Добыча знаний

Данные, хранящиеся в корпоративных сетях, зачастую являются основой для принятия важных решений, влияющих на работу или даже выживание компаний. Получение требуемого и отсечение информационного шума становятся определяющими для конкурентоспособности

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 исследование изменения объема информации в мире за год. Оно под руководством профессоров Питера Лаймана и Хола Вэриена с 2000 года проводится в Калифорнийском университете в Беркли. Ученые пришли к выводу, что на протяжении трех лет, предшествующих 2002 году, количество информации, произведенной человечеством, удвоилось. А в самом 2002 году в мире появились пять экзабайт (миллионов терабайт) новой информации. Для сравнения приведем данные об объеме фонда библиотеки Конгресса США, где хранятся 19 млн книг и 56 млн рукописей: он составляет около десяти терабайт.

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 вкупе с оптическими цифровыми носителями приходится лишь 7 % информации.

Инструменты для корпоративных массивов

Итак, на жестких дисках отдельных компьютеров или на серверах в корпоративной сети накапливаются огромные массивы документов, навигация в которых по понятным причинам затруднена. Для обеспечения комфортности работы с такими массивами документы обычно пытаются классифицировать, распределить их по тематическим папкам или каталогам. 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 в свое время широкое распространение в Интернете. В отличие от Сети, где данные в основном представлены как html-файлы, поиск производится в другой среде. Ведь в корпоративных системах преимущественно используются форматы офисных приложений и систем документооборота. Наряду с поиском большое значение приобретают задачи группировки тематически близких документов, автоматического реферирования, перевода, выявления ключевых понятий, проведения нечеткого поиска.

Средства поиска

Рассмотрим некоторые популярные системы поиска для корпоративных сетей.

Универсальная поисковая система mnoGoSearch (*www.mnogosearch.org*) предназначена для интернет- или интранет-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 способна хранить данные во всех популярных реляционных СУБД. Существуют версии для Linux и Windows.

Полнотекстовая персональная поисковая система «Ищейка» (*www.isleuthhound.com*) обладает возможностями поиска документов и файлов на русском и английском языках. Она воспринимает запросы во всех словоформах и с любыми падежными окончаниями (то есть поддерживает морфологический поиск) и способна автоматически распознавать основные кодировки текста – ASCII, ANSI, Unicode. 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 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, а также поиск по атрибутам файлов, накопление «популярных» запросов и т. п.

Серверный «Следопыт 1.5» (*www.medialingua.ru*) — мощная поисковая система, предоставляющая возможность поиска нужной информации на отдельном вэб-сайте или сервере корпоративной интрасети. 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 Microsoft Outlook (причем как сами сообщения, так и вложения). В системе реализован морфологический поиск, то есть для каждого слова учитывается парадигма. Фильтр для формата pdf при работе с русским языком является в «Следопыте» одним из лучших.

> Свыше 90 % носителей информации представлены магнитными носителями, преимущественно жесткими дисками

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).

Основное назначение программы Data Search 6.0 (*www.dtsearch.com*) – поиск информации на локальном компьютере. Система имеет английский интерфейс и работает под управлением операционных систем Windows 9x/Me/NT/2000. Она состоит из следующих модулей: dtSearch Desktop 6.0 – 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, а также форматы документов Microsoft Office (Word, Excel, PowerPoint) и WordPerfect. Поддерживается кодировка Unicode. Допускаются несколько видов поиска, а именно морфологический и фонетический поиск, а также поиск синонимов и поиск в словах с орфографическими ошибками.

Система полнотекстового поиска CROS 4.01 (*www.cronos.ru*) предназначена для накопления и обработки текстовых документов различных форматов. Хранение документов в базах данных системы обеспечивает умень-

ТЕЛЕКОМ-ИНФО

Информация в корпоративной среде

Большая часть информации приходится на компьютеры. Это позволяет утверждать, что в значительной мере информация хранится в корпоративных сетях. Поэтому не вызывает удивления ощутимое на рынке высоких технологий присутствие продуктов для построения информационных хранилищ. Последние ориентируются на интеграцию данных, представленных в сетях организаций, а также на интеллектуальный поиск и обобщение информации. Прежде всего требуется обработка неструктурированной информации (текстовые документы, электронные таблицы, сообщения электронной почты). Ведь по существующим оценкам, неструктурированные данные, главным образом текст, составляют не менее 90 % информации, с которой компаниям приходится иметь дело. И только 10 % приходится на структурированные данные, хранимые, как правило, в реляционных СУБД и в системах документооборота.

