ПРОФЕССИОНАЛЬНАЯ РАБОТА

Д.В. ЛАНДЭ

ПОИСК ЗНАНИЙ В INTERNET

- New Media — новая информационная среда
- Проблемы и феномены Internet
- Информационно-поисковые системы в Internet
- Топология Web-пространства
- Скрытый Web
- Интеграция Web-контента на основе XML
- Text Mining — глубинный анализ текстов
- Инструменты конкурентной разведки
- Информационные фракталы

Издательство Диалектика
www.dialektika.com

Книги серии "Профессиональная работа"

Книги этой серии предназначены для специалистов средней и высокой квалификации, желающих получить глубокие знания в области практического использования тех или иных программных средств и технологий. Изложение материала строится на предоставлении читателю большого объема специализированной технической информации и углубленном практическом подходе к вопросам реализации. Как правило, в книгах этой серии приводится специфическая информация, которой нет ни в каких других книгах, а также присутствует большое количество реальных примеров, благодаря которым пользователь даже с незначительным личным опытом сможет быстро достичь высокого профессионального уровня.

Состав серии

1. *Галисеев Г.В. Компоненты в Delphi 7. Профессиональная работа*
2. *Клюшин Д.А. Полный курс C++. Профессиональная работа*
3. *Минько А.А. Статистический анализ в MS Excel. Профессиональная работа*
4. *Сергеев А.П. HTML и XML. Профессиональная работа*

Книги серии "Решение практических задач"

Книги этой серии ориентированы на пользователей компьютеров из самых различных сфер — бизнеса, науки, производства, образования. Главное что их объединяет — это стремление профессионально освоить уже существующие прикладные компьютерные пакеты и технологии, что позволит им добиться максимальной эффективности в использовании компьютеров для решения стоящих перед ними задач. Изложение материала строится на предоставлении читателю большого объема необходимой теоретической и практической информации с акцентом на методах решения конкретных прикладных задач, которые могут его интересовать. Серия предназначена для читателей, желающих получить глубокие практические знания в области эффективного использования различных программных средств и технологий в своей повседневной профессиональной деятельности.

Состав серии

1. *Васильев А.Н. Научные вычисления в Microsoft Excel. Решение практических задач*
2. *Сингаевская Г.И. Функции в Excel. Решение практических задач*

ПРОФЕССИОНАЛЬНАЯ РАБОТА

Д.В. Ландэ

Поиск знаний в INTERNET

Диалектика

Москва • Санкт-Петербург • Киев
2005

ББК 32.973.26-018.2.75
Л22
УДК 681.3.07

Компьютерное издательство "Диалектика"

Зав. редакцией *А.В. Слепцов*

По общим вопросам обращайтесь в издательство "Диалектика" по адресу:
info@dialektika.com, http://www.dialektika.com

Ландэ, Д.В.

Л22 Поиск знаний в Internet. Профессиональная работа.: Пер. с англ. — М. : Издательский дом "Вильямс", 2005. — 272 с. : ил. — Парал. тит. англ.

ISBN 5-8459-0764-0 (рус.)

Книга посвящена современным подходам к получению новых знаний на основе анализа информационного пространства сети Internet и методам обработки информационных потоков с целью выявления значимых тенденций, понятий, феноменов, их взаимосвязей. Анализируются проблемы и феномены Internet, топология Web-пространства, методы доступа к информации в "скрытом" Web, рассматриваются особенности различных информационно-поисковых систем и средства интеграции Web-контента на основе XML. Большое внимание в книге уделено новому направлению обработки текстовой информации — "глубинному анализу текстов" (Text Mining), объединяющему в себе технологические и методологические подходы контент-анализа, компьютерной лингвистики и искусственного интеллекта.

Книга ориентирована на широкий круг читателей, интересующихся современными информационными технологиями. При этом она будет полезна и аналитикам, которые с помощью инструментов Text Mining смогут повысить эффективность и качество своей работы.

ББК 32.973.26-018.2.75

ISBN 5-8459-0764-0 (рус.)

Оглавление

Содержание

Предисловие

Эта книга для тех, кто интересуется методами получения новых знаний на основе анализа современного информационного пространства, а также способами обработки информационных потоков с целью выявления тенденций, новых понятий, феноменов, взаимосвязей.

Одно из определений знаний, которое дает энциклопедический словарь Webster, следующее: состояние осведомленности о чем-то или обладание информацией. Именно эта трактовка знаний наиболее близка к проблематике данной работы. Объем данных, из которого приходится выискивать крупицы необходимой, актуальной, готовой к немедленному использованию информации для решения проблем, обусловливает актуальность и значимость самого процесса поиска знаний.

Если знания — это сила, то сегодня первоочередная задача — найти эту силу. При этом поиск знаний, в отличие от простого поиска информации, при котором зачастую не учитывается семантика запросов, должен предоставлять пользователю только действительно актуальную информацию, наиболее точно соответствующую его потребностям, и вместе с тем адекватную исходному запросу. Если при обычном информационном поиске пользователь в конечном итоге знает, что он может получить, то при поиске знаний он должен получить нечто до сих пор ему неизвестное и познать его.

О сложности такого процесса говорит, например, недавнее исследование, проведенное фирмой Reuters среди 1300 менеджеров, которое показало, что "менеджеры чувствуют, что не могут эффективно работать без получения большого объема информации, но эта тяжелая загрузка данными, часто не имеющими никакого отношения к делу, снижает эффективность их работы и препятствует нормальному функционированию корпоративной машины". Это состояние было названо "синдромом информационной усталости", что свидетельствует об избытке информации и недостатке знаний. Из опрошенных фирмой Reuters менеджеров, 38% утверждают, что "тратят много времени, пытаясь найти нужную информацию". По оценкам экспертов, около 79% журналистов обращаются к Internet в поисках новостей и лишь 20 % находят ту информацию, которая им необходима. Все они на самом деле ищут именно знания.

В последнее время о поиске знаний пишут достаточно много. Появилось новое направление в обработке текстовой информации — "глубинный анализ текстов" (Text Mining). Это направление, скорее технологическое, чем научное, включило в себя все реальные, реализуемые на практике результаты исследований в области контент-анализа и компьютерной лингвистики, которая, как и теория баз знаний, интенсивно развивалась в 70–80-е годы прошлого века.

Сегодня прагматичные подходы, свойственные технологии Text Mining, могут применяться как студентами при написании обзорных курсовых работ, так и маркетологами при анализе рынков, политиками, бизнесменами, учеными — всеми, кто активно участвует в современных информационных, политических и бизнес-процессах.

Методы Text Mining уже используются в таких основных областях, как:

- политические исследования — геополитика, анализ предвыборной и выборной ситуации, деятельность партий, общественных организаций, отдельных политических деятелей и т.д.;
- конкурентная разведка — обобщенный анализ деятельности конкурентов, их PR-активности, клиентской базы;
- анализ рынков — выявление основных тенденций в производстве и потреблении товаров и услуг определенных видов, в политике фирм, участвующих в рынках, ареалах;
- анализ новых технологий — в различных сферах науки, бизнеса, безопасности;
- образование, культура.

Несмотря на то что книга ориентирована на широкий круг читателей, интересующихся современными информационными технологиями, хочется верить, что она будет также полезна и аналитикам, которые с помощью методологии Text Mining или отдельных ее компонентов смогут повысить эффективность и качество своей работы.

Введение

Количество информации, обрушивающейся на человека в современном мире, обусловливает актуальность задачи отделения действительно важных сведений от информационного шума. Человек, группа людей, информационная служба, профессиональные эксперты-аналитики уже не могут пропускать через себя потоки информации, которые изливаются на них сегодня электронными медиа. Зачастую даже опытные эксперты не могут выделить главного, не находят сведений, необходимых для принятия решений, в результате чего действия как отдельных людей, так и коллективов или даже государств становятся неадекватными реальной обстановке.

Таким образом, самая главная проблема современных коммуникаций — это извлечение действительно ценных сведений из информационных потоков; другими словами, получение знаний из информации.

Обилие информации уже давно воспринимается как нечто само собой разумеющееся. Количественные оценки ее суммарного объема как таковые вряд ли могут стать поводом для особых размышлений. Но если подобные показатели подвергнуть структурному анализу, то полученные результаты могут оказаться весьма неожиданными.

Возьмем, к примеру, исследование изменения объема информации в мире за год [54]. С 2000 года оно проводится в Калифорнийском университете в Беркли под руководством профессоров Питера Лаймана и Хола Вэриена. Ученые пришли к выводу, что на протяжении трех лет, предшествующих 2002 году, количество информации, произведенной человечеством, удвоилось. А в самом 2002 году в мире было произведено пять экзабайт (миллионов терабайт) информации. Для сравнения приведем данные об объеме фонда библиотеки Конгресса США, где хранится 19 млн книг и 56 млн рукописей: он составляет около десяти терабайт информации. В упомянутом исследовании информация структурировалась по типам носителей. Оказалось, что лидерство прочно удерживают магнитные носители, доля которых превышает 90%. Из них большую часть составляют жесткие диски. На кино, фото, печатные издания и другие бумажные документы вместе с оптическими цифровыми носителями приходится лишь 7% информации.

Очевидно, что лишь человеческого опыта в данной информационной ситуации становится уже недостаточно. Сама среда поступления информации определяет и возможные реальные подходы к ее обработке. Только мощные возможности информационной техники — компьютеров, сетей — в совокупности со специальным программным обеспечением могут оказаться той панацеей, которая спасет нас от информационного хаоса. В свое время казались очень перспективными системы искусственного интеллекта, экспертные системы со своими парадигмами фреймов и правил — баз знаний. То ли в 80-х годах двадцатого столетия не до конца сформировалась общественная потребность в широком использовании таких систем, то ли недостаточными были мощности компьютеров, то ли не доработаны были теоретические и алгоритмические основы таких систем, но бум их популярности в конце 80-х годов закончился. За прошедшее с тех пор время наряду с бурным технологическим процессом (до сих пор не опровергнут закон

Мура) сложилось понимание того, что для решения проблемы информационного хаоса больше всего подходят технологии, порожденные некогда таким направлением, как контент-анализ, и сегодня получившие названия Data Mining и Text Mining. В настоящее время существуют достаточно развитые системы, реализующие эти направления. Практически все самые известные производители программного обеспечения предлагают на рынке системы глубинного анализа данных и текстов (у компании Oracle — это Oracle Text, у IBM — Intelligent Miner for Text, у SAS — Text Miner).

Следует отметить, что большая часть информационного потока — это неструктурированная текстовая информация, в то время как значительная часть электронной информации, порожденной путем использования современных СУБД, — это численные фактографические данные. Если обработка таких данных позволяет использовать уже отработанные методы и погружать потоки данных в СУБД, то задача анализа текстовой информации открывает широкое поле для применения новейших методик и технологий, таких как XML, лингвистические, эмпирические, статистические подходы. В настоящее время уже определено несколько задач, стоящих перед технологией Text Mining, — это автоматическая классификация, кластеризация, выявление смысловых взаимосвязей отдельных фрагментов и понятий, выраженных в тексте, а также составление осмысленных рефератов, резюмирующих знания, содержащиеся в текстовых массивах больших объемов. Возможно, эти технологические подходы в случае массового применения смогут облегчить ориентацию человека в постоянно расширяемом информационном поле, позволят ему адекватнее реагировать на происходящие события, уверенно принимать важные решения на основе концентрации знаний.

Развитие вычислительной техники и компьютерных сетей способствовало появлению систем, назначение которых — поиск в массивах полнотекстовых документов. К таким документам можно отнести, например, статьи, нормативные акты, реферативные описания, тексты брошюр, диссертаций, монографий. До определенного времени полнотекстовые информационно-поисковые системы использовались преимущественно специалистами, круг которых был не очень широк, — архивные работники, сотрудники библиотек, ученые, аналитики.

Появление и развитие сети Internet в корне изменило ситуацию. Сегодня информационные ресурсы Сети составляют около десяти миллиардов документов (Web-страниц), к которым возможен свободный доступ любого пользователя. Естественно, чтобы найти необходимую информацию в этой крупнейшей полнотекстовой базе данных, необходимо использовать очень мощные поисковые средства, которые в зачаточном состоянии уже существуют, развиваются и конкурируют друг с другом на рынке информационных технологий.

Сегодня миллионам пользователей Internet известны такие системы, как Google, Yahoo, AllTheWeb, AltaVista, каждая из которых охватывает несколько миллиардов Web-документов. Мы стали свидетелями "информационного взрыва", в результате которого менее чем за 10 лет мало кому известная технология полнотекстового поиска стала повседневным инструментом миллионов людей.

В связи с этим первая глава книги — "New Media" — посвящена Internet и ее информационному подпространству World Wide Web. В этой главе описывается топология этого подпространства, а также средства навигации в нем и эволюция этих средств — от простейших наборов ссылок и каталогов до многофункциональных порталов.

Вторая глава посвящена поисковым системам, процессу поиска информации и его отдельным звеньям, а также включает трактовки таких фундаментальных понятий информационного поиска, как полнота и релевантность. Кроме того, эта глава содержит информацию о практической стороне использования процедур поиска, особенностях формирования запросов к различным информационно-поисковым системам с использованием слов, словоформ, фрагментов текстов, а также о поиске с учетом структуры документов, морфологии, подобия.

Третья глава охватывает вопросы ориентации в новостной информации, представленной в Сети. Для такого поиска используется специальный класс информационно-поисковых систем — системы мониторинга контента Internet, на основе которых строятся современные службы синдикации новостей.

Вопросам современного унифицированного представления информации в перспективном формате гипертекстовой разметки XML, а также технологическим решениям, построенным на основе идеологии XML, посвящена четвертая глава "XML — язык разметки и модель данных".

Технологиям выявления знаний в текстовых массивах с использованием как классических, так и новых, интеллектуальных подходов к анализу информации посвящена пятая глава "Технология Text Mining".

Шестая глава посвящена очень популярному сегодня направлению использования технологии Text Mining — конкурентной разведке, которая заключается в сборе и аналитической обработке информации, необходимой для принятия оптимальных управленческих решений. Очень важно, что при этом конкурентная разведка выполняется строго в рамках правовых норм.

Седьмая, заключительная, глава книги содержит обзор общих закономерностей, присущих информационным системам, в частности таких, как правило Парето, законы Зипфа и Брэдфорда и так далее, что должно дать читателю некоторое обобщенное представление о тенденциях и подходах, обсуждаемых в книге.

Дмитрий Ландэ, сентябрь 2004 года

Глава 1

New Media

Internet, появившись вначале как феномен новых технологий, породила мощный инструмент специфического воздействия на сознание человека, получивший название "New Media" [10].

1.1. Общая информация об Internet

Internet более чем за 30 лет своего существования вышла за пределы военных лабораторий США (где она родилась в рамках проекта ARPANET) и научных кругов [11] и к настоящему времени стала одним из самых известных явлений современности.

Благодаря чему же произошло это, почему из сотен компьютерных сетей именно Internet получила такое развитие? Ответов несколько.

1. Высокая технологичность, надежность и расчет на работу сети в любых, даже экстремальных условиях.
2. Открытость протоколов (правил), их доступность каждому.
3. Вследствие этого — поддержка как широким кругом пользователей, так и крупнейшими производителями программного и аппаратного обеспечения.
4. И последнее, на чем можно остановиться, — способность системы к саморазвитию, саморасширению. Это объясняется тем, что чем больше ресурсов вовлекается в Сеть, тем она становится интереснее и полезнее пользователям, круг которых в результате растет. Есть и другая причина — постоянное снижение расходов на работу в Internet.

Internet-ресурсы сегодня — это, прежде всего, объемы — свыше 10 млрд документов на более 50 млн Web-сайтов. По заявлению аналитической компании Cyveillance (`http://www.cyveillance.com`), темпы роста Сети составляют 7 млн новых страниц в день. По прогнозам, "центр роста" Internet сейчас уходит из США. Динамика роста объемов информационных ресурсов в Сети настолько велика (для сравнения, можно отметить — количество Web-сайтов в 1998 году составляло около 1 млн), что методы решения задачи обеспечения навигации в ресурсах Internet кажутся далеко не очевидными.

Сегодня каждый пользователь New Media на собственном опыте "ощущает" один из самых больших парадоксов этой среды: "полезной информации становится все больше, но найти что-то определенное все сложнее".

Как гласит опубликованный аналитической службой Netcraft Web Server Survey (`www.netcraft.com`) отчет (рис. 1.1), количество Web-сайтов в Internet в 2004 году достигло 50 млн, а темпы увеличения их числа составляют 1,7 млн

в месяц. Количество же отдельных документов (страниц), размещенных на этих сайтах, составляет около 10 млн. Заметим, речь идет о ресурсах открытой части Internet, доступной информационно-поисковым системам. О гораздо большем объеме ресурсов "скрытого" Web речь пойдет ниже.

Рис. 1.1. Кривая роста числа Web-сайтов (данные службы Netcraft)

При этом даже самые крупные информационно-поисковые системы в мире охватывают в своих индексах не более 30–40% доступных ресурсов. Было бы логичным, чтобы владельцы некоторых систем подобного типа, договорившись, попытались охватить лишь определенные "вертикальные" фрагменты Сети, совместно решая задачу полного охвата ресурсов и обеспечивая качественную навигацию в своих областях. Однако такая модель утопична, а тенденции на рынке глобальных информационно-поисковых систем никак нельзя назвать радужными. Реалии таковы: новизна охватываемой информации падает, навигационные сервисы в основной своей массе не улучшаются, а количество самих глобальных информационно-поисковых систем (за редким исключением не ставших порталами, решающими другие задачи) стремительно растет.

Свалка или клондайк?

Эффективное использование традиционных поисковых систем достигается только в случае обращения их к относительно стабильной части информационного пространства. Но парадокс заключается как раз в том, что Internet в основном таковым *не является*.

С точки зрения обновляемости информации, все Internet-пространство можно условно разделить на две составляющие — стабильную и динамическую. Стабильная составляющая содержит информацию "долговременного" плана, например монографии, галереи, коллекции или архивы. Динамическая составляющая включает постоянно обновляемые или новые ресурсы. Небольшая часть этой составляющей вливается затем в стабильную, в то время как большая часть "исчезает" из Сети.

В свою очередь, информационные потребности пользователей можно условно разделить на две части — "знания и понятия" и "новости". Очевидно, что первая часть потребностей в большей мере удовлетворяется стабильной составляющей Internet, в то время как потребности в новостях могут найти свое удовлетворение только в динамической составляющей New Media.

1.2. New Media и СМИ

На сегодняшний день New Media де-факто заняла место в ряду других средств массовой информации (СМИ). Любое СМИ, будь то печатное издание, радиостанция или телеканал, обладает своими техническими возможностями. Для различных видов СМИ эти параметры разные. Характеристиками потенциала печатного издания служат его тираж, формат, число страниц и т.д. Для радиостанции или телеканала — это частота вещания, мощность передатчика, область охвата. Internet обладает своим техническим потенциалом: пропускной способностью каналов, количеством подключенных компьютеров, их характеристиками и т.д. Реализация потенциала New Media, так же как и в случае традиционных СМИ, выражается в посещаемости, популярности, аудитории и ее направленности.

В качестве подтверждения важности роли Internet как средства массовой информации приведем пример — пожар на Останкинской телевизионной башне в августе 2000 года, обусловивший бурный всплеск интереса к Internet в России. То, что стало катастрофой для российского ТВ, заметно подтолкнуло развитие Рунета. Число посещений ленты новостей сайта РИА "РосБизнес-Консалтинг" (`www.rbc.ru`) 28 августа составило 6 млн, что в 6 раз превышает средний ежедневный трафик (рис. 1.2).

Internet-газета "Lenta.ru" (`www.lenta.ru`) сообщила, что за тот же день было зафиксировано 63,5 тыс. посещений, что в 2 раза превышает обычный показатель. Однако аналитики заявляют, что сдвиг в сторону СМИ в Internet предшествовал пожару в Останкино. Они говорят о том, что спрос на сетевую информацию вырос чуть ранее — 11 августа того же года, подогретый трагедией на подлодке "Курск".

Действительно, катастрофы, скандалы и крупные спортивные события способствуют популярности Internet. После терактов 11 сентября 2001 года суммарная аудитория сетевых СМИ увеличилась в два раза. К примеру, трафик только крупнейших российских новостных сайтов резко вырос почти в три раза и составил около 15%.

Естественно, Internet-издания превосходят по оперативности всевозможные печатные издания. Ленты новостей Web-сайтов содержат самую оперативную информацию, публикуемую в режиме "реального времени". Этим в основном и объясняется стремление традиционных СМИ к интеграции с New Media.

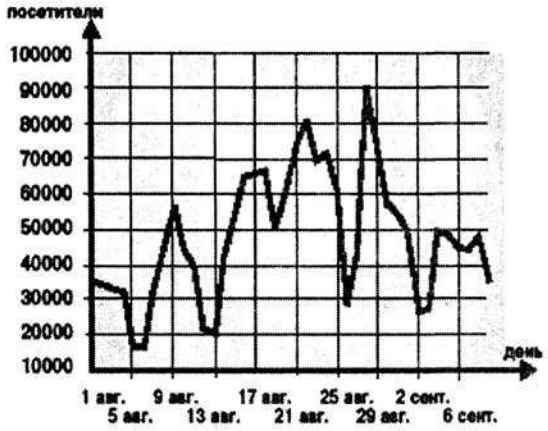


Рис. 1.2. Пожар в Останкино "спровоцировал" интерес к Internet. Статистика сайта РБК

New Media как рекламная среда

Согласно исследованию, проведенному аналитической компанией JupiterResearch, объем рынка он-лайновой рекламы, включая контекстную и баннерную, к 2009 году вырастет по сравнению с 2004 годом почти вдвое и составит 16,1 млрд долларов. По данным этой компании в 2003 году продажи превысили 6,6 млрд долларов, в 2004 году составят порядка 8 млрд долларов. В частности, рекламодатели более чем в два раза увеличат расходы на размещение платных ссылок в результатах поиска информационно-поисковых систем по определенным ключевым словам — с 2,6 млрд долларов в 2004 году до 5,5 млрд долларов в 2009 году.

По прогнозам, в 2008 году рынок Internet-рекламы впервые превысит рынок рекламы в печатных изданиях. Рост продаж будет вызван рядом факторов: увеличением числа Internet-пользователей, ростом популярности Internet-сервисов, а также созданием новых, более совершенных и точных инструментов рекламы во Всемирной Сети. "Данный рынок феноменально вырос за последние несколько лет, — комментирует аналитик JupiterResearch Нэйт Эллиот (Nate Elliott). — Теперь он становится более сформировавшимся".

По другим данным, предоставленным аналитической фирмой E-Marketer, в 2004 году объем затрат на Internet-рекламу в США впервые превысит рекордный уровень 2000 года и составит 9,1 млрд долларов. В 2000 году на пике неоправдавшегося бума Internet-экономики затраты на рекламу в Internet в США составили 8,1 млрд долларов. В 2003 году этот показатель достиг 7,3 млрд долларов. Это означает, что в настоящее время рост американского рынка Internet-рекламы составляет 25%, что больше, чем в любом другом секторе рекламной индустрии.

В России, по данным Ассоциации Коммуникационных Агентств России (АКАР), в 2003 году сегмент Internet-рекламы составил 18 млн долларов и также является самым быстрорастущим. Он растет в два раза быстрее, чем весь рекламный рынок. Сегодня годовой прирост в этом сегменте в России составляет свыше 60%.

Вместе с тем, темпы годового роста данного рынка все же постепенно замедлятся: с 65% в 2003 году до 11% в 2009 году. Поэтому крупнейшие информационно-

поисковые службы уже сегодня четко видят необходимость развиваться в разных направлениях, предоставляя разнообразные услуги как пользователям, так и рекламодателям.

СМИ в Internet и сетевые СМИ

Сегодня принято различать два понятия: СМИ в Internet и сетевые СМИ. В свое время вместе с переносом СМИ в Internet зародился процесс создания в Сети изданий, электронные версии которых дополняли (а порой и заменяли) традиционные. СМИ в Internet зачастую представляют собой прямую репликацию традиционных средств массовой информации на Web-серверах. Лишь немногие средства массовой информации, даже имея свое "представительство" в Internet, смогли "перешагнуть" рамки традиционного представления своей информации и стать полноценными сетевыми СМИ.

При этом миф о негативном влиянии сетевых СМИ на популярность традиционных прототипов на практике не нашел своего подтверждения, скорее справедливо обратное.

Сетевые СМИ — это новый тип носителей информации, изначально ориентированный на Internet, учитывающий многие нюансы представления информации в New Media. Как правило, выпуск традиционным СМИ полноценного сетевого варианта требует не только изменения форматов и формы подачи информации, но и определенной семантической корректировки материалов. Сетевым СМИ присущи два огромных преимущества: оперативность и интерактивность. Вторая особенность подразумевает возможность самостоятельного "выстраивания маршрута" при чтении материалов издания, используя механизм гиперссылок или встроенные поисковые системы. Вместе с тем, бытовавший ранее миф о "миграции" читательской аудитории "бумажных" СМИ в New Media и негативном влиянии сетевых СМИ на популярность традиционных прототипов в действительности не нашел своего подтверждения, скорее справедливо как раз обратное.

1.3. Гипертекст и WWW

Гипертекст, появившийся как форма гиперсвязи между отдельными фрагментами текста, настолько же древнее понятие, как и письменность. Библия, с ее сложным употреблением аннотаций и комментариев, — один из древнейших примеров гипертекста. Словари и энциклопедии также могут рассматриваться как сети из текстовых блоков, соединенных ссылками.

В XX веке (1945) Ванневер Буш (Vannevar Bush) создал первую фотоэлектрическую память и вычислительное устройство Memex (memory extension), представляющее собой справочник, реализованный путем гиперссылок в пределах документа. Тед Нельсон (Ted Nelson) в 1965 году ввел термин "гипертекст" и создал гипертекстовую систему Xanadu с двухсторонними гиперсвязями.

В 1980 году Тим Бернерс-Ли (Berners-Lee), консультант CERN (Европейская организация ядерных исследований), написал программу, позволяющую создавать и просматривать гипертекст и реализующую двунаправленные связи между документами в коллекции [69]. В 1990 году для поддержки документации, циркулирующей в CERN, Бернерс-Ли начал работу над графическим интерфейсом пользователя (GUI) для гипертекста. Эта программа была названа "WorldWideWeb" (рис. 1.3). К 1992 году уже были созданы такие программные реализации GUI, как Erwise и Viola.

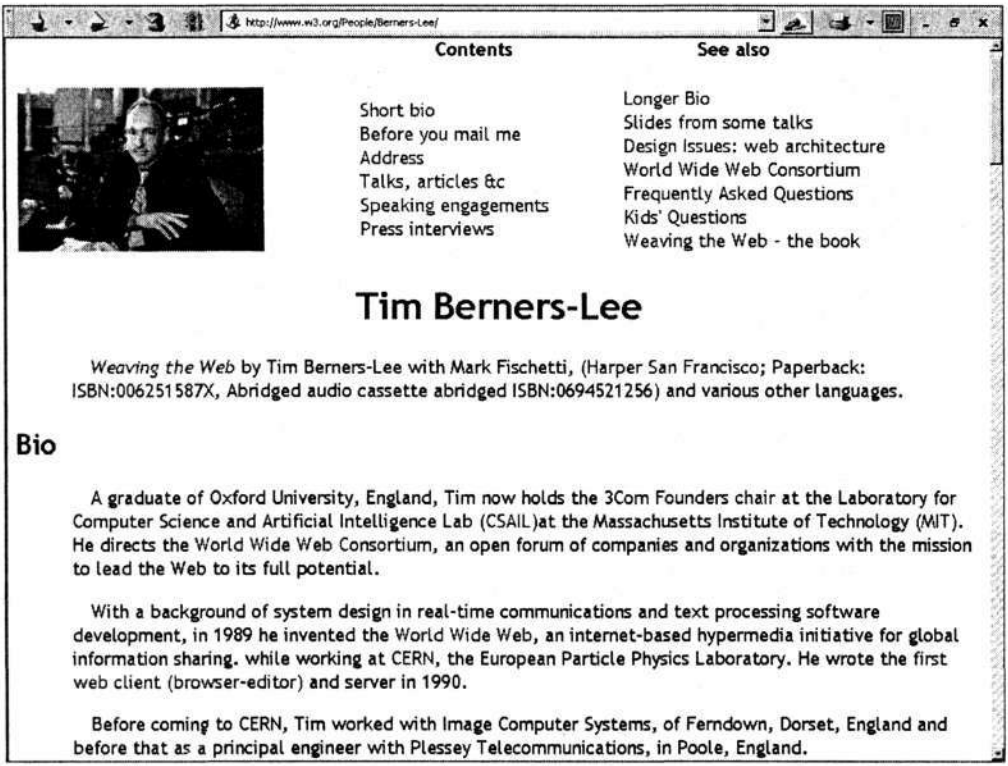


Рис. 1.3. Персональная Web-страница отца-основателя WWW

В феврале 1993 года Марк Андрессен (Mark Andressen) из NCSA (Национальный Центр Суперкомпьютерных приложений США, www.ncsa.uiuc.edu) закончил начальную версию программы визуализации гипертекста Mosaic для популярного графического интерфейса Xwindow System под UNIX. Одновременно CERN развивал и улучшал HTML — язык гипертекстовой разметки текстов, и HTTP — протокол передачи гипертекста, а также сервер обработки гипертекстовых документов — CERN HTTPD.

С тех пор гипертекстовое пространство стало активно развиваться. В 1993 году гипертекстовый трафик составлял от 0,1% до 1% всего Internet-трафика. К концу 1993 года существовало несколько сотен HTTP-серверов. Год 1994 стал переломным: была основана Mosaic Communications Corporation (позже Netscape), состоялась первая конференция WorldWideWeb и MIT совместно с CERN основали Консорциум WorldWideWeb (W3C).

1.4. Интеграция информационных ресурсов

Конечно, большинство СМИ, представленных в Internet, находят своего потребителя. Однако если рассматривать всю совокупность сетевых СМИ как некую общность по отношению к конкретному пользователю (или группе таковых), то обнаруживается ряд проблем, связанных с полнотой, релевантностью и оперативностью получения новостей.

Пользователи зачастую часами "зависают" в Internet, обходя сотни сайтов с целью получения новостей по определенной тематике (или предметной области). В этом поиске традиционные каталоги и поисковые системы оказывают лишь косвенную помощь: они указывают адреса сайтов соответствующей тематики. Однако ни одна из традиционных универсальных поисковых систем не поможет в поиске актуальных новостей — период индексации таких систем составляет от недели до нескольких месяцев. Тем не менее количество уникальных сообщений на новостных Web-сайтах в российском и украинском сегментах Internet превышает 100 тыс. записей в сутки. Неудивительно, что во всем мире, в том числе и в странах восточной Европы, начали создаваться службы интеграции новостей.

Для предоставления тематического (соответствующего специальным запросам) контента из Internet в корпоративные сети или порталы американская служба Moreover (www.moreover.com) обеспечивает сбор данных с 7 тыс. источников в режиме реального времени, классифицируя информацию, которая обновляется каждые 15 минут.

В 2002 году популярная система Internet-поиска Google запустила свой новостной сервис — Google News, который интегрирует информацию с 4500 различных сайтов. Данные рассортированы по нескольким категориям, таким как международные новости, деловой мир, шоу-бизнес, технологии и спорт. "Новости — естественное продолжение нашей миссии", — заявил представитель компании Марисса Майер. Новости в системе отбираются в зависимости от времени их публикации, популярности источника информации и количества появившихся в Internet статей на данную тему.

Одна из самых перспективных в Сети служб интеграции новостей NewsIsFree (www.newsisfree.com) охватывает свыше 12 тыс. источников (в том числе и несколько десятков российских и украинских). Основная особенность службы NewsIsFree — это полная интеграция с XML, в частности с RSS (рис. 1.4.).

Российское агентство Интегрум (www.integrum.ru) обеспечивает сбор электронных версий коммерческих и новостных информационных продуктов. Доступ к данным в Интегрум обеспечивается с помощью информационно-поисковой системы Артефакт, основанной на уникальных морфологических алгоритмах. В 5200 базах данных службы содержится свыше 300 млн документов.

Известный российский поисковый портал Яndex открыл проект Яndex.Новости (http://news.yandex.ru), с которым в настоящее время сотрудничают свыше 130 партнеров — Internet-изданий. Для сбора новостей в службе используется формат RSS 2.0 (Realy Simple Syndication).

Система интеграции новостей InfoStream (http://infostream.ua) обеспечивает интеграцию информации более чем с 800 сайтов. Ядром системы является полнотекстовая информационно-поисковая система InfoReS, обеспечивающая рассылку релевантной информации по электронной почте, непосредственный доступ пользователей к оперативным и ретроспективным базам данных, а также возможность аналитической обработки и обобщения информации. Персонализация интерфейса пользователей, работающих в режиме он-лайн, реализуется на основе современных технологий, ориентированных на формат RSS.

Интеграция сетевых новостей на неплохом уровне выполняется в России также службами ЗАГОЛОВКИ.РУ (www.zagolovki.ru) и Webscan (www.webscan.ru), а в Украине в рамках проектов Медиа-Хвыля (www.media-wave.com.ua) и Паук новостей (www.topnews.com.ua).

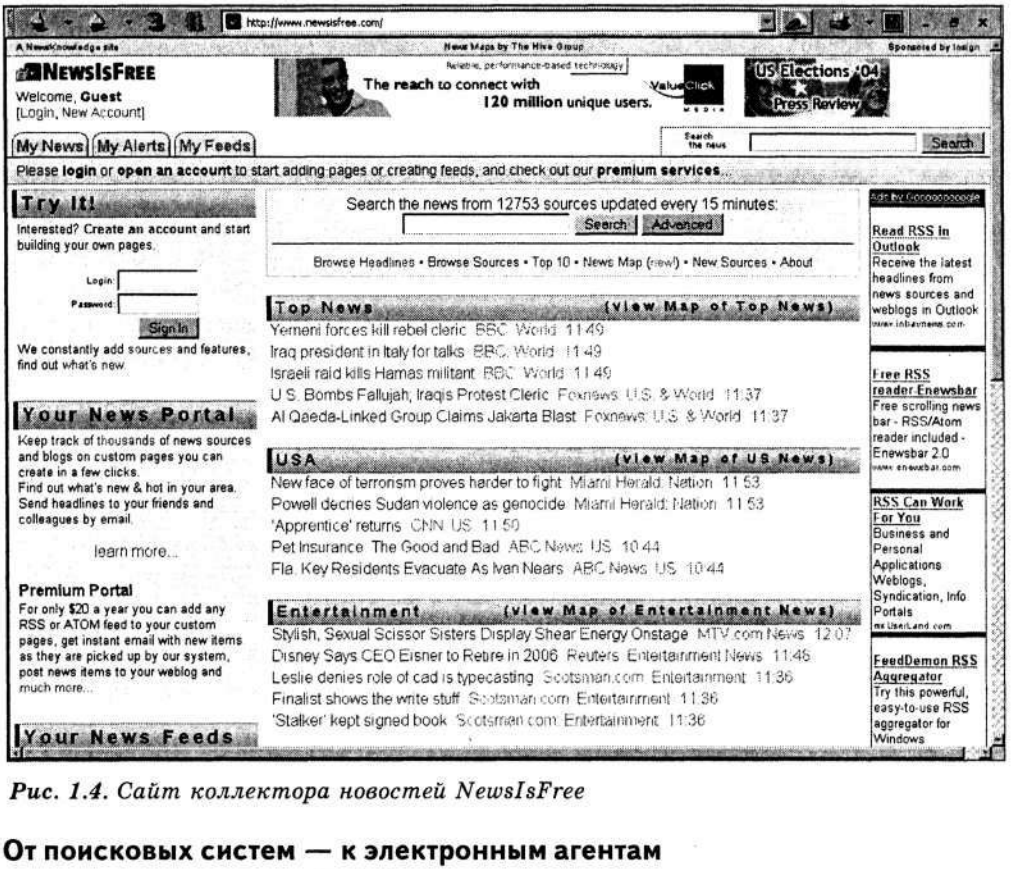


Рис. 1.4. Сайт коллектора новостей NewsIsFree

От поисковых систем — к электронным агентам

Николас Негропонте (Nicholas Negroponte) из MIT еще несколько лет назад на страницах "Wired" (www.wired.com) заметил, что будущее принадлежит электронным агентам по сбору информации. Некоторые инструменты фильтрации информационного потока сегодня можно видеть в Internet на серверах названных выше проектов [14].

Вместе с тем, фундаментальные разработки в этом направлении начались лишь с развитием XML-технологий. Обычно поиск, фильтрация и сбор информации в Internet, во-первых, сопряжены с необходимостью отвлечения соответствующих человеческих ресурсов и оплаты дополнительных временных затрат, а во-вторых, требуют достаточной квалификации персонала и, к сожалению, не могут учитывать всех особенностей структуры Сети и представления информации в ней. Это, в свою очередь, не делает полученную выборку информации из Internet репрезентативной.

При этом информационный поток, "потребляемый", например, организацией из Internet, носит, как правило, выраженную предметную окраску, характеризуемую областью интересов данной организации. Один из вариантов сокращения общих расходов на сбор и фильтрацию информации — выделение специального персонала для выполнения функций ее сбора, селекции и "доводки". Однако поиск и предварительная обработка информации в ручном режиме — достаточно трудоемкий процесс, который не всегда позволяет достичь желаемого эффекта.

Решение перечисленных задач возможно путем создания автоматических и автоматизированных систем сканирования, фильтрации и анализа информации, так называемых своеобразных "интеллектуальных посредников" между пользователем или корпоративной информационной системой организации и Internet. Подобная система выполняет всю "черновую" работу по сбору и селекции информации из Сети и создает документальную базу данных, специфицированную предметной областью заказчика. Загрузка информации в базу данных сопровождается ее категоризацией и частичным "обогащением". Для последующей информационно-аналитической работы конечному пользователю корпоративной информационной системы предоставляются эффективные средства навигации и поиска информации в созданной документальной базе данных.

1.5. Топология Web-пространства

Сегодня каждый пользователь на своем опыте может почувствовать в действии один из самых больших парадоксов Internet — "полезной информации в WWW становится все больше, но найти что-то необходимое — все сложнее".

Традиционные средства "учета" информационных ресурсов Сети — каталоги и информационно-поисковые системы — уже сегодня не справляются с задачей поиска информации, поставленной в общем виде. Эффективными оказываются лишь узко тематические (или региональные) каталоги и поисковики.

Вместе с тем, в отличие от обычного хранилища информации, Web-пространство характеризуется большим количеством скрытых в нем неявных экспертных оценок, реализованных в виде гиперссылок. Именно гиперссылки оказались базой для построения модели Web-пространства.

Для большего охвата информационных ресурсов средствами информационно-поисковых систем необходимо учитывать архитектуру всего Web-пространства, но именно этой информацией никто ранее не владел. Близкой к реальности математической модели не существовало до 1999 года.

В ноябре 1999 года один из руководителей Института поиска и анализа текстов, входящего в исследовательское подразделение IBM, Андрей Бредер (Andrei Broder) и его соавторы из компаний AltaVista, IBM и Compaq совершили прорыв, математически описав "карту" ресурсов и гиперсвязей существующего пространства World Wide Web [42–44]. Исследования опровергли расхожее мнение, будто Internet — это единое густое пространство. Проследив с помощью поискового механизма AltaVista свыше 200 млн Web-страниц и несколько миллиардов ссылок, размещенных на этих страницах, ученые пришли к следующим выводам о структуре Web-пространства. По их мнению, эта структура в действительности соответствует ориентированному графу с топологией "галстука-бабочки" (Bow Tie), в котором вершины соответствуют страницам, а ребра — соединяющим страницы гиперссылкам. Анализ структуры связей между отдельными Web-страницами, выполненный в рамках этой модели, позволил обнаружить следующее.

1. Центральное ядро (28% Web-страниц) — компоненты сильной связности (SCC) или узел галстука. Сюда относятся Web-страницы, связанные так тесно, что, просто следуя по гиперссылкам, из любой из них в конечном счете можно попасть на любую другую.

2. "Отправные" Web-страницы (IN) (22% Web-страниц) — они содержат гиперссылки, которые в конечном счете ведут к ядру, но из ядра к ним попасть нельзя.
3. "Оконечные" Web-страницы (OUT) (столько же — 22%) — к ним можно прийти по ссылкам из ядра, но нельзя вернуться назад в ядро.
4. "Отростки" (еще 22% Web-страниц) — полностью изолированы от центрального ядра: это либо "мысы", связанные гиперссылками со страницами любой другой категории, либо "перешейки", соединяющие две Web-страницы, не входящие в ядро.

Указанные четыре основных множества, в сумме составляющие более 90% всех исследованных Web-страниц, каждая из которых топологически относится к одной компоненте связности, и обусловили название полученной модели (рис. 1.5) — Bow Tie ("галстук-бабочка"). Помимо этого, в Web существуют и "острова", которые вообще не пересекаются с остальными ресурсами Internet. Единственный способ обнаружить ресурсы этой группы — знать их адрес. Никакие поисковые машины не смогут найти эти острова, если они в прошлом каким-то образом не соединялись с другими частями Internet.

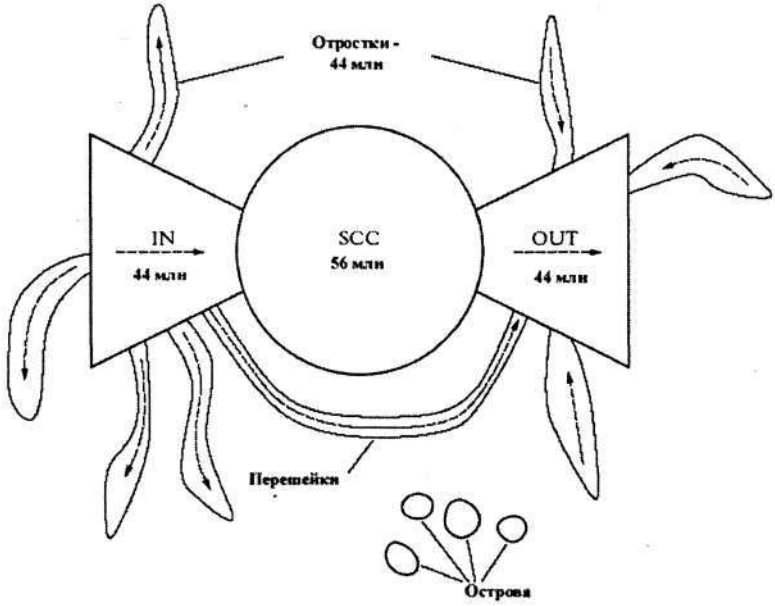


Рис. 1.5. Модель Bow Tie

Исследователи обнаружили, что пропорции этих четырех категорий в течение нескольких месяцев оставались неизменными, несмотря на значительное увеличение общего объема Web-ресурсов.

Были исследованы такие параметры данной модели, как среднее количество сайтов, через которые связываются любые два сайта гиперссылками, а также распределение входящих и исходящих ссылок. Было показано, что распределение полустепеней захода и исхода вершин графа Web-пространства подчиняется степенному

закону, т.е. вероятность того, что соответствующая степень вершины равна i, пропорциональна $1/i^k$ (для входящих ссылок $k \approx 2{,}1$, а для исходящих $k \approx 2{,}45$).

По словам исследователей, эксперимент выявил гораздо более детальную и сложную картину: значительная область WWW вообще отделена от других крупных частей, — говорится в отчете компаний. С большой степенью вероятности случайно выбранные Web-страницы окажутся никак не связаны. Если же путь все-таки существует, среднее количество щелчков, необходимых для переходов между ними, составляет 16. А если этот путь двусторонний, то среднее число промежуточных щелчков сокращается до семи.

Топология и характеристики модели оказались примерно одинаковыми для различных подмножеств Web-пространства, подтверждая тем самым наблюдение о том, что "Web — это фрактал", т.е. свойства структуры Bow Tie всего Web-пространства также верны и для его отдельных подмножеств. Таким образом, алгоритмы, использующие информацию о структуре Web-пространства, предположительно будут работать и на отдельных его подмножествах.

Информация о структуре Web-пространства уже достаточно широко используется при решении многих задач, например, для оптимизации эффективности механизмов сканирования, при анализе и прогнозе его развития, при построении новых Web-сервисов.

Полученные в результате исследований сведения заставили заново взглянуть и на стратегии Web-серфинга. Теория Bow Tie поясняет динамический характер Сети и позволяет получить представление о некоторых особенностях сложной организации WWW. Благодаря полученным результатам, уже сегодня может быть создан инструментарий, способный превратить Web-пространство в систему двустороннего движения. "Сейчас трафик по существу односторонний. Если бы браузер был наделен средствами серфинга в обратном направлении, это открыло бы доступ к гораздо большему числу ресурсов", — заявил по этому поводу представитель IBM Нам Ламор (Nam LaMore).

1.6. Навигация в Internet

Традиционные средства навигации в Web-пространстве — это каталоги и поисковые системы [21]. Причем первыми появились Web-каталоги, как психологически наиболее приближенные к образу мышления человека. Действительно, каталоги в принципе не требуют от пользователя ввода какой-либо информации с клавиатуры — достаточно воспользоваться гиперссылками, чтобы найти необходимую информацию. Трудно представить традиционный "бумажный" каталог, содержащий несколько миллионов ссылок. Точно так же трудно ориентироваться в электронном Web-каталоге, не используя дополнительных возможностей, главной среди которых является возможность ввода "своего" запроса с клавиатуры.

Рост объема Web-ресурсов привел к появлению и бурному росту информационно-поисковых серверов в Сети. Сегодня наиболее развитые системы навигации в Internet обладают свойствами как Web-каталогов, так и информационно-поисковых серверов. Среди таких систем — мировые лидеры Google, Yahoo, AltaVista, Alltheweb. В России лидирующее положение занимают системы Яndex, Rambler и Aport. В Украине первые каталоги появились в 1995 году, а поисковики — в 1997. В настоящее время известно до десятка украинских информационно-поисковых серверов и около пятидесяти каталогов. Среди лидирующих систем можно назвать UAport и META.

Web-каталоги и информационно-поисковые серверы (и их симбиоз) стали прародителями нового типа Web-сервиса — порталов, т.е. "ворот в Internet".

В то же время сегодня прослеживается эволюция порталов от поисковых машин и каталогов до самостоятельных, насыщенных информацией и самодостаточных Web-ресурсов. Порталы, как новые объекты WWW, возникли в 1998 году. Основная идея их создания заключалась в стремлении, наряду с возможностями навигации в Сети, предоставить пользователю максимальный уровень сервиса, сделать так, чтобы каждый свой сеанс работы в Internet он начинал именно с данного ресурса.

По мнению специалистов компании McKinsey, в Internet-бизнесе можно выделить три основных момента: привлечение нового пользователя на сайт (attraction); превращение посетителя в клиента (conversion), которое достигается, если сайт настолько интересен пользователю, что он проводит там значительное количество времени; и, наконец, необходимость сделать так, чтобы у посетителя после ухода с Web-сайта оставались причины вернуться туда (retention). В этом плане типичный портал пытается не только привлечь пользователей удобными средствами навигации в Internet, но и "удержать" их, предоставляя на своем ресурсе максимум необходимой и полезной информации. Таким образом, порталы представляют собой объединение средств навигации и информационных служб, однако это далеко не полная их характеристика. Портал представляет собой сайт, организованный как системное многоуровневое объединение разных ресурсов и сервисов. Как правило, такой сайт совмещает в себе разнообразные функции, предлагает разноплановые информационные ресурсы и различные сервисы (поиск, рубрикаторы, финансовые индексы, информация о погоде и т.д.). С момента появления первых порталов основные функции "ворот в Internet" существенных изменений не претерпели: это средства реализации поиска данных, общения, новостная часть, торговля и службы приложений. Таким образом, можно дать следующее определение понятию "портал": *сайт (или совокупность сайтов), обеспечивающий удовлетворение основных потребностей пользователей путем реализации услуг (сервисов) в следующих областях: информация, бизнес, общение, а также предоставления инструментария, необходимого пользователю для продвижения собственного контента в рамках портала*. В соответствии с данным определением портал должен включать четыре основных типа сервисов.

1. Информационный сервис — все, что помогает найти (при необходимости) и получить информацию.
2. Сервис реализации бизнес-функций — все то, что ориентировано непосредственно на продажу товаров/услуг.
3. Инструментарий пользователя — все, что помогает ему создавать и продвигать свой контент в сети, прежде всего бесплатный хостинг и бесплатный e-mail, рейтинги, баннеры, "анонсировщики" и др.
4. Сервис обеспечения общения (community) — все, что направлено на удовлетворение потребности в общении.

Различают "вертикальные" и "горизонтальные" порталы. *Вертикальный портал* — это обладающий всеми качествами портала видовой или тематический сайт, ориентированный на один тип информационного наполнения. *Горизонтальный портал* — это поливидовой и политематический сайт, обладающий

всеми качествами Web-портала. Горизонтальный портал может включать в себя несколько видовых или тематических вертикальных порталов.

При отсутствии четкого определения понятия "портал" и в связи с инвестиционной привлекательностью данного направления деятельности в области Internet, порталами зачастую называют сайты, не удовлетворяющие некоторым, а зачастую и многим признакам порталов. В этой связи в качестве примеров приведем наиболее характерные проекты, которые признаны во всем мире. Это, прежде всего, названные выше зарубежные каталоги и поисковики — Yahoo!, AltaVista, Google.

Среди поисковых систем, эволюционирующих в порталы, заслуживают внимания также Lycos (`http://www.lycos.com`) и Excite (`http://www.excite.com`) — рис. 1.6. Другими общеизвестными путями пришли к "порталности" такие сайты, как Microsoft (`http://www.microsoft.com`) и AOL (`http://www.aol.com`).

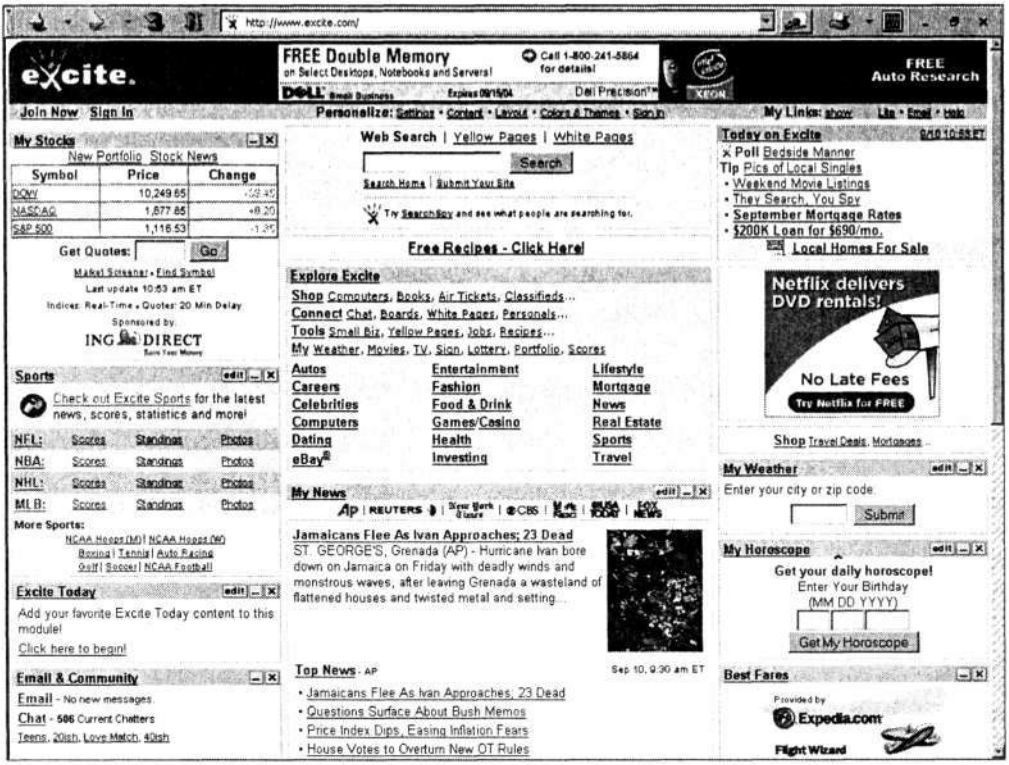


Рис. 1.6. Excite — поисковая система, ставшая порталом

Конечно же, создание порталов требует долгосрочных инвестиций. При отсутствии в СНГ фондового рынка, на котором котировались бы акции Internet-компаний, инвесторам приходится ориентироваться на текущие финансовые показатели проекта и на прогнозы его развития в будущем. Несмотря на убыточность едва ли не любого Internet-проекта в первые несколько лет его существования, некоторые из них котируются довольно высоко.

Сегодня стремление рекламодателей к повышению отдачи от своих вложений в он-лайновый маркетинг подтверждается и результатами опроса, проведенного

в начале 2004 года компанией Forrester Research: заказчики практически единодушно высказали свое предпочтение вертикальным Internet-порталам перед web-структурами общего характера. Вертикальные порталы, такие как CBS Sportsline, CNNfn, Garden.com, CNET, нацелены на конкретную категорию контента, сферу торговли или сегмент аудитории и готовы предоставлять широкий набор услуг для определенной целевой группы. Горизонтальные же порталы, такие как AOL, Yahoo!, MSN, AltaVista, предоставляют лишь набор базового контента, коммуникационных и торговых услуг.

Несмотря на то что в настоящее время три портала — America On Line, Yahoo! и MSN — забирают около 15% всего сетевого трафика и являются получателями 45% всех денег за он-лайновую рекламу, рекламодатели замечают, что реклама на вертикальных порталах более эффективна. Поэтому ожидается, что в ближайшее время вертикальные порталы будут являться получателями свыше половины общих расходов на рекламу в Internet.

Сколько же должно быть порталов, в частности, в русскоязычной части Internet или в Украине? Многие аналитики считают, что для России, например, достаточно 2–3 десятков порталов, больше не потребуется ни пользователям, ни инвесторам. Если не принимать во внимание качества российских "порталов", то уже сегодня эта цифра превышена в сто раз — разработчики просто используют модную терминологию. С другой стороны, мировой опыт показывает, что на 1000 сайтов должен приходиться один навигатор (каталог, поисковик или портал). На основании последних расчетов, на украинскую часть Сети должно приходиться 10–20 порталов. Что интересно, именно около 20 Web-сайтов позиционируют себя здесь как "порталы", зачастую "вертикальные".

1.7. Информационно-поисковые системы

Первые полнотекстовые информационно-поисковые системы (Fulltext Retrieval System) появились в начале компьютерной эры. Назначением этих систем был поиск в библиотечных каталогах, архивах, массивах документов, таких как статьи, нормативные акты, рефераты, тексты брошюр, диссертаций, монографий. Вначале информационно-поисковые системы (ИПС) использовались преимущественно в библиотечном деле и в системах научно-технической информации.

В 1966 году 16-ю американскими библиотеками с целью установления стандартного формата для электронных каталогов была начата реализация проекта МАРК, обеспечившего переход к унифицированному обмену электронными данными, что способствовало эффективной организации баз данных библиографических каталогов. Внедрение стандартного библиографического формата позволило библиотекам объединить усилия в работе над электронными каталогами. В 1972 году получил международное признание стандарт МАРК-2 [39], на основе которого были созданы многие национальные стандарты [6] (рис. 1.7).

В начале 1970-х годов коммерческие компьютерные службы уже предоставляли возможность интерактивного поиска в тематических базах данных Национальной медицинской библиотеки и Министерства образования США. При этом некоторые из этих служб существуют и сегодня — основанная еще в 1965 году система ДИАЛОГ, входящая в настоящее время в корпорацию Thomson, сегодня обеспечивает своим клиентам доступ к сотням базам данных.

В настоящее время информационные ресурсы только сети Internet составляют свыше десятка миллиардов документов (Web-страниц), к которым возможен

свободный доступ любого пользователя. Естественно, чтобы найти необходимую информацию в этой крупнейшей распределенной полнотекстовой базе данных, необходимо использовать самые мощные ИПС. Такие системы существуют и конкурируют друг с другом на современном рынке информационных технологий.

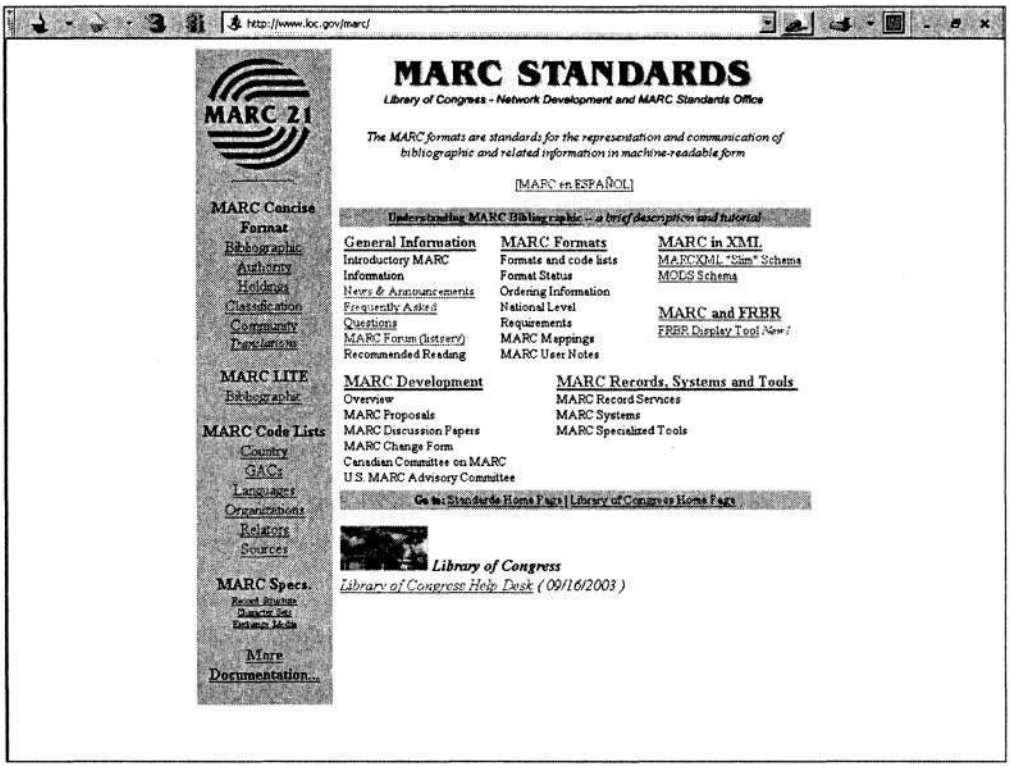


Рис. 1.7. *Версии стандарта МАРК популярны во всем мире*

В начале 1990-х годов для унификации информационных систем был разработан важный международный стандарт Z39.50 — информационно-поисковый протокол для библиографических систем [72]. В 1994 году университет Джорджии запустил пилотный проект "ГАЛИЛЕЙ" с использованием Site-Search — пакета программ Огайского центра в стандарте Z39.50 (рис. 1.8). Стандарт Z39.50 положен в основу службы поиска распределенной информации в Internet — системы WAIS (Wide Area Information Service) [63].

Сегодня миллионам пользователей Internet известны такие информационно-поисковые системы, как Google, Yahoo, AltaVista, AllTheWeb, каждая из которых охватывает свыше миллиарда Web-документов. За прошедшее десятилетие технология полнотекстового поиска стала повседневным инструментом миллионов пользователей. При этом далеко не все лидеры информационного рынка осознали эту тенденцию десятилетие назад.

"Недостаточные инвестиции Microsoft в технологию Internet-поиска были непростительной ошибкой компании, но она работает над тем, чтобы наверстать упущенное. Говорят, что Microsoft успевает везде, но вот вам пример того, где мы не успели", — заявил CEO корпорации Стив Баллмер, выступая в начале

2004 года перед аудиторией менеджеров по маркетингу и представителей СМИ на пятой ежегодной конференции Microsoft по рекламе в Редмонде. Microsoft с трудом протискивается на одну из самых оживленных территорий в WWW и пока отстает от своих главных конкурентов. При этом Баллмер заявил, что в ближайшие 12 месяцев команда разработчиков Microsoft должна предложить поисковую технологию нового поколения.

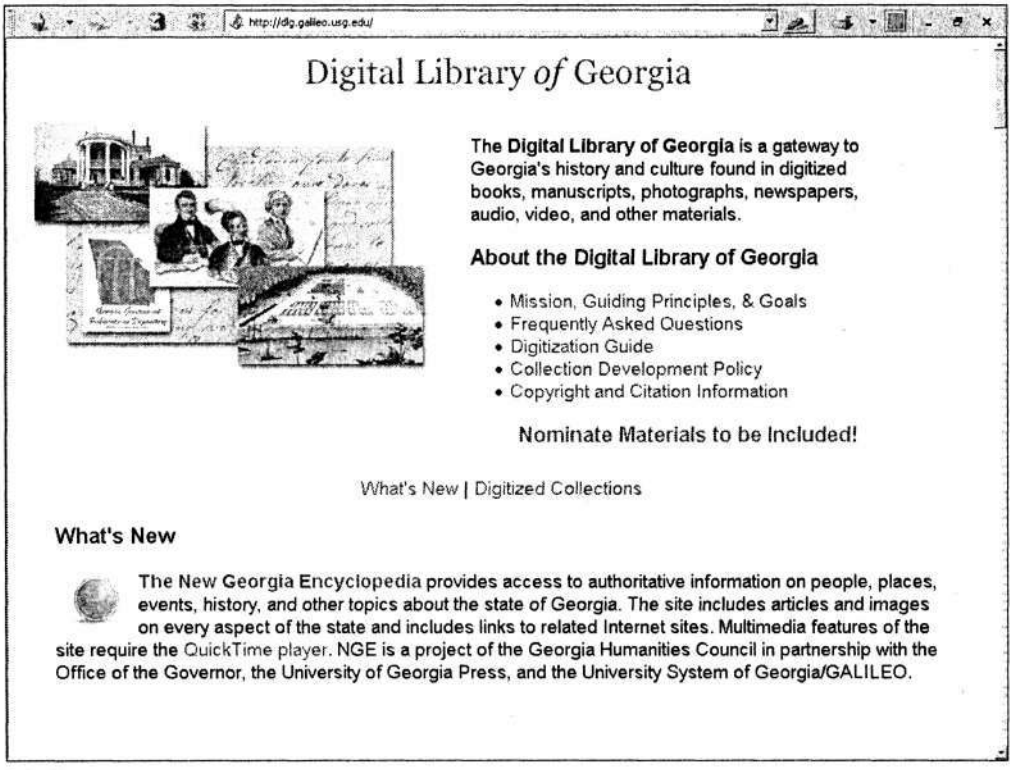


***Рис. 1.8.** Сайт проекта "ГАЛИЛЕЙ" сегодня*

Для охвата поиска в новостной части Internet Microsoft уже сегодня приступила к тестированию агрегатора новостей MSN Newsbot. Сейчас поисковая база MSN Newsbot формируется по результатам сканирования четырех тысяч источников. Представители Microsoft заявляют, что преимущество MSN Newsbot состоит не столько в ширине охвата Internet, сколько в персонализации результатов поиска.

Для нахождения информации в Internet, чаще всего представленной в формате HTML, используются специальные средства — сетевые информационно-поисковые системы. Очень часто возникает вопрос: как соотносятся эти системы, работающие с потоками документов в форматах гипертекстовой разметки, и популярные сегодня реляционные системы управления базами данных (СУБД)? Решают ли СУБД такого класса задачи поиска информации в текстовом формате, и будут ли они эффективны в случае гипертекстовых документов?

Современные реляционные СУБД поддерживают обработку запросов в стандарте SQL, позволяющем проводить поиск в рамках реляционной модели. Иными

словами, стандартные средства этих систем обеспечивают эффективный поиск по совокупностям формализованных полей в рамках двухмерной таблицы. Полнотекстовый же поиск сводится к нахождению отдельных слов, их комбинаций, словосочетаний в рамках одного или нескольких текстовых полей (ячеек или тэгов), порой значительного размера.

В то время как промышленные СУБД предлагаются в качестве основ для конкретных, иногда очень масштабных приложений, информационно-поисковым системам, построенным на основе СУБД, присуща эффективность текстового поиска в достаточно узко очерченном фрагменте. Причина кроется в том, что подобные ИПС просто не предназначены для решения этой специфической задачи. Поэтому несмотря на постоянное совершенствование информационно-поисковых систем, встроенных в СУБД, с сожалением, приходится констатировать их непригодность для решения задач глобального поиска информации в Internet-ресурсах.

Если в контексте этого утверждения рассматривать программное обеспечение таких СУБД, как Oracle, Adabas, Informix, DB2, с одной стороны, и программно-технологические решения сетевых ИПС, таких как Alltheweb, AltaVista, Google, Yahoo!, Exite, с другой стороны, то проследить корреляцию между этими понятиями действительно трудно. Стоит отметить, что необходимость решения задачи полнотекстового поиска для навигации в сетевых ресурсах является всего лишь одной из предпосылок становления нового подхода к представлению информации в Internet, и об этом пойдет речь ниже.

В отличие от реляционных СУБД, у систем полнотекстового поиска не существует стандартизированного языка запросов. У каждой системы этого типа существует свой способ задания критериев поиска. Очень часто языки запросов поисковых систем приближены к SQL, однако каждой из них присущ ряд индивидуальных особенностей, связанных с такими моментами, как

- интерпретация операций, зависящих от порядка расположения слов в тексте (операций контекстной близости слов и др.);
- реализация вычисления близости, т.е. определения соответствия найденных документов запросам (релевантности) для представления результатов поиска;
- применение нестандартных функций, требующих, например, использования методов искусственного интеллекта (нахождение документов по принципу подобия, построение дайджестов из фрагментов документов и др.).

В различных полнотекстовых ИПС различаются архитектуры, структуры данных, алгоритмы их обработки, методологии организации поиска.

1.8. "Скрытый" Web

1.8.1. Очередной феномен Internet

В Internet информации куда больше, чем можно найти с помощью традиционных информационно-поисковых систем. Чаще всего пользователь выходит на необходимые ему новые источники в Сети через поисковые системы-брэнды, ставшие для многих "де-факто" стандартными. Однако, кроме видимой для поисковых систем части Web-пространства, существует огромное количество страниц, которые ими не охватываются. При этом доступ пользователя к таким

ресурсам в принципе возможен (хотя иногда "слегка прикрыт" паролями). Как правило, эти Web-страницы доступны в Internet, однако выйти на них трудно, а порой невозможно, если не знать точного адреса. Эти ресурсы уже десять лет как имеют собственное название — "скрытый" (deep) Web [71], которое ввел Джилл Иллсворт (Jill Ellsworth) в 1994 году, обозначив им источники, недоступные для обычных поисковых систем. Сегодня такие ресурсы называют также "невидимым" (invisible) Web. Они чаще всего охватывают динамически формируемые Web-страницы, содержание которых хранится в базах данных и доступно лишь по запросам пользователей.

В 2000 году американская компания BrightPlanet (`www.brightplanet.com`) опубликовала сенсационный доклад, в котором утверждается, что в Web-пространстве в сотни раз больше страниц, чем их удалось проиндексировать самыми популярными поисковыми системами. Эта же компания разработала программу LexiBot, которая позволяет сканировать некоторые динамические Web-страницы, формируемые из баз данных, и, запустив ее, получила неожиданные данные. Выяснилось, что для традиционных поисковых систем огромная часть Сети просто невидима (рис. 1.9).

Рис. 1.9. *Часто задаваемые вопросы по скрытому Web на сайте BrightPlanet*

Напомним, что в ноябре 1999 года Андрей Бредер и его соавторы из компаний AltaVista, IBM и Compaq разработали структурную модель ресурсов и гиперсвязей Web, опровергнув мнение, что Internet — это единое связное пространство.

Мы уже обсуждали выше топологию этой модели, получившей название Bow Tie. Здесь же мы еще раз остановимся на "островах", которые не пересекаются с остальными ресурсами Сети. Единственный способ обнаружить ресурсы этой группы — точно знать их адрес. Поисковые машины в принципе не находят этих островов, если они в прошлом каким-то образом не соединялись с другими частями Internet. Именно этот факт объясняет недостатки модели Бредера — он исследовал в основном страницы открытого (поверхностного) Web, к тому же отбирая их, видимо, не совсем случайно. Поэтому, если процентное соотношение первых четырех составляющих "поверхностного" Web можно признать верным, "острова" в реальности будут более объемными, чем в модели. Согласно исследованиям компании BrightPlanet, число скрытых (но не секретных) Web-страниц во много раз превышает количество видимых. Доступные сегодня посредством традиционных информационно-поисковых систем 10 млрд Web-страниц — это лишь видимая крупица. Непознанных, скрытых ресурсов Сети в сотни (!) раз больше. Это, прежде всего, динамически генерируемые страницы, файлы нераспознаваемых поисковыми системами форматов, информация из многочисленных баз данных. В результате исследований также было выявлено немало интересных особенностей "скрытого" Web. Так, например, известно, что средняя его страница на 27% компактнее средней страницы из поверхностной части Web-пространства.

1.8.2. Типы скрытых ресурсов

Для того чтобы определить, какие из ресурсов невидимы для поисковых систем, следует рассмотреть принцип работы типового индексатора-робота таких систем. Эти программы-роботы, как правило, посещают Web-страницы по известным заранее адресам, анализируют их содержание и выделяют гиперссылки, идущие от них. Обычно, обработав текущую страницу, выделив ключевые слова и некоторые поля, робот переходит по адресам, найденным на ней, сканирует последующие страницы, выделяет новые адреса и т.д. Обычно, если робот определяет, что в данный момент обращается к динамической странице, он останавливает свою работу. Эта тактика выбрана в предположении, что чаще всего для получения осмысленного ответа из баз данных требуется осмысленный запрос, а большинству из роботов чужды элементы интеллекта, даже искусственного. В результате "скрытый" Web охватывает в первую очередь содержимое он-лайновых баз данных, доступных в сети. Динамической является и быстро обновляемая информация — новости, конференции, он-лайновые журналы.

Конечно, есть и явные "острова" по Бредеру, на которые не указывают никакие гиперссылки и от которых никаких гиперссылок не исходит. Защищенные паролями коммерческие Web-сайты также попадают в категорию "скрытого" Web — о материалах этих сайтов большинство пользователей никогда не узнают лишь с помощью поисковых систем. Однако относительное количество таких сайтов невелико. Например, среди крупнейших сайтов "скрытого" Web платными являются только 10% ресурсов, хотя именно они включают важнейшие издательства и базы данных.

Основатель BrightPlanet Майкл Бергман (Michael K. Bergman) выделил 12 разновидностей "скрытых" Web-ресурсов (`www.leidenuniv.nl/ub/biv/specials.htm`), относящихся к классу он-лайновых баз данных. В списке оказались как традиционные базы данных (патенты, медицина и финансы), так и публичные ресурсы — объявления о поиске работы, чаты, библиотеки, справочники. Бергман причислил к "скрытым" ресурсам и специализированные поисковые системы,

которые обслуживают определенные отрасли или рынки, базы данных которых не включаются в глобальные каталоги традиционных поисковых служб.

К "скрытому" Web также относятся многочисленные системы интерактивного взаимодействия с пользователями — системы помощи, консультирования, обучения, требующие участия людей для формирования динамических ответов от серверов. К ним также можно отнести и закрытую (полностью или частично) информацию, доступную пользователям Сети только с определенных адресов, групп адресов, иногда городов или стран. К "скрытой" части Сети многие причисляют и Web-страницы, зарегистрированные на бесплатных серверах, которые индексируются, в лучшем случае, лишь частично — поисковые системы во избежание рекламного спама не стремятся обходить их в полном объеме.

Недавно появилась категория так называемых "серых" сайтов, функционирующих на основе динамических систем управления контентом (Dynamic Content Management Systems). В поисковых системах обычно ограничивается глубина индексирования таких сайтов во избежание возможного циклического просмотра одних и тех же страниц.

И конечно же, "скрытыми" оказываются и Web-сайты, создатели которых не оповещают кого-либо о создании этих ресурсов.

Безусловно, основной формат данных, с которым работают традиционные поисковые системы в Internet, — это HTML, причем статическая его часть. С другими форматами у многих поисковых систем имеются различные проблемы. К примеру, наличие различных версий формата PDF (Adobe Portable Document Format), а также особенности хранения инкапсулированных графических изображений заставляют считать сетевые ресурсы, представленные в этом формате, "скрытыми". Тем не менее некоторые современные поисковые системы уже вполне сносно индексируют документы в PDF-формате. К "скрытым" форматам принято относить также и Flash, широко использующийся для обеспечения визуальных эффектов на Web-сайтах.

Кроме того, для нашего пользователя наверняка "скрытой" можно признать большую часть гигантского китайского сегмента Internet. Например, малопопулярный в Европе и Америке китайский поисковый портал Baidu (`www.baidu.com`) в 2004 году опередил Google по объему трафика и стал четвертым в мире Web-ресурсом по этому показателю. Еще одна китайская поисковая система, `3721.com`, заняла седьмое место. Эти данные по ранжированию привела исследовательская компания Alexa, речь о которой пойдет ниже. Портал `Baidu.com` стал крупнейшей в мире поисковой системой на китайском языке и охватывает более 95% китайских пользователей Сети.

1.8.3. Базы данных "скрытой" Сети

Пожалуй, самыми большими из известных ресурсов "скрытого" Web являются базы данных служб Dialog и LexisNexis.

Одной из крупнейших мировых служб информационного поиска является американская компания Dialog (`http://www.dialog.com`), созданная при поддержке NASA и до 1988 года принадлежавшая аэрокосмической фирме Lockheed. Сегодня Dialog принадлежит корпорации Thomson (США) — одному из всемирных лидеров в области предоставления интегрированных информационных решений (рис. 1.10). Корпорация Thomson имеет свыше 20 миллионов пользователей в 130 странах мира.

Рис.1.10. *Сайт службы Dialog*

Сервисом компании Dialog также пользуются в более чем 100 странах мира. Образованная в 1965 году как первая в мире он-лайновая информационно-поисковая служба, Dialog фактически определила современные стандарты управления информацией. На сегодняшний день она включает такие продукты и сервисы, как Dialog®, Dialog Profound®, Dialog DataStar®, Dialog NewsEdge® и Dialog Intelliscope, которые обеспечивают доступ к более 1,4 млрд документов через Internet или сети intranet. При этом в компании Dialog определяют свои ресурсы как часть "скрытого" Web (Deep Web), заявляя, что содержат полезной, не дублирующейся информации в 500 (!) раз больше, чем доступно с помощью традиционных информационно-поисковых систем. Коллекция баз данных службы Dialog содержит 900 баз данных, доступных 700 000 пользователям, которые только за один час прочитывают свыше 17 млн документов из этих баз данных.

Основанная в 1973 году, крупнейшая в мире он-лайновая служба LexisNexis (`http://www.lexisnexis.com`) предоставляет своим пользователям юридическую, политическую, коммерческую, новостную, регистрационную и другую информацию (рис. 1.11). С 1979 года система баз данных LexisNexis — первая в мире служба полнотекстового поиска. В настоящее время эта служба охватывает свыше 35 000 источников информации, содержащих в совокупности более 4,6 млрд документов с глубиной ретроспективы до 200 лет. Каждый час в базы данных LexisNexis добавляется 57 500 документов. LexisNexis представлена сегодня в 20 странах, пользователи сервиса находятся в более чем 100 странах.

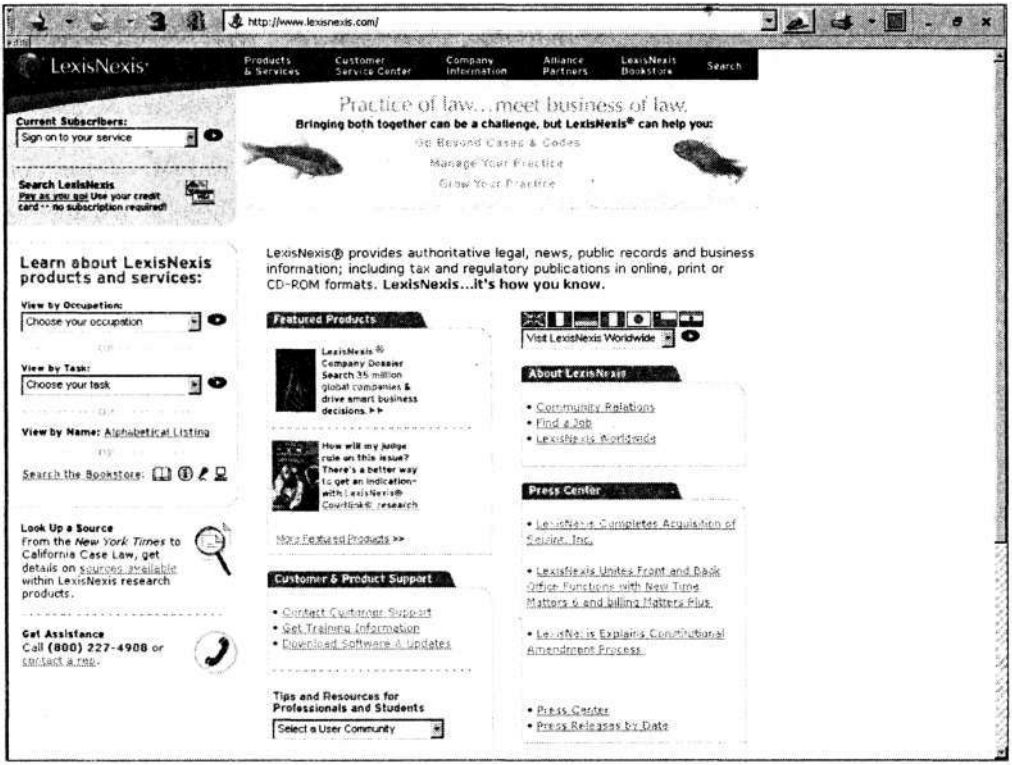


Рис. 1.11. *LexisNexis — крупнейшая он-лайновая информационная служба*

К коммерческим базам данных "скрытого" Web можно отнести и информационные ресурсы крупнейших мировых информационных агентств, уже много лет работающих на рынке финансовой информации, таких как Reuters, Tenfore, Dow Jones Telerate, Bloomberg.

С другой стороны, в "скрытом" Web существует множество альтернатив коммерческим базам данных. Среди них, например, сайт `www.10kwizard.com`, предлагающий доступ к полным текстам корпоративных документов, хранящихся в Комиссии США по ценным бумагам и биржам. Существуют тысячи баз данных "скрытого" Web, свободно доступные для пользователей, но чаще всего не охватываемые традиционными поисковыми системами.

Приведем еще несколько примеров.

- Educator's Reference Desk (`http://www.askeric.org`) — этот ресурс содержит свыше двух тысяч учебных планов, несколько тысяч ссылок на образовательные документы, а также ссылки, представляющие собой запросы к архиву. С этого сайта обеспечивается доступ к базе данных ERIC — крупнейшему источнику информации по проблемам образования, а также к полнотекстовым дайджестам, составляемым экспертами.
- Nuclear Explosions Database (`http://www.ga.gov.au/oracle/nukexp_query.html`) — австралийская база данных по географии. Для работы с системой достаточно перейти в режим "Online Tools", после чего будет представлен список баз данных и карт.

- PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) — с ресурса обеспечивается доступ к свыше 14 млн ссылок системы MEDLINE, включая ссылки на полные тексты статей и информационные ресурсы. Имеется возможность перехода к службе PubMed Central (PMC), к свободно доступному архиву статей (свыше 90 тысяч) из научных журналов. Обеспечивается также доступ к глобальной поисковой системе NCBI, охватывающей базы данных по естествознанию.
- LookSmart's FindArticles (http://www.findarticles.com) — база данных FindArticles — доступный через Web-интерфейс архив, содержащий 2,8 млн статей из более чем 500 источников, накапливаемый с 1998 года.

1.8.4. Сталкеры в скрытом пространстве

"Скрытый" Web представляет собой гигантский репозитарий документов, звуков, изображений, фильмов и т.п. Безусловно, если большая часть этой информации не доступна традиционным поисковым системам, то существует потребность в специальных инструментах поиска "скрытого" контента. Эти инструменты включают каталоги, метапоисковые сайты, доступные через Web базы данных, а также большое количество глобальных, региональных и специальных поисковых систем.

Для поиска в "скрытой" Сети, а именно в том ее сегменте, который составляют базы данных, сегодня уже существуют некоторые специализированные ресурсы. Среди них, например, системы BigHub (www.bighub.com) и InvisibleWeb (www.invisible-web.net) компании IntelliSeek (рис. 1.12).

Сайт Invisible Web включает в себя каталог баз данных, большинство из которых не проиндексированы известными поисковыми машинами. При введении запроса этот сайт выдает ссылки на ресурсы, с помощью которых поиск необходимой информации станет наиболее оптимальным. На этом сайте Криса Шермана (Chris Sherman) и Гари Прайса (Gary Price) собраны коллекции ссылок на различные базы данных, среди которых содержится немало уникальных ресурсов — например, сборник спичей политиков и бизнесменов. Программный пакет BullsEye компании IntelliSeek осуществляет поиск более чем в 800 сетевых ресурсах.

Лидером среди навигаторов в "скрытом" Web является сайт CompletePlanet (www.completeplanet.com) компании BrightPlanet. Этот сайт является крупнейшим каталогом, насчитывающим свыше 100 тыс. ссылок. Компания BrightPlanet также создала персональную утилиту для поиска в он-лайновых базах данных — LexiBot, которая может обеспечивать поиск в нескольких тысячах поисковых систем "скрытого" Web. Метапоисковый пакет DeepQueryManager (DQM) этой же компании обеспечивает поиск по 55 тыс. "скрытых" Web-ресурсов.

Сайт Direct Search (http://www.freepint.com/gary/direct.htm), созданный Гари Прайсом, также обеспечивает поиск в базах данных "скрытого" Web. На сайте содержится ссылка на лучшие ресурсы ценовой информации (MySimon.com), финансовой информации (FinancialFind.com), а также ссылки на информацию из научно-популярных журналов и научных баз данных по биотехнологиям (Biolinks.com).

В Internet есть и другие сайты-навигаторы, а также специализированные программы поиска. Например, поисковая система по университетским архивам, библиотекам и книгам — Infomine Multiple Database Search (http://infomine.ucr.edu/search.phtml); каталог информационных сайтов, которые уникальны в своих областях, — BUBL LINK (bubl.ac.uk/link); полнотекстовый поиск по содержанию всех книг — Amazon.com.

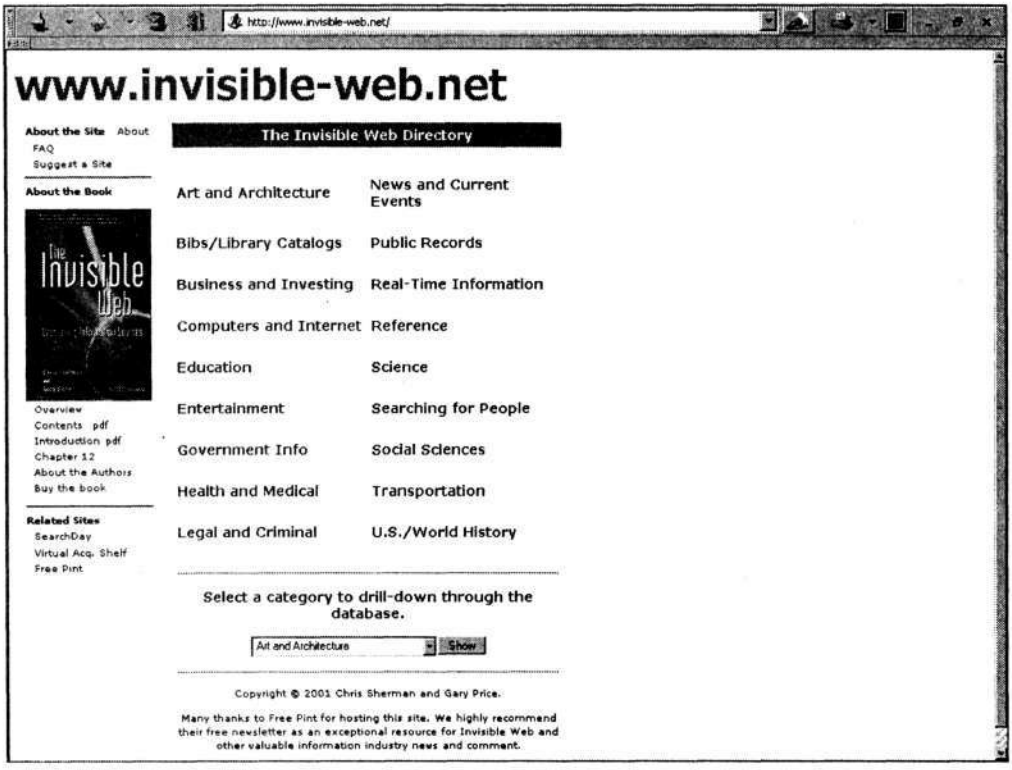


Рис. 1.12. Сайт InvisibleWeb

Особенность большинства "скрытых" ресурсов — в их узкой специализации. Для поиска в них используются те же механизмы, что и для "поверхностного" Web, однако чаще всего роботы поисковых систем для "скрытого" Web включают уникальные для каждого такого ресурса модули доступа к данным.

1.8.5. "Скрытый" Web в каталогах

Каталоги, как глобальные, так и специальные, могут содержать ссылки на "скрытые" ресурсы, прежде всего базы данных. Приведем несколько самых известных примеров.

- Портал WebData.com на первый взгляд ничем не отличается от других подобных ресурсов, однако содержит гиперсылку "Add Your Database" (добавить Вашу базу данных), говорящую о том, что на данном портале можно зарегистрировать базу данных — часть "скрытого" Web.
- Librarians' Index to the Internet (http://lii.org) — каталог, содержащий свыше 14 000 Internet-ресурсов. LII также включает ссылки на "скрытые" в Web-пространстве базы данных. У владельцев таких баз данных есть возможность поместить соответствующую гиперссылку в этом каталоге на свой ресурс (в LII есть ссылка "and databases" (добавить базу данных).
- FindLaw (`http://www.findlaw.com`) — один из наиболее популярных в мире юридических Web-сайтов, представляющий собой огромный каталог

правовых ресурсов, содержащий аннотированный список свободно доступных баз данных нормативно-правовых документов, для которых данный ресурс является "точкой входа".

- InfoMine (http://infomine.ucr.edu) — ресурс, содержащий ссылки на 120 000 документов, представленных в 9 аннотированных базах данных. Этот каталог позиционирует себя как "виртуальную библиотеку Internet-ресурсов", ориентированную на студентов и исследователей-профессионалов.
- About.com (http://www.about.com) — портал, охватывающий тысячи снабженных комментариями ссылок на Web-ресурсы, в том числе и на ресурсы "скрытого" Web (имеется ссылка "Invisible Web"). На портале предоставляется возможность поиска в каталоге. Ресурс также включает несколько статей по проблематике "невидимого" Web: "What is the Invisible Web?", "Finding the Invisible Web", "Top Places to Search the Invisible Web" и др.

К разряду каталогов можно также отнести следующие коллекции ссылок, поисковые системы и "скрытые" базы данных.

- Direct Search (http://www.freepint.com/gary/direct.htm) — ресурс, содержащий ссылки на ресурсы "скрытого" Web. Например, присутствует ссылка на сайт ResourceShelf (http://www.resourceshelf.com), обеспечивающий поиск в блогах (сетевых журналах) и новостных сообщениях.
- The Invisible Web Directory (http://www.invisible-web.net) — Web-сайт Шермана и Прайса (Chris Sherman & Gary Price), соавторов термина "Invisible Web".
- Profusion (http://www.profusion.com) — сайт компании Intelliseek, первой создавшей каталог "невидимого" Web InvisibleWeb.com. ProFusion; это модифицированная метапоисковая система, позволяющая выбирать области поиска в "вертикальных" (тематических) разрезах.
- CompletePlanet (http://www.completeplanet.com) — сайт корпорации BrightPlanet Corporation, который охватывает свыше 70 000 поисковых баз данных и специальных поисковых систем.

1.8.6. Системы поиска в "скрытом" Web

Традиционная поисковая система чаще всего может выдать адрес базы данных, но не укажет, какие конкретно документы содержатся в ней. Типичный пример — информационно-поисковые системы по украинскому (http://www.rada.gov.ua) или российскому (http://www.kodeks.ru) законодательству. Тысячи документов из баз данных становятся доступны только после входа в систему, а роботы стандартных поисковых систем не в состоянии заиндексировать контент баз данных. Многие поисковые системы, как глобальные, так и локальные, описаны на сайтах Search Engine Watch (http://www.searchenginewatch.com) и Search Engine Showdown (http://www.searchengineshowdown.com). На этих сайтах приведены, среди прочих, и поисковые системы "скрытого" Web.

- Singingfish (http://www.singingfish.com) — эта поисковая система обеспечивает поиск аудио- и видеофайлов, представленных на Web-сайтах.

- Scirus (http://www.scirus.com) — поисковая система по представленным в Internet научным материалам, включая статьи из журналов и отчеты. Со страницы расширенного поиска (Advanced Search) доступны многочисленные тексты из баз данных EBSCO и ProQuest.
- UFOSeek (http://www.ufoseek.com) — поисковая система по материалам о паранормальных явлениях и НЛО.

Качественный и полноценный поиск информации в "скрытом" Web возможен и с использованием таких специализированных коммерческих баз данных, как Dialog, ProQuest, Web of Science. Но эти базы данных, ввиду своей платности, сами являются объектами "скрытого" Web.

1.8.7. Информация в различных форматах

Информация, представленная в форматах, отличных от HTML, для многих поисковых систем оказывается недоступной, хотя сегодня ситуация меняется в корне. Например, популярная система Google (http://www.google.com) уже обеспечивает поиск в документах, представленных в форматах MS PowerPoint, DOC, RTF, Postscript, PDF, а также обеспечивает преобразование этих файлов в текстовый формат. Поиск документов разнообразных форматов доступен в этой системе как из режима расширенного поиска в Google (Advanced Search), так и из "простого" поиска — достаточно использовать в запросе команду "filetype:", уточнив поиск выражением "filetype:pdf".

Знаменитая служба Yahoo! сегодня уже не только каталог, но и полнофункциональная поисковая система. Поисковая система Yahoo! Search (http://www.yahoo.com), как и Google, обеспечивает выдачу текстовых копий документов, размещенных в Internet в форматах Word, Excel, PowerPoint и PDF, а также RSS/XML-фидов (новостных лент и блогов — "живых журналов").

