэтому, уже в процессе создания ЭБ данных наукометрических исследований необходимо предусматривать разработку средств обеспечения согласованного понимания индикаторов в пределах как отдельных групп, так и всего сообщества пользователей этой ЭБ. Разработка средств формирования согласованного понимания является одной из ключевых задач проблемы концептуализации, от решения которой во многом зависят результаты наукометрических исследований. При этом, речь не идет обязательно о двухэтапном процессе: сначала личностная концептуализация, а затем коллективная. Целью разработки этих средств является формирование с самого начала совместной деятельности именно согласованного понимания данных наукометрических исследований и новых индикаторов.

В интересах разработки средств обеспечения согласованного понимания данных и индикаторов необходимо определить причины возникновения проблем, связанных с непониманием и/или различным пониманием разными пользователями ЭБ смысла новых индикаторов. Перечислим основные причины, используя работу [19].

Во-первых, названия значительного числа индикаторов, используемых в наукометрических исследованиях, не обладают той степенью конвенциональности, которая позволила бы определить значения этих индикаторов в процессе диалога пользователей ЭБ. Даже наличие дефиниции не всегда помо-

гает согласованному пониманию смысла индикаторов.

Во-вторых, существуют индикаторы, зависящие от нескольких параметров, изменение которых существенно изменяет их численные значения (например, значения индикаторов цитирования зависят от глубины ретроспективы используемого массива научных статей) и/или от выбора варианта используемого алгоритма вычисления, что может иногда изменить и смысл индикатора (например, учет самоцитирования и ссылок соавторов меняет значения и смысл индексов цитирования авторов).

В-третьих, численные значения индикаторов могут зависеть от числа, структуры и наполнения записей в используемых нормативных файлах (например, значения индексов цитирования зависят от конкретного списка журналов, используемых в процессе мониторинга и оценки), а также от текущего состояния семантического словаря ЭБ, включающего дефиниции индикаторов, ссылки на алгоритмы их вычисления, схемы информационных ресурсов ЭБ и используемых нормативных списков.

Приведенный перечень причин, не претендующий на полноту, иллюстрирует необходимость информирования пользователей ЭБ, проводящих наукометрические исследования, о зависимости смысла и прагматических аспектов использования индикаторов от разных видов обеспечения ЭБ. Например, пользователям ЭБ может понадобиться следующая информация о видах обеспечения ЭБ:

- реализованные алгоритмы вычисления индикаторов (алгоритмическое обеспечение),
- глубина ретроспективы используемого массива научных статей (информационное обеспечение),
- изменение наполнения семантического словаря (лингвистическое обеспечение),
- действующие нормативные акты и документы в сфере мониторинга и оценивания НТД (нормативное обеспечение).

Приведем пример, когда появление новых нормативных актов привело к существенному изменению прагматических аспектов использования удельного веса научных статей ученых России в общем числе статей, опубликованных в ведущих мировых журналах (см. табл. 1). В Стратегии развития науки и инноваций в РФ на период до 2015 года в качестве индикатора оценки степени достижения целевых результатов реализации этой Стратегии предложено использовать удельный вес научных статей ученых России [20]. В Стратегии указаны следующие целевые (запланированные) значения этого индикатора: 2004 - 2,7%, 2005 - 2,9%, 2006 - 3,1%, 2007 - 3,2%, 2008 - 3,4%, 2009 - 3,5%, 2010 - 4,0%.

Как видно из табл. 1 известные в настоящее время фактические значения этого индикатора меньше целевых: 2004 - 2,17% и 2005 - 2,03% [8]. Важно отметить, что до утверждения Стратегии удельный вес научных статей позиционировался только как одна из характеристик сферы науки, а после ее ут-

верждения он дополнительно приобрел роль индикатора степени достижения целевых результатов Стратегии.

С онтологической точки зрения удельный вес научных статей является вершиной двух разных отношений. В первом случае эта вершина входит в отношение «тип-тип» (удельный вес научных статей - это характеристика сферы науки), а во втором случае она входит в отношение «тип-роль» (удельный вес научных статей - это индикатор оценки степени достижения целевых результатов). Различие заключается в том, что удельный вес научных статей является характеристикой сферы науки до утверждения Стратегии, в течение ее срока действия и после его окончания, а его использование в качестве индикатора степени достижения целевых результатов определено только в течение срока действия Стратегии [21].

Вернемся к проблеме концептуализации и попытаемся сформулировать один из возможных методологических подходов к ее решению, а также определить, как и в какой форме описать и визуализировать влияние вариантов реализации разных видов обеспечения ЭБ на формирование согласованного понимания индикаторов и прагматические аспекты их использования. При этом желательно, чтобы выбранная форма описания была понятна пользователям ЭБ. Кроме того, одни и те же информационные ресурсы сферы науки, например библиографические описания научных статей, могут использоваться в ЭБ разных систем мониторинга. В таком случае, каким же образом можно сопоставлять значения и смысловое содержание индикаторов, вычисляемых с помощью информационных ресурсов ЭБ разных систем мониторинга, чтобы можно было взаимно верифицировать результаты наукометрических исследований, полученные разными коллективами?

Предлагаемый подход к решению проблемы концептуализации существенно использует идеи Г.П. Щедровицкого и Н.Г. Алексева, «связанные с противопоставлением двух планов, или аспектов изучения мышления – плана образов (или знаний) и плана "процессов" (или деятельности) ...». В работе [22] сформулирована задача «операционно-деятельностного анализа понятий и знаний, позволяющего, исходя из формы какого-либо сложившегося понятия или знания, сводить его к системе операций и действий, порождающих содержание этого понятия или знания (выделено мною. – И.З.)» [23].

Предлагаемый в докладе подход имеет два основных отличия. Во-первых, в нем учитывается то, что для индикаторов нередко отсутствуют конвенциональные формы представления их смысла, то есть, общепринятые названия индикаторов. Во-вторых, в нем учитывается то, что смысловые значения новых индикаторов не являются устоявшимися и общепринятыми. И, тем не менее, идея Г.П. Щедровицкого и Н.Г. Алексева о системе операций и действий, порождающих содержание понятий, вероятно, может быть использована в качестве ме-

тодологической основы решения проблемы концептуализации.

Общепринятое название и четкий смысл у нового индикатора могут отсутствовать, но если разработано соответствующее лингвистическое обеспечение ЭБ данных наукометрических исследований, то ее процессы, содержание и реализация видов обеспечения со временем помогут сформировать согласованное понимание этого индикатора в процессе совместной деятельности экспертов и других групп пользователей ЭБ.

Важно отметить, что, во-первых, процессы формирования согласованного понимания отчасти являются управляемыми; во-вторых, обеспечивая формирование согласованного понимания индикаторов, ЭБ выполняют важную и социально значимую функцию фиксации новых знаний в предметной области наукометрии.

Одной из форм описания влияния содержания и реализации видов обеспечения ЭБ на формирование согласованного понимания индикаторов могут быть полидоменные модели видов обеспечения ЭБ, фиксирующие стадии и результаты процесса концептуализации индикаторов. Их описанию посвящена работа [24].

Эти модели, которые позволяют описывать влияние разных видов обеспечения ЭБ на значения, смысл индикаторов и прагматические аспекты их использования, включают следующие компоненты [25]:

- 1) информационный компонент в форме набора схем информационных ресурсов ЭБ;
- 2) математический компонент в виде множества функций, используемых для вычисления значений индикаторов;
- 3) алгоритмический компонент, необходимый для определения значений параметров функций на основе информационных ресурсов ЭБ;
- 4) информационно-математический компонент, определяющий связи между параметрами функций и теми информационными ресурсами, которые необходимы для определения значений параметров;
- 5) лексико-семантический компонент, включающий используемые в ЭБ классификаторы, семантические словари и тезаурусы;
- 6) аналитический компонент в форме набора критериев принятия решений, построенных на основе индикаторов, методики построения критериев и правил их применения пользователями ЭБ;
- 7) нормативно-правовой компонент модели.

В полидоменной модели для каждого индикатора предлагается фиксировать в роли его денотата следующие составляющие: 1) схемы соответствующих ему информационных ресурсов, 2) функции для вычисления его значений, 3) алгоритмы для определения параметров функций на основе информационных ресурсов ЭБ, 4) связи между схемами, функциями, алгоритмами, параметрами функций и теми информационными ресурсами, которые необ-

ходимы для определения значений этих параметров [25].

4 Пример индикатора

В начале доклада была приведена таблица, которая содержит значения традиционного индикатора сферы науки «Удельный вес научных статей ученых

Проценты (%)

России». Был также упомянут новый индикатор, основная цель разработки которого заключается в том, чтобы оценить распределение научных статей в зависимости от возраста их авторов. Это даст возможность оценивать публикационную активность каждой возрастной группы ученых.

В настоящее время разработка нового индикатора ведется в рамках создания Информационной системы обеспечения реализации и мониторинга Программы фундаментальных научных исследований РАН. Идея разработки индикатора для оценивания распределение научных статей в зависимости от возраста их авторов возникла в процессе разработки средств вычисления традиционного индикатора «Возрастное распределение сотрудников научного коллектива (подразделения, организации)» (см. рис. 1).

На этом рисунке приведены значения индикатора в виде графика процентного распределения сотрудников одного из научных коллективов ИПИ РАН по 14 возрастным группам (20 – 24, 25 – 29 и далее до группы 85 – 89 лет).

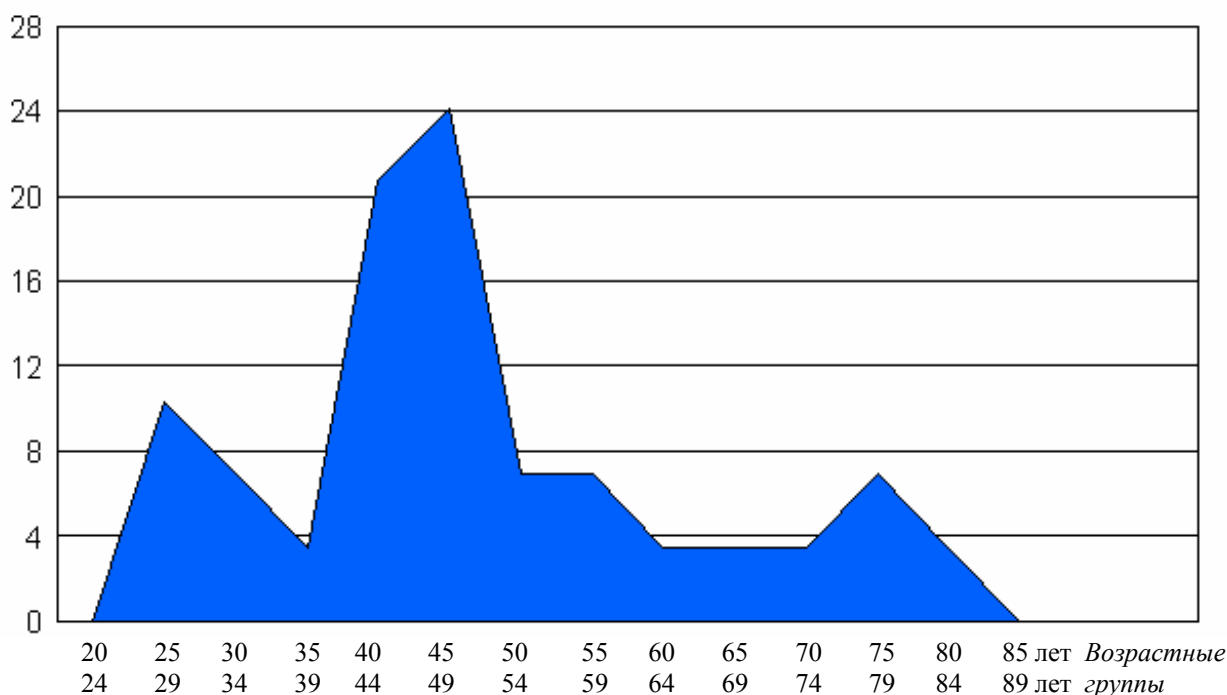


Рис. 1. График возрастного распределения сотрудников научного коллектива

Для каждой группы сотрудников вычисляется ее доля в общей численности научного коллектива. На рисунке видно, что для этого коллектива наибольшую долю имеет возрастная группа 45 – 49 лет (24%). Далее по убыванию следуют следующие пять возрастных групп: 40 – 44 года (20,7%), 25 – 29 лет (10,3%), 50 – 54 года, 55 – 59 лет и 75 – 79 лет (по 6,9%). В настоящее время имеется возможность построения графиков возрастного распределения с шагом от одного года и более³⁾.

Для определения публикационной активности возрастных групп, сейчас разрабатываются средства вычисления значений этого нового индикатора. Кроме этого, разрабатываются средства вычисления индикаторов патентной активности и участия в научных мероприятиях также с определением доли каждой возрастной группы.

В презентации доклада будут описаны основные стадии разработки индикатора публикационной активности возрастных групп, будет продемонстрирована эволюция семантики этого индикатора и тех факторов, которые повлияли на изменение его смыслового содержания.

5 Заключение

Процесс концептуализации индикаторов охватывает следующие сущности: концепты, информация (информационные ресурсы), данные, три категории кодов и цифровые данные. В предлагаемой системе терминов эти сущности в соответствии с их природой, зафиксированной в дефинициях терминов, отнесены к трем разным сферам (средам) и двум разным источникам их происхождения – человек и цифровая техническая система. По смыслу перечисленные термины не пересекаются (имеют изолированные семантические поля в рамках рас-

³⁾ Программы вычисления значений этого индикатора разработаны В.В. Косариком.

смотренной системы терминов), что позволяет более четко описывать процессы личностной и коллективной концептуализации через взаимодействие соответствующих сущностей. Изолированность семантических полей исключает родовидовые и синонимические отношения между этими терминами, что необходимо для более четкого описания проблемы концептуализации.

Предлагаемый подход к решению этой проблемы основан на идеи Г.П. Щедровицкого Н.Г. Алексеева о системе операций, порождающих содержание понятий. В качестве основного средства решения проблемы предложено использовать полидоменные модели видов обеспечения ЭБ, фиксирующие стадии и результаты процесса концептуализации индикаторов [25].

Основное назначение полидоменной модели как интегрального описания процесса концептуализации индикаторов заключается в том, что она делает «прозрачными» для экспертов и других пользователей ЭБ структурные, процедурные и содержательные элементы разных видов обеспечения ЭБ, которые вовлечены в процесс вычисления значений любого индикатора.

Предложенный подход к решению проблемы концептуализации индикаторов опробуется в настоящее время в процессе создания Информационной системы обеспечения реализации и мониторинга Программы фундаментальных научных исследований государственных академий наук на 2008 - 2012 годы.

Программа утверждена распоряжением Правительства РФ от 27 февраля 2008 г. № 233-р. В ней определены восемь индикаторов эффективности ее реализации. Для каждого индикатора заданы пять целевых значений для каждого из 2008 - 2012 годов. При этом предполагается, что число индикаторов, методика их определения и их целевые значения могут изменяться (корректироваться и уточняться). Кроме индикаторов эффективности, в утвержденной Программе заданы параметры финансирования на 2008 - 2012 годы для каждого направления научных исследований. Например, для РАН параметры финансирования заданы для 88 направлений на период 2008 - 2012 годы.

Разработка полидоменной модели для концептуализации индикаторов и других показателей эффективности реализации Программы на 2008 - 2012 годы начата в ИПИ РАН с нормативного (совместно с ЦЭМИ РАН), информационного и лексико-семантического видов обеспечения ЭБ. Последний компонент включает семантический словарь и ряд классификаторов.

В заключение отметим, что полидоменные модели могут использоваться не только для описания процессов концептуализации индикаторов. Они могут быть применены для описания всей функциональности ЭБ, используемых в наукометрических исследованиях, и системы мониторинга в целом.

Литература

- [1] Candela L., Castelli D., Pagano P. et al. Setting the Foundations of Digital Libraries: The DELOS Manifesto // D-Lib Magazine, March/April 2007, Vol. 13, No 3/4.
- [2] Lavoie B., Henry G., Dempsey L. A Service Framework for Libraries // D-Lib Magazine, July/August 2006, Vol. 12, No 7/8.
- [3] Shvaiko P., Euzenat J. In Brief: Ontology Matching // D-Lib Magazine, December 2005, Vol. 11, No 12.
- [4] Mankita I., Meltzer E. A Handful of Things: Calisphere's Themed Collections from the California Digital Library // D-Lib Magazine, May 2006, Vol. 12, No 5.
- [5] Dillon A. Spatial-Semantics: How Users Derive Shape from Information Space // Journal of the American Society for Information Science, 2000, Vol. 51, No 6, pp. 521-528 (D-Lib Magazine опубликовал аннотацию этой статьи по адресу: www.dlib.org/dlib/march00/03JASIS_TOC.html).
- [6] Зацман И.М. Терминологический анализ нормативно-правового обеспечения создания систем мониторинга в сфере науки // Экономическая наука современной России. – 2005, №4. – С. 114–129.
- [7] National Science Board, Science and Engineering Indicators – 2006. Two volumes. Arlington, VA: National Science Foundation, 2006.
- [8] National Science Board, Science and Engineering Indicators – 2008. Two volumes. Arlington, VA: National Science Foundation, 2008.
- [9] Зацман И.М., Курчавова О.А., Галина И.В. Информационные ресурсы и индикаторы для оценки инновационного потенциала направлений научных исследований // Системы и средства информатики. Вып. 18. – М.: Наука, 2008.
- [10] Зацман И.М., Курчавова О.А. Термины для описания процессов представления научно-технических знаний в цифровой среде // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». Вып. 7(14). – М.: РГГУ, 2008. – С. 164–170.
- [11] Buddenberg R. Toward an Interoperability Reference Model - http://web1.nps.navy.mil/~budden/lecture.notes/interop_RM.html (состояние страницы на 12.02.2008).
- [12] Зацман И.М., Веревкин Г.Ф. Информационный мониторинг сферы науки в задачах программно-целевого управления // Системы и средства информатики. Вып. 16. – М.: Наука, 2006. – С. 164-189.
- [13] Planning for Performance and Evaluating Results of Public R&D Programs: Meeting the OMB PART Challenge (Workshop Report). – Washington: The Washington Research Evaluation Network, 2004.
- [14] Соломоник А. Синтаксис в знаковых системах. – Минск: МЕТ, 2007.

- [15] Гак В.Г. Лексическое значение слова // Языкознание : Большой энцикл. слов. – М. : Большая рос. энцикл., 1998. – С. 261–263.
- [16] Vossen P. (ed.) EuroWordNet General Document (Version 3) (<http://www.illc.uva.nl/EuroWordNet/docs/GeneralDoc>).
- [17] McArthur D. Information, its forms and functions: The elements of semiology. – Lewinton: The Edwin Mellen Press, Ltd., 1997.
- [18] Уфимцева А.А. Знак языковой // Языкознание : Большой энцикл. словарь. – М. : Большая рос. Энцикл., 1998. – С. 167.
- [19] Зацман И.М., Кожунова О.С. Семантический словарь системы информационного мониторинга в сфере науки: задачи и функции // Системы и средства информатики. Вып. 17. – М. : Наука, 2007. – С. 124–141.
- [20] Стратегия развития науки и инноваций в РФ на период до 2015 года. Утверждена Межведомственной комиссией по научно-инновационной политике, протокол от 15 февраля 2006 г. № 1 (<http://mon.gov.ru/science-politic/conception/strateg.zip>).
- [21] Лукашевич Н.В. Описание понятий-ролей в лингвистических и онтологических ресурсах // Тезисы постерных докладов 9-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007. – Переславль-Залесский : Изд-во «Ун-т города Переславля», 2007. – С. 81–89.
- [22] Щедровицкий Г.П., Алексеев Н.Г. О возможных путях исследования мышления как деятельности // Докл. АПН РСФСР. 1957. № 3; 1958. № 1, 4; 1959. № 1, 2, 4; 1960. № 2, 4, 5, 6; 1961. № 4, 5; 1962. № 2-6.
- [23] Пископелъ А.А. К творческой биографии Г.П. Щедровицкого (1929–1994) // Щедровицкий Г.П. Избр. труды. – М. : Шк. культ. полит., 1995. – С. XIII–XXXVII.
- [24] Зацман И.М. Полидоменные модели в системах оценки инновационного потенциала и результативности научных исследований // Труды международной конференции Диалог-2006 «Компьютерная лингвистика и интеллектуальные технологии». – М. : Изд-во РГГУ, 2006. – С. 178–183.
- [25] Зацман И.М. Полидоменные модели электронных библиотек систем мониторинга сферы науки // Труды 8-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006. – Ярославль : Ярославский государственный ун-т, 2006. – С. 75–81.

Conceptualization of Scientometric Data in Scientific Digital Libraries

I.M. Zatsman

The problem of science indicator conceptualization is considered. Indicator conceptualization processes are analyzed in terms of teamwork of digital library users carrying out scientometric research. A number of new terms, the semiotic foundations for a *science indicator conceptualization problem*, and semiotic approach to its solution are discussed. The science indicator conceptualization problem is formulated as a special case of the cognitive interoperability of digital library users.

Definitions are formulated for new terms “personal indicator conceptualization” and “coordinated indicator conceptualization”. As it follows from these definitions, digital libraries play a key role in indicator conceptualization processes. The key role implies that fixing of indicator significance and an evolution of indicator semantics for the time is reflected in a digital library lingware, classification systems, semantic dictionaries and thesauruses.

Проблемы технологий создания систем смысловой обработки данных*

© В.Б. Барахнин

А.М. Федотов

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет

bar@ict.nsc.ru

fedotov@sbras.ru

Аннотация

В работе обсуждаются перспективы развития процесса смысловой обработки данных как технологии, при этом в качестве источника данных рассматриваются интернет-документы достаточно произвольной структуры, что принципиально отличает описываемый подход от концепции Semantic Web.

1 Введение

Проблема доступа к информации является одной из основных проблем, возникающих в деятельности научного исследователя. Любой научный процесс порождает огромные объемы данных, и работать с ними становится все сложнее по мере того, как гигабайты данных превращаются в терабайты. Так, еще в начале 1960-х годов американский историк и социолог науки Дерек де Солла Прайс на основании исследований развития науки в течение последних 200 лет выявил следующую эмпирическую закономерность [1]: *любой достаточно большой сегмент науки в нормальных условиях растет экспоненциально, то есть любые параметры науки за определенный промежуток времени удваиваются*. Эта закономерность получила название *закона экспоненциального роста науки*.

Происшедшее за последние 10-15 лет бурное развитие высоких технологий в области передачи и обработки информации, в частности, создание современных телекоммуникационных систем (прежде всего сети Интернет), привело к появлению принципиально новых возможностей организации практический всех этапов научно-информационного процесса, что, в свою очередь, обусловило качественный рост информационных потребностей научного сообщества. Отсюда следует необходимость разработки и создания новых инструментальных средств и алгоритмов для сбора и анализа данных,

содержащихся в разнообразных интернет-документах научной тематики.

Однако при решении поставленной задачи возникают трудности, зачастую носящие принципиальный характер.

Во-первых, распределенное хранение информации требует *интероперабельности* (т.е. обеспечения взаимодействия) разнородных информационных источников. Различают два уровня интероперабельности: семантический и технический [2], причем в последнем иногда выделяют синтаксический уровень [3]. Семантическая интероперабельность, заключающаяся в использовании согласованных стандартов метаданных (обзор которых применительно к научным информационным системам приведен в [3]), как правило, соблюдается. Проблемы возникают на уровне технической интероперабельности, точнее, согласования моделей данных и форматов их представления (что относится к синтаксической интероперабельности).

При этом очень важно иметь в виду следующее обстоятельство. Основным источником электронных документов в настоящее время является сеть Интернет. Однако ее развитие сети изначально носит децентрализованный характер, поэтому выработка сколько-нибудь сложных стандартов представления информации – не более чем благое пожелание. Разумеется, определенным группам разработчиков сайтов, относящихся к той или иной предметной области, удастся договориться об использовании единых стандартов и технологий, позволяющих, в частности, интегрировать ресурсы соответствующих сайтов, однако можно с уверенность утверждать, что имеются сайты, содержащие важную информацию из данной предметной области, но не соответствующие этим стандартам. Поэтому, говоря об интероперабельности, затрагивающей широкий спектр ресурсов Интернет, необходимо априорно рассчитывать лишь на минимальную их стандартизацию.

Во-вторых, в настоящее время наука об обработке информации, особенно в ее прикладном аспекте, несколько отстает от соответствующих аппаратно-программных средств, хотя еще А.Н. Колмогоров показал, что данные представляют информационную ценность лишь тогда, когда они являются составной частью некоторой модели реального мира и

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

связаны с другими данными. Хотя данные представляют информационную ценность лишь тогда, когда они являются составной частью некоторой модели реального мира и связаны с другими данными [4, 5]. Как отмечал А.А. Ляпунов [6], «информация всегда относительна, она зависит от того, какой информационной системой она воспринимается». Аналогичное отставание наблюдается и в прикладных исследованиях, посвященных извлечению из информации *знаний*, понимаемых как структурированная (связанная причинно-следственными и иными отношениями) информация [7]. Разработчикам программных средств обработки данных зачастую недостает понимания того обстоятельства, что конечная цель работы, связанной с применением информационных технологий – *понимание* того или иного явления, а не получение каких-либо чисел, гистограмм, отдельных фактов и т.п.

Одна из основных причин сложившейся ситуации заключается в том, что с появлением в середине 1970-х годов персональных компьютеров появились мощные средства визуализации информации, вследствие чего были практически остановлены научные изыскания в области интеллектуального поиска, которые возобновились лишь в середине 1990-х с появлением сети Интернет, приведшем к распределенному хранению информации. В настоящее время в этой области получены важные результаты (см., например, монографии [8, 9] и др.), однако эти разработки, лежащие, как правило, в рамках концепции Semantic Web [10], обычно опираются на неявное предположение о возможности широкого распространения более или менее подробной стандартизации представления информации. Проблема заключается в том, что наработки консорциума W3 носят лишь *рекомендательный* характер, а объявить их *стандартами* могут только организации, имеющие соответствующий статус, например ISO или ГОСТ, поэтому реальное развитие большинства ресурсов Интернет, в том числе научной направленности, идет без учета соответствующих рекомендаций.

Особо подчеркнем, что при разработке *собственных* ресурсов разработчикам информационных систем следует придерживаться таких стандартов метаданных, которые обеспечивают ориентированность на семантические технологии описания и использования информации, в т.ч. Semantic Web, однако при обработке *чужих* информационных ресурсов надо быть готовым к худшему.

Ресурсы, разработанные без учета рекомендаций консорциума W3, зачастую не могут быть обработаны с использованием онтологий сложной структуры, включающих правила вывода (аксиомы), поскольку «в настоящее (и ближайшее) время ни одна из существующих систем автоматической обработки текстов, извлечения знаний из текстов не может обеспечить такой уровень точности и полноты получения информации из текстов, на которых надежно можно было обосновывать работу таких правил вывода» [11]. Заметим, что это утверждение при-

надлежит коллективу создателей крупнейшего тезауруса русского языка РуТез, предназначенного для автоматической обработки больших текстовых коллекций, который представляет собой иерархическую сеть 50 тысяч понятий с более чем 125 тысячами слов и выражений (статистика относится к 2007 году [12]).

В итоге разработчики прикладных информационных систем, интегрирующих разнородные интернет-ресурсы, стоят перед дилеммой: или опираться на мощный (хотя и созданный еще в досетевую эпоху) аппарат для манипулирования с *данными*, но не с *информацией*, либо использовать результаты последних теоретических исследований, которые, однако, не всегда позволяют обрабатывать огромные пласты информации, доступной через сеть Интернет (внутренние проблемы концепции Semantic Web, связанные, например, со сложностью создания онтологий верхнего уровня, в данной работе мы не рассматриваем).

Таким образом, возникает третья проблема: разработка методологии изучения интернет-ресурсов достаточно произвольной структуры с целью вовлечения данных, содержащихся в этих ресурсах, в сферу научных исследований.

Резюмируя сказанное выше, можно сделать вывод о том, что комплексное решение указанных проблем возможно лишь при осмыслении процесса обработки компьютерной информации как *технологии*. Заметим, что аналогичное осмысление другой области кибернетики – вычислительного моделирования – было осуществлено в начале 1980-х годов в работах Н.Н. Яненко [13] и А.А. Самарского [14] и стало важной вехой в развитии прикладной математики.

2 Проблема стадий переработки информации

В соответствии с [15] будем понимать под технологией определенную последовательность методов обработки, изготовления, изменения состояний и свойств сырья или материалов в процессе производства продукции. Иными словами, любая технология по своей сути – инструмент, применяемый для превращения потребляемых факторов в продукцию, или, вообще говоря, для достижения планируемых результатов (см. [16]).

Сошлемся еще на одно, пожалуй, наиболее краткое из определений технологии: «технология – способ преобразования данного в необходимое» (см., например, [17]), которое подтверждает, что применительно к поставленной задаче по-настоящему технологичным можно назвать лишь тот подход, который способен «перерабатывать» максимально широкие пласты интернет-ресурсов научной тематики, о чем шла речь выше.

Что же выступает исходным материалом для технологии переработки информации? Ответ, на первый взгляд, очевиден: сама информация. Однако и на вопрос о конечном продукте напрашивается

тот же ответ! Разумеется, человек, владеющий теоретическими основами информатики, по некотором размышлении ответит, что исходным материалом служат данные, а конечным продуктом – знания (или, по крайней мере, семантическая информация). Тем не менее, приведенный пример показывает, что проблемы возникают уже на терминологическом уровне.

Поскольку существует множество подходов к понятию «информация» с философских, социологических, биологических, физико-математических или кибернетических позиций ([7], с. 393), включая так называемую «техническую» теорию информации, которая является, по сути, теорией передачи и хранения данных, постольку можно обнаружить десятки порой противоречащих друг другу определений того, что является информацией или знанием. Вряд ли существует некая «абсолютная» точка зрения, с которой возможно было бы судить о том, какое из многочисленных определений является «более правильным». Поэтому цель проводимого ниже экскурса в историю терминологии заключается, прежде всего, в том, чтобы уточнить соответствующие определения применительно к той области информатики, которая изучает процессы взаимных преобразований данных, информации и знаний, установив при этом основания выбора определений, принятых именно в этой области.

Реальное осознание сложностей в преодолении разрыва между шенноновским понятием информации и концепцией семантической информации как средства социальной коммуникации возникло в середине 1960-х гг. (см., например, работы У. Шрамма [18] и Ю.А. Шрейдера [19], причем Ю.А. Шрейдер показал, что о количестве семантической информации в данном сообщении есть смысл говорить лишь применительно к конкретному приемнику сообщения).

Попытка «телеологического» описания особенностей восприятия сообщения субъектом была предпринята Р. Акоффом и Ф. Эмери [20]. Они предложили классифицировать сообщения по видам изменений в получателе, которые делятся на несколько типов (при этом сообщение может принадлежать сразу к нескольким типам):

- 1) *информация* (изменения вероятности выбора);
- 2) *инструкция* (изменения в эффективности выбора);
- 3) *мотивация* (изменения в удельных ценностях).

Тем самым Р. Акофф и Ф. Эмери, по-видимому, одними из первых специалистов в области информатики обратили на многоуровневость восприятия сообщения получателем и сделали попытку описать эти уровни.

Наконец, в начале 1980-х годов немецким исследователем В. Гиттом была предложена пятиуровневая модель информации [21], наиболее полно отражающая различные аспекты термина «информация». Ее уровни (снизу вверх):

- *статистика*;
- *синтаксис*;
- *семантика*;
- *практика*;
- *апобетика* (результат).

Анализируя эту модель, нетрудно видеть, что ее нижний уровень соответствует шенноновскому значению термина «информация», три последующих – семиотической триаде (синтагматика – семантика – прагматика), а верхний (пятый) уровень носит метафизический характер. При этом наличие в некотором сообщении информации высокого уровня влечет за собой наличие информации всех низших высоких уровней, но, разумеется, не наоборот (еще раз напомним: объем информации зависит, в том числе, от характеристик адресата, причем это касается всех уровней информации).

Следует отметить, что модель В. Гитта не получила широкого распространения (во многом потому, что он пытался с ее помощью, делая акцент на пятый уровень, доказать невозможность самопроизвольного возникновения такой сложной информации как генетический код, что явно противоречит общепринятым в современной науке представлениям). Тем не менее, с начала 1980-х годов семиотическая триада заняла прочное место в кибернетике, о чем свидетельствуют соответствующие статьи в «Словаре по кибернетике» [24], хотя в первое время семиотическая терминология применялась, скорее, при описании языка (понимаемого как частный случай знаковой системы) в целом, нежели при анализе отдельных сообщений. Однако к настоящему моменту описание непосредственно информации с помощью семиотической терминологии получило широкое распространение в отечественной литературе.

Важно подчеркнуть, что семиотический подход фактически использован при определении базисных понятий в фундаментальной монографии [7], изданной ВИНТИ. Данные понимаются в ней (в соответствии с традиционным подходом) как факты и идеи, представленные в символической форме, позволяющей проводить их передачу, обработку и интерпретацию, *информация* – как смысл, приписываемый данным на основании известных правил представления фактов и идей. Структурированная информация, образующая систему, составляет *знания*.

Исходя из этого понимания терминов «данные», «информация», «знания», которого мы будем придерживаться в дальнейшем, можно сказать, что *данные соответствуют синтаксическому уровню сообщения, информация (в узком смысле!) – семантическому, а знания – прагматическому*.

Таким образом, проделанный нами анализ показывает, что создание технологий компьютерной обработки информации невозможно без анализа стадий процесса переработки информации, иными словами, без должного интеллектуального обеспечения технологии, которое, разумеется, основано на широких возможностях современных аппаратных и программных средств.

