На правах рукописи

Бакулева Марина Алексеевна

МОДЕЛИ И АЛГОРИТМЫ АВТОМАТИЗАЦИИ ПРОЕКТИРОВАНИЯ СТРУКТУР ХРАНИЛИЩ ДАННЫХ ДЛЯ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ЧИСЛОВЫХ ПОКАЗАТЕЛЕЙ

Специальность
05.13.12 – Системы автоматизации проектирования
(технические системы)

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Рязань – 2007

Работа выполнена на кафедре систем автоматизированного проектирования вычислительных средств (САПР ВС) ГОУ ВПО «Рязанский государственный радиотехнический университет».

Научный руководитель: кандидат технических наук, доцент Телков Игорь Анатольевич.

Официальные оппоненты:

доктор технических наук, профессор
Цветков Игорь Анатольевич
кандидат технических наук, доцент
Пресняков Александр Николаевич

Ведущая организация: ГНИИ ИТТ «Информика», г. Москва.

Защита состоится «___»_____2007 г. в ___ часов на заседании диссертационного совета Д212.211.02 в ГОУ ВПО «Рязанский государственный радиотехнический университет» по адресу: 390005, г. Рязань, ул. Гагарина, д. 59/1.

С диссертацией можно ознакомиться в библиотеке Рязанского государственного радиотехнического университета.

Автореферат разослан «___» _____ 2007 г.

Отзывы на автореферат в двух экземплярах, заверенные печатью организации, просим направлять по адресу: 390005, г. Рязань, ул. Гагарина, д. 59/1, Рязанский государственный радиотехнический университет.

Ученый секретарь
диссертационного совета
кандидат технических наук, доцент И.А. Телков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Одним из основных компонентов современных информационных систем являются средства содержания и манипулирования массивами разнородных данных. По мере увеличения объемов и сложности хранимых данных и по мере их интеграции растет потребность и в средствах автоматизированного проектирования способных сократить сроки разработки и внедрения новых методов обработки столь многообразной и неструктурированной информации.

Одной из основных задач, решаемых в крупных информационных системах, является предоставление аналитической информации, необходимой для принятия решений. В процессе анализа данные должны поступать к потребителю в режиме реального времени. Если же данные собираются из разных источников, то, во-первых, отчет готовится недопустимо долго, а во-вторых, другие приложения, работающие с этими же реляционными СУБД во время выполнения отчета, будут работать значительно медленнее. Решением проблемы производительности является создание специализированной базы данных (БД) — хранилища данных (ХД), — предназначенной исключительно для обработки и анализа информации.

ХД являются основным источником данных оперативно-аналитических подсистем (On-line analytical processing system — OLAP-систем) современных систем поддержки принятия решений (СППР). Создание средств автоматизации проектирования структур ХД и предварительной обработки данных для анализа является важной задачей, поскольку от скорости и корректности этого процесса напрямую зависит способность информационной системы поддерживать (сопровождать) процессы принятия решений.

В работах Р. Кимбала, Б. Инмона, М.С. Куприянова, Б. Маркова рассматриваются методы построения ХД на основе реляционной модели. В работах А. Кэмбла, Д. Селко метод построения модели основан на теории множеств. Многомерная модель данных приводится в работе Э. Франкони, У. Сатлера, Л. Черняка. Некоторые из перечисленных методов применяются для автоматизации формирования многомерных представлений данных.

Однако на данный момент не решена конкретная задача формализации процесса трансформации реляционной БД, нормализованной и зачастую распределенной архитектуры, в удобную с точки зрения анализа форму ХД. Потребность оперативной обработки данных ХД (содержащих порядка 10^7 записей) приводит к необходимости создания специализированных аналитических решений. Для достижения приемлемого быстродействия (время ответа не более 5 секунд) необходимо разработать способ представления данных, адаптированный к условиям оперативного анализа.

Таким образом, в теории и практике существует важная научно-техническая задача создания моделей и алгоритмов автоматизации проектирования структур ХД, способствующих разработке эффективных ХД, а также алгоритмов OLAP на основе более адаптированного к анализу представления данных. С учетом вышесказанного разработка моделей ХД,

алгоритмов автоматизации построения ХД и алгоритмов повышения производительности аналитических подсистем с исходными данными реляционной структуры является актуальной задачей.

Объект исследования. Объектом исследований, проводимых в рамках диссертационной работы, является ХД и его взаимодействие с системами оперативного анализа числовых данных.

Цель работы. В процессе анализа темы диссертационной работы было выявлено, что задача проектирования хранилищ данных включает в себя требование обеспечения оперативности аналитических подсистем, работающих с хранилищами. Поэтому целями диссертационной работы являются:

- ускорение процесса проектирования ХД;
- повышение скорости аналитической обработки данных ХД.

