БИБЛИОСФЕРА, 2005, № 1, с. 63–71

Мастер-класс

УДК 025.4.03
ББК 78.30; 73

ЛИНГВИСТИЧЕСКИЕ СРЕДСТВА ИНФОРМАЦИОННОГО ПОИСКА В ИНТЕРНЕТЕ

© В.П. Захаров, 2005

Библиотека Российской академии наук
199034, г. Санкт-Петербург, Биржевая линия, 1

В лекции излагается современный взгляд на некоторые ключевые аспекты теории и практики информационного поиска. Большое внимание уделено информационно-поисковым языкам, их свойствам, поиску по свободному тексту. Даны подробные характеристики поисковым системам Интернета, а также лингвистическим механизмам формирования поисковых предписаний.

Ключевые слова: информационно-поисковые системы; информационно-поисковый язык; WWW поисковые системы; методы анализа текста; вербальные поисковые системы.

Можно сказать, что в основе всей человеческой цивилизации лежит совершенствование средств накопления и обработки знаний. Еще до вступления человечества в XXI в. начался век информационный. Рост числа научных публикаций, уже в начале XX в. осознаваемый как кризис[1], в 50–60-х гг. получил название «информационного взрыва». Однако настоящий взрыв был еще впереди – и связан он с появлением и развитием Интернета, который явился той «производительной силой», которая (по Марксу) за считанные годы поменяла так называемые «производственные отношения», в данном случае мы имеем в виду всемирную систему социальных коммуникаций. Мир вступил в эру безбумажной, электронной информации всех видов. Простота и доступность, можно сказать демократичность, средств создания и распространения электронных данных привели к тому, что каждый житель развитых стран получил возможность неограниченно заявить о себе на страницах всемирной электронной «прессы» – и многие не преминули этим воспользоваться, в том числе и те, кому нечего сказать. И как средство «борьбы» с этим лавинообразным потоком нужной и ненужной информации сегодня выступают информационно-поисковые системы (ИПС). Считается, что до 70% той информации, которую сегодня «потребляют» пользователи Интернета, они находят через поисковые системы.

Современные ИПС впечатляют своей простотой: пользователь вводит ключевые слова и получает в ответ документы на заданную тему. Однако простота эта кажущаяся: за всем этим стоят сложные специальные методы и алгоритмы поиска.

Характерная особенность информационных потоков в Интернете – это то, что подавляющая часть информационных массивов представляет собой полнотекстовые документы на естественном языке. А естественными языками, как известно, занимается наука языкознание, или, по-новому, лингвистика.

С самого начала появления информационно-поисковых систем важнейшим компонентом в них был язык, но… язык информационно-поисковый. Нужно сказать, что для ускорения и упрощения процедуры отбора из массива документов по их содержимому давно появились различные формы сокращенной записи содержания документов – библиографические описания, аннотации, рефераты. *Информационно-поисковый язык* (ИПЯ) также является механизмом представления основного содержания документов и запросов. Информационно-поисковые языки – это знаковые системы, со своим алфавитом, лексикой, грамматикой и правилами пользования. Хорошо известны такие «недостатки» естественно-языковых знаков, как омонимия, синонимия, многозначность. В ИПЯ эти недостатки

[1] «Издание массы книг и научных трудов становится бедствием, так как солидное, ценное и необходимое все чаще исчезает в огромном потоке ненужных изданий, и надвигается такая угроза, что все достойные внимания издания захлебнутся в этом потоке», – писал немецкий исследователь Адольф фон Харнак еще в 1911 г. (Цит. по: *Фабиан Б.* Книги, библиотеки и гуманитарные исследования / Отв. ред. В.П. Леонов. – СПб., 1996. – С. 258).

снимаются. Основными требованиями к информационно-поисковым языкам являются однозначность между планом выражения и планом содержания (каждая лексическая единица соотносится с одним понятием, и наоборот – каждое понятие имеет уникальное имя, и, как следствие, каждая запись на ИПЯ имеет только один смысл), достаточная семантическая сила (способность полно и точно фиксировать все существенное в содержании документов и запросов), открытость (возможность корректировки и пополнения языка). С другой стороны, все информационно-поисковые языки так или иначе создавались и создаются на базе естественных языков.

Постепенно наряду с понятием ИПЯ стал использоваться термин «*лингвистическое обеспечение*» (ЛО) информационно-поисковых систем, обозначающий весь комплекс языковых и логических средств и методов, используемых в ИПС для обеспечения основной задачи информационного поиска – сопоставления содержания документов и запросов. К ним относятся собственно ИПЯ, критерий смыслового соответствия, командные языки, методики индексирования, методики составления тезаурусов, вспомогательные средства создания и ведения ИПЯ и т.п.

История развития документальных ИПС последних десятилетий характеризуется как раз отказом от создания и использования «классических» ИПЯ в пользу развития языков и систем, получивших название «бестезаурусных», или систем поиска по свободному тексту (free-text searching systems). Особенностью их является, с одной стороны, отказ от лексического контроля и тем самым от учёта парадигматики, а с другой – широкое использование контекста и синтагматических связей.

Помимо внутренних причин этому способствовали внешние технологические факторы, заключающиеся в появлении информационной инфраструктуры по «производству» неструктурированной машиночитаемой информации. При этом процессы создания и использования информации, по-другому – процессы первичной семантической обработки документов и запросов – оказались разорванными. Кардинальные изменения в информационной сфере произошли в связи с развитием Интернета и резким ростом объемов документальной информации в электронном виде. Для современных поисковых систем, особенно в Интернете, характерна ориентация на поиск по полным текстам документов, представленных на естественных языках в виде гипертекста, структурированного средствами языков гипертекстовой разметки электронных документов (HTML, XML).

За короткий период существования Сети и сервиса WWW поисковые системы прошли большой путь развития. Многие проблемы, стоящие перед разработчиками ИПС в Интернете, не решены до сих пор /1/. Немалую проблему представляет изменчивость Сети. И если появление новых ресурсов можно считать естественным процессом (характеризующимся, правда, невиданными ранее скоростью и объемами), то частое изменение документов, как их содержания, так и местоположения в Сети, массовое их исчезновение представляют труднорешаемую проблему.

Число веб-страниц (документов) в Сети точно никому не известно и, по оценкам, достигает 5–6 млрд (конец 2003 г.). Это число удваивается каждые 8–18 мес. (ежедневно прибывает несколько миллионов веб-страниц!) /2/. В последнее время ряд ИПС начинает индексировать документы в форматах, отличных от «штатных» форматов Интернета (html, txt), после чего объемы баз данных поисковых систем (и их трудности!) возрастают еще больше.

Естественно, значительная часть документов оказывается неучтенной (не заиндексированной поисковыми системами). Американские исследователи С. Лоренс и К. Жиль полагают, что даже для лучших систем эта часть составляет от одной трети до половины /3/. Результаты работы Лоренса и Жиля подтверждаются и другими исследованиями (например, К. Бхарата и А. Бродера /4/ из Центра системных исследований фирмы Digital). Большое количество веб-страниц порождается в момент обработки запросов на основе информации, хранящейся на серверах в виде баз данных (динамические веб-страницы). И объем таких документов растет с каждым годом. Для них появились выражения «невидимый веб» (invisible Web) или «глубинный веб» (deep Web) /5/. Как бы то ни было, действительность такова, что поисковые службы заведомо страдают неполнотой.

Большая (как лингвистическая, так и программная) проблема – многоязычие информационного пространства Интернета. Примерно около 50% информации в Сети представлено на английском языке, вторая половина – на всех остальных языках, количество которых увеличивается по мере распространения сетевых технологий. Эта проблема касается и обработки документов, и составления и обработки запросов, и собственно работы модулей поиска и выдачи информации.

Эффективность функционирования ИПС во многом зависит от заложенного в них лингвистического обеспечения – методов, алгоритмов,

базирующихся на лингвистических моделях и машинных словарях. В современных ИПС используются словарные, морфологические, прагматические, статистические и другие методы анализа текста.

Лингвистические средства информационного поиска в Интернете частично «упрятаны» внутрь самих ИПС, частично реализуются на уровне пользователя. Нельзя сказать, что с появлением Интернета и бурным вхождением его в практику информационного обеспечения появилось нечто принципиально новое с точки зрения теории информационного поиска, чего бы не было раньше. Однако уровень сложности задач (поиск информации в больших объемах разнородных документов) и уровень требований, предъявляемых ко всем видам обеспечения, возрос, и сегодняшние ИПС, работающие в Сети, пытаются соответствовать им. Причем развитие ЛО ИПС идет по пути как усложнения, так и упрощения. Последнее объясняется, на наш взгляд, финансовыми причинами и невостребованностью отдельных элементов ЛО в условиях массового спроса на услуги систем.

Если попытаться классифицировать ИПС сети Интернет по типу ИПЯ, то большинство систем можно отнести к одному из двух типов: 1) ИПС дескрипторно-вербального типа, 2) классификационные ИПС.

Основным инструментом поиска в Интернете следует считать вербальные поисковые системы, индексирующие (по крайней мере, претендующие на это) всё интернет-пространство. Их называют еще словарными системами, хотя классический дескрипторный словарь, или тезаурус, практически во всех системах отсутствует. Это тип систем посткоординатного типа, который «противостоит» предкоординируемым ИПС – классификационным (тематическим каталогам, directories). В английской литературе за вербальными системами закрепился термин «search engines». По-русски обычно используют название «поисковые машины», хотя нам больше нравится «поисковые системы».

К числу главных поисковых систем этого типа (в первую очередь, по объему базы данных), по состоянию на конец 2003 г., можно отнести следующие: Google, Fast Search (AllTheWeb& Lycos), AltaVista, Direct Hit, HotBot, Inktomi, Teoma, WiseNut. Все они отличаются от других (в лучшую сторону) объемом базы данных, языком запросов, алгоритмами ранжирования и другими особенностями. Полнота поисковой базы и оперативность индексирования веб-сайтов являются главной проблемой всех ИПС в Интернете. Как правило, системы с бóльшим объемом базы дают в результате поиска и большее количество документов. Среди российских систем главными являются три системы: Яндекс (Yandex), Рамблер (Rambler), Апорт (Aport).

Поисковые системы вербального типа состоят из нескольких частей, главные из которых следующие.

Робот (robot) – система, обеспечивающая просмотр (сканирование) Интернета, создание и поддержание инвертированного файла (поисковой базы данных). Этот программный комплекс является основным источником информации о состоянии информационных ресурсов Сети.

Поисковая база данных – так называемый *индекс* (англ. index database) – специальным образом организованная база данных, включающая прежде всего инвертированный файл, который состоит из лексических единиц проиндексированных веб-документов и содержит другую разнообразную информацию о лексемах (в частности, их позиция в документах), документах и сайтах в целом.

Клиент – пользовательский программный интерфейс для общения пользователя с поисковым аппаратом: системой формирования запросов и просмотра результатов поиска. Реализуется в виде экранных форм, обрабатываемых веб-броузерами.

Поисковая система – подсистема поиска, обеспечивающая обработку поискового предписания пользователя, поиск в поисковой БД и выдачу результатов поиска пользователю.

Главная содержательная проблема, решаемая при индексировании, заключается в том, какие термины приписывать документам, откуда их брать. Роботы разных систем решают этот вопрос по-разному. Хотя индексируются полные тексты веб-документов, далеко не всегда все термины из них попадают в индекс. Применяются списки запрещенных слов (stop-words), многие системы индексируют лишь часть документа (обычно начальную), есть роботы, которые обрабатывают только часть веб-страниц с одного и того же сайта. И тем не менее объем поисковых индексов глобальных ИПС уже сегодня измеряется терабайтами.

Обычно при индексировании и ранжировании результатов поиска используются различные «значимые» элементы гипертекстовой разметки: ссылки, заголовки, заглавия, аннотации, списки ключевых слов и тому подобные данные тэга META (поле keywords и др.).

Значительная часть работы по семантической нормализации лексики в ИПС вербального типа (а точнее, по устранению последствий отказа от

лексического контроля на входе) возлагается на пользователя в процессе составления запроса и итеративного поиска. Поэтому большое значение имеют языки запросов ИПС, те возможности, которые в них закладываются.

Языки запросов представляют собой сложные объекты и объединяют собственно ИПЯ и критерий смыслового соответствия, а также могут содержать в себе требования к интерфейсу выдачи. Обобщенная структурная модель языка запросов включает следующие элементы:

1. Собственно поисковые элементы (термины, выражающие информационную потребность, и т.п.).

2. Средства морфологической нормализации текстовых элементов запроса.

3. Поисковые (булевские) операторы.

4. Средства линейной грамматики (операторы расстояния, позиционные операторы).

5. Дополнительные условия поиска:

- поиск в определенных полях (частях) документа,
- ограничение области поиска по языку, региону, дате создания документа и т.п.

6. Средства управления критерием смыслового соответствия.

7. Требование на сортировку (ранжирование) выдаваемых результатов поиска.

8. Требования к форме представления результатов поиска:

- вид выдаваемых результатов,
- количество выдаваемых документов и т.п.

Задача морфологической нормализации лексических единиц (словоформ) в документах и в запросах может решаться разными путями:

1) отказ от морфологической нормализации;

2) автоматический морфологический анализ и последующая генерация стандартной (канонической) формы (лемматизация).

