РОССИЙСКИЙ НАУЧНЫЙ ЭЛЕКТРОННЫЙ ЖУРНАЛ
Электронные библиотеки
Издается с 1998 года
ISSN 1562-5419

Электронные библиотеки - 2000 - Том 3 - Выпуск 1

Взаимодействие пользователя с крупными массивами электронных документов в системе PCBIRS

В.Ю. Бугаев
ВНИИ физико-технических и радио технических измерений

Двадцатый век решил проблему коммуникации людей на планетарном уровне, но усугубил проблему информационного хаоса. Мир уже сегодня готов буквально завалить нас информацией, представленной в электронном виде. Достаточно упомянуть Internet и огромное количество изданий на компакт-дисках. Разработчики OCR грозят перевести все бумажные документы в электронные эквиваленты. И если учесть, что персональный компьютер в состоянии хранить гигабайты, то накопить большой объем информации - не проблема даже у себя дома. Вопрос только в том, что с этим делать дальше?

Похоже компьютер - самый подходящий протез, который человечество придумало для решения этих проблем. Но, судя по основным направлениям развития компьютерных технологий, которые сложились к настоящему времени, складывается впечатление, что в голове самое важное - уши.

Действительно, львиная доля проектов связана с развитием средств коммуникации и обеспечения доступа к огромным информационным ресурсам, которые беспорядочно растут с ужасающей скоростью. В тоже время культура потребления информации до сих пор не претерпела существенных изменений и мало чем отличается от офисного разложения файлов по папкам. Если у пользователя нет инструмента, позволяющего ориентироваться и манипулировать информацией в собственных массивах, он скорее всего будет смотреть на возможность дополнительного получения информации из внешних источников, как на потенциальную, но в принципе совершенно бесполезную. Поэтому все острее встает задача создания инструментов, которые не только расширят возможности доступа к различным источникам, но и позволят человеку "переваривать" огромные информационные залежи.

Как накапливать информацию, чтобы она не превращалась в мусор, а приносила пользу? Как, взаимодействуя с крупными информационными источниками, формировать собственное информационное пространство?

Это не только проблема пользователя, но и проблема практически всех агентств, владеющих крупными информационными ресурсами, и предоставляющих доступ широкой аудитории. Казалось бы они не должны заниматься этими вопросами (своих проблем хватает), но нужно отдавать себе отчет, что в результате количество потребителей, обрадованных вначале возможностями доступа к новым источникам, может резко сократиться.

Очевидно, что развитие методов доступа должно идти параллельно с развитием методов, позволяющих конечному пользователю усваивать крупные объемы информации. Это обстоятельство на мой взгляд должно учитываться и при разработке таких проектов, как электронные библиотеки.

Одним из наиболее эффективных методов потребления информации из внешних источников является построение конечным пользователем собственных баз данных. Не нужно забывать, что чаще всего пользователь обращается к внешним источникам информации не потому, что ему нечего почитать. Одной из задач накопления информации является ее переработка, и как результат, формирование новых информационных массивов для последующих публикаций.

СУБД и ИПС

Если проблема накопления и хранения информации снимается аппаратными средствами и средствами операционной системы, то проблема обеспечения поиска и манипулирования информацией, включения ее в тот или иной алгоритм принятия решений, целиком ложится на соответствующее программное обеспечение.

Проблема существенно упрощается, если информация поступает и в дальнейшем хранится в виде структурированных баз данных. В этом случае удается формализовать большинство операций обработки, поскольку манипулирование самими данными заменяется задачей манипулирования их именами.

Возможность формализации процедур обработки структурированной информации во многом определяет направления развития систем управления базами данных (СУБД), которые сложились к настоящему времени. Многочисленные СУБД, представленные на современном рынке программных продуктов, отличаются друг от друга методами хранения и доступа, сервисом для конечного пользователя, а объединяет их общая идея работы со структурами. Причем наибольшей популярностью пользуется так называемая реляционная модель, в которой информация может быть представлена в виде множества связанных таблиц (иерархические и сетевые архитектуры, по-видимому, следует рассматривать, как варианты реализации реляционной модели).