Основные задачи. В работе поставлены следующие основные задачи:

- разработать математическую модель реляционного ХД;
- разработать математическую модель многомерного ХД;
- разработать математическую модель операций над многомерным ХД;
- разработать алгоритмы, автоматизирующие процесс построения ХД на основе предложенных моделей;
- разработать математическую модель структуры данных в ХД, отвечающую требованиям оперативной обработки запросов;
- разработать алгоритмы OLAP на основе предложенной структуры;
- разработать программную систему, реализующую алгоритмы автоматизации проектирования ХД и оперативный анализ данных ХД.

Методы исследования. Для решения поставленных задач использован аппарат тензорной алгебры, кратномасштабного анализа, вейвлет-преобразований и сигнатурного поиска.

Научная новизна. В диссертационной работе предлагается решение поставленных задач. Научная новизна состоит в следующем:

- предложены новые модели реляционного и многомерного ХД на основе тензорного аппарата;
- разработан метод аналитической обработки данных ХД, основанный на вейвлет – преобразовании;
- разработан алгоритм автоматизации процесса проектирования ХД на основе разработанных моделей. Основным преимуществом данных алгоритмов является снижение временных затрат на разработку ХД и автоматизация рутинных операций по описанию разнообразных связей исходной БД и проецированию соответствующих отношений в ХД;
- разработаны алгоритмы агрегации и анализа данных ХД. Отличие от известных заключается в использовании кратномасштабного анализа;

- разработаны алгоритмы ретроспективного анализа. Отличие от известных алгоритмов, прежде всего – в большей производительности, полученной за счет применения сигнатурного поиска.

Практическая ценность и внедрение результатов работы. Практическая ценность полученных результатов состоит в следующем:

- Создан комплекс программных средств на основе разработанного алгоритма автоматизации проектирования структур ХД, позволяющий сократить сроки проектирования ХД. На основе данного комплекса в научно – производственном предприятии «Эльф 4М» (г. Рязань) произведена реконструкция системы информационного обеспечения производственной деятельности предприятия.
- Созданный комплекс программных средств по обработке данных ХД является инструментом, обеспечивающим высокую скорость аналитической обработки большого числа хранимых данных. Внедрение разработанного комплекса значительно повысило эффективность работы научно – производственного предприятия «Эльф 4М», масштабы производства которого охватывают 12 стран (БД содержит ≈500000 записей). Созданный программный комплекс предоставляет руководителю актуальную информацию о темпах производства, географии сбыта, а также векторе развития предприятия.
- Созданный программный комплекс используется в ООО «Торгтек» (г. Рязань) для оперативной обработки отчетной и аналитической информации.
- Результаты, полученные в диссертационной работе, представляют часть НИР (НИР № 10-06Г (РНТП 3.2.3.7637) «Разработка нормативной базы, информационного обеспечения и регламентов открытой информационно-образовательной среды для дистанционной подготовки, переподготовки и повышения квалификации специалистов в области ИПИ (CALS) и CASE-технологий», НИР 11-06Г (РНТП 3.2.3.7652) «Интегрированная автоматизированная информационная система управления качеством образования ВУЗа»), проводимых РГРТУ.

Достоверность. Достоверность научных положений и полученных результатов диссертационной работы подтверждается математическими обоснованиями и доказательствами, а также результатами проведенных экспериментов.

Апробация результатов диссертации. Результаты, полученные в рамках работы над диссертацией, докладывались на 10-й Всероссийской научно-технической конференции студентов, молодых ученых и специалистов «Новые информационные технологии в научных исследованиях и образовании», Рязань, 2005 г.; 2-й Международной научно-практической конференции «Информационная деятельность: проблемы науки и практики», Киев, 2005 г.; 14-й Международной научно-технической конференции «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций», Рязань, 2005 г.;

11-й Всероссийской научно-технической конференции студентов, молодых ученых и специалистов «Новые информационные технологии в научных исследованиях и образовании», Рязань, 2006 г.; 12-й Всероссийской научно-технической конференции студентов, молодых ученых и специалистов «Новые информационные технологии в научных исследованиях и образовании», Рязань, 2007 г.

Публикации. Основные результаты диссертации опубликованы в 11 работах, из них 2 работы опубликованы в изданиях, рекомендованных ВАК.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, библиографического списка (82 источника), изложенных на 131 странице (содержит 10 таблиц, 53 рисунка), и 2 приложения. Общий объем диссертации 144 страницы.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, изложены цель и соответствующие ей задачи, приведена структура работы.

Первая глава посвящена рассмотрению вопросов проектирования ХД. Подчеркивается важная роль хранилища в СППР как основного источника данных для оперативного анализа (OLAP). Далее описываются отличия традиционных БД от ХД, главным из которых является упразднение требований нормализации. Преимущество по обеспечению оперативности заключается в том, что ХД, как правило, не имеют никакого отношения к третьей нормальной форме, соответственно описание ХД средствами реляционной алгебры, которая создавалась как инструмент нормализации БД, становится достаточно проблематичным. Поэтому далее приводится описание и анализ существующих подходов к моделированию ХД (обобщенная многомерная модель Энрико Франкони и Ананда Кэмбла, схема «звезда» и «снежинка», куб Грея, модель Кабиббо и Торлоне, многомерная модель фактов Голфарелли) с выводом о том, что работа по созданию модели ХД, объединяющей все этапы его проектирования является актуальной и востребованной.