Специализированная система Gigablast (http://www.gigablast.com) предназначена исключительно для поиска по документам в форматах Word, Excel и PDF. Эта система выдает по запросу кэшированные (архивные) копии документов в исходных форматах, при этом обеспечивает булевый поиск и выдачу версионных копий документов, которые были размещены в Сети, но затем, возможно, удалены.

1.8.8. Скрытые новостные ресурсы

Текст новостей тоже традиционно относился к "скрытой" Сети, однако в последние годы все крупнейшие поисковые сайты разработали эффективные инструменты поиска оперативно обновляемых новостных сообщений — это, например, " Яndex.Новости" (http://news.yandex.ru), Google News (http://news.google.com) или Uaport (http://uaport.net/UAnews). Служба Google News автоматически собирает новости из нескольких тысяч источников, обновляя свои базы данных каждые 15 минут. Существуют и другие зарубежные службы интеграции новостей, например NewsIsFree, Topix.net и Daypop (www.daypop.com). В России крупнейшими интеграторами новостей являются системы Integrum (http://www.integrum.ru) и Webscan (http://www.webscan.ru), в Украине — InfoStream (http://infostream.ua) и WebObserver (http://webobserver.info).

Многие сайты на своих страницах публикуют новости, как собственные, корпоративные, так и общеотраслевые. Если на сайте не реализован статический

механизм архивации старых сообщений, то, даже будучи помещенными в архив, доступный из Internet, эти сообщения рискуют оказаться в зоне "скрытого" Web.

Материалы публикаций попадают в разряд "невидимого" Web и в том случае, если они защищены паролями как средствами обеспечения оплаты или просто сбора статистики о читателях. Многие аналитики (в частности, аналитик IDC Джеймс Левин) признают, что для изданий значительно выгоднее публиковать усеченную бесплатную версию своих материалов — это обеспечит их популярность в Internet, попадание изданий в индексы популярных поисковых систем.

1.8.9. "Скрытый" архив "поверхностного" Web

Парадоксально, но как один из ресурсов "скрытого" Web можно рассматривать и архив ресурсов открытого Web-пространства. Такой архив — Internet Archive — с 1996 года создает компания Alexa (www.alexa.com). Сегодня объем базы данных Alexa превышает 500 Тбайт. Новые страницы в настоящее время попадают в хранилище со скоростью 1 Тбайт в день. Технология хранилища Alexa включает ряд современных средств управления гигантским документальным хранилищем. Например, с помощью технологии Alexa выполняется кластеризация Web-ресурсов, т.е. формирование коллекций документов, близких по тематикам. Особый интерес у пользователей сервиса Alexa вызывает "Машина времени" (Wayback Machine), открывающая доступ к временным срезам Web-пространства. Одно из наиболее интересных практических применений этой технологии — восстановление документов, некогда опубликованных в Web-пространстве, но впоследствии удаленных. При этом рост "скрытого" Web грозит серьезными пробелами в хранилище системы, связанными с увеличивающимся количеством сайтов, эксплуатирующих различные технологии управления контентом, динамической публикацией документов из баз данных и т.п.

Аналогичный проект — Informedia (www.informedia.com), но относящийся только к одному типу информации (аудиовизуальной), разрабатывается в институте Карнеги Меллона. Informedia появился в 1996 году в рамках инициативы Digital Library Initiative. С тех пор к проекту в роли спонсоров присоединились многие компании, в том числе Microsoft, Intel, CNN, Boeing и даже Visa. В рамках проекта разрабатываются технологии распознавания образов и речи.

1.8.10. Подходы к решению проблемы "скрытого" Web

Чем быстрее растет Web-пространство, тем хуже оно охватывается традиционными каталогами и поисковыми машинами. Ввиду роста количества Web-сайтов и порталов, использующих в своей работе хранящуюся в базах данных информацию, динамических систем управления контентом, появлением новых версий форматов представления информации, "скрытый" сегмент Web растет очень интенсивно. С одной стороны, Internet как огромное хранилище увеличивает объем информации, доступной "в принципе", но с другой стороны — растет информационный хаос, увеличивается энтропия сетевого информационного пространства. Все меньшая часть информационных ресурсов становится доступной пользователям реально. Объем "скрытого" Web, содержащего полезную для пользователей, но слабодоступную информацию, в сотни раз превышает "поверхностную" часть. Иными словами, традиционные средства охвата информационных ресурсов не справляются с задачей поиска большей части информации.

Эффективными оказываются лишь тематические каталоги и поисковики — сталкеры в мире "скрытого" Web.

Спасти ситуацию могут и новые возможности унификации обмена информацией в Internet. Одним из первых проектов консорциума W3C в этой области стал "Семантический Web", речь о котором пойдет ниже. Основная идея проекта заключается в следующем: Web-серверы должны не только визуализировать, но и использовать данные, чтобы программы разных производителей могли эффективно работать с контентом.

Для решения задачи интеграции новостной информации было создано несколько форматов описания данных на основе XML. Самый распространенный формат получил название RSS, что означает Really Simple Syndication, Rich Site Summary. Сегодня экспорт данных в формате RSS осуществляют крупнейшие порталы, включая CNN, BBC News, Amazon, CNet News, MSNBC, The Register, Wired и т.д.

Аналитики отмечают, что только в начале 2004 года пользователи Internet по-настоящему открыли для себя технологию RSS. Сегодня для работы с данными в формате RSS разрабатываются все новые программы, сайты и поисковые системы, которые все более востребованы пользователями. Эти программы приоткрывают завесу над динамично обновляемой частью "скрытого" Web.

Глава 2

Поиск в Internet

2.1. Характеристики ИПС

Основополагающими характеристиками качества функционирования информационно-поисковых систем являются полнота и релевантность [12]. Сегодня информации в Сети появляется больше, чем ее успевают проиндексировать поисковые системы. Это означает, что информационный хаос увеличивается, и существующие подходы не соответствуют требованиям растущего информационного пространства. Вместе с тем, чем больше ресурсов соответствующего профиля включает база данных системы, тем выше должна быть ее полнота. Сегодня в области сетевых ИПС идет жесткая конкурентная борьба, связанная с этим аспектом. В 2002 году система Alltheweb неожиданно вышла на первую позицию по охвату сетевых ресурсов и, соответственно, была признана лучшей сетевой ИПС в мире, проиндексировав в своей базе данных 2,1 млрд Web-страниц. В настоящее время лидерство вернулось системе Google (свыше 4,2 млрд Web-страниц).

Два аспекта полноты

Понятие полноты динамической базы данных из Internet тесно связано с оперативностью обновления информации [17]. Сеть Internet представляет собой своеобразный "живой организм", — здесь постоянно добавляются новые ресурсы, удаляются устаревшие, некоторые документы меняют адреса, некоторые модифицируются. Созданная однажды база данных является "слепком" состояния информационных ресурсов Сети на конкретный момент. Если база данных ИПС не будет обновляться постоянно и оперативно, имеющиеся в ней ссылки на документы станут мертвыми, т.е. по адресам, представленным в этих ссылках, документы могут либо не существовать, либо будут размещены с совершенно другим содержанием. Кроме того, отсутствие оперативности и обновления баз данных не позволит пользователю отслеживать последние изменения в его предметной области.

Полнота охвата ресурсов Сети — это один из двух аспектов характеристики качества сетевой информационно-поисковой системы. Второй аспект связан с полнотой информации, предъявляемой пользователю по его запросу к ИПС (рис. 2.1). Если предположить, что по запросу пользователя Q в базе данных находится P документов, соответствующих этому запросу, а предъявлено для просмотра всего N документов, то полнота ИПС определяется по формуле: $П = (N/P) \times 100\%$. В случае, если $П$ оказывается больше 100%, очевидно, что пользователю выдано минимум $N - P$ документов, не соответствующих его запросу, т.е. нерелевантных.

Рис. 2.1. Два аспекта полноты охвата ИПС ресурсов Internet

Релевантность и пертинентность

Под релевантностью понимается формальное соответствие информации, выдаваемой системой, запросу. Если по запросу пользователя получено N документов, представляющих собой объединение двух множеств документов: соответствующих запросу (пусть их количество — N_1) и не соответствующих (их количество — N_2), т.е. $N = N_1 + N_2$, тогда релевантность, как степень соответствия, определяется по формуле $P = (N_1/N) \times 100\%$, а шум — по формуле $S = (N_2/N) \times 100\% = 100\% - P$. Если же обозначить количество соответствующих запросу документов в исходном массиве как R, то отношение $(N_1/R) \times 100\%$ будет определять полноту поиска.

Проиллюстрируем эти понятия. Допустим, исходный массив содержит 100 документов, из которых 50 соответствуют запросу. Если в результате поиска будет выдан всего один документ, который при этом будет соответствовать запросу, релевантность выдачи будет равна 100%, шум — 0, а полнота — всего 2%. В другом, крайнем, случае, если будут выданы все 100 документов, релевантность результатов поиска составит 50%, шум — 50%, а полнота — все 100%. Подобные рассуждения позволяют констатировать следующее: чем выше в системе релевантность, тем ниже полнота, и, соответственно, чем ниже релевантность, тем полнота выше.

Это определение характерно для формальной релевантности, однако на практике используется другое, неформальное понятие — пертинентность [23]. Для пользователя пертинентность, как соотношение объема полезной для него информации к общему объему полученной информации, имеет решающее значение (рис. 2.2). При этом следует учитывать, что формальный запрос к системе является предметом творческого осмысления информационной потребности и не всегда точно отражает последнюю. Неумение большинством пользователей правильно формулировать запросы и получать приемлемые объемы отклика породило в конце XX века мнение об Internet как об огромной информационной свалке.

Средства повышения пертинентности в современных системах, помимо возможностей уточнения формулировки запросов, включает и весовые критерии, позволяющие ранжировать найденные документы и выдавать пользователю для просмотра наиболее весомые документы либо вообще ограничиваться выдачей не более заданного числа наиболее весомых документов. В последнем случае, естественно,

страдает полнота выдачи, т.е. при этом полнота и релевантность являются антагонистическими характеристиками — чем выше релевантность, тем ниже полнота и наоборот. Проблеме релевантности, а особенно пертинентности, уделяется большое внимание в современных системах. Так, например, служба Google реализовала алгоритмы достижения неформальной релевантности, и именно благодаря этому в свое время стала самой популярной системой в Internet.

Рис. 2.2. Релевантность и пертинетность

Как сообщается на сайте CNET News.com, в 2003 году калифорнийская компания Google — создатель одноименной поисковой системы — запатентовала метод определения релевантности Web-страниц, отбираемых по запросу пользователя. Это свидетельство стало для Google первым. Ранее компания подала еще три заявки с просьбой выдать патенты на методы и технологии поиска страниц по нечетко определенным запросам и на основе анализа их посещаемости.

Кроме характеристик полноты и релевантности, для пользователей ИПС большое значение имеют такие характеристики, как скорость обработки запросов, получения отклика от системы, достоверность отклика (например, оцениваемая по ее источникам), а также дополнительные сервисы — возможность нахождения документов, подобных уже имеющимся (like this), подключения автоматических переводчиков и, конечно же, уточнения запроса непосредственно после выполнения процедуры поиска.

2.2. Лингвистическое обеспечение ИПС

На всех этапах развития полнотекстовых ИПС лингвистическое обеспечение играло важную роль. Именно средства лингвистики выступают интерфейсами между естественным языком и формальными поисковыми механизмами ИПС.

Лингвистическое обеспечение включает такие основные элементы:

- языки представления данных в ИПС, которые определяют архитектуру, синтаксис и семантику представления информации в базах данных ИПС;
- информационно-поисковый язык, т.е. язык, на котором обращается пользователь к системе, чтобы получить интересующий его отклик.

Современные информационно-поисковые языки включают поддержку булевых операторов (И, ИЛИ, НЕ), операторов контекстной близости, средств управления приоритетами операторов, естественных языков и, наконец, языков разметки, на которых представлены документы-первоисточники.

Большое значение в современных полнотекстовых ИПС уделяется морфологическому анализу, т.е. автоматическим средствам обработки отдельных слов, как в текстах исходных документов, так и в запросах пользователей.

Стоп-слова

Большинство естественных языков имеет так называемые вспомогательные слова типа артиклей и предлогов, которые входят в большинство документов и не влияют на процесс выявления документов, удовлетворяющих информационным потребностям пользователей, занимающихся поиском. Такие слова (например, a, an, the, on для английского языка) называются "стоп-словами".

Поисковые системы обычно не включают стоп-слов в свой индекс, однако учитывают при сквозной нумерации слов, что позволяет выполнять поиск фраз, содержащих "стоп-слова", например "чай с молоком" (хотя имеется ненулевая вероятность появления различных фраз, содержащих значимые слова на определенных местах).

Исключение стоп-слов из индекса ведет к его существенному сокращению и повышению эффективности работы. Однако некоторые запросы, состоящие только из стоп-слов (типа "to be or not to be"), в этих случаях уже не пройдут. Неудобство вызывают и некоторые случаи полисемии (многозначности слова в зависимости от контекста). Например, в одних случаях английское слово "can" как вспомогательный глагол должно быть включено в список стоп-слов, однако как существительное оно часто несет большую содержательную нагрузку.

Морфемный анализ

При построении базы данных из массива документов (в случае сетевых ИПС такими документами выступают отдельные Web-страницы) формируется индекс из всех слов, входящих в эти документы, иногда за исключением так называемой "незначащей лексики" — предлогов, артиклей, частиц и т.д.

Файл незначащей лексики представляет собой стоп-словарь системы. Построенный словарный индекс системы во многих реализациях ИПС лемматизируется, т.е. все слова приводятся к каноническим формам, например существительные — к именительному падежу, глаголы — к инфинитивной форме и т.д. Это особенно характерно для славянских языков, для которых, в отличие, например, от английского, специфично достаточно много словоизменений. Построение индекса системы на основе лемматизированной лексики во многих случаях оправдано, но в системах, ориентированных на профессиональную работу, ориентация только на такой подход является спорной. В системах, работающих с учетом морфологии, лемматизации должны подвергаться и запросы пользователей, т.е. если в исходном документе присутствует словосочетание "белые ночи", то в индексе системы в этом случае имеются слова "белый" и "ночь". Если бы запрос пользователя "белые" и "ночи" был передан на вход поискового механизма без преобразования, то исходный документ не был бы найден, однако если данный запрос подвергнуть лемматизации, то он примет следующий вид: "белый" и "ночь", и исходный документ будет найден.

Для систем, рассчитанных на непрофессионалов (а таких большинство), лемматизация поискового индекса и запроса очень удобна. Например, задав в качестве аргумента поиска слово "конфета", пользователь получит ссылки даже на те документы, в которых это слово используется в различных формах, например "конфеты", "конфетами", "конфет". Более того, представленное в запросе слово "люди" обеспечит поиск и по слову "человек".

Однако для профессионального поиска лемматизация не всегда пригодна, так как она может лишить поиск гибкости. Отсечение окончаний может увеличить количество документов, выдаваемых по запросу пользователей, однако может привести и к выдаче нерелевантных документов. Например, при использовании известного алгоритма Портера для отбора английских слов, одинаковая основа "univers" будет выделена в таких различных словах, как university и universal.

Рассмотрим конкретный пример. При поиске документов, в которых должна быть фамилия "Тарасюк", использовалась одна из систем со встроенными возможностями принудительного морфемного анализа. В результате обработки соответствующего запроса было найдено 32 документа, среди которых 31 документ относился к творчеству Тараса Шевченко и к мероприятиям на Тарасовой горе и лишь один (!) документ оказался релевантным. Таким образом, на практике лемматизация далеко не всегда увеличивает число пертинентных документов. Механизм отсечения окончаний, полезный в некоторых случаях и неэффективный в других, в информационно-поисковых системах должен при необходимости подключаться или отключаться.

Недаром, например, такая служба, как AltaVista, вообще не занимается морфологической обработкой текста. Все слова для нее — лишь последовательности символов.

Профессиональные запросы к традиционным системам

Традиционные системы пакетного поиска, обеспечивающие, например, рассылку результатов по электронной почте, не предполагают интерактивного взаимодействия с конечным пользователем, поэтому им присуща полнота, которая сродни избыточности.

Так, профессиональный запрос к системе Интегрум по теме "Услуги связи" выглядит следующим образом:

"услуги связи" или "междугородные переговоры" или "телефонные переговоры" или "мобильная связь" или "фиксированная связь" или "сотовая связь" или "сотовый оператор" или "средства связи" или "телефонная связь" или "спутниковая связь" или "космическая связь" или GPS или ростелеком или связьинвест или госкомсвязь или госкомтелеком или госсвязьнадзор или телекоммуникации или электросвязь или АТС или ГТС или минсвязи или "министерство связи" или "волоконно-оптическая линия связи" или ВОЛС

В системе InfoStream для реализации точной рассылки сообщений по теме "Мобильная связь" применяется такой запрос:

(((мобильн~связ) | (мобільн~зв'яз) | (сотов~связ) | (стільник~зв'яз) | (беспроводн~связ) | (бездрот~зв'яз) | (бесперебойн~связ) | (безперебійн~зв'яз) | j2me]| ems]| 3g]| gprs]| ggsn]| sgsn]| sms]| mms]| ems]| bluetooth]| mms]| tdma]| multipoint]| pcs]| cdma]| ofdm]| vpn]| wap]| umts]| gsm)&((моб~телефон)| (стільник~телефон)| (сотов~телефон))) ! this.is

Вместе с тем, очевидно, что для работы в интерактивном режиме такие запросы неприемлемы. Пользователь желает ввести 1-2 слова и получить то, что ему необходимо. Тут на помощь могут прийти только интеллектуальные, семантические методы.

Тезаурус

Еще при появлении первых ИПС возникла дискуссия, предметом которой стало использование в качестве индексов систем автоматически формируемых словарей или подключение заранее подготовленных словарных массивов, снабженных рядом дополнительных атрибутов, — тезаурусов.

В тезаурусах каждой лексической единице приписывается небольшой пояснительный текст — словарная статья и ссылки на другие слова этого словарного массива. Содержательно ссылки могут означать следующее: синонимию, противопоставление отдельных слов, подчиненность и т.д. Структура наполнения тезауруса регламентируется соответствующими стандартами — ISO 2788, ГОСТ 7.25-80 (для одноязычных тезаурусов) и ГОСТ 7.24-90 (для многоязычных тезаурусов) [7].

Формирование поискового индекса во многих ИПС выполняется по правилам построения тезаурусов, в которые были включены такие типы лексических единиц:

- отдельные слова (существительные, прилагательные, глаголы, наречия);
- словосочетания;
- лексически весомые компоненты сложных слов;
- аббревиатуры;
- сокращения слов и словосочетаний.

В тезаурусах различные формы лексических единиц приводятся к каноническим формам. Кроме того, лексическим единицам приписываются указатели, которые соответствуют стандарту ISO 2788. В соответствии с этим стандартом определяются такие основные виды ссылок:

- смотри — USE;
- синоним — UF (used for);
- выше — BT (broader term);
- ниже — NT (narrower term);
- ассоциация — RT (related term).

При формировании поискового индекса системы на основе тезауруса каждое слово из документов, входящих в базу данных ИПС, анализируется на вхождение в тезаурус. Особый смысл имеет использование тематических тезаурусов для специализированных баз данных, однако сегодня остается открытым вопрос построения политематического тезауруса и индекса ИПС на его основе. Хотя следует отметить, что при формировании баз данных на основе Web-сайтов Internet именно политематический тезаурус представляет самый большой интерес.

Другой подход, который чаще всего используется сегодня, основан на механизмах автоматического построения поискового индекса системы на основе входящих в документальный массив слов. Этот подход предполагает отказ от использования тезауруса или, по крайней мере, лишь минимальное его использование для второстепенных целей. В настоящее время этот подход считается более технологичным.

В самом начале истории ИПС противники использования тезаурусов приводили как один из основных аргументов то, что объемы тезаурусов не позволяют хранить их в памяти машины. В те времена объемы текстовых баз данных были относительно небольшими и автоматические индексы систем были на порядок менее объемными, чем соответствующие тезаурусы. Сегодня ситуация изменилась в корне — с одной стороны, объемы промышленных носителей информации позволяют хранить практически неограниченное количество тезаурусов, а с другой стороны, объемы текстовых баз данных настолько велики, что их индексы зачастую превышают объемы тезаурусов. Все это дает основание предполагать, что соотношение тезаурусных и бестезаурусных систем, сложившееся в результате инертности внедрения новых технологий, в недалеком будущем изменится.

2.3. Семантические методы

В последнее время в технологии поиска все чаще стали внедряться элементы контент-анализа — методологии, возникшей в конце XIX–начале XX вв. Эта методология, изначально ориентированная на применение в психологии и социологии, сегодня все чаще используется в различных автоматизированных системах. Различают количественный и качественный контент-анализ. Если качественный контент-анализ базируется на глубоком лингвистическом и семантическом анализе отдельных предложений и всего текста, то основой количественного контент-анализа являются статистические подходы.

В последнее время получили развитие такие направления контент-анализа, как "Data Mining" и "Text Mining", которые предполагают автоматическое выявление из текстовых массивов нового смысла, новых данных, феноменов, фактов-знаний. Все чаще возникают попытки привлечения методов контент-анализа, а точнее Text Mining, в реальные поисковые системы. И эти попытки не умозрительны — они обусловлены объемами и темпами роста Сети. Во многие современные сетевые поисковые системы внедрены такие компоненты, как:

- автоматическая группировка документов по определенному заранее классификатору;
- автоматическое определение новых, не заданных заранее классов на основе неструктурированных или слабо структурированных документов;
- ранжирование документов по смысловой релевантности;
- выявление семантически подобных документов — поиск подобных документов на основе эталона;
- автоматический анализ и смысловое преобразование запросов пользователей.

Группировка результатов поиска

В свое время создатели службы Oingo реализовали технологию выявления "смысла" слов путем построения обучаемой внутренней семантической сети. Сегодня наиболее интересной кажется технология, предлагаемая службой AltaVista (`http://www.av.com`), обеспечивающая для реализации режима уточнения поиска (Refine Your Search) автоматическое определение классов и последующую группировку (кластеризацию) откликов ИПС в соответствии ними.

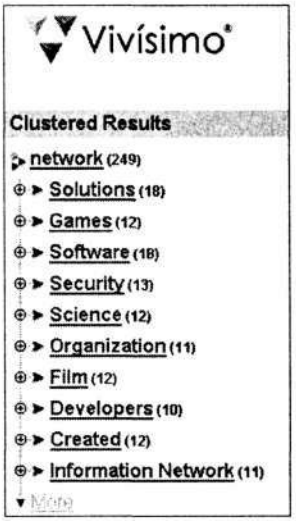


Рис. 2.3. Классы понятий поисковой системы Vivisimo

Например, в результате обработки запроса "network" (сеть) она предлагает следующие классы документов: Management; Solution; Catholic Church; Christian Organization; Domain Names; Blog; Economy; Moving; Project. В этой системе, как и в большинстве остальных, активизация соответствующего класса приводит к уточнению первоначального запроса.

Большинство же из современных интеллектуальных систем обеспечивает группировку своих откликов по заранее определенным классификаторам. Так, система Vivisimo (`http://www.vivisimo.com`) определила для запроса "network" такие классы: Solutions; Games; Software; Security; Science; Organization; Film; Developers; Created; Information Network (рис. 2.3). Служба *Lycos* в режиме "Narrow Your Search" при этом определила такие классы: Carton Network; Dish Network; Food Network; Network Marketing; Home Shopping Network; Network Security. А система *Google* по этому же запросу выдала всего два класса: "Computers>Consultants>Network" и "Computers>Software>Operation System>Network".

"Сюжетный" подход

При поиске новостной информации всегда возникает задача нахождения и объединения в сюжетные темы документов, описывающих одни и те же события и ранжирования сюжетов по некоторым признакам, что должно обеспечить не только выявление самой важной темы, но и "веерное", многоаспектное освещение всех наиболее значимых событий.

Эта задача решается во многих системах, но с использованием различных подходов и алгоритмов. При этом неизменной остается технологическая цепочка: построение семантической сети из документов, кластеризация — автоматическое выявление наиболее взаимосвязанных групп (т.е. сюжетов), "взвешивание" этих сюжетов и наглядная визуализация самых важных из них.

При выделении сюжетных цепочек для определения попарной текстуальной близости текстов, как правило, используются алгоритмы выявления похожих документов, ставшие уже традиционными в поисковых системах. Так, матрица попарной близости документов обрабатывается алгоритмами кластеризации. Выделенные классы документов и представляют собой сюжетные цепочки (рис. 2.4).

Для предъявления пользователям сюжеты должны быть ранжированы. Основные факторы, влияющие на ранжирование по важности, — оперативность информации и размер сюжетной цепочки. Под оперативностью понимается некоторая функция от времени публикации всех сообщений в сюжете, а размер сюжета отражает общий интерес к конкретной теме. Во всех этих подходах центральная задача состоит в отождествлении сообщений, относящихся к одному сюжету, и выявлении "непересекающихся" сюжетов. Для результирующего отображения каждого отдельно взятого сюжета используются отобранные по содержательной близости документы из различных источников, отсортированные в хронологическом порядке.

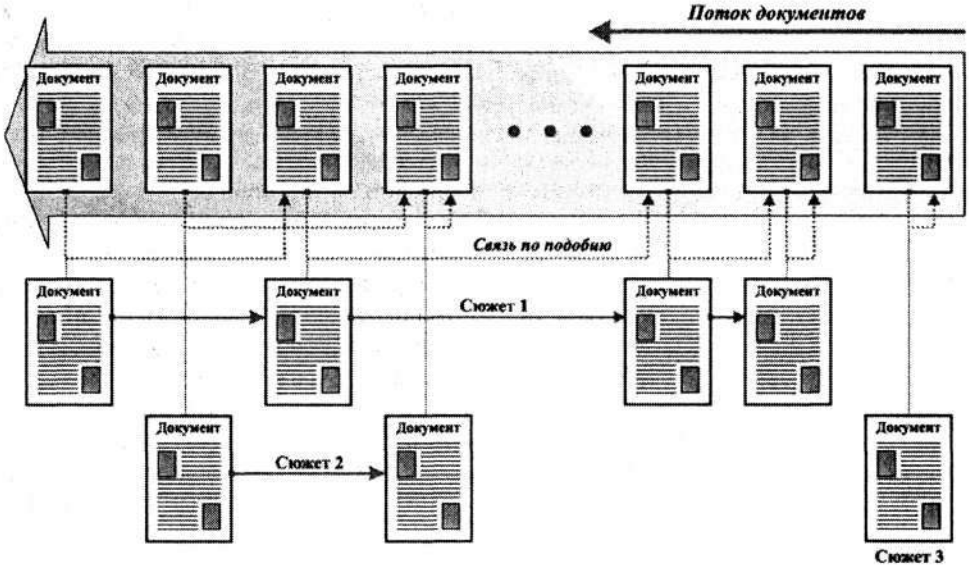


Рис. 2.4. Построение сюжетных цепочек

При этом сюжеты могут представлять собой дайджесты, интегрирующие общие места документов по теме, а также уникальную информацию, содержащуюся в отдельных документах. Реферирование сюжета в этом случае сводится не к *свертыванию* информации, а к построению *расширенной* версии, по сравнению с любым документом из сюжетной цепочки.

Например, в системе Яndex.Новости (`http://news.yandex.ru`) для этого строится матрица попарной близости документов, которая обрабатывается алгоритмом кластеризации с эмпирически подобранными параметрами (в частности, радиусом метрики близости). Для того чтобы увеличить связность крупных сюжетов, в системе Яndex.Новости дополнительно используется кластеризация второго уровня, обеспечивающая сбор атомарных кластеров в более крупные. В результате внедрения этой системы, все сообщения в результатах поиска на сайте Яndex.Новости сгруппированы по сюжетам (рис. 2.5), при этом ранжирование построено на стандартных для Яndex принципах ранжирования сгруппированной выдачи. Оно основано на числе и ранге новостей внутри новостных сюжетов, при этом ранг отдельной новости определяется как ее свежесть с учетом приоритетов текстуального совпадения.

В результате функционирования технологии выявления сюжетов, на сайте `www.yandex.ru` представлены пять главных новостей за последний час, а на сайте `news.yandex.ru` — новости с цитатными аннотациями, а также еще 10 новостей, упорядоченных по важности.

В системе InfoStream (`http://infostream.ua`) тематическая близость документов определяется на основе нормированных последовательностей наиболее весомых ключевых слов, входящих в каждый документ. Последовательности подобных (с определенным коэффициентом близости, превышающим некоторый установленный эмпирически уровень) документов образуют цепочки. При этом каждый документ попадает в какую-нибудь цепочку, даже состоящую только из

него самого. Затем цепочки "взвешиваются" по длине и оперативности, после чего пользователю предъявляется определенное количество самых важных тематических сюжетов. Для репрезентации сюжетной цепочки заголовки документов также "взвешиваются" относительно ключевых слов, соответствующих сюжету, а затем из всех заголовков выбираются наиболее весомые для отображения.

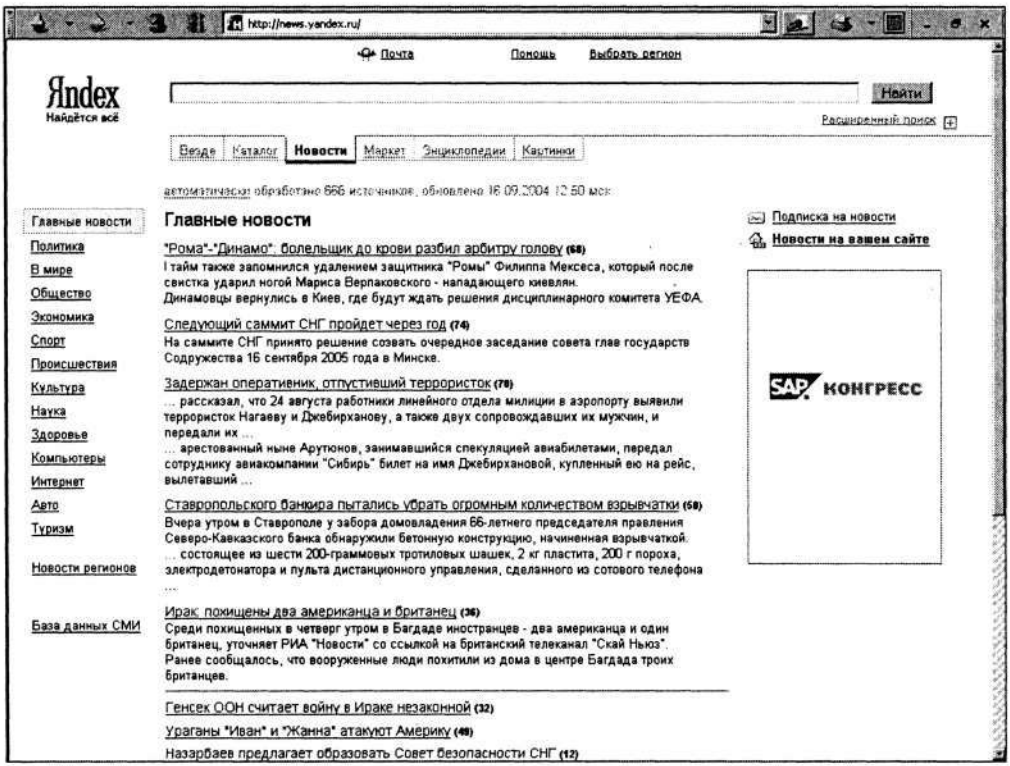


Рис. 2.5. Реализация сюжетных цепочек в системе Яndex.Новости

Следует обратить внимание, что задача автоматического построения качественных тематических сюжетов на основе потоков сетевой новостной информации сегодня практически решена. Например, полностью автоматические средства системы InfoStream, обрабатывая поток новостной информации, превышающий 25 000 документов в сутки, обеспечивают полноту свыше 80% и точность около 95%.

2.4. Этапы поисковой процедуры

Итак, как показано выше, процедура поиска имеет вполне определенную этапность — от определения информационной потребности и области поиска до анализа результатов и выбора пертинентных объектов. Приведем еще одну аналогию, которая относится к шахматному искусству. Начало шахматной партии — дебют — обеспечивает развитие фигур на доске и определяет стратегическую канву будущей партии. Несмотря на то что шахматы допускают миллиарды последова-

тельностей ходов, количество дебютов, на самом деле, ограничено несколькими сотнями. Точно так же, как в шахматном искусстве, в искусстве поиска можно определить первый этап — дебют. На этой фазе определяется цель поиска, его стратегия и область проведения (поисковые серверы, каталоги, тематические порталы).

Информационные потребности пользователя могут относиться к разным областям, которые могут быть как узкоспециализированными, так и достаточно типовыми. На практике основная часть информационных потребностей приходится именно на типовые области применения:

- поиск отдельных Web-страниц;
- поиск новостей;
- поиск людей и организаций;
- поиск литературных произведений;
- поиск программного обеспечения;
- поиск музыкальных произведений;
- поиск графических изображений;
- поиск видеоинформации;
- поиск коммерческой информации.

Вторым этапом в шахматах является миттельшпиль. При хорошо разыгранном дебюте и определенной стратегической направленности партии, наибольшее значение на этом этапе уделяется многовариантному анализу и тактическим решениям. В этом случае шахматист-профессионал просчитывает в уме несколько десятков вариантов (из миллионов возможных). Лишние неэффективные варианты он просто не рассматривает, руководствуясь логическими образами, заложенными на уровне подсознания.

Точно так же вторая, оперативная, часть поисковой процедуры предполагает многовариантность подходов и решений при формализации запросов в процессе их отработки. В этом случае также аналитик-профессионал приходит к необходимости использования весьма ограниченного числа поисковых серверов, каталогов и отдельных web-ресурсов для решения своей задачи.

Основной задачей второго этапа является формирование эффективных запросов к ИПС. Наибольшую проблему при формировании запросов представляет то, что на каждом поисковом сервере используется свой информационно-поисковый язык (ИПЯ), несмотря на то что у различных языков этого типа много общего, — например, схожий набор булевых операций. В настоящее время не существует единого стандарта, подобного стандарту языка SQL для СУБД, хотя на протяжении многих лет ведутся попытки такой стандартизации.

Последняя часть шахматной партии — эндшпиль — заключается в поиске вариантов при очень ограниченном количестве ресурсов (фигур). В этом случае количество вариантов, как правило, значительно более скромное, чем на втором этапе, и их правильный выбор определяет результат всей партии.

Точно так же третий этап поиска в сети Internet является определяющим, — от его реализации зависит, будет ли найденное решение пертинентно. На этом этапе пользователь работает с конечными документами, полученными в виде отклика ИПС. От правильного выбора набора документов-первоисточников зависит результат работы всех трех этапов поисковой процедуры.

Полученные в результате обработки запросов отклики ИПС требуют, с одной стороны, скрупулезной работы пользователей-аналитиков и, с другой стороны, развитых средств автоматизации аналитической работы, обеспечивающих:

- итеративное уточнение запросов;
- поиск по подобию;
- ранжирование выдаваемых документов;
- построение графических отчетов, визуализацию.

2.5. Процесс поиска непосредственно

Целью создания ИПС является предоставление пользователю возможности поиска информации по интересующей его тематике, выраженной специальными запросами. Различные ИПС имеют собственные языки запросов или, как их еще называют, информационно-поисковые языки (ИПЯ), позволяющие в той или иной мере описывать предметные области пользователей. Очевидно, что составление запросов должно базироваться на этих ИПЯ, однако сам процесс составления запросов допускает многовариантность и является своего рода искусством.

В качестве еще одного из аналогов процесса поиска в Internet можно рассмотреть сбор грибов в лесу во всей его этапности. Грибник, отправляясь за грибами, анализирует состояние погоды в определенное время года, климатическую зону и, в соответствии с этим, определяет, какие грибы можно найти. Он знает и свои потребности: какие грибы ему нужны, чтобы, например, их можно было засушить.

Точно так же при поиске в Internet следует четко определить информационные потребности, необходимую ретроспективу информации, круг поисковых серверов, специализирующихся на индексировании подобной информации, и даже предусмотреть заранее возможный результат, подобрав несколько известных документов сходной тематики. По приходу в лес грибник выбирает ту его часть, где могут расти те грибы, которые он предполагает собрать. Например, подосиновики следует искать в березово-осиновой роще, белые грибы — в дубраве или смешанном лесу, а маслята — в посадках молодого соснового леса. Точно так же пользователь Internet должен определить необходимые ему поисковые серверы и каталоги.

Грибник по знакомым ему образам определяет грибные места и практически интуитивно выходит на объект своего поиска. При этом он, конечно же, не формирует в явном виде запрос — поисковое предписание. Запрос содержится у него в подсознании, и составлен он на языке образов, хотя формально его можно сформулировать так: *"сосновый лес" и "солнечная погода" и "два дня назад прошел дождь" и "расстояние от дерева не более 5 м" и "восточная сторона" и т. д.*

Точно так же, выбрав необходимые поисковые ресурсы, пользователь Internet составляет поисковое предписание, соответствующее интересующей его тематике. Только при этом он осознанно формирует запрос на ИПЯ.

Когда грибник находит грибы, он их рассматривает, определяет их виды, выделяет требуемые ему, срезает и помещает в корзинку, при этом анализируя качество грибов и не оставляя у себя испорченных или червивых грибов даже полезных видов.

Пользователь Internet, анализируя отклик ИПС, выбирает ссылки на документы, которые, по его мнению, действительно соответствуют его информационным потребностям. Далее он выходит непосредственно на первоисточники, анализирует их и копирует себе только ту информацию, которая является наиболее полезной для него.

Как видим, процесс поиска в сети Internet имеет много общего с поиском в житейском понимании этого смысла, только на более высоком уровне виртуализации. Как и любой поиск, поиск в Internet является искусством, и ему, как и многим видам искусства, присуща многовариантность и творческий подход. Поиск в Internet можно рассматривать и с точки зрения его этапности.

2.6. Запросы пользователей

Казалось бы, с развитием технологических возможностей современные поисковые системы должны обеспечить гарантированное нахождение информации, однако "ленивые" пользователи все же очень часто недовольны качеством их работы. Основная масса пользователей не хочет прикладывать особых интеллектуальных усилий при формировании критериев поиска. Удивительно низким оказывается процент использования запросов, усложненных хотя бы одним логическим или контекстным оператором. Если и используются операторы, то это, в основном, булевы AND и OR. Доля использования операторов контекстной близости и логического отрицания (NOT) не превышает 1-2%. В то же время реализация отработки сложных запросов (которых пока не более 20%) и определяет эффективность использования времени, проводимого пользователем в Internet [22].

Согласно исследованию, проведенному OneStat.com в 2004 году (табл. 2.1), большинство поисковых запросов в Сети состоят из двух слов — 32,58% от общего количества. Из трех слов состоит 25,61% запросов и лишь 19,02% запросов состоит из одного слова. Для сравнения еще в апреле 2003 года процент поисковых запросов из одного слова составлял 24,76%.

Таблица 3.1. Распределение запросов по количеству слов, полученное аналитической службой OneStat

Количество слов в запросе	*Количество запросов в процентах*
1	19,02
2	32,58
3	25,61
4	12,83
5	5,64
6	2,32
7	0,98
8 и более	1,02

Эти результаты можно признать утешительными, так как они наглядно свидетельствуют о том, что человек в Internet постепенно становится все более разумным.

Среди поисковых запросов год от года преобладают все более сложные конструкции — чтобы найти что-то конкретное, пользователям приходится прибегать к все более сложным поисковым запросам.

Кроме того, очень большое значение имеет ранжирование результатов поиска, т.е. порядок следования документов, предъявляемых пользователю. Так, исследователи из IST, проанализировав характер свыше 450 тыс. запросов, выданных за сутки поисковой системе alltheweb.com, обнаружили, что пользователи чаще всего

просматривают первые три ссылки, полученные по запросу, очень быстро оценивают найденные сайты и еще быстрее разочаровываются в результатах. Другие исследования показали, что 75% пользователей удовлетворяются первыми 10–15 результатами поиска. И только 20% просматривают результаты на второй странице и менее 5% добираются до третьей и последующей страниц с результатами поиска.

Для ввода сложных запросов требуется использование булевых и контекстных операторов, скобок, указание полей и тому подобное, что недоступно для среднестатистического пользователя. Поисковые службы обычно создают два интерфейса — простой (по умолчанию) и расширенный (называемый в разных системах детальным, мощным или профессиональным), однако главная задача коммерческих поисковых служб как раз и заключается в удовлетворении информационных потребностей среднестатистического пользователя.

Назовем лишь некоторые возможности языков запросов наиболее популярных систем — возможности, которые есть в распоряжении пользователей, но которые используются в очень небольшой части. Во всех современных системах реализованы булевы операторы AND, OR и NOT, а также работа со скобками. Однако в двух из них — AltaVista и Excite — оператор NOT записывается в виде "AND NOT"; таким образом подчеркивается его бинарность (в математической логике оператор NOT в чистом виде унарный). В режимах простого поиска булевы операторы реализуются не всегда указанием их в явном виде. Например, во многих поисковых системах пробел между словами запроса по умолчанию воспринимается как оператор AND (Alltheweb, Google, META и UAport). В то же время при указании опций типа "any of the words" пробел в таких системах воспринимается как OR. Кроме того, в Alltheweb допускается использование операторов "+" и "-" перед словами фактически как синонимов операторов AND и NOT соответственно.

Создается впечатление, что при поиске очень редкого слова любая ИПС дает хороший результат. Действительно, если слово или словосочетание редкое, то документов выдается, как правило, немного, они все быстро просматриваются без особых затрат на аналитическую работу. Правда, все усложняется, если система не находит ни одного документа. В этом случае следует обратиться к другим системам или изменить критерии поиска.

В свое время Google затеяла сетевую игру, смысл которой состоит в том, что игроки называют словосочетания из двух осмысленных слов, которые встречаются в Internet точно один раз. Тысячи энтузиастов принялись отыскивать такие словосочетания. Как только словосочетание "срабатывало", человек объявлялся победителем на текущий момент, и его имя публиковалось на сайте вместе со словосочетанием. В результате это словосочетание встречалось в Сети уже два раза — один раз в оригинале и второй раз в таблице победителей.

Иное дело, когда на, казалось бы, логично сформулированный запрос выдается тысяча документов, имеющих слабое отношение к информационным потребностям. В этом случае рекомендуется применить два метода: первый — кардинальный — полностью переформулировать запрос, изменив представление о возможном поисковом образе, второй — уточнить запрос с помощью добавления еще одного условия с применением операции конъюнкции (оператора логического "И"). Второй путь реализуется в большинстве систем опцией "искать в найденном". В этом случае, не изменяя логики предыдущего запроса, а лишь уточняя его, можно добиться удовлетворительных результатов, например, если словосочетанию "стол деревянный" соответствуют 500 откликов, то уточнение "обеденный" приведет к двум десяткам документов.

2.7. Поиск подобных документов

Рассмотрим случай, когда в результате поиска по запросу найдено избыточное количество документов, но при просмотре первых страниц результатов поиска найдено несколько пертинентных документов. Естественно, у пользователя возникает желание найти еще документы (или ссылки на них), сходные с ними по содержанию, не затрачивая интеллектуальных усилий на анализ и составление запроса.

В результате многие ИПС реализовали опции "найти подобное", "find similar", "like this". Однако этот режим не всегда ведет к удовлетворительным результатам при целевом поиске, но иногда приводит к получению полезных документов, имеющих косвенное отношение к теме первичного запроса. Что означает "подобный документ" и по каким критериям это определяется, зачастую остается загадкой для пользователя. Один из подходов к ее решению может быть таким: каждое значимое, по мнению системы, слово ранжируется по какому-то критерию, из наиболее весомых слов автоматически формируется запрос, рассматриваемый как новый критерий поиска. Такой режим реализован во многих современных ИПС, например, на серверах Excite, Google и Яndex.

2.8. Ранжирование откликов

Ранжирование выдаваемых документов, в отличие от предыдущей опции, имеет большое значение в работе современных ИПС. Инструменты повышения пертинентности в современных системах, помимо возможностей уточнения формулировки запросов, предусматривают использование весовых критериев, что позволяет ранжировать найденные документы и выдавать пользователю для просмотра наиболее весомые документы либо вообще ограничиваться выдачей не более заданного числа наиболее весомых документов. Следует отметить, что в современных системах проблеме релевантности, а особенно пертинентности, уделяется все большее внимание. Яркий пример — служба Google, которая реализовала алгоритмы достижения неформальной релевантности, благодаря чему в настоящее время стала самой популярной системой в Internet.

Ранжирование выдаваемых документов может выполняться по дате создания/обновления документа, по степени важности (многие системы оценивают важность документов по весовым критериям или по количеству ссылок на них, т.е. по цитированию). Ранжирование по дате имеет особое значение при поиске новостных сообщений средств массовой информации и информационных агентств.

Ранжирование по индексу цитирования, аналогичное оценке значимости научных публикаций в традиционной научной среде, впервые ввела Google, продемонстрировавшая эффективность такого подхода для Web-пространства.

2.9. Поиск по словам и словоформам

Остановимся подробнее на особенностях формирования запросов к поисковым системам.

Все поисковые системы обеспечивают поиск хотя бы по одному слову. Средства навигации в Internet, не обеспечивающие такого поиска, называются каталогами, коллекциями ссылок и т.п.

Иначе дело обстоит с усечениями слов. Некоторые системы рассматривают все слова запроса как правые усечения. У других известных систем возможность

поиска по усечениям попросту не реализована (Google, Alltheweb, Рамблер). Однако в большинстве систем для маскирования правого усечения слова достаточно поставить символ "*" (AltaVista, Яndex).

Некоторые системы не чувствительны к регистрам букв в словах запросов. К таким системам относится Alltheweb, Google и UAport. При этом система UAport не делает различий даже между буквами одинакового написания латыни и кириллицы, что в некоторых случаях упрощает ввод запросов. Однако в большинстве приведенных выше систем "чувствительность" к регистрам включается при употреблении хотя бы одной прописной буквы в слове запроса.

Поиск по словоформам является результатом серьезного лингвистического анализа и реализован в русскоязычных системах Апорт, Яndex и Рамблер, а также в украинской системе МЕТА. К примеру, в системе Апорт, независимо от того, в какой грамматической форме указано слово в запросе, оно находится в базе данных во всех своих формах. В этой системе запрос "ребенок шел" эквивалентен запросу "дети идут".

В системах Яndex и Рамблер, если слово участвует в запросе, учитываются также все его формы. Для поиска по конкретному слову, а не всем словоформам, перед ним ставится символ "!" (Яndex) или оно берется в кавычки (Рамблер).

Портал Рамблер встроил лингвистическую поддержку украинского языка в свою поисковую машину, в результате пользователям стали в полной мере доступны Internet-ресурсы на украинском языке. Если раньше запросы на украинском языке Рамблер понимал буквально, то сейчас он способен осуществлять морфологический анализ запросов, определять грамматическую форму слова и выдавать корректный результат.

Поисковая машина Рамблер распознает язык поискового запроса и формирует адекватную выдачу. Например, для запроса "поисковая система" основным языком является русский, так что документы, содержащие формы соответствующего украинского слова ("системі" и "системою"), в процессе поиска рассматриваться не будут. По запросу же "пошукова система", наоборот, не будут учитываться документы, содержащие слова "системы" и "системе". Предполагается, что при поиске по запросу на украинском языке преимущество будут получать сайты, популярные среди украинских пользователей Internet.

2.10. Логические операторы

Для ввода сложных запросов требуется использование булевых и контекстных операторов, скобок, указание полей и т.п. Хотя для большинства случаев (по статистике 70% запросов состоят из одного слова) этого не требуется. Поэтому поисковые службы обычно создают два интерфейса — простой (по умолчанию) и расширенный (называемый в разных системах детальным, мощным или профессиональным). Но есть и такие системы, которые с помощью одного и того же механизма позволяют вводить, а затем обрабатывают простые и сложные запросы, обеспечивая пользователей руководствами различного уровня сложности.

Во всех современных системах реализованы булевы операторы AND, OR и NOT, а также работа со скобками. Однако в двух из них — AltaVista и Excite — оператор NOT записывается в виде AND NOT, что подчеркивает его бинарность (в математической логике оператор NOT в чистом виде является унарным). В режимах простого поиска булевы операторы реализуются не всегда указанием

их в явном виде. Например, во многих поисковых системах пробел между словами запроса по умолчанию воспринимается как оператор AND (Allthenews, Google, META и UAport). В то же время при указании опций типа any of the words, пробел в таких системах воспринимается как OR. Кроме того, в Alltheweb допускается использование перед словами операторов + и - фактически как синонимов операторов AND и NOT соответственно. Точно так же используются эти операторы в AltaVista, Excite, Lycos и Апорт. Можно отметить, что у самой популярной сегодня системы Google — самый лаконичный набор логических операторов: +, OR и –.

2.11. Операторы контекстной близости

Большинство профессиональных поисковых систем обеспечивает выполнение операций контекстной близости, одна из реализаций которой — поиск выражений в кавычках.

Например, в системе Google реализована только возможность поиска по фразам в кавычках, в AltaVista реализован оператор NEAR (~), обеспечивающий нахождение документов, у которых два слова находятся на расстоянии не более 10 слов. В системе Lycos функции контекстной близости получили наибольшее развитие и реализованы с помощью четырех операторов: ADJ, NEAR, FAR и BEFORE. Оператор ADJ обеспечивает близость двух слов в тексте в любом порядке, а оператор NEAR позволяет находить документы, в которых слова-операнды удалены не более, чем на 25 слов. FAR — оператор, противоположный по смыслу оператору NEAR, т.е. он исключает близость терминов запроса в пределах 25 слов текста документа, а оператор BEFORE похож на оператор ADJ, только с учетом порядка встречаемости терминов в тексте. Оригинально решен вопрос контекстной близости в системе Рамблер. Значение ограничения контекста в этой системе можно изменять конструкцией (*число*, *запрос*), где *число* — любое положительное число, а *запрос* — любой корректный запрос, состоящий более чем из одного слова. Таким образом, по запросу (2, красная роза) будут найдены только те документы, в которых между словами "красная" и "роза" хотя бы раз не стоит ни одного слова. В системе Яndex режим контекстного поиска называется "поиском с расстоянием". В общем виде ограничение по расстоянию задается в строке данных выражением вида */(n m)*, где *n* — минимальное, а *m* — максимальное допустимое расстояние. В системе Апорт существует два вида ограничения по расстоянию: в словах w*N*(...), где *N* — число слов, и в предложениях s*N*(...), где *N* — число предложений. В этой системе также подвергаются интеллектуальной обработке выражения в кавычках. Например, запрос "яблоки на снегу" эквивалентен запросам "яблоки и снег", "яблоки под снегом", "яблоко снег".

Большинство из названных систем способно реализовать контекстный поиск заключенной в кавычки фразы (Google, Alltheweb, AltaVista, Lycos и др.). Такая способность — это реализация неявно указанных с помощью кавычек операторов контекстной близости.

2.12. Поиск по параметрам

Отдельного рассмотрения заслуживает возможность поиска по параметрам документов, которая позволяет ограничивать диапазон поиска значениями URL, датами, заглавиями и т.п. Чаще всего получить такую возможность можно из

режима расширенного поиска, в котором для ввода значений отдельных параметров предлагается весь диапазон возможностей Web-интерфейса.

Например, в системе Alltheweb в запросах можно указать параметры, обеспечивающие поиск по таким элементам:

- URL — url: (например, по запросу url:energ будут найдены документы, в URL которых имеется строка "energ");
- ссылки на страницы сайтов — link:;
- доменные имена — site: (например, site:ua обеспечит нахождение документов из украинского домена);
- заголовки — title:.

В этой системе допустим поиск, кроме всех вариантов текстовых файлов, еще трех типов файлов — PDF, MS Word, Flash.

В системе AltaVista имеются все приведенные для Alltheweb возможности (параметру site: в AltaVista соответствует host:); кроме того, в режиме расширенного поиска обеспечивается поиск по датам (с явным указанием диапазона поиска либо с указанием типа "искать за последние 8 месяцев"). Этот режим в системе традиционно называется "Web-археологией".

В Google обеспечивается поиск по сайту (site:), определение ссылок на сайт (admission site:), поиск по ценам, например DVD player $250..350, странам, датам, доменам и т.д. В поле ввода запроса можно вводить и арифметические выражения, используя интерфейс Google как калькулятор, что, конечно же, подчеркивает своеобразность данной системы (например, по запросу 4^2 будет выведен результат 16).

Тернистый путь прогресса

Синтаксис запросов к популярным поисковым системам в последнее время значительно упростился. Вместе с тем, качество откликов постоянно улучшается, несмотря на лавинообразный рост ресурсов Сети.

Традиционные подходы к поиску, основанные на использовании логических операторов, потерпели крах одновременно с бумом Web-технологий. Первые скрипки в поисковых системах стали играть не инструменты индексирования баз данных и организации логического поиска, а новые семантические алгоритмы. Можно признать, что пионером в этом стала компания Google, сделавшая ставку на ранжирование выдачи и алгоритмы, основанные на цитируемости.

Незавидна роль традиционных систем искусственного интеллекта в этой "семантической революции". Системы, основанные на базах знаний, в большинстве своем не выдержали силы потока Internet-информации. При этом речь идет не столько об объемах, сколько о политематичности и динамике, т.е. о постоянном обновлении информации, которое, к тому же, не имеет очевидной тематической направленности и регулярности.

При этом возник новый класс систем, который все же позволяет справляться с проблемой "размерности" Сети. Как один из удивительных феноменов, сегодня можно рассматривать тот факт, что содержательные, семантически наполненные результаты формируются без непосредственного привлечения методов искусственного интеллекта, объемных баз знаний и даже экспертов как таковых, лишь путем использования частотно-лингвистических и эвристических методов. И сегодня эффективно работают в основном системы, базирующиеся именно на таких методах.

2.13. Популярные сетевые информационно-поисковые службы

Безусловно, для обеспечения полноты поиска необходимо знать степень охвата информационных ресурсов Internet поисковыми системами. Сегодня ведущими по охвату информационных ресурсов Internet являются поисковые системы Google и Alltheweb. Вместе с тем, даже эти системы охватывают всего лишь третью часть существующих Web-страниц. Количество поисковых серверов, охватывающих Internet, а не отдельные его части, ограничено несколькими десятками. Лидерами здесь являются такие поисковые машины, как:

- `http://www.google.com`
- `http://search.yahoo.com`
- `http://www.ask.com`
- `http://www.alltheweb.com`
- `http://www.altavista.com`
- `http://www.lycos.com`

Среди российских поисковых серверов особого внимания заслуживают три — это Яndex (`http://www.yandex.ru`), Рамблер (`http://www.rambler.ru`) и Апорт (`http://www.aport.ru`). В Украине две лидирующие поисковые системы: META (`http://meta.ua`) — по стабильной части украинского сегмента Сети, и UAport (`http://uaport.net`) — по новостной части.

2.13.1. Крупнейшие зарубежные службы

Google

В январе 1996 года будущие основатели Google, студенты Сергей Брин и Ларри Пэйдж, начали совместную работу над поисковой системой под названием BackRub. В сентябре 1998 года ими была основана компания Google. Название поисковой системы Google было образовано в результате игры букв в слове "googol". Этим компания хотела подчеркнуть свое намерение индексировать и обрабатывать большие объемы информации.

К 2000 году служба Google заняла лидирующее положение на рынке сетевых поисковых систем; трафик к ней непрерывно растет в течение шести лет. В 2002 году Google на короткое время отдала первенство по объему поискового индекса системе Alltheweb, но в настоящее время вновь заняла устойчивое первое место, охватывая свыше 4 млрд документов, и осуществляет более 200 млн поисковых операций в день. Поисковая машина Google позволяет искать как без учетов специфики алфавитов и языков, так и с учетом особенностей свыше 97 языков (рис. 2.6).

Компания является лидером поискового рынка во всем мире. В США ее предпочитают 34,7% пользователей, тогда как в мире доля Google на рынке англоязычного поиска достигает 43,3%. Большинство пользователей службы находятся за пределами США. Самым близким преследователем Google является компания Yahoo!, до недавнего времени также применявшая поисковую технологию Google, но в начале 2004 года сменившая ее на собственную систему.

Сегодня 95% всех поисковых операций в Сети в США осуществляется через эти две компании, Google и Yahoo!, либо напрямую, либо через другие сайты, использующие их технологию. Множество компаний используют поисковую технологию Google в своих сервисах, например Интернет-провайдер America Online и российский холдинг Mail.ru.

Рис. 2.6. Google — самая популярная сетевая поисковая система в мире

Google позволяет проводить поиск в таких сегментах, как обычные Web-документы, изображения, телеконференции Usenet, новости, а также в собственном каталоге.

Очень удобной функцией является *cache*. Благодаря этой функции пользователь может просмотреть проиндексированную страницу, даже если она удалена или сервер, на котором расположена страница, недоступен. Так, в середине 2002 года правительство КНР временно запретило доступ китайских пользователей к Google именно из-за наличия этой функции, поскольку система в полном объеме предоставила контент сайтов, зафильтрованный по политическим мотивам.

Yahoo! Search

Традиционно служба Yahoo! позиционировалась как развитый каталог Web-ресурсов. Однако в апреле 2003 года, после поглощения компании Inktomi и приобретения Overture, компания Yahoo! стала обладательницей всех основных поисковых технологий на рынке, кроме технологии Google. В результате

этой компании принадлежат такие поисковые службы, как Inktomi, Altavista и FAST. Сама Yahoo! на базе технологий Overture и Inktomi разработала глобальную поисковую систему Yahoo! Search (`http://search.yahoo.com`) и прекратила использование на своем сайте поисковой системы основного конкурента — Google. (Примечательно, что Yahoo! останется одним из главных акционеров Google: ей принадлежит около 2,4% бизнеса конкурента.)

В новую поисковую систему (рис. 2.7) встроены функции по работе с информационными каналами в форматах XML/RSS. Кроме того, Yahoo! Search обладает уникальными технологиями борьбы со спамом, с помощью которых осуществляется фильтрация избыточных ссылок и поискового мусора. Помимо Web-страниц, с помощью Yahoo! Search возможен поиск изображений, новостей, товаров. Естественно, возможен поиск и в собственном каталоге Yahoo!. Компания активно использует поисковую систему для привлечения дополнительных доходов путем платного занесения ссылок в базу данных, размещения контекстной рекламы в результатах поиска, а также лицензирования поисковых технологий.

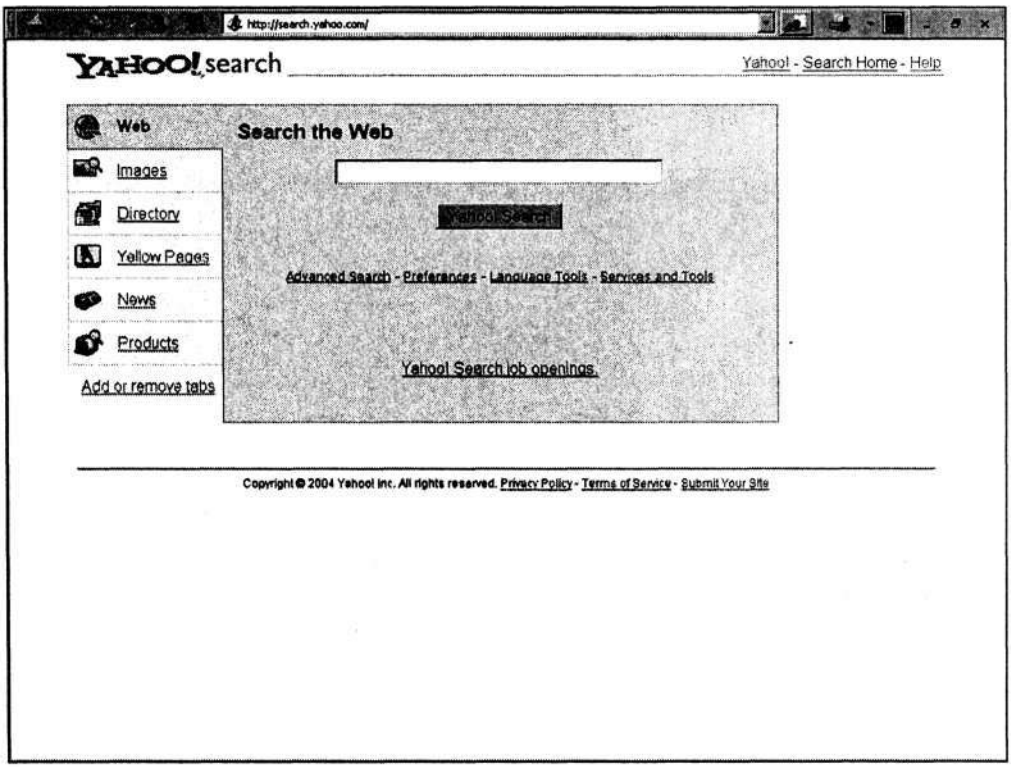


Рис. 2.7. Новая поисковая система от Yahoo!

Коммерческая инициатива Yahoo! — Content Acquisition Program — позволяет владельцам сайтов добиться более полного индексирования своих ресурсов в Yahoo! Search и более оперативного обновления информации в поисковой базе данных. Content Acquisition Program предполагает включение информации в базу данных на платной основе.

Само собой разумеется, что от бесплатного включения сайтов в базу данных поисковой системы Yahoo! также не отказывается. По словам вице-президента Yahoo! по поиску Тима Кадогана (Tim Cadogan), 99% ресурсов в базе данных Yahoo! Search будут индексироваться бесплатно. Однако никаких гарантий в случае бесплатного индексирования компания Yahoo! дать не может.

Ask Jeeves

Поисковая система Ask Jeeves (`http://www.ask.com`) — одна из лидирующих в области информационного поиска — обслуживает более 16 млн пользователей в месяц. Главной особенностью Ask Jeeves считается способность работать с запросами на естественном языке (рис.2.8). Кроме того, в результаты поиска попадает не только информация из автоматически обновляемой базы данных, но и ссылки, подобранные вручную.

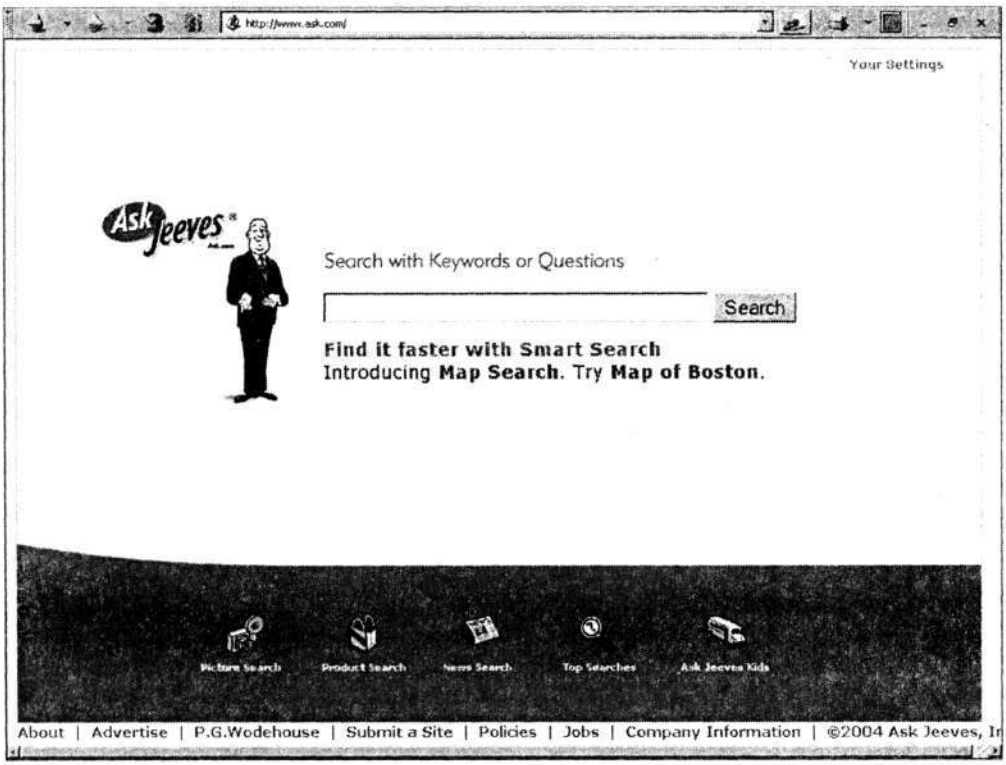


Рис. 2.8. Ask.com — один из самых популярных поисковиков в США

Система Ask Jeeves способна распознать некоторое количество популярных вопросов пользователей, таких, например, как расшифровка аббревиатур, указания о том, как добраться в определенное место, даты праздников и т.д. Для поиска информации на эти темы Ask Jeeves предоставляет специализированные интерактивные средства. В частности, если спросить о дате какого-либо праздника, то на первом месте среди результатов окажется именно ответ на данный вопрос. Запросы для поиска карт, изображений или новостей также распознаются автоматически. Очень часто, вместо того чтобы переходить на вспомогательные

страницы поиска с помощью закладок, достаточно набрать запрос на естественном языке. Например, в ответ на запрос map of Russia (карта России) на странице результатов выводятся сначала картинки, а затем уж и обычные ссылки.

В свое время компания Ask Jeeves приобрела компанию Teoma (http://www.teoma.com) и внедрила у себя одноименную поисковую технологию. Поисковая технология Teoma использует в качестве критерия релевантности принцип, аналогичный используемому в Google, — число ссылок на данный ресурс с других страниц. Однако Teoma при этом учитывает еще и тематику отдельных сайтов, ссылающихся на данную страницу, что обеспечивает большую точность и избирательность поиска.

Вице-президент управления производством компании Ask Jeeves Джим Ланзоне так охарактеризовал коммерческую деятельность службы: "Наша компания извлекает прибыль из размещения рекламы. Существуют три основных способа размещения рекламы: графический (рекламные баннеры), оплачиваемое размещение (спонсорские ссылки) и новая программа, получившая название «Paid Inclusion» (плата вносится за включение сайта в индекс поисковой системы)".

Недавно поисковая система Ask Jeeves объявила о внедрении локального поиска, лицензировав соответствующие базы данных у компании CitySearch. Компания Ask Jeeves является партнером Google как поставщик оплаченных рекламных ссылок.

По информации Nielsen NetRatings по состоянию на июнь 2004 года, как отмечает SearchengineJournal, Ask Jeeves является девятым по популярности сайтом в американской части Internet с количеством уникальных посетителей более 32 млн.

Alltheweb

Поисковая служба Alltheweb была основана в Норвегии в 1997 году компанией Fast Search and Transfer. В 1999 году в результате партнерства с компанией Dell был создан поисковый сервер http://www.alltheweb.com (рис. 2.9). В 2002 году на некоторое время была достигнута главная цель поисковой службы — создана самая большая база данных Web-документов объемом свыше 2 млрд записей. Но позже первенство все же было упущено.

Сегодня поисковая технология Alltheweb, получившая название Fast, считается наиболее близкой по своим возможностям к Google, признанной лидером среди сетевых ИПС. Alltheweb отличается высокой скоростью, время ее ответа на поисковый запрос не более 0,05 секунд. Система Alltheweb обеспечивает поиск Web-документов, новостей, изображений, видео, аудио, файлов на FTP-серверах.

В начале 2003 года компания Overture Services Inc., специализирующаяся на размещении рекламы в результатах поиска, приобрела службу Alltheweb у компании Fast Search and Trans. Сама же компания Overture с октября 2003 года принадлежит Yahoo!.

AltaVista

Служба AltaVista (http://www.altavista.com или http://www.av.com) появилась в 1995 году и вначале принадлежала компании Digital Equipment (прежний адрес службы http://www.altavista.digital.com). После этого AltaVista перешла компании Compaq, а затем выделилась в отдельную фирму. В апреле 2003 года была куплена компанией Overture Services Inc., принадлежащей в настоящее время Yahoo!. Преимущество этой системы — развитые, мощные средства сложного поиска. Одно из самых слабых мест системы — недостаточная

актуальность ее базы данных. Со страниц сервера, кроме поиска HTML-файлов, возможен поиск графических изображений, музыкальных произведений в формате MP3, видеоклипов, а также текущих новостей (рис. 2.10).

Рис. 2.9. *Лаконичный интерфейс поисковой системы Alltheweb*

AltaVista была и остается одной из самых популярных поисковых служб. "Мы считаем, что AltaVista — это по-прежнему очень сильный брэнд, у которого есть своя группа пользователей", — заявил вице-президент по инжинирингу Yahoo! Пху Хоанг. По словам Хоанга, AltaVista рассматривается как своеобразный полигон для испытания новых поисковых технологий. По его словам, "применение найдется и для других поисковых систем, ставших в последнее время собственностью Yahoo!".

Lycos

Информационно-поисковый сервер Lycos (`http://www.lycos.com`) существует с мая 1994 года и является старейшей, представленной в Internet поисковой системой с широким кругом пользователей. Основатели службы Lycos — Carnegie Mellon University и Lycos Inc. Базовая страница Lycos, кроме интерфейса самой поисковой системы, содержит около 20 справочников (рис. 2.11). Предусмотрен режим расширенного поиска с использованием операций И, ИЛИ и НЕ при выборе зоны поиска (заглавие, адрес, ссылка, весь текст), а также работы со словосочетаниями. В системе имеется возможность поиска в HTML-документах, новостях и в электронных магазинах.

Рис. 2.10. Классическая система поиска в Web — AltaVista

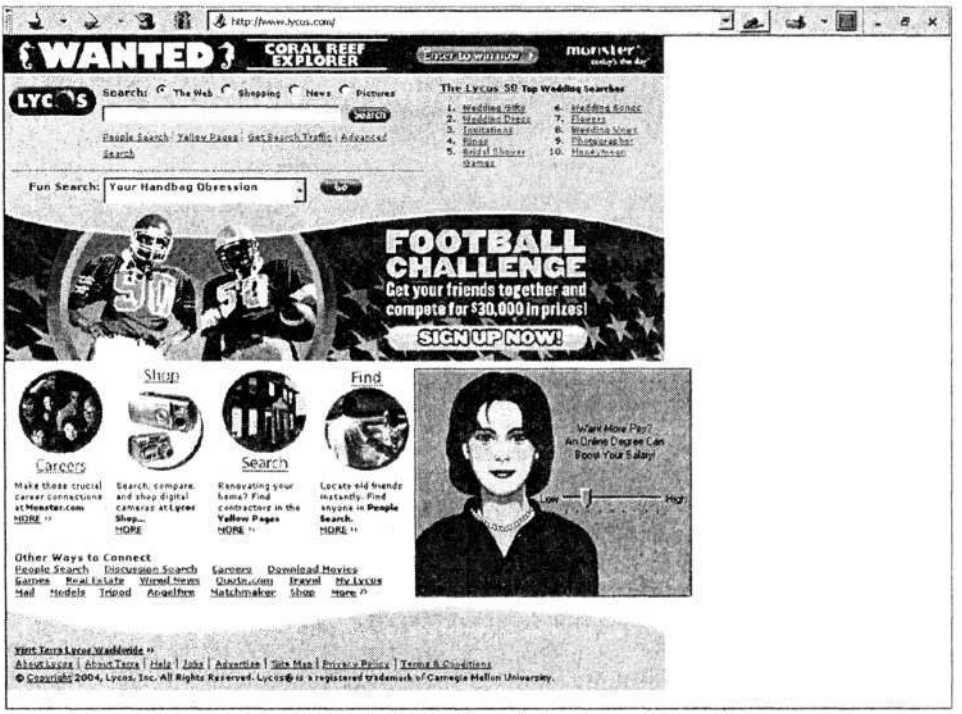


Рис. 2.11. Поисковый портал Lycos

30 октября 2000 года, на пике бума "доткомов", служба Lycos была куплена испанской компанией Terra Networks. В августе 2004 года служба была перекуплена корпорацией Daum, специализирующейся на коммуникациях и смежных сферах и содержащей самый крупный Internet-портал в Южной Корее.

2.13.2. Службы поиска в российском сегменте Сети

Яndex

Летом 1996 года руководство и разработчики этой поисковой системы пришли к выводу, что развитие самой технологии важнее и интереснее, чем создание прикладных продуктов на базе поиска. Исследования рынка показали своевременность и большие перспективы поисковых технологий. Тогда в Internet и появился "Яndex". Поисковая машина Yandex.Ru была анонсирована компанией CompTek в сентябре 1997 года. Слово "Яndex" было придумано за несколько лет до этого и означает "Языковой index", или, если по-английски, "Yandex" — "Yet Another indexer".

Основными особенностями Yandex.Ru были и остаются проверка уникальности документов (исключение копий в разных кодировках), учет морфологии различных языков, поиск с учетом расстояния, оценка релевантности (рис.2.12).

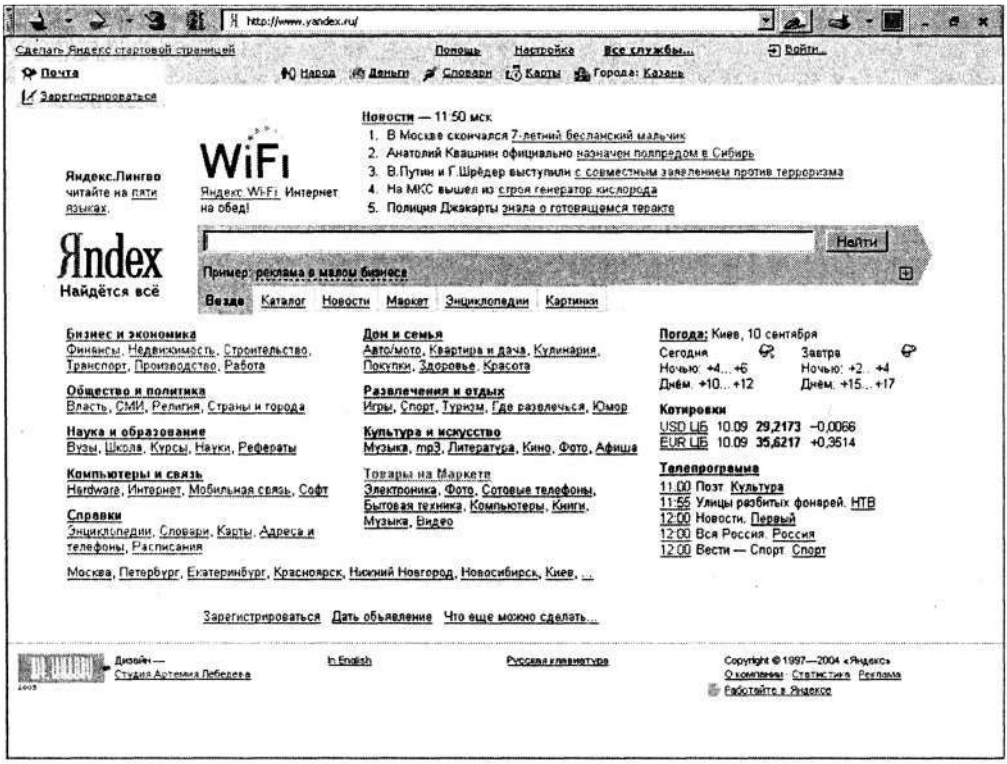


Рис. 2.12. Яndex — лидер в русскоязычном сегменте Сети

Сегодня Яndex представляет собой полнотекстовую поисковую систему, обеспечивающую поиск в таких сегментах русскоязычного Internet, как Web-документы, изображения, товары и услуги (маркет), новости, собственный каталог.

В марте 2004 года в Яndex были реализованы новые поисковые возможности. По словам разработчиков, система теперь может учитывать социальную структуру Сети — она умеет отличать мнение людей от технической, вспомогательной и рекламной информации, т.е. лучше распознавать, какой ресурс является авторитетным в своей области. Яndex автоматически определяет, в каком городе находится компьютер, с которого поступил запрос, и, если уточнение по региону имеет смысл, предлагает повторить поиск, ограничив его сайтами данного региона. В системе реализована очистка результатов поиска от дубликатов. Пользователь избавлен от повторения в списке найденного почти одинаковой информации. Поиск поддерживает шесть языков: к русскому и английскому добавились украинский, белорусский, французский и немецкий. Язык документов и сайтов определяется автоматически, а ограничить область поиска нужным языком можно в настройках или при расширенном поиске.

Летом 2001 года сайт Яndex, согласно данным исследовательских компаний Комкон-2 и Gallup Media, стал самым большим ресурсом в Рунете по объему аудитории. 5 ноября 2002 года компания Яndex вышла на самоокупаемость.

Сегодня ежедневная аудитория Яndex (включая зарубежных пользователей) составляет около двух миллионов человек, ежемесячная — более 12 миллионов.

Rambler

В 1996 году программист Дмитрий Крюков написал уникальную российскую поисковую систему для ресурсов Internet, которая сразу же была введена в эксплуатацию по адресу `http://rambler.ru`. Со временем был образован Internet-холдинг Rambler, который в настоящее время занимает одну из ведущих позиций в России. Поисковая машина Rambler работает с учетом морфологии русского и английского языков, сама определяет тематику запросов (например, запрос, в котором упомянут "амидопирин" или "клиника", автоматически распознается как "медицинский").

В июне 2003 года компания Rambler запустила новую версию поисковой системы, которая отличается высокой скоростью поиска и оперативностью обновления индекса. Rambler понимает живой язык, опознает общепринятые сокращения и аббревиатуры (рис. 2.13).

В настоящее время поисковая машина Rambler реализует полноценную лингвистическую поддержку уже трех языков — русского, английского и украинского. В системе реализован механизм ассоциаций, который помогает пользователям быстрее и точнее формулировать свои запросы. Так, после выполнения поиска по запросу пользователя перед ним открывается страница, на которой найденные документы расположены в порядке убывания релевантности, а также появляется строка "У нас также ищут". В этой строке приведено несколько слов и словосочетаний, ассоциативно связанных с исходным запросом. Например, на слово "релевантность" в строке "У нас также ищут" выдаются результаты "толковый словарь, словарь релевантность, словарь иностранных слов, релевантный ... еще". Если щелкнуть мышью на слове "еще", отображается целый блок ассоциаций, более развернутый. Щелкнув на любом слове из списка и уточнив тем самым запрос, пользователь может продолжить поиск.

Апорт

Поисковый сервер Апорт, принадлежащий Golden Telecom Inc., появился в Сети в 1996 году. В октябре 2000 года официально был представлен "Апорт 2000". Оригинальной особенностью системы Апорт является учет "ранга страницы", который характеризует ее популярность. Он вычисляется по количеству ссылок на

ресурс со страниц других Web-сайтов. Обработка запроса при этом ориентируется на гипертекстовую структуру Сети. Это — реальный пример использования коллективного разума владельцев отдельных Web-сайтов. Система ранжирования результатов поиска Page Rank учитывается с весовыми коэффициентами: вес ссылки с популярного сайта выше, чем вес ссылки с менее популярного, т.е. разработчики Апорта удачно использовали некоторые идеи, впервые реализованные в системе Google. В конечном итоге при выдаче результатов поиска в Апорт одними из первых выдаются сайты, название которых в службах реальных имен является синонимом со словами запроса или совпадает с ними (рис. 2.14).

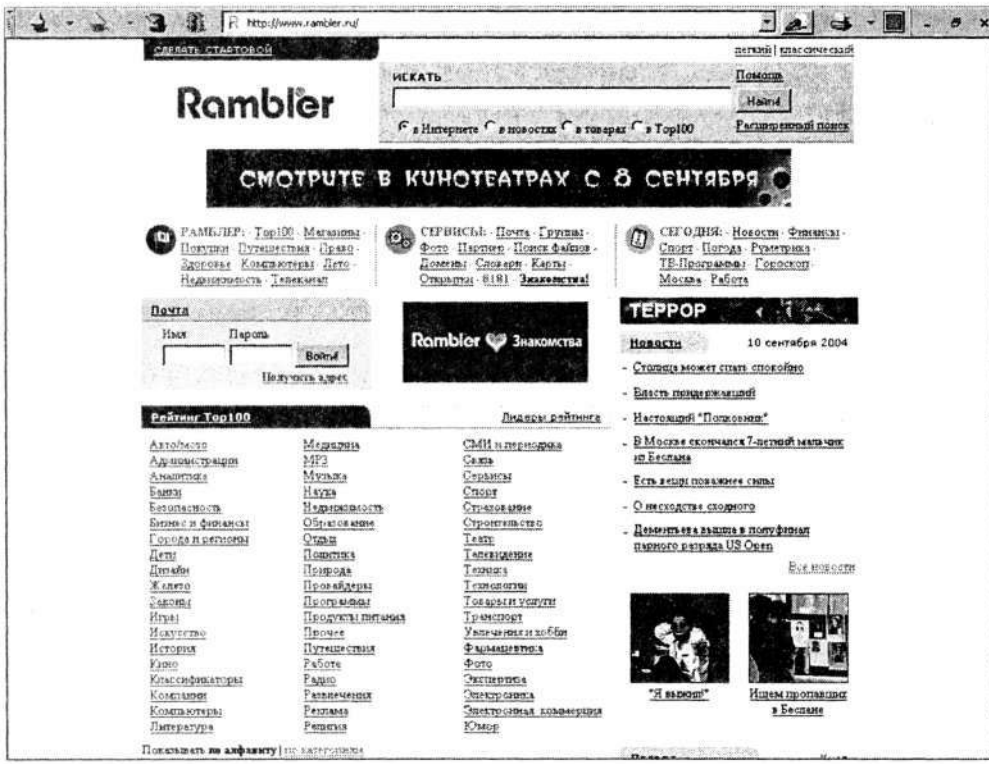


Рис. 2.13. Rambler — ветеран российского поиска

Поиск в системе Апорт осуществляется в таких сегментах Internet, как Web-сайты, рефераты, товары, работа, знакомства, MP3, новости, энциклопедия "Кругосвет".

У системы Апорт есть ряд ключевых особенностей — в качестве результатов поиска она предоставляет не разрозненный набор страниц с разных сайтов, а достаточно осмысленный их список, причем часто — с названием и описанием.

2.13.3. Крупнейшие украинские службы

META

Украинская поисковая система META (`http://meta.ua`) была основана в 1998 году. С 2001 года META работает на новом поисковом ядре, создание которого стало возможным благодаря участию в проекте компании SigmaBleyzer.

Рис. 2.14. Апорт — поисковая система и каталог

Алгоритмы вычисления меры соответствия документов запросу в системе META учитывают не только количество слов в документе, но и частоту этого слова во всем обрабатываемом пространстве документов, близость и порядок слов, различные признаки форматирования. В системе META не используется технология "стоп-слов". Разработчиками предполагается, что это приближает систему к обработке запросов на естественном языке. Например, при запросе "крем от загара" большинством поисковых систем предлог "от" не будет учитываться при поиске, и в первых результатах будут выданы документы со словосочетанием "крем для загара". Система META обеспечивает возможность поиска с учетом закономерностей изменений русских и украинских слов. Результаты поиска в META могут быть представлены как в традиционной форме, так и сгруппированными по сайтам (рис. 2.15).

Система META позволяет искать по таким сегментам, как Web-сайты, новости, реестр (каталог системы), прайс-листы, рефераты и книги.

UAport

Интернет-холдинг UAport (`http://uaport.net`), созданный в 2001 году, объединил основные сетевые проекты компании ElVisti. UAport полностью включил в свой состав поисковую систему ElVisti (`http://el.visti.net`), первую из украинских поисковых систем, которая была представлена в Internet с 1997 года (рис. 2.16).

В качестве программного ядра в UAport используется полнотекстовая информационно-поисковая система InfoRes, обеспечивающая поиск с учетом логических операторов и оператора контекстной близости (с возможностью задания расстояния между отдельными словами).

Рис. 2.15. META — поиск в украинском сегменте Internet

Рис. 2.16. UAport — Интернет-холдинг и поисковый портал

В Интернет-холдинге UAport получили развитие и новое современное представление самые популярные поисковые службы ElVisti, которые стали основой формирования таких разделов:

- Net.UAport.net — информационно-поисковая система по украинским Web-ресурсам;
- Каталог.UAport.net — тематический и региональный каталоги Web-ресурсов;
- Новости.UAport.net — раздел, в котором благодаря возможностям технологии InfoStream(r) доступна (в том числе и в формате RSS) новостная лента объемом свыше 20 000 сообщений в сутки из сотен информационных источников;
- Медиа.UAport.net — раздел, в котором в свободном доступе представлены информационные материалы украинских СМИ;
- ИТ.UAport.net — раздел информационных технологий — "вертикальная" информационно-поисковая система по тематике информационных технологий;
- Бизнес.UAport.net — основой данного раздела является поисковый прайс-каталог по товарам и услугам, охватывающий данные свыше 15 000 фирм.

2.14. Поиск информации в корпоративных сетях

На жестких дисках отдельных компьютеров или на серверах корпоративной сети накапливаются огромные массивы документов, навигация в которых по понятным причинам затруднена. Для обеспечения комфорта работы с такими массивами документы обычно пытаются классифицировать, распределить по тематическим папкам или каталогам [18]. Эта процедура трудоемкая и, что самое главное, не исключает возможности внесения дополнительных ошибок.

Понятно, что создать информационную среду, инкапсулирующую разнородные объекты, непросто. Естественным выходом из этой ситуации оказались полнотекстовые информационно-поисковые системы, получившие широкое распространение в Internet. В отличие от Сети, где данные в основном представлены как HTML-файлы, поиск в корпоративной сети производится в другой среде [5]. Ведь в этом случае преимущественно используются форматы офисных приложений и систем документооборота. Наряду с поиском в корпоративной сети, большое значение приобретают задачи группировки тематически близких документов, автоматического реферирования, перевода, выявления ключевых понятий, проведения нечеткого поиска.

2.14.1. Популярные ИПС

Рассмотрим некоторые популярные системы поиска для корпоративных сетей.

mnoGoSearch

Универсальная поисковая система mnoGoSearch (`mnogosearch.org`) предназначена для Internet- или intranet-серверов. Она индексирует информацию, которая сканируется по локальным дискам или в соответствии с протоколами HTTP, FTP, NNTP. Система работает с документами в форматах `.html`, `.txt`,

.doc, .pdf. В запросах воспринимаются различные формы слов и логические операторы. Результаты запросов можно настраивать с помощью html-шаблонов. Система mnoGoSearch может хранить данные во всех популярных реляционных СУБД. Существуют версии для Linux и Windows (рис. 2.17).

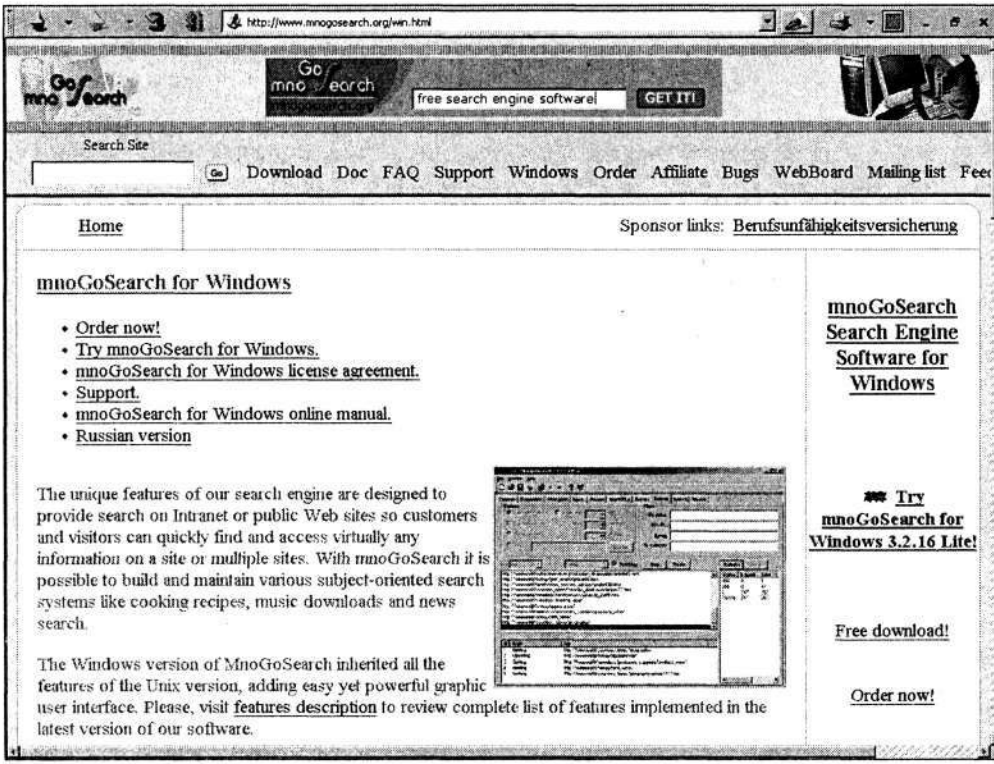


Рис. 2.17. Поисковая система mnoGoSearch

"Ищейка"

Полнотекстовая персональная поисковая система "Ищейка" (http://www.isleuthhound.com) обладает возможностями поиска документов и файлов на русском и английском языках (рис. 2.18). Она воспринимает запросы со всеми словоформами и с любыми падежными окончаниями (т.е. поддерживает морфологический поиск) и способна автоматически распознавать основные типы кодировки текста — ASCII, ANSI, Unicode. В "Ищейке" заложена возможность просмотра краткой выдержки (аннотации) из найденного документа. Предполагается работа с документами форматов .txt, .rtf, .doc, .html.

При первом запуске на основе заданного массива документов, "Ищейка" создает и индексирует базу данных, которая представляет собой зону поиска, состоящую из каталогов. В пределах этой зоны и производится поиск документов и файлов.

Система допускает организацию собственных хранилищ данных из неструктурированной информации, создание до пятидесяти зон поиска с индексированием неограниченного количества файлов, накопление "популярных" запросов и т.п.