3 Системный подход – основа технологии обработки информации

Какие же качественно новые возможности предоставляют мощности современных компьютеров и языков манипулирования данными? В классических *информационно-поисковых системах* (ИПС) основным элементом (или же логической единицей хранения) являлась *запись* в базе данных, представлявшая собой поисковый образ документа [22]. При этом важно отметить, что записи не имели непосредственной связи друг с другом, что резко суживало возможности ИПС. В частности, автоматизированные системы, способные строить даже простые категорические силлогизмы (для чего требуется наличия в системе связей между терминами силлогизма), отнесены ([23, с.149-150]) к особому классу *информационно-логических систем*. Одной из наиболее очевидных практических проблем, порожаемых отсутствием связей между записями, является невозможность установить наличие (или отсутствие) связи между собственным именем и предполагаемым его конкретным носителем, даже если информация о предполагаемом носителе и присутствует в ИПС [23, с. 137]. Тем самым ИПС полностью оправдывали свое название – они выдавали в качестве продукта переработки данных именно информацию, но не знания.

Развитие алгоритмических, программных и аппаратных средств информатике привело в 1980-е годы к возможности создания *интеллектуальных информационных систем*, в которых компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека. Интеллектуальные системы (ИнтС) функционируют по следующей схеме [7]:

ИнтС = РИС + ИПС + ИнИн+ АП,

где РИС – рассуждающая информационная система (формализующая правила логического вывода), ИнИн- интеллектуальный интерфейс (диалог, графика и т.д.).

Более развитые ИнтС должны обладать и механизмом пополнения базы данных, функционируя по схеме

ИнтС = РИС + ИПС + ИнИн,

где АП – автоматическое извлечение данных из текстов и соответствующее пополнение базы данных посредством этих фактов.

Таким образом, интеллектуальная система обладает по сравнению с обычной ИПС новыми возможностями, позволяя удовлетворить квалифицированного пользователя в соответствии со схемой «документ – факт – рассуждение» [7, с. 343], то есть *интеллектуальные информационные системы позволяют не только извлекать из данных информацию, но и получать новые знания.*

На основании сказанного можно сделать вывод, что функционирование интеллектуальной информационной системы основано на двух противоположных процессах: *при пополнении ИнтС новыми сведениями происходит преобразование семантиче-*

ской информации в данные, однако непосредственно потребности пользователя удовлетворяет обратный процесс – извлечение из данных нужной пользователю информации и знаний.

Таким образом, для наиболее эффективного функционирования ИнтС целесообразно рассматривать в качестве логической единицы хранения *документ*, понимаемый как информационный ресурс, имеющий (по определению [24]) уникальный идентификатор и обладающий некоторой структурой и содержанием.

Разумеется, документ – информационный ресурс представляет собой поисковый образ исходного документа, причем в некоторых случаях содержание последнего может входить в поисковый образ в качестве одного из элементов (это противоречит ограничению из классической монографии [25], но из контекста следует, что подобное ограничение было вызвано необходимостью уменьшения объема поисковых образов с целью уменьшения трудоемкости процесса их обработки). С другой стороны, поисковый образ документа тоже является документом (описывающим исходный документ), поэтому далее, где это не вызовет недоразумения, мы будем использовать термин «документ» в значении «поисковый образ исходного документа». С другой стороны, в фундаментальных работах по информатике и кибернетике [22, 25], вышедших, в том числе в конце 1980-х годов, поисковый образ документа не рассматривается даже в качестве вторичного документа.

Важнейшей особенностью данного подхода является использование для описания документов метаданных иерархической структуры. Наиболее общий характер имеют метаданные, задающие структуру документа, т.е. описывающие метаданные более низкого уровня (атрибуты документа), которые определяют содержание документа (см. рис. 1). Наконец, значения этих атрибутов является фактически метаданными по отношению к исходному документу. Отсюда следует *важнейшая отличительная черта предлагаемого подхода к построению информационных систем: работа не с данными, а исключительно с метаданными.*

Важно подчеркнуть, что документ может входить в качестве значения некоторого элемента метаданных другого документа. Подробнее, любой документ d_i массива данных представляется как $d_i = \langle m_i^{j,k} \rangle$, где $m_i^{j,k}$ – значения элементов мета-

данных M^j из набора единого набора метаданных $M = \bigcup M^j$, k – количество значений (с учетом повторов) соответствующего элемента метаданных в описании документа. Если же документ $d_{i'}$ входит в качестве значения элемента M^j метаданных документа d_i , то можно говорить о связи между этими $m_{i,i'}^{l,k}$ – атрибуты этой связи, являющиеся значениями соответствующих элементов метаданных.

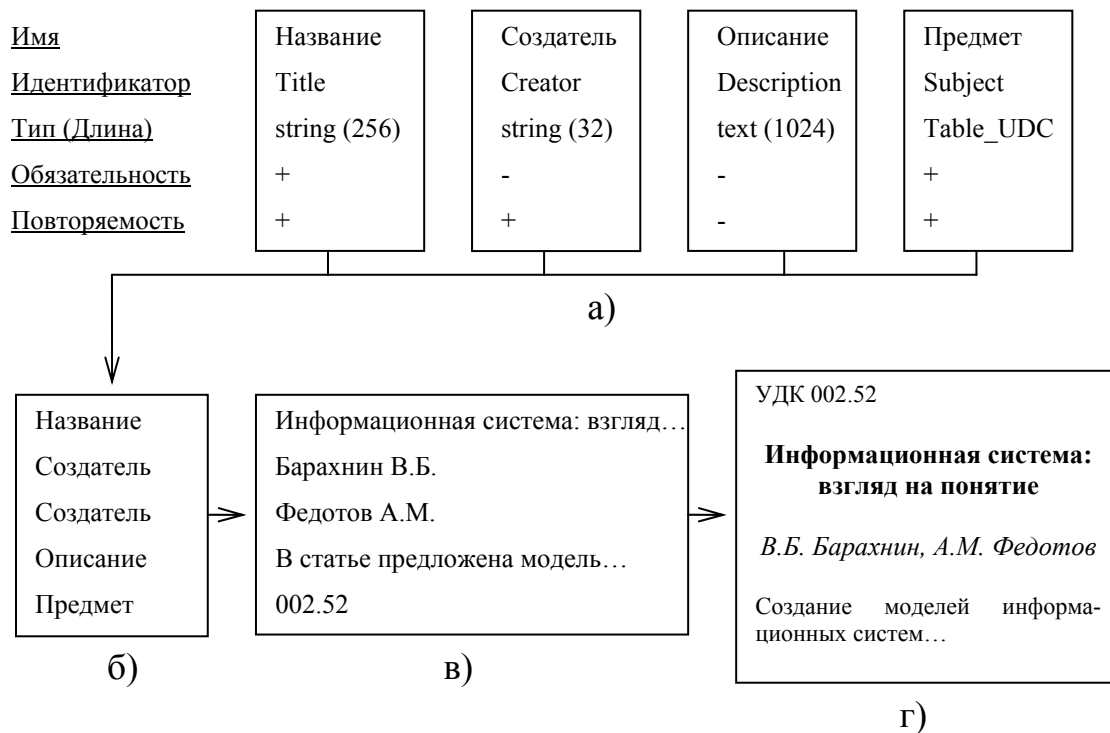


Рис. 1. Иерархия метаданных документа: а) структура, б) атрибуты, в) содержание, г) документ

Таким образом, наличие внутренних связей между элементами массива данных позволяет рассматривать его как некоторую *систему* и анализировать его с использованием методов общей теории систем (заметим, что классическое определение системы «множество объектов вместе с отношениями между объектами и между их атрибутами» [26] основано на тех же понятиях, что и, например, реляционная модель данных).

Соответствующий анализ был проведен нами в [27, 28]. Перечислим основные выводы этих статей, имеющие отношение к технологическим аспектам обработки данных.

Прежде всего, отметим, что с использованием системного подхода в [27] удалось дать обоснованную формулировку информационных потребностей научного сообщества и предложить реально выполнимую схему их удовлетворения, учитывающую необходимость компромисса между качеством решения поставленной задачи и разумными сроками ее выполнения. Последний принцип давно является основополагающим в другой отрасли кибернетики – прикладной математике (см., например, [29]), при этом улучшение результата возможно с течением времени и достигается, применительно к информационной системе, посредством расширения массива данных (как посредством добавления новых документов, так и расширением структуры уже существующих).

Модель данных в системе смысловой обработки данных строится посредством задания классов документов K_i (т.е. множеств документов, описываемых при помощи одних и тех же элементов метаданных M_i), и типов возможных связей между классами $M^j < K_i, K_i >$ с указанием элементов метадан-

ных $M_{i,j}^j$, описывающих атрибуты соответствующих связей, т.е. для построения модели данных используется комбинация иерархической и реляционной моделей, что сближает ее с инфологическими моделями [30]. Анализ иерархии метаданных, приведенной на рис. 1, позволяет сделать важный вывод: *описание массива данных посредством метаданных* наделяет их, в том числе, семантикой, воспринимаемой в среде социальных коммуникаций, т.е. *делает данные информацией* (в узком значении этого слова).

Как же добиться возможности реализации следующего технологического шага – получения *новых* (т.е. явно не содержащихся в исходном массиве данных) *знаний*? Очевидно, необходима, как минимум, хорошая структуризация данных, предусматривающая, в частности, достаточно большое количество поисковых признаков, образующих поисковый образ документа, причем соответствующие документы-описания должны быть объединены в *каталог*.

Кроме того, в информационно-поисковом языке, используемом при создании каталога, должны присутствовать средства выражения имманентных отношений между предметами, т.е. язык должен обладать парадигматическими отношениями (примером языка, не обладающего этими отношениями, может служить система унитаров – набора одиночных ключевых слов (в редких случаях словосочетаний)). Средством же выражения парадигматических отношений является *онтология* предметной области или ее *тезаурус*, причем граница применения этих терминов весьма размыта (как отмечено в [31], «...еще недавно сегодняшняя Онтология именовалась Тезау-

русом», что иллюстрируют, например, тезаурусы по науковедению и лексикографии [32], которые ввиду своей структурной сложности с сегодняшней точки зрения явно представляются онтологиями). Таким образом, наличие онтологии (тезауруса) в качестве составной части информационно-поискового языка, используемого при создании каталога, является необходимым и достаточным условием (см. [28]) возможности получения из данных, уже преобразованных в информацию, новых знаний.

Попутно заметим, что именно каталог является наиболее естественной формой унификации представления данных, тем самым служа достаточно простым средством решения отмеченной во введении проблемы синтаксической интероперабельности.

Наконец, рассмотрение массива данных как системы позволяет уделить особое внимание ее динамическим характеристикам, поскольку «...отдельные уровни системы обуславливают определенные аспекты ее поведения, а целостное функционирование оказывается результатом взаимодействия всех ее сторон и уровней» [33].

4 Технология автоматизации обработки интернет-документов

Важнейшим вопросом является пополнение информационной системы новыми интернет-документами. Опыт создания информационных систем научной направленности показывает, что подобные системы могут успешно развиваться лишь в случае актуализации содержащейся в них информации самими пользователями этих систем. Более того, поскольку в интеллектуальных информационных системах компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека, при этом происходит автоматизированное пополнение базы данных. В силу указанных обстоятельств при работе с интеллектуальными информационными системами многих пользователей, возможности систем резко возрастают.

Так как пользователи, принимающие участие в актуализации информации, могут находиться в разных регионах России и даже мира, то практическое взаимодействие таких программных систем с внешним миром в плане занесения в них новых данных целесообразно организовывать преимущественно (или даже почти исключительно) через веб-интерфейс.

Отметим, что обработка интернет-документов имеет ряд специфических особенностей, отличающих их каталогизацию от каталогизации полиграфических изданий. В частности, каждую публикацию в составе электронного журнала, сборника и т.п. целесообразно представлять как отдельный документ. Это существенно облегчает процесс поиска пользователем нужной информации, позволяя вести атрибутивный поиск отдельных статей по авторам, названию, классификационным признакам, ключевым словам и т.п.

Ввиду того, что информационная система работает не с данными, а исключительно с метаданными, сбор интернет-документов сводится к сбору их метаданных, к тому же непосредственное копирование документов может вызвать серьезные вопросы относительно соблюдения авторских прав.

Однако, как уже было отмечено ранее, значительная доля интернет-документов не содержит метаданных, явно заданных в соответствии с какими-либо рекомендациями, позволяющими проводить автоматическое извлечение, вследствие чего возникает необходимость разработки технологии частичной автоматизации извлечения метаданных из документов произвольной структуры.

Так как однородные документы, размещенные на одной сайте, имеют однородную структуру, то наиболее целесообразно использовать алгоритмы, использующую информацию о гипертекстовой разметке обрабатываемых документов (см., например [34, 35]), при этом надо иметь в виду, что документ может не обладать xml-разметкой и не содержать метаданные в мета-теге, поэтому следует ориентироваться только на html-разметку.

Один из возможных алгоритмов решения задачи частичной автоматизации процесса извлечения метаданных разработан и изложен нами в [36]. Алгоритм, основанный на типичном для интеллектуальных информационных систем человеко-машинном взаимодействии, сводится к выполнению последовательных операций:

- 1) создание шаблона для обрабатываемого сайта;
- 2) создание списка адресов, где расположены документы;
- 3) обработка документов;
- 4) поддержание актуальности информации.

Следует обратить особое внимание на извлечение таких метаданных, как классификационные признаки (т.е. коды того или иного классификатора) документа и ключевые слова. Без этих элементов метаданных ценность каталожного описания документа минимальна, поскольку в описанной ситуации процесс поиска документа человеком или его обработка рассуждающей информационной системой может опираться только на простую проверку вхождения тех или иных терминов в текст документа.

К сожалению, даже журнальные статьи далеко не всегда содержат ключевые слова и классификационные признаки. И даже в тех случаях, когда эти признаки указаны, классификатор, используемый журналом, может не соответствовать классификатору каталога. Так, в некоторых математических журналах используется классификатор УДК, в то время как в международном математическом сообществе более распространен классификатор MSC2000.

Разумеется, наиболее качественно решить задачу классификации может эксперт-человек, поэтому, прежде всего, следует проверить, внесена ли информация о полиграфической версии статьи в ту или иную электронную библиографическую базу

данных удаленного доступа, в которой документы классифицированы в соответствии с нужным классификатором. Так, в среде математиков очень популярна база данных журнала «Zentralblatt MATH» (<http://www.zentralblatt-math.org/MATH/home>), содержащая более 2 миллионов записей. Статью в этой базе можно однозначно идентифицировать по ISSN журнала, его номеру и страницам, на которых расположена статья. К сожалению, не все электронные версии журналов содержат номера страниц полиграфических версий статей, поэтому при отсутствии сведений о страницах в процессе идентификации следует опираться на фамилии автора (авторов) в латинской транскрипции.

Подчеркнем, что полная репликация метаданных документа из библиографической базы не может служить эффективной заменой процессу непосредственного извлечения метаданных из интернет-документа, поскольку в большинстве случаев библиографические базы не содержат сведений об url-адресе электронной версии документа.

Процесс определения метаданных документа с использованием удаленной библиографической базы также может быть частично автоматизирован [36].

Наконец, если классификационные признаки документа отсутствуют как в нем самом, так и в библиографических базах удаленного доступа, то требуется провести автоматическую классификацию документа, исходя непосредственно из его содержания. Для решения этой задачи был разработан и реализован алгоритм автоматической классификации (кластеризации) документов на основании меры их сходства, задаваемой с использованием атрибутов их библиографического описания. [37]. Его отличительными особенностями являются, во-первых, использование в процессе координатного индексирования документа не отдельных слов, входящих в словарь предметной области, а терминов-словосочетаний, образующих ее тезаурус; во-вторых, подсчет меры сходства на основании не только координатного индекса текста документа, но и ключевых слов (в узко-библиографическом понимании), а также сведений об авторах документа; и, в-третьих, применение продукционных правил, позволяющих изменять весовые коэффициенты, соответствующие тем или иным атрибутам библиографического описания в формуле задания меры сходства на основании апостериорной достоверности значений этих атрибутов.

5 Заключение

В данной работе намечены первые шаги в направлении осмысления процесса смысловой обработки данных, содержащихся в интернет-документах достаточно произвольной структуры, как технологии. Показано, что в основе этой технологии должна лежать представление о массиве данных как о системе, описываемой с использованием метаданных посредством комбинации иерархиче-

ской и реляционной моделей, благодаря чему между элементами системы (поисковые образы документов) устанавливаются внутренние связи. Описание массива данных посредством метаданных делает данные информацией, а наличие онтологии (тезауруса) в качестве составной части информационно-поискового языка, используемого при создании каталога, является необходимым и достаточным условием возможности получения из данных, преобразованных в информацию, новых знаний. Установлено, что применение методов общей теории систем открывает дополнительные возможности исследования технологии смысловой обработки данных. Предложена технология автоматизации извлечения метаданных (в т.ч. классификационных признаков) из интернет-документов.

Описанные технологии были использованы при создании сайта СО РАН www.sbras.ru, занимающего (по данным за июль 2008 года) согласно рейтингу Webometrics [38] наивысшее среди российских сайтов 54-е место в мире и 18-е – в Европе (всего в рейтинг-лист входят 500 ведущих сайтов университетов и научно-исследовательских центров всего мира), а также ряда связанных с этим сайтом информационных систем.

