На правах рукописи

МАКАРОВА Екатерина Сергеевна

ФУНКЦИИ АНАЛИТИКИ В ВЕБ-ПРИЛОЖЕНИЯХ НА ОСНОВЕ СИТУАЦИОННО-ОРИЕНТИРОВАННЫХ БАЗ ДАННЫХ

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата технических наук

Уфа – 2013

Работа выполнена на кафедре автоматизированных систем управления ФГБОУ ВПО «Уфимский государственный авиационный технический университет»

Научный руководитель — д-р техн. наук, проф.
МИРОНОВ Валерий Викторович

Официальные оппоненты — д-р техн. наук, проф.
МАРТЫНОВ Виталий Владимирович
зав. кафедрой экономической информатики,
ФГБОУ ВПО «Уфимский государственный авиационный технический университет»

канд. техн. наук
АЛИМБЕКОВА Софья Робертовна
начальник отдела управления проектами
ООО «НИИ технических систем «Пилот»»

Ведущая организация — ГБОУ ВПО «Башкирская академия государственной службы и управления при Президенте Республики Башкортостан»

Защита диссертации состоится «06» декабря 2013 г. в 10.00 часов на заседании диссертационного совета Д-212.288.07 при Уфимском государственном авиационном техническом университете по адресу: 450000, г. Уфа, ул. К. Маркса, 12

С диссертацией можно ознакомиться в библиотеке университета

Автореферат разослан «___» 2013 г.

Ученый секретарь
диссертационного совета
д-р техн. наук, проф. И. Л. Виноградова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Современный этап развития баз данных характеризуется активным исследованием нереляционных подходов, которые более эффективны при решении новых задач, в частности, возникающих при создании веб-приложений (документно-ориентированные базы данных, базы данных NoSQL и др.). Примером таких баз данных, рассматриваемых в данной работе, являются развиваемые на кафедре автоматизированных систем управления УГАТУ так называемые ситуационно-ориентированные базы данных (СОБД). В основе СОБД лежит динамическая модель предметной области, с состояниями которой ассоциированы данные в виде XML-документов. Такая организация данных делает удобным использование СОБД при построении веб-приложений, управляемых встроенными динамическими моделями (Model Driven Approach).

Современные веб-приложения развиваются в направлении реализации функций аналитики (Web OLAP), предоставляющих пользователям гибкий, простой и удобный доступ к гиперкубам данных с разной степенью детализации. Это требует использования технологий OLAP в нетрадиционных условиях. Если традиционные, ориентированные на OLAP, хранилища данных (Data Warehouse) имеют специальную организацию данных в виде многомерной модели (Multidimensional Model) или особой реляционной модели данных (структуры типа «звезда» или «снежинка»), то в данном случае приходится использовать базы данных, ориентированные на традиционные задачи OLTP и формировать гиперкубы «на лету». Эти вопросы пока недостаточно исследованы в концептуальном плане, а также в плане моделей, методов, алгоритмического и программного обеспечения, как сами по себе, так и применительно к СОБД. Задача реализации аналитики на основе СОБД дополнительно усложняется за счет того, что база данных имеет иерархическую (нереляционную) документно-ориентированную структуру и отсутствуют методы формирования гиперкубов данных из совокупности XML-документов.

Таким образом, возникает актуальная научно-техническая задача создания научно обоснованного подхода к реализации функций аналитики в веб-приложениях, основанных на СОБД, в которых в определенных состояниях динамической модели формируются «на лету» гиперкубы данных на основе данных из хранилища XML-документов.

Степень разработанности темы исследования. OLAP (On-Line Analytical Processing – «аналитическая обработка в реальном времени») – технология обработки данных, заключающаяся в подготовке суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу. Термин OLAP был предложен в 1993 году Э. Коддом в противовес термину OLTP (On-Line Transaction Processing – «обработка транзакций в реальном времени»). Кодд сформулировал 12 классических правил OLAP. В последние годы OLAP находит все большее применение в информационно-аналитических системах, в том числе ориентированных на Web.

В настоящее время существует большое количество работ, посвященных OLAP, выполненных как зарубежными, так и отечественными авторами, в том числе работы А. Бергегра, Э. Спирли, Р. Кимбалла, Б. Инмона, Э. Томсена, а также А. А. Барсегян, М. С. Куприянова, Н. Б. Паклина, А. А. Демидова, В. А Филиппова и др. В УГАТУ различные аспекты OLAP-технологий исследовались Ю. С. Кабальновым (концепция многомерных баз данных применительно к объектам нефтедобычи и использование OLAP в системах реального времени), С. В. Павловым и О. И. Христодуло (многомерные информационные объекты и их приложение к геоинформационным системам), Г. Г. Куликовым, В. В. Антоновым и Д. В. Антоновым (модели описания хранилища данных) и др.

СОБД предложены и исследуются в УГАТУ В. В. Мироновым и его учениками. Впервые понятие СОБД было введено в работе В. В. Миронова, Н. И. Юсуповой и Г. Р. Шакировой. Ранее различные аспекты иерархических ситуационных моделей, лежащих в основе СОБД, исследовались в рамках кандидатских диссертаций Ю. Б. Головкина, Р. А. Ярцева, Л. Е. Гончар, О. Н. Сметаниной, А. Н. Ситчихина, Р. Ф. Ахметшина, Т. А. Гарифуллина, Г. Р. Шакировой. Применение встроенных динамических моделей и XML как основы для разработки интернет-приложений рассматривалось в кандидатских диссертациях К. Э. Маликовой, А. С. Гусаренко.

Актуальность обозначенной проблемы определила цель и основные задачи исследования.

Объектом исследования являются веб-приложения на основе СОБД, подходы к их построению.

Предметом исследования являются способы формирования «на лету» гиперкубов данных для задач Web OLAP на основе хранилища XML-документов СОБД.

Целью исследования является обеспечение функций аналитики и снижение трудоемкости программирования OLAP-функциональности в веб-приложениях, основанных на СОБД.

Задачи исследования:

1. Разработать подход к проектированию многомерной модели данных для решения «на лету» задач Web OLAP на основе базы данных OLTP.
2. Проанализировать особенности многомерной модели OLAP, возникающие при построении ее на основе ER-модели OLTP.
3. Разработать методологические и лингвистические средства реализации функций аналитики в веб-приложениях на основе СОБД.
4. Разработать программное обеспечение для реализации функций аналитики в веб-приложениях на основе СОБД и подтвердить результаты на исследовательском прототипе реального веб-приложения.