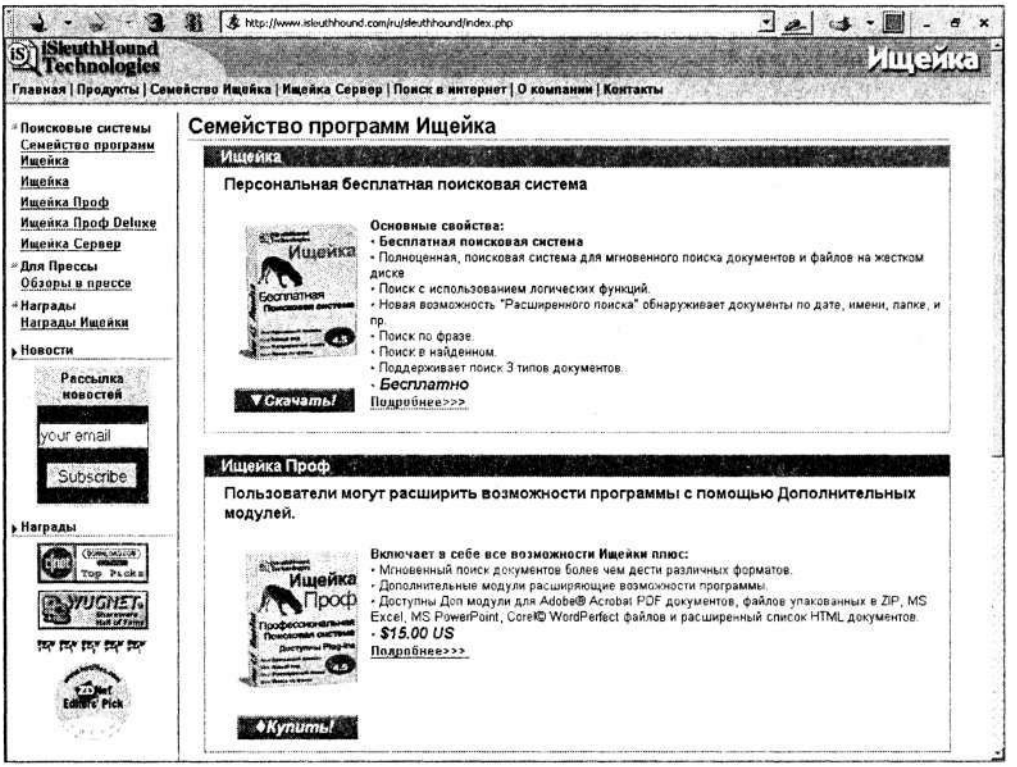


Рис. 2.18. Web-страница системы "Ищейка"

Серверный "Следопыт"

Серверный "Следопыт" (www.medialingua.ru) — мощная поисковая система, предоставляющая возможность поиска нужной информации на отдельном Web-сайте или сервере корпоративной интрасети (рис. 2.19). Поиск осуществляется по содержанию документов и их атрибутам, а также по размеру, имени, дате создания, по отправителю или получателю почтового сообщения. Программа может обрабатывать файлы практически всех форматов: .doc, .rtf, .html, .xls, .pdf, .zip, .pst, а также папки (как сами сообщения, так и вложения) Microsoft Outlook. В системе реализован морфологический поиск, т.е. для каждого слова учитывается вся парадигма. Фильтр для формата .pdf при работе с русским языком является в "Следопыте" одним из лучших.

Полнотекстовый поиск под Microsoft SQL Server 2000 в "Следопыте" реализован для русского и английского языков (подразумевается возможность динамического отслеживания изменений в базе данных и обновления полнотекстового индекса Change Tracking, которая появилась в Microsoft SQL Server 2000).

Data Search

Основное назначение программы Data Search 6.0. (www.dtsearch.com) — поиск информации на локальном компьютере (рис. 2.20). Система имеет английский интерфейс и работает под управлением операционных систем Windows 9x/Me/NT/2000. Она состоит из следующих модулей: dtSearch Desktop 6.0 —

Рис. 2.19. Серверный "Следопыт"

главный интерфейс программы, dtSearch Indexer — индексатор документов, dtSearch Index Library Manager — менеджер библиотек индексов, dtSearch CD Wizard — индексатор данных, находящихся на CD. Data Search позволяет создавать один общий индекс для нескольких компьютеров в локальной сети.

Система поддерживает поиск документов разных типов, включая `.zip`, `.rtf`, `.pdf`, `.html`, `.xml`, документы Microsoft Office (Word, Excel, PowerPoint) и WordPerfect. Поддерживается кодировка Unicode. Допускается несколько видов поиска, а именно морфологический и фонетический, а также поиск синонимов и слов с орфографическими ошибками.

CROS

Система полнотекстового поиска CROS 4.01 (`www.cronos.ru`) предназначена для накопления и обработки текстовых документов различных форматов (рис. 2.21). Хранение документов в базах данных системы обеспечивает уменьшение в два-три раза необходимого объема дисковой памяти. Предусмотрено автоматическое определение форматов документов Microsoft Word версий 6.0, 7.0, 97, 2000, а также документов формата `.rtf` и `.html`. Помимо этого определяется тип кодировки (DOS, Win, КОИ8, Unicode).

Система CROS обеспечивает навигацию по найденным документам, способна работать в локальной сети и поддерживает защиту информации от несанкциони-

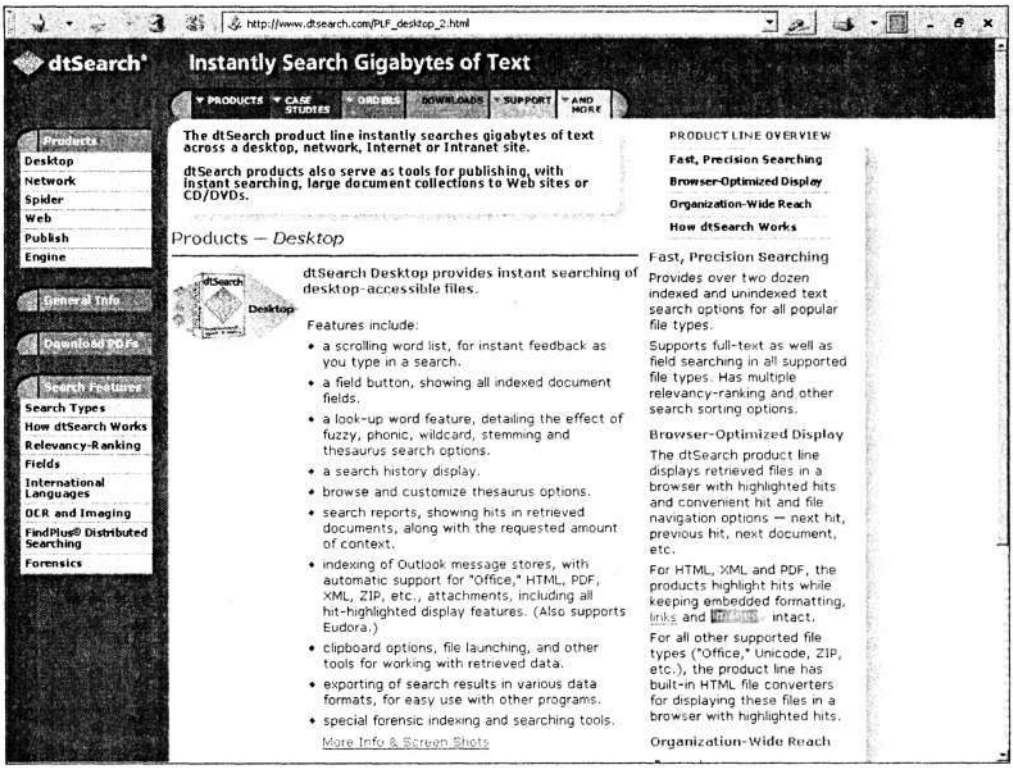


Рис. 2.20. Система поиска на локальном компьютере Data Search Desktop

рованного доступа. При этом отсутствуют ограничения на количество иерархических областей поиска, осуществляется сортировка найденных документов по дате, имени, типу и атрибутам, которые задаются самим пользователем.

Greenstone

Система Greenstone (`www.greenstone.org`) представляет собой Open Source-решение для создания "цифровых библиотек", поддерживаемое ЮНЕСКО (рис. 2.22). Естественно, она включает поиск с предварительным индексированием по документам всех популярных форматов и, прежде всего, `.doc` и `.pdf`, которые могут быть представлены и в заархивированном виде. Система создает каталог документов, конвертирует их в html-формат, а затем обеспечивает удаленный доступ к библиотеке посредством браузера.

Google Search Appliance

Программно-аппаратный комплекс Google Search Appliance обеспечивает поиск документов в рамках корпоративных сетей. Джон Пискителло, менеджер Google по продуктам, определил эту систему как "естественный шаг для компании, которая постоянно стремится предложить пользователям новые способы доступа к информации". По его словам, пришлось учитывать возрастающие требования, включая поиск в границах, определенных корпоративными межсетевыми экранами, и это заставило Google разработать новые решения.

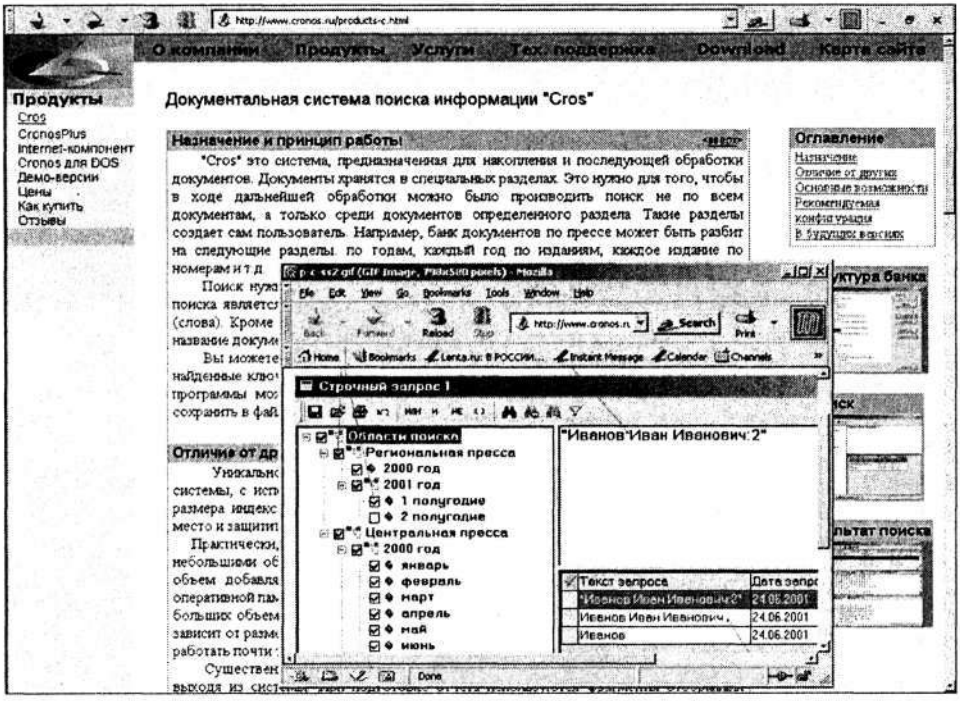


Рис. 2.21. Система полнотекстового поиска CROS

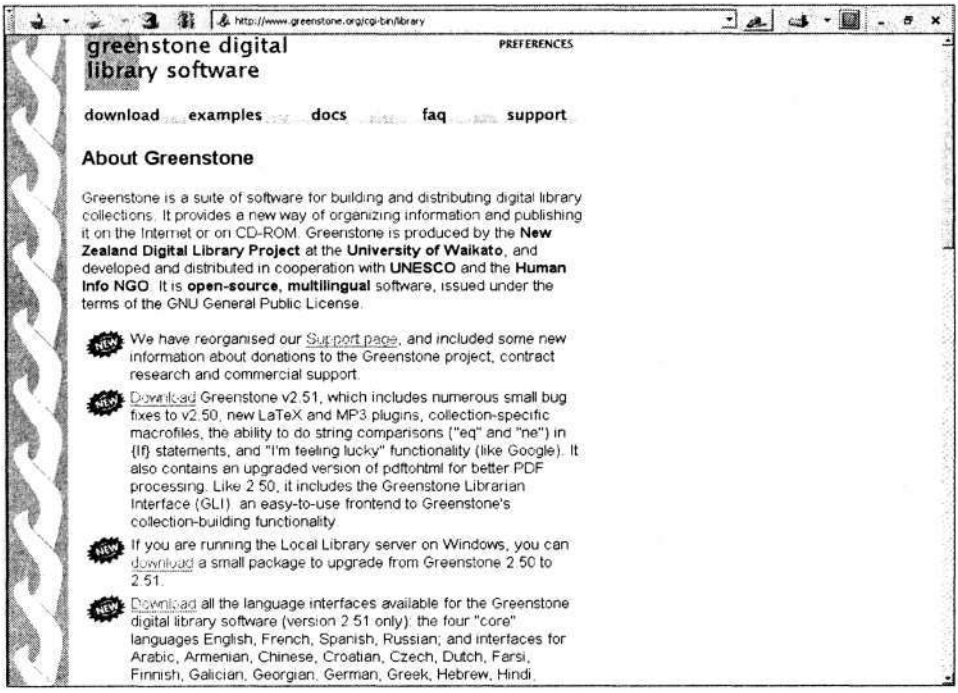


Рис. 2.22. Greenstone — бесплатное UNIX-решение

Поисковые устройства этой компании используют в своей работе армия США, администрация калифорнийского города Сан-Диего, фармацевтический гигант Pfizer, корпорации Boeing, Procter & Gamble, Cisco Systems и др.

Поисковый механизм комплекса обеспечивает работу более чем с двумястами типами файлов (естественно, включая .html, .pdf, .doc). При этом осуществляется учет синонимов при полнотекстовом поиске по запросам и возможна работа более чем с пятьюдесятью естественными языками.

Google Search Appliance поддерживают функции поиска защищенной информации, находящейся на закрытых серверах. При этом пользователь может обратиться к защищенному документу лишь при наличии у него соответствующих полномочий доступа.

2.14.2. Новый уровень обработки сетевой информации

RetrievaWare

Информационно-поисковая система RetrievaWare (www.convera.com) представляет собой средство полнотекстового и атрибутивного поиска (рис. 2.23). К документам, с которыми способна работать система RetrievalWare, относятся тексты в различных форматах и кодировках, электронные таблицы, базы данных, почтовые сообщения и т.п. — всего более двухсот форматов. Система обладает дополнительным инструментарием, позволяющим настроиться на поддержку документов специфических форматов. Объем архива при необходимости может измеряться терабайтами.

Архитектура системы RetrievalWare позволяет ей работать как через локальную корпоративную сеть, так и через Internet. Серверная часть системы поддерживает все распространенные серверные платформы, а клиентским местом может быть любой компьютер, имеющий графический Web-браузер. Система обладает возможностью работы в различных многопроцессорных и распределенных многосерверных конфигурациях.

Попытки анализа больших объемов неструктурированных или слабо структурированных данных очень часто усложняют процесс принятия решений. Если широкий спектр поисковых систем достаточно легко справляется с "простым" полнотекстовым поиском, то для подобного анализа нужны технологии совсем другого типа, представленные системами извлечения знаний (Knowledge Mining). Стоимость внедрения таких систем составляет сотни тысяч долларов.

Итак, основная задача — выявление знаний в массивах неструктурированных данных с целью их использования в процессе принятия решений. Чтобы добиться этого, необходимо сделать информацию доступной для анализа, выявить классы понятий и сопоставить их с документами.

Как правило, информационные массивы преобразуются такими системами в хранилища данных (Data Warehouse) или корпоративные порталы знаний — интегрированные информационные репозитарии, доступные для оперативного обобщения и анализа. Часто такие хранилища являются самообучаемыми за счет использования статистических байесовских алгоритмов. Последние обеспечивают адаптацию критериев группирования документов. Большую роль играют и "отклики" реальных пользователей.

За счет предварительной обработки информации, проводимой на этапе формирования хранилищ данных, значительно повышается эффективность таких

процессов, как интеллектуальный анализ данных, глубинный анализ текстов и обнаружение новых знаний в текстах. Как неожиданную производную этих процессов можно назвать появление средств, упрощающих поиск для пользователя, таких как реализация нечеткой логики запросов (нечеткого поиска), средств построения функциональных информационных портретов, визуализации семантических связей и т.д. В свою очередь, эти возможности напрямую связаны с распознаванием образов, поиском мультимедийных данных, анализом речевого ввода.

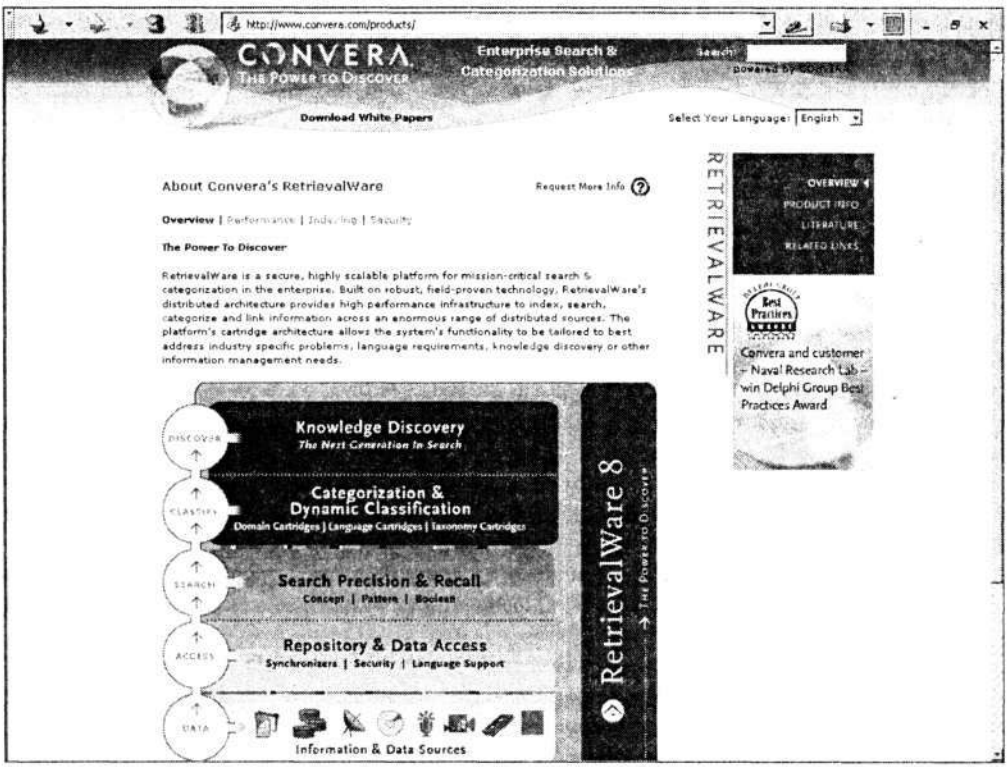


Рис. 2.23. RetrievaWare компании Convera

Яndex.Server

Яndex.ServerStandard 3.2 (`http://company.yandex.ru/technology/products/yandex-server.xml`) представляет собой системный сервис для организации полнотекстового поиска информации в заданной коллекции документов. Он предназначен для работы с текстами как в локальной, так и в глобальной сети. Система не содержит лицензионных ограничений на число индексируемых документов, их размер или суммарный размер индекса и позволяет индексировать документы как через HTTP-соединение, так и чтением локальной файловой системы. Яndex.Server Standard представляет результаты поиска во встроенном дизайне.

Яndex.Server 3.0 состоит из двух основных логических частей: индексатора и поискового сервера. Индексатор анализирует документы, среди которых должен проводиться поиск, и сохраняет информацию о них в специальных индексных файлах.

Обычно используется режим работы, при котором не создаются заново индексные файлы, а отрабатывается информация только по изменившимся, новым и удаленным документам. Поисковый сервер после запуска находится в постоянном ожидании запросов, которые могут быть представлены на естественном языке. Поиск может осуществляться с учетом морфологии языка, в одной или нескольких коллекциях документов.

Яndex.Server 3.2 поддерживает форматы `.html`, `.xml`, `.rtf`, `.pdf`, `.doc`, `.mp3` и многие другие. Содержимое индексируемых документов также может быть получено при обращении к произвольной базе данных, в частности MySQL и MS SQL Server.

Система предоставляет возможность кластеризации результатов поиска (группирует найденные документы в соответствии с внешними атрибутами), а также ранжирует результаты (сортирует документы по степени соответствия запросу).

InfoStream

Поиск в корпоративной сети, реализуемый на UNIX-платформах, выполняется с помощью корпоративного решения на основе технологии мониторинга контента InfoStream (`infostream.com.ua`). Эта технология позволяет обрабатывать данные в форматах Microsoft WORD (версии 2000, 97, 6), `.rtf`, `.pdf` и всех текстовых форматах (простой текст, `.html`, `.xml`). Системы на основе InfoStream в настоящее время функционируют под управлением операционных систем FreeBDS, Linux и Solaris.

На основе InfoStream создана система управления документальным информационным хранилищем, в котором реализуется интегрированная информационно-поисковая среда на основе Web-решений. С ее помощью обеспечивается доступ к электронным документам, размещенным на компьютерах в корпоративной сети, в режимах поиска, навигации по компьютерам/каталогам, просмотра как оригиналов документов, так и их текстовых образов. Комплекс обеспечивает интерактивный полнотекстовый поиск информации по сложным запросам, состоящим из ключевых слов, логических и контекстных операторов, разнообразное ранжирование результатов поиска. Предоставляется возможность уточнения результатов поиска с помощью механизма "информационных портретов".

2.14.3. Порталы знаний

По данным недавно проведенного исследования, сотрудники компаний могут тратить до трех часов в день на поиски информации, которые зачастую оказываются безрезультатными, вследствие чего тысяча крупнейших фирм США ежегодно теряет 2,5 млрд долларов.

Именно для решения этой проблемы созданы и продолжают создаваться корпоративные поисковые системы и порталы знаний [3] как среды для эффективного поиска знаний и обмена ими, инструменты, которые представляют собой совокупность технологических решений для выявления, хранения, классификации, обработки и распространения знаний.

В настоящее время широко используется система IBM Lotus Discovery Server — программный продукт, предназначенный для управления знаниями в корпоративных порталах, для нахождения экспертов, идентификации связей и общего управления интеллектуальным капиталом (рис. 2.24). Lotus Discovery Server является логическим продолжением ранее популярного программного продукта Lotus Raven — системы построения корпоративных порталов знаний.

Благодаря возможности анализа информации, хранящейся в организации, Lotus Discovery Server в состоянии указывать области экспертных знаний и подразумеваемые знания сотрудников, находя и организуя динамические связи между информацией, людьми и их деятельностью.

Рис. 2.24. Lotus Discovery Server — инструментарий порталов знаний

Современные порталы знаний [9] обеспечит решение целого комплекса задач, среди которых — сбор информации об объектах, определение связи между объектами, выявление тенденций. Функциональные возможности таких систем позволят проводить многофакторные динамические исследования, выполнять диагностику и прогнозирование развития ситуации. В дополнение к возможностям глубинного анализа данных и текста, в порталах знаний широко используется человеческий опыт — знания экспертов в процессах выявления, сохранения и эффективного использования знаний.

Около пяти лет назад по заказу группы аналитиков Гарвардского университета российские разработчики из "Инфорус" создали систему Avalanche, которая в процессе поиска формирует модель предметной области в виде набора "умных папок", каждая из которых знает, что в нее должно попадать. Наполнением папок занимается специализированный робот, который запускается с компьютера "хозяина" и приносит только то, что у него просили. Это одно из первых эффективных решений на базе современной технологии глубинного анализа текстов.

Очень близка по идеологии и технология компании Vivisimo, в рамках которой результаты Internet-поиска распределяются по папкам-категориям, которые система создает автоматически. Достигается это за счет лексического сопоставления запросов и результатов поиска.

Естественно, свое применение Vivisimo сразу же нашла в корпоративных сетях и порталах знаний. Рауль Валдес-Перес (Raul Valdes-Perez), один из учредителей Vivisimo, сравнил систему с очень умным библиотекарем, который мгновенно находит нужную книгу в море неупорядоченной информации.

2.15. Поисковые программно-аппаратные комплексы

Многим корпоративным пользователям необходим оперативный доступ к полным базам данных определенных информационно-поисковых систем, отвечающих их информационным потребностям, что требует создания специального механизма локального копирования (кэширования) баз данных ИПС. Одной из первых эту ситуацию почувствовала известная своим поисковым сервисом американская компания Google, поисковый механизм которой заинтересовал ряд корпоративных пользователей. Выйдя на рынок с аппаратным поисковым сервером Google Search Appliance [46], компания стала пионером в новой области — создании кэширующих информационно-поисковых серверов. Устройство Google Search Appliance предназначено для подключения к сетям предприятий и реализует функции поиска информации как внутри этих сетей, так и в Internet. Это небольшое сетевое устройство (недорогой сервер), оснащенное программным обеспечением Google, позволяет находить на корпоративных серверах различные документы, начиная от сообщений электронной почты и заканчивая программными кодами. Оно позволяет находить документы HTML, PDF, PostScript, Microsoft Office и еще приблизительно двухсот других форматов.

В числе иных особенностей Google Search Appliance отмечаются функции кэширования поисковых страниц, сервера-посредника, группировки результатов поиска, поддержки 28 языков и метатегов. Кроме того, сервер обладает достаточно широкими возможностями администрирования. По заявлению компании, особенностью Google Search Appliance является весьма гибкая настройка поиска, благодаря которой заказчики могут задавать его параметры в соответствии со своими потребностями. Google Search Appliance предлагается в двух моделях — GB-1001 для малых и средних фирм (от 20 тыс. долларов; охват до 150 тысяч документов) и GB-8008 для крупных корпораций (250 тыс. долларов; "просматривает" миллионы документов (рис. 2.25)).

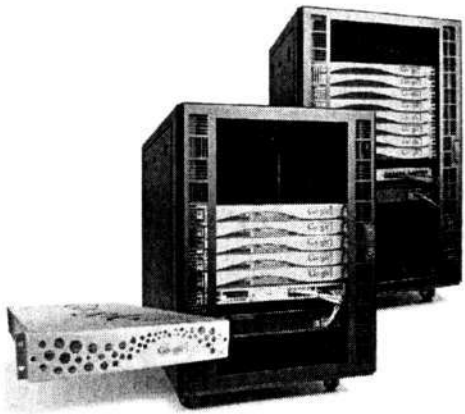


Рис. 2.25. *Линейка моделей Google Search Appliance*

Компания Google со своим аппаратным решением вышла на рынок, на котором уже имеются со своими программными системами такие компании, как Verity, Ask Jeeves и Altavista.

Google отличается от них тем, что предлагаемая ею поисковая система — это не только программа, но и устройство, которое может устанавливаться за корпоративный брандмауэр и которое можно настроить на поиск документов во внутренней базе данных.

Google сразу же смогла продать несколько своих устройств, причем среди первых его заказчиков числится корпорация National Semiconductor. Поисковая машина выпускается в двух версиях. Для сравнения, аналогичная продукция Altavista стоит от 30 тысяч долларов, и такая корпоративная поисковая система способна просматривать от 30 тысяч документов и, теоретически, до бесконечности.

Еще одной известной информационно-поисковой системой, реализованной в виде аппаратного решения, способного хранить в своем кэше свыше миллиона документов, является ThunderStone Search Appliance (последняя версия 5.0., `http://www.thunderstone.com`). Эта система позволяет хранить и индексировать данные, получаемые по протоколам HTTP, HTTPS, FTP, Gopher или просто из файлового сервера локальной сети.

Рис. 2.26. *Одноюнитовый сервер ThunderStone Search Appliance*

В Украине в Информационном центре "ЭЛВИСТИ" разработана технология автоматического мониторинга новостей в Internet InfoStream. Эта технология обеспечивает сканирование информации в режиме реального времени из нескольких сотен источников — украинских и зарубежных Web-сайтов сети Internet.

Для корпоративных пользователей на основе технологии InfoStream построено аппаратно-программное решение InfoStream Port, которое обеспечивает доступ к базам данных оперативной и ретроспективной информации в корпоративной сети. Программно-технологическое обеспечение InfoStream Port включает как компоненты утилиту обмена данными с информационным хранилищем (кэшем) ElVisti и полнотекстовую информационно-поисковую систему InfoReS. Информационное хранилище способно хранить и обеспечивать интерактивный доступ к более чем 10 млн документов, размещенных на одноюнитовом сервере Prime LAN 1900 на базе процессора Intel Pentium IV (рис. 2.27).

Информационное обеспечение системы у корпоративного заказчика строится на основе использования информационного кэша, формируемого на технической площадке провайдера. Система InfoStream Port работает по такой схеме:

- информация в соответствии с регламентом поступает из кэша информационного провайдера ElVisti на сервер InfoStream Port;

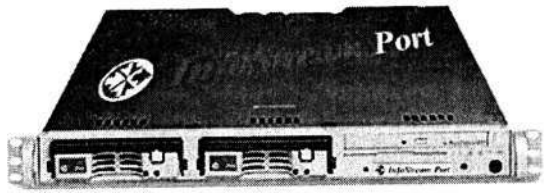


Рис. 2.27. Корпоративное решение InfoStream Port

- на сервере происходит формирование и индексирование оперативных и ретроспективных баз данных;
- со стороны корпоративных пользователей обеспечивается доступ к этим базам данных через Web-интерфейс.

Благодаря высоким поисковым характеристикам, оперативности доступа к информации со стороны корпоративного пользователя в сочетании с невысокой ценой (менее 10 тыс. долларов), это решение является полезным инструментом в работе информационно-аналитических служб.

Глава 3

Системы интеграции Internet-контента

Internet представляет собой гигантское хранилище информации, объем которой удваивается каждый год. По экспертным оценкам, количество новостей только в украинском и российском сегментах Internet превышает 100 тыс. сообщений в сутки.

Очевидно, что такое разнообразие информации может быть полезным лишь при эффективном доступе к ней, что оказывается не просто осуществить на практике. Как уже упоминалось, по оценкам экспертов, около 79% журналистов обращаются к Internet в поисках новостей и лишь 20% из них находят необходимую информацию.

Попробуем проанализировать, почему традиционный поиск в Сети может оказаться неэффективным, и рассмотрим существующие уже решения, справляющиеся с подобной задачей гораздо лучше.

3.1. Статическая и динамическая составляющие Web-пространства

Как уже было сказано, все Internet-пространство можно с достаточной долей условности разделить на две составляющие — стабильную и динамическую. Стабильная составляющая Сети содержит информацию "долговременного" плана, в то время как динамическая включает постоянно обновляемые ресурсы. Некоторая часть динамической составляющей со временем вливается в стабильную, однако большая ее часть "исчезает" из Сети или попадает в сегмент "скрытого" Web-пространства, не доступного пользователям с помощью информационно-поисковых систем.

В традиционной сетевой поисковой системе информационное пространство, состоящее из стабильной и новостной частей и индексируемое этой ИПС, меняет свое содержимое через *N* дней: некоторые новостные документы уходят в стабильную часть в виде архивов, а остальные исчезают.

В этом случае пользователь при обращении к ИПС получает соответствующие запросу ответы из стабильной части, устаревшие ссылки из новостной части и ничего из обновленной новостной части.

Пользователь часто часами проводит время в Сети, посещая сотни сайтов с целью получения данных по определенной тематике. Ведь ни одна из традиционных поисковых систем в достаточном объеме не помогает в поиске актуальной новостной информации, находящейся в динамической части Сети.

Решение этой задачи требует создания своеобразного интеллектуального посредника между пользователем и Internet. Подобный посредник (или агент новостей) должен выполнять всю "черновую" работу по сбору и селекции информации и обеспечивать предпосылки для создания документальной базы данных.

Принцип индексирования, используемый посредником, несколько отличается от аналогичного принципа традиционных поисковых систем: индексируется не все пространство Internet, а только его новостная часть.

При этом, за счет относительно небольшого объема этих данных, частота индексирования выбирается достаточно малой — от нескольких минут до нескольких часов (в зависимости от источника).

В результате через *N* дней обрисовывается такая ситуация: пользователь получает необходимые ответы по новостной и "устаревшей" новостной части, подтвержденные документами из собственной архивной базы данных, но не получает полной выборки документов из стабильной части информационного наполнения Сети.

Таким образом, проблема получения полной информации из Сети в идеале может быть решена путем использования двух инструментов — традиционных ИПС и системы агентов новостей.

3.2. Недостатки традиционного поиска

Владельцы Web-сайтов уже давно осознали, что новостная информация привлекает посетителей, а потому количество источников такого рода информации в Сети постоянно возрастает, что само по себе усложняет задачу поиска необходимых данных.

Можно сказать, что извечная проблема поиска информации сегодня получила новое звучание: "поиск информации в неограниченной, неоднородной динамической информационной среде", или, если перефразировать, — "поиск иголки в стоге сена".

Традиционные поисковые системы предлагают лишь частичное решение этой проблемы. Периоды индексации у них составляют от недель до нескольких месяцев. И несмотря на то, что практически все известные поисковые порталы (Google, Yahoo!, AltaVista, Lycos и другие) имеют новостные разделы, они, сами по себе, уже многих не устраивают. Традиционным подходам к организации поиска сетевой информации присущи такие недостатки, как низкая оперативность, зависимость от набора источников и ограниченность спектра этих источников, средние поисковые возможности, отсутствие средств уведомления о появлении новых данных.

Одна из проблем нахождения информации в Сети обусловлена основным форматом, в котором представлена эта информация, — HTML. Этот формат был разработан, в первую очередь, для решения задач отображения содержания на каждом конкретном Web-ресурсе, поэтому он не всегда удобен для автоматической обработки информации, в том числе и организации поиска. В результате информация в Internet оказалась ориентирована, прежде всего, на отдельные сайты и очень слабо приспособлена для автоматизированного обобщения, классификации и аналитической обработки.

При импортировании в Web-ресурс информации с другого сайта (включении новостных сообщений и т.п.) возникает вопрос однотипного представления их содержания (контента). Если этот вопрос не решается, то изменение HTML-оформления сайта-источника приводит к необходимости одновременной модификации программного обеспечения на всех сайтах, которые принимают от него информацию.

Итак, объективно назрела необходимость использования некоего унифицированного формата представления данных. Сегодня в качестве такового все чаще используется XML или его подмножество — RSS (об этих форматах речь пойдет ниже). XML представляет собой метаязык, т.е. язык, на базе которого можно определять новые языки. Он предназначен не только для организации обмена данными в Web, но и для распознавания их семантики. В отличие от HTML, XML предназначен для представления информации в "чистом" виде, предполагая структурную, а не оформительскую разметку данных.

3.3. Невизуальный Web

Сегодня Internet — это огромное хранилище информации, интегрированный доступ к динамической составляющей которого — новостным ресурсам — затруднен. Разнообразие информации, в том числе и новостных сообщений, в Сети не может быть полезным на практике при отсутствии эффективного доступа.

Поэтому, как было сказано выше, возникла необходимость в использовании унифицированного формата данных на сайтах, стандарта, обеспечивающего однотипный обмен данными в Internet. В качестве такого унифицированного формата все шире используется язык eXtensible Markup Language (XML) и его диалекты.

Одним из первых проектов унификации обмена данными в Internet стал "Семантический Web" (Semantic Web) [40, 59]. Основная идея проекта, задуманного в консорциуме W3C Тимом Бернерсом-Ли и его коллегами (рис. 3.1), за ключалась

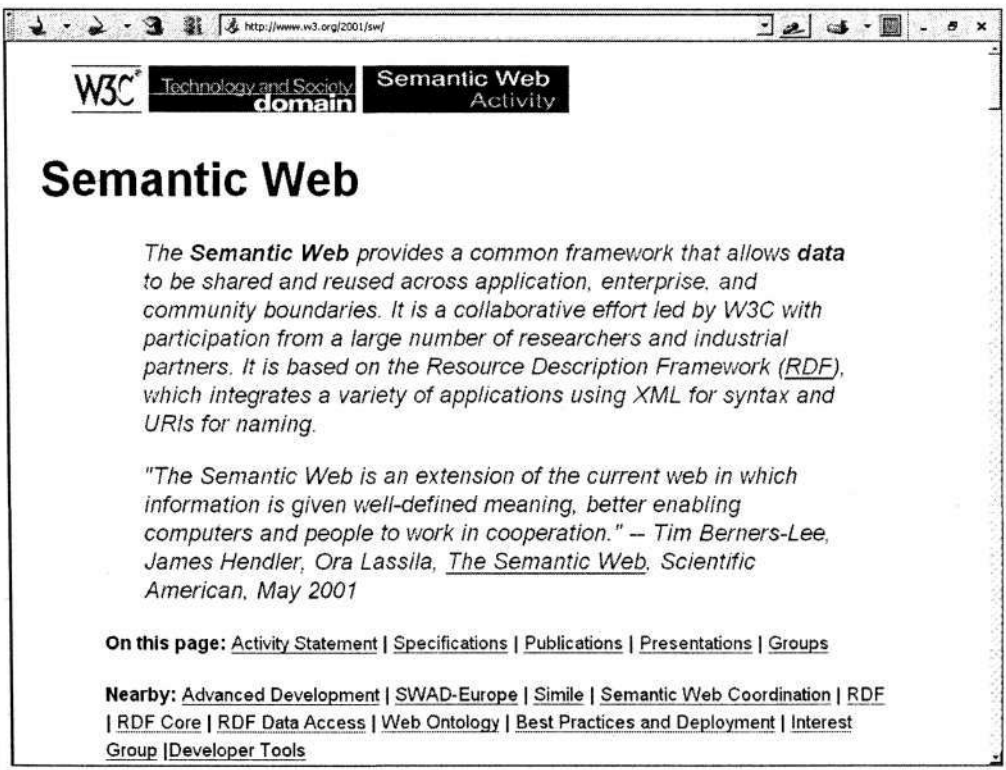


Рис. 3.1. Проект консорциума W3C "Семантический Web"

в такой организации данных: Web-серверы должны не только визуализировать, но и использовать их, чтобы программы разных производителей могли эффективно работать с Web-контентом. Именно для "Семантического Web" были разработаны спецификации XML, предусматривающие разделение средств визуализации и смыслового содержания.

В основу "Семантического Web" были положены три ключевых элемента:

- спецификация XML, позволяющая определить синтаксис и структуру документов;
- механизм описания ресурсов — Resource Definition Framework (RDF), обеспечивающий модель кодирования для значений, определенных в онтологии;
- система онтологий, позволяющая определять термины (или понятия) и отношения между ними.

"Семантический Web" также использует другие технологии и концепции, в частности универсальные идентификаторы ресурсов, цифровые подписи, системы логического вывода, обычные протоколы Internet и т.д.

XML представляет собой метаязык, т.е. язык, на базе которого можно определять новые языки. Но он предназначен не только для организации обмена данными в Web, но и для распознавания семантики этих данных. В отличие от HTML, XML обеспечивает представление информации в чистом виде, предполагая ее структурную, а не оформительскую разметку. При этом потребовались стандарты не только для синтаксической формы документов, но и для их семантического наполнения. В результате консорциумом W3C были разработаны стандарты языков XML и RDF, которые совместно позволяют поддерживать семантическую совместимость в Сети.

Вместе с тем, формально элементы разметки (теги) XML оторваны от определения их смыслового наполнения. Поэтому параллельно с XML была начата разработка стандарта для схемы описания источников RDF — языка формального описания содержимого Web-сайтов в рамках единого стандарта.

RDF является языком общего назначения для описания информации в Web. RDF-документы представляют собой совокупность RDF-предложений, состоящих из троек элементов: *ресурс — именованное свойство — значение свойства* (или *объект — атрибут — значение атрибута*). Ресурсом может выступать понятие, которому можно приписать некоторый URI (Uniform Resource Identifiers). Значение свойства или атрибута — это его контент, т.е. содержимое.

Спецификации RDF обеспечивают поддержку тегов, позволяющих определять любые понятия (например, тегами `PRICE` и `INVOICE` можно пользоваться для обозначения цены и счета соответственно). Следует заметить, что данным в формате RDF присваиваются дескрипторы, которые могут определяться в отдельных файлах определения типов документов (Document Type Definitions — DTD). Сегодня практически в каждой отрасли знаний имеется свой, постоянно расширяющийся список DTD. На основе XML и RDF был создан формат RSS, специально предназначенный для организации информационных коммуникаций как между людьми, так и между серверами [57].

Предполагается, что третий элемент "Семантического Web" — онтологии — будет играть определяющую роль в обработке знаний в Сети, а также в их совместном использовании приложениями. При этом онтология определяется как

система, состоящая из набора понятий и набора утверждений этих понятиях, на основе которых можно строить классы, объекты и отношения. Онтология определяет семантику конкретной области и способствует установлению связей между значениями элементов предметной области.

В рамках "Семантического Web" предлагается в среде RDF описывать онтологии на языке RDF.

3.4. Синдикация новостной информации

Оптимальное решение, способное помочь ориентироваться в новостной информации Internet, сегодня предоставляют информационные службы нового типа — системы синдикации новостей. Под синдикацией в данном случае понимается сбор информации в Internet и последующее распространение ее фрагментов в соответствии с потребностями пользователей [16, 25]. Кроме того, службы синдикации обеспечивают публикацию одних и тех же данных на различных сайтах (в том числе предназначенных для карманных компьютеров и мобильных телефонов).

Службы синдикации обеспечивают публикацию соответствующих потребностям пользователей данных на различных Web-страницах, сайтах и порталах (в том числе на Internet-ресурсах для карманных компьютеров и мобильных телефонов), а также доставку информации пользователям.

Технология синдикации Internet-новостей включает в себя "обучение" программ сбора структуре выбранных источников (Web-сайтов), непосредственное сканирование информации, приведение ее к общему формату (в последнее время — к XML), а также ее классификацию и доставку пользователям различными путями (e-mail, Web, WAP, SMS и т.д.).

3.5. От "поисковиков" — к "интеграторам"

Необходимость сетевой интеграции новостей несколько лет назад осознали известные сетевые поисковые службы — Google, Excite, Lycos, AltaVista. На первых этапах они заключили соглашения с крупнейшими информационными агентствами, такими как Reuters, Associated Press, CNN и другие, и стали предоставлять возможности поиска и просмотра новостных сообщений. Таким образом у пользователя впервые появилась возможность бесплатно находить и просматривать новости реального (а не только "виртуального") мира в Сети. Старейший навигационный портал Yahoo! также не обошел стороной идею интеграции новостей, создав службу Daily News (`http://dailynews.yahoo.com`), объединяющую информацию нескольких десятков информационных агентств и обеспечивающую графическое и мультимедийное представление отдельных тематических областей.

Практически одновременно с освоением традиционными СМИ виртуального пространства Internet и с настоящей экспансией он-лайновых изданий стали возникать службы, обеспечивающие обобщенное представление информации со страниц сетевых СМИ на своих сайтах, а также "проталкивание" (push-технология) информации, якобы интересующей пользователей, в рабочие области их браузеров. Получила развитие технология "персональных информационных агентов", обеспечивающих клиентскую часть систем интеграции новостей.

Идея интеграции новостей породила несколько технологий, имеющих общие корни. Один из мировых лидеров на рынке интеграции Internet-контента,

компания MoreOver (http://www.moreover.com) опубликовала свою технологическую схему интеграции контента, которая в настоящее время "де-факто" признана типовой (рис. 3.2).

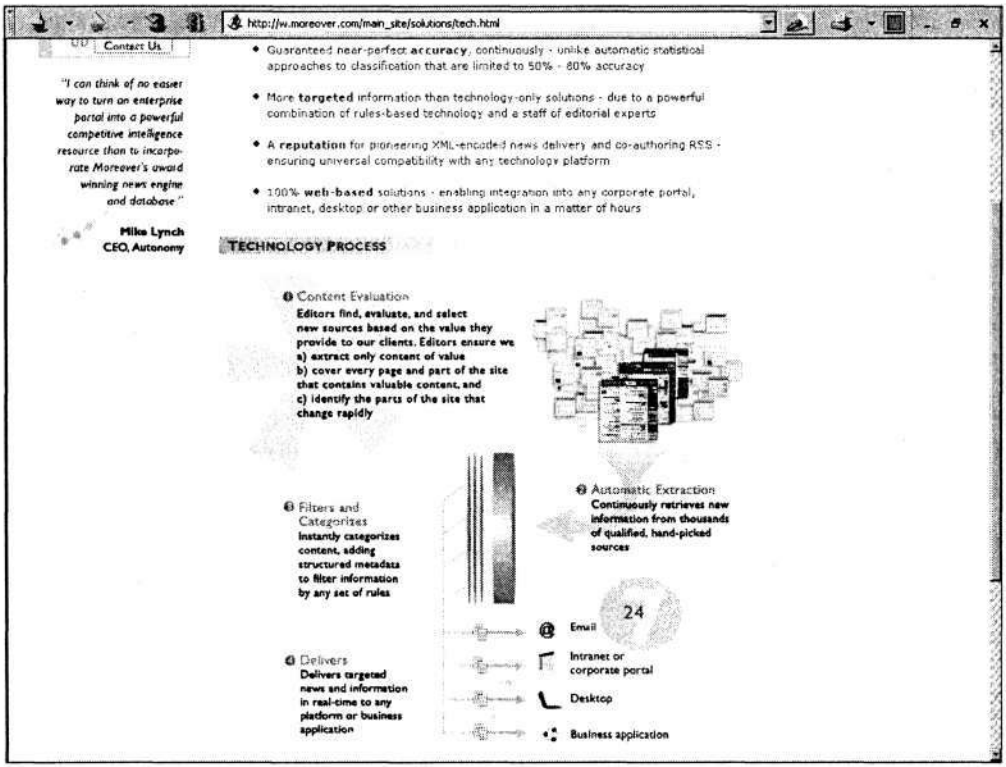


Рис. 3.2. Технология интеграции новостей компании MoreOver

На первом уровне типовой системы интеграции контента выполняется оценка источников информации: специалисты-эксперты находят необходимые сайты-источники, анализируют их содержание, форматы представления данных, частоту обновления результатов. Конечный результат их работы — шаблоны (или макроописания) информационных ресурсов.

На втором уровне осуществляется автоматизированный сбор обновлений с целевых Web-сайтов — непрерывное сканирование материалов из тысяч источников с помощью программ-роботов.

На третьем уровне обрабатывается контент, что выполняется в несколько этапов.

1. Приведение собранной информации к выбранному внутрисистемному формату (как правило, это подмножество XML).
2. Систематизация информации, добавление к ней метаданных для обеспечения последующей фильтрации.
3. Категоризация.
4. На последнем этапе осуществляется доставка релевантных информационных материалов пользователям путем:

- рассылки результатов избирательного распределения сообщений по электронной почте;
- публикации на определенных страницах Web-сайтов (как открытых, так и закрытых, корпоративных);
- загрузки в intranet-сети (информационные хранилища или корпоративные порталы);
- подачи информации на вход разнообразным бизнес-приложениям.

3.6. Форматы синдикации новостей

Для решения задачи синдикации новостей было создано несколько форматов описания данных на основе XML. Самый распространенный формат получил название RSS [29, 30], что означает Really Simple Syndication, Rich Site Summary, хотя изначально он назывался RDF Site Summary. Смысл всех этих аббревиатур заключается в простом способе обобщения и распределения информационного наполнения Web-сайтов, т.е. синдикации контента.

Изначально формат RSS создавался компанией Netscape для портала Netcenter как одно из первых XML-приложений, но затем он стал использоваться на многих других сайтах. Сегодня практически все ведущие новостные сайты, — "Живые журналы", работающие в Internet, — используют RSS в качестве инструмента оперативного представления своих обновлений. Например, сегодня экспорт в RSS осуществляют крупнейшие порталы, включая CNN, BBC News, Amazon, CNet News, MSNBC, The Register, Wired и т.д.

RSS действительно обеспечивает согласованный способ резюмировать содержимое Web-сайтов. Кроме того, его применение позволило администраторам новостных сайтов, он-лайновых дневников-блогов, форумов и других часто обновляемых Web-ресурсов представить информацию в унифицированном виде.

По признанию многих экспертов, 2004 год стал "Годом RSS", т.е. в этом году началось широкое внедрение этого формата. При этом только в начале 2004 года Internet-пользователи по-настоящему открыли для себя все прелести технологии RSS. Сегодня для работы с новостями в формате RSS разрабатываются все новые программы, сайты и поисковые системы, которые все более востребованы, в частности, пользователями карманных компьютеров.

Итак, RSS — это формат данных и технический стандарт, который обеспечивает интегрированный доступ к новостной информации, представленной на Web-сайтах, и специально создан для обмена их контентом.

Развитие RSS началось с версии 0.90, разработанной компанией Netscape, но его посчитали слишком сложным, и Netscape разработала упрощенную версию — 0.91, которую, после бума портальных технологий, передала компании UserLand Software. Это самый простой и доступный стандарт, который применяется сегодня, когда требуется несложный экспорт заголовков. Одновременно еще одна организация — RSS-DEV Working Group — создала свою версию RSS (1.0), близкую к исходной версии RSS 0.90 и максимально приближенную к стандарту RDF; эта версия предоставляет больше возможностей, чем все 0.9x, — например, допускает расширение при помощи модулей. Компания же UserLand решила развить ветвь 0.9x и создала версии 0.92, 0.93, 0.94, которые позволяют представлять метаданные, и, наконец, 2.0. При этом RSS 2.0 — это не новая версия RSS 1.0, а логическое продолжение ветви 0.9x. В ней также добавлена поддержка модулей. В настоящее

время существует 7 независимых версий RSS — 0.90, 0.91, 0.92, 0.93, 0.94, 1.0, 2.0. Эти версии отличаются друг от друга, хотя все они ориентированы на один тип информации и содержат одинаковые базовые поля. При этом многие считают все версии, кроме 2.0, устаревшими и "отмененными", но это далеко не так. Пока еще самой популярной является RSS 0.91. Что же касается версии 0.94, то ее спецификаций не сохранилось даже на авторском сайте UserLand. Так, по адресу `http://backend.userland.com/rss094` находится спецификация версии RSS 2.0.

Спецификации отдельных версий формата RSS приведены на таких Web-страницах:

- RSS 0.90: `http://www.purplepages.ie/RSS/netscape/rss0.90.html`
- RSS 0.91: `http://my.netscape.com/publish/formats/rss-spec-0.91.html`
- RSS 0.92: `http://backend.userland.com/rss092`
- RSS 0.93: `http://backend.userland.com/rss093`
- RSS 1.0: `http://web.resource.org/rss/1.0`
- RSS 2.0: `http://backend.userland.com/rss`

Во всех версиях RSS есть некоторые особенности, но, как уже говорилось, объединяет их ориентация на один тип информации, вследствие чего они содержат общие базовые поля: основной блок данных (`channel`), который содержит такие атрибуты, как "заглавие канала" (`title`), "ссылки" (`link`), "данные о языке сообщений" (`language`) и "логотип" (`image`), после них идет список самих сообщений, где в каждом пункте (`item`) указывается заголовок (`title`), краткое описание (`description`) и ссылка на новость (`link`). Кроме того, каждый RSS-файл начинается обязательными элементами `xml` и `rss`. Первый из этих элементов содержит атрибуты `version` (версия) и `encoding` (кодировка).

Среди множества необязательных элементов RSS можно назвать самые распространенные — язык (`language`), авторское право (`copyright`), категория информации (`category`), дата и время публикации сообщения (`pubDate`), программа, которая использовалась для создания файла (`generator`), картинка, которую следует показывать наряду с текстовой информацией (`image`).

Кроме заголовка блока данных, в формате RSS предусмотрено описание отдельных информационных элементов (`item`). Каждый элемент — это отдельная статья или краткая аннотация и ссылка на полную версию статьи. Канал (`channel`) может содержать любое число элементов, содержащих только два обязательных вложенных элемента — название (`title`) и описание (`description`). Кроме того, часто используются такие вложенные элементы: ссылка на первоисточник (`link`), категория (`category`), комментарий (`comments`) и автор (`author`).

В качестве примера новостного канала формата RSS 0.91 можно привести динамический файл, формируемый по адресу `http://uaport.net/cgi-bin/infostream.rss` (обзор основных событий дня "Електронні Вісті"). Он имеет такой вид:

```
 <?xml version="1.0" encoding="windows-1251" ?>
<!DOCTYPE rss PUBLIC "-//Netscape Communications//DTD RSS 0.91//EN"
"http://my.netscape.com/publish/formats/rss-0.91.dtd">
<rss version="0.91">
<channel>
```

```
<title>Електронні Вісті</title>
<language>ru</language><image>
<title>Електронні Вісті</title>
<url>http://www.elvisti.com/images/export/elvisticom3_88x31.gif</url>
<link>http://www.elvisti.com</link>
<width>88</width>
<height>31</height>
</image>

<item><title>РАДАР СЛЕДИТ ЗА КОСМИЧЕСКИМ МУСОРОМ</title>
<description>В японской префектуре Окаяма с 6 апреля начал работать радар
с дистанционным управлением, основная функция которого состоит в
отслеживании перемещения космического мусора.<criptions>
<link>http://elvisti.com/2004/04/06/sci-tech.shtml#3</link>
<item>

<item><title>В ИВАНО-ФРАНКОВСКОЙ ОБЛАСТИ КУРИЦА СНЕСЛА ЯЙЦО ВЕСОМ 143 Г</title>
<description>В селе Делиев Галицкого района Ивано-Франковской области
курица снесла яйцо весом 143 г. </description>
<link>http://elvisti.com/2004/04/06/misc.shtml</link>
</item>

<item><title>В США БОЛЕЕ 60% КОРПОРАЦИЙ В 1990-Е ГОДЫ НЕ ПЛАТИЛИ НАЛОГИ</title>
<description>Более 60% американских корпораций в период бума
американской экономики с 1996 по 2000 годы не платили налоги
в государственную казну, сообщило Главное бюджетно-контрольное
управление США.</description>
<link>http://elvisti.com/2004/04/06/biz.shtml#2</link>
</item>

<item><title>СЕДЬМОЕ АПРЕЛЯ - ВСЕМИРНЫЙ ДЕНЬ ЗДОРОВЬЯ</title>
<description>В нынешнем году по рекомендации ВОЗ этот день пройдет под
лозунгом "Безопасность на дорогах зависит от каждого из нас".<criptions>
<link>http://elvisti.com/2004/04/06/health.shtml#2</link>
</item>

</channel>
</rss>
```

Помимо формата RSS, недавно появился формат Atom (`http://www.mnot.net/drafts/draft-nottingham-atom-format-02.html`), пока окончательно не утвержденный, но используемый на крупнейшем поисковом портале Google, что предопределяет его популярность. Открытый стандарт Atom (актуальная версия — 3.0) совершенствуется командой программистов из IBM, Google и других компаний.

По информации журнала New Scientist, сегодня компания Google снова рассматривает возможность использования в некоторых своих сервисах формата RSS, хотя еще в феврале 2004 года сообщила, что для генерации и доставки срочных сообщений подписчикам он-лайнового журнала Blogger Google будет использовать формат Atom, разработанный в 2003 году компанией IBM.

И Atom, и RSS имеют свои преимущества. Atom обеспечивает подписчикам большую гибкость в выборе профиля, поскольку поддерживает больше метаданных, и является открытым стандартом, развитие которого далеко от завершения.

Он позволяет рассылать не только сами сообщения, но и комментарии читателей, что формату RSS пока не под силу.

Как и RSS, Atom является подмножеством XML. Приведем пример файла в этом формате, чтобы подчеркнуть его близость с RSS:

```
<?xml version="1.0" encoding="utf-8"?>
<feed version="0.3" xmlns="http://purl.org/atom/ns#">
  <title>Наименьший возможный фид в формате Atom 3.0</title>
  <link rel="alternate" type="text/html"
href="http://diveintomark.org/"/>
  <modified>2004-04-09T18:30:02Z</modified>
  <author>
    <name>Иванов Петр</name>
  </author>
  <entry>
    <title>Atom 0.3 пример</title>
    <link rel="alternate" type="text/html"
     href="http://uaport.ua/2004/04/09/atom03"/>
    <id>tag:uaport.ua,2004:4.2397</id>
    <issued>2004-04-09T08:29:29-04:00</issued>
    <modified>2004-04-09T18:30:02Z</modified>
  </entry>
</feed>
```

Дэйв Уинер (Dave Winer), один из главных разработчиков RSS, недавно призвал всех разработчиков объединить свои усилия и разработать единый формат, совместимый как с RSS, так и с Atom, чтобы слить конкурентные стандарты в единое целое. "Новый формат можно назвать RSS/Atom, — заявил Уинер. — Он бы имел всю функциональность, которую разработчики Atom обещают внедрить. Максимально авторитетный формат получил бы наиболее полную поддержку от всех разработчиков." Уинер предлагает, чтобы в RSS/Atom было как можно меньше отличий от RSS 2.0.

3.7. OPML — формат для хранения списка RSS-фидов

Еще один диалект XML — язык OPML (Outline Processor Markup Language) — используется для описания совокупности RSS-фидов. Его спецификация размещена по адресу `http://opml.scripting.com/spec`. С помощью OPML обеспечивается эффективный унифицированный обмен списками RSS-фидов.

Для того чтобы обеспечить удобную подписку с помощью RSS-агрегаторов сразу на несколько фидов, разработан специальный механизм, также ориентированный на формат XML. Список RSS-фидов заносится в файлы специальных форматов, имя и путь (адрес в сети Internet) к которым указывается в программах-агрегаторах при подписке. Применение этих файлов возможно в большинстве современных агрегаторов и не зависит от операционных систем, в которых эти агрегаторы работают.

В настоящее время для создания списка RSS-фидов применяется два основных, базирующихся на XML, открытых формата: OCS (Open Content Syndication) и OPML (Outline Processor Markup Language). Наиболее распространенным является формат OPML, спецификация которого приведена по адресу `http://www.opml.org/spec`. По этому же адресу можно подписаться на рассылку бюллетеня, в котором публикуются новости, относящиеся к спецификации и применениям OPML.

Рассмотрим, для примера, файл, описывающий RSS-каналы системы InfoStream® на портале UAport (http://uaport.net/feeds.opml).

```
<?xml version="1.0" encoding="windows-1251" ?>
<opml version="1.0">
<head>
    <title>InfoStream News</title>
    <ownerName>InfoStream.ua</ownerName>
    <ownerEmail>stream@visti.net</ownerEmail>
</head>

<body>
<outline text="АПК"
    htmlUrl="http://uaport.net/UAnews?rub=01"
    language="ru" title="http://uaport.net/UAnews?rub=01"
    type="rss"
    xmlUrl="http://uaport.net/cgi-bin/infostream.rss?rubr01" />
<outline text="Банковская сфера"
    htmlUrl="http://uaport.net/UAnews?rub=02"
    language="ru" title="http://uaport.net/UAnews?rub=02"
    type="rss"
    xmlUrl="http://uaport.net/cgi-bin/infostream.rss?rubr02" />
<outline text="Экономика"
    htmlUrl="http://uaport.net/UAnews?rub=03"
    language="ru" title="http://uaport.net/UAnews?rub=03"
    type="rss"
    xmlUrl="http://uaport.net/cgi-bin/infostream.rss?rubr03" />
<outline text="Недвижимость"
    htmlUrl="http://uaport.net/UAnews?rub=05"
    language="ru" title="http://uaport.net/UAnews?rub=05"
    type="rss"
    xmlUrl="http://uaport.net/cgi-bin/infostream.rss?rubr05" />

</body>
</opml>
```

Первая строка OPML-файла несет информацию о том, что этот файл размечен на подмножестве формата XML версии 1.0, а также о том, что в нем используется кодировка Windows 1251. Вторая строка содержит обязательный атрибут `version`. Данный OPML-файл состоит из двух основных частей: заголовка (`head`) и тела (`body`). Заголовок может содержать следующие основные теги, описывающие данный OPML-документ:

- `<title>` — заголовок документа, представляющего собой список RSS-фидов;
- `<dateCreated>` — дата создания документа в формате RFC 822 (с указанием дня недели и часового пояса);
- `<dateModified>` — дата модификации документа;
- `<ownerName>` — владелец документа;
- `<ownerEmail>` — электронный почтовый адрес владельца.

Остальные элементы, упомянутые в спецификации, имеют второстепенное значение.

Тело OPML-документа теоретически состоит из неограниченного количества тегов. Также теоретически допускается использование тегов других типов, однако они чаще всего не поддерживаются популярными программами-агрегаторами. Каждый тег может содержать следующие основные атрибуты:

- `type` — тип элемента; имеет значение `rss` для фидов;
- `title` — название RSS-фида, которое соответствует тегу `<title>` элемента `<channel>` в формате RSS;
- `description` — краткое описание фида, которое соответствует тегу `<description>` элемента `<channel>` в формате RSS;
- `language` — язык документа;
- `xmlUrl` — гиперссылка на фид в виде RSS;
- `htmlUrl` — гиперссыка на HTML-страницу данного фида, которая соответствует тегу `<link>` в `<channel>` для RSS.

3.8. Источники новостного контента

Основным применением RSS в настоящее время являются новостные фиды (feed). Фид — это файл в формате RSS, в который записывается новостной контент Web-ресурса. Если есть необходимость оперативно отслеживать изменения на содержащем фид сайте, то можно делать это с помощью программы-агрегатора, не посещая самого сайта с помощью стандартных программ-браузеров.

Ниже приведены адреса самых популярных в Internet RSS-фидов:

```
http://w.moreover.com/categories/ocs/ocsdirectory.rdf
http://10.am/extra/ocsdirectory.php
http://www.newsisfree.com/ocs/directory.xml
http://blogspace.com/rss/feeds/converted.ocs
http://www.groksoup.com/ocs/ocsdirectory.xml
http://theweb.startshere.net/channels.phtml?format=OCS
http://myrss.com/catalog/ocs04.rdf
http://www.syndic8.com/xml.php
```

В настоящее время в русскоязычной части Internet представлены тысячи RSS-фидов, наиболее популярные из которых такие:

- NEWSru.com — `http://www.newsru.com/plain/rss/all.xml`
- Газета.ru - Все новости (RSS) — `www.gazeta.ru/export/gazeta_rss.xml`
- Lenty.RU — `http://www.lenty.ru/export/bestnews.rss`
- Подробности — `http://www.podrobnosti.com.ua/export`
- Lenta.ru — `http://lenta.ru/l/r/EX/import.rss`
- Полит.РУ — `http://www.polit.ru/rss/index.xml`
- Портал "Юридическая Россия" — `http://law.edu.ru/rss/news.rss`
- Водка он-лайн — `http://vodka.com.ua/export/rss.xml`
- Портал "ПлейМобайл" — `http://playmobile.ru/news/rss`
- 3Dnews — `http://www.3dnews.ru/expnews/rss/newsrss.xml`

Обширный список RSS-фидов русскоязычного сегмента Internet находится по адресу http://my.yandex.ru/rss.opml; приведем лишь некоторые, наиболее интересные новостные фиды:

- Аргументы и Факты — http://www.aif.ru/info/rss.php?magazine=aif
- АвтоОБЗОР — http://auto.obzor.ru/news/autonews.xml
- АвиаПорт.Ру — http://www.aviaport.ru/news/yandex_export.xml
- Деловая Хроника — http://www.chronicle.ru/l/r/EX/rsschannel.xml
- K2Kapital — http://ad.k2kapital.com/cbp/mynetscape/mynews.news
- Linux.org.ru — http://images.linux.org.ru/getrss.php3
- PalmQ Online — http://www.palmq.net/backend.php
- СПОРТ сегодня — http://www.sports.ru/sports_docs.xml
- TRAVEL.RU. Все о путешествиях — http://www.travel.ru/inc/side/yandex.rdf
- АПК-Информ — http://www.apk-inform.com/yandextr.php
- ФОНТАНКА.РУ — http://www.fontanka.ru/_transmission_for_yandex.thtml
- IMA Press. Тема дня — http://www.ima-press.ru/rss.php?newsblock=theme&limit=1
- Журнал "Итоги" — http://www.itogi.ru/WebExport.nsf/Anons/itogi.xml
- Остров. Новости Донбасса — http://www.ostro.org/yandex.php
- ПОЛИТ.РУ — http://www.polit.ru/rss/index.xml?yandex_mode=1
- PRAVDA.Ru — http://export.pravda.ru/yandex.txt
- PR NEWS (все пресс-релизы компаний) — http://www.prnews.ru/yandex/business.asp
- Энциклопедия поисковых систем — http://www.searchengines.ru/news/news.rdf
- Сетевой журнал — http://www.setevoi.ru/weekly/export1.txt

На сегодня существует уже множество служб синдикации новостей, которые предоставляют тематические фиды, построенные на основе использования многочисленных источников. Такой фид, к примеру, доступен на портале UAport (http://uaport.net) и позволяет получить интегрированный доступ к потоку украинских и российских новостных сообщений, собираемому системой InfoStream. С помощью RSS-шлюза системой InfoStream предоставляется унифицированный доступ к информации более чем с 600 Web-сайтов, сгруппированной по тематикам, языкам, странам, источникам. Объем этой информации сегодня превышает 20 000 сообщений в сутки. RSS-каналы UAport могут генерироваться системой по собственным запросам пользователей к поисковой системе.

Рассмотрим функциональность отдельных служб синдикации новостей, предоставляющих информацию в формате RSS.

Moreover

Для интеграции соответствующего запросам пользователей контента в корпоративные сети или порталы служба Moreover (http://www.moreover.com) использует собственное решение — Connected Intelligence. Прием информации в систему от 6500 источников в режиме реального времени происходит каждые 15 минут, сообщения классифицируются и группируются по темам.

На сайте Moreover содержатся сведения о технологических подходах к интеграции новостей, которые были созданы в этой службе и де-факто стали стандартами в системах мониторинга. Определена следующая технологическая цепочка: сначала выполняется оценка информационного содержания Web-ресурса и построение конфигурационных профилей, описывающих данный ресурс. Затем Web-ресурсы автоматически сканируются в соответствии с профилями и происходит преобразование информации в формат XML с добавлением метатегов. После этого осуществляется классификация информации и ее распределение по запросам пользователей. На последнем этапе происходит вывод и доставка информации клиентам.

В июле 2003 года технология Moreover была интегрирована в новостной портал Yahoo!, с сайта которого (http://news.yahoo.com) возможен доступ к информации из 3500 источников (рис. 3.3).

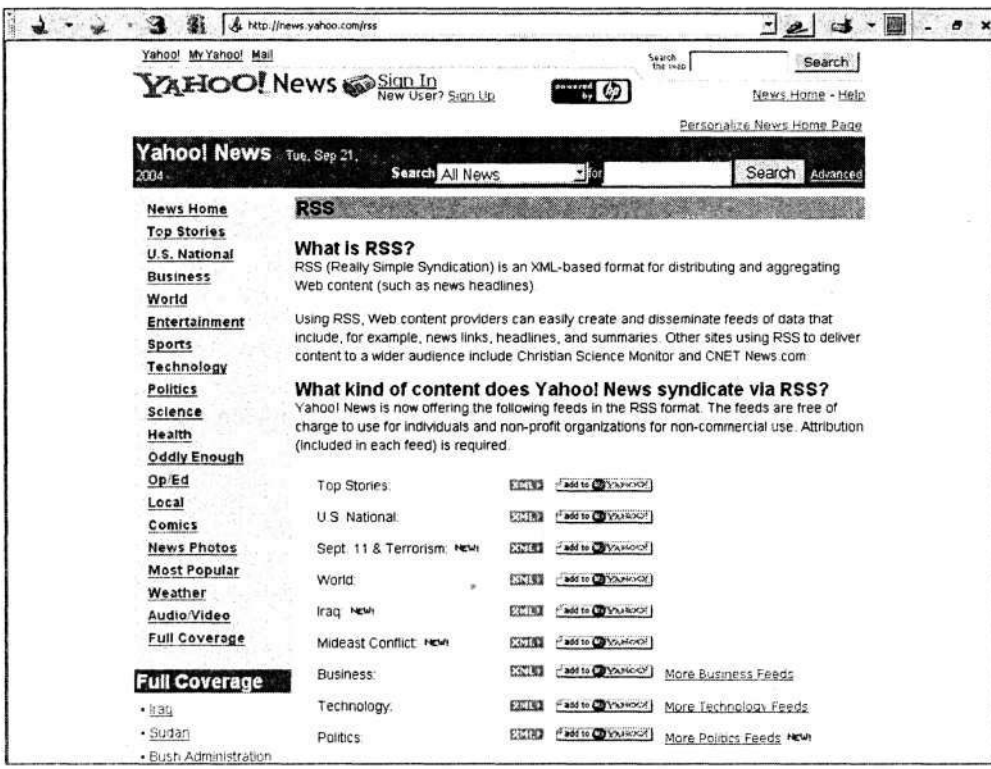


Рис. 3.3. RSS-фиды новостной службы Yahoo

Google

В 2002 году популярная поисковая система Google запустила свой новостной сервис — Google News (http://news.google.com), который охватывает информацию с 4500 различных сайтов за последние 30 дней. Данные на сайте системы отсортированы по нескольким категориям, таким как международные новости, деловой мир, шоу-бизнес, технологии и спорт.

Новости в системе отбираются в зависимости от времени их публикации, популярности источника информации и количества статей, появившихся в Internet, на данную тему. Компания Google — популяризатор и один из разработчиков формата Atom, применяемого, в основном, в блогах.

Вместе с тем, компания Google с подозрением относится к широким возможностям RSS-синдикации, углядев в этой технологи возможности для нарушений авторских прав. Так, недавно Google запретила британскому Web-мастеру использовать результаты поиска в системе Google News на другом сайте в виде RSS-фида. Британский программист Джулиан Бонд создал сценарий на языке PHP, который берет введенный пользователем запрос, направляет его на Google News, а результат выдает в формате RSS. Полученный результат можно использовать в любом RSS-агрегаторе. Сам скрипт под названием gnews2rss можно найти на сайте http://www.voidstar.com/gnews2rss.php. По словам Бонда, основной протест со стороны Google вызвал не сам скрипт, а использование его для формирования новостной ленты на постороннем сайте. Сам скрипт попрежнему доступен в Internet, и его можно использовать в программах-агрегаторах. Тем не менее в письме Бонду в Google указывали на то, что предпочтительным вариантом является применение службы Google News Alerts.

NewsIsFree

Одна из самых перспективных в Сети служб синдикации новостей NewsIsFree (http://www.newsisfree.com) охватывает свыше 9000 источников (в том числе российских и украинских). Сообщения обновляются каждые 15 минут и группируются по 15 основным категориям (http://www.newsisfree.com/sources/browse). Примечательно, что режим поиска в RSS-ресурсах обеспечивается поисковым механизмом компании Google. Основная особенность службы NewsIsFree — это полная интеграция с XML, в частности с RSS 0.91. Большинство разделов сайта службы содержит ссылки Syndicate, активизация которых приводит к отображению кода разделов в формате XML.

Несмотря на то что основой информационных ресурсов, охватываемых службой, являются англоязычные источники, NewsIsFree сегодня крупнейший интегратор и русскоязычных RSS-фидов, каталог которых доступен по адресу http://newsisfree.com/sources/bylang/?lang=ru.

MSDN

Учитывая существующие в мире тенденции, служба MSDN (http://msdn.microsoft.com) также приступила к публикации своих новостей в формате RSS, выбрав версию 2.0 (рис. 3.4). Ниже приведен список некоторых тем и адресов новостных фидов MSDN:

- .NET Framework — http://msdn.microsoft.com/netframework/rss.xml
- ASP.NET — http://msdn.microsoft.com/asp.net/rss.xml

- Longhorn — http://msdn.microsoft.com/longhorn/rss.xml
- Mobile and Embedded — http://msdn.microsoft.com/mobility/rss.xml
- MSDN Subscriber Download — http://msdn.microsoft.com/subscriptions/rss.xml
- Office — http://msdn.microsoft.com/office/rss.xml
- Security — http://msdn.microsoft.com/security/rss.xml
- Visual Basic — http://msdn.microsoft.com/vbasic/rss.xml
- Visual C# — http://msdn.microsoft.com/vcsharp/rss.xml
- Visual C++ — http://msdn.microsoft.com/visualc/rss.xml
- Visual FoxPro — http://msdn.microsoft.com/vfoxpro/rss.xml
- Visual J# — http://msdn.microsoft.com/vjsharp/rss.xml
- Visual Studio — http://msdn.microsoft.com/vstudio/rss.xml
- Web Services — http://msdn.microsoft.com/webservices/rss.xml
- Windows Embedded — http://msdn.microsoft.com/embedded/rss.xml
- XML — http://msdn.microsoft.com/embedded/xml.xml

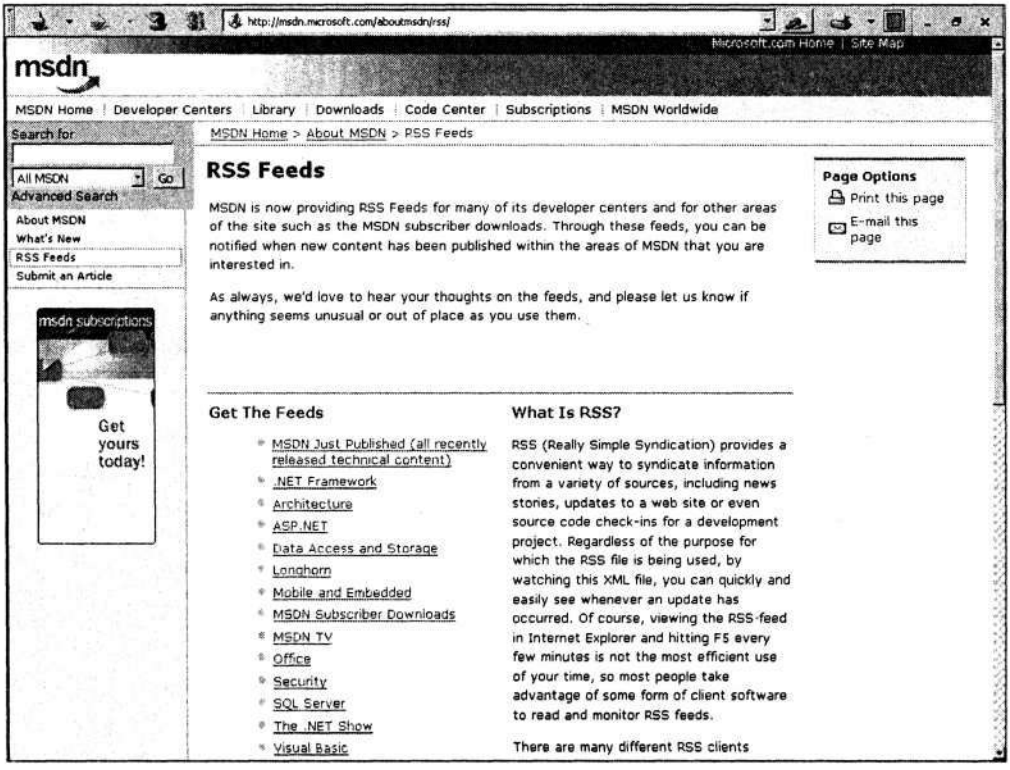


Рис. 3.4. Новостные RSS-фиды от MSDN

Яndex.Новости

Служба Яndex открыла проект Яndex.Новости (http://news.yandex.ru), к которому в настоящее время присоединилось свыше 500 Internet-изданий. Новости сортируются по десяти категориям, существует возможность поиска с указанием раздела и времени публикации новости. Поиск новостей возможен как по всем источникам, так и по заданным пользователем. Имеется также возможность поиска за произвольный период времени. Для сбора и экспорта новостей используется формат RSS 2.0.

Сегодня бесплатная служба синдикации новостного контента Яndex представляет такие основные каналы:

- Главные новости — http://news.yandex.ru/index.rss
- Политика — http://news.yandex.ru/politics.rss
- В мире — http://news.yandex.ru/world.rss
- Общество — http://news.yandex.ru/society.rss
- Экономика — http://news.yandex.ru/business.rss
- Спорт — http://news.yandex.ru/sport.rss
- Происшествия — http://news.yandex.ru/incident.rss
- Культура — http://news.yandex.ru/culture.rss
- Здоровье — http://news.yandex.ru/health.rss
- Компьютеры — http://news.yandex.ru/computers.rss
- Internet — http://news.yandex.ru/internet.rss
- Авто — http://news.yandex.ru/auto.rss

InfoStream

Разработанная в Информационном центре ЭЛВИСТИ система InfoStream® (http://infostream.ua) обеспечивает персонализацию интерфейса пользователей, работающих в режиме он-лайн, т.е. сохранение их постоянных запросов и организацию подписки, что реализуется на основе современной технологии RSS 0.91.

Для получения тематической ленты InfoStream (RSS-фида) в соответствующее поле RSS-агрегатора следует ввести адрес в формате:

http://uaport.net/cgi-bin/infostream.rss?<*ЗАПРОС*>

где в качестве значения <*запрос*> можно ввести слово или словосочетание на языке запросов информационно-поисковой системы InfoReS.

На основе технологии InfoStream созданы такие новостные каналы:

- Агропром — http://uaport.net/cgi-bin/infostream.rss?rubr01
- Банки — http://uaport.net/cgi-bin/infostream.rss?rubr02
- Экономика — http://uaport.net/cgi-bin/infostream.rss?rubr03
- Экономика Украины — http://uaport.net/cgi-bin/infostream.rss?rubr04
- Недвижимость — http://uaport.net/cgi-bin/infostream.rss?rubr05
- Биржи — http://uaport.net/cgi-bin/infostream.rss?rubr06

- Инвестиции — http://uaport.net/cgi-bin/infostream.rss?rubr07
- Приватизация — http://uaport.net/cgi-bin/infostream.rss?rubr08
- Нормативные акты — http://uaport.net/cgi-bin/infostream.rss?rubr09
- Оборона, Конверсия — http://uaport.net/cgi-bin/infostream.rss?rubr10
- Официальная хроника — http://uaport.net/cgi-bin/infostream.rss?rubr11
- Криминал — http://uaport.net/cgi-bin/infostream.rss?rubr12
- Обзоры прессы — http://uaport.net/cgi-bin/infostream.rss?rubr13
- Связь — http://uaport.net/cgi-bin/infostream.rss?rubr14
- Экология — http://uaport.net/cgi-bin/infostream.rss?rubr15
- Энергетика — http://uaport.net/cgi-bin/infostream.rss?rubr16
- Медицина — http://uaport.net/cgi-bin/infostream.rss?rubr17
- Наука и техника — http://uaport.net/cgi-bin/infostream.rss?rubr18
- Компьютеры — http://uaport.net/cgi-bin/infostream.rss?rubr19
- Астрология — http://uaport.net/cgi-bin/infostream.rss?rubr20
- Культура — http://uaport.net/cgi-bin/infostream.rss?rubr21
- Катастрофы — http://uaport.net/cgi-bin/infostream.rss?rubr22
- Образование — http://uaport.net/cgi-bin/infostream.rss?rubr23
- Внешнеэкономическая деятельность — http://uaport.net/cgi-bin/infostream.rss?rubr25
- Масс-медиа — http://uaport.net/cgi-bin/infostream.rss?rubr26
- Калейдоскоп — http://uaport.net/cgi-bin/infostream.rss?rubr27
- Религия — http://uaport.net/cgi-bin/infostream.rss?rubr28
- Спорт — http://uaport.net/cgi-bin/infostream.rss?rubr29
- Туризм — http://uaport.net/cgi-bin/infostream.rss?rubr30
- Транспорт — http://uaport.net/cgi-bin/infostream.rss?rubr31
- Автотранспорт — http://uaport.net/cgi-bin/infostream.rss?rubr32
- Политика — http://uaport.net/cgi-bin/infostream.rss?rubr34
- Страхование — http://uaport.net/cgi-bin/infostream.rss?rubr35

3.9. Системы поиска RSS-фидов

Для нахождения RSS-фидов существуют многочисленные списки и каталоги, однако объемы существующих RSS-ресурсов таковы, что пользователям уже недостаточно десятка-другого рубрик первого уровня, имеющихся в каталогах. Как всегда, в подобных случаях на помощь приходят информационно-поисковые системы, которые позволяют находить как целые RSS-фиды, так и отдельные сообщения по ключевым словам. Поэтому в Internet появились поисковые сайты по RSS-фидам.

Одним из первых был создан сервис Feedster.com, который, кроме непосредственного поиска, позволяет подписаться на его результаты в формате RSS. В настоящее время Feedster обрабатывает 500 тысяч RSS-сообщений в сутки (рис. 3.5).

Еще одна поисковая система доступна на сайте http://Assimilatethe.net; она охватывает свыше 3500 RSS-ресурсов. Система ищет по заголовкам и описаниям RSS-сообщений. В базе данных системы Assimilatethe сейчас порядка 193 000 сообщений.

Как известно, RSS — самый распространенный формат для "живых журналов", т.е. блогов (сокращение от слова Weblog). Для поиска по блогам также существуют сотни каталогов и поисковых систем. Среди основных поисковых систем по блогам можно назвать следующие:

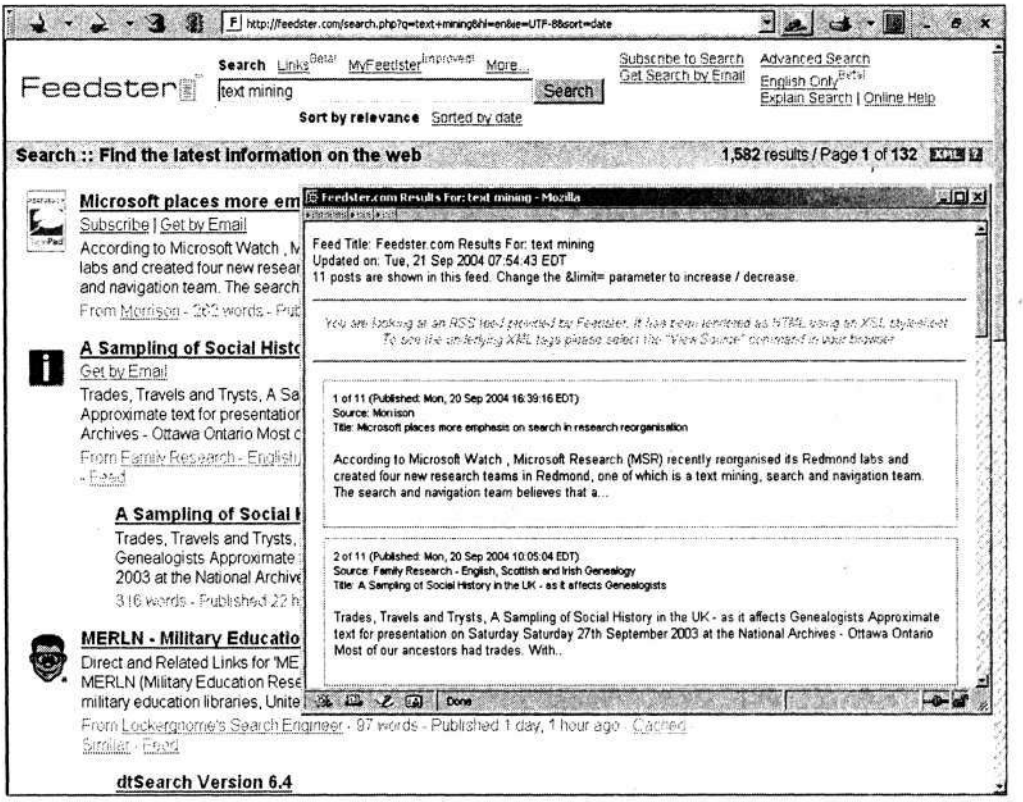


Рис. 3.5. Результаты поиска в системе Feedster

- DayPop — http://www.daypop.com
- Blog Search Engine — http://blogsearchegine.com
- Feedster — http://www.feedster.com
- BlogStreet — http://www.blogstreet.com
- Blogarama — http://blogarama.com/in.php?ID=2080
- Globe of Blogs — http://www.globeofblogs.com

- BlogDex — http://blogdex.media.mit.edu
- Weblogs.com — http://weblogs.com
- BlogWise — http://www.blogwise.com
- BlogHop — http://www.bloghop.com
- BlogUniverse — http://www.bloguniverse.com

3.10. Агрегаторы

Пользователи, конечно же, могут читать RSS-файлы с помощью стандартных Web-браузеров, что, однако, сопряжено с просмотром XML-разметки и полным отсутствием всякого оформления. Но за это и боролись создатели формата RSS. А вот для интерпретации этого формата существует бесчисленное множество программ, созданных, в основном, за последние два-три года. Это означает, что пользователи могут получить доступ к данным в формате RSS с помощью специальных программ. Эти программы называются *RSS-агрегаторами* и в наглядном виде отображают содержание RSS-фидов.

Программа-агрегатор позволяет собирать RSS-файлы с Web-сайтов, одновременно следить за появлением на них новостей и читать содержание этих новостей. Программы-агрегаторы (их еще называют RSS-парсерами) выполняют синтаксический разбор данных, представленных в формате RSS, после чего могут реализовать любые действия по отношению к этим данным, — например, отсылать их по электронной почте либо отображать на определенном Web-сайте. Сегодня наиболее популярны агрегаторы, позволяющие собирать RSS-данные с разных Web-сайтов вместе.

Feedreader

Feedreader — это свободно распространяемая программа для Windows (http://www.feedreader.com), позволяющая читать данные в формате RSS версий 0.9, 0.91, 1.0, а также различную информацию от таких систем, как Dublin Core и Slashback (стандарты описания метаданных информационных ресурсов Сети). Утилита очень удобна в использовании, обеспечивает работу с информацией на русском и украинском языках и обладает широким кругом сервисных возможностей (рис. 3.6). FeedReader версии 2.5 можно загрузить по адресу http://sourceforge.net/project/showfiles.php?group_id=70179; размер файла инсталляции — 1,2 Мбайт.

Feedreader — типичный RSS-агрегатор, интерфейс которого напоминает интерфейс почтовых программ. У пользователя, знакомого с почтовыми клиентами, работа с программой не вызывает затруднений. Остановимся подробнее на самых необходимых возможностях этой программы.

Для настройки подписки на RSS-фид пользователю следует активизировать опцию New и ввести следующую информацию:

- адрес RSS-фида;
- название фида (оно может быть определено пользователем);
- периодичность обращения к фиду на Web-сайте для обновления.

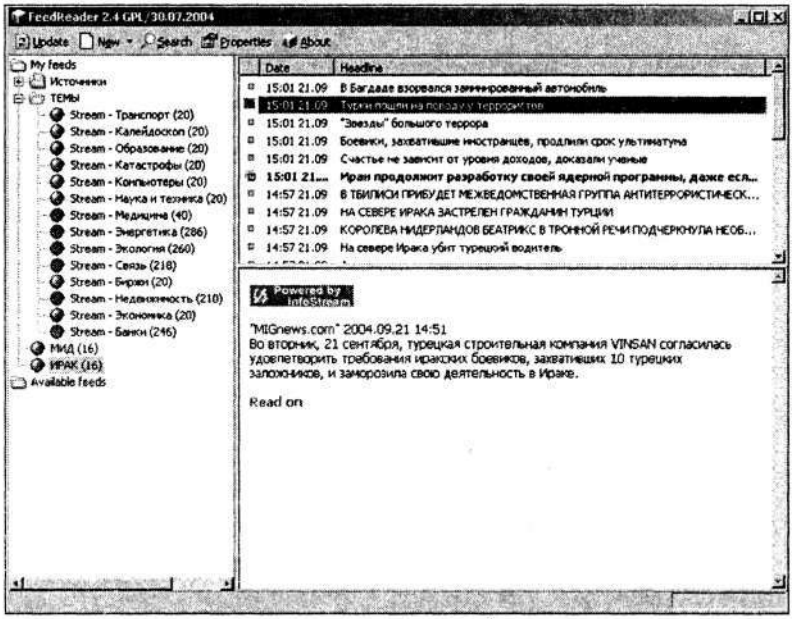


***Рис. 3.6.** Интерфейс агрегатора Feedreader напоминает почтовую программу*

При этом имеется возможность изменить кодировку, размер шрифтов, поместить фид в отдельную папку и сгруппировать фиды.

Для управления подпиской существуют дополнительные опции, доступные в контекстном меню, отображаемом при щелчке правой кнопкой мыши на конкретном фиде:

- обновление фида (списка активных сообщений);
- отметка всех сообщений как уже прочитанных;
- удаление списка сообщений;
- изменение свойств подписки, включая тему, периодичность и др.

Для получения полного текста сообщения (на которое есть ссылка — "link"), заголовок и аннотация которого вызвали интерес, следует выполнить одно из следующих действий:

- дважды щелкнуть левой кнопкой мыши на заголовке сообщения;
- щелкнуть на ссылке Read on в поле аннотации;
- щелкнуть на соответствующей кнопке, расположенной перед заголовком сообщения;
- щелкнуть правой кнопкой мыши на заглавии сообщения — в результате можно открыть текст сообщения в новом окне браузера;
- щелкнуть на ссылке на первоисточник, что позволит перейти на соответствующий сайт в Internet.

FeedDemon

Feed Demon представляет собой коммерческую программу (www.feeddemon.com), обеспечивающую удобную работу с RSS версии 2.0. Имеется возможность попробовать работу программы в "триальном" режиме. Утилита работает в среде Windows, корректно обращается с русской и украинской кодировками, обеспечивает поиск и фильтрацию информации фидов. Триал-версию FeedDemon 1.0 можно найти по адресу http://www.feeddemon.com/download/dloadhandler.asp?file=feeddemon-trial.exe; размер файла инсталляции — 2,3 Мбайт. Удобный интерфейс агрегатора позволяет легко отслеживать и читать свежие фиды. Feed Demon позволяет также представлять содержимое новостных лент в виде своеобразной газеты.

Приступить к использованию программы можно немедленно после инсталляции, так как пользователь сразу же начинает получать рассылки с сайтов Rollingstone.com, Scripting News, Sladshot, Wired, Yahoo! и др. Программа позволяет сохранять сообщения (News Bins) и отслеживать их по ключевым словам, запуская функцию Watches. Отдельные RSS-фиды можно перенаправлять в тематические списки или каналы. FeedDemon также позволяет проводить поиск и читать новости в автономном режиме.

Для подписки на фиды в программе следует ввести URL источника или импортировать файл OPML. Цена FeedDemon 1.0 составляет $29,95.

Abilon и ActiveRefresh

Это два агрегатора от одного производителя — компании ActiveRefresh (http://www.activerefresh.com/download.php). Бесплатная программа Abilon вполне подходит для среднего пользователя; она проста и надежна, отличается высокой скоростью и малой ресурсоемкостью (339 Кбайт). Она обладает возможностью закачки новых каналов с сайтов MoreOver, MyRss и NewsIsFree. Однако ей не хватает возможностей глобального поиска и сжатия информации.

В отличие от Abilon, ActiveRefresh — это платная программа, обеспечивающая полную реализацию концепций компании и позволяющая агрегировать обычные Web-сайты, импортировать с них новости, представленные в HTML, следить за почтовыми ящиками, проводить глобальный поиск и т.д.

Syndirella0.9b

Программа Syndirella может как показывать информацию с обычных Web-страниц, так и отображать данные, представленные в формате RSS. Программа реализована на платформе .NET, функционирует в среде операционных систем Windows и требует установки Internet Explorer версии 5.0 или выше. Для работы программы необходимо инсталлировать библиотеку Microsoft .NET Framework runtime версии 1.0 (20 Мбайт). Однако если этот компонент уже установлен, то сама программа Syndirella займет всего 250 Кбайт. Адрес для загрузки: http://www.yole.ru/projects/syndirella.

Прочие программы для среды Windows

Кроме перечисленных выше, сегодня большую популярность получили еще два агрегатора для работы под Windows — Awasu и Beaver. Особенность бесплатной программы Awasu (http://www.awasu.com) заключается в ее возможности объединять потоки множества новостных сайтов и блогов. Beaver (http://www31/brinkster.com/toolmaker) принимает фиды форматов RSS/RDF и имеет привычный интерфейс в стиле Outlook Express.

K.R.S.S. 2.6

KDE's Rich Site Summary viewer — приложение для Linux, позволяющее отображать данные в формате RSS на экране в виде HTML-страниц. Есть возможность настраивать способ отображения информации с использованием технологии CSS (Cascading Style Sheets) и устанавливать специальные фильтры новостей. Адрес для загрузки программы: `http://krss.sourceforge.net/downloads.html`; размер файла — 394 Кбайт.

Liferea

В последнее время для ОС Linux большую популярность приобрел агрегатор Liferea (`http://liferea.sourceforge.net`). Он поддерживает многочисленные форматы новостных фидов, основанные на XML, — такие как RSS, RDF, Atom, Echo, PIE, а также OCS и OPML для списков фидов. Эта программа распространяется с библиотекой GTK2.

Opera 7.5 и прочее

Норвежская компания Opera Software (`http://www.opera.com`) выпустила новую версию браузера Opera 7.5, в которой появился встроенный RSS-агрегатор. Доступ к нему организован через интерфейс почтового клиента.

В настоящее время создаются (и уже созданы!) многочисленные инструментальные средства для разработки программ работы с RSS-данными. Например, для разработки программ-парсеров на языке Perl создан модуль `XML::RSS`, который загружается с сайта `http://search.cpan.org`.

Встраиваемые в Internet Explorer инструментальные панели от Dogpile (`http://www.dogpile.com/info.dogpl/tbar`) и HotBot Desktop (`http://www.hotbot.com/tools/desktop`) поддерживают технологии RSS и Atom. С помощью этих средств заголовки сайтов, поддерживающих RSS, просматриваются прямо в окне браузера.

Одна из самых заметных особенностей интерфейса будущей версии ОС Windows-Longhorn заключается в наличии многофункциональной боковой панели (Sidebar). На нее может быть помещена любая информация — от часов и списка контактов до новостей, импортируемых в формате RSS. При этом средства настройки панели включены в состав инструментария разработчиков и поддаются настройке с их стороны.

3.11. Новые подходы

С помощью современной RSS-технологии пользователи Internet получили надежный и простой доступ к ресурсам оперативной информации с Web-сайтов Сети. Перспективность и популярность RSS как стандарта обусловлена, прежде всего, его доступностью и простотой. Сегодня практически все ведущие информационные сайты в мире и "живые журналы", работающие в Internet, используют RSS как инструмент оперативного представления обновлений своих ресурсов.

Еще один аспект применения RSS-технологии стал актуален в связи с массовым распространением невостребованных рассылок по электронной почте — спама. Действительно, электронная почта привлекательна и для спамеров. Нередко списки электронных адресов подписчиков новостей на сайтах и порталах становятся добычей взломщиков, что делает подписку через e-mail достаточно рискованным занятием.

Поэтому можно предположить, что на смену рассылкам придет использование RSS-фидов. В отличие от рассылок по электронной почте, где доставка инициируется администраторами сайтов после того, как подписчик оставил им свой адрес, в случае использования RSS пользователь сам вводит адрес необходимого ему RSS-фида в программу-агрегатор. Эта программа периодически проверяет, не изменилось ли содержание RSS-фида, и при наличии изменений автоматически закачивает его содержимое. Главным преимуществом RSS-технологии здесь является то, что пользователь сам принимает решение о получении каждого конкретного сообщения.

Все большую популярность RSS-технология приобретает у владельцев Web-ресурсов (не только новостных, но и коммерческих) еще и благодаря своей экономичности — не требуется никаких средств для борьбы со спамом, нет необходимости в фильтрации писем и в управлении рассылкой. При этом все, кому это необходимо, получают желаемую информацию о важных событиях, корпоративных анонсах, обновлениях Web-сайтов и пр.

Индустрия рекламы также не осталась в стороне от использования технологии RSS. Хотя RSS последних версий допускает вставку гиперссылок и изображений, однако как рекламный носитель она несколько уступает электронным письмам в HTML-формате. В настоящее время в Сан-Франциско создается первая он-лайновая рекламная сеть RSSAds, которая базируется на внедрении текстовой рекламы в заголовки RSS версий 0.90, 0.91, 1.0, 2.0 и Atom. Основателям этой сети удалось разработать систему подсчета рекламных показов — как только RSS-клиент обращается на сервер с запросом, система учитывает это событие. RSSAds планирует продавать рекламу, используя разнообразные модели оплаты: за количество показов, за время показов, за количество кликов, а также за размещение рекламы в заголовках.

Системы синдикации Internet-новостей решают проблему нахождения необходимой информации, но оставляют без внимания такие задачи, как обобщение данных — их обработку и анализ. Одним из самых перспективных направлений обобщения информационных потоков в настоящее время является метод "глубинного анализа текстов" (Text Mining). Применительно к новостным потокам его идеологию можно сформулировать как постоянное, воспроизводимое во времени выполнение их содержательного анализа. Непрерывная аналитическая обработка сообщений является самой характерной особенностью этого метода, который позволяет формировать автоматические дайджесты, выявлять новые понятия и их взаимосвязи, рассчитывать разнообразные рейтинги. Именно системы такого типа смогут избавить пользователей от дублирующейся информации, информационного шума, позволят выявлять главные тенденции, находить коррелирующие события. По прогнозам аналитической компании IDC, спрос на подобные системы существенно возрастет в течение ближайших 4-5 лет. Ожидается, что в 2005 году прибыль от продажи таких систем составит 1,5 млрд долларов США, а в 2006 году они будут доминировать в сфере анализа информации.

3.12. Информационные ресурсы для мобильных устройств

3.12.1. Wireless Application Protocol

Wireless Application Protocol (WAP) — это открытый протокол и технический стандарт, разработанный по инициативе фирмы "Unwired Planet" (сегодня ее название "OpenWave"). WAP обеспечивает передачу информации из Internet на

дисплеи мобильных телефонов [26]. Широкое применение этого стандарта привело к тому, что в настоящее время можно подключиться и обмениваться информацией через Internet непосредственно с мобильных телефонов, без посредничества компьютера. WAP существенно расширяет набор таких услуг мобильной связи, как обыкновенные звонки и короткие текстовые сообщения (SMS), позволяя внедрять сервисы, аналогичные тем, которые предлагаются в World Wide Web. Internet на экране мобильного телефона чем-то напоминает телетекст на миниатюрном телевизоре, однако интерактивность, т.е. возможность не только получать, но и вводить информацию, существенно расширяет сферу применения WAP.

Первые мобильные телефоны с поддержкой WAP появились на рынке в середине 1999 года. С 2001 года практически все производители выпускают мобильные телефоны с поддержкой WAP-протокола. По прогнозам бюллетеня Analysys, число пользователей мобильных устройств с поддержкой протокола WAP к 2005 году превысит 370 млн человек. В настоящее время для доступа к беспроводным сетям 66% предприятий пользуются мобильными ПК, 24% — карманными ПК, 21% — мобильными телефонными аппаратами с поддержкой протокола WAP.

Сегодня происходит широкое внедрение нового стандарта скоростного обмена данными в сетях мобильной связи — General Packet Radio Service (GPRS), согласно которому данные передаются пакетами по каналам, свободным от голосового трафика. Сети GPRS могут поддерживать максимальную скорость 107,2 Кбит/с (в то время как GSM — всего 9,6 Кбит/с). Поэтому они обеспечивают эффективную высокоскоростную работу в Internet и постоянное пребывание на линии. Трафик здесь подсчитывается не по времени соединения, а по объему информации, что дополнительно стимулирует к работе с лаконичными, информационно насыщенными ресурсами.

Надо отметить, что по протоколу WAP на мобильный телефон содержание Web-ресурсов сети Internet непосредственно не передается. В Internet информация представлена в виде HTML-страниц, работа с которыми предполагает быстрые коммуникации, мощные процессоры, большие объемы памяти компьютеров, большие экраны. Мобильные же телефоны обладают медленными процессорами, небольшими объемами памяти и совсем маленькими дисплеями. Поэтому для WAP используется свой специализированный язык разметки — Wireless Markup Language (WML), более простой и строгий, чем HTML. В WAP также применяются язык сценариев — WMLScript (упрощенная версия JavaScript), а также собственный формат растровых изображений — Wireless Bitmap (WBMP).

3.12.2. WAP-ресурсы

WAP — это, прежде всего, мобильная технология доступа к информационным ресурсам. Сегодня происходит бурное насыщение ниши сетевых WAP-ресурсов. Количество русскоязычных WAP-сайтов составляет уже несколько тысяч и постоянно увеличивается. Для ориентации в этих ресурсах уже мало традиционных каталогов типа Wapper (`http://www.wapper.ru/catalog`), MOBIL (`http://www.MOBIL.ru/wap.php` — рис. 3.7) или WapGate (`http://www.wapgate.ru`). Как и для англоязычных (всемирных) ресурсов, возникла необходимость создания "WAP-поисковиков", одним из лучших среди которых в настоящее время является "Яndex" (`http://wap.yandex.ru`).

Сегодня многие банки стараются предоставлять в режиме он-лайн тот же набор услуг, что и в обычных отделениях (денежные переводы, пополнение карточного счета, погашение потребительского кредита). Одна из главных тенденций развития

Internet-банкинга — его использование "в одном флаконе" с другими возможностями удаленного доступа: call-центром (когда клиент, пользуясь Internet-банкингом, тут же консультируется с сотрудником банка по телефону) и Mobile-банкингом (банковские услуги через мобильный телефон при помощи технологии WAP).

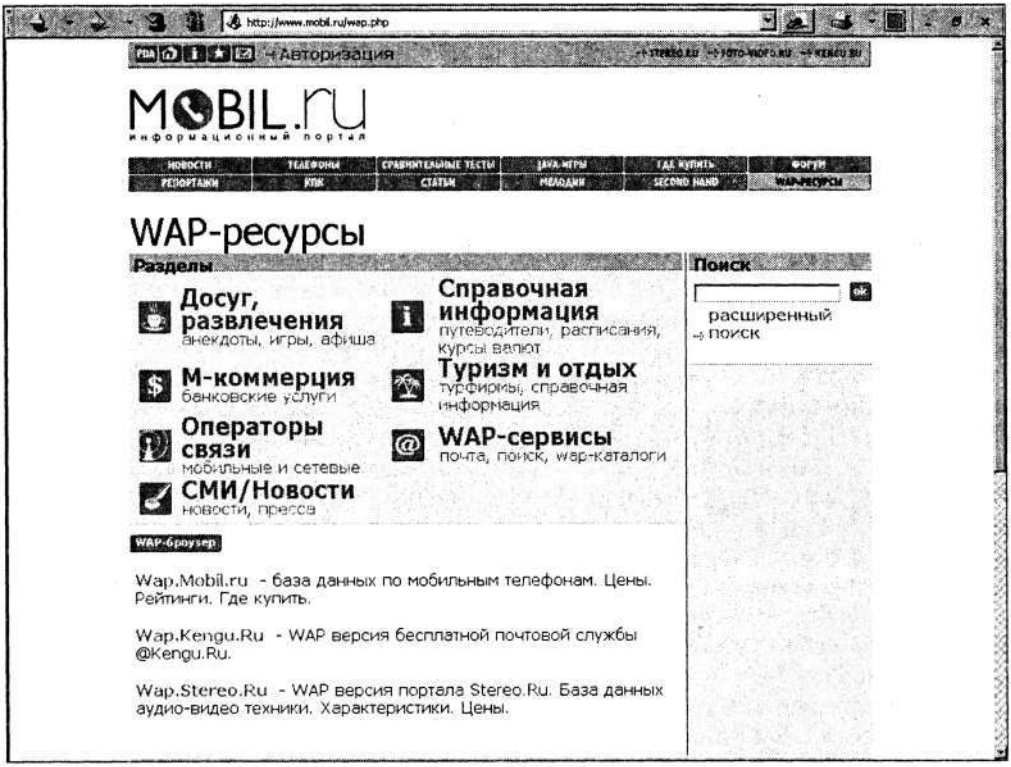


Рис. 3.7. MOBIL.RU — каталог русскоязычных WAP-ресурсов

Наиболее полезны и удобны услуги WAP, связанные с доступом к электронной почте. Благодаря им пользователь может в любой момент просмотреть свежую корреспонденцию на дисплее мобильного телефона. В последнее время такая услуга появилась у крупнейших бесплатных почтовых серверов (`wap.mail.ru`, `wap.imail.ru`, `wap.newmail.ru`). Наряду с электронной почтой, сервер `wap.beep.ru` предлагает виртуальный ежедневник, в котором можно распланировать свой день, а затем получать соответствующие напоминания.

Для ориентации в мире WAP можно воспользоваться поисковыми серверами — например, `wap.yandex.ru` или `www.wapgate.ru` (второй ресурс является обычным Web-сайтом, с которого можно, кроме того, просматривать WAP-ресурсы). На WAP-сайтах сегодня находится не так уж мало информации по различным темам. Так, поклонников карманных ПК наверняка заинтересует сайт `wap.handy.ru`. Любители игр в режиме on-line оценят сервер `www.wirelessgames.com/index.wml`, зарегистрировавшись на котором, можно поиграть прямо с экрана мобильного телефона. В мире WAP не забыты и такие традиционные для Internet средства общения, как chat (`wap.chat.ru`) и ICQ (`www.wapicq.com`).

Ниже приведен список еще нескольких интересных ресурсов средств массовой информации в русско- и украиноязычном WAP-пространстве.

- РосБизнесКонсалтинг — http://wap.rbc.ru. Новости политики и культуры, фондовые индексы, котировки акций.
- Polit.RU — http://wap.polit.ru. Новости российской политики.
- RevKom — http://wap.revkom.ru. Новости мира связи, коммуникаций, цифровых устройств.
- RosWeb — http://wap.rosweb.ru. Сборник последних международных и российских новостей.
- UAport — http://wap.uaport.net. Украинский интегратор сетевых новостей Internet-холдинг UAport.
- UAtoday — http://uatoday.net/wap. Газета "Украина сегодня" в WAP-формате. Новости в реальном времени.
- IT-News — http://wap.it-news.net.ua. Новости информационных технологий.
- Апорт — http://wap.aport.ru. Новости прессы, электронных СМИ.
- БиЛайн — http://wap.beeline.ru. Новости.
- Med-news — http://wap.bmed-news.kiev.ua. Страница медицинских новостей.
- Финансовые новости — http://www.audit-it.ru/wap. Налоговые новости, курсы валют от ЦБ РФ.
- GALA.NET — http://wap.gala.net. Новости.
- BBC-news — http://www.bbc.co.uk/mobile/mainmenu.wml. Деловая информация и новости, а также данные о котировках.
- Экономические новости — http://wap.mfd.ru. Экономические новости для финансистов и аналитиков.
- МТС — http://wap.mts.ru. Финансовые и бизнес-новости, биржевые индексы, новости РБК и АФП.
- Сетевая Лаборатория — http://wap.netlab.ru. Компьютерные новости.
- "Новая газета" — http://wap.novayagazeta.ru. Обзор и анализ событий в России. WAP-ресурс.
- Nursat — http://wap.nursat.kz. Новости агентства "Интерфакс-Казахстан".
- Аудит-ИТ — http://www.audit-it.ru/wap. Новости налогообложения, бухучета, аудита.

3.12.3. Реализация WAP-протокола

Схема работы WAP-сервисов включает три основных компонента: WAP-микробраузер, WAP-шлюз и WAP-сервер (рис. 3.8).

В качестве микробраузера может выступать мобильный терминал или программный эмулятор. WAP-шлюз взаимодействует с микробраузером, используя стек протоколов WAP. Шлюз переводит полученные от пользователя запросы, представлен-

ные в бинарном виде, в принятый в World Wide Web формат HTTP-сообщений. При этом провайдеры информации в качестве WAP-сервера могут использовать любой HTTP-сервер, например Apache, применяя все существующие наработки для создания сервисов и администрирования. Когда загрузка информационного блока с WAP-сервера завершается, WAP-шлюз компилирует элементы WML в компактную бинарную форму, что позволяет обеспечить большую скорость обмена информацией.

Рис. 3.8. Общая схема WAP-технологии

Кроме стандартных HTTP-серверов, для реализации WAP-серверов могут использоваться и специальные разработки, среди которых наиболее известен сервер компании Nokia, объединяющий функции сервера и шлюза.

В качестве коммерческих WAP-шлюзов наиболее известны продукты компаний Nokia и Ericsson. Nokia Artuse WAP Gateway, связывающий Internet (или intranet) и мобильные сети, обеспечивает для мобильной связи доступ к различным Internet-сервисам, а также позволяет применять устройства с поддержкой WAP для доступа к Web-приложениям.

Ericsson же предлагает два продукта — Ericsson WAP/Gateway Proxy и Jambala WAP Gateway. Оба пакета представляют собой комплексные решения для организации на базе операторов мобильной связи доступа к WAP-сервисам. Они включают сервер-шлюз с возможностью компиляции HTML-страниц в WML, интерфейс для WTA, поддержку SMS-шлюза.

Кроме продуктов от Nokia и Ericsson, существует несколько продуктов других производителей. Среди них наибольший интерес представляет проект Kannel, который был основан компанией Wapit Ltd. в 1999 году с целью разработки свободно распространяемого WAP-шлюза для UNIX-платформ. В данный момент на сайте Kannel (`http://www.kannel.org`) можно бесплатно получить полноценный WAP- и SMS-шлюз для Linux RedHat 6.1 или Debian с исходными кодами. Существует версия Kannel под Windows, адаптированная под эту ОС фирмой Wapme (`http://kannel.dev.wapme.net`).

3.12.4. WML и микробраузеры

Язык WML основан на модели описания языков XML (extensible Markup Language), поэтому первой строкой в любом файле должно быть указание на документ описания DTD (Document Type Definition) для данного языка. Официальная спецификация WML разработана и поддерживается WAP Forum, производственным консорциумом, основанным Nokia, Phone.com, Motorola и Ericsson. Эта спецификация определяет синтаксис, переменные и элементы, используемые в файлах формата WML. Исходя из этого, начало любого WML-документа должно выглядеть так:

```
<?xml version="1.0"?>
<!DOCTYPE wml PUBLIC "-//WAPFORUM//DTD WML 1.1//EN"
"http://www.wapforum.org/DTD/wml_1.1.xml">
```

Указанный DTD описывает все обязательные теги и элементы стандарта WML. Сам WML-код должен быть заключен в теги. Поскольку WML был разработан для устройств с низкой пропускной способностью и маленьким дисплеем, в качестве одной из составляющих используемого дизайна была применена концепция дек (колод) и карт. Отдельный WML-документ называется декой (deck), а интерактивное взаимодействие с пользователем осуществляется с помощью карт (card). Преимущество такой реализации заключается в том, что несколько экранов могут быть загружены в буфер мобильного телефона за один раз.

Карты можно рассматривать как прямую аналогию с HTML-страницами. Так же, как и HTML-страницы, карты имеют заголовок, определяемый параметром `title`, и внутренний идентификатор (`id`), по которому формируются гиперссылки на карту. На экране браузера может отображаться только одна карта, но поскольку она обычно достаточно мала по объему и на запрос ее уходит больше времени, чем на загрузку, близкие по смыслу карты объединяются в деки, что существенно ускоряет работу. Отдельный WML-файл представляет собой отдельную деку.

Описание текстовых элементов внутри карты начинается с парного тега "*абзац*". С этим тегом можно задавать следующие параметры: `align="`*выравнивание*`"` и `mode="wrapmode"`. Параметр выравнивания может принимать значения `right`, `center` или `left`, а параметр `mode` определяет, будет ли текст автоматически переноситься на экране (значение `wrap`) или задействуется скролинг (значение `nowrap`). Для форматирования текста может использоваться несколько тегов, аналогичных HTML: `<b> </b>`, `<i></i>`, `<u></u>`.

Стандарт WML предусматривает собственный формат изображений для отображения в браузерах — WBMP. Это двухцветное изображение с специфическим алгоритмом сжатия. Для преобразования обычных цветных растровых файлов в WBMP существует несколько средств: например, по адресу `www.rcp.co.uk/distributed/Downloads` можно получить плагин Photoshop для экспорта в этот формат. На Web-странице `http://www.netec.de/downloads/wap_pictus.htm` приведена программа WAP Pictus для конвертирования изображений в формат WBMP. Кроме того, конвертор изображений в формат WBMP содержится, например, в WAP-эмуляторе для ПК Deck-it WAP Previewer. Для присоединения изображения к карте служит тег `img` с параметрами, идентичными аналогичному HTML-тегу.

На сегодняшний день существует несколько микробраузеров, т.е. программ, встраиваемых в мобильные телефоны для работы по WAP-протоколу с WML-документами. Самый известный микробраузер создан компанией Unwired Planet (UP); он применяется в WAP-телефонах Motorola, Alcatel, Samsung. Nokia и Ericsson имеют собственные решения; кроме того, существуют модели телефонов (Benefon, Sony), использующие микробраузер корпорации Microsoft — Mobile Explorer.

Недавно представители компании Nokia заявили о том, что, крупнейший производитель модулей памяти для компьютеров, компания Samsung Electronics согласилась лицензировать исходный код браузера для мобильных устройств и технологию обмена сообщениями Smart Messaging. Оригинальная версия программы поддерживает WML 1.3, XHTML и WAP CSS и отвечает спецификациям WAP Forum и W3C. Использование реализованных в программном продукте технологий позволит удовлетворить всем требованиям заказчиков. Браузер для мобильных устройств от Nokia представляет собой платформо-независимый продукт, отвечающий требованиям OEM.

Существующие микробраузеры имеют свои особенности. Если говорить о параметрах, имеющих значение при разработке сервисов, то, прежде всего, для разных микробраузеров следует учитывать различия в максимальных размерах загружаемого за один сеанс блока информации (деки). Кроме того, не везде решены проблемы русификации (не говоря уже об украинизации) микробраузеров. Следует отметить, что в настоящее время в этой области повсеместно применяется кодировка символов `utf-8`, что, однако, не учитывается, например, во многих телефонах от Motorola. Для обеспечения работы с этими моделями информационным провайдерам до сих пор приходится применять транслитерацию в WML-документах.

3.12.5. Эмуляторы WAP

Доступ к WAP-ресурсам возможен не только с мобильных телефонов. Оперативная и лаконичная информация WAP-сайтов доступна также с помощью программ-эмуляторов для ПК, например Deck-It (`www.pyweb.com`), M3Gate (`www.numeric.ru/m3gate/r_index.htm`), WinWAP (`www.winwap.org`). Эти программы позволяют просматривать WAP-серверы с обычных компьютеров, подключенных к Internet. Кроме того, просмотр WAP-серверов возможен и с помощью специализированных Web-сайтов сети Internet (как правило, для этого в специальное поле эмулятора достаточно ввести адрес WAP-сайта).

Универсальный Web-браузер Opera на данный момент — единственный продукт данного класса, поддерживающий протокол WAP и обеспечивающий просмотр страниц, написанных на языке WML.

Одной из лучших программ-эмуляторов WAP для ПК является Klondike WAP Browser Personal Edition 1.5 компании Apache Software (рис. 3.9). Этот эмулятор позволяет посещать WAP-сайты и сайты, автоматически перекодируемые сервером с HTML в WML. С помощью Klondike WAP Browser можно также провести тестирование любого WAP-сайта в режиме офф-лайн.

Эмулятор M3Gate 0.5 компании Numeric Algorithm Laboratories поддерживает работу по стандарту WAP 1.1 с WML и WMLScript, изображениями WBMP и PNG, различными кодировками, позволяет отлаживать WAP-приложения, просматривать WML-код документов. M3Gate интегрируется с Internet Explorer или Netscape Navigator и автоматически запускается при попытке открыть WAP-ресурс. WinWAP — простой в использовании WAP-эмулятор, который работает через HTTP и WAP-шлюзы. С его помощью можно просматривать исходный код, переменные, Cookie. Включает в себя множество стандартных для Web-браузеров возможностей: печать, поддержку технологии drag and drop, использование закладок, изменение размера окна, поиск текста и др. Имеется версия для WinWAP Pocket PC.

Практически все производители мобильных телефонов имеют свои версии WAP-эмуляторов для персональных компьютеров.

Достаточно обширный список WAP-эмуляторов для ПК можно найти в Internet по адресу `http://www.wapgate.ru/soft/?s=2`.

Кроме программных WAP-эмуляторов (их еще называют WAP-браузерами), можно воспользоваться специальными Web-серверами, доступными с обычного компьютера, подключенного к Internet. Эти сайты (например, `www.wapgate.ru`) представляют возможность просмотра WAP-страниц путем указания WAP-адресов в специальных окнах ввода. Сегодня такие сайты открывают перед пользователями Internet новый срез информационных источников — кратких и оперативных, доступных к тому же с помощью привычного интерфейса.

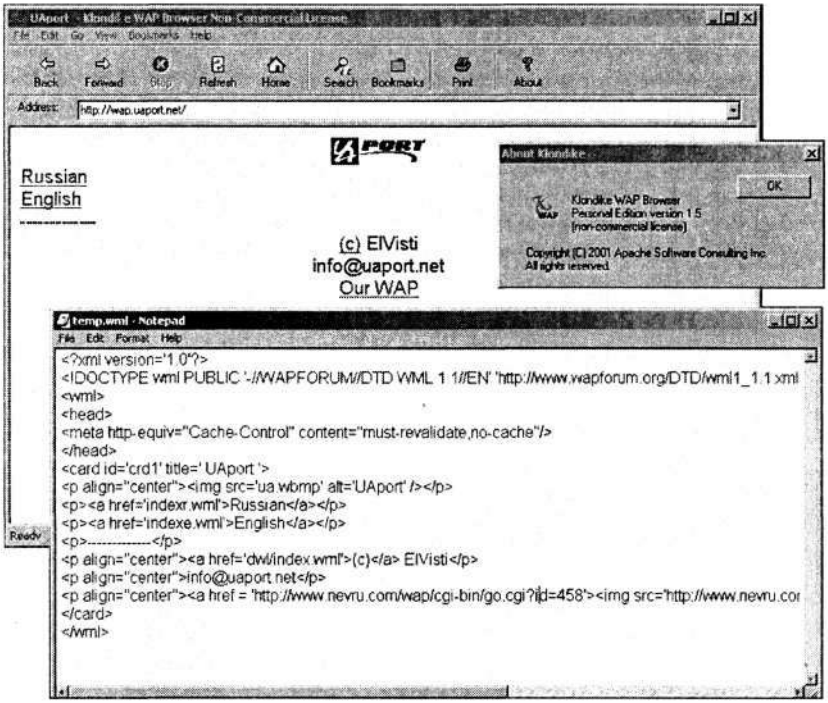


Рис. 3.9. Исходный текст WML, вызванный из эмулятора Klondike WAP Browser

Рис. 3.10. Wapsilon и его "скин" — телефон Nokia 7210

WAP-эмулятор Wapsilon v 2.4, который представлен на сайте wapsilon.com, ориентирован, прежде всего, на Nokia 7210 (рис. 3.10).

WAP-эмулятор Gelon (разработчик Gelon.net) обеспечивает отображение WML-страниц с помощью обширного набора "скинов": Ericsson R320, Nokia 7110, Nokia 6210, Siemens C35, Siemens M35, Siemens S35, Motorola A6188, Motorola P7389 (рис. 3.11).

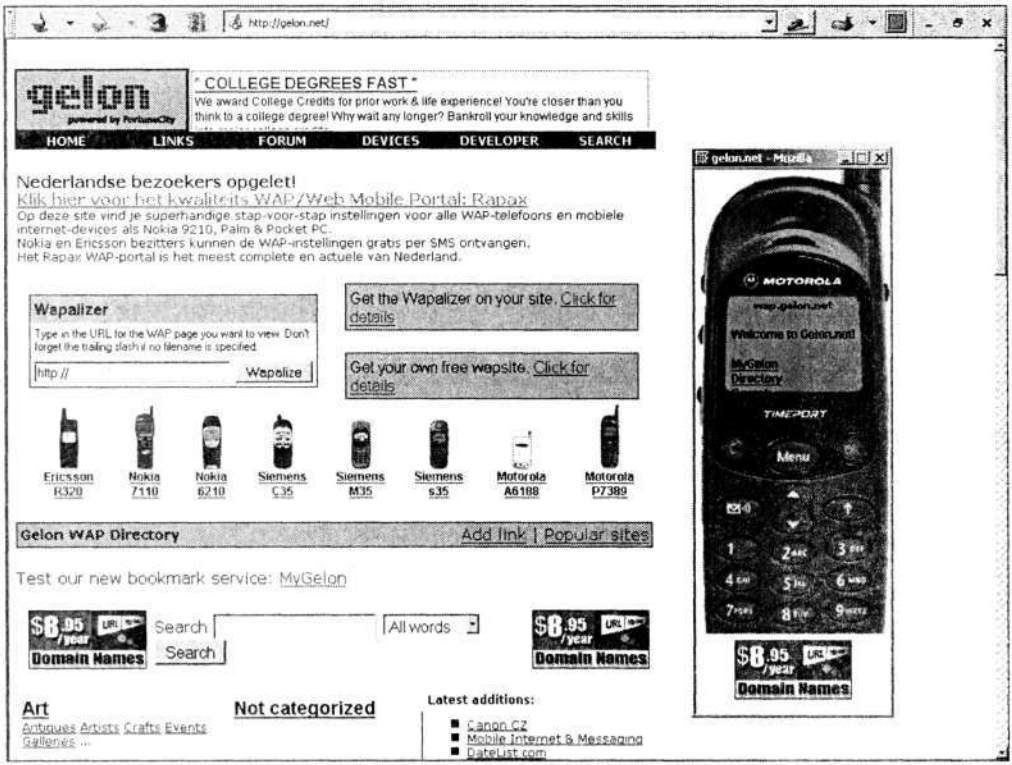


Рис. 3.11. Gelon — сайт-эмулятор WAP с множеством "скинов"

Еще один WAP-эмулятор для телефонов Nokia 3310, 6210, 7110, 9110, построенный на основе Java-аплетов, можно найти на украинском сайте "Мобильный портал", находящемся по адресу http://www.mobile-portal.kiev.ua/wap.phtml?t=emul.

Эмулятор, расположенный по адресу TatTag.com, обладает способностью настраиваться на различные кодировки, что очень удобно при работе с многочисленными вариантами кириллических кодировок (рис. 3.12).

3.12.6. Проблемы и перспективы WAP

На сегодняшний день существует несколько проблем, ограничивающих распространение WAP-технологии. Прежде всего, такая связь с Internet остается дорогой и медленной. Кроме того, прогресс слабо коснулся возможностей отображения информации на небольших дисплеях мобильных устройств. Через год после рождения WAP стали появляться сомнения в наличии достаточного числа

потребителей, необходимого для окупаемости данной технологии. Абоненты сотовых сетей все еще редко используют свои телефоны для выхода в Internet. Компания Nielsen Norman Group провела в Лондоне "драйв-тест" технологии WAP, выдав добровольцам телефоны с WAP. По результатам исследования, Nielsen Norman рекомендовала операторам сотовой связи следующее: "Не тратьте деньги на внедрение тех услуг, которыми никто не будет пользоваться. Мы документально зафиксировали, что сфера применения WAP весьма узка".

***Рис. 3.12.** Эмулятор TagTag.com*

На смену телефонам с поддержкой WAP приходят другие компактные устройства, поддерживающие иные принципы работы в Internet. Например, компания Interactive Intelligence объявила о создании нового протокола передачи данных Mobilite 1.0, который по всем основным показателям опережает WAP и позволяет развивать и реализовать корпоративные решения с использованием карманных ПК, работающих под управлением Pocket PC, Palm OS и других систем. По словам Дональда Брауна, исполнительного директора Interactive Intelligence, протокол WAP прекрасно подходит лишь для решения относительно простых задач, поскольку обладает ограниченными функциональными возможностями. В отличие от WAP, программное обеспечение протокола Mobilite поддерживает все известные платформы, на которых работают карманные ПК.

Кроме того, во всем мире сейчас началось внедрение технологии скоростного обмена данными в сетях мобильной связи — General Packet Radio Service

(GPRS), с помощью которой можно работать в Internet, принимать почту, отправлять цифровые фотоснимки, играть в он-лайновые игры. Причем все это относительно недорого. Трафик в GPRS подсчитывается не в привычных для абонента минутах соединения, а в байтах — за объем переданной информации. Первым мобильным аппаратом, поддерживающим GPRS, была Motorola P7386i. Из удачных моделей стоило бы выделить Siemens S45, ME45, Nokia 8310 и несколько моделей от Ericsson.

Несмотря на критику WAP, этой технологии суждено стать стандартом "де-факто" — просто потому, что ею уже пользуется очень много людей. В принципе, у WAP-технологии неплохие перспективы, так как в комитет по разработке входят такие компании, как Alcatel, Matsushita, Swisscom Motorola, Nokia, Philips, Qualcomm, T-Mobil Samsung, Intel, NEC, Siemens, Fujitsu, IBM, Ericsson Psion Software, AT&T Wireless Services, BellSouth Cellular Corporation, Sonera, Telenor, Telstra, Vodafone, BT Cellnet, Sprint PCS, Telia Mobile. Большинство аналитиков сходятся во мнении, что в течение нескольких лет (где-то 2003–2005 гг.) число мобильных телефонов с доступом к Internet в мире превысит число подсоединенных к Internet ПК. Согласно прогнозу компании Strategy Analytics, уже в 2003 году 95% мобильных телефонов, выпущенных в США и Западной Европе, будут способны работать по протоколу WAP. Только в 2003 году в эфире было около миллиарда цифровых мобильных телефонов и новых WAP-оптимизированных устройств.

В настоящее время становится актуальной задача создания мобильных сетей третьего поколения (3G), которая заключается не только в разработке мобильных телефонов лучшего качества. Каналы связи 3G должны обеспечивать работу мультимедийных приложений, включая видеотелефонию (компания Orange уже работает над подобным проектом), видео по требованию и другие формы широкополосной связи. Менеджер компании Ericsson по связям с общественностью Питер Бодор (Peter Bodor) считает недостаточный успех WAP следствием того, что эта технология обманула ожидания пользователей, чего с 3G быть не должно. "Разочарование в WAP вызвано неспособностью индустрии оправдать ожидания, причем главной проблемой стала ее медлительность, — говорит он. — С 3G работать в Internet будет так же легко, как и из дома, к тому же локально-ориентированные услуги сделают эту работу более персонифицированной."

Недавно более десятка ведущих коммуникационных компаний, включая AT&T Wireless, Cingular Wireless, mm02, NTT DoCoMo, Telefonica Moviles, Vodafone, Fujitsu, Matsushita, Mitsubishi Electric, Motorola, NEC, Nokia, Samsung, Sharp, Siemens, Sony Ericsson, Toshiba и Symbian, объявили об инициативе, направленной на создание максимально "однородного" рынка услуг мобильной связи. Данная инициатива предполагает, что ее участники будут разрабатывать свое ПО для мобильных систем следующих поколений (включая терминальные клиентские модули и серверные решения) строго в соответствии со спецификациями основных органов по стандартизации, таких как 3GPP.

Несмотря на прогнозы скептиков, предсказывавших уход WAP с рынка еще в 2003 году и рассматривающих эту технологию как переходную, на июнь 2004 года зафиксирован значительный рост числа WAP-сайтов (эти данные предоставила Mobile Data Association). Их количество выросло на 42% по сравнению с аналогичным периодом 2003 года и достигло 1,1 млн. Наибольшей популярностью сегодня пользуются WAP-сайты и порталы крупных мобильных операторов, основанные на WAP-технологии.

С помощью мобильных устройств с WAP-протоколом пользователи уже сегодня получают надежный и простой доступ к ресурсам сети Internet, справочной и оперативной информации, поступающей со всего мира.

3.12.7. Доступ к сетевому контенту с КПК

В течение нескольких последних лет во всем мире продолжается бум беспроводных (wireless) технологий. Естественно, не последнюю очередь в этом играют мобильные устройства [19]. На сегодня известно два основных типа устройств, применяемых в беспроводных технологиях. Во-первых, это мобильные телефоны, во-вторых — "наладонники", они же КПК (карманные персональные компьютеры) или PDA (Personal Digital Assistent) двух конкурирующих типов — Palm и Pocket PC.

Широкое распространение уже в течение нескольких лет получили наладонники, представляющие собой, как правило, бесклавишные компьютеры карманного формата, которые обладают такими же потенциальными возможностями, как и обычные персональные компьютеры (ПК), если не принимать во внимание несколько непривычные и, откровенно говоря, весьма ограниченные возможности ввода данных, а также небольшие дисплеи этих устройств. Рабочие части экранов КПК имеют разрешение всего 160x160, 240x320 или, что пока очень редко, 320x320 точек (монохромных или цветных). Сенсорные экраны КПК позволяют осуществлять рукописный ввод информации. Например, у КПК Palm часть экрана занимает область, где можно рисовать символы, которые автоматически преобразуются в распознанные буквы. Практически для всех КПК разработаны и "крупногабаритные" клавиатуры, которые, однако, не входят в базовые поставки наладонников.

Что же касается конкуренции различных моделей, то можно отметить, что среди обычных пользователей наиболее популярны сегодня Palm и его клоны, а среди корпоративных пользователей все большую популярность получает модель Pocket PC, которая гарантирует полную совместимость с серверами, настольными компьютерами и корпоративными сетями. Обеспечивая завидное единообразие Pocket PC, корпорация Microsoft установила жесткий контроль над дизайном, процессорами, размерами и разрешением дисплеев.

В то же время американский журнал Business Week отмечает, что в 2003 году именно клоны Palm (КПК, работающие на базе операционной системы Palm OS) стали самыми новаторскими (Tapwave Zodiac, Sony CLIE PEG-UX50, Handspring Treo 600).

С КПК — в Сеть

Как известно, есть, как минимум, две основные причины, тормозящие широкое распространение доступа в Internet с мобильных устройств. Во-первых, скорость передачи данных не всегда высокая, и, во-вторых, это неудобства, возникающие при просмотре Web-страниц, созданных для больших экранов.

Известно, что для мобильных телефонов, работающих по протоколу WAP, разработчики соответствующих сайтов реализуют структурирование и сокращение объемов данных, которые приходится загружать на мобильные устройства. Этот же принцип используется и в методике Web-клиппинга (Web clipping) для наладонников.

Существуют два подхода к Web-клиппингу — серверный и локальный. В случае серверного клиппинга, который в настоящее время применяется реже, используется специализированный прокси-сервер, который принимает запросы со

стороны Palm, получает данные с Web-сайтов, а затем пересылает "карманнику" сжатый ответ. С помощью специального приложения пользователь генерирует локальный запрос в специальном формате PQA на получение данных из Сети. Этот запрос передается на прокси-сервер, который определяет, что именно требуется, переходит на соответствующие Web-сайты и извлекает необходимые данные. Прокси-сервер сжимает данные и пересылает их обратно на Palm, где ответ отображается с помощью приложения обработки запросов.

Локальный Web-клиппинг реализуется программными приложениями на самих Palm. Например, недавно выпущен программный пакет Mobile Internet Kit (MIK), позволяющий использовать PDA для работы с множеством ресурсов Сети. Он включает около 450 программ, в том числе реализующих Web-клиппинг для доступа с Palm к таким ресурсам, как `Amazon.com`, `Britannica`, `MapQuest.com` и многим другим. Кроме того, MIK включает приложения для полноценной работы с электронной почтой, SMS-сервисом, WAP-сайтами.

Корпорация Microsoft выработала собственный подход, предложив новое программное обеспечение для Pocket PC — Pocket Internet Explorer для Pocket PC. Это решение идеологически близко к локальному Web-клиппингу. Браузер Pocket Internet Explorer, входящий в состав ПО Pocket PC корпорации Microsoft, предлагает доступ к содержимому оригинальных страниц Web-сайтов, поскольку способен переформатировать страницы так, чтобы они наилучшим образом отображались на цветном экране размером 320x240 пикселей, которым оснащен Pocket PC.

3.12.8. Информационные ресурсы для КПК

С развитием технологии GPRS стали активно разрабатываться и КПК-версии самых разнообразных Web-сайтов Internet. При создании таких ресурсов всегда учитываются две основные особенности наладонников — ограниченные размеры экрана и скудные возможности клавиатурного ввода. Интерфейс типичного PDA-сайта, как правило, представляет собой основное меню, перемещаясь по которому, посетитель выбирает ту позицию (раздел, тематику или группу товаров), которая его интересует. Несмотря на это, создатели как правило стремятся к тому, чтобы PDA-версии принципиально не отличались от Web-сайтов по своему наполнению (контенту).

Рис. 3.13. Мобильный портал от Yahoo!

В последнее время все основные мировые поисковые системы и интеграторы контента представили PDA-варианты своих Web-сайтов (например, Google, Yahoo, Яndex, Lenta.ru, Newsru и многие другие).

PDA-поисковик от Google, оптимизированный под Palm, находится по адресу `http://www.google.com/palm`. Правила и условия работы с этой информационно-поисковой системой приведены по адресу `http://www.google.com/wireless/pda.html`.

Другой популярный поисковый портал Yahoo! также имеет PDA-версию, расположенную по адресу `http://wap.oa.yahoo.com` (рис. 3.13).

У пользователей имеется возможность персонализации, работы с электронной почтой, получения оперативных новостей, информации о финансах, спорте, погоде и кинофильмах.

PDA-вариант поисковика Яndex находится по адресу `http://yandex.ru/yandsearch?useie5=1`. Основной интерфейс представляет собой окно для ввода запроса. В результате обработки запроса на PDA-сайте выводятся заголовки релевантных документов (по 10 документов на страницу), которые представляют собой гиперссылки на соответствующие Web-ресурсы.

Известный российский новостной сайт LENTA.RU имеет PDA-версию, размещенную по адресу `http://pda.lenta.ru`, газета "Известия" — `http://pda.izv.info`, питерский филиал "Московского комсомольца" — `http://www.mk-piter.ru/pda`, газета "Деловая панорама" — `http://pda.dpw.ru`.

Создаются и специализированные поисковые сайты, ориентированные на пользователей КПК, посвященные карманным компьютерам, программам для них, мобильным устройствам и технологиям. Так, в России поисковая система PDANewsCollector.ru дополняет поисковый сервис возможностями оперативного просмотра новостей большинства популярных российских и украинских сайтов, посвященных тематике мобильных устройств.

Обширный список русскоязычных ресурсов для КПК от Антона Носика находится по адресу `http://pda.lenta.ru/info/pdalinks.htm`. Приведем лишь некоторые, наиболее информативные из них.

- `pda.utro.ru` — Утро.Ру, карманная версия.
- `palm.newsru.com` — новости от NewsRu.Com.
- `pda.gismeteo.ru` — погода от GISmeteo.Ru.
- `r0.ru` — облегченный Рамблер.
- `pda.mail.ru` — карманный Mail.Ru.
- `r0.ru/_lgmail.html` — карманная почта Рамблера.
- `www.google.com/palm` — поиск в Google для PDA.
- `www.hpc.ru/pda/links` — каталог сайтов для КПК/PDA
- `www.77.ru/ppc.php` — карта Москвы и адресный поиск.
- `www.pocketpcrussia.com/pda` — форум для владельцев КПК/PDA.
- `pda.translate.ru/?lang=ru` — он-лайновый переводчик.
- `pda.eda-server.ru` — информационно-поисковый сервер по продуктам питания.
- `pda.lenta.ru` — версия новостного сайта Lenta.ru.
- `pda.izv.ru` — газета "Известия".
- `www.mk-piter.ru/pda` — питерский филиал газеты "Московский комсомолец".
- `pda.dpw.ru` — газета Деловая панорама".

Еще один интересный ресурс, адаптированный под наладонники, — это он-лайновый переводчик, который находится по адресу `http://pda.translate.ru/?lang=ru`.

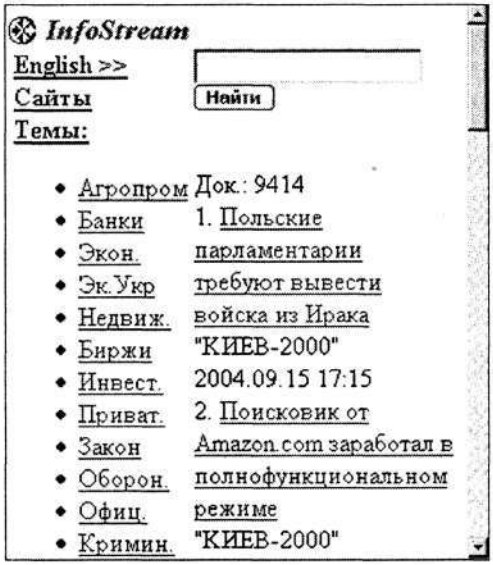


Рис. 3.14. Новостной сайт *pda.uaport.net*

Появляются и Internet-магазины, ориентированные на "мобильных" клиентов. Так, например, недавно в России компания Topman.Ru открыла PDA-версию своего магазина — www.pda.topman.ru. В пресс-релизе, посвященном этому событию, особое внимание обращается на такой технологический момент: "клиенту, посещающему электронный магазин с помощью КПК, не нужно использовать прокрутку, усложнявшую процедуру ознакомления и выбора товара".

Один из крупнейших русскоязычных новостных ресурсов для КПК размещен по адресу http://pda.uaport.net (рис. 3.14). Сайт предназначен для предоставления пользователям мобильных устройств оперативной информации и представляет собой PDA-версию новостного раздела поискового портала UAport (http://infostream.com.ua/rss). Оригинальная технология InfoStream-клиппинга обеспечивает просмотр с экрана мобильных устройств информации с сотен Web-сайтов сети Internet. Доступ к этому информационному PDA-сайту бесплатный и не требует регистрации.

3.12.9. Эмуляция мобильности

Информация сайтов, предназначенных для КПК, доступна как с помощью многочисленных программ-эмуляторов для ПК, так и непосредственно с экранов Web-браузеров. Действительно, бывает полезным поработать с компактной информацией, представленной на WAP- или PDA-сайтах, находясь в стационарных условиях у монитора мощного компьютера. Эмуляторы полезны разработчикам специализированных Internet-устройств, программистам КПК, а также потенциальным покупателям, выбирающим своего "карманного питомца".

Для наладонников типа Palm существует ряд эмуляторов, однако имеется лишь один несомненный лидер среди программ данного класса — Palm OS Emulator (POSE), разработка которого в свое время патронировалась Palm Inc. (рис. 3.15). Программа свободно распространяется и пригодна для разнообразных Palm-платформ. Исходные коды Emulator Application 3.5 для Windows, Mac OS, Unix доступны на sourceforge.net/rojects/pose. (Скачать их также можно, например, с palm.com.ua/?files&id=913.) С помощью POSE легко моделируется несколько десятков разновидностей КПК, при этом для работы программе требуются "образы" прошивки (ROM) соответствующих КПК. Для эмуляции внешних образов наладонников предусмотрено использование разнообразных "скинов". В процессе работы на виртуальном КПК можно инсталлировать практически любые программные коды, предназначенные для работы на соответствующих Palm.

В последних версиях Web-браузера Opera под Windows и Linux (начиная с седьмой) реализована технология Small-Screen Rendering (`www.opera.com/products/smartphone/smallscreen`). При нажатии клавиш <Shift+F11> включается режим эмуляции экрана мобильного устройства. Как и в случае с WAP-сайтами, в этом браузере обеспечивается полноценная эмуляция экранов КПК. При этом исключается так называемая "горизонтальная прокрутка", являющаяся настоящим бичем в случае просмотра Web-ресурсов с мобильных устройств.

В отличие от WAP-сайтов, информация которых адекватно не отображается основными браузерами настольных ПК, страницы сайтов для наладонников представлены в формате HTML (пусть и не всегда последней версии; например, большинство браузеров для Palm-устройств корректно интерпретирует лишь HTML версии 2). Браузер Pocket Internet Explorer настолько приближен к IE для ПК, что практически исключает потребность в использовании соответствующего эмулятора.

Рис. 3.15. Эмулятор POSE

3.12.10. RSS-формат на КПК

Технологию RSS News на сегодняшний день поддерживает все большее количество новостных Web-сайтов. Напомним, что формат RSS, являясь унифицированным подмножеством современного языка разметки XML, обеспечивает согласованный способ резюмирования содержимого Web-сайтов. Его применение предоставляет администраторам сайтов новостей, он-лайновых дневников (weblog), форумов и других часто обновляемых Web-ресурсов простой метод подачи информации о происходящих событиях.

RSS можно рассматривать и как формат, ориентированный, прежде всего, на публикацию и обеспечение экспорта новостей. После того как информация преобразована в формат RSS, ориентированная на него программа может загружать новые версии Web-сайтов и выполнять определенные действия, например автоматически обновлять список актуальных информационных сообщений. Такие программы называют RSS-агрегаторами. Они выполняют синтаксический разбор (парсинг) данных, представленных в формате RSS, после чего могут реализовать любые действия по отношению к этим данным, опираясь на полученные результаты. Данные в формате RSS представляют собой каналы или фиды (feed-файлы), в которых записывается новостная информация Web-сайта. Соответственно, если есть необходимость оперативно отслеживать изменения на Web-сайте, не посещая его, то можно подписаться с помощью программы-агрегатора на этот фид (конечно, если он существует).

Широчайшее распространение формата RSS (на одном только сайте NewsIsFree представлено свыше 6 тыс. каналов с различных, в том числе российских и украинских, новостных сайтов) обусловило появление множества программ-агрегаторов, ориентированных на КПК. Владельцы же КПК, установив на свои устройства RSS-агрегаторы, могут эффективно просматривать новостные файлы в RSS-формате.

Для платформы Palm OS наиболее популярной, пожалуй, является программа Hand RSS компании Stand Alone. С помощью этого "карманного агрегатора" владельцы устройств, работающих под управлением Palm OS, могут читать новости в формате RSS с новостных лент BBC News, CNet News, Fark.com, MSNBC, Salon.com, The Register, Wired и т.д. Скачать демоверсию программы или купить ее можно по адресу `http://www.standalone.com/cgi/prc_request.cgi`.

В качестве еще одного эффективного агрегатора можно назвать программу Quick Palm RSS Reader (`remus.manilasites.com`). Из специализированных для Pocket PC можно назвать агрегатор новостей в RSS/RDF PocketFeed (`http://www.furrygoat.com/Software`). Пятнадцатидневную демоверсию еще одной программы для этой платформы (PocketPC 2002 и Windows Mobile 2003) — PocketRSS 1.3 — можно скачать на сайте `http://www.happyjackroad.com/AtomicDB/pocketpc/pocketRSS/pocketRSS.asp`.

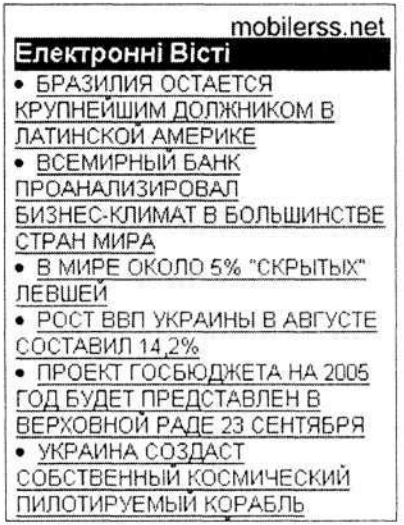


Рис. 3.16. Эмулятор MobileRSS

Но вовсе необязательно устанавливать программу-агрегатор прямо в наладоннике. Как и для WAP-сайтов, в этом случае также существуют серверные решения, выполняющие всю работу по интерпретации RSS-фидов и преобразованию результатов в формат, пригодный для КПК. Один из лучших сайтов подобного назначения — MobileRSS (`mobilerss.net` — рис. 3.16).

Для работы с этим бесплатным сервером необходима формальная авторизация. Зарегистрированный клиент вводит и активизирует адреса необходимых ему RSS-фидов, после чего просматривает их в свободном режиме. Этот зарубежный сервис, помимо прочего, обеспечивает и корректную работу с кириллическими шрифтами.

3.12.11. Игрушка или рабочий инструмент

По многим причинам (прежде всего, благодаря внедрению технологии GPRS, обеспечивающей скоростной доступ в Internet) наладонники завоевывают внимание всех целевых групп — от детей и домохозяек до корпоративных клиентов. Так, число пользователей КПК в Европе составляет 4% от всего населения. Производители "карманников" за счет недополучения прибыли пытаются развить рынок КПК, захватить на нем лидирующие позиции. В минувшем году они поставили своей целью обеспечить продажу полнофункциональных карманных устройств по цене ниже $300. Так, современная модель Palm Zire продается всего за 130 евро. В Palm Inc. полагают, что такой агрессивный маркетинг позволит удвоить число

европейцев, пользующихся КПК. Hewlett-Packard объявила о выпуске новых моделей iPAQ, работающих под управлением ОС Pocket PC производства Microsoft (рис. 3.17). Ожидается, что розничная цена iPAQ h1910 в США составит $299. Чуть раньше о намерении выпустить свой наладонник на основе Pocket PC заявила компания Dell. Ее детище будет продаваться по демпинговой цене $199.

Во многих странах наладонники широко применяются различными ведомствами. К примеру, в армии США стандартом являются КПК Palm. В московской милиции с помощью КПК обеспечивается доступ к различным информационным массивам, включающим текстовую и графическую информацию ("Розыск лиц", "Паспорта", "Оружие", "Угон" и др.).

Сегодня стало действительно велением времени получение мобильного, оперативного и дешевого доступа к актуальной информации. С помощью PDA-устройств, таких как Palm или Pocket PC, пользователи получают простой и надежный доступ к самой актуальной информации из Internet/intranet-сетей в удобном, компактном виде.

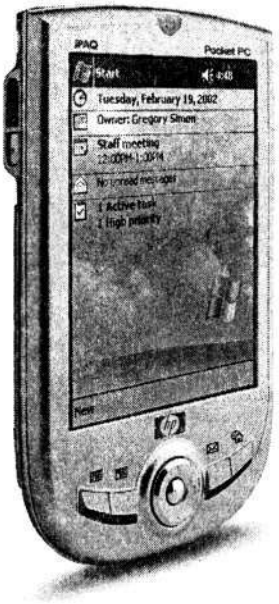


Рис. 3.17. *Pocket PC от Hewlett-Packard*

3.13. Службы доставки новостей по электронной почте

Сегодня, когда информационное пространство World Wide Web развивается феноменальными темпами, что может быть реальной альтернативой онлайновым методам получения актуальной информации из Internet? Казалось бы, ставшие традиционными методы доставки информации по электронной почте утрачивают свою актуальность. Однако это не совсем так. Крупнейшие информационные службы рассматривают электронную почту как один из самых надежных каналов распространения своего контента.

3.13.1. История сервиса

Электронная почта (e-mail), безусловно, была первым механизмом доставки информации в Internet. Это первый сервис Сети, возникшей свыше трех десятилетий тому назад. Сразу же, наряду с персональной перепиской, возникла необходимость информирования групп людей, которая нашла свое выражение в так называемых списках рассылок (Mail lists), реализующих рассылку сообщений одновременно по нескольким электронным адресам. До настоящего времени практически все ведущие информационные агентства мира используют технологию списков рассылки, поддерживаемую сотнями специализированных пакетов программ, интегрируемых с биллинговыми системами и с системами документооборота информационных служб. Пользователь, подписываясь на такую рассылку (как платную, так и бесплатную), точно знает, на какую тему он будет получать электронные письма. Другое дело, когда его "подписывают" несанкционированно, после чего следует получение сотен невостребованных писем рекламного

характера. Такие письма называют спамом, и об этом явлении речь пойдет чуть ниже. Электронной почте, как средству доставки контента, посвящены многочисленные сетевые публикации. Пожалуй, наиболее полным, энциклопедическим источником по этой теме в Рунете можно назвать "Библиотеку электронной почты" (`http://www.mailinfo.ru` — рис. 3.18).

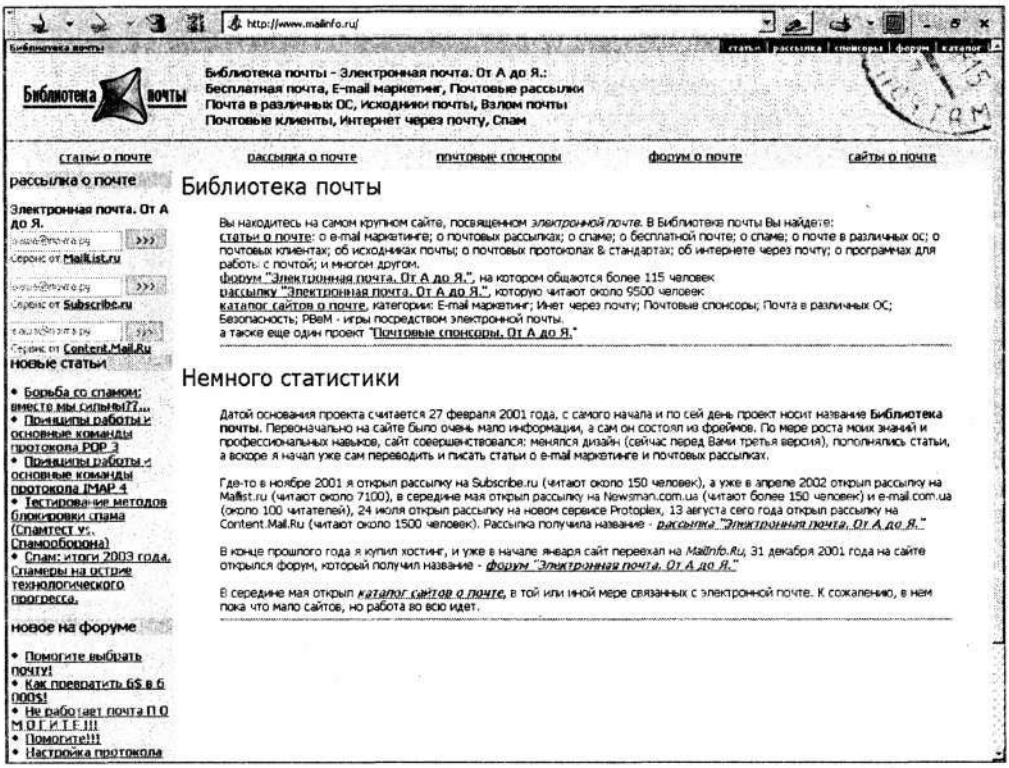


Рис. 3.18. Библиотека электронной почты

В Библиотеке электронной почты содержатся статьи о маркетинге e-mail, почтовых рассылках, спаме, бесплатной почте, почтовых клиентах, протоколах и стандартах и о многом другом. На сервере также размещен форум "Электронная почта. От А до Я", на котором постоянно общаются свыше 100 пользователей. Рассылку с этого сайта читают около 10 000 подписчиков. На сервере также приведен каталог сайтов об электронной почте, содержащий такие категории, как маркетинг e-mail, Internet через почту, почта в различных ОС, безопасность, игры посредством электронной почты и др.

3.13.2. Система телеконференций Usenet

Практически одновременно с Internet появилась технология телеконференций Usenet, которая обеспечила возможность общения и информирования групп пользователей Сети, используя в качестве транспорта преимущественно электронную почту. Абонент может подписаться на рассылку информации из телеконференций, используя как для подписки, так и для получения информации

e-mail, а также в рамках технологии Usenet размещать собственную информацию в уже существующих телеконференциях или создавать собственные. В настоящее время в Сети на Usenet-серверах (их еще называют News-серверами) существуют десятки тысяч доступных абонентам телеконференций. Телеконференции Usenet размещаются на большинстве Internet-узлов посредством реализации механизма обмена данными, с помощью которого каждый подписчик имеет возможность обратиться к интересующей его телеконференции, получить поступающие новости или послать свою информацию средствами электронной почты. При этом Usenet в целом может рассматриваться как динамическая распределенная база данных.

Протокол как набор правил и команд, в соответствии с которым в сети Internet обеспечивается доступ к телеконференциям Usenet, называется NNTP (Network News Transport Protocol). При работе с электронной почтой в режиме off-line используются стандартные шлюзы между системой новостей и электронной почтой, а также серверами новостей. Эти серверы (шлюзы "mail-news") имеют свои электронные адреса, по которым им пишутся письма-запросы. Как правило, провайдер, к которому подключаются абоненты, предоставляет им такую услугу, как доступ к своему почтовому серверу новостей и NNTP-серверу. Но как получить доступ к информации телеконференций Usenet, которые не представлены на серверах конкретного провайдера? Для этого достаточно указать в настройках агента новостей имя другого сервера новостей (NNTP-сервера), к которому обеспечивается свободный доступ. Такие серверы, "открытые для всех" (или Public News-Server), существуют в Internet в большом количестве, а их актуальный перечень можно найти, например, на таких Web-страницах: `http://www.greenline.it/news.htm` или `http://www.newsservers.net/free_news_servers` (рис. 3.19).

Для поиска открытых серверов новостей существует поисковый механизм, находящийся по адресу `http://freenews.maxbaud.net`, где в качестве запроса вводится название группы новостей. В результате обработки запроса на этом сервере выводится список открытых серверов, поддерживающих запрашиваемую телеконференцию Usenet.

Средства Usenet оказались достаточно эффективным инструментом для распространения платной информации, поэтому в свое время получили широкое распространение телеконференции коммерческой службы ClariNet, распространяющей на платной основе информацию ведущих мировых информационных агентств.

В Украине уже десять лет существует электронная газета ElVisti.Info, которая содержит около 30 новостных телеконференций, доступных на платной основе пользователям украинских Internet-узлов (рис. 3.20). Разделы ElVisti.Info поступают в виде телеконференций, доступных с помощью различных серверов новостей в режиме полной подписки (feed). Объем поступающей на компьютер пользователя текстовой информации из ElVisti.Info составляет до 5 Мбайт в сутки.

Что же касается бесплатных, слабо модерируемых телеконференций, то они в последнее время стремительно теряют свою актуальность, так как стали площадкой для спамеров, которые, с одной стороны, публикуют в них гигабайты коммерческой рекламы, а с другой стороны, пытаются отлавливать в них электронные адреса наивных пользователей, использующих телеконференции по их прямому назначению.

Несмотря на то, что технология Usenet, появившись значительно раньше Web-технологии, имеет собственные механизмы и средства доступа к информации, сегодня можно констатировать, что, наряду с электронной почтой, и Web-

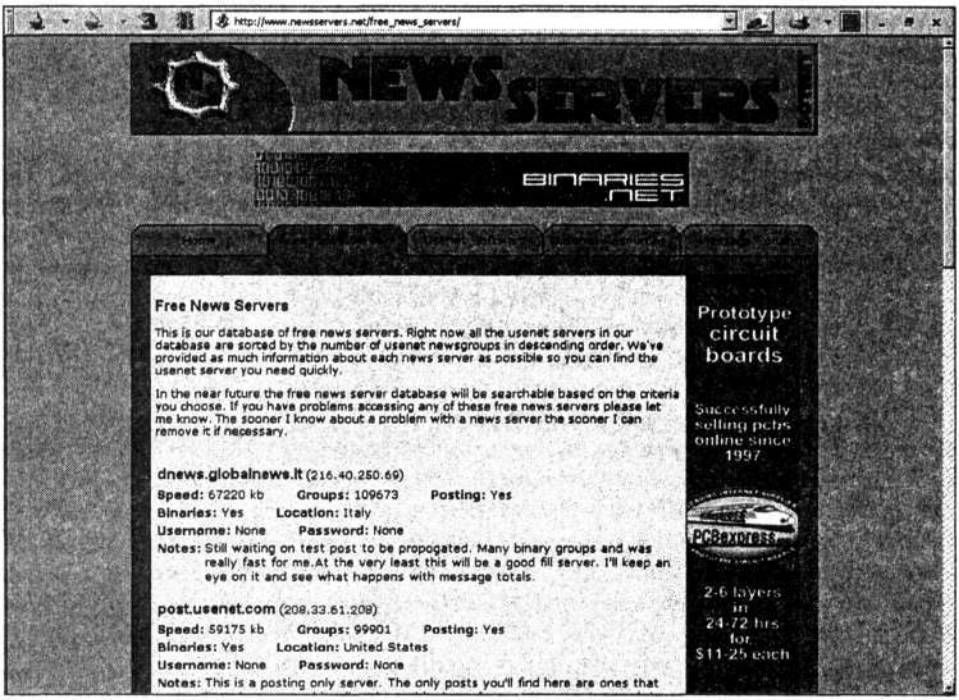
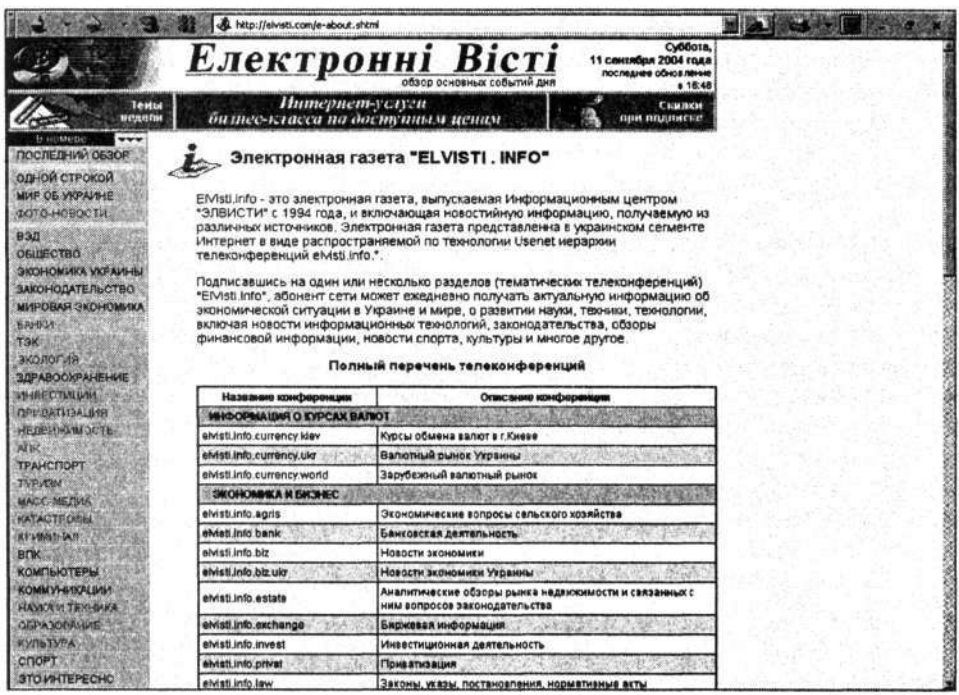


Рис. 3.19. Список доступных NNTP-серверов на сайте NewsServers



Название конференции	Описание конференции
ИНФОРМАЦИЯ О КУРСАХ ВАЛЮТ	
elvisti.info.currency.kiev	Курсы обмена валют в г.Киеве
elvisti.info.currency.ukr	Валютный рынок Украины
elvisti.info.currency.world	Зарубежный валютный рынок
ЭКОНОМИКА И БИЗНЕС	
elvisti.info.agris	Экономические вопросы сельского хозяйства
elvisti.info.bank	Банковская деятельность
elvisti.info.biz	Новости экономики
elvisti.info.biz.ukr	Новости экономики Украины
elvisti.info.estate	Аналитические обзоры рынка недвижимости и связанных с ним вопросов законодательства
elvisti.info.exchange	Биржевая информация
elvisti.info.invest	Инвестиционная деятельность
elvisti.info.privat	Приватизация
elvisti.info.law	Законы, указы, постановления, нормативные акты

Рис. 3.20. Телеконференции ElVisti.Info

механизмы (engine) представляют доступ к ресурсам Usenet. Сегодня существуют тысячи серверов, обеспечивающих интеграцию этих двух технологий. Многие из таких серверов, кроме того, обеспечивают поисковые возможности, которых так не хватало системе Usenet на этапе ее становления. В качестве примера специализированного поискового сервера по телеконференциям Usenet можно назвать NewzBot (`http://www.newzbot` рис. 3.21).

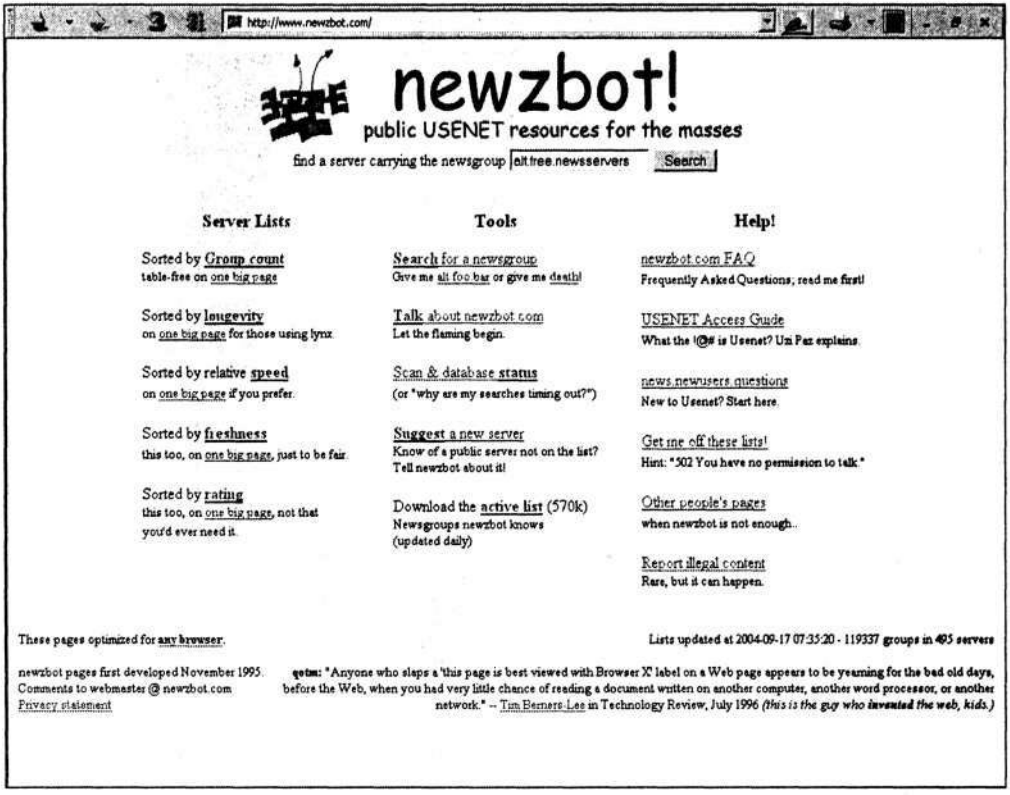


Рис. 3.21. *Специализированная поисковая система NewzBot*

Кроме того, многие универсальные сетевые поисковые службы обеспечивают полноценную работу с Usenet. В частности, крупнейшая поисковая служба Google со своего сайта "Google Groups" (`http://groups.google.com`) обеспечивает как просмотр списка и актуальности заполнения отдельных телеконференций, так и полноценный информационный поиск (рис. 3.22).

3.13.3. Доставка новостей с отдельных сайтов

Едва ли не стандартной функцией на современных Web-серверах и порталах является подписка на новости. Пользователь может зарегистрироваться на том или ином сервере и указать свой электронный адрес для получения необходимой, как ему кажется, рассылки. В распоряжении администраторов сайтов имеется широкий спектр программ рассылки; кроме того, они могут обратиться к специализированным службам рассылки, отдавая эту функцию на аутсорсинг.

Для пользователя подписка может привести к нескольким возможным "мелким неприятностям". Во-первых, электронный адрес абонента может указать не он сам, а, например, "доброжелатель" (которым может оказаться и назойливый администратор Web-сайта). Существует несколько подходов, позволяющих обойти такую принудительную подписку. Конечно же, стоит подписываться только на новости с авторитетных, надежных Web-сайтов. Администраторы корректных подписок обеспечивают возможность отказа от получения новостей как через Web-интерфейс, так и с помощью электронного письма. С другой стороны, корректная рассылка начинается всегда с обмена письмами, подтверждающими подписку.

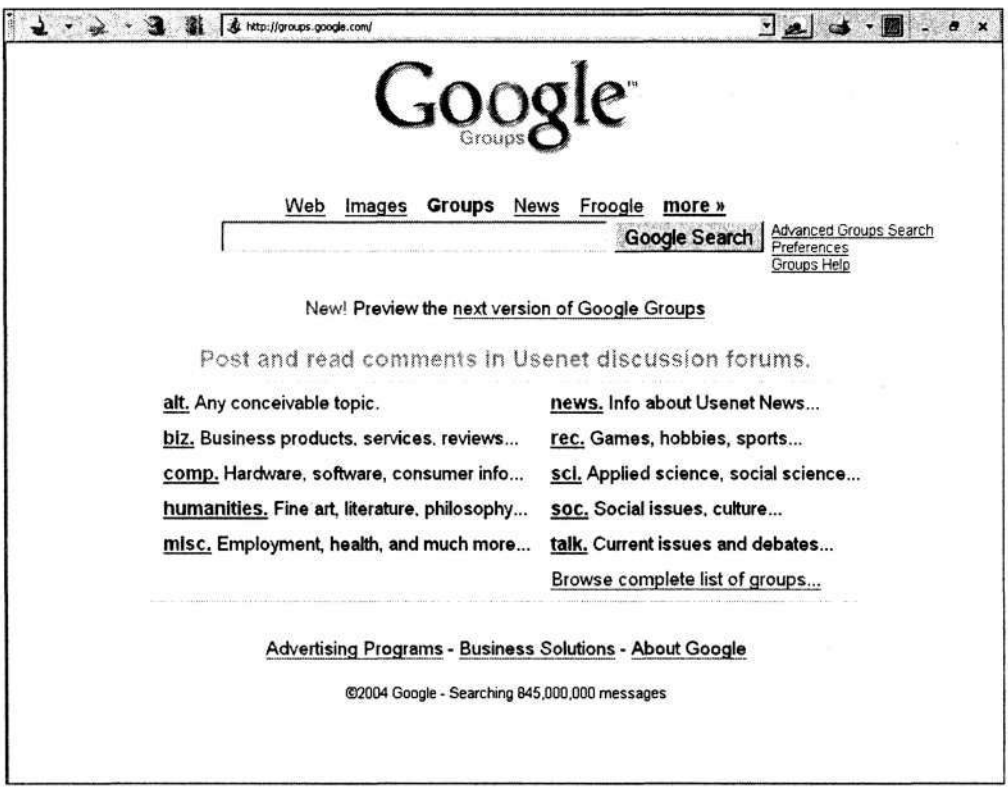


Рис. 3.22. Сервер Google Groups

Во-вторых, даже на самых надежных Web-сайтах возможна утечка информации. Электронные адреса подписчиков могут попасть в руки спамеров, и тогда на подписчика могут обрушиться потоки невостребованной рекламы.

И, наконец, в-третьих, указанная в явном виде возможность отказа от рассылки не всегда срабатывает, то ли по техническим причинам, то ли по "забывчивости" администратора рассылки. Абонент может годами получать обновления с сайта мебельного магазина, хотя необходимое ему кресло уже давно было куплено. Грубо говоря, предоставление пользователю возможности отписки от новостей не входит непосредственно в область бизнес-интересов администраторов Web-сайтов, а относится лишь к элементарной сетевой этике.

Пополнение баз данных

Механизмы доставки информации по электронной почте широко используются в технологиях поддержки баз данных, которые ведутся на локальных компьютерах пользователей и регулярно пополняются. Например, для обновления баз данных справочников и классификаторов Российского Информационного Комплекса (сервер "Ваш информационный мир офф-лайн" — http://info.rosinfocom.ru/nws_distribution.php) при установленной программной оболочке достаточно подписаться на рассылку новостей.

Эта же технология используется для обновления законодательных баз данных. Например, в Украине таким образом пополняются нормативно- правовые базы данных компаний ЛИГА (http://www.liga.kiev.ua) или НАУ (http://nau.kiev.ua).

3.13.4. Специализированные службы рассылки новостей

Несмотря на бурное развитие Web-технологий, сегодня лучший способ бесплатно и без особых усилий найти целевую аудиторию в Internet — это создание собственной почтовой рассылки. В России и Украине есть несколько серверов служб бесплатных почтовых рассылок, где можно открыть рассылку, а также организовать для нее рекламу. Предоставляемыми службами рассылок могут воспользоваться все, кому есть чем делиться с широкой аудиторией, при соблюдении некоторых условий, в том числе обеспечивая периодичность выпусков.

Службы рассылок предоставляют такие сервисные возможности:

- как правило, удобные механизмы рассылок, которые постоянно развиваются;
- организация доставки писем подписчикам в различных форматах и кодировках, формирование архивов, ведение статистики, поддержка рассылок на различных языках;
- возможность для подписчиков управления всеми аспектами своей подписки.

Количество подписчиков — самый очевидный показатель популярности рассылки. Рейтинг различных рассылок определяется по признаку, называемому "он-лайн-активностью подписчиков" или просто "активностью", которая рассчитывается как отношение количества зафиксированных чтений рассылки к тиражу рассылки.

В настоящее время в русскоязычном сегменте Сети существует три ведущие службы почтовых рассылок:

- http://subscribe.ru
- http://content.ru
- http://maillist.ru

Стоит упомянуть и об украинской службе http://newsman.com.ua. Некоторая статистика по названным службам приведена в табл. 3.1.

Без сомнения, рассылки Городского кота (http://subscribe.ru) — это самый популярный в Рунете сервис. Информационный канал Subscribe.Ru — старейший и один из крупнейших в русскоязычном сегменте Сети. Для продвижения новых рассылок, примерно раз в неделю служба рассылает новости тиражом около 500 тыс. писем с описаниями новых рассылок. Служба Subscribe.ru предоставляет форму для поиска в архивах рассылок, которые хранятся на самом сервере (рис. 3.23).

Таблица 3.1. Статистические показатели важнейших почтовых рассылок России и Украины

Служба почтовых рассылок	Subscribe.ru	Content.ru	MailList.ru	Newsman.com.ua
Дата открытия	1998	2000	1999	1999
Количество подписчиков	2 240 000	2 630 000	500 000	54 000
Количество рассылок	17 600	12 000	15 000	1700

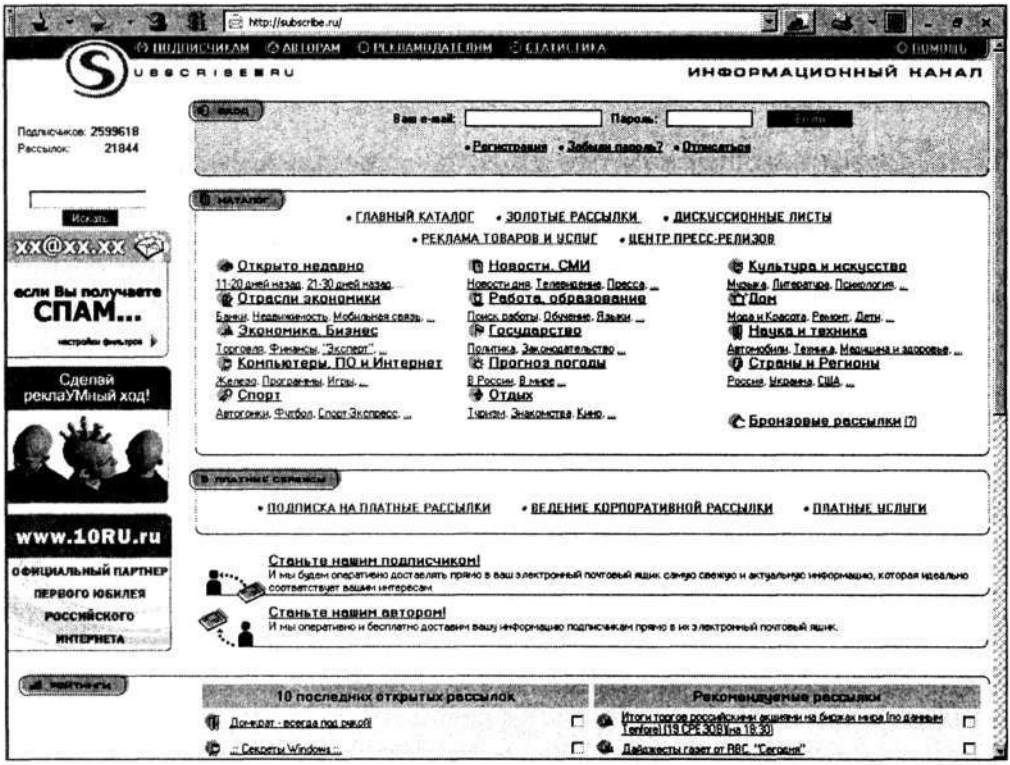


Рис. 3.23. Служба Subscribe.Ru

Помимо бесплатных услуг, служба Subscribe.ru предоставляет ряд платных. Платные услуги подразделяются на собственные платные услуги ("Рассылки без рекламы", "Персональная отправка", "POP-ящик для подписки" и другие) и платные рассылки — информация от нескольких десятков информационных агентств, партнеров службы.

Очень несложно открыть рассылку и на `http://content.ru` — рис. 3.24. Количество рассылок на этом сервере превышает 7 тыс. Здесь отсутствует классификация рассылок, поэтому их ранжирование реализовано только в зависимости от популярности, определяемой тиражом.

На сервере компании Agava SoftWare (`http://maillist.ru` — рис. 3.25) рассылку также открыть достаточно просто, однако она будет зарегистрирована в каталоге только после согласования с администраторами сервера, которые самое большое значение придают названию рассылки. Действительно, удачное

название очень важно при создании собственной рассылки. Даже плохая рассылка может иметь более высокую популярность благодаря удачному названию. На MailList.ru возможно создание коммерческих рассылок, в рамках которых заказчик получает ряд преимуществ по сравнению с владельцем обычных рассылок: в коммерческих рассылках отсутствует реклама службы рассылок, заказчик может полностью управлять содержимым своей рассылки, самостоятельно устанавливать обратный адрес и изменять другие настройки.

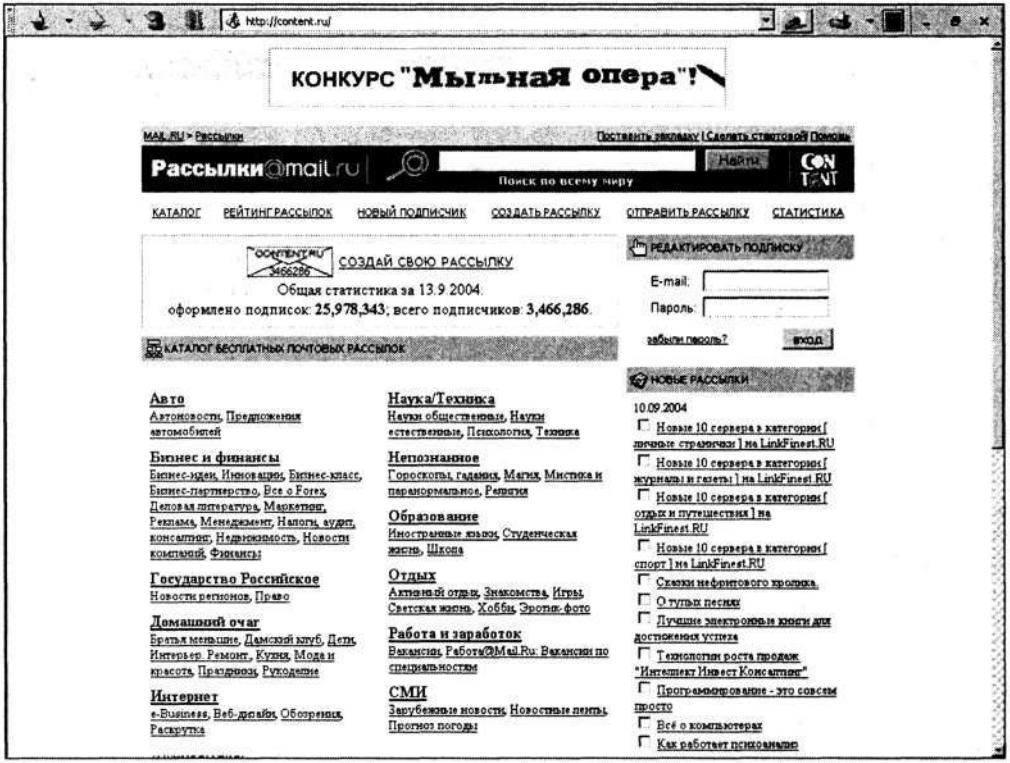


Рис. 3.24. Служба Content.ru

Для рассылки тиража очень удобно использовать content.ru и maillist.ru. На этих серверах доступна статистика по количеству. Службы subscribe.ru и content.ru предоставляют специальную кнопку определения тиража для размещения на сайте.

Естественно, большинство из служб рассылки существуют за счет рекламы, размещаемой в некоммерческих рассылках их клиентов. Вместе с тем возможны и коммерческие рассылки, за которые платит клиент-рекламодатель, который вправе использовать в рассылках исключительно собственные рекламные баннеры.

3.13.5. Интеграция новостей с целью рассылки

В настоящее время в связи с развитием наполнения Web-пространства новостной информацией возникла необходимость интеграции динамичного новостного контента с большого числа сайтов для обеспечения доступа к нему как через

Web-интерфейс, так и средствами электронной почты. Во всем мире стали появляться специализированные информационные службы, занимающиеся сбором и обработкой новостей с Internet-изданий с целью их дальнейшей рассылки.

Рис. 3.25. Служба MailList.ru

Например, известный российский поисковый портал Яndex недавно открыл проект Яndex.Новости (http://news.yandex.ru — рис. 3.26), к которому в настоящее время присоединились несколько сотен Internet-изданий. Главной особенностью Яndex.Новости как открытого публичного сервиса является наличие тем, которые объединяют содержательно близкие новости с различных сайтов.

Посетитель Яndex.Новостей может воспользоваться тематическими разделами (все полученные новости группируются по нескольким рубрикам), а также подписаться на новости определенной тематики или соответствующие конкретному поисковому запросу новости. Поиск новостей возможен как по всем источникам, так и по отдельным источникам, заданным пользователем. Имеется также возможность поиска за произвольный период времени, а также подписка на получение анонсов новостей по электронной почте.

Крупнейшее в России агентство по интеграции новостей Интегрум (http://www.integrum.ru — рис. 3.27) обеспечивает сбор в единый массив электронных версий коммерческих, статистических и новостных информационных продуктов. Контент-механизм службы является авторской разработкой агентства — это лингвистическая поисковая система Артефакт, построенная на использовании сложных морфологических алгоритмов. Сервис агентства "Персональная газета" обеспечивает создание и функционирование запросов-

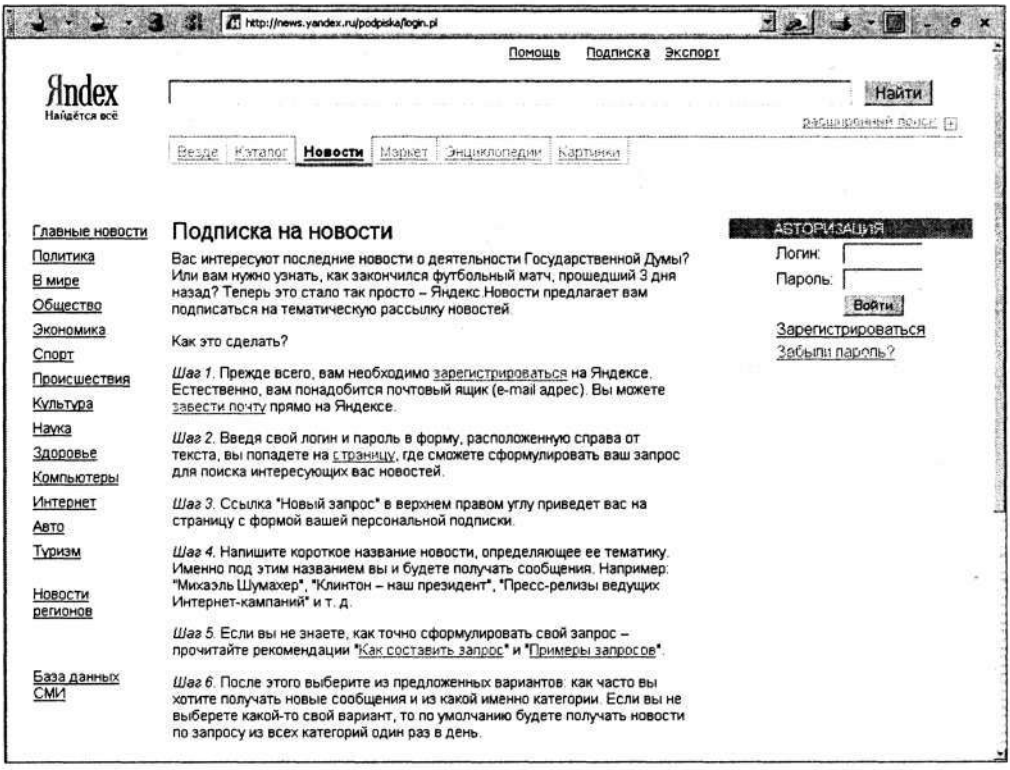


Рис. 3.26. Яndex.Новости

роботов, осуществляющих автоматический поиск и доставку по электронной почте материалов подписчикам по заданным ими запросам — ключевым словам и логическим операторам. Сервис имеет развитую систему настроек отбора по контексту и источникам информации. Каждый запрос обрабатывается системой Артефакт от одного до трех раз в сутки. В результате выбираются предварительно загруженные в базу данных документы, соответствующие запросам, которые высылаются пользователям по электронной почте.