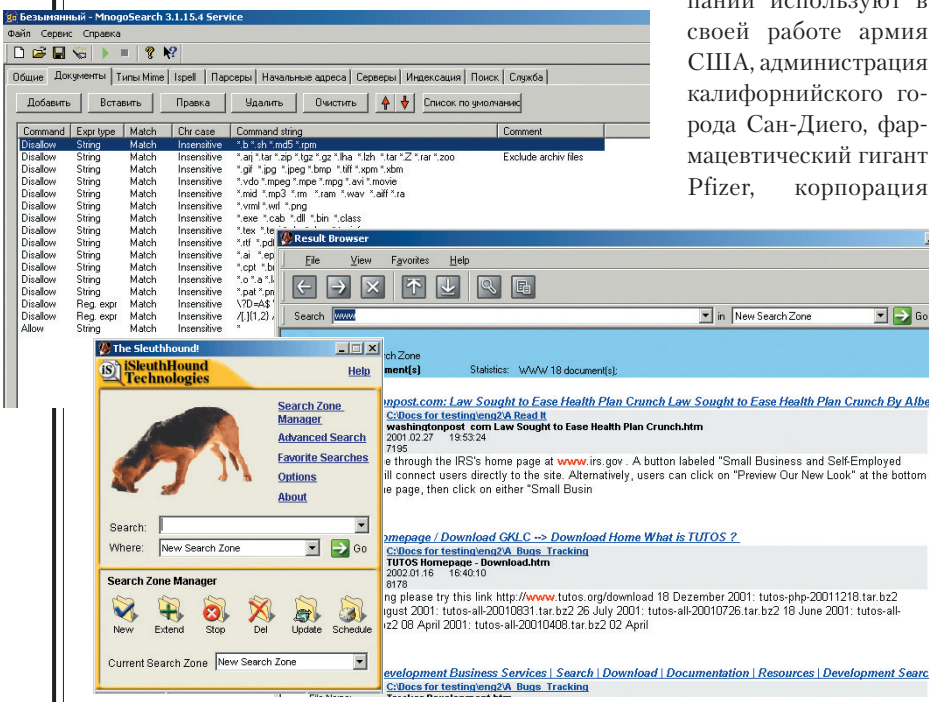
шение в два-три раза необходимого объема дисковой памяти. Предусмотрено автоматическое определение форматов документов Microsoft Word версий 6.0, 7.0, 97, 2000, а также rtf и html. Помимо этого определяется тип кодировки (DOS, Win, КОИ8, Unicode).

CROS обеспечивает навигацию по найденным документам, способен работать в локальной сети и поддерживает защиту информации от несанкционированного доступа. При этом отсутствуют ограничения на количество иерархических областей поиска, осуществляется сортировка найденных документов по дате, имени, типу, а также по атрибутам, которые задаются самим пользователем.

Система Greenstone (*www.greenstone.org*) представляет собой Open Source-решение для создания «цифровых библиотек». 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.

Программно-аппаратный комплекс Google Search Appliance обеспечивает поиск документов в рамках корпоративных сетей. Джон Пискителло, менеджер Google по продуктам, определил это устройство как «естественный шаг для компании, которая всегда 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.

Поисковые устройства этой компании используют в своей работе армия США, администрация калифорнийского города Сан-Диего, фармацевтический гигант Pfizer, корпорация Boeing, Procter & Gamble, Cisco Systems и др. Темпы роста продаж этих устройств за последний год составили 200 %.



Системы корпоративного поиска способны работать с широким разнообразием форматов, а также обладают возможностью организации большого количества зон поиска

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

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

Новый уровень обработки информации

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 добычи знаний (Knowledge Mining). Стоимость внедрения таких систем составляет сотни тысяч долларов.

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

Как правило, информационные массивы преобразуются такими системами в хранилище 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 для 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.

Обогатители знаний

Сегодня на рынке корпоративных систем все большую известность получает технология компании Autonomy (*www.autonomy.com*), которая позиционируется как инструментарий для автоматизированного управления информационными потоками. Основные научные принципы Autonomy базируются на информационной теории Клода Шеннона, байесовых вероятностях и нейронных сетях. Концепция адаптивного вероятностного моделирования позволяет системе Autonomy идентифицировать шаблоны в тексте документа и автоматически определять подобные шаблоны в массиве других документов.

