На правах рукописи

Бадмаева Ксения Владимировна

МЕТОДИКА АДАПТИВНОГО УПРАВЛЕНИЯ ПРОЦЕССОМ ПРОЕКТИРОВАНИЯ И РАЗВИТИЯ СПЕЦИАЛИЗИРОВАННЫХ ХРАНИЛИЩ ДАННЫХ

05.13.01 – Системный анализ, управление и обработка информации
(информатика, вычислительная техника и управление)

А В Т О Р Е Ф Е Р А Т
диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2009

Работа выполнена в Институте вычислительного моделирования СО РАН, г. Красноярск

Научный руководитель: кандидат технических наук
Исаева Ольга Сергеевна

Официальные оппоненты: доктор технических наук, профессор
Ковалев Игорь Владимирович

кандидат физико-математических наук
Цыганок Дмитрий Алексеевич

Ведущая организация: Институт систем энергетики
им. Л.А. Мелентьева СО РАН, г. Иркутск

Защита состоится 2 октября 2009 года в 14-00 на заседании диссертационного совета ДМ 212.099.06 при Сибирском федеральном университете по адресу: г. Красноярск, ул. Киренского, 26, УЛК 115.

С диссертацией можно ознакомиться в научной библиотеке Сибирского федерального университета по адресу: г. Красноярск, ул. Киренского, 26, Г 274.

Автореферат разослан 1 сентября 2009 года.

Ученый секретарь
диссертационного совета Р.Ю. Царев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования

Необходимость оперативной аналитической обработки информации в задачах организационного управления в разных прикладных областях – в здравоохранении, социальной, экономической сферах и других – требует эффективной организации больших объемов данных, поступающих из разнородных источников. Проблемы согласованности данных, оперативности выполнения запросов и обеспечения доступа к информации могут быть решены с использованием технологии хранилищ данных (data warehouses). Чтобы повысить производительность и обеспечить эффективное применение технологии OLAP (On-Line Analytical Processing) требуется решать задачи оптимизации и развития хранилищ данных, учитывая изменяющиеся внешние условия.

Разработке и внедрению хранилищ данных (далее также «хранилищ») посвящены работы зарубежных ученых Б. Инмона, Р. Кимбалла, Э. Спирли, С. Риззи, Л. Кабиббо, Р. Торлоне, М. Голфарелли, Д. Грея, Д. Теодоратоса и др. Из отечественных исследований на эту тему следует отметить работы А.А. Сахарова и подход Л.В. Массель, заключающийся в расширенном применении репозитория хранилища данных в рамках построения ИТ-инфраструктуры системных исследований. Несмотря на обширную методологическую базу, существует необходимость создания формализованных методов и алгоритмов, обеспечивающих корректировку модели данных при изменении условий эксплуатации с целью повышения производительности с сохранением актуальности хранилища. Сложность создания таких методов и алгоритмов обуславливается тем, что для задач разработки хранилищ данных характерна сильная зависимость процесса проектирования и его результата от особенностей конкретной предметной области и опыта проектировщика.

Значительного повышения производительности можно добиться, сохраняя в хранилище обобщенные данные – материализованные представления (materialize views). Выбор данных для материализации определяется интуитивно или после учета статистики, собранной сервером хранилища данных в процессе работы пользователей с аналитической системой. Для обоснованного включения материализованных представлений в хранилище данных при проектировании и развитии не достаточно существующих методов оптимизации модели данных. Актуальна и востребована разработка методов, позволяющих формализовать выбор материализованных представлений для повышения производительности хранилища данных на всех стадиях его существования, включая самые ранние стадии эксплуатации и развития.

Объект исследования: специализированные хранилища данных.

Предмет исследования: методическое обеспечение процессов проектирования, эксплуатации и развития специализированных хранилищ данных.

Цель диссертационного исследования: повышение производительности специализированных хранилищ данных за счет создания методики адаптивного управления процессом проектирования и развития хранилищ данных, позволяющей реагировать на изменение условий эксплуатации.

Для достижения поставленной цели решены следующие **задачи:**

- исследование существующих методов разработки хранилищ данных и алгоритмов материализации представлений, обоснование актуальности и целесообразности разработки новых методов и алгоритмов;
- разработка методики адаптивного управления процессом проектирования и развития специализированных хранилищ данных на основе известных и предлагаемых методов и алгоритмов;
- разработка метода построения модели общей стоимости материализации представлений на основе информации о предметной области;
- разработка алгоритма выбора релевантных представлений, включающего определение релевантных элементов множества представлений и выбор представлений для материализации;
- программная реализация хранилища медико-демографических данных и информационной системы «Анализ медико-демографических процессов», проектирование модели данных хранилища социально-экономических показателей развития региона, построение банка данных интернет-системы спортивно-оздоровительного движения «Игры народов планеты».

Методы исследования, применяемые в работе, основаны на теории адаптивных систем, методах системного анализа, теории принятия решений и технологии хранилищ данных.

Новые научные результаты и положения, выносимые на защиту

1. Предложенная методика адаптивного управления процессом проектирования и развития хранилищ данных обеспечивает оригинальную формализацию процессов проектирования и развития модели данных, учитывая условия эксплуатации и специализированную информацию о предметной области, что позволяет повысить производительность хранилищ данных.

2. Новый метод формирования модели общей стоимости материализации обеспечивает оценку релевантности представлений на основе информации о предметной области.

3. Оригинальный алгоритм выбора релевантных представлений для повышения производительности хранилища данных позволяет принимать решения о материализации при отсутствии статистической информации о работе хранилища.

Теоретическая значимость

Результаты, полученные при выполнении диссертационной работы, имеют существенное значение для проектирования хранилищ данных и позволяют повышать их производительность за счет учета условий эксплуатации. В отличие от других подходов к проектированию хранилищ данных, предложенные методы и алгоритмы обеспечивают корректировку модели данных даже при отсутствии накопленной статистической информации о работе хранилища.

Практическая значимость

Практическим результатом диссертационной работы является информационно-аналитическая система «Анализ медико-демографических процессов», построенная на основе специализированного хранилища данных. Система внедрена в промышленную эксплуатацию в Красноярском краевом медицинском информационно-аналитическом центре, что подтверждается актом о внедрении. Разработанные подходы и алгоритмы могут применяться для широкого круга предметных областей по отдельности, комплексно или в сочетании с различными существующими алгоритмами поиска представлений, алгоритмами вычисления общей стоимости материализации и методами проектирования хранилищ данных.

Исследование выполнено в соответствии с планами научно-исследовательских работ Института вычислительного моделирования СО РАН по программе СО РАН – проект 4.3.1.4 «Гибридные информационно-аналитические методы, системы и технологии (№ гос. регистрации 01.2.007 09439); по программе фундаментальных исследований Президиума РАН – проект № 14.7 «Методы и средства OLAP-моделирования»; по грантам Президента для ведущих научных школ № НШ-3428.2006.9, № НШ-3431.2008.9 и гранту РФФИ № 05-07-90244-в – «Развитие технологий хранилищ данных и оперативной аналитической обработки (OLAP) в задачах здравоохранения».

Достоверность и обоснованность полученных результатов обеспечивается корректно проведенным анализом и применением средств тех-

нологии хранилищ данных и теории принятия решений, анализом литературы и существующих разработок, обоснованием постановки задач диссертационной работы, результатами успешного применения предложенных методов и алгоритмов, а также результатами анализа эффективности применения предложенных подходов.

Личный вклад автора

Все результаты и положения, выносимые на защиту, получены непосредственно автором.

Апробация работы

Основные результаты диссертационной работы, а также результаты конкретных прикладных исследований представлены на VI Всероссийской конференции с международным участием «Новые информационные технологии в исследовании сложных структур» (Шушенское, 2006), на Международной научно-практической конференции «Общественное здоровье: инновации в экономике, управлении и правовые вопросы здравоохранения» (Новосибирск, 2005), на IX и X Всероссийской научно-практической конференции «Проблемы информатизации региона» (Красноярск, 2005, 2007), на конференции молодых ученых ИВМ СО РАН (Красноярск, 2005, 2006, 2007), на IV и V Всероссийской конференции «Всесибирский конгресс женщин-математиков» (Красноярск, 2006, 2008), на XIV Байкальской Всероссийской конференции «Информационные и математические технологии в науке и управлении» (2009).

Публикации

По результатам диссертационного исследования опубликовано 14 работ, в том числе 4 статьи в журналах из списка изданий, рекомендуемых ВАК для публикации результатов кандидатских и докторских диссертаций.

Структура и объем работы

Диссертация состоит из введения, трех глав, заключения и списка использованных источников. Основное содержание работы изложено на 148 страницах текста, содержит 26 рисунков и 12 таблиц. Список использованных источников включает 127 наименований, в том числе 74 иностранные публикации и 28 Internet-источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** показана актуальность разработки методов и алгоритмов, выполняющих корректировку модели хранилища данных для обеспечения приемлемой скорости работы пользователей с информацией на всех этапах эксплуатации хранилища. Методы формирования модели данных должны быть применимы на этапе проектирования при отсутст-

вии накопленной статистики о работе хранилища и в процессе его развития при поступлении новых данных предметной области или дополнительного дискового пространства.

Определены цели и задачи диссертационной работы, раскрыта новизна, практическая значимость и апробация результатов диссертации. Сформулированы положения, выдвигаемые на защиту, определен непосредственный вклад автора в решаемые задачи.

В **главе 1** представлены результаты исследования задач проектирования хранилищ данных, рассмотрены существующие подходы к проектированию хранилищ и алгоритмы выбора представлений для материализации (Б. Инмон, Р. Кимбалл, Э. Спирли, С. Риззи, Л. Кабиббо, Р. Торлоне, М. Голфарелли, Д. Грей, Д. Теодоратос, А. А. Сахаров и др.).

Показано, что повышение производительности хранилищ данных может быть достигнуто за счет сохранения в модели данных предварительных вычислений материализованных представлений. Обоснованное включение материализованных представлений при проектировании и эксплуатации повышает эффективность работы хранилища.

Проанализированы преимущества и ограничения существующих алгоритмов выбора представлений для материализации. Ограничением применимости существующих алгоритмов является использование статистической информации о работе хранилища данных, которая накапливается в процессе эксплуатации и отсутствует на ранних этапах проектирования. В этой связи особую значимость и актуальность приобретает разработка методов выбора представлений для материализации при отсутствии накопленной статистики. За счет возможности применения таких алгоритмов при проектировании и в процессе развития хранилища, достигается желаемая скорость получения информации на всех этапах эксплуатации хранилища данных.

Проведенный анализ литературы и программных разработок показал актуальность исследования и позволил обосновать задачи диссертации.

В **главе 2** представлена методика адаптивного управления процессом проектирования и развития специализированных хранилищ данных, ориентированная на достижение эффективной работы с данными при изменении условий эксплуатации. Предложенная методика разработана на основе применения принципа адаптации и методов системного анализа. Методика обеспечивает комплексное решение задач проектирования специализированных хранилищ данных и состоит из 3 основных этапов.

Этап 1. Исследование предметной области: выделение измерений, показателей и методик расчета. Определение источников данных, частоты их обновления. Изучение структуры исходных данных для после-

дующей разработки методов унификации и верификации данных. Определение релевантных данных с помощью методик расчета показателей предметной области.

Этап 2. Формирование модели хранилища данных: создание таблиц фактов и таблиц измерений. Включение в модель хранилища данных дополнительных структур, используя алгоритм выбора релевантных представлений. Формирование репозитория метаданных.

Этап 3. Эксплуатация хранилища данных и накопление статистики для последующего его развития с использованием алгоритма выбора релевантных представлений и рекомендаций автоматических средств системы управления базами данных (СУБД).

Методика адаптивного управления основана на использовании алгоритма выбора релевантных представлений (алгоритм релевантности). Предложенный алгоритм включает определение релевантных элементов множества представлений и выбор представлений для материализации.

Функциональная модель процесса адаптивного управления процессом проектирования и развития специализированного хранилища данных представлена на рисунке 1.