Вообще, приведение словоформ к каноническому виду необязательно. Его можно избежать, если рассматривать каждую словоформу как отдельную лексическую единицу (ЛЕ) – на уровне текста и входа в инвертированный файл. При этом формальное отождествление разных словоформ одной лексемы обычно обеспечивается на этапе составления запросов (поисковых предписаний) и поиска механизмом, получившим название «маскирования» (wild cards) или «усечения» (truncation). Этот метод, применяющийся во многих системах, заключается в следующем: в поисковом предписании указывается лишь часть слова, а механизм поиска находит в документах все слова, имеющие точно такую же часть. Как правило, это начальная часть слова – основа или квазиоснова, в этом случае говорят о правом усечении. Часть слова, которая в поисковом предписании опускается, отбрасывается, обычно обозначается специальным символом (чаще всего звездочкой (*). Например, если требуется найти документы со словами *sing, singer, singers, singing*, то в запросе задается *sing**. Но есть системы, которые поддерживают и левое усечение, и внутреннее (маскирование). Например, на запрос **хлоран* будут выданы все документы по химии, содержащие сложные слова со второй частью «*хлоран*» (*гексахлоран, метилхлоран* и т.п.), а на запрос *colo*r* будут выданы документы, содержащие то же слово в написании *colour*. Маскирование с точностью до количества символов, которые опускаются, может задаваться с помощью знаков вопроса (одного или нескольких).

Использование словоформ в качестве самостоятельных ЛЕ не только ведет к значительному увеличению объема машинных массивов (инверсных файлов), но и вызывает определенные неудобства и сложности для пользователей, связанные с необходимостью выделения основ на уровне запросов и проистекающими из этого возможными ошибками /6/. Поэтому целесообразно иметь в ИПС механизм морфологической нормализации. Алгоритмы автоматического морфологического анализа сегодня реализованы в основных ИПС русскоязычного Интернета (Яндекс, Рамблер, Апорт). Так, на запрос со словом *плохой* будут выданы документы, содержащие словоформы *плохая, плохого, плохих, плохими* и т.п., а также *хуже* и *худший.*

Большинство систем сегодня базируется на булевой (логической) модели поиска. Запрос в этих системах представляет собой булево выражение – набор логических переменных (поисковых терминов), объединенных логическими операторами с учетом правил поискового синтаксиса. Логические переменные получают значение «истина» или «ложь» в зависимости от вхождения или невхождения терминов запроса в соответствующий документ. Тогда и все булево выражение в результате его вычисления в процессе сравнения запроса с документом всегда получает значение «истина» или «ложь». Если «истина» – документ признается релевантным запросу, если «ложь» – нерелевантным.

Основные **булевские операторы**, используемые в ИПС: **AND, OR, NOT**.

На запрос с булевским выражением с оператором *AND* выдаются документы, содержащие оба (все) поисковые элементы, объединённые этим оператором. Оператором AND объединяются поисковые элементы, описывающие каж-

дый аспект данного запроса. Оператор AND сужает множество результатов поиска и уменьшает число релевантных документов по сравнению с поиском по каждому отдельному поисковому элементу. В теории множеств этому оператору соответствует операция пересечения.

На запрос с булевским выражением с оператором *OR* выдаются документы, содержащие хотя бы один из поисковых элементов, объединённых этим оператором. Оператор OR расширяет результаты поиска и увеличивает число релевантных документов по сравнению с поиском по каждому отдельному поисковому элементу. Оператором OR, как правило, объединяются поисковые элементы, находящиеся в отношении поисковой синонимии (т.е. все лексические синонимы и/или условные синонимы, описывающие один и тот же аспект запроса). Таким образом, для того чтобы признать какой-либо документ соответствующим данному аспекту запроса, достаточно обнаружить в нём хотя бы один из поисковых элементов, описывающих этот аспект. В теории множеств этому оператору соответствует операция объединения.

Оператор *NOT* – одноместный оператор, но часто понимается как AND NOT. Этот оператор удаляет из массива (как правило, это массив документов, релевантных левой части запроса) все документы, содержащие поисковый элемент, стоящий справа от оператора NOT. Как результат, выдаются все оставшиеся документы. В теории множеств этому оператору соответствует операция дополнения. Пользоваться оператором NOT следует только тогда, когда мы точно уверены, что любое употребление поискового элемента в документе свидетельствует о нерелевантности документа запросу.

В подавляющем большинстве случаев логическая формула запроса представляет собой конъюнктивную нормальную форму – конъюнкцию дизъюнкций (AND-выражение, объединяющее OR-группы). В свою очередь, каждая OR-группа может представлять собой сложное выражение. Например, поисковое предписание по теме «Исследование и анализ информационных потоков» может выглядеть следующим образом:

[Исследование OR Анализ OR Модель OR (Количественная and мера) OR Критерий OR (Ранговое and распределение) OR (Закон and Ципфа) OR (Закон and Бредфорда) OR Параметр OR Цитируемость OR (Частотное and распределение) OR (Распределение and Лотки) OR (Показатель and рассеяния) OR (Частота and терминов)]

AND

[(Информационный and поток) OR (Документальный and поток) OR (Периодическое and издание) OR (Продолжающееся and издание) OR (Рассеяние and информации) OR (Распределение and публикаций) OR (Поток and публикаций) OR (Массив and публикаций)]

Пример несколько упрощенный: на самом деле словосочетания, заключенные в круглые скобки, представляют собой не простые AND-выражения внутри OR-групп, а устойчивые словосочетания, задаваемые специальными грамматическими операторами (условно показаны как *and* строчными буквами).