Научная новизна результатов *в целом* заключается в идее создания оснащенных OLAP-функциональностью веб-приложений на основе СОБД, в которых в определенных состояниях динамической модели из хранилища XML-документов «на лету» формируется гиперкуб данных и пересылается в OLAP-клиент пользователя.

По существу новизна результатов определяется следующими отличительными признаками:

1) *подход к проектированию многомерной модели данных*: за основу берется ER-модель базы данных OLTP и в ней дополнительно вычленяются в отдельные сущности атрибуты, существенные в плане анализа; после чего связи типа «многие-ко-многим» рассматриваются как потенциальные гиперкубы (атрибуты связей – как меры гиперкуба, а связываемые сущности – как измерения гиперкуба).

2) *свойства многомерной модели данных, возникающие при построении ее на основе ER-модели OLTP*: при выполнении операции сведения показателей по измерениям в многомерной модели данных зачастую обнаруживаются функциональные зависимости между измерениями; операция сведения по функционально-зависимым измерениям является особой формой фильтрации, при которой игнорируются те идентифицирующие факт-координаты, для которых привязанные неидентифицирующие факт-координаты не попадают в множество укрупнения.

3) *методологические и лингвистические средства реализации функций аналитики в веб-приложениях на основе СОБД*: из XML-документов серверного хранилища СОБД в определенных состояниях формируется «на лету» контент-таблица, которая отправляется в OLAP-клиент пользователя; построчное формирование контент-таблицы основано на вложенных циклах обработки множеств однотипных XML-документов из хранилища СОБД, причем строки контент-таблицы формируются путем XSL-трансформации XML-документов хранилища, динамически загружаемых в DOM-объекты; введены новые элементы динамической модели СОБД, позволяющие в ходе ее интерпретации формировать строки контент-таблицы путем сканирования хранилища XML-документов: элемент цикла, обеспечивающий сканирование однотипных папок, и элемент селекции документа, обеспечивающий выбор экземпляров однотипных документов для загрузки в DOM-объекты в ходе сканирования.

4) *программное обеспечение для реализации функций аналитики в веб-приложениях на основе СОБД*: предложенные элементы динамической модели реализованы в виде программных модулей в составе интерпретатора динамической модели СОБД, ориентированных на контент-таблицы в формате CSV и OLAP-клиент FlexMonster Pivot Table & Charts Component.

Теоретическая и практическая значимость результатов

Значение результатов для теории (методологии) веб-приложений состоит в углублении представлений о том, каким образом на основе СОБД может быть обеспечена OLAP-функциональность, и как она может быть реализована путем формирования «на лету» гиперкубов из XML-документов.

Значение результатов для практики построения веб-приложений заключается в том, что они дают инструментарий для обеспечения OLAP-функциональности СОБД, который позволяет значительно снизить трудоемкость программирования.

Методология и методы исследования. В работе использовались технологии и методы построения веб-приложений, объектно-ориентированного программирования, OLAP, системного анализа, теории множеств, ситуационного управления, иерархических моделей и концептуального моделирования баз данных.

Положения, выносимые на защиту:

1. Подход к проектированию многомерной модели данных для решения «на лету» задач Web OLAP на основе базы данных OLTP, *отличающийся* тем, что за основу берется ER-модель базы данных OLTP, в ней вычленяются в отдельные сущности атрибуты, представляющие интерес в плане анализа, затем связи типа «многие-ко-многим» преобразуются в гиперкубы (атрибуты связей – в меры гиперкуба, а связываемые сущности – в измерения гиперкуба), и *позволяющий*, проанализировав структуру имеющейся базы данных OLTP, построить на ее основе модель OLAP, потенциально возможную для реализации «на лету», после чего выбрать для построения те гиперкубы, которые актуальны в плане задач анализа.

2. Свойства многомерной модели, возникающие при построении ее на основе ER-модели базы данных OLTP, *отличающиеся* тем, что при выполнении операции сведения показателей по измерениям в многомерной модели данных зачастую обнаруживаются функциональные зависимости между измерениями, что приводит к появлению эффекта фильтрации, при котором игнорируются те идентифицирующие факт-координаты, для которых привязанные неидентифицирующие факт-координаты не попадают в множество укрупнения, и *позволяющие* избежать аномалий при проектировании многомерных моделей OLAP, формируемых «на лету».

3. Методологические и лингвистические средства реализации функций аналитики в веб-приложениях на основе СОБД, *отличающиеся* тем, что формирование контент-таблицы, отправляемой в OLAP-клиент пользователя, выполняется на основе вложенных циклов обработки множеств однотипных XML-документов из хранилища СОБД, для реализации которых введены новые элементы динамической модели: элемент цикла, обеспечивающий сканирование однотипных папок, и элемент селекции документа, обеспечивающий выбор экземпляров однотипных документов для загрузки в DOM-объекты в ходе сканирования, и *позволяющие* избежать ошибок при написании программного кода, поскольку OLAP-отчеты специфицируются в динамической модели СОБД.

4. Программное обеспечение для реализации функций аналитики в веб-приложениях на основе СОБД, *отличающееся* тем, что предложенные элементы динамической модели реализованы в виде модулей в составе интерпретатора динамической модели СОБД, ориентированных на контент-таблицы в формате CSV и OLAP-клиент FlexMonster Pivot Table & Charts Component, и *позволяющее* снизить трудоемкость программирования при реализации OLAP-функциональности в веб-приложениях, основанных на СОБД. Для рассмотренных в работе примеров было установлено сокращение объема программного кода до 7 раз.

Степень достоверности и апробация результатов. Достоверность результатов подтверждена путем разработки программного обеспечения, основанного на предложенном подходе, и практическом применении его в исследовательском прототипе реального веб-приложения на основе СОБД.

Разработанное программное обеспечение внедрено в научно-производственной фирме «РД-технология» и в ФГБОУ ВПО «УГАТУ».

Основные результаты работы были представлены на 9 конференциях российского и международного формата обсуждения.

Результаты получены в рамках плановых исследований в области СОБД, проводимых на кафедре автоматизированных систем управления УГАТУ при поддержке РФФИ (гранты №№ 10-07-00167-а и 13-07-00011).

Публикации. По теме диссертации опубликовано 13 работ общим объемом 7 печатных листов, в том числе 2 статьи в рецензируемом научном журнале из перечня ВАК; 4 публикации выполнены в единоавторстве. Разработанное программное обеспечение защищено свидетельством о государственной регистрации программы для ЭВМ.

