На правах рукописи

МИН ТХЕТ ТИН

МЕТОДИКА ФОРМИРОВАНИЯ РЕЛЯЦИОННЫХ ТАБЛИЦ НА ОСНОВЕ ИНФОРМАЦИИ ТАБЛИЧНОГО ВИДА

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2015

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования (ФГБОУ ВПО) «Московском государственном техническом университете им. Н.Э. Баумана» на кафедре «Компьютерные системы и сети» ИУ-6.

Научный руководитель:	Доктор технических наук, доцент, профессор кафедры «Компьютерные системы и сети», МГТУ имени Н.Э. Баумана Брешенков Александр Владимирович
Официальные оппоненты:	Доктор технических наук, профессор, профессор кафедры «Экономики городского хозяйства» ГАОУ ВПО Московский городской университет управления Правительства Москвы (МГУУ) Данчул Александр Николаевич. Кандидат технических наук, доцент, начальник отдела ЗАО «Всесоюзный институт волоконно-оптических систем связи и обработки информации» Самарев Роман Станиславович.
Ведущая организация:	Открытое акционерное общество «Государственный научно- исследовательский институт приборостроения» (ОАО «ГосНИИП»)

Защита диссертации состоится « 27 » апреля 2015 г. в 12.00 часов на заседании диссертационного совета Д212.125.01 при Московском авиационном институте (национальном исследовательском университете) – МАИ по адресу: 125993, г. Москва, А-80, ГСП-3, Волоколамское шоссе, д. 4.

С диссертацией можно ознакомиться в библиотеке Московского авиационного института (национального исследовательского университета) – МАИ

Отзывы, заверенные печатью, просьба высылать по адресу: 125993, г. Москва, А-80, ГСП-3, Волоколамское шоссе, д.4, МАИ, Ученый совет МАИ

Автореферат разослан « _____» _____2015 г.

Ученый секретарь
диссертационного совета Д212.125.01
кандидат технических наук, доцент А.В.Корнеенкова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. В настоящее время трудно переоценить значение компьютерных информационных систем. А коль скоро базы данных (БД) являются ядром информационных систем, в полной мере это относится и к БД. Это детально и убедительно доказывается в соответствующей научно-популярной и технической литературе. Более того, в паспорте специальности 05.13.11 (Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей) отмечается:

−необходимость разработки и исследования в области программных средств организации и управления обработкой данных и знаний;

−необходимость создания прикладного математического обеспечения, программных средств автоматизации разработки программ;

−актуальность разработки программных средств обработки данных и знаний в ВМ, ВК и КС;

−актуальность разработки методов проектирования систем управления базами данных (СУБД) и базами знаний (СУБЗ), в том числе распределенными СУБД и СУБЗ.

Собственно понятие информации глобальное и охватывает все сферы человеческой деятельности от вербального общения между людьми до работы в интернете. А данные − это информация, представленная в регламентированном виде. К сожалению, не всю информацию можно строго регламентировать и использовать в реляционных базах данных (РБД). Поэтому работы в этом направлении представляют интерес. В диссертации рассматривается информация табличного вида (ИТВ). В качестве примеