На правах рукописи

Кадэй Тхэй

ПРЕДСТАВЛЕНИЕ И ОБРАБОТКА XML-БАЗ ДАННЫХ

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Автор: 

Москва – 2009

Работа выполнена в Московском инженерно-физическом институте (государственном университете)

Научный руководитель: доктор технических наук, профессор
Щукин Борис Алексеевич

Официальные оппоненты: доктор технических наук, профессор
Ветошкин Владимир Михайлович
кандидат технических наук, доцент
Новиков Валерий Ареанович

Ведущая организация: Негосударственное образовательное учреждение Московский институт повышения квалификации «Атомэнерго» (НОУ МИПК «Атомэнерго»)

Защита диссертации состоится 01 июля 2009 г. в 14 часов 00 минут на заседании диссертационного совета Д 212.130.03 при Московском инженерно-физическом институте (государственном университете) по адресу:
115409, Москва, Каширское шоссе, 31, тел.:(495) 324-84-98, 323-95-26.

С диссертацией можно ознакомиться в библиотеке МИФИ.

Автореферат разослан «__» мая 2009 г.

Отзывы в двух экземплярах, заверенные печатью организации, просьба направлять по адресу:115409, Москва, Каширское шоссе, д.31, диссертационный совет, Шумилову Ю.Ю.

Ученый секретарь
диссертационного совета

д.т.н., профессор  Шумилов Ю.Ю.

Общая характеристика работы

Актуальность работы

Возрастающее использование XML-технологий привело к тому, что стали накапливаться значительные архивы XML-документов, поэтому в последнее время все больше стали говорить о создании XML баз данных, так как частое конвертирование XML-документов в структуры реляционных баз не эффективно из-за слишком большой разницы в структурах. К настоящему моменту создано уже несколько «native» (родных, созданных именно для XML) XDMS, и они непрерывно совершенствуются.

Однако XML-документы не однородны, их можно разделить на две большие группы: документы, ориентированные на данные и документы со смешанным контентом. Если для управления базами документов со смешанным контентом использование «native» XDMS несомненно оправдано, то управление базами документов, ориентированных на данные, которые широко используются в коммерческой и производственной практике, вызывает дискуссии.

Дело в том, что значительная часть документа – разметка без всяких изменений повторяется из документа в документ, что существенно увеличивает объем базы. В диссертации разработаны алгоритмы, позволяющие разделить разметку и собственно данные, что существенно сокращает объем базы и позволяет использовать для работы с ней стандартные методы DBMS, построенных на базе модели данных Pick UDM.

Цель работы

Целью диссертации является исследование и разработка методов и инструментальных программных средств отображения XML-документов в структуры, определяемые моделью Pick UDM, а также разработка экспериментальных приложений, работающих с XML-базами данных.

Для достижения поставленной цели в диссертации решены следующие задачи:

1. Проанализированы современные методы создания и использования XML-баз данных в информационных системах.
2. Проанализированы инструментальные средства работы с XML-базами данных, как создаваемые с «нуля» - «native» XDMS, так и встраиваемые в современные реляционные DBMS.
3. Разработаны алгоритмы отображения XML-схем в структуры, определяемые моделью Pick UDM.
4. Разработаны алгоритмы загрузки XML-документов в базу, управляемую в соответствии с моделью Pick UDM.
5. Разработаны экспериментальные приложения, работающие с XML-базами данных.

Научная новизна

1. Разработаны алгоритмы отображения XML-схем в структуры, определяемые моделью Pick UDM.
2. Сформулированы условия, при выполнении которых отображение XML-документов осуществляется без декомпозиции последних.
3. Разработаны алгоритмы загрузки XML-документов в базу, управляемую в соответствии с моделью Pick UDM.
4. Разработаны экспериментальные программные средства, продемонстрировавшие эффективность разработанных алгоритмов.

Практическая ценность

Разработанные алгоритмы и программные средства могут быть использованы в следующих областях:

1. Создания баз XML-документов с ориентацией на данные, работающих под управлением DBMS с моделью данных Pick UDM.
2. Создание систем динамического гипертекста.

3. Разработанные в диссертации подходы, алгоритмы и программные средства использовались в учебном процессе для студентов Союза Мьянма, проводимом на кафедре «Кибернетика» МИФИ. Автор лично проводил занятия со студентами.

Основные научные результаты, представляемые к защите

1. Предлагаемый подход и алгоритмы для отображения XML-схем в структуры, определяемые моделью Pick UDM.
2. Алгоритмы и программные средства загрузки XML-документов и оформления в виде XML-документа ответа на запрос к базе данных в соответствии с заданной XML-схемой.
3. Алгоритмы и программные средства, позволяющие проводить динамическую разметку текстов и связывать с этой разметкой возможность выполнения определенных операций.

Апробация работы

Основные результаты диссертации докладывались и обсуждались на следующих научных конференциях и семинарах.

1. На научной сессии МИФИ 2008 г; Москва, МИФИ;
2. XII московская международная телекоммуникационная конференция студентов и молодых ученых – 2009 г.
3. На Всероссийской межвузовской научно-технической конференции студентов и аспирантов «Микроэлектроника и информатика - 2008 г; Москва, МИЭТ;
4. На международном научно-техническом семинаре «Современные технологии в задачах управления, автоматики и обработки информации» 2007, 2008 гг. (Алушта).
5. Опубликованы в 2009 году в журнале «Безопасность информационных технологий».