Литература

- [1] Прайс Д. Малая наука, Большая наука // Наука о науке. – М. : Прогресс, 1966. – С. 281–385.
- [2] Фейгин Д. Концепция SOA // Открытые системы. – 2004. – № 6. – http://www.osp.ru/os/2004/06/184447/_p1.html.
- [3] Бездушный А.Н., Кулагин М.В., Серебряков В.А., Бездушный А.А., Нестеренко А.К., Сысоев Т.М. Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии. – 2005. – Т. 10. – Специальный выпуск. – С. 29–48.
- [4] Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. – Т. I. – Вып. 1. – С. 3–11.
- [5] Колмогоров А.Н. Теория информации и теория алгоритмов. – М. : Наука, 1987.
- [6] Ляпунов А.А. О соотношении понятий материя, энергия и информация // Ляпунов А.А. Проблемы теоретической и прикладной кибернетики. – Новосибирск : Наука, 1980. – С. 320–323.
- [7] Арский Ю.М., Гиляревский Р.С., Туров И.С., Черный А.И. Инфосфера: Информационные структуры, системы и процессы в науке и обществе. – М. : ВИНТИ, 1996.
- [8] Krogstie J., Halpin T., Siau K. Information Modeling Methods and Methodologies. – Idea group publishing, 2005.
- [9] Staab S., Stuckenschmidt H. (Eds.) Semantic Web and Peer-to-Peer, Decentralized Management and Exchange of Knowledge and Information. – Springer, 2006.
- [10] Semantic Web. <http://www.w3.org/2001/sw/>.

- [11] Добров Б.В., Лукашевич Н.В., Сеницын М.Н., Шапкин В.Н. Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска // Труды Седьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2005). – Ярославль, 2005. – С. 70–79.
- [12] Лукашевич Н.В. Описание понятий-ролей в лингвистических и онтологических ресурсах // Сборник тезисов постерных докладов Девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007). – Переславль-Залесский, 2007. – С. 81–89.
- [13] Яненко Н.Н. Методологические вопросы современной математики // Вопросы философии. – 1981. – № 8. – С. 60–68.
- [14] Самарский А.А. Задачи прикладной математики на современном этапе развития // Коммунист. – 1983. – № 18. – С. 31–42.
- [15] Жукова Е.А., Мелик-Гайказян И.В. Философские проблемы технологий и феномен Hi-Tech // Философия математики и технических наук. – М.: Академический Проект, 2006. – С. 557–586.
- [16] Желены М. Управление высокими технологиями // Информационные технологии в бизнесе. Энциклопедия. – СПб.: Питер, 2002. – С. 81–89.
- [17] Технология // Тезаурус по образованию и педагогике. – Институт информатизации образования в составе Московского государственного гуманитарного университета им. М.А. Шолохова. http://www.mgopu.ru/ininfo/r1_thesaurus.htm#technology.
- [18] Schramm W. Information Theory and Mass Communication // In Communication and Culture. N.Y.: Holt, Rinehart & Winston, 1966. P. 521–534.
- [19] Шрейдер Ю.А. О семантических аспектах теории информации // Информация и кибернетика. – М.: Советское радио, 1967. – С. 15–47.
- [20] Ackoff R., Emery F. On Purposeful Systems. – Ch. – N.Y.: Aldine – Atherton, 1972.
- [21] Gitt W. Ordnung und Information in Technik und Natur // In: Gitt W. (Hrsg.): Am Anfang war die Information. Grdfeling: Resch KG, 1982. – S. 171–211.
- [22] Словарь по кибернетике. – Киев: Гл. ред. Украинской Советской Энциклопедии им. М.П. Бажана, 1989.
- [23] Михайлов А.И., Черный А.И., Гиляревский Р.С. Научные коммуникации и информатика. – М.: Наука, 1976.
- [24] Berners-Lee T., Fielding R., Masinter L. Uniform Resource Identifiers (URI). Generic Syntax. RFC 2396. – <http://www.ietf.org/rfc/rfc2396.txt/>.
- [25] Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики. – М.: Наука, 1968.
- [26] Холл А.Д., Фейджин Р.Е. Определение понятия системы // Исследования по общей теории систем. – М.: Прогресс, 1969. – С. 252–282.
- [27] Барахнин В.Б., Леонова Ю.В., Федотов А.М. К вопросу о формулировке требований для построения информационных систем научно-организационной направленности // Вычислительные технологии. – 2006. – Т. 11. – Специальный выпуск. – С. 52–58.
- [28] Барахнин В.Б., Федотов А.М. Информационная система: взгляд на понятие // Вестник НГУ. Сер.: Информационные технологии. – 2007. – Том 5. – Вып. 2. – С. 12–19.
- [29] Бахвалов Н.С. Численные методы. – М.: Наука, 1970.
- [30] Langefors B. Infological models and information user views // Information Systems. – 1980. – № 5. – P. 17–32.
- [31] Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям, т. 1. – Аксаково, 2001. – С. 184–188.
- [32] Никитина С.Е. Семантический анализ языка науки. – М.: Наука, 1987.
- [33] Садовский В.Н. Система // Философский энциклопедический словарь. – М.: Советская Энциклопедия, 1983. – С. 610–611.
- [34] Crescenzi V., Mecca G., Merialdo P. Roadrunner: Towards automatic data extraction from large web sites // In: The VLDB Journal. – Rome, 2001. – P. 109–118.
- [35] Sahuguet A., Azavant F. Building intelligent web applications using lightweight wrappers // Data Knowledge Engineering. – 2001. – V. 36 – № 3. – P. 283–316.
- [36] Барахнин В.Б., Ведерников В.В. Алгоритм автоматической каталогизации статей, опубликованных в электронных версиях научных журналов // Труды Всероссийской научной конференции «Научный сервис в сети Интернет: технологии параллельного программирования». – Новороссийск, 2006. – С. 277–279.
- [37] Барахнин В.Б., Нехаева В.А., Федотов А.М. О задании меры сходства для кластеризации текстовых документов // Вестник НГУ. Сер.: Информационные технологии. – 2008. – Т. 6. – Вып. 1. – С. 3–9.
- [38] Top 100 R&D European Institutes. – http://www.webometrics.info/top100_r&d_europe.asp

Technologies for construction of intelligence system for data analysis

V.B. Barakhnin, A.M. Fedotov

The paper describes technologies for construction of intelligence system for data analysis.

* Работа выполнена при частичной поддержке РФФИ: проекты 06-07-89060, 06-07-89038, 07-07-00271; президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.

Решение некоторых задач Text Mining при помощи концептуальных графов

© М.Ю. Богатырев, В.В. Тюхтин

Тульский государственный университет
okkambo@mail.ru

Аннотация

В работе задача извлечения ассоциативных правил из текстовых данных сформулирована на концептуальных графах – семантических моделях текстов. Показано, что способом решения данной и других задач Text Mining является кластеризация концептуальных графов. Данный подход применен в проекте экспериментальной электронной библиотеки научных статей.

1 Введение

Известной тенденцией развития электронных библиотек является их превращение в *системы поддержки знаний*. Поддержка знаний означает решение на данных библиотеки специфических задач обработки текстов с целью построения объектов, трактуемых как *знания* в рамках определенных моделей знаний.

Методы, модели и алгоритмы, применяемые в подобных задачах обработки текстов, образуют направление, известное как Text Mining [11].

Это направление связано с другим известным направлением, Data Mining – извлечением знаний из данных [3]. Справедливо считать, что Text Mining – это Data Mining применительно к текстовым данным. До сих пор нет устоявшегося русского термина – перевода для Text Mining, поэтому мы будем употреблять англоязычный термин.

Среди методов Text Mining можно выделить методы, применяемые непосредственно к текстам как неструктурированным данным, и методы, использующие структурированные модели текстов.

Одной из востребованных здесь структурированных моделей текста является *концептуальный граф* [10]. Вместе с *концептуальными решетками* концептуальные графы относятся к *концептуальным структурам*, которые являются одним из формальных представлений знаний. Концептуальные графы как атомарные объекты моделирования, в

отличие, например, от ключевых слов текста, предоставляют информацию о структуре предложения, его логике и семантике, что важно в задачах Text Mining.

Концептуальный граф – это двудольный направленный граф, состоящий из двух типов узлов: *концептов* и *концептуальных отношений*. Концептуальный граф соответствует одному предложению текста. Концептами концептуального графа являются части речи – существительные, прилагательные, глаголы, наречия. Концептуальные отношения не являются элементами текста непосредственно, но строятся, например, на основе анализа семантических ролей предложения [5].

В работах [13–15] рассмотрена методика применения концептуальных графов в качестве объектов хранения электронных библиотек, рассмотрен специальный, эволюционный подход к задачам кластеризации концептуальных графов. В данной работе концептуальные графы применяются к решению некоторых задач Text Mining.

2 Задачи Text Mining

В задачах Text Mining строятся *кластеры*, *ассоциации*, анализируются *особенности* текстов, *подобие* текстов и т.д. [4], [9]. Традиционные методы обработки текстов используют ключевые слова или векторные модели текста, что требует значительных вычислительных ресурсов при обработке больших текстовых коллекций.

Концептуальный граф как модель сложнее, чем набор ключевых слов, но компактнее, чем вектор, построенный на тексте, например, в методах латентно-семантического анализа [5]. Это позволяет эффективно применять концептуальные графы в задачах Text Mining.

Применение концептуальных графов привносит определенную специфику в задачи Text Mining. Далее рассмотрены две задачи – построение ассоциативных правил и поиск особенностей в тексте, решения которых основаны на решении задачи кластеризации концептуальных графов.

Всякая технология Text Mining включает в себя два этапа:

- подготовку данных;
- решение задач Text Mining на подготовленных данных.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

Рассмотрим реализации указанных этапов с применением концептуальных графов.

2.1 Подготовка данных

Применительно к текстам, подготовка данных означает замену неструктурированных текстов их структурированными моделями. В данном случае текстам должны быть сопоставлены концептуальные графы.

Проблема построения концептуальных графов хорошо известна [2], [6], [9] и не имеет общего решения, ввиду семантической неоднозначности концептуальных графов.

Следующий хрестоматийный пример демонстрирует такую неоднозначность: если спросить, о чем повествует предложение «Студент читает книгу»? то ответами могут быть «о студенте», «о чтении», «о книге». Поэтому возможны три варианта концептуальных графов Направления дуг графа

$$\begin{aligned} [\text{Студент}] \leftarrow (\text{Агенса}) \leftarrow [\text{Читает}] \rightarrow \\ \rightarrow (\text{Пациенса}) \rightarrow [\text{Книга}] \end{aligned} \quad (1)$$

выделяют концепт [Читает] как основной, поэтому данный граф определяет, что рассматриваемое предложение повествует, прежде всего, о чтении.

Здесь использован *вербоцентрический* подход к построению концептуальных графов и линейная форма их представления [10]: в прямоугольные скобки заключаются концепты, в круглые – отношения.

На практике применяют ручное и автоматическое построение концептуальных графов. Ручное построение разрешает возможные семантические неоднозначности и считается наиболее корректным. Однако, для создания полнофункциональных систем на основе концептуальных графов, особенно систем, работающих в реальном времени, принципиальное значение имеет автоматическое построение графов.

Известны подобные англоязычные системы [2], [6]. Нами был разработан и опробован алгоритм построения концептуальных графов для предложений русского языка [15]. Для реализации алгоритма необходима подсистема синтаксического разбора предложений, например, система DWARF [17].

2.2 Построение ассоциативных правил

Построение ассоциативных правил является классической задачей Data Mining и заключается в следующем.

Пусть $R = \{r_1, \dots, r_n\}$ – множество объектов, $T = \{t_1, \dots, t_m\}$ – множество транзакций, такое, что каждая транзакция есть эксперимент по обнаружению объектов из R и представляет собой бинарный вектор длины n , в котором i -я компонента равна 1, если объект r_i присутствует в транзакции. Пусть

$X \subset R, Y \subset R \setminus X$ – подмножества объектов в множестве R .

Ассоциативное правило на R есть клауза:

$$X \Rightarrow Y \quad (2)$$

Правило (2) поддерживается с *доверием* γ , если

$$\frac{\sum_{j=1}^m t_j[X \wedge Y]}{\sum_{j=1}^m t_j[X]} \geq \gamma \quad (3)$$

и имеет *распространенность* σ , если

$$\frac{\sum_{j=1}^m t_j[X \wedge Y]}{m} \geq \sigma \quad (4)$$

Здесь $t_j[X \wedge Y]$ – транзакция, в которой обнаружены объекты X и объекты Y , $t_j[X]$ – транзакция, в которой обнаружены объекты X .

Неравенство (3) обычно трактуется как вероятность того, что появление объектов X влечет появление объектов Y , что вполне вписывается в статистический подход к построению ассоциативных правил [3]. Правила (2) также часто интерпретируют как правила типа «если – то», не учитывая, что $\Rightarrow$ – знак метаимпликации.

Логическая интерпретация правил (2) как клауз применяется реже, но при использовании концептуальных графов эта интерпретация – «от общего к частному» – как раз актуальна.

Ассоциативное правило на концептуальных графах определяется следующим образом. Пусть $G = \{g_1, \dots, g_n\}$ – множество концептуальных графов. *Обобщенным графом* для графа g назовем граф $\hat{g}$, полученный из графа g применением к нему конечного числа операций *разъединения* и *обобщения* [10]. Операции разъединения применяются к дугам графа, а операции обобщения – к его концептам и отношениям.

Граф (1) является обобщенным графом для графа, который соответствует, например, предложению «Студент увлеченно читает интересную книгу».

Обобщенный граф может быть порожден не одним графом, а множеством графов $\{g_i\}$.

Ассоциативным правилом на концептуальных графах назовем клаузу

$$\hat{g} \Rightarrow g, (b, v), \quad (5)$$

где b, v – доверие и распространенность правила, вычисляемые следующим образом:

$$\frac{N[\hat{g}, g]}{N[\hat{g}]} \geq \alpha, \quad (6)$$

$$\frac{N[g]}{n} \geq \beta \quad (7)$$

Здесь $N[\hat{g}, g]$ – функция, возвращающая число графов из G , имеющих графы $\hat{g}$ и g в качестве под-

графов, аналогично интерпретируются $N[\mathcal{G}]$ и $N[g]$.

Множество транзакций как результат экспериментов по обнаружению объектов в данном случае отсутствует. Роль таких экспериментов играет операция обобщения графов.

Семантическая интерпретация правила (5) состоит в упомянутом уже принципе «от общего к частному»: обобщенный граф $\hat{g}$ задает общий смысл, графы g_i (б, в) – детали. Соответственно, один обобщенный граф может порождать несколько правил с различными величинами доверия и пространственности.

Определив ассоциативные правила в виде (5), мы сталкиваемся с проблемой их построения.

Покажем, что решение данной проблемы может быть найдено как решение задачи кластеризации.

В определении обобщенного графа не указан принцип, в соответствии с которым выполняются операции разъединения и обобщения. Эти операции, в действительности, служат моделью выделения на множестве G подграфов, обладающих некоторыми общими свойствами.

Такие свойства можно задать бинарным отношением близости графов в соответствие с некоторой мерой.

Пусть

$$C_1 \subset C_2 \subset \dots \subset C_k \quad (8)$$

– иерархия кластеров на множестве концептуальных графов, построенная с использованием некоторой меры близости графов.

Обозначим $g_i^{(i)} \in C_i$ i -й граф, принадлежащий кластеру C_i , $\mathcal{G}_s^{(i)}$ – его обобщенный граф.

Утверждение 1. Наличие кластеризации (8) достаточно для того, чтобы среди множества ассоциативных правил, существующих на множестве концептуальных графов, существовали правила вида

$$\mathcal{G}_s^{(i)} \Rightarrow g_r^{(j)}(\alpha, \beta) \quad (9)$$

где $i > j$, индексы s, r не обязательно различны. Другими словами:

- кластеризация концептуальных графов позволяет получать обобщения концептуальных графов;
- иерархия кластеров на множестве концептуальных графов задает некоторое подмножество ассоциативных правил.

Решение задачи кластеризации определяется применяемой мерой близости.

В работе [13] рассматриваются меры близости концептуальных графов на основе известных коэффициентов Дайса – концептуальной близости d_c и относительной близости d_r :

$$d_c = \frac{2n(g_c)}{n(g_i) + n(g_j)}, \quad (10)$$

$$d_r = \frac{2m(g_c)}{m_{g_c}(g_i) + m_{g_c}(g_j)}, \quad (11)$$

где $g_c = g_i \cap g_j$, $n(g_i)$ – число концептов графа g_i , $m(g_c)$ – число отношений концептуального графа g_c , $m_{g_c}(g_i)$ – число отношений концептуального графа g_i , для которых хотя бы одна из вершин принадлежит графу g_c . Там же рассмотрены модификации указанных мер, учитывающих размеры сравниваемых графов, и наличие общезначимых концептов.

Применяя меры (6), (7), можно выделить на множестве G кластеры, графы в которых близки в смысле наличия у них одного обобщенного графа. Такой кластеризации можно сопоставить операцию разъединения.

Операция обобщения сложнее. Ее результатом может быть замена концептов и отношений на «более общие». Это требует введения *иерархии типов* концептов $\mathbf{T}_C$ и отношений $\mathbf{T}_R$.

Вершина каждой иерархии представлена *универсальным концептом* и *универсальным отношением* [10] в контексте решаемой задачи. Соответственно, внизу иерархий находятся типы *абсурдного концепта* и *абсурдного отношения*. Место типа в иерархии определяется *порядком* типа. Отношения в иерархии типов отношений характеризуются также *арностью отношения*.