Обрабатывая шаблоны строк в документах, система Autonomy определяет корреляцию образов и выявляет закономерности среди больших массивов документов. При этом не учитываются никакие специфичные правила (в том числе и лингвистические). Поскольку система не базируется на предопределенных ключевых словах, она может работать с любыми языками.

Одним из программных продуктов Autonomy является пакет Portal-in-a-box, который помимо традиционных функций агрегирования информации из разнородных источников имеет и средства для решения такой проблемы, возникающей при построении порталов, как систематизация неструктурированных данных. Очевидно, что группировка документов по категориям и создание их метаописаний требует немалых редакторских усилий. Portal-in-a-box в этом случае полностью автоматизирует процессы категоризации информации, ее реферирования и расстановки гиперссылок.

Несмотря на высокую цену внедрения (несколько сотен тысяч долларов), у Autonomy – свыше 400 корпоративных клиентов, в том числе и British Telecom, France Telecom, General Motors, Reuters, BBC, British Airways и др.

Информационно-поисковая система RetrievalWare (*www.convera.com*) представляет собой средство полнотекстового и атрибутивного поиска. К документам, с которыми RetrievalWare способна работать, 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 RetrievalWare позволяет работать с системой как через корпоративную локальную сеть, так и через Интернет. Серверная часть системы поддерживает все распространенные серверные платформы, а клиентским местом может быть любой компьютер, имеющий графический вэб-браузер. Система обладает возможностью работы в различных многопроцессорных и распределенных многосерверных конфигурациях.

Источником информации может быть файловая система, системы уп-

> Задача – выявить в неструктурированных данных знания, чтобы использовать их в процессе принятия решений

ТЕЛЕКОМ-ИНФО

Разработка информационных ресурсов

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

Александр Линден, директор компании Gartner Research, выделил четыре основных вида приложений технологий Text Mining:

✓ Классификация текста, в которой используются статистические корреляции для построения правил размещения документов в предопределенные категории. В современных системах классификация применяется, например, в таких задачах: группировка документов в сетях интранет, размещение документов в определенные папки, избирательная доставка новостей подписчикам.

✓ Кластеризация, базирующаяся на признаках документов, использующая лингвистические и математические методы без использования предопределенных категорий. Кластеризация широко применяется при реферировании больших документальных массивов, определении взаимосвязанных групп документов, для упрощения визуализации информации, выявления дубликатов или близких по содержанию документов.

✓ Семантические сети или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения навигации. Используемая при этом визуализация является ключевым звеном при представлении схем неструктурированных текстовых документов. Она используется как средство представления контента всего массива документов, а также для реализации навигационного механизма, который может применяться при исследовании документов и их классов.

✓ Извлечение фактов состоит в получении некоторых фактов из текста с целью улучшения классификации, поиска и кластеризации.

Можно назвать еще несколько задач технологии Text Mining, например, прогнозирование и нахождение исключений, то есть поиск объектов, которые своими характеристиками выделяются из общей массы. Все эти задачи находят свое воплощение в современных корпоративных хранилищах.

Решения для глубинного анализа текстов и обнаружения в них новых знаний базируются на оценке смысла слов естественного языка и определении связей между понятиями, обозначаемыми этими словами

равления базами данных (MS SQL, ORACLE, Sybase, прочие СУБД), почтовые системы (Microsoft Exchange, Lotus Notes и т. п.), системы управления документами (Documentum EDMS, FileNET Panagon и т. п.), узлы корпоративной сети и Интернета, а также электронный архив Excalibur File Room – средство организации доступа к бумажным документам.

Лежащая в основе системы технология адаптивного распознавания образов базируется на нейронных сетях для обработки информации и действует как самоорганизующаяся система, которая выделяет в массиве хранимой информации и индексирует бинарные образы. К преимуществам применения этой технологии для поиска текстовой информации можно отнести осуществление нечеткого поиска, языковую независимость, малые объемы индексных файлов.