Это операторы для словосочетаний, которые могут быть отнесены к грамматическим средствам ИПЯ. Фактически это оператор AND с контекстными ограничениями (условиями) на расстояние между терминами и/или на порядок их следования (последнее из всех основных систем постепенно исчезло).

В ИПС различают устойчивые (жесткие) словосочетания («phrase» – все слова стоят рядом) и разрывные (нежесткие). В некоторых системах имеются средства «вычисления» словосочетаний в документах с учетом расстояния и порядка слов между элементами словосочетаний (Апорт, Яндекс, AltaVista). Длина такой нежесткой синтагмы может быть как постоянной (Alta Vista, Рамблер), так и переменной (Апорт, Яндекс, задается в запросе). Например, выражение запроса *climate ONEAR/5 change* (в старой версии Lycos) означает требование на выдачу документов, содержащих оба термина (*climate* и *change*) в данном порядке и находящихся в тексте на расстоянии не более чем *5* слов.

Главный контекстный оператор – это phrase. Он есть практически во всех системах. Это оператор для устойчивых словосочетаний, когда два (или более) слова запроса в документе должны стоять рядом (с точностью до отброшенных стоп-слов). Устойчивые словосочетания чаще всего задаются в кавычках. Для систем в сети Интернет с документами большого объема, где в одном документе могут быть представлены различные темы, использование этих операторов очень важно. Словосочетания должны использоваться обязательно, когда слово-определитель (слова, если их несколько) не просто сужает объем основного поискового термина, но образует в сочетании с ним новое понятие (соответствующее отдельному денотату). Например: *«rain in Spain», «Gettysburg Address», «big bad wolf», «редкие животные», «Красная книга», «Белая книга», «желтая пресса», «железная дорога»* и т.п.

В ряде систем имеются дополнительные условия поиска, или ограничения области поиска. Эти средства также называются «поиск по полям». В их числе можно назвать ограничение по месту (поиск только по тем документам, которые

находятся на серверах с заданным доменным именем), ограничение по дате создания или регистрации электронного документа, поиск по элементам гиперссылок (по тексту (якорю) ссылки или по адресу), поиск по заглавию, по специальным объектам (апплеты, графические файлы), по полю комментария к графическим файлам и т.д.

К лингвистическому обеспечению также можно отнести средства ранжирования результатов поиска по релевантности. Различные поисковые системы используют различные алгоритмы ранжирования, однако основные принципы определения релевантности следующие:

- количество слов из запроса в текстовом содержимом документа;
- элементы (теги), в которых эти слова располагаются (повышенный вес имеют теги заголовков, поля META, гиперссылок и т.п.);
- местоположение искомых слов в документе (чем ближе к началу, тем выше значимость термина);
- удельный вес слов (относительная частота), относительно которых определяется релевантность, в общем количестве слов документа.

Эти принципы применялись и применяются практически всеми поисковыми системами. Кроме того, в последнее время активно учитываются и внелингвистические критерии:

- «время жизни» – как долго веб-страница находится в базе поискового сервера;
- индекс цитируемости – как много ссылок на данный документ идет с других веб-страниц, зарегистрированных в базе ИПС;
- индекс популярности – как часто пользователи обращались к данному документу.

В завершение данного раздела перечислим характерные элементы языков запросов разных поисковых систем, естественно, далеко не все, и их распределение по основным системам.

Логические операторы по умолчанию (на месте пробела)

AND: AllTheWeb, HotBot, Google, MSN Search, Lycos, WiseNut, Teoma, Рамблер, Апорт, Яндекс

OR: AltaVista , Excite

Логические операторы и выражения в запросе:

and, or, скобочные выражения: AltaVista Advanced, HotBot, Excite, MSN Search, Рамблер (также &), Апорт (также +, &, И, и), Яндекс (в специфической форме: & (and в пределах предложения), && (and) и | (or))

not: HotBot, Excite, MSN Search, Google (знак «-»), Рамблер (также «!»), Апорт (также НЕ), Яндекс (в специфической форме: ~)

and not: AltaVista Advanced, Excite

AND, OR (только прописными): AltaVista Simple, Excite

Только OR: Google

OR в виде скобочной записи *(термин 1 термин 2)*: AllTheWeb

Контекстные операторы (близость, расстояние)

Phrase: AltaVista, Google, HotBot, Excite, MSN Search, Lycos, AllTheWeb, WiseNut, Teoma, Рамблер, Апорт, Яндекс

NEAR: AltaVista, Рамблер (по умолчанию и в расширенной форме)

Расстояние в словах: Апорт, Яндекс

Расстояние в предложениях: Яндекс

Морфологическая нормализация

Усечение: AltaVista, HotBot, MSN Search, NBCi, iWon, Апорт, Рамблер

Автоматическое усечение: Yahoo!