Публикации

Результаты диссертации опубликованы в 6 печатных трудах, в том числе в одной статье в журнале, который включен ВАК РФ в перечень ведущих рецензируемых научных журналов и изданий, и тезисах докладов в сборниках трудов конференций.

Структура и объем работы

Диссертация содержит 4 главы, введение и заключение, 48 рисунков, 1 таблица и 5 приложений.

Общий объем – 128 страниц машинописного текста. Список использованных источников содержит 58 наименования.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, её научная новизна и практическая значимость, сформулирована цель работы.

В главе 1 рассматриваются основные положения XML-технологии, анализируется проблема создания XML-баз данных, рассматриваются подходы к решению этих задач на базе «native» XDMS и специальных средств, встраиваемых в современные реляционные DBMS. Поднимается проблема места XML-баз данных в современных информационных системах. В конце первого раздела диссертации поставлена цель и конкретные задачи диссертационного исследования.

XML-технология разработана под эгидой консорциума W3C и рассматривалась как технология следующего поколения Internet. Однако сама технология и идеи, заложенные при ее создании, оказались столь востребованы, что она превратилась в нечто вполне самостоятельное, так как XML-документы одинаково понятны как человеку, так и машине и могут использоваться не только в сфере Web-приложений.

Для человека имена тегов и атрибутов играют как структуро-образующую, так и семантическую роль, которые позволяют ему правильно

интерпретировать входящие в XML-документ данные. Для машины – это всего лишь один из способов иерархической структуризации данных: теги выполняют чисто синтаксическую функцию.

Данные, оформленные (размеченные тегами) в соответствии с правилами XML, называют XML-документами. Логическая модель таких данных – это разновидность модели полуструктурированных данных.

Существует два типа размеченных данных, которые называют XML-документами.

- правильно оформленные документы (well-formed);
- правильные документы (valid).

Документ считается правильно оформленным, если при его разметке не нарушены правила порождения тегов. Правильным считается документ, теги которого и вся структура определяются предписывающей XML-схемой документа.

Это значит, что при создании правильно оформленного документа могут использоваться любые имена тегов, а для правильного документа – только имена, определенные предписывающей XML-схемой.

Можно дать и другое определение. Документ считается правильно сформированным, если по его тексту можно сгенерировать описывающую XML-схему. Документ считается правильным, если он принадлежит множеству документов, порождаемых предписывающей XML-схемой.

Фактически XML - это метаязык разметки (рис.1). На его основе создаются описания других языков, которые непосредственно используются для разметки документов. Это реализуется с помощью XML схем. Язык, используемый для создания XML схем, может быть разработан без использования синтаксиса XML, в этом случае схема XML-документа не будет XML-документом. Таким языком является Document Type Definition (DTD).

В настоящее время отдается предпочтение языкам создания XML схем, использующим синтаксис XML. Примером такого языка является W3C XML

Schema, определенного в спецификации, принятой в качестве рекомендации W3C. Поэтому XML-схема, созданная на языке W3C XML Schema сама является XML-документом. Однако для решения задачи, которая решается в диссертации, а именно отображения XML-схемы в последовательность атрибутов (описывающую схему в модели Pick UDM) удобнее работать с XML-схемой в синтаксисе DTD. При этом ограничимся только рассмотрением отображения элементов и атрибутов.

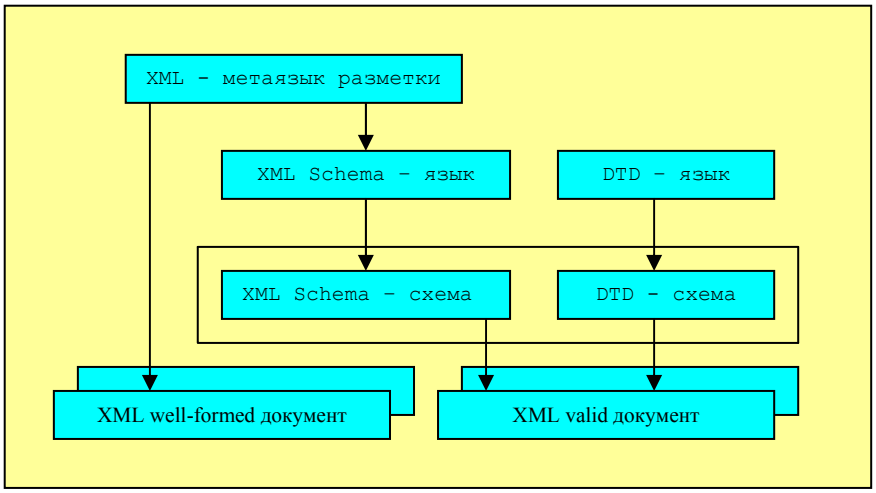

Рис.1. XML - метаязык разметки

XML-схема может быть представлена в виде дерева (рис.2).