Система InfoStream® (`http://infostream.ua` — рис. 3.28), разработанная украинской компанией ElVisti, предназначена для автоматизированного сбора новостной информации с сайтов, ее обработки, систематизации и обеспечения доступа к ней. Если пользователь хочет получать новостную информацию в режиме избирательного распространения информации по интересующей его тематике (она определяется на языке запросов с помощью ключевых слов, логических операторов, операторов контекстной близости и скобок) по e-mail, SMS или встроить постоянную подборку в свою Web-страницу, то к его услугам сервис InfoStream Client.

Сегодня системой InfoStream охватывается ежедневно свыше 20 тыс. документов из более чем 600 информационных источников, перечень которых постоянно растет. Сведения о новых информационных источниках поступают как непосредственно от разработчика, так и от пользователей сервисов InfoStream. В результате реализуется эффективный механизм обратной связи между службой сопровождения системы и пользователями.

Рис. 3.27. Персональная газета службы Интегрум

Рис. 3.28. Служба InfoStream Client

3.13.6. Спам — альтернатива востребованной рассылке

Открытость и доступность электронной почты позволила наводнить Сеть множеством ненужных или малополезных посланий (рекламные сообщения, предложения о посещении тех или иных ресурсов и др.). Очень часто почтовые ящики доверчивых пользователей превращаются в мусорную яму, оказавшись заваленными невостребованными письмами — спамом [24]. Наиболее часто спам трактуется как непрошеное рекламное сообщение или информация, рассылаемые по электронной почте в личные почтовые ящики или телеконференции. Наряду с попытками четкого юридического определения, назовем несколько действий, которые пользователи Internet воспринимают как рассылку спама.

Во-первых, это рассылка (массовая или индивидуальная) почтовых сообщений без предварительного желания адресата получать подобную корреспонденцию. Во-вторых, это подписка пользователя Internet на список рассылки без его ведома. В-третьих, — помещение в телеконференции Usenet, дискуссионные листы, гостевые книги сообщений, не имеющих отношения к их тематикам (off-topic).

Пока спам способен приносить своим создателям хоть какие-то деньги, это явление в Internet будет процветать. Массовый характер несанкционированной рассылки информации сегодня представляет собой одну из самых серьезных проблем Сети, поскольку на обработку, просмотр и удаление непрошеных писем тратится все больше времени. Так, согласно исследованиям компании MessageLabs, британские фирмы из-за спама недополучают $4,6 млрд прибыли. По их информации, каждый сотрудник компании ежедневно тратит 10 минут на просмотр и удаление спама. Это составляет 95% от общего времени просмотра почты. Борьба со спамом обходится недешево — работодатели теряют по $472 на каждого работника в год.

3.13.7. Перспективы технологий доставки новостей

Электронная почта — это всего лишь один из способов доставки пользователю персонализированного контента. С одной стороны, этот способ не самый прогрессивный, но, безусловно, самый надежный. Сегодня доставку оперативной информации на компьютеры пользователей в режиме он-лайн обеспечивает новый класс программ — агенты новостей. Настройка взаимодействия агентов новостей с Сетью, ориентированная на интересы и потребности каждого конкретного пользователя, делает их очень востребованным продуктом, особенно сегодня, когда на повестке дня стоит персонализация Internet. Клиентские агенты новостей представляют собой приложения, устанавливаемые на компьютерах пользователей и облегчающие поиск и доставку необходимой информации. Они используют, как правило, целый набор традиционных поисковых систем и каталогов.

Пожалуй, наиболее известной утилитой этого класса является Copernic 2000. Программа позволяет искать информацию, "паразитируя" на таких поисковых машинах, как AltaVista, DejaNews, Euroseek, Excite, HotBot, Infoseek, Lycos, Yahoo!, одновременно используя более 30 информационных ресурсов. Некоторые агенты новостей предполагают наличие сразу двух частей — клиентской и серверной. На серверной части выполняется сбор и обработка информации, а на клиентской — настройка и запуск агента. Среди таких систем можно выделить известную российскую систему News Alert, которая постоянно сканирует новости из российских Web-ресурсов и обеспечивает их избирательное распространение в режиме реального времени. Клиентская программа News Alert имеет размер всего 50 Кбайт, она оповещает о новостях и показывает их анонсы по желанию пользователей.

В последние несколько лет стремительную популярность набирают технологии, ориентированные на использование формата RSS (Rich Site Summary), основное назначение которого — обеспечение однотипного обмена данными в такой сложной системе, как Internet. RSS обеспечивает согласованный способ резюмировать содержимое и обновления Web-сайтов. Именно возможность однотипного экспорта новостей обусловила появление нового типа новостных агентов — RSS-агрегаторов, представляющих собой клиентские программы доступа к данным в формате RSS. С помощью этих программ, например FeedReader (`http://www.feedreader.com`) или Syndirella (`http://www.yole.ru/projects/syndirella`), пользователи могут получать доступ к RSS-данным с помощью интерфейсов, которые удивительно напоминают интерфейсы почтовых программ. При этом аналогия интерфейсов подтверждается и теми преимуществами, которые традиционно присущи электронной почте, а именно:

- информация поступает по запросу пользователя после соединения с соответствующим сервером, а на сервере накапливается по мере поступления;
- информация из почтового ящика может накапливаться в архивах пользователя;
- информация в электронном почтовом ящике максимально персонализирована.

Глава 4

XML — язык разметки и модель данных

Для эффективного представления и последующего поиска знаний в Internet в идеале следовало бы связать воедино информационные ресурсы, позволить программным приложениям установить связи между объектами и обмениваться информацией на одном языке [20].

В связи с этим в 1996 году консорциумом World Wide Web (`http://www.w3.org`) была предпринята попытка приступить к проектированию расширяемого языка разметки, который сочетал бы в себе гибкость и мощность промышленного стандарта для издательских приложений — языка SGML (Standard Generalized Markup Language) и был совместим с уже получившим повсеместное распространение языком HTML. Новый язык получил название XML (Extensible Markup Language) [33, 62], в феврале 1998 года его версия XML 1.0 была принята как рекомендация W3C (рис. 4.1).

На языке HTML составлена большая часть Web-сайтов и отдельных Web-страниц в Internet. Тем не менее этот язык не лишен многих ограничений по отображению символов, не входящих в стандартные таблицы кодировок. Например, в нем отсутствуют возможности для отображения музыкальных нот или математических формул.

Кроме того, изначально HTML не предполагал серьезного интерактивного взаимодействия. Возможности подключения CGI-форм или JavaScript-приложений лишь частично решают эту проблему. Поэтому приходится согласиться с тем, что для HTML известная парадигма "What You See Is What You Get" справедлива в интерпретации: "Вы получите не более того, что Вы видите".

Именно для решения проблем внешнего и смыслового описания документов и была в свое время начата разработка нового стандарта хранения и обеспечения доступа к информации (прежде всего, сетевой), который получил наименование XML — "расширяемый язык разметки". Благодаря своей простоте и расширяемости, XML практически сразу же получил широкое распространение и послужил основой для создания других специализированных языков разметки — CML, MathML, WML и десятков других. Даже традиционный HTML можно рассматривать как дочерний язык XML, поскольку его описание также укладывается в стандарт XML.

Необходимо отметить, что прямым прародителем языка XML был все же не HTML, а стандартный общий язык разметки SGML — "язык для описания языков", т.е. метаязык. В свое время HTML был определен на основе SGML, который достаточно сложно автоматизируется ввиду наличия широких "оформительских" возможностей. Создатели языка XML убрали эти изобрази-

тельные "излишества", в результате чего пришли к простому, эффективному и наглядному метаязыку. В отличие от большинства подобных языков, XML одинаково понятен и человеку, и программам обработки — парсерам.

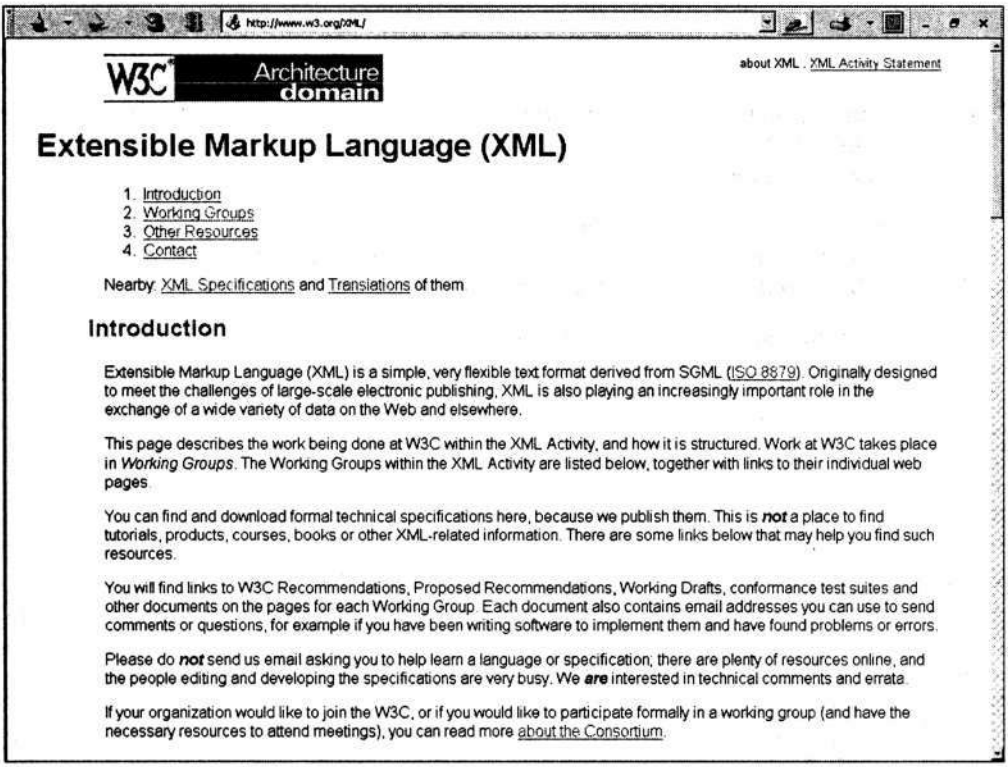


***Рис. 4.1.** Страница XML на сайте консорциума World Wide Web*

При создании XML-документа нет ограничений, связанных с фиксированным набором элементов разметки, — в этом языке не существует заранее определенного набора тегов. Все необходимые для описания документа теги разработчик может определять самостоятельно. При этом теги XML не предлагают решения вопроса, как информация должна выглядеть на экране. При публикации можно применять множества правил, собранных в листы стилей для автоматического форматирования документов. Для этого был разработан специальный расширяемый язык стилей (Extensible Stylesheet Language, или XSL).

XML-документ представляет собой дерево вложенных открывающих и закрывающих тегов, каждый из которых может включать в себя несколько пар "атрибут-значение". При этом не существует фиксированного словаря тегов и набора их допустимых комбинаций. В XML 1.0 описание правил построения документа выполняется с помощью специального языка определения типа документа DTD (document type definition), накладывающего ограничения на используемые теги и указывающего, каким образом должна быть организована их вложенность внутри документа. В DTD определяется грамматика, т.е. допустимые комбинации и вложенность имен тегов и атрибутов.

В XML применяется два вида указания на определение DTD, использованное для конкретного документа.

- Декларации внутренних подмножеств DTD-определений, помещаемых посредственно в XML-документ. При этом команда-определение DTD заключается в квадратные скобки, например `<!DOCTYPE rootElement [declarations]>`.
- Ссылки на внешние DTD-определения, например `<!DOCTYPE wml PUBLIC "-//WAPFORUM//DTD WML 1.1//EN" "http://www.wapforum.org/DTD/wml_1.1.xml">`.

Указание `PUBLIC` во втором случае свидетельствует о том, что DTD является общедоступным стандартом, в частности в данном примере это стандарт языка WML.

Язык DTD позволяет определить логическую структуру документа, т.е.

- задать порядок следования элементов;
- определить вложенность элементов;
- установить количество возможных элементов;
- установить типы атрибутов;
- определить сущности и нотации.

Однако языку DTD присущи серьезные недостатки, а именно ограниченность описания типов данных и синтаксис, отличный от XML.

Поэтому в настоящее время W3C активно работает над тем, чтобы заменить язык DTD на новый стандарт — использование XML-схем (XML Schema). Техническая рекомендация "XML-схема" была принята в 2001 году (`http://www.w3.org/TR/xmlschema-formal`).

Ведущие производители программного обеспечения во всем мире приняли концепцию XML-схем и внедрили ее в своих продуктах. Так, корпорация Microsoft на условиях бесплатного лицензирования предоставила разработчикам доступ к XML-схемам, используемым в Microsoft Office 2003. В настоящее время по адресу `www.microsoft.com/office/xml/default.mspx` доступны следующие XML-схемы:

- SpreadsheetML для Microsoft Office Excel 2003;
- FormTemplate Schemas для Microsoft Office InfoPath 2003;
- WordprocessingML для Microsoft Office Word 2003;
- DataDiagramingML для Microsoft Office Visio 2003.

По мнению представителей корпорации, доступность схем значительно облегчит реализацию поддержки возможности обмена данными между разрабатываемыми программами и приложениями офисного пакета.

Безусловное преимущество XML — это использование им современного стандарта кодировки символов Unicode, который позволяет в одном документе комбинировать тексты, написанные на основных языках мира (в том числе имеется поддержка кириллицы). Тем самым XML дает возможность с легкостью обмениваться информацией вне национальных границ.

4.1. XML как модель данных

Итак, XML — это удобный формат обмена данными, язык описания документов. При этом, как уже было замечено, оказалось, что спецификацию XML можно рассматривать еще и как реализацию иерархической модели данных. На основе этой модели можно строить системы управления XML-базами данных, причем первые же попытки показали высокую эффективность таких систем. В свое время аналитик Yankee Group Роб Перри отметил: "XML-документы могут содержать мегабайты текстовых строк и тегов XML. Такие документы имеют иерархическую структуру. И лучше оставить эту иерархию в неизменном виде, ведь поиск в этом случае окажется гораздо более эффективным". Даже отдельный XML-файл сам по себе — уже готовая иерархическая база данных. Дело остается лишь за языком запросов, основой для которого тоже может служить сам XML.

Представление данных в виде XML-документов является естественным отражением структуры реальных документов. Представлять данные как XML-документы значительно легче, чем помещать их в реляционные таблицы, которые проще трактовать лишь как фрагменты документов. При этом манипулировать данными с помощью присущей XML-технологии гиперсвязности пользователю удобнее, чем ключами, используемыми в реляционной модели. Поэтому сегодня стоит вопрос не о том, возможно ли создание XML-СУБД, а о том, можно ли заменить такими системами традиционные реляционные СУБД.

При создании СУБД нового поколения для XML-модели данных (их часто называют native — естественными) разработчики исходят из необходимости отражения операций, которые совершаются с документами в реальной жизни. Такие базы данных в качестве модели используют XML-модель данных. С одной стороны, именно такие базы данных намного лучше подходят для хранения XML-контента, чем широко распространенные сейчас реляционные СУБД. В то же время такие компании, как Oracle, IBM и Microsoft, разрабатывают технологии, улучшающие возможности работы их реляционных СУБД с XML-документами. При этом, конечно, следует отличать XML-ориентированные базы данных от реляционных и смешанных, которые также основаны на реляционной модели, но поддерживают обмен данными на языке XML (такие БД также называют постреляционными).

Под языком XML подразумевается совокупность трех тесно связанных стандартов. Сюда входит сам XML как средство описания структуры документов; XSL — как средство преобразования XML-документов для отображения; а также XLL-расширяемый язык связывания документов. Последний позволяет устанавливать многонаправленные ссылки и ссылаться не на весь документ, а на его конкретные элементы. Кроме того, в XML структурированные данные документа отделены от описания способа их логико-графического представления в виде списков, параграфов, таблиц, диаграмм и так далее, а определение логического представления отделено от задания конкретного внешнего вида (стиля).

Для разработчиков приложений существует спецификация программного интерфейса XML OM (например, компания Microsoft использовала его в виде Document Object Model — DOM).

Работа с XML-данными заранее неизвестной структуры — это принципиальная особенность XML-ориентированных СУБД, выгодно отличающая их от реляционных СУБД.

XML-ориентированные СУБД обеспечивают значительно большую скорость выполнения транзакций, что объясняется, с одной стороны, меньшими затратами

на выполнение преобразований данных, а с другой стороны, способом управления памятью — по алгоритму двоичного дерева. В XML-СУБД данные могут быть записаны без предварительного ввода описаний, при этом они сразу же становятся доступными пользователям за счет унификации обработки XML-тегов. XML-СУБД также характеризуются простотой разработки приложений, что обусловлено естественностью и простотой их архитектуры.

Кроме того, технология XML-СУБД полностью соответствует концепции корпоративных хранилищ данных. Обычно такие комплексы физически представляют собой совокупность нескольких, связанных между собой баз данных, содержащих, помимо прочего, аналитическую информацию, необходимую для принятия решений. В XML-базах данных каждый документ может быть интерпретирован как аналитический, т.е. представляющий собой ответ на запрос, непосредственно аналитический расчет, входную или выходную форму. Возможность преобразования XML-документов в реляционное представление и обратно позволяет подсоединить к XML-СУБД реляционные базы данных из уже имеющихся корпоративных хранилищ. Кроме того, технология XML предоставляет возможность создания децентрализованных баз данных, логически объединенных через Internet/intranet.

Следует отметить, что отсутствие до последнего времени унифицированных стандартов для XML-запросов и достаточного числа примеров масштабного развертывания XML-баз данных пока еще сдерживает массовый переход к технологии XML — большинство корпоративных пользователей по-прежнему предпочитают хранить данные в реляционных СУБД. Однако внедрение этих технологий дает немалые шансы на существенный выигрыш в оперативности обработки информации и естественности ее представления. Для ускорения подобного перехода многие производители реляционных СУБД добавляют поддержку XML к своим продуктам.

Так, в СУБД Oracle версии 9i (`www.oracle.com`) появилось множество новых возможностей, включая поддержку методов Data Mining (глубинный анализ данных), персонализацию и работу с XML. Теперь сервер Oracle является не только объектно-реляционным, но и позволяет хранить и обрабатывать XML-данные. В свою очередь, компания IBM (`www.ibm.com`) создала расширение своей СУБД DB2 Universal Database для представления и обработки слабоструктурированных данных DB2 XML Extender, предназначенное, прежде всего, для работы с XML-данными. Кроме того, IBM объявила о создании новой технологии Xperanto, призванной облегчить поиск информации в реляционных базах данных, в документах формата XML, в плоских (текстовых) файлах, электронных таблицах и прочих источниках данных, как если бы они были организованы в единую базу данных (рис. 4.2). В своей работе Xperanto использует язык XQuery для поиска информации в XML-документах. За исключением языка XQuery, все технологии, применяемые в Xperanto, уже применялись в последней версии СУБД IBM DB2.

4.2. XML-поиск и языки запросов

Параллельно с XML была начата разработка стандарта для метаданных — схемы описания источников (Resource Description Framework, или RDF). Предполагается, что в будущем узлы метаданных RDF, распространенные по всей Сети, обеспечат значительно более высокое качество и скорость поиска данных в Internet, т.е. спецификация RDF будет играть роль концептуальной схемы базы данных Глобальной сети. Согласно архитектуре Семантического Web, которую разрабатывает и продвигает W3C, RDF представляет собой "связующее звено"

между XML-документами и высокоуровневыми средствами, обеспечивающими поиск и навигацию на основе логических утверждений. Основной задачей при разработке RDF была необходимость определения механизма описания ресурсов, который не делал бы никаких предположений относительно специфики предметной области, но был бы удобным для описания и обработки сведений о любой области.

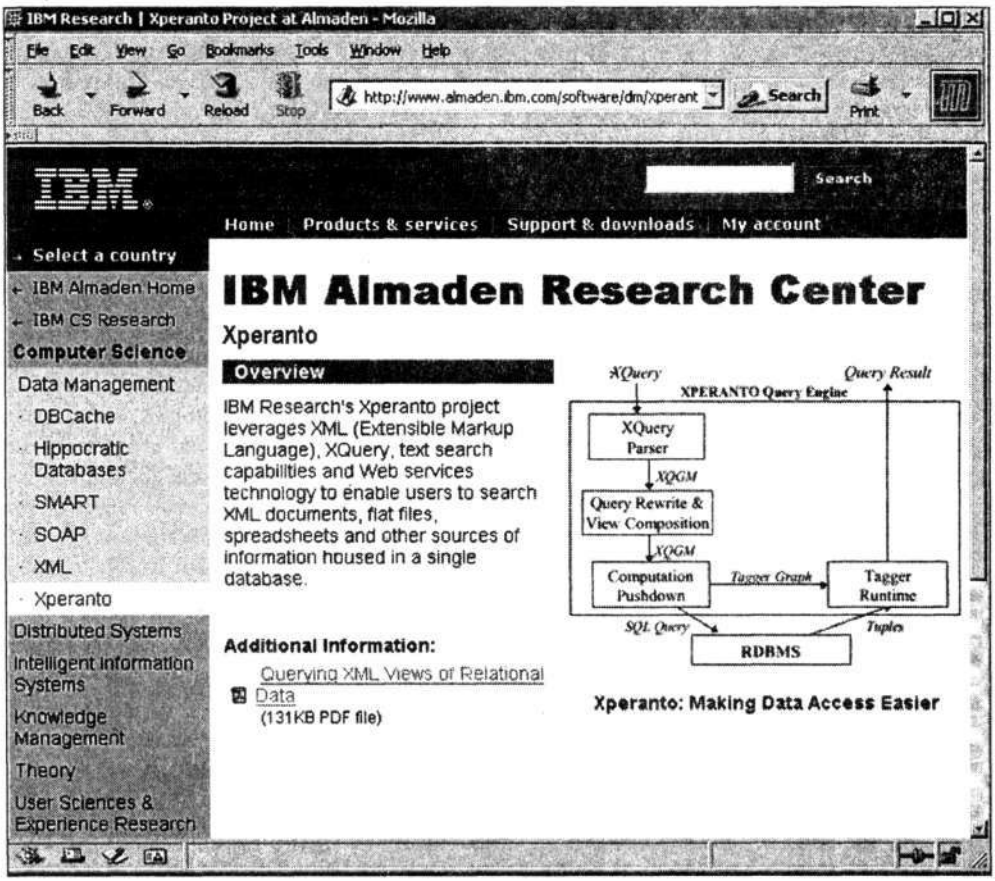


Рис. 4.2. Проект Xperanto корпорации IBM

Модель Resource Description Framework, имеющая статус рекомендации W3C, имеет своей целью стандартизовать определение и использование метаданных, описывающих ресурсы Web. Кроме того, RDF используется и для представления данных. Базовый строительный блок в RDF — тройка "объект-атрибут-значение", которую часто записывают в виде A(O,V), где O — объект, A — атрибут со значением V. RDF позволяет менять местами объекты и значения. Изначально в RDF для записи метаданных применялся синтаксис XML. Однако существуют и другие формы RDF-описаний, например в виде наборов троек. Приведем пример записи в формате RDF.

```
HasName
('http://dwl.visti.net',
"Dmitry Lande")
```

```
authorOf
('http://dwl.visti.net/',
 'http://dwl.visti.net/art/abstr-st/index1.html')
hasPrice
('http://dwl.visti.net/art/abstr-st/index1.html', "$12")
```

Кроме того, RDF допускает форму представления, в которой любое выражение RDF в тройке может быть объектом или значением.

```
hasName
('http://dwl.visti.net',
"Dmitry Lande")
authorOf
('http://dwl.visti.net',
'http://dwl.visti.net/art/abstr-st/index1.html')
hasPrice
('http://dwl.visti.net/art/abstr-st/index1.html',
"$12")
```

Гипертекстовая среда обеспечивает и другие способы нахождения нужной информации. Так, сетевые гиперссылки также предполагается усилить в рамках XML-технологии путем ввода специальных ссылок (Xlink), которые должны обеспечить переход по списку возможных назначений. Xlink позволит использовать непрямые связи, которые указывают на данные в центральной базе данных, а не связывают сами Web-страницы. В случае изменения адреса страницы ее автор сможет актуализировать все связи, которые указывают на нее, изменяя всего одну запись.

Естественным методом нахождения информации в Internet является отработка поисковых предписаний на специальном языке запросов. Вопрос стандартизации языка запросов все еще остается открытым. При этом вполне естественным кажется использование наработок, полученных при стандартизации реляционной модели. Язык SQL уже на протяжении многих лет является промышленным стандартом управления данными. Однако его прямое использование недопустимо, прежде всего, из-за отличий иерархической модели XML от реляционной. Как уже говорилось, XML является более естественным форматом, имеющим несколько дополнительных степеней свободы. Метаданные XML содержатся в самом документе, его элементы могут быть вложенными, причем в документе допускается несколько элементов с идентичными тегами — в XML возможно позиционное обращение, благодаря поддержке иерархической структуры данных, тогда как в реляционной модели не допускается несколько столбцов с одинаковыми названиями.

Вполне естественно, что появилось много работ по созданию языков запросов и моделей данных для XML-документов, которые реализуются в различных XML-СУБД. Консорциум World Wide Web, к примеру, предложил язык запросов XQuery, который пока еще не находит повсеместного применения, так как до последнего времени он не был признан полноценным стандартом.

Вместе с тем, основа стандарта, описывающего извлечение данных из XML-документов, уже существует. Рекомендация Xpath предполагает иерархическое представление структуры документа, а в роли XML-документа может выступать все дерево документов Web-сайта или даже всей Сети. Но, несмотря на простоту и удобство доступа к XML-элементам, документам или сайтам, эта рекомендация описывает только логическую, а не физическую структуру хранения документов. В Xpath для определения критериев выборки данных используются фильтры, семантически подобные выражению WHERE в языке SQL. Кроме того, XPath не

способен в полном объеме решить вопросы связи между документами, а также организации структуры вывода данных.

Почти все языки запросов, претендующие на промышленное использование, обеспечивают формирование запросов, состоящих из трех частей: оператор шаблона, оператор фильтра и оператор конструктора. Кроме того, все они содержат конструкции для выражения вложенности и упорядоченности. При этом представление запросов очень часто синтаксически близко к языку XML. Рассмотрим реализацию этих свойств на примере популярных языков XML-запросов: QL, YATL, Lorel и XQL. Из всех этих языков только XQL не поддерживает операторов конструктора. Вместе с тем, XQL может применять фильтры к элементам и атрибутам, а также к инструкциям обработки, комментариям и к внешним ссылкам.

Язык XML-QL был спроектирован в AT&T Labs как язык запросов непосредственно к XML-данным для решения задач их выборки, интеграции и преобразования. В качестве основной прикладной задачи рассматривается электронный обмен данными. В запросе XML-QL шаблоны и фильтры появляются в операторе `WHERE`, а конструктор — в операторе `CONSTRUCT`. XML-QL поддерживает упорядоченность элементов на уровне модели данных, группировку результатов запроса по значению элементов, внешние и внутренние соединения.

Язык YATL основан на механизме правил, обладает мощными средствами реструктуризации и шаблонами для создания и использования новых элементов. YATL поддерживает графический интерфейс — программисты не пишут непосредственно YATL-программы, они генерируются автоматически по графической спецификации. В запросе на языке YATL конструктор появляется в операторе `make`, шаблон — в операторе `match`, а фильтры — в операторе `where`. YATL обладает такими основными свойствами: представление модели данных на различных уровнях абстракции, обработка данных из различных источников, обработка мультимножеств и списков на основе группировки и упорядочения, возможность настройки программ под входные данные.

Язык Lorel изначально был спроектирован в Стэндфордском университете в рамках проекта Lorel, а затем был модифицирован для работы с XML. Это дружественный язык с простым синтаксисом в стиле SQL. Основным преимуществом Lorel является совместное использование средств преобразования типов и механизма построения пути запроса над нестрогой структурой данных (general path expression). Lorel реализует собственную (не XML) модель данных, являющуюся расширением объектной модели, включающей в себя в виде узлов элементы данных XML. В запросе Lorel конструктор задается оператором `select`, а шаблоны — оператором `from`. В операторе `where` одновременно используются шаблоны и фильтры. Значение операторов `from` и `where` представляет собой отношение, преобразующее переменные в кортежи значений, удовлетворяющие условиям. Одной из наиболее сильных сторон языка Lorel являются богатые возможности навигации по "неопределенным" схемам с использованием шаблонов.

Концепция языка XQL основана на преобразовании документов. Сам язык является естественным расширением синтаксиса XSL и представляет собой спецификацию для поиска в документе отдельных элементов или узлов с соответствующими характеристиками. XQL не имеет операторов конструирования. Вместо них результат запроса определяется выражениями шаблона. Язык XQL обеспечивает обработку операторов сравнения, логических операторов, наборов ключевых слов, расширенных кванторами общности и существования, множественных операторов объединения и пересечения.

Все названные языки обеспечивают неплохой компромисс между выразительной силой, простотой и функциональностью и служат основой для построения других, более приближенных к предметным областям языков запросов.

4.3. XML-решения для хранения данных

Наряду с внедрением поддержки XML в уже существующие решения, производители СУБД проявляют все больший интерес к созданию "чистых" XML-решений. Благодаря этому появилась возможность в полной мере использовать одно из главных преимуществ этого языка — возможность оперировать иерархическими массивами данных и метаданных напрямую, без преобразования в реляционный формат. Поэтому ниже будут рассмотрены продукты, которые целиком следуют оригинальной философии XML.

XML-сервер Tamino

Самый известный — XML-сервер Tamino компании Software AG (`www.softwareag.com` — рис. 4.3). В свое время компания получила широкую известность благодаря своей СУБД Adabas. Сегодня основной вектор деятельности компании направлен на развитие XML-технологий, и, прежде всего, на ее новый брэнд — XML-сервер Tamino, который был разработан с учетом существующих линий продуктов Software AG. Он интегрируется с такими системами, как Bolero, EntireX, Adabas и Natural.

Рис. 4.3. XML-сервер Tamino

Несмотря на прежние наработки, имеющиеся у компании-производителя в области СУБД, в основе Tamino лежит XML-архитектура. Базовой технологией этого сервера является X-машина — технология хранения информации, служащая для хранения XML-документов и обслуживания запросов к ним. При обработке XQL-запроса X-машина может обращаться как к внутреннему репозиторию XML-данных (XML Store), так и к внешним источникам через X-Node и серверные расширения.

Концепция X-узла (X-Node), реализованная в Tamino, является интерфейсом для доступа к существующим традиционным базам данных. Кроме того, в ядре Tamino также представлены средства полнотекстового поиска, которые позволяют создавать интеллектуальные поисковые машины, обеспечивающие поиск с учетом структуры документа, что делает Tamino особенно привлекательным для Internet.

Tamino имеет собственный репозиторий XML-данных и позволяет строить виртуальные хранилища данных над разнородными источниками. В качестве декларативного языка запросов Tamino поддерживает подмножество языка XQL. Для поддержки работоспособности SQL-приложений в состав Tamino входит SQL-процессор (для языка SQL версии 2), обеспечивающий также и среду хранения реляционных данных. Еще один компонент — Tamino Manager — представляет собой инструментарий для администрирования объектов Tamino в Internet.

Кроме прочего, сервер Tamino включает XML-СУБД, позволяющую оперировать XML-объектами. Эта СУБД обеспечивает доступ ко всем популярным типам данных, в том числе и из традиционных СУБД, а также их конвертацию в XML-объекты. В Tamino использована философия открытых СУБД, сервер содержит стандартные интерфейсы: OLE DB, DCOM, ODBC и JDBC. Tamino поддерживает XML V.1.0, язык ссылок XLL, таблицы стилей XSL и подмножество языка XQL, а также концепцию пространства имен XML. Для описания схем XML-данных используется язык описания схем Tamino (Tamino Schema Language). Tamino предоставляет возможность генерации схем по DTD-описанию.

Ipedo XML Information Hub

В основу иерархической СУБД Ipedo XML Database 2.0 компании Ipedo положен язык XML, что позволило ускорить доступ к информации в формате XML и устранить необходимость ее преобразования в формат реляционной базы данных.

Основным компонентом информационного узла от компании Ipedo является СУБД Ipedo XML Database (`www.ipedo.com` — рис. 4.4). Эта система работает внутри основной памяти сервера баз данных, при этом она способна на XML-основе осуществлять поиск и генерацию Web-страниц. В Ipedo XML Database используется стандарт запросов W3C Xpath, который позволяет производить запросы к базе XML-документов непосредственно в синтаксисе XML. Кроме того, эта СУБД включает в себя механизм XSLT, объединяющий доступ к данным и их трансформацию в единый процесс. К числу плюсов этого продукта стоит отнести его работу на J2EE-совместимых Web-серверах приложений под управлением Windows 2000, Windows NT, Sun Solaris и Red Hat Linux.

СУБД Ipedo XML Database включает функцию поиска в свободной форме, которая позволяет находить текст, игнорируя включенные в символьные последовательности теги XML. Кроме того, система обеспечивает поиск информации, представленной в графическом формате Scalable Vector Graphics, базирующемся на спецификациях XML. На СУБД возложена также задача преобразования

XML-документов в форматы других приложений. К примеру, один и тот же каталог динамически может быть преобразован в формат HTML или WML, поддерживаемый беспроводными устройствами.

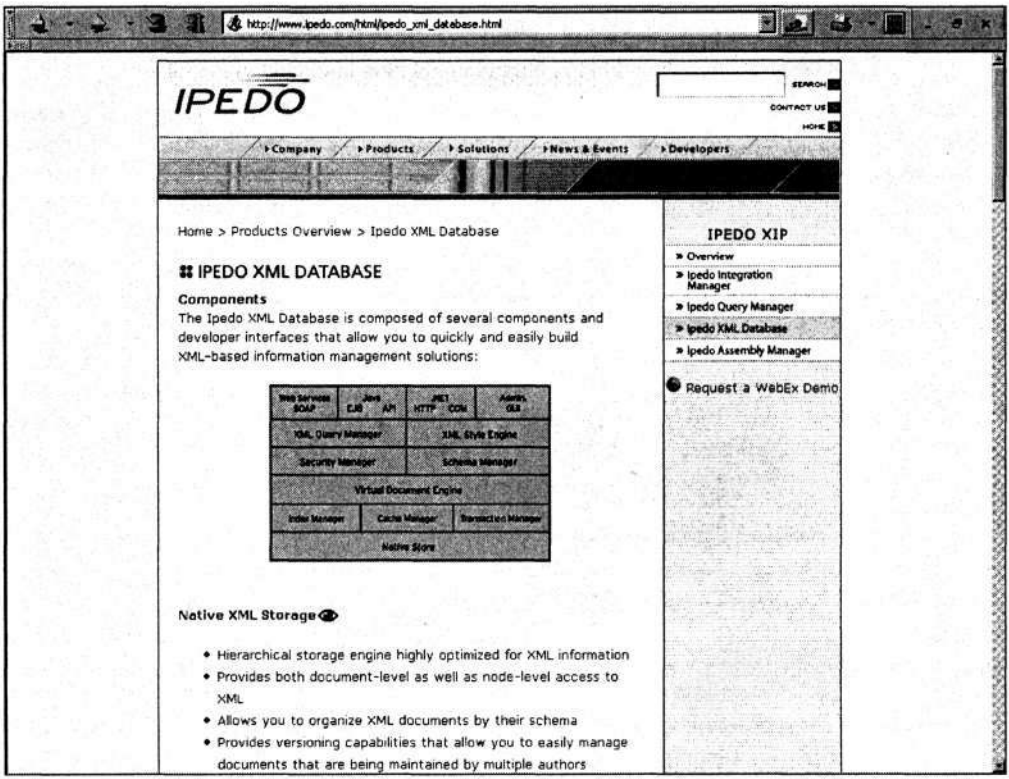


Рис. 4.4. *СУБД Ipedo XML Database*

"Преобразование данных может оказаться весьма сложной задачей для приложений, поскольку требует интенсивных вычислений и больших затрат процессорных ресурсов, — пояснил президент компании Ipedo Тим Мэттьюз. — Мы интегрировали в свою СУБД и механизм Xquery, стандартный язык запросов на базе XML, разработку спецификаций которого консорциум W3C планирует завершить в ближайшее время."

Sonic XML Server

СУБД Sonic XML Server (ранее известная как eXcelon), в отличие от большинства других XML-СУБД, способна работать с отдельными сегментами XML-документов, а не только с документами целиком (рис. 4.5). Сервер eXcelon, в первую очередь, предназначен для создания Internet-приложений. Этот сервер был разработан на базе объектной СУБД ObjectStore компании Object Design и обеспечивает унифицированное ведение данных в XML-формате, извлекая их из различных источников, конвертируя в XML и сохраняя в собственном репозитории (XML Store). XML-данные, сохраненные в репозитории, индексируются для обеспечения эффективного доступа к ним.

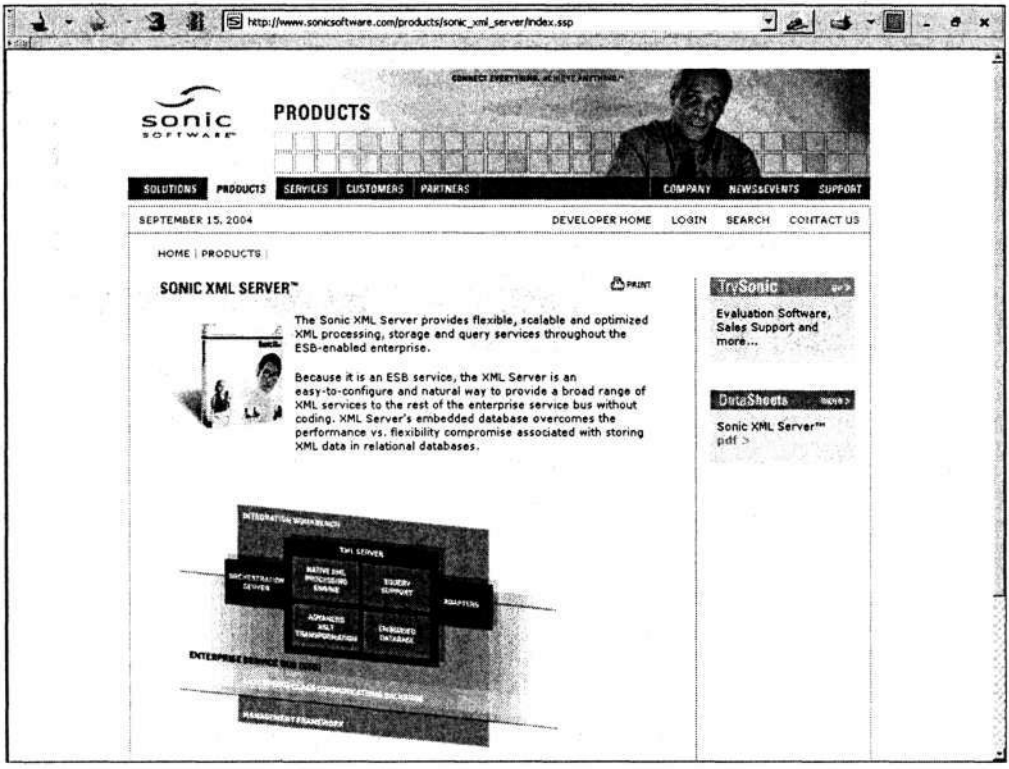


Рис. 4.5. СУБД Sonic XML Server

Следует отметить, что СУБД Sonic XML Server базируется на оригинальной модели Sonic ESB (enterprise service bus), которая обеспечивает обработку и хранение информации, а также выполнение запросов к базе данных.

XQEngine

XQEngine (`www.fatdog.com`) представляет собой полнотекстовую поисковую систему для XML-документов. Она позволяет осуществлять поиск по коллекции XML-документов, используя в качестве запроса логические выражения, составленные из ключевых слов, которыми определяется искомая информация, в комбинации с XQL-запросом. В интерфейсе реализованы методы, которые имеют непосредственное отношение к поисковым системам. Так, например, здесь применяются функции для задания так называемых стоп-слов, указания размера, местоположения и других параметров индексного файла, а также имеется возможность определять, что из документа войдет в индексный файл. XQEngine реализует задание запроса с помощью языка XQL, что при реальной работе приводит к необходимости знать структуру XML-документа.

XMS

Компания NeoCore (`www.neocore.com` — рис. 4.6) позиционирует СУБД XMS (XML Management System) как самоконструирующуюся и естественную (native). По заявлению разработчиков, продукт нацелен больше на поиск информации,

чем на поиск отдельных документов. При этом достигается более высокая скорость обработки каждого элемента XML, хранящегося в естественном виде с сохранением иерархии отношений. Система XMS обеспечивает самоописание структуры базы данных, не содержащей строк и колонок. Эта база данных автоматически индексируется, а также предоставляет широкие возможности расширения и модификации. Язык запросов здесь также базируется на XQuery и Xpath, обеспечивая поиск документов, их фрагментов и отдельных элементов.

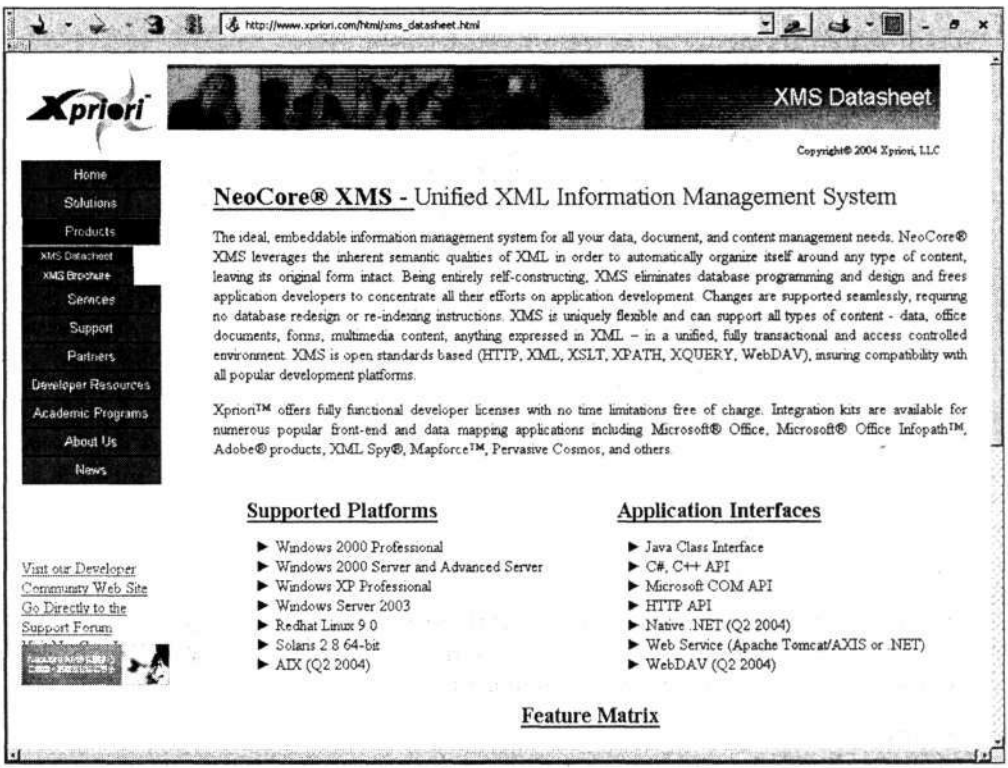


***Рис. 4.6.** СУБД XML Management System (XMS)*

Cache

Постреляционная СУБД компании InterSystems (`www.intersystems.com`) включает в себя многомерный сервер данных Cache для доступа к реляционным данным через SQL и сервер приложений Cache.

Помимо совместимости с предыдущими продуктами InterSystems и с традиционными реляционными СУБД, комплекс Cache позволяет автоматически проецировать данные в XML-документы. Такой "проецируемый" XML может как использоваться в виде файлов, так и служить для формирования контента в режиме on-line. Проекция объектов может быть также использована для формирования DTD. XML-документы могут быть автоматически трансформированы в объекты, для реализации бизнес-логики приложений. Система также позволяет пользователям разрабатывать собственные серверы или создавать собственные механизмы XML-импорта.

4.4. Корпоративные и офисные приложения для XML

Office 2003 и InfoPath от Microsoft

В октябре 2002 года корпорация Microsoft сообщила о начале бета-тестирования очередной версии Office 11 (сегодня это Office System 2003). XML стал вторым форматом для файлов этого программного пакета. CEO Microsoft Стив Баллмер рассказал о продукте с кодовым названием XDocs, который призван облегчить ввод и систематизацию данных в электронных формах, связав офисные документы с источниками данных на серверах. В настоящее время этот продукт называется Microsoft InfoPath (рис. 4.7).

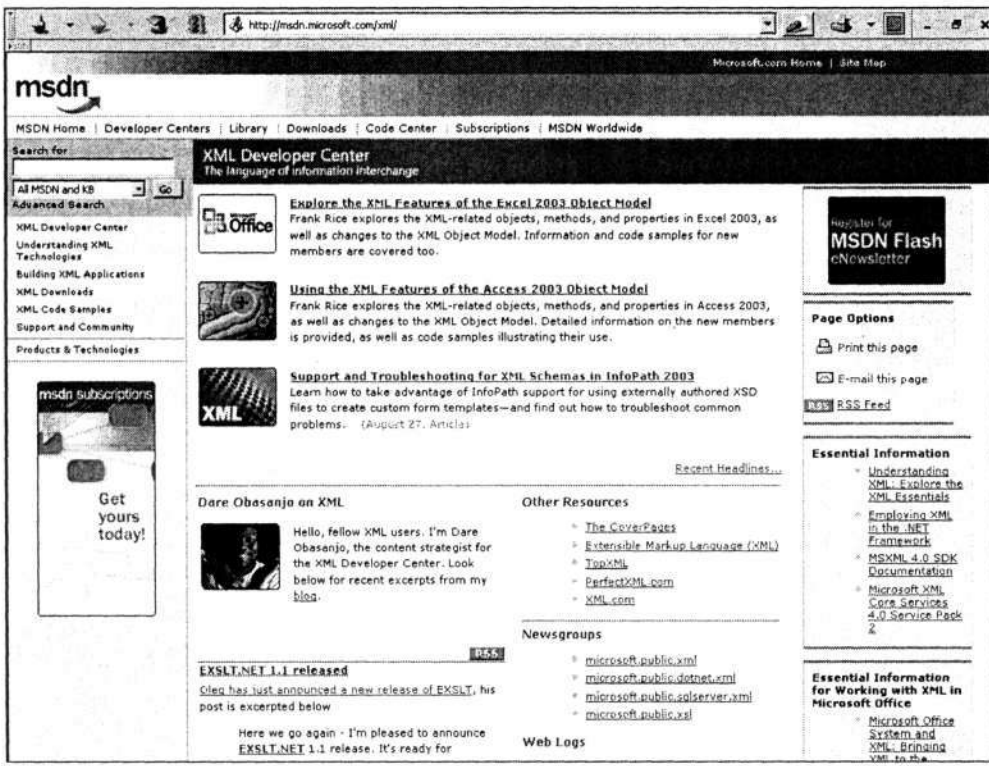


Рис. 4.7. XML-решения от корпорации Microsoft

Встроенная поддержка XML в Office 2003 позволяет пользователям работать в знакомой офисной среде, но при этом создавать и сохранять документы XML, даже не зная о том, что они работают с XML. Это значит, что пользователи без дополнительных курсов обучения могут работать с XML, используя привычные инструменты. Для работы с XML в Microsoft Office вовсе необязательно владеть основами программирования, тем не менее одно из новых дополнений в Microsoft Word 2003 — это возможность просматривать теги XML. Когда пользователь открывает XML-документ и начинает работу, любые XML-теги отображаются в виде скобок вокруг слова, фразы или абзаца.

Приложение Microsoft InfoPath позиционируется в Office в качестве инструмента для конечных пользователей приложений, предназначенного для управления взаимодействием с заказчиками (CRM) и планирования корпоративных ресурсов (ERP). Microsoft InfoPath рассматривается как механизм извлечения данных, предпочтительный для приложений Office. InfoPath — это, прежде всего, приложение для представления разнородной информации в стандартном XML-формате. Приложение можно применять как для ввода XML-данных, так и для извлечения их из базы данных. По словам аналитика Meta Group Дэвида Йокельсона (David Yockelson), преимущество XDocs (сегодня — это Microsoft InfoPath) в том, что он поможет собрать большое количество взаимосвязанных данных в табличные формы, устраняя разграничение между электронными таблицами, документами Word, формами Access, а также между способами ввода информации во все эти приложения. InfoPath поддерживает любые СУБД, совместимые с XML, — например, СУБД производства Microsoft и Oracle.

В то же время в Microsoft ведется разработка нового, ориентированного на XML языка программирования. Один из главных разработчиков платформы .Net Дон Бокс недавно высказал мнение, что "языки программирования должны либо эволюционировать, либо прекратить свое существование". По словам Бокса, XML и Web-сервисы заставляют программистов манипулировать данными, в то время как современные языки оптимизированы для работы с объектами. Бокс заявил, что "если мы сделаем Web-сервисы похожими на CORBA, мы упустим многие возможности". (Для справки: CORBA представляет собой объектно-ориентированную технологию взаимодействия приложений друг с другом.) В XML-сервисах понятиям объектов, функций, компонентов не отводится столь важной роли, поэтому новое поколение .Net-языков должно претерпеть ряд эволюционных изменений. Разработку "ориентированного на данные" языка программирования Бокс охарактеризовал как одну из наиболее интересных задач в ближайшие пять лет.

Обобщая сказанное, можно сделать вывод, что технологии XML, реализованные в Microsoft Office 2003, могут быть использованы для организации взаимодействия конечных пользователей с унаследованными системами и создания новых решений, органично интегрируемых в бизнес-процессы заказчиков.

Решения от Adobe

Недавно компания Adobe также объявила о выпуске ПО, которое расширяет формат PDF возможностями XML. Это — приложения Document Server и Adobe Document Server for Reader Extensions, предназначенные для автоматизации корпоративного документооборота, которые могут интегрироваться в системы управления ресурсами предприятий (ERP), системы автоматизации работы с клиентами (CRM) и СУБД. Приложение Adobe Document Server поддерживает команды на языке XML, а также стандарт Extensible Style Language Formatting Objects (XSL-FO), описывающий стандартные методы обработки XML-документов. Приложение Adobe Document Server for Reader Extensions после соответствующей настройки позволяет конечным пользователям заполнять, сохранять и отправлять электронные формы, используя Acrobat Reader 5.1, так как XML предлагает структуру не только для форматирования документов, но и для внедрения в документы метаданных. Это позволяет хранить в файлах описания процессов извлечения исходной информации и генерировать на основе этих описаний различные версии документов для внутреннего или внешнего использования, для архивирования, пересылки по электронной почте или вывода на печать.

4.5. Настоящее и обозримое будущее XML

XML-ориентированные СУБД могут стать весьма полезным дополнением к традиционным средствам управления базами данных, особенно если XML станет доминирующим стандартом представления сложно структурированных данных. Например, по мнению аналитиков компании Meta Group, почти 85% крупных организаций в ближайшие три года планируют перевести все свои Web-ресурсы в формат XML.

В статье Джеффа Моуда "Готовность к взлету" [31] рассказано, как BBC США поместили сервер баз данных на XML в центре новой глобальной системы доступа к служебным документам на основе Web. Это решение уже приносит отдачу, позволив ведомству сэкономить на одной из своих оборонных систем более 800 тыс. долл. в год. Внедрением этой технологии в службах BBC стал проект, реализованный в начале 2001 года с целью обеспечения поискового Web-доступа к полутораста тысячам страниц технической документации системы раннего обнаружения и наведения AWACS (Airborne Warning and Control System), связанной с самолетами наблюдения E-3.

На сегодняшний день язык XML признан во всем мире в качестве основы для развития электронной коммерции, но пока еще на рынке нет брэндов XML-СУБД, безусловных кандидатов для широкого применения. Это объясняется многими причинами, в том числе и тем, что не завершена работа над двумя важнейшими стандартами платформы XML — XPointer и XLink. Кроме того, широкому внедрению XML-технологий препятствует большая инертность администраторов корпоративных информационных хранилищ, данные в которых сегодня в большинстве своем представлены в SQL-таблицах.

Вместе с тем, все же ожидается лавинообразный переход к XML-технологиям, который породит ряд серьезных проблем. Существующие сегодня проблемы безопасности еще более усугубятся в случае массового внедрения XML. Например, ввиду открытости стандартов может приобрести невиданный размах кража информации, "широкие перспективы" откроются и для хакеров. Среда Internet, безусловно, станет более удобной для пользователей, чего нельзя однозначно утверждать по отношению к дизайнерам. Требования к дизайнерам XML-сайтов значительно возрастут, поскольку они должны будут не только заботиться о визуальном восприятии предоставляемой информации, но и конструировать сложные системы определений типов документов (DTD или схем), деревьев данных, структур гиперсвязей, метаданных и листов стилей.

Безусловно, широкое распространение языка XML стимулирует развитие электронной коммерции. XML был изначально задуман для обмена документами, и становится все более очевидным, что электронная коммерция в перспективе будет в значительной степени опираться на поток договоров, записанных в миллионах XML-документов в Internet, средой хранения которых будет эта сеть.

XML-ориентированные СУБД могут стать реальным дополнением к реляционным системам. Так, наблюдаемые тенденции развития указывают, что язык XML де-факто становится стандартом представления сложно структурированных данных. По прогнозам экспертов, мировой рынок серверного ПО на XML будет стабильно развиваться в последующие 5 лет, поскольку спрос на технологии интеграции остается стабильным. IDC прогнозирует, что этот выросший спрос на XML-продукты позволит увеличить доход в данном секторе рынка до $3,5 млрд

уже к 2006 году. Темпы роста, скорее всего, останутся интенсивными благодаря увеличению функциональности XML в приложениях и серверах. Согласно IDC, XML-базы данных и серверные решения, которые уже серьезно воздействуют на рынок, будут продолжать свой рост в дальнейшем благодаря присущим им функциональным возможностям и простоте интеграции с платформами для электронного бизнеса.

Вполне можно предположить, что XML-СУБД в перспективе позволят решить задачу прямого отражения реальных документов и алгоритмов в базах данных на описательном уровне. Тем самым уйдет в историю привлечение проектировщиков к созданию баз данных, — подобно тому, как в свое время при появлении реляционных СУБД отпала необходимость непосредственного участия системотехников в процессе создания подобных комплексов. Таким образом, в перспективе можно надеяться, что для создания баз данных больше не понадобится дорогостоящий труд специалистов, а среда хранения данных станет еще ближе их владельцам и пользователям.

Реальность состоит в том, что имеется множество информационных приложений (в том числе и приложения полнотекстового поиска), для которых данные не укладываются естественным образом в реляционную модель. Для них язык XML и сопутствующие ему стандарты представления семантики данных могут быть действительно революционным прорывом в области представления и, соответственно, поиска данных.

Глава 5

Основы технологии Text Mining

Сегодня в информационных хранилищах, распределенных по всему миру, собраны терабайты текстовых данных. Сырые неструктурированные данные составляют большую часть информации, с которой имеют дело пользователи. Найти в таких данных нечто ценное можно лишь посредством специализированных технологий. Развитие информационных ресурсов Internet многократно усугубило проблему информационной перегрузки.

Исследовательская служба Cyveillance сообщила, что еще в 2001 году общее количество страниц в Internet превысило 4 млрд. Средний размер Web-страницы — 10 Кбайт, среднестатистическая страница содержит 20–25 внутренних ссылок, 5-6 внешних и 14-15 изображений. Если к этому добавить массивы неструктурированных документов в корпоративных файловых системах и базах данных, то легко видеть, почему многие организации заинтересованы в технологиях автоматизированного анализа и классификации информации, представленной на естественном языке. Ведь по существующим оценкам, неструктурированные данные, главным образом текст, составляют не менее 90% информации. И лишь 10% приходится на структурированные данные, загружаемые в реляционные СУБД.

"Люди будут искать то, что они знают, обращаясь к документальным репозиториям. Однако они вообще не будут или просто не смогут выражать запросом то, чего они не знают, даже имея доступ к собранию документов", — заметил Джим Нисбет, вице-президент компании Semio, которая является одним из ведущих производителей систем "добычи данных" (Data Mining). "Метод эффективного анализа текста — Text Mining, — используя вычислительные мощности, позволяет выявить отношения, которые могут приводить к получению новых знаний пользователем."

Задача Text Mining — выбрать ключевую и наиболее значимую информацию для пользователя [15]. Таким образом, ему будет незачем самому "просеивать" огромное количество неструктурированной информации. Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта, технологии Text Mining как раз и предназначены для проведения смыслового анализа, обеспечения навигации и поиска в неструктурированных текстах. Применяя системы класса Text Mining, пользователи в принципе должны получить новую ценную информацию, т.е. знания.

В конце 2000 года ЦРУ опубликовало документ "Анализ плана стратегических инвестиций разведсообщества" (Strategic Investment Plan for Intelligence Community Analysis — http://www.cia.gov/cia/reports/unclass_sip — рис. 5.1). В этом документе разведчики признают, что ранее не использовали полностью возможности открытых источников, и теперь работа с ними должна стать "высшим приоритетом для инвестиций". Иначе говоря, в ЦРУ резонно по-

лагают, что брать информацию из открытых источников безопаснее и дешевле, чем пользоваться разведданными. Технология глубинного анализа текста — Text Mining — и представляет собой тот самый инструментарий, который позволяет анализировать большие объемы информации в поисках тенденций, шаблонов и взаимосвязей, способных помочь в принятии стратегических решений. Кроме того, Text Mining — это новый вид поиска, который, в отличие от традиционных подходов, не только находит списки документов, формально релевантных запросам, но и позволяет получить ответ на просьбу: "Помоги мне понять смысл, разобраться с этой проблематикой".

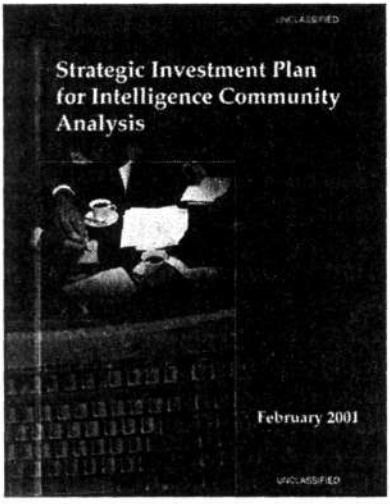


Рис. 5.1. Документ ЦРУ "Анализ плана стратегических инвестиций разведсообщества"

Клод Фогель (Claude Vogel), один из основателей легендарной компании Semio, используя аналогию с библиотекой поясняет: "Технология Text Mining открывает перед читателем книги с подчеркнутой необходимой ему информацией. Сравните это с выдачей читателю кипы документов и книг, в которых где-нибудь содержится нужная читателю информация, однако найти ее читателю будет непросто". Процесс осмысленного поиска является далеко не тривиальным, часто в коллекции документов имеется только намек на необходимую информацию. Необходимы мощные интеллектуальные возможности, чтобы найти то, что требуется. В названии технологии слово *mining* ("добыча руды") выступает как метафора отыскания глубоко "зарытой" информации.

Следует заметить, что технологии глубинного анализа текста исторически предшествовало создание технологии глубинного анализа (добычи) данных (Data Mining), методология и подходы которой широко используются и в методах Text Mining. Для "добычи текстов" вполне справедливо определение, данное для добычи данных одним из ведущих мировых экспертов Григорием Пятецким-Шапиро из GTE Labs: "Процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности". Как и большинство когнитивных технологий, Text Mining — это алгоритмическое выявление прежде неизвестных связей и корреляций в уже имеющихся текстовых данных.

Оформившись в середине 90-х годов XX века как направление анализа неструктурированных текстов, технология Text Mining сразу же взяла на вооружение методы классической добычи данных, такие как классификация или кластеризация. В Text Mining появились и дополнительные возможности, такие как автоматическое реферирование текстов и выявление феноменов, т.е. понятий и фактов. Возможности современных систем Text Mining могут применяться при управлении знаниями для выявления шаблонов в тексте, для автоматического "проталкивания" информации или ее распределения по интересующим пользователей профилям, а также для создания обзоров документов. Технологии Text

Mining, кроме того, присуща абсолютная объективность — в ней отсутствует субъективизм, свойственный человеку-аналитику.

Важный компонент технологии Text Mining связан с извлечением из текста его характерных элементов или свойств, которые затем могут использоваться в качестве метаданных документа, ключевых слов, аннотаций. Другая важная задача состоит в отнесении документа к некоторым категориям из заданной схемы систематизации. Text Mining также обеспечивает новый уровень семантического поиска документов.

5.1. Основные элементы Text Mining

В соответствии с уже сложившейся методологией, к основным элементам Text Mining относятся реферирование (summarization), выявление феноменов (feature extraction), классификация (classification), кластеризация (clustering), ответ на вопросы (question answering), тематическое индексирование (thematic indexing) и поиск по ключевым словам (keyword searching). Также в некоторых случаях указанный набор дополняют средства поддержки и создания таксономии (taxonomies) и тезаурусов (thesauri).

Александр Линден (Alexander Linden), директор компании Gartner Research, выделил четыре основных вида приложений технологии Text Mining.

1. Классификация текста, в которой используются статистические корреляции для построения правил размещения документов в предопределенные категории.
2. Кластеризация, базирующаяся на признаках документов; используются лингвистические и математические методы без применения предопределенных категорий. Результат — таксономия или визуальная карта, которая обеспечивает эффективный охват больших объемов данных.
3. Построение семантической сети или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения поиска и навигации.
4. Извлечение фактов, цель которого — получение некоторых фактов из текста с целью улучшения классификации, поиска и кластеризации.

Так сложилось, что чаще всего решаемая в Text Mining задача — это классификация, т.е. отнесение объектов базы данных к заранее определенным категориям. Фактически задача классификации — это вариант классической задачи распознавания, когда система по обучающей выборке относит новый объект к той или иной категории. Особенность же системы Text Mining состоит лишь в том, что количество таких объектов и их атрибутов может быть очень большим; поэтому должны быть предусмотрены интеллектуальные механизмы оптимизации процесса классификации. В существующих сегодня системах классификация применяется, например, для решения таких задач, как группировка документов в intranet-сетях и на Web-сайтах, размещение документов в определенные папки, сортировка сообщений электронной почты, избирательное распространение новостей подписчикам и пр.

Вторая задача — кластеризация — состоит в выделении компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти

признаки и разделить объекты по подгруппам. Решение этой задачи, как правило, предшествует задаче классификации, поскольку позволяет определить группы объектов. Различают два основных типа кластеризации — иерархическую и бинарную (двоичную). Иерархическая кластеризация заключается в построении дерева кластеров, в каждом из которых размещается небольшая группа документов. Пример утилиты двоичной кластеризации можно найти на сервере корпорации IBM по адресу `http://www.software.ibm.com/data/iminer/fortext`. Двоичная кластеризация обеспечивает группировку и просмотр документальных кластеров по ссылкам подобия. В один кластер помещаются самые близкие по своим свойствам документы. В процессе кластеризации строится базис ссылок от документа к документу, основанный на весах и совместном употреблении определяемых ключевых слов. Сегодня кластеризация широко применяется при реферировании больших документальных массивов или определении взаимосвязанных групп документов, а также для упрощения процесса просмотра при поиске необходимой информации, для нахождения уникальных документов из коллекции, для выявления дубликатов или очень близких по содержанию документов.

Можно назвать еще несколько задач, которые могут решаться средствами технологии Text Mining, — например, прогнозирование, которое состоит в том, чтобы предсказать по значениям одних признаков объекта значения остальных. Еще одна задача — нахождение исключений, т.е. поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Для этого сначала выясняются средние параметры объектов, а затем исследуются те объекты, параметры которых наиболее сильно отличаются от средних значений. Как правило, поиск исключений проводится после классификации или кластеризации, для того чтобы выяснить, насколько последние были точны.

Несколько отдельно от кластеризации стоит задача поиска связанных признаков (полей, понятий) отдельных документов. От прогнозирования эта задача отличается тем, что заранее неизвестно, по каким именно признакам реализуется взаимосвязь; цель именно в том и состоит, чтобы найти связи между отдельными признаками. Эта задача сходна с кластеризацией, но выполняется не по множеству документов, а по множеству присущих документу признаков.

И наконец, для обработки и интерпретации результатов Text Mining большое значение имеет визуализация данных, что подразумевает обработку структурированных числовых данных. Однако визуализация также является ключевым звеном при представлении схем неструктурированных текстовых документов. В частности, современные системы класса Text Mining могут осуществлять анализ больших массивов документов и формировать предметные указатели понятий и тем, освещенных в этих документах. Визуализация обычно используется как средство представления контента всего массива документов, а также для реализации навигационного механизма, который может применяться при исследовании документов и их классов.

5.2. Контент-анализ

Сегодня весьма актуальной является задача мониторинга ресурсов Internet, которая тесно связана с достаточно популярным в последние десятилетия контент-анализом. Это перспективное направление развития систем сетевой интеграции рассматривается сегодня многими экспертами как контент-мониторинг,

появление которого вызвано, прежде всего, задачей систематического отслеживания тенденций и процессов в постоянно обновляемой сетевой информационной среде. Контент-мониторинг — это содержательный анализ информационных потоков с целью получения необходимых качественных и количественных срезов, который производится постоянно, т.е. на протяжении не определяемого заранее промежутка времени. Важнейшей теоретической основой контент-мониторинга является контент-анализ, — понятие, достаточно "заезженное" социологами.

Контент-анализ начинался как количественно-ориентированный метод анализа текстов для изучения массовых коммуникаций. Впервые этот метод был применен в 1910 году социологом Максом Вебером (Max Weber), чтобы проанализировать охват прессой политических акций в Германии (рис. 5.2). Американский исследователь средств коммуникации Гарольд Лассвелл (Harold Lasswell) в 30–40-е годы использовал подобную методику для изучения содержимого пропагандистских сообщений военного времени. В 1943 году Абрахам Каплан (Abraham Kaplan) увеличил фокус контент-анализа от статистической семантики (значения текстов) политических дискуссий до анализа значений символов (семиотики). Во время Второй мировой войны растущая популярность семиотики способствовала использованию качественно-ориентированного контент-анализа для изучения "идеологических" аспектов в таких жанрах, как телевизионные шоу и коммерческая реклама. Ряд современных исследований с применением методологии контент-анализа включает, наряду с анализом текста, и анализ изображений.

Начиная с 60-х годов, с появлением средств автоматизации и текстов в электронном виде, получил начальное развитие контент-анализ информации больших объемов — баз данных и интерактивных медиасредств. Традиционное "политическое" использование современных технологий контент-анализа дополнено неограниченным перечнем рубрик и тем, охватывающих производственную и социальную сферы, бизнес и финансы, культуру и науку, что сопровождается большим количеством разнородных программных комплексов. При этом выделилось направление, получившее самостоятельное развитие — Data Mining, все еще не имеющее устойчивого русского термина-эквивалента. Так, даже выше в этой главе использовались сразу два перевода этого термина: "добыча данных" и "глубинный анализ данных".

Рис. 5.2. *Макс Вебер (1864–1920)*

Под Data Mining понимается механизм обнаружения в потоке данных интересных новых знаний, таких как модели, конструкции, ассоциации, изменения, аномалии и структурные новообразования. Большой вклад в развитие контент-анализа внесли психологические исследования в области феноменологии, ведущая идея которой заключается в обращении к каждодневному миру через различные явления (phenomena) в фактических ситуациях. С феноменологией неразрывно связаны имена ее основателя Эдмунда Хассерла (Edmund Husserl) и нашего современника Амадео Джиорджи (Amadeo Giorgi).

Однозначная трактовка понятий необходима, прежде всего, в технических системах. Развитие технологических систем невозможно без стандартизации. В качестве примера можно привести операционную систему UNIX, определение

стандартов на которую в рамках ISO (POSIX) привело к преобладанию клонов этой системы на серверных платформах. Понятие же контент-анализа, имеющее корни в психологии и социологии, сегодня пока не имеет однозначного определения. Это порождает ряд проблем, важнейшая из которых заключается в том, что программные системы, построенные на основе различных подходов к контент-анализу, будут несовместимы. Приведем лишь некоторые существующие определения контент-анализа.

- Контент-анализ — это методика объективного качественного и систематического изучения содержания средств коммуникации (Д. Джери, Дж. Джери).
- Контент-анализ — это систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника (Д. Мангейм, Р. Рич).
- Контент-анализ — это качественно-количественный метод изучения документов, который характеризуется объективностью выводов и строгостью процедуры и состоит в квантификационной обработке текста с дальнейшей интерпретацией результатов (В. Иванов).
- Контент-анализ состоит в нахождении в тексте определенных содержательных понятий (единиц анализа), выявлении частоты их встречаемости и соотношения с содержанием всего документа (Б. Краснов).
- Контент-анализ — это исследовательская техника для получения результатов путем анализа содержания текста о состоянии и свойствах социальной действительности (Е. Таршис).

Большинство из приведенных определений конструктивны, т.е. являются процедурными. Из-за разных начальных подходов они порождают различные, а порой и противоречащие друг другу алгоритмы. Принятые в современной литературе различные подходы к пониманию контент-анализа поддаются полностью оправданной критике. Так, высказываются сомнения в информационной насыщенности частотных характеристик в плане определения элементов, весомых с точки зрения содержания. Также подчеркивается игнорирование роли контекста. Однако, несмотря на многообразие трактовок контент-анализа, большое прикладное значение методологии все же позволяет избежать многих противоречий. Объединение средств и методов, их естественный отбор путем многократной оценки полученных результатов позволяют выделять и подтверждать знания, выявлять фактическую силу и полезность инструментария.

Диапазон методов и процедур, касающихся самого процесса контент-анализа, весьма широк. К примеру, при подготовке исследования выполняются следующие действия:

- описание проблемной ситуации, поиск цели исследования;
- уточнение объекта и предмета исследования;
- смысловое уточнение понятий;
- эмпирическая интерпретация понятий;
- описание процедур регистрации свойств и явлений;
- предварительный целостный анализ объекта;

- определение общего плана исследования;
- определение типа выборки и т.д.

Методы сбора данных также многообразны:

- наблюдение;
- анкетный опрос;
- интервью;
- телефонный опрос;
- накопление совокупности писем;
- получение потока документов Сети.

Для отбора информации применяются такие методы:

- гнездовой;
- квотная выборка;
- неслучайная выборка;
- метод нетипичных представителей;
- метод "снежного кома";
- стихийная выборка;
- случайная выборка;
- одно- и многоступенчатая выборка;
- районированная (расслоение) выборка;
- систематическая выборка и т.д.

В контент-анализе применяются такие математические методы, как:

- дисперсионный анализ для выявления влияния отдельных, независимых факторов на наблюдаемый признак;
- кластерный анализ для классификации объектов и описывающих их признаков;
- логлинейный анализ для статистической проверки гипотезы о системе одновременных парных и множественных взаимосвязей в группе признаков;
- причинный анализ для моделирования причинных отношений между признаками с помощью систем статистических уравнений;
- регрессионный анализ для исследования регрессионной зависимости между зависимыми и независимыми признаками;
- факторный анализ для получения обобщенной информации о структуре связи между наблюдаемыми признаками изучаемого объекта на основе выделения скрытых факторов;
- корреляционный анализ для выявления зависимости между числовыми случайными величинами, одна из которых зависит и от ряда других случайных факторов.

От поиска информации — к поиску знаний

В последнее время происходят изменения подходов к форме и семантике взаимодействия пользователей с поисковыми системами в Internet. Через десятилетие после возникновения первых поисковых серверов в Сети оказалось, что надежды на интеллект пользователя при формулировке запросов были тщетны. Это замечание относится как к обычным пользователям, так и к пользователям-профессионалам. В результате современные поисковые системы сами все более интеллектуализируются, включают семантические инструменты, пытаясь выявлять информационные потребности пользователей и учитывать их при поиске.

Сегодня естественно желание пользователя видеть достаточно короткий список классов, в который попадают все возвращенные информационно-поисковой системой документы. Пользуясь этой классификацией, пользователь сможет существенно сузить границы своего поиска. При этом к классификации предъявляются такие два основных требования:

- классы должны содержать близкие по смысловому признаку документы;
- этот признак должен быть основой названия класса, которое должно восприниматься пользователем.

Новые подходы к организации поиска заставляют заново взглянуть на модели представления информации в базах данных поисковых машин и методы автоматической группировки, применяемые при поиске информации в сети Internet.

5.3. Модели поиска

В настоящее время используется несколько подходов к представлению информации в базах данных для обеспечения последующего поиска этой информации [65, 67]. Рассмотрим два наиболее популярных подхода. Первый базируется на теории множеств, а второй на векторной алгебре. Оба подхода достаточно эффективны на практике, однако у них есть общий недостаток, который следует из основного упрощающего предположения, заключающегося в том, что смысл документа, его основное содержание определяется множеством ключевых слов — терминов и понятий, входящих в него. Конечно же, такие подходы частично ведут к потере содержательных оттенков текстов, зато позволяют выполнять быстрый поиск и группировку документов по формальным признакам. Сегодня эти подходы — самые популярные. Следует заметить, что существуют и другие методы, например семантические, в рамках которых делаются попытки выявить смысл текста за счет анализа грамматики текста, использования баз знаний и различных тезаурусов, отражающих семантические связи между отдельными словами и их группами. Очевидно, что такие подходы требуют больших затрат на поддержку баз знаний и тезаурусов для каждого языка, тематики и вида документов.

5.3.1. Булева модель поиска

Булева модель является классической и широко используемой моделью представления информации, базирующейся на теории множеств, и, следовательно, моделью информационного поиска, базирующейся на математической логике. Популярность этой модели связана, прежде всего, с простотой ее реализации, позволяющей индексировать и выполнять поиск в массивах документов большого объема. В настоящее время популярным является объединение булевой модели

с алгебраической векторно-пространственной моделью представления данных, что обеспечивает, с одной стороны, быстрый поиск с использованием операторов математической логики, а с другой стороны — качественное ранжирование документов, базирующееся на весах входящих в них ключевых слов.

В рамках булевой модели документы и запросы представляются в виде множества морфемных основ ключевых слов, будем их в дальнейшем называть термами. Пусть документальный массив C состоит из множества документов $d_1, \ldots, d_n$, а документ d_i содержит множество различных термов $T(d_i)$. Обозначим через $T = \bigcup_{i=1,\ldots,n} T(d_i)$ словарь массива C, представляющий собой множество всех термов, встречающихся в документах из C, и через $T(d_i)$ — словарь документа d_i. В булевой модели запрос пользователя представляет собой логическое выражение, в котором ключевые слова (термы запроса) связаны логическими операторами AND, OR и NOT. В различных поисковых системах в Internet пользователи могут пользоваться умолчаниями, не используя в явном виде логических операций, а просто перечисляя ключевые слова. Чаще всего по умолчанию предполагается, что все ключевые слова соединяются логической операцией AND — в этих случаях в результаты поиска включаются только те документы, которые содержат одновременно все ключевые слова запроса. В тех системах, в которых пробел между словами приравнивается к оператору OR, в результаты поиска включаются документы, в которые входит хотя бы одно из ключевых слов запроса.

При использовании булевой модели база данных включает индекс, организуемый в виде инвертированного массива, в котором для каждого терма из словаря базы данных содержится список документов, в которых этот терм встречается.

В индексе могут храниться также значения частоты вхождения данного терма в каждом документе, что позволяет сортировать список по убыванию частоты вхождения. Классическая база данных, соответствующая булевой модели, организована таким образом, чтобы по каждому терму можно было быстро получить доступ к соответствующему списку документов. Кроме того, структура инвертированного массива обеспечивает его быструю модификацию при включении в базу данных новых документов. В связи с этими требованиями, инвертированный массив часто реализуется в виде B-дерева.

Существует несколько подходов к формированию архитектуры поисковых систем, соответствующих булевой модели и нашедших свое воплощение в реальных системах. Одной из наиболее удачных реализаций структуры базы данных информационно-поисковой системы на мэйнфреймах фирмы IBM была признана модель данных системы STAIRS (Storage and Information Retrieval System), которая, благодаря изначально удачным архитектурным решениям до сих пор продолжает развиваться. База данных информационно-поисковых систем этой традиционной архитектуры состоит из следующих основных таблиц [27]:

- текстовой, содержащей текстовую часть всех документов;
- таблицы указателей текстов, включающей указатели местонахождения документов в текстовой таблице, а заодно и форматные поля всех документов;
- словарной, содержащей все уникальные слова, встречающиеся в полях документов, т.е. те слова, по которым может осуществляться поиск. Слова могут быть связаны в синонимические цепочки;
- инверсной, содержащей списки номеров документов и координаты всех вхождений отдельных слов в полях документов.

Процессы, происходившие при поиске информации в базе данных STAIRS, сегодня реализуются средствами современных СУБД и ИПС документального типа. Поиск термина в базе данных осуществляется следующим образом.

1. Происходит обращение к словарной таблице, по которой определяется, входит ли слово в состав словаря базы данных, и если входит, то определяется ссылка на цепочку появлений этого слова в документах.
2. Выполняется обращение к инверсной таблице, по которой определяются координаты всех вхождений терма в текстовую таблицу базы данных.
3. По номеру документа происходит обращение к записи таблицы указателей текстов. Каждая запись этого файла соответствует одному документу в базе данных.
4. По номеру документа осуществляется прямое обращение к фрагменту текстовой таблицы — документу — и последующий его вывод.
5. В случае, когда обрабатывается выражение, состоящее не из одного слова, а из некоторого словосочетания, в результате отработки поиска по каждому слову запроса формируется массив записей, соответствующих вхождению этого термина в базу данных. После окончания формирования массивов результатов поиска происходит выявление релевантных документов путем выполнения теоретико-множественных операций над записями этих массивов.

5.3.2. Векторно-пространственная модель

Большинство известных информационно-поисковых систем и систем классификации информации в той или иной мере основываются на использовании векторной модели описания данных (Vector Space Model) [66, 68]. Векторная модель является классической алгебраической моделью. В рамках этой модели документ описывается вектором в некотором евклидовом пространстве, в котором каждому используемому в документе терму ставится в соответствие его весовой коэффициент (значимость), который определяется на основе статистической информации о его вхождении в отдельном документе или в документальном массиве. Описание запроса, который соответствует необходимой пользователю тематике, также представляет собой вектор в том же евклидовом пространстве термов. В результате для оценки близости запроса и документа используется скалярное произведение соответствующих векторов описания тематики и документа.

В рамках этой модели с каждым термом t_i в документе d_j (и запросе q) сопоставляется некоторый неотрицательный вес w_{ij}. Таким образом, каждый документ и запрос могут быть представлены в виде k-мерного вектора $\|w_{ij}\|_{i=1,\ldots,k}$, где k — общее количество различных термов во всех документах. Согласно векторной модели, близость документа d_i к запросу q оценивается как корреляция между векторами их описаний. Эта корреляция может быть вычислена как скалярное произведение соответствующих векторов описаний. При этом весовые коэффициенты отдельных термов можно вычислять множеством различных способов.

Один из возможных простейших (но эффективных) подходов — использовать в качестве веса терма w_{ij} в документе d_i нормализованную частоту его использования $freg_{ij}$ в данном документе.

$$w_{ij} = tf_{ij} = freq_{ij} / \max_i freq_{il}$$

Этот подход не учитывает частоту вхождения отдельного терма во всем информационном массиве, так называемую дискриминационную силу терма. Поэтому в случае, когда доступна статистика использований термов во всем информационном массиве, более эффективно следующее правило вычисления весов:

$$w_{ij} = tf \times idf_{ij} = tf_{ij} \times log\ N / n_i,$$

где n_i — число документов, в которых используется терм t_j, а N — общее число документов в массиве.

Обычно значения весов w_{ij} нормируются (дополнительно делятся на квадратный корень из суммы весов всех термов, входящих в документ), что позволяет рассматривать документ как ортонормированный вектор. Такой метод взвешивания термов имеет стандартное обозначение — *tf×idf*, где *tf* указывает на частоту использования термина в документе (term frequency), а *idf* — на величину, обратную числу документов массива, содержащих данный терм (inverse document frequency).

Когда возникает задача определения тематической близости двух документов или документа и запроса, в этой модели используется простое скалярное произведение $sim(d_1, d_2)$ двух векторов $\|w_{i1}\|_{i=1,..,k}$ и $\|w_{i2}\|_{i=1,..,k}$, которое, очевидно, соответствует косинусу угла между векторами-образами документов d_1 и d_2. Очевидно, $sim(d_1, d_2)$ принадлежит диапазону [0, 1]. Чем больше величина $sim(d_1, d_2)$ — тем более близки документы d_1 и d_2. Для любого документа d_i имеем $sim(d_i, d_i) = 1$. Аналогично мерой близости запроса q г документу d_i считается величина $sim(q, d_i)$.

Векторно-пространственная модель представления данных автоматически обеспечивает системам, построенным на ее основе, такие возможности:

- обработку сколь угодно больших запросов;
- простую реализацию режима поиска документов, подобных уже найденным;
- сохранение результатов поиска в некотором виртуальном массиве с последующим уточняющим поиском в нем.

5.3.3. Гибридные модели поиска

Несмотря на то что приведенные выше модели являются классическими, в чистом виде они применяются только в моделях систем. На практике чаще всего используются гибридные подходы, в которых объединены возможности булевой и векторно-пространственной моделей и зачастую добавлены оригинальные методы семантической обработки информации. Чаще всего в информационно-поисковых системах процедура поиска выполняется в соответствии с булевой моделью, а результаты ранжируются по весам в соответствии с моделью векторного пространства.

5.4. Группировка текстовых данных

Названные выше модели представления данных обладают общим недостатком, связанным с большой размерностью как векторного пространства (векторная модель), так и множества (булева модель). Для обеспечения эффективной работы необходимо сгруппировать как подмножества термов, так и тематически подобные документы. Только в этом случае может быть обеспечена обработка информационных массивов в режиме реального времени. В этом случае на помощь приходят два основных приема группировки — классификация

и кластеризация. Здесь классификация — это отнесение каждого документа к определенному классу с заранее известными признаками, полученными на этапе обучения. Число классов строго ограничено.

Тематические каталоги, построенные с участием людей (например, Yahoo! или Open Directory), приводят к естественному вопросу: а не могут ли подобные каталоги быть построены автоматически? Один из путей решения этой проблемы — кластеризация, т.е. автоматическая группировка тематически близких документов.

При кластеризации гипертекстовых документов возникают некоторые осложнения, связанные с множественностью выбора алгоритмов кластеризации. Разные алгоритмы используют различные алгоритмы подобия при наличии большого количества признаков.

Гипертекст достаточно богат возможностями: текстовые блоки, теги разметки, URL-адреса, имена доменов в URL, подстроки в URL, которые могут быть значащими словами, и т.д. Как в этом случае определить меру подобия таким образом, чтобы достичь хорошей кластеризации?

Как только класс определен методом кластеризации, возникает необходимость его сопровождения, так как Сеть постоянно изменяется и растет. В этом случае на помощь приходит классификация. Механизм классификации сначала обучается на основе выявления признаков документов, которые соответствуют определенным темам. На этой стадии определяются корреляции между отдельными признаками, после чего механизм становится способен классифицировать новые документы.

Классификация и кластеризация представляют собой две противоположные крайности в отношении человеческого участия в процессе группировки документов.

Механизм классификации обычно обучается на отобранных документах только после того, как заканчивается стадия автоматического выявления классов (кластеров).

Кластеризация — это разбиение множества документов на кластеры, представляющие собой подмножества, смысловые параметры которых заранее неизвестны. Количество кластеров может быть произвольным или фиксированным. Если классификация предполагает приписывание документам определенных, известных заранее признаков, то кластеризация — это более сложный процесс, который предполагает не только приписывание некоторых признаков, но и выявление самих этих признаков-классов.

Итак, основная идея современных методов кластеризации — снижение размерности пространства признаков, по которым происходит классификация документов. В то время как классификация документов заключается в автоматическом определении тематики документа по заданному множеству возможных тематик, задачей кластеризации документов является автоматическое выявление групп семантически подобных документов. Однако, в отличие от классификации, тематическая ориентация этих групп не задана заранее. Иными словами, цель кластеризации некоторого множества документов состоит в выделении подмножеств (кластеров), где все документы, попавшие в один кластер, в определенном смысле будут близки друг другу. Иначе говоря, кластер можно рассматривать как группу документов со схожими признаками. Цель всех методов кластеризации заключается в том, чтобы схожесть документов, попадающих в кластер, была максимальной, семантической.

Числовые методы кластеризации базируются на определении кластера как множества документов, 1) значения семантической близости между любыми двумя элементами которого не меньше определенного порога или 2) значения

близости между любым документом множества и центроидом этого множества не меньше определенного порога. Под центроидом кластера в этом случае понимается вектор, который вычисляется как среднее арифметическое векторов всех документов кластера. Нечисловые семантические методы кластеризации не накладывают таких ограничений на кластеры, однако в результате применения большинства семантических методов в полученных множествах приведенные условия близости, как правило, выполняются.

Начальным пространством признаков обычно является пространство термов, которое сжимается в результате анализа большого массива документов. Для проведения такого анализа используются различные подходы — весовой, вероятностный, семантический и т.д., определяющие правила классификации.

В области информационного поиска кластеризация применяется для решения двух задач — группировки документов и результатов поиска.

При использовании векторно-пространственной модели представлений данных в информационно-поисковых системах всегда актуальна задача снижения размерности, что должно повысить скорость обработки и выполнения быстрого поиска по заданному векторному образу запроса релевантных ему векторных представлений документов. Если разбить все множество документов на кластеры, содержащие семантически близкие друг другу документы, то можно реализовать следующую процедуру: сравнить образ запроса с центроидами ("типичными представителями" — осредненными значениями векторов из кластера), выбрать кластеры, центроиды которых наиболее близки запросу, после чего сравнить запрос со всеми документами в выбранных кластерах.

Таким образом, процедурно все множество документов разбивается на несколько кластеров, каждый из которых содержит множество близких друг другу документов, и для каждого кластера находится центроид — документ, образ которого расположен наиболее близко к геометрическому центру кластера. В этом случае поиск по запросу разбивается на два этапа. Вначале запрос сопоставляется с центроидами всех кластеров и определяются кластеры, образы центроидов которых наиболее близки образу запроса. Далее поиск проводится исключительно в выбранных кластерах.

5.4.1. Кластеризация

В результате выполнения поисковой процедуры пользователю предъявляются списки документов, как правило, упорядоченные по убыванию соответствия запросу. В результате неизбежных неточностей при ранжировании результатов поиска, такой вид представления не всегда оказывается удобным.

И тогда на помощь приходит кластеризация результатов поиска, которая позволяет представить полученные результаты в обобщенном виде, что упрощает выделение области, соответствующей информационным потребностям пользователя [73].

В этом случае используют два класса методов кластеризации — иерархический и неиерархический. Наиболее популярны сегодня методы иерархической кластеризации, которые благодаря своей простоте широко применяются в современных информационных системах.

При иерархической кластеризации (снизу вверх либо сверху вниз) формируется дерево кластеров. При иерархической кластеризации снизу вверх два документа, попавшие в один кластер, будут принадлежать одному и тому же кластеру и на более высоких уровнях иерархии. При использовании кластеризации сверху вниз документы, попавшие в различные кластеры, будут принадлежать различным

кластерам на более низких иерархических уровнях. Иначе говоря, принятое один раз решение о принадлежности документов одному (кластеризация снизу вверх) или разным (кластеризация сверху вниз) кластерам в дальнейшем не пересматривается, что обеспечивает вычислительную простоту и эффективность метода.

Методы неиерархической кластеризации обеспечивают качественную кластеризацию за счет более сложных алгоритмов. Для этих методов, как правило, имеется некоторая пороговая функция качества кластеризации, максимизация которой достигается за счет распределения документов между отдельными кластерами.

5.4.2. Тематическая близость

Теоретически предполагается, что тематика документа определяется его словарным запасом, а тематическая близость термов характеризуется тем, насколько часто эти термы используются в документах одной и той же тематики. Отметим, что это не всегда подразумевает обязательное использование этих термов в одних и тех же документах.

Обозначим тематическую близость двух термов w_i и w_j как $FSR(w_i, w_j)$. Вычисление оценок тематической близости термов и, как следствие, задание функции FSR выполняются по результатам анализа использования термов в массиве документов, которыми описываются тематики. По исходному массиву документов строится матрица A, строки которой отражают распределение термов по документам. В качестве оценки тематической близости двух термов используется скалярное произведение соответствующих строк этой матрицы. Таким образом, для вычисления оценок близости между всеми парами термов достаточно вычислить матрицу AA^T.

Такой подход аналогичен классическим методам представления информации, основанным на векторно-пространственной модели. Поэтому ему присущи следующие недостатки:

- не определяет зависимости между термами, которые используются в документах одной и той же тематики, но редко встречаются вместе;
- случайные неточности и зависимости оказывают существенное влияние на получаемые оценки и негативно влияют на точность метода;
- размер матрицы A очень велик — использование этой матрицы весьма ресурсоемко.

Дальнейшим развитием такого подхода является использование так называемого латентно-семантического анализа (LSA). По матрице AA^T строится ее аппроксимация $\hat{A}\hat{A}^T$, где $\hat{A}$ — это аппроксимация A, полученная методом латентно-семантического анализа (подробнее на этом мы остановимся далее).

Функция тематической близости двух термов $FSR(w_1, w_2)$ однозначно задается матрицей $\hat{A}\hat{A}^T$:

$$FSR(\mathrm{w_i}, \mathrm{w_j}) = \hat{A}\hat{A}^T[\mathrm{w_i}, \mathrm{w_j}].$$

Отметим, что матрица $\hat{A}\hat{A}^T$ имеет размерность k, где k — это выбранная при аппроксимации желаемая размерность пространства тематик. Таким образом, при данном подходе трудоемкость вычисления тематической близости двух термов составляет $O(k)$, т.е. она не зависит от количества анализируемых документов и размера общего словаря.

Таблица взаимосвязей понятий

В качестве основы для группировки документов в информационном массиве можно рассмотреть понятия (не отдельные термины, а некоторые семантические сущности), которые, теоретически, можно выразить языком запросов. Точно так же, как и в случае отдельных термов, кластеризация документов сопоставляется с кластеризацией понятий, при этом понятия более точно отражают тематические свойства документов. Конечно же, это достигается за счет усложнения алгоритмической части кластеризации. Построение таблиц взаимосвязей понятий (ТВП) базируется на языковых средствах информационно-поисковой системы, а также методах кластерного анализа. Семантическое значение понятий определяется на основе информационно-поискового языка.