Структура и объём диссертации. Диссертационная работа состоит из введения, четырех глав, заключения, выводов по диссертационной работе, изложенных на 126 страницах машинописного текста, списка литературы из 119 наименований и 2 приложений на 9 страницах, содержит 34 рисунка и 5 таблиц. Всего в работе 149 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы исследования, формулируются цель работы и решаемые задачи, отмечается научная новизна и практическая значимость выносимых на защиту результатов.

В первой главе выполняется анализ существующих подходов к проектированию веб-приложений, оснащенных функциями OLAP, и обзор известных OLAP-клиентов. Проводится сравнительная характеристика технологий OLAP и OLTP, в процессе которой определяются достоинства и недостатки каждой из них. Обсуждается идея применения возможностей OLAP в системах, ориентированных на OLTP. Рассматриваются СОБД, и обосновывается для них потребность в функциях аналитики применительно к выбранной предметной области. Формулируется цель исследования, и ставятся задачи, решаемые для достижения поставленной цели.

Во второй главе рассматриваются особенности проектирования многомерной концептуальной модели данных на основе ER-модели OLTP-ориентированной базы данных. Предлагается использовать полученную многомерную модель данных для построения веб-приложения, основанного на СОБД и выполняющего OLAP-функции.

Архитектура веб-приложения OLAP на основе СОБД. Веб-приложение выполняется на сервере (Web server), для доступа к которому через Интернет на клиентской машине используется веб-браузер (Web browser), выполняющий роль «тонкого клиента» (рис. 1). В ответ на клиентский запрос веб-

сервер генерирует и отправляет в браузер клиентскую страницу (Client Page), сформированную на основе данных из СОБД (SODB) в соответствии с текущим состоянием. СОБД включает в себя динамическую модель (HSM), память ассоциированных данных (ADM), память текущего состояния (CSM). Память ассоциированных данных может быть представлена набором XML-документов, XML-архивами или реляционной базой данных.

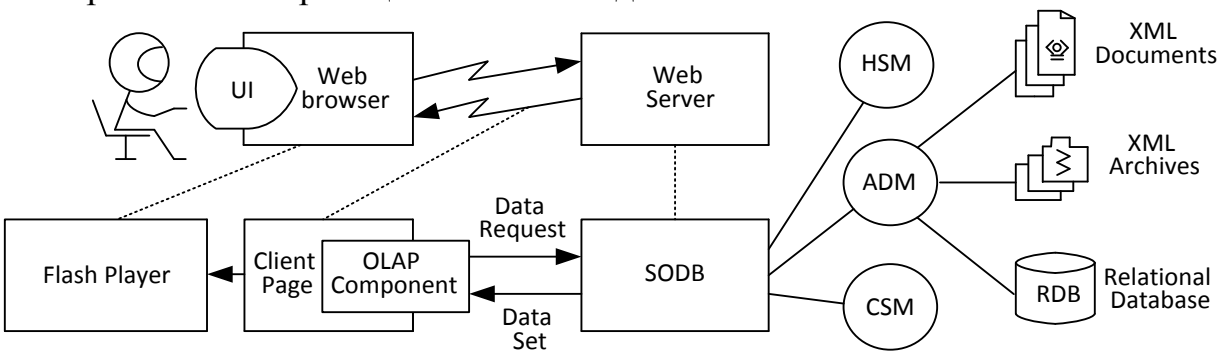

Рисунок 1 – Архитектура веб-приложения с OLAP-возможностями, основанная на СОБД

Графическая нотация для ER- и MD-моделей. ER-модель (модель сущность-связь) предлагается задавать с помощью нотации, изображенной на рис. 2.

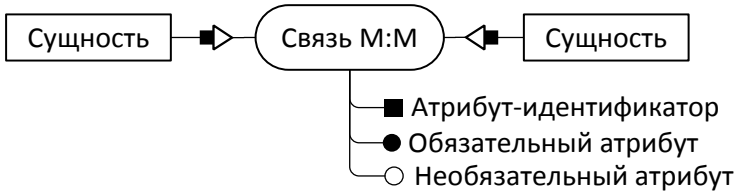

Рисунок 2 – Элементы ER-модели

Сущность изображается на диаграмме в виде прямоугольника. Связь типа 1:M («один-ко-многим») изображается с помощью стрелки в виде сцепленных треугольника и квадрата, направленной от сущности-родителя к сущности-ребенку. Идентифицирующая связь обозначается темным квадратом, а неидентифицирующая – светлым; обязательная связь обозначается темным треугольником, а необязательная – светлым.

Связь типа M:M («многие-ко-многим») изображается с помощью овала с присоединенными к нему 1:M-связями, идущими от связываемых сущностей. Атрибуты сущностей и M:M-связей отображаются с помощью выносных линий. Атрибуты-идентификаторы помечаются темным квадратом (символ ключа), а остальные атрибуты – кружком (символ однозначности). Темный кружок – обязательный атрибут, светлый – необязательный.

Средства графического задания многомерной модели данных (MD-модели) несколько беднее, чем в случае ER-модели (рис. 3).

Гиперкуб на MD-диаграмме изображается символом куба, с которым соединены символы измерений. Атрибуты измерений показываются на выносных линиях, исходящих из символов измерений, а меры (показатели) – на выносных линиях, исходящих из символа куба.

Если на измерениях куба заданы иерархии, то самый верхний уровень иерархии изображается в виде овала, промежуточные уровни – в виде трапеций, а нижний уровень – символом измерения. Иерархии одного измерения могут иметь общие нижние уровни.

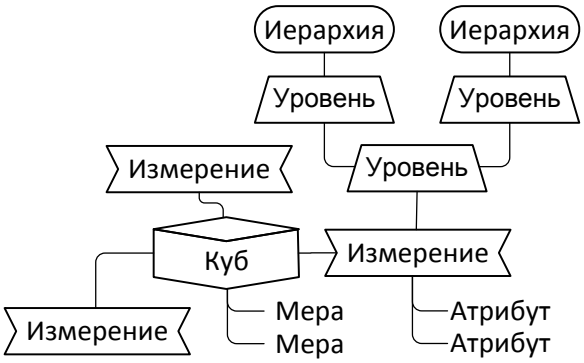

Рисунок 3 – Элементы MD-модели

Переход от ER- к MD-модели. Идея преобразования ER-модели в MD-модель заключается в том, что каждую связь типа «многие-ко-многим» следует рассматривать как потенциальную факт-сущность MD-модели.

