На правах рукописи

Попов Сергей Геннадьевич

ПОСТРОЕНИЕ ОПТИМАЛЬНОГО РЕПОЗИТОРИЯ АТРИБУТОВ И ОТНОШЕНИЙ ДЛЯ ИНТЕГРАЦИИ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

СПЕЦИАЛЬНОСТЬ

05.13.11 — Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

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

Санкт-Петербург — 2010

Работа выполнена в государственном образовательном учреждении высшего профессионального образования «Санкт-Петербургский государственный политехнический университет».

Научный руководитель:	доктор технических наук, профессор Заборовский Владимир Сергеевич
Официальные оппоненты:	доктор технических наук, профессор Гаврилова Татьяна Альбертовна
	кандидат технических наук Волков Михаил Владимирович
Ведущая организация:	Санкт-Петербургский институт информатики и автоматизации Российской академии наук

Защита состоится « *23* » *декабря* 2010 года в *16* часов на заседании диссертационного совета Д 212.229.18 ГОУ ВПО «Санкт-Петербургский государственный политехнический университет» по адресу: 195251, Санкт-Петербург, Политехническая ул., д. 21, к. 9, ауд. 325.

С диссертацией можно ознакомиться в фундаментальной библиотеке ГОУ ВПО «Санкт-Петербургский государственный политехнический университет» по адресу: 195251, Санкт-Петербург, Политехническая ул., д. 29, Главное здание.

Автореферат разослан « ____ » _____ 2010 года.

Учёный секретарь
диссертационного совета,
к.т.н., доцент Васильев А. Е.

Общая характеристика работы

Актуальность темы диссертационной работы обусловлена значимостью проблемы минимизации объёмов метаданных и времени исполнения алгоритмов интеграции независимых реляционных баз данных.

Под интеграцией подразумевается решение комплекса задач по объединению наборов данных из не связанных между собой реляционных баз. В приложении к реляционным базам данных под метаданными понимается набор имён атрибутов схем. Оптимальное управление интеграцией данных обеспечивает минимизацию объёмов служебных данных и времени формирования интегрирующих запросов к объединённым данным. На практике реализация алгоритмов управления метаданными с заранее известными временами исполнения позволяет применять их в высоконагруженных системах интеграции данных информационно-управляющих систем.

Вопросы интеграции реляционных баз данных рассматриваются в сфере управления базами данных, интеллектуального анализа и аналитической обработки данных, интеграции разнородных данных и нашли практическое применение в программных пакетах, обеспечивающих федеральную интеграцию данных и формирование открытых репозиториев метаданных. Разработками в данной области занимаются специалисты: Л. А. Калиниченко, М. Ш. Цаленко, Ю. Г. Карпов, С. А. Ступников, М. Р. Когаловский, С. Д. Кузнецов, Б. А. Новиков, Д. Мейер, М. Франклин, Г. Гарсия Молина, Д. Ульман, а также ведущие IT компании: Microsoft, Oracle, IBM.

В настоящее время основным способом интеграции данных из реляционных баз является управление данными на основе метаданных. К метаданным относятся имена БД, отношений, атрибутов, описания функциональных, в том числе и многозначных, зависимостей, которые, в совокупности, составляют схему БД.

К трём главным проблемам интеграции данных на основе репозиториев можно отнести многообразие внутренних представлений репозиториев в различных РСУБД, распространение репозитория на одну РСУБД и отсутствие возможностей достоверной оценки времени функционирования алгоритмов управления репозиторием. Выявленные ограничения приводят к высокой трудоёмкости управления крупными интеграционными проектами на основе реляционных СУБД различных производителей в условиях больших объёмов интегрируемых данных. Решением данной проблемы является применение подсистемы интеграции реляционных баз данных на основе независимого, оптимизируемого на этапе эксплуатации реляционного репозитория с полным набором операторов управления схемами независимых баз данных.

Целью работы является построение оптимального репозитория метаданных на основе алгоритмов управления схемами интегрируемых баз данных с прогнозируемым временем отклика на произвольном наборе атрибутов и отношений интегрируемых баз данных.

Для достижения цели в диссертационной работе поставлены и решены следующие задачи:

- предложена классификация схем репозитория на основе анализа числа реляционных отношений и функциональных зависимостей;
- разработана методика оптимизации репозитория данных на этапе эксплуатации с гарантированными оценками времени выполнения операций;
- разработан функционально полный набор операций, обеспечивающий согласованную последовательность преобразований схемы репозитория;
- разработаны алгоритмы управления схемами репозитория, обеспечивающие адаптацию его структуры к изменяющимся наборам структур интегрируемых данных;
- разработана архитектура и реализована подсистема интеграции реляционных баз данных на основе реляционного репозитория и операторов управления схемами.

Объектом исследования являются схемы реляционных баз данных, схемы репозитория и алгоритмы управления данными в нём.

Предметом исследования является организация изменения схемы репозитория на этапе эксплуатации и алгоритмы управления атрибутами и отношениями в объединённой базе данных.

