УДК 002:681.3
ББК 78.305.8
К 68
В.Н. Коробков

Концепция построения систем управления документами в электронных библиотеках

(Рецензирована)

Аннотация:

В статье рассматриваются принципы построения электронных библиотек, выделяются проблемы связанные с переходом документов на электронный вариант, выделяются принципы проектирования систем управления документами и дается концепция их построения применительно к электронным библиотекам.

Ключевые слова:

Электронная библиотека, система управления документами, информационно-поисковая система, электронный документ, принципы построения СУД, вычислительная система, клиент-серверная технология.

С развитием вычислительной техники, в особенности глобальных вычислительных систем, перед человечеством стала задача отказа от традиционных бумажных способов хранения информации и перехода к электронным документам.

Целью любой библиотеки является формирование фонда произведений печати в соответствии с профилем комплектования, и информационное обслуживание читателей и абонентов. Круг читателей любой библиотеки обычно ограничен размерами города и близлежащих районов. Поиск любой информации осуществляется вручную, с использованием алфавитных и тематических каталогов, что может занять от пары минут до нескольких часов времени.

Создание электронных библиотек призвано устранить проблемы классических библиотек. В частности, время поиска электронного документа и время реакции на запрос гораздо меньше, чем при работе с бумажными документами. Увеличивается возможность доступа к редким литературным источникам, доступ к которым часто разрешен только в архивах. Повышается надежность хранения информации, тогда как бумага подвержена многим внешним факторам: старение, опасность нагрева или охлаждения. Кроме того, очень трудно сделать копию всего архива бумажных документов на случай непредвиденных обстоятельств. В случае с электронными носителями все происходит наоборот. Компактность, быстрота и дешевизна копирования позволяет делать и хранить столько копий информации, сколько необходимо для обеспечения надежности.

Однако переход к электронным библиотекам не является гладким и безупречным, как это может показаться на первый взгляд. Можно выделить следующие основные проблемы:

— перевод информации с бумажных носителей в электронный формат;

— автоматизированное формирование коллекций документов;

— создание информационно-поисковых систем обработки информации;

— разработка методов построения систем управления документами.

В данной статье будут рассмотрены пути решения последней, из перечисленных, проблемы, а именно, изложена концепция построения систем управления электронными документами.

Условно, информацию можно разделить на два вида: структурированная и неструктурированная. В первом случае предполагается, что за ее хранение и управление отвечают базы данных и прикладные информационные системы. Неструктурированная информация – это неупорядоченные, ни по каким признакам, документы или упорядоченные частично. Сегодня пришло понимание необходимости автоматизации хранения и обработки неструктуриро-

ванной информации, так как ее объемы такие, что обрабатывать ее вручную уже не представляется возможным. Для решения подобных задач используются системы управления документами (СУД). Однако подходы и концепции построения таких систем имеют определенные отличия. Рассмотрим основные принципы построения и функционирования систем управления документами.

Выделяют следующие принципы построения СУД:

1. Масштабируемость системы. Оно означает, что система может работать как с одним пользователем, так и с 10000, как с 10 документами, так и с 10 миллионами. При увеличении нагрузки на систему можно сменить сервер, на котором работает система. Поэтому при построении решения нужно руководствоваться принципом поддержки максимально возможного количества операционных систем, если это не удается, необходима поддержка, по крайней мере, операционных систем семейства Windows. Чтобы обеспечить переносимость данных, желательна поддержка многоплатформенных серверов баз данных, таких как Sybase, Oracle, Microsoft, Informix.

2. Распределенность. Основные проблемы при работе с информацией возникают при удаленном обращении к ней. Это значит, что архитектура систем документооборота должна поддерживать взаимодействие распределенных площадок. Причем распределенные площадки могут объединяться самыми разнообразными по скорости и качеству каналами связи.

3. Открытость. Система должна аккуратно вписываться в уже существующие или новые приложения.