Но при этом открытым остается один чрезвычайно важный вопрос: **как исходные сообщения, которые чаще всего поступают в виде свободных текстов, превращаются собственно в базу, т.е. в хранилище информации в виде структур связанных данных?**

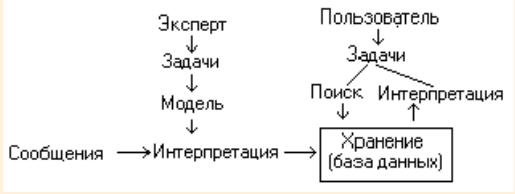
К сожалению, общего формализма для решения этой проблемы не существует, и проектировщику баз данных приходится полагаться в основном на интуицию и практический опыт, которые складываются в процессе изучения предметной области и продиктованы кругом решаемых задач.

Не случайно в последнее время широко обсуждаются ограниченные возможности реляционной модели для хранения и доступа к информации в базах данных. В основном это связано с тем, что само понятие "данные" претерпевает существенные изменения. С одной стороны все чаще в качестве данных выступают довольно сложные объекты: тексты, таблицы, графические изображения, органы управления и т.д. С другой стороны основная парадигма структурированных баз носит довольно ограниченный характер, поскольку предполагает, что семантика должна быть описана на стадии проектирования и может храниться отдельно от данных. К тому же вопрос о том, всегда ли возможно любую информацию объективно представить в виде конечного множества связанных таблиц простых данных, до сих пор остается открытым.

Поэтому не прекращаются попытки выйти за пределы реляционного подхода (пост- реляционные, объектно-реляционные, объектно-ориентированные и др. базы данных). Следует заметить, что использование в качестве данных сложных объектов само по себе еще не означает отхода от реляционной модели. Возможность поддержки функций, восстанавливающих объект для просмотра (например, мультимедийный файл), является очень важным свойством конкретной СУБД, но все это не противоречит реляционной модели и никак ее не расширяет до тех пор, пока множество таких объектов может быть описано с помощью своих параметров и представлено в виде конечного числа связанных таблиц. Здесь скорее речь идет о способе реализации реляционной модели, чем об отходе от нее.

Ограниченность реляционного подхода становится понятным, когда данные не имеют четко выраженной и фиксированной семантики, которая в значительной степени может зависеть от способов интерпретации информации конкретным человеком. С этой ситуацией чаще всего проходится иметь дело при обработке текстов. Дело в том, что каждое предложение связанного текста в состоянии породить множество данных, семантику которых зачастую невозможно отделить от общего контекста. К тому же смысл данных зависит от субъективного взгляда, продиктованного кругом решаемых задач (один и тот же текст каждый может прочитать по-своему и выделить в нем то, что представляет наибольший интерес). И хотя теоретически любой текст можно разбить на множество бинарных отношений типа: **данное - значение**, практически при построении реляционной базы данных из всех возможных отношений выделяют только те, которые предполагается в дальнейшем использовать для решения вполне конкретного круга задач. Условная схема функционирования такой базы может быть представлена на рисунке.

Рис.1.



Как правило, создание базы ложится на плечи эксперта, который разрабатывает модель данных и в соответствии с этой моделью обрабатывает исходные сообщения (в общем случае они могут представлять сложные документы), которые он интерпретирует и преобразует в структуры для дальнейшего хранения. Проектируя базу, эксперт должен быть уверен заранее, что по мере поступления информации, будут меняться только значения данных, а структура останется неизменной.

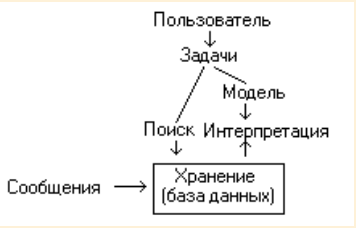
Пользователь, который нуждается в информации для решения собственных задач, может работать только в рамках модели, определенной экспертом, причем он не имеет возможности понять, как те или иные данные попали в информационное хранилище, поэтому его собственная интерпретация результатов поиска в значительной степени привязана к мнению эксперта.

Несмотря на явное противоречие, такой подход во многих случаях является вполне оправданным для

ряда технологий со сложившимися и достаточно четкими стереотипами обработки информации (бухгалтерский учет, банковские операции, пассажирские перевозки и т.д.). Но при этом всегда нужно помнить, что такие базы - это вторичный продукт предварительного анализа первичной информации, при котором процедура формализации и абстрагирования может привести не только к обеднению, но и к искажению семантики данных, вырванных из контекста исходных сообщений.

Другой подход к формированию баз данных заключается в хранении всех исходных сообщений в том виде, в котором они поступают.

Рис.2.