Таблица взаимосвязей понятий, которая строится как статистический отчет, отражающий близость (совместное вхождение в документах) отдельных понятий из реального мира, — это симметричная матрица $A = \|a_{ij}\|$, элементы которой a_{ij} — это коэффициенты взаимосвязи соответствующих пар понятий. Коэффициент a_{ii} соответствует количеству документов входного информационного потока, которые включают понятие (термины или словосочетания, представленные на языке запросов, соответствующие понятию) i, а коэффициент a_{ij}, где $(i \neq j)$, — количеству документов во входном потоке, которые одновременно соответствуют понятиям i и j.

Предполагается, что качественные признаки вполне адекватно выражаются информационно-поисковым языком. Как показывает практика, это решение в большинстве случаев является достаточно эффективным и оперативным (реализуется быстро просчитываемыми алгоритмами).

Для переупорядочения понятий с целью выявления блоков — множеств наиболее взаимосвязанных понятий — применяется алгоритм кластерного анализа. Например, для выделения двух таких блоков необходимо выделить два понятия-полюса (соответствующих, например, индексам k и l), наиболее тесно связанных с другими понятиями, но минимально связанных между собой. Формально эти условия можно записать таким образом:

$$\Sigma(a_{ik} - a_{kk}) \rightarrow max$$

$$\Sigma(a_{il} - a_{ll}) \rightarrow max$$

$$a_{kl} \rightarrow min,\ k \neq l$$

Остальные понятия (например, понятие i) относятся к блоку k, если $a_{ik} > a_{il}$. В противном случае понятие i будет отнесено к блоку l.

При визуализации ТВП ее отдельные ячейки, соответствующие отдельным элементам матрицы A, отображаются различными оттенками серого цвета (в зависимости от значений коэффициентов взаимосвязи a_{ij}).

Процедура построения таблицы взаимосвязей понятий предназначена для практического выявления взаимосвязанных понятий, их перегруппировки, визуализации и фрагментации входного документального массива.

Процедура построения ТВП принимает на своем входе два потока — документальный массив и таблицу понятий (ТП), строки которой представляют собой названия понятий и запрос на информационно-поисковом языке, соответствующий этому разделу. На первом этапе построения таблицы взаимосвязей понятий должен быть построен текстовый файл взаимосвязей понятий, который соответствует матрице $A = \{a_{ij}\}$, где a_{ij} — коэффициенты взаимосвязей понятий i и j.

В файле, который соответствует матрице A, первая строка будет соответствовать первому понятию и заполняется коэффициентами взаимосвязей с другими понятиями.

Коэффициент a_{ii} будет соответствовать количеству документов во входном массиве, которые соответствуют понятию i, а коэффициент a_{ij} — количеству документов, которые одновременно соответствуют понятиям i и j. Алгоритм определения элементов матрицы A такой:

1) все элементы матрицы A устанавливаются равными;
2) осуществляется попытка чтения очередного документа из входного массива. Если эта попытка успешна, то происходит переход к п. 3, иначе — к п. 4;
3) для каждой пары i и j происходит проверка соответствия входной записи понятиям i и j. Если соответствие установлено, то коэффициент a_{ij} увеличивается на единицу, после чего выполняется переход к п. 2;
4) если был обработан хотя бы один документ, то построенная таблица взаимосвязей понятий считается сформированной.

На втором этапе построения ТВП выполняется перегруппировка понятий в зависимости от значений элементов матрицы A. Перегруппировка происходит путем одновременной перестановки строк и столбцов этой матрицы с целью сведения ее к блочно-диагональному виду. Диагональные блоки соответствуют кластерам обобщенных понятий.

На третьем этапе процедуры происходит визуализация ТВП для удобного представления взаимосвязей понятий.

На последнем этапе осуществляется формирование типовых запросов для последующей группировки документов, т.е. реализуются механизмы фрагментации документального массива.

5.4.3. Вероятностная модель

Рассматриваемая модель поиска базируется на теоретических подходах байесовских условных вероятностей. Основным подходом вероятностной модели является вероятностная оценка веса термов в документе. С другой стороны, в качестве оценки соответствия документа запросу используется вероятность того, что пользователь признает документ релевантным.

При описании вероятностной модели, как и ранее, используется словарь массива, включающий все термы, встречающиеся хотя бы в одном документе из информационного массива. С документом сопоставляется вектор $x=(t_1, \ldots, t_n)$, компонента i которого равна 1, если терм i входит в данный документ, и 0 — в противном случае. Здесь, как и ранее, терм задается своим порядковым номером в словаре, а n — общее количество термов в словаре коллекции. Далее будем считать фиксированным некоторый запрос q. Обозначим через W_1 событие, состоящее в том, что рассматриваемый документ релевантен запросу q, а через W_2 — событие, состоящее в том, что рассматриваемый документ не релевантен запросу q. В этом случае $P(W_i|x)$ — вероятность того, что для документа x наступает событие W_i. Зная эту вероятность, можно использовать следующее правило, используемое при поиске: если $P(W_1|x) > P(W_2|x)$, то документ, представленный вектором x, релевантен запросу q. Теорема Байеса позволяет перейти к вероятностям, значения которых удобнее оценить так:

$$P(\mathrm{W_i}|\mathrm{x}) = P(\mathrm{x}|\mathrm{W_i})\, P(\mathrm{W})/P(\mathrm{x}).$$

В вероятностной модели используется упрощение, заключающееся в предположении (в общем случае, неточном) независимости вхождения в документ любой пары термов.

В этом случае

$$P(\mathrm{x}|\mathrm{W_i}) = P(\mathrm{x}_1|\mathrm{W}i) \times \ldots \times P(\mathrm{x_n}|\mathrm{W_i}).$$

Если использовать следующие обозначения: $p_i = \mathrm{P}(x_i = 1|w_1)$, $q_i = \mathrm{P}(x_i = 1|w_2)$, то

$$P(\mathrm{x}|\mathrm{W}_1) = \Pi_{\mathrm{i}\,=\,1,\,\ldots,\,\mathrm{n}}\ \mathrm{p_i}^{\mathrm{xi}}(1 - \mathrm{p_i})^{1\,\cdot\,\mathrm{xi}}$$

и

$$P(\mathrm{x}|\mathrm{W}_2) = \Pi_{\mathrm{i}\,=\,1,\,\ldots,\,\mathrm{n}}\ \mathrm{q_i}^{\mathrm{xi}}(1 - \mathrm{q_i})^{1\,\cdot\,\mathrm{xi}}.$$

Неравенство, определяющее релевантность документа запросу, можно переписать следующим образом:

$$log\ (P(\mathrm{x}|\mathrm{W}_1)\ P(\mathrm{W}_1)/P(\mathrm{x}|\mathrm{W}_2)\ P(\mathrm{W}_2)) > 0.$$

Пусть N — общее число документов в информационном массиве; R — число документов, релевантных запросу q; n_i — число документов, в которых имеется терм с номером i; r_i — число документов, релевантных запросу q и включающих терм с номером i. В этих обозначениях $p_i \approx r_i/R$ и $q_i \approx n_i - r_i/N - R$. В качестве веса терма с номером i в документе, представленном вектором x, можно взять величину

$$\mathrm{W(i)} = log(\mathrm{r_i} \times (\mathrm{N} - \mathrm{R} - \mathrm{n} + \mathrm{r_i})/(\mathrm{n_i} - \mathrm{r_i})(\mathrm{R} - \mathrm{r_i})).$$

При выполнении информационного поиска, благодаря режиму обратной связи по релевантности, можно итеративным путем уточнять вес термов. В начале поиска вес терма i вычисляется по формуле:

$$\mathrm{W(i)} = log(\mathrm{N} - \mathrm{n_i}/\mathrm{n_i}) \approx log\ (\mathrm{N}/\mathrm{n_i}).$$

Затем на каждой итерации поиска можно определять множество документов, отмеченных пользователем как соответствующие его информационным потребностям. Их общее число можно принять за некоторую оценку величины R, а число отмеченных документов, содержащих термы с номером i, служит основой оценки величины r_i.

Подход к решению проблемы спама

Приведенная выше вероятностная модель неожиданно нашла широкое применение в такой актуальной сфере информационного поиска, как борьба с несанкционированной рассылкой электронной почты, получившей название "спам". Непрошеные рекламные рассылки по электронной почте являются одной из наиболее серьезных проблем Internet. Нередки случаи, когда спам проникает сквозь фильтры, а обычные письма, напротив, оказываются в папке со спамом. Важное направление борьбы со спамом заключается в совершенствовании и интеллектуализации спам-фильтров.

Американский исследователь и программист Пол Грэм (Paul Graham), ранее известный как разработчик электронного магазина Viaweb Store, известного в настоящее время как Yahoo! Store, опубликовал в Internet статью "A Plan for Spam" (`http://www.paulgraham.com/spam.html`) [47], весьма подробно описывающую эффективный метод борьбы со спамом. Этот метод основывается на теории вероятностей и использует для фильтрации спама алгоритм Байеса [48]. Суть этого метода состоит в статистической фильтрации — применении математической теоремы Байеса к входящим электронным письмам. Эта теорема

позволяет вычислить вероятность некоторого события на основе статистики совершения этого события в прошлом.

С точки зрения математической статистики, выявление спама — это типичная задача выбора из двух гипотез. Если обозначить H_0 гипотезу того, что электронное сообщение является спамом, а H_1 — что оно спамом не является, то опровержение гипотезы H_0 означает принятие H_1 и, соответственно, наоборот.

Качество критерия выявления спама определяется вероятностями принятия и опровержения каждой из гипотез в зависимости от того, какая из гипотез верна. Обычно этот критерий характеризуется вероятностями ошибок. Ошибка первого рода заключается в отвержении истинной гипотезы о спаме (H_0). Эта ошибка обозначается буквой α.

$$\alpha = P(\mathrm{H}_1/\mathrm{H}_0)$$

Ошибка второго рода (β) — принятие гипотезы о спаме, когда на самом деле верна ее альтернатива.

$$\beta = P(\mathrm{H}_0/\mathrm{H}_1)$$

Таким образом, α — уровень значимости критерия, а 1 – β — его мощность. Обычно уровень значимости α выбирается заранее (например, 0,5%), а мощность стараются сделать максимальной.

Метод Байеса подразумевает использование статистической оценочной базы — двух наборов электронных писем, один из которых составлен из спама, а другой — из обычных писем. При создании этой базы подсчитывается количество вхождений каждого отдельного термина в каждом наборе, и на основании этого для каждого терма вычисляется оценка того, что письмо, содержащее этот термин, является спамом.

В разработанном Грэмом прототипе фильтра каждому встречающемуся в электронной переписке слову или тегу присваивается значение вероятности его наличия в спаме. Грэм разработал алгоритм отсева спама, основанный на формуле Байеса:

$$P(\mathrm{H_i}|\mathrm{U_i}\mathrm{I}) = P(\mathrm{H_i}|\mathrm{I}) \times P(\mathrm{U_i}|\mathrm{H_i},\mathrm{I})/P(\mathrm{U_i}|\mathrm{I}),$$

где $P(H_i|I)$ — начальная вероятность того, что гипотеза H верна, исходя из имеющегося опыта I; $P(U_i|I)$ — вероятность наблюдения события U_i, исходя только из опыта I; $P(U_i|H_i,I)$ — вероятность наблюдения события U_i, исходя как из опыта I, так и из гипотезы H_i; $P(H_i|U_iI)$ — постериорная вероятность истинности гипотезы H_i на основании опыта I и полученных экспериментальных наблюдений U_i.

Оценка принадлежности конкретного слова спаму измеряется по шкале от 0 до 1. Значение 0 означает отсутствие спама, а 1 — 100% уверенность в том, что это слово принадлежит к спаму.

Пусть письмо содержит n слов с оценками $S1...Sn$. Тогда общая оценка принадлежности письма к спаму S может быть вычислена, например, по следующей формуле:

$$S = S1 \times S2 \times ... \times \mathrm{Sn}/(S1 \times S2 \times ... \times \mathrm{Sn} + (1 - \mathrm{S1}) \times (1 - \mathrm{S2}) \times ... \times (1 - \mathrm{Sn})).$$

Полученная оценка определяет условную вероятность принадлежности электронного письма к спаму на основании существующей оценочной базы.

Следует отметить, что математики называют применяемый Грэмом метод "наивным" байесовским, поскольку принимается заведомо неверная гипотеза о независимости появления отдельных слов в письме.

Как уже говорилось, в разработанном Грэмом прототипе фильтра каждому встречающемуся в электронной переписке слову или тегу присваивается значение вероятности его наличия в спаме. На основе этих вероятностей с помощью байесовского подхода для электронного письма вычисляется вероятность того, что данное письмо является спамом. Высокая вероятность присваивается как словам вроде sexy или promotion, так и термам ff0000 — код ярко-красного цвета в языке HTML. Соответственно, низкая вероятность соответствует профессиональным терминам или просто редко используемым в рекламе словам. Именно переход от условных вероятностей того, что слова, входящие в письмо, относятся к спаму, к вычислению вероятности того, что данное письмо является спамом, и реализуется формулами Байеса.

Для статистической фильтрации спама не требуется вычисления оценки письма по всем входящим в него словам. Выбираются лишь наиболее значимые с точки зрения оценок. Уровень значимости определяется тем, насколько оценка слова отличается от нейтральной.

Эвристическим параметром для статистической фильтрации спама является количество слов, по которым оценивается электронное письмо. Пол Грэм предложил в качестве такого параметра число 15.

В процессе испытания системы фильтрации спама Грэм пропустил через нее 8000 писем, половина из которых являлась спамом. В результате через фильтры смогли просочиться лишь 0,5% рекламных сообщений (ошибка первого рода), а количество ошибочных срабатываний фильтра на основе байесовского подхода оказалось нулевым (мощность критерия оказалась стопроцентной!).

По мнению Грэма, для того чтобы система была действительно эффективной, она должна поддерживать возможность индивидуальной настройки, поскольку терминология, используемая в электронной переписке разными людьми, отличается. Если же пользователь будет регулярно помечать рекламные письма как СПАМ, то программа сможет накопить достаточно информации для эффективной фильтрации электронной почты.

Отличия технологии статистической фильтрации от технологии фильтрации на основе отдельных признаков заключаются в следующем.

- Особенностью статистической технологии является возможность индивидуальной автоматической настройки фильтра — разные люди используют в электронной переписке различную лексику. Настройка фильтра производится по результатам статистического анализа существующего у пользователя архива электронной почты.
- В обоих случаях вычисляется "вес" письма. Однако при использовании метода учета отдельных признаков "вес" письма вычисляется только на основе признаков спама, что в результате часто приводит к ложному принятию решения (ошибка второго рода).
- В алгоритме Байеса наборы признаков определяются объективно — в результате статистического анализа реальных архивов писем. Получаемые наборы признаков оказываются весьма нетривиальными и эффективными. Например, в качестве "плохого" признака может появиться строка "0Xffffff" — ярко-красный цвет; а в качестве "хорошего" — номер телефона или другие персональные данные.

Грэм разработал вариант своего фильтра на созданном им самим языке Arc (вариант LISP). В свою очередь, группа энтузиастов в настоящее время работает над проектом spambayes (`http://spambayes.sourceforge.net`), целью которого является разработка спам-фильтра на основе байесовского алгоритма на языке Python.

5.4.4. Латентно-семантический анализ

Латентно-семантический анализ, или индексирование, (LSA/LSI) — это теория и метод извлечения "скрытых" контекстно-зависимых значений термов и структуры семантических взаимосвязей между ними путем статистической обработки больших наборов текстовых данных [52]. Этот метод широко используется в области поиска и в задачах классификации информации.

Данный подход позволяет автоматически распознавать смысловые оттенки слов в зависимости от контекстов их использования. Он использует выявленные показатели тематической близости термов (см. выше), которые затем применяются для вычисления оценок тематической близости документов.

Метод LSA широко применяется в факторном анализе. Задачей факторного анализа является выделение главных факторов из пространства элементарных. В большинстве случаев задача нахождения главных факторов решается с помощью алгебраического метода главных компонентов и сингулярного разложения матриц. В случае информационного поиска под факторами понимаются некоторые семантические сущности, которые зачастую не имеют определенных названий, выбор которых — открытая задача.

Матричный латентно-семантический анализ

Математический аппарат данного метода базируется на сингулярном разложении матриц. Метод позволяет выявить скрытые семантические связи при обработке больших массивов документов.

В качестве исходной информации LSA использует ту же матрицу, что и в векторно-пространственной модели. Элементы этой матрицы содержат значения частоты использования отдельных термов в документах.

Из матричного анализа известно, что любая прямоугольная матрица A может быть разложена в произведение трех матриц: $A = U\Sigma V^T$, таких, что матрицы U и V состоят из ортонормированных колонок, а Σ — диагональная матрица сингулярных значений, диагональные элементы которых являются сингулярными числами матрицы A, т.е. неотрицательными квадратными корнями собственных чисел матрицы A^TA. Не умаляя общности, можно считать, что $\Sigma_{11} \geq \Sigma_{22} \geq \ldots, \Sigma_{nn}$. Естественно, что порядок расположения собственных векторов матриц AA^T и A^TA соответствует выбранному порядку расположения сингулярных чисел.

Наиболее распространенный вариант LSA основан на использовании разложения матрицы по сингулярным значениям, благодаря чему исходная матрица разлагается во множество из k ортогональных матриц, линейная комбинация которых является неплохим приближением исходной матрицы.

Доказано, что такое разложение обладает замечательной особенностью: если в Σ оставить только k наибольших сингулярных значений, а в матрицах U и V только соответствующие этим значениям колонки, то произведение получившихся матриц Σ_{lsa}, U_{lsa} и V_{lsa} будет наилучшим приближением исходной матрицы A матрицей ранга, не превышающего k: $A \cong A = U_{lsa}\Sigma_{lsa}V_{lsa}$. Здесь расстояние между матрицами A и B

задается выражением $\Sigma_{ij}(A_{ij} - B_{ij})^2$. Обозначим через U_k подматрицу матрицы U, образованную ее первыми k столбцами, через V_k — подматрицу матрицы V, образованную ее первыми k столбцами, а через Σ_k — подматрицу матрицы Σ, образованную ее первыми k строками и столбцами. Очевидно, что $A_k = U_k \Sigma_k V^T_k$. Другими словами, матрица A_k является оптимальной малоранговой аппроксимацией матрицы A. Иными словами, если в качестве A используется матрица связи термов и документов, то матрица $\hat{A}$, содержащая только k первых линейно независимых компонентов A, отражает основную структуру скрытых зависимостей, присутствующих в исходной матрице, и одновременно не содержит шума.

Выбор же наилучшей размерности k для LSA — это открытая исследовательская проблема. В идеале k должно быть достаточно велико для отображения всей реально существующей структуры данных и в то же время достаточно мало, чтобы не учитывать шума, т.е. случайных зависимостей.

Для целей поиска особое значение играют матрицы U_k и V^T_k. Строки матрицы U_k рассматриваются как образы термов в k-мерном вещественном пространстве. Аналогично столбцы матрицы V^T_k рассматриваются как образы документов в том же k-мерном вещественном пространстве. Эти векторы задают искомое представление термов и документов в k-мерном пространстве скрытых факторов.

При пополнении новым документом d информационного массива, для которого уже проведено сингулярное разложение, можно не вычислять разложение заново. Достаточно аппроксимировать его, вычисляя образ нового документа на основе ранее вычисленных образов термов и весов факторов. Пусть d — вектор весов термов нового документа (новый столбец матрицы A), тогда его образ можно вычислить по формуле: $d` = \Sigma_k^{-1} U^T_k d$.

Если q — запрос пользователя — есть вектор размерности m, i-й элемент которого равен 1, когда терм с номером i входит в запрос, и 0 — в противном случае, тогда образ запроса q в пространстве латентных факторов будет иметь вид: $q` = q^T U_k \Sigma_k^{-1}$.

Теперь мера близости запроса q и документа d оценивается величиной скалярного произведения векторов $q`$ и $V^T_k\{d\}$. Здесь $V^T_k\{d\}$ обозначает d-столбец матрицы V^T_k. Так как на практике матрица A чаще всего сильно разрежена, для эффективной работы с ней используются специальные алгоритмы.

Анализ гипертекстовых ссылок

В Internet помощь в определении авторитетности источника может оказать анализ топологии ссылок между документами. Два основанных на связях алгоритма ранжирования Web-страниц, PageRank и HITS (hyperlink induced topic search), были развиты в 1996 году в Станфордском Университете Лэрри Пейджем (Larry Page) и Сергеем Брином (Sergey Brin) [41] и в центре IBM Almaden Джоном Клейнбергом (Jon Kleinberg).

Оба алгоритма предназначены для решения "проблемы избыточности", свойственной широким запросам, а также для добавления точности результатам поиска на основе методов семантических сетей. PageRank подсчитывает общий "авторитет" документа, в то время как HITS определяет "авторитет" документа для конкретной темы.

Как одно из приложений метода лантентно-семантического анализа рассмотрим модель гипертекстовой структуры Web-пространства, которая включает критерий

ранжирования Web-страниц — PageRank. Одним из наиболее часто используемых форматов для представления документов в Internet является HTML, который позволяет создавать гипертекстовые документы, связанные гиперссылками. Сегодня множество исследований посвящены анализу структуры сети, образованной посредством гиперссылок с одних Web-страниц на другие. Например, популярная поисковая система Google обеспечивает относительно высокую точность поиска за счет использования собственного алгоритма ранжирования документов, предоставляемых пользователю в ответ на его запрос. Система Google сохраняет для всех индексируемых документов информацию о ссылках одних документов на другие и ранжирует документы в соответствии с показателем их цитирования, который грубо можно оценить числом ссылок, ведущих к данному документу из других.

***Рис. 5.3.** Лэрри Пейдж и Сергей Брин*

Один из подходов к оценке показателя цитирования документов основан на использовании сингулярного разложения матриц взаимосвязи. Для заданного массива документов все множество ссылок между ними можно представить графом G, каждая вершина которого соответствует отдельному документу, а ориентированное ребро из вершины i в вершину j свидетельствует о наличии в документе i ссылки на документ j. Граф G можно представить матрицей $A = \|a_{ij}\|$, в которой элемент a_{ij} равен 1, если из вершины i выходит ребро в вершину j, в противном случае элемент a_{ij} равен 0.

Рассмотрим матрицу $B = AA^T$. Ее элемент b_{ij} равен числу документов, которые содержат одновременно ссылки как на документ i, так и на документ j. Таким образом, матрицу B можно рассматривать как матрицу подобия авторитетных документов.

Аналогично для матрицы $C = A^TA$ ее элемент c_{ij} равен числу документов, на которые одновременно ссылаются как документ i, так и документ j. Следовательно, C можно рассматривать как матрицу подобия индексных документов.

Используя сингулярное разложение $A = USV^T$, мы можем рассматривать главное направление для матрицы B — первый столбец матрицы V — как характеристический вектор для множества авторитетных документов. Иными словами, чем больше значение величины $\|v_{1i}\|$, тем больше степень авторитетности документа i.

Аналогично первый столбец матрицы U можно рассматривать как характеристический вектор для множества индексных документов.

Одна из важнейших составляющих успеха Google — высокая пертинентность, т.е. соответствие между ожидаемым результатом поиска и фактически полученным. Это достигается за счет выделения наиболее подходящих запросу Web-страниц и их удобной группировки при выдаче. Для ранжирования результатов поиска Google использует индекс PageRank, близкий по идеологии к литературному индексу цитирования — количеству ссылок с других документов на данный документ. Индекс является реализацией метода латентно-семантического индексирования. Но индекс PageRank, в отличие от литературного индекса цитирования, не считает все упоминания и ссылки равными. Он учитывает больше факторов и определяется более сложным путем.

Обработчик ссылок системы (URLresolver) читает сформированный индексатором файл ссылок, конвертирует относительные URL в абсолютные, помещает текст ссылки в предварительный индекс и устанавливает значение параметра `docID` для того документа, на который ссылка указывает. Еще одна задача обработчика URLresolver — составление базы данных связей между документами. В дальнейшем эта база используется для определения ранга документа, т.е. значения его параметра `PageRank`.

Наиболее известным расширением индекса цитирования в WWW является PageRank, который определяет важность Web-страницы A рекурсивно на основе информации о страницах T, ссылающихся на страницу A.

Рассмотрим некую Web-страницу A. Пусть имеется n страниц $T = \{T1, T2, \ldots, Tn\}$, цитирующих данный документ, и $C(A)$ — общее число ссылок с Web-страницы A на другие документы. Пусть d (damping factor) — это вероятность того, что пользователь, просматривая какую-либо Web-страницу из множества T, перейдет на страницу A по ссылке, а не набирая ее URL по каким-либо другим причинам. Обычно значение d близко к числу 0,85. Тогда вероятность продолжения Web-серфинга без использования гиперссылок, путем ручного ввода адреса (URL) со случайной страницы, будет равна $1 - d$. Индекс PageRank — $PR(A)$ — для страницы A вычисляется по формуле: $PR(A) = (1 - d) + d(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$. Таким образом индекс легко подсчитывается простым итерационным алгоритмом.

Принцип подсчета ранга Web-страницы PageRank состоит в следующем. Рассматривается процесс, при котором пользователь сети Internet открывает случайную Web-страницу, с которой переходит по случайно выбранной гиперссылке на другую страницу. Затем, переместившись на другую Web-страницу, он снова активизирует случайную гиперссылку и так далее, постоянно переходя от страницы к странице, никогда не возвращаясь. Иногда ему такое блуждание надоедает, и он снова переходит на случайную Web-страницу — не по ссылке, а набрав вручную некоторый URL. В этом случае вероятность того, что блуждающий в Сети пользователь перейдет на некоторую определенную Web-страницу, — это ее ранг PageRank. PageRank Web-страницы тем выше, чем большее число страниц ссылается на нее и чем эти страницы популярнее.

Формально модель PageRank не учитывает динамику развития WWW и анализирует граф с некоторой статической структурой. Однако теоретический анализ показал устойчивость получаемых рангов по отношению к изменениям, касающимся ресурсов с невысоким рангом.

Немного раньше, чем использование ранга PageRank; был предложен локальный (т.е. основанный на запросе) алгоритм учета популярности — HITS, в котором учитывается запрос, позволяющий выбрать подграф из гипертекстовой сети. Из этого подграфа выделяются два вида узлов: "первоисточники" — авторитетные страницы, на которые ведут ссылки с многих других страниц, и страницы-посредники (хабы), которые содержат множество ссылок на страницы, соответствующие запросу. Алгоритм HITS заключается в выборе подмножества Web-пространства на основе запроса и определении лучших первоисточников и посредников по результатам анализа этого подмножества. Подмножество строится путем расширения множества найденных по запросу страниц за счет добавления всех страниц, связанных с ними путем, состоящим из заданного числа ссылок (на практике — одной или двух). Затем для каждого документа рекурсивно вычисляется его значимость как первоисточника a_p и посредника h_p по формулам:

$$a_p = \Sigma h_q,$$
$$h_p = \Sigma a_q.$$

Алгоритм HITS предназначен для выявления множества наиболее авторитетных страниц, определяемых главными собственными векторами X^TX и XX^T (X обозначает матрицу взаимосвязи узлов — инциденций рассматриваемого графа). При этом предполагается, что процедура формирования анализируемого множества страниц влечет доминирование страниц нужной тематики в этом множестве.

Как некоторое расширение стандартного алгоритма HITS рассматривается алгоритм Probabilistic HITS (PHITS), использующий условные вероятности $P(c|z)$ и $P(z|d)$ для описания зависимостей между наличием ссылки c, латентным (скрытым) фактором z и документом d.

Для вычисления рангов необходимо задать количество факторов z, и тогда $P(c|z)$ будет характеризовать качество страницы как "первоисточника" в контексте тематики z. В ситуациях, когда в множестве Web-страниц нет явного доминирования тематики запроса, PHITS ведет себя лучше HITS.

Несмотря на различия данных алгоритмов, общее у них то, что авторитетность (вес) узла зависит от веса других узлов, а уровень "посредника" — от того, насколько авторитетны соседние узлы. Кроме того, оба алгоритма используют вычисления собственных векторов для матриц взаимосвязи (инциденций) соответствующих Web-страниц. Расчет авторитетности отдельных документов сегодня широко используется в таких приложениях, как определение порядка сканирования документов, ранжирование результатов поиска, формирование тематических сюжетов и т.д. Формулы расчета авторитетности постоянно совершенствуются. Предполагается, что применение этих алгоритмов в будущем станет еще более эффективным, так как гиперссылки между документами постоянно оптимизируются, с одной стороны, учитывая предпочтения пользователей, а с другой стороны, явно ориентируясь на существующие методы их обработки поисковыми системами.

Вероятностное латентно-семантическое индексирование (PLSI)

Это метод выделения скрытых факторов, характеризующих значение отдельных термов и документов из заданного массива документов. В отличие от традиционного, данный метод основан на вероятностном подходе. Метод вероятностного латентно-семантического индексирования ставит своей задачей выявление латентных, скрытых факторов (тем), присутствующих в информационном массиве и связанных с его документами и словами.

Как и в предыдущем случае, рассмотрим матрицу A связи n документов $d_1, \ldots, d_n$ и m термов $t_1, \ldots, t_m$. Пусть число основных тем в документальном массиве будет k и им соответствуют k факторов $z_1, \ldots, z_k$ (зачастую k задается пользователем заранее). Сопоставим с фактором z_i вероятность $P(z_i)$ того, что случайно выбранный из данной коллекции документ точнее всего характеризуется данным фактором z_i. Итак, $\Sigma_i P(z_i) = 1$.

Обозначим через $P(d|z_i)$ вероятность того, что для заданного фактора z_i из всех документов именно документ d лучше всего характеризуется фактором z_i. Тогда $\Sigma_d P(d|z_i) = 1$. Аналогично обозначим через $P(t|z_i)$ вероятность того, что для заданного фактора z_i из всех термов именно терм t лучше всего характеризуется фактором z_i. Тогда $\Sigma_t P(t|z_i) = 1$.

Вероятность случайного выбора документа d и терма t, таких, что терм t встречается в документе d, можно оценить как $P(d|t) = \Sigma_{i=1,\ldots,k} P(z_i)P(d|z_i)P(t|z_i)$

Зафиксировав число скрытых факторов k, именно с помощью метода PLSI можно оценить следующие величины.

- $P(z_i)$ — вероятность того, что случайно выбранный из коллекции документ наиболее тесно связан с фактором (в наибольшей степени соответствует теме) z_i.
- $P(d_j|z_i)$ — вероятность того, что наиболее тесно связанный с данным фактором z_i документ — это d_j.
- $P(t_j|z_i)$ — вероятность того, что для данного фактора z_i наиболее тесно связано с ним слово — это t_j.

Наблюдаемая частота вхождения терма t в документ d задается величиной $tf(d, t)$. В соответствии с принципом максимального правдоподобия, упомянутые вероятности определяются исходя из условия максимизации функции:

$$L = \Sigma_d \Sigma_t tf(d, t)\ log\ P(d, t),$$

где внешняя сумма берется по всем документам, а внутренняя — по всем термам словаря.

Стандартной процедурой для оценки значений упомянутых вероятностей является итеративный алгоритм, на каждой итерации которого выполняется два шага — оценка и максимизация.

Данный алгоритм обеспечивает сходимость функции L к некоторому локальному максимуму. Эксперименты показывают, что сходимость достигается после нескольких десятков итераций.

Аппроксимация образа нового документа в пространстве факторов

Обозначим через $D`$ подмножество документов из информационного массива $D$, в котором производится поиск, а через $W`$ — множество всех термов в документах, вошедших в $D`$. Предположим также, что подмножество $D`$ может рассматриваться в качестве представительной выборки документов из коллекции D, где представлены все темы, отраженные в полном массиве. Иными словами, произвольный документ d из D содержит значительное число термов из $W`$. Пусть $n(d, w)$ — число вхождений терма w в документ d.

Рассмотрим систему линейных алгебраических уравнений:

$$P(\mathrm{d}, \mathrm{w}) = \Sigma_{z \in Z} P(\mathrm{z}) P(\mathrm{w}|\mathrm{z}) P(\mathrm{d}|\mathrm{z}),$$

где $w \in W`$, $n(d, w) > 0$. В качестве неизвестных рассматриваются величины $P(d|z)$, $z \in Z$. При этом значения величин $P(z)$, $P(w|z)$, $z \in Z$, $w \in W`$ получены в результате применения PLSI к множеству документов $D`$. В результате $P(d, w)$ аппроксимируется по формуле $P(d, w) \approx (1/|D`|)(n(d, w)/\mathrm{length}(d))$, где $\mathrm{length}(d)$ — это количество слов из $W`$, имеющихся в документе d. Полученное значение является наилучшим приближением по методу наименьших квадратов образа документа d в пространстве факторов Z.

Расширение запроса пользователя

В результате выявления тематической принадлежности документов, отмеченных пользователем в результате первичной процедуры поиска как релевантные, возможно расширение запроса словами из отмеченных документов. В случае, когда информационный массив, в котором выполняется поиск, содержит небольшое число релевантных запросу документов, расширение запроса, основанное на использовании обратной связи, эффективнее традиционного матричного контекстного анализа. Расширение запроса пользователя на заданное число (k) слов на основе отмеченных пользователем документов (множество S) происходит следующим образом.

1. Для всех слов t из словаря, которые встречаются в документах из S, вычисляется их вес $\mathrm{weight}(t) = \Sigma_{d \in S, z \in Z} P(z) P(d|z) P(t|z)$.
2. Множество слов t упорядочивается по убыванию весов $\mathrm{weight}(t)$.
3. Из построенного списка выбираются первые k слов.

При этом, если документ d входит в множество $D`$, значения величин $P(z)$, $P(d|z)$ и $P(w|z)$ уже известны (вычислены при применении PLSI к множеству документов $D`$). В противном случае неизвестные величины $P(d|z)$, $z \in Z$ оцениваются с помощью соответствующего алгоритма.

Метод суффиксных деревьев

Изначально метод суффиксных деревьев (Suffix Tree Clustering) был разработан для быстрого поиска подстрок в строках. Суффиксное дерево — это дерево, содержащее все суффиксы строки. Оно состоит из вершин, ветвей и суффиксных указателей, с помощью которых добиваются высокой (линейной) скорости построения дерева. Ветви дерева обозначаются отдельными буквами или частями суффиксов строки. Суффикс, соответствующий определенной вершине, можно получить путем объединения букв, которые находятся на ветвях, начиная от корневой вершины и заканчивая данной. Сегодня идеология суффиксных деревьев применяется для кластеризации результатов работы информационно-поисковой системы. К достоинствам этого метода можно отнести высокую скорость работы ($O(n)$), наглядность представления результатов, а также вычислительную простоту.

При построении дерева вначале подвергаются отчистке от пунктуации документы, получаемые от поисковой системы, затем осуществляется приведение слов к каноническим формам (лемматизация) и т.д. После этого для найденных

документов строится дерево, но в этом случае ветвям приписываются термы (слова или словосочетания), а не буквы, как в традиционном методе. В результате вершинам дерева соответствуют фразы, которые можно получить, объединив все термы, находящиеся на ветвях, ведущих от корня к данной вершине дерева. В вершинах дерева, имеющих потомков, расположены ссылки на документы, в которых встречается фраза, соответствующая вершине. Множества документов, на которые указывают эти ссылки, образуют базовые кластеры. Затем происходит укрупнение базовых кластеров и получение окончательного набора кластеров.

Кластеры укрупняются по следующему алгоритму: пусть B_m и B_n — базовые кластеры, $|B_m|$, $|B_n|$ — их размеры, а $|B_n \cap B_m|$ — количество общих документов для этих кластеров. Тогда, если $|B_n \cap B_m| / |B_m|$ и $|B_n \cap B_m| / |B_n|$ превышают определенный порог, например 0,5, базовые кластеры объединяются в один общий кластер.

K-means

В основе метода K-means лежит итеративный процесс стабилизации центроидов кластеров. Основная идея метода заключается в итеративном достижении изменений центроидов кластеров, после чего процесс кластеризации считается завершенным. Теоретическая скорость работы алгоритма линейна, т.е. составляет $O(n)$, где n — число документов в информационном массиве.

В начале применения метода априорно выбираются начальные центроиды для множества документов. Например, из множества документов случайным образом выбираются k документов, где k равно требуемому числу кластеров. (В этом методе необходимо явно указывать требуемое число кластеров.) Начальные кластеры можно выбрать и на основе байесового оценивания и нахождения подходящего для данного информационного массива числа кластеров и их центроидов.

После этого все документы распределяются по кластерам, причем каждый документ может попасть только в один кластер, центроид которого наиболее близок к данному документу. Затем центроиды кластеров пересчитываются, и если они не изменились, т.е. стабилизировались, то процесс кластеризации завершается.

Метод "папок поиска"

Метод "папок поиска" (Custom Search Folders) не связывается с определенным алгоритмом кластеризации, а представляет собой множество подходов, общее у которых — попытка кластеризовать результаты поиска и представить на Web-сайте кластеры в удобном для пользователей виде.

Суть этого метода (скорее технологии) заключается в том, что пользователь может сузить результат поиска посредством того, что будет рассматривать объекты, распределенные по папкам-кластерам, автоматически формируемым в результате поиска. Достигается это за счет лексического анализа результатов поиска и запросов. Такой подход позволяет преобразовать страницы результатов поиска в интуитивно понятную древовидную структуру папок, т.е. пользователь после проведения первичного поиска по своему запросу может выбрать одну из предложенных папок, тем самым сузив область поиска. Папки чаще всего имеют иерархическую структуру, что дает возможность еще больше конкретизировать результаты поиска. Распределение по папкам происходит в режиме реального времени по ходу предоставления пользователю результатов поиска. По сути, папки выступают центроидами кластеров, с которыми затем соотносятся документы. Очевидно, для оперативного распределения документов по папкам,

заранее должна быть построена матрица близости документов (типа $tf \times idf$), расчет которой обычно требует существенного времени. В результате технология обладает высокой скоростью работы и большой наглядностью. Метод "папок поиска" в настоящее время нашел широкое применение и реализован на сотнях Web-сайтов, представленных в Internet.

Одна из первых удачных реализаций была представлена на сервере Vivisimo (`http://www.vivisimo.com`). Подход Vivisimo предполагает анализ текста, в частности статей новостных ресурсов, позволяющий выделить ключевые слова и фразы. При этом предполагается, что читатель ищет статьи на определенную тему. "Это правильное решение — предпринять шаги в направлении дальнейшей персонализации, пойти еще дальше, чем Google", — считает Рауль Вальдес-Перес, президент компании Vivisimo, специализирующейся на кластерных технологиях. При этом Вальдес-Перес подчеркнул, что интересы читателей новостных сайтов куда более изменчивы, чем предпочтения посетителей книжных магазинов. Компания Vivisimo на своем сайте представила поисковую систему, в которой обобщаются ("кластеризуются") ссылки на статьи в соответствии с их темами. По мнению главы Vivisimo, такая стратегия предоставляет посетителю значительную свободу: "Нет лучшего устройства для персонализации, чем сама персона", — замечает он.

В частности, система Vivisimo состоит из трех модулей, первый из которых, Knowledge Writer (Фиксатор знаний), поддерживает базу синонимов, акронимов и различных вариантов лексических единиц. Основная задача этого интеллектуального модуля — "подстроиться" под имеющиеся данные для корректной разбивки и сортировки по категориям. Второй модуль, Web-Based Administration (Администратор ресурсов Web), является интерфейсом настройки системы и управления ею, а третий, Organized Content from Multiple Sources (Обработчик упорядоченного содержимого множественных источников), позволяет проводить поиск по нескольким ресурсам одновременно.

Свое применение Vivisimo уже нашла в корпоративных сетях и Web-сервисах. У Vivisimo имеются достаточно мощные аналоги, один из которых — система графического представления результатов поиска Grokker. В отличие от Vivisimo, Grokker является не автономной поисковой системой, а модулем для поискового брэнда Google.

Приведем еще несколько примеров. Австралийский поисковый сервер Mooter (`http://www.moother.com`) избрал собственный визуальный подход к предоставлению результатов поиска по обрабатываемым запросам. Вместо стандартных "плоских" результатов в виде списка, Mooter группирует результаты поиска по категориям. Например, при вводе словосочетания "Semantic Web" (семантический Web) пользователю будут представлены группы категорий, относящихся к этому понятию. В данном случае это XML, Internet, World Wide Web, conference и т.д. Если эти результаты пользователя не устроят, то он может просто воспользоваться ссылкой "следующие кластеры" (next clusters), как показано на рис. 5.4.

Поисковый сервер iBoogie (`http://www.iboogie.com`) тоже группирует результаты поиска, но отображает их иначе — в виде, близком к используемому проводником Windows: справа представлены списки найденных документов, а слева категории (кластеры) для просмотра.

Как и другие современные поисковые системы, сервер iBoogie предоставляет возможность выбора большого количества критериев поиска: MP3/аудио, изображения, видео, поиск в директориях (рис. 5.5).

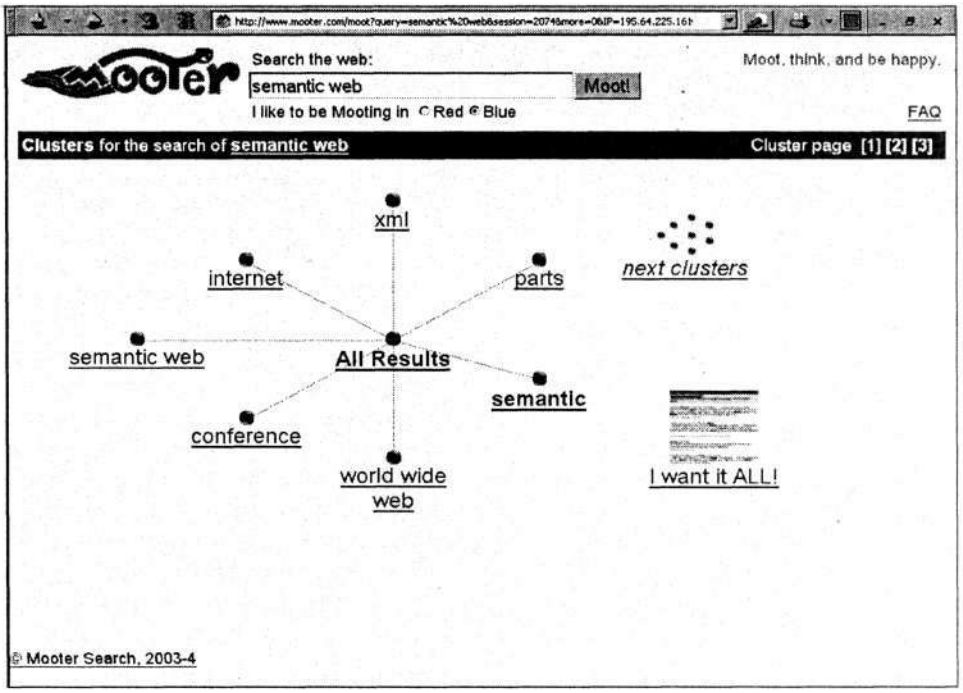


Рис. 5.4. Кластеры на сервере Mooter

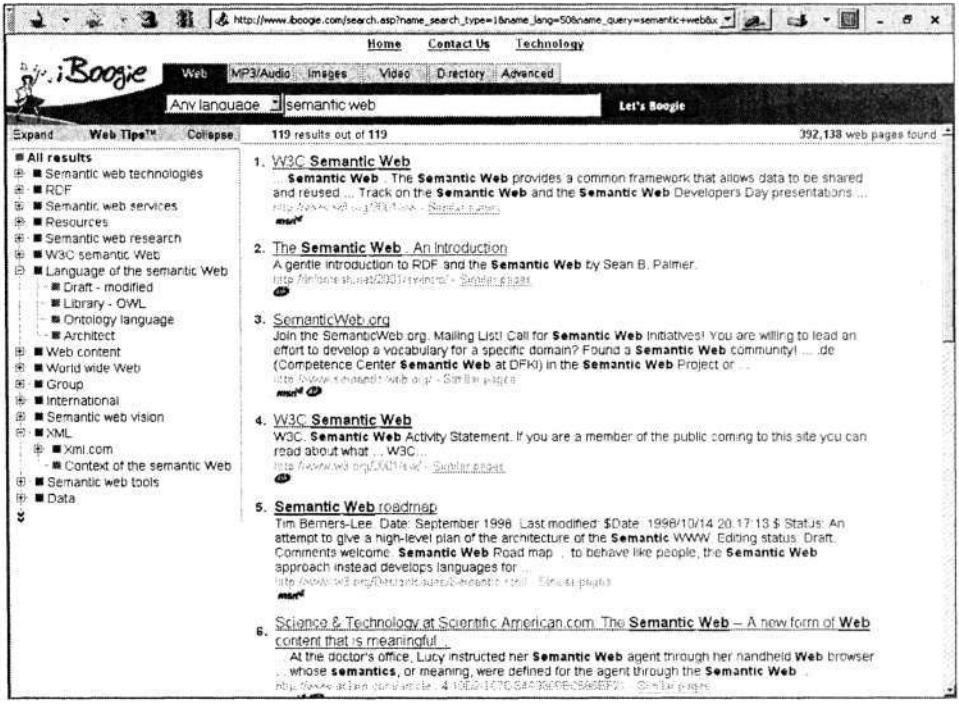


Рис. 5.5. Интерфейс поисковой системы iBoogie

Поисковая система WiseNut (`http://www.wisenut.com`), разработанная компанией LookSmart, обеспечивает группировку результатов поиска по различным категориям, которые отображаются под строкой запроса. Например, по запросу "Text Mining" система определяет такие релевантные категории: Workshop Text, Knowledge Management, Document Warehousing и др. (рис. 5.6).

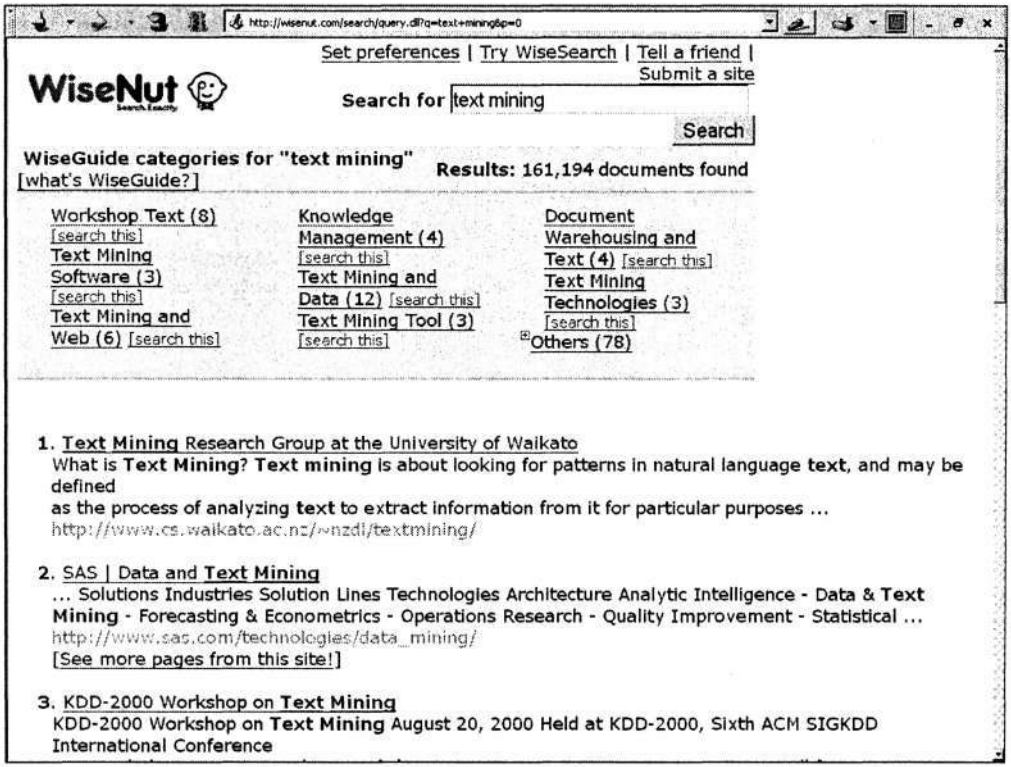


Рис. 5.6. Результаты поиска в WiseNut

В то же время уже на первой странице отклика ИПС будут также представлены результаты поиска в виде традиционного списка. Пользователь может для каждого найденного сайта просмотреть его описание, за которым следует адрес и гиперссылка для быстрого просмотра. Перейдя по этой ссылке, пользователь прямо под результатом увидит экранную копию сайта, с которого легко можно перейти к оригиналу.

5.5. Автоматические ответы на вопросы

Системы автоматических ответов на вопросы пользователей, задаваемых на естественном языке, задумывались еще на заре кибернетики. Некоторые практические наработки в этой области были получены в эпоху расцвета идеологии экспертных систем (80–90-е годы XX века). Однако, несмотря на большой спрос, технологические проблемы не позволили создать промышленные системы этого класса. В последние годы в связи с развитием технологии Text Mining о таких системах говорят все чаще, прежде всего в связи с возможностью их практической

реализации. Нередко в качестве базы знаний в этих системах предполагается использовать ресурсы Internet, обработанные современными средствами глубинного анализа текстов. Кстати, с самого начала развития технологий Text Mining служба получения ответов на вопросы (Question Answering) рассматривалась как их органическая составляющая.

По мнению многих экспертов, ожидается настоящая революция в области поиска в Internet. И эта революция практически заявит о себе, когда в Сети появятся системы, которые смогут давать прямые и четкие ответы на произвольные вопросы пользователей. Сегодня корпорация Microsoft уже пытается создать первую реально работающую систему, способную отвечать на вопросы пользователей. Работы в этом направлении ведутся в исследовательском центре корпорации (Microsoft Research) доктором Эриком Бриллом (Eric Brill), недавно опубликовавшим совместно с Руди Сорикутом (Radu Soricut) алгоритм работы такой системы в статье "Автоматические ответы на вопросы: по ту сторону от фактоидов" (`http://research.microsoft.com/tmsn/Papers/camera-ready-QA.doc`).

В соответствии с этим алгоритмом, вопрос пользователя поступает модулю Question2Query, переводящему его в запрос на информационно-поисковом языке. При этом на основе статистических подходов из строки, т.е. вопроса пользователя, выделяются и нормируются ключевые слова, которые затем и становятся основой запроса.

Например, из вопроса

```
How do herbal medications differ from conventional drugs?
(Чем медикаменты растительного происхождения отличаются от обычных
препаратов)
```

система выделяет ключевые слова

```
"How do" "herbal medications" "differ from" "conventional" "drugs",
```

после чего обращается к традиционной поисковой системе (модуль Search Engine) с запросом: `"differ from" & "herbal medications"`.

После получения откликов от традиционных поисковых систем первые *N* документов (наиболее релевантные, ранжированные) обрабатываются модулем фильтрации, который выполняет дополнительный поиск и выделяет наиболее релевантные фрагменты из этих документов. Результаты фильтрации поступают на модуль AnswerExtraction, который по весовому алгоритму выбирает необходимое для ответа слово или предложение (рис. 5.7).

В настоящее время разработана уже первая версия системы, получившая название "Ask MSR", которая способна не только проводить поиск в Сети, но и извлекать из найденных Web-страниц полезную информацию, текст с фактами, которые используются для ответа на вопрос пользователя. При этом ответ системы представляет собой одно слово или предложение. Например, если задать системе вопрос: "Когда родилась Мерлин Монро?", то алгоритмы сначала проанализируют структуру вопроса, определят объект поиска, преобразуют вопрос в поисковый запрос, отправят его на обычную ИПС (MSNSearch или Google), получат результаты, а потом интеллектуально отфильтруют найденные страницы и выдадут требуемый ответ. В настоящее время система Ask MSR является всего лишь моделью, однако уже имеются планы по выводу ее на рынок под названием AnswerBot.

Параллельно группа исследователей под руководством доктора Брилла работает над развитием алгоритмов системы, дополняя их элементами искусственного интеллекта. Модель системы "Ask MSR" для создания собственной базы знаний

проанализировала свыше миллиарда Web-страниц, выбрав 2,3 млн адресов часто задаваемых вопросов (FAQ). В соответствии с алгоритмом работы системы, ее база знаний аккумулируется в модуле Training Corpus. В результате система уже сейчас способна моделировать ответ, который выдается пользователю на его вопрос. Существующая модель Ask MSR пока обеспечивает корректные ответы только на 40% вопросов, что, тем не менее, признается сегодня большим успехом.

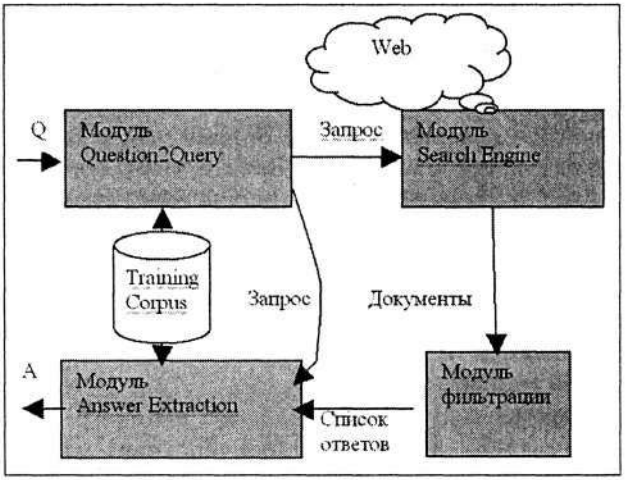


Рис. 5.7. Алгоритм доктора Брилла

5.6. Реализация систем Text Mining

В настоящее время многие ведущие производители программного обеспечения предлагают свои продукты и решения в области Text Mining. Как правило, это масштабируемые системы, в которых реализованы различные математические и лингвистические алгоритмы анализа текстовых данных. Они имеют развитые графические интерфейсы, богатые возможности визуализации и манипулирования данными, предоставляют доступ к различным источникам данных и функционируют в архитектуре клиент/сервер. Вот несколько примеров таких систем.

- Intelligent Miner for Text (IBM)
- PolyAnalyst, WebAnalyst (Мегапьютер Интеллидженс)
- Text Miner (SAS)
- SemioMap (Semio Corp.)
- Oracle Text (Oracle)
- Knowledge Server (Autonomy)
- Galaktika-ZOOM (корпорация Галактика)
- InfoStream (ИЦ "ЭЛВИСТИ")

Ниже мы рассмотрим эти системы более подробно.

5.6.1. Intelligent Miner for Text

Этот продукт фирмы IBM (`http://www-3.ibm.com/software/data/iminer/fortext`) представляет собой набор отдельных утилит, запускаемых из командной строки, или скриптов, выполняемых независимо друг от друга. Эта система является одним из лучших инструментов глубинного анализа текстов. Она содержит следующие основные утилиты (Tools) для построения приложений управления знаниями.

- Language Identification Tool — утилита, предназначенная для автоматического определения языка, на котором составлен документ.
- Categorisation Tool — утилита классификации, предназначенная для автоматического отнесения текста к некоторой категории (входной информацией на обучающей фазе работы этого инструмента может служить результат работы следующей утилиты — Clusterisation Tool).
- Clusterisation Tool — утилита кластеризации, предназначенная для разбиения большого множества документов на группы по стилю, форме, различных частотных характеристиках выявляемых ключевых слов (рис. 5.8).
- Feature Extraction Tool — утилита определения нового, предназначенная для выявления в документе новых ключевых слов (собственные имена, названия, сокращения) на основе анализа заданного заранее словаря.
- Annotation Tool — утилита "выявления смысла" текстов и составления рефератов, предназначенная для формирования аннотаций к исходным текстам.

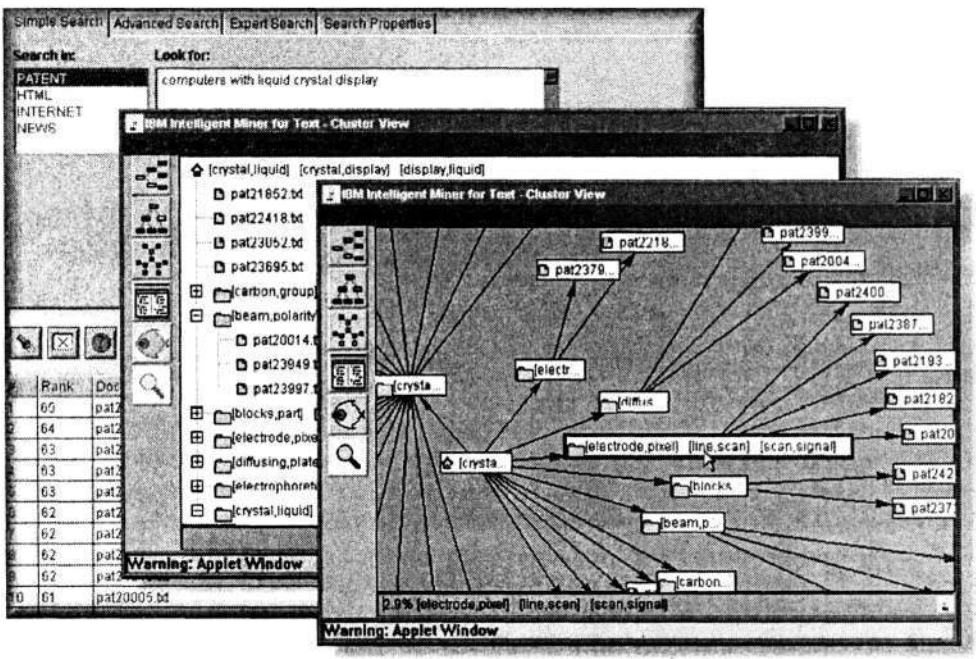


Рис. 5.8. Визуализация кластеров в IBM Intelligent Miner for Text

Пакет IBM Intelligent Miner for Text объединяет мощную совокупность инструментов, базирующихся, в основном, на механизмах поиска информации (information retrieval), что является спецификой всего продукта. Система включает ряд базовых компонентов, которые имеют самостоятельное значение вне пределов технологии "добычи текстов" — это информационно-поисковая система Text Search Engine, утилита сканирования Web-пространства Web crawler, Net Question Solution (решение для поиска в локальном Web-сайте или на нескольких intranet/Internet-серверах, Java Sample GUI), набор интерфейсов Java Beans для администрирования и организации поиска на основе Text Search Engine.

Intelligent Miner for Text как продукт IBM включен в комплекс "Information Integrator for Content" для СУБД DB2 в качестве средства Information Mining ("глубинного анализа информации"). Стоимость продуктов разных уровней семейства Intelligent Miner составляет от 18 до 75 тыс. долларов.

5.6.2. PolyAnalyst

Решение PolyAnalyst российской компании "Мегапьютор" (`http://www.megaputer.com`) может применяться для автоматизированного анализа числовых и текстовых баз данных с целью обнаружения ранее неизвестных, нетривиальных, полезных и доступных пониманию закономерностей (рис. 5.9).

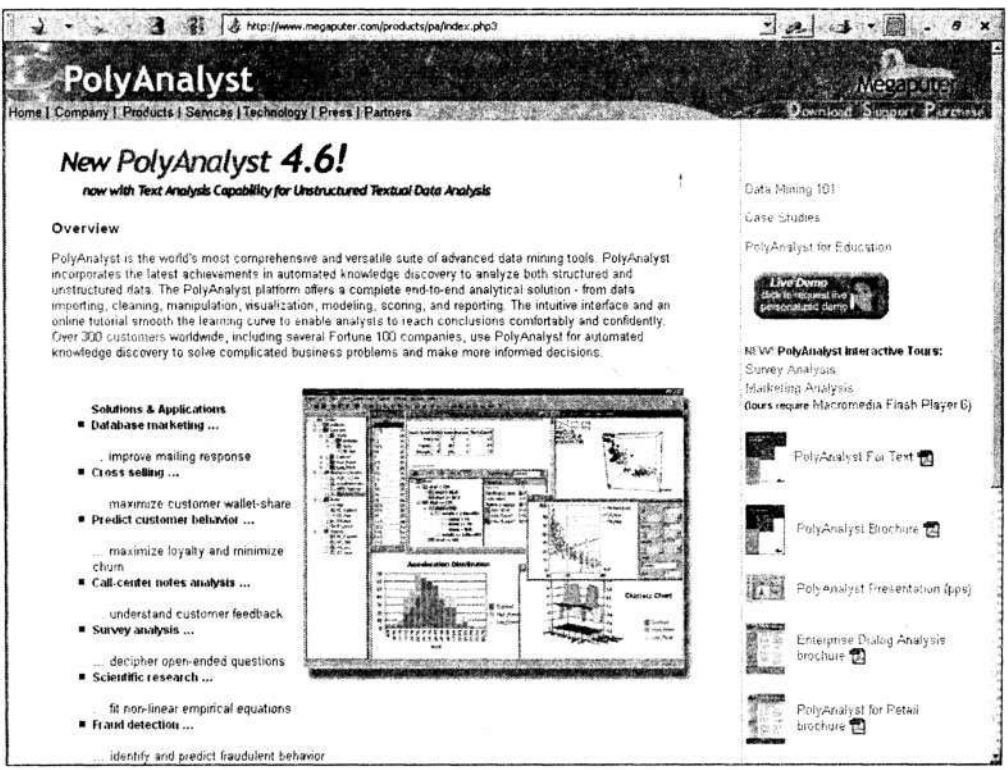


Рис. 5.9. Новая версия системы PolyAnalyst

По своей природе PolyAnalyst является клиент-серверным приложением. При этом пользователь работает с программой PolyAnalyst Workplace. Математические же модули выделены в серверную часть — PolyAnalyst Knowledge Server. Такая архитектура предоставляет естественную возможность для масштабирования системы — от однопользовательского варианта до корпоративного решения с несколькими серверами.

PolyAnalyst работает с разными типами данных. Это — числа, логические переменные, текстовые строки, даты, а также свободный текст. PolyAnalyst может обрабатывать исходные данные из различных источников, — например, файлы Microsoft Excel 97/2000, файлы ODBC-совместимых СУБД, файлы данных системы SAS, файлы Oracle Express и IBM Visual Warehouse.

В состав PolyAnalyst входит система TextAnalyst (`http://www.megaputer.com/products/ta/index.php3`), которая решает следующие задачи Text Mining: создание семантической сети большого текста, подготовка резюме текста, поиск по тексту и автоматическая классификация и кластеризация текстов. Построение семантической сети — это поиск ключевых понятий текста и установление взаимоотношений между ними. По такой сети можно не только понять, о чем говорится в тексте, но и осуществить контекстную навигацию. Подготовка резюме — это выделение в тексте предложений, в которых чаще других встречаются значимые для этого текста слова. В 80% случаев этого вполне достаточно для получения представления о тексте. Для поиска информации в системе предусмотрено использование запросов на естественном языке. По запросу строится уникальная семантическая сеть, которая при взаимодействии с сетью документа позволяет выделить нужные фрагменты текста. Кластеризация и классификация проводятся стандартными методами добычи данных.

Система TextAnalyst рассматривает Text Mining в качестве отдельного математического аппарата, который разработчики программного обеспечения могут встраивать в свои продукты, не опираясь на платформы информационно-поисковых систем или СУБД. Основная платформа для применения системы — MS Windows 9x/2000/NT. Существует плагин TextAnalyst для браузера Microsoft Internet Explorer.

Благодаря технологии эволюционного программирования и другим интеллектуальным алгоритмам, PolyAnalyst с успехом применяется в различных бизнес-задачах, в социологических исследованиях, в прикладных научных и инженерных задачах, в банковском деле, в страховании и медицине.

WebAnalyst

Система WebAnalyst (`http://www.megaputer.com/products/wa/index.php3`) — также продукт "Мегапьютер Интеллидженс" — представляет собой интеллектуальное масштабируемое клиент-серверное решение для компаний, желающих максимизировать эффект анализа данных в Web-среде. Сервер WebAnalyst функционирует как экспертная система сбора информации и управления контентом Web-сайта. Модули WebAnalyst решают три задачи: сбор максимального количества информации о посетителях сайта и запрашиваемых ими ресурсах; исследование собранных данных; и генерация персонализированного, на основе результатов исследований, контента. Решение этих задач в совокупности должно, по мнению разработчиков системы, привести к максимизации количества новых посетителей Web-сайта и сохранению уже имеющихся, а следовательно,

к увеличению популярности ресурса. Помимо этого, WebAnalyst способен интегрировать возможности Text Mining напрямую в Web-сайт организации. Это позволяет организовать индивидуализированный, автоматизированный и целевой маркетинг, автоматический поиск и реализацию перекрестных продаж, а также расширить набор данных, настраиваемых пользователем. По сути, WebAnalyst представляет собой интеллектуальный сервер приложений электронной коммерции. Техническая платформа та же, что и у PolyAnalyst.

5.6.3. Text Miner

Американская компания SAS Institute выпустила систему Text Miner для сравнения определенных грамматических и словесных рядов в письменной речи (`http://www.sas.com/technologies/analytics/datamining/textminer`). Система Text Miner весьма универсальна, поскольку может работать с текстовыми документами различных форматов — в базах данных, файловых системах и даже в Web. Text Miner обеспечивает логическую обработку текста в среде мощного пакета SAS Enterprise Miner. Это позволяет пользователям обогащать процесс анализа данных, интегрируя неструктурированную текстовую информацию с существующими структурированными данными, такими как возраст, доход и характер покупательского спроса.

Пример успешного применения логических возможностей Text Miner демонстрирует компания Compaq Computer Corp., которая в настоящее время тестирует Text Miner, анализируя более 2,5 гигабайт текстовых документов, полученных по e-mail и собранных представителями компании. Ранее обработать такие данные было практически невозможно.

Программа Text Miner позволяет определить, насколько правдив тот или иной текстовый документ. Обнаружение лжи в документах производится путем анализа текста и выявления изменений стиля письма, которые могут возникать при попытке исказить или скрыть информацию. Для поиска таких изменений используется принцип, заключающийся в поиске аномалий и трендов среди записей баз данных без выяснения их смысла. При этом в Text Miner включен обширный набор документов с различной степенью правдивости, структура которых принимается в качестве шаблонов. Каждый документ, "прогоняемый" на детекторе лжи, анализируется и сравнивается с этими эталонами, после чего программа присваивает документу тот или иной индекс правдивости. Особенно полезной программа может стать в организациях, получающих большой объем электронной корреспонденции, а также в правоохранительных органах для анализа показаний наравне с детекторами лжи, действие которых основано на наблюдении за эмоциональным состоянием человека.

Интересен пример применения Text Miner в медицине: в одной из американских национальных здравоохранительных организаций было собрано свыше 10 тыс. врачебных записей о заболеваниях сердца, собранных из клиник по всей стране. Анализируя эти данные с помощью Text Miner, специалисты обнаружили некоторые административные нарушения в отчетности, а также смогли определить взаимосвязь между сердечно-сосудистыми заболеваниями и другими недугами, которые не были определены традиционными методами. Вместе с тем, компания SAS отмечала, что выпускает свой продукт Text Miner, в основном, для привлечения внимания бизнес-интеллигенции.

5.6.4. SemioMap

Это продукт компании Entrieva, созданный в 1996 году ученым-семиотиком Клодом Фогелем (Claude Vogel). В мае 1998 года SemioMap был выпущен как промышленный комплекс SemioMap 2.0 — первая система Text Mining, работающая в архитектуре клиент/сервер (`http://www.entrieva.com/entrieva/products/semiomap.asp?Hdr=semiomap`). Система SemioMap состоит из двух основных компонентов — сервера SemioMap и клиента SemioMap. Работа системы протекает в три этапа.

1. *Индексирование.* Сервер SemioMap автоматически читает массивы неструктурированного текста, извлекает ключевые фразы (понятия) и создает из них индекс.
2. *Кластеризация понятий.* Сервер SemioMap выявляет связи между извлеченными фразами и строит из них лексическую сеть ("понятийную карту") на основе данных о совместном их использовании.
3. *Графическое отображение и навигация.* Визуализация карт связей, обеспечивающих быструю навигацию по ключевым фразам и связям между ними, а также возможность быстрого обращения к конкретным документам (рис. 5.10).

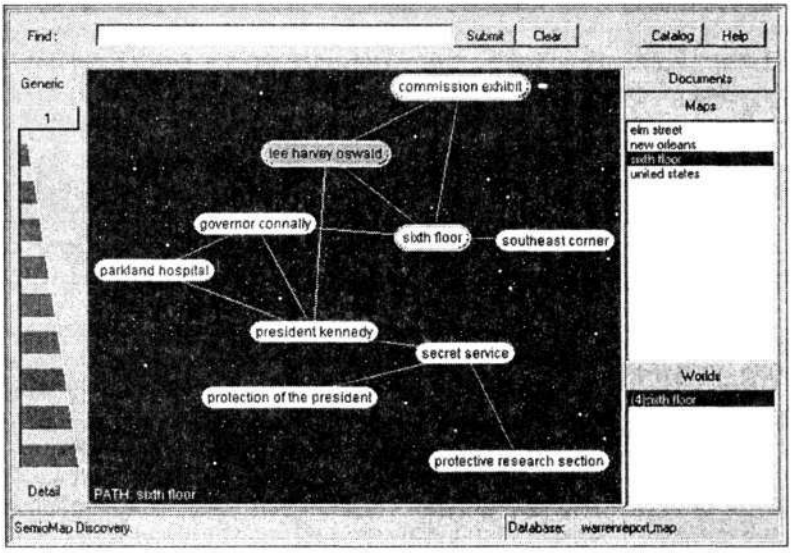


Рис. 5.10. Семантическая карта SemioMap

SemioMap поддерживает разбиение материала по "папкам" и создание отдельной базы данных для каждой папки. Связи между понятиями, которые выявляет SemioMap, базируются на совместной встречаемости фраз в абзацах исходного текстового массива.

Центральным блоком SemioMap является *лексический экстрактор* — программа, которая извлекает фразы из текстовой совокупности и выявляет совместную встречаемость этих фраз (их взаимные связи). Лексический экстрактор базируется на патентованной технологии SEMIOLEX. Она реализует идеи вычислительной семиотики, науки о знаках в языковой коммуникации, разработанной Клодом Фогелем.

5.6.5. InterMedia Text, Oracle Text

Начиная с Text Server в составе СУБД Oracle 7.3.3 и картриджа interMedia Text в Oracle8i, средства Text Mining являются неотъемлемой частью продуктов Oracle. В Oracle9i эти средства были существенно развиты и получили новое название — Oracle Text [70] (`http://technet.oracle.com/products/text/content.html`). Это программный комплекс, интегрированный в СУБД и позволяющий эффективно работать с запросами, относящимися к неструктурированным текстам. При этом обработка текста сочетается с возможностями, которые предоставлены пользователю для работы с реляционными базами данных. В частности, при написании приложений для обработки текста стало возможным использовать язык SQL.

Основной задачей, на решение которой нацелены средства Oracle Text, является поиск документов по их содержанию — словам или фразам, которые при необходимости комбинируются с использованием булевых операций. Результаты поиска ранжируются по релевантности, с учетом частоты использования слов запроса в найденных документах. Для повышения полноты поиска Oracle Text предоставляет ряд средств расширения поискового запроса, среди которых можно выделить следующие: расширение слов запроса всеми морфологическими формами, что реализуется привлечением знаний о морфологии языка; расширение слов запроса близкими по смыслу словами за счет подключения тезауруса — семантического словаря; а также расширение запроса словами, близкими по написанию и звучанию — нечеткий поиск и поиск созвучных слов. Нечеткий поиск целесообразно применять при поиске слов с опечатками, а также в тех случаях, когда возникают сомнения в правильном написании фамилии, названия организации и т.п.

Система Oracle Text обеспечивает тематический анализ текстов на английском языке. В ходе обработки текст каждого документа подвергается процедурам лингвистического и статистического анализа, в результате чего определяются его ключевые темы и создаются тематические резюме, а также общее резюме-реферат.

Все описанные средства могут использоваться совместно, что поддерживается языком запросов в сочетании с традиционным синтаксисом языка PL/SQL для поиска документов. Oracle Text предоставляет возможность работать с современными реляционными СУБД в контексте сложного многоцелевого поиска и анализа текстовых данных.

5.6.6. Autonomy IDOL Server

Архитектура IDOL (Intelligent Data Operating Layer) сервера компании Autonomy (`http://www.autonomy.com`), известной своими разработками в области статистического контент-анализа, объединяет интеллектуальный парсинг по шаблонам со сложными методами контекстного анализа и извлечения смысла для решения задач автоматической классификации и организации перекрестных ссылок. Основное преимущество системы Autonomy — мощные интеллектуальные алгоритмы, основанные на статистической обработке. Эти алгоритмы базируются на информационной теории Клода Шеннона, байесовых вероятностях и нейронных сетях [55]. Концепция адаптивного вероятностного моделирования (APCM) позволяет системе Autonomy идентифицировать шаблоны в тексте документа и автоматически определять подобные шаблоны во множестве других документов.

Важный момент в системе Autonomy IDOL Server — это возможность анализа текстов и идентификации ключевых концепций в пределах документов путем

анализа корреляции частот и отношений терминов со смыслом текста (рис. 5.11). Система Autonomy использует уникальную технологию анализа шаблонов (нелинейная адаптивная цифровая обработка сигнала) для извлечения из документов смысла и определения характеристик, содержащихся в текстах. АРСМ позволяет идентифицировать уникальные "сигнатуры" смысла текста, а также создавать агенты концепций, с помощью которых ищутся подобные по смыслу записи на Web-сайтах, в новостях, архивах электронной почты и в других документах. Поскольку система не базируется на предопределенных ключевых словах, она может работать с любыми языками.

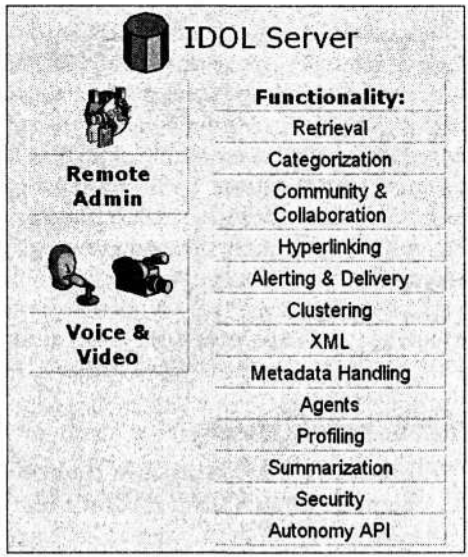


Рис. 5.11. Архитектура Autonomy IDOL Server

Функциональность системы Autonomy включает такие основные возможности:

- автоматическая классификация;
- кластеризация;
- автореферирование;
- автоматическое проставление гиперссылок;
- автоматическое создание профилей (информационных портретов);
- генерация таксонометрических деревьев;
- создание метаданных и манипулирование ими;
- интеллектуальная обработка XML-данных;
- персонализация;
- поиск.

Ядро системы агентов Autonomy — это механизм динамического рассуждения (DRE), основанный на технологии обработки шаблонов, в которой используются методы нейронных сетей. В DRE используется концепция адаптивного вероятностного моделирования для реализации четырех главных функций: выявление концепции, создание агента, обучение агента и стандартный поиск текста. DRE воспринимает запросы на естественном языке или термины, связанные булевыми операторами, и возвращает список документов, упорядоченных по релевантности запросу. Этот механизм является основой для всех продуктов системы агентов от Autonomy. Описание сервера IDOL компании Autonomy приведено по адресу `http://www.autonomy.com/content/Products/IDOL`.

5.6.7. Galaktika-ZOOM

Система Galaktika-ZOOM — продукт российской корпорации "Галактика" (`http://zoom.galaktika.ru/content.htm`). Основное назначение системы — интеллектуальный поиск по ключевым словам с учетом морфологии русского

и английского языков, а также формирование информационных массивов по конкретным аспектам. При этом объемы информации могут достигать сотен гигабайт. Именно ориентация на большие информационные объекты — сообщения и статьи СМИ, отраслевую печать, нормативную документацию, деловую переписку и материалы внутреннего документооборота предприятия, информацию из Internet — и составляет главную особенность продукта. При этом система предоставляет определенный инструментарий для анализа объективных смысловых связей отобранных данных и формирования "образа" проблемы — многомерной модели в информационном потоке в форме ранжированного списка значимых слов, употребляемых совместно с темой проблемы. Большое внимание в системе уделено выявлению тенденций динамики развития изучаемой проблемы. Система содержит конверторы чаще всего встречающихся текстовых форматов: простой текст, RTF, DOC, HTML. Система Galaktika-ZOOM функционирует в среде ОС Windows 2000.

5.6.8. InfoStream

Технология InfoStream® (`http://infostream.ua`) была создана для охвата и обобщения больших динамических информационных массивов, непрерывно генерируемых в Сети. Методы, применяемые в системе, тесно связаны с методологией контент-анализа, проводимого непрерывно во времени.

Технология InfoStream ориентирована, прежде всего, на работу с Web-документами, однако в корпоративной реализации она позволяет обрабатывать данные в форматах офисных систем — MS Word (DOC, RTF), PDF — и других текстовых форматах (простой текст, XML и пр.). Системы на основе технологии InfoStream в настоящее время функционируют на платформах таких операционных систем, как FreeBDS, Linux, Solaris, Microsoft .NET.

5.7. Text Mining не только для спецслужб

Весной 2001 года ЦРУ представило широкой публике свои технологии "добычи данных", используемые для поиска информации в публикуемых текстах, радио- и телепередачах. Отдел современных информационных технологий, входящий в состав управления науки и техники Центрального разведывательного управления США, продемонстрировал общественности технологии "извлечения текстовых данных" ("Text- и Data Mining"), используемые для поиска значимой информации в огромной массе документов, а также в радио- и телепередачах на различных языках. Поиск ведется как по систематизированным, так и по случайным источникам, причем объектами поиска являются тексты в печатных изданиях и в цифровом виде, графические изображения, аудиоинформация на 35 языках. Для отсеивания аудиоинформации используется методика Oasis, которая распознает речь и превращает ее в текст. При этом технология позволяет отделять мужские голоса от женских, а также голоса, принадлежащие разным людям, и записывать их в виде диалогов. Однако методика Oasis позволяет выделять из аудиопотока только те голоса или ту конкретную информацию, которая заложена в настройках поиска.

Еще одна компьютерная технология под названием Fluent позволяет подразделениям ЦРУ искать информацию в текстовых документах. Эта технология подразумевает поиск по ключевым словам, причем вводится слово или сочетание на английском языке, которое тут же переводится на ряд других

языков, и найденная информация из базы данных на разных языках поступает исследователю после автоматического перевода. Такая программа, как Text- и Data Mining, позволяет автоматически создавать предметные указатели для текстовых документов, а также получать данные по частоте употребления тех или иных слов в документах. Эти технологии ЦРУ использует сегодня для отслеживания незаконных финансовых операций и наркотрафика.

Названными выше технологиями занимается отдел Advanced Information Technology (Директората науки и технологии ЦРУ). "Мы развиваемся не так быстро, чтобы поспеть за стремительным ростом информационных потоков, стекающихся сюда каждый день, — сказал директор AIT Ларри Ферчайлд (Larry Fairchild). — Мы должны снабжать сотрудников технологией, которая поможет им справиться с гигантскими объемами оперативно обрабатываемых данных."

В плане профессионального использования инструментов Text Mining ЦРУ — далеко не монополист. По прогнозам аналитической компании IDC, спрос на подобные программы существенно возрастет в течение ближайших 4-5 лет. Так, к 2005 году ожидается повышение прибылей от такого ПО с 540 млн долларов (в 2002 году) до полутора миллиардов. Такие возможности, как экспресс-анализ найденной информации, информационная разведка (выявление разрозненной прямой и косвенной информации по некоторой проблеме), формирование и ведение тематических досье с возможностью выявления тенденций и взаимосвязей персон, событий, процессов, уже используются рядом крупных предприятий и наверняка будут востребованы в дальнейшем.

Как утверждает эксперт Алессандро Занаси (Alessandro Zanasi), ранее сотрудник META Group., к 2006 году такого рода программы станут доминирующими при анализе информации от клиентов в компаниях любого уровня, будь то телефонные центры поддержки, Internet-агентства или аналитические агентства. Кадровые отделы будут использовать программы этого класса для поиска резюме, подходящих по сложной сетке показателей, а маркетинговые службы найдут применение таким программам в качестве анализаторов ситуации на рынке, отслеживающих тенденции, положение конкурентов и другие показатели на основе информации из самых разных источников — новостных лент, отчетов о НИР, обзоров, патентов.

5.8. Автоматическое реферирование

Экспоненциальный рост темпов производства информации, безусловно, существенно снижает эффективность обработки информации традиционными методами. С самого начала компьютерной эры создавались программы автоматизированной обработки текстов, реализующие индексирование, аннотирование, реферирование, фрагментирование и другие формы информационного анализа и синтеза.

Такие программы, с одной стороны, способствуют расширению информационного пространства, а с другой — являются единственным инструментом, который потенциально может обеспечить охват современных информационных ресурсов. Особенно большое значение приобрела задача *автоматического реферирования* (Automatic Text Summarization) [1, 35, 64] — составление кратких изложений материалов, аннотаций или дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов, и генерация на их основе лаконичных и информационно емких отчетов. Сегодня потребность в автоматическом реферировании текстов стабильно возрастает.

Вместе с тем, нишу систем автоматического реферирования нельзя считать заполненной. Большинство процессов создания аннотаций еще неэффективны, сохраняется необходимость в масштабируемых методологиях и программах. Учитывая бурный рост технологий глубинного анализа текстов (Text Mining), ожидается большой прогресс и в области автореферирования. Однако, несмотря на то что отдельные производители уже создали системы автореферирования, порождаемые сегодня объемы информации не позволяют оперативно получать аннотации с необходимой полнотой и релевантностью.

На сегодня существует множество путей решения задачи, которые достаточно четко подразделяются на два направления, — квазиреферирование и краткое изложение содержания первичных документов. Квазиреферирование основано на экстрагировании фрагментов документов, т.е. выделении наиболее информативных фраз и формировании из них квазирефератов.

Краткое изложение исходного материала основывается на выделении из текстов с помощью методов искусственного интеллекта и специальных информационных языков наиболее существенной информации и порождении новых текстов, содержательно обобщающих первичные документы.

Конечно, применяя такой подход, можно получать более сложные аннотации, которые в принципе могут содержать информацию, дополняющую исходный текст. Благодаря опоре на формальное представление семантики исходного документа, подобные системы теоретически могут быть настроены на очень высокую степень сжатия, необходимую, например, для рассылки сообщений на мобильные устройства. Иначе говоря, главное различие между средствами реферирования состоит в том, *что* они, в сущности, формируют набор выдержек или краткое изложение документа.

Все существующие промышленные системы класса Text Mining включают средства автореферирования, которые являются неотъемлемыми компонентами таких систем. Одна из базовых процедур систем этого класса — автоматическое формирование дайджестов — представляет собой автореферирование на основе большого количества документов. Для дайджеста отбираются документы, в которых наиболее явно отражены тенденции всего входного потока. Можно утверждать, что такие дайджесты должны в наибольшей степени соответствовать информационным потребностям пользователя, по запросу которого формируется этот входной информационный поток.

Предполагается, что на основании реферата, составляющего по объему незначительную часть исходного текста, пользователи смогут составить обоснованное заключение о первичном документе, затратив на это значительно меньше усилий, в сравнении с его полным прочтением [8]. Как правило, при автореферировании объем реферата должен составлять от 5 до 30% исходного текста. Подготовка документов, представляющих собой аннотации из нескольких источников, т.е. дайджестов, предполагает еще большую степень сжатия. При этом анализ качества реферирования — это отдельная и очень важная задача, на которую зачастую не удается получить однозначного ответа. Как показывает практика, даже люди редко приходят к согласию относительно качества передачи основного смысла в одном и том же реферате.

Все существующие промышленные системы класса Text Mining включают средства автореферирования, которые являются неотъемлемыми компонентами таких систем.

5.8.1. Квазиреферирование

Несмотря на большую популярность методов искусственного интеллекта, в области автоматического реферирования сегодня можно констатировать тот факт, что получение семантически наполненных результатов оказалось возможным и без привлечения баз знаний и правил. Вместе с тем, разработчики средств автореферирования все больше внимания уделяют гибридным системам, успешно объединяя статистические методы и методы искусственного интеллекта.

Большинство систем автореферирования сегодня использует вариации статистических методов анализа, зачастую игнорируя при этом лингвистическую взаимосвязанность и семантику естественного языка. В большинстве известных систем автоматическое реферирование по сути является экстрагированием, т.е. квазиреферированием. Развитый синтаксический разбор и применение баз знаний или хотя бы тезаурусов встречаются очень редко.

Предложения, характеризующиеся как "обрывки", — например, начинающиеся со слов "При этом...", "Во-вторых..." и т.д., зачастую просто игнорируются подобными системами. В наиболее развитых на сегодня системах реферирования учитывается зависимость предложений друг от друга, что обеспечивает связность результирующих аннотаций — подбираются группы взаимосвязанных предложений, которые для достижения большей связности слегка изменяются на стыках.

Еще одно направление, заключающееся в формировании изложений на основе использования баз знаний и являющееся в целом более перспективным, в настоящее время, к сожалению, представлено лишь экспериментальными исследованиями — до широкой реализации дело еще не дошло.

Квазиреферирование сводится к экстрагированию (извлечению) из документов минимальных релевантных фрагментов. При этом, по сравнению с кратким изложением, оно обладает той особенностью, что основывается на анализе поверхностно-синтетических отношений лексических единиц в тексте, выраженных в нем и не требующих обращения к семантическим процессам, изучения которых пока еще недостаточно для описания свойств любого текста.

Квазиреферирование предполагает акцент на выделение характерных фрагментов методом сопоставления фразовых шаблонов, в результате чего выделяются блоки наибольшей лексической и статистической релевантности. Автоматическое определение частот использования отдельных слов и сочетаний в исходном документе позволяет определять абзацы и предложения, в которых тематика документа представлена наиболее точно. Создание итогового документа в данном случае представляет собой просто соединение выбранных фрагментов. Формируемый квазиреферат при этом производит впечатление связного текста, что значительно облегчает его восприятие. Однако качество реферирования при этом во многом зависит от жанра обрабатываемого текста. Гладкость и содержательность квазиреферата также зависит и от других особенностей исходного текста. Так, для больших текстов, монографий или интервью построение качественного реферата из фрагментов исходного документа без учета семантических закономерностей практически невозможно.

Основу аналитического этапа квазиреферирования составляет процедура вычисления весовых коэффициентов для каждого блока текста в соответствии с такими характеристиками, как расположение этого блока в оригинале, частота появления в тексте, частота использования в ключевых предложениях, а также некоторые другие показатели.

В рамках квазиреферирования выделяют три основных направления, применяемые совместно в современных системах.

- Статистические методы, основанные на оценке информативности различных элементов текста по частоте использования, которая служит основным критерием информативности слов, предложений или фраз.
- Позиционные методы, опирающиеся на предположение о том, что информативность элемента текста находится в зависимости от его позиции в документе.
- Индикаторные методы, основанные на оценке элементов текста исходя из наличия в них специальных слов и словосочетаний — так называемых *маркеров важности* [2] ("в заключение", "было отмечено, что..." и пр.), характеризующих их смысловую значимость. Иначе говоря, индикаторные методы обеспечивают оценку фраз первичного документа на основе специальных словарей маркеров.

Следует отметить, что для русского языка, например, существуют словари маркеров, включающие свыше 1500 лексических единиц внетематической лексики, а также формулы выбора, отражающие требования к вторичным документам, получаемым путем экстрагирования фраз на основе индикаторных методов. Эти элементы лексического аппарата обеспечивают достаточно точную идентификацию фрагментов исходного текста.

5.8.2. Алгоритмы автореферирования

Большинство алгоритмов автоматического реферирования документов предполагают три основных этапа: анализ исходного текста, определение весомых фрагментов (предложений или целых абзацев) и формирование вывода.

Первый этап начинается с выделения из исходного текста лексических единиц (слов или словосочетаний), их взвешивания по некоторым критериям и определения массива самых весомых. При этом сначала выполняется выделение из исходного текста всех лексических единиц и построение из них последовательного словарного массива. При этом каждой лексической единице присваивается предварительный коэффициент, зависящий от ее расположения в исходном тексте. Затем выполняется их нормализация с помощью средств автоматического морфологического анализа (в настоящее время это уже решенная проблема). Морфологический анализ решает задачу приведения всех слов к каноническому виду. Цель морфологического анализа состоит в выделении основ слов, т.е. словоформ с отсеченными окончаниями, а также при необходимости в подключении синонимических цепочек для отдельных слов. Для выполнения последующего семантического анализа каждой словоформе ставятся в соответствие значения грамматических категорий (род, падеж, число).

На этом этапе также выполняется удаление из словарного массива слов, не несущих явной смысловой нагрузки. Для этого применяются программные средства, основанные на использовании так называемого "стоп-словаря". Затем все лексические единицы массива сортируются, и определяется их частота появления. При этом каждой из лексических единиц присваивается весовой коэффициент, который определяется как результат учета нескольких составляющих: частоты появления, тематического словаря (определяемого, например, тематикой

запроса пользователя) и "плюс-словаря", включающего наиболее важную лексику общего назначения.

Последний этап при формировании массива лексических единиц заключается в выборе некоторого ограниченного количества самых весомых терминов. Полученный массив лексических единиц, кроме задачи автореферирования, в дальнейшем может быть полезен и при различных лингвистических исследованиях текста.

Определение веса фрагментов (предложений или абзацев) исходного текста выполняется по алгоритмам, разработанным еще в 60–70-е годы XX века и ставшим уже традиционными. Общий вес текстового блока на этом этапе вычисляется по формуле:

$$\text{Weight} := \text{Location} + \text{KeyPhrase} + \text{StatTerm}.$$

Здесь коэффициент *Location* определяется расположением блока в исходном тексте и зависит от того, где появляется данный фрагмент — в начале, в середине или в конце, а также используется ли он в ключевых разделах текста, например в заключении. Ключевые фразы (*KeyPhrase*) представляют собой резюмирующие конструкции-маркеры типа "в заключение", "в данной статье", "согласно результатам анализа" и т.п. Весовой коэффициент ключевой фразы может зависеть также от оценочного термина, например "отличный". Статистический вес текстового блока (*StatTerm*) вычисляется как нормированная по длине этого блока сумма весов входящих в него терминов — слов и словосочетаний. После выявления определенного (заданного коэффициентом необходимого сжатия) количества текстовых блоков с наивысшими весовыми коэффициентами, они объединяются для построения квазиреферата.

