На правах рукописи

СЕМЧЕНКОВ Сергей Юрьевич

АЛГОРИТМЫ ПРОЕКТИРОВАНИЯ СИСТЕМ МНОГОМЕРНОГО АНАЛИЗА ДАННЫХ, ОСНОВАННЫХ НА OLAP ТЕХНОЛОГИИ

Специальность 05.13.11. «Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Рязань 2010

Работа выполнена на кафедре вычислительной и прикладной математики ГОУВПО «Рязанский государственный радиотехнический университет»

Научный руководитель:	доктор технических наук, профессор Каширин Игорь Юрьевич
Официальные оппоненты:	доктор технических наук, профессор Шибанов Александр Петрович
	кандидат технических наук, доцент Швечков Виталий Александрович
Ведущая организация:	ОАО "Корпорация "Фазотрон-НИИР" - НИИ "Рассвет"

Защита состоится 20 октября 2010 г. в 12 часов на заседании диссертационного совета Д212.211.01 в Рязанском государственном радиотехническом университете по адресу: 390005, г. Рязань, ул. Гагарина, 59/1.

С диссертацией можно ознакомиться в библиотеке ГОУВПО «РГРТУ».

Автореферат разослан « 15 » сентября 2010 г.

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

Ученый секретарь
диссертационного совета
канд. техн. наук, доцент В.Н. Пржегорлинский

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Современный уровень развития аппаратных и программных средств сделал возможным повсеместное ведение баз данных оперативной информации на разных уровнях управления предприятием. Однако накопления оперативной информации недостаточно для получения релевантной информации, позволяющей руководителю принимать важные управленческие решения и формировать стратегию развития предприятия на основе ключевых показателей. Решение этой проблемы видится ученым во внедрении OLAP технологии. OLAP – технология обработки информации, позволяющая агрегировать информацию из нескольких источников данных в виде многомерных представлений, а также выполнять аналитические запросы пользователя, включая составление и динамическую генерацию отчетов.

Суть этой технологии заключается в формировании единого источника информации, содержащего согласованные и непротиворечивые данные, полученные в ходе извлечения, преобразования и переработки данных из баз данных, содержащих накопленную к текущему моменту оперативную информацию. Как правило, OLAP системы содержат не все данные из систем оперативной обработки данных, а только те, которые имеют отношение к основным ключевым показателям, характеризующим деятельность предприятия. Весомые результаты в работах по OLAP системам связаны с такими учеными, как Н. Караянидис, Д. Педерсен, Р. Агравал, М. Гольфарелли, Р. Торлоне, Д. В. Ивлев, П. П. Ишенин, А.К. Дорожкин.

В рамках OLAP технологии различными группами специалистов разработано большое количество программных продуктов, реализующих многомерную модель данных. Среди этих продуктов можно выделить OLAP Option to Oracle Database фирмы Oracle, Microsoft Analysis Services фирмы Microsoft, Palo фирмы Jedox, Mondrian фирмы Pentaho. Удобство использования конечной системы, ее масштабируемость, производительность и функциональность зависят от средств, предназначенных для автоматизированного проектирования такого рода систем. Опыт разработки систем и эксплуатации реальных продуктов для автоматизированного проектирования позволил выявить следующие проблемы, решение которых является наиболее актуальным.

1. Отсутствие адаптивной подстройки под конкретного пользователя. Аналитические системы, созданные на базе OLAP технологии, строятся на основе предметно-ориентированного подхода, то есть для решения конкретных задач пользователя. При первоначальном проектировании аналитической системы невозможно учесть все интересы пользователей, что приводит к созданию избыточных для конкретного пользователя структур хранения данных. Несмотря на большое количество отчетов, получаемых с помощью OLAP технологии, пользователя, в зависимости от текущей ситуации, интересует ограниченный набор срезов. Существующие системы не учитывают это

обстоятельство, в результате чего время, необходимое для принятия решения, существенно увеличивается.

2. Существенное снижение производительности системы при увеличении числа пользователей. Под масштабируемостью будем понимать функцию, описывающую зависимость характеристики производительности (время выполнения запроса, пропускная способность) от размеров системы (количества оборудования, объема хранения данных, количества поступающих запросов). Применение OLAP технологии решает проблему невысокой производительности систем оперативной обработки данных при выполнении запросов на выборку из большого количества таблиц. Однако количество пересылаемых данных в клиент-серверной архитектуре аналитических систем остается большим, что приводит к резкому увеличению времени выполнения запроса при увеличении количества запросов пользователей.

На основании сказанного можно сделать вывод об актуальности выбранной темы диссертационной работы.

Цель диссертационной работы состоит в разработке и исследовании формализмов, позволяющих уменьшить время выполнения запросов пользователя за счет адаптивной подстройки системы под изменяющиеся интересы пользователя.

Для достижения поставленной цели необходимо решить следующие **основные задачи**.

1. Разработка математического формализма, позволяющего адекватно описывать гиперкубы OLAP систем и операции над многомерными кубами.

2. Разработка алгоритмов преобразования многомерного куба к регулярной структуре для автоматизации внесения корректных изменений в гиперкуб.

3. Разработка модели пользователя, позволяющей учитывать интересы различных групп пользователей, выделяя для них соответствующие подкубы.

4. Разработка алгоритма декомпозиции многомерного куба на основе предложенных формализмов.