Пример древовидной структуры, представленной на рис.2, отражает логику представления XML-документа в основной памяти. Эта структура лежит в основе Document Object Model (DOM) – рекомендации консорциума W3C, предназначенной для манипулирования элементами XML-документа в программном тексте.

Фактически DOM определяет "интерфейс, не зависящий от платформы и языка, который позволяет программам и скриптам динамически получать доступ и обновлять содержимое, структуру и стили документов".

По существу, это древовидная структура данных, находящаяся в основной памяти, дополненная набором методов для доступа и редактирования XML документа. Т.е. DOM позволяет выполнять все операции по обработке XML-данных: можно не только читать данные, но и

модифицировать содержимое XML-документа, вставляя туда новые теги, удаляя и изменяя их.

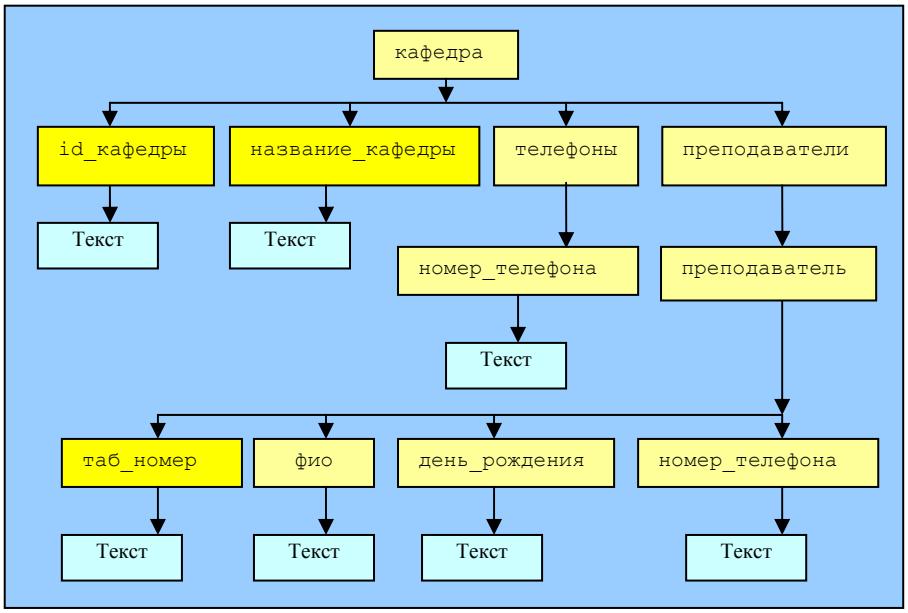

Рис.2. Дерево, представляющее XML-схему

Термин «модель данных» в литературе по базам данных и реальной практике интерпретируется по разному: как средство для моделирования и как результат моделирования. Для обозначения результата моделирования, под которым обычно понимается предписывающая схема базы данных.

Как средство моделирования модель данных обязана включать:

- средства для декларации структуры данных;
- средства для манипулирования данными;
- средства для декларации ограничений целостности.

Например, в теоретической реляционной модели схема базы данных определяется как совокупность схем отношений, манипулирование данными осуществляется средствами алгебры или исчисления, с помощью этих же средств декларируются ограничения целостности. В практических реляционных базах, которые иногда называют SQL-базами данных, схема базы определяется как совокупность структур таблиц, манипулирование данными и декларация ограничений целостности осуществляется средствами языка SQL.

По аналогии, схема XML-базы данных должна определяться совокупностью XML-схем документов, манипулирование данными XML-базы данных – средствами специального языка, в качестве которого в настоящее время предлагается использовать XQuery. По крайней мере, базовые средства для декларации целостности по ссылкам должны быть определены в этом языке.

На самом деле в традиционном понимании термина «база данных» XML-баз данных пока нет, в реальности есть базы XML-документов. Может быть поэтому в аббревиатуре названия систем управления ими часто опускают букву «B»: вместо XDBMS используют аббревиатуру XDMS. Однако термин «XML-база данных» прижился, поэтому в тексте диссертации используется также аббревиатура XML DBMS, хотя в реальности он обозначает систему управления базой XML-документов.

В современных реляционных системах однородные данные объединены в таблицы, аналогично, в объектных базах однородные данные разбиваются по классам. В XML-базах данных в большинстве своем хранятся XML-документы, у которых нет предписывающей XML-схемы. Все эти документы сваливаются в единое хранилище, они могут быть определены на разных пространствах имен, что, в общем случае, существенно затрудняет работу с ними.

XML-документ, сохраняемый в базе, самодостаточен – в нем должна присутствовать вся информация, которую он представляет, никаких ссылок вне документа быть не должно. Действительно, пусть XML-документ представляет некоторый коммерческий счет на продукцию. Возьмем один из многих продуктов, представленных в счете. Все его атрибуты: код, наименование, цена, скидка и т.д. должны присутствовать как реальные данные. В этом смысле XML-база данных будет информационно избыточна: в ней должен храниться информационный образ реального счета.