Основой технологии семантического поиска является использование семантических сетей, описывающих смысл слов естественного языка и связи между обозначаемыми ими понятиями. Реализована также поддержка русской морфологии. Семантическая сеть словаря этого языка включает в себя около 40 тыс. семантических групп в базовом варианте. Это позволяет пользователю вводить запрос на естественном языке, предоставив системе самой искать все документы, контекст которых совпадает с контекстом запроса. Применение семантики позволяет учитывать общий контекст документа.

Модуль аннотирования в системе RetrievalWare, который позволяет строить аннотации документов в виде связного текста, построен на базе сервера аннотирования ML NetLibretto компании «Медиалингва».

В список компаний и организаций, пользующихся этой системой, входят ABC News, Encyclopedia Britannica, Microsoft, Sun Microsystems, Всемирный банк, ФАПСИ, Центральный Банк России, «Лукойл» и др.

Яndex.Server Standard 3.0 (*www.yandex.ru*) 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 сетях. 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 при чтении локальной файловой системы.

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 3.0 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.

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

Решение PolyAnalyst российской компании «Мегапьютер» (*www.megaputer.ru*) – это система, предназначенная для автоматического и полуавтоматического анализа числовых и текстовых баз данных с целью обнаружения в них ранее неизвестных, нетривиальных, практически полезных и доступных пониманию закономерностей, которые необходимы для принятия оптимальных решений в бизнесе и в других областях человеческой деятельности.

По своей природе PolyAnalyst является клиент/серверным приложением. Пользователь работает с клиентской программой PolyAnalyst Workplace. Математические 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, категориальные переменные, текстовые строки, даты, а также свободный текст. PolyAnalyst может обрабатывать исходные данные из различных источников, таких как файлы Microsoft Excel 97/2000, любая ODBC-совместимая СУБД, SAS data files, Oracle Express, IBM Visual Warehouse.

Модули PolyAnalyst используют различные алгоритмы Data и Text Mining, в том числе модуль Text Categorizer – каталогизатор, который позволяет автоматически создать иерархический древовидный каталог имеющихся текстов и пометить каждый узел этой древовидной структуры как наиболее индикативный для относящихся к нему текстов.

Модуль Link Terms обеспечивает связь понятий. Он позволяет выявлять связи между понятиями, встречающимися в текстовых полях изучаемой базы данных, и представлять их в виде графа, который может быть использован для выделения записей, реализующих выбранную связь. Модуль Link Analysis выявляет корреляционные и антикорреляционные связи между значениями категориальных и булевых полей.

Благодаря уникальной технологии «эволюционного программирования» и другим интеллектуальным алгоритмам, PolyAnalyst с успехом применяется в различных типах бизнес-задач, в социологических исследованиях, в прикладных научных и инженерных задачах, в банковском деле, в страховании и медицине.

PolyAnalyst получил широкое распространение в мире, среди ее пользователей Boeing, 3M, Chase Manhattan Bank, Dupont, Siemens.

Ядром механизма обработки контента InfoStream (*infostream.com.ua)* является полнотекстовая информационно-поисковая система InfoReS. Технология InfoStream позволяет создавать полнотекстовые базы данных и осуществлять поиск информации, формировать тематические информационные каналы, автоматически рубрицировать информацию, формировать дайджесты, таблицы взаимосвязей понятий (относительно встречаемости их в сетевых публикациях), гистограммы распределения весовых значений отдельных понятий, а также динамики их встречаемости по времени. С помощью InfoStream можно 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 таких операционных систем, как FreeBSD, Linux, Solaris.

Технологии InfoStream позволяют создать комплекс поддержки документального информационного хранилища, в котором реализуется интегрированная информационно-поисковая среда на основе вэб-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 помо-

щью механизма «информационных портретов».

Порталы знаний

По данным недавно проведенного исследования, сотрудники компаний могут тратить до трех часов в день на поиск информации, который зачастую оказывается безрезультатным. Вследствие этого тысяча крупнейших фирм США ежегодно теряет $2,5 млрд.

Именно для решения данной проблемы созданы и продолжают создаваться корпоративные поисковые системы и порталы знаний, представляющие среду для эффективного поиска и обмена знаниями. Это инструменты, представляющие собой совокупность технологических решений для выявления, хранения, классификации, обработки и распространения знаний.