Основные методы исследования. В качестве методов исследования применялись методы теории оптимизации, использовался аппарат теории множеств, общей и реляционной алгебр, методы анализа алгоритмов.

Результаты, выносимые на защиту, и их научная новизна. Предлагаемая диссертация содержит следующие результаты:

- разработана методика оптимизации реляционного репозитория метаданных отличающихся от известных совместным использованием 2 критериев оптимизации, что позволяет получить оптимальный репозиторий, в котором минимизированы объём данных и время доступа;
- разработан полнофункциональный набор алгоритмов управления схемами репозитория позволяющий обрабатывать любую из существующих реализаций схем репозитория;
- показано, что для произвольной схемы репозитория объём и время обработки данных линейно зависит от числа атрибутов объединяемых баз, что позволяет эффективно управлять репозиторием с произвольным набором независимых баз данных.

Теоретическая значимость работы заключается в разработке методики построения оптимального репозитория с целью интеграции реляционных баз данных на основе полнофункционального набора операций над схемами средствами реляционной алгебры. Разработанная методика управления схемами позволяет эффективно интегрировать независимые базы данных.

Практическая значимость работы состоит в разработке подсистемы интеграции, состоящей из инструментальных средств анализа схем интегрируемых баз данных, компоненты реконфигурации репозитория, графического редактора доопределения связей и универсального интерфейса доступа к данным средствами системы интеграции.

Предложенный подход позволяет сохранить неизменными данные, операторы и схемы интегрируемых БД, что позволяет реализовать операторы управления схемами независимо от интегрируемых БД и получать подсистемы интеграции с заранее известными временами отклика. В ходе практической реализации получены следующие результаты:

- реализация методики обеспечивает синтез репозитория независимых баз данных для заданного времени доступа к данным в репозитории, что позволяет проектировать систему с известными эксплуатационными характеристиками;
- разработанные алгоритмы позволяют повысить производительность подсистемы интеграции баз данных при неизменных требованиях к аппаратной компоненте;
- реализованный универсальный пользовательский интерфейс, обеспечивает единую технологию доступа к независимым реляционным базам данных.

Разработанные инструментальные средства нашли применение в качестве составных частей системы интеграции независимых реляционных баз данных.

Реализация результатов работы. Разработанная информационная система внедрена в качестве подсистемы интеграции реляционных баз данных в информационной системе управления учебным процессом Санкт-Петербургского государственного политехнического университета и электронной системе мониторинга технологических компетенцией предприятий машиностроительного комплекса Северо-Западного региона.

Апробация работы. Научные результаты и основные положения работы докладывались и обсуждались на конференциях: Международной научно-методической конференции «Высокие интеллектуальные технологии и интеграция знаний в образовании и науке» (Санкт-Петербург 2005), международной научно-методической конференции «Высокие интеллектуальные технологии и инновации в образовании и науке» (Санкт-

Петербург, 2006), международной научно-практической конференции «Высокие интеллектуальные технологии в образовании и науке» (Санкт-Петербург, 2010).

Публикации. По теме диссертации автором опубликовано 14 работ, объёмом 5,81 п.л. в том числе в изданиях, рекомендованных ВАК - 2 работы, объёмом 1,05 п.л.

Личный вклад автора. Все основные результаты работы диссертации получены автором самостоятельно.

Структура и объём диссертационной работы. Диссертация состоит из введения четырёх глав и заключения, изложена на 132 страницах, включая перечень литературы из 70 наименований, 60 рисунков и 4 таблицы. К диссертации добавлено приложение на 4 листах, содержащее схемы работы предложенных и реализованных в диссертации алгоритмов.

Содержание работы

Во введении приводится обоснование актуальности работы, формируется цель, описывается предмет, приводится краткое содержание диссертационной работы, формулируются результаты, выносимые на защиту, отмечается их актуальность, новизна и практическая значимость.

Первая глава содержит обзор систем управления репозиториями схем реляционных баз данных, описание недостатков существующих способов интеграции и путей их устранения, формулировку критериев классификации репозиториев схем, и постановку задачи интеграции данных на основе оптимального реляционного репозитория.

Исследования методов интеграции данных, организованных на различных моделях данных — реляционной, сетевой, иерархической - показали, что, если под моделью данных $\mathfrak{M}_i$ подразумевается четвёрка вида $\langle Sch_i, Ms_i, O_i, Mo_i \rangle$, где Sch_i — все возможные схемы в модели $\mathfrak{M}_i$, отображаемые ЯОД, O_i, — все возможные операторы ЯМД $\mathfrak{M}_i$, $Ms_i \colon Sch_i \to B_i$ — семантическая функция ЯОД $\mathfrak{M}_i$, а $Mo_i \colon O_i \to [B_i \to B_i]$ — семантическая функция ЯМД $\mathfrak{M}_i$, B_i — i–е состояние из пространства состояний данных модели $\mathfrak{M}_i$.