Для поиска в базах XML-документов в настоящее время определен в качестве стандарта язык XQuery, в него же встроены средства для

модификации данных. XQuery очень гибкий язык, допускающий совместное использование с SQL. Этот язык стандартизирован консорциумом W3C, в его разработке принимали участие ведущие специалисты компьютерной отрасли. XQuery позволяет извлекать древовидные данные, трансформировать их и генерировать в качестве результата опять же древовидные данные. Это позволяет строить эффективные решения в области Интернет-приложений: функции, написанные на XQuery, могут генерировать непосредственно XHTML-страницы (или фрагменты страниц).

Более глубокую обработку найденных XML-документов можно проводить средствами языка программирования, например, Java, используя API на базе объектной модели документа. Структурное преобразование XML-документов целесообразно проводить средствами языка XQuery, а также средствами XSLT.

В заключение первой главы ставится задача диссертации.

XML-документы принято делить на две большие группы: «ориентированные на данные» и «ориентированные на документы» («data centric» и «document centric»). Более подробно эти группы будут описаны в дальнейшем, сейчас же заметим, что для документов первой группы главное – это данные, в основном – коммерческие, и их иерархическая структуризация, а для документов второй группы главное – разметка содержания. Это значит, что документы первой группы имеют дело с данными, являющиеся объектом хранения традиционных систем баз данных, а архивы соответствующих XML-документов состоят из документов электронной коммерции, B2B систем и т.д. Для документов первой группы характерно наличие предписывающих XML-схем, то есть они относятся к категории «valid». Архивы XML-документов второй группы – это содержательные документы интернет-сайтов, например документы XHTML. Разметка документов, так называемая смешанная разметка, осуществляется с ориентацией на определенный словарь, но как таковой предписывающей XML-схемы часто нет, так как она мало информативна.

Разумеется, как те, так и другие документы могут быть объектами хранения в XML-базах данных, однако для документов первой группы вполне могут подойти и традиционные DBMS с моделью данных, предполагающей более глубокую иерархию по сравнению с реляционной.

В качестве такой модели данных целесообразно исследовать модель Pick UDM, так как построенная на ее основе XDMS TigerLogic запатентована и обеспечивает столь высокую производительность при работе с XML-базами. XDMS TigerLogic в России в настоящее время недоступна, однако доступна DBMS D3, на которой можно промоделировать разработанные алгоритмы и методики. Таким образом, в диссертации предполагается решить следующие задачи:

1. Исследовать возможность работы с базой XML-документов исключительно средствами DBMS D3;
2. Исследовать ограничения на структуру XML-документа, позволяющую непосредственно загружать данные в область данных, а теговую структуру в словарь;
3. Разработать средства оформления отчетов на запросы к базе данных D3 в виде XML-документа со структурой, задаваемой предписывающей XML-схемой;
4. Разработать экспериментальную базу данных и систему для работы с ней в среде DBMS D3 и в среде XML DBMS.

В главе 2 развивается подход к отображению XML-документов в структуры, определяемые моделью Pick UDM. Выделяется подкласс XML-документов «ориентированных на данные», для которых такое отображение реализуется естественно и эффективно.

На базе модели Pick UDM разработаны СУБД D3 (TigerLogic) UNIVERSE (IBM), UNIDATA (IBM) и многие другие. В отличие от реляционной, эта модель поддерживает многозначность, единственный тип данных – строки переменной длины и определяет атрибуты как функциональные преобразования от данных, хранящихся в базе.

В соответствии с этой моделью база данных рассматривается как совокупность файлов. Стандартно «файл базы данных» состоит из «словаря», используемого для хранения атрибутов и «области данных» - для собственно данных, хотя и те и другие организованы абсолютно одинаково. С каждым словарем может быть связано несколько областей данных, в том числе и ни одной. Это делает модель Pick UDM более удобной, по сравнению с реляционной, для отображения XML-документов на структуры данных, определяемые этой моделью (рис.3), так как основной компонент файла – запись имеет практически неограниченную длину.

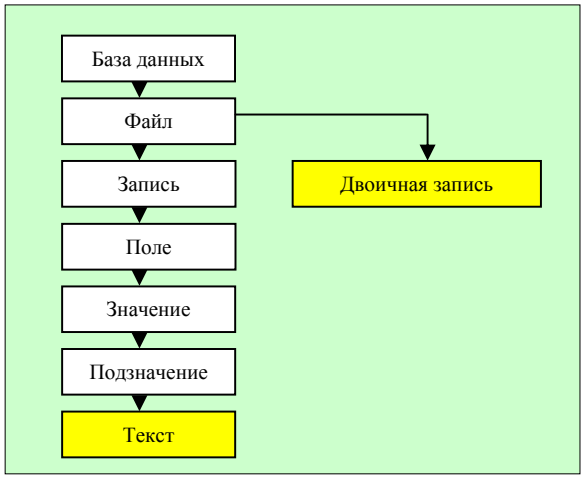

Рис.3. Структура данных модели Pick UDM

Для отображения XML-документов будем говорить только о записях с текстовыми данными. Любая запись, будучи считанной из файла базы данных в некоторую переменную «основной памяти», например, Item принимает структуру динамического массива, к элементам которого можно обращаться по целочисленным индексам: Item<i, j, k>, i – номер поля записи, j – номер значения, k – номер подзначения. Слово «динамический» означает, что границы измерений априорно не определены и могут быть сколь угодно большими.