4. Модульность и технологичность. Не всегда необходимо внедрять весь комплекс работы с документами в организации сразу. Это может быть вызвано разными причинами: от нехватки средств, до неспособности организации резко перестроить свою работу. Поэтому система должна состоять из модулей, каждый из которых позволяет решить ту или иную задачу, причем эти модули могут быть без особого труда добавлены в работающую систему в произвольной последовательности. И что самое главное они по возможности должны быть независимы друг от друга, при сохранении глубокой интеграции между ними.

К основным задачам систем управления документами относятся:

— регистрация поступающей информации (заполнение необходимых атрибутов документа);

— организация хранения документов (обеспечение хранения произвольного количества документов на разнообразных носителях);

— автоматизация операций с документами (просмотр, создание, копирование, уничтожение);

— организация индексирования и быстрого поиска документов (поддержка индексов различных типов);

— обеспечение безопасности документов (контроль доступа к информации);

— организация коллективной работы с информацией (сокращение времени ожидания запрошенных данных);

— организация распределенных хранилищ документов.

При создании систем управления документами, часто возникает вопрос о поддержке распределенного в пространстве хранилища данных. Под требованием распределенности понимается следующее:

1. Доступ удаленных пользователей к хранилищу информации:

— через локальную сеть;

— посредством Интернет соединения с WWW-сервером;

— через службу доставки сообщений (электронная почта).

2. Взаимодействие нескольких хранилищ и одновременный доступ пользователей к информации, расположенной в разных архивах. Такого рода взаимодействие может быть построено на двух основных принципах:

— взаимное тиражирование хранилищ – технология, которая позволяет каждому пользователю на любом рабочем месте иметь доступ к информации со всех других рабочих мест. Однако, это приемлемо только при небольших объемах информации, и практически не реализуемо для хранилищ документов серьезных размеров (порядка сотни Гигабайт и выше).

— технология распределенного доступа – система регистрирует несколько хранилищ и как бы создает одно глобальное информационное пространство. Пользователь делает один

запрос к глобальному хранилищу и система сама делит запросы по реальным хранилищам, собирает с них ответы и выдает консолидированный результат пользователю.

Рассмотренные принципы построения СУД позволяют сделать вывод, что такие системы обладают широким набором различных функций. Однако, при использовании СУД в электронных библиотеках более целесообразно иметь систему, которая не переполнена излишними функциями, а реализует специализированные функции, направленные на эффективную работу электронных библиотек.

В современных условиях представления информации в глобальных сетях обойтись без использования Web-технологий практически невозможно. Использование таких технологий позволяет работать с данными через обычные Web-браузеры, а они могут быть размещены на самых разнообразных клиентских платформах. Тем самым оказывается отчасти решенной проблема работы в гетерогенной сетевой среде. Для работы в Web-среде необходимо наличие сети поддерживающей TCP/IP-протоколы. Наличие TCP/IP позволяет легко интегрировать систему с другими информационными сервисами, например, электронной почтой. При этом автоматически решается проблема масштабируемости, так как для технологий TCP/IP нет никакой разницы в том, где расположены ресурсы и интерфейс пользователя: локально, в рамках корпоративной сети, или распределены по глобальной сети. При использовании Web-технологии у СУД появляются серверные компоненты, отвечающие за доступ к документам через обычный Web-браузер.

Исходя из всего вышесказанного, можно сделать вывод о необходимости разработки системы управления документами, обладающей специализированными функциями, отвечающими требованиям электронных библиотек, и позволяющей осуществлять работу через Интернет.

Определим методы реализации основных функций системы управления документами и сформулируем концепцию создания системы.

Система должна обладать следующими функциями:

— формирование и сохранение информации с применением алгоритмов сжатия данных;

— организация хранилища данных на сервере;

— обеспечение доступа (просмотр, редактирование) к документам;

— добавление информации в систему, удаление из системы;

— автоматическое выделение ключевых слов из текста документа;

— быстрый поиск данных по ключевым словам;

— автоматическая классификация данных для организации каталога.