5. Разработка новой архитектуры OLAP систем, учитывающей принципы декомпозиции и кластеризации данных на основе пользовательских интересов.

6. Программная реализация алгоритмов проектирования систем многомерного анализа данных.

Методы исследования. Разработка и исследование проводились с использованием теории алгебраических систем, теории реляционных баз данных, методов объектно-ориентированного проектирования.

Научная новизна работы состоит в следующем.

1. Разработана новая математическая модель систем многомерного анализа данных на основе понятий базового и многомерного куба. Основным преимуществом модели является произвольная последовательность выполнения операций без необходимости выполнения объединения с другими кубами.

2. Разработаны алгоритмы преобразования многомерного куба к регулярной структуре, позволяющие выполнять корректное вычисление агрегированных показателей, избегая множественного наследования.

3. Разработан алгоритм иерархической декомпозиции многомерных кубов, предназначенный для автоматизации проектирования OLAP систем и оптимизации их структуры.

4. Разработана модель пользователя OLAP систем, с помощью которой можно определить общие интересы группы пользователей, составив для них унифицированную концептуальную иерархию потребностей.

5. Разработана новая архитектура OLAP систем. Основным преимуществом архитектуры является снижение количества запросов пользователей к центральному серверу и сокращение времени выполнения запросов пользователя.

Практическая значимость. На основе разработанных теоретических результатов были получены алгоритмы проектирования регулярных структур многомерной модели данных, а также разработаны принципы комбинированного выполнения запросов к OLAP серверу. Эффект от внедрения этих принципов выражается в сокращении интенсивности запросов к центральному серверу до 30 % и уменьшении времени выполнения запросов пользователя в среднем на 40 %. Результаты диссертации нашли отражение в реальной программной системе CuDBIS v. 1.02, предназначенной для оптимизации структуры многомерного куба.

Апробация результатов диссертации. Основные результаты диссертационной работы были представлены на следующих конференциях.

1. МНТК «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций». Рязань, РГРТА, 2005 г.

2. МНТК «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций». Рязань, РГРТУ, 2008 г.

3. Всероссийская НТК «Новые информационные технологии в научных исследованиях и образовании». Рязань, РГРТУ, 2008 г.

4. Всероссийская НМК «Методы обучения и организация учебного процесса в вузе». Рязань, РГРТУ, 2009 г.

5. Всероссийская заочная НТК «Информационные технологии в науке, проектировании и производстве». Нижний Новгород, 2009 г.

6. Всероссийская НПК «Информационные технологии в науке, экономике и образовании». Бийск, Бийский технологический институт, 2009 г.

7. Всероссийская НТК «Научная сессия ТУСУР-2009». Томск, Томский государственный университет систем управления и радиоэлектроники, 2009 г.

Публикации. По теме диссертации было опубликовано 14 работ, из них 7 тезисов докладов международных и всероссийских конференций, 4 статьи в межвузовских сборниках, 2 статьи в журналах из списка ВАК, одно свидетельство об официальной регистрации программы.

Внедрение результатов работы. Результаты исследования внедрены в форме информационно-аналитического интернет-сервиса в ООО «Интертех», специализирующемся на продаже потребительской электроники, аудио-, видео- и бытовой техники, а также в учебный процесс ГОУВПО «Рязанский государственный радиотехнический университет».

ОСНОВНЫЕ ПОЛОЖЕНИЯ, ВЫНОСИМЫЕ НА ЗАЩИТУ

1. Новый формализм описания многомерных кубов OLAP систем на основе понятия «базового куба».
2. Алгоритм иерархической декомпозиции многомерного куба OLAP систем.
3. Алгоритм устранения несбалансированности иерархии измерений куба.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность выбора темы диссертации, формулируется цель исследований, научная новизна и практическая ценность основных результатов.

В первой главе «Проблема адаптивного анализа данных в OLAP технологии» формулируются цели и задачи OLAP технологии как инструмента для агрегирования данных из нескольких источников и динамической генерации отчетов.

Рассмотрены основные понятия OLAP систем, проведена их классификация. В зависимости от способа организации данных в многомерных кубах выделяют следующие виды систем.

1. MOLAP системы – исходные и агрегированные данные хранятся в многомерных структурах. 2. ROLAP системы – исходные данные хранятся в реляционной БД, а агрегированные – в служебных таблицах той же БД. 3. HOLAP системы – гибридная архитектура, объединяющая ROLAP и MOLAP. Для каждого способа приведено краткое описание, указаны достоинства и недостатки.

Приведены основные требования к OLAP системам, рассмотрены различия OLAP и OLTP систем, приводящие к необходимости наличия отдельной многомерной СУБД, интегрирующей данные из внешних источников и обрабатывающей аналитические запросы пользователей системы.

Проанализированы существующие подходы к формальному описанию OLAP систем. Рассмотрены следующие формальные модели: модель Агравала, модель Ли, Ванга, модель Датта, Томаса, модель Гиссенса, модель Каббибо-Торлоне. Для всех моделей приведено описание представления структуры элементов многомерной модели и операций над многомерным кубом. Подробно рассмотрены наиболее распространенные программные реализации

технологии аналитической обработки данных: OLAP Option to Oracle Database фирмы Oracle, Microsoft Analysis Services фирмы Microsoft, Palo фирмы Jedox, Mondrian фирмы Pentaho, Cognos TM1 фирмы IBM.

Отсутствие учета аномалий, возникающих в иерархии измерений, приводит к некорректному вычислению агрегированных показателей. Кроме того, перечисленные модели предполагают однородность многомерного пространства с точки зрения пользователя. Отсутствие учёта интересов пользователя приводит к невозможности дополнительной оптимизации производительности.

На основе проведенного анализа сформулированы цель и задачи диссертации.

Во второй главе «Формальное описание адаптивных OLAP систем» разрабатывается формализм гиперкубов OLAP систем, позволяющий адекватно описывать и преобразовывать многомерные кубы для последующего их использования в информационных аналитических системах.

Множество-носитель всех гиперкубов Λ представляется в виде декартова произведения множеств: $\Lambda = \Theta \times \Psi \times V \times Y$, где Θ – множество всех измерений многомерного пространства, Ψ – множество возможных уровней всех измерений, V – множество возможных значений всех измерений, Y – множество возможных значений всех ячеек многомерного куба. Связи между подмножествами множества-носителя задаются с помощью бинарных и тернарных отношений, соответствия выражаются с помощью сечений отношений.

Основным элементом модели является *базовый куб*, содержащий наиболее детализированные данные. Базовый куб C_b может быть представлен системой кортежей $\langle D_b, L_b, R_b \rangle$:

1) $D_b = \langle D_{b1}, D_{b2}, \ldots, D_{bq}, M_b' \rangle$ – кортеж измерений, $D_{bi} \in \Pi p_1(\Lambda)$, $i = 1 \ldots q, M_b' \in \Pi p_1(\Lambda)$, где M_b' – измерение, представляющее показатель куба;

2) $L_b = \langle DL_{b1}, DL_{b2}, \ldots, DL_{bq}, ML_b' \rangle$ – кортеж уровней измерений, $DL_{bi} \in \Pi p_2(\Lambda), i = 1 \ldots q, ML_b' \in \Pi p_2(\Lambda)$, где ML_b' – это уровень измерения показателя куба;

3) R_b – множество значений ячеек куба в виде кортежей $x = \langle v_1, v_2, \ldots, v_q, m_x \rangle$, где $v_i \in \Pi p_3(\Lambda), i = 1 \ldots q, m_x \in \Pi p_4(\Lambda)$.

Многомерный куб C может быть представлен системой кортежей $\langle C_b, D, L, R \rangle$.

1. C_b – это базовый куб.

2. $D = \langle D_1, D_2, \ldots, D_n, M' \rangle$ – кортеж измерений куба, $n \le q, D \subseteq D_b$.

M' – это измерение показателя куба.

3. $L = \langle DL_1, DL_2, \ldots, DL_n, ML' \rangle$ – кортеж уровней измерений. $L \subseteq \Pi p_2(\Lambda)$.

4. R – это множество значений ячеек куба в виде кортежей $x = \langle v_1, v_2, \ldots, v_q, m_x \rangle$, $R \subseteq \Pi p_{(3,4)}(\Lambda)$.

Для получения информации из базы данных посредством гиперкуба используются соответствующие операции. Все операции над многомерными кубами можно разбить на простейшие – повышение уровня, применение функции, проекция, выборка – и операции, основанные на базе простейших, - навигация и срез. Аргументами каждой из операций являются исходный куб и куб-шаблон, задающий параметры операции. Результатом операции является новый куб $C' = Operation(C, C^{\sigma})$, где Operation – выполняемая операция. Рассмотрим операции более подробно.

1. Операция *повышения уровня* $C' = \varphi(C, C^{\sigma})$ заключается в том, что значения измерений, уровень которых необходимо повысить, заменяются значениями, соответствующими более высокому уровню этих измерений, значения остальных измерений не изменяются.

2. Операция *применения функции* $C' = \theta(C, C^{\sigma})$ состоит в получении агрегированных значений на основе детализированных с помощью функции агрегации.

3. Операция *проекции* $C' = \pi(C, C^{\sigma})$ – это удаление измерения из многомерного куба при сохранении измерения в базовом кубе.

4. Операция *выборки* $C' = \rho(C, C^{\sigma})$ позволяет выделить подмножество из исходного многомерного куба.

5. Операция *навигации* $C' = \eta(C, C^{\sigma}) = \theta(\varphi(C, C^{\sigma 1}), C^{\sigma 2})$ является операцией, основанной на базе простейших операций. Навигация – это изменение уровня выбранного измерения с последующей генерацией нового куба с использованием операции *применение функции*.

6. Операция *среза* $C' = \chi(C, C^{\sigma}) = \theta(\pi(C, C^{\sigma 1}), C^{\sigma 2})$ – это удаление выбранного измерения с последующей агрегацией измерений с использованием выбранной пользователем функции агрегации.

Для преобразования многомерных кубов к оптимизированным формам разработана операция иерархической декомпозиции. Декомпозиция куба основана на разбиении куба на подкубы в соответствии с иерархией измерений. Декомпозиция применима только к кубу или подкубу, содержащему наиболее детализированные данные. Алгоритм декомпозиции состоит из следующих основных этапов.

1. Нумерация всех вершин всех измерений многомерного куба.

Принцип нумерации должен поддерживать отношение «родитель-потомок». Это означает, что должно существовать отображение, которое индексу каждой вершины иерархии ставит в соответствие индекс родительской вершины. Принцип нумерации должен быть одинаков для всех измерений.

2. Сортировка измерений по убыванию количества уровней. Результатом этого этапа является кортеж, содержащий в отсортированном виде (по убыванию) количество уровней всех измерений. Количество этапов декомпозиции определяется наибольшим количеством уровней среди всех измерений многомерного куба.

3. Разбиение отсортированного кортежа, полученного на предыдущем этапе, на подмножества кортежей, содержащих одинаковые элементы. Мощность каждого подмножества будет определять количество измерений, по которым будет происходить декомпозиция.

4. На каждом этапе декомпозиции для каждого измерения, по которому проводится декомпозиция, определяется количество составных частей этого измерения в соответствии с количеством значений на текущем уровне измерения. Под разбиением измерения на составные части понимается выделение подмножества значений этого измерения. На основании разбиения измерения можно определить количество подкубов, получаемых на каждом этапе декомпозиции.

5. Построение множества подкубов, являющихся результатом декомпозиции, на измерениях с ограниченным (в соответствии с предыдущим этапом) количеством элементов.

Для получения корректных результатов при агрегировании недостаточно введения понятия базового куба и определения многомерного куба через базовый. Иерархическая структура каждого измерения многомерного куба должна удовлетворять следующим требованиям:

1) все измерения многомерного куба должны быть попарно независимы, а показатель должен полностью определяться набором значений терминальных уровней иерархий измерений;

2) запрещается неполнота иерархий всех измерений;

3) в многомерном кубе не должно быть несбалансированных иерархий;

4) в многомерном кубе должно отсутствовать множественное наследование в иерархии измерения.

Декомпозиция многомерного куба дает возможность построения модели пользователя OLAP систем, учитывающей его интересы и потребности. Пространство интересов пользователя можно представить с помощью концептуальной иерархии потребностей. Концептуальная иерархия представляет собой множество понятий, упорядоченных с помощью древовидной структуры.

Отображением концептуальной иерархии потребностей в многомерной модели является многомерный куб интересов пользователя. Измерениями многомерного куба интересов служат понятия, являющиеся терминальными вершинами в концептуальной иерархии потребностей пользователя, а также

дополнительное временное измерение. Показателем многомерного куба интересов является количество переходов пользователя к конкретной потребности. Интересы каждого пользователя представляются в виде динамически изменяющегося во времени n-мерного куба.

Для каждого интересующего пользователя понятия можно определить функцию потребности k-го пользователя $f_k(t, d_i)$ как ранг интереса, изменяющийся во времени. Аргументами функции потребности являются конкретный момент времени t и интересующая пользователя потребность, соответствующая терминальной вершине концептуальной иерархии потребностей. Рассматривая функцию потребности во времени, для каждого k-го пользователя можно составить пространство интересов, состоящее из функций потребностей:

$$IS_k = \{f_k(t, d_1), f_k(t, d_2), ..., f_k(t, d_i), ...\},$$

где IS (interest space) – пространство интересов k-го пользователя. Сформировать соответствующий интересам конкретного пользователя многомерный куб можно с помощью операций выборки, проекции и композиции многомерных кубов.

Из различных поддеревьев можно получить одну унифицированную концептуальную иерархию, отражающую обобщенные интересы группы пользователей. Для кластеризации интересов пользователей используется представление кластера в виде гистограммы, показывающей количество вхождений каждого интереса пользователя в кластер (рис. 1).

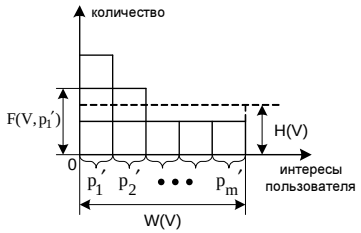


Рис. 1. Гистограмма кластера

Пусть требуется выполнить кластеризацию n кубов интересов $\{C_1', C_2', \ldots, C_n'\}$. Куб можно представить в виде множества интересов $\{p_1', p_2', \ldots, p_m'\}$. Тогда кластеризация заключается в нахождении множества $\{C_1'', C_2'', \ldots, C_k''\}$ такого, что

$$C_1'' \cup C_2'' \cup \ldots \cup C_k'' = \{C_1', C_2', \ldots, C_n'\},$$

причем $C_i'' \neq \varnothing\ (i = 1 \ldots k)$, $C_i'' \cap C_j'' = \varnothing, 1 \le i \le k, 1 \le j \le k, i \neq j$.

Для каждого кластера V можно определить следующие характеристики: 1) W(V) – ширина кластера, равная в данном случае мощности множества уникальных потребностей каждого пользователя; 2) F(V, p) – количество вхождений элемента p в кластер V; 3) S(V) – площадь гистограммы кластера, равная сумме количеств каждого элемента кластера $\sum_{p_i \in U(V)} F(V, p_i)$, где U(V) – множество уникальных интересов; 4) H(V) – высота кластера, равная отношению площади к ширине: H(V)=S(V)/W(V); 5) градиент кластера –

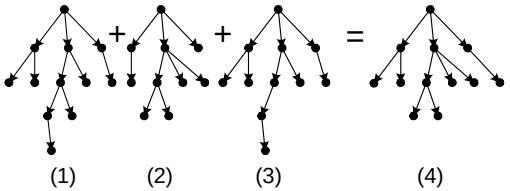
$G(V) = H(V) / W(V) = S(V) / W^2(V)$. Градиент кластера – это характеристика, с помощью которой можно осуществить сравнение нескольких разбиений, причем большее значение градиента означает лучшее разбиение.

Формула для вычисления глобальной функции стоимости имеет следующий вид:

$$T_r(V) = \frac{\sum_{i=1}^{k} \frac{S(V_i)}{W^r(V_i)} x |V_i|}{\sum_{i=1}^{k} |V_i|}.$$

Параметр r – коэффициент отталкивания, который регулирует уровень сходства кубов интересов пользователей внутри кластера.

Коэффициент отталкивания r подбирается пользователем, при этом чем больше r, тем меньше уровень сходства, что соответствует большему количеству кластеров. Процесс кластеризации заключается в создании новых кластеров или добавлении новых кубов интересов в один из существующих кластеров, а критерием выбора конкретного действия является максимизация значения глобальной функции стоимости.



На основе концептуальных иерархий потребностей всех пользователей можно с помощью интеграции разнородных деревьев определить общие интересы различных групп пользователей (рис. 2).

Рис. 2. Интеграция поддеревьев интересов

В третьей главе «Автоматизированное проектирование и оптимизация многомерных структур в OLAP системах» на основе предложенных формализмов разработаны архитектура и алгоритмы адаптивных OLAP систем с учетом принципов декомпозиции и кластеризации данных на основе пользовательских интересов.

Для декомпозиции измерений используется следующий принцип нумерации значений измерения. Вершины самого верхнего уровня, соответствующие наиболее агрегированным данным, будем обозначать одним индексом, причем нумерация начинается с нуля. Вершины, находящиеся на следующем уровне иерархии, обозначаются двумя индексами, разделенными точкой. При этом первый индекс совпадает с индексом родительской вершины, а нумерация второго индекса начинается с нуля. Вершины третьего уровня будут обозначаться тремя индексами, соответствующими предкам на первом и втором уровнях и т.д. Схема алгоритма декомпозиции представлена на рис. 3. Для проверки несбалансированности иерархии будем пользоваться условием неравенства нулю количества терминальных вершин.

Схема алгоритма приведения иерархии к сбалансированной приведена на рис. 4.

Этот алгоритм приводит иерархию конкретного измерения D_i к сбалансированному виду путем искусственного добавления потомков ко всем нетерминальным вершинам, не принадлежащим последнему уровню иерархии.

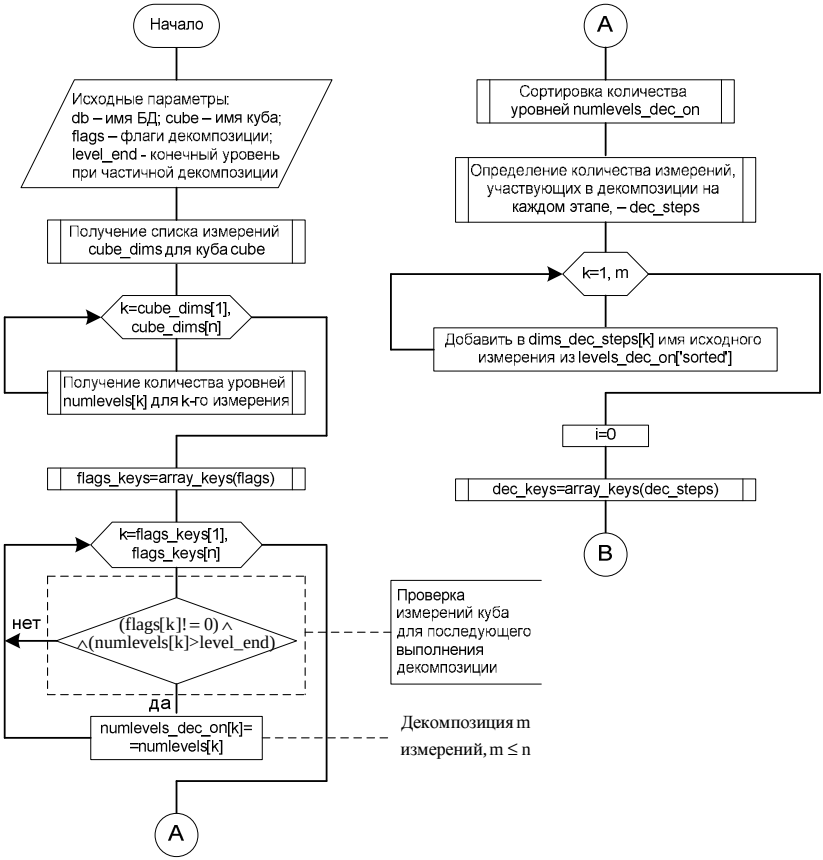


Рис. 3. Схема алгоритма декомпозиции многомерного куба

Добавление потомков в данном случае является рекурсивным, то есть к нетерминальной вершине добавляется один потомок, к которому, в свою очередь добавляется еще одна вершина в качестве потомка и т.д. Процесс добавления потомков будет завершен после добавления $|\text{Сеч}_{D_i}(f_l)|-k$ потомков, где $|\text{Сеч}_{D_i}(f_l)|$ – общее количество уровней в иерархии, k – номер уровня вершины, из-за которой образовалась несбалансированность иерархии. Для устранения аномалии множественного наследования используется

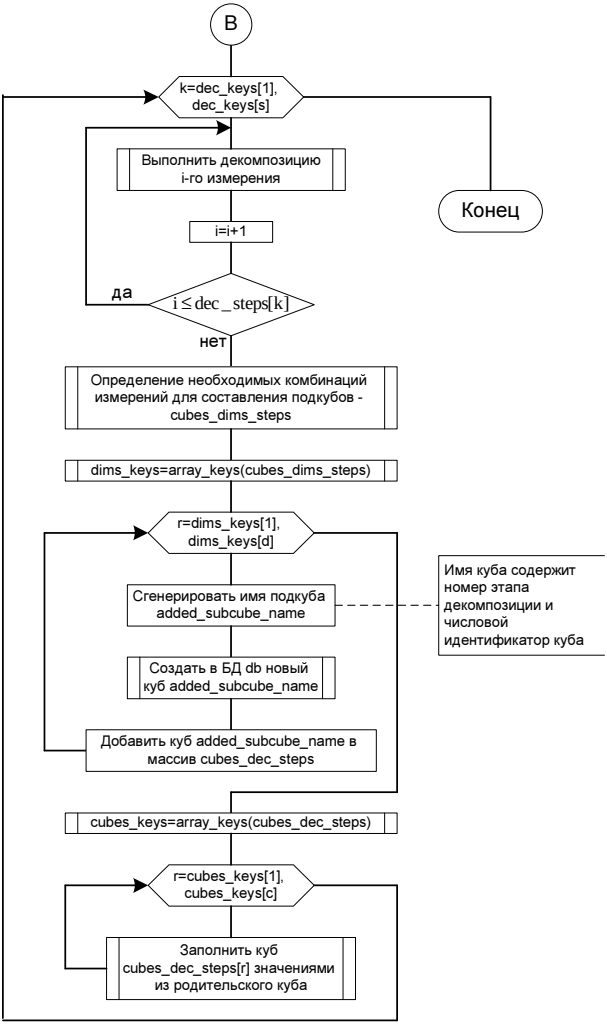


Рис. 3 (окончание). Схема алгоритма декомпозиции многомерного куба

присоединенный куб. Идея присоединенного куба состоит в том, что ячейки этого куба содержат значения некоторых измерений. Присоединенный куб состоит из трех измерений: разделяемое измерение, классификационное измерение и ссылочное измерение. Разделяемое измерение является общим для присоединенного куба и того многомерного куба, к которому относится данный многомерный куб. Такая конструкция удобна тем, что позволяет обеспечить автоматическое добавление, обновление и удаление элемента измерения многомерного куба. Классификационное измерение присоединенного куба позволяет классифицировать или ранжировать свойства сущностей,

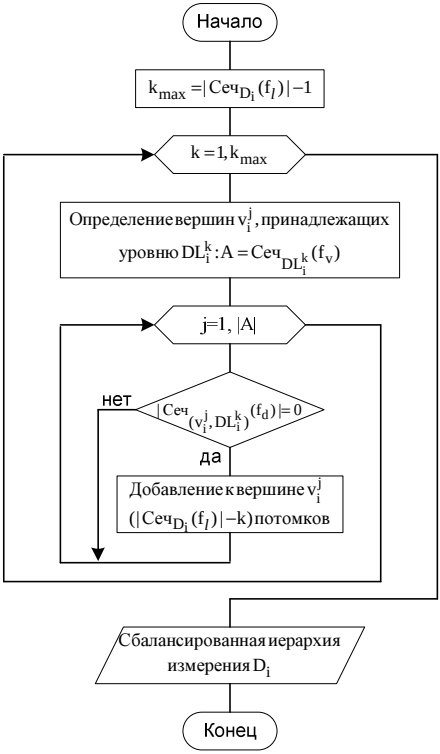


Рис. 4. Схема алгоритма устранения несбалансированности иерархии

относящихся к разделяемому измерению.

Ссылочное измерение определяет одно или несколько измерений, значения индексов которых будут находиться в ячейках присоединенного куба.

Общая структура алгоритма устранения аномалии множественного наследования состоит из пяти шагов:

1) формирование копии разделяемого измерения;

2) проведение изменений по устранению аномалии над оригиналом разделяемого измерения;

3) формирование значений классификационного и ссылочного измерений;

4) построение и заполнение присоединенного куба на основании информации, полученной из сохраненной копии разделяемого измерения, а также результатов выполнения двух предыдущих шагов;

5) удаление копии разделяемого измерения.

Присоединенный куб формируется на основе преобразованного разделяемого измерения, не содержащего аномалий. Для заполнения ячеек присоединенного куба может использоваться информация о начальной структуре

иерархии, поэтому необходимо формирование копии разделяемого измерения. Разрешение аномалии может осуществляться двумя способами.

1. Создание отдельной родительской вершины для значения измерения, которое вызывает аномалию. В этом случае пользователю необходимо самостоятельно сформировать элементы классификационного измерения. В качестве значений ячеек многомерного куба указывается индекс, который выражает взаимосвязь между различными вершинами разделяемого измерения.

2. Выделение приоритетной родительской вершины для значения измерения, которое вызывает аномалию. В этом случае классификационное измерение формируется на основе альтернативных иерархий. В качестве значений ячеек многомерного куба указывается индекс родительской вершины, который выражает отношение принадлежности к альтернативным классификациям.

В четвертой главе «Принципы программной реализации адаптивных систем автоматизированного проектирования многомерного анализа данных в программной системе CuDBIS» для исследования и сравнительного анализа предложенных формализмов и алгоритмов разработан программный комплекс CuDBIS (**Cu**be **D**ecomposition **B**ased on **I**nterest **S**pace – декомпозиция куба, основанная на пространстве интересов).

Эксперименты проводились в организации ООО «Интертех», торгующей бытовой техникой. Результаты экспериментов представлены на рис. 5, 6.

В ходе экспериментов было зафиксировано снижение количества запросов в единицу времени к центральному серверу до тридцати процентов. Снижение количества запросов к центральному серверу произошло вследствие того, что запросы пользователей, попадающие в область куба интересов пользователя, обрабатываются локально. Аналитическая обработка данных при использовании локального сервера происходит существенно быстрее, так как локальный многомерный куб, построенный на базе куба интересов, не содержит лишних для этого пользователя срезов и измерений. Время выполнения запроса пользователя уменьшилось в среднем на сорок процентов. Уменьшение времени выполнения запроса можно объяснить наличием локально обрабатываемых запросов. Уменьшение количества запросов к центральному серверу приводит к снижению нагрузки на него, а следовательно, к уменьшению времени выполнения запроса и для тех пользователей, для которых оптимизации не применяются.

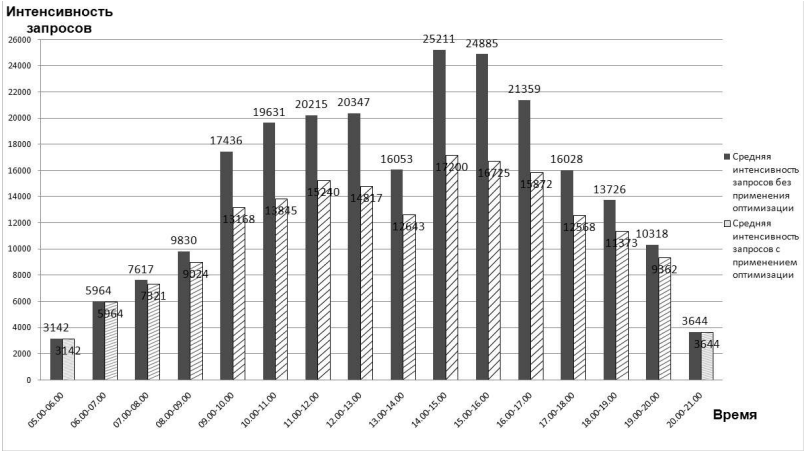


Рис. 5. Зависимость средней интенсивности запросов к серверу от времени

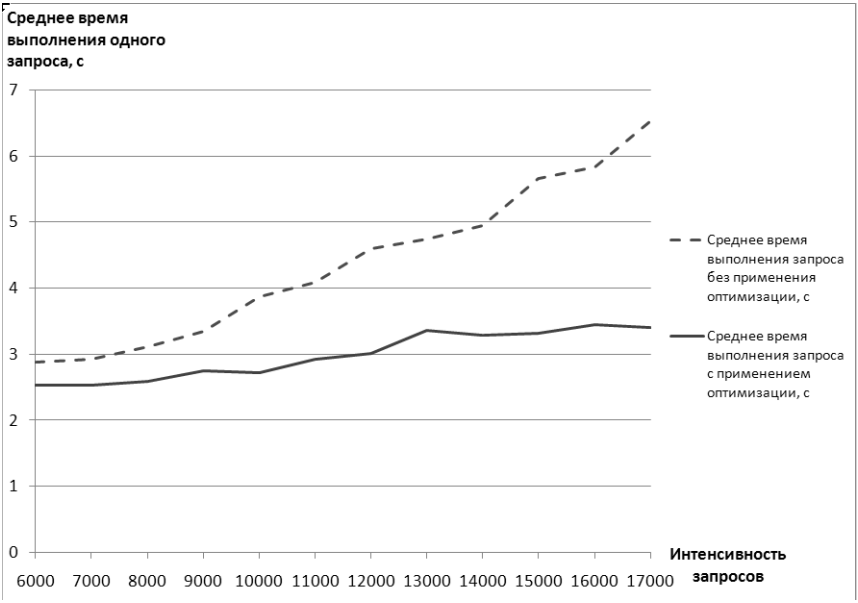


Рис. 6. Зависимость среднего времени выполнения запроса от количества запросов в единицу времени

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Проведен сравнительный анализ современных подходов к формальному описанию OLAP систем. Приведена классификация программных продуктов, реализующих технологию многомерного анализа данных. Разработана новая математическая модель систем многомерного анализа данных на основе понятий базового и многомерного куба. Основным преимуществом модели является произвольная последовательность выполнения операций без необходимости выполнения объединения с другими кубами. Проанализированы и показаны на конкретных примерах оптимизирующие и структурирующие свойства операций для преобразования гиперкубов.

2. Разработаны алгоритмы преобразования многомерного куба к регулярной структуре: определения неполноты иерархии, устранения аномалий несбалансированности и множественного наследования. Указанные алгоритмы позволяют выполнять корректное вычисление агрегированных показателей.

3. Разработан алгоритм иерархической декомпозиции многомерных кубов, предназначенный для автоматизации проектирования OLAP систем и оптимизации их структуры, а также позволяющий эффективно выполнять запросы, содержащие ограничения на значения измерений. Предложенный алгоритм предоставляет возможность построения модели пользователя OLAP систем, учитывающей его интересы и потребности.

4. Рассмотрены вопросы кластеризации пользователей по интересам, разработана модель пользователя OLAP систем, с помощью которой можно определить общие интересы группы пользователей, составив для них унифицированную концептуальную иерархию потребностей. Этим достигается оптимизация запросов, связанных с построением срезов, по времени.

5. Разработана новая архитектура OLAP систем, учитывающая принципы декомпозиции и кластеризации данных на основе пользовательских интересов, а также позволяющая производить реструктуризацию иерархической структуры для устранения аномалий в ней. Основным преимуществом архитектуры является снижение количества запросов пользователей к центральному серверу и сокращение времени выполнения запросов пользователя.

6. Определены условия применимости комбинированной схемы с центральным и локальными серверами. Получено экспериментальное подтверждение эффективности декомпозиции многомерного куба, основанной на интересах пользователя, и клиент-серверной архитектуры с комбинированным исполнением запросов. Эффект от внедрения выражается в сокращении интенсивности запросов к центральному серверу до 30 % и уменьшении времени выполнения запросов пользователя в среднем на 40 %.

ПУБЛИКАЦИИ ПО ОСНОВНЫМ РЕЗУЛЬТАТАМ ДИССЕРТАЦИИ

1. Семченков С.Ю. Вопросы использования OLAP систем для анализа информации // Проблемы передачи и обработки информации в сетях и системах телекоммуникаций: материалы 14-й международной научно-технической конференции (Рязань, 6-8 декабря 2005 г.). – Рязань, 2005. – С. 179-180.

2. Семченков С.Ю. Вопросы организации промежуточной области хранения для OLAP систем // Математическое и программное обеспечение вычислительных систем. – 2008. – С. 139-143.

3. Семченков С.Ю. Особенности применения OLAP систем: проблемы и актуальные подходы // Математическое и программное обеспечение вычислительных систем. – 2006. – С. 83-86.

4. Каширин И.Ю., Семченков С.Ю. Интерактивная аналитическая обработка данных в современных OLAP-системах // Журнал «Бизнес-информатика». Москва, 2009. – № 8(02). – С. 12-19.

5. Семченков С.Ю. Принципы реализации иерархической структуры измерений в OLAP системах // Математическое и программное обеспечение вычислительных систем. – 2007. – С. 49-57.

6. Семченков С.Ю. Вопросы организации детализированных и агрегированных данных в OLAP системе // Проблемы передачи и обработки информации в сетях и системах телекоммуникаций: материалы 15-й международной научно-технической конференции (Рязань, 13-15 февраля 2008 г.). – Рязань, 2008. – Ч. 2. С. 72-73.

7. Семченков С.Ю. Применение OLAP технологий в управлении качеством учебного процесса // Методы обучения и организация учебного процесса в вузе: материалы всероссийской научно-методической конференции (Рязань, 3-5 февраля 2009 г.). – Рязань, 2009. – С. 177-179.

8. Семченков С.Ю. Построение куба интересов пользователя в OLAP системах // Научная сессия ТУСУР-2009: материалы всероссийской научно-технической конференции студентов, аспирантов и молодых ученых (Томск, 12-15 мая 2009 г.). – Томск, 2009. – С. 219-222.

9. Семченков С.Ю. Вопросы моделирования интересов пользователя в OLAP системах // Информационные технологии в науке, проектировании и производстве: материалы XXVI всероссийской заочной научно-технической конференции (Нижний Новгород, апрель 2009 г.). – Нижний Новгород, 2009. – С. 1-2.

10. Горюнов И.В., Семченков С.Ю. Методология разработки систем информационной поддержки образовательного процесса в вузе на основе принципов всеобщего менеджмента качества (TQM) с использованием OLAP-технологии //Вестник РГРТУ. – 2008. – №4 (выпуск 26). – С. 69-74.

11. Семченков С.Ю. Операция декомпозиции многомерного куба в OLAP системах // Информационные технологии в науке, экономике и

образовании: материалы всероссийской научно-практической конференции (Бийск, 16-17 апреля 2009 г.). – Бийск, 2009. – С. 290-293.

12. Семченков С.Ю. Принципы построения регулярной структуры измерений в OLAP системах // Математическое и программное обеспечение вычислительных систем. –2009. – С. 136-140.

13. Семченков С.Ю. Применение OLAP сервера Palo для анализа данных // Новые информационные технологии в научных исследованиях и образовании: материалы XIII всероссийской научно-технической конференции студентов, молодых ученых и специалистов (Рязань, 14-16 мая 2008 г.). – Рязань, 2008. – С. 1-2.

14. Семченков С.Ю. CuDBIS v. 1.02. Свидетельство о регистрации программы для ЭВМ № 2009613357 от 26 июня 2009 г.

СЕМЧЕНКОВ Сергей Юрьевич

АЛГОРИТМЫ ПРОЕКТИРОВАНИЯ СИСТЕМ МНОГОМЕРНОГО АНАЛИЗА ДАННЫХ, ОСНОВАННЫХ НА OLAP ТЕХНОЛОГИИ

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать _____ . Формат бумаги 60x84 1/16.
Бумага офисная. Печать трафаретная. Усл. печ. л. 1,0.
Тираж 100 экз.
Рязанский государственный радиотехнический университет.
390005, г. Рязань, ул. Гагарина, д. 59/1.
Редакционно-издательский центр РГРТУ.