

СТАТИСТИЧЕСКИЕ И АНАЛИТИЧЕСКИЕ ВОЗМОЖНОСТИ БИБЛИОТЕКИ БУДУЩЕГО

Рохварг С.Л.

Корпорация «Индустрия ИНТЕЛЛЕКТА»

"Мы никогда не понимаем, какие сокровища перед нами.
Знаешь почему? Потому что люди вообще не верят в сокровища."
Пауло Коэльо "Алхимик"

1. Куда мы движемся.

Для начала давайте обратим внимание на несколько моментов:

- **количество** информации с каждым днем увеличивается в геометрической прогрессии;
- целевая информация в этом потоке практически теряется и на ее поиск уходит все больше времени (снижается **качество** информации);
- **обработка информации** становится гораздо более важным процессом, чем просто поиск, из-за увеличения количества и снижения качества информации;
- значительную роль в процессе поиска и обработки информации играет **скорость** и **удобство поиска** информации.

Таким образом, библиотека становится не только местом хранения, а источником для поиска и обработки информации. Так как значительное количество информации находится в электронной форме, и, со временем, это соотношение все больше будет меняться не в пользу печатных изданий, сейчас уже все понимают необходимость создания баз данных для хранения имеющейся и накопления новой информации. Дальше мы будем говорить не просто о библиотеках – а о системах для поиска и обработки информации (на примере информационной системы «АТЛАС»). Попробуем сформулировать принципы такой информационной системы.

2. Принципы построения информационной системы.

Мы согласились, что основными требованиями к информационной системе будут не только хранение информации, но ее поиск и обработка. Поэтому важное место в реализации такой системы получают статистические функции системы, функции оперативной аналитической обработки (OLAP) и функции системы добычи данных (data mining). Список требований, сформулированных нами для информационной системы, выглядит так:

1) **СКОРОСТЬ** - означает, что система должна обеспечивать выдачу большинства ответов пользователям в пределах приблизительно пяти секунд. При этом самые простые запросы обрабатываются в течение одной секунды и очень немногие - более 20-ти секунд. Недавнее исследование в Нидерландах показало, что конечные пользователи воспринимают процесс неудачным, если результаты не получены по истечении 30 секунд. Они способны нажать "Ctrl+Alt+Del", если система не предупредит их, что обработка данных требует больше времени. Даже если система предупредит, что процесс будет длиться существенно дольше, пользователи, могут отвлечься и потерять мысль, при этом качество анализа страдает. Скорость у нас обеспечивается как разными серверами, так и тематическим разделением базы на физические части.

2) **ДОСТУПНОСТЬ** – означает, что система позволяет работать с ней пользователю в любой момент времени. Это обеспечивается распределенной

базой данных, несколькими серверами; пользователь получает доступ к тому серверу, который меньше загружен в данный момент, причем при этом учитывается предпочтение по трафику (например, пользователь из России быстрее получит доступ к российскому серверу, аналогично для пользователей Украины), по времени (например, время загрузки серверов для пользователей Северной Америки – ночь для России и Украины) и т.д.

3) **Точность** – возможность поиска, статистического и аналитического модулей выделять необходимую информацию, находить оптимальные пути для ее получения, предоставлять возможности по обработке информации для обеспечения максимально требуемого результата.

4) **Конфиденциальность** – означает, что пользователь получает доступ только к предназначенной ему информации, и к его информации остальные пользователи получают доступ только по его разрешению. Сюда же входит и многопользовательская поддержка, и гибкая система управления правами пользователя.

5) **Информативность** – означает, что в случае отсутствия в базе необходимой информации, в системе должна быть возможность получить (или, по крайней мере, предпринять возможность поиска) из внешних источников – например, из Интернета.

6) **Актуальность** – информация в системе должна своевременно обновляться и дополняться. Время выхода информации в печатном виде не должно быть раньше электронного.

7) **Универсальность** – означает, что система должна быть неприхотлива в установке и использовании, в отсутствии программных конфликтов и остаточных эффектов от ее работы. В частности, у нас это – система не нуждается в установке, может встраиваться в другие системы обмена данными, пользуется стандартными механизмами для работы с базами данных и т.д.