В диссертационной работе рассматриваются следующие дополнительные ограничения:

- модель данных единственна и является реляционной: $\mathfrak{M}_{Rel}$;
- схемы базы данных выбираются из одного множества реляционных схем Sch_{Rel};
- операторы O_{Rel_i} ЯОД модели $\mathfrak{M}_{Rel}$, являются разномощностными реализациями диалектов языков структурированных запросов,

основанных на реляционной алгебре.

Учитывая, что в современных СУБД репозитории реализованы в форме реляционных баз данных и уже содержат некоторые функции интеграции внутри одной СУБД, например в MySQL и Oracle, а операторы ЯОД управляют данными в этой схеме, то с практической точки зрения следует рассматривать ЯОД СУБД и базу данных репозитория как единую подсистему управления схемой хранимых БД.

Основными показателями эффективности применения интеграционной компоненты репозитория является объем хранимых метаданных и время выполнения операций управления репозиторием, которые зависят от схемы репозитория и набора исходных метаданных.

Анализ репозиториев современных СУБД показал, что их схемы фиксированы в соответствии со стандартами. Интеграция произвольных схем баз данных в репозиторий с фиксированной стандартом схемой, в общем случае, не обеспечивает гарантированной оптимальности хранения и управления схемами на этапе эксплуатации интегрирующего приложения.

Решением задачи повышения эффективности управления метаданными в ситуации заранее неизвестных наборов интегрируемых баз, является реконфигурация схемы репозитория в процессе эксплуатации системы с целью выбора оптимальных характеристик его функционирования.

Сформулируем новое требование оптимальности функционирования подсистемы интеграции. Это требование расширяет понятие правильного отображения моделей данных метода построения интероперабельных систем, введённое Калиниченко А. Л. В совокупности с требованиями интегрируемости, открытости и полнофункциональности выполнение требования оптимальности обеспечивает эффективность функционирования подсистемы на этапе эксплуатации.

Интегрируемость — формирование биективного отображения схем $\sigma\colon Sch_j \to Sch_i$, с целью объединения независимых баз данных.

Открытость — формирование набора отображений семантических функций диалектов ЯОД модели $\mathfrak{M}_{rel}$ $Ms_i\colon Sch_i \to B_i$, что обеспечивает отображение набора функций интеграции произвольных реляционных баз данных.

Полнофункциональность — формирование полного набора семантических функций подсистемы интеграции Ms_{Uni}, что аналогично предоставлению универсального интерфейса к функциям управления схемами и средствам интеграции.

Оптимальность — минимизация занимаемого объёма данных в репозитории Sch_{Uni} и времени доступа в процессе реализации функций Ms_{Uni} интеграции баз данных.

Под интеграцией независимых реляционных баз данных понимается

построение базы данных $BD^{(Uni)}$ со схемой $Sch^{(Uni)}$, которая содержит отношения

$$\begin{array}{l} R_1^{(Uni)}\left(C_{1_1}^{(Uni)}, C_{1_2}^{(Uni)}, \ldots, C_{1_{x_1}}^{(Uni)}\right) \\ R_2^{(Uni)}\left(C_{2_1}^{(Uni)}, C_{2_2}^{(Uni)}, \ldots, C_{2_{x_2}}^{(Uni)}\right) \\ \ldots \\ R_l^{(Uni)}\left(C_{l_1}^{(Uni)}, C_{l_2}^{(Uni)}, \ldots, C_{l_{x_f}}^{(Uni)}\right), \end{array} \quad (1)$$

языком манипулирования данными $O^{(Uni)}$, основанным на реляционной алгебре $Rel(T)$.

Доопределение связей в базах данных. Функциональные отношения интегрированной базы данных $R_l^{(Uni)}$ формируются двумя способами: включением исходных функциональных отношений интегрируемых баз данных и дополнением новыми. Второй набор отношений отражает новые семантические связи между базами данных и вводится извне, определяя новую схему интегрируемой базы данных.

Представим схему интегрируемых баз данных в форме ориентированного псевдографа $G(V, N, L)$, где V — множество атрибутов отношений схемы s_i, а N — множество рёбер, маркированных идентификатором отношения включения атрибута в отношение, L — множество троек вида (X_r, r, Y_r), маркированных идентификатором функциональных зависимостей исходных баз данных s_i, причём X_r — множество атрибутов детерминантов, Y_r — множество атрибутов, функционально зависящих от детерминанта, r — маркер функционального отношения. Тогда доопределение связей в базе данных состоит в формировании графа $G'(V, N, L')$, такого что $L' = (L \cup \{X', r', Y'\} | X'_i \subset s_i, Y'_j \subset s_j, i \neq j)$.

Постановка задачи оптимизации схемы репозитория. Пусть задан произвольный набор схем независимых интегрируемых баз данных $S = \{s_1, s_2, \ldots, s_n\}$, объединённых в набор $Sch^{(Uni)}$, на котором доопределены новые зависимости в $G'(V, N, L')$, а множество всех возможных схем репозитория баз данных $M = \{m_1, m_2, \ldots, m_{2^n}\}$, где n - число атрибутов в схеме $Sch^{(Uni)}$.

