

На правах рукописи



Громов Алексей Юрьевич

**АЛГОРИТМЫ ПРОЕКТИРОВАНИЯ СХЕМ РЕЛЯЦИОННЫХ
БАЗ ДАННЫХ, СОДЕРЖАЩИХ РАНЖИРУЕМЫЕ АТТРИБУТЫ**

**05.13.11 – Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей**

**АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук**

Рязань 2012

Работа выполнена в ФГБОУ ВПО «Рязанский государственный радиотехнический университет»

Научный руководитель: кандидат технических наук, доцент
Баранчиков Алексей Иванович

Официальные оппоненты: Скворцов Сергей Владимирович,
доктор технических наук, профессор,
ФГБОУ ВПО «Рязанский
государственный радиотехнический
университет», профессор кафедры
САПР вычислительных средств.

Миловзоров Александр Владимирович,
кандидат технических наук, доцент,
ОГБОУ ДПО «Рязанский институт
развития образования», проректор по
научно-методической работе.

Ведущая организация: ФГБОУ ВПО «Владимирский
государственный университет имени
Александра Григорьевича и Николая
Григорьевича Столетовых»

Защита состоится «21» ноября в 12 часов на заседании
диссертационного совета Д 212.211.01 в ФГБОУ ВПО «Рязанский
государственный радиотехнический университет» по адресу: 390005, г.
Рязань, ул. Гагарина, 59/1.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВПО
«Рязанский государственный радиотехнический университет»

Автореферат разослан « » октября 2012 г.

Ученый секретарь
диссертационного совета
к.т.н., доц.



В.Н. Пржегорлинский

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Современное информационное общество нуждается в надёжных средствах разграничения доступа к информации. Не все современные системы управления базами данных (СУБД) обеспечивают достаточный уровень разграничения прав доступа к отдельным объектам реляционных баз данных (РБД). Исключение составляют СУБД, поддерживающие мандатную схему управления доступом (например, СУБД Линтер и Oracle).

В большинстве современных СУБД развита дискреционная схема с разграничением прав доступа пользователей на уровне таблиц. В случае необходимости разграничения прав доступа на уровне атрибутов данные функции реализуются в прикладных программах, с помощью которых обеспечивается интерфейс конечного пользователя с ресурсами РБД.

Возникает задача обеспечения разграничения прав доступа к объектам РБД на уровне атрибутов средствами СУБД, поддерживающих дискреционную схему управления доступом с разграничением прав доступа на уровне таблиц.

Задачу разграничения прав доступа можно расширить. Более широким классом задач являются задачи ранжирования атрибутов отношений с точки зрения ограничений, накладываемых на сущности в предметной области (ПрО). В число этих задач входит размещение частей РБД по узлам распределённой системы, для чего необходима вертикальная фрагментация схемы РБД, с сохранением согласованности информации.

В работе используются следующие обозначения, касающиеся атрибутов, зависимостей, отношений, признаков и их множеств: атрибут, зависимость, отношение и признак будут обозначаться заглавными буквами латинского алфавита, а их множества и подмножества — заглавными буквами латинского алфавита, выделенными жирным шрифтом.

Пусть множество $A = \{A_1, A_2, \dots, A_g\}$ атрибутов, выявленных при семантическом анализе предметной области, можно разделить на группы по некоторым признакам $Y = \{Y_1, Y_2, \dots, Y_n\}$, тогда входное множество A можно разделить на подмножества $A_{Y_1}, A_{Y_2}, \dots, A_{Y_n}$ такие, что $A = A_{Y_1} \cup A_{Y_2} \cup \dots \cup A_{Y_n}$. Множество A будем называть множеством ранжируемых атрибутов по множеству признаков Y .

Задача, включающая аспекты разграничений прав доступа к объектам РБД и проектирования структуры распределённой РБД сформулирована, как проектирование логической структуры РБД с ранжируемыми атрибутами с точки зрения ограничений предметной области. В случае обеспечения требований конфиденциальности эти ограничения выглядят как требование хранить в одном отношении атрибуты одного уровня конфиденциальности. В случае построения

структуры распределённой РБД эти ограничения выглядят как требование хранить в одном отношении атрибуты одного признака (конфиденциальности, быстродействия, территориального размещения и др.).

Таким образом, необходимо разработать алгоритм, позволяющий строить корректные схемы РБД $R = \{R_1, R_2, \dots, R_m\}$ на основе выявленных в предметной области множеств ранжируемых атрибутов A и семантических зависимостей $U = F \cup MV \cup J$ [функциональных зависимостей (F), многозначных зависимостей (MV) и зависимостей соединения (J)].

Для решения поставленной задачи необходимо:

а) выявить атрибуты и семантические зависимости предметной области;

б) сформулировать ограничения предметной области. Здесь под ограничениями понимаются правила хранения информации конкретной предметной области (например, бизнес-правила, правила, связанные с конфиденциальностью информации, или хранением информации в распределённой РБД), которые не могут быть выражены ограничениями целостности, привычными для разработчиков РБД;

в) выделить множество признаков Y атрибутов A , основанных на сформулированных ограничениях. Разработчик логической структуры РБД, на основе сформулированных ограничений выделяет признаки атрибутов $Y = \{Y_1, Y_2, \dots, Y_n\}$, по которым последние можно разделить на группы $A_{Y_1}, A_{Y_2}, \dots, A_{Y_n}$;

г) построить корректную схему РБД $R = \{R_1, R_2, \dots, R_k, \dots, R_m\}$, в которой отношения R_k ($k=1, m$) не должны содержать атрибуты разных признаков, с сохранением согласованности данных.