В следующем разделе приведены некоторые примеры иерархий типов.

Наличие иерархий позволяет применить иные меры близости, измеряющие близость двух концептов как сумму расстояний от них до другого концепта с более высоким порядком в иерархии типов. Концепты разных иерархий типов могут быть связаны отношениями в конкретных графах.

Данная идея не нова [8], ее применение позволяет модифицировать меру (10) следующим образом.

Если два графа не имеют общих концептов, но в них есть концепты, принадлежащие одной иерархии типов, то эти концепты «засчитываются» в меру близости:

$$d_c = \frac{2n(g_c) + a(\sum_k f(c_{k,i}, c^0) + \sum_l f(c_{l,j}, c^0))}{n(g_i) + n(g_j)}, \quad (12)$$

где $c_{k,i}$, $c_{l,j}$ – k -й и l -й концепты, соответственно, i -го и j -го графа, $f(c_{k,i}, c^0)$ – функция расстояния между концептами графов и ближайшим общим концептом в иерархии, a – масштабирующий множитель. В простейшем варианте функция $f(c_{k,i}, c^0)$ возвращает натуральное число – разницу в уровнях иерархии между концептами c^0 и $c_{k,i}$.

Поддержка иерархии типов отношений не столь актуальна, ввиду того отношение по своей природе

семантически локально. В самом деле, например, отношения *агенса*, *пациенса*, *атрибута*, *цели* и т.п. действуют в рамках одного предложения и не являются наследниками каких-то более общих отношений.

2.3 Выделение особенностей в тексте

Плоская кластеризация множества концептуальных графов разбивает его на подмножества, слабо связанные с точки зрения применяемой меры близости графов. Поэтому плоская кластеризация может служить средством выделения тем (особенностей) в тексте. Данный подход известен [7], однако мы воспользуемся более общим подходом, использующим понятие регулярности [1].

Назовем концептуальный граф *регулярным*, если он является обобщенным графом хотя бы для одного другого графа. На практике свойство регулярности на концептуальных графах поддерживается множествами графов, имеющих обобщенные графы.

Назовем концептуальный граф *особенным*, если он не имеет ни одного регулярного графа.

Утверждение 2. Пусть G – множество концептуальных графов. Если имеет место кластеризация вида

$$G = \{C_1 \subset C_2 \subset \dots \subset C_m, C_n\} \quad (13)$$

то в кластере C_n существуют особенные графы.

Таким образом, особенный кластер, не входящий в иерархию (8), содержит особенные графы.

Множество особенных графов на множестве G соответствует фрагментам текста, имеющим особенность.

Поиск таких фрагментов, выпадающих из общей содержательной картины текста, представляет практический интерес.

Утверждения 1, 2 следует принимать с определенными оговорками. Оба утверждения являются достаточными, то есть не гарантируют необходимость существования декларируемых ассоциативных правил и особенных графов. Тем не менее, утверждения 1, 2 важны тем, что они условно связывают язык концептуальных графов, элементы которого здесь рассмотрены, с операцией кластеризации.

Таким образом, кластеризация становится экспериментальным инструментом решения рассмотренных задач Text Mining.

3 Проект электронной библиотеки

Рассмотренные принципы и решения применены в экспериментальном проекте электронной библиотеки научных статей.

Вместе с реляционной базой данных, содержащей тексты и их параметры, в системе поддерживается XML - база данных концептуальных графов. Выбор данного формата базы обусловлен тем, что языки концептуальных графов [10], [13] используют XML.

В системе возможны ручное и автоматическое построение концептуальных графов. Концептуаль-

ные графы строятся только для аннотаций статей, что позволяет иметь достаточно компактные графы. Аннотации статей загружаются с сервера [18]. Также выполняется обработка электронного ресурса статей конференция RCDL [19].

В настоящее время в системе реализованы следующие функции:

- полнотекстовый поиск в статей;
- построение концептуальных графов;
- кластеризация концептуальных графов;
- визуализация результатов кластеризации в виде дендрограмм (актуальна только в экспериментах с небольшим числом характерных графов).

Приведем некоторые характерные результаты экспериментального решения рассмотренных задач.

В экспериментах исследовались возможности эволюционного подхода к решению задачи кластеризации на концептуальных графах.

Эволюционный подход, описанный в работах [13, 14, 16], основан на применении генетических алгоритмов оптимизации при поиске экстремума целевой функции, отражающей меру близости концептуальных графов друг другу. Использовались меры близости (10) – (12).

Исследовались варианты кодирования популяций в схемах генетических алгоритмов, алгоритмы рекомбинации, влияние вероятностных параметров алгоритма на сходимость.

Моделирование семантики аннотаций текстов научных статей при помощи концептуальных графов позволило обнаружить особые типы аннотаций, важные как средства репрезентативности корпуса текстов научной тематики – это аннотации *декларативного* типа и аннотации *повествовательного* типа.

Аннотация декларативного типа содержит небольшое число предложений. Концептуальная близость предложений невелика. Относительная близость может быть высокой за счет применения типичных оборотов речи, характерных для подобного типа аннотаций.

Аннотация повествовательного типа содержит большее число предложений, чем аннотация декларативного типа. Концептуальная близость предложений высокая или фрагментарно высокая, когда концептуально связаны несколько предложений, составляющих фрагмент текста.

В эксперименте, результаты которого представлены на рис. 1, 2 выбрана длинная повествовательная аннотация из 11 предложений (пронумерованных от 0 до 10):

0. We assume that the modality (i.e., number of local optima) of a fitness landscape is related to the difficulty of finding the best point on that landscape by evolutionary computation (e.g., hillclimbers and genetic algorithms (GAs)).
1. We first examine the limits of modality by constructing a unimodal function and a maximally multimodal function.

2. At such extremes our intuition breaks down.
3. A fitness landscape consisting entirely of a single hill leading to the global optimum proves to be hard for hillclimbers but apparently easy for GAs.
4. A provably maximally multimodal function, in which half the points in the search space are local optima, can be easy for both hillclimbers and GAs.
5. Exploring the more realistic intermediate range between the extremes of modality, we construct local optima with varying degrees of “attraction” to our evolutionary algorithms.
6. Most work on optima and their basins of attraction has focused on hills and hillclimbers, while some research has explored attraction for the GA's crossover operator.
7. We extend the latter results by defining and implementing maximal partial deception in problems with k arbitrarily placed global optima.
8. This allows us to create functions with multiple local optima attractive to crossover.
9. The resulting maximally deceptive function has several local optima, in addition to the global optima, each with various size basins of attraction for hillclimbers as well as attraction for GA crossover.
10. This minimum distance function seems to be a powerful new tool for generalizing deception and relating hillclimbers (and Hamming space) to GAs and crossover. Все предложения аннотации логически связаны.

Другая аннотация состоит из одного предложения:

11. This paper describes an initial version of a library of sharable and reusable medical ontological theories, organized according to a proposed classification of ontologies.

Длинная повествовательная аннотация содержит большое число различных слов, которые могут также встречаться во второй короткой аннотации.

Тем не менее, необходимо отделить единственное, последнее по номеру предложение 11 от остальных.

Обе меры близости, использованные в экспериментах рис. 1, 2, обеспечивают успешную кластеризацию: граф с номером 11 представляет собой отдельный кластер. Из рис. 2 видна иерархичность полученной кластеризации и соответствие ее виду (13). Это следствие «повествовательности» первой аннотации из 11 предложений, проявляющейся во «вложенности смыслов» предложений. Соответственно, на полученных кластерах имеют место ассоциативные правила вида (5). Граф с номером 11 –

особенный. Ему соответствует предложение, образующее *особенную аннотацию*.

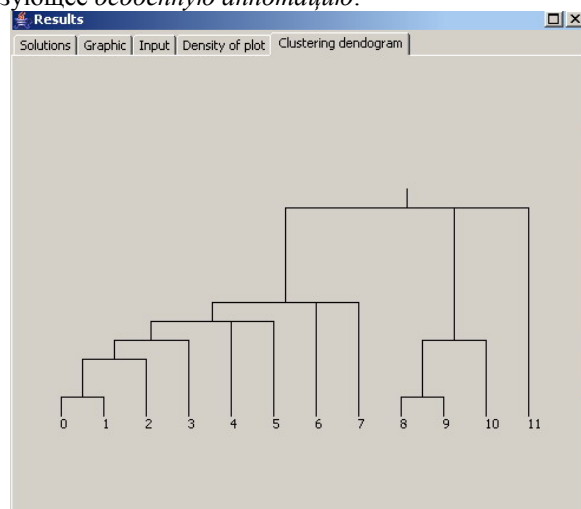


Рис. 1. Результат кластеризации генетическим алгоритмом с использованием концептуальной меры близости в виде дендрограммы (внизу показаны номера концептуальных графов)

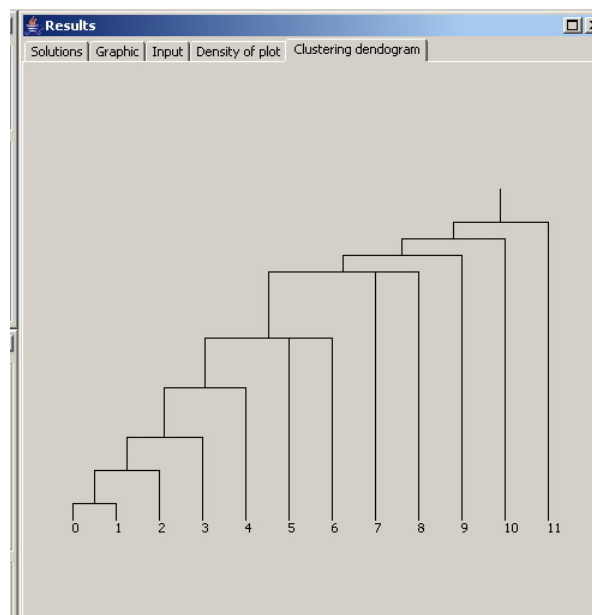


Рис. 2. Результат кластеризации генетическим алгоритмом с использованием реляционной меры близости

4 Выводы и дальнейшие исследования

Таким образом, применяя кластеризацию концептуальных графов, можно решать некоторые задачи Text Mining на текстах, моделируемых при помощи концептуальных графов. При этом реляционная мера близости концептуальных графов полнее отображает семантическую близость текстов.

Другим, интересным инструментом анализа концептуальных графов является аппарат математиче-

ской логики, предложенный еще в пионерских работах Р. Совы [10]

Применение данного аппарат позволит исследовать ассоциативные правила как клаузы, что следует из их строгого определения (5)

Это позволит получать математически строгие результаты, имеющие новые приложения

Литература

- [1] Agrawal, R., A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant. The Quest Data Mining System, Proc. of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August, 1996.
- [2] Boytcheva, S. Dobrev, P. Angelova, G. CGExtract: Towards Extraction of Conceptual Graphs from Controlled English. Lecture Notes in Computer Science № 2120, Springer 2001.
- [3] D. Hand, H. Mannila, P. Smyth Principles of Data Mining. MIT Press, Cambridge, MA, 2001, 445 p.
- [4] Feldman D., Hirsh M., Mining Associations in Text in the Presence of Background Knowledge. - Proc. of the 2nd International Conference on Knowledge Discovery (KDD-96), Portland, 1996.
- [5] Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics 28(3), 2002, p.p. 245--288.
- [6] Hensman, S., Dunnion, J. Applying Verbnet For Automatic Semantic Role Identification. Обработка текста и когнитивные технологии : Сб. / Под ред. Полякова В.Н., Соловьева В.Д. – Варна: 2003 (Вып.8), с. 471–481.
- [7] Knorr, N. Algorithms for Mining Distance-based Outliers in Large Datasets, Proc. of the International Conference on Very Large Data Bases (VLDB'98), Newport Beach, CA, 1998.
- [8] Foo, N., Garner, B. J., Rao, A., Tsui, E. Semantic distance in conceptual graphs. In: Conceptual structures: current research and practice book contents. Ellis Horwood Upper Saddle River, NJ, USA, 1992. P. 149–154.
- [9] Montes-y-Gomez, M., Gelbukh, A., Lopez-Lopez, M. Text Mining at Detail Level Using Conceptual Graphs. Lecture Notes In Computer Science; Vol. 2393. P. 122–136.
- [10] Sowa R., Conceptual Graphs: Draft Proposed American National Standard, International Conference on Conceptual Structures ICCS-99, Lecture Notes in Artificial Intelligence 1640, Springer 1999.
- [11] Tan, A. Text Mining: The state of the art and challenges, Proc. of the Workshop Knowledge Discovery from advanced Databases PAKDDD-99, April 1999.
- [12] Wermelinger M. Conceptual Graphs and First Order Logic. - 3 rd International Conference on Conceptual Structures. Lecture Notes in Artificial Intelligence, Vol 954. – Springer Verlag, 1995, p. 323–337.
- [13] Богатырев М.Ю., Латов В.Е., Столбовская И.А. Применение концептуальных графов в системах поддержки электронных библиотек. – Электронные библиотеки: перспективные методы и технологии, электронные коллекции // Труды Девятой Всероссийской научной конференции RCDL'2007 (Переславль-Залесский, Россия, 15–18 октября 2007). – Т. 2, С. 104–110.
- [14] Богатырев М.Ю., Латов В.Е., Столбовская И.А., Тютин В.В. Эволюционный подход к задаче кластеризации на концептуальных графах и его применение в системах поддержки электронных библиотек. – Математические методы распознавания образов. Тринадцатая Всероссийская конференция : Сб. докладов. – М. : МАКС Пресс, 2007. – 668 с. – С. 464–468.
- [15] Богатырев М.Ю., Тютин В.В. Построение и анализ концептуальных графов в системах поддержки электронных библиотек. – Известия ТулГУ. Сер. Технические науки, 2008 (в печати).
- [16] Богатырёв М.Ю., Латов В.Е. Исследование генетических алгоритмов кластеризации. – Известия. ТулГУ. Сер. Математика Механика. Информатика. Том 8, вып. 3 . Информатика. – Тула, 2002. – С. 101–107.
- [17] Электронный ресурс: Cognitive Technologies – Интеллектуальные технологии управления. <http://www.cognitive.ru>
- [18] Электронный ресурс: Scientific Literature Digital Library <http://citeseer.ist.psu.edu/>
- [19] Электронный ресурс: Труды конференций Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции <http://rcdl.ru/>

Solving Some Text Mining Problems with Conceptual Graphs

M.Y. Bogatyrev, V.V. Tuhtin

The problem of finding Associated Rules and other Text Mining problems formulated on Conceptual Graphs. It is shown that Clustering is appropriate way to solve these problems.

RCDL'1999–RCDL'2008: DL, VDL, Semantic Web/GRID, GRID...

© В.П. Шириков

Объединенный институт ядерных исследований (ОИЯИ)
shirikov@jinr.ru

Аннотация

Данная статья должна восприниматься как личный короткий авторский обзор достижений в области разработки и реализации электронных библиотек, представлявшихся участниками ежегодных конференций RCDL (в основном российскими) в последние 10 лет. Все оценки и соображения в тексте обзора приведены в соответствие с личным авторским пониманием известных технологий, необходимых для реализации электронных библиотек и в других областях e-Science и ее приложений.