У нас есть возможности работы с базой в 5 вариантах:

1. работа с локальной версией базы – база находится на сервере клиента, локальные версии – на рабочих станциях клиента;
2. работа с локальной версией базы – база находится на сервере клиента, в сервер клиента встраивается интерфейс на РНР для доступа к базе, на рабочих станциях клиента - через веб-интерфейс - Internet Explorer;
3. работа с Интернет-версией базы – база находится на нашем ближайшем сервере, локальные версии – на рабочих станциях клиента;
4. работа с Интернет-версией базы – база находится на нашем ближайшем сервере, на рабочих станциях клиента - через веб-интерфейс - Internet Explorer;
5. работа с Интернет-версией базы – база находится на нашем ближайшем сервере, в сервер клиента встраивается интерфейс на РНР для удаленного доступа к нашей базе, при этом клиент получает все возможности поиска и обработки информации, пользователи клиента – доступ с его сайта.

Во всех вариантах обновление информации (репликация) идет с ближайшего нашего сервера (по Интернету) или путем получения обновлений в файлах (на дисках, почтой, FTP-доступом и т.д.). В каких случаях предпочтительно использование локальных версий клиентов? Если для клиентских компьютеров ограничен доступ в Интернет, если существуют жесткие ограничения по размеру или типам трафика, если надо отдельно считать трафик или время пользования соответствующими сервисами.

В нашей информационной системе существует возможность клиента самому добавлять рубрики и пополнять базу своей информацией (только для

локальной версии). Причем, если эта информация пересекается с информацией в базе, можно настроить систему на обмен информацией нужной тематики.

8) **УДОБСТВО** – свойства системы, которые позволяют неподготовленному (или плохо подготовленному) пользователю пользоваться системой, максимально приблизить вид получаемой информации к оригиналу (это, например, предоставление полнотекста), приблизить эргономические характеристики системы к естественным человеческим. Подробнее про эргономику и удобство пользовательских интерфейсов можно прочитать на сайте www.usetronics.ru.

9) **СТОИМОСТЬ** – минимализация затрат на внедрение и использование системы обработки информации. Это в наших условиях достаточно актуально. У нас это означает использование бесплатной версии базы данных (за основу взят Firebird), использование своих механизмов обмена и обработки информацией, своего сервера и локальной версии оболочки.

3. Статистические возможности системы хранения и обработки информации.

Статистические возможности системы хранения и обработки информации можно рассматривать в двух смыслах: как статистику пользования базами данных и как статистическая обработка информации в базе.



Статистическая обработка информации допускает анализ по частоте встречающихся запросов, по структуре запросов к базе данных, по характеру и типам информации, по потребителям и источникам информации, по количеству и качеству тематической информации (это достигается присвоением информации «коэффициента аналитичности»), по статистическим выборкам из базы данных. К статистической обработке относятся возможности по сортировке и группировке результатов поиска:

- (1) % совпадения результата;
- (2) частоте попадания слов поисковой фразы в поиске;
- (3) величине документа;
- (4) частотности нахождения в поисковых фразах (по статистике поиска);
- (5) по любому из полей поиска.

С помощью этих статистических данных можно делать выводы по данным, отсутствующим в аналитическом разделе, - например, вероятность и мнения конкретной группы пользователей по данному вопросу. Статистические возможности по учету пользования информацией делятся у нас на несколько типов:

1. Статистика пользователя информации (за период):
 - 1.1. использованный трафик в базе данных;
 - 1.2. количество статей, к которым получен доступ;
 - 1.3. время работы в базе данных;
 - 1.4. время подписки на доступ к базе данных;

- 1.5. остаток денег (трафика, статей, времени, входов) на счету (баланс);
- 1.6. сумма скидок;
- 1.7. история подписок на доступ к базе данных;
- 1.8. история оплат доступа к базе данных;
- 1.9. количество и время входов в базу данных (история пользования).