Автоматическая нормализация: Апорт, Яндекс, Рамблер

Автоматическая нормализация множественного числа: Northern Light[2]

Автоматическое усечение до основы слова: HotBot, MSN Search

Чувствительность к регистру[3]:

Всегда: AltaVista (Advanced and Power, AltaVista Simple (если термины в кавычках)

Частично (с точностью до прописных): HotBot, MSN Search

Поиск по полям[4]

title: AltaVista, Northern Light, AllTheWeb, HotBot, Lycos, MSN Search, Апорт, Яндекс, Рамблер

intitle: Google

allintitle: Google

url: AltaVista, Northern Light, Fast Advanced Search, Lycos Advanced, Апорт, Яндекс, Рамблер

inurl: Google

allinurl: Google

link: AltaVista, Google, Fast Advanced Search, Lycos Advanced, MSN Search, Апорт, Яндекс

host: AltaVista

domain: HotBot, MSN Search

site: Google

anchor: AltaVista, Fast Advanced Search, Апорт, Яндекс

image: AltaVista

related: Google

others: AltaVista, Northern Light, HotBot, MSN Search

Ограничение области поиска

По дате: AltaVista Advanced, Northern Light, HotBot, MSN Search, Апорт, Яндекс, Рамблер

По языку: AltaVista, Northern Light, AllTheWeb, Excite, Google, HotBot, MSN Search, Lycos, WiseNut, Яндекс, Рамблер

---

[2] В настоящее время система закрыта для свободного использования.

[3] В последнее время почти из всех систем эта возможность исчезла.

[4] По ограничениям объема приводится только часть полей без пояснений. Подробности см. в справочных подсистемах соответствующих ИПС.

По домену: AllTheWeb Advanced Search, HotBot, Excite, MSN Search, Lycos, Яндекс, Рамблер

По типу данных внутри документа: HotBot, MSN Search

По глубине внутри сайта: HotBot

Индексирование с использованием стоп-слов

Не используются (в инвертированный файл включаются все слова): AltaVista Advanced, AllTheWeb, Lycos, Яндекс.

Используются: AltaVista Simple, HotBot, Excite, MSN Search, Lycos, Апорт.

Используются с сохранением возможности поиска по стоп-словам: Google, Teoma, WiseNut

Ранжирование

По релевантности: все

По дате: Яндекс, Рамблер

По сайту: Excite, Google, Рамблер.

Здесь приведен обзор лишь основных элементов языков запросов и лишь для некоторых систем. Дополнительно во многих системах существуют различные другие возможности, например режим установки так называемого семейного фильтра, при котором из результатов поиска исключаются документы неприличного содержания. И многое другое.

Мы рассмотрели лингвистическое обеспечение поисковых систем. В то же время сами системы могут выступать и как инструменты лингвистического анализа. Индексы поисковых систем (инвертированные файлы) – это, по сути, не что иное, как конкордансы к текстам. Результаты же поиска в ИПС в виде кратких описаний документов часто содержат контексты, в которых в найденных документах искомые слова встретились. Отличие лишь в том, что конкордансы обычно составляются к конкретному произведению или группе произведений (например, все тексты одного и того же автора), в то время как ИПС Интернета индексируют все доступное множество электронных документов.

Главный материал лингвистического анализа – язык, зафиксированный в виде речевых произведений, – в Интернете представлен в огромном объеме и разнообразии и непосредственно доступен для машинной обработки. Этот факт представляет для лингвистов большую ценность, так как ранее на перевод текстов в машинную форму приходилось тратить много времени, сил и денег.

Во всех лингвистических исследованиях существенное значение имеет проблема выборки и репрезентативности анализируемого материала. В настоящее время в научный оборот лингвистов все шире входит понятие «корпуса текстов» (КТ), на базе которых развивается корпусная лингвистика /7/.

В ряде случаев данные Интернета и баз данных поисковых систем можно рассматривать как текстовые корпусы.

Но при этом полезно представлять, как эти базы строятся и, соответственно, учитывать эти особенности в исследованиях. Главная содержательная проблема при индексировании вебсайтов заключается в том, какие термины приписываются документам, откуда они берутся. Не все термины из документов и не всегда попадают в индексы. Активно применяются списки запрещённых слов (stop-words), которые в индекс не попадают – это общая, служебная лексика (предлоги, союзы и т.п.) и незначащие слова. Многие системы индексируют лишь часть документа (обычно начальную), есть роботы, которые обрабатывают только часть веб-страниц с одного и того же сайта. Знание того, как работают роботы, каковы их технические характеристики, полезно и для создателей веб-документов, и для составителей запросов при проведении поисков. Подробное описание работы роботов можно найти в Сети /8/. Сведения о большом количестве роботов (более 200) можно почерпнуть из базы данных The Web Robots Database /9/.