Внешне эта схема выглядит проще, но эта простота является кажущейся. Если в первом случае в результате поиска пользователь находит множество интересующих его данных, то во втором он находит множество исходных сообщений, содержащих эти данные. Если при этом возникает необходимость включения данных в те или иные алгоритмы последующей обработки, то процесс интерпретации существенно усложняется, так как требует использования дополнительных неформальных механизмов их извлечения из найденных текстов.

С другой стороны эта схема исключает те противоречия, о которых говорилось выше. К тому же способ хранения и поиска информации никак не зависят от модели данных, которые у пользователя могут меняться в зависимости от возникающих задач. Это предоставляет большую гибкость особенно при работе с крупными информационными массивами и исключает необходимость предварительной структуризации и формализации информации, что является одним из самых дорогостоящих этапов создания баз. Следует также учитывать тот факт, что большая часть информации (письма, протоколы, нормативные акты, указы, законы и т.п.) вообще не поддается какой-либо априори разумной структуризации и может быть представлена только в виде свободных текстов.

СУБД для работы со структурированными данными обычно относят к разряду фактографических систем. Программные продукты, реализующие вторую схему, относятся к разряду документальных, и представляют собой информационно поисковые системы (ИПС). Оба типа существенно отличаются не только по техническому воплощению, но и по результатам взаимодействия с конечным пользователем.

Разработчики ИПС пытаются убедить нас в том, что сама по себе возможность из миллионов документов отыскать всего несколько тысяч, уже невероятное счастье. Разработчики СУБД говорят: дайте нам структуру, заполните ее непротиворечивыми данными и убедитесь сами, какие чудеса вытворяют SQL запросы.

Действительно, в фактографических системах пользователь в результате запроса получает набор так называемых кортежей с релевантными данными и имеет возможность получить ответы на интересующие его вопросы.

В документальных ИПС пользователь в результате выполнения запроса на контекстный поиск получает список релевантных документов и имеет возможность получить ответ только на один вопрос: какие документы содержат необходимые данные? Ему еще предстоит выделить и связать подходящие данные, прежде чем он получит ответы на вопросы, которые его волнуют. Если найденных документов много, то задача становится практически невыполнимой. Единственный выход – автоматизированный семантический анализ текстов. Но нужно признать тот факт, что на сегодняшний день приемлемых результатов в этом направлении еще не получено. В лучшем случае ИПС предлагают лишь различные статистические методы ранжирования документов по степени релевантности. Но несмотря на тень наукообразия, которая наводится на эти методы, они все-таки носят оттенок хиромантии. Оценка значимости слов только по их частотным характеристикам без оценки их значимости для "вопрошающего" не выдерживает критики. Тем не менее иногда эти методы приводят к неплохим результатам. В таких случаях говорят: "и на том спасибо".

Следует заметить, что на практике пользователю приходится иметь дело как со структурированной, так и с полнотекстовой информацией. Использование для этого различных программных средств и различных стереотипов обработки вызывает массу неудобств. Поэтому возникает необходимость иметь программный продукт, который объединял бы возможности как структурированных СУБД, так и ИПС.

PCBIRS=СУБД + ИПС

Рассмотрим, как проблема потребления информации конечным пользователем решается в системе PCBIRS.

Информационно- поисковая аналитическая система PCBIRS (см. "Мир ПК" 12/97,с. 54, Мир ПК 8/99,с.76,

http://www.chat.ru/~birs) предназначена для работы на ПК с большими информационными массивами, которые могут состоять из множества произвольных текстовых документов, или представлять структурированные базы данных. Для тех и других PCBIRS обеспечивает методы быстрого контекстного поиска информации, единый стереотип анализа и последующей обработки, которые базируются на технологии автоматической лексической индексации текстов и структур данных.

С технической точки зрения PCBIRS представляет систему управления документально-фактографическими базами данных. В отличие от традиционных (реляционных) СУБД, PCBIRS ориентирована на работу с информацией, которую, в известном смысле, следует рассматривать как сырье или первоисточник для дальнейшей обработки.

Информацию из внешних источников предлагается хранить в базах данных, которые пользователь может создавать по своему усмотрению. Причем в базах могут храниться либо сами документы, либо только ссылки на источники.

PCBIRS в основном реализует те стереотипы обработки, которые связаны с поиском и анализом первичной информации, представленной в виде свободных текстов. Основная идея состоит в обеспечении возможности хранить и манипулировать информацией в том виде, в каком она поступает, а необходимые структуры данных получать динамически (виртуально) в зависимости от решаемой задачи и цели (в терминологии PCBIRS это виртуальные списки, которые создаются на множестве найденных документов).