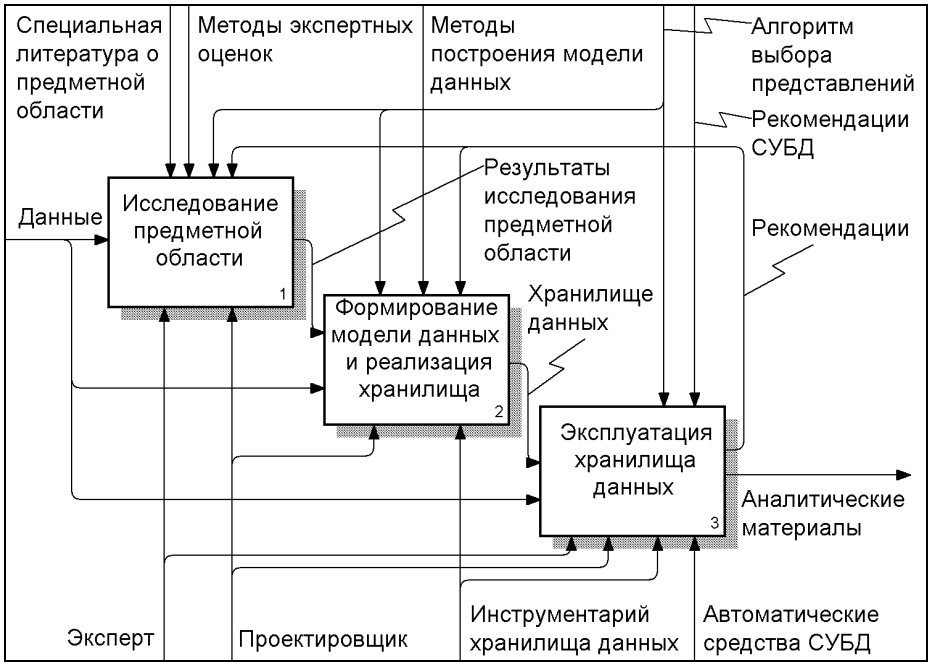

Рисунок 1 – Процесс адаптивного управления процессом проектирования и развития специализированных хранилищ данных

Модель описывает основные этапы проектирования и роли специалистов, участвующих в процессе. «Проектировщик» объединяет роли программиста, администратора хранилища и специалиста, проводящего исследование предметной области. «Эксперт» выступает в роли специалиста предметной области и/или пользователя хранилища данных. Спецификация управления процессом проектирования и развития специализированных хранилищ данных представлена в нотации таблицы переходов состояний.

В процессе эксплуатации хранилища данных автоматическими средствами СУБД по полям таблиц хранилища собираются и обрабатываются статистические сведения, на основе которых формируются рекомендации по внесению изменений в модель данных. При поступлении новой информации предметной области или дополнительного дискового пространства, в случае отсутствия рекомендаций СУБД, решение об изменении модели хранилища данных принимается на основе выполнения алгоритма релевантности.

Для реализации алгоритма релевантности в работе предложен метод формирования модели общей стоимости материализации представлений в хранилище данных. Модель хранилища данных можно представить в виде многомерной модели данных и множества материализованных представлений: $MDW = < MDDB, MV >$, где $MDDB = < D, F >$ – многомерная модель данных, D – множество таблиц измерений, F – множество таблиц фактов; $MV = \{ V_i \}$ – множество материализованных представлений, $i = \overline{1, s}$, s – количество материализованных представлений. Каждому представлению соответствует множество формирующих его атрибутов: $V_t = \{ A_t \}$, $A_t \subseteq A$, $t = \overline{1, s}$, A – множество атрибутов таблиц измерений и фактов.

Критерием оптимальности хранилища данных с дополнительными структурами является стоимость материализации представлений, оценка которой впервые предложена Харинараяном в 1996. В общем случае стоимость материализации множества представлений MV складывается из стоимости обслуживания и стоимости выполнения запросов и формируется, используя информацию о рабочей нагрузке:

$$Cost(MV) = c_1 \cdot CQ + c_2 \cdot CM \ , \qquad (1)$$

где c_1, c_2 – весовые коэффициенты; CQ – стоимость выполнения запросов на множестве материализованных представлений MV; CM – стоимость обслуживания материализованных представлений.

Вместо элемента CQ из (1), который не может быть вычислен на начальном этапе проектирования из-за отсутствия статистики о работе хранилища данных предлагается использовать элемент CV, рассчитываемый по формуле:

$$CV(MV) = \sum_{i=1}^{N} fm(V_i)Ca(V_i)\,, \qquad (2)$$

где $fm(V_i) = \sum_{j \in A_i} fm_j(a_i) / |A_i|$ – частота использования результатов V_i представления; A – множество всех атрибутов таблиц измерений и фактов; $A_i \subseteq A$ – множество атрибутов входящих в V_i представление; $|A_i|$ – количество атрибутов в V_i представлении; $fm_j(a_i) = \frac{|Ma_i|}{N}$ – частота встречаемости атрибута a_i, где $|Ma_i|$ – количество методик расчета показателей предметной области, в которых встречается a_i, N – количество рассматриваемых расчетных методик; $Ca(V_i)$ – стоимость формирования V_i представления, необходимого для реализации расчетной методики M_i, используя материализованные представления или исходные данные.

Предложенная формула (2) обеспечивает возможность вычисления общей стоимости материализации представлений в процессе проектирования и развития хранилища данных при отсутствии накопленной статистической информации о его работе. Для решения задачи повышения производительности хранилища данных, предложены алгоритмы определения и выбора релевантных представлений, оценивающие стоимость материализации разработанным способом.

На этапе проектирования выполняется исследование предметной области и определятся набор методик расчета показателей $M = \{M_i\}$, $i = \overline{1, N}$, предназначенных для анализа данных. Методики включают в себя правила вычисления показателей предметной области. Каждой методике предметной области ставится в соответствие числовое значение $S(M_i)$, отражающее ее семантическую важность. Частота использования данных в хранилище $F(M_i)$ оценивается через временной период обновления результатов расчетных методик, связанный с поступлением новых данных. Для этого необходимо определить наибольший период обновления данных для расчета методик – T, характерный для данной предметной области и вычислить частоту применения результатов каж-

дой методики за этот период. Частота применения каждой методики расчета показателей предметной области определяется методами экспертных оценок.

Входными данными алгоритма определения релевантных представлений является множество методик предметной области с рассчитанной семантической важностью и частотой использования: $M = \{M_i\}$, $i = \overline{1, N}$; $S(M_i) \in [0; \sigma]$; $F(M_i) \in [0; \theta]$, где σ – некоторое максимальное значение, соответствующее наибольшей важности; θ – наибольшая частота использования за период *T*.