Особенности построения и структура индекса напрямую связаны с языком запросов и возможностями поисковых систем. Наиболее важными с точки зрения лингвистического анализа текстового материала представляются следующие *особенности ИПС*:

- «грамотная» работа со словоформами – способность ИПС отождествлять разные словоформы одной и той же лексемы, по-другому, порождать каноническую форму – лемму, и возможность выделять среди множества словоформ конкретную форму;
- поиск слов с заданным или произвольным усечением, как правым, так и левым;
- индексирование полных текстов в полном объеме без исключения. Многие системы, как уже говорилось, не включают в индекс служебную и незначимую лексику;
- работа со словосочетаниями – учет расстояния между элементами словосочетаний и порядка их следования;
- различение больших и малых букв.

Также важно, какую информацию и в каком виде можно извлечь из выходных интерфейсов ИПС. Интерфейс выдачи (форма представления результатов) у разных систем включает такие параметры, как-то: статистика слов из запроса, количество найденных документов, количество найденных сайтов, количество документов на странице с результатами поиска, средства управ-

ления сортировкой документов в выдаче, описание сайта, с которого взят соответствующий документ, описание документа. Последнее, в свою очередь, содержит в своем составе заглавие документа, URL (адрес в Сети), размер документа (объем), дату создания, кодировку, аннотацию (краткое содержание), выделение в аннотации слов из запроса, указание на другие релевантные веб-страницы того же сайта, ссылку на рубрику каталога, к которой относится найденный документ или сайт, коэффициент релевантности, другие возможности поиска (поиск похожих документов, поиск в найденном). Из всех этих реквизитов для задач лингвистического исследования наибольший интерес представляют частотные характеристики. Следует различать два типа частот: пословную и подокументную. Сведения о количестве языковых единиц в разных системах и разных режимах поиска могут относиться как к словоформам, так и к лексемам. Некоторые системы ведут журнал запросов с возможностью повторных поисков и выдачей статистики по запросам. Полезной и интересной возможностью является также отнесение документов к тематическим классам.

На нескольких примерах покажем возможности поисковых систем для получения лингвостатистических данных о частоте использования тех или иных слов или словосочетаний. В принципе, нас, как правило, интересуют относительные частоты, а для этого достаточно проведения сравнительных поисков в рамках одной ИПС. Однако для того чтобы убедиться в достоверности данных и показать особенности разных систем, мы выбрали для эксперимента пять ИПС, наиболее популярных и обладающих наиболее развитым лингвистическим обеспечением. В первую очередь, это российские ИПС Яндекс, Рамблер и Апорт. Возможно, наиболее мощный лингвистический аппарат имеет ИПС Артефакт (фирма «Интегрум-ТЕХНО», г. Москва), однако эта система является коммерческой, и ее база данных по составу заметно отличается от других. Из западных систем, в большинстве своем не обладающих развитыми лингвистическими средствами анализа текстового материала, мы взяли хорошо известные ИПС Google и AltaVista. Кратко охарактеризуем особенности этих систем применительно к нашему эксперименту (наличие или отсутствие соответствующих возможностей помечено знаками «+» и «–») (таблица).

«Поиск по лексемам» означает, что результат сравнения слов документов и запросов признается положительным при наличии в документе любой формы слова из запроса, что обеспечивается механизмом автоматической лемматизации.

«Поиск по словоформам» означает, что результат сравнения документов и запросов признается положительным при наличии в документе словоформы, точно совпадающей со словом из запроса, что происходит при отсутствии автоматической лемматизации или обеспечивается особым механизмом учета словоформ.

«Частота подокументная» означает, что в результате поиска выдается сообщение о количестве релевантных документов, т.е. документов, содержащих данное слово (словоформу) или словосочетание.

«Частота пословная» означает, что в результате поиска дополнительно выдаются сведения об общем количестве словоупотреблений данной лексемы (независимо от формы) или конкретной словоформы в поисковой базе данных (индексе).

Характеристика поисковых систем

Характеристика	Яндекс	Рамблер	Апорт	Google	AltaVista
Поиск по лексемам	+ (однословный запрос или логическая формула)	+	+	–	–
Поиск по словоформам	+ (в синтагмах: однословный запрос в кавычках или словосочетание в кавычках)	–	+	+	+
Учет синтагм (неразрывных словосочетаний)	+	+	+	+	+
Учет больших и малых букв	+ (в синтагмах)	–	–	–	–*
Частота пословная	+	–	–	–	–
Частота подокументная	+	+	+	+	+

* Ранее большие и малые буквы различались; в ныне работающей версии эта возможность отсутствует.