В настоящее время широко используется система Lotus Discovery Server – программный продукт, предназначенный для управления знаниями в корпоративных порталах. Он предполагает нахождение и идентификацию связей, а также общее управление интеллектуальным капиталом. Благодаря возможности анализа информации, хранящейся в организации, Lotus Discovery Server может определять области экспертных знаний и подразумеваемые знания сотрудников, находя и организуя динамические связи между информацией, людьми и их деятельностью.

Современные порталы управления знаниями обеспечивают решение целого комплекса задач, среди которых сбор информации об объектах, определение связи объектов, выявление тенденций. Функциональные возможности таких систем позволяют 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.

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 просили. Это – одно из первых эффективных решений на базе современной технологии глубинного анализа текстов.

Очень близким к Avalanche по идеологии является подход компании Vivisimo, в рамках которого результаты интернет-поиска распределяются по папкам-категориям, автоматически создаваемым системой. Достигается это за счет лексического сопоставления запросов и результатов поиска.

Естественно, свое применение Vivisimo сразу же нашла в корпоративных сетях и вэб-сервисах. Рауль Валдес-Перес, один из учредителей Vivisimo, сравнил систему с очень умным библиотекарем, который мгновенно находит нужную книгу в море неупорядоченной информации.

Перспективы обработки информации

Сегодня данные, представленные на компьютерах корпоративных сетей, зачастую являются основой для принятия важных решений, влияющих на работу или даже на выживание компаний. Интенсивная информатизация государственных органов и коммерческих структур, растущая доступность инструментария для сбора и мониторинга данных ведут к изобилию информации, в котором может утонуть работа практически любой организации.

Эффективный поиск, вовремя предлагающий необходимые сведения, сопровождаемый способностью избавляться от информационного шума, оказывается решающим фактором для повышения конкурентоспособности. Без поисковых систем, систем анализа текстов и систем добычи знаний любые серьезные информационные начинания завтра будут обречены на провал.

Естественно, эти технологии широко используются «силовиками». В прошлом году свои технологии «добычи данных», применяемые для поиска информации в текстах, радио- и телепередачах, публично представило ЦРУ. Оказалось, что объектами поиска спецслужбы являются тексты, опубликованные 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. Методика позволяет выделять из аудиопотока только те голоса или ту конкретную информацию, которая заложена в настройках поиска. Еще одна технология, Fluent, позволила ЦРУ искать информацию в текстовых документах, причем запрос вводится на английском языке и тут же переводится на целый ряд других языков, а найденная информация из базы данных на разных языках поступает исследователю после автоматического перевода.

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 продаж такого ПО до $1,5 млрд (в 2002 году – $540 млн). А в 2006 году такие системы будут доминировать при проведении анализа информации от клиентов в компаниях любого уровня, будь то контакт-центры и службы поддержки, интернет-агентства или аналитические агентства. ●

Без поисковых систем, систем анализа текстов и систем добычи знаний любые серьезные информационные начинания обречены на провал

Дмитрий Ландэ,
Dwl@visti.net

ТИЖДЕНЬ ШИФРОВИХ ТЕХНОЛОГІЙ

17-20 лютого

виставковий центр
КИЇВЕКСПОПЛАЗА
(метро "Нивки", вул. Салютна, 2-Б)

ВИСТАВКА ДЛЯ

**технічних директорів,
головних інженерів,
начальників
технічних відділів
інженерів-зв`язківців**

expoTEL 2004

третя міжнародна виставка
корпоративних телекомунікаційних мереж

Замовлення запрошень,
програма виставки на сайті **www.expotel.ua**

Програма виставкових заходів

Вперше в Україні!

18-19 лютого — З'їзд IT-директорів України

Повна інформація щодо програми з`їзду та умови участі на сайті: **www.cio.org.ua**

- Показова експозиція корпоративної інформаційної системи АНТК ім. Антонова.
  17 лютого — конференція "Досвід впровадження передових інформаційних технологій на наукоємних виробництвах України".
- **18 лютого** — конференція "Сучасні телекомунікаційні рішення для бізнесу".
- **19 лютого** — конференція "Бездротове майбутнє України".
- Семінари та презентації учасників.

організатор: "Євроіндекс"

медіа-партнер: Мир связи

інформаційні спонсори виставки: Коммуникации и сети ТЕЛЕКОМ; Дзеркало тижня ЗВ'ЯЗОК

онлайн-партнери тижня цифрових технологій: "Ай-ті Веа"; "Ай-ті-сі онлайн"