Множество методик расчета показателей предметной области разбивается на классы по частоте использования. Количество классов *K* для разбиения методик выбирается на основе исследования предметной области с помощью экспертных оценок. Алгоритм определения релевантных представлений выполняется по шагам:

Шаг 1. Разбить множество расчетных методик на классы по частоте их использования.

Пусть *K* – количество классов, на которое разбивается интервал частоты применения расчетных методик:

$$[0; \theta] = [0; portion_1) \cup \ldots \cup [portion_{K-1}; \theta],$$

тогда каждая из частей интервала определяет отдельный класс:

$$M_i \in P_1, \text{ если } F(M_i) \in [portion_{K-1}; \theta],$$

$$M_i \in P_j, \text{ если } F(M_i) \in [portion_{j-1}; portion_j),$$

$$M_i \in P_K, \text{ если } F(M_i) \in [0; portion_1),$$

где $P = \{P_1, \ldots, P_j, \ldots, P_K\}$ – множество классов частот использования расчетных методик, $i = \overline{1, N}$, $j = \overline{2, K-1}$.

Шаг 2. Сформировать упорядоченное по релевантности множество расчетных методик предметной области.

Формируются классы методик расчета показателей предметной области по частоте и семантической важности. Входные данные: $M = \{M_i\}$, $i = \overline{1, N}$, $S(M_i) \in [0; \sigma]$, $F(M_i) \in [0; \theta]$, *K*. Требуется выделить *s* уровней на интервале $[0; \sigma]$:

$$[0; \sigma] = [0; level_1) \cup [level_1; level_2) \cup \ldots \cup [level_{s-1}; \sigma].$$

Количество уровней *s* и границы их интервалов определяются экспертами. В первую очередь необходимо учесть элементы, имеющие наибольшую оценку семантической важности. Правила разбиения определяют

классы расчетных методик по уровням семантической важности и классам частоты использования:

$$M_i \in R_j^{\,1}, \text{ если } F(M_i) \in P_j \text{ и } S(M_i) \in [level_{s-1}; \sigma],$$

$$M_i \in R_j^{\,2}, \text{ если } F(M_i) \in P_j \text{ и } S(M_i) \in [level_{s-2}; level_{s-1}),$$

$$M_i \in R_j^{\,s}, \text{ если } F(M_i) \in P_j \text{ и } S(M_i) \in [0; level_1), \; j = \overline{1, K}.$$

Выходные данные: $R_M = \{R_1^{\,1}, R_1^{\,2}, ..., R_1^{\,s}, \ldots, R_K^{\,1}, R_K^{\,2}, ..., R_K^{\,s}\}$ множество классов расчетных методик, упорядоченное по уменьшению релевантности и частоты использования. Пошаговое построение R_M с промежуточным выделением его элементов позволит выполнять перерасчет множества релевантных представлений за счет изменения или исключения из рассмотрения уровней семантической важности.

Шаг 3. Построить множество классов атрибутов методик.

Предлагаемый метод предназначен для выявления релевантных элементов предметной области. Для определения множеств атрибутов, участвующих в формировании расчетных методик, элементы множества R_M группируются по частоте использования: $R_1^1 = \{R_1^{\,1}, R_1^{\,2}, ..., R_1^{\,s}\}$,…, $R_M^K = \{R_K^{\,1}, R_K^{\,2}, ..., R_K^{\,s}\}$. Множество релевантных атрибутов M_A формируется, используются предложенные в работе алгоритмы «пересечения» и «исключения» (рисунок 2).

В алгоритме «исключения» элементы множества атрибутов, участвующих в формировании наиболее важных расчетных методик, вычисляются по формуле $M_1^A = A_1$, $M_j^A = A_j \setminus \bigcup_l A_l$, где $l = \overline{1, (j-1)}$, $j = \overline{2, K}$, A_j, A_l – множества атрибутов методик R_M^j и R_M^l соответственно.

Шаг 4. Упорядочить по релевантности множество представлений.

В результате выполнения алгоритма оценки релевантности атрибутов построено множество атрибутов M^A. Количество элементов множества зависит от выбранного алгоритма. Для алгоритма «исключения» количество элементов совпадает с количеством классов расчетных методик *K*, заданным экспертным путем: $M^A = \{M_i^A\}$, где $i = \overline{1, K}$.

Релевантность представлений определяется входящими в них атрибутами. На основе множества M^A формируется множество представлений

$M^{CV} = \{M_i^{CV}\}$, где M_i^{CV} – множество всевозможных сочетаний элементов $M_i^A \cup M_{i-1}^A$.

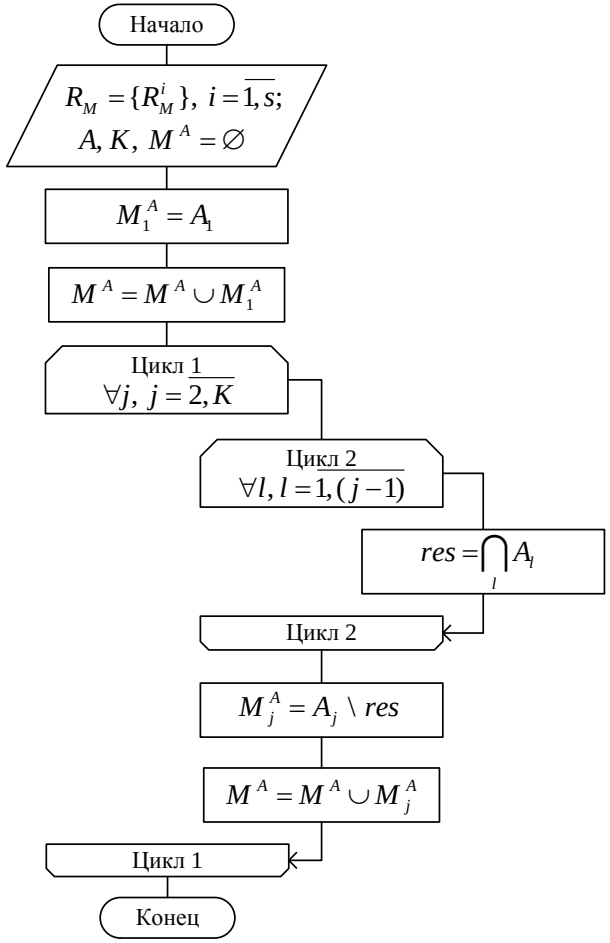

Рисунок 2 – Блок-схема алгоритма «исключения»

На рисунке 3 представлена блок-схема алгоритма выбора представлений для материализации. Особенностью алгоритма является использование упорядоченного по релевантности множества представлений M^{CV} как множество кандидатов на материализацию, что позволяет учесть требования пользователей, сократить пространство поиска и получить решение без накопленной информации о работе хранилища данных.