Для каждой схемы $m_j \in M$ определяется объём данных, как

$$V_{s_i} = \sum_{i=1}^{n} size(at(s_i)). \quad (2)$$

Для каждой схемы $m_j \in M$ определяется время выполнения команд $Z(S) = \{z_{Select}(s_i), z_{Insert}(s_i), z_{Delete}(s_i), z_{Integrity}(s_i)\}$ с частотами исполнения $\alpha_Q = \{\alpha_{Select}, \alpha_{Insert}, \alpha_{Delete}, \alpha_{Integrity}\}$, как

$$T_{V,s_i} = \sum_{i=1}^{k} time(V, m_i) + \sum_{j=1}^{p} time(V, m_i), \quad (3)$$

где k — число отношений, p — число зависимостей в схеме в схеме m_i, а функция $time()$ вычисляет время выполнения операции на каждом отношении схемы s_i.

Определим меру $|X| = \alpha V_{s_i} + \beta T_{V,S_i}$, при условии $\alpha + \beta = 1$.

Тогда задача оптимизации подсистемы интеграции баз данных на основе репозитория определяется как:

$$M \xrightarrow[min|X|]{} s_i \quad (4)$$

Нормированные параметры α и β выбираются пользователем исходя из стратегии оптимизации и используются при построении меры $|X|$ в минимизируемом выражении 4 для выбора оптимальной схемы s_i.

В второй главе описана методика оптимизации схемы реляционного репозитория с оценкой максимального времени исполнения алгоритмов, определён полнофункциональный набор операторов: добавления, удаления, интеграции и декомпозиции схем БД в реляционном репозитории, приведено исследование свойств схемы репозитория, приведены оценки времени выполнения операций, изложены результаты исследования оценки объёмов данных и времени выполнения запросов к репозиторию.

Схема методики оптимизации репозитория и оценки времени отклика подсистемы интеграции на этапе эксплуатации, основанной на реляционном репозитории базы данных приведена на рисунке 1.

Объектами управления в информационной подсистеме репозитория являются схемы баз данных $s = \{K(A, db, T, F), Lf_i, f_j | \forall f_i, F_j \in F, i \neq j\}$ из множества схем S, операции управления схемами репозитория $Z(S)$ и консолидированные запросы на выборку данных $q = \{q_1, q_2, \ldots, q_n\}$ к данным из объединённого репозитория.

С целью ограничения числа вариантов реализации схемы репозитория сгенерированы и классифицированы все варианты схем. Построение всех вариантов схем репозитория реализовано при помощи алгоритмов построения замыкания функциональных зависимостей $FSCHEME^+$ и всех возможных схем репозиториев $M = \{m_1, \ldots, m_{2^n}\}$, основанных на аксиомах пополнения и транзитивности Армстронга. Программная реализация формирования множества наборов позволила сформировать 64 набора функциональных зависимостей, которые были классифицированы по числу отношений и связей между ними.

В качестве критерия классификации схем репозитория предложено правило размещения атрибутов по функциональным отношениям. Тогда