Модель Pick UDM не предполагает обязательную предписывающую схему, которая однозначно диктует структуру записей файлов базы данных. Так как единственным типом данных является текст, то его интерпретация может быть самой разнообразной. Поиск в базе данных может быть проведен

с ориентацией на произвольную описывающую схему, состоящую из атрибутов одного или нескольких словарей.

Возникает несколько задач, связанных с отображением XML-документов на структуры, определяемые моделью Pick UDM. Остановимся только на двух:

1. Конвертировать совокупность XML-документов, построенных в соответствии с некоторой XML-схемой, в файл базы данных.
2. Получить ответ на запрос к базе данных, построенной в соответствии с моделью Pick UDM, в форме XML-документа с заданной XML-схемой.

Решение связано с разработкой алгоритма отображения XML-схемы в записи определения атрибутов, сохраняемых в словаре, и разработке алгоритма загрузки информационной составляющей XML-документа в записи области данных (рис. 4).

В случае XML-документа произвольной структуры его информационная составляющая отображается в несколько записей области данных. Практически важны случаи, когда XML-документ отображается в единственную запись. Эти случаи можно описать, сформулировав следующее утверждение:

Утверждение. Для отображения XML-документа в виде одной записи необходимо, чтобы в дереве, построенном по преобразованной XML-схеме этого документа, на каждом пути из корня дерева в висячую вершину было не более двух модификаторов «+». (Модификатор «*» рассматривается как «+»).

Большинство экономических документов удовлетворяют этому ограничению.

Рассмотрим алгоритм отображения XML-схемы в структуру, определяемую моделью Pick UDM. Считаем, что XML-схема представлена в

синтаксисе DTD. Каждому элементу (тегу) и каждому атрибуту элемента необходимо поставить в соответствие атрибут модели Pick UDM.