Конечно, преимущество методов квазиреферирования заключается в простоте их реализации. Однако выделение текстовых блоков, не учитывающее взаимоотношений между ними, часто приводит к формированию бессвязных рефератов. Некоторые предложения могут оказаться пропущены либо в них могут встречаться слова или фразы, которые невозможно понять без предшествующего, но пропущенного в автореферате текста. Попытки решить эту проблему, в основном, сводятся к исключению таких предложений из рефератов. Реже делаются попытки разрешения ссылок с помощью методов лингвистического анализа. В ряде человеко-машинных подходов создаются специальные интерфейсы, с помощью которых можно определить наличие смыслового разрыва или "висящего" слова. Очевидно, что такой подход не годится для сколько-нибудь массовой обработки текстов.

5.8.3. Дайджесты

Дайджест представляет собой аннотированный текст, построенный на основе анализа нескольких документов. При составлении дайджестов методы автореферирования одного документа распространяются на массив из большого количества документов. Вместе с тем, дайджест можно также рассматривать как аннотированный источник гиперссылок на документы, лежащие в его основе.

При формировании дайджестов методами квазиреферирования практически невозможно получить связный текст. Объединение рефератов каждого из документов неизбежно будет содержать избыточную несвязную информацию. Однако при условии составления автореферата, состоящего из определенного количества анонсов входных документов и разделенного на подразделы в соответствии с этими документами, описанный выше метод оказывается вполне приемлемым.

Как и в случае квазиреферирования одного текстового документа, на первом этапе формирования дайджеста происходит отбор наиболее весомых лексических единиц, входящих в массив исходных документов (входной информационный поток), на основании которых строится словарь системы.

Выбор исходных документов из входного массива построения дайджеста осуществляется также с учетом их весов. Вес каждого документа определяется с учетом нормированной по длине документа суммы весов отдельных слов, входящих в этот документ. Этап выбора документов для дайджеста состоит из таких шагов, как определение веса каждого документа, сортировка входного потока документов по весам, определение смысловых дублей документов по статистическим критериям, отбрасывание документов, непригодных для построения дайджестов (недопустимых типов документов, например обзоров), а также смысловых дублей (выявляемых по частотным алгоритмам). Последний этап выбора документов для формирования дайджеста заключается в выборе заранее определенного количества самых весомых документов из отсортированного и отфильтрованного на предыдущих этапах массива.

Статистический алгоритм выявления дублирующихся документов из входного потока может базироваться, например, на определении цепочек ключевых слов и частот их использования для отдельных документов и последующем сравнении их между собой всех таких цепочек исходных документов.

Последний этап синтеза дайджеста заключается в выделении из отобранных документов самых значимых предложений и построении из них единого текста, разделенного на подразделы. Для этого к каждому из отобранных документов может применяться описанный выше алгоритм квазиреферирования.

Отобранные документы представлены в дайджесте заранее заданным количеством весомых предложений. В случае формирования дайджестов на основе динамически изменяющейся информации из Internet, автоматически формируется гипертекстовое представление самого дайджеста, который можно рассматривать как самостоятельный документ, обладающий ссылками на документы-первоисточники в Сети.

Приведенная выше процедура обеспечивает формирование дайджеста, отражающего основные тенденции, представленные в исходном информационном массиве. Вместе с тем, имеет смысл формирование "веерного" многоаспектного дайджеста, отражающего наряду с главной тенденцией несколько других аспектов, игнорируемых в дайджестах первого типа. Многоаспектный дайджест можно построить, базируясь на технологических решениях, применяемых при предыдущем подходе, при реализации следующего алгоритма.

- **1 этап.** Построение дайджеста, отражающего основную тенденцию.
- **2 этап.** Удаление из входного информационного потока документов, соответствующих тенденции, определенной на предыдущем шаге.
- **3 этап.** Построение дайджеста, отражающего основную тенденцию оставшейся части информационного потока.
- **4 этап.** Объединение полученных дайджестов.
- **5 этап.** При необходимости (исходя из требуемых объемов результирующего дайджеста) выполняется переход к этапу 2.

5.8.4. Поисковые образы документов

Задача полнотекстового поиска, в последнее время ставшая особенно актуальной в связи с развитием ресурсов Internet, предполагает проведение поиска документов, в том числе и больших объемов, с использованием весовых критериев и логических операторов. Вместе с тем, проведение поиска по всему тексту может оказаться неэффективным, — например, в романе Л.Н. Толстого "Война и мир" можно найти большинство лексем русского языка. В таких случаях проблему точности решает поиск по аннотированным текстам. Иначе говоря, вместо поиска по полным текстам оказывается целесообразным выполнять поиск по аннотациям — поисковым образам документов.

При этом методы квазиреферирования легко настроить для обработки крупных массивов информации. Хотя квазиреферат часто для больших текстов оказывается образованием, лишь отдаленно напоминающим исходный текст и при этом зачастую не воспринимаемым человеком, именно как поисковый образ документов, содержащий взвешенные ключевые слова и фразы, он может приводить к вполне адекватным результатам при полнотекстовом поиске. Поэтому можно прогнозировать, что статистические методы реферирования, квазиреферирование получат широкое распространение в области автоматического индексирования.

5.8.5. Информационные портреты

Портрет можно рассматривать как модель реального объекта (или субъекта), выраженную его наиболее узнаваемыми чертами. Как в связи с задачами автореферирования, так и для решения других аналитических задач возникает потребность оценить содержание документа, получить его "информационный портрет", т.е. статистически значимую совокупность информационных характеристик. В большинстве из существующих реализаций такой портрет состоит из статистически значимых слов и выражений, сопровождающих упоминание объекта.

Например, в качестве информационного портрета темы, соответствующей запросу, можно рассматривать множество ключевых слов, наиболее точно (по статистическим и смысловым алгоритмам) отражающее информацию, получаемую в результате поиска по данному запросу. Построение информационных портретов в реально функционирующих системах выполняется на базе эмпирических и статистических методов, основу которых, как и в случае автореферирования, составляют частотно-лингвистические алгоритмы.

С помощью информационного портрета в ИПС может детализироваться и уточняться критерий поиска. Информационный портрет может быть реализован как отдельная семантическая карта или как таблица на экране с результатами поиска. Чаще всего в этих случаях для уточнения запроса определенным словом из информационного портрета достаточно просто активизировать гиперссылку, соответствующую этому слову. Для уточнения запроса сразу несколькими словами из информационного портрета часто используется механизм установки флажков опций (checkbox), находящихся рядом со словами в информационном портрете.

5.8.6. Программы автореферирования

На рынке существует достаточно большое количество программ автореферирования, реализующих преимущественно статистические алгоритмы. Одним из первых проектов такого типа была система Inxight Summarizer, созданная в 1995 году в Исследовательском центре корпорации Ксерокс в Пало Альто. Эта

система параллельно использовала несколько известных алгоритмов реферирования и оценки качества рефератов. Кроме того, она распространялась не только в виде готовой программной системы, но и в виде модулей реферирования в составе библиотек для платформ Win32 и Solaris.

Компания British Telecommunications Laboratories для экспериментальной онлайн-платформы TranSend в свое время создала Prosum — систему реферирования, реализованную в виде cgi-сценария, встраиваемого в страницы Web-сайтов.

В текстовом процессоре Microsoft Word 2000 реализована функция Автореферат (AutoSummarize), которая обеспечивает формирование рефератов из фраз исходного текста, наиболее информативных с точки зрения вхождения в них высокоранговых для данного текста слов. При этом пользователь может устанавливать относительный размер реферата (коэффициент сжатия первичного документа) от 50 до 10% исходного объема. Полученный в результате текст реферата является лишь наброском, и пользователю его всегда необходимо будет дополнительно корректировать — сама совокупность фраз в реферате не обеспечивает его смыслового единства. Для аннотирования текстов на русском языке существует компонент системы ОРФО 5.0, выпускаемой компанией "Информатик" (www.informatic.ru), и программа Либретто 1.0 компании "МедиаЛингва" (www.medialingua.ru), выполняющая аннотирование русских и английских документов. Обе эти программы могут быть встроены в среду Microsoft Word. Уже устаревший, но вполне работоспособный вариант Либретто можно получить по адресу: http://www.vlz.ru/books/pcmag/b.htm. В программе Либретто коэффициент сжатия задается пользователем. Программа имеет два режима: собственно аннотирование и выделение ключевых слов. В режиме аннотирования из текста отбираются предложения, в наибольшей степени характеризующие его содержание. В режиме выделения ключевых слов производится выборка из текста наиболее информативных слов — построение его информационного портрета.

В настоящее время компания "МедиаЛингва" предлагает на рынке другую систему — "Аннотатор SDK 1.0" (http://www.medialingua.ru/annotator.html), представляющую собой набор средств, предназначенный для автоматического аннотирования документов любого объема и степени сложности на русском и английском языках (рис. 6.12). В этой системе для каждого предложения входного текста на основе вероятностных моделей, дополненных лингвистическими методами и словарями, вычисляются весовые коэффициенты.

Из наиболее значимых и независимых предложений составляется реферат настраиваемого размера. Для придания реферату большей связности исходные предложения могут быть переформулированы. В результате получается аннотация, более связанная, чем построенная в режиме обычного квазиреферирования. Кроме того, пакет Аннотатор SDK 1.0 обеспечивает построение некоторого подобия информационных портретов — системой обеспечивается выделение в исходном тексте наиболее информативных ключевых слов. Система включает набор интерфейсов (API) для использования в сторонних приложениях, написанных на языках программирования C/C++, Visual Basic и Java.

Российская компания AGCProduct выпустила бесплатную программу Content Analyzer (последняя версия, 0.52, доступна по адресу http://www.agcproduct.com/rus/downloads/#ca), предназначенную для автореферирования и построения информационных портретов Web-страниц в режиме он-лайн (рис. 5.13). Эта программа обеспечивает анализ Web-страниц из Internet или с локального диска на русском и английском языках.

Рис. 5.12. Аннотатор SDK 1.0 — конструктор рефератов

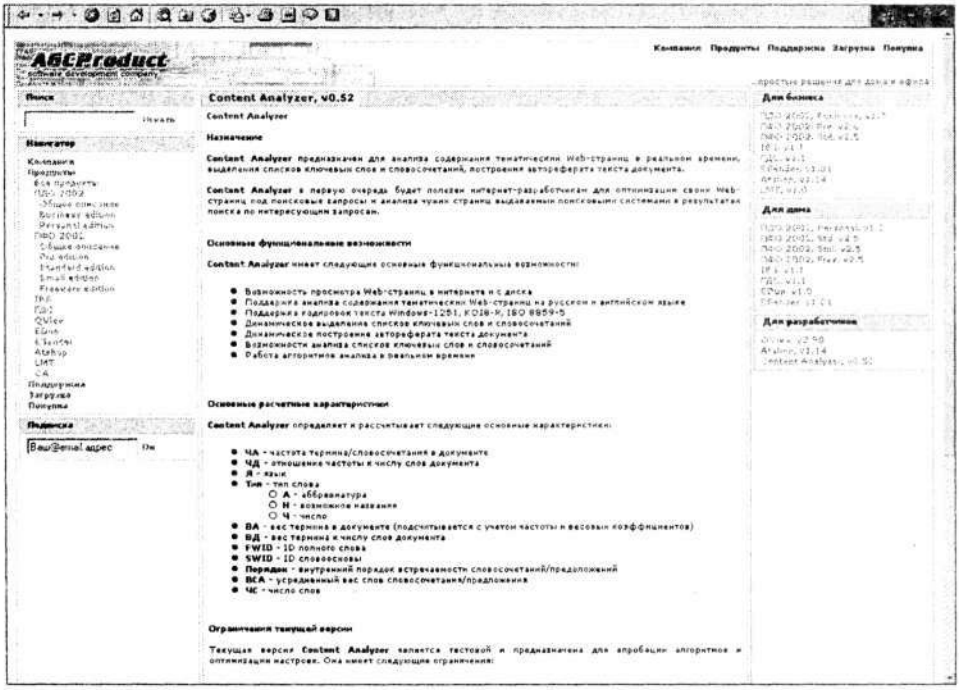


Рис. 5.13. Программа Content Analyzer

В свое время российская компания "Микросистемы" (MicroSystems) выпустила программу TextReferent. В настоящее время эта программа распространяется бесплатно. TextReferent предоставляет возможность составления квазирефератов — наборов предложений исходного текста, которые содержат ключевые термины (рис. 5.14). Система позволяет настраивать "подробность" получаемого реферата.

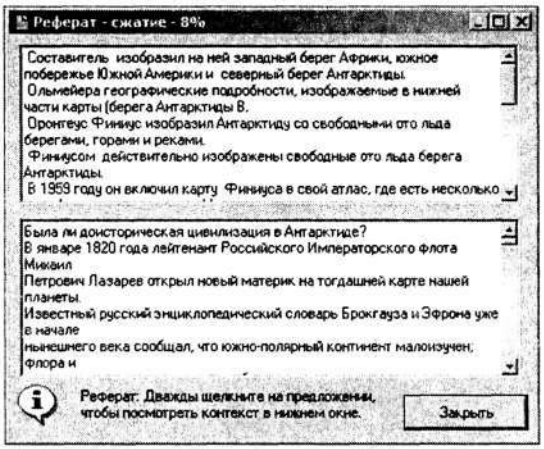


Рис. 5.14. *Интерфейс программы TextReferent*

Система TextAnalyst была разработана этой же компанией как инструмент для анализа содержания текстов, смыслового поиска информации и формирования электронных архивов. В то время как TextReferent реализует лишь одну из функций системы TextAnalyst — автоматическое составление реферата, полная версия системы предоставляет пользователю возможности выделения именных групп и построения на их основе семантической сети — структуры взаимосвязей между именными группами. Система TextAnalyst обеспечивает следующее.

- Анализ содержания текста с автоматическим формированием семантической сети с гиперссылками — получение смыслового портрета текста в терминах основных понятий и их смысловых связей.
- Анализ содержания текста с автоматическим формированием тематического древа с гиперссылками — выявление семантической структуры текста в виде иерархии тем и подтем.
- Смысловой поиск с учетом скрытых смысловых связей слов запроса со словами текста.
- Автоматическое реферирование текста из наиболее информативных фраз (рис. 5.15).
- Кластеризацию информации — анализ распределения материала текстов по тематическим классам.
- Автоматическую индексацию текста с преобразованием в гипертекст.
- Ранжирование всех видов информации о семантике текста с возможностью варьирования детальности.

- Автоматическое/автоматизированное формирование полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

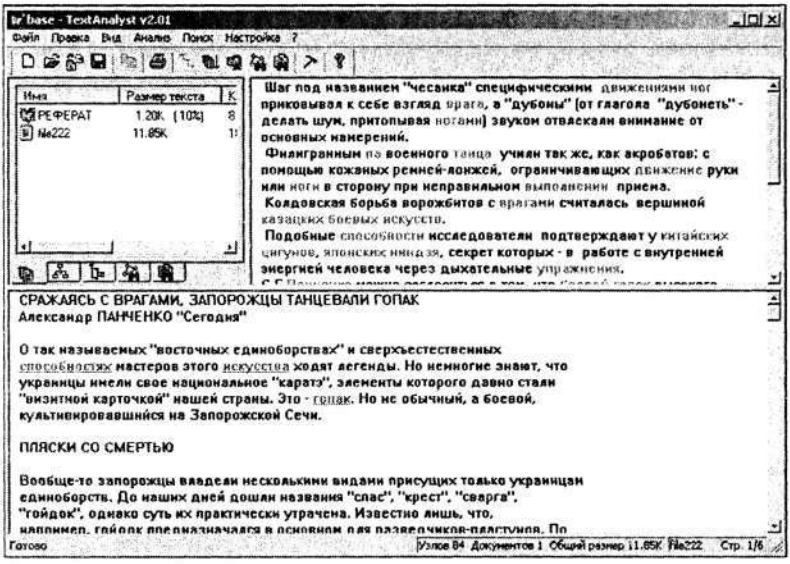


Рис. 5.15. Реферирование — лишь одна из возможностей системы TextAnalyst

Возможности систем квазиреферирования ограничены выделением и выбором оригинальных фрагментов из исходного документа и соединением их в короткий текст. Подготовка же краткого описания предполагает *изложение*, т.е. краткий пересказ содержания текста. Вместе с тем, когда рынок недорогих традиционных систем автоматического реферирования был практически заполнен, появились новые гибридные решения, построенные на основе синтаксических анализаторов.

Мощные современные системы

Обе, ставшие уже традиционными программы TextAnalyst и TextReferent сегодня нашли свое развитие в мощной системе PolyAnalyst новой компании Megaputer. Сегодня пакет PolyAnalyst — это один из немногих коммерческих продуктов, в котором реализованы методы Text Mining — глубинного анализа текстовой информации. Система PolyAnalyst предназначена для автоматического анализа числовых и текстовых данных с целью обнаружения в них ранее неизвестных, нетривиальных, практически полезных и доступных пониманию закономерностей, необходимых для принятия оптимальных решений во многих областях человеческой деятельности. Благодаря технологии эволюционного программирования и другим математическим алгоритмам, PolyAnalyst сочетает в себе высокую производительность с относительно невысокой стоимостью. Современная система Text Analysis, входящая в пакет PolyAnalyst, представляет собой средство формализации неструктурированных текстовых полей в базах данных. При этом текстовое поле представляется как набор булевых признаков, основанных на наличии и/или частоте данного слова, устойчивого словосочетания или понятия (с учетом отношений синонимии и "общее-частное") в данном

тексте. При этом появляется возможность распространить на текстовые поля всю мощь алгоритмов Text Mining, реализованных в системе PolyAnalyst. Кроме того, этот метод может быть использован для лучшего понимания текстовых компонентов данных за счет автоматического выделения наиболее ключевых понятий.

Мощные современные системы класса Text Mining, включающие элементы автореферирования, разработаны сегодня такими производителями, как IBM (Intelligent Miner), Silicon Graphics (SGI Miner), Integral Solutions (Clementine), SAS Institute (SAS). Неотъемлемой частью продуктов Oracle сегодня также являются средства глубинного анализа текстов, которые появились еще в составе Oracle 7.3.3 (Text Server) и в Oracle8i (interMedia Text). Для решения проблемы обработки текстовой информации на русском языке в Oracle Text компанией Гарант-Парк-Интернет был разработан модуль Russian Context Optimizer (RCO), предназначенный для совместного использования с interMedia Text (или Oracle Text). Помимо поддержки русскоязычной морфологии, RCO включает в себя дополнительные средства нечеткого поиска, тематического анализа и реферирования документов. В состав системы входит мощный программный комплекс RCO КАОТ, который обеспечивает автоматический анализ содержания полнотекстовых документов и поддержку рабочего места аналитика (рис. 5.16).

Рис. 5.16. Страница RCO КАОТ на сервере Russian Context Optimizer

Одна из первых промышленных программ автореферирования с элементами семантического анализа — Extractor — была создана в Институте информационных технологий Национального исследовательского Совета Канады. Она обеспечивает

выделение из входного исходного текста наиболее информативных именных групп. Основной модуль Extractor используется в программных продуктах фирм ThinkTank Technologies и Tetranet, а также в некоторых поисковых системах.

В настоящее время аннотирована технология анализа текстовой информации WebFountain, разрабатываемая в исследовательском центре IBM Almaden Research Center (рис. 5.17). Эта технология ориентирована на анализ слабоструктурированных и неструктурированных данных, получаемых из Internet [49].

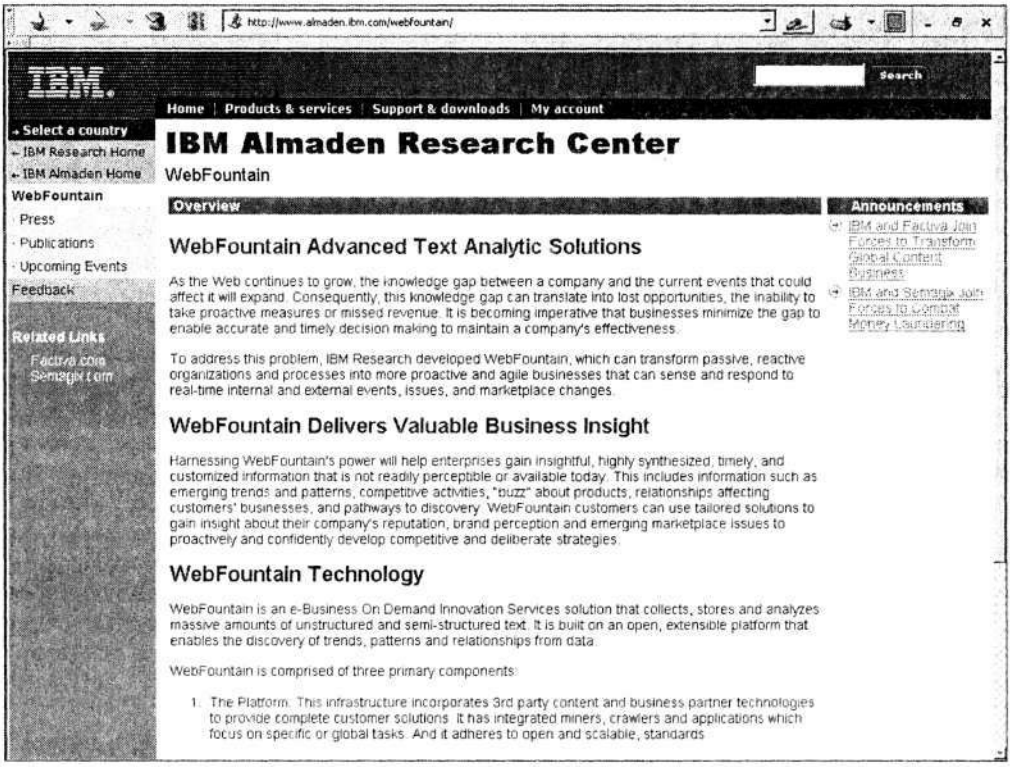


Рис. 5.17. *WebFountain — новый проект корпорации IBM*

WebFountain обрабатывает не только статические страницы, но и доступные корпоративные базы e-mail, каналы IRC, Живые журналы (Web-логи или просто блоги), электронные доски объявлений, специализированные хранилища бизнес-информации, а также новостные ленты и периодику. Основные новшества WebFountain заключаются в технологиях контент-анализа и структурирования информации, которые обеспечивают точный тематический поиск. При этом сам контент-анализ выполняется специальными модулями-аннотаторами, которые могут разрабатываться сторонними компаниями и не являются неизменной частью системы WebFountain. Реализация конкретных аннотаторов зависит от вида информации, на работу с которой настраивается система. Аннотатор вкладывает в исходные документы свои "знания", дополняя их исходные тексты специальными XML-тегами, содержащими расширенную информацию о понятиях, встречающихся в текстах. Иначе говоря, аннотатор связывает значение отдельных слов с некоторой дополнительной релевантной информацией. Это делается для того, чтобы последующая обработка

текста велась уже с учетом дополнительных сведений о понятиях — словах или словосочетаниях, которые в нем встречаются. В дальнейшем текст, увеличившийся по объему в несколько раз за счет тегов с дополнительными сведениями, поступает на обработку модулей анализа информации и синтеза выходных форм.

5.8.7. Автореферирование на основе семантических методов

Подход, опирающийся на методы искусственного интеллекта, исходит из предположения, что если удается определить семантику текста, то аннотация будет более качественной. Используемые при этом базы знаний должны постоянно поддерживаться в актуальном состоянии и сопровождаться экспертами [50]. Для подготовки рефератов при таком подходе требуются мощные информационные ресурсы для обработки данных на естественных языках, в том числе базы грамматических правил и словари для синтаксического разбора естественно-языковых конструкций. Для реализации этого метода нужны многочисленные справочники, отражающие понятия, ориентированные на предметные области, необходимые для анализа и определения наиболее важной информации.

В отличие от частотно-лингвистических методов, обеспечивающих квазиреферирование, подход, основанный на базах знаний, опирается на автоматизированный качественный контент-анализ, состоящий, как правило, из трех основных этапов. Первый — сведение исходной текстовой информации к заданному числу фрагментов, т.е. единиц значения, которыми являются категории, последовательности и темы. На втором этапе производится поиск регулярных связей между единицами значения, после чего наступает третий этап — формирование выводов и обобщений. Строится структурная аннотация, представляющая содержание текста в виде совокупности концептуально связанных единиц значения.

Семантические методы формирования рефератов-изложений предполагают два основных подхода: метод синтаксического разбора предложений и методы, опирающиеся на понимание естественного языка. В первом случае используются деревья разбора текста. Процедуры автореферирования манипулируют непосредственно деревьями, выполняя перегруппировку и сокращение ветвей на основании структурных критериев. Такое упрощение обеспечивает построение автореферата — структурную "выжимку" исходного текста.

Второй подход основывается на системах искусственного интеллекта, в которых на этапе анализа также выполняется синтаксический разбор текста, но синтаксические деревья не порождаются. В этом случае формируются семантические структуры, которые накапливаются в базе знаний. В частности, известны модели, позволяющие производить автореферирование текстов на основе психологических ассоциаций сходства и контраста. В базах знаний избыточная и не имеющая прямого отношения к тексту информация устраняется путем отсечения некоторых концептуальных подграфов. Затем информация подвергается агрегированию методом слияния графов или обобщения. Для этих преобразований выполняются манипуляции логическими предположениями, выделяются определяющие шаблоны в текстовой базе знаний. В результате преобразования формируется концептуальная структура текста — аннотация, т.е. концептуальные "выжимки" из текста.

Многоуровневое структурирование текста с использованием семантических методов позволяет подходить к решению задачи реферирования различными методами.

1. Путем удаления малозначащих смысловых единиц. Преимуществом метода является гарантированное сохранение значащей информации, недостатком — низкая степень сжатия.

2. Посредством сокращения смысловых единиц, т.е. заменой их основной лексической единицей, выражающей основной смысл.
3. Гибридным способом, заключающимся в уточнении реферата с помощью статистических методов, с использованием семантических классов, особенностей контекста и синонимических связей.

Хотя использование семантических методов при реферировании зачастую приводит к потере некоторых второстепенных смысловых элементов, однако они не снижают качество реферата — его точность, компактность и связность.

К программам, базирующимся на семантических методах, можно отнести, например, интерактивный анализатор текстов DictaScope (`http://www.dictum.ru/onlinedictascope.htm`). Программа ориентирована на работу с текстами на русском языке и позволяет в пошаговом режиме наблюдать процесс автоматического построения дерева смысловых связей (рис. 5.18). Интерфейс программы разработан для применения ее в учебном процессе лингвистических и филологических отделений. Модули анализатора DictaScope могут быть использованы в информационно-поисковых системах, в системах извлечения знаний и системах автоматического реферирования.

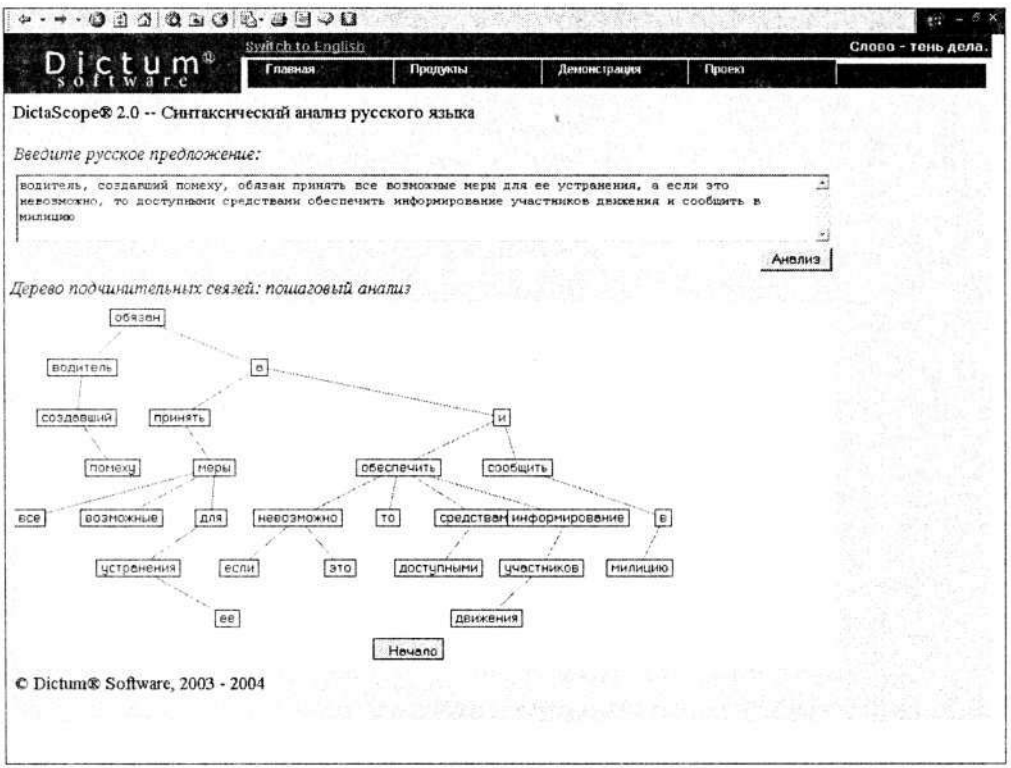


Рис. 5.18. Интерактивный анализатор текстов DictaScope

К немногим системам, построенным на основе когнитивных подходов, можно отнести систему Астарта российской компании Cognitive Technologies Ltd. (`http://www.cognitive.ru/products/astarta.htm`), предназначенную для

эффективного сбора, обработки и анализа неструктурированной информации (рис. 5.19). Эта программа обеспечивает следующие возможности.

- Сбор информации из таких источников, как Web-сайты, новостные ленты и т.д.
- Автоматическое определение тематики документов, т.е. автоматическое отнесение документа к тем или иным рубрикам. При этом реализовано автоматическое обучение рубрикатора, т.е. автоматическое построение списка терминов и понятий, определяющих принадлежность документа к рубрике.
- Формирование широкого спектра информационных отчетов (рефератов и дайджестов) по разнообразным критериям. Стиль отчетов определяется пользователем системы, в распоряжении которого предоставлено хранилище шаблонов, содержащее различные варианты представления дайджестов.

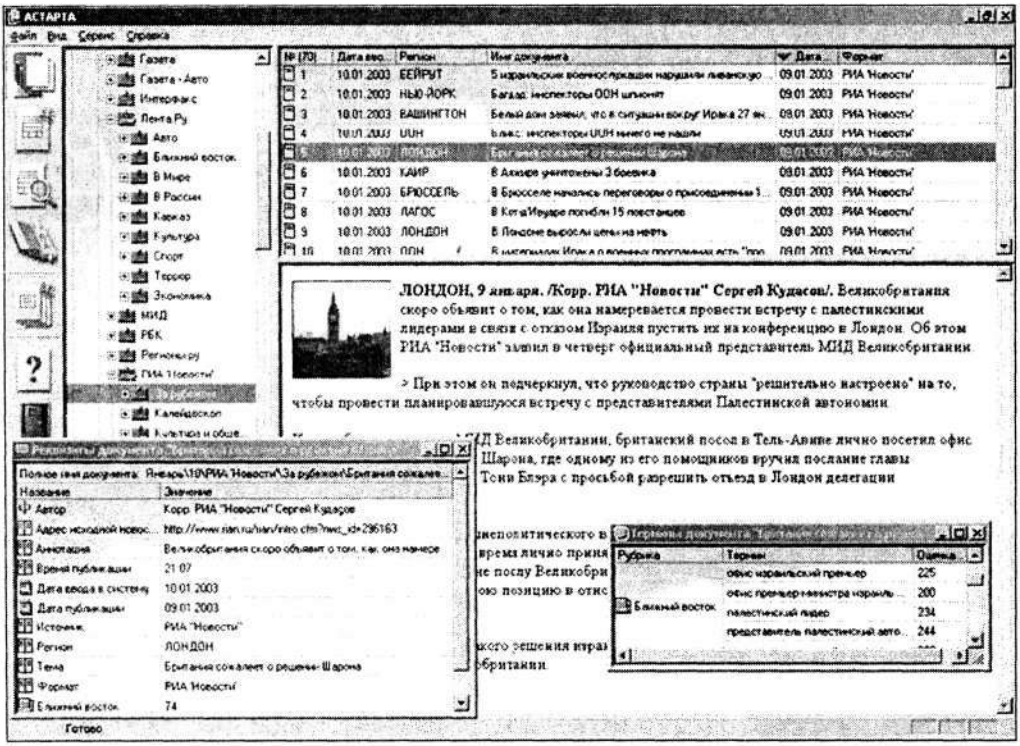


Рис. 5.19. Web-сайт системы "Астарта"

5.8.8. Перспективы автореферирования

По сравнению с традиционными подходами, использование технологий Text Mining при анализе ресурсов Internet уже сегодня обеспечивает, наряду с включением рабочих мест пользователей в динамическое информационное пространство, получение оперативных количественных и качественных аналитических срезов, что ранее было практически невозможным.

Кроме того, растущий объем мультимедийной информации в Сети делает ее также очень важным объектом для обработки средствами реферирования. Технологии автореферирования должны обрабатывать данные разного типа на этапах анализа и синтеза, реализуя интеграцию информации всех типов. Следует заметить, что это направление находится лишь в самом начале своего развития, но уже достигнуты определенные успехи. Например, в новой версии Яndex.Новостей сегодня появилась группировка по сюжетам не только текстовых сообщений, но и фото-, аудио- и видеофайлов.

Все сообщения в результатах поиска на сайте Яndex.Новости сгруппированы по сюжетам, при этом ранжирование построено на стандартных для Яndex принципах ранжирования сгруппированной выдачи. Оно строится на количестве и рангах отдельных сообщений внутри сюжетов, при этом ранг одной новости определяется ее оперативностью с учетом совпадений ключевых слов. В результате функционирования технологии выявления сюжетов на сайте `www.yandex.ru` представлены пять главных новостей за последний час, а на сайте `news.yandex.ru` эти новости представлены с цитатными аннотациями, а также имеется еще 10 новостей, следующих по важности.

Хотя сегодня подходы, не предполагающие использования методов искусственного интеллекта, будут доминировать, однако системы, основанные на экспертных системах, в ближайшее время смогут получить большее распространение в тех областях, для которых существуют разработанные лингвистические механизмы и базы знаний. Важно учитывать, что для работы с этими источниками нужны специалисты-эксперты, обладающие широкими познаниями в своей области.

Глава 6

Инструментарий конкурентной разведки

Без глубинного анализа информации, неуправляемые потоки которой сегодня, скорее, искажают реальную ситуацию на рынках, невозможно успешное ведение бизнеса.

У конкурентной разведки (competitive intelligence) [4], которая заключается в сборе и аналитической обработке информации, необходимой для принятия оптимальных управленческих решений руководством высшего звена компаний при ведении конкурентной борьбы, в последнее десятилетие появилось и развилось до невиданных ранее масштабов новое информационное поле — Web-пространство сети Internet.

Конкурентная разведка должна позволять получать данные о рынках сбыта, конкурентах, партнерах, контрагентах, новых технологиях, нормативных актах. При этом, в отличие от промышленного шпионажа, конкурентная разведка проводится строго в рамках правовых норм. Сбор и обработка информации при конкурентной разведке принципиально отличаются тем, что используют исключительно легитимные методы. Следует отметить, что конкурентная разведка на западе уже представляет собой отдельную область экономики.

Сегодня, по оценкам экспертов, Internet по количеству информации находится на первом месте, опережая СМИ, отраслевые издания и получаемые от коллег новости (по 15%), специальные обзоры (10%), закрытые базы данных (8%). При этом в открытых источниках и специализированных базах данных, доступных в Internet, содержится большая часть информации, необходимой для проведения конкурентной разведки, однако остается открытым вопрос ее нахождения и эффективного использования. Последние исследования информационного Web-пространства показали, что доступные через традиционные информационно-поисковые системы 10 млрд Web-страниц — это лишь "поверхностная" крупица. Непознанных, скрытых (deep, invisible) ресурсов Сети в сотни раз больше. Это, прежде всего, динамически генерируемые страницы, файлы разнообразных форматов, информация из многочисленных баз данных, которые представляют собой самый большой интерес для конкурентной разведки.

Существенно возросший объем информации в Internet затруднил поиск и выбор действительно нужных сведений. Ведь сама по себе информация, которая не служит для принятия решений, является беспредметной, а следовательно, несущественной. Традиционные сетевые информационно-поисковые системы не в полной мере справляются с задачей поиска информации, необходимой для решения задач конкурентной разведки. Эффективным дополнением их оказываются специализи-

рованные системы, широко распространенные в настоящее время. Это объясняет острую необходимость интеграции информации из различных источников.

Шеф разведслужбы флота США в годы второй мировой войны, адмирал Захария говорил: "Большая часть полезной информации — 95% — не является секретом". Мы можем заметить, однако, что не вся открытая "несекретная" информация является хорошо доступной, скорее наоборот. Доступ к необходимой в каждом конкретном случае информации является сложной задачей [36].

Естественно, при проведении конкурентной разведки отправной точкой считается не информационный шум, а исследуемый объект. Поэтому хотя использование информационного пространства Internet можно считать очень перспективным, одновременно следует учитывать и слабые стороны Сети: большой уровень недостоверности информации, неструктурированность необходимых данных и, как следствие, сложность их поиска. Но в целом возможности сети Internet оцениваются всеми экспертами в области конкурентной разведки достаточно высоко.

6.1. Задачи конкурентной разведки

Система конкурентной разведки должна позволять руководству, а также аналитическому и маркетинговому отделам компании не только оперативно реагировать на изменения ситуации на рынках, но и оценивать дальнейшие возможности своего развития. Основная цель систем конкурентной разведки — переход от традиционного метода интуитивного принятия решений на основе недостаточной информации к управлению, основанному на знаниях.

Конкурентная разведка в современных условиях выполняется для достижения двух основных целей — снижение рисков и обеспечение безопасности сделок, а также приобретение конкурентных преимуществ. Современная система конкурентной разведки должна позволять не только осуществлять мониторинг информации, но и моделировать стратегию конкурентов, выявить их партнеров, поставщиков, уяснить условия их сотрудничества.

Основные задачи систем конкурентной разведки относятся к нахождению и обобщению информации о конкурентах, рынках, товарах, бизнес-тенденциях и операциях по следующим объектам.

- Партнеры, акционеры, смежники, союзники, контрагенты, клиенты, конкуренты (личности и компании).
- Объединения компаний, слияния, поглощения, кризисные ситуации и т.п.
- Кадровый состав как своей компании, так и партнеров, конкурентов и т.д., а также кадровые изменения, их динамика.
- Торговый оборот, бюджет и его распределения по пунктам.
- Заключенные договора, достигнутые соглашения или договоренности.

Интерес при проведении конкурентной разведки вызывает не только непосредственная сфера деятельности компаний, но и сферы их влияния и интересов. Эти знания могут применяться, например, для оказания влияния на позиции партнеров и оппонентов в ходе деловых переговоров. Большое значение имеет информация, относящаяся к политике конкурентов, их намерениям, их сильным и слабым сторонам, продукции и услугам, ценам, рекламным кампаниям, другим параметрам рынка.

6.2. Источники информации и базы данных

Сегодня для конкурентной разведки основными источниками информации служат Internet, пресса, а также открытые базы данных. Очень популярны среди специалистов по конкурентной разведке базы данных государственных и статистических органов, торгово-промышленных палат, органов приватизации и т.д. Большую пользу приносят и отдельные доступные базы данных других органов власти. В последнее время все более популярны становятся базы данных на основе архивов СМИ, в том числе и сетевых. В России, например, большой популярностью пользуется крупнейшая архивная база данных СМИ службы Интегрум, содержащая несколько сотен миллионов документов. С помощью другой российской базы данных Лабиринт, составленной на основе публикаций ведущих бизнес-изданий, можно получить обширную информацию о конкретных персонах, организациях и компаниях.

Традиционно конкурентная разведка опирается на следующие источники информации: опубликованные документы открытого доступа, которые содержат обзоры товарного рынка, информацию о новых технологиях, создании партнерств, слияниях и приобретениях, объявлениях о рабочих вакансиях, о выставках и конференциях и т.п. Широко используются сведения, находящиеся в документах, уже имеющихся в компаниях, ведущих конкурентную разведку, результаты маркетинговых исследований, информация, полученная на конференциях, при общении с клиентами и коллегами. Большая часть этих данных попадает в сетевую прессу, пресс-релизы или публикуется на корпоративных Web-сайтах.

Как правило, для успешного ведения конкурентной разведки должен быть создан и непрерывно поддерживаться банк данных, включающий такие основные базы данных:

- конкуренты (действующие и потенциальные);
- информация о рынке (тенденции, номенклатурная, ценовая, адресная информация);
- технологии (продукты, выставки, конференции, ГОСТы, качество);
- ресурсы (сырье, человеческие и информационные ресурсы);
- законодательство (международные, центральные, региональные и ведомственные нормативно-правовые акты);
- общие тенденции — политика, экономика, региональные особенности, социология, демография.

Система конкурентной разведки, использующая Internet как один из источников информации, должна настраиваться под специфику деятельности компании. Она должна включать в себя соответствующую классификацию, гибкие механизмы поиска, оперативной доставки данных, а также качественной оценки информации. Одной из самых важных задач анализа информации является определение ее достоверности, т.е. задача анализа и фильтрации шума и ложной информации. Без таких оценок всегда есть риск принять неверные решения. После анализа достоверности информации должны следовать оценки ее точности и важности. Главным критерием достоверности данных на практике является подтверждение информации другими источниками, заслуживающими доверия. Например, не всегда стоит доверять "желтой" прессе или информации, поступившей от недостаточно профессиональных источников, которыми кишит Internet.

6.3. Подходы к анализу контента

Процесс конкурентной разведки можно рассматривать как построение сети из исследуемых объектов и связей между ними. Результаты должны представлять собой аналитическую информацию, которая может быть использована для принятия решений. Аналитическая информация может быть представлена в виде представленных наглядно схем — семантических сетей, дайджестов, наборов сюжетных линий, взаимосвязей ключевых понятий, компаний, лиц, технологий и т.п.

Задачи конкурентной разведки породили спрос на специальные информационные технологии, обеспечивающие возможность извлечения и обработки необходимой информации, что, в свою очередь, вызвало поток предложений систем со стороны разработчиков программного обеспечения.

Сегодня решать задачи конкурентной разведки на основе информации из Internet помогают общедоступные и специальные программы и сервисы. Например, в последнее время приобрели популярность так называемые "персонализированные разведпорталы", способные отбирать информацию по самым узким, специфическим вопросам и темам и предоставлять ее заказчикам.

В настоящее время декларированы технологии и системы "компьютерной конкурентной разведки", идея которых заключается в автоматизации и ускорении процессов извлечения необходимой для конкурентной борьбы информации из открытых источников и ее аналитической обработки.

При ведении конкурентной разведки все более широкое применение находят новые направления науки и технологий, получившие названия: "управление знаниями" (knowledge management) и "обнаружение знаний в базах данных" (knowledge discovery in databases), или Data и Text Mining — глубинный анализ данных или текстов.

Если системы управления знаниями реализуют идею сбора и накопления всей доступной информации как из внутренних, так и из внешних источников, то технологии Data Mining и Text Mining, как уже было показано выше, позволяют выявлять неочевидные закономерности в данных или текстах — так называемые латентные (скрытые) знания. В целом, эти технологии еще определяют как процесс обнаружения в "сырых" данных ранее неизвестных, но полезных знаний, необходимых для принятия решений. Системы этого класса позволяют осуществлять анализ больших массивов документов и формировать предметные указатели понятий и тем, освещенных в этих документах.

Характерная задача конкурентной разведки, обычно реализуемая в системах Text Mining, — это нахождение исключений, т.е. поиск объектов, которые своими характеристиками сильно выделяются из общей массы. Еще один класс важных задач, решаемых в рамках технологии Text Mining, — это моделирование данных, ситуационный и сценарный анализ, а также прогноз.

Для обработки и интерпретации результатов Text Mining большое значение имеет визуализация. Часто руководитель не всегда адекватно воспринимает предлагаемую ему аналитическую информацию, особенно если она не вполне совпадает с его пониманием ситуации. В связи с этим служба конкурентной разведки должна стремиться представлять информацию в виде, адаптированном к индивидуальному восприятию заказчика. Любопытно, что ЦРУ предоставляло Рональду Рейгану ежедневную информацию в виде видеофильма, который снимали каждый день, поскольку бывший киноактер воспринимал такую подачу информации наиболее адекватно.

Визуализация обычно используется как средство представления контента всего массива документов, а также для реализации механизма навигации по семантическим сетям, который может применяться при исследовании как отдельных документов, так и их классов.

6.4. Некоторые примеры

Для качественного проведения конкурентной разведки методами анализа текстов из Internet необходимо сформулировать цели, построить базы данных для наблюдений и проведения исследований, сформулировать требуемые запросы. Заметим, что не следует ограничиваться одной информационно-поисковой системой, даже для анализа такой информации, как Internet-ресурсы. Рекомендуем использовать лучшие глобальные и специальные информационно-поисковые системы, такие как Google, Yahoo! или Яndex (`http://www.yandex.ru`). Для специальных потребностей рекомендуется также использовать законодательные, адресно-номенклатурные, ценовые базы данных, доступные как в Internet, так и в локальных версиях.

Запросы к ИПС

Покажем, как формируются относящиеся к конкурентной проблематике запросы, на примере поисковых предписаний к информационно-поисковой системе InfoStream (`http://infostream.ua`).

Обычно поиск информации о компании или персоне всегда начинается с указания различных способов написания названия компании или полного имени персоны. Часто поиска в оперативных и ретроспективных данных по таким "примитивным" запросам вполне достаточно, однако задача усложняется, если необходимо исследовать состояние отдельной отрасли, отдельного региона или даже целой страны. В таких случаях в соответствии с проблематикой строятся запросы, которые затем итеративно уточняются.

В качестве примера назовем ряд проблем, поставим им в соответствие запросы и рассмотрим найденные фрагменты текстов, публикуемые различными источниками, которые затем можно будет использовать для построения собственных аналитических справок.

Ниже приведены уточняющие запросы, относящиеся к финансовому положению компаний:

```
Уставной~капитал~/2/долл
Уставной~фонд~/2/грн
Финансовое~положение
принадлежит~/2/акций
```

В первых двух запросах подразумевается нахождение документов, в которые входят фрагменты, содержащие словосочетания "уставной капитал" или "уставной фонд", с указанием значения в долларах или гривнах ("~/2/" на языке запросов означает расстояние в 2 или менее слов между выражениями).

В результате поиска получены тексты, содержащие такие фрагменты:

```
Альфа-частицы украинского циркония
...в 1998 году было создано совместное предприятие "ТВЭЛ-Энергия"
с уставным капиталом 1 млн. долл. СП была создана на троих: "ТВЭЛ",
"Энергоатом" и близкое к "Интерпайпу" украинско-андоррское предприятие АМП
(руководили им братья Петр и Сергей Устенко), которое позже фигурировало
в скандале вокруг неудачного приобретения 25%-го пакета харьковского
"Турбоатома". Позже 14,55% из этого пакета оказались в собственности
```

подконтрольного К. Григоришину белизского оффшора Parminter Group...
"Российский сайт ядерного нераспространения" 2004.07.20

За январь-июнь финансовый результат Приватбанка составил 144,7 млн. грн.
...За январь-июль текущего года финансовый результат КБ "ПриватБанк" (Днепропетровск) составил 144,7 млн. грн. Об этом УНИАН сообщили в банке. Как говорится в сообщении, на 1 июля 2004 года чистые активы ПриватБанка вырос до 14 млрд. 192 млн. грн. (на 1 января 2004 года - 9 млрд. 842 млн. грн.), собственный капитал - 1 млрд. 202 млн. грн., уставный фонд - 700 млн. грн...
УНИАН 2004.07.22

Куда делись деньги Parmalat
...Бонди не заявляет прямо, что банки участвовали в схемах Parmalat, но утверждает, что реальное финансовое положение компании можно было бы легко определить, сравнив информацию, предоставляемую компанией, и данные независимых аналитиков относительно ее облигационных займов. "Операторы финансового рынка знали о беспорядке, который царил в Parmalat", - говорится в отчете...
"Ведомости" 2004.07.26

ФГИУ объявил конкурс по продаже 93,07% ОАО "Криворожский железорудный комбинат"
...Правительство Украины 1 марта 2002 года передало 100% акций ОАО "Комсомольское рудоуправление" (Донецкая обл.), входившего в состав ГАК "Укррудпром", в управление ОАО "Мариупольский металлургический комбинат им. Ильича" сроком на пять лет с правом их дальнейшего выкупа. ОАО "ММК им. Ильича" - одно из трех крупнейших металлургических предприятий Украины. ЗАО "Ильич-Сталь" принадлежит более 90% акций комбината, остальные - физическим и юридическим лицам...
"Finance.com.ua" 2004.07.23

Информация о слияниях и приобретениях в той или иной сфере бизнеса, позволяющая следить за экспансией конкурентов в новые рыночные ниши, может быть получена в результате отработки таких уточняющих запросов:

приобрел/2/акций
приобрел~/2/пакет~акций *(допустимо, например,* "контрольный пакет акций"*)*
продал~/2/пакет~акций
(слияние~компаний) & (акций,активов)

Выполнение этих уточняющих запросов позволило получить документы, содержащие следующие фрагменты:

Гута-банк открылся
...Уже 12 июля "Гута" начала принимать заявления от вкладчиков на выдачу денег, 16 июля ВТБ приобрел 85,8% акций Гута-банка и получил большинство в совете директоров, а спустя неделю, в минувшую пятницу, Гута-банк возобновил работу...
"Ведомости" 2004.07.26

Покупатель "Укртелекома" получит в управление часть пакета акций ОАО, закрепленного в госсобственности
... Промышленный инвестор, который приобрел пакет акций "Укртелекома", имеет право получить в управление пакет акций компании в размере до половины пакета акций ОАО "Укртелеком", которые закреплены в государственной собственности. Условия передачи в управление

промышленному инвестору этого пакета акций определяются в соответствующем договоре. Такова одна из норм Положения о порядке подготовки и проведения открытых торгов по продаже пакета акций открытого акционерного общества "Укртелеком", утвержденного Приказом Фондом госимущества от 29.06.2004 номер 1256 (зарегистрирован в Минюсте 16 июля 2004 года под номером 893/9492)...
"Подробности" 2004.07.23

"Альфа-Эко" купила весомый аргумент на переговорах с Sun Interbrew
...На "Патре", однако, говорят, что сделка с акциями была. По словам руководителя пресс-службы предприятия Сергея Салыгина, топ-менеджмент завода "продал контрольный пакет акций структурам "Альфа-Эко". Сумма сделки не разглашается. По оценкам аналитиков, она могла составить $20-25 млн...
"Рынок продуктов питания" 2004.07.23

"Силовые машины" не объединились с ОМЗ
...Запланированное слияние компаний "Силовые машины" и ОМЗ не состоялось, пишут "Ведомости". По информации газеты, причиной этому стал пропуск владельцем "Силовых машин" - холдингом "Интеррос" - срока оплаты своей доли в ОМЗ акциями "Силовых машин"...
"ПОЛИТ.РУ" 2004.07.21

Для выявления публикаций об изменении финансового состояния и банкротствах можно использовать такие уточняющие запросы:

```
выпуск~/2/акций
(увеличить~уставной)&(фонд,капитал)
повысить~/1/долю~/1/акций
снизить~/1/долю~/1/акций
продать~/2/акций
объявить~/2/банкротство
```

Отработка указанных запросов позволила найти такие документы:

Нефтяная концентрация
...Л.Кучма также отдал поручение в максимально быстрые сроки провести общее собрание акционеров "Укрнафты", чтобы оно могло принять решение о повышении уставного фонда компании путем дополнительного выпуска акций, сохранив при этом за государством 50%+1 акция и передав в уставный фонд "Укрнафты" госпакеты акций "Укртатнафты" и НПК "Галичина"...
Газета "День" 2004.07.23

Страховая компания "Веста" планирует увеличить уставный фонд
...ГКЦБФР зарегистрировала новый уставный фонд страховой компании "Веста" в размере 6,5 млн. грн. Однако компания не собирается останавливаться на достигнутом. Еще до конца года планируется увеличение уставного фонда до 7 млн. грн. А в начале следующего года компания планирует поднять УФ до 11-12 млн. грн...
"UABanker" 2004.07.23

"Составляются новые списки"
...По словам аналитика банка "Зенит" Сергея Суверова, в этой ситуации о вкладчиках должен позаботиться главный регулирующий орган - Центробанк: например, дать банку "Диалог-Оптим" стабилизационный кредит или продать контрольный пакет акций кредитного учреждения, как это было сделано в случае с Гута-Банком. В пик банковского кризиса, когда с проблемами столкнулся даже Альфа-Банк, глава Центробанка

Сергей Игнатьев заявил, что проблем у банка нет и свои обязательства перед вкладчиками он выполнит...
"ГАЗЕТА" 2004.07.23

"ЮКОС" поднимает цены на нефть
...Судя по всему, продажи "Юганскнефтегаза" не избежать. Глава Федеральной службы по финансовым рынкам Олег Вьюгин даже рассказал, как собираются реализовывать на бирже основной актив "ЮКОСа": небольшими порциями. А это, как известно, главная "дойная корова" "ЮКОСа", добывающая для компании 60 процентов нефти. И цена одних только запасов "Юганскнефтегаза" почти в 9 раз превышает сумму налоговых претензий к компании. Потерять ее - значит лишиться львиной доли прибыли. Вслед за чем остается лишь объявить о банкротстве. По словам руководителей "ЮКОСа", это может произойти уже через три недели...
"Российская газета" 2004.07.23

Методы контент-мониторинга

Методы контент-мониторинга — это адаптация классических методов контент-анализа к условиям динамических информационных массивов, например потоков информации из Internet.

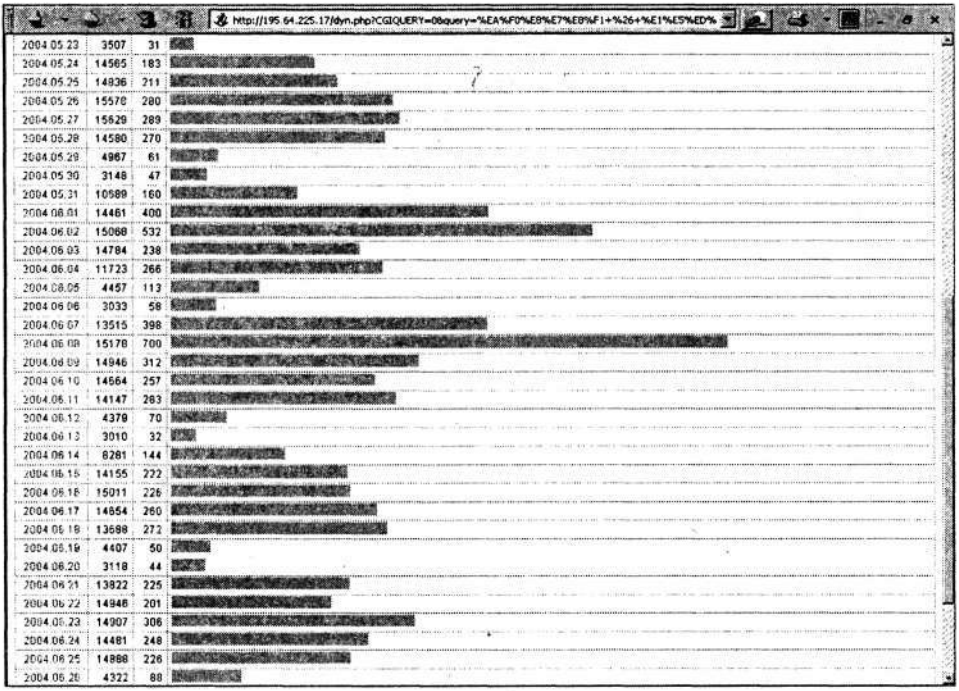
1. Типичная задача контент-мониторинга — построение диаграмм динамики появления понятий по времени. Рассмотрим, как в системе InfoStream отслеживались кризисные явления на рынке нефтепродуктов Украины в июне 2004 года (рис. 6.1). Для этого был составлен запрос "кризис & бензин & Украина", который был введен через Web-интерфейс системы.

   Из приведенной диаграммы видно, что массовое появление сообщений о кризисных явлениях произошло 25-го мая (в то время как сами цены на бензин резко возросли лишь 1-го июня) и завершилось 24-25 июня, в то время как цены стабилизировались 15-го. Безусловно, оперативное получение такого типа данных должно было помочь аналитикам при построении краткосрочных прогнозов.

2. Аналогично можно проводить мониторинг финансового рынка. К примеру, простой запрос "падение ~ курса ~ доллара", относящийся к фрагменту информационного потока, классифицированного рубрикой "Банки", за два периода выдал диаграммы, свидетельствующие о динамике падения курса доллара США (рис. 6.2). Как видим, при наличии случайного появления релевантных запросу документов и до 20 сентября, лишь начиная с 21 сентября 2004 года процесс носит явно выраженный характер (21 документ по сравнению с 6–8 в обычные рабочие дни).

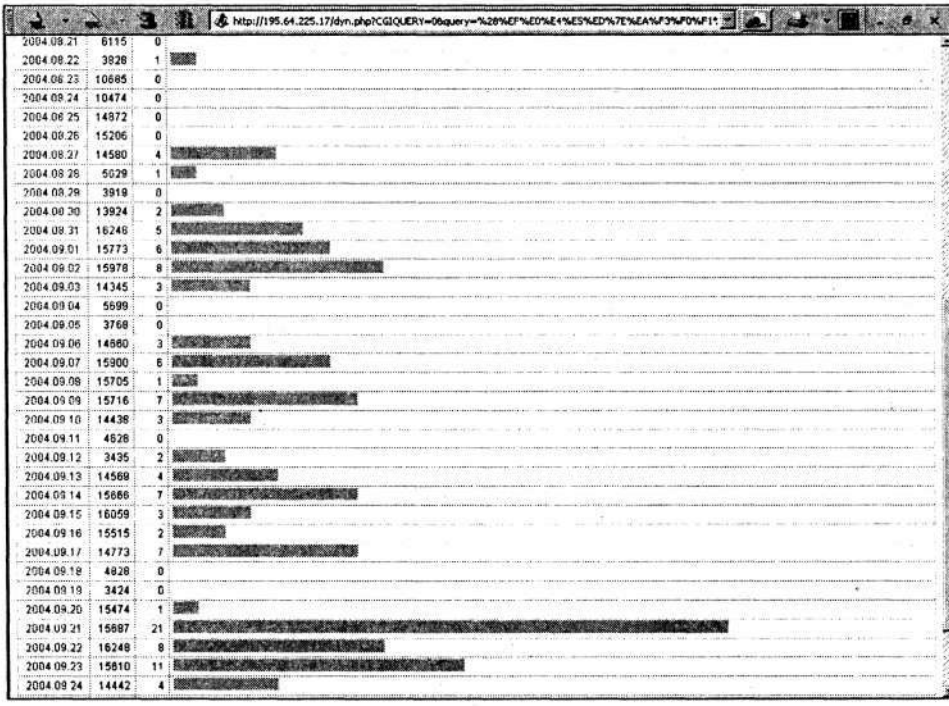
3. На примере рынка нефтепродуктов рассмотрим, как из массивов текстовой информации из Internet могут быть выявлены сюжетные цепочки из документов, содержащих максимальное количество ценовой информации по данному рынку (рис. 6.3).

   Для получения основных сюжетов, относящихся к рынку нефтепродуктов, можно ввести запрос "(нефтепродукты | бензин) & цены", уточнив его специальными признаками "numb.medium | numb.large", означающими в системе InfoStream средний или высокий уровень заполненности документов цифровой информацией (рис. 6.4).



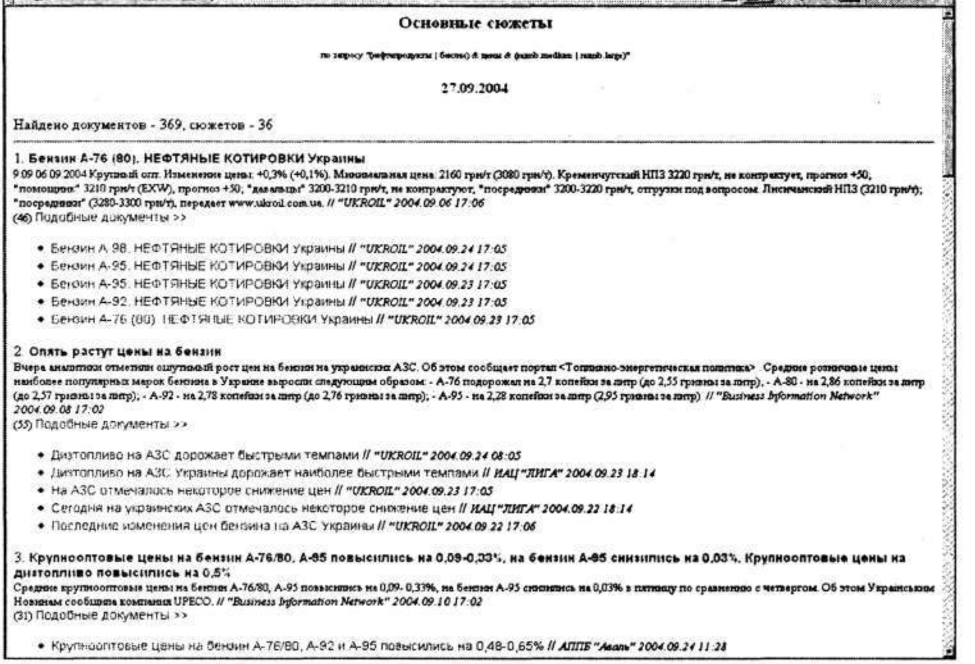
2004.05.23	3507	31
2004.05.24	14565	183
2004.05.25	14836	211
2004.05.26	15570	280
2004.05.27	15629	289
2004.05.28	14580	270
2004.05.29	4967	61
2004.05.30	3148	47
2004.05.31	10589	160
2004.06.01	14461	400
2004.06.02	15068	532
2004.06.03	14784	238
2004.06.04	11723	266
2004.06.05	4457	113
2004.06.06	3033	58
2004.06.07	13515	398
2004.06.08	15178	700
2004.06.09	14946	312
2004.06.10	14564	257
2004.06.11	14147	283
2004.06.12	4378	70
2004.06.13	3010	32
2004.06.14	8281	144
2004.06.15	14155	222
2004.06.16	15011	226
2004.06.17	14654	260
2004.06.18	13698	272
2004.06.19	4407	50
2004.06.20	3118	44
2004.06.21	13822	225
2004.06.22	14948	201
2004.06.23	14907	306
2004.06.24	14481	248
2004.06.25	14868	226
2004.06.26	4322	88

Рис. 6.1. *Динамика появления понятия*



2004.08.21	6115	0
2004.08.22	3828	1
2004.08.23	10865	0
2004.08.24	10474	0
2004.08.25	14872	0
2004.08.26	15206	0
2004.08.27	14580	4
2004.08.28	5029	1
2004.08.29	3918	0
2004.08.30	13924	2
2004.08.31	16248	5
2004.09.01	15773	6
2004.09.02	15978	8
2004.09.03	14345	3
2004.09.04	5699	0
2004.09.05	3768	0
2004.09.06	14660	3
2004.09.07	15900	6
2004.09.08	15705	1
2004.09.09	15716	7
2004.09.10	14438	3
2004.09.11	4628	0
2004.09.12	3435	2
2004.09.13	14569	4
2004.09.14	15666	7
2004.09.15	16059	3
2004.09.16	15515	2
2004.09.17	14773	7
2004.09.18	4828	0
2004.09.19	3424	0
2004.09.20	15474	1
2004.09.21	15687	21
2004.09.22	16248	8
2004.09.23	15810	11
2004.09.24	14442	4

Рис. 6.2. *"Падение курса доллара" в динамике*



http://stream2.visti.net/dig.php?CGIQUERY=0&dtb=138&lng=RUS&query=%28%ED%E5%F4%F2%E5%E

Основные сюжеты

по запросу "(нефтепродукты | бензин) & цены & (rank medium | rank large)"

27.09.2004

Найдено документов - 369, сюжетов - 36

1. Бензин А-76 (80). НЕФТЯНЫЕ КОТИРОВКИ Украины

9.09 06.09.2004 Крупный опт. Изменение цены: +0,3% (+0,1%). Минимальная цена: 2160 грн/т (3080 грн/т). Кременчугский НПЗ 3220 грн/т, не контрактует, прогноз +50, "помощник" 3210 грн/т (EXW), прогноз +50; "давальцы" 3200-3210 грн/т, не контрактуют, "посредники" 3200-3220 грн/т, отгрузки под вопросом. Лисичанский НПЗ (3210 грн/т); "посредники" (3280-3300 грн/т), передает www.ukroil.com.ua. // *"UKROIL" 2004.09.06 17:06*

(46) Подобные документы >>

- Бензин А 98. НЕФТЯНЫЕ КОТИРОВКИ Украины // *"UKROIL" 2004.09.24 17:05*
- Бензин А-95. НЕФТЯНЫЕ КОТИРОВКИ Украины // *"UKROIL" 2004.09.24 17:05*
- Бензин А-95. НЕФТЯНЫЕ КОТИРОВКИ Украины // *"UKROIL" 2004.09.23 17:05*
- Бензин А-92. НЕФТЯНЫЕ КОТИРОВКИ Украины // *"UKROIL" 2004.09.23 17:05*
- Бензин А-76 (80). НЕФТЯНЫЕ КОТИРОВКИ Украины // *"UKROIL" 2004.09.23 17:05*

2. Опять растут цены на бензин

Вчера аналитики отметили ощутимый рост цен на бензин на украинских АЗС. Об этом сообщает портал <Топливно-энергетическая политика>. Средние розничные цены наиболее популярных марок бензина в Украине выросли следующим образом: - А-76 подорожал на 2,7 копейки за литр (до 2,55 гривны за литр), - А-80 - на 2,86 копейки за литр (до 2,57 гривны за литр); - А-92 - на 2,78 копейки за литр (до 2,76 гривны за литр); - А-95 - на 2,28 копейки за литр (2,95 гривны за литр) // *"Business Information Network" 2004.09.08 17:02*

(55) Подобные документы >>

- Дизтопливо на АЗС дорожает быстрыми темпами // *"UKROIL" 2004.09.24 08:05*
- Дизтопливо на АЗС Украины дорожает наиболее быстрыми темпами // *ИАЦ "ЛИГА" 2004.09.23 18:14*
- На АЗС отмечалось некоторое снижение цен // *"UKROIL" 2004.09.23 17:05*
- Сегодня на украинских АЗС отмечалось некоторое снижение цен // *ИАЦ "ЛИГА" 2004.09.22 18:14*
- Последние изменения цен бензина на АЗС Украины // *"UKROIL" 2004.09.22 17:06*

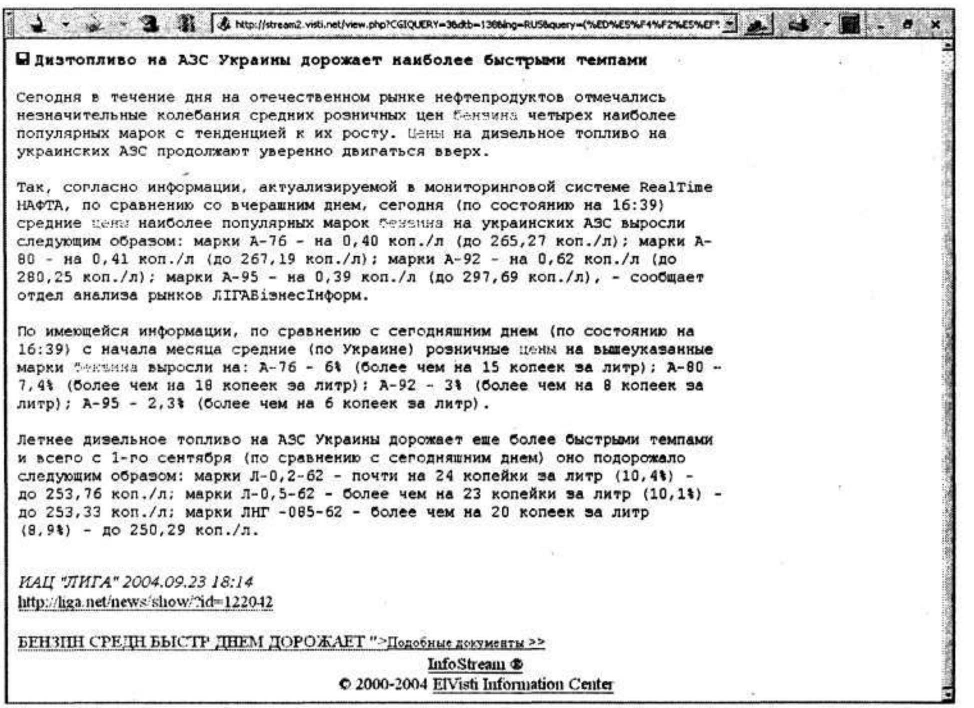
3. Крупнооптовые цены на бензин А-76/80, А-95 повысились на 0,09-0,33%, на бензин А-95 снизились на 0,03%. Крупнооптовые цены на дизтопливо повысились на 0,5%

Средние крупнооптовые цены на бензин А-76/80, А-95 повысились на 0,09- 0,33%, на бензин А-95 снизились на 0,03% в пятницу по сравнению с четвергом. Об этом Украинським Новинам сообщила компания UPECO. // *"Business Information Network" 2004.09.10 17:02*

(31) Подобные документы >>

- Крупнооптовые цены на бензин А-76/80, А-92 и А-95 повысились на 0,48-0,65% // *АППБ "Аваль" 2004.09.24 11:28*

Рис. 6.3. *Цепочка основных сюжетов*



http://stream2.visti.net/view.php?CGIQUERY=3&dtb=138&lng=RUS&query=(%ED%E5%F4%F2%E5%EF

Дизтопливо на АЗС Украины дорожает наиболее быстрыми темпами

Сегодня в течение дня на отечественном рынке нефтепродуктов отмечались незначительные колебания средних розничных цен бензина четырех наиболее популярных марок с тенденцией к их росту. Цены на дизельное топливо на украинских АЗС продолжают уверенно двигаться вверх.

Так, согласно информации, актуализируемой в мониторинговой системе RealTime НАФТА, по сравнению со вчерашним днем, сегодня (по состоянию на 16:39) средние цены наиболее популярных марок бензина на украинских АЗС выросли следующим образом: марки А-76 - на 0,40 коп./л (до 265,27 коп./л); марки А-80 - на 0,41 коп./л (до 267,19 коп./л); марки А-92 - на 0,62 коп./л (до 280,25 коп./л); марки А-95 - на 0,39 коп./л (до 297,69 коп./л), - сообщает отдел анализа рынков ЛIГАБiзнесIнформ.

По имеющейся информации, по сравнению с сегодняшним днем (по состоянию на 16:39) с начала месяца средние (по Украине) розничные цены на вышеуказанные марки бензина выросли на: А-76 - 6% (более чем на 15 копеек за литр); А-80 - 7,4% (более чем на 18 копеек за литр); А-92 - 3% (более чем на 8 копеек за литр); А-95 - 2,3% (более чем на 6 копеек за литр).

Летнее дизельное топливо на АЗС Украины дорожает еще более быстрыми темпами и всего с 1-го сентября (по сравнению с сегодняшним днем) оно подорожало следующим образом: марки Л-0,2-62 - почти на 24 копейки за литр (10,4%) - до 253,76 коп./л; марки Л-0,5-62 - более чем на 23 копейки за литр (10,1%) - до 253,33 коп./л; марки ЛНГ -085-62 - более чем на 20 копеек за литр (8,9%) - до 250,29 коп./л.

ИАЦ "ЛИГА" 2004.09.23 18:14
http://liga.net/news/show/?id=122042

БЕНЗИН СРЕДН БЫСТР ДНЕМ ДОРОЖАЕТ ">Подобные документы >>

InfoStream ®
© 2000-2004 ElVisti Information Center

Рис. 6.4. *Документ с ценовой информацией*

После этого достаточно перейти в режим "Сюжеты" и проанализировать документы и ссылки, выданные системой. Режим "Сюжеты" не только предусматривает обработку запросов в рамках булевой модели, но и добавляет учет весовых критериев, выдавая лишь наиболее весомые цепочки документов. Поэтому обеспечивается достаточно высокий уровень соответствия выдаваемых документов и потребности, выраженной запросом.

6.5. Конкурентная разведка и "скрытый" Web

Как уже было замечено ранее, необходимой (в том числе и для конкурентной разведки) информации в Internet значительно больше, чем ее охватывают универсальные поисковые машины. Предполагается, что, в отличие от "познаваемой" части Internet, "скрытая" часть информации оказывается в сотни раз более объемной.

К разряду "скрытого" Web, например, относится и крупнейшая в мире полнотекстовая он-лайновая информационная система Lexis-Nexis, которая содержит более 2 млрд документов с глубоким архивом до 30 лет по бизнес-информации и более 200 лет по юридической информации. Каждую неделю в архивы добавляется еще 14 млн документов. В отличие от неструктурированных массивов "поверхностного" Web, пользователи Lexis-Nexis могут использовать мощные инструменты поиска для получения достоверной и классифицированной информации.

Приведем еще один пример зарубежной базы данных из "скрытого" Web. Корпорация ChoicePoint недавно предоставила сервис Auto TrackXP, вошедший в список двадцати крупнейших "скрытых" сайтов мира (по рейтингу BrightPlanet). Auto TrackXP представляет собой базу данных объемом 30 Тбайт, охватывающую практически все аспекты гражданской жизни США. База данных системы Auto TrackXP содержит информацию практически о каждом гражданине США. Сайт `TestProfiles.com` — часть службы ChoicePoint Online — содержит личные характеристики и сведения о компетентности граждан США. Например, чтобы определить, не завладел ли человек чужими документами, на основе системы организован платный сервис ProCheck, позволяющий сопоставить информацию из различных источников и государственных каталогов.

Для частных любителей составления "досье" ChoicePoint предлагает более скромный, но не менее любопытный набор сервисов (`www.choicetrust.com`). Подозрительные пациенты с помощью Doctor Check имеют возможность самостоятельно выбрать или проверить квалификацию врачей 40 различных специализаций. Отчет, получаемый с помощью системы, может, например, служить для страховой компании поводом в отказе выдачи полиса.

Система широко используется как легальный ресурс для задач конкурентной разведки. Вместе с тем, сегодня американцы повсеместно выражают возмущение, обнаруживая существование подобных сервисов, усматривая в этом нарушение своих гражданских прав.

6.6. Перспективы систем конкурентной разведки

Актуальность конкурентной разведки в последнее время значительно возросла. Это связано с такими процессами, как глобализация экономики, а следовательно и конкуренции, виртуализация экономики, развитие информационных технологий.

Широкому внедрению систем компьютерной конкурентной разведки способствуют и законодательные акты многих стран мира. Так, в США еще

в 1996 году был принят Закон о свободе информации, который обязал федеральные ведомства обеспечить гражданам свободный доступ ко всей своей информации. Ограничения касаются лишь материалов, имеющих отношение к национальной обороне, личных и финансовых документов, а также документов правоохранительных органов. Отказ в доступе к информации можно обжаловать в суде. Информация должна быть представлена в десятидневный срок, а споры разрешаются в течение 20 дней.

Об актуальности конкурентной разведки на основе Internet-ресурсов говорят многочисленные публикации, тренинги, конференции. Например, в ходе состоявшейся в 2004 году VIII международной выставки-конгресса "Высокие технологии. Инновации. Инвестиции", проходившей в Санкт-Петербурге, был представлен проект системы углубленного поиска и анализа данных в Internet T2 Business Analysis Console (T2 BANC), разработки российской компании ГиперМетод. Система T2 BANC является специализированным механизмом изучения конкурентного окружения и прогнозирования развития ситуаций на базе информации таких открытых ресурсов Internet, как поисковые серверы, отраслевые каталоги, коммерческие базы данных, сайты компаний, новостные порталы и т.д. Система T2 BANC обеспечивает формирование структурированного запроса, извлечение информации из различных баз данных, анализ и нахождение взаимосвязей. Результатом работы системы является предоставление клиенту систематизированных сведений, интересных как организациям из государственного сектора, так и инвестиционным фондам, коммерческим и консалтинговым компаниям.

Сегодня задачи конкурентной разведки стимулируют развитие систем управления знаниями, глубинного анализа данных и текстов. С другой стороны, наиболее развитые из этих систем в явном виде содержат аналитические блоки, специально ориентированные на задачи конкурентной разведки. Поэтому у пользователей имеется широкий выбор средств автоматизации аналитической деятельности, причем уровни функциональности таких систем могут быть очень разнообразными — от простых информационно-поисковых программ, необходимых на этапе становления систем конкурентной разведки, до дорогих и ресурсоемких систем управления знаниями и глубинного анализа данных и текстов [32].

Среди самых развитых систем управления знаниями, применяемых для решения задач конкурентной разведки, нельзя не назвать систему Hummingbird Enterprise™ канадской компании Hummingbird (рис. 6.5). Среди множества компонентов системы можно выделить Hummingbird Portal — платформу, позволяющую интегрировать информацию из информационного хранилища и приложения в едином Web-интерфейсе. Эта платформа, как и ранее названный портал IBM Lotus, является полнофункциональным порталом знаний.

Еще одна флагманская платформа для конкурентной разведки — это система американской корпорации Documentum, предназначенная для управления неструктурированной информацией, хранящейся в виде файлов различных форматов (рис. 6.6). Система Documentum (EMC Platform) основана на трехуровневой архитектуре, включающей хранилище содержания — репозитарий, службу управления содержанием — контент-сервер и клиентские приложения для работы с неструктурированными данными, Web-контентом, XML-документами, мультимедиа-данными. Репозитарий системы обеспечивает как безопасность, так и открытость хранения контента, позволяет объединять корпоративные данные в единую корпоративную информационную среду.

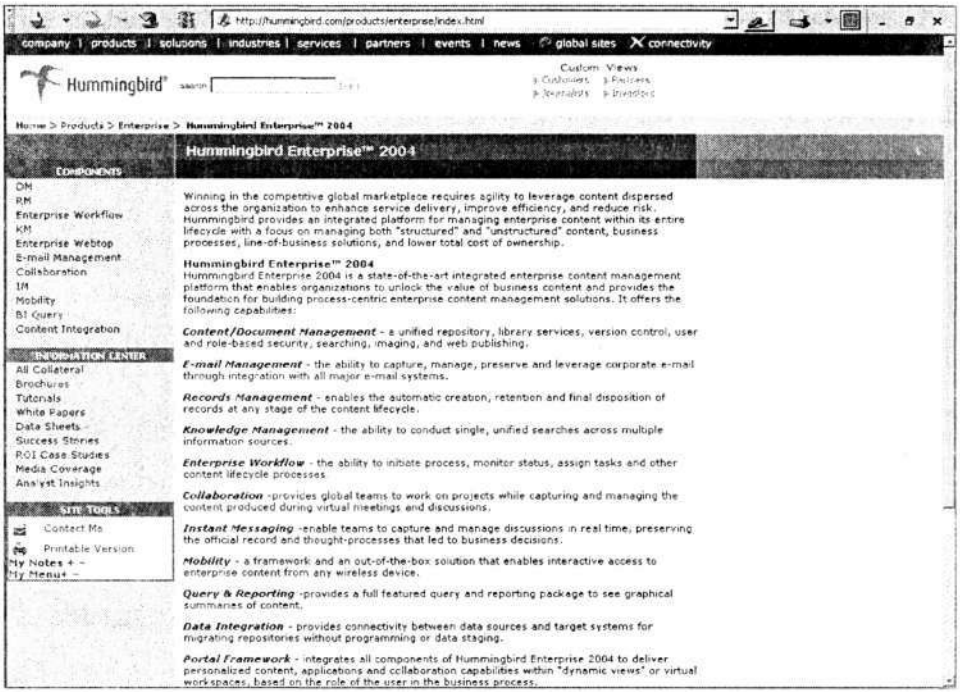


Рис. 6.5. Hummingbird Enterprise™

Рис. 6.6. Система Documentum — основа для построения портала знаний

Для решения информационно-аналитических задач в настоящее время также широко используется система Cognos Business Intelligence корпорации Cognos (рис. 6.7). Решение Cognos BI базируется на идеологии OLAP. Одна из особенностей системы — это ее возможность интеграции с компонентами других информационных систем, в том числе необходимых для проведения бизнес-разведки систем финансово-экономического планирования и управления клиентской базой. В этом случае обеспечиваются широкие возможности сбора и консолидации данных из внутренних и внешних источников. Говоря о системе Cognos как о лидирующей в области Business Intelligence, следует отметить, что под этим термином понимается набор инструментальных средств анализа данных и их визуализации, в отличие от Competitive Intelligence (конкурентной разведки), которая, как видим, является очень широким направлением информационной деятельности.

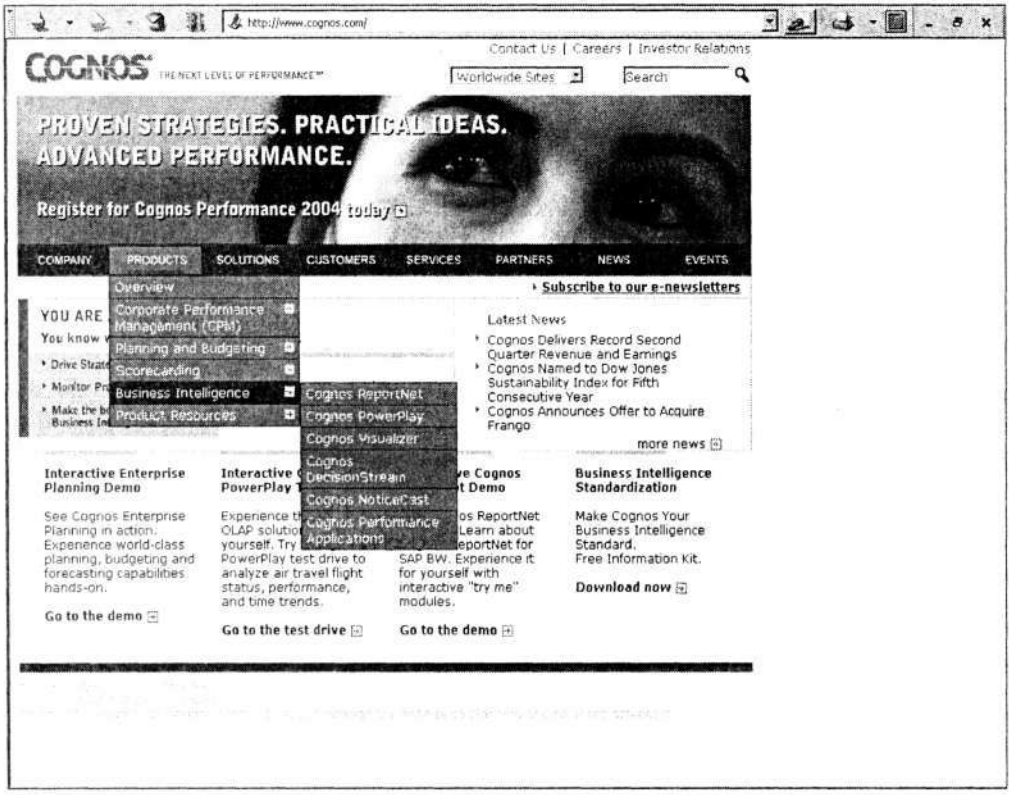


Рис. 6.7. OLAP-система Cognos

Не все из названных систем являются доступными, ввиду их стоимостных характеристик. Вместе с тем, отдельные задачи конкурентной разведки могут быть частично решены общедоступными средствами. Использование новых подходов, а также открытых, доступных и относительно недорогих информационных источников позволяет уже сегодня эффективно поддерживать принятие решений не только в стратегических областях.

Глава 7

Закономерности, присущие информационным системам

Практика построения и использования современных информационных систем еще раз подтверждает некоторые закономерности из области системотехники и статистического анализа. Причем учет этих закономерностей практически гарантирует успех как при поиске информации, так и при построении самих информационных систем.

7.1. Правило Парето

Швейцарский ученый Вильфредо Парето (Vilfredo Federico Damaso Pareto) в свое время предпринял попытку математически обосновать взаимозависимость различных экономических и социальных факторов, к которым сегодня можно было бы отнести и Internet.

К основным трудам В. Парето принято относить двухтомный "Курс политической экономии" [61] (1897), "Учение политической экономии" (1906) и "Трактат по общей социологии" (1916).

Анализируя общественные процессы, он рассматривал социальную среду как пирамиду, наверху которой находятся немногие люди, составляющие элиту. В результате кропотливых исследований ученый сформулировал математическую зависимость между величиной дохода и количеством получающих его лиц. В 1906 году Парето установил, что 80% земли в Италии принадлежит лишь 20% ее жителей.

В результате обобщения обширного статистического материала Парето пришел к выводу, что параметры этого полученного им распределения примерно одинаковы и не различаются принципиально в разных странах и в разное время. "Кривая распределения доходов отличается замечательной устойчивостью, она меняется незначительно, хотя обстоятельства времени и места сильно преображаются", — писал Парето в "Социалистических системах". Распределение доходов по Парето описывается уравнением $N = A/X^{p+1}$, где X — величина дохода, N — численность людей с доходом, равным или выше X, A и p — коэффициенты уравнения. В математической статистике это распределение получило имя распределения Парето, при этом имеют место естественные ограничения на коэффициенты: $X \geq 1$, $p > 0$.

Распределение Парето обладает свойством *устойчивости* (stable distribution), т.е. сумма двух случайных переменных, имеющих распределение Парето, также будет иметь это распределение.

Ученый показал, что замеченное им правило применимо и в многих других областях, и сформулировал правило, называемое "Законом Парето", или

"Принципом 80/20". На практике полезна такая трактовка правила Парето: первые 20% усилий дают первые 80% желаемого результата. Необходимо только найти требуемые ресурсы и реализовать их. Например, при информационном поиске достаточно определить 20% необходимых ключевых слов, что позволит найти 80% требуемых документов, а затем расширить поиск или воспользоваться опцией "найти похожие" для полного решения задачи. Эта важная закономерность сегодня формулируется по-разному:

- 80% функциональности приходится на 20% модулей;
- 80% работы выполняет 20% людей;
- 80% посещений Web-сайта приходится лишь на 20% его Web-страниц;
- 80% пива выпивает 20% людей.

Применительно к современной экономике приведенный выше закон Парето сегодня трактуется так: "20% потребителей покупают 80% товаров определенной марки, представляя обобщенную группу целевых потребителей данного товара". Поэтому компании при продвижении своих продуктов и услуг на рынок проводят маркетинговые мероприятия, ориентируясь именно на эти 20% клиентов ("стрельба по целям"), а не на весь рынок в целом ("стрельба по площадям"), поскольку такая стратегия рыночной деятельности более эффективна.

Джозеф Джуран отметил универсальность применения принципа Парето к любой группе причин, вызывающих те или иные последствия, благодаря чему большая часть последствий вызывается малым количеством причин. Анализ Парето ранжирует отдельные области по их значимости или важности, позволяя выявить и устранить в первую очередь те причины, которые вызывают наибольшее количество проблем (несоответствий).

При реализации систем массового обслуживания, в том числе и поисковых систем, необходимо учитывать то, что наиболее сложными функциональными возможностями системы, на реализацию которых уйдет 80% и более трудозатрат, в конечном счете будут пользоваться не более чем 20% ее пользователей. В качестве яркой иллюстрации этого правила можно привести реальный пример из практики Internet: анализ запросов пользователей к поисковой системе, язык запросов которой обладает массой синтаксических и семантических характеристик, показал, что свыше 80% этих запросов состоят не более чем из трех слов.

В частности, для информационно-поисковых систем, если их создатели хотят ориентироваться на широкий круг пользователей Internet (т.е. непрофессионалов), достаточно реализовать относительно узкий спектр самых важных поисковых функций, которые удовлетворят 80% будущих пользователей. Для того чтобы удовлетворить остальные 20% пользователей (как раз сюда попадают профессионалы), в подобной системе потребуется реализовать различные усложнения поиска, например опцию "расширенный поиск". Профессиональный поиск требует серьезной проработки поискового аппарата — такой, которая реализована в сложных системах анализа контента Сети. Подобных систем во всем мире всего несколько десятков.

Цена одного процента результатов

Определенный интерес представляет рассмотрение пошагового применения принципа Парето. Так, если на первом шаге, прилагая 20% усилий, можно получить 80% результата, то на втором шаге, применяя 20% (от оставшихся 80%)

усилий, можно достичь 80% от оставшихся на первом шаге 20% результатов, т.е. 16%. Это означает, что за первых два шага, применив 20 + 16 = 36% усилий, можно получить 96% результатов!

Очевидно, что на N-м шаге, применив в сумме $(1 - 0{,}8N) \times 100\%$ усилий, можно получить $(1 - 0{,}2N) \times 100\%$ результатов. Следовательно, на третьем шаге, применяя еще 20% от оставшихся 64% или потратив в сумме менее 50% усилий, можно получить более 99% результатов.

Если предположить, что система имеет 99% необходимых возможностей и ее создали за 10 человеко-лет, то на практике для доведения функциональности системы до уровня 100% потребуется еще не менее 10 человеко-лет. Таким образом, цена последнего процента равна цене всей системы, работающей с 99% функциональности!

Это — очевидное следствие закона Парето в интерпретации "причина-следствие". Повышение до 100% функциональности системы, работающей на уровне 99% предельных возможностей (следствие), потребует удвоения усилий (причины). Конечно же, это соотношение достаточно приблизительно, но основная тенденция прекрасно видна на диаграмме или графике функции распределения Парето.

Возможно, приведенное выше математическое обоснование этой закономерности не достаточно строго (хотя и весьма наглядно), однако "эффект одного процента" на практике встречается повсеместно.