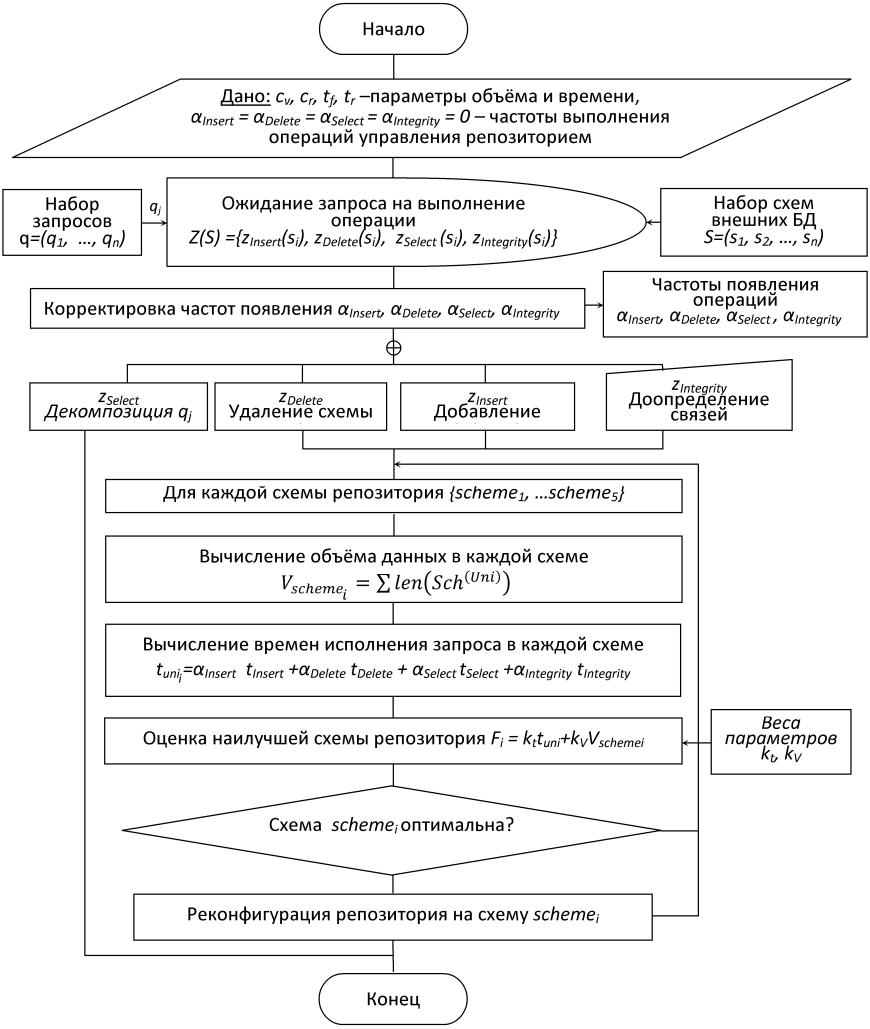

Рисунок 1. Схема алгоритма оптимизации репозитория

все схемы репозитория классифицируются по числу отношений. В результате анализа сформировано 5 классов схем. Общность схем каждого класса определяется равным числом операций просмотра всех отношений и естественного соединения отношений по первичному ключу.

Для управления схемами в репозитории для каждой операции раз-

работаны четыре алгоритма управления схемами. Особенностью реализации алгоритмов является использование метаданных в качестве исходных данных, что позволяет применять их без изменений для любого класса репозитория.

1. Добавление новой схемы в репозиторий z_{Insert}.

   Require: s_i; {схема базы для добавления}
   Ensure: $DB_{Uni} + s_i$;
   1: **for all** $_{Uni} \in scheme_i$ **do**
   2: **for all** $T_s \in s_i$ **do**
   3: $T_s \to T_{Uni}$;
   4: **for all** $F_s \in s_i$ **do**
   5: $F_s \to F_{Uni}$;

2. Удаление схемы из репозитория с удалением доопределенных связей z_{Delete}.

   Require: $s_i, X_{r_{s_i,s_j}}$; {схема базы для удаления и набор доопределяемых связей}
   Ensure: $DB_{Uni} - s_i$;
   1: **for all** $_{Uni} \in scheme_i$ **do**
   2: **for all** $T_s \in s_i$ **do**
   3: $T_s \nrightarrow T_{Uni}$;
   4: **for all** $F_s \in s_i$ **do**
   5: $F_s \nrightarrow F_{Uni}$;
   6: **for all** $F_s \in X_{r_{s_i,s_j}}$ **do**
   7: $F_s \nrightarrow X_{Uni}$;

3. Доопределение семантических связей $z_{Integrity}$.

   Require: $X_{r_{s_i,s_j}}$; {набор доопределяемых связей}
   Ensure: $DB_{Uni} + X_{r_{s_i,s_j}}$;
   1: **for all** $F_s \in X_{r_{s_i,s_j}}$ **do**
   2: $F_s \to X_{Uni}$;

4. Декомпозиция схемы z_{Select}.

   Require: $q_i(atr_1, atr_2, \ldots, atr_n)$; {запрос на декомпозицию данных}
   Ensure: $F(s_1, s_2, \ldots, s_i)$; {набор атрибутов из разных схем баз данных}
   1: **for all** $T_q \in q_i$ **do**
   2: **for all** $_{Uni} \in scheme_i$ **do**
   3: **for all** $t_{Uni_i} \bowtie t_{Uni_j} \in scheme_i$ **do**
   4: $F_{s_i} \leftarrow atr_k$;

Пусть в тестовой схеме внешней базы данных s_i определены: r_s — число атрибутов в схеме s_i, c_V — коэффициент увеличения объёма данных при отсутствии нормализации, t_r — время управления отношением,

t_f — время управления зависимостью, l_s — число отношений запросе, c_t — коэффициент времени на формирование избыточности в отношении. Проанализировав приведённые алгоритмы с учётом свойств схем репозиториев каждого класса, получим аналитические оценки времени исполнения и объёма данных в каждом классе схем. Результаты анализа времени исполнения и оценки объёма данных для различных вариантов реализации схем репозитория приведены в таблице 1.

Таблица 1. Анализ характеристик классов схем репозиториев

Класс схемы	Оценка объёма, V	Добавление, z_{insert}	Удаление, z_{delete}	Декомпозиция, z_{Select}
1 отношение	$6c_v r_s$	$6c_t t_r$	$t_r + t_f$	$5c_t * l_s * t_r$
2 отношения 1 зависимость	$\frac{r_s}{2c_v}$	$4c_t t_r + t_f$	$2t_r + 2t_f$	$4c_t * l_s * t_r + t_f$
3 отношения 2 зависимости	$\frac{r_s}{3c_v}$	$6c_t t_r + 2t_f$	$3t_r + 2t_f$	$3c_t * l_s * t_r + 2t_f$
4 отношения 3 зависимости	$\frac{r_s}{4c_v}$	$8c_t t_r + 3t_f$	$4t_r + 3t_f$	$2c_t * l_s * t_r + 3t_f$
5 отношений 4 зависимости	$\frac{r_s}{5c_v}$	$10c_t t_r + 4t_f$	$5t_r + 4t_f$	$l_s * t_r + 4t_f$