Степень разработанности темы. Разработке подходов и алгоритмов построения схем РБД, а также всем аспектам данных процессов посвящено достаточно большое количество трудов. Основные теоретические положения данного направления были заложены в 70-80-е годы 20-го века. Наиболее известными авторами данной тематики являются Э. Кодд, К. Дейт, Д. Мейер, Х. Дарвен, Д. Ульман, Д. Уидом, А. Ахо, И. Хит, Р. Фейгин, Ф. Бернштейн, Ж. Риссанен, М. Стоунбрейкер и другие.

Из отечественных авторов можно выделить Кузнецова С.Д., Пушкинова А.Ю, Грушо А.А., Новосельского В.Б.

В работе Грушо А.А. приведен метод декомпозиции отношений БД в стандартные базовые отношения реляционной модели с использованием классификационных ограничений. В работах Новосельского В.Б. рассматривается генетический подход к проектированию распределённых БД.

В работах Ф. Бернштейна предложен алгоритм синтеза схем РБД. В работах Д. Мейера решаются недостатки алгоритма синтеза, касающиеся

необходимости наличия отношения в результирующей схеме РБД **R**, содержащего универсальный ключ отношений, а также учета атрибутов, не использующихся в семантических зависимостях $U_i \in U$ предметной области. В работах И. Хита и Р. Фейгина рассмотрены вопросы, связанные с декомпозицией отношений без потерь информации. Ж. Риссанен в своих трудах рассматривает проблемы, связанные с потерей зависимостей при декомпозиции без потерь отношений R_k , и предлагает вариант решения данной задачи. К. Дейт на основе результатов Риссанена приводит методику декомпозиции отношений на независимые проекции.

Цель работы — расширение функциональности алгоритмов построения схем РБД за счет учета ранжируемых атрибутов, выделенных на основе ограничений предметной области.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:

- провести теоретические исследования в области построения схем РБД с учетом ранжируемых атрибутов;
- разработать алгоритмы, основанные на нормализации отношений для построения схем РБД **R** с учётом ранжируемых атрибутов **A**, а также оценить их временную сложность;
- разработать алгоритмы, основанные на синтезе отношений для построения схем РБД **R** с учётом ранжируемых атрибутов **A**, а также оценить их временную сложность;
- разработать алгоритм генерации предметных областей для экспериментальных исследований, направленных на оценку временных характеристик и проверку сходимости предложенных алгоритмов построения схем РБД, учитывающих ранжируемые атрибуты.

Методы исследования. Основные результаты работы получены с использованием математического аппарата реляционной алгебры, теории вероятностей, теории множеств, теории информации и теории планирования эксперимента.

Публикации. По теме диссертации опубликовано 10 работ, в том числе 6 статей, 5 из которых в журналах, рекомендованных ВАК, и 3 тезиса доклада в материалах всероссийских и международных научно-технических конференций. В Федеральной службе по интеллектуальной собственности, патентам и товарным маркам зарегистрирована одна программа для ЭВМ (свидетельство № 2012614219).

Апробация работы. Результаты диссертационной работы докладывались на 4 конференциях: 13-й всероссийской научно-технической конференции студентов, молодых ученых и специалистов «Новые информационные технологии в научных исследованиях и в образовании» (Рязань, 2008, 1 доклад), 15-й научно-технической конференции студентов,

молодых ученых и специалистов «Новые информационные технологии в научных исследованиях и в образовании» (Рязань, 2010, 1 доклад), II международной научно-технической конференции "Технологии разработки информационных систем ТРИС-2011" (Таганрог, 2011, 1 доклад), всероссийской научно-технической конференции с международным участием «Компьютерные и информационные технологии в науке, инженерии и управлении» (КомТех- 2011) (Таганрог, ТТИ ЮФУ 2011, 1 доклад).

Научная новизна диссертации заключается в том, что в ней разработаны алгоритмы построения схем РБД, позволяющие учитывать ограничения предметных областей, связанные с ранжируемостью атрибутов.

Достоверность научных положений подтверждается:

- корректным использованием математического аппарата реляционной алгебры, теории вероятностей, теории множеств, теории информации и теории планирования эксперимента;
- экспериментальными исследованиями разработанных алгоритмов построения схем РБД с помощью предложенного алгоритма генерации формализованных предметных областей.

Практическая ценность работы:

- предложены метод и алгоритмы проектирования схем РБД, позволяющие на основе выявленных ограничений предметной области строить корректные схемы, удовлетворяющие условиям нормализации и учитывающие наличие ранжируемых атрибутов;
- разработан алгоритм, позволяющий генерировать формализованные предметные области с заданными количественными характеристиками;
- на основе предложенных алгоритмов разработана программа для ЭВМ, позволяющая: строить схемы РБД с учетом ранжируемых атрибутов; автоматизировать процесс исследования алгоритмов построения схем РБД.

Основные положения, выносимые на защиту:

- алгоритмы построения схем РБД на основе нормализации отношений до третьей и четвертой нормальной формы (ЗНФ и 4НФ), отличающиеся от классического алгоритма нормализации возможностью учета ранжируемых атрибутов;
- алгоритмы построения схем РБД на основе синтеза отношений, удовлетворяющих нормальной форме Бойса - Кодда (НФБК), отличающиеся от классического алгоритма синтеза возможностью учета ранжируемых атрибутов;
- алгоритмы построения схем РБД на основе синтеза отношений, удовлетворяющих четвертой и пятой нормальной форме (4НФ и 5НФ), отличающиеся от классического алгоритма синтеза возможностью учета

многозначных зависимостей, зависимостей соединения и ранжируемых атрибутов.

Реализация и внедрение результатов работы

Результаты работы внедрены:

- в ООО «Эконом» при проектировании РБД, используемой несколькими территориально распределенными офисами;
- в ООО «ОЛВиД» при проектировании РБД с атрибутами, ранжируемыми с точки зрения конфиденциальности;
- в ООО «АРГО ТУР» при проектировании распределенной РБД;
- в учебный процесс ФГБОУ ВПО «Рязанский государственный радиотехнический университет» при обучении студентов направлений 230100 и 010500 по дисциплинам «Теория проектирования РБД», «Базы данных и СУБД» и «Программирование клиентских приложений БД».

Структура и объем работы

Диссертация состоит из введения, 4 глав, заключения и списка использованных источников из 101 наименования, изложенных на 193 страницах, содержит 35 рисунков и 22 таблицы. В приложении приведены документы о внедрении и практическом использовании результатов диссертации и свидетельство о регистрации программы для ЭВМ.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность темы, определены цели и задачи исследований.

В первой главе «Анализ алгоритмов и подходов к проектированию схем реляционных баз данных» приведён краткий обзор наиболее известных на текущий момент моделей данных: *инвертированных таблиц, иерархической модели, сетевой модели, реляционной модели, объектно-ориентированных моделей и модели данных SQL*. Для решения задачи фрагментации схем РБД **R** наиболее предпочтительна реляционная модель данных, основными преимуществами которой, в контексте решаемой задачи, являются мощный аппарат реляционной алгебры и обоснованность фрагментации (вертикальной и горизонтальной) схем РБД.

Рассмотрены алгоритмы построения схем реляционных баз данных, чаще всего применяемые на практике: *алгоритм нормализации, алгоритм синтеза, ортогональное проектирование* и способы, основанные на *семантических моделях*, таких как *ER-диаграммы* и *диаграммы классов языка UML*. Проанализированы достоинства и недостатки способов проектирования. Алгоритм синтеза обладает рядом преимуществ, в числе которых: минимальность множества отношений **R** в выходной схеме РБД **R**; полиномиальная сложность алгоритма; выходные схемы отношений удовлетворяют нормальной форме Бойса - Кодда. Рассмотрено применение

РБД в оперативной аналитической обработке данных *OLAP* (OnLine Analytical Processing) и глубинном анализе данных (Data mining).

Проведён анализ схем разграничения доступа к информации в РБД: *дискреционной схемы, мандатной схемы и ролевой схемы*. Выявлены проблемы, связанные с разграничением доступа на уровне атрибутов в случае использования СУБД, поддерживающих только дискреционную модель доступа на уровне таблиц.

Рассмотрены основные требования к проектированию распределённых РБД. Обоснована актуальность расширения функциональности алгоритмов проектирования схем РБД, в которых атрибуты могут быть разделены на группы по признакам, сформулированным на основе ограничений предметной области, не связанных с целостностью данных.

Во второй главе «Разработка алгоритмов построения схем РБД с ранжируемыми атрибутами» предложен метод построения схем РБД $R=\{R_1, R_2, \dots, R_m\}$ с учетом ранжируемых атрибутов (рисунок 1).

На основе ограничений предметной области, не связанных с целостностью данных, разработчик логической структуры РБД формулирует множество признаков $Y=\{Y_1, Y_2, \dots, Y_n\}$ и подмножества атрибутов $A_{Y_i} (i=\overline{1,n})$ с данными признаками, образующие множество $A=A_{Y_1} \cup A_{Y_2} \cup \dots \cup A_{Y_n}$.