Далее формулируются основные проблемы проектирования.

1. Руководителю, принимающему решения, необходимы самые разнообразные отчеты, причем каждый раз новые. Однако, даже хорошо подготовленный пользователь, успешно справляясь с операциями поиска и сортировки, не может правильно объединять таблицы. Существуют разнообразные инструменты (например, Crystal Reports, Reporting Services, Component One, Fast Report Studio), интерфейс которых достаточно прост, чтобы непрофессионалы в области информационных технологий могли готовить отчеты. Однако в этом случае структура ХД должна быть понятна пользователям.

2. Формирование нового заранее не определенного запроса – сложная квалифицированная работа с большими временными затратами. Данные хранилищ, как правило, представляют собой результаты таких запросов. Для описания нерегламентированных запросов заполнения ХД необходимо четко

представлять архитектуру исходной БД, в то время как разработчик ХД и администратор БД в общем случае не одно и тоже лицо.

3. Обработка запросов к хранилищу должна быть проведена с высокой производительностью, желательно в реальном масштабе времени. Поэтому должна быть обеспечена приемлемая скорость выполнения сложных аналитических запросов, для которых необходима последовательная обработка тысяч или миллионов записей.

Для решения задач автоматизации рассматривается возможность применения современной методология проектирования ХД — Dimensional. Однако проектирование ХД на основе модели Dimensional требует выбора так называемого «центрального вопроса», на практике таких вопросов может быть достаточно много, следовательно, под каждый из них необходимо создавать отдельное ХД. Это приводит к неоднократному выполнению сложной квалифицированной работы по описанию структуры будущего ХД, маршрута извлечения данных из множества реляционных таблиц (при этом надо хорошо представлять структуру исходной БД) и запроса на заполнение ХД. Очевидно, что автоматизация этих процессов значительно разгрузит разработчика и сократит сроки внедрения. Таким образом, автоматизация проектирования ХД является важной и актуальной задачей.

Далее в первой главе подчеркивается, что решение проблемы обеспечения высокой производительности обработки запросов к хранилищу является приложением задачи автоматизации. После сбора и предварительной обработки данные хранилища используются OLAP–системами. Скорость аналитической обработки зависит от способа представления данных. Поэтому далее проводится сравнительный анализ существующих структур данных с выводом о том, что необходимо разработать более приспособленную к анализу структуру данных, лишенную выявленных недостатков. На основе разработанной структуры должны быть предложены алгоритмы OLAP, учитывающие ее особенности.

В заключительной части главы приводится обоснование выбора инструментария для решения поставленных задач.

Во второй главе разработана единая математическая модель для описания цепочки преобразований: БД – ХД. Получение данного представления является начальным этапом алгоритма автоматизации.
Поскольку достаточно трудно описывать структуры ХД на абстрактных отношениях, то построение математических моделей рассматривается на конкретном примере (рисунок 1). Приведенный пример представляет собой усеченную версию БД предприятия, где осуществлялось внедрение.

В разработанном математическом представлении БД каждой сущности ставится в соответствие тензор. Сущности «Сотрудники» будет соответствовать тензор $F_{к}^{nx}$, сущности «Требования» – тензор N_{l}^{drb}, «Выпуск» – P_{jl}^{k}, «Изделие» – T_{j}^{acv}, «Сортировка» – W_{c}^{hs}, «Заказчики» – Q_{b}^{fy}, «Поставщики» – L_{p}^{mz}, «Детали» – M_{g}^{tp}, «Производство» — E_{jg}^{i}. Связь между сущностями БД,

которая в реляционной модели осуществляется через первичные ключи, в данном случае будет определяться наличием у тензоров одноименных индексов. В предлагаемой модели ковариантным (нижним) индексам соответствуют возможные ключи, по которым можно определить другие зависимые данные, обозначенные контравариантными (верхними) индексами. Таким образом, тензорная модель БД приведенной структуры имеет вид:

$$(F_r^{nx}, N_l^{drb}, P_{jl}^k, T_j^{acv}, W_c^{hs}, Q_b^{fy}, L_p^{mz}, M_g^{tp}, E_{jg}^i).$$