Наша конференция – заключительная в 10-летнем цикле российских конференций по электронным библиотекам (DL). Еще на самой первой 1999-го года в разных докладах само понятие DL трактовалось по-разному: либо только как «зеркальное» Интернет-отображение ресурсов традиционных библиотек и средств их использования (как это было в докладе авторов из Технологического Университета в Мюнхене), либо уже в той расширенной трактовке, которая тогда была дана в докладе М.Р. Когаловского и преобладает сейчас: как хранилища (коллекции) знаний и данных общего и специального назначения, реализованные в сетевой среде с применением современных технологий и стандартов формирования, модификации и средств использования информации. Теме создания и использования информационной модели, способной отобразить структуру и семантику таких гетерогенных коллекций, на этой же конференции был посвящен доклад Л.А. Калиниченко "Integration of Heterogeneous Semistructured Data Models in the Canonical One": сразу заметим, что своеобразным логическим продолжением данной работы из последних стала доложенная с соавторами на конференции GRID'2008: "Application Driven Mediation Middleware Grid Infrastructure for Problem solving over multiple Heterogeneous Distributed Information Resources". К 1999-му году Web с первыми версиями Semantic Web был достаточно оснащен технологиями, минимально необходимыми для реализации распределенных информационных систем, в том числе виртуальных DL и коллекций разного типа, поэтому во второй половине 1990-х годов и в последующем

выполнено много работ по Web-ориентированным системам DL, доложенных, в частности, на конференциях RCDL и «Научный сервис в сети Интернет» (можно упомянуть хотя бы работы В.А. Серебрякова и его соавторов по созданию инфраструктуры единого научного информационного пространства РАН). Среди необходимых или признанных технологий реализации в последние годы и российскими, и иностранными участниками RCDL (см. презентацию Т. Risse для конференции RCDL'2007, http://rcdl2007.pereslavl.ru/en/doc/Risse_rcdl_tutorial.pdf), упоминаются следующие:

Grid, Pear-to-Pear, SOA, Semantic Web с его средствами применения онтологий. В указанной презентации подчеркивается тенденция перехода от понятия DL как интегрированной системы с централизованным управлением к динамически конфигурируемой федерации DL-сервисов и информационных коллекций, когда возникает понятие виртуальной DL как коллекции данных от разных контент-провайдеров без централизованного управления и с распределенными гетерогенными сервисами. Естественно, что предоставляемые каждым провайдером сервисы должны поддерживать поисковые средства для различных типов информации (разных типов мультимедийных данных, поиска по контексту, многоязыкового доступа), индексирование, аннотирование документов, предоставление регистров метаданных и ресурсов; в составе DMS виртуальной DL могут ресурсоемкие сервисы (например, для извлечения характеристических деталей из image/audio/video-документов, генерации метаданных, автоматического редактирования сложных документов). С учетом перечисленных выше требований и был реализован, в частности, проект BRICKS (<http://www.brickcommunity.org>), научным и техническим директором которого является Т. Risse. В своей презентации он останавливается в основном на характеристике концептуальной модели и ее реализации в проекте, прикладной задачей стала интеграция ресурсов в общую и разделяемую DL, куда входят в качестве локальных материалы электронных музеев, архивы исторических документов и другие виды электронной памяти (цифрового наследия культурного многообразия: Building Resources for Integrated Cultural Knowledge Services); узлы (сайты) созданной по проекту сети BNnet оснащаются свободно доступными программными интерфейсами-«кирпичиками» (bricks), через которые без централизованного управления они

взаимодействуют друг с другом и используют доступные ресурсы для работы с контентом и метаданными по принципу «равный с равным» (P2P); каждый узел может знать напрямую только одно подмножество других узлов, но если хочет использовать внешние по отношению к этому подмножеству ресурсы, то может послать запрос к одному из известных ему и тот сформулирует запрос. T. Risse не акцентирует свое внимание на возможном применении своих средств в рамках Grid-структур, по существу это выглядит как Web-ориентированная реализация, сама по себе не нуждающаяся в явном виде в каких-то сервисах из набора Grid-middleware, поэтому упоминание им Grid-технологий в указанном списке можно понимать по-разному.

Моей целью не был обзор тех многочисленных работ российских авторов, которые решали чем-то похожие задачи с применением технологий из приведенного списка и докладывали их на прошедших конференциях RCDL, и я хочу только в основном остановиться на отношениях в развитии программных средств Grid-middleware и средств DL/VDL (т.е. как они сложились в прошедшие годы), поскольку их независимое использование уже сейчас в целом ряде случаев оказывается или неоправданным, или невозможным.

Так получилось, что хотя к тому же 1999-му году концепция Grid и его технологий была сформулирована и появились первые реализации его middleware в пакетах типа GT1, но она была представлена как концепция использования распределенных вычислительных ресурсов (BP) и в значительной мере предусматривала только создание нового уровня системы распределенной потоковой обработки задач: Web как информационную систему и Grid как вычислительную систему объединяло на технологическом уровне только общее использование сетевых протоколов нижнего уровня, поэтому изначально системное Grid-middleware своими сервисами обеспечивало только поиск компьютерных вычислительных ресурсов для выполнения задачи (job, понятие не из области информационных систем) в соответствии с ее требованиями в сопровождающем ее JDL-файле. Конечно, большинство задач порождает или использует какие-то данные, сосредоточенные в DBMS или файловых системах, поэтому выполнение требований на предоставление таких данных задаче могло частично выполняться сервисом DMS из состава базового middleware, но в основном даже и сейчас работа с распределенными данными – это проблема приложений, пользующихся возможностями того сервисного программного обеспечения, которое "on top of Grid-middleware": к нему можно отнести, например, сервисы OGSA-DAI/DQP, изначально разработанные для пользователей британской архитектуры GridPP и AstroGrid и включенные в пакеты уровня GT3/GT4, а также использованные для работы с распределенными базами данных. Конечно, развитие технологий распределенных информационных систем в чем-то обязано той стандартизации, которая при создании Grid-

архитектур была отработана для формирования VO/IVO и создания SOA, да и ПД (Пространств Данных). В материалах прошлогодней конференции RCDL'2007 хотелось бы в связи с этим отметить работу А.В. Жучкова, А.В. Кравченко и Н.В. Твердохлебова (авторов из ИХФ РАН) «Сервис-ориентированный Грид-подход к информационным задачам в пространствах данных виртуальных организаций». На примере перечня требований к системе DSSP (Data Space Support Platform), охарактеризованной в материалах ACM SIGMOD в 2005–2006 гг., авторы указанной работы подробно расписали те средства, которые заложены в рамках пакета GT4 и могли бы использоваться достаточно эффективно для реализации многих элементов перечня. Рассматривая пример разработанного ими сервиса высокого уровня для обеспечения информационной поддержки проведения исследований по разработке вакцин, они отмечали, что его реализация без использования Grid-технологий была бы затруднительна по двум причинам: вычислительная сложность процедуры лексического анализа (здесь средства GT4 позволяют распределить работу по подходящим узлам VO) и отсутствие необходимости в централизованном управлении. Заметим, что и в этой работе отмечается роль сервисов OGSA-DAI/DQP, независимо успешно применяемых также в совместных работах ИПИ РАН (В.Н. Захаров, Л.А. Калининченко) и SAO РАН в рамках деятельности IVOA и RVO: эти работы также докладывались на наших конференциях RCDL и симпозиуме по онтологическому моделированию, упоминаемому в конце обзора. Кстати, значимость и признание полезности упомянутых сервисов подтверждается и тем, что они были использованы в продукте IBM Sphere Information Integrator для обеспечения доступа к виртуальной базе данных (как набору физических разнотипных СУБД) через универсальный SQL или XML-запрос, автоматически преобразуемый в формат, понимаемый той или иной физической СУБД: в этой работе помогают программные Java-посредники Wrappers между IBM Sphere и источниками данных. Ключевой идеей «виртуализации» реализации подобного "Data Grid" была естественная необходимость избавить прикладную задачу от необходимости знать – где физически размещены данные, в которых она нуждается, и как они структурированы: это отмечается и в докладе на RCDL'2007 авторов из ИФХ РАН, и независимо в моем обзоре на прошлогодней конференции «Научный сервис в сети Интернет».

Говоря об информационных сервисах "on top of Grid-middleware", сосредоточенного в пакетах типа GT4 и обслуживающего разные Grid-архитектуры, нельзя не остановиться о том, что происходит в структуре EGEE (European Grid for E-science), в рамках которой функционирует и развивается и структура RDIG (Russian Data Intensive Grid). Изначально прикладными областями, ради которых начинал развиваться EGEE, были определены науки о Земле, физика высоких энергий, биоинформатика и медици-

на, астрофизика. Сейчас примерно 68% пропускной способности EGEE и его ресурсов тратится на задачи, связанные с подготовкой к запуску в ЦЕРН (Женева) нового ускорителя LHC и основных экспериментальных установок на канале его пучков (т.е. на физику высоких энергий). Учитывая уникальность готовящихся экспериментов и заинтересованность в них стран-участниц ЦЕРН и его коллаборантов во многих странах мира (в том числе России), большинство существующих в них Grid-структур старается реализовать средства совместимости с EGEE и его базовым программным Grid-middleware (к июлю 2008 г. это стал набор gLite версии 3.1). Его разработка частично шла с учетом имевшегося в пакетах GT, но многое создано в рамках деятельности групп в ЦЕРН как головной организации по проекту EGEE, регулирующей все отношения с консорциумом из 91 представителя от институтов и стран, участвующих в проекте и комплектовании его ресурсов. Распределенные компьютерные ресурсы и средства хранения информации EGEE позволяют обрабатывать сейчас ежедневно более 100 000 задач, к моменту запуска LHC ожидается как минимум удвоение их возможностей. Будучи ориентированным разработчиками изначально на Linux-платформы, EGEE приспособляется и к использованию Windows потребителями его ресурсов. Поскольку возможности EGEE и его базового gLite-middleware уже представили значительный интерес для разработчиков ресурсоемких информационных систем, хотелось бы обратить внимание на два связанных между собой проекта, которые имеют кураторов в группе EGE (ЦЕРН) и ESA (European Space Agency). Речь идет о проекте DILIGENT (A Digital Library Infrastructure on Grid Enabled Technology, <http://www.diligentproject.org>) и его системе сервисов "gCube on top of gLite", техническим и научным координатором которого является Донателла Кастелли из CNR-ISTI (Пиза, Италия), и его продолжении с 2008 г.: проекте D4Science (<http://www.d4science.eu>). Я хочу в данном обзоре как-то отобразить причину того интереса, который был проявлен к этим проектам со стороны руководства ПК RCDL'2008 и RCDL Steering Committee, планировавшего организацию тьюториала на нашей конференции.

Одной из первых прикладных целей DILIGENT было создание сервисов для проекта SAPIR (Search in Audio Visual Content Using Peer-to-Peer IR) как части проекта Chorus (средства Multimedia Content Search Engines), т.е. для задачи создания в интересах этих проектов нового типа представления и поиска данных, отсутствовавших в традиционно используемых поисковых системах типа Google и Yandex, ограниченных средствами текстовой формы запросов и текстового описания данных, ассоциирующегося с мультимедийным контентом. Среди примеров желаемого, приводимых на сайтах этих проектов – поиск для пользователя мобильного устройства информации об объекте после ввода в качестве запроса фотографии этого объекта или поиск текста песни по ее мелодии (заметим кстати, что проблема нахождения в базе данных по

образу достаточно не новая и для российских авторов, достаточно упомянуть из последних работ хотя бы сообщение на нашей RCDL'2008 от авторов из Петрозаводского университета «Система поиска в электронной коллекции изображений петроглифов Карелии»). Сама по себе проблема подготовки базы мультимедийных данных для того, чтобы по описаниям и представлениям их фрагментов можно было удовлетворять такие запросы, например, для модифицированного Web-приложения Flickr (см. <http://www.flickr.com/tour> по использованию цифрового архива фотографий и обмена ими) связана с обработкой каждой фотографии для представления и описания характерных деталей (черт лица, например), и это достаточно ресурсоемкая задача с вычислительной точки зрения. Во второй половине 2007 г. с этой целью с применением сервисов "gCube on top of gLite" был проведен на инфраструктуре EGEE 16-недельный прогон (data challenge) по обработке 37 млн фотографий из on-line базы данных Flickr, сгенерировано около 112 млн текстовых и image-объектов (примерно 5TB данных); ресурс-брокером в EGEE PPS (Pre-Production Service) было пропущено более 44 тыс. задач (jobs), в каждой обрабатывалось примерно 1000 фотографий. Составленная коллекция

(см. <https://twiki.cern.ch/twiki/bin/view/DILIGENT/DiligentFlickrDL>) и должна использоваться проектом SAPIR для развития новой крупномасштабной технологии контекстно-ориентированной классификации и выборки данных, расширяющей пределы возможностей традиционных поисковых механизмов. Почему DILIGENT попал в сферу интересов группы EGE ЦЕРН? Да может хотя бы потому, что исходная информация о событиях, регистрировавшихся на экспериментальных установках при ускорителях, изначально представлялась в виде обычных фотографий и затем их электронных подобию, а астрономия и астрофизика до сих пор используют фотографии, которые ежемесячно публикуются в издании CERN Courier или размещаются на сайте www.sai.msu.su/apod, но это уже мои домыслы... Во всяком случае, в программе прошедшей в октябре 2007 г. международной конференции EGEE'2007 была и секция по проекту DILIGENT как средству для интеграции Grid и DL, что, как отмечалось, заложит основы нового поколения инфраструктуры использования знаний в электронизированной науке (e-Science) для разных областей исследований и в промышленности. Интерес ESA к EGEE и сервисам типа "gCube on top of gLite" также понятен: накапливается спутниковая информация о Земле, лет через 10 это примерно 12 Петабайт данных, нужна оперативная информация для мониторинга и систем охраны окружающей среды, контроля состояния ее ресурсов и результатов их использования, а это, в частности, одна из главных задач упомянутого проекта D4Science, базирующегося на использовании сервисов DILIGENT. Отмечу сразу, что на прошлой конференции RCDL'2007 был представлен родственный по теме доклад Е.Б. Кудашева и А.Н. Филонова «Технология и стандарты интеграции сервисов и баз данных дистанционного исследования Земли из космоса», и сейчас

на RCDL'2008 в докладе тех же авторов делается обзор текущего состояния развития технологий в области распределенных информационных систем для данных дистанционного зондирования, а также возможные пути развития подобных систем в ближайшие годы, технические аспекты реализации систем трех поколений: от проекта первого поколения INTAS IRIS с участием авторов доклада до инициативы GMES HMA, находящейся в стадии реализации и сфере интересов и NASA, и ESA, и прикладной науки о Земле.

Теперь немного более подробно о сервисах gCube (см. материалы по ссылкам <http://www.gcube-system.org/architecture/overview.html> и <http://www.gcube-system.org/architecture/services/services.htm>).

В самом общем определении gCube позволяет исследователям динамически, по требованию (on-demand) создавать информационно-вычислительные среды (Virtual Research Environments, VREs), агрегируя и формируя контент-ресурсы, прикладные сервисы и компьютерные ресурсы как за счет собственных у прикладных проектов, так и за счет имеющихся в EGEE. В составе gCube достаточно средств для мониторингирования использования разделяемых ресурсов с гарантией их оптимального распределения и эксплуатации, а также легкого создания Web-порталов для VREs, через которые пользователи могут иметь доступ к контенту и сервисам; предоставляется также набор типовых для DL функций (поиск, аннотирование, формирование, визуализация документов и др.).

Не все виды сервисов были изначально подробно описаны в указанных выше разделах сайта, например, сервис vdlgenerator (Virtual Library Creation and Management), дающий возможность пользователям или их сообществам создавать собственные DLs: он позволяет специфицировать набор критериев, отражающих ожидаемые характеристики новой DL, и затем определяет сервисы и информационные источники, нужные для обеспечения для обеспечения требуемых свойств DL и ее контента. Само формирование и функционирование DL обеспечивается выбором доступных DHNs (Diligent Hosting Nodes), Grid-узлов: они, как уже было сказано, берутся из собственных компьютерных ресурсов проектов или получаются от сервисов gLite (его брокера) с учетом требований на свойства (Property requirements) нужного ресурса (тип операционной системы, мощность CPU, размер дисковой памяти, наличие нужного software: например, конкретной библиотеки программ общего или специального назначения). Ключевую роль играют такие сервисы общего назначения (Collective layer services), как Keeper (с его распределяемым набором программных компонент, нужных для создания и поддержки функционирования динамических DL) и DIS (Distributed Information System): все ресурсы и их состояние регистрируются здесь и их статус при необходимости опрашивается и корректируется другими сервисами системы через универсальный API-интерфейс.

Возвращаясь к теме обзора и итога работ, выполненных российскими участниками конференций RCDL и опубликованных в трудах, доступных на

сайтах этих конференций, нельзя не отметить, что многие их важные итоговые работы были доложены в мае этого года в Звенигороде на симпозиуме «Онтологическое моделирование: состояние и направления исследований и применения», организованном ИПИ РАН и Московской секцией ACM SIGMOD, поэтому судить о состоянии дел на текущий момент по проблематике RCDL нельзя без учета материалов этого симпозиума по ссылкам:

<http://synthesis.ipi.ac.ru/synthesis/ontologyprogram>

<http://synthesis.ipi.ac.ru/synthesis/publications>

<http://www.cemi.rssi.ru/mei/articles/koga08-1.pdf>

В заключение хочется отметить следующее личное впечатление: как-то подзабытой оказалась та концепция Semantic Grid, которая была сформулирована еще в 1999 г. Keith G.J. effry и затем детализована в 2002 г. в статье David De Rouge et al "The Semantic Grid: a Future e-Science Infrastructure" (<http://www.semanticgrid.org/documents/semgrid-journal/semgrid-journal.pdf>), где авторы предсказывают, что программная среда компьютеризированной науки и все Grids должны будут включать в себя трехуровневую систему сервисов:

Data/Computation Services, средства размещения данных и их транспортировки между обрабатывающими программами, обеспечение вычислительных и сетевых ресурсов;

Information Services, средства представления, запоминания и доступа к информации, управление ею;

Knowledge Services, средства накопления, представления, обновления, «публикации» (сетевой распространения) знаний для помощи ученому в его исследовательском процессе.

Все положения демонстрируются детальным формализованным примером цикла полной автоматизации обработки экспериментальных данных в сетевой компьютерно-аппаратной среде (от начала поступления данных на анализ до подведения итогов результата обработки научным сообществом) с применением конкретного перечня сервисов каждого из указанных уровней, подчеркивается роль семиуровневой системы онтологий для нормального функционирования всей клиент-сервисной структуры приведенного примера.

Что-то до сих пор не видно и сейчас в нашей практике примеров подобной реализации.

RCDL'1999–RCDL'2008: DL, VDL, Semantic Web/GRID, GRID...

V.P. Shirikov

This article must be imaged as private author's short review of researches in the field of digital libraries, presented by RCDL community and some foreign specialists for our RCDL's in the last 10-years period. Of course, author's estimations in this review were done in the main in accordance with private understanding of technologies, necessary to be taken in account for realizations of digital libraries and in other fields of e-Science and its applications.