Особенности проектирования MD-модели:

1) **Многокубовая модель.** В случае, если в исходной ER-модели имеется несколько M:M-связей, то в результате замены их гиперкубами получается многокубовая модель, в которой отдельные кубы взаимосвязаны.

Предположим, что исходная ER-модель содержит две M:M-связи: A и B, причем A является родителем для B. При переходе к MD модели (рис. 4) связи превращаются в кубы A и B, связываемые сущности – в измерения соответствующих кубов, а связь типа «родитель-ребенок» между связями A и B – в указатель наследования (стрелка A → B).

В результате дочерний куб B наследует измерения родительского куба A, а также меры родительского куба – тоже в форме измерений. То есть, доступ к показателям куба B возможен и по измерениям, и мерам родительского куба A.

Кроме того, куб A, имеющий в качестве наследника куб B, может использовать его сводные показатели в качестве своих мер. Здесь куб A помимо своей «родной» меры MA1 использует меру MB1 в сводном виде – например, в виде суммы ∑MB1.

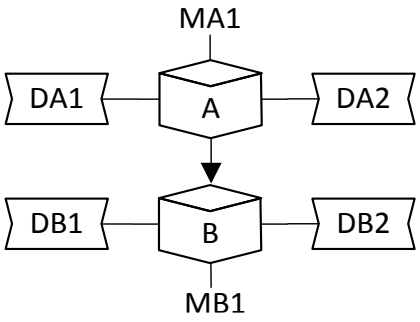

Рисунок 4 – Взаимосвязь кубов в MD-модели

2) **Иерархии в измерениях.** Связи 1:M задают иерархии в измерениях кубов (рис. 5). Здесь сущности-измерения DA1 и DA2 имеют родителей – сущности C1 и C2, которые, в свою очередь, имеют родителей D1 и D2. При переходе к MD-модели родительские сущности становятся частью измерений DA1 и DA2 в виде уровней D1 и C1, D2 и C2 иерархий H1 и H2 соответственно.

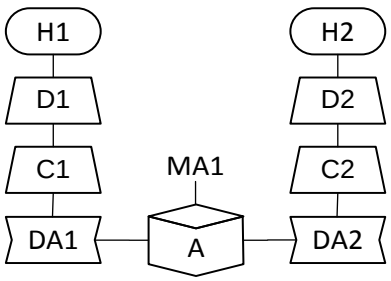

Рисунок 5 – Иерархии в измерениях

3) **«Аппендиксы» в ER-модели.** В ER-модели возможны фрагменты, в которых сущности связаны в цепочки «родитель-ребенок», но при этом не

участвуют в M:M-связях. Такие фрагменты будем называть «аппендиксами». Аппендиксы обычно можно игнорировать в ER-модели, поскольку из них нет доступа к M:M-связям или к сущностям измерений в направлении от родителя к ребенку и поэтому они не участвуют в формировании кубов, измерений кубов или иерархий измерений.

4) **Вырожденные измерения.** Встречаются случаи, когда какой-либо атрибут M:M-связи представляет интерес в плане использования его в качестве измерения гиперкуба. В таком случае целесообразно выделить этот атрибут в самостоятельную сущность, позволяя проводить по нему многомерный анализ данных. Такие измерения называются вырожденными.

5) **Обобщение кубов.** Исходная ER-модель может содержать несколько M:M-связей, которые связывают между собой одни и те же сущности, отражая их различные аспекты взаимодействия. При переходе к MD-модели более удобно иметь укрупненные кубы, позволяющие выполнять анализ по различным аспектам. Целесообразно обобщить кубы, базирующиеся на общих измерениях, сделав из них один укрупненный куб и добавив при необходимости новое измерение, позволяющее селектировать срезы исходных кубов.

6) **Хронологические данные.** Хранилище данных, как правило, содержит измерение времени, позволяющее анализировать данные по временным интервалам. В базах данных, ориентированных на OLTP, возникает проблема анализа по измерению времени, поскольку в ней хранятся версии данных, соответствующие текущему состоянию. Для обеспечения возможности анализа по измерению времени в OLTP базе данных необходимо предусмотреть:

- накопление необходимых для анализа версий хронологических данных вместе с информацией о времени;
- возможность селекции нужных версий хронологических данных при построении кубов «на лету».

В третьей главе рассматриваются особенности агрегации данных в многомерном кубе при наличии измерений, между которыми имеются функциональные зависимости.

В гиперкубе для адресации показателей предусмотрены измерения, содержащие множества координат, которые могут быть организованы в виде иерархий. Во многих работах на измерения накладываются ограничения взаимной независимости по аналогии с евклидовой системой координат. Однако многомерное моделирование многих предметных областей требует использования измерений, находящиеся в функциональной зависимости друг от друга.

Ключевую роль в OLAP-системах играют операции агрегирования, или сведения, т. е. процедуры автоматического формирования меньшего количества результирующих значений (агрегатов) из большего количества исходных значений. Стандартные агрегации, основанные на суммировании, подсчете количества значений, вычислении минимального и максимального значений показателей, достаточно просты для теоретического анализа благодаря свойству аддитивности сводного показателя (меры). Однако во многих случаях требуется нестандартная, неаддитивная агрегация. Современные OLAP-системы предусмат-

ривают возможность как аддитивного, так и неаддитивного агрегирования, что обуславливает необходимость исследований в общем виде.

В общем случае куб данных имеет несколько измерений, задающих систему координат пространства данных. Каждое измерение представляет собой конечное множество координат. Совокупность таких координат по всем измерениям куба образует кортеж, однозначно идентифицирующий ячейку. Ячейка – это часть данных, получаемая путем определения одного элемента в каждом измерении многомерного массива. В каждой ячейке размещены в общем случае несколько мер – показателей, хранящихся в кубе. Как правило, в роли показателей выступают некоторые числовые значения.

Укрупненные координаты посредством группирования задают иерархии. В каждом измерении гиперкуба задается множество факт-координат, т. е. однородных элементов измерения с наибольшей степенью детализации. Укрупненные координаты могут использоваться в качестве координат измерений для адресации ячеек гиперкуба точно так же, как обычные факт-координаты. Им в гиперкубе соответствуют так называемые агрегированные или сводные ячейки.