Эти возможности используются в первую очередь клиентами, пользователями информации, по этим данным делаются выводы о полезности данных и частоте их использования.

2. Статистика источника информации (за период):

- 2.1. использованный трафик в базе данных (по информации этого источника);
- 2.2. количество статей, к которым получен доступ;
- 2.3. время работы в базе (с информацией этого источника);
- 2.4. время подписки на доступ к информации этого источника;
- 2.5. количество денег (трафика, статей, времени, входов) на счету (баланс) – оплачено пользователями за эту информацию;
- 2.6. сумма скидок, предоставленных на эту информацию;
- 2.7. история подписок на доступ к базе данных;
- 2.8. история оплат доступа к базе данных;
- 2.9. количество и время входов в базу данных (история пользования).

Эти разделы статистики полезны всем издательствам, которые с нами сотрудничают, другим базам данных, с которыми у нас есть соглашения о предоставлении информации, и авторам тематических материалов.

3. Статистики менеджеров, бухгалтеров, операторов, администраторов – т.е. непосредственно тех людей, которые работают с базой данных изнутри, пополняют ее, следят за ее целостностью и работают с потребителями информации.

Эти данные нужны, например, менеджерам для предоставления скидок клиентам и формирования тематических подборок.

4. Аналитические возможности системы хранения и обработки информации



Мы разделили аналитические возможности системы на следующие виды:

1. **Предварительная обработка данных** – сюда входит автоматическое аннотирование, выделение микро- и макро-цитат, автоматическая рубрикация статей.

2. **Аналитический отбор данных** – возможности элементов анализа при поиске и выделении информации по данной тематике, применение аналитических возможностей к результатам поиска.
3. **Анализ по базе данных** – работа с фактическим материалом, использование рейтинга аналитичности при формировании выводов по информации, содержащейся в базе данных, и доступной из других источников.
4. **Анализ информации пользователя** – возможность для пользователя ввода своей информации и верификации ее по материалам базы данных и информации из сторонних источников, формирование выводов корреляции этой информации с имеющейся в базе.
5. **Сложные и специфические методы анализа** – возможности добавления методов анализа, предложенных пользователем, формирование аналитических материалов, по его данным. Подбор этих материалов; сюда входит также возможность составления аналитических обзоров профессиональными аналитиками в данной области по заказу клиента, и размещение этой информации в его целевой рубрике.

Мы уже говорили, что аналитические возможности информационной системы играют едва ли не самую главную роль для пользователя. Поэтому здесь мы попытались реализовать некоторые принципы OLAP (подробнее о них можно прочитать на сайте www.olap.ru) по модели Кодда:

- 1) многомерное представление данных – то есть работа с данными строится по многим, не зависящим друг от друга, параметрам;
- 2) интуитивное манипулирование данными – пользователь получает возможность манипулирования данными прямо в месте их нахождения (например, в ячейке);
- 3) 4 модели анализа (Категориальный, Толковательный, Умозрительный и Стереотипный) – используются нами при поиске данных и аналитических обработках информации;
- 4) Прозрачность - пользователь получает искомые результаты, не заботясь об их источнике, в базе есть возможность поиска информации по сторонним источникам (например, Интернет или базы данных аналогичной тематики);
- 5) сохранение результатов (возможность сохранения, экспорта результатов работы как для поиска, аналитики, так и по заказным темам, и последующая работа с ними);
- 6) гибкость формирования отчетов. Возможность стыковки с другими системами формирования отчетов. Множество форматов для экспорта данных дает возможность импорта и экспорта практически в любую систему хранения и обработки данных, при этом формы отчетов легко подстраиваются под нужный образец. Возможность разработки специальных отчетов (заказная аналитика).

Разрабатываются механизмы добычи данных (так называемое «окно фактов») («data mining»). Собственно поиск, хранение и обработка фактических данных займут в базе свое определенное место. Огромное значение для скорости и точности получения результата имеет механизм поиска. Анализ структуры текста (например, поиск по рекламным блокам), логический и понятийный поиск, ведение истории поиска (как поисковых запросов, так и результатов поиска), статистики поиска, ведения поиска по всем признакам информации – значительно ускоряют, облегчают и делают поиск целенаправленным. Помогают в этом автоматическое определение гиперссылок, имен и названий.

