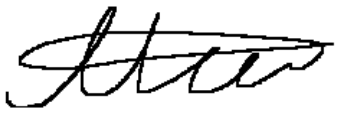
На правах рукописи 

Мосин Сергей Владимирович

МЕТОДЫ И АЛГОРИТМЫ ФОРМИРОВАНИЯ МНОГОМЕРНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПРОМЕЖУТОЧНЫХ ПРЕДСТАВЛЕНИЙ

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

Автореферат

диссертации на соискание ученой степени

кандидата физико-математических наук

Челябинск – 2017

Работа выполнена в ФГБУН «Институт математики им. С. Л. Соболева Сибирского отделения Российской академии наук» (Новосибирск)

Научный руководитель: **Зыкин Сергей Владимирович**
доктор технических наук, профессор,
ФГБУН «Институт математики им. С. Л. Соболева Сибирского отделения Российской академии наук»,
заведующий лабораторией Методов преобразования и представления информации

Официальные оппоненты: **Кузнецов Сергей Дмитриевич,**
доктор технических наук, профессор,
ФГБУН «Институт системного программирования Российской академии наук» (Москва),
главный научный сотрудник

Костенецкий Павел Сергеевич,
кандидат физико-математических наук, доцент,
ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)» (Челябинск),
Руководитель лаборатории «Суперкомпьютерное моделирование»

Ведущая организация: ФГБОУ ВО «Сибирский государственный университет телекоммуникаций и информатики» (Новосибирск)

Защита состоится 4 октября 2017 г. в 12:00 часов на заседании диссертационного совета Д 212.298.18 на базе ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)» по адресу: 454080, г. Челябинск, пр. Ленина, 76, ауд. 1001.

С диссертацией можно ознакомиться в библиотеке Южно-Уральского государственного университета и на сайте: http://susu.ac.ru/ru/dissertation/d-21229818/mosin-sergey-vladimirovich.

Автореферат разослан «___» _____ 2017 г.

Ученый секретарь
диссертационного совета М.Л. Цымблер

Общая характеристика работы

Актуальность темы. Базы данных играют значительную роль в выполнении задач хранения и обработки накопленной информации. Эти технологии развиваются уже порядка полувека и претерпели большое количество изменений. Перед многими организациями стоит серьезная проблема обработки и анализа данных с требованием минимальной задержки по времени. Сам факт наличия доступа к большим объемам информации не гарантирует возможности делать какие-либо выводы о закономерностях, скрытых в данных. Для решения этой задачи нужны особые методы представления и обработки информации.

Самым распространенным подходом к решению этой проблемы в настоящее время является технология оперативной аналитической обработки данных OLAP (online analytical processing). В основе этой технологии лежит построение многомерного (гиперкубического) представления данных. Выделяют 3 основных типа технологии OLAP по способу организации базы данных, лежащей в основе многомерной модели данных: Relational OLAP (ROLAP), Multidimensional OLAP (MOLAP), Hybrid OLAP (HOLAP).

Подход ROLAP базируется на использовании исходной реляционной СУБД на всем этапе формирования и хранения многомерной модели данных. Существенным недостатком такого подхода является ограничения на схему БД: она должна быть в форме «звезды» или «снежинки», что нарушает принцип независимости данных.

Многомерное представление данных, сформированное по технологии MOLAP, постоянно хранится и обновляется периодически из исходной базы данных. Преимуществом такого подхода является минимальный отклик системы на запросы пользователя, так как гиперкубическое представление оптимизировано специально для выполнения таких запросов. Очевидный недостаток заключается в дублировании данных. Информация должна храниться как в исходной базе данных, так и в многомерном представлении.

Наконец, технология HOLAP пытается совместить лучшие показатели из двух предыдущих подходов. Данные хранятся в исходных таблицах, а заранее просчитанные агрегированные величины записываются в многомерные таблицы. Преимущество заключается в достижении более высоких средних показателей скорости выполнения запросов, по сравнению с ROLAP, и меньшим объемом дублированных данных. Недостатками, соответственно, является на-

личие, пусть и меньшего чем в MOLAP, дублирования данных, а также более низкая средняя скорость выполнения запросов по сравнению с MOLAP.

Общим недостатком перечисленных подходов является отсутствие автоматизации процесса построения гиперкубического представления данных. Пользователю необходимо вручную осуществлять задание размерностей и мер гиперкуба, а также ограничений на размерности, что требует привлечения специалиста по OLAP. Этот процесс необходимо осуществлять каждый раз при формировании нового гиперкуба, что во многом определяет ограниченность использования технологии оперативной аналитической обработки данных.

В соответствие с этим актуальной является задача разработки автоматизированной системы формирования гиперкубического представления данных из исходного реляционного представления, позволяющей строить итоговое представление с минимальным участием пользователя. Данная система также должна использовать сохраненные в процессе формирования гиперкубов данные для выполнения построений следующих гиперкубов и проведения операций над данными.

Целью данной работы является разработка методов и алгоритмов формирования многомерного представления данных из реляционного представления при наложении логических ограничений на размерности и при использовании сохраненных данных. Для достижения этой цели были выполнены следующие **задачи**:

1. Исследовать и обосновать признаки выполнения свойства соединения без потерь информации для оптимизации алгоритма формирования контекстов.
2. Разработать способ формирования «Таблицы Соединений», осуществляющий построение, начиная с наименьших комбинаций отношений контекста и обосновать эквивалентность с существующим методом.
3. Разработать и обосновать методы аналитического определения возможности использования сохраненных данных.
4. Разработать алгоритмы использования сохраненных данных при формировании многомерных данных, а также при анализе данных.

5. Реализовать программное обеспечение, формирующее многомерное представление из исходного реляционного представления с использованием сохраненных данных.
6. Провести вычислительные эксперименты, подтверждающие эффективность предложенных подходов.

Научная новизна заключается в исследовании подходов к аналитическому сравнению областей истинности логических ограничений, разработке новых методов и оптимизированных алгоритмов формирования многомерных данных, разработке программного обеспечения для преобразования реляционной базы данных к многомерному представлению.

Теоретическая значимость. Исследованы свойства и методы сравнения областей истинности логических ограничений на основе логики предикатов и реляционной алгебры. Данные результаты могут быть использованы в дальнейшем при исследовании свойств логических формул реляционной алгебры и для усовершенствования систем хранения и анализа накопленной информации.

Практическая значимость заключается в разработке программного обеспечения, формирующего гиперкубическое представление из исходного реляционного представления. Данная система реализует теоретические принципы сравнения областей истинности логических формул при использовании сохраненных данных с целью снижения объема передаваемой информации и увеличения скорости работы. Проведено сравнение данной программы с существующими аналогами и выявлены сильные и слабые стороны всех систем.

Методология и методы исследования. Работа была выполнена с использованием методов межмодельных коммутативных преобразований, теории проектирования реляционных баз данных, логики предикатов, реляционной алгебры.

Степень достоверности результатов. Достоверность научных результатов, полученных в работе, подтверждается строгими математическими доказательствами. Теоретические построения подтверждены экспериментами, проведенными в соответствии с общепринятыми методиками.

Результаты работы могут быть использованы в научных исследованиях в области баз данных, а также при разработке прикладных программ, ставящих задачи переиспользования результатов реляционных запросов.

Апробация работы. Основные результаты работы докладывались на следующих конференциях и семинарах:

1. XVI Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL), 13 – 16 октября 2014 г., Дубна
2. VI Международная молодежная научно-практическая конференция с элементами научной школы «Прикладная математика и фундаментальная информатика» (ПМиФИ), 23 – 30 апреля 2015 г., Омск
3. Международная IEEE Сибирская конференция по управлению и связи (SIBCON), 21 – 23 мая 2015 г., Омск
4. VI Всероссийская научно-техническая конференция «Россия молодая - передовые технологии в промышленность!», 10 – 11 ноября 2015 г., Омск

Публикации. Основные результаты по теме диссертации изложены в 9 печатных изданиях, 4 из которых изданы в журналах, рекомендованных ВАК, 2 – в изданиях, индексированных в Scopus и Web of Science. В рамках выполнения диссертационной работы получено одно свидетельство Роспатента об официальной регистрации программ для ЭВМ и баз данных. В работе [7] Зыкину С.В. принадлежит постановка задачи, Полуянову А.Н. – исследование свойства существующего соединения, Мосину С.В. – исследование методов построения многомерных данных (стр. 401-407); в работе [6] Зыкину С.В. принадлежит постановка задачи, Полуянову А.Н. – разработка алгоритма проверки свойства существующего соединения, Мосину С.В. – разработка алгоритмов построения многомерных данных (стр. 976-985); в работе [3] Зыкину С.В. принадлежит постановка задачи, Полуянову А.Н. – доказательство достаточности условия существующего соединения, Мосину С.В. – анализ алгоритмов построения многомерных данных (стр. 122-128); в работах [1, 5, 9] Зыкину С.В. принадлежит постановка задачи, Мосину С.В. – все полученные результаты.

Объем и структура работы. Диссертация состоит из введения, трех глав, заключения и списка литературы. Объем диссертации **111** страниц текста. Список литературы содержит **109** наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава, «Подходы к реализации OLAP-технологии», посвящена анализу различных видов систем OLAP, а также обзору научных исследований в области современных подходов к реализации этой технологии. Особое внимание уделяется методам формирования и способам хранения многомерных данных, использованию сохраненных данных.

Вторая глава, «Автоматизация формирования гиперкуба», посвящена исследованию принципов формирования многомерного представления данных. Описываются и обосновываются оптимизированные алгоритмы построения контекстов и формирования представления данных «Таблица Соединений». Рассмотрены логические ограничения на размерности, приведены методы сравнения их областей истинности, а также алгоритмы повторного использования сохраненных данных.

Для автоматизации формирования гиперкуба предлагается использовать последовательность преобразований

$$RRD \Rightarrow TJ \Rightarrow GC,$$

где *RRD* – реляционное представление данных, *TJ* – «Таблица Соединений», *GC* – гиперкуб.

Рассмотрим формализацию задачи. Пусть задана схема базы данных $\Re_1 = \{R_1, R_2, \ldots, R_k\}$, полученная в результате нормализации отношений. Отношения R_i определены на множестве атрибутов $U = \{A_1, A_2, \ldots, A_k\}$. Пусть $\langle R_i \rangle$ – схема отношения, множество атрибутов, на которых определено отношение R_i. Предположим, что схема $\Re$ является редуцированной, то есть не существует двух отношений, таких, что $\langle R_i \rangle \subseteq \langle R_j \rangle$ при $i \neq j$. Кортеж $t[X]$ – совокупность значений атрибутов $A_j \in X \subseteq \langle R_i \rangle$, заданных в кортеже $t \in R_i$. Неопределенное значение *NULL* атрибута A_j в кортеже t не равно любому другому значению, в том числе другому неопределенному значению.

Многомерное представление будем задавать в виде совокупности размерностей: $\{D_1, D_2, \ldots, D_k\}$, где D_l – множество расширенных имен атрибутов: $R_i.A_j$, $A_j \in \langle R_i \rangle$, M – множество мер, также заданных в виде расширенных имен атрибутов. Для каждой размерности задается ограничение в виде логической формулы F_l.

Пусть *DEP* – множество зависимостей (функциональных, многозначных, включения, соединения), $C = R_1, R_2, \ldots, R_q$ – произвольное подмножество отношений реляционной БД.

***Определение* 1.** *Множество C будем называть контекстом, если оно удовлетворяет свойству СБПИ на зависимостях DEP, реализованных в C.*

***Теорема* 1.** *Множество отношений C обладает свойством СБПИ, если существует отношение $R_i \in C$, замыкание первичного ключа которого совпадает со всем множеством атрибутов отношений множества C.*

Пусть $m = \{R_1, R_2, \ldots, R_q\}$ – произвольное множество отношений и $\langle C_m \rangle = \langle R_1 \rangle \cup \langle R_2 \rangle \cup \cdots \cup \langle R_q \rangle$.

***Теорема* 2.** *Множество отношений $C_{m+1} = \{R_1, R_2, \ldots, R_q, R_{q+1}\}$ не обладает свойством СБПИ на DEP, если зависимость $Z \twoheadrightarrow X(Y)$ не выводима из DEP, где $X \subseteq \langle C_m \rangle$, $Y \subseteq \langle R_{q+1} \rangle$ и $\langle C_m \rangle \cap \langle R_{q+1} \rangle \subseteq Z$.*

Совокупность отношений, по которым строится гиперкуб, должна удовлетворять свойству СБПИ. Следовательно, дальнейшая задача состоит в дополнении множества C^0 отношениями из $\Re$, чтобы результирующее множество отношений удовлетворяло свойству СБПИ на множестве зависимостей. Задача алгоритма заключается в последовательной генерации контекстов без зацикливания. Для сокращения количества перебираемых вариантов предлагается сделать этот перебор направленным.

***Определение* 2.** *Контекстом приложения называется контекст, сформированный на основе базового множества отношений C^0.*

Логические формулы, по которым проводится селекция, будем рассматривать в дизъюнктивной нормальной форме. В общем случае формула F имеет вид

$$F = K_1 \vee K_2 \vee \cdots \vee K_l, \quad (1)$$

$$K_i = T_1^i \,\&\, T_2^i \ldots \&\, T_p^i(i), i = 1, \ldots, l, \quad (2)$$

где $T_j^i, i = 1, \ldots, l, j = 1, \ldots, p(i)$ - предикаты, в которых явным образом специфицированы расширенные имена атрибутов $R_i.A_j$ (атрибут A_j в отношении R_i)

Обозначение. *Множество атрибутов, входящих в формулу, выражает размерность формулы и обозначается* $\langle F \rangle$.

$\langle F \rangle = \{R_1^F.A_1^F, \ldots, R_k^F.A_k^F\}, k -$ *количество атрибутов, входящих в формулу* *(3)*

Введем в рассмотрение множество $\mathcal{A} = \{(a_1, \ldots, a_n) \mid a_i \in Dom(A_i), i = 1, \ldots, n\}$, где $Dom(A_i)$ - множество всех допустимых значений атрибута A_i.

***Определение* 3.** *Областью истинности логической формулы* F*, заданной (1), (2), (3), является множество, определяемое по следующему правилу:* $M(F) = \{a \in \mathcal{A} \mid F(a) = TRUE\}$.

***Определение* 4.** *Проекцией логической формулы* F*, заданной (1), (2), (3), на множество атрибутов* X *называется логическая формула* $F[X], \langle F[X] \rangle = X$*, в которой все термы, содержащие атрибуты* $R_i^F.A_i^F \notin X$*, заменены на тривиальный терм TRUE.*

Рассмотрим преобразование реляционной БД: $\Re = \{R_1, R_2, \ldots, R_k\}$ в «Таблицу Соединений» *TJ* со схемой (S,g), где S – схема, определенная на множестве атрибутов $U = \{A_1, A_2, \ldots, A_n\}$, g – вектор вхождения кортежей отношений длины k. Определим принцип формирования кортежей $t \in s$, где s – реализация (множество кортежей) таблицы *TJ*. Рассмотрим все возможные сочетания без повторений отношений $R_1, R_2, \ldots, R_k$, удовлетворяющие свойству СБПИ. Пусть множество отношений $C' = \{R_{mas(1)}, R_{mas(2)}, \ldots, R_{mas(p)}\}$ – контекст, где *mas* – целочисленный массив из p номеров отношений текущего сочетания, и c' – его реализация, ограниченная операцией селекции σ_F с логической формулой F:

$$c' = \sigma_F(R_{mas(1)}[W_{mas(1)}] \bowtie R_{mas(2)}[W_{mas(2)}] \bowtie \cdots \bowtie R_{mas(p)}[W_{mas(p)}]), \quad (4)$$

где $V_{mas(i)}$, p – количество отношений в контексте, V_i – множество атрибутов $A_j \in \langle R'_i \rangle$, для которых выполнено: либо существует размерность D_l такая, что $R'_i.A_j \in D_l$, либо $R'_i.A_j \in M$, либо $R'_i.A_j \in KM$, либо существует $R'_v \in C_0$ такое, что $A_j \in \langle R'_v \rangle$ и $i \neq v$, либо существует логическая формула F_l и $A_j \in \langle F_l \rangle$, $V_{mas(i)} \subseteq W_{mas(i)}, i = 1, 2, \ldots, p$.

Для каждого кортежа $u \in c'$ формируем кортеж t по следующим правилам: $t[A_j] = u[A_j]$, если атрибут A_j принадлежит хотя бы одному отношению соединения, и $t[A_j] =$ *emp* в противном случае, где *emp* – пустое значение. Каждому кортежу поставим в соответствие битовый вектор $g(t) = (g_1(t), g_2(t), \ldots, g_k(t))$, где $g_j(t) = 1$, если отношение R_j участвует в текущем соединении, и $g_j(t) = 0$ в противном случае.

Рассмотрим отношение частичного порядка над кортежами $t \in s$.

***Определение* 5.** *Кортеж $t \in s$ является менее определенным или равным кортежу $t' \in s$, когда для любого атрибута A_i выполнено условие: если $t[A_i] \neq t'[A_i]$, то $t[A_i] =$ emp и $g_j(t') \geq g_j(t), j = 1, \ldots, k$, причем $t[A_i] = t'[A_i]$, если A_i принимает значение NULL в обоих кортежах. В этом случае будем писать $t \prec t'$, назовем кортеж t подчиненным кортежу t' и оба этих кортежа будем считать сравнимыми.*

Пусть $R_(L) = \{R_{mas(1)}, R_{mas(2)}, \ldots, R_{mas(p)}\}$, где $L = (mas(1), mas(2), \ldots, mas(p))$, и *mas* – целочисленный массив номеров p отношений. Определим операцию проекции на множестве s.

***Определение* 6.** *Проекция $\pi_{R(L)}(s)$ есть совокупность кортежей $u[R(L)]$, определенных на множестве всех атрибутов отношений $R(L)$, где для каждого $u[R(L)]$ существует кортеж $t \in s$, такой, что $u[R(L)] = t[R(L)]$ и $g_{mas(i)}(t) = 1, i = 1, 2, \ldots, p$.*

Логическое ограничение $F(t)$ для каждой размерности будем представлять в виде 1, 2. Если какой-либо предикат T_j^i не определен на кортеже t, то он аннулируется – заменяется значением *TRUE*, если в этом конъюнкте еще есть не аннулированные предикаты, в противном случае – *FALSE*. Такая подстановка позволяет оставить в s кортежи, для которых пока не определены некоторые атрибуты или отсутствуют связанные по значениям кортежи в других отношениях, что также является предметом анализа информации. Формула F после подстановки будет принимать только два значения: *TRUE* и *FALSE*.

***Теорема* 3.** *Представление s всегда существует и единственно независимо от порядка удаления подчиненных кортежей.*

***Теорема* 4.** *Для любого множества отношений $R^* = \{R_1^*, R_2^*, \ldots, R_q^*\} \subseteq C'$, удовлетворяющего свойству СБПИ, где C' – контекст и s – таблица соединения, соответствующая C', выполнено:*

$$\pi_{R^*(L)}(s)[Z] = \sigma_F(R_1^*[Z_1] \bowtie R_2^*[Z_2] \bowtie \cdots \bowtie R_q^*[Z_q]),$$

где $V_i \subseteq Z_i \subseteq W_i$, $Z = Z_1 \cup Z_2 \cup \cdots \cup Z_q$, L – вектор номеров отношений в R^*.

Рассмотрим использование рассмотренных таблиц для формирования результирующего представления данных *GC*, используемого при проведении различных видов многомерного анализа. Совокупность таблиц данных $T_i, i = 0, 1, 2, \ldots, d$, определенных на множествах атрибутов D_i соответственно. Таблица T_0 соответствует контексту приложения и определена на множестве атрибутов $D_0 = D_1 \cup D_2 \cup \cdots \cup D_d \cup M$. При этом:

$$T_0 = s_0[D_0], \tag{5}$$

где s_0 – таблица соединения для контекста приложения C_0

Для формирования каждой размерности ($1 \leq i \leq d$) в зависимости от потребностей пользователя используется одна из трех последующих формул:

1. $T_i = s_i[D_i]$, где s_i – таблица соединения для контекста C_i
2. $T_i = \pi_R^*(L)(s_0)[D_i]$, где $C_i = \{R_1', R_2', \ldots, R_q'\}$ – пустой контекст, L – вектор номеров отношений пустого контекста.
3. $T_i = \sigma_{F_i}(R_1'[W_1] \bowtie R_2'[W_2] \bowtie \cdots \bowtie R_p'[W_p])[D_i]$, если $C_i = R_1', R_2', \ldots, R_p'$ – псевдоконтекст, $0 < i \leq d$, F_i –логическое ограничение на кортежи, остальные обозначения и ограничения совпадают с формулой 4.

Введем определения, позволяющие различать дублирующиеся значения мер от повторяющихся различных значений.

***Определение* 7.** *Множество атрибутов KM_j будем называть ключом атрибута меры $R_i.A_j \in M$ в «Таблице Соединений» TJ, если $KM_j \subseteq \langle TJ \rangle$, зависимость $KM_j \to R_i.A_j$ выводима на множестве функциональных зависимостей,*

и не существует выводимой зависимости $Y \to R_i.A_j$, где $Y \subset KMj$. Пусть $KM = KM_1 \cup KM_2 \cup \cdots \cup KM_h$, где $h = |M|$, общий ключ для всех мер гиперкуба.

***Определение* 8.** *Значение атрибута меры $t[A_j]$, где $A_j \in M$, для текущего кортежа $t \in s_0$ дублирует значение $t'[A_j], t' \in s_0$, если:*

1. $t[A_j] = t'[A_j]$
2. $t[D_i] = t'[D_i], i = 1, 2, \ldots, d$
3. $t[KMj] = t'[KMj]$

Сохраненные результаты запросов обозначим $P = \{P_1, P_2, \ldots, P_m\}$, где $P_v = \pi_{X_v}(\sigma_{F_v}(R_1^v \bowtie R_2^v \bowtie \cdots \bowtie R_{s(v)}^v))$, $s(v)$ – количество отношений в базе данных, использованных при формировании представления P_v, π_{X_v} - операция проекции по множеству атрибутов X_v, σ_{F_v} - операция селекции с логическим ограничением на кортежи F_v. Целевое выражение, которое надо будет получить из представлений P, запишем в виде:

$$P^* = \pi_{X^*}(\sigma_{F^*}(R_1^* \bowtie R_2^* \bowtie \cdots \bowtie R_l^*)$$

***Теорема* 5.** $P^* \subseteq \pi_{X^*}(\sigma_{F^*[X_v]}(P_v))$ *, если:*
а) $X^* \subseteq X_v$
б) $\{R_1^v, \ldots, R_{s(v)}^v\} \subseteq \{R_1^*, \ldots, R_l^*\}$
в) $M(F^*) \subseteq M(F_v)$.

***Теорема* 6.** $P^* = \pi_{X^*}(P_v)$ *, если:*
а) $X^* \subseteq X_v$
б) $\{R_1^v, \ldots, R_{s(v)}^v\} = \{R_1^*, \ldots, R_l^*\}$
в) $M(F^*) = M(F_v)$.

***Теорема* 7.** $P^* \subseteq \pi_{X^*}(\sigma_{F^*[X]}(P_1 \bowtie \cdots \bowtie P_n))$, *где* $X = \bigcup_{v=1}^{n} X_v$ *если:*
а) $X^* \subseteq X$
б) $\bigcup_{v=1}^{n} \{R_1^v, \ldots, R_{s(v)}^v\} = \{R_1', \ldots, R_{s'}'\} \subseteq \{R_1^*, \ldots, R_l^*\}$
в) $M(F^*) \subseteq M(F_v), v = 1, \ldots, n$.

***Теорема* 8.** $P^* = \pi_{X^*}(P_1 \bowtie \cdots \bowtie P_n)$, *где* $X = \bigcup_{v=1}^{n} X_v$ *если:*
а) $X^* \subseteq X$, $X_v \supseteq \langle \bowtie_{i=1}^{s(v)} R_i^v \rangle \cap (\bigcup_{\substack{w=1 \\ w \neq v}}^{n} \langle \bowtie_{i=1}^{s(w)} R_i^w \rangle), v = 1, \ldots, n$

б) $\bigcup_{v=1}^{n}\{R_1^v,\ldots,R_{s(v)}^v\}=\{R'_1,\ldots,R'_{s'}\}=\{R_1^*,\ldots,R_l^*\}$
в) $M(F^*)=M(F_1\&\ldots\&F_n)$.

Аналитическое сравнение формул позволяет определить возможность использования приведенных выше теорем 5, 6, 7, 8, не прибегая к запросам на сервер базы данных. В работе приводятся возможные случаи, когда такое сравнение может быть выполнено аналитически.

В публикациях рассматриваются различные подходы к получению оценки мощности результата операции естественного соединения.

Пусть $|R_i| = N_i, i = 1,\ldots,M$ – мощности отношений, $A(C(m)) = \{A_1, A_2, \ldots, A_p\}$ – множество атрибутов, принадлежащих, по крайней мере, одному пересечению $R_i \cap R_j$, где $i,j \in C(m)$. $C(m) = \{j_1,j_2,...,j_m\}$ – совокупность номеров отношений, являющаяся сочетанием без повторений по m элементов из множества $I = \{1,2,...,M\}$, $|S(C(m))||$ – мощность естественного соединения отношений $R_i, i \in C(m)$. $k(A_j,C(m))$ – количество различных значений атрибута A_j, являющихся общими для всех отношений R_i, $i \in C(m)$, которые содержат атрибут A_j. В частности, $k(A_j,i)$ – количество различных значений атрибута A_j только в реализации R_i. $R_i \cap A(C(m)) = X_i(C(m))$. Тогда оценка мощности соединения отношений с индексами $C(m)$ принимает следующий вид:

$$|S(C(m))| = \frac{\prod_{j=1}^{p} k(A_j,C(m)) \prod_{i\in C(m)} N_i}{\prod_{i\in C(m)} \prod_{A_j\in X_i(C(m))} k(A_j,i)}. \qquad (6)$$

Далее эта оценка улучшена для случая с логическими ограничениями.

В третьей главе, «Реализация программного обеспечения системы», описывается архитектура и принципы работы разработанного программного обеспечения, осуществляющего преобразование исходной реляционной модели данных в целевое гиперкубическое представление.

Далее приводится обзор наиболее популярных на сегодняшний день решений по построению и работе с гиперкубическими представлениями данных. В качестве таких систем были выбраны продукты от Microsoft и Oracle. Проведены сравнительные испытания, демонстрирующие прирост производитель-

ности при использовании предлагаемого в данной работе программного обеспечения в сравнении с упомянутыми аналогами.

В **заключении** перечисляются результаты, полученные в итоге выполнения исследования, проводится сравнение с наиболее близкими работами по данной тематике, даются рекомендации по использованию разработанной технологии и программного обеспечения, рассматриваются направления дальнейших исследований.

Основные результаты диссертационной работы

1. Разработан и аналитически исследован алгоритм направленного перебора для формирования контекстов.
2. Разработан оптимизированный алгоритм формирования представления данных «Таблица Соединений».
3. Предложен и исследован оригинальный метод сравнения областей истинности логических ограничений при анализе сохраненных (кэшированных) данных.
4. Разработаны алгоритмы повторного использования сохраненных данных и вычисления недостающих данных на основе сравнения областей истинности.
5. Реализовано программное обеспечение, формирующее многомерное представление из исходного реляционного представления с использованием сохраненных данных.
6. Проведены вычислительные эксперименты, подтверждающие эффективность предложенных подходов.

Публикации по теме диссертации

Публикации из перечня ВАК

1. *Mosin S.*, *Zykin S.* Truth space method for caching database queries // Modeling and Analysis of Information Systems. 2015. Т. 22, № 2. С. 248–258. DOI: 10.18255/1818-1015-2015-2-248-258. индексирована в MathSciNet.

2. *Мосин С.В.* Сравнение областей истинности запросов к реляционной базе данных // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2016. Т. 5, № 1. С. 85–99. DOI: 10.14529/cmse160108.
3. *Зыкин С.В.*, *Мосин С.В.*, *Полуянов А.Н.* Технология раздельного формирования многомерных данных // Вестник Донского государственного технического университета. 2016. Т. 85, № 2. С. 114–129. DOI: 10.12737/19696.
4. *Мосин С.В.* Алгоритм использования кэша запросов к реляционной базе данных // Вестник СибГУТИ. 2017. № 1. С. 47–57.

Публикации в изданиях, индексируемых в Scopus и Web of Science

5. *Mosin S.V.*, *Zykin S.V.* Using logical formulas for caching uniform RDB queries // 2015 International Siberian Conference on Control and Communications (SIBCON), Omsk, May 21–23, 2015. 2015. С. 852–857. DOI: 10.1109/sibcon.2015.7147151.
6. *Zykin S.*, *Mosin S.*, *Poluyanov A.* Technology of separate formation of multidimensional data with lists of measure values // 2015 International Siberian Conference on Control and Communications (SIBCON), Omsk, May 21–23, 2015. 2015. С. 975–986. DOI: 10.1109/sibcon.2015.7147176.

Статьи в изданиях, индексируемых в РИНЦ

7. *Зыкин С.В.*, *Мосин С.В.*, *Полуянов А.Н.* Технология формирования многомерных данных // Труды RCDL 2014 (XVI Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL), Омск, 13–16 Октября, 2014 г.). 2014. С. 399–408.
8. *Мосин С.В.* Использование логических формул для кэширования универсальных запросов к реляционной базе данных // Прикладная математика и фундаментальная информатика (VI Международная молодежная научно-практическая конференция с элементами научной школы «Прикладная математика и фундаментальная информатика» (ПМиФИ), Омск, 23–30 Апреля, 2015 г.). 2015. С. 240–249.

9. *Мосин С.В.*, *Зыкин С.В.* Аналитическая обработка кэшированных данных // Россия молодая: передовые технологии – в промышленность (VI Всероссийская научно-техническая конференция «Россия молодая – передовые технологии в промышленность!», Омск, 10–11 Ноября, 2015 г.). 2015. С. 26–30.

Свидетельства о регистрации программ и баз данных

10. *Мосин С.В.* Свидетельство Роспатента об официальной регистрации программы для ЭВМ «PyRO» № 2017613344 от 15.03.2017, правообладатель: Мосин Сергей Владимирович.

Работа выполнялась при финансовой поддержке Отделения математических наук Российской академии наук, грант № II.2.П/I.1-7, а также при поддержке гранта РФФИ № 12-07-00066-а.