На рис. 6 изображено некоторое измерение X гиперкуба. Каждой укрупненной координате x* некоторого уровня иерархии измерения X соответствует множество $\Omega(x^*) = \{x_1, x_2, x_3, \ldots\}$ дочерних координат нижестоящего уровня этой иерархии. Пусть $Z = \{Z_1, Z_2, \ldots\}$ – множество остальных измерений гиперкуба, а z – кортеж текущих координат этих измерений, то есть $z = \langle z_1, z_2, \ldots\rangle$, $z_1 \in Z_1$, $z_2 \in Z_2$, … . Для заданного кортежа z координатам $x_1, x_2, x_3, \ldots$ соответствуют ячейки с координатами $\langle x_1, z\rangle$, $\langle x_2, z\rangle$, $\langle x_3, z\rangle$, …, а укрупненной координате x* – сводная ячейка с координатами $\langle x^*, z\rangle$. Указанные ячейки содержат значения некоторого показателя F:

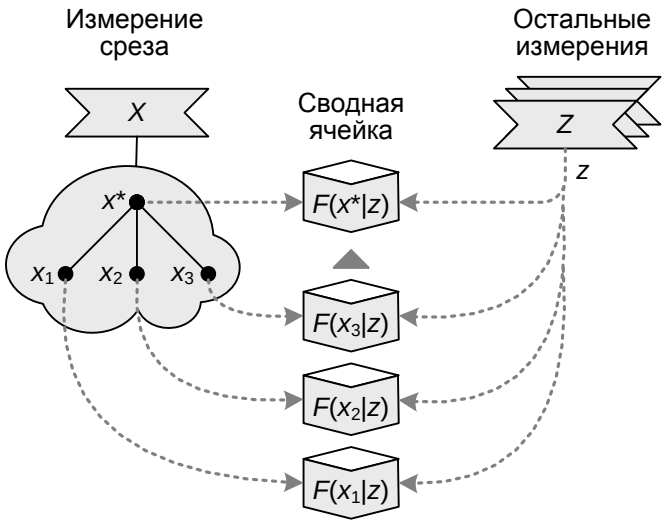

Рисунок 6 – Схема формирования сводных ячеек

$\langle x_1, z\rangle \rightarrow F(x_1|z), \langle x_2, z\rangle \rightarrow F(x_2|z), \langle x_3, z\rangle \rightarrow F(x_3|z), \ldots, \langle x^*, z\rangle \rightarrow F(x^*|z)$.

Тогда должна иметь место зависимость: $F(x^*|z) = \text{Aggr}\ \{F(x_1|z), F(x_2|z), \ldots\} = \text{Aggr}\ F(y|z)$, $y \in \Omega(x^*)$, где Aggr – некоторая функция агрегации, задающая правила вычисления агрегированных показателей сводных ячеек на основе исходных показателей укрупняемых ячеек.

Глобальное сведение показателей – это вычисление сводного показателя (меры) для некоторой сводной ячейки путем непосредственного агрегирования показателей множества факт-ячеек, соответствующих этой сводной ячейке.

Локальное сведение показателей – это вычисление меры некоторой сводной ячейки путем агрегирования показателей ячеек нижестоящих уровней иерархии.

Сведение по измерениям – процедура, предполагающая определенную последовательность агрегирования локально дочерних показателей. А именно, сначала показатели агрегируются по координатам одного измерения, затем полученный агрегат, не зависящий от этого измерения, агрегируется по координатам другого измерения, далее – по координатам третьего измерения и т. д. до последнего измерения, при агрегации по координатам которого получается искомый сводный показатель.

Формально эту процедуру можно представить следующим образом. Пусть необходимо выполнить локальное сведение показателей для ячейки с координатами $\vec{t}$. Для этого необходимо выполнить агрегацию локально дочерних ячеек $\mathrm{Aggr}F(\vec{t}^{\,*})$ при $\vec{t}^{\,*} \in \Omega^*(\vec{t})$. Пусть гиперкуб имеет N измерений и множество измерений упорядочено. Запишем

$$\vec{t}^{\,*} = (t_1^{\,*}, t_2^{\,*}, ..., t_N^{\,*}); \Omega^*(\vec{t}) = \Omega_1^{\,*}(\vec{t}) \times \Omega_2^{\,*}(\vec{t}) \times ... \times \Omega_N^{\,*}(\vec{t}), \quad (1)$$

где $\Omega_i^{\,*}(\vec{t})$ – множество локально дочерних координат ячейки $\vec{t}$ по измерению i, i = 1, 2, …, N, а «×» – символ операции декартова произведения множеств. Тогда сведение по измерению означает, что

$$\mathrm{Aggr}\, F(\vec{t}^{\,*}) = \mathrm{Aggr}^{(1)}\, \mathrm{Aggr}^{(2)} \ldots \mathrm{Aggr}^{(N)} F(t_1^{\,*}, t_2^{\,*}, ..., t_N^{\,*}). \quad (2)$$

Для аддитивных функций агрегирования, во-первых, сведение показателей всегда можно представить в виде сведения по измерениям, во-вторых, порядок следования измерений не имеет значения.

В теории многомерных баз данных оперируют понятиями функциональной зависимости и детерминанта. Они используется для определения связей между измерениями и показателями многомерной модели данных.

Измерение называется **идентифицирующим** для некоторой меры, если идентификатор факт-координат этого измерения входит в состав детерминанта функциональной зависимости факт-показателя этой меры, и **неидентифицирующим** в противном случае. Совокупность идентификаторов идентифицирующих измерений является идентификатором для показателя меры; показатель не зависит функционально от своих неидентифицирующих измерений. Вместе с тем, каждой факт-ячейке соответствуют единственные факт-координаты неидентифицирующих измерений, поэтому факт-координаты неидентифицирующих измерений функционально зависят от совокупности факт-координат идентифицирующих измерений как от детерминанта.

