На правах рукописи

Гудков Кирилл Сергеевич

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ УПРАВЛЕНИЯ ОБРАБОТКОЙ ИНФОРМАЦИИ В КОРПОРАТИВНЫХ АВТОМАТИЗИРОВАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Специальность 05.13.18 – математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2012

Работа выполнена на кафедре управляющих и информационных систем Московского физико-технического института (государственного университета)

Научный руководитель: доктор физико-математических наук, профессор БОНДАРЕНКО Александр Викторович

Официальные оппоненты: ВИЗИЛЬТЕР Юрий Валентинович, доктор физико-математических наук, ст. н.с., подразделение 3000 Государственного научно-исследовательского института авиационных систем, начальник подразделения

БОНДАРЕВ Александр Евгеньевич, кандидат физико-математических наук, Институт прикладной математики им. М.В. Келдыша РАН, старший научный сотрудник

Ведущая организация: Вычислительный центр имени А.А. Дородницына РАН

Защита диссертации состоится «_____» _____ 2012 года в _____ ч. _____ мин. на заседании диссертационного совета Д 212.156.05 при Московском физико-техническом институте (государственном университете) по адресу: 141700, Московская область, г. Долгопрудный, Институтский пер., д. 9, ауд. 903 кпм.

С диссертацией можно ознакомиться в библиотеке Московского физико-технического института (государственного университета).

Автореферат разослан «_____» _____ 2012 года

Ученый секретарь
диссертационного совета
Д 212.156.05

Федько Ольга Сергеевна

Общая характеристика работы

В работе проводится комплексное исследование проблемы управления обработкой нормативно-справочной информации с применением современной технологии математического моделирования и вычислительного эксперимента. Строятся математические модели для импорта внешних справочников, интеграции внутрикорпоративных справочников и тиражирования справочников. Доказанные в рамках моделей теоремы иллюстрируются вычислительными экспериментами и используются при разработке алгоритмов и комплекса программ. Комплекс программ для решения задачи управления обработкой нормативно-справочной информации включает автоматизированную систему импорта данных из внешних источников, интегратор справочников внутренних источников, систему репликации баз данных. В ходе решения задачи импорта данных был разработан, обоснован и протестирован с применением современных компьютерных технологий эффективный вычислительный метод нахождения изменений между различными версиями справочников на основе красно-чёрных деревьев. В ходе решения прикладной проблемы управления обработкой нормативно-справочной информации были применены математическое моделирование, численные методы и комплексы программ.

Актуальность темы. В настоящее время корпоративные автоматизированные информационные системы, как правило, не обходятся без использования нормативно-справочной информации. Существует три метода её хранения: централизованный, децентрализованный и смешанный. В последние годы сформировалась устойчивая тенденция к отделению функциональности по управлению обработкой нормативно-справочной информации от функциональности автоматизированных информационных систем по управлению обработкой прочих корпоративных данных. Для

государственных ведомств и крупных корпораций характерно наличие центральной базы данных и разветвлённой структуры дочерних баз данных, расположенных на территориально удалённых участках корпоративной автоматизированной информационной системы. Именно централизованная структура хранения нормативно-справочной информации рассматривается в диссертационной работе. Справочники формируются в консолидированной базе данных нормативно-справочной информации, откуда тиражируются в дочерние базы данных территориально удалённых участков корпоративной автоматизированной информационной системы.

В качестве технологии тиражирования нормативно-справочной информации используется система репликации баз данных, способная функционировать в гетерогенной среде. Специфика нормативно-справочной информации накладывает дополнительные требования на системы синхронизации данных, которые на сегодняшний день не были всесторонне исследованы. Исследование этого вопроса позволит выбрать подходящую для тиражирования нормативно-справочной информации систему репликации баз данных применительно к гетерогенной среде.

Использование системы репликации позволяет обеспечить согласованность данных между консолидированной базой данных нормативно-справочной информации и дочерними базами данных территориально удалённых участков корпоративной автоматизированной информационной системы. Поэтому для актуальности, полноты и непротиворечивости нормативно-справочной информации в корпоративной автоматизированной информационной системе необходимо обеспечить её актуальность, полноту и непротиворечивость в консолидированной базе данных нормативно-справочной информации. Решение этой задачи зависит от источника справочников. При формировании справочников консолидированной базы данных нормативно-справочной информации на основе справочников из внешних источников возникают вопросы, связанные с выбором подходящих справочников из внешних источников, поддержки

синхронизации данных консолидированной базы данных с ними, модификации их структуры для соответствия принятым корпоративным стандартам. В настоящее время не существует чётких механизмов решения перечисленных задач, а одним из частых подходов является адаптация справочников корпоративной автоматизированной информационной системы к справочникам из внешних источников, а не наоборот. Разработка математической модели импорта справочников из внешних источников и реализация на её основе комплекса программ позволят упростить управление импортом нормативно-справочной информации. В случае формирования справочников консолидированной базы данных нормативно-справочной информации на основе справочников внутренних источников возникают вопросы, связанные со слиянием содержащейся в них нормативно-справочной информации и с устранением противоречий в данных. Полная автоматизация этих процессов вряд ли возможна, но частичная автоматизация для справочников определённого вида позволяет упростить управление обработкой нормативно-справочной информации.

Предлагаемые в диссертационной работе методы для решения перечисленных проблем могут применяться при первичном внедрении системы управления обработкой нормативно-справочной информации, при слиянии и поглощении корпоративных информационных систем.

Применение рекомендуемых в диссертационной работе подходов позволит:

- избежать финансовых потерь, связанных с неактуальностью, противоречивостью и неполнотой данных;
- построить отчётность, соответствующую предъявляемым к ней требованиям достоверности и актуальности;
- принимать на основе этой отчётности правильные управленческие решения;
- повысить интеграцию бизнес-процессов.

Цель работы. Целью диссертационной работы является решение проблемы управления обработкой нормативно-справочной информации в корпоративных автоматизированных информационных системах с помощью математического моделирования, численных методов и комплексов программ.

Задачи исследования. Основные задачи диссертационной работы:

1. создание математических моделей для управления обработкой нормативно-справочной информации в корпоративных автоматизированных информационных системах;
2. выявление с помощью моделей соответствующих характеристик вычислительных алгоритмов и комплексов программ;
3. разработка комплекса программ для интеграции справочников внутренних источников, импорта справочников из внешних источников, тиражирования информации между территориально удалёнными участками корпоративной автоматизированной информационной системы.

Методы исследования. В работе использовались методы теории баз данных, реляционной алгебры, теории репликации баз данных, теории множеств, теории графов, вычислительной математики и прикладной математической статистики.

Научная новизна полученных результатов. Научная новизна диссертационного исследования состоит в следующем:

1. Для управления обработкой нормативно-справочной информации в корпоративных автоматизированных информационных системах предложены новые математические модели на основе реляционной алгебры, позволяющие осуществлять операции импорта и экспорта нормативно-справочной информации, обеспечивая согласованность, актуальность и полноту данных.
2. Разработан метод импорта данных в консолидированную базу данных нормативно-справочной информации, использующий двухступенчатый

механизм переноса данных, когда вначале выделяются требуемые данные справочников из внешних источников и они представляются в промежуточном формате, а затем осуществляется перенос данных в консолидированную базу данных нормативно-справочной информации, что позволяет упростить процесс согласования форматов данных.

3. Разработан вычислительный метод нахождения изменений между версиями справочников, использующий красно-чёрные деревья и позволяющий обеспечивать более высокую скорость поиска изменений по сравнению с другими известными методами.

Практическая значимость исследования. Созданные модели, алгоритмы и программное обеспечение могут быть использованы для импорта справочников из внешних источников, интеграции справочников внутренних источников, а также для тиражирования справочников в любой корпоративной автоматизированной информационной системе.

Положения, выносимые на защиту.

1. Математические модели управления обработкой нормативно-справочной информации в корпоративных автоматизированных информационных системах. Указанные модели обеспечивают согласованность, актуальность и полноту данных при решении задач импорта и экспорта нормативно-справочной информации.
2. Метод, алгоритм и программная реализация импорта данных в консолидированную базу данных нормативно-справочной информации. Указанный метод позволяет упростить процесс согласования форматов данных справочников из внешних источников и консолидированной базы данных нормативно-справочной информации.
3. Метод на основе красно-чёрных деревьев, его алгоритм и программная реализация, а также результаты вычислительных экспериментов для поиска различий между версиями справочников. Указанный метод

работает на 15% быстрее метода, использующего AVL-деревья, и до трёх раз быстрее метода, использующего хэш-таблицы.

Апробация работы. Основные результаты работы докладывались, обсуждались и получили одобрение специалистов на следующих конференциях:

- L, LI, LII научных конференциях Московского физико-технического института (государственного университета), (Долгопрудный, 2007, 2008, 2009),
- XVI международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов-2009», (Москва, МГУ, 2009),
- VI международной научно-практической конференции «Ключевые проблемы современной науки – 2010», (Болгария, София, 2010),
- VII международной научно-практической конференции «Актуальные научные достижения – 2011», (Чехия, Прага, 2011),
- юбилейной всероссийской научно-технической конференции «Моделирование авиационных систем», (Москва, 2011),
- а также на научных семинарах базовой кафедры МФТИ «Управляющие и информационные системы», научных семинарах ВЦ РАН и на научно-техническом совете ФГУП «ГосНИИАС» (Москва, 2011-2012).

Доклады на L и LI научных конференциях МФТИ, как лучшие в секции, были отмечены дипломами победителя.

Публикации. Основные положения работы отражены в 11 публикациях, в том числе двух, [6, 7], в издании из списка, рекомендованного ВАК РФ.

Структура и объём диссертации. Диссертация состоит из введения, шести глав, заключения и списка использованных источников. Объём работы составляет 133 страницы. Список использованных источников содержит 92 наименования.

Краткое содержание работы

Во введении даётся общая характеристика работы.

В главе 1 рассматриваются четыре направления в управлении обработкой нормативно-справочной информации: хранение и использование нормативно-справочной информации; выбор структуры справочников; заполнение справочников; тиражирование нормативно-справочной информации. Даётся обзор методов хранения информации. Обосновывается выбор реляционных систем управления базами данных для управления консолидированной базой данных нормативно-справочной информации и дочерними базами данных территориально удалённых участков корпоративной автоматизированной информационной системы. Формулируются задачи, решение которых возможно благодаря выбору структуры справочников: поддержка иерархии данных в реляционных базах данных, поддержка репликации данных, поддержка исторических данных. Рассматривается вопрос выбора источников заполнения консолидированной базы данных нормативно-справочной информации (КБД НСИ). На основе анализа публикаций предложены следующие рекомендации по управлению обработкой нормативно-справочной информации с точки зрения наполнения данных:

1. Корпоративные справочники должны формироваться на основании данных открытых внешних источников во всех случаях, когда эти данные удовлетворяют корпоративным требованиям надёжности, актуальности и полноты.
2. Внешние справочники должны пройти предварительную обработку, чтобы их структура соответствовала потребностям корпоративной автоматизированной информационной системы.
3. Однотипные внутренние справочники подразделений предприятия должны быть объединены с устранением существующих противоречий

между ними. После объединения подразделения предприятия должны полностью прекратить использование прежних версий справочников и перейти к использованию объединённых справочников.

Проводится сравнительный анализ существующих систем репликации данных с точки зрения их применимости к тиражированию нормативно-справочной информации. В конце главы приводится краткое описание реляционной алгебры – математического аппарата, который используется при создании собственных математических моделей управления обработкой нормативно-справочной информации в корпоративных автоматизированных информационных системах.

В главе 2 рассматривается задача переноса данных справочников из внешних источников в таблицы консолидированной базы данных нормативно-справочной информации. В общем случае её решение состоит из шести этапов:

1. загрузки справочников из открытых источников;
2. разархивации данных;
3. преобразования форматов данных внешних источников к промежуточному формату данных, который используется автоматизированной системой импорта данных;
4. выделения изменений в справочниках из внешнего источника, произошедших между соседними сеансами синхронизации данных;
5. подготовки изменений к переносу в консолидированную базу данных нормативно-справочной информации;
6. переноса изменений.

На первых двух этапах формируется множество OD справочников из внешних источников. В консолидированной базе данных используется лишь часть этих справочников $SOD = \{od_1,...,od_N\} \subseteq OD$. На основе справочников из множества SOD формируется множество справочников в промежуточном формате $ID = \{id_1,...,id_{4N}\} = ID_{new} \cup ID_{old} \cup ID_{changes}$. В множестве ID_{new}

содержится текущая версия справочников, а в множестве ID_{old} содержится сохранённая предыдущая версия справочников. Структура справочников совпадает со структурой таблиц консолидированной базы данных нормативно-справочной информации, наполнение справочников совпадает с наполнением справочников из внешних источников. Множество $ID_{changes}$ разбивается на два подмножества: $\begin{cases} ID_{changes} = ID_{changesadd} \cup ID_{changesdel} \\ ID_{changesadd} \cap ID_{changesdel} = \varnothing \end{cases}$, где $\begin{cases} idcd_n \in ID_{changesdel} \\ idcd_n = ido_n - idn_n \forall n \end{cases}$ и $\begin{cases} idca_n \in ID_{changesadd} \\ idca_n = idn_n - ido_n \forall n \end{cases}$. Таким образом, множество $ID_{changes}$ содержит изменения, которые необходимо внести в консолидированную базу данных нормативно-справочной информации. Множество справочников консолидированной базы данных $RCDB$ разбивается на два подмножества: $\begin{cases} RCDB = DRCDB \cup CRCDB \\ DRCDB \cap CRCDB = \varnothing \end{cases}$, где $DRCDB = \{dr_1, ..., dr_K\}$ - справочники, формируемые на основе внешних данных, содержащихся в множестве ID, а $CRCDB = \{cr_1, ..., cr_K\}$ - вспомогательные для репликации таблицы.

Посредством l^k обозначено число справочников внешних источников, отвечающих заданному справочнику $dr_K \in DRCDB$. Тогда для синхронизации данных между внешним источником и консолидированной базой данных необходимо для каждого dr_K выполнить:

1. Перенос l^k-справочников в промежуточном формате из множества ID_{new} в множество ID_{old}. Функция, выполняющая данный этап, обозначена $F_{Manipulate} : ID_{new} \to ID_{old}$.
2. Преобразование l^k-справочников из множества SOD к промежуточному формату. Результаты фиксируются в множестве ID_{new}. Функция, выполняющая данный этап, обозначена $F_{OI} : SOD \to ID_{new}$.

3. Сравнение l^k-справочников из множества ID_{new} с l^k-справочниками из множества ID_{old}. Результаты фиксируются в множестве $ID_{changes}$. Функция, выполняющая данный этап, обозначена $F_{GetChanges} : HID \to HID_{changes}$, где множество $HID \subseteq ID_{new} \times ID_{old}$ состоит из пар $(idn_n, ido_n) \leftrightarrow od_n$, а множество $HID_{changes} \subseteq ID_{changesadd} \times ID_{changesdel}$ состоит из пар $(idca_n, idcd_n) \leftrightarrow od_n$.
4. Модификация справочника dr_k в соответствии с $2l^k$-справочниками из множества $ID_{changes}$. Результаты фиксируются в множестве $DRCDB$. Множество $SID_{changes} = \{sid_1, ..., sid_K\} \subseteq 2^{HID_{changes}}$ включает в себя подмножества пар изменений, отвечающие заданным справочникам dr_k. Множество $CIDDR$ определено следующим образом: $CIDDR \subseteq SID_{changes} \times DRCDB$. Каждая тройка в составе $CIDDR$ включает в себя справочник консолидированной базы данных и изменения, которые над ним необходимо сделать. С учётом введённых обозначений, функция, выполняющая данный этап, определена следующим образом: $F_{IR} : CIDDR \to DRCDB$.

Функция, выполняющая задачу синхронизации справочников внешних источников и справочников консолидированной базы данных, может быть представлена в виде суммы:

$$F = \sum_k \sum_l F_{Manipulate}(idn_l) * F_{OI}(sod_l) * F_{GetChanges}(idn_l, ido_l) * F_{IR}(idca_l, idcd_l, dr_k).$$

Далее в главе 2 математическая модель иллюстрируется на примере импорта данных российского административно-территориального деления. После этого рассматриваются особенности реализации функций $F_{Manipulate}$, F_{OI}, $F_{GetChanges}$ и F_{IR} в рамках построения автоматизированной системы импорта данных из внешних источников. Вычисление $F_{Manipulate}$ представляет

лишь технические сложности. Вычисление F_{OI} не может быть полностью автоматизировано. Шаги, предпринятые к частичной автоматизации, описаны в тексте диссертационной работы. Приведём алгоритм вычисления функции $F_{GetChanges}$:

1. Данные из справочника $ido \in ID_{old}$ заносятся в красно-чёрное дерево.
2. Осуществляется линейный проход по кортежам справочника $idn \in ID_{new}$. Для каждого кортежа проверяется, содержится ли он в красно-чёрном дереве. Если нет, то он присоединяется к справочнику $idca \in ID_{changesadd}$.
3. Данные из справочника $idn \in ID_{new}$ заносятся в красно-чёрное дерево.
4. Осуществляется линейный проход по кортежам справочника $ido \in ID_{old}$. Для каждого кортежа проверяется, содержится ли он в красно-чёрном дереве. Если нет, то он присоединяется к справочнику $idcd \in ID_{changesdel}$.

Автоматизированная система импорта данных из внешних источников позволяет использовать 3 режима вычисления F_{IR}, каждый из которых может оказаться полезным для конкретной практической задачи:

1. по файлам изменений создаются SQL-сценарии для добавления или удаления записей, которые затем выполняются;
2. по мере обработки файлов изменений SQL-сценарии создаются в памяти и выполняются;
3. перенос данных осуществляется при помощи сервисов Microsoft SQL Server (DTS, SSIS).

В главе 3 рассматривается задача интеграции справочников внутренних источников и формирования на их основе таблиц консолидированной базы данных нормативно-справочной информации. Выводится общая формула для объединения справочников с совпадающими естественными первичными ключами и различным списком атрибутов: $T(A_1,...,A_n,B_1,...,B_m,C_1,...,C_k) = \sigma_{C_1}(R) \cup \sigma_{C_2}(\rho_{S(A_1,...,A_n,C_1,...,C_k)}(S)) \cup \sigma_{C_3}($

$R(A_1,...,A_n,B_1,...,B_m) \times \rho_{S(A_1,...,A_n,C_1,...,C_k)}(S))$, где $C_1 = (a_1...a_n \notin S)$, $C_2 = (a_1...a_n \notin R)$, $C_3 = (R.a_1...a_n = S.a_1...a_n)$. Далее рассматривается обобщение на случай применения суррогатных первичных ключей при наличии атрибутов, которые можно использовать в качестве естественных первичных ключей. В конце главы обсуждается решение проблем, связанных с наличием противоречий в исходных данных.

В главе 4 рассматривается задача тиражирования изменений, произошедших в консолидированной базе данных нормативно-справочной информации, в базы данных территориально удалённых участков корпоративной автоматизированной информационной системы.

Корпоративной автоматизированной информационной системе ставится в соответствие ориентированный граф $G(V,E)$, где V - это множество участков информационной системы, а E - множество каналов связи между ними. Рассматривается информационная система, имеющая «звёздную топологию»:

$(\exists! v^0 \in V : \forall v \neq v^0 \rightarrow (v^0, v) \in E) \,\&\, (\forall v^1 \neq v^0 \forall v^2 \neq v^0 \rightarrow (((v^1, v^2) \notin E) \,\&\, ((v^2, v^1) \notin E)))$.

v^0 - это участок с консолидированной базой данных, $v^i, i \neq 0$ - территориально удалённый участок корпоративной автоматизированной информационной системы.

Множества $TOAD^i = \{td_0^i, td_1^i ..., td_{P^i}^i, ...\}$ содержат времена согласования данных участка с консолидированной базой данных v^0 и территориально удалённого участка $v^i, i \neq 0$. Множество $TOA = \{t_0, t_1, ..., t_J, ...\}$ содержит времена синхронизации консолидированной базы данных и внешних источников. Взаимное расположение этих времён можно представить следующим образом:

$t_0, ..., t_{j_0}, td_0^i, t_{j_0+1}, ..., t_{j_1}, td_1^i, t_{j_1+1}, ..., t_{j_p}, td_p^i, t_{j_p+1}, ..., t_{j_P}, td_P^i, t_{j_P+1}, ..., t_J, ...$ Функции из множеств $SFC^i = \{SFC_1^i, ..., SFC_K^i\}$ выполняют изменения в справочниках из

множества $DRCDB$ в интервале времени (td_p^i, td_{p+1}^i). Любая из этих функций может быть представлена следующим образом: $SFC_k^i = sfc_1^i * ... * sfc_{j_{p+1}-j_p}^i$. Если $j_{p+1} = j_p$, то SFC_k^i - тождественное преобразование. Функции из множеств $SFCS^i = \{SFCS_1^i, SFCS_2^i, ..., SFCS_{M_i}^i\}$ выполняют необходимые для синхронности данных изменения в справочниках s_m^i территориально удалённого участка v^i. Любая из этих функций может быть представлена следующим образом: $SFCS_m^i = sfcs_1^i * sfcs_2^i * ... * sfcs_{j_{p+1}-j_p}^i$. Если $j_{p+1} = j_p$, то $SFCS_m^i$ - тождественное преобразование. Связь между справочниками в смежные моменты синхронизации выражается следующим образом:

$$\begin{cases} dr_k(td_{p+1}) = SFC_k^i(dr_k(td_p)) = sfc_1^i * ... * sfc_{j_{p+1}-j_p}^i * dr_k(td_p) \\ s_m^i(td_{p+1}) = SFCS_m^i(s_m^i(td_p)) = sfcs_1^i * sfcs_2^i * ... * sfcs_{j_{p+1}-j_p}^i * s_m^i(td_p) \end{cases}$$

Любая из функций sfc и $sfcs$ производит два типа изменений над справочниками: добавление кортежа и удаление кортежа. Модификация кортежа – это суперпозиция перечисленных операций. Поэтому любая из функций sfc и $sfcs$ изоморфна отношению, в котором к столбцам операнда добавлен ещё один целочисленный столбец, означающий тип операции. Справедлива теорема 4.1 об изоморфизме.

Теорема 4.1. Изменения, происходящие в справочниках в составе КБД НСИ, могут быть представлены в реляционных таблицах.

Предлагается следующий порядок тиражирования информации:

1. Изменения SFC_k^i, происходящие в консолидированной базе данных, отражаются в изоморфных им справочниках $cr_k \in CRCDB$. В конкретной реализации системы репликации применительно к гетерогенной среде для этого используются триггеры.
2. Справочники s_m^i и dr_k связаны при помощи операторов проекции, переименования и выбора реляционной алгебры:

$s_m^i = \sigma_C(\pi_{A_1A_2...A_N}(\rho_{S(A_1A_2...A_L)}(dr_k)))$. Связь между таблицей cs_m^i, изоморфной $SFCS_m^i$, и справочником $cr_k \in CRCDB$ выглядит следующим образом: $cs_m^i = \sigma_C(\pi_{A_1A_2...A_N}(\rho_{S(A_1A_2...A_L)}(cr_k)))$. В результате, таблицы cs_m^i формируются как наборы данных в памяти на сервере репликации. Сервер репликации расположен на участке корпоративной автоматизированной информационной системы с центральной консолидированной базой данных нормативно-справочной информации.

3. Сформированные таблицы передаются клиенту репликации, установленному в территориально удалённом участке корпоративной автоматизированной информационной системы. В конкретной реализации системы репликации данных для этого могут использоваться DCOM, сокеты поверх TCP/IP или HTTP.
4. Каждая из функций $SFCS_m^i$ получается на основе изоморфизма с таблицей cs_m^i.
5. Справочники s_m^i изменяются при помощи функций $SFCS_m^i$.

Далее в главе 4 предлагается один из возможных методов решения проблем масштабируемости и готовности при построении системы репликации баз данных. После этого предлагается два подхода к обеспечению территориально удалённых участков корпоративной информационной системы совпадающим программным обеспечением: тиражирование требуемых для него справочников и использование Web-приложений.

В главе 5 проводится сравнительный анализ разработанных алгоритмов с точки зрения их производительности и создаваемой нагрузки на каналы связи на основе теоретических оценок. В первую очередь оценивается целесообразность хранения предыдущей версии справочников. Доказываются теоремы 5.1 и 5.2.

Теорема 5.1. Время работы и объём передаваемых к КБД НСИ данных алгоритма, использующего хранение предыдущей версии справочника, меньше времени работы и объёма передаваемых к КБД НСИ данных алгоритма, не использующего его.

Теорема 5.2. Объём передаваемых данных между КБД НСИ и территориально удалёнными участками корпоративной автоматизированной информационной системы меньше при использовании алгоритма с хранением предыдущей версии справочников.

Далее проводится сравнение импорта справочников в консолидированную базу данных нормативно-справочной информации с последующим их тиражированием и непосредственного применения автоматизированной системы импорта данных на каждом из участков информационной системы. Доказывается теорема 5.3.

Теорема 5.3. Объём передаваемых по сети данных при использовании КБД НСИ и системы репликации данных меньше, чем при использовании автоматизированной системы импорта данных из внешних источников на каждом из территориально удалённых участков корпоративной автоматизированной информационной системы, причём разность объёмов увеличивается с ростом их числа и размеров используемых справочников.

Далее проводится сравнение производительности авторского алгоритма $F_{GetChanges}$ на основе использования красно-чёрных деревьев с альтернативными подходами: использованием специализированных программных продуктов, теоретических алгоритмов, а также авторского алгоритма с использованием альтернативных структур данных.

В главе 6 рассматривается задача создания в корпоративной автоматизированной информационной системе справочников международного административно-территориального деления. На её основе проводится анализ разработанных алгоритмов по результатам компьютерного моделирования.

При создании справочников международного административно-территориального деления необходимо обеспечить поддержку исторической информации, поддержку иерархической информации и поддержку возможности репликации данных. Предлагается ко всем таблицам, работающим с историческими данными, добавить поле, содержащее время создания кортежа, и таблицу-дубликат, содержащую суррогатный первичный ключ, поля исходной таблицы, поле, содержащее время удаления кортежа и специфичные для конкретного случая дополнительные поля. Тогда отношение, соответствующее содержимому справочника на заданный момент времени, вычисляется по следующей формуле: $T = \pi_{A_1 A_2 \ldots A_N}(\sigma_{C_1}(R) \cup \sigma_{C_2}(H))$, где $C_1 = StartDate < CurrentDate$, $C_2 = IsMin \,\&\, (StartDate < CurrentDate) \,\&\, (DeleteDate > CurrentDate)$. R используется для обозначения исходной таблицы, H - для обозначения таблицы-дубликата. Предикат $IsMin$ проверяет, является ли запись самой ранней из удовлетворяющих условию $(DeleteDate > CurrentDate)$. Для поддержки работы с иерархическими данными предлагается модификация существующего подхода, основанного на добавлении родительского идентификатора $ParentID$ и дочернего идентификатора ID. В диссертационной работе рекомендуется использование LRO-репликации. При её использовании для поддержки репликации не требуется изменять структуру справочников.

Целесообразность хранения предыдущей версии справочников и целесообразность использования центральной консолидированной базы данных нормативно-справочной информации подтверждаются в диссертационной работе результатами вычислительных экспериментов.

Для выбора структуры данных, наиболее эффективной с точки зрения скорости работы использующего её авторского алгоритма вычисления $F_{GetChanges}$, было проведено статистическое сравнение линейного списка, бинарного дерева поиска, хэш-таблицы, AVL-дерева и красно-чёрного

дерева. Результаты эксперимента – это количественные данные, то есть замеры времени работы алгоритмов на конкретных данных международного административно-территориального деления. Предполагается разбиение этих данных на группы в зависимости от качественного параметра – типа используемого алгоритма. После прохождения логарифмического преобразования данные компьютерных экспериментов прошли тесты на гомогенность дисперсии Левена и на нормальность распределения Д'Агостино. В результате, к ним стало возможным применить дисперсионный анализ Фишера, который показал значимость различий между группами. Применение критериев Ньюмана-Кейлса и Тьюки позволило расположить алгоритмы по порядку скорости их работы: красно-чёрное дерево, AVL-дерево, хэш-таблица, бинарное дерево поиска, линейный список. Вычислительные эксперименты и их обработка методом наименьших квадратов позволили получить численные оценки сложности алгоритмов. Результаты компьютерного моделирования оказались согласованы с ожиданиями на основе теоретических оценок, полученными с использованием теории сложности.

В заключении приведены основные результаты работы.

Основные результаты работы

1. Предложены математические модели управления обработкой нормативно-справочной информации в корпоративных автоматизированных информационных системах на основе реляционной алгебры. Показано, что их использование при импорте и экспорте нормативно-справочной информации обеспечивает согласованность, актуальность и полноту данных в корпоративных автоматизированных информационных системах.
2. Разработан метод импорта данных в консолидированную базу данных нормативно-справочной информации, использующий двухступенчатый

механизм переноса данных, когда вначале выделяются требуемые данные справочников из внешних источников и они представляются в промежуточном формате, а затем осуществляется перенос данных в консолидированную базу данных нормативно-справочной информации, что позволяет упростить процесс согласования форматов данных. Предложены рекомендации по выбору справочников из внешних источников и интеграции справочников внутренних источников.