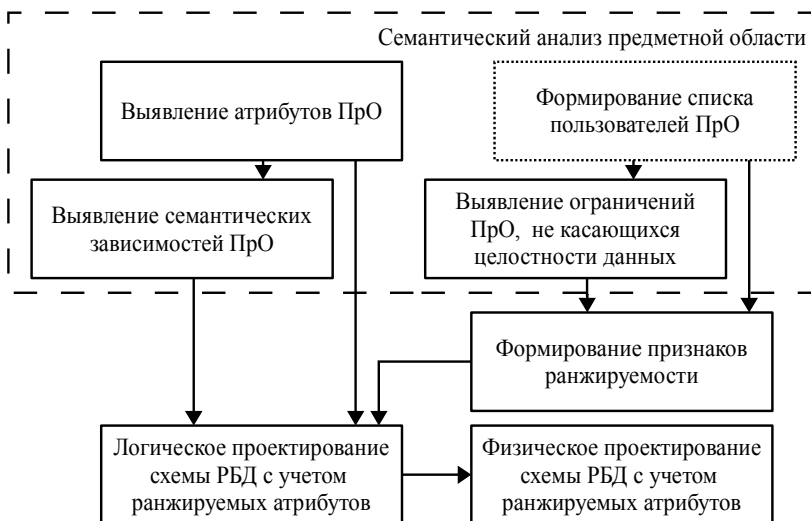


Рисунок 1 - Метод построения схем РБД с учетом ранжируемых атрибутов

Разработаны требования к формированию множества признаков Y :

- признаки должны быть сформулированы таким образом, чтобы множества атрибутов различных признаков не пересекались: $A_{Y_i} \cap A_{Y_j} = \emptyset, i \neq j, i = \overline{1, n}, j = \overline{1, n}$;

- количество признаков должно быть минимальным, но достаточным для выполнения всех ограничений предметной области, не связанных с согласованностью данных.

Разработаны алгоритмы, основанные на классических методах нормализации и синтеза, позволяющие разделять схему РБД на вертикальные фрагменты с сохранением согласованности данных. Алгоритмы разрабатывались с постепенным наращиванием мощности входных данных (входного слова параметров).

Входные и выходные данные разработанных алгоритмов

Вход: множество зависимостей U , включающее подмножества F-зависимостей F , MV-зависимостей MV и J-зависимостей J ($U = F \cup MV \cup J$); множество ранжируемых атрибутов $A = \{A_1, A_2, \dots, A_g\}$; множество признаков ранжируемости $Y = \{Y_1, Y_2, \dots, Y_n\}$ и множества атрибутов каждого признака $A_{Y_1}, A_{Y_2}, \dots, A_{Y_n}$, причем $A = A_{Y_1} \cup A_{Y_2} \cup \dots \cup A_{Y_n}$.

Выход: полная схема базы данных $R = \{R_1, R_2, \dots, R_m\}$ с учетом ранжируемых атрибутов.

Общий принцип действия разработанных алгоритмов построения схем РБД:

Шаг 1. Для каждого входного ключевого атрибута K_i схемы R вводится избыточность в виде атрибутов K_i' , таких что $K_i \rightarrow K_i'$ и $K_i' \rightarrow K_i$, то есть $K_i \leftrightarrow K_i'$; $K_i \in K$, $K_i' \in K'$, где K – множество входных ключевых атрибутов, K' – множество атрибутов, введенных для однозначного определения ключевых входных атрибутов.

Шаг 2. Обработка входного множества семантических зависимостей U , целью которой является введение новых классов эквивалентности E_F , необходимых для построения отношений R_k , отвечающих за согласованность данных при вертикальной фрагментации отношений.

Шаг 3. Построение схемы реляционной базы данных $R = \{R_1, R_2, \dots, R_k, \dots, R_m\}$ по классическим алгоритмам. Схемы отношений R_k ($k = \overline{1, m}$) приводятся к максимально возможной нормальной форме (3НФ, НФБК, 4НФ, 5НФ) с учетом найденных в результате анализа предметной области семантических зависимостей: функциональных (F-зависимостей), многозначных (MV-зависимостей) и зависимостей соединения (J-зависимостей). Это требуется для сохранения независимости проекций при дальнейшей декомпозиции отношений.

Шаг 4. Обработка полученной схемы баз данных R для учёта признаков ранжируемости Y за счет дополнительной декомпозиции

отношений R_k .

Шаг 5. Полученная схема базы данных $R=\{R_1, R_2, \dots, R_m\}$ подвергается обработке, заключающейся в задании соответствия между атрибутами K_i и K'_i в отношениях, ключевые атрибуты которых имеют разные признаки Y_j ($j=\overline{1,n}$).

На шаге 4 возможны различные варианты подмножеств ранжируемых атрибутов:

- отношение R_k содержит атрибуты A_i одного признака Y_j ;
- отношение R_k состоит из двух или более подмножеств атрибутов $A_1 = \{A_{11}, A_{12}, \dots, A_{1t}\}$, $A_2 = \{A_{21}, A_{22}, \dots, A_{2r}\}$, $\dots$, $A_p = \{A_{p1}, A_{p2}, \dots, A_{pq}\}$ (t, r и q – количество атрибутов в соответствующих множествах) разных признаков Y_j : $\{A_1, A_2, \dots, A_h, \dots, A_p\} \in R_k$, $\{A_{h1}, A_{h2}, \dots, A_{hL}\} \in A_{Y_j}$, где A_{Y_j} – подмножество атрибутов признака Y_j , L – количество атрибутов в множестве A_h , причем $A_{Y_j} \subseteq A_h$.

Если отношение содержит атрибуты двух и более признаков, то возможны различные варианты его обработки, в зависимости от вида ключа K_k отношения R_k , а также признаков ранжируемости неключевых атрибутов:

- все зависимые атрибуты одного признака; ключи отношения разных признаков;
- есть неключевые атрибуты разных признаков; ключи отношения одного признака;
- есть неключевые атрибуты разных признаков; ключи отношения разных признаков.

Введено понятие S-усиленной нормальной формы. Отношение находится в усиленной с точки зрения конфиденциальности атрибутов третьей нормальной форме (S-усиленной 3НФ или 3SNF), если оно находится в 3НФ и все неключевые атрибуты отношения относятся к одному уровню конфиденциальности.

Понятие S-усиленной нормальной формы может быть определено для других нормальных форм, в том числе для четвертой нормальной формы (4НФ), нормальной формы Бойса — Кодда (НФБК) и форм более высоких порядков.

Разработаны алгоритмы, основанные на нормализации отношений:

- алгоритм построения схем РБД на основе нормализации отношений до третьей нормальной формы, отличающийся от классического алгоритма нормализации возможностью учета ранжируемых атрибутов;
- алгоритм построения схем РБД на основе нормализации отношений до четвертой нормальной формы, отличающийся от классического алгоритма нормализации возможностью учета ранжируемых атрибутов.

Введено понятие R-усиленной нормальной формы. Отношение r со схемой R удовлетворяет R-усиленной нормальной форме, если в R содержатся атрибуты одного признака Y_j . Если отношение R нормализовано, то оно удовлетворяет соответствующей R-усиленной нормальной форме (R-усиленной 3НФ, R-усиленной 4НФ и т. д.).

Схема РБД $R = \{R_1, R_2, \dots, R_k, \dots, R_m\}$ удовлетворяет R-усиленной нормальной форме, если все схемы отношений R_k удовлетворяют R-усиленной нормальной форме.

Разработаны алгоритмы, основанные на синтезе отношений:

- алгоритм построения схем РБД R на основе синтеза отношений R_k , удовлетворяющих нормальной форме Бойса - Кодда, отличающийся от классического алгоритма синтеза возможностью учета ранжируемых атрибутов двух уровней (целесообразно использовать в случае наличия только двух уровней ранжируемости атрибутов);

- алгоритм построения схем РБД R на основе синтеза отношений R_k , удовлетворяющих четвертой нормальной форме, отличающийся от классического алгоритма синтеза возможностью учета многозначных зависимостей и ранжируемых атрибутов двух уровней;

- алгоритм построения схем РБД R на основе синтеза отношений R_k , удовлетворяющих нормальной форме Бойса - Кодда, отличающийся от классического алгоритма синтеза возможностью учета ранжируемых атрибутов. Алгоритм приведён в виде псевдокода, основанного на языке «Pascal»:

begin

Поиск неизбыточного редуцированного покрытия G для F ;

Поиск классов эквивалентности E_F для G ;

for каждая F-зависимость G_i вида $U \rightarrow V$ из $E_F(U)$ **do**

for каждая F-зависимость $G_j \neq G_i$ вида $W \rightarrow Z$ из $E_F(U)$ **do**

if U прямо определяет W **then**

begin $G = G - \{U \rightarrow V\}$; $G = G - \{W \rightarrow Z\} + \{W \rightarrow VZ\}$; **end**;

for каждый класс эквивалентности E_F **do**

$CF = CF + \{(U_1, U_2, \dots, U_f) \rightarrow V\}$;

for каждая CF-зависимость CF_i вида $(U_1, U_2, \dots, U_h, \dots, U_f) \rightarrow V$ **do**

if U_{hj} — перемещаемый атрибут в CF_i **then**

begin $U_h = U_h - U_{hj}$; $V = V + U_{hj}$; **end**;

if V_j — посторонний атрибут в CF_i **then** $V = V - V_j$;

Построение схемы базы данных $R = \{R_1, R_2, \dots, R_k, \dots, R_m\}$;

Обработка ранжируемых атрибутов каждой схемы R_k из R ;

end.

Где U, V, W и Z — подмножества входного множества атрибутов;

– алгоритм построения схем РБД **R** на основе синтеза отношений R_k , удовлетворяющих пятой нормальной форме, отличающийся от классического алгоритма синтеза возможностью учета ранжируемых атрибутов, а также многозначных зависимостей и зависимостей соединения:

begin

Поиск избыточного редуцированного покрытия **G** для **F**;

Поиск классов эквивалентности E_F для **G**;

Минимизация покрытия **G**;

Построение множества CF-зависимостей **CF**;

Редуцирование множества CF-зависимостей **CF**;

Построение схемы базы данных $R = \{R_1, R_2, \dots, R_k, \dots, R_m\}$;

for каждая схема R_k из **R** **do**

for каждая MV-зависимость MV_i вида $U \rightarrow \rightarrow V|Z$ из **MV** **do**

if MV_i применима к R_k **then** $R = R - R_k + \{UV, UZ\}$;

for каждая схема R_k из **R** **do**

for каждая J-зависимость J_i вида $*(U, V, \dots, Z)$ из **J** **do**

if J_i применима к R_k **then** $R = R - R_k + \{U, V, \dots, Z\}$;

Обработка ранжируемых атрибутов каждой схемы R_k из **R**;

end.

Предложенные алгоритмы, основанные на синтезе отношений, обладают полиномиальной временной сложностью.

В третьей главе «Генерация формализованной предметной области» формулируется задача предоставления входных данных для алгоритмов построения схем РБД, решение которой необходимо для проведения экспериментальных исследований, направленных на оценку временных характеристик и проверку сходимости разработанных во **второй главе** алгоритмов. Для решения поставленной задачи предлагается алгоритм генерации предметных областей с различными количественными характеристиками.

Дается обзор основных (в контексте решаемой задачи) параметров предметных областей и формулируются требования к алгоритму генерации предметных областей:

– сгенерированный набор данных должен включать множество атрибутов $A = \{A_1, A_2, \dots, A_g\}$; множество семантических зависимостей **U**, включающих данные атрибуты, состоящее из подмножеств F-зависимостей **F**, MV-зависимостей **MV** и J-зависимостей **J**, причём $U = F \cup MV \cup J$; множество признаков ранжируемости $Y = \{Y_1, Y_2, \dots, Y_n\}$ и соответствующие множества атрибутов каждого признака $A_{Y_1}, A_{Y_2}, \dots, A_{Y_n}$, причём $A = A_{Y_1} \cup A_{Y_2} \cup \dots \cup A_{Y_n}$;

– допускается наличие атрибутов, не входящих в какие-либо

зависимости;

– должна иметься возможность управления количественными характеристиками генерируемых формализованных предметных областей.

В соответствии с принципами теории планирования эксперимента, касающимися рандомизации и отказа от полного перебора возможных уровней факторов эксперимента, в качестве инструмента рандомизации количественных параметров формализованной предметной области используются законы распределения случайной величины. Дается краткий обзор наиболее распространенных законов распределения случайной величины и обоснование выбора равномерного, нормального, показательного законов и закона Рэлея для задания количественных параметров генерируемых предметных областей.

Для реализации псевдослучайных последовательностей выбранных законов распределения используются обратные функции (преобразование Н.В. Смирнова) с применением стандартной равномерной последовательности [для показательного распределения по формуле (1), нормального распределения по формуле (2) и распределения Рэлея по формуле (3)]:

$$x_i = \frac{-1}{\lambda} \ln r_i, \quad (1)$$

где x_i - реализация псевдослучайной величины, λ — параметр показательного распределения, r_i - реализация случайной величины, распределённой равномерно на интервале (0,1);

$$x_i = \sqrt{\frac{12}{n}} \left(\sum_{i=1}^n r_i - \frac{n}{2} \right); \quad (2)$$

$$x_i = a \sqrt{-2 \ln r_i}, \quad (3)$$

где n – параметр генерирования случайных чисел, a – параметр масштаба.

На основе описанных способов получения псевдослучайных последовательностей чисел с различными законами распределения предлагается алгоритм генерации формализованной предметной области с заданными параметрами.

Входные и выходные данные алгоритма генерации предметных областей

Вход: параметры законов распределения для задания количественных характеристик генерируемой формализованной предметной области.

Выход:

- множество атрибутов $A = \{A_1, A_2, \dots, A_g\}$;
- множество признаков ранжируемости $Y = \{Y_1, Y_2, \dots, Y_n\}$ и соответствующие множества атрибутов каждого признака

$A_{Y_1}, A_{Y_2}, \dots, A_{Y_n}$, причем $A = A_{Y_1} \cup A_{Y_2} \cup \dots \cup A_{Y_n}$;

– множество семантических зависимостей U , состоящее из подмножеств F-зависимостей F , MV-зависимостей MV и J-зависимостей J , причём $U = F \cup MV \cup J$ (F, MV, J — непересекающиеся множества семантических зависимостей).

Алгоритм заключается в генерации множеств атрибутов и признаков ранжируемости предметной области, на основе которых формируются множества зависимостей (функциональных, многозначных и зависимостей соединения) и используется в четвертой главе для экспериментальных исследований, направленных на оценку временных характеристик и проверку сходимости разработанных алгоритмов построения схем РБД.

Временная сложность предложенного алгоритма генерации является полиномиальной.

В четвертой главе «Экспериментальные исследования разработанных алгоритмов» с использованием основных принципов теории планирования эксперимента формулируются цели исследований:

– оценка корректности синтезированных схем реляционных баз данных, включающая:

- проверку синтезированной схемы РБД на удовлетворение условиям нормализации;
- проверку синтезированной схемы РБД на удовлетворение условиям наличия ранжируемых атрибутов;
- оценка временных характеристик исследуемых алгоритмов в зависимости от различных характеристик входных данных;
- сравнение разработанных алгоритмов, основанных на синтезе отношений, с классическим алгоритмом синтеза с точки зрения временных затрат;
- сравнение длительности обработки запросов к спроектированным схемам РБД с учётом ранжируемых атрибутов и без него.

Исходя из поставленных целей формулируются отклики экспериментов, которые можно разделить на качественные и количественные.