Формально функциональные зависимости между измерениями можно выразить следующим образом. Пусть $\vec{t} = \langle \vec{p}, \vec{q} \rangle$ – кортеж факт-координат, где $\vec{p}$ – факт-координаты идентифицирующих измерений, а $\vec{q}$ – неидентифицирующих. Факт-показатели функционально зависят только от идентифицирующих измерений, поэтому $\vec{p}$ задает единственную факт-ячейку, и, следовательно, единственную соответствующую $\vec{p}$ совокупность $\vec{q}$, т. е. существует

функция G такая, что $\vec{q}$ = G($\vec{p}$). Тогда для произвольных факт-координат $\vec{p}$ и $\vec{q}$ имеет место следующее соотношение:

$$f(\vec{p},\vec{q}) = \begin{cases} f(\vec{p}, G(\vec{p})) = \phi(\vec{p}), \text{если} G(\vec{p}) = \vec{q}, \\ \text{Null в противном случае} \end{cases} \quad (3)$$

Наличие функциональных зависимостей между измерениями приводит к тому, что сведение показателей выполняется не во всем пространстве факт-ячеек, а лишь в подпространстве связанных факт-ячеек.

Пусть для подсчета сводного показателя выполняется агрегация $\text{Aggr}_{\vec{q}\in\Omega_Q} f(\vec{p},\vec{q})$ по некоторому подмножеству Ω_Q факт-координат $\vec{q}$ неидентифицирующих измерений для фиксированных факт-координат $\vec{p}$ идентифицирующих измерений. Для фиксированных координат $\vec{p}$ имеются соответствующие им координаты неидентифицирующих измерений $G(\vec{p})$, которым, в свою очередь, соответствует не более одной факт-ячейки. В соответствии с формулой (2) и с учетом того, что элементы, имеющие Null-значения, игнорируются при агрегировании, получаем

$$\underset{\vec{q}\in\Omega_Q}{\text{Aggr}}\, f(\vec{p},\vec{q}) = \begin{cases} f(\vec{p}, G(\vec{p})) = \phi(\vec{p}), \text{если} G(\vec{p}) \in \Omega_Q, \\ \text{Null в противном случае.} \end{cases} \quad (4)$$

В четвертой главе рассматривается организация функций аналитики в веб-приложениях, функционирующих на основе серверных СОБД и OLAP-клиентов.

Выбранный в данной работе OLAP-клиент поддерживает загрузку анализируемых данных из различных видов хранилищ в формате CSV. Исходные данные для анализа, загружаемые в OLAP-клиент для формирования сводной таблицы, предлагается называть контент-таблицей. Помимо контент-таблицы, содержащей собственно анализируемые данные, в OLAP-клиент могут передаваться метаданные, задающие структуру сводной таблицы, отображаемой пользователю (рис. 7).

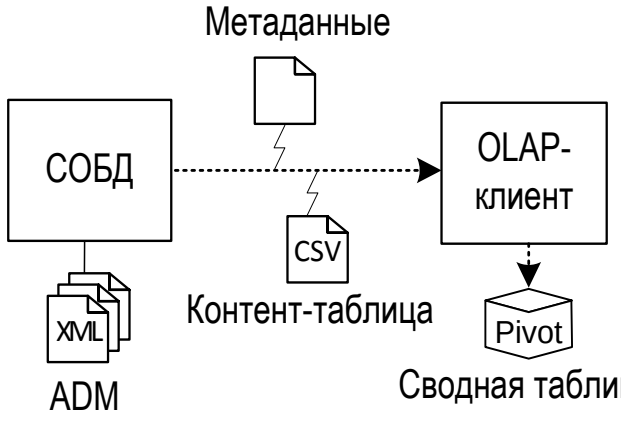

Рисунок 7 – Взаимодействие СОБД и OLAP-клиента с помощью контент-таблицы

CSV-формат предписывает плоскую структуру контент-таблицы: первая строка задает имена столбцов, последующие строки задают построчные кортежи значений. Следовательно, исходный гиперкуб данных, предоставляемый пользователю для анализа в виде сводной таблицы, должен быть преобразован в плоскую форму. Чтобы задать столбцы контент-таблицы

нужно выписать через запятую заголовки интересующих мер и измерений, помечая их определенным образом.

В общем случае хранилище XML-документов (ADM) имеет расщепленную структуру в том смысле, что однотипные данные хранятся в нескольких однотипных документах, размещенных в разных папках. В связи с этим для извлечения данных из множества документов необходимо организовать циклы перебора требуемых экземпляров в множествах однотипных папок и документов. Для придания динамики статическая модель ADM дополняется элементами-действиями трех видов: цикла (элемент for), сохранения промежуточных значений (конструкция вида x → $y, где x – элемент данных, извлекаемый из документа, $y – временная переменная, в которой сохраняется значение извлеченного элемента данных) и вывода результата (элемент echo).

На рис. 8 приведена концептуальная схема формирования строк контент-таблицы с использованием элементов-действий. Обработка начинается с корневой папки A. Первым обрабатывается документ X, из которого извлекается элемент данных «a» и сохраняется во временной переменной $a. Затем обрабатывается элемент цикла «I», который последовательно запускает на обработку папки B[i].

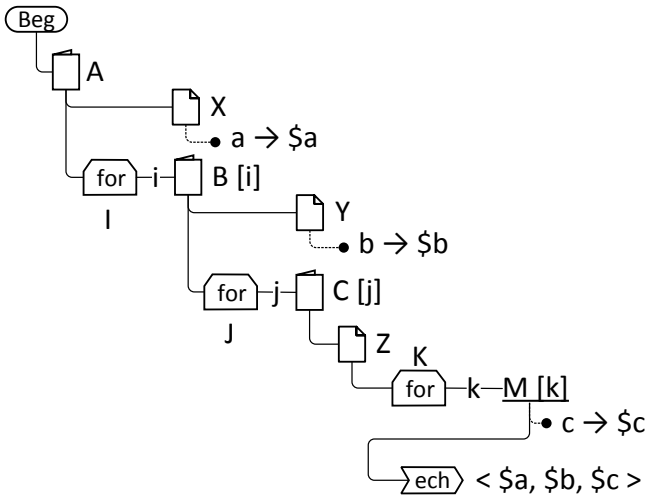

Рисунок 8 – Схема формирования строк контент-таблицы

При обработке папки B[i] сначала обрабатывается вложенный документ Y, из которого извлекается элемент данных b и сохраняется во временной переменной $b. Далее обрабатывается элемент цикла J, в результате чего последовательно запускается обработка папок C[j].

При обработке папки C[j] обрабатывается вложенный документ Z, для которого, в свою очередь, обрабатывается элемент цикла K. Результатом этого является последовательный запуск обработки экземпляров многозначного XML-элемента M[k]. При обработке экземпляра M[k] сначала извлекается XML-атрибут «c» и сохраняется во временной переменной $c. Далее обрабатывается элемент вывода, в результате чего в выходной поток выводится сформированная строка контент-таблицы <$a, $b, $c>.