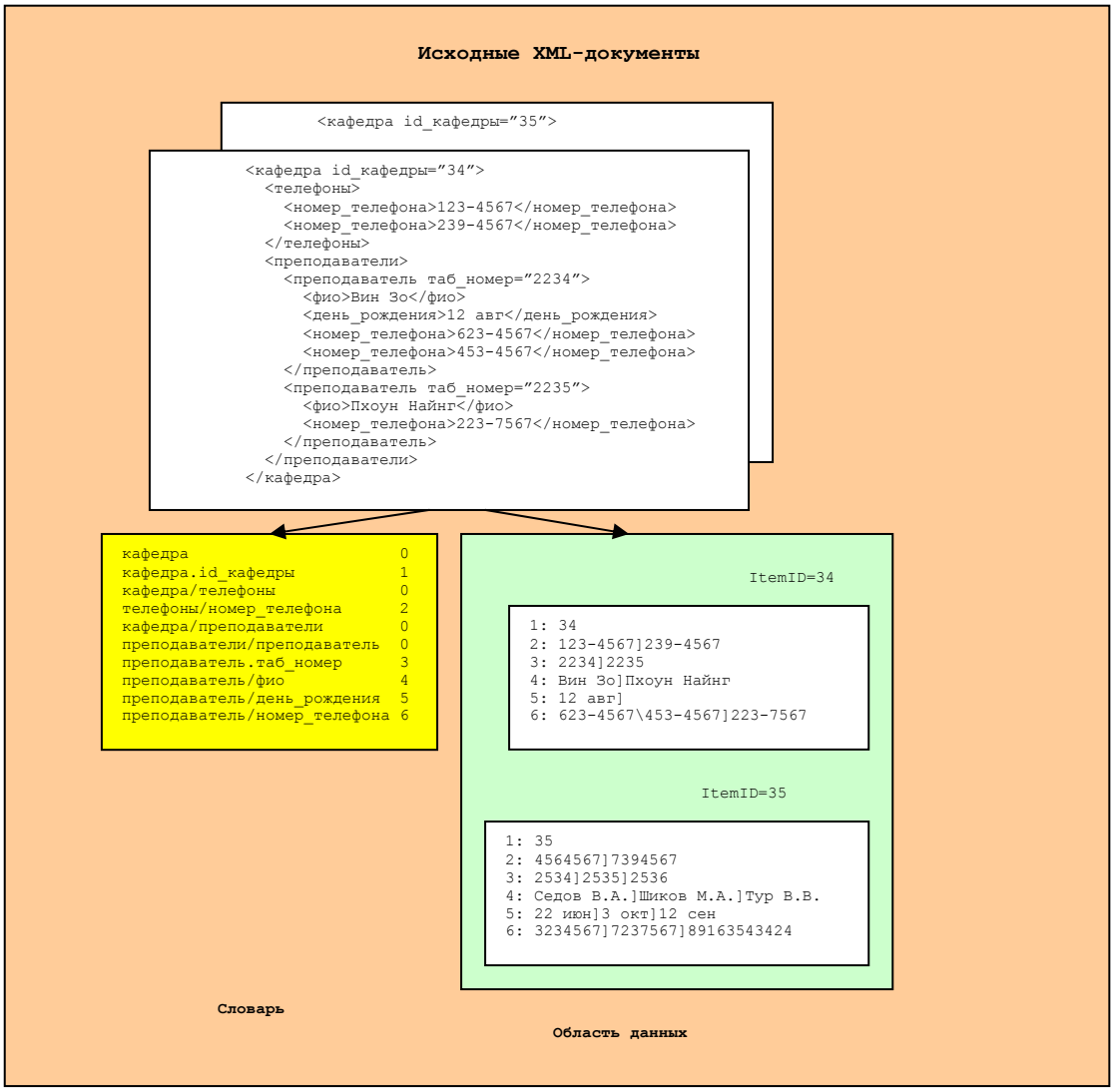

Рис.4. Отображение XML-документов

Все элементы в XML-схеме необходимо расположить в стандартном иерархическом порядке в соответствии с принципом «сверху – вниз, слева - направо». Декларация атрибутов элемента должна следовать непосредственно за декларацией элемента, причем атрибут типа ID должен быть декларирован первым.

Рассмотрим путь от корня дерева, представляемого XML-схемой, до его висячей вершины. Примерами таких путей могут служить (см. рис.4).

`кафедра/преподаватели/преподаватель.табномер`

`кафедра/преподаватели/преподаватель/фио`

Первый путь заканчивается именем атрибута, а второй – именем негруппового элемента. Эти пути можно интерпретировать как имена соответствующих данных.

Для того чтобы иметь возможность полностью восстановить исходный XML-документ, сохраненный в базе данных в «разобранном» виде, необходимо каждому «подпути» поставить в соответствие атрибут модели Pick UDM:

```
кафедра
кафедра/преподаватели
кафедра/преподаватели/преподаватель
кафедра/преподаватели/преподаватель.таб_номер
кафедра/преподаватели/преподаватель/фио
```

При этом всем групповым тегам ставятся в соответствие виртуальные атрибуты модели Pick UDM, которые не имеют собственных значений и которым традиционно ставится в соответствие поле с номером 0. Собственные значения имеют атрибуты модели Pick UDM, соответствующие висячим вершинам дерева, представляющего XML-схему. Им должны соответствовать конкретные номера полей записи области данных.

При предположении, что имена групповых тегов не входят в другие группы, имена атрибутов можно сократить и представить в виде:

```
кафедра
кафедра/преподаватели
преподаватели/преподаватель
преподаватель.таб_номер
преподаватель/фио
```

Таким образом, имена атрибутов модели Pick UDM образуются следующим образом:

- Корневому элементу XML-документа ставится в соответствие атрибут модели с именем корневого элемента, например, `<кафедра> -> кафедра`.

- Дочернему элементу XML-документа ставится в соответствие атрибут модели, имя которого образуется конкатенацией имени родителя, символа «/» и имени дочернего элемента:
  `<преподаватель> -> преподаватели/преподаватель.`
- Атрибуту XML-документа ставится в соответствие атрибут модели, имя которого образуется конкатенацией имени элемента, символа «.» и имени атрибута:
  `<преподаватель табномер> -> преподаватель.табномер.`

Формирование атрибутов модели Pick UDM производится последовательной обработкой строк XML-схемы, при этом считается, что корневой элемент всегда групповой. Всем групповым тегам ставятся в соответствие виртуальные атрибуты, которые не имеют собственного значения и которым традиционно ставится в соответствие поле с номером 0. Негрупповым тегам и атрибутам последовательно приписываются номера полей.

Загрузка XML-документа в базу данных выполняется при условии, что внутреннее представление XML-схемы предварительно построено.

Считается, что XML-документ представлен текстовым файлом с расширением .xml в каталоге windows, unix или linux. С точки зрения модели Pick UDM — это одна запись, в которой i-ая строка файла представляет значение i-ого поля. Алгоритм загрузки предполагает, что XML-документ представлен в нормализованном виде, то есть его строки соответствуют строкам дерева элементов, получаемым при визуализации в браузере.

Предлагаемая технология позволяет создать компактную базу XML-документов, в которой средствами СУБД D3 решаются все вопросы модификации и поиска. Восстановление XML-документа в виде .xml файла или оформление ответа на запрос к базе в виде XML-документа с XML-схемой, отличной от исходной, выполняется с несущественными затратами времени.

Существуют две задачи получения XML-документа из базы данных, построенной в соответствии с моделью Pick UDM:

1. оформление в виде XML-документа динамического массива;
2. оформление в виде XML-документа ответа на произвольный запрос к базе данных.

В первом случае получение XML-документа производится в соответствии с заданной внутренней XML-схемой. Во втором случае искомое решение будет получено, если результаты произвольного запроса представить в виде динамического массива. Эта задача полностью решается стандартными средствами DBMS D3 путем промежуточной выдачи результатов запроса в отдельный файл. Значения вычисляемых атрибутов при этом записываются в реальные поля записи, которая при считывании превращается в динамический массив.

В главе 3 рассмотрены вопросы отображения XML-документов со смешанным контентом в структуры, определяемые моделью Pick UDM. Разрабатывается идея динамической разметки линейного текста и связывания с этой разметкой определенных операций, в частности, выполнения гипертекстовых переходов, выполнения операций с базой данных и т.д.