Из рассмотрения функций, которыми должна обладать система, можно сделать вывод, что основными задачами, которые необходимо решить при проектировании СУД являются следующие: представление информации в системе, организация хранилища информации, организация каталога и поиск данных в соответствии с информационными потребностями пользователя.

Для хранения внутренней информации системы, такой как индексы, структура каталога с данными о вхождении в него документов и другой дополнительной информации, целесообразно использовать систему управления базами данных (СУБД). Выбор объясняется наличием такой характеристики реляционных баз данных, как небольшое время выборки конкретной записи из миллионов других, что очень важно в условиях большого количества информации в библиотеках и высоких требований к быстродействию системы. Малое время выборки из базы данных достигается путем создания индекса к таблице по одному или нескольким из ее полей. Индексы создаются посредством СУБД и обычно реализуются с применением алгоритма сбалансированного двоичного дерева.

Необходимо решение также проблемы хранения самой информации. На сегодняшний день применяется два подхода к организации хранения электронных документов. Первый состоит в том, что собственно тело документов хранится в файловой системе, второй предусматривает хранение документов в реляционной или специализированной базе данных. Второй подход хотя и обладает большей степенью защиты документов, но несет в себе ряд следующих ключевых недостатков:

— трудности с поддержкой носителей информации, отличных от жестких дисков (немного СУБД поддерживает магнитооптические и другие накопители) и практическая невозможность построения гетерогенных систем хранения;

— при работе с приложениями, в которых создаются и изменяются электронные документы, тела документов в любом случае проходят через файловую систему, а так как приложение не умеет работать напрямую с базами данных, это означает удвоение числа операций записи и считывания с жесткого диска. При больших размерах тел документов это серьезно влияет на скорость работы.

Таким образом, хранение информации необходимо осуществлять в файловой системе с разработкой формата представления документа.

Для обеспечения эффективного поиска документов в системе управления документами необходимо наличие информационно-поисковой подсистемы. Для ее реализации можно использовать методы, применяемые в классических информационно-поисковых системах. Для быстрого поиска по ключевым словам можно использовать методы информационно-поисковых систем словарного типа. Вся информация должна быть проиндексирована по ключевым словам, причем выделение ключевых слов должно производиться автоматически. Поиск данных осуществляется по индексу.

Логически документы должны быть представлены в виде тематического каталога (распределены по рубрикам). Классификацию документов по рубрикам целесообразно производить автоматически с использованием возможностей искусственных нейронных сетей.

Схематично система управления документами может быть отображена в виде представленном на рисунке 1.

