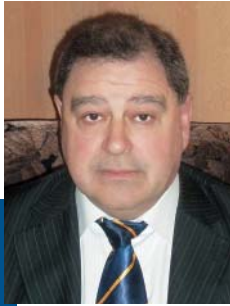


Корпоративная информация: особенности поиска



Е. Коржов

директор компании «Текон»

Рассмотрены характерные проблемы, возникающие при поиске документов в больших объемах текстовой информации. Организация доступа к данным напрямую зависит от технологий и программ, обеспечивающих скорость и качество обработки информации.

Общая тенденция развития ИТ-приложений — обрабатывать данные, которые все более «богаты» по смысловому содержанию (контенту) и, одновременно, все менее организованы по структуре. В порядке уменьшения «организованности» и увеличения «смысловой ценности» данные можно разделить на:

- ❖ простые структурированные;
- ❖ сложно структурированные;
- ❖ полу (частично) структурированные — XML-документы, сообщения EDI и e-mail;
- ❖ неструктурированные — текст, графика, видео, звукозаписи и др.

Реляционная модель, лежащая в основе большинства современных СУБД, позволяет успешно обрабатывать простые структурированные данные, представимые в виде строк таблиц. Некоторые СУБД (например, Oracle) обеспечивают объектно-реляционные возможности, так что приложения могут работать с комплексными структурированными данными (коллекции, ссылки, определенные пользователем типы и др.). Такие технологии, как Oracle Streams Advanced Queuing, позволяют работать с сообщениями и другими частично-структурированными данными.

Неструктурированные данные не могут быть разбиты на отдельные компоненты (поля, подполя и т.п.). Именно неструктурированные данные обеспечивают тот «прирост» объемов информации, который радует продавцов хранилищ данных и озадачивает ИТ-руководителей организаций. Их важность определяет тот факт, что, по мнению экспертов до 90% объема корпоративной информации представлено в виде текста (если, конечно, это не телестудия, фирма звукозаписи или рекламное агентство). По некоторым оценкам сотрудники компании тратят до половины рабочего времени на решение проблем, связанных с управлением неструктурированной информацией.

ПРОБЛЕМЫ ПОИСКА ТЕКСТОВОЙ ИНФОРМАЦИИ

Идеализированные модели поисковых алгоритмов, на которых студентов начинают учить программирование, внушают иллюзию простоты задач поиска. В реальной жизни быстрый

поиск подходящих документов в больших объемах данных — это одна из важнейших и сложнейших задач, решаемых сегодня с помощью ИТ. Среди моментов, которые сильно усложняют поиск текстовой информации на практике: разнообразие источников данных (базы данных, почтовые системы, web-страницы, файловые системы), разные форматы представления данных, оперативность получения информации (иногда минуты и секунды), объемы хранимых данных, ошибки (проблемы грамотности, массовый перевод из звуковой формы в текст), языки (часто в запросе сочетаются слова на нескольких языках) и др.

Для текстового поиска характерно, что ищется не точное совпадение, а «похожее». Поисковые системы давно уже не дают простых ответов («найдено» и указатель на место, или «не найдено»). В качестве ответа на запрос теперь выдается не единственный результат, а множество результатов (выборка), в той или иной степени близких к теме запроса.

Качество поиска. Есть два основных показателя качества поиска текстовой информации. Релевантность показывает, насколько близки полученные по запросу документы к искомому (больше релевантность — меньше «мусора» в результатах поиска). Полнота показывает, сколько подходящих документов не попало в «ответ». Собственно проблемы поиска связаны с балансировкой этих двух показателей:

- ❖ можно просто ввести в качестве запроса весь текст, что гарантирует полную релевантность — совпадение будет полным, если текст будет найден (вот только зачем его искать);
- ❖ можно просто включить в ответ все документы, что гарантирует полноту (если документ есть, его не пропустят).

Понятно, что ни тот, ни другой вариант не подходят, истина где-то между ними.

Инструменты поиска можно поделить на три группы: поиск на локальном компьютере, глобальные поисковые интернет-системы, корпоративные решения.

Локальный поиск. С поиском на отдельном персональном компьютере все внешне просто: вводится имя искомого файла (полностью или частично) — для поиска в оглавлении, или часть текста — для поиска в документах. Обычно применяется простой перебор, усложняемый разными форматами текста и его архивацией.

Глобальный поиск. Огромные объемы информации, распределенная структура ее хранения делают простой поиск просмотром текста не просто неэффективным, но невозможным. Поэтому в последнее время активно ведутся разработки по улучшению глобального поиска. Этому

способствует ряд моментов, облегчающих поиск в Интернет:

- ❖ практически отсутствуют проблемы секретности (в отличие от проблем безопасности);
- ❖ редко выполняется поиск конкретного документа;
- ❖ релевантность обычно определяется на основе анализа количества ссылок.

Корпоративные системы. Сложность задач этого направления в том, что необходимо не только решать задачи первых двух групп (локального и глобального поиска), но и учитывать ряд дополнительных особенностей корпоративного поиска:

- ❖ информация распределена по разнородным корпоративным источникам и защищена политиками безопасности;
- ❖ сотрудники имеют роли, определяющие их уровень допуска к информации;
- ❖ практически отсутствуют перекрестные ссылки во внутренних документах;
- ❖ важность документа определяется на основе подходов, отличных от тех, которые используются при глобальном поиске, а эффективные ответы не связаны с индексами популярности.