Облегчает обработку информации автоматическое аннотирование и цитирование текста, автоматическая рубрикация и тематические подборки. Вся информация хранится в системе в двух форматах – HTML и PDF (с сохранением аутентичности расположения элементов документа). Информацию можно просмотреть как в оригинальном виде (полнотекст) так и в оптимизированном для просмотра в Интернете. Используя механизм фильтров (настраиваемый пользователем), Вы всегда будете видеть только ту информацию, которая Вам нужна.

5. Дополнительные (сервисные) возможности системы хранения и обработки информации.

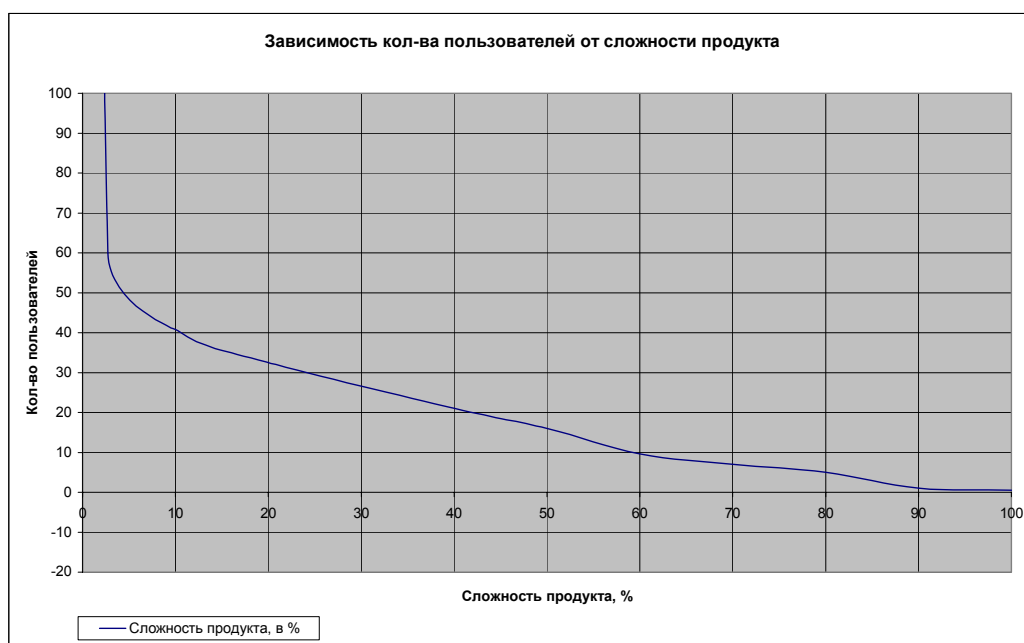
К дополнительным сервисным возможностям относятся:

1. Автоматический перевод с и на язык, используемый пользователем.
2. Автоматическая проверка орфографии и предложение подходящих вариантов пользователю.
3. Хранение в базе данных информации пользователя, возможность обмена с основной базой, причем на информацию пользователя распространяются все сервисные возможности системы.
4. Возможность экспорта из базы в самых распространенных форматах данных:
 - (1) PDF (аутентичный вид страницы);
 - (2) HTML (только текст или текст с изображениями);
 - (3) MHTML (все содержимое страницы в одном файле);
 - (4) TXT (без графической информации);
 - (5) RTF (как с изображениями, так и без них);
 - (6) DOC;
 - (7) JPG (графическое аутентичное изображение страницы).
5. Автоматическое формирование рассылок с выборками информации.
6. Ведение подборки по какому-либо параметру («своя газета»).
7. Ведение подборки типа «новостная лента».
8. Возможность локальным пользователям удаленно или локально обновлять информацию по какой-либо тематике (механизм репликаций).
9. Настройка автоматических уведомлений о появлении новой информации.
10. Выбор типа обмена информацией: по почте, по электронной почте, Интернет, локально, FTP и т.д.