Планируются и проводятся эксперименты, направленные на получение оценок вероятностных характеристик временных затрат разработанных алгоритмов в зависимости от вида распределения количественных параметров входных предметных областей. Число опытов N , необходимых для соблюдения доверительной вероятности $\beta = 0.95$ ($t_\beta = 1.960$), рассчитывается по формуле (4):

$$N \geq \frac{\tilde{D} \cdot t_\beta^2}{\varepsilon^2}, \quad (4)$$

где $\tilde{D}$ - оценка дисперсии временных характеристик алгоритмов, которая рассчитывается по формуле (5):

$$\tilde{D} = \frac{\sum_{i=1}^N (X_i - \tilde{\mu})^2}{N - 1}. \quad (5)$$

Погрешность ε выбирается равной 5 % от значения оценки математического ожидания $\tilde{\mu}$ количественных откликов эксперимента, рассчитываемого по формуле (6):

$$\tilde{\mu} = \frac{\sum_{i=1}^N X_i}{N}. \quad (6)$$

Проводится оценка доверительных интервалов вероятностных характеристик μ и D временных затрат разработанных алгоритмов и исследуются зависимости временных характеристик от каждого фактора эксперимента.

На рисунке 2 приведены диаграммы зависимостей временных характеристик алгоритма построения схем РБД на основе синтеза отношений с учетом ранжируемых атрибутов от значений количественных параметров генерируемых предметных областей при равномерном распределении уровней факторов эксперимента.

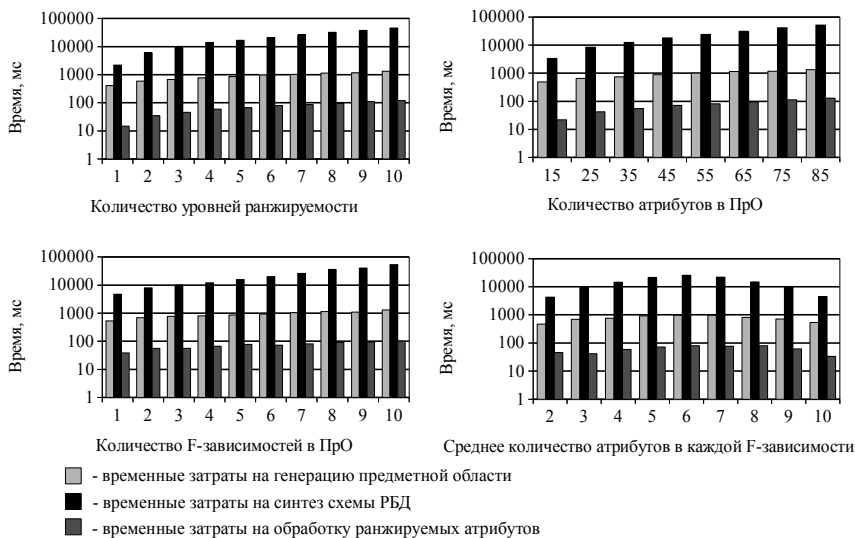


Рисунок 2 — Результаты эксперимента с равномерным распределением уровней факторов

По результатам проведённых экспериментов сделаны выводы:

- предложенные алгоритмы построения схем РБД с учетом ранжируемых атрибутов позволяют строить корректные схемы баз данных **R**, удовлетворяющие как условиям нормализации, так и условиям ранжируемости атрибутов;
- временные затраты на построение схем РБД с использованием разработанных алгоритмов являются приемлемыми в контексте решаемой задачи, так как другие этапы проектирования информационных систем могут составлять дни и недели;
- с точки зрения использования принципа рандомизации при задании уровней факторов наиболее предпочтительным является равномерный закон распределения, так как при условии достаточного числа опытов позволяет проработать наибольшее количество комбинаций уровней факторов эксперимента.

Для проведения эксперимента, направленного на исследования временных затрат на запросы к спроектированным схемам РБД, генерируются тестовые предметные области, на основе которых проектируются схемы РБД по классическим и разработанным алгоритмам, содержащие 1000, 5000, 25000, 50000 и 100000 кортежей. Из результатов эксперимента на выборку видно, что с ростом количества хранимых кортежей время, затрачиваемое на соединение таблиц в случае более сильной фрагментации схемы БД, растёт медленнее времени, затрачиваемого на выборку данных (рисунок 3).

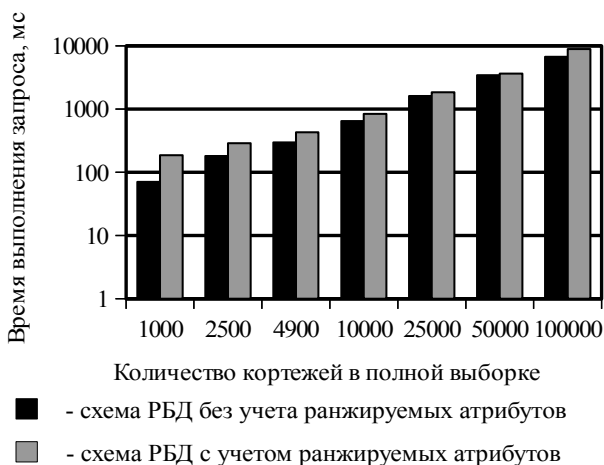


Рисунок 3 — Временные затраты на полную выборку данных из РБД

В **заключении** подведены итоги проведённой работы и сформулированы основные научные и практические результаты.