3. Разработан метод выделения изменений в версиях справочников на основе красно-чёрных деревьев. Показано, что при его использовании достигается более высокая скорость выделения изменений в справочниках, чем при использовании других известных методов.
4. На основе предложенных в диссертационной работе математических моделей и методов разработаны вычислительные алгоритмы и реализующий их комплекс программ.

Список публикаций по теме диссертации

1. *Бондаренко А.В., Гудков К.С.* Математическое моделирование миграции нормативно-справочной информации в корпоративных информационных системах // Моделирование авиационных систем: Сб. аннотаций докладов / НИИАС. – М., 2011. – С. 110-111.

2. *Бондаренко А.В., Гудков К.С.* Создание таблиц нормативно-справочной информации на основе разнородных внешних справочников // Модели и методы обработки информации: Сб.ст. / МФТИ. – М., 2009. – С. 148-152.

3. *Гудков К.С.* Выделение изменений в версиях открытых баз данных при построении автоматизированной системы импорта внешних справочников // Основни проблеми на съвременната наука - 2010. Том 22 Съвременни технологии на информации Математика Здание и архитектура. – София, 2010. – С. 16-19.

4. *Гудков К.С.* Консолидация нормативно-справочной информации в распределённых информационных системах // Современные проблемы фундаментальных и прикладных наук. Часть VII. Управление и прикладная математика: Труды 51-й научной конференции МФТИ. / МФТИ. – М., 2008. – С. 86-88.

5. *Гудков К.С.* Математическая модель управления нормативно-справочной информацией в распределённых информационных системах // Современные проблемы фундаментальных и прикладных наук. Часть VII. Управление и прикладная математика: Труды 52-й научной конференции МФТИ. / МФТИ. – М., 2009. – С. 123-125.

6. *Гудков К.С.* Математическая модель управления справочниками административно-территориального деления стран СНГ в корпоративных информационных системах // Прикладная информатика. – 2010. – № 5(29). – С. 117-124.

7. *Гудков К.С.* Механизмы интеграции внутрикорпоративных справочников // Прикладная информатика. – 2011. – № 6(36). – С. 14-22.

8. *Гудков К.С.* Моделирование импорта данных разнородных внешних справочников в консолидированную базу данных нормативно-справочной информации // Актуальные проблемы гуманитарных и естественных наук. – 2009. – № 9. – С. 11-14.

9. *Гудков К.С.* Оценка времени работы одного алгоритма, находящего разность в версиях открытых внешних справочников // Aktuální vymoženosty vědy - 2011. Díl 20. Technické vědy. Moderní informačni technologie. – Прага, 2011. – С. 59-62.

10. *Гудков К.С.* Решение проблемы готовности в рамках построения системы репликации баз данных // Современные проблемы фундаментальных и прикладных наук. Часть VII. Управление и прикладная математика: Труды 50-й научной конференции МФТИ. / МФТИ. – М., 2007. – С. 62-64.

11. *Гудков К.С.* Управление внешней нормативно-справочной информацией в распределённых информационных системах // Материалы XVI Международной конференции студентов, аспирантов и молодых учёных "Ломоносов-2009", секция "Вычислительная математика и кибернетика". / МГУ. – М., 2009. – С. 23.

В работах с соавторами [1, 2] лично соискателем выполнено следующее:

1. Предложены математические модели управления обработкой нормативно-справочной информации в корпоративных автоматизированных информационных системах на основе реляционной алгебры.
2. Разработан метод импорта данных в консолидированную базу данных нормативно-справочной информации.

Гудков Кирилл Сергеевич

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ УПРАВЛЕНИЯ ОБРАБОТКОЙ ИНФОРМАЦИИ В КОРПОРАТИВНЫХ АВТОМАТИЗИРОВАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Автореферат

Подписано в печать 12.03.2012.
Формат 60x84 1/16. Усл. печ. л. 1,0.
Тираж 80 экз. Заказ № 303.

ФГУП Государственный научно-исследовательский институт авиационных систем
125319, Москва, ул. Викторенко, 7