Представленные результаты позволили оценить объем данных и время выполнения алгоритмов при выполнении операций управления схемами в репозиториях разных классов, что позволило прогнозировать временные характеристики управления репозиторием при известном распределении времени исполнения запросов, и представить оценку среднего времени исполнения запроса как:

$$t_{uni} = \alpha_{Insert} * t_{Insert} + \alpha_{Delete} * t_{Delete} + \alpha_{Select} * t_{Select} \qquad (5)$$

Максимальное время исполнения запроса декомпозиции $t_{uni} = t_{Select} \max len(q_i)$, где $len(q_i)$ — функция определения числа естественных соединений в схеме репозитория $scheme_i$, а функцию оптимизации в форме нахождения минимума на дискретном наборе значений времени и

объёмов данных с весовыми коэффициентами k_t и k_V соответственно:

$$|X| = k_t t_{uni} + k_V * V_{scheme} \quad (6)$$

Предложенная функция применяется в алгоритме принятия решения о необходимости изменения структуры репозитория баз данных.

Третья глава содержит описание подсистемы интеграции на основе оптимального репозитория, её функции, структурную и функциональную схемы подсистемы, описание реализации графического интерфейса доопределения связей и универсального интерфейса пользователя генератора отчётов с использованием данных реляционного репозитория.

Методика интеграции независимых РБД на основе оптимального репозитория реализует предложенный полнофункциональный набор операторов управления схемами, включающий в себя средства управления репозиторием, взаимодействия со внешними базами данных и прикладным ПО системы. Управление схемами внешних баз данных предполагает извлечение, преобразование и анализ схем — создание модели управления схемами, и обработку запросов пользователей в этой модели. Совокупность модели и методов и составляет автоматизированную систему интеграции баз данных.

На вход подсистемы поступают описания схем баз данных, запросы пользователей на интеграцию данных и критерии оптимизации подсистемы. Выходом подсистемы является описание декомпозиции интегрированных запросов пользователей с целью обращения ко внешним базам данных с учётом требований к объёму данных и времени доступа к репозиторию.

Подсистема позволяет пользователю до исполнения запроса оценивать временные затраты на декомпозицию запроса к независимым базам данных, что позволяет строить системы интеграции с заранее известными временами отклика.

В результате анализа информационных процессов выявлены и описаны 7 функций системы, свойства двух внешних источников данных и двух информационных хранилищ, выделены 10 управляющих потоков и 12 потоков данных.

Подсистема оптимизации репозитория организована по модульному принципу и состоит из модулей:

- интерфейса со внешними базами данных. Модуль осуществляет взаимодействие с внешними базами данных, преобразуя описание схемы с конкретного диалекта ЯМД РСУБД во внутренний формат программы;
- анализа оптимальности репозитория. Модуль вычисляет значение критерия эффективности текущего состояния базы данных, и в

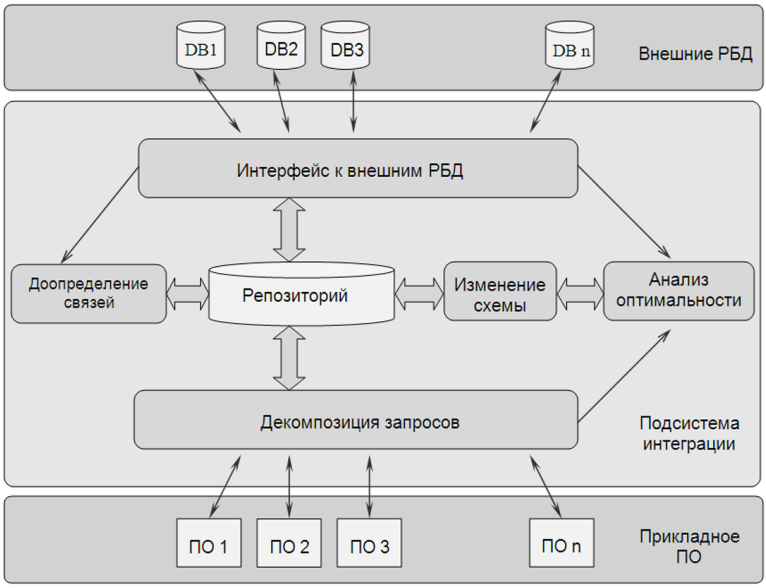

Рисунок 2. Структурно-функциональная схема подсистемы интеграции РБД

случае нахождения иного оптимального решения, формирует запрос на реконфигурацию репозитория;

– изменения схемы репозитория. Модуль принимает запрос на изменение структуры репозитория и формирует набор команд на языках управления ЯМД$_{Uni}$ и ЯОД$_{Uni}$ с целью преобразования структуры репозитория;
– доопределения связей. Модуль взаимодействует с пользователем в интерактивном режиме с целью формирования новых семантических связей между атрибутами внешних баз данных;
– декомпозиции запросов пользователя. Модуль на входе получает запрос пользователя и осуществляет выделение частей запроса, относящихся к каждой независимой базе данных.

Структурно-функциональная схема подсистемы интеграции РБД приведена на рисунке 2.

Разработанная оптимизирующая подсистема интеграции данных на основе реляционного репозитория схем баз данных, обеспечивает единую форму представления репозитория для всех РБД, полнофункцио-

нальный универсальный язык управления содержимым репозитория с возможностью прогнозирования объёма и времени доступа к схемам.

В четвёртой главе содержится описание применения подсистемы интеграции независимых реляционных баз данных на основе оптимального репозитория при построении информационно-управляющих систем.

Программная реализация подсистемы интеграции применена в системе управления учебным процессом Санкт-Петербургского государственного политехнического университета и электронной системе мониторинга технологических компетенцией предприятий машиностроительного комплекса Северо-Западного региона.

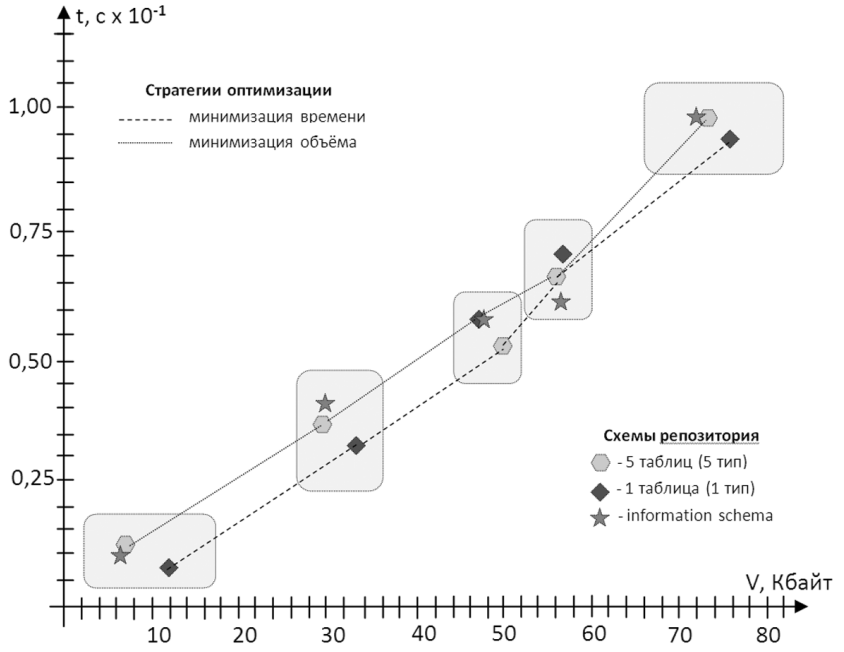

Рисунок 3. Траектория изменения схемы репозитория в процессе интеграции подсистем системы управления учебным процессом СПбГПУ

В процессе функционирования подсистемы оптимизации при каждом добавлении новой схемы интегрируемой базы производилась оценка занимаемого объёма и времени исполнения операции декомпозиции запроса. Результаты вычислений, траектория изменения состояния репозитория information_schema и траектория изменения двух схем репозитория

при реализации стратегий «минимальный объём» и «минимальное время декомпозиции» приведены на рисунке 3. Из представленного графика видно, что в процессе добавления новых схем при реализации стратегий «минимальный объём» и «минимальное время» происходит переключение схемы репозитория с 1 типа на 5 и обратно. Такое переключение связано с добавлением новой схемы, имена атрибутов которой повторяют имена атрибутов других схем репозитория.

Заключение содержит описание результатов, полученных в работе.

Основные результаты работы

В диссертации получены следующие научные и практические результаты:

1. Проведён обзор существующих репозиториев реляционных СУБД, в результате чего установлено, что в существующих СУБД отсутствуют средства оптимизации схем реляционных репозиториев. Отсутствие средств оптимизации обусловлено отсутствием эффективных средств решения задачи интеграции, которая стала актуальна в последние 10-15 лет.
2. Сформулировано новое требование оптимальности репозитория, расширяющее существующий метод интеграции баз данных на основе отображения моделей данных.
3. Сформулирована постановка задачи оптимальной интеграции, новизна постановки связана с тем, что задача интеграции рассматривается для произвольных схем независимых баз данных в отличие от классического подхода к интеграции схем нормализованных баз.
4. Выработаны критерии классификации схем реляционных репозиториев. В основу классификации положены следующие критерии: число отношений и число функциональных зависимостей. Выделены 5 классов схем репозитория. Каждый класс объединяет схемы с одинаковыми оценками объёма данных и времени доступа к ним.
5. Разработана методика оптимизации схемы репозитория, которая синтезирует репозиторий с заданным временем исполнения алгоритмов преобразования и объёмов данных.
6. Предложен функционально полный набор операций, обеспечивающий согласованную последовательность преобразований схемы репозитория. Предложенный набор содержит 4 операции управления схемами, позволяющие впервые решать эту задачу на практическом уровне.
7. Реализованы четыре алгоритма управления схемами, инвариантные к текущей схеме репозитория.