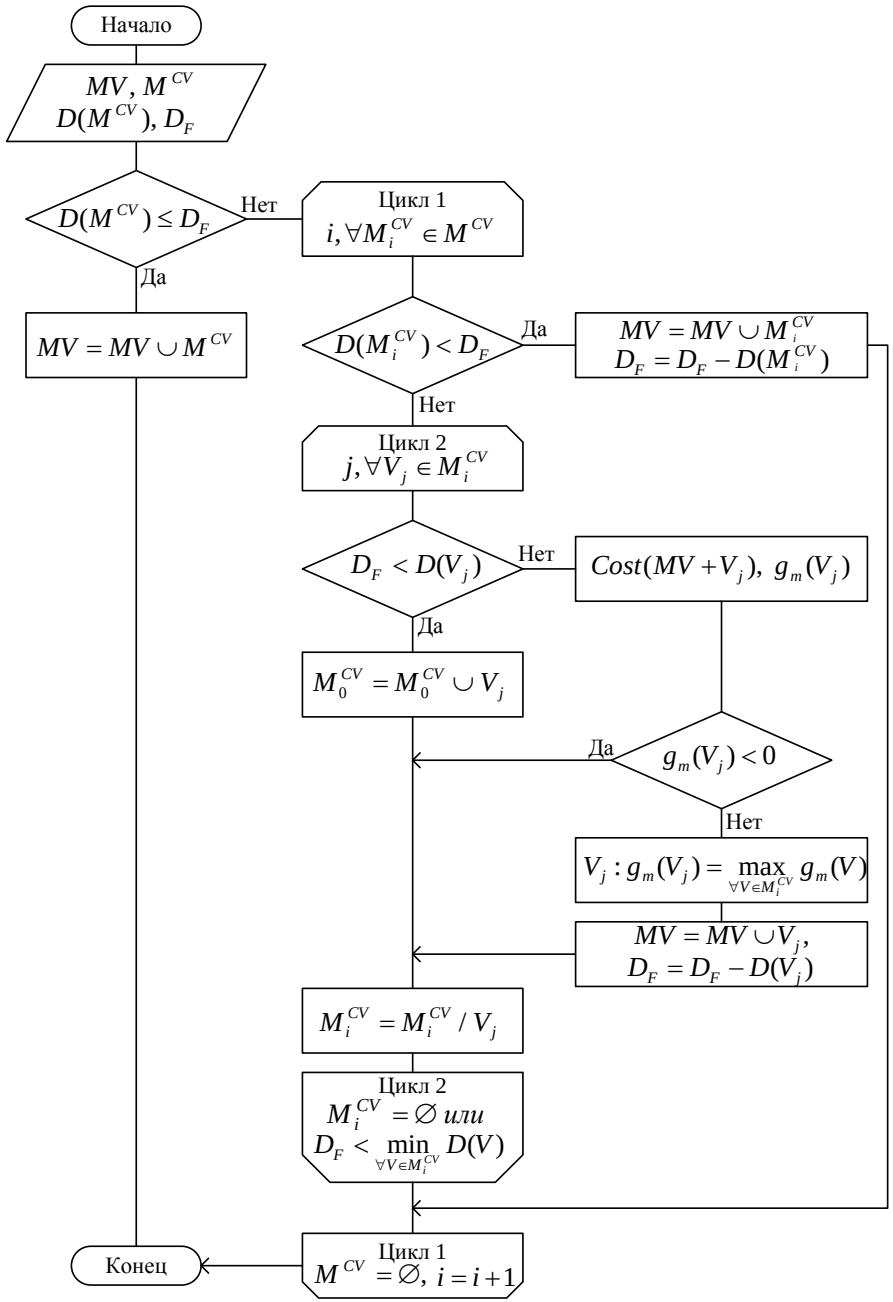

Рисунок 3 – Блок-схема алгоритма выбора представлений

Входными данными алгоритма выбора представлений для материализации являются: множество материализованных представлений *MV* (может быть пустым); множество представлений кандидатов на материализацию M^{CV}, построенное по алгоритму выбора релевантных представлений; объем дискового пространства D_F, выделенного для материализации представлений и пространство, необходимое для материализации всех кандидатов $D(M^{CV})$. Другие обозначения, используемые в блок-схеме: $Cost(MV+V_j)$ – общая стоимость материализации при добавлении V_j представления к *MV*; $V \in M^{CV}$ – представление из множества M^{CV}; $D(V_j)$ – пространство, занимаемое представлением V_j. В алгоритме используется показатель, позволяющий определить выгоду материализации в пространстве, занимаемом представлениями *MV*, который предложен М.-Ч. Хунгом и рассчитывается по формуле:

$$g_m(V_j) = \left(Cost(MV) - Cost(MV+V_j)\right) / D(V_j) \qquad (3)$$

Алгоритм останавливается, после того как множество *MV* сформировано или когда исчерпано доступное дисковое пространство.

Алгоритм релевантности используется в зависимости от изменений внешних условий, инициировавших его вызов – поступление новой информации, выделение или сокращение дискового пространства, при недостатке статистической информации о работе хранилища данных. Методика адаптивного управления процессом проектирования и развития специализированных хранилищ данных предполагает применение алгоритма релевантности. Использование алгоритма релевантности обеспечивает развитие хранилища и повышает его производительность начиная с ранних этапов эксплуатации за счет обоснованного включения в модель данных материализованных представлений.