Прилагаются копии актов о внедрении результатов работы и данные о регистрации программы для ЭВМ.

Основные результаты работы

- Проведён анализ алгоритмов и подходов построения схем РБД с учетом ранжируемых атрибутов.
- Сформулирована задача расширения функциональности алгоритмов построения схем РБД для учета ранжируемых атрибутов.
- Разработаны алгоритмы, основанные на нормализации отношений, учитывающие наличие ранжируемых атрибутов в предметной области.
- Разработаны алгоритмы, основанные на синтезе отношений, учитывающие наличие ранжируемых атрибутов в предметной области.
- Проведена оценка временной сложности разработанных алгоритмов и доказана их сходимость.
- Проведён анализ задачи генерирования формализованных предметных областей, необходимых для экспериментальных исследований алгоритмов построения схем РБД.
- Разработан алгоритм, позволяющий генерировать формализованные предметные области с заданными количественными характеристиками и автоматизировать процесс исследования временных характеристик алгоритмов построения схем РБД.
- На основе предложенных алгоритмов разработана программа для ЭВМ, позволяющая строить схемы РБД с учетом наличия ранжируемых атрибутов в предметной области.

Публикации по теме диссертации

Работы, опубликованные в научных журналах, входящих в перечень ведущих рецензируемых журналов и изданий ВАК РФ

1. Баранчиков А.И., Громов А.Ю. Построение схем реляционных баз данных в четвертой нормальной форме для информационных систем с конфиденциальной информацией // Известия ЮФУ. Технические науки. Тематический выпуск «Компьютерные и информационные технологии в науке, инженерии и управлении». №5(118). 2011 г. С. 65-69.

2. Баранчиков А. И., Громов А. Ю. Алгоритм синтеза реляционной базы данных, учитывающий атрибуты различной степени секретности // Системы управления и информационные технологии: науч.-техн. журн. N3(37). Москва - Воронеж, 2009. С. 25 – 27.

3. Баранчиков А. И., Громов А. Ю. Алгоритм построения схемы реляционной базы данных на основе множества многозначных и функциональных зависимостей, учитывающий атрибуты различной

степени секретности // Системы управления и информационные технологии: науч.-техн. журн. N1(43). Москва - Воронеж, 2011. С. 53 – 56.

4. Баранчиков А.И., Громов А.Ю. Построение схем реляционных баз данных с n-ранжируемыми атрибутами // Системы управления и информационные технологии: науч.-техн. журн. N2.1(48). Москва - Воронеж, 2012. С.114-117.

5. Баранчиков А.И., Громов А.Ю. Алгоритм генерации формализованной модели предметной области // Вестник РГРТУ: науч.-техн. журн. №3 (выпуск 33). Рязань, 2010. С. 45-49.

Работы, опубликованные в межвузовских сборниках научных трудов

6. Баранчиков А.И., Громов А.Ю. Алгоритм построения схемы реляционной базы данных, содержащей атрибуты различной степени секретности // Информатика и прикладная математика. Рязан. гос. ун-т им. С.А. Есенина, 2008. С. 7-12.

Работы, опубликованные в сборниках научных трудов международных и всероссийских конференций

7. Баранчиков А.И., Громов А.Ю. Алгоритм синтеза схемы реляционной базы данных, содержащей атрибуты различной степени секретности // Новые информационные технологии в научных исследованиях и в образовании: тез. докл. 13-й Всеросс. науч.-техн. конф. Рязан. гос. радиотехн. ун-т. Рязань, 2008. С. 125-126.

8. Баранчиков А.И., Громов А.Ю. Алгоритм построения формализованной модели предметной области // Новые информационные технологии в научных исследованиях и в образовании: тез. докл. 15-й науч.-техн. конф. Рязан. гос. радиотехн. ун-т. Рязань, 2010. С. 66-67.

9. Баранчиков А.И., Громов А.Ю. Алгоритм построения схем реляционных баз данных на основе многозначных и функциональных зависимостей с n-ранжируемой конфиденциальностью атрибутов // Технологии разработки информационных систем ТРИС-2011: тез. докл. II Междунар. науч.-техн. конф. Таганрог, 2011. С. 113.

Регистрация программы для ЭВМ

10. Громов А.Ю., Баранчиков А.И. Свидетельство о государственной регистрации программы для ЭВМ «SynthesisFR» №2012614219 // Федеральная служба по интеллектуальной собственности, патентам и товарным маркам. Зарегистрировано в Реестре программ для ЭВМ 12 мая 2012 г.

Громов Алексей Юрьевич

**АЛГОРИТМЫ ПРОЕКТИРОВАНИЯ СХЕМ РЕЛЯЦИОННЫХ БАЗ
ДАННЫХ, СОДЕРЖАЩИХ РАНЖИРУЕМЫЕ АТТРИБУТЫ**

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать 03.10.12. Формат бумаги 60x84 1/16.
Бумага офсетная. Печать трафаретная. Усл. печ. л. 1,0.
Тираж 100 экз.
Рязанский государственный радиотехнический университет.
390005, Рязань, ул. Гагарина, 59/1.
Редакционно-издательский центр РГРТУ.