6. Прогнозы развития информационных систем.

Попытаемся продумать пути развития информационных систем. Поскольку компьютерные системы развиваются в направлении миниатюризации, а тенденции создания переносной техники говорят о глобализации и объединении в одном устройстве наибольшей функциональности (например, производители часов серьезно обеспокоены конкуренцией со стороны производителей мобильных телефонов), видно, что в скором времени большинство функций будут объединены в одном устройстве (карманном компьютере). Это устройство будет выполнять функции как мобильного телефона, компьютера, так и коммуникационного устройства для получения информации из баз данных. Главная роль в процессе обмена информацией будет принадлежать источникам информации – издательствам, библиотекам - онлайн-базам данных. Чтобы дать возможность

пользователям работать с такими базами, поставщикам придется удовлетворять требованиям, сейчас еще не настолько актуальным – это, в первую очередь, **удобство, скорость и доступность**. Если сейчас пользователь еще может позволить себе нажать 5 кнопок и подождать 2 минуты, чтобы найти информацию на странице, то при резком увеличении количества пользователей и переноса функций доступа в мобильные устройства, необходимо будет продумать новые эргономичные интерфейсы и механизмы для пользования информационными системами. Для массовых систем необходимо уже сейчас строить распределенные сети для доступа к информации, древовидные серверные структуры. Пользователь не будет работать с «продуктом», прогресс идет по пути упрощения, чтобы предложить что-то на рынок, надо показать то, чем все уже пользуются, но в новом качестве (пример - «электронная книга», но с полным отображением страницы, и с эффектом перелистывания). Успехом будут пользоваться продукты, для которых не является обязательным серьезное обучение, которые можно будет использовать неподготовленному человеку, и делать это будет удобно.



Как мы видим из графика, с ростом сложности продукта, резко падает количество пользователей. И обратный процесс – при доступности интерфейса и функций системы количество пользователей растет, хотя, естественно, оно никогда не станет больше целевой аудитории этой системы.

7. Выводы.

Подведем итоги. Полной реализации всех методов подбора и анализа данных в информационных системах такого уровня на рынке сейчас нет. Мы строим информационную систему, наиболее полно отвечающую возможностям как аналитиков, научных работников, профессионалов в области работы с информацией, так и рядовых пользователей, людей, ежедневно нуждающихся в информации по своей профессиональной деятельности. **МЫ ПРЕДЛАГАЕМ** на рынке целую систему баз данных, объединенных сервисными возможностями, разных по направленности, но обладающих общими свойствами обработки

данных. Это «АТЛАС Аналитика» - аналитические статьи, «АТЛАС Наука» - научные материалы, «АТЛАС Мир» - пресса на иностранных языках и в переводе, «АТЛАС Пресса» - пресса стран СНГ, «ATLAS Science» - пресса стран СНГ в переводе на английский и другие языки международного общения и многие другие.

Мы попытались сделать систему, наиболее приближенную к системам, которые будут использоваться в ближайшем будущем, с учетом положительных и отрицательных моментов уже существующих продуктов. То, над чем мы сейчас работаем, Вы можете увидеть на сайтах www.atlas-w.com и www.atlas-center.org. Естественно, далеко не все еще реализовано, впереди еще много работы, в том числе, вместе с Вами. Только пользователи информационной системы «АТЛАС» могут указать нам точный путь развития. Но уже сейчас видно, в каком направлении надо двигаться, какие свойства системы развивать. Мы благодарны всем нашим пользователям за тот вклад, который они уже внесли в разработку нашей информационной системы. И надеемся, что они и дальше будут нам в этом помогать.

Аннотация

В докладе рассматриваются вопросы развития информационных систем, требования к ним и их сегодняшние возможности. Формулируются принципы построения информационной библиотечной системы на примере информационной системы «АТЛАС». Делается попытка спрогнозировать необходимые элементы систем для поиска и обработки информации будущего, и заложить основы для этого развития в разработке системы уже сегодня.