Документы со смешанной разметкой – это широкий класс XML-документов, например, практически все тексты XHTML. Однако в XHTML теги разметки используются исключительно для целей представления, например, какой то фрагмент строки нужно выделить курсивом, что достигается заключением его в теги . С точки зрения диссертации смешанная разметка интересна как простейший способ внесения в текст некоторой семантики.

Элемент со смешанным контентом декларируется в DTD строкой, построенной в соответствии со следующим шаблоном:

```
<!ELEMENT имя_родителя (#PCDATA|имена_дочерних_элементов)*>
```

Элемент содержит текст, который может (но не обязан) включать фрагменты (до, между, после планарного текста), размеченные тегами с

именами дочерних элементов. Эти теги могут располагаться в произвольной последовательности, произвольное количество раз, некоторые из них в реальном XML-документе могут не появиться вообще. С другой стороны, в реальном XML-документе не может появиться тег не декларированный в DTD.

XML-документ, ориентированный на документы, обычно не имеет предписывающей XML-схемы. Вполне достаточно того, чтобы теги были расставлены в соответствии с правилами языка XML. С другой стороны, если документ правильно оформлен, то по его тексту всегда можно сгенерировать описывающую XML-схему.

На практике оказывается очень востребованным простейший вариант элементов со смешанным контентом, в котором все входящие дочерние элементы являются конечными. В этом случае элементы разметки можно выделить в отдельное поле, поставив в соответствие текстовой составляющей специальный тег, например, слово «текст».

На рис. 5 для XML-документа, в поле 002 помещена последовательность тегов, а в поле 003 оставлен собственно текст. И теги и фрагменты текста разделены символом «]» – разделителем значений. На этих полях можно определить атрибуты, связанные зависимостью управляющий-подчиненный, которые позволят четко сопоставить каждому тегу его текстовое содержание.

001 call

002 текст]operator]текст]fdi]текст]adi]текст] dict_atr]текст]dict_atr]текст]dict_atr]текст

003 При вызове подпрограммы никакие параметры не указываются, но в любом случае параметр подпрограмме передается, поэтому один параметр должен быть задан в операторе]Subroutine]. При вызове подпрограммы из]записи описания файла] в подпрограмму передается вся запись. Если подпрограмма вызывается из]записи описания атрибута], в качестве параметра ей передается значение вызывающего атрибута. Подпрограммы могут быть вызваны из атрибутов корреляций]Correlatives], входных преобразований]Input Conversion] или выходных преобразований]Output Conversion] записей описания файла или записей описания атрибута.

Рис.5. Отделение тегов XML-документа от контента

В главе 4 посвящен экспериментальной проверке предлагаемых подходов к работе с XML-базами данных на основе разработанного Web-приложения. Анализируется технология JSF как технология практического построения Web-приложений на основе MVC. Разработана XML-база данных для хранения и поиска документов со смешанным контентом и продемонстрированы результаты поиска на основе SQL и XQuery.

Предположение, что смешанная разметка возникает в результате вторичной разметки XML-документов с четкой статической структурой, информация в которых представляется в виде значительных фрагментов планарного текста. С этими документами могут работать разные люди и каждый оценивает содержимое документа под углом собственной точки зрения. По мере работы с документом отдельные фрагменты текста подвергаются дополнительной разметке с помощью тегов из некоторого ограниченного множества, специфичного для специалиста определенного профиля. Дополнительно размеченный документ может быть сохранен в собственной XML-базе данных и в дальнейшем использован для целей анализа, поиска прецедентов и т.д. В настоящее время выполнение подобной семантико-прагматической функции скорее всего, прерогатива человека, но в дальнейшем эта работа может быть будет выполняться интеллектуальными роботами.

В данной главе речь пойдет о построении системы, которая может использоваться при редактировании XML-документов, настраиваться на различные предметные области, использоваться для поиска в XML-базе данных и т.д.

В диссертации, на примере создания базы данных по сбойным ситуациям и описанию способов их устранения, продемонстрирована целесообразность создания таких гибридных систем, совмещающих SQL и XML данные. Проблема создания базы данных по сбойным ситуациям часто встает перед фирмами, производящими обслуживание некоторого оборудования. Функционально системы, поддерживающие такие базы, должны обеспечить

средства для описания особенностей сбойной ситуации и способов устранения ее, средства для формирования поисковых запросов и средства для удаленного доступа к базе данных, обычно через Интернет.

Функционал «Разметка»

Разметка выполняется на Web-клиенте и начинается с неразмеченного планарного текста. Априорно считается, что текст заключен между корневым тегом, например, `<repair>` и его закрывающим аналогом `</repair>`. Текст может быть набран непосредственно в окне или выбран из XML-базы данных. Заметим, что это «valid» XML-документ.

При выполнении разметки курсором выделяется фрагмент текста, ограниченный пробелами, и вызывается список допустимых тегов. Выбирается требуемый тег и «кликом» осуществляется разметка, то есть выбранный тег и его закрывающий аналог обрамляют выделенный фрагмент. Если теперь внутри размеченного фрагмента снова выделить некоторый текст, то при его разметке будут вызываться теги второго уровня, то есть теги определенные в XML-схеме документа в элементе, которым вводится обрамляющий тег.

Чтобы аннулировать уже сделанную разметку надо полностью выделить фрагмент текста, находящегося между открывающим и закрывающим тегами и «кликнуть». После подтверждения разметка удаляется.