Preservation Planning with Plato*

© Hannes Kulovits, Andreas Rauber

Vienna University of Technology, Vienna, Austria
{kulovits, rauber}@ifs.tuwien.ac.at

Abstract

Due to the rapid technological changes in today's information landscape the preservation of digital information has turned into a pressing challenge affecting everybody from large cultural heritage institutions, via industry and small and medium enterprises, to private users wanting to preserve their holiday photos and email communications. Several examples of considerable data loss because of taking the longevity of digital objects for granted have drawn the public's attention to the problem that digital material does not last forever. Also a comprehensive survey among professional archivists conducted by SNIA's 100 Year Archive Task Force arrived at the result that 'Digital information is at risk of being lost' [15] and underlines the growing awareness of the urgency of digital preservation. The challenge of maintaining the physical readability is just one among multitude other problems one comes along when dealing with this issue.

A lot of different strategies, i.e. preservation actions, have been proposed to tackle this challenge. However, which strategy to choose, and subsequently which tools to select to implement it, is a crucial decision which poses significant challenges. The decision must be based on a well-documented and profound analysis of the requirements and performance of the tools taken into consideration.

This tutorial will introduce the major challenges that digital preservation activities face. It will then discuss the advantages and disadvantages of the various strategies that exist to tackle these problems. The OAIS reference model, an ISO standard describing the functional model as well as the information model of an archival information system will be presented, before focusing on a specific task in digital preservation, namely planning a sound strategy for preserving one's digital assets. To this end a solid workflow that leads to a profound, well-documented decision that

one can be held accountable for, will be discussed. Additionally, we will present Plato, a software tool that is supporting this workflow, automating the core steps in this endeavour.

1 Introduction

Digital objects have become the dominant way that we create, shape, and exchange information. They increasingly contain essential parts of our cultural, intellectual and scientific heritage; they form a central part of our economy, and increasingly shape our private lives. The ever-growing heterogeneity and complexity of digital file formats together with rapid technological changes turn the preservation of digital information into a pressing challenge. The challenge is to keep electronic data accessible, viewable, and usable for the future, to ensure the survival of our digital artefacts when the original software or hardware to interpret them correctly becomes unavailable [16].

Digital preservation denotes the process of keeping electronic material accessible and usable for a certain period of time. The Digital Preservation Coalition defines it as *the series of managed activities necessary to ensure continued access to digital materials*, and moreover, *refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technical change* [7]. The preservation problem of digital data can be seen as twofold: (1) The physical media the data is recorded on as bit stream must be preserved. The lifespan of data written on today's prevalent media such as DVD or CD-ROM is considered only a couple of years, so reading from them might already be a serious problem in a decade. (2) As digital data is merely a series of binary codes the way to interpret this stream must be preserved. Digital objects require specific programs (and program versions) to open and render them, these in turn require a set of specific libraries and an operating system, which in turn runs on and supports a specific type of hardware components. If any of these is lost, a digital object cannot be rendered anymore and is left to be a useless concatenation of binary data. Even the serendipity of a storage medium being capable of retaining digital data for a millennium is worthless if the means of interpreting it is lost. A good overview of the challenges in Digital Preservation and of preservation strategies is provided in the accompanying document to the

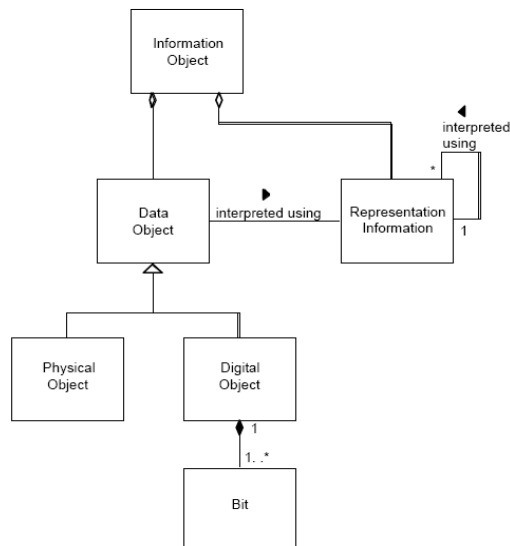


Figure 2 Information Object in the OAIS Model

UNESCO charter for the preservation of the digital heritage [16].

The remainder of this paper is organized as follows: The next section gives an overview of the OAIS reference model as a conceptual framework for an archival system. Section 3 provides an overview of digital preservation, its challenges, and different preservation strategies. Section 4 outlines the Planets Preservation Planning approach and describes the steps of the planning workflow. Section 5 describes the Planets preservation planning tool Plato and its application, and Section 6 summarizes this work.

2 The OAIS Reference Model

In digital preservation standards are very important not only because they promote openness but also for interoperability and sustainability reasons. A generic standard is the Open Archival Information System (OAIS) reference model which was published in 2002 by the Consultative Committee for Space Data Systems (CCSDS). It has proven to be a very useful high-level reference model, describing participants, roles and responsibilities as well as the exchange of information. It does not recommend or specify any particular implementation but identifies and defines certain terms and components that might be involved in an archival information system. Because of its growing acceptance in the community, the OAIS model is the most common framework for digital preservation systems. The reference model was adopted as a standard and registered under ISO 14721:2003 [6]. It defines an OAIS as ‘... *an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community*’ over an indefinite period of time. Furthermore, the OAIS model ‘... *provides a framework for describing and comparing different long term preservation strategies and techniques.*’

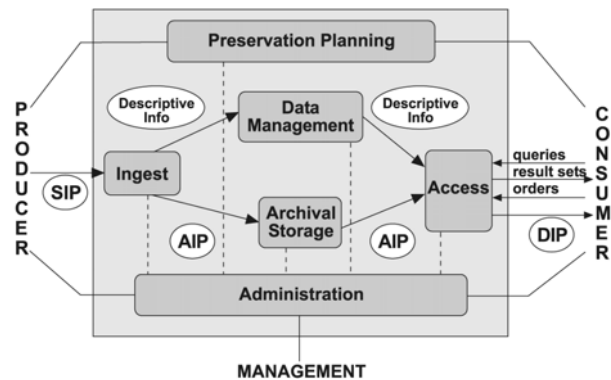


Figure 1 OAIS Reference Model

Figure 1 shows the main functional components and information packages of the OAIS model with the Producer, Consumer, and Management as the main stakeholders of the whole system. While the Producer provides the information to be preserved and the Consumer obtains particular preserved information of interest, the Management role sets overall OAIS policy within a broader policy domain.

The fundamental concept of the OAIS model is the concept of an Information Object consisting of Data and Representation Information whereas the Information Object can be either physical or digital. Representation Information is the information necessary to be capable of fully interpreting the data. Figure 2 illustrates the concept of the Information Object.

The OAIS model differentiates three information package variants: (1) Submission Information Package (SIP) which is negotiated between the Producer and the OAIS, and sent to the system by a Producer. (2) Archival Information Package (AIP) which is used for preservation and comprises a complete set of Preservation Description Information (PDI) for the Content Information. (3) Dissemination Information Package (DIP) which includes a part or all of one or more AIPs and is delivered to the user community by the OAIS.

When a producer submits a digital object into the system, it has to be packaged together with required metadata as a Submission Information Package (SIP). The *Ingest* module provides the services and functions to accept SIPs from Producers. It further performs quality assurance and generates the AIP complying to the archive's standards. Ingest also extracts descriptive information from the AIPs and coordinates updates to *Archival Storage* and *Data Management*. Archival Storage stores, maintains and retrieves AIPs, while Data Management populates, maintains and accesses descriptive information about archived objects as well as administrative data. Every action inside the archive that affects the object is added to the metadata of the AIP. The *Access* component is responsible for supporting consumers in finding, requesting and receiving information stored in the system. Access functions include access control, coordinating requests, generating responses as DIPs and delivering the responses to Consumers.

The *Preservation Planning* entity monitors the environment and provides recommendations to ensure the long-term accessibility of the stored information. This includes monitoring and evaluation of the archive and periodical recommendations on archival updates for migration. A central component is the development of preservation strategies and standards as well as packaging designs and plans.

3 Digital Preservation

3.1 Preservation Strategies

In recent years several different ways of preserving a digital collection have been evaluated, developed, and deployed. Most research on actual solutions is focused on two prevailing preservation strategies – migration and emulation.

Today, **migration** is the most common preservation strategy for reasons as the comparatively small initial effort necessary and price in terms of cost per object. Migration refers to updating a digital object by converting it from one hardware or software generation to another generation or representation. The main goal of migration is to transform to a more suitable or stable representation which is better suited for long-term access. When talking about file formats this might be a newer version of the same file format or a different format. For example, a text document stored in Microsoft's document format MS Word 97 can be migrated either to a newer more up-to-date version (e.g. MS Word 2007) or to PDF/A [5] which is considered more stable. Migration is subdivided into four types varying in the amount of changes applied to the object which in turn result in different levels of possible information loss [2]: (1) Refreshment, (2) Replication, (3) Repackaging, and (4) Transformation. The method involving the lowest risk of information loss is *Refreshment* where the physical medium is replaced by another physical medium with the same storage-mapping infrastructure. The bit stream stored on the medium is transferred without altering it and is used in order to prevent the data from being lost due to media decay. *Replication* is similar to Refreshment. It doesn't alter the Packaging Information but may require changes in the mapping infrastructure in Archival Storage. *Repackaging* causes some changes in the bits and bytes of the Packaging Information. *Transformation* poses the highest risk because changes to the actual content are made.

Basically there are three approaches to the migration of a collection of digital objects: Firstly, the whole collection can be converted to a file format that is standardised, with the goal of reducing the complexity of maintenance. The second option is to keep the original bit stream and migrate when the user accesses the object (migration on demand). Lastly, 'migration within the same format' performs migration every time the underlying format is at risk or changes significantly.

The main challenge when migrating from one format to another is to only transform the logical representation, ensure authenticity and consistency and

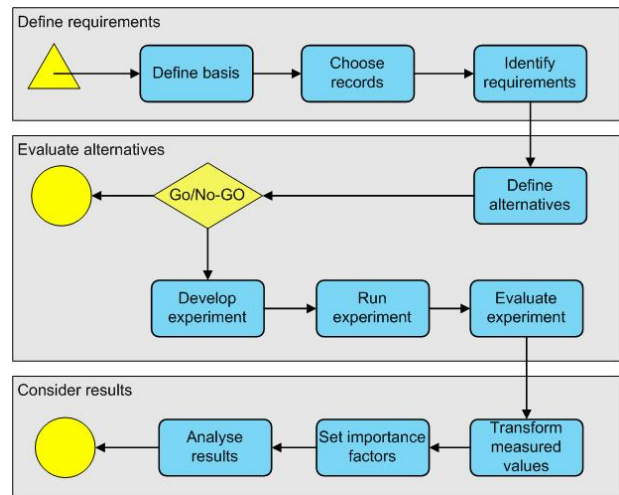


Figure 3 PLANETS Preservation Planning Approach

preserve all essential characteristics for the original object. The Council of Library and Information Resources investigated several migration projects and identified different kinds of risks [8]. As migration has to be performed continuously, i.e. at all times a file format is considered to be at risk, Jeff Rothenberg points out six problems of a migration approach [11]: (1) Labour intensive, (2) Time consuming, (3) Expensive, (4) Error prone, (5) Risky and (6) Non-scalable.

In contrast to migration, **emulation** operates on the environment of the object rather than the object itself. Emulation denotes the recreation of functionality of systems (software and/or hardware) which is needed to render, access, or edit a certain document to overcome technological obsolescence. In digital preservation this means mostly the emulation of a certain version of a software system needed to access a file in an obsolete version or format. Jeff Rothenberg [11] envisions a framework of an ideal preservation surrounding. The Universal Virtual Computer (UVC) concept [4] uses elements of both, migration and emulation, allowing digital objects to be reconstructed in their original appearance. The UVC is independent of any existing hardware or software; it simulates a basic architecture including memory, register and rules. An emerging approach of emulation is modular emulation. It imitates the hardware environment by emulating the components of the hardware architecture. Each hardware component can be seen as an emulator of the component and the components are assembled in order to create a full emulation process.

Emulation is considered the better means for preserving complex objects as the more complex digital objects get, the more loss can occur in the migration process, i.e. the more significant properties can get lost. While having the advantage of leaving the original file untouched without any modifications, writing an emulator is a very complex and time-consuming problem. Additionally there will probably be a point in future where users no longer know how to interact with today's applications such as a certain word processing

software. Several projects have been working on the development of emulation approaches. The BBC Domesday project for instance implemented an emulation strategy to preserve access to the digitised version of a 900-year-old Britain book called the Domesday book [9].

Although not a desirable approach **data archaeology** also has been proposed as a preservation strategy especially because sometimes it might be essential to rescue neglected digital data of potentially vital importance. Also unexpected disasters may necessitate the application of special techniques to recover data already considered lost. An overview of different approaches to data archaeology can be found in [10].

3.2 Authenticity of Digital Objects

One of the most challenging problems in conjunction with digital data is that they can be altered easily, without leaving any trace or a clue what the object was like before. Compared to the analogue world a text written on a piece of paper and altered ex post, the original will in some cases be restorable or at least the change can be identified (that it took place) and thus becomes a part of the object. A user who wants to access a digital object at some point in the future must be assured that there have not been any unauthorised changes, neither accidental nor intentional by a malicious manipulation. This means that the digital object must be exactly what it purports to be, i.e. must be *authentic*, and assert confidence in the identity and integrity of it. Digital preservation therefore also has to deal with authentication of digital material in order to prevent it from being maliciously altered or corrupted without knowing it. Various threats to authenticity exist: (1) Decay of storage medium. (2) Different kinds of malicious attacks. (3) Unintentional damage to the object. (4) Natural disaster. (5) Business failure, e.g. company taking care of objects runs out of money. (6) Multiple versions of an electronic document may easily exist and circulate without being discovered.

4 Preservation Planning

4.1 Overview

As already discussed different strategies such as migration and emulation come into question for preserving digital objects such as electronic documents. However, the decision for a specific tool e.g. for format migration or an emulator, as well as appropriate parameter settings for these tools, is very complex. The process of evaluating potential solutions against specific requirements and building a plan for preserving a given set of objects is called preservation planning.

The Planets Preservation Planning approach allows the assessment of all kinds of preservation actions against individual requirements and the selection of the most suitable solution. It enforces the explicit definition of preservation requirements and supports the appropriate documentation and evaluation by assisting

in the process of running preservation experiments. It is based on work performed in the DELOS Digital Preservation cluster introduced in [12] and described in more detail in [13].

4.2 PLANETS Preservation Planning Workflow

Figure 3 provides an overview of the workflow of the Planets Preservation Planning approach, which was described in [14]. *Define requirements* describes the scenario, the collection that is being considered as well as institutional policies and obligations. Then the requirements and goals for a preservation solution in a given application domain are defined. In the so-called objective tree, high-level goals and detailed requirements are collected and organised in a tree structure. *Evaluate alternatives* consists of the definition and evaluation of potential preservation alternatives. Alternatives are therefore identified, including technical settings and required resources for running the experiments. The Go/No-Go-Decision enforces a review of the work in the previous steps. In the experiments the preservation alternatives are applied to the sample records. The final step of the second phase is the evaluation of the experimental outcomes against the requirements and goals defined in the first phase. In the third phase *Consider Results*, results of the experiments are aggregated to make them comparable, importance factors are set and the alternatives are ranked. The stability of the final ranking is analysed with respect to minor changes in the weighting and performance of the individual objectives using Sensitivity Analysis. After this consideration a clear and accountable recommendation can be made for one of the alternatives.

Define Basis

In the first step, the preservation scenario is described in a semi-structured way including the collection to be considered. Information about the collection includes details about the objects, number of objects in the collection, and legal requirements for handling the records. Moreover, the environment is described in which the preservation process takes place including institutional policies for preservation.

Choose records

In this step, a small number of representative sample records from the collection is taken. The samples, usually between 5 and 10 objects, are used for evaluating the preservation alternatives.

Identify Requirements

The goal of this step is to define clearly the requirements and goals (objectives) for a preservation solution in a given domain. High-level goals are specified, collect detailed requirements, and organise them into a tree structure, referred to as the tree of objectives or shortly, 'objective tree'. While the resulting trees usually differ according to specific preservation context, some general principles can be

observed. At the top level, the objectives can usually be organised into four main categories:

Object characteristics – describe the visual and contextual experience a user has when dealing with a digital object. Subdivisions may be ‘Appearance’, ‘Content’, ‘Structure’ and ‘Behaviour’, with lowest level characteristics being colour depth, image resolution, forms of interactivity, macro support, or embedded metadata.

Record characteristics – denote the technical foundations of a digital record, the context, interrelationships among records and metadata. A record can include one or more objects.

Process characteristics – refer to the preservation process. These include usability, complexity, or scalability.

Costs – have a significant influence on the choice of a preservation strategy and may usually be divided into technical and personnel costs or start-up and operational expenditures.

The objective tree is usually created in a workshop with experts from different domains contributing to the requirements gathering process. The tree is independent from the preservation actions that are considered. It models the requirements, not the actions to be taken. The tree documents the individual preservation requirements of an institution for a given collection of objects. Typical trees may contain from 50 up to several hundred objectives, usually organised in four to six hierarchy levels.