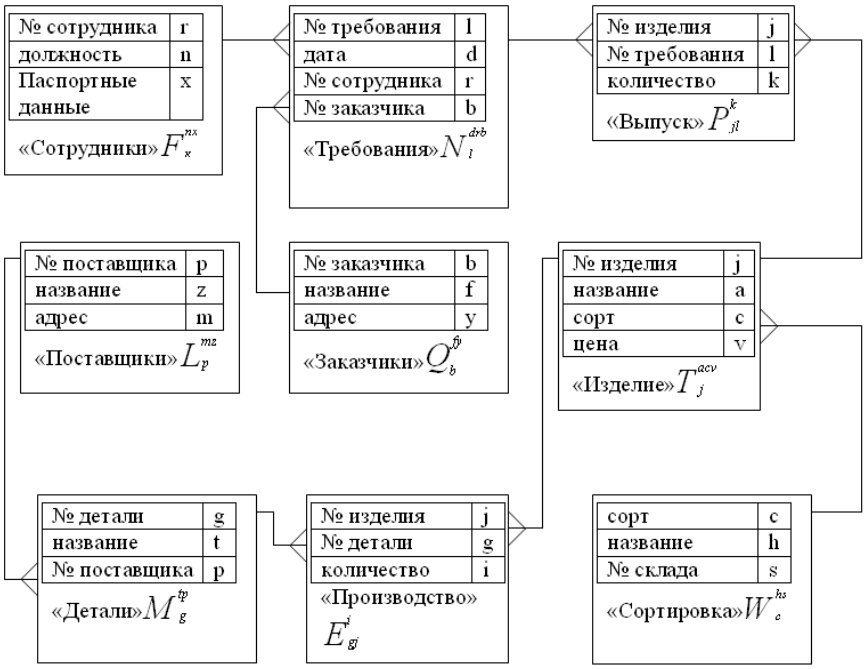

Рисунок 1 «Структура БД»

Поскольку наполнение ХД происходит посредством обработки запросов БД, необходимо представить математическую модель работы СУБД.

Рассмотрим тензорную модель простого запроса, то есть обращение к одной сущности. Так как представлением тензора является *N*–мерная матрица, которая при *N*=0 называется константой, при *N*=1 – вектором и т.д., то обработку запросов можно описать этими матрицами, представляющими тензор. Например, запрос к сущности «Выпуск» (рисунок 1).

В тензорной модели БД данной сущности поставлен в соответствие тензор P_{jl}^k. В матричном представлении тензор P_{jl}^k будет иметь вид:

$$\left\| k_{jl} \right\|_{n \times n} = \begin{Vmatrix} k_{11} & k_{12} & . & k_{1n} \\ k_{21} & k_{22} & . & k_{2n} \\ . & . & . & . \\ k_{n1} & k_{n2} & . & k_{nn} \end{Vmatrix},$$

где элементы k_{jl} отображают значения атрибута «количество», зависящие от ключевых атрибутов «№ изделия» (индекс *j*) и «№ требования» (индекс *l*) сущности «Выпуск» (рисунок 1).

Поток запросов к сущности БД обозначается $e^{\wedge}$, под «^» подразумевается индекс (индексы), обозначающий входные данные запроса. В общем случае $e^{\wedge}$

можно описать матрицей, в которой единичные элементы расположены на позициях равных значению «^».

Тогда тензорная модель запроса описывается выражением: $e^{jl} * P^{k}_{jl} = e^{k}$, которое имеет следующее матричное отображение:

$$\left\{ \left\| e_{jl} \right\|_{n\times n} \times \left\| k_{jl} \right\|_{n\times n} = \begin{Vmatrix} 0 & 0 & . & 0 \\ 0 & 0 & . & 0 \\ . & . & 1_{jl} & . \\ 0 & 0 & . & 0 \end{Vmatrix} \times \begin{Vmatrix} k_{11} & k_{12} & . & k_{1n} \\ k_{21} & k_{22} & . & k_{2n} \\ . & . & k_{jl} & . \\ k_{n1} & k_{n2} & . & k_{nn} \end{Vmatrix} = k_{jl} \right.$$

Замечание: знак **«*»** в тензорных уравнениях (выражениях) обозначает тензорную операцию умножения со сверткой (свертка производится по одноименным индексам).

Аналогично можно описать простые запросы ко всем сущностям БД. Таким образом, получим тензорную модель обработки простых запросов в виде системы:

$$\begin{cases} e^{b} * Q^{fy}_{b} = e^{y} * e^{f} \\ e^{p} * L^{mz}_{p} = e^{m} * e^{z} \\ e^{r} * F^{nx}_{r} = e^{n} * e^{x} \\ e^{l} * N^{drb}_{l} = e^{d} * e^{r} * e^{b} \\ e^{jl} * P^{k}_{jl} = e^{k} \\ e^{j} * T^{ac}_{j} = e^{a} * e^{c} \\ e^{g} * M^{tp}_{g} = e^{t} * e^{p} \\ e^{jg} * E^{i}_{jg} = e^{i} \\ e^{c} * W^{hs}_{c} = e^{hs} \end{cases}$$

Все обозначения приведены в соответствии с рисунком 1.

Для заполнения хранилищ данных, как правило, необходимо выполнение сложных запросов, в которых происходит обращение к нескольким сущностям. В этом случае каждый запрос будет описываться системой тензорных уравнений, обобщение которых представляет собой тензорную модель работы реляционной СУБД.