Индексация. Для предприятий и компаний с гигантскими объемами неструктурированных «знаний» простой просмотр всего текста каждого существующего документа занимает огромное количество времени. Поэтому, чтобы быстро находить в тексте нужную информацию, его нужно предварительно «разметить». Иногда это делается в самом документе (гипертекстовая структура), иногда в отдельных файлах (индексах), которые и используются в дальнейшем при поиске.

Ключевые слова. Если база текстовых данных содержит несколько десятков тысяч документов, то быстро найти информацию, даже тщательно подобрав в запросе ключевые слова, очень трудно. Придется просматривать в полученной выборке документ за документом, добавляя новые ключевые слова и их комбинации — и так до достижения соответствия. Причем совсем не факт, что пользователь самостоятельно сможет подобрать нужное сочетание ключевых слов (или вспомнить его в дальнейшем).

Кроме «традиционного», есть еще несколько усовершенствованных видов поиска по ключевым словам: с учетом морфологии (строения слов), нечеткий (учитывает возможность ошибок и опечаток), фонетический (учитывает сходные по звучанию слова) и синонимический (учитывает похожие по смыслу слова). Как вариант, в некоторых системах можно указать в качестве аргумента поиска документ, и искать «похожие» на него.

ТИПОВЫЕ ПРОБЛЕМЫ КОРПОРАТИВНЫХ ПОЛЬЗОВАТЕЛЕЙ

Поиск «по содержанию». Скорость поиска информации в больших объемах данных является важным фактором. Речь идет не о скорости работы самой системы-поисковика (поисковой системы), а о времени поисковой сессии (первоначального запроса, уточнения или поиска новых ключевых слов). Основные проблемы связаны с неудачным выбором ключевых слов и просмотром ненужных документов, полученных в списке результатов запроса. Сократить время можно, указывая в качестве шаблона поиска документ — с поиском близких по содержанию.

Проблема «близнецов». В базе данных или информационной системе предприятия могут содержаться документы из различных источников, содержащие похожую или идентичную информацию. Один и тот же текст может быть с разными заголовками, с небольшими изменениями или дополнениями, что вносит определенную путаницу при его использовании. Несколько сотрудников могут хранить у себя на компьютерах одинаковые документы. В некоторые они могут внести правки, комментарии и пр., некоторые — использовать как образец для подготовки новых документов. Плюс резервные копии на сервере, плюс пересылаемые как вложение в e-mail. Все это вызывает появление множества очень похожих документов (или просто полных копий).

Чтобы упростить поиск необходимо избавить информационную систему от ненужных дублей. Решить эту проблему можно, сравнивая поступающие в базу документы с уже содержащимися в ней, выявляя дубликаты и «нейтрализуя» их.

Консолидация информации. Крупные предприятия вынуждены затрачивать огромные средства на совмещение информации из различных систем (например, проектной, технологической и финансовой документации).

Современные технологии поиска и структуризации информации могут являться консолидирующим элементом для различных информационных систем на предприятии. Поиск и автоматическая классификация документов позволяют структурировать информационные составляющие любого крупного предприятия под управлением одной программы — без перевода документов и данных в какой-либо единый формат. Вся информация, доступная для индексирования и дальнейшего поиска может быть распределена, структурирована и отображена в удобном виде.

РЕШЕНИЕ ЯЗЫКОВЫХ ПРОБЛЕМ

И русский и украинский языки сложно построены и даже в учебниках грамматики отмечается множество исключений, уточнений и противоре-

чий. Повседневная речь делает языковые проблемы еще сложнее: профессиональные сленги (использование слов в нетрадиционном смысле) и региональные диалекты (использование «нетрадиционных» слов), быстрый рост словаря, отягощенный заимствованиями из иностранных языков, падение грамотности, многозначность слов и многовариантность выражения понятий (например, «покупка компаний», «поглощение компаний», «приобретение компаний»).

Все это вызывает необходимость от поисковых инструментов применения сложных лингвистических технологий (а не просто «дружественного пользователю интерфейса»). Подобные развитые возможности обработки текста обеспечивают продукты компании Oracle, однако у них есть один недостаток — полный набор возможностей обеспечен только для текстов на английском языке (и частично на других).

Для преодоления указанного недостатка фирмой «Текон» совместно с ее партнерами — «ЭР СИ О» (Россия) и «Трайидент Софтвр» (Украина) — был создан продукт Ukrainian Context Optimizer (UCO). В нем использованы технологии и алгоритмы, которые прошли проверку на практике и успешно работают в самых разных отраслях.

В результате разработки была найдена и реализована в программном коде наиболее рациональная система описания украинской морфологии, которая обеспечила максимальное быстродействие при минимальном объеме хранимых лингвистических данных. Так, общий объем словаря в 115 тысяч слов (около 4 миллионов словоформ) и данных, необходимых для анализа неизвестных слов, не превышает 10 Мб. При этом на современных процессорах обеспечивается разбор 200 тысяч известных слов в секунду или около 40 тысяч неизвестных.

Сегодня UCO for Oracle — единственный на рынке продукт, позволяющий значительно расширить возможности Oracle Text при работе с базами данных, содержащими документы на украинском языке. Продукт предназначается для отделов автоматизации производства среднего и крупного бизнеса, ИТ-подразделений государственных учреждений, системных интеграторов и разработчиков приложений, использующих возможности информационного поиска. UCO for Oracle задействует такие технологии, как лексико-грамматический и статистический анализ текста, алгоритмы автоматической классификации, рубрицирования и реферирования; нечеткого поиска, реализуя для украинского языка все существующие в OracleText возможности. Существуют версии продукта для ОС Windows и различных UNIX-платформ (SUN Solaris, SCO UnixWare, Compaq Tru64 Unix, HP UX, IBM AIX).

Евгений Коржов