Таким образом, значения $a являются одинаковыми для каждой строки контент-таблицы, а значения $b – одинаковыми для групп строк, сформированных в пределах каждой папки B. На один документ Z приходится несколько сформированных строк по числу экземпляров XML-элемента M.

Формирование контент-таблицы средствами динамической модели СОБД обеспечивается с помощью так называемых динамических DOM-объектов. Для порождения DOM-объектов в соответствующих состояниях sta дерева HSM в

качестве дочерних размещаются dom-элементы **dom** , которые, в свою очередь, в качестве дочерних могут содержать элементы – источники **src** и приемники **rcv** данных. Элементы-источники специфицируют загрузку XML-документов из ADM в DOM-объект, а элементы-приемники – выгрузку XML-содержимого DOM-объекта (возможно, подвергнутого XSL-трансформации) в выходной поток, в ADM или в другой DOM-объект.

Для упрощения циклической обработки множеств однотипных документов, размещенных в различных папках, в HSM введено два дополнения: элемент цикла **for** и элемент селекции документа **doc**.

Элемент цикла располагается в HSM как дочерний элемент состояния (рис. 9). Атрибут scanSubDirs этого элемента задает родительскую папку, дочерние папки которой подвергаются циклической обработке. Спецификация пути в этом атрибуте может содержать ссылки на переменные охватывающих циклов и на элементы селекции. Атрибут scanVar задает переменную цикла, в которой хранится имя текущей папки (документа).

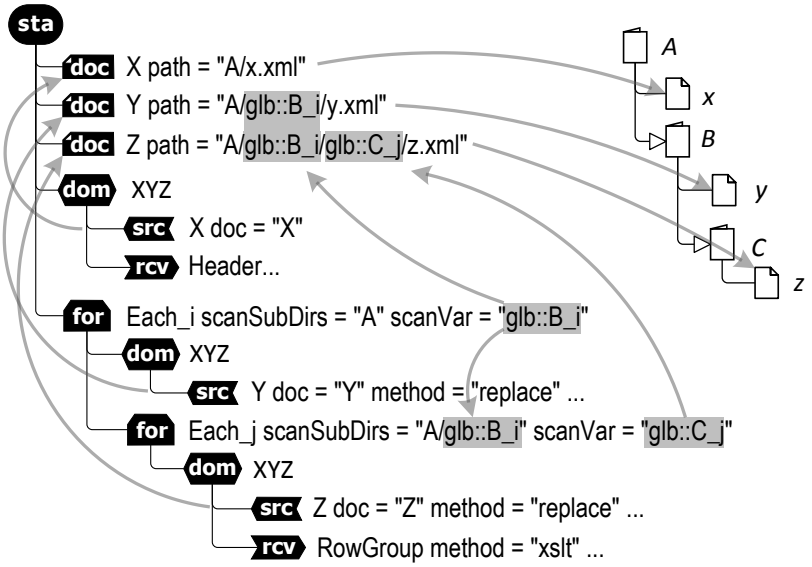

Рисунок 9 - Пример динамической модели HSM формирования контент-таблицы с использованием элементов цикла и селекции документов

Элемент селекции располагается в HSM как дочерний элемент состояния. Атрибут path этого элемента задает путь к папке или документу в ADM. Важной особенностью этого атрибута является то, что спецификация пути может содержать ссылки на переменные циклов и на другие элементы селекции. Это обеспечивает гибкий доступ к папкам и документам в ходе выполнения цикла.

Реализация предлагаемой концепции выполнялась в рамках информационной системы, обслуживающей деятельность диссертационных советов вуза. Информационная система представляет собой веб-приложение, в базе данных которого хранятся сведения о диссертационных советах вуза, о диссертантах и другие сведения, относящиеся к диссертационному процессу. Веб-приложение функционирует на основе СОБД, обеспечивающей хранение данных в виде XML-документов и взаимодействие с пользователями через браузеры.

В рамках исследовательского прототипа системы были реализованы 4 OLAP-отчета, отражающие активность диссертантов, участников диссертационного совета и организаций в диссертационном процессе.

Проведенный расчет оценки эффективности в плане сокращения объема программного кода показал, что применение модельно-ориентированного подхода к разработке веб-приложений и, в частности, введение в динамическую модель дополнительных элементов (DOM-элементы и элементы цикла), позволяет сократить объем кода до 7 раз для рассмотренных примеров.

В заключении изложены основные результаты работы.

В приложениях приведены: модуль обработки элемента селекции документа doc динамической модели СОБД и листинг XSL-преобразований для формирования отчета об активности организаций диссертационных советов вуза.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ РАБОТЫ

В работе представлены новые научно обоснованные информационно-технологические решения и разработки в области управления XML-данными в ситуационно-ориентированных базах данных, имеющие значение для развития веб-технологий, ориентированных на OLAP. При этом получены результаты:

1. Подход к проектированию многомерной модели данных для решения «на лету» задач Web OLAP на основе базы данных OLTP, *отличающийся* тем, что за основу берется ER-модель базы данных OLTP и в ней дополнительно вычленяются в отдельные сущности атрибуты, значимые в плане анализа, после чего связи типа «многие-ко-многим» рассматриваются как потенциальные гиперкубы (атрибуты связей – как меры гиперкуба, а связываемые сущности – как измерения гиперкуба).

Это *позволяет* проанализировать структуру имеющейся базы данных OLTP и построить на ее основе модель OLAP, потенциально возможную для реализации «на лету», после чего выбрать для построения те гиперкубы, которые актуальны в плане задач анализа.

2. Свойства многомерной модели OLAP, возникающие при построении ее на основе ER-модели OLTP, *отличающиеся* тем, что при выполнении операции сведения показателей по измерениям в многомерной модели данных зачастую обнаруживаются функциональные зависимости между измерениями, что приводит к появлению эффекта фильтрации, при котором игнорируются те идентифицирующие факт-координаты, для которых привязанные неидентифицирующие факт-координаты не попадают в множество укрупнения.

Это *позволяет* избежать аномалий при проектировании многомерных моделей OLAP, формируемых «на лету».

3. Методологические и лингвистические средства реализации функций аналитики в веб-приложениях на основе СОБД, *отличающиеся* тем, что формирование контент-таблицы, отправляемой в OLAP-клиент пользователя, выполняется на основе вложенных циклов обработки множеств однотипных XML-документов из хранилища СОБД, для реализации которых введены новые элементы динамической модели: элемент цикла, обеспечивающий сканирование однотипных папок, и элемент селекции документа, обеспечивающий выбор эк-

земпляров однотипных документов для загрузки в DOM-объекты в ходе сканирования.