В **главе 3** описаны результаты применения разработанного методического и алгоритмического обеспечения для решения задачи построения специализированного хранилища медико-демографических данных. Разработка специализированного хранилища медико-демографических данных выполнялась на основе методических рекомендаций специалистов Красноярского медицинского информационно-аналитического центра (КМИАЦ). Для создания информационной системы оперативного анализа медико-демографических данных использовался аналитический программный комплекс «Аналитик» (разработка ИВМ СО РАН), включающий систему генерации и администрирования хранилищ данных «Ме-

неджер хранилища» и аналитическую инструментальную среду, функционирующую под управлением СУБД Oracle 10g (рисунок 4).

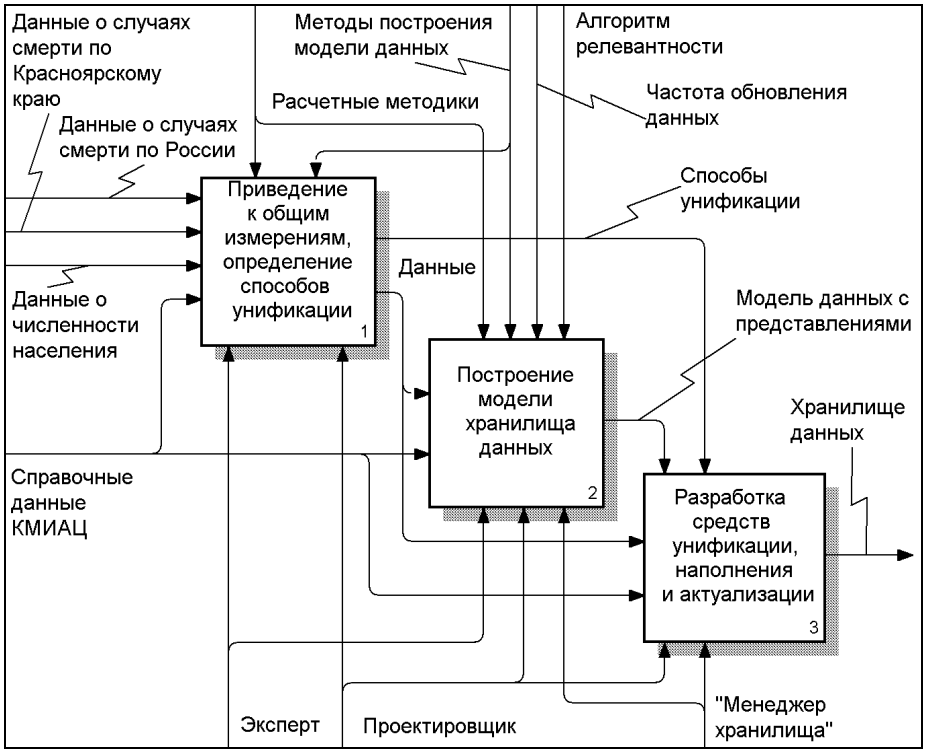

Рисунок 4 – Функциональная модель разработки специализированного хранилища медико-демографических данных

Данные, необходимые для анализа медико-демографических процессов, содержатся в информационных источниках, накопленных за длительный период времени и имеющих разные форматы хранения. Источники данных содержат записи о случаях смерти по территориям Красноярского края, данные по России (Госкомстат РФ) и данные о численности населения (демографический справочник, ведется специалистами КМИАЦ).

Проектирование модели данных хранилища выполняется по расширенной пространственной модели Р. Кимбалла, ориентированной на реализацию в реляционной базе данных. По проекту пространственной модели хранилища данных осуществляется ее физическая реализация (рисунок 5). Для повышения скорости обработки данных, выполняется вы-

бор и построение дополнительных структур модели данных – материализованных представлений, содержащих агрегированные результаты.

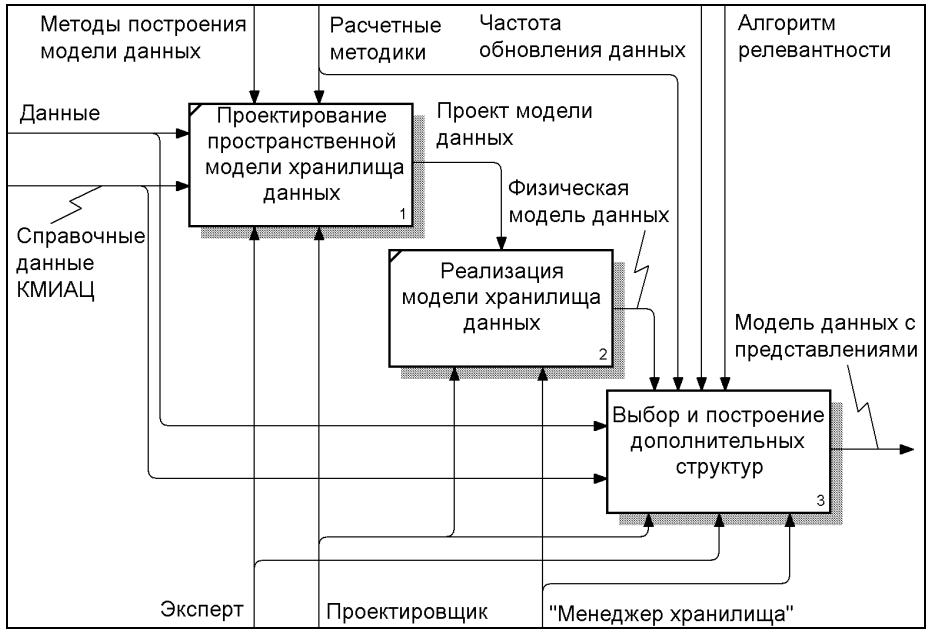

Рисунок 5 – Построение модели хранилища данных

Выбор представлений для материализации осуществляется на основе алгоритма релевантности, используя информацию о расчетных методиках предметной области, частоте обновления источников данных и сведения о физической реализации основных информационных объектов пространственной модели. Далее осуществляется добавление выбранного множества представлений для материализации к основным структурам модели хранилища данных (рисунок 6).

В соответствии с реализованной моделью хранилища данных с дополнительными структурами и определенными на этапе исследования предметной области требованиями к преобразованию данных, разрабатываются средства унификации, наполнения и актуализации хранилища. Исходная информация пополняется ежегодно. Для сокращения трудозатрат по обработке вновь поступивших данных созданы ETL-пакеты, которые содержат совокупность присоединенных процедур, выполняющихся в заданном порядке.