Размеченный текст сохраняется в XML-базе данных и может неоднократно модифицироваться.

Функционал «Поиск»

Поиск может производиться как по SQL-столбцам, так и по XML-столбцам. Для последнего используется XQuery.

Для поиска по SQL-столбцам на экранной форме клиента обозначаются названия SQL-атрибутов и поля для ввода соответствующих значений.

Для поиска по XML-фрагменту на экране клиентской формы выделено три поля. Первое поле предназначено выбора тега первого уровня, второе

поля – для выбора тега второго уровня, третье поле предназначено для искомой текстовой составляющей, в частности, это может быть слово «ANY», которое обозначает «любой текст». В этом случае интересно само присутствие соответствующей разметки.

Значение всех полей передаются на сервер приложений, где формируется поисковая строка, которая затем передается на выполнение на сервер базы данных. Ответ на запрос выдается в виде XML-документа, для визуального представления которого написана программа на XSLT.

Рассмотрены вопросы построения программной оболочки для Web-приложения, выполняющего подобные функции. Анализируется технология JSF как технология практического построения Web-приложений на основе MVC. Разработана XML-база данных для хранения и поиска документов со смешанным контентом и продемонстрированы способы поиска XML-документов на основе использования языков SQL и XQuery.

В отличие от подхода, использованного в главе 3, когда для разметки использовались только конечные теги, здесь рассмотрен общий случай, когда смешанная разметка существенным образом зависит от обрамляющего тега. Это предполагает однозначную ориентацию либо на создание базы под управлением «native» XDMS, либо на использование встроенных XML-столбцов в реляционных системах.

К сожалению, сравнивать реальные характеристики приложений не представляется возможным, так как первое работает с энциклопедией, состоящей из 2640 объемных записей и занимающей несколько мегабайт, а второе – только с тестовой базой данных из нескольких десятков сбойных операций. При работе над диссертацией для экспериментов использовалась DBMS DB2 v9, которая обеспечивает широкий набор функций для работы с XML-документами.

В заключении приведены основные результаты диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

При выполнении данной работы получены следующие основные результаты:

1. Проанализированы существующие подходы к созданию XML-баз данных. В результате анализа показано, для создания баз XML-документов с ориентацией на данные совершенно не обязательно использовать специальные XDMS.
2. Проанализированы подходы к созданию современных XDMS и показано, что системы, построенные на базе модели Pick UDM демонстрируют повышенную производительность и масштабирование.
3. Разработан алгоритм преобразования XML-схемы документов, ориентированных на данные, в схему базы данных, определяемую моделью Pick UDM.
4. Разработан алгоритм загрузки XML-документов в базу данных, работающую под управлением DBMS с моделью Pick UDM.
5. Разработаны программные средства, реализующие перечисленные алгоритмы.
6. Проведена экспериментальная проверка работы предложенных алгоритмов
7. Разработана многотерминальная система, позволяющая создавать динамический гипертекст в среде DBMS D3. Система использована в учебном процессе кафедры 22 МИФИ при обучении студентов Союза Мьянма.
8. Разработано экспериментальное Web-приложение с XML-базой данных, использующее разработанные в диссертации алгоритмы и ориентированное на создание баз XML-документов со смешанным контентом и выполнение поисковых операций.

Результаты работы показывают, что поставленные цели диссертации можно считать достигнутыми. Эксперименты подтвердили теоретические разработки, предложенные в диссертации, и показали возможность использования баз данных, построенных на базе модели Pick UDM, для эффективного хранения и обработки XML-документов, ориентированных на данные.

По теме диссертации опубликованы следующие работы:

1. Кадэй Тхэй, Щукин Б.А, Безопасность хранения XML-документов // «Безопасность информационных технологии», 2009 г. №1, с 45-49.
2. Кадэй Тхэй, Вин Зо, Моделирование процесса взаимодействия локальных систем при их интеграции // «Современные технологии в задачах управления, автоматики и обработки информации: Труды XVI Международного научного технического семинара», Алушта, с 42.
3. Кадэй Тхэй, Обработка RDF данных средствами реляционных СУБД // «Современные технологии в задачах управления, автоматики и обработки информации: Труды XVII Международного научного технического семинара», Алушта, с 230.
4. Кадэй Тхэй, Работа с базами данных, содержащими XML-документы // «Научная сессия МИФИ-2008. Сборник научных трудов. В 15 томах. Т.11. Программное обеспечение технологии» М.: МИФИ, 2008, с 78.
5. Кадэй Тхэй, Включение XML-столбцов в реляционные таблицы // «Микроэлектроника и информатика – 2008. XV-ая Всероссийская межвузовская научно-техническая конференция студентов и аспирантов: Тезисы докладов» М.: МИЭТ, 2008, с 195.
6. Кадэй Тхэй, Приложения с XML-базами данных // «Научная сессия МИФИ-2009. XII Московская международная телекоммуникационная конференция студентов и молодых ученых «МОЛОДЕЖЬ И НАУКИ». Тезисы докладов. В 2-х частях. Ч. 2.» М.: МИФИ, 2009, с 89-90.