Основная задача, которая решалась при создании PCBIRS – обеспечение прозрачности при работе с крупными массивами как полнотекстовой, так и структурированной информации, поскольку любой достаточно крупный информационный массив для пользователя в конечном итоге представляется неким "черным ящиком". Как правило, необходимо выполнить запрос на поиск интересующей информации, причем всегда имеется возможность не получить ответа. Значит ли это, что нужной информации нет, или сам запрос сформулирован неудачно? Нельзя ли получить представление о содержимом всего информационного массива в целом, не читая документов базы данных, и попытаться понять, о чем там в основном идет речь? Оказывается можно. В PCBIRS для этого предусмотрен целый ряд средства, которые избавляют пользователя не только от слепого поиска, но позволяют извлекать и анализировать данные, содержащиеся в найденных документах.

Одним из таких средств является поддержка динамических понятийных классификаторов.

Пользователю предоставляется возможность формулировки области своих интересов в виде списков понятий, которые он может накапливать и проецировать на различные информационные массивы для получения обзора содержания последних. Эти списки представляют то, что психологи называют установкой.

Поскольку динамическая классификация даже крупных информационных массивов в сотни мегабайт занимает, как правило, несколько секунд, пользователь может свободно менять свою точку зрения, фиксировать различные множества документов и проецировать на них другие списки понятий, что в конечном итоге представляет мощное средство поисковой навигации. С одной стороны, еще до формулировки запросов он видит, что же содержит данный информационный массив (с его точки зрения), с другой стороны, выполнив запрос на поиск документов, у него имеется возможность понять, что же содержит найденное множество.

Кроме тривиальной подсветки поисковых терминов, PCBIRS позволяет гибко варьировать формы просмотра документов, формировать различные списки для отображения информации (например, списки актуальных фраз), что в конечном счете сокращает время принятия решения по поводу оценки релевантности найденных документов.

Таким образом, ориентация в крупных массивах на персональном компьютере перестает казаться неразрешимой проблемой. Задача сводится к методам, с помощью которых пользователь отображает область своих интересов, используя лексику естественного языка.

PCBIRS позволяет строить из источников вторичные базы данных и готовить из них различные приложения, которые могут быть оформлены, как электронные издания для публикации на CD или в сети Internet.

Некоторые технические характеристики PCBIRS 3.2

Основной единицей хранения и объектом поиска информации является документ. Каждый документ проходит индексацию по содержанию: автоматически строится словарь поисковых терминов для обеспечения высокой скорости поиска документов, которая практически не зависит от объемов баз данных. Средняя скорость индексации текстового массива в пакетном режиме для компьютеров Pentium 133 Мгц составляет 10Мб/мин, при этом скорость поиска по отдельным терминам на массивах в сотни мегабайт составляет 0.01-0.1с. (в зависимости от частоты встречаемости термина). Объем словаря для баз свыше 100 мегабайт составляет 1-3% для текстов на русском языке. Каждая база данных может хранить до 500000 документов, объем базы до 4Гб при хранении текстов непосредственно в базе, или до 16Гб при хранении ссылок на источники. Несколько баз данных на

логическом уровне могут объединяться в качестве подбаз и выглядеть для пользователя, как одна база. Количество подбаз в базе не ограничено.

PCBIRS функционирует на IBM совместимых компьютерах с процессорами не ниже 486 под управлением системы WINDOWS версии не ниже 3.1 (WINDOWS '95,'98,NT). Для непрерывной индексации информационных массивов объемом до 2Гб необходима память не менее 32 Мб.

ДОКУМЕНТЫ

Могут иметь или не иметь внутреннюю структуру хранения информации, содержат:

- произвольные тексты;
- отдельные данные;
- графические изображения;
- таблицы;
- кнопки;
- параметры внешних функций.

ИСТОЧНИКИ ИНФОРМАЦИИ

- **Диалоговый ввод** документов с клавиатуры или из внешних приложений (технология аннотирования файлов);
- **Пакетный ввод из файлов**:
- формат помеченных строк (входной формат PCBIRS);
- базы данных PCBIRS;
- базы данных в формате DBASE IV;
- файлы указатели на текстовые фрагменты (PMT файлы);
- линейно размеченные тексты в ASCII или ANSI кодировке;
- произвольные текстовые файлы в ASCII или ANSI кодировке;
- множество HTML файлов ;
- подключение собственных функций чтения источников;

Пакетная индексация информационных массивов предполагает хранение в базах PCBIRS либо текстов, либо только ссылок на источники информации. При загрузке структурированных документов имеется гибкая возможность отображения структуры источника на структуру базы.