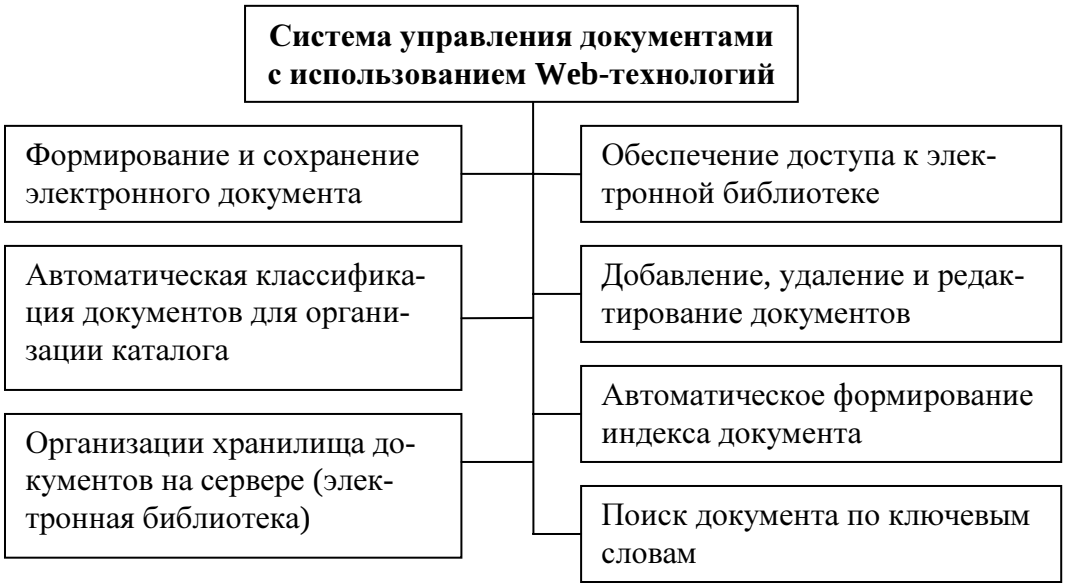

Рисунок 1. Система управления документами

При создании электронной библиотеки следует выбрать: сервер хранения баз данных, тип хранения данных, протокол для передачи данных между сервером и клиентами, и приложения для сервера и клиента.

Для организации современных корпоративных сетей и систем с клиент- серверной архитектурой используется XML в качестве протокола передачи данных и язык Java при построении приложений для сервера и клиентов. Приложения на Java являются эффективными и надежными для корпоративной системы.

При построении электронной библиотеки можно использовать компонент JDBC в качестве базы данных для хранения документов, для системы запросов к базе данных, а также ответов с сервера баз данных использовать протокол SOAP, и сервер для управления базами данных – Apache 2.

Структура электронной библиотеки представлена на рисунке 2.

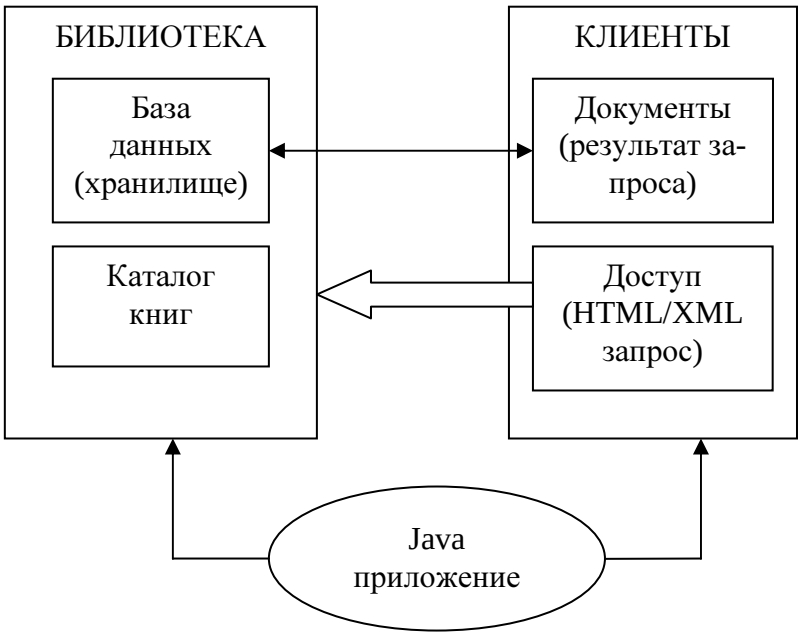

Рисунок 2. Структура электронной библиотеки

В настоящее время разработаны и используются разнообразные подходы к управлению данными, создан целый ряд международных, национальных и индустриальных стандартов в этой области, многие из которых находят применение в коллекциях электронных библиотек. Однако до сих пор остается много открытых вопросов, касающихся проблем построения электронных библиотек и управления информацией в них.

Примечания:

1. Гринберг И., Гарбер Л. Разработка новых технологий информационного поиска. Открытые системы, 1999.
2. Кормен Т., Ривест Р. Алгоритмы, построение и анализ. М.: МЦНМО , 2002.
3. Бишоп Дж. Эффективная работа: Java 2. СПб.: Питер, 2002.