8. Проведено исследование зависимости объёма данных от исходного набора схем в каждом классе репозитория. Исходный набор схем характеризуется числом отношений, числом атрибутов в каждом отношении и числом функциональных зависимостей. В результате анализа установлено, что объем данных линейно зависит от вышеперечисленных параметров. Линейность зависимости позволяет использовать все пять классов репозиториев при любых наборах исходных схем.
9. Проведено исследование зависимости времени исполнения операций из полнофункционального набора от исходного набора схем в каждом классе репозитория. В результате анализа установлено, что рост времени линейно зависит от исходного набора схем, что позволяет применять алгоритмы для произвольной схемы репозитория.
10. Разработана архитектура подсистемы интеграции реляционных баз данных на основе репозитория. Подсистема интеграции включает модули управления схемами, реконфигурации репозитория, универсальный генератор отчётов и редактор графического представления схем, реализующий полнофункциональный набор операций управления репозиторием.
11. Разработано программное обеспечение, реализующие подсистему интеграции. Подсистема интеграции данных включена в интегрированную информационную систему управления учебным процессом СПбГПУ и электронной системе мониторинга технологических компетенцией предприятий машиностроительного комплекса Северо-Западного региона, что подтвердило правильность предложенной методики.

Публикации по теме диссертации

Научные статьи, опубликованные в изданиях, рекомендованных перечнем ВАК:

1. *Курочкин М. А., Попов С. Г.* Метод интеграции независимых баз данных // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Серия «Информатика. Телекоммуникации. Управление.» — СПб.: Изд-во СПбГПУ, 2006. — №4(46). — С. 28-32. *С. Г. Попову принадлежит описание требования оптимальности и постановка задачи интеграции реляционных баз данных.*

2. *Попов С. Г.* Методика построения оптимального репозитория схем реляционных баз данных // Научно-технические ведомости Санкт-

Петербургского государственного политехнического университета. Серия «Информатика. Телекоммуникации. Управление.» — СПб.: Изд-во СПбГПУ, 2010. — №5(108). — С. 91–98.

Научные статьи, опубликованные в иных изданиях:

1. *Криулин К. Н., Попов С. Г., Рафиков Ш. М.* Автоматизированная система управления контингентом СПбГПУ. // Высокие интеллектуальные технологии в высшей школе: Материалы XII Международной научно-методической конференции. — СПб.: Изд-во СПбГПУ, 2005. — С.65–71.

2. *Орлов Е. А, Попов С. Г.* Разработка подсистемы автоматизированной генерации структуры базы данных. // XXXIII Неделя науки СПбГПУ: Материалы Всероссийской межвузовской научно-технической конференции студентов и аспирантов. — СПб.: Изд-во Изд-во СПбГПУ, 2005. — Ч.XII. — С.21–23.

3. *Попов С. Г., Слюньков Н. В.* Структура автоматизированной системы управления рабочими учебными планами. // X Всероссийская конференция по проблемам науки и высшей школы. — СПб, Изд-во СПбГПУ, 2006. — С.69-73.

4. *Курочкин М. А., Попов С. Г.* Методология проектирования системы управления учебным процессом университета. // Фундаментальные исследования в технических университетах: Материалы X Всерос. конф. по проблемам высшей школы. — СПб.: Изд-во СПбГПУ, 2006. — С.48-52.

5. *Курочкин М. А, Попов С. Г.* Постановка задачи интеграции независимых реляционных баз данных. // Международная научно-методическая конференция «Высокие интеллектуальные и инновации в образовании и науке». — Изд-во. СПбГПУ, 2006. — Т1. — С.161-165.

6. *Попов С. Г.* Постановка задачи автоматического построения графического представления схемы реляционной базы данных произвольной структуры // XXXVII Неделя науки СПбГПУ: Материалы Всероссийской межвузовской научной конференции студентов и аспирантов. — СПб.: Изд-во СПбГПУ, 2008. — Ч.XVII. — С.46–84.

7. *Курочкин М. А., Попов С. Г., Курочкин Л. М., Тимофеев Д. А., Радкевич М. М.* Концепция построения электронной системы мониторинга технологических компетенцией предприятий Санкт-Петербурга // Высокие интеллектуальные технологии в образовании и науке: Сб. тезисов международной научно-практической конференции. — СПб.: Изд-во СПбГПУ, 2009. — С.56-67.

8. *Попов С. Г.* Исследование вариантов реализации схем реляционного репозитория схем баз данных. // Высокие интеллектуальные технологии в образовании и науке: Материалы XVII Международной научно-методической конференции. — СПб.: Изд-во СПбГПУ, 2010. — С. 71–74.