Objective trees were initially created with post-it notes on a flip chart. While this is convenient for certain environments, an alternative way has been introduced and the feedback on it has been very positive. This involves the use of mind-mapping software, usually projected onto a large screen to provide an overview, to allow multiple stakeholders working on the tree. A mind-mapping software that has been used several times and has proved very helpful is FreeMind¹ which is open source and can be downloaded and used freely.

Having defined the objectives, the next step is to assign measures to each of the objectives in the tree, which provides metrics to determine how successful a requirement is met. Wherever possible, the metrics should be objectively quantifiable (e.g. € per year, frames per second). In some cases, (semi-) subjective scales are necessary, for example degrees of openness and stability, support of a standard, degree of file format acceptance within different communities.

Define alternatives

Different preservation strategies, using for example migration tools or emulators, are selected. A detailed description of each preservation alternative is provided. The description includes the software environment and parameter settings of the tool, in order to ensure a clear understanding of the alternative and allow a later re-evaluation of the planning process. For each defined

alternative, the amount of work, time, and money required for running experiments is estimated.

Go/No-Go

Some experiments need a considerable amount of effort and required resources to run the experiments, for example experiment with great number of alternatives or high cost of hardware and software to run the experiments. Feasibility of the proposed alternatives is determined in this step by considering the defined requirements, the selected preservation alternatives, and estimated resources. The result is a decision for continuing the evaluation process or a justification of the abandonment or postponement of certain alternatives.

Develop experiment

In the experiments, the preservation alternatives are applied to the previous defined sample records. The results of the experiments are later evaluated against the goals and requirements of the objective tree. In order to run repeatable tests, it is important to document all relevant experiment settings. This stage produces a specific development plan for each experiment, which includes the workflow, the software and hardware systems used for the experiments, and the mechanisms to capture the results. All items needed for the experiment will be developed and/or installed and tested, including copies of the sample objects, software packages and programs, and mechanisms for capturing the results.

Run experiment

Experiments are designed to test one or more aspects of a specific preservation alternative when applied to the previously defined sample records. Running an experiment produces results, for example converted computer files, revised metadata, and measured workload of the hardware. The results are evaluated in the next step.

Evaluate experiments

The results of the experiments are evaluated to determine the degree to which the requirements defined in the objective tree were met. Therefore, the leaf objectives defined in the objective tree are evaluated with the defined measurement unit. For each alternative, the outcomes of this stage are measured performance values for each leaf in the objective tree.

Transform measured values

The measurements taken in the experiments might all be measured on different scales. In order to make these comparable, they are transformed to a uniform scale using transformation functions. These transformation functions can define thresholds or injective mathematic functions to map the measured values to the uniform scale. The resulting scale ranges from zero to five. A value of zero denotes an unacceptable result and thus serves as a dropout criterion for the whole preservation alternative.

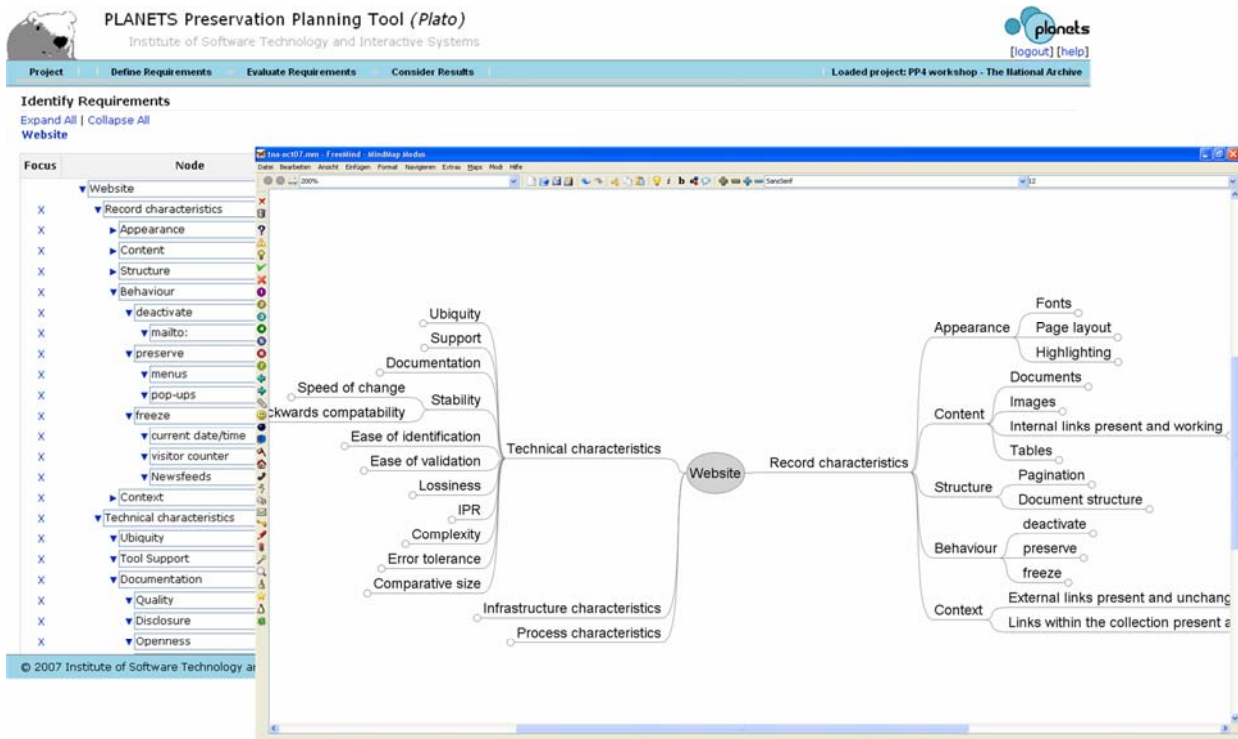


Figure 4 Requirements Definition in FreeMind and subsequent import into Plato

Set importance factors

Not all of the identified objectives are equally important and different degrees of conformance of a solution are accepted in different objectives. This step assigns importance factors to each objective depending on the specific preferences and requirements in the scenario.

Analyse results

In this step, the performance measures for the individual objectives are aggregated to one single comparable value for each alternative. The following two methods are mostly used:

Sum – The measured performance values, as transformed by the transformation functions, are multiplied by the weighting factor. These values are summed up to a single comparable value per alternative. Leaf values that score zero (measured performance under required minimum threshold) have no decisive effect on the final root value.

Multiplication – The first step here is to multiply the comparable value per leaf by the weight of that leaf. The results are then multiplied throughout the tree for the whole alternative. The multiplication method highlights alternatives with drop out values, as these alternatives with leaf values zero have a final root value of zero.

We thus obtain aggregated performance values for every part of the objective tree for each alternative, including an overall performance value at the root level. A first ranking of the alternatives can be done based on the final root values associated with each alternative. This ranking is based on the specific requirements of

the preservation context. It forms the basis for a documented and accountable selection of a specific preservation alternative. Furthermore, an analysis of all parts of the objective tree can identify the strengths and weakness of an alternative.

5 Preservation Planning Tool (Plato)

The preservation planning tool Plato is a web-based software tool that implements the workflow depicted in Figure 3 with the aim to provide support to the user in the complex decision making process of finding the optimal preservation strategy. Plato enables the evaluation of potential preservation strategies against defined requirements by conducting experiments defined by the planner on representative sample objects. Plato includes additional external services to automate this process. Furthermore it extends the workflow with a fourth phase in which an executable preservation plan is created, based on the well-documented recommendation.

The software itself is a J2EE web application that relies on open frameworks such as Java Server Faces and AJAX for the presentation layer and Enterprise Java Beans for the backend. It is integrated in an interoperability framework that guarantees loose coupling of services and registries through standard interfaces and provides common services such as user management, security, and a common workspace.

Based on this technical foundation, the aim is to create an interactive and highly supportive software environment that advances the insight of preservation planners and enables proactive preservation planning. In principle, there are three aspects to consider:

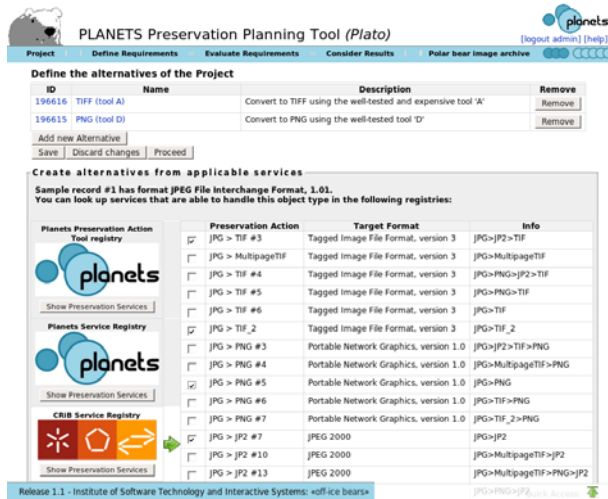


Figure 5 Discovery of Preservation Actions in Plato

(1) Integrating registries for information discovery; (2) Integrating services for preservation action and characterisation of objects; and (3) Proactively supporting the planning with a knowledge base that holds reusable patterns and templates for requirements recurring in different planning situations. The right choice of samples that are representative for the collection under consideration is essential, as any skewed representation might lead to wrong results. Collection profiling services based on characterisation services and format registries inform the selection process and ensure the right stratification of samples. Risk assessment services further assist by quantifying both the inherent risks of object formats and the salient risks present in the objects which are of particular relevance to a specific file format, such as the number of pages for some document formats or the presence of transparency layers in images. The specification of requirements in a tree structure is often done in a workshop setting. This is supported by both a flexible web interface as depicted in Figure 4 and a direct tree import from the openly available mind-mapping software FreeMind. The knowledge base provides recurring fragments and templates, such as process requirements for an archival institution or essential object characteristics for electronic documents in a library, to assist in the process of tree creation. Service discovery is the prime issue during the next step of defining alternatives to consider for evaluation. Starting from the sample objects and their formats, the system queries available registries of preservation actions and looks up applicable tools such as emulators of the original environment or migration tools that can handle the provided input format. Figure 5 shows potential atomic and chained migration services for migrating JPEG files to several different file formats. The Planets registry moreover holds information on benchmark evaluation results produced by experiments carried out in the Planets Testbed, which provides a controlled environment for preservation experiments [3]. Preservation action tools that are accessible through a web service are directly invoked during the execution of

Expand All | Collapse All
ONB Master thesis > Object characteristic

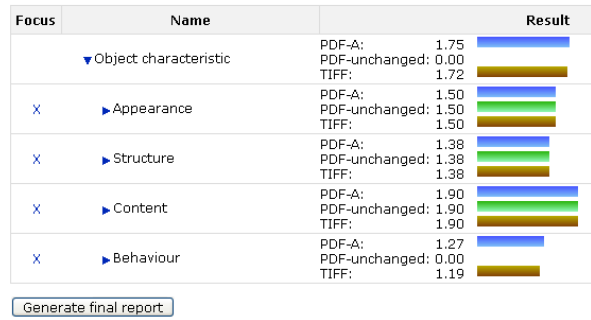


Figure 6 Visualisation of Evaluation Results

experiments on the sample objects; other tools such as emulators have to be executed externally. The evaluation of experiments is probably the most complex and, so far, least automated step in preservation planning. Until today, most of the judgement, e.g. if a migration tool accurately preserves the colour model of an image or the line breaks in a document, has to be carried out manually by looking at the rendered objects. However, characterisation services are available that can measure some of the essential characteristics of objects such as the dimensions of images. In contrast to characterisation tools like JHove, the extensible characterisation languages (XCL) [1] do not attempt to extract a set of characteristics from a file, but instead are able to express the complete informational content of a file in a format independent model. Comparison services specify measurable properties as well as property-specific metrics and their implementation as algorithms in order to identify degrees of equality between two objects. This is in principle independent of the applied strategy, i.e. migration or emulation. The compared objects can be both the original and a migrated object, or the original object in two different environments. To allow comparison and evaluation, a mapping is created between the requirements specified in the objective tree and the characteristics that can be measured and compared automatically by the available characterisation tools. This mapping partly stems from the knowledge base, but can be adapted by the user. Both XCL and other characterisation tools such as JHove are integrated in the evaluation of experiments. This also includes risk assessment services which compare the risk scores of objects resulting from the application of preservation actions against the scores of the original samples. The transformation of measured values to a uniform scale as needed for the aggregation of results and the importance weighting of requirements are supported by the knowledge base. Analysis of results is facilitated by a dynamic and flexible visualisation as depicted in Figure 6, where the planner can choose between different aggregation methods and dynamically configure the information content to analyse the strengths and weaknesses of the alternatives considered. Based on this analysis, a well-documented and solid recommendation for a solution can be made. This recommendation forms the basis for building a

preservation plan in the fourth stage. A preservation plan contains a description of the context and the decision taken, including the complete evidence base. This evidence base comprises a thorough description of the planning context and environment, ranging from the institution's mission statement via user group characteristics and policies to the collection at hand (documented in a collection profile). Moreover it contains the chosen sample objects, the requirements and additional documentation as well as considered solutions and the evaluation results. The plan furthermore contains cost indications and triggers for re-iterating the planning, and as a core part it entails a preservation action plan. If the applied strategy and its deployment support it, this is an executable workflow accessing distributed services. During the fourth stage, the planner may select a subset of the criteria used for evaluating solutions to be applied automatically with each preservation action as a mechanism for quality assurance. The corresponding characterisation actions which are used for property extraction and validation are then included in the executable preservation plan.

The first version of Plato is available² and freely accessible to the public. It implements the described workflow [14] and provides partial service integration such as file format identification. Subsequent versions will include a wider set of services for preservation actions and characterization. The final version will include the creation of an executable preservation plan.

6 Conclusions

Starting from the challenges of digital preservation and the OAIS reference model this paper presented the Planets Preservation Planning tool Plato. It supports the planner in evaluating available solutions against the specific requirements of a particular situation. In order to arrive at a profound and well-documented decision for a specific preservation strategy Plato implements a solid workflow that has been validated in a series of case studies. The procedure can be applied to any class of strategy, be it migration, emulation or any other approaches.

References

- [1] C. Becker, A. Rauber, V. Heydegger, J. Schnasse, and M. Thaller. A generic XML language for characterising objects to support digital preservation. In Proc. 23rd Annual ACM Symposium on Applied Computing (SAC'08) (Fortaleza, Brazil, March 16–20 2008), vol. 1, ACM, pp. 402–406.
- [2] CCSDS. Reference Model for an Open Archival Information System (OAIS) – blue book, 2002. <http://public.ccsds.org/publications/archive/650x0b1.pdf>. Accessed: July 1st, 2008.
- [3] A. Farquhar, and H. Hockx-Yu. Planets: Integrated services for digital preservation. *Int. Journal of Digital Curation* 2, 2 (2007), 88–99.
- [4] J. Hoeven, R. Van Der Diessen, and K. Van En Meer (2005). Development of a universal virtual computer

- (UVC) for long-term preservation of digital objects. *Journal of Information Science*, Vol. 31 (3), 196–208.
- [5] ISO. Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A) ISO/CD 19005-1, 2004.
- [6] ISO. Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003), 2003.
- [7] M. Jones and N. Beagrie. *Preservation Management of Digital Materials: A Handbook*. Digital Preservation Coalition, 2002. <http://www.dpconline.org/graphics/handbook>, accessed: July 1st, 2008.
- [8] G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H. Walters, and A. R. Kenney. Risk management of digital information: A file format investigation. CLIR Report 93, CLIR, June 2000.
- [9] P. Mellor. Camileon: Emulation and BBC domesday. *RLG DigiNews*, 7(2), April 2003.
- [10] S. Ross, and A. Gow. *Digital Archaeology: rescuing neglected and damaged data resources*, London: South Bank University, Library Information Technology Centre, www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf, Accessed: July 10th, 2008.
- [11] J. Rothenberg. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Council on Library and Information Resources, January 1999. http://www.clir.org/pubs/reports/rothenberg/content_s.html.
- [12] C. Rauch, and A. Rauber. (2004). Preserving digital media: Towards a preservation solution evaluation metric. In *Proceedings of the 7th international conference on asian digital libraries (ICADL 2004)* (p. 203–212). Berlin, Heidelberg: Springer.
- [13] S. Strodl, A. Rauber, C. Rauch, H. Hofman, FrancaDebole, and G. Amato. (2006). The DELOS testbed for choosing a digital preservation strategy. In *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL 2006)* (p. 323–332). Berlin, Heidelberg: Springer.
- [14] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. (2007). How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)* (p. 29–38). New York, NY, USA: ACM.
- [15] The 100 Year Archive Task Force. The 100 year archive requirements survey (2007). http://www.snia.org/forums/dmf/programs/ltacsi/100_year, accessed: July 4th, 2008.
- [16] UNESCO charter on the preservation of digital heritage. (Adopted at the 32nd session of the General Conference of UNESCO).

* Work presented in this paper is partially supported by European Community under the Information Society Technologies (IST) Programme of the 6th FP for RTD – Project IST-033789.

¹ <http://freemind.sourceforge.net>

² <http://olymp.ifs.tuwien.ac.at:8080/plato>