Итак, покажем возможности поисковых систем для получения лингвостатистических данных. Начнем наши лингвистические изыскания с вопроса, как следует называть программу просмотра веб-страниц (англ. browser): «броузер» или «браузер». В нормативных словарях русского языка это слово отсутствует. Поиск в Яндексе дает следующие результаты. Статистика слов: броузер – 472 847, браузер – 997 666; запросов за месяц: броузер – 2 150, браузер – 5 335. Из этих данных можно сделать вывод, что написание «браузер» в настоящее время утверждается как языковая норма.

Следующий пример. Анализ поисковой базы Яндекса показывает, что наряду с написанием «офсайд» (27 168 словоупотреблений) в русском языке также достаточно широко используется написание «оффсайд» (9 867 словоупотреблений). Синонимичное словосочетание «вне игры», по данным Яндекса, используется примерно в два раза чаще (34 217 документов), чем «офсайд» (19 106 документов), что можно объяснить, по-видимому, его частым использованием в переносном смысле.

Подобные изыскания каждый лингвист может провести, не тратя времени на сбор текстового материала.

***

И все же лингвистический компонент современных ИПС отстает от технического и программного. Стоит задача создания лингвистического обеспечения на новом уровне. Требуются новые теоретические и практические разработки в области информационных систем, в первую очередь, в направлении их «интеллектуализации» /10/. В то же время можно сказать, что ряд теоретических положений и исследований, похороненных, казалось бы, вместе с системами последней четверти XX в., находят применение уже сейчас. В качестве примера можно назвать многочисленные методы и модели, предложенные в свое время Дж. Солтоном /11/.

Список литературы

1. См.: *Halasz F.C.* Seven issues for the next generation of hypermedia systems // Communication of the ACM. – 1988. – Vol. 31, N 7. – P. 836–852.
2. Internet World. – 1998. – Sept. 28.
3. *Lawrence S., Giles C.L.* Searching the World Wide Web // Science magazine. – 1998. – April 3. – P. 98–100. См.: http://www.neci.nj.nec.com/homepages/lawrence/websize.html; *Lawrence S., Giles C.L.* Accessibility and distribution of information on the web // Nature. – 1999. – Vol. 400, N 6740, July 8. – P. 107–109.
4. http://www.research.digital.com/SRC/whatsnew/sem.html
5. *Bergman M.* The deep Web: surfacing hidden value // BrightPlanet LCC. – 2000, July. (http://www.completeplanet. com/tutorials/deepweb/)
6. *Белоногов Г.Г., Кузнецов Б.А., Новоселов А.П.* Автоматизированная обработка научно-технической информации. Лингвистические аспекты // Итоги науки и техники. Информатика. – 1984. – Т. 8. – С. 87–89.
7. *Leech G.* The State of Art in Corpus Linguistics // English Corpus Linguistics / K. Aimer, K. Altenberg (eds.) – London, 1991. – P. 8–29.
8. См., в частности, http://info.webcrawler.com/mak/projects
9. http://www.robotstxt.org/wc/active/html/. Имеется русский перевод описания на сайте WebClub.Ru (http://www. webclub.ru)
10. См.: *Лахути Д.Г.* Проблемы интеллектуализации информационно-поисковых систем: Дис. в виде науч. докл. … д-ра техн. наук. – М., 1999. – 56 с.; *Рубашкин В.Ш.* Представление и анализ смысла в интеллектуальных информационных системах. – М., 1989; *Финн В.К.* Информационные системы и проблемы их интеллектуализации // НТИ. Сер. 1. – 1984. – № 1. – С. 1–14; *Chen H., Lynch K.J.* Automatic construction of networks of concepts characterizing document databases // IEEE transactions on systems, man and cybernetics. – 1992. – Vol. 22, N 5. – P. 885–902; *Chen H., Tobun D.Ng., Martinez J., Schatz B.R.* A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system // JASIS. – 1997. – Vol. 48, N 1. – P. 17–31; *Ingwersen P.* Information retrieval interaction. – London, 1992; *Ingwersen P.* Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory // J. of documentation. – 1996. – Vol. 52, N 1. – P. 3–50.
11. *Солтон Дж.* Динамические библиотечно-информационные системы. – М.: Мир, 1979; *Сэлтон Г.* Автоматическая обработка, хранение и поиск информации. – М.: Сов. радио, 1973; *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing & Management. – 1988. – Vol. 24, N 5. – P. 513–523.

Материал поступил в редакцию 18.11.2004 г.

Сведения об авторе: *Захаров Виктор Павлович – кандидат филологических наук, тел. 921-630-24-21 mobil, e-mail: vz@laz.usr.pu.ru*