ОПЕРАЦИИ С БАЗАМИ ДАННЫХ

- Создание и реорганизация баз данных в диалоговом режиме;
- Обеспечение авторской защиты информации (copyright, доступные операции с информацией);
- Разграничение уровней доступа к информации при работе в локальной сети;
- Подключение/отключение баз в качестве подбаз, создание тем для совместной работы с множеством баз в многооконном режиме;
- Диалоговый и пакетный ввод, корректировка, удаление документов в базе с немедленной или отложенной индексацией текстов и обеспечением контроля ввода информации;
- Выгрузка документов из базы, слияние баз данных.

ЯЗЫК ЗАПРОСОВ

- Включает поисковые термины (слова, числа, даты, спец термины) и условия их вхождения в искомые документы (операторы логики AND, OR, NO, XOR, NOT, контекстной близости NEAR, CTX, SEGM, ограничения области поиска). Допускается четкое и нечеткое маскирование терминов слева, справа, внутри, запись четких и нечетких вариантов слов;
- Поиск по числам и числовым интервалам допускает автоматический перевод размерностей (метры, сантиметры и т.д.).

ВЫПОЛНЕНИЕ ЗАПРОСОВ

- Составление запросов и их выполнение в диалоговом режиме;
- Выделение терминов запроса непосредственно в просматриваемых документах, из словаря базы данных и частотного словаря;
- Выбор и выполнение запросов, хранящихся в базах данных;
- Создание и выполнение постоянных запросов в виде понятий и иерархических классов в каталогах Динамическое подключение каталогов запросов к базе данных и выполнение пакета запросов;
- Сохранение и уточнение запросов, автоматическое повторение запросов в режиме фоновой подкачки документов;
- Сквозной поиск в нескольких, распределенных в локальной сети, базах;

АНАЛИЗ РЕЗУЛЬТАТОВ ПОИСКА

- Представление документов в различных формах, перестановка, переключение фрагментов, таблицы отчета, компактный текст, мониторинг актуальных фраз из списка найденных документов, частные формы, бланки;
- Подсветка актуальных терминов запроса для любой формы представления документов;
- Просмотр частотного словаря документов с фильтрацией по определенной тематике и быстрым скроллингом на актуальные термины;
- Быстрый скроллинг на фрагменты в документах;
- Просмотр документов в отсортированном порядке с выбором критериев сортировки;
- Вспомогательные списки быстрого просмотра;
- Виртуальные списки данных из текстов документов, отображение их в виде диаграмм, графиков, таблиц с возможностью расчетов и дополнительного отбора документов в зависимости от содержащихся в них данных;
- Фиксация множества документов для последующих запросов;
- Инвертирование списка документов для просмотра документов нерелевантных запросу;
- Удаление классов и отдельных нерелевантных документов из списка найденных;
- Вывод документов и виртуальных списков на печать, в файлы, экспорт в другие WINDOWS приложения;
- Ограничение вывода со стороны авторской защиты.

РАЗРАБОТКА ПРИЛОЖЕНИЙ

- Объединение баз данных в тему для совместной обработки в многооконном режиме;
- Связывание документов одной или нескольких баз по содержанию (динамический гипертекст);
- Передача параметров из текстов документов и вызов внешних DOS или WINDOWS приложений;
- Средство программирования и диалоговой отладки (встроенный макроязык **BML**) развитых приложений;
- Подключение дополнительных команд на кнопки интерфейса PCBIRS, создание собственных навигаторов поиска и обработки информации, подготовка электронных изданий, подготовка электронных изданий на компакт дисках и публикации в сети Internet.

Об авторе

Бугаев Виталий Юрьевич - к.ф.-м.н, руководитель лаборатории ВНИИ Физико-технических и радиотехнических измерений, автор информационно-поисковой аналитической системы PCBIRS (см. Мир ПК 12/1997, Мир ПК 8/1999), www.chat.ru/~birs, телефон: (095)535-08-52,e-mail: bgv@ftri.extech.msk.su

---