Например, чтобы получить вполне надежные (свыше 99% полноты) результаты при поиске информации в Сети, достаточно найти 50% необходимых ключевых слов. С другой стороны, можно рассмотреть процесс создания информационно-поисковых систем в Internet. Казалось бы, вновь появившаяся система вот-вот должна превзойти такие брэнды, как Yahoo!, AltaVista или Google, и осталось реализовать совсем немного, всего несколько процентов ее функциональности, однако можно с уверенностью прогнозировать, что этого, скорее всего, не произойдет — ведь понадобятся дополнительные капиталовложения, превосходящие средства, уже вложенные в создание "рабочей модели" новой системы. Конечно же, возможны и исключения. Например, в 2001 году стремительно вышла на первые позиции рынка европейская система поиска Alltheweb, ставшая одним из лидеров по объемам проиндексированных баз данных (2,1 млрд документов; больше тогда было только у Google — 3 млрд).

7.2. О переходе количества в качество

Если система достигла 99% своей идеальной функциональности, то дальнейшие попытки ее совершенствования ведут, в лучшем случае, к повышению качества сопровождения уже реализованных функций. Поэтому, если построить график, отмечая по оси абсцисс затраченные на развитие системы ресурсы, а по оси ординат — уровень ее функциональности, то график будет иметь вид кривой, у которой вначале наблюдается резкий подъем, а затем она быстро стабилизируется на уровне чуть ниже 100% (еще раз можно обратиться к распределению Парето). В реальной жизни бывают случаи, когда после длительного процесса стабилизации происходит резкий взлет этой кривой выше уровня 100%, т.е. график принимает вид зигзага. С чем же может быть связан такой подъем, когда функциональность резко превышает "идеальную" 100-процентную? Ответ очевиден —

этот феномен связывается с появлением новых подходов и взглядов на ставшие уже традиционными устоявшиеся процессы.

Реализация новых подходов приводит к появлению новой, даже не предполагавшейся ранее функциональности. В качестве примера этой закономерности можно привести развитие сети Internet, которая до начала 90-х годов прошлого века рассматривалась, прежде всего, как компьютерная сеть передачи данных, а уж затем как хранилище информационных ресурсов. Несмотря на то что существовали такие информационные службы, как Usenet, FTP и Gopher, до 90-х годов Сеть решала свои главные задачи, просто обеспечивая электронную связь между научными, общественными, государственными организациями и частными лицами. К этому времени Сеть существовала уже свыше 15 лет и фактически стабилизировалась в своем развитии. Феномен появления и развития Web-технологий привел к тому, что за следующие 10 лет Internet стала крупнейшим информационным ресурсом в мире, число абонентов которой превысило миллиард человек!

7.3. Закон Зипфа

Интеллектуализация информационных систем базируется на мощном математическом фундаменте, обеспечивающем "понимание" текстов компьютерными программами. В науке уже давно известны закономерности, свойственные всем текстам, с учетом которых (явным или неявным образом) были построены многие современные информационно-поисковые системы, а также системы автоматической классификации и глубинного анализа текстов (Text Mining).

К таким закономерностям, наряду с правилом Парето, прежде всего, следует отнести закон, который корреспондируется с уже упомянутым правилом. При статистическом описании распределения слов по частоте их употребления в тексте (как, впрочем, и в документальных потоках) используются так называемые ранговые распределения. (Ранг — это, например, порядковый номер слова в списке, где все слова упорядочены по возрастанию относительных частот.)

В 1949 году профессор филологии из Гарварда Джордж Зипф (George K. Zipf) [74] собрал достаточный статистический материал и экспериментально показал, что распределение слов естественного языка подчиняется закону, который можно сформулировать следующим образом. Если к какому-либо достаточно большому тексту составить список всех используемых в нем слов, а затем проранжировать эти слова — расположить их в порядке убывания частоты вхождения в данном тексте и пронумеровать в возрастающем порядке, — то для любого слова произведение его порядкового номера в этом списке (ранга) и частоты его вхождения в тексте будет величиной постоянной. Ученый следующим образом описал обнаруженные им закономерности распределения слов в текстах на английском языке.

- Небольшое количество слов, таких как "the" или "and", имеют очень высокий ранг (левый "рог" диаграммы).
- Среднее количество слов имеет средний ранг (средняя часть диаграммы).
- Большое количество слов имеет очень низкий ранг (правый "рог" диаграммы).

Таким образом, $f \times r = c$, где f — частота вхождения слова в тексте, r — ранг (порядковый номер) слова в списке, c — эмпирическая постоянная величина. Эту закономерность зависимости частоты от ранга называют первым законом

Зипфа. Таким образом, было установлено, что зависимость количества слов с данной частотой встречаемости в документе от частоты описывается гиперболой с параметрами, постоянными для *всех текстов* в пределах одного языка. Значение константы Зипфа в разных языках различно, но внутри одной языковой группы оно остается неизменным. Так, для английских текстов константа Зипфа равна приблизительно 0,1, а для русского и украинского языков — приблизительно 0,06–0,07 [34]. Это означает, что самое популярное слово в английском языке (the) употребляется в 10 раз чаще, чем слово, стоящее на десятом месте, в 100 раз чаще, чем сотое, и в 1000 раз чаще, чем тысячное.

Новые статистические методы обработки текстов, аналогичные тем, которые применил Зипф, активно исследуются и сегодня. Так, недавно в статье журнала Physical Review Letters ученые из римского университета La Sapienza предложили оригинальный метод автоматического определения авторов литературных произведений с помощью свободно распространяемой программы сжатия данных. Итальянские ученые обнаружили скрытые возможности для анализа строк данных в обычной программе Gzip, которая сжимает файлы, в том числе и текстовые, путем поиска повторяющихся фрагментов. Находя и распознавая в тексте определенные комбинации символов, программа сжатия классифицирует их и уменьшает размер файла, включая в архивный файл лишь основные компоновочные блоки, "кирпичики" данных, из которых состоит исходный текст, и инструкции, следуя которым можно заново его "собрать".

Эмануэль Кальоти (Emanuele Caglioti), адъюнкт-профессор математики и один из авторов отчета, утверждает, что процесс сжатия данных, используемый программой, также может играть ключевую роль в распознавании незнакомых текстовых файлов. Как пишет Кальоти, когда программа-архиватор сжимает данные, "она узнает кое-что о файле". В частности, она определяет файловую энтропию — минимальное число битов, необходимых для сжатия файла. "Если вы сжимаете файл, скажем, состоящий из английского текста, то, пока Gzip его читает, она изучает статистику английского языка", — объяснил г-н Кальони. Если добавить еще один файл на английском, то это существенно не изменит размера файла, так как базовый компонент — его энтропия — уже известен. Но если второй файл будет, к примеру, на итальянском языке, то процесс придется начать заново, и программа определит новую энтропию. Для обработки файла на итальянском потребуется больше места, так как это другой язык.

Как выяснил г-н Кальоти и его сотрудники, тот же принцип и процесс можно использовать для распознавания автора текста. В своем исследовании ученые использовали 90 текстов 11 итальянских авторов, и в 93% случаев программа правильно классифицировала маленькие отрывки текстов по авторам. Оказывается, можно смело говорить о том, что процесс сжатия данных вполне допустимо использовать и в других целях. "Кроме распознавания текстов, его можно использовать для сравнения Web-страниц и нахождения среди них одинаковых", — сказал он.

В соответствии с алгоритмами сжатия и законами Зипфа, слова с высоким рангом хорошо сжимаются, а с низким (редкие) — наоборот, плохо. По-видимому, каждой зоне рангового распределения Зипфа соответствует свой коэффициент сжатия. При этом состав и разнообразие лексики каждого конкретного автора достаточно своеобразны и хорошо проявляются на значительных объемах информации.

Зипф сформулировал еще одну закономерность, близкую по смыслу к своему первому закону. Он определил, что частота и количество слов, входящих в текст

с этой частотой, также взаимосвязаны. Если построить диаграмму, отложив по одной оси частоту вхождения слова, а по другой — количество слов, входящих в текст с данной частотой, то получившаяся кривая будет сохранять свои параметры для всех текстов в пределах одного языка. Иными словами, на каком бы языке текст ни был написан, форма кривой Зипфа останется неизменной — отличаться могут лишь коэффициенты. Эта закономерность получила название второго закона Зипфа — закон "количество–частота".

Известный математик Беноит Мандлеброт (Benoit Mandlebrot) предложил теоретическое обоснование закона Зипфа, полагая, что можно сравнивать язык текста с кодированием. Исходя из требований минимальной стоимости сообщений, Мандельброт математическим путем пришел к аналогичной первому закону Зипфа зависимости $f \times r^e = c$, где e — близкая к единице переменная величина, которая может изменяться в зависимости от свойств текста и языка. Постоянство коэффициента e сохраняется только в центральной зоне диаграммы распределения. Участок распределения с $e = \text{const}$ называется центральной зоной рангового распределения. По относительной величине той или иной зоны на подобном графике можно судить о характеристиках рассматриваемой в тексте области знаний.

Очевидно, что наиболее значимые слова лежат в средней части диаграммы. График с обширной средней частью (центральной зоной распределения) относится к достаточно широкой области знаний. Центральная зона содержит термины, наиболее характерные для данной области знаний, которые в совокупности выражают ее специфичность, отличие от других наук, "охватывают ее основное содержание". Иначе говоря, основа лексики конкретной области знаний сосредоточена в центральной зоне рангового распределения. Это правило может успешно использоваться на практике для выделения значащих слов в тексте. От того, как задан диапазон значимых слов, зависят результаты текстового поиска.

Малая величина центральной зоны распределения свидетельствует об оригинальности области знаний, к которой относится построенное ранговое распределение и т.д. Зона левой части диаграммы распределения содержит наиболее общеупотребительные термины. Слова из левой области в основном оказываются предлогами, местоимениями, в английском языке — артиклями и т.п. В зоне усечения (правой части диаграммы) сосредоточены термины, сравнительно редко употребляемые в конкретной области знаний. Эти слова чаще всего не имеют решающего смыслового значения, однако для динамичных отраслей науки характерна увеличенная правая часть диаграммы.

Например, при статистическом анализе слов из образовательных стандартов России были выделены основные значимые слова и термины выбранной предметной области. Анализ стандарта высшего профессионального образования по специальности "математика", проведенный сотрудниками МИФИ В. Уроженко и В. Сергиевским, позволил выделить набор значимых слов, определяющих круг знаний, которыми владеет специалист-математик. В результате в область значимых слов (свыше 30 раз) попали такие слова: *функция*, *теорема*, *пространство*, *метод*, *решение*, *уравнение*. Как оказалось, эти слова нельзя рассматривать в отрыве от контекста — в каждой области знаний они имеют различное смысловое значение. Поэтому эти слова не могут претендовать на определение круга знаний по выбранной специальности. Слова же, определяющие специальность, встречаются в тексте 1–3 раза.

Законам Зипфа удовлетворяют не только слова из одного текста, но и слова из различных текстовых массивов, библиотек, архивов радиопередач и т.д. Мало

того, законам Зипфа удовлетворяют практически все объекты современного информационного пространства. Например, множество данных свидетельствуют, что само Web-пространство следует распределению Зипфа, если в качестве параметров, вместо слов, рассматривать Web-страницы, которые, в свою очередь, ранжировать по популярности (частоте обращений), поскольку этот показатель можно рассматривать как некоторый аналог "полезности". Рассматривая график распределения для конкретного Web-сайта, можно увидеть практически полное совпадение наблюдаемых закономерностей, за исключением правой части графика. Это отклонение, по-видимому, связано с тем фактом, что Web-сайт является достаточно динамичной системой, не способной инициировать запросы к наименее интересным страницам (в данном случае справедлива поправка Мандлеброта).

Многие исследования показывают, что законам Зипфа подчинены также и запросы работников различных организаций к Web-пространству. Следовательно, работники чаще всего посещают небольшое количество сайтов, при этом достаточно большое количество остальных Web-ресурсов посещается лишь один-два раза.

С другой стороны, каждый Web-сайт получает большую часть посетителей, пришедших по гиперссылкам из небольшого количества сайтов, а из всего остального Web-пространства на него приходит лишь небольшая часть посетителей. Таким образом, объем входящего трафика от ссылающихся Web-сайтов также подчиняется распределению Зипфа. Кстати, по данным аналитической службы Taylor Nelson Sofres, самым эффективным способом привлечения посетителей на Web-сайт являются рекомендации друзей и знакомых. Именно так находят сайты более 18% пользователей Internet. В 13% случаев для этого используются гиперссылки, в 10% — поисковые машины. В целом, около 98% посетителей сайта, довольных его содержанием, скоростью работы, наличием поисковых функций и возможностями персонализации, рекомендуют сайт своим близким. Все эти закономерности могут эффективно использоваться, например, при построении систем кэширования Web-трафика, а также при оптимизации конструкции кэш-систем.

Не так давно Джон Клайнберг из Корнеллского университета предложил свой способ фильтрации информации, позволяющий выявлять наиболее актуальные для каждого конкретного момента времени проблемы, обозначенные в текстах. Этот способ базируется на анализе больших объемов текстовой информации. Когда происходит какое-либо важное событие, о нем начинают активно писать, что приводит к своеобразным "скачкам" в частоте употребления тех или иных слов.

Клайнберг разработал алгоритм, позволяющий анализировать частоту использования того или иного слова, т.е. выполнять ранжирование слов по частоте вхождения. На выходе алгоритм представляет собой рейтинг слов, на основании которого можно делать выводы о популярности той или иной темы и производить сортировку информации.

Чтобы испытать свою разработку, ученый решил проанализировать тексты всех президентских докладов о положении в США (State of the Union addresses) начиная с 1790 года. В итоге получилось, что в период Войны за независимость американских колоний часто употреблялись слова militia ("ополчение") и British ("британский"), а в период с 1947 по 1959 годы наблюдался "скачок" в использовании слова atomic ("атомный"). Таким образом, ученому удалось доказать работоспособность системы.

7.4. Закономерность Брэдфорда

В информатике и математической лингвистике для описания эмпирических ранговых распределений используются и другие статистические распределения, к самым популярным из которых можно отнести распределение знаменитого математика XX века Валодди Вейбулла (E. H. Waloddi Weibull, 1887–1979) [32].

Интересующимся читателям напомним некоторые сведения о статистических распределениях. К основным характеристикам статистических распределений относятся функция и плотность распределения. Для случайной величины x ее функцией распределения называется $F_x(z) = P(x < z)$, где P — это вероятность и $z \in (-\infty, +\infty)$. Плотность распределения (определяемая для непрерывных функций распределения) представляет собой производную от функции распределения: $f_x(z) = F'_x(z)$, $z \in (-\infty, +\infty)$. Обычно, говоря о каком-либо известном статистическом распределении, ограничиваются формулой плотности распределения.
Например, стандартное распределение Парето-Зипфа имеет следующую функцию плотности (для положительного параметра c): $f(x) = c/x^{c+1}$, $x \geq 1$, $c > 0$, где c — параметр распределения. Распределение Вейбулла выражается формулой: $f(x) = (c/b)\cdot((x - \theta)/b)^{c-1}\cdot e^{-((x-\theta)/b)^c}$, где b — параметр масштаба, c — параметр формы, θ — параметр положения.

Частным случаем законов Зипфа и Вейбулла также является закономерность Брэдфорда, связанная с распределением не слов в текстах, а статей, документов или Web-сатраниц, соответственно, в рамках тематических каталогов, баз данных или Web-сайтов.

Основной смысл закономерности С. Брэдфорда (химика, который в свое время исследовал количество публикаций в научных журналах) заключается в следующем: если научные журналы расположить в порядке убывания числа помещенных в них статей по конкретному предмету, то полученный список можно разбить на три зоны таким образом, чтобы количество статей в каждой зоне по заданному предмету было одинаковым. Эти три зоны составляли:

- профильные журналы, непосредственно посвященные рассматриваемой тематике (ядро);
- журналы, частично посвященные заданной области;
- журналы, тематика которых весьма далека от рассматриваемого предмета.

С. Брэдфорд установил, что, по сравнению со второй зоной, количество журналов в третьей зоне будет примерно во столько раз больше, во сколько раз число наименований во второй зоне больше, чем в ядре. Иными словами,

$$P3 / P2 = P2 / P1 = N,$$

где $P1$ — число журналов в 1-й зоне, $P2$ — во 2-й, $P3$ — в 3-й зоне. Однако из приведенной формулировки не совсем ясно, как определяется число журналов, образующих ядро, а также чему равна величина N. На эти вопросы и позволяет ответить анализ свойств ранговых распределений (например, Зипфа или Вейбулла).

Б. Викери уточнил модель С. Брэдфорда [37]. Он выяснил, что журналы, ранжированные (выстроенные) в порядке уменьшения в них статей по конкретному вопросу, можно разбить на любое нужное число зон (K). При этом

$$Pn / P(n-1) = P(n-1) / P(n-2),$$

где $n > 2$ и $n < K + 1$.

Закономерность Брэдфорда изначально рассматривалась как специфический случай распределения Зипфа для системы периодических изданий по науке и технике. Исходя из реалий развития сети Internet, ее можно рассматривать как закономерность, относящуюся к ранговому распределению Web-сайтов, относительно вхождения в них Web-страниц, релевантных некоторой области знаний.

Очевидно, что закономерность Брэдфорда (как и закон Зипфа) можно использовать и при построении словарей ключевых слов по некоторой тематике. Если на основе анализа текстов документов построить частотный словарь по некоторой тематике, то в нем также можно выделить такие области: 1 — наиболее часто используемые слова с самыми малыми рангами, куда входят главным образом служебные слова; 2 — общеупотребительные слова; 3 — тематическая лексика (среднечастотные слова); 4 — межотраслевая лексика (редко употребляемые слова). Естественно, для построения тематического словаря наибольший интерес представляет третья область.

7.5. Прогноз Мура и информационная сфера

В заключение обсуждения назовем еще одну закономерность, которая родилась как прогноз развития технологии микросхем, но все шире вторгается во все сферы жизни. В 1965 году один из учредителей компании Intel Гордон Мур предсказал, что плотность транзисторов в интегральных схемах и, соответственно, производительность микропроцессоров будут удваиваться каждый год. В течение трех последних десятилетий этот прогноз, названный "законом Мура", более или менее выполнялся, хотя достаточно быстро был скорректирован — удвоение должно происходить каждые два года. Например, в феврале 2003 года в Сан-Хосе на ежегодном весеннем форуме Intel (IDF) исполнительный директор компании Крейг Барретт заявил, что в настоящее время прогноз Мура продолжает действовать, в результате чего сохраняется высокая потребность в новейших технологиях. Кроме того, в соответствии с известным законом Мура, к 2010 году "железо" самого современного компьютера превзойдет по своим возможностям человеческий разум, а затем, в самом ближайшем будущем, это станет по силам и программному обеспечению.

Еще недавно было принято считать, что закон Мура относится исключительно к микросхемам, потому что Гордон Мур — из Intel. При этом предполагалось, что в сфере коммуникаций и Internet закон Мура не действует, так как эти области часто построены на использовании более старых технологий, не способных масштабироваться на таком же уровне, как и современные вычислительные технологии. Несмотря на небольшой спад на рынке высоких технологий, который длится уже около трех лет, развитие коммуникационного оборудования, широкого спектра устройств, таких как оптические, сенсорные, механические и даже биологические, все-таки подтвердило, что прогноз Мура распространяется на все большее количество областей.

Сегодняшнее расширение Internet, стремительный рост объемов пересылаемых данных, развитие электронной коммерции и беспроводной связи, а также

внедрение цифровых технологий в бытовую технику можно рассматривать как следствие все того же закона Мура [13].

Было замечено, что рост документальной информации, вполне подчиняясь закону Мура, также носит экспоненциальный характер, а именно кривая роста числа документов может быть описана уравнением вида $y = Ae^{kt}$, где y — количество документов; t — время (в годах); A — количество документов в начале отсчета (при $t = 0$); k — некоторый коэффициент. Процесс экспоненциального роста информации не сулит ничего хорошего ввиду стремительного увеличения хаоса и накопления энтропии. Средства автоматизации обработки и сетевые технологии способствуют многократному дублированию информации, т.е. эффекту автоматического порождения новых документов на основании существующих.

Американские исследователи — профессоры Калифорнийского университета в Беркли Питер Лайман и Хол Вэриен [54] — пришли к выводу, что за три года объем информации, производимой человечеством, удваивается.

В 2003 году в мире было заархивировано свыше 5 млрд гигабайт новой информации, а электронным путем передано примерно 18 млрд гигабайт информации, из которых 17,3 млрд — через телефонные линии. Идея "офиса без бумаги" оказалась иллюзией — объемы бумажных архивов за последние три года выросли на 43%. К таким выводам пришли профессоры и студенты Школы управления информацией Университета Беркли, которые провели соответствующие исследования.

Принимая во внимание все население Земли, на одного человека в среднем за год пришлось примерно 800 Мбайт новых данных. Проще говоря, такое же количество информации содержится в книгах, сложенных в стопку высотой 10 м. Человечество всего за один год создало столько информации, что ею можно было заполнить 500 тыс. библиотек Конгресса США. Количество хранимых данных, по сравнению с 1999 годом, когда проводились такие же исследования, возросло на 30%.

Наиболее распространенным в мире средством хранения данных являются жесткие диски. Результаты исследований гласят, что количество информации, хранящейся на постоянно увеличивающих свои объемы дисках, по сравнению с 1999 годом, возросло на 114%. Кроме того, исследователи серьезно развеяли миф о том, что архивы (причем как бумажные, так и пленочные) будут постепенно переводиться в цифровую форму. По результатам исследований, устойчивый тренд наблюдается исключительно в области фотографии — количество сделанных во всем мире отпечатков в 2002 году, по сравнению с 1999, снизилось на 9%.

Развитие коммуникационных возможностей способствует росту количества доступной через Сеть информации, появлению технологий немедленной публикации идей, комментариев, дневников, фотографий и т.д. С другой стороны, увеличение объемов доступного контента приводит к росту инновационной деятельности, все больше знаний, необходимых для исследовательских работ, публикуется в Internet, что способствует технологическому прогрессу, на котором базируется прогноз Мура.

7.6. Фракталы и информационное пространство

Термин *фрактал* образован от латинского слова fractus — дробный, состоящий из фрагментов. Он был предложен Бенуа Мандельбротом в 1975 году для обозначения нерегулярных самоподобных математических структур. Популярная сегодня фрактальная геометрия получила свое название лишь в 1977 году благодаря книге Мандельброта "The Fractal Geometry of Nature". В его работах

использованы научные результаты многих ученых, работавших в этой же области (прежде всего, Пуанкаре, Кантора, Хаусдорфа). Определение фрактала, данное Мандельбротом, звучит так: "Фракталом называется структура, состоящая из частей, которые в каком-то смысле подобны целому".

Таким образом, одним из основных свойств фракталов является *самоподобие*. В самом простом случае небольшая часть фрактала содержит информацию о всем фрактале. Строгое определение самоподобных множеств было дано Дж. Хатчинсоном в 1981 году. Он назвал множество самоподобным, если оно состоит из нескольких компонентов, подобных ему, т.е. компонентов, получаемых афинными преобразованиями: поворотом, сжатием и отражением исходного множества.

7.6.1. Примеры абстрактных фракталов

Мандельброт предложил не только определение фракталов, но также и алгоритм построения одного из них, получивший название в честь этого ученого (рис. 7.1). Алгоритм построения фрактала Мандельброта основан на итеративном вычислении по формуле:

$$Z[i+1] = Z[i] \times Z[i] + C,$$

где $Z[i]$ и C — комплексные переменные. Итерации выполняются для каждой стартовой точки C прямоугольной или квадратной области, представляющей собой подмножество комплексной плоскости. Итерационный процесс продолжается до тех пор, пока $Z[i]$ не выйдет за пределы окружности заданного радиуса, центр которой лежит в точке $(0, 0)$, или после достаточно большого числа итераций. В зависимости от количества итераций, в течение которых $Z[i]$ остается внутри окружности, можно установить цвет точки C. Если $Z[i]$ остается внутри окружности в течение достаточно большого количества итераций, итерационный процесс прекращается, и эта точка окрашивается в черный цвет. Множеству Мандельброта принадлежат точки, имеющие черный цвет, т.е. те, которые в течение бесконечного числа итераций не уходят в бесконечность.

Так как количество итераций соответствует номеру цвета, точки, находящиеся ближе к множеству Мандельброта (черного цвета), имеют более яркий цвет.

Построение другого фрактального множества, называемого "снежинкой Коха", начинается с правильного треугольника, длина стороны которого равна единице. Сторона треугольника в этом случае считается базовым звеном. Далее на любом шаге итерации каждое звено заменяется на образующий элемент — ломаную, состоящую по краям из отрезков длиной 1/3 длины звена, между которыми размещаются две стороны правильного треугольника со стороной, равной 1/3 длины звена. Получаемое в результате итерационного процесса фрактальное множество (кривая n-го поколения при любом конечном n называется предфракталом, а при n, стремящемся к бесконечности, кривая Коха становится фракталом) представляет собой линию бесконечной длины, ограничивающую конечную площадь. Действительно, при каждом шаге число сторон результирующего многоугольника увеличива-

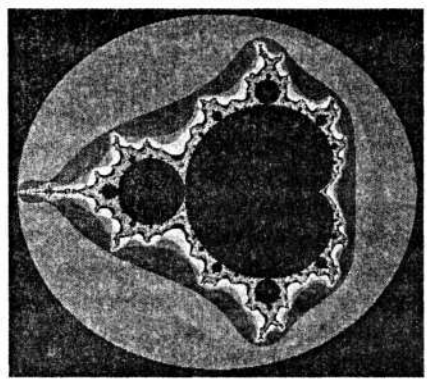


Рис. 7.1. Множество Мандельброта

ется в 4 раза, а длина каждой стороны уменьшается только в 3 раза, т.е. длина многоугольника на n-й итерации равна $3 \times (4/3)^n$ и стремится к бесконечности с ростом n. Первые шаги построения этого фрактала изображены на рис. 7.2.

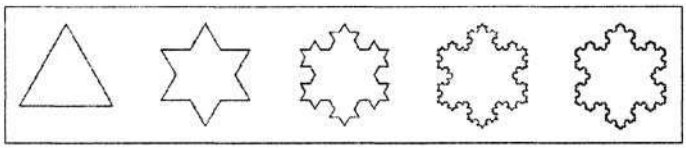


Рис. 7.2. Снежинка Коха

Площадь под кривой, если принять площадь образующего треугольника за единицу, равна

$$S = 1 + \frac{1}{3}\sum_{k=0}^{\infty}\left(\frac{4}{9}\right)^k = 1{,}6.$$

Таким образом, площадь под снежинкой Коха в 1,6 раза больше площади образующего ее треугольника.

В середине 80-х годов появился метод IFS (Iterated Functions System — система итерируемых функций), который получил распространение как простое средство построения фрактальных структур. IFS реализуется как система функций, отображающих одно многомерное множество на другое. Простейшая IFS представляет собой аффинные преобразования плоскости:

$$X' = A \times X + B \times Y + C$$
$$Y' = D \times X + E \times Y + F$$

В 80-х годах американские ученые Майкл Барнсли и Алан Слоан предложили некоторые идеи, основанные на теории динамических систем, для сжатия и хранения графической информации. Они назвали это методом фрактального сжатия информации. В результате был создан алгоритм, который позволил сжимать некоторые виды графической информации в 500–1000 раз. При этом исходное изображение разбивается на фрагменты, каждый из которых кодируется несколькими аффинными преобразованиями. Закодировав какой-то фрагмент изображения двумя аффинными преобразованиями, его можно определить с помощью 12-ти коэффициентов. Если задаться какой-либо начальной точкой и запустить итерационный процесс, то через несколько десятков итераций совокупность полученных точек будет описывать закодированный фрагмент изображения.

Ниже приведены некоторые значения коэффициентов, которые позволяют получать конкретные геометрические образы фракталов (взято из http://ilab24.narod.ru/fract/frifs1.html). Отметим, что дополнительно к 6-ти коэффициентам каждого афинного преобразования добавляется 7-й нормирующий коэффициент, пропорциональный площади, занимаемой соответствующим звеном фрактала.

```
Кристалл {
.69697 -.481061 -.393939 -.662879 2.147003 10.310288 .747826
.090909 -.443182 .515152 -.094697 4.286558 2.925762 .252174
}
Цветок {
0 -.5 .5 0 -1.732366 3.366182 .333333
.5 0 0 .5 -.027891 5.014877 .333333
```

```
0 .5 -.5 0 1.620804 3.310401 .333333
}
Кривая Коха {
.307692 0 0 .294118 4.119164 1.604278 .151515
.192308 -.205882 .653846 .088235 -.688840 5.978916 .253788
.192308 .205882 -.653846 .088235 .668580 5.962514 .253788
.307692 0 0 .294118 -4.136530 1.604278 .151515
.384615 0 0 -.294118 -.007718 2.941176 .189394
}
Дерево {
0 0 0 .50 0 0 .05
.42 -.42 .42 .42 0 .2 .4
.42 .42 -.42 .42 0 .2 .4
.1 0 0 .1 0 .2 .15
}
Салфетка Сперанского {
.5 0 0 .5 0 .5 .333333
.5 0 0 .5 -.25 0 .333333
.5 0 0 .5 .25 0 .333333
}
Дракон {
.5 -.5 .5 .5 0 0 .5
 -.5 -.5 .5 -.5 1.5 -.5 .5
}
Гоблины {
.47 .17 -.17 .47 0 0 .45
 -.11 .77 -.77 -.11 1.11 .77 .55
}
```

Использование IFS для сжатия обычных изображений (например, фотографий) основано на выявлении локального самоподобия, в отличие от фракталов, где наблюдается глобальное самоподобие и нахождение IFS не слишком сложно. По алгоритму Барнсли происходит выделение в изображении пар областей, меньшая из которых подобна большей, и сохранение нескольких коэффициентов, кодирующих преобразование, переводящее большую область в меньшую. Требуется, чтобы множество "меньших" областей покрывало все изображение. Восстанавливающий алгоритм должен применять каждое преобразование к некоторому фрагменту, принадлежащему области, соответствующей применяемому преобразованию.

Фракталы с большой точностью описывают многие физические явления и природные образования: горы, турбулентные течения, ветви деревьев, кровеносные сосуды, форма которых очень далека от простых геометрических фигур. Мандельброт в свое время заметил: "Почему геометрию часто называют холодной и сухой? Одна из причин заключается в ее неспособности описать форму облака, горы, дерева или берега моря. Облака — это не сферы, горы — не конусы, линии берега — это не окружности, и кора не является гладкой, и молния не распространяется по прямой. Природа демонстрирует нам не просто более высокую степень, а совсем другой уровень сложности."

Роль фракталов в машинной графике сегодня достаточно велика. Они приходят на помощь, например, когда требуется с помощью нескольких коэффициентов задать линии и поверхности очень сложной формы. С точки зрения машинной графики, фрактальная геометрия незаменима при генерации "квазиприродных" изображений. Фактически найден способ легкого представления сложных неевклидовых объектов, образы которых весьма похожи на природные.

7.6.2. Фракталы из жизни

Сложное устройство береговых линий — один из лучших примеров проявления фракталов в природе. Действительно, на километровом отрезке побережье выглядит столь же изрезанным, как и на стокилометровом.

Недавно Бернард Саповаль из Политехнической школы в Палезо (Франция) и его коллеги создали компьютерную модель эрозии побережья. В модели вещество разрушалось либо под прямым воздействием волн, либо медленным "выветриванием", когда минералы растворялись в воде. Побережье было разделено на равные участки, причем в модели типы камней на этих участках выбирались случайным образом. Модель показала, что изначально гладкая береговая линия стремительно приобретает неровный профиль с выступами и множеством отделенных от берега островов, приближаясь в результате к привычному фрактальному профилю. Образовавшийся при моделировании берег очень напоминал Восточное побережье США. Ученые полагают, что им удалось обнаружить основное воздействие — изменение эрозионной силы самим побережьем.

Сегодня при моделировании рельефа широко используются двухмерные стохастические фракталы, которые получаются в том случае, если в итерационном процессе случайным образом менять какие-либо его параметры. Идеальным примером случайного фрактала в природе является береговая линия.

Опыт показывает, что длина береговой линии L зависит от масштаба l, с которым проводятся измерения, и увеличивается с уменьшением последнего по степенному закону:

$$L = \Lambda l^{-\alpha}, \qquad \Lambda = const .$$

Так, например, для побережья Великобритании $\alpha \approx 0{,}3$. Число раз N, которое измерительный масштаб l укладывается вдоль побережья, равно: $N = L/l = \Lambda l^{-(1+\alpha)}$, т.е. фрактальная размерность береговой линии Великобритании — степенной показатель с обратным знаком — равна $1 + \alpha \approx 1{,}3$.

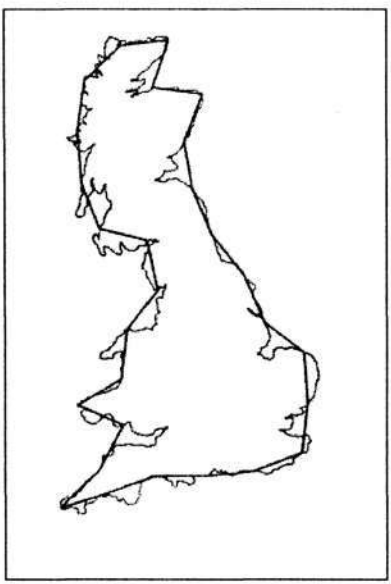


Рис. 7.3. Береговая линия побережья Великобритании

Процессы, происходящие в живой природе, также часто ассоциируются с фракталами. Пожалуй, самый яркий пример — растения или животные, которые развиваются согласно данным такого носителя биологической информации, как ДНК.

В 2004 году в Ньюфаундленде биологом Ги Нарбонн из университета Кингстона (Канада) была открыта редкая ископаемая природная структура фрактального типа. Были найдены следы организмов, живших на Земле около 575 миллионов лет назад, не относившихся ни к растениям, ни к животным. Эти организмы назвали *рангеоморфами*, они были неспособны двигаться и не имели репродуктивных органов, а размножались, создавая новые ответвления. Организмы собирались в фрактальные структуры из

разветвляющихся частей. Как выяснилось, каждый ветвящийся элемент фрактальных структур состоял из множества трубок, удерживаемых вместе полужестким органическим скелетом организмов. Нарбонн обнаружил рангеоморфы, собранные в несколько разных форм. Фрактальный рисунок представляется достаточно сложным, но, по словам исследователя, сходство организмов друг с другом делало достаточным простой геном для создания новых свободно плавающих ответвлений и соединения ответвлений в более сложные структуры.

Уже около полувека в биологии известен закон, который утверждает, что многие свойства организмов, от продолжительности жизни и количества детенышей до скорости обмена веществ, пропорциональны массе тела в степени $n/4$, где n — целое. При этом сама природа закона более полувека оставалась загадкой. На первый взгляд, вместо четверки должна быть тройка, поскольку масса пропорциональна кубу размера тела.

Несколько лет назад объяснение было найдено. Дело в том, что пронизывающие каждый организм сети — кровеносная у животных или капиллярная у растений — обладают свойствами фракталов. Фрактальность этих сетей как раз и приводит к добавлению еще одного "измерения" у живых организмов.

Вся Вселенная, в соответствии с гипотезой российского физика Сергея Хайтуна, является фракталом, причем единственным известным в природе, полностью удовлетворяющим классическому определению. В физике давно известен факт, что плотность космических объектов стремительно падает с увеличением их размеров. Еще в 50-х годах советские физики-теоретики пришли к выводу, что "бесконечная" плотность Вселенной равна нулю. Эта идея и новейшие представления о фрактальности Вселенной подтверждают друг друга. Дело в том, что плотность всякого фрактала, расположенного в трехмерном пространстве, тождественно равна нулю. Классические фракталы обладают "всюду пустой" структурой, которая, при проникновении в нее, "расширяется" до бесконечности. Реальные же системы, естественно, не позволяют бесконечного углубления в свою структуру; поэтому на каком-то конечном этапе реальная структура теряет свой "фрактальный" вид, а значит, реальные структуры лишь "фракталоподобны".

Позволяя — из-за своей бесконечности — бесконечное проникновение в свою структуру, Вселенная, судя по всему, является единственным "настоящим" фракталом, имеющим нулевую бесконечную плотность.

7.6.3. Информационные фракталы

В настоящее время информационное пространство в целом, ввиду его объемов и динамики изменения, принято рассматривать как стохастическое. В большинстве моделей информационного пространства изучаются структурные связи между тематическими множествами, входящими в это пространство. При этом численные характеристики этих множеств подчиняются гиперболическому закону (с возможными степенными поправками). Сегодня в моделировании информационного пространства все чаще используется фрактальный подход, базирующийся на свойстве самоподобия информационного пространства, т.е. сохранение внутренней структуры множеств при изменениях их размеров или масштабов их рассмотрения извне.

Самоподобие информационного пространства выражается, прежде всего, в том, что, при почти обвальном росте этого пространства в последние десятилетия, гиперболические частотные и ранговые распределения, получаемые в таких разрезах, как, например, источники и авторы, практически не меняют своей формы. Следовательно, применение теории фракталов при анализе информаци-

онного пространства позволяет с общей позиции взглянуть на эмпирические законы, составляющие теоретические основы информатики. Например, тематические информационные массивы сегодня представляют развивающиеся самоподобные структуры, т.е. являются стохастическими фракталами. В информационном пространстве возникают, растут и формируются кластеры документов, отражающие современные процессы коммуникации.

Закономерности, открытые такими учеными, как Зипф, Брэдфорд, Лотки и другие, в полной мере свидетельствуют о самоподобии информационного пространства. С другой стороны, самоподобие (скейлинг) можно рассматривать и как следствие общих структурных закономерностей информационного пространства.

Свойства самоподобия фрагментов информационного пространства наглядно демонстрирует новый интерфейс, представленный на Web-сайте службы News Is Free (`http://newsisfree.com`) в режиме бета-тестирования. На этом сайте отображается состояние информационного пространства в виде ссылок на источники и отдельные сообщения. При этом учитывается два основных параметра отображения — ранг популярности и "свежесть" информации. Укрупненное представление отдельных источников и/или документов — наиболее популярных и актуальных, приведено на рис. 7.4.

Средних по популярности документов, конечно, значительно больше. При сохранении общей структуры происходит "дробление" источников — как показано на рис. 7.5.

И наконец, когда предельный ранг популярности, а также "свежести" повышается, дробление уже не позволяет без особых усилий читать названия источников и идентифицировать отдельные документы (рис. 7.6).

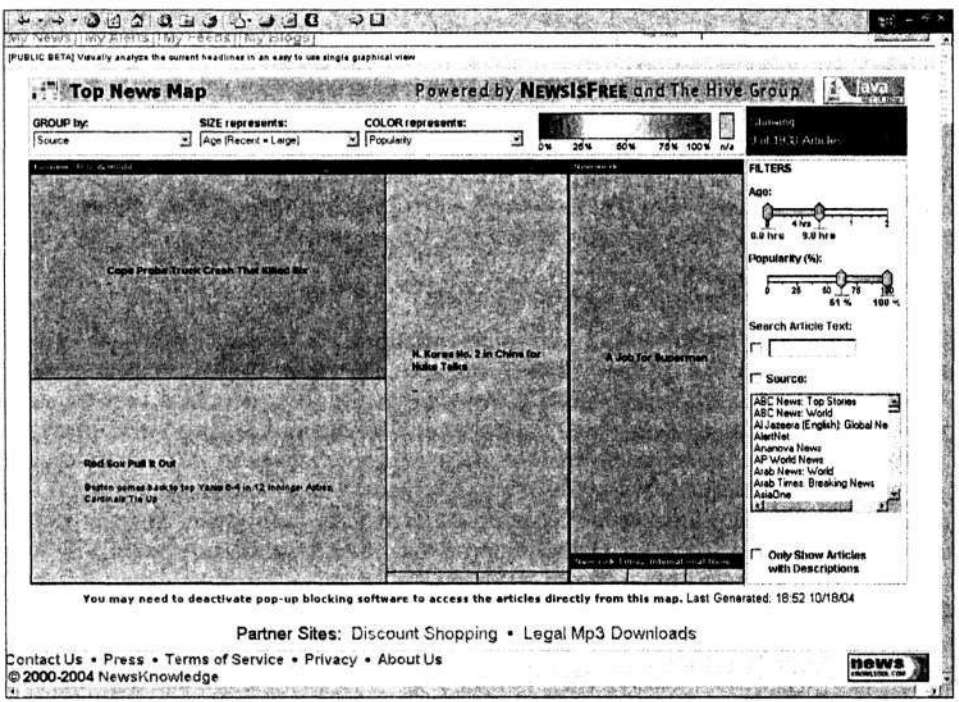


Рис. 7.4. Небольшой кластер популярных изданий средней "свежести"

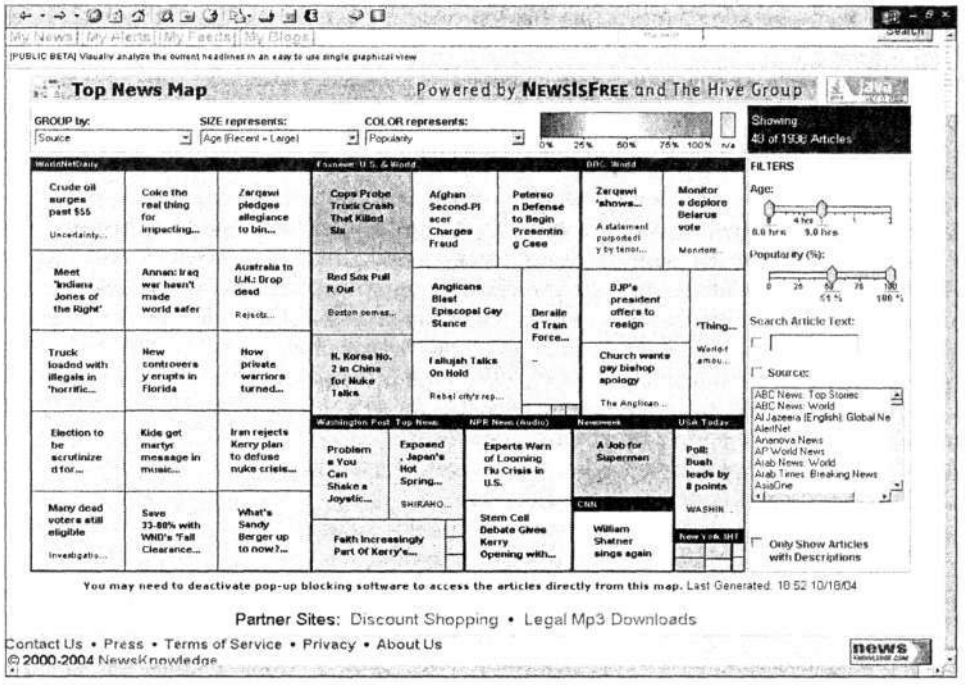


Рис. 7.5. Кластер источников средней популярности и средней "свежести"

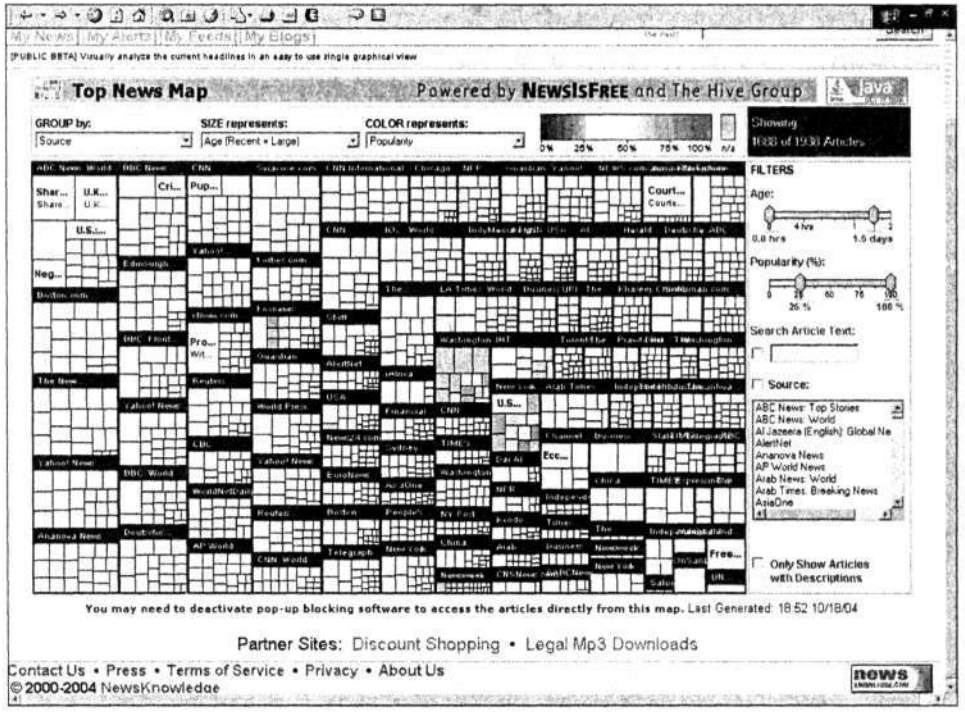


Рис. 7.6. Кластер источников низкой популярности "3-й категории свежести"

При этом, очевидно, что три последние иллюстрации в целом хорошо демонстрируют свойство подобия информационного пространства.

Приведем одну из фрактальных моделей информационного пространства, основанную на подходе, называемом *диффузионно-ограниченной агрегацией*. Эта стохастическая модель широко применяется для процессов, распространенных в природе. Ее обычно определяют следующим образом.

Представим себе многомерную сферу (окружность в двухмерном случае) достаточно большого радиуса, на поверхности которой время от времени в случайных местах появляются частицы, которые затем диффундируют внутрь сферы. В центре сферы находится так называемый "зародыш". При столкновении с ним диффундирующая частица "прилипает" к нему и больше не движется (попадает в "архив"). Затем с этим образованием сталкивается следующая, выпущенная с поверхности сферы частица, и так до бесконечности. Поток частиц с поверхности сферы будем считать достаточно малым, так что столкновениями диффундирующих частиц друг с другом можно пренебречь.

В результате образуется очень пористая структура, проекция которой на двухмерную поверхность показана на рис. 7.7. Большие поры внутри "экранируются" отростками достаточно большой длины. По мере роста структуры число пор и их размеры увеличиваются.

В природе так, например, растут кораллы, кристаллы, снежинки, опухоли. Перенос этой модели на информационное пространство можно интерпретировать таким образом. Каждой размерности исходной сферы приписывается определенная тематика, а роль "зародыша" играет исходный информационный массив. При пополнении информационного массива новый документ, размещенный в определенном месте на поверхности сферы, стремится к ядру, пересекается с некоторой ветвью и увеличивает ее. Проекция такой модели на плоскость вполне соответствует структуре, представленной на рис. 7.7. Что может дать подобная модель? Самое главное, она может служить эффективным алгоритмом группировки объектов, способным выявлять новые темы (ветви — кластеры), служащие в дальнейшем основой для новой уточненной классификации.

Web-пространство, являясь, пожалуй, самой динамичной частью информационного пространства, характеризуется большим количеством скрытых в нем неявных экспертных оценок, реализованных в виде гиперссылок. В ноябре 1999 года Андрей Бредер (Andrei Broder) и его соавторы из компаний AltaVista, IBM и Compaq построили модель ресурсов и гиперсвязей Internet. Исследования опровергли расхожее мнение, будто WWW — это единое густое пространство. В рамках этой модели было обнаружено постоянное соотношение между отдельными его частями. Топология и характеристики модели, получившей название Bow Tie ("галстук-бабочка"), оказались примерно одинаковыми для различных фрагментов полного Web-пространства, под-

Рис. 7.7. Фрактал, полученный в процессе диффузионно-ограниченной агрегации

утверждая тем самым наблюдение о том, что Web — это фрактал, т.е. свойства структуры всего пространства также верны и для его отдельных подмножеств. Сегодня фрактальные особенности WWW уже достаточно широко используются при решении таких задач, как оптимизация механизмов сканирования, анализ и прогноз развития информационных ресурсов, построение новых Web-сервисов.

7.7. Проблемы и феномены Internet

С появлением и развитием сети Internet и ее информационных ресурсов в корне изменились тенденции и темпы роста информационного пространства. Некоторые явления, не известные ранее, стали проявляться с достаточной очевидностью. В совокупности эти явления образуют группу новых понятий информационного динамично изменяемого пространства — феномены современных информационных потоков [28].

1. *Прогресс в области производства информации ведет к снижению общего уровня информированности.*

Полезной информации все больше, но найти ее все сложнее. Вследствие этого традиционные информационно-поисковые системы постепенно стали утрачивать свою актуальность. Причина этого не столько в физических объемах информационных потоков, сколько в их динамике, т.е. в постоянном систематическом обновлении информации, которое к тому же далеко не всегда имеет очевидную регулярность. Современные информационно-поисковые системы уже не в состоянии угнаться за обновлениями существующих сайтов, а также за учетом контента новых, постоянно создаваемых сайтов. Периоды индексации универсальных систем составляют от двух-трех недель до нескольких месяцев. Интеграторы новостей, учитывающие ничтожно малую часть сетевых источников, также не справляются с этой задачей с гарантированной полнотой и точностью.

Количество новостных сообщений, публикуемых в сети Internet во всем мире, превышает 1 000 000 в сутки. Крупнейшие сетевые интеграторы новостей обрабатывают ежесуточно десятки тысяч сообщений. Ситуация резкого роста темпов производства информации породила ряд проблем.

- непропорциональный рост "информационного шума" ввиду слабой структурированности информации;
- появление паразитной информации (невостребованной, получаемой в качестве несанкционированных "приложений", например, к электронным письмам);
- несоответствие формально релевантной (уместной, относящейся к делу) информации действительным потребностям;
- многократное дублирование информации (типичный пример — публикация одного и того же сообщения в разных изданиях).

2. *Новые сетевые службы, охватывая порой в 1000 раз меньше источников, значительно эффективнее решают проблемы пользователей.*

Необходимость сетевой интеграции новостей несколько лет назад осознали известные сетевые поисковые службы. На первых этапах они заключили соглашения с такими крупнейшими информационными агентствами, как Reuters, Associated Press, CNN и другие, и стали предоставлять доступ в режиме поиска

и просмотра новостных сообщений. Таким образом, у пользователя впервые появилась возможность бесплатно находить и просматривать новости реального (а не только "виртуального") мира в Сети. Например, старейший навигационный портал Yahoo! создал службу Daily News (`http://dailynews.yahoo.com`), объединив информацию нескольких десятков агентств и обеспечив графическое и мультимедийное представление отдельных тематических областей.

3. *Интенсивность роста объема шумовой информации многократно превышает интенсивность роста информации полезной.*

Преобладание шумовой информации обусловливает необходимость подходов, аналогичных стохастическим критериям, применяемым при разделении сигналов и шумов. Вместе с тем, текстовый характер информации порождает новые семантические методы, которые уже сегодня успешно применяются. Вспомним Пола Грема и его высказывание о том, что ахиллесова пята спамеров — это текстовое содержание электронных сообщений.

4. *Важные сообщения многократно дублируются в экспоненциально растущем количестве сайтов, в то время как количество заслуживающих внимания источников растет не такими большими темпами и скорее всего линейно.*

Дело в том, что серьезные источники информации — это объекты реальной жизни, в то время как сайты в своей совокупности представляют виртуальное пространство, которое развивается по другим законам. Преодоление использования явно дублирующейся информации не представляет проблем, однако дублирующиеся по смыслу сообщения выявляются не так легко, здесь на помощь приходят алгоритмы, аналогичные алгоритмам построения информационных портретов, их сопоставления, сравнения и вероятностной оценки. Очевидно, что такие подходы требуют очень больших вычислительных мощностей при учете общего роста объемов потоков. Серьезное упрощение задачи может быть получено за счет применения содержательных методов, например при ранжировании первоисточников, определении и выделении тематических информационных каналов, экспертном формировании словарей значимых слов и т.п.

5. *Устранение дублирующихся сообщений в информационных потоках требуется далеко не всегда.*

Существует ряд задач, в которых используется факт дублирования текстов сообщений в различных источниках, например при определении важности сообщения (например, если сообщение многократно дублируется на сайтах и в СМИ) или при определении эффективности PR-компании (подсчет републикаций пресс-релизов и др.).

6. *Управление информационными потоками, построенное на основе учета закономерностей их формирования, особенностей републикации отдельных сообщений, динамики использования отдельных понятий и даже индексирования отдельных сообщений различными поисковыми системами, в настоящее время является мощнейшим инструментом влияния за счет механизмов обратной связи.*

Информация, возникающая на сайтах — в виртуальном пространстве, — становится доступной все большему количеству людей — пользователей сети Internet, но сегодня даже не это главное. Она становится доступной журналистам и аналитикам, политикам и бизнесменам, которые эффективно используют

ее в реальной жизни — в публикациях в СМИ, при принятии решений, проведении маркетинговых мероприятий.

Web-пространство продолжает расширяться, хотя темп его роста замедлился по сравнению с 1994–2000 годами. Доступ к известной информации из Web-пространства теперь технически значительно упрощен; однако объемы информации растут, что, в свою очередь, усложняет поиск. Большие поисковые системы редко предоставляют возможности глубокого анализа массивов документов, потому что процессы семантической обработки текстовой информации намного затратнее, чем регулярное сканирование данных и их индексация. По всей видимости, в ближайшем будущем ожидается включение лингвосемантических компонентов в лидирующие поисковые системы.

Извлекаемая из сетевых документов информация будет интегрироваться с информацией из других источников — баз данных и баз знаний, словарей, проблемно-ориентированных каталогов. При этом объединяющими форматами данных, по-видимому, будет XML и связанные с ним стандарты описания метаданных.

В конечном счете, информационный поиск и глубинный анализ гипертекста в любом случае расширятся на естественный язык и будут интегрированы с результатами исследований в области компьютерной лингвистики, которая автономно развивается уже десятилетия.

Вместе с тем, один из основателей Google Сергей Брин на проходящей недавно конференции "Supernova" заявил, что еще в течение пары веков(!) никакие машинные комплексы не догонят человека по способности искать, сортировать и оценивать информацию.

Руководитель аналитического проекта IST Джим Янсен высказал более оптимистичный прогноз, заявив, что "нынешние поисковые системы неплохи, но следует их улучшать и, возможно, создавать специальные тематические поисковики по определенным направлениям".

Несмотря на увеличение возможностей современных информационно-поисковых систем, они, в основном, еще не способны настраиваться на информационные потребности отдельных пользователей — слабо учитываются (если учитываются) персональные закладки, история запросов и т.п.

Все это обусловливает развитие общедоступных систем в направлении персонализации, основанной на построении пользовательских профилей, архивов поисковых сессий и так далее, и должно будет обеспечивать каждого пользователя информацией, соответствующей его информационным потребностям.

Глоссарий

ARPANET (Advanced Research Projects Agency Network)
Сеть Управления перспективных исследований

Глобальная исследовательская сеть с коммутацией пакетов; предшественница Internet. Основана в 1969 году под эгидой Агентства перспективных исследований Министерства обороны США (Defense Department's Advanced Projects Research Agency). В сети ARPANET впервые были реализованы многие из сетевых принципов, которые используются сегодня. Завершила свое существование в 1990 году.

HTML (HyperText Markup Language)
Язык гипертекстовой разметки

Стандартный язык для описания содержания и структуры гипертекстовых документов. HTML-документы представляют собой текстовые файлы со встроенными специальными командами (разметкой), которые, как правило, отмечают определенную область текста. HTML состоит из независящих от программного обеспечения и аппаратной платформы команд, описывающих структуру гипертекстовых документов. HTML является прикладной разновидностью языка SGML. Используется в WWW для создания Web-страниц.

HTTP (HyperText Transport Protocol)
Протокол передачи гипертекста

Протокол, предназначенный для общения клиента и сервера в среде WWW. Обеспечивает передачу Web-страниц по Internet.

Internet Интернет, Сеть

Internet — глобальная информационная сеть, части которой логически взаимосвязаны единым адресным пространством, основанным на протоколе TCP/IP. Internet состоит из множества взаимосвязанных компьютерных сетей и обеспечивает удаленный доступ к компьютерам, электронной почте, доскам объявлений, базам данных и дискуссионным группам.

RFC (Request for Comments)
Запрос для комментариев

Совокупность публикуемых документов, в которых излагаются стандарты, проекты стандартов и принципиально согласованные идеи по деятельности Internet. Эти документы фактически регламентируют функционирование Internet. Первый RFC вышел в 1969 году. Общее число RFC на сегодня превышает две тысячи.

RSS (Realy Simple Syndication)
Формат RSS

Формат данных и технический стандарт, который обеспечивает интегрированный доступ к новостной информации, представленной на Web-сайтах сети Internet. RSS — это разновидность XML, формат, специально созданный для обмена контентом Web-сайтов. Популярность RSS как стандарта обусловлена его

доступностью и простотой. Сегодня практически все ведущие информационные сайты в мире, "живые журналы", работающие в Сети, используют RSS как инструмент оперативного представления своих обновлений.

RSS-фид, RSS-канал (RSS-feed)

Основным применением RSS в настоящее время являются новостные фиды. Фид — это файл в формате RSS, в который записывается, например, новостной контент Web-сайта.

SQL (Structured Query Language)
Язык структурированных запросов

Язык системы управления базой данных, использующий соответствующие команды и синтаксис для управления процессом взаимодействия и обработки данных в базе данных.

WAIS (Wide Area Information Service)
Служба поиска распределенной информации

1. WAIS-протокол Internet, позволяющий осуществлять поиск информации в Internet в соответствии с библиографическим стандартом Z39.50.
2. Информационно-поисковая система, построенная в соответствии с WAIS-протоколом. В настоящее время WAIS-системы повсеместно интегрируются с средой WWW.

WAP (Wireless Application Protocol)
Протокол приложений беспроводной связи

Протокол беспроводного доступа к информационным и сервисным услугам глобальной сети Internet непосредственно с мобильных телефонов. Основное преимущество WAP заключается в том, что для работы в сети абоненту не нужны дополнительные устройства — компьютер, модем и прочее — достаточно одного мобильного аппарата с поддержкой WAP.

Web-портал (Web-Portal)

WWW-сервис, в основе которого лежит идея создания унифицированного интерфейса для эффективного доступа к информации и объединения в одном месте большой группы Internet-сервисов. Главная тенденция Web-порталов в настоящее время состоит в конвергенции Web-информации с приложениями настольных систем.

Web-сайт (Web-Site)

Происходит от английского "site" (участок). Является совокупностью Web-страниц, объединенных и связанных по смыслу или ссылкам и размещенных на каком-либо сервере в Internet. Доступ к Web-сайту обеспечивается с использованием протокола HTTP.

Web-страница, Web-документ (Web-page)

Составная часть Web-сайта. Обычно Web-страница — это электронный документ, который может содержать текст, изображения, Java-аплеты и другие Web-элементы. Web-страница может быть статическая или динамически сгенерированная.

WWW (World Wide Web, Web)
Всемирная паутина

Графический сервис в среде Internet, предназначенный для гипертекстового связывания мультимедиа-документов в сети. Устанавливает универсальные информационные связи между этими документами независимо от их физического размещения в сети. Для загрузки Web-документов и других данных с WWW-серверов используется протокол HTTP.

XML (Extensible Markup Language)
Расширяемый язык разметки

Стандарт языка разметки, принятый консорциумом W3C в феврале 1998 года. Главные его особенности заключаются в возможности расширения набора тегов, используемых для разметки документов; возможности задания структуры документа, правильность которой верифицируется браузером; в отделении средств разметки по содержанию от разметки, ориентированной на представление документов. Для решения второй задачи предназначены дополнительные специальные языки описания стилей документов — CSS и XSL.

XML-СУБД (XML DBMS)

СУБД нового поколения для хранения и обработки XML-данных (их часто называют native — естественными), при построении которых исходят из необходимости отражения операций, которые совершаются с документами в реальной жизни. В качестве модели данных в XML-СУБД используется XML-модель. Следует отличать XML-ориентированные базы данных от реляционных, поддерживающих обмен данными на языке XML.

Автоматическое реферирование (Summarization)

Автоматическое формирование краткого изложения исходного текстового материала либо путем выделения фрагментов информационного наполнения и последующего их соединения, либо методом генерации текста на основании выявления знаний из оригинала.

Булева модель (Boolean model)

Модель поиска, опирающаяся на булевы операции — пересечения, объединения и вычитания множеств.

Булевы операторы (Boolean operators)

Логические операторы, позволяющие создавать логические выражения: "И", "ИЛИ", "НЕ". Используются при составлении сложных запросов к информационно-поисковым системам.

Векторно-пространственная модель (Vector Space Model, VSM)

Модель информационного поиска, рассматривающая документы и запросы как векторы в пространстве слов, а релевантность — как расстояние между ними.

Вероятностная модель (Prohibited Model)

Это модель информационного поиска, рассматривающая релевантность как вероятность соответствия данного документа запросу на основании вероятностей соответствия слов данного документа идеальному ответу.

Вертикальный Web-портал, вортал (Vertical Web-portal, vortal)

Специальные Web-сайты, к которым обращаются пользователи, интересующиеся какой-либо определенной темой. Слово "вертикальный" подразумевает отношение к одной тематике. Технология вертикальных порталов — "порталов" — получает все большее распространение в корпоративных сетях.

Весовой коэффициент (Weighting)

Коэффициент, приписываемый лексической единице в документе и учитываемый при вычислении числового значения релевантности. Весовой коэффициент может зависеть от расположения лексической единицы в документе, абзаце, предложении. Кроме того, весовой коэффициент непосредственно зависит от смысла лексической единицы, ее соответствия тематике поисковой системы, частоты использования в документе. Весовые коэффициенты могут приписываться лексическим единицам как в индексе информационно-поисковой системы, так и в запросах пользователя.

Выделение основы слова (Stemming)

Обеспечивает возможность поиска слова не только в строго заданном виде, но и во всех его морфологических формах. Например, слову "программа" будут соответствовать "программе", "программный" и т.д.

Гиперсвязь (ссылка) (Hyperlink)

Связь между отдельными компонентами информации. Применяется для реализации ссылок, сделанных внутри одного объекта на другой объект. Ссылка, как правило, делается из объекта, размещенного на HTML-странице, на другой объект, который может находиться на произвольном FTP- или WWW-сервере.

Гипертекст (Hypertext)

Документы, содержащие связи с другими документами (или имеющие внутренние связи). Гипертекстовый документ представляет собой специальным образом размеченную текстовую информацию. При отображении гипертекстовых документов отдельные элементы текста могут служить ссылками на другие документы. Механизм ссылок, дополняющий текстовую информацию, является неотъемлемой частью гипертекста. Web-страницы, как правило, представляют собой гипертекстовые документы, написанные с использованием языка гипертекстовой разметки HTML.

Глубинный анализ данных (Data Mining)

1. Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности (G. Piatetsky-Shapiro, GTE Labs).
2. Data Mining — это процесс выделения (selecting), исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе (SAS Institute).

Глубинный анализ текста (Text Mining)

Text Mining — это алгоритмическое выявление прежде неизвестных связей и корреляций в уже имеющихся текстовых данных. Важная задача технологии Text Mining — извлечение из текста его характерных элементов или свойств,

которые могут использоваться в качестве метаданных документа, ключевых слов, аннотаций. Другая важная задача состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации. Технология Text Mining также обеспечивает новый уровень семантического поиска документов.

Глубинный, скрытый Web (Deep Web)

Кроме видимой для поисковых систем части WWW-пространства, существует огромное количество страниц, которые ими не охватываются. Как правило, эти Web-страницы доступны в Internet, однако выйти на них невозможно, если не знать точного адреса. Эти ресурсы имеют собственное название — Deep Web. В их число входят и динамически формируемые Web-страницы, содержание которых хранится в базах данных и доступно лишь по запросам пользователей.

Дайджест (Digest)

Аннотированный текст, адаптированный к проблемной области пересказ или экстрагирование основных фактов из нескольких документов.

Добыча знаний (Knowledge Mining)

Это метод получения знаний в форме композиции новых фактов на основе анализа неформализованных и неструктурированных информационных массивов.

Закон Зипфа

Одна из основных закономерностей, широко применяемая в теории и практике информационного поиска. Профессор Джордж Зипф экспериментально показал, что если для какого-либо достаточно большого текста составить список всех используемых в нем слов, а затем ранжировать эти слова по частоте вхождения, то для любого слова произведение его порядкового номера-ранга в этом списке и частоты вхождения в тексте будет величиной постоянной.

Запрос (Query)

Определенным образом составленный набор слов, словосочетаний и служебных символов, характеризующий информацию, которую хочет найти пользователь.

Запрос по примеру (поиск подобного)
Query-By-Example (find similar)

Запрос на поиск документов, релевантных найденным ранее. Каждая информационно-поисковая система, в которой предусмотрена данная функция, интерпретирует и реализует такой запрос по-своему. При этом значительно облегчается процесс формирования поискового запроса, однако результат его обработки часто бывает непредсказуемым.

Индекс ИПС, Индекс (Index IRS)

Индекс информационно-поисковой системы представляет собой определенным образом организованную совокупность данных, где содержатся поисковые образы всех документов базы данных. Является основной составляющей архитектуры информационно-поисковой системы, обеспечивающей возможность оперативного поиска и доступа к информации.

Информационно-поисковая система
IRS (Information Retrieval System)

Система, предназначенная для обеспечения поиска и отображения документов, представленных в базах данных. Ядром информационно-поисковой системы

(ИПС) является поисковый механизм — программный модуль, который осуществляет поиск по запросу. ИПС, интегрированные с Web-технологиями, являются основой построения информационно-поисковых Web-серверов.

Классификация (Classification)

Процесс определения принадлежности информационного ресурса к предопределенным категориям.

Кластеризация (Clusterization)

Один из методов анализа данных, позволяющих упорядочивать многомерные наблюдения, например документы в пространстве слов или рубрик. Целью кластеризации является образование групп схожих между собой объектов, порождение новых классов для последующей классификации.

Ключевое слово (Keyword)

1. Отдельный термин, используемый в запросах к информационно- поисковым системам.
2. Дескриптор, отдельное слово или словосочетание, используемое при ручном или автоматизированном индексировании документов перед погружением в ИПС.

Консорциум W3C (World Wide Web Consortium, W3C)

Международный индустриальный консорциум, образованный в 1994 году первоначально в рамках CERN при поддержке DARPA и Европейской комиссии. В настоящее время W3C поддерживается совместно Лабораторией информатики Массачусетсского технологического института (США), INRIA (Франция) и университетом Кейо (Япония). Целью создания W3C была разработка общих протоколов, позволяющих расширить доступность и эффективность ресурсов World Wide Web, а также руководство эволюцией системы. В задачу W3C входит, прежде всего, разработка рекомендаций по новым технологиям, а также реализующих их спецификаций, имеющих статус стандарта консорциума.

Контент (Content)

Содержательная часть информационных ресурсов.

Контент-анализ (Content analys)

Метод получения выводов путем анализа содержания текстовой информации. Чаще всего реализуется как систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника.

Контент-мониторинг (Content monitoring)

Систематическое, непрерывное во времени сканирование и контент-анализ информационных ресурсов.

Концептуальный поиск (Concept search)

Поиск документов, имеющих прямое отношение к указанному поисковому термину, а не просто содержащих его.

Латентно-семантический анализ (Latent Semantic Analysis, LSA)

Это теория и метод извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных. Латентно-

семантический анализ основывается на идее, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые в значительной степени позволяют определить похожесть смысловых значений слов и множеств слов между собой. В качестве исходной информации LSA использует матрицу "термы-на-документы", содержащую частоты использования данного терма в данном документе.

Лемматизация (Lemmatization)

Реконструкция основной формы изменяемых частей речи, приведение слов к исходной (канонической) форме — лемме. Если это существительное — то к номинативу (именительному падежу), если глагол — к инфинитивной форме, и т.д.

Логический поиск (Boolean search)

Поиск информации с использованием логических (булевых) операторов.

Матрица близости между документами (Similarity Matrix)

Матрица, отражающая количественную взаимосвязь отдельных документов на основании одновременного вхождения в эти документы одних и тех же лексических единиц (слов).

Мониторинг (Monitoring)

Непрерывный во времени процесс сканирования информационных ресурсов с целью их дальнейшего содержательного анализа.

Операторы сравнения (Comparison Operators)

Математические операторы, используемые для сравнения значений в поисковых выражениях. Применяются при формировании запросов, содержащих данные, представленные в виде чисел или дат либо возможных границ их изменения (интервалов).

Пертинентность (Pertinent)

Характеристика степени соответствия содержания документа задаче исследования.

Поиск по ключевым словам (Keyword search)

Поиск документов, которые содержат указанные пользователем ключевые слова.

Поиск по словосочетаниям (Phrase search)

Поиск документов, которые в точности содержат указанное пользователем словосочетание или фрагмент текста.

Поиск с расстоянием (Proximity search)

Поиск, при котором пользователь указывает, на каком "расстоянии" друг от друга должны располагаться ключевые слова в искомых документах. Обычно расстояние измеряется в количестве отдельных слов, расположенных между ключевыми словами. При поиске может указываться как максимально возможное расстояние, так и диапазон.

Поисковое устройство (Search Appliance)

Информационно-поисковая система, поставляемая заказчику в виде автономного аппаратно-программного комплекса, как правило сервера с установленным программным обеспечением ИПС.

Поисковый Web-сервер (Retrieval Web-server)

Web-сервер, предназначенный для поиска информации в Internet. Как правило, запрос на поиск информации указывается в виде выражения, содержащего ключевые слова, путем заполнения простой или расширенной формы запроса. Полученный от поискового сервера результат представляет собой отсортированный список адресов Web-страниц, формально удовлетворяющих поисковому запросу.

Поисковый механизм (Search Engine)

Основной компонент любой информационно-поисковой системы. Программный модуль, осуществляющий поиск в базе данных по запросу (поисковому предписанию) пользователя.

Полнота, охват (Recall)

Отношение количества релевантных документов в отклике информационно-поисковой системы к общему количеству релевантных документов в исходном массиве.

Полнотекстовая поисковая система (Full-text search engine)

Информационно-поисковая система, которая при составлении индекса охватывает все слова в тексте документа (иногда за исключением стоп-слов) и учитывает порядок их расположения по отношению друг к другу.

Портал знаний (Knowledge Portal)

Портал знаний предприятия — это результат эволюции портала под влиянием современной концепции управления знаниями. Портал знаний не только предоставляет средства доступа к информации, но и позволяет пользователям взаимодействовать друг с другом, помогая связывать информацию с коллективным пониманием. Портал знаний корпорации дает возможность принимать оптимальные решения, поскольку сочетает приобретенные знания с информацией и служит "самодокументирующимся" центром обучения.

Программный агент (Digital agent)

Программный объект, который выполняет некие упреждающие действия в соответствии с задачами, делегированными человеком. В частности Web-агенты или роботы информационно-поисковых серверов в Internet занимаются сканированием информации с Web-сайтов, маршрут к которым они определяют по специальным алгоритмам.

Проект МАРК (MARC)

Проект, который был начат в 1966 году 16 библиотеками США для разработки стандарта формата обмена библиографическими записями в электронном виде. В 1972 году модернизированный стандарт МАРК-2 получил международное признание.

Протокол (Protocol)

Совокупность семантических и синтаксических правил, которые определяют поведение компьютеров и программ во время их взаимодействия в сети. Область действия протоколов простирается от регламентации порядка посылки битов по физическим линиям до форматов сообщений электронной почты.

Ранжирование (Ranking)

Упорядочение результатов поиска — отклика поисковой системы — по некоторым критериям, например по дате публикации документов или по релевантности.

Расширенный запрос (Query Expansion)

Позволяет детализировать поисковый запрос и области поиска, задаваемые для информационно-поисковой системы. Например, может указывать диапазон дат документов, язык представления информации, тип и т.д. Применяется для более точного задания критериев поиска.

Релевантность (Relevancy)

Мера того, насколько точно документ, найденный информационно-поисковой системой, отвечает запросу пользователя. Обычно выражается в числовой форме. Единых взглядов на это понятие нет. Далеко не всегда документ, отмеченный информационно-поисковой системой как наиболее релевантный по формальным признакам, будет таковым по мнению самого пользователя.

Робот, (Spider, Crawler, Bot, Robot)

Неотъемлемая составляющая поисковой системы в Internet. Согласно заданному сценарию, посещает Web-страницы, копирует и индексирует полностью или частично их содержимое и далее следует по ссылкам, найденным на данной странице. Информация, полученная роботом в результате обхода серверов Internet, заносится в индекс информационно-поисковой системы.

Семантический Web (Semantic Web)

Проект консорциума W3C, в рамках которого предлагается способ сделать информацию в Сети более доступной, что, в свою очередь, позволит создать интеллектуальное программное обеспечение, которое могло бы искать в Web необходимые данные, выявляло их семантику, создавало перекрестные ссылки и использовало эти данные для решения практических задач. Одина из основных концепций Семантического Web — ориентация на формат XML.

Сервер баз данных (Database Server)

Система программного обеспечения, имеющая средства обработки на языке баз данных. Обеспечивает выполнение различных операций, таких как создание, модификация, извлечение и другие по отношению к данным, содержащимся в базах данных.

Синдикация контента (Content syndication)

Под синдикацией в данной книге понимаются технологии сбора информации в Internet и последующее распространение ее фрагментов в соответствии с потребностями пользователей. Службы синдикации обеспечивают одновременную публикацию одних и тех же данных на различных страницах, сайтах и мобильных устройствах (в том числе в карманных компьютерах и мобильных телефонах), а также доставку информации пользователям.

Система управления базами данных, СУБД (Database management system, DBMS)

Комплекс программных и лингвистических средств общего или специального назначения, реализующий поддержку создания баз данных, централизованного управления и организации доступа к ним различных пользователей в условиях принятой технологии обработки данных.

Спам (SPAM)

Спам — это непрошеное рекламное сообщение, сетевой мусор, рассылаемый по электронной почте в личные почтовые ящики или телеконференции. Рассылка спама считается нарушением этикета и правил применения компьютерных сетей.

Стандарт Z39.50

Информационно-поисковый протокол для библиографических систем. Положен в основу службы поиска распределенной информации в Internet — WAIS.

Стоп-слова (Stop words)

Слова, исключаемые из индекса системы и/или запроса пользователя. Отдельные информационно-поисковые системы, для сокращения размеров индекса и увеличения производительности, не включают в индекс часто встречаемые на Web-страницах слова. К стоп-словам обычно относятся предлоги, междометия и другие сочетания, которые не несут содержательного смысла.

Сценарий (Script)

Обычно под сценарием понимается набор команд на интерпретирующем языке программирования. Термин используется применительно к короткому интерпретирующемуся коду, написанному на таких языках, как JavaScript, Perl и др.

Тезаурус (Thesaurus)

Словарь лексических единиц (ключевых слов), основанный на лексике естественного языка и отражающий семантические отношения между лексическими единицами. Используется в информационно-поисковых системах для уточнения сферы действия поискового запроса.

Формат Atom

Формат, открытый стандарт для генерации и доставки срочных сообщений подписчикам он-лайновых журналов, близкий по функциональности к RSS. Atom совершенствуется компаниями IBM и Google. Окончательно не утвержденный формат Atom обеспечивает подписчикам большую, чем RSS, гибкость, поскольку поддерживает больше метаданных.

Формат OPML

OPML — это формат для хранения списка RSS-фидов. Он разработан на основе XML для того, чтобы обеспечить удобную подписку с помощью RSS-агрегаторов сразу на несколько фидов.

Чувствительность к регистру (Case sensitivity)

Обычно применяется по отношению к поисковым запросам. Поисковые системы, чувствительные к регистру, различают заглавные и строчные буквы в терминах поискового выражения. При использовании чувствительной к регистру поисковой системы эти термины будут восприниматься неодинаково.

Литература

1. *Автоматизация* индексирования и реферирования документов // Информатика. Сер. "Итоги науки и техники". — М.: ВИНИТИ, 1983. — Т.7. — 246 с.
2. *Блюменау Д.И., Гендина Н.И., Добронравов И.С., Лахути Д.Г., Леонов В.П., Федоров Е.Б.* Формализованное реферирование с использованием словесных клише (маркеров) // НТИ. — Сер. 2. — 1981. — № 2. — С. 16–20.
3. *Боровикова О. И., Загорулько Ю. А.* Организация порталов знаний на основе онтологий // Российский НИИ Искусственного Интеллекта, Институт систем информатики СО РАН (http://www.dialog-21.ru/archive_article.asp?-param=7527&y=2002&vol=6078).
4. *Дудихин В.В.* Конкурентная разведка в Интернет // АСТ, 2004. — 229 с.
5. *Гордиенко И.* Sancta simplicitas или... Корпоративные поисковые машины // CIO. — 2002. — №6-7 (http://www.cio-world.ru/offline/2002/6/19864).
6. ГОСТ 7.52-85. Система стандартов по информации, библиотечному и издательскому делу. Коммуникативный формат для обмена библиографическими данными на магнитной ленте. *Поисковый образ документа.* — Москва, 1985.
7. ГОСТ 7.24-90. Система стандартов по информации, библиотечному и издательскому делу. *Тезаурус информационно-поисковый многоязычный.* Состав, структура и основные требования к построению. — Москва, 1990.
8. *Горькова В.И., Борохов Э.А.* Реферат в системе научной коммуникации. Направления совершенствования лингвистических и структурных характеристик // Информатика. Сер. "Итоги науки и техники". — М.: ВИНИТИ, 1987. — Т.11. — 232 с.
9. *Граммер Дж.* Портал знаний предприятия // DM Review, 2000, март (http://www.it2b.ru/it2b2.view8.page7.html).
10. *Григорьев А.Н., Ландэ Д.В.* NEW MEDIA — новая информационная среда // Сети и телекоммуникации. — 2000. — № 4. — С. 18–22
11. *Гусев В.С.* Освоение Internet: Самоучитель. СПб: Диалектика /Вильямс, 2003. — 304 с.
12. *Гусев В.С.* Поиск в Internet: Самоучитель. СПб: Диалектика /Вильямс, 2004. — 336 с.
13. *Кириченко К.М, Герасимов М.Б.* Обзор методов кластеризации текстовых документов // Материалы международной конференции Диалог'2001, (http://www.dialog-21.ru/Archive/2001/volume2/2_26.htm).
14. *Кларк Д.* Закон Мура останется в силе // Ведомости. — 2003. — №11 (http://www.silicontaiga.ru/home.asp?artId=2066).
15. *Ландэ Д.В.* Агенты новостей в сети Интернет // CHIP/Украина. — 2001. — №5. — С. 108–111.
16. *Ландэ Д.В.* Добыча знаний // Телеком. — 2004. —№1-2. — С. 36–42.
17. *Ландэ Д.В.* Интернет-старатели // Мир связи. — 2002. — №8. — С. 38–42.
18. *Ландэ Д.В.* Искать и не сдаваться // CHIP/Украина. — 2004. — №5. — С. 84–87.

19. *Ландэ Д.В.* Ловцы данных // CHIP/Украина. — 2004. — №3. — С. 72–75.
20. *Ландэ Д.В.* Мобильный информатор // CHIP/Украина. — 2004. — №2. — С. 80–83.
21. *Ландэ Д.В.* На границе стихий // CHIP/Украина. — 2003. — №5. — С. 72–77.
22. *Ландэ Д.В.* Навигация в Сети: каталоги — поисковики — порталы // InternetUA. — 2000. — №1. — С. 43–47.
23. *Ландэ Д.В.* О чем говорят запросы пользователей к поисковым серверам // Сети и телекоммуникации. — 1999. — №4. — С. 19–21.
24. *Ландэ Д.В.* Поисковые системы: поле боя — семантика // Телеком. — 2004. — №4. — С. 44–50.
25. *Ландэ Д.В.* Проклятье сети // Мир связи. — 2002. — №12. — С. 46–50.
26. *Ландэ Д.В.* Эффективный сбор новостей // InternetUA. — 2003. — №9. — С. 16–19.
27. *Ландэ Д.В.* WAP: прибытие вовремя // CHIP/Украина. — 2002. — №3. — С. 86–90.
28. *Ландэ Д.В., Зубок В.Ю.* Информационно-поисковый сервер InfoReS для работы в среде WWW // Компьютеры плюс программы. — 1996. — №5. — С. 65–69.
29. *Ландэ Д.В., Литвин А.Б.* Феномены современных информационных потоков // Сети и бизнес. — 2001. — №1. — С. 14–21.
30. *Ландэ Д.В., Морозов А.Ю.* Редкостный Синтез Сайтов // Мой компьютер. — 2003. — **248** №25 . — С. 44–46.
31. *Ландэ Д.В, Морозов А.Ю.* Читайте новости, батенька! // CHIP/Украина. — 2004. — №7. — С. 82–85.
32. *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Основы информатики — М.: Наука, 1968. — 756 с.
33. *Моуд Дж.* Готовность к взлету // PC Week-Россия. — 2002. — №9 (`http://www.pcweek.ru/Year2002/N9/CP1251/Strategy/chapt2.htm`).
34. *Печенкин И.А.* Информационные технологии на службе разведки. Обзор современных программных средств обеспечения принятия управленческих решений. Защита информации // Конфидент. — 2004. — № 4. — С. 2–15.
35. *Питц-Моултис Н., Кирк Ч.* XML / Пер. с англ. — СПб.: БХВ-Петербург, 2001. — 736 с.
36. *Попов А.* Поиск в Интернете — внутри и снаружи // Intrnet. — 1998. — №2 (`http://www.citforum.ru/pp/search_03.shtml`).
37. *Хан Удо, Мани Индерджиет.* Системы автоматического реферирования. (`http://www.osp.ru/os/2000/12/067.htm`).
38. *Хант Ч., Зартарьян В.* Информация — основа успеха: разведка на службе вашего предприятия. — Киев: "Укрзакордонвизасервис", 1992. — 160 с.
39. *Чурсин Н.Н.* Популярная информатика. — К.: Техника, 1982. — 158 с.
40. *Тан И Цзе.* Цифровые агенты меняют мир программ // Computerworld. — 2001. — №43 (`http://www.osp.ru/cw/2001/43/038_1.htm`).
41. *Auerbach F.* Das Gesetz der Bevolkerungskonzentrationen // Peterman's Mittelungen. — 1913. — V.59. — P. 74–76.

42. *Avram H.D., Knapp J.F., andRather L.J.* The MARC II Format: A Communications Format for Bibliographic Data // Library of Congress, Washington, D.C.,Jan. 1968.

43. Tim Berners-Lee, James Hendler, Ora Lassila, (http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2), Scientific American, May 2001 (http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21).

44. *Brin S., Page L.* The anatomy of a large scale hypertextual web search engine // Proc. 7th WWW, 1998 (http://www-db.stanford.edu/pub/papers/google.pdf).

45. *Broder A., Henzinger M.* Algorithmic aspects of information retrieval on the web. Source Handbook of massive data. Kluwer Academic Publishers Norwell, MA, USA, 2002. — P. 3–23.

46. *Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener.* Graph structure in the web (http://www.almaden.ibm.com/cs/k53/www9.final).

47. *Chakrabarti Soumen.* Mining the web. Discovery knowledge from hypertext data // Publisher: Morgan Kaufmann, 2002. — 344 p.

48. *Goertzel B.* Meaning is a fuzzy Web of patterns: Semiotics/autonomy feedback in the Webmind Internet AI system // Proceedings of the 1998 IEEE International Symposium on Intelligent Control, Piscataway, NJ, USA,98CH36262, 1998. — P. 689–693.

49. ***Google*** Search Appliance for Intranets // Google Inc., 2004 (http://www.google.com/appliance/pdf/ds_GSA_intranets.pdf).

50. *Graham P.* A Plan for SPAM, 2002 (http://www.paulgraham.com/spam.html).

51. *Graham P.* Better Bayesian Filtering, 2003 (http://www.paulgraham.com/better.html).

52. *D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A Tomkins, J. Zien.* How to build a WebFountain: an architecture for very large-scale text analytics // IBM Systems Journal, March, 2004.

53. *Hahn U., Reimer U.* Knowledge-Based Text Summarization: Salience and Generalization Operators for Knowledge-Based Abstraction / Advances in Automatic Text Summarization, I. Mani and M. Maybury, eds. // MIT Press, Cambridge, Mass., 1999. — P. 215–232.

54. *Kleinberg Jon.* Bursty and Hierarchical Structure in Streams // Data Mining and Knowledge Discovery, October 2003. — Volume 7 Issue 4.

55. *Landauer T.K., Foltz P.W., Laha, D.* An introduction to latent semantic analysis // Discourse Processes, 1998. — Volume 25. — P. 259–284.

56. *Lasswell, Harold D., Nathan Leites, and Associates.* The Language of Politics: Studies in Quantitative Semantics // New York: George Stewart Publisher, 1949.

57. *Lyman P., Varian Hal R.* How much information 2003? (http://www.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf).

58. *Lynch M.* Autonomy Knowledge Server // Neurodynamics Ltd, Cambridge UK, 1998.

59. *Mack R., Ravin Y., Byrd R.J.* Knowledge portals and the emerging digital knowledge workplace // IBM Systems Journal, Dec. 2001.

60. http://www.xml.com/pub/au/164. What is RSS? 2002. (http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html).

61. *Mark T. Maybury.* Extraction of Knowledge from Unstructured Text // MITRE Corporation, 2001 (http://www.mitre.org/work/tech_papers/tech_papers_01/maybury_unstructured/ maybury_unstructured.pdf).

62. *Eric Miller, Ralph Swick, Dan Brickley, Brian McBride, Jim Hendler, Guus Schreiber, Dan Connolly.* Semantic Web. (http://www.w3.org), (http://www.csail.mit.edu), ERCIM, (http://www.keio.ac.jp), (http://www.w3.org/2001/sw).

63. *Paijmans Hans.* Indexing Texts with Smart // — Linux Gazette, 1997. Issue 13.

64. *Pareto V.* Cours d'economie politique // Rouge, Lausanne et Paris, 1897.

65. *Quin Liam.* Extensible Markup Language (XML) (http://www.w3.org/XML).

66. *RFC 1625 — WAIS over Z39.50-1988.* Network Working Group Request for Comments: 1625. M. St. Pierre, J.Fullton, K.Gamiel, J.Goldman, B.Kahle, J.Kunze, H.Morris, F.Schiettecatte, 1994 (http://www.faqs.org/rfcs/rfc1625.html).

67. *Salton G. et al.* Automatic Text Structuring and Summarization // Information Processing & Management. — 1997. — V. 33. — №2. — P. 193–207.

68. *Salton G., Buckley C.* Improving retrieval performance by relevance feedback // Journal of the American Society of Information Science. — 1990. — 4 41: — P. 288–297.

69. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing and Management. — 1988. — 24 : — P. 513–523.

70. *Salton G., McGill M.J.* Introduction to Modern Information Retrieval // New York [etc.] : McGraw-Hill, 1983.

71. *Salton G., Wong A., Yang C.* A Vector Space Model for Automatic Indexing // Communications of the ACM, 1975. — 18(11) : — P. 613–620.

72. *Steele B., Kleinberg J.* Buzzwords of history show the way to Web searches. (http://www.news.cornell.edu/Chronicle/03/AboutChron.html), 2003 (http://www.news.cornell.edu/Chronicle/03/2.20.03/AAAS.Kleinberg.html).

73. *Text Mining* With Oracle Text // Oracle White Papers (http://www.oracle.com/technology/products/text/pdf/10gR1text_mining.pdf).

74. *The Deep Web:* Surfacing Hidden Value // BrightPlanet.com LLC, 2000. — 35 p. (http://www.dad.be/library/pdf/BrightPlanet.pdf).

75. *The Z39.50* Information Retrieval Standard. Part I: A Strategic View of Its Past, Present and Future / Clifford A. Lynch, D-Lib Magazine, April 1997, ISSN 1082–9873 (http://www.dlib.org/dlib/april97/04lynch.html).

76. *Yang Y. and Pederson J.* Feature selection in statistical learning of text categorization // In Proc. of the ICML'97. — 1997. — P. 412–420.

77. *Zipf G.K.* Human Behavior and the Principle of Least Effort // Addison-Wesley, Cambridge: Univer. Press, 1949.

Предметный указатель

K

L

M

N

O

P

R

S

T

U

Л

М

Н

О

П

Р

С

Т

Ф

Ц

Э

Я

Научно-популярное издание

Дмитрий Владимирович Ландэ

Поиск знаний в Internet
Профессиональная работа

Литературный редактор	*Е.Д. Давидян*
Верстка	*В.И. Бордюк*
Художественный редактор	*В.Г. Павлютин*
Корректоры	*Л.А. Гордиенко, О.В. Мишутина, Л.В. Чернокозинская*

Издательский дом "Вильямс".
101509, Москва, ул. Лесная, д. 43, стр. 1.

Подписано в печать 12.01.2005. Формат 70×100/16.
Гарнитура Times. Печать офсетная.
Усл. печ. л. 21,93. Уч.-изд. л. 19,38.
Тираж 3000 экз. Заказ № 75.

Отпечатано с диапозитивов в ФГУП "Печатный двор"
Министерства РФ по делам печати,
телерадиовещания и средств массовых коммуникаций.
197110, Санкт-Петербург, Чкаловский пр., 15.

ПРОФЕССИОНАЛЬНАЯ РАБОТА

Д.В. ЛАНДЭ

ПОИСК ЗНАНИЙ В INTERNET

Книга посвящена современным подходам к получению новых знаний на основе анализа информационного пространства сети Internet и методам обработки информационных потоков с целью выявления значимых тенденций, понятий, феноменов, их взаимосвязей.

Большое внимание в книге уделено новому направлению обработки текстовой информации — "глубинному анализу текстов" (Text Mining), объединяющему в себе технологические и методологические подходы контент-анализа, компьютерной лингвистики и искусственного интеллекта.

Книга ориентирована на широкий круг читателей, интересующихся современными информационными технологиями. При этом она будет полезна и аналитикам, которые с помощью инструментов Text Mining смогут повысить эффективность и качество своей работы.

ОБ АВТОРЕ

Дмитрий Владимирович Ландэ — кандидат технических наук, заместитель директора Информационного центра "ЭЛВИСТИ" (ElVisti). Автор более 100 публикаций по тематике информационных технологий. Руководитель нескольких Internet-проектов, научный руководитель и один из авторов информационно-поисковой системы InfoReS® и системы контент-мониторинга InfoStream®.

ISBN 5-8459-0764-0

05034

9 785845 907646

ДИАЛЕКТИКА

www.dialektika.com