Для примера, приведенного на рисунке 1, эта модель будет иметь вид:

$$\begin{cases} e^{l} * N^{r}_{l} * F^{nx}_{r} = e^{nx} \\ e^{l} * N^{b}_{l} * Q^{fy}_{b} = e^{fy} \\ e^{l} * N^{d}_{l} = e^{d} \\ e^{j} * e^{l} * P^{k}_{jl} = e^{k} \\ e^{j} * T^{ac}_{j} = e^{ac} \\ e^{g} * e^{j} * E^{i}_{jg} = e^{i} \\ e^{g} * M^{p}_{g} * L^{mz}_{p} = e^{mz} \\ e^{g} * M^{t}_{g} = e^{t} \end{cases}$$

Тогда математическое описание одной из структур ХД (например, с данными о выпуске изделий), соответствующего схеме «звезда», имеет вид системы тензорных уравнений:

$$\begin{cases} e^{j} * P^{k}_{jl} = P^{k}_{l} = e^{l} * k \\ e^{l} * N^{drb}_{l} = e^{drb} \end{cases}.$$

Тензорная модель возможной структуры ХД (например, с данными о производстве), соответствующего схеме «снежинка», имеет следующий вид:

$$\begin{cases} e^j * E^{\,i}_{jg} = E^{\,i}_{g} = e^g * i \\ e^g * M^{\,tp}_{g} = e^{tp} \\ e^p * L^{\,mz}_{p} = e^{mz} \\ e^j * T^{\,acv}_{j} = e^{acv} \\ e^c * W^{\,hs}_{c} = e^{hs} \end{cases}$$

Представленные математические модели позволяют облегчить и ускорить процесс проектирования ХД. Для автоматизации этого процесса необходимо разработать алгоритм, основанный на предложенном математическом обеспечении. Результатом работы алгоритма должны быть возможные структуры ХД, получаемые из исходной БД. Таким образом, исходные данные представляют собой структуру БД и атрибуты, соответствующие количественным показателям процессов, отраженных в БД. В разработанном алгоритме исходная БД задается тензорной моделью, атрибуты – последовательностью контравариантных индексов.

В третьей главе диссертации описано кратномасштабное представление данных ХД, позволяющее значительно повысить производительность оперативного анализа, и представлены соответствующие алгоритмы аналитической обработки данных. Следует подчеркнуть, что данные, подвергаемые обработке в ХД и последующему анализу, как правило, являются числовыми значениями, то есть представляют численные характеристики рассматриваемого процесса. Поэтому последовательность анализируемых данных в общем случае представляет собой числовой ряд. Для его преобразования к более удобному для анализа кратномасштабному представлению используются вейвлеты Хаара (в диссертации приводится обоснование выбора этого базиса). Пусть временной ряд *W(t)* отображает численные показатели, содержащиеся в таблице фактов ХД. Мощность данного ряда $|W(t)| = n$, тогда количество уровней иерархии *p* вычисляется по формуле $p = \log_2 n$. Кратномасштабное представление данных в базисе Хаара выполняется по схеме, представленной на рисунке 2.

<table>
<tr><td>$w_{0,1}$</td><td rowspan="2">$w_{1,1} = \frac{w_{0,1} + w_{0,2}}{2}$</td><td rowspan="4">$w_{2,1} = \frac{w_{1,1} + w_{1,2}}{2}$</td><td></td></tr>
<tr><td>$w_{0,2}$</td><td></td></tr>
<tr><td>$w_{0,3}$</td><td rowspan="2">$w_{1,2} = \frac{w_{0,3} + w_{0,4}}{2}$</td><td></td></tr>
<tr><td>$w_{0,4}$</td><td></td></tr>
<tr><td>…</td><td>…</td><td>...</td><td>…
$w_{p,m}$
…</td></tr>
<tr><td>$w_{0,n-3}$</td><td rowspan="2">$w_{1,\frac{n}{2}-1} = \frac{w_{0,n-3} + w_{0,n-2}}{2}$</td><td rowspan="4">$w_{2,\frac{n}{4}} = \frac{w_{1,\frac{n}{2}-1} + w_{1,\frac{n}{2}}}{2}$</td><td></td></tr>
<tr><td>$w_{0,n-2}$</td><td></td></tr>
<tr><td>$w_{0,n-1}$</td><td rowspan="2">$w_{1,\frac{n}{2}} = \frac{w_{0,n-1} + w_{0,n}}{2}$</td><td></td></tr>
<tr><td>$w_{0,n}$</td><td></td></tr>
</table>

Рисунок 2 Кратномасштабное представление ряда

где $w_{p,m}$ — элемент с номером *m* на уровне разложения *p*.

Прежде чем данные ХД будут использованы OLAP–системой, производится предварительный подсчет числовых показателей – агрегирование. В диссертации доказано, что использование кратномасштабного представления данных позволяет ***значительно сократить время выполнения агрегирования***. Разработано четыре алгоритма агрегации: алгоритм агрегации по условным диапазонам, алгоритм разбиения на тетрады, алгоритм кратномасштабного погружения, алгоритм кратномасштабного разбиения.

Во второй части главы приводятся алгоритмы расчета аналитических показателей (тренд и периодичность), алгоритмы кратномасштабного анализа и алгоритм ретроспективного анализа.

Важной особенностью кратномасштабного разложения является возможность беглого анализа динамики бизнес процесса и возможность обратной оценки, то есть можно просмотреть укрупненный масштаб (нижние ряды разложения) данных с целью получения обобщенной и емкой картины исследуемого бизнес процесса. Данные этого представления будут прямой проекцией исходного ряда, поэтому сравнение чисел верхних уровней иерархии однозначно определяет соотношения между соответствующими диапазонами на нижнем уровне. В этом и заключается суть кратномасштабного анализа. Каждый последующий уровень иерархии обобщает информацию нижних уровней, представляя тем самым целостную картину. Данные идеи послужили основой алгоритмов кратномасштабного анализа.

Системы поддержки принятия решений должны обладать средствами предоставления пользователю данных о подобных последовательностях изменения исследуемого параметра в прошлом с целью получения вариантов стратегий для принятия оптимального решения. К тому же установление закономерности в таких последовательностях позволяет с некоторой долей вероятности предсказывать появление событий в будущем, что позволяет принимать более правильные решения. Такая задача называется ретроспективным анализом. Соответствующий алгоритм основывается на вейвлет – преобразовании анализируемых данных.

Основное преимущество представленных алгоритмов заключается в значительном ***выигрыше по быстродействию***, что особенно критично для систем оперативного анализа. Преимущество по сравнению с другими известными алгоритмами достигается, прежде всего, за счет применения кратномасштабнго представления, полученного по средствам вейвлет – преобразований данных.

Четвертая глава посвящена экспериментальной проверке работы разработанных алгоритмов. Разработанная программная система, подтверждающая корректность представленных алгоритмов, состоит из двух частей. Первая часть реализует алгоритм автоматизации проектирования ХД. Практическое значение разработанного алгоритма заключается в сокращении срока проектирования ХД. Соответствующий программный продукт ориентирован на использование администратором информационного обеспечения. Очевидно, что такие пользователи хорошо разбираются в

структуре администрируемой БД, но, как правило, достаточно нечетко представляют себе структуру будущего ХД. Разработанный программный продукт предоставляет пользователю все возможные варианты структуры ХД, которые можно построить на основе БД при заданных именах анализируемых значений. Практическая значимость данного программного продукта подтверждается успешным внедрением и эксплуатацией на предприятиях города.

Вторая часть разработанной программной системы представляет собой комплекс программных средств для агрегации и анализа данных ХД. Целью проводимого эксперимента является эмпирическое подтверждение преимуществ разработанных алгоритмов по быстродействию.

Алгоритмы расчета аналитических показателей сравниваются с энциклопедичным методом наименьших квадратов. В ходе эксперимента фиксировалось количество анализируемых данных (мощность исходного числового ряда), соответствующее время одного прогона алгоритма по показаниям встроенного программного таймера и количество элементарных операций. В результате получены следующие зависимости (рисунок 3):

Рисунок 3 Результаты экспериментальной оценки быстродействия предложенного алгоритма и известного метода

Программа, реализующая алгоритмы агрегации данных, сравнивалась с программным продуктом, производящим последовательное суммирование. В ходе эксперимента фиксировалась длина диапазона агрегации (количество данных, входящих в данный диапазон), соответствующее время одного прогона программ по показаниям встроенного программного таймера и количество элементарных операций. В результате получены следующие зависимости (рисунок 4):

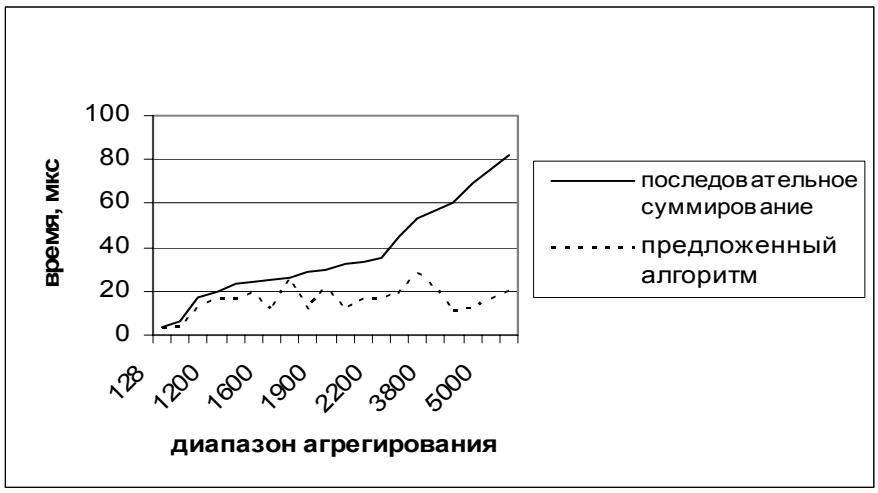

Рисунок 4 Результаты экспериментальной оценки быстродействия предложенного алгоритма агрегации и алгоритма последовательного суммирования

Алгоритмы кратномасштабного анализа являются специфическими, так как, по сути, являются следствием применения вейвлет – преобразования данных. В ходе эксперимента соответствующий программный продукт сопоставлялся программам, производящим последовательное сравнение данных анализируемого диапазона. Показания таймера и встроенного счетчика элементарных операций обнародовали следующие зависимости (рисунок 5):

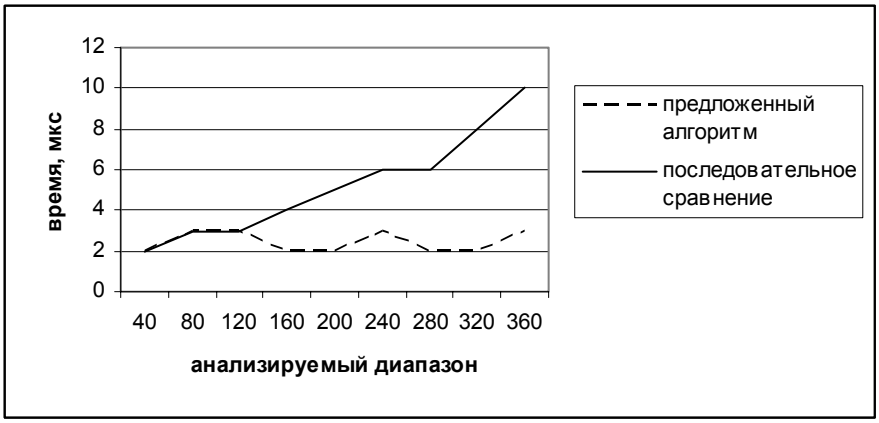

Рисунок 5 Результаты экспериментальной оценки быстродействия алгоритма кратномасштабного анализа

Программа, реализующая алгоритм ретроспективного анализа, сравнивались с программным продуктом, основанным на переборе. В ходе эксперимента фиксировалось количество данных последовательности – запроса, соответствующее время одного прогона программы и показания встроенного счетчика элементарных операций. В результате получены следующие зависимости (рисунок 6):

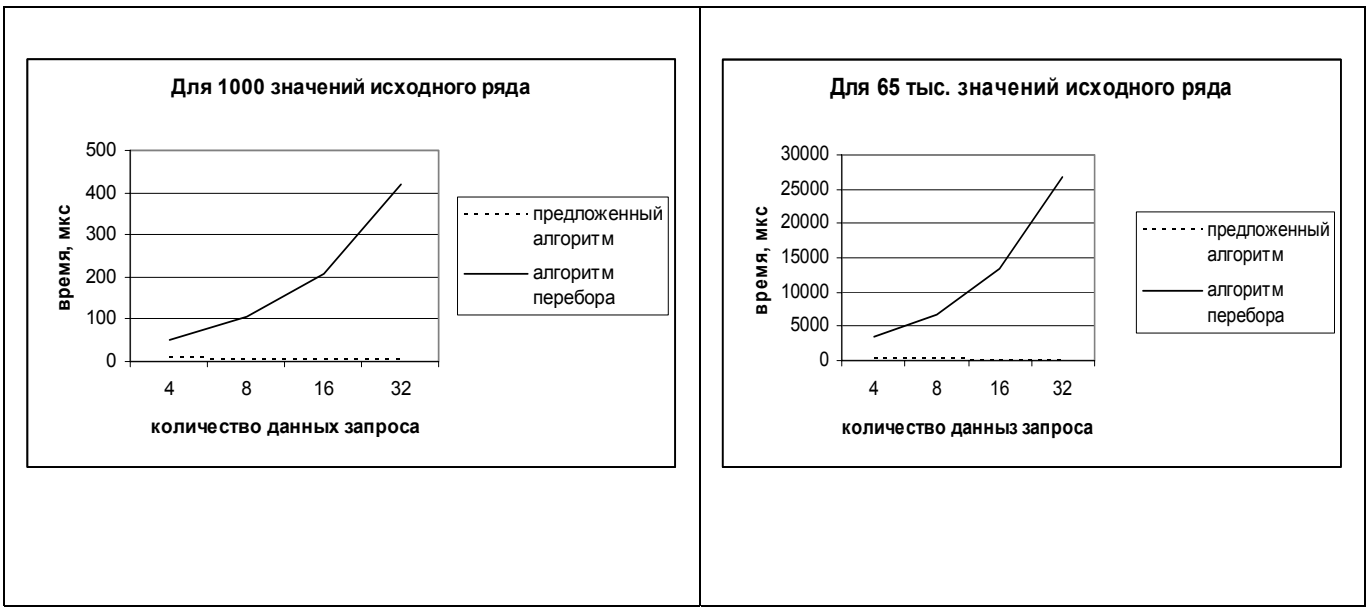

Рисунок 6 Результаты экспериментальной оценки быстродействия алгоритма ретроспективного анализа

Анализ полученных зависимостей показывает, что быстродействие и сложность разработанных алгоритмов по своим значениям приближены к константе, то есть практически не зависят от количества обрабатываемых данных. Эти преимущества достигаются за счет использования кратномасштабной структуры данных.

В заключении сформулированы основные результаты, полученные в диссертационной работе.

В приложении приводятся пояснения тензорных уравнений, справочная информация по теории кратномасштабного анализа, а также представлены копии актов о внедрении.

Основные результаты работы

Главным итогом диссертационной работы является разработка моделей и алгоритмов, позволяющих снизить трудоемкости проектирования ХД и обеспечить высокую производительность аналитических приложений, взаимодействующих с ХД.

В ходе выполнения работы:

1. Выполнен анализ современных работ в области новых информационных технологий хранения и обработки данных, рассмотрены существующие модели и методы проектирования ХД, проанализированы возможности применения в ходе автоматизированного проектирования ранее разработанных моделей и методов.
2. Разработано единое математическое описание исходных БД и ХД, что позволило моделировать процесс трансформации данных из нормализованной структуры БД в структуру, подчиненную условиям оперативного анализа.

3. Разработан алгоритм автоматизации проектирования ХД, снижающий временные затраты разработчика на построение из исходной БД множества возможных структур ХД.
4. Разработаны модели, позволяющие отображать многомерный характер данных, и моделировать операции присущие многомерному представлению информации.
5. Разработана модель данных, внедрение которой значительно повышает производительность работы с ХД.
6. На основе новой модели данных разработаны алгоритмы, позволяющие значительно ускорить процесс получения аналитических показателей.
7. Разработан программный комплекс, подтверждающий экспериментально преимущества разработанных алгоритмов.
8. Разработана и реализована программная система, позволяющая автоматизировать проектирование ХД и получать альтернативные структуры для выбора оптимальной с точки зрения задач анализа.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Бакулева М.А., Телков И.А. Применение современных математических методов для поиска информации в базах данных библиотек. Библиотековедение. Информационная деятельность: проблемы науки и практики: Материалы второй Международной научно-практической конференции. К., 2005. Ч.1. – С.165–167

2. Бакулева М.А. Тензорная модель работы реляционной СУБД// Информационные технологии в проектировании. Межвузовский сборник научных трудов. Рязань: РГРТА, 2004. – С.39–43

3. Бакулева М.А. Математическая модель построения многомерной БД. Информационные технологии и телекоммуникации в образовании и науке. Межвуз. сб. научн. трудов. Рязань: РГРТА, 2005. – С.23–26.

4. Бакулева М.А Применение вейвлет-преобразований в задачах поиска информации для систем поддержки принятия решений. Новые информационные технологии в научных исследованиях и в образовании. Материалы десятой научно-технической конференции студентов, молодых ученых и специалистов. Рязань: РГРТА, 2005. – С.103–104

5. Бакулева М.А Математическое представление операций манипулирования измерениями при проектировании многомерной базы данных. Проблемы передачи и обработки информации в сетях и системах телекоммуникации. Материалы четырнадцатой международной научно-технической конференции. Рязань, 2005. – С. 125–126

6. Бакулева М.А Применение вейвлет-преобразований в задачах поиска информации. Информационные технологии и телекоммуникации в образовании и науке. Межвуз. сб. научн. трудов. Рязань: РГРТА, 2006. –С.26–31

7. Бакулева М.А Анализ данных на основе вейвлет-преобразований в системах поддержки принятия решений. Новые информационные технологии в научных исследованиях и в образовании. Материалы одиннадцатой научно-

технической конференции студентов, молодых ученых и специалистов. Рязань: РГРТА, 2006. – С.92–93

8. Бакулева М.А Применение вейвлет-преобразований для представления данных хранилища. Вестник РГРТА. Научно-технический журнал. Выпуск 18. Рязань: РГРТА, 2006. – С.80–86

9. Бакулева М.А Применение кратномасштабного представления в хранилищах данных. Новые информационные технологии в научных исследованиях и в образовании. Материалы одиннадцатой научно-технической конференции студентов, молодых ученых и специалистов. Рязань: РГРТА, 2007. – С.250

10. Бакулева М.А, Бакулев А.В Применение вейвлет-преобразования для анализа данных хранилища. Вестник РГРТУ. Научно-технический журнал. Выпуск 21. Рязань: РГРТУ, 2007. – С.57–60

11. Телков И.А., Бакулева М.А. Разработка математической модели многомерной базы данных. Сборник материалов Всероссийского конкурса инновационных проектов аспирантов и студентов по приоритетному направлению развития науки и техники "Информационно-телекоммуникационные системы" / Под редакцией А.О. Сергеева. — М.: ГНИИ ИТТ "Информика", 2005. — 132 с.

Подписано в печать 27.09.2007 г. Формат 60×84 1/16.
Бумага для множительных аппаратов. Печать офсетная.
Гарнитура Times. Усл. печ. л. 1,0
Уч.-изд. л. 1,0. Тираж 100 экз.

Рязанский государственный радиотехнический университет
390005, Рязань, ул. Гагарина, д.59/1

Редакционно-издательский центр РГРТУ