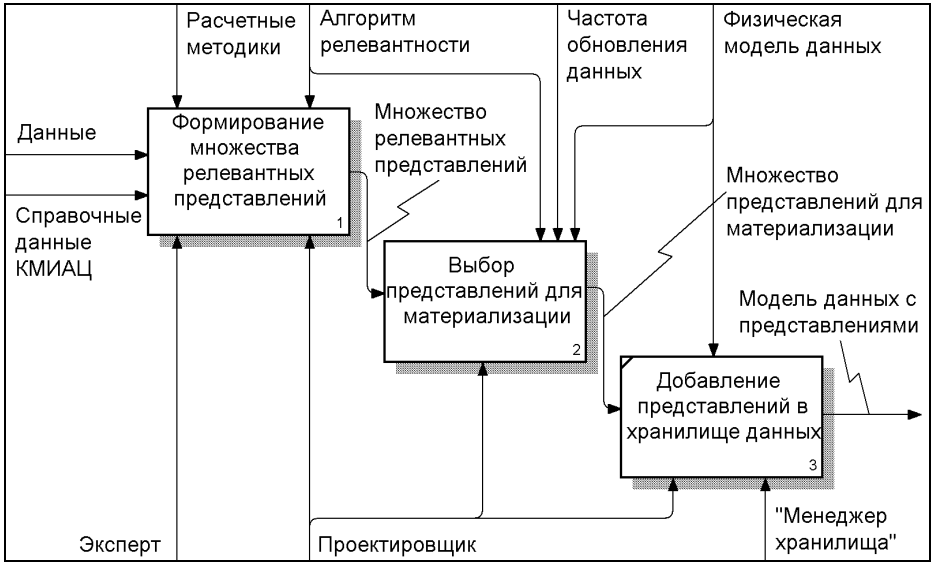

Рисунок 6 – Выбор и построение дополнительных структур данных

Применение предложенного в работе алгоритма выбора представлений рассмотрено на примере группы расчетных методик половозрастной смертности. Каждая методика включает обязательные для рассмотрения измерения «пол», «возраст» и факт «число умерших». Узлы решетки представляют результат сгруппированных по заданным измерениям данных таблицы фактов. Для упрощения записи введены обозначения узлов в соответствии с входящими в них измерениями: Y – год, R – район, T – тип поселения, P – причина смерти, M – место смерти, обязательные измерения (пол, возраст) объединяются под одной буквой – S. В структурной решетке (рисунок 7) цветом выделены представления, входящие в множество материализованных представлений, построенное с помощью предложенного алгоритма.

Выполнено сравнение результатов применения предложенного алгоритма выбора представлений для материализации и алгоритма Хунга. Вычислительная эффективность алгоритмов выбора представлений для материализации определяется количеством вычислений общей стоимости материализации. Показано, что для выполнения предложенного алгоритма требуется меньшее количество вычислений общей стоимости материализации. При усилении ограничения дискового пространства предложенный алгоритм позволяет найти решение, которое лучше соответствует требованиям пользователей.

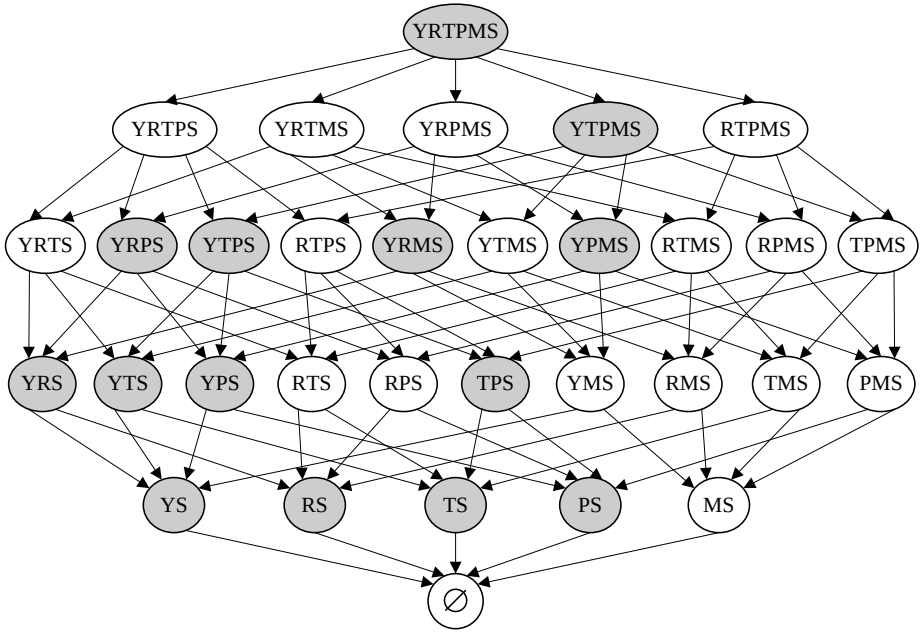

Рисунок 7 – Результат выполнения алгоритма
в структурной решетке

Выполнен сравнительный анализ характеристик существующих подходов к проектированию и развитию хранилища данных. Преимущество предложенной методики адаптивного управления процессом проектирования и развития состоит в том, что в отличие от других подходов, обеспечивается формализованное включение в модель хранилища данных материализованных представлений при отсутствии накопленной статистической информации о хранимых данных. Предложенная методика обеспечивает комплексное решение задач проектирования и развития хранилищ и является естественным дополнением к уже существующим и активно применяемым методам и подходам к проектированию хранилищ данных.

В **заключении** приведены результаты исследований, проведенные автором диссертационной работы, сформулированы выводы. Анализ практической значимости и перспективности полученных результатов позволяет сделать выводы о целесообразности применения разработанных методических и алгоритмических средств для создания специализированных хранилищ данных в задачах информационно-аналитической поддержки принятия решений.

Разработанная методика адаптивного управления применима при создании и развитии специализированных хранилищ данных для повышения их производительности. Методика предоставляет формализованные способы развития модели хранилища данных при изменении условий эксплуатации и появлении новой информации или дополнительного дискового пространства независимо от наличия накопленной статистики о работе хранилища данных.

В **Приложении 1** представлен фрагмент модели хранилища данных социально-экономических показателей развития региона.

В **Приложении 2** приведена интернет-система спортивно-оздоровительного движения «Игры народов планеты».