Это *позволяет* повысить наглядность и избежать ошибок при написании программного кода, поскольку использование моделей высокого уровня абстракции дает возможность специфицировать OLAP-отчеты в динамической модели СОБД, автоматически формируемые «на лету» в определенных ситуациях.

4. Программное обеспечение для реализации функций аналитики в веб-приложениях на основе СОБД, *отличающееся* тем, что предложенные элементы цикла и селекции документа реализованы в виде модулей в составе интерпретатора динамической модели СОБД, ориентированных на контент-таблицы в формате CSV и OLAP-клиент FlexMonster Pivot Table & Charts Component.

Это *позволяет* снизить трудоемкость программирования при реализации OLAP-функциональности в веб-приложениях, основанных на СОБД. Для рассмотренных в работе примеров было установлено сокращение объема программного кода до 7 раз.

Перспективы дальнейшей разработки темы. В рамках дальнейших исследований планируется разработка концепций, методов, алгоритмов и программного обеспечения, позволяющих загружать в OLAP-клиент пользователя исходные данные в различных форматах.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

В рецензируемых журналах из списка ВАК

1. Агрегация показателей в OLAP-кубе при сведении по зависимым измерениям / В. В. Миронов, Е. С. Макарова // Вестник УГАТУ. 2012. № 3. С. 180–187.

2. Проектирование концептуальной модели данных для задач Web-OLAP на основе ситуационно-ориентированной базы данных / В. В. Миронов, Е. С. Макарова // Вестник УГАТУ. 2012. № 6. С. 177–189.

Зарегистрированные программы для ЭВМ

3. Свид. о гос. рег. программы для ЭВМ № 2013614961. Интерпретатор динамических моделей ситуационно-ориентированных баз данных с функциями веб-анализа с использованием встроенного OLAP-клиента / В. В. Миронов, Е. С. Макарова. Зарег. 23.05.2013. М.: Роспатент, 2013.

В других изданиях

4. Информационная система анализа результатов экзаменационной сессии на основе OLAP-технологий / Е. С. Макарова // Актуальные проблемы науки и техники: 4-я Всеросс. зимн. шк.-сем. аспирантов и молодых ученых. Уфа: Изд-во «Диалог», 2009. Т. 1. С. 349–353.

5. Сведение по зависимым измерениям / Е. С. Макарова // Концептуальные модели баз данных. Многомерные модели / В. В. Миронов, Н. И. Юсупова. Уфа: УГАТУ, 2010. С. 52-57.

6. Проектирование многомерной базы данных для анализа учебного процесса на основе ER-модели / Е. С. Макарова, Т. Ф. Файзрахманов // Актуальные

проблемы науки и техники: 5-я Всеросс. зимн. шк.-сем. аспирантов и молодых ученых. Уфа: УГАТУ, 2011. Т. 2. С. 245–249.

7. Реализация рекурсивных измерений в реляционных моделях для многомерного анализа данных / Т. Ф. Файзрахманов, Е. С. Макарова // Актуальные проблемы науки и техники: 5-я Всеросс. зимн. шк.-сем. аспирантов и молодых ученых. Уфа: УГАТУ, 2011. Т. 3. С. 92–96.

8. Преобразование OLTP-ориентированной модели в OLAP-модель хранилища данных / Е. С. Макарова // Актуальные проблемы науки и техники: 6-я Всеросс. зимн. шк.-сем. аспирантов и молодых ученых. Уфа: УГАТУ, 2011. Т. 1. С. 188–191.

9. Сведение показателей OLAP-куба при наличии функционально зависимых измерений / Е. С. Макарова // Актуальные проблемы науки и техники: 7-я Всеросс. зимн. шк.-сем. аспирантов и молодых ученых. Уфа: УГАТУ, 2012. Т. 1. С. 157–160.

10. Построение OLAP-кубов на основе OLTP-ориентированной модели для анализа данных образовательного процесса / В. В. Миронов, Е. С. Макарова // Интеллектуальные технологии обработки информации и управления: сб. науч. тр. Междунар. молодеж. конф. Уфа: Аркаим, 2012. Т. 1. С. 167–170.

11. Вычисление мер OLAP-куба при наличии зависимых измерений / В. В. Миронов, Е. С. Макарова // Труды 14-й междунар. конф. по выч. наукам и инфор. техн. (CSIT’2012). Уфа-Гамбург-Норвежские Фьорды, 2012. Т. 1. С. 93-99.

12. Процедура разработки многомерной модели данных для задач OLAP и интеллектуального анализа / В. В. Миронов, Е. С. Макарова // Искусственный интеллект и его приложения : сб. материалов 3 Межвузовск. научн.-иссл. сем. с междунар. участием. Магнитогорск: МаГУ, 2012. С. 85–95.

13. Построение OLAP-модели хранилища данных на основе ER-модели / В. В. Миронов, Е. С. Макарова // Повышение эффективности использования информационных технологий в государственном и муниципальном управлении: сб. тр. Всеросс. науч.-практ. конф. с междунар. участием. Уфа: БАГСУ, 2012. С. 95-98.

14. Принципы взаимодействия компонентов веб-приложения для решения задач Web OLAP / Е. С. Макарова // Актуальные проблемы науки и техники: 8-я Всеросс. зимн. шк.-сем. аспирантов и молодых ученых. Уфа: УГАТУ, 2013. Т. 1. С. 220–224.

Диссертант  Е. С. Макарова

МАКАРОВА Екатерина Сергеевна

ФУНКЦИИ АНАЛИТИКИ В ВЕБ-ПРИЛОЖЕНИЯХ НА ОСНОВЕ СИТУАЦИОННО-ОРИЕНТИРОВАННЫХ БАЗ ДАННЫХ

Специальность 05.13.11 – Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Подписано к печати 29.10.2013. Формат 60×84 1/16.
Бумага офисная. Печать плоская. Гарнитура Таймс.
Усл. печ. л. 1,0. Уч.–изд. л. 1,0.
Тираж 100 экз. Заказ №569.

ФБГОУ ВПО Уфимский государственный авиационный технический университет
Центр оперативной полиграфии
450000, Уфа-центр, ул. К.Маркса, 12