В **Приложении 3** приведены результаты эксперимента анализа скорости доступа к данным.

В **Приложении 4** приведена копия акта о внедрении результатов диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. На основе анализа предметной области и обзора научных публикаций обоснована актуальность разработки методов и алгоритмов, обеспечивающих повышение производительности хранилищ данных.

2. Предложена методика адаптивного управления процессом проектирования и развития специализированных хранилищ данных. Данная методика позволяет повышать производительность хранилищ данных за счет учета условий эксплуатации и использования алгоритма выбора релевантных представлений для материализации.

3. Разработан метод формирования модели общей стоимости материализации представлений на основе информации о предметной области, обеспечивающий использование алгоритмов выбора представлений для материализации при отсутствии накопленной статистической информации о работе хранилища данных.

4. Разработан алгоритм выбора релевантных представлений, учитывающий требования пользователей и позволяющий принимать решения о включении материализованных представлений в хранилище данных на основе информации о предметной области.

5. Предложенные методические и алгоритмические средства применены для построения специализированного хранилища медико-демографических данных. На основе специализированного хранилища данных разработана информационная система «Анализ медико-демографических процессов».

6. Полученные результаты также использованы в проектировании модели данных хранилища социально-экономических показателей развития региона и построении банка данных интернет-системы спортивно-оздоровительного движения «Игры народов планеты».

Дальнейшее развитие предложенной методики адаптивного управления предполагает интеграцию с существующими подходами и средствами развития структуры хранилища данных для расширения области действия алгоритмов и повышения их универсальности.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

Основные результаты исследований по теме диссертации, опубликованные автором К.В. Бадмаевой (Шалдыбиной) в изданиях, рекомендуемых ВАК для публикации результатов кандидатских и докторских диссертаций:

1. Бадмаева К.В. Алгоритм оценки релевантности представлений для материализации в специализированном хранилище данных / К.В. Бадмаева // Вестник Сибирского государственного аэрокосмического университета им. академика Решетнева. Выпуск 1(22). В 2 частях. Ч.2. – 2009. – С. 60-64.

2. Бадмаева К.В. Проектирование специализированного хранилища показателей социально-экономического развития региона / Т.Г. Пенькова, Л.Ф. Ноженкова, К.В. Бадмаева // Вестник КрасГАУ. – 2007. – № 14. – С. 122-128.

3. Шалдыбина К.В. Проблемы построения специализированного хранилища демографических данных / О.С. Исаева, К.В. Шалдыбина // Вестник Красноярского государственного университета, «Физико-математические науки». – 2006. – № 1. – С. 222-227.

4. Шалдыбина К.В. Технологические особенности оперативного анализа медико-демографических данных / О.С. Исаева, К.В. Шалдыбина // Вестник Томского государственного университета. Приложение. – 2006. – № 18. – С. 169-174.

Публикации в прочих изданиях:

5. Шалдыбина К.В. Методы подготовки демографических данных для оперативного анализа / К.А. Виноградов, А.В. Шульмин, О.С. Исаева, К.В. Шалдыбина // Материалы I Международной научно-практической конференции «Общественное здоровье: инновации в экономике, управлении и правовые вопросы здравоохранения». Т.2. / Сибмедиздат НГМА. – Новосибирск, 2005. – С. 86-88.

6. Бадмаева К.В. Формирование стоимостной модели для проектирования хранилищ данных / К.В. Бадмаева // Материалы VII всероссийской

научно-практической конференции «Современные информационные технологии в науке, образовании и практике» / ОГУ. – Оренбург, 2008. – С. 332-341.

7. Бадмаева К.В. Методика адаптивного проектирования специализированных хранилищ данных / К.В. Бадмаева // Труды XIV Байкальской конференции «Информационные и математические технологии в науке и управлении». Часть III. – Иркутск: ИСЭМ СО РАН, 2009 – С. 214-221.

8. Бадмаева К.В. Интернет средства мониторинга состояния здоровья населения / К.В. Бадмаева // Материалы пятого Всесибирского конгресса женщин-математиков / РИО СФУ. – Красноярск, 2008. – С. 38-43.

9. Бадмаева К.В. Проектирование и реализация модели данных системы ведения банка информации спортивно-оздоровительного движения «Игры народов планеты» / К.В. Бадмаева // Материалы десятой Всероссийской научно-практической конференции «Проблемы информатизации региона» ПИР-2007. В 2 т. Т.2. / Сиб. федер. ун-т; Политехн. ин-т. – Красноярск, 2007. – С. 77-84.

10. Шалдыбина К.В. Структурирование данных для мониторинга социально-экономического развития региона / Т.Г. Пенькова, К.В. Шалдыбина //«Открытое образование». Приложение. 2006. – С. 171-174.

11. Шалдыбина К.В. Построение хранилища данных для мониторинга демографических процессов / К.А. Виноградов, А.В. Шульмин, О.С. Исаева, К.В. Шалдыбина // Материалы девятой Всероссийской научно-практической конференции «Проблемы информатизации региона» ПИР-2005. В 2 т. Т.2. / ИПЦ КГТУ. – Красноярск, 2005. – С. 32-38.

12. Бадмаева К.В. Проектирование и реализация системы ведения мирового банка информации для спортивно-оздоровительного движения «Игры народов планеты» / К.В. Бадмаева // Материалы конференции молодых ученых / ИВМ СО РАН. – Красноярск, 2007. – С. 38-42.

13. Шалдыбина К.В. Модель хранилища данных для анализа показателей развития региона / Т.Г. Пенькова, К.В. Шалдыбина // Материалы конференции молодых ученых / ИВМ СО РАН. – Красноярск, 2006. – С. 78-82.

14. Шалдыбина К.В. Средства накопления, унификации и актуализации данных для анализа медико-демографических процессов / К.В. Шалдыбина // Материалы конференции молодых ученых / ИВМ СО РАН. – Красноярск, 2005. – С. 75-79.

Бадмаева Ксения Владимировна

Методика адаптивного управления
процессом проектирования и развития
специализированных хранилищ данных

Автореферат диссертации

Подписано в печать 31 августа 2009 г.
Формат 60×84/16
Усл. печ. л. 1. Тираж 100 экз.

Отпечатано на ризографе ИВМ СО РАН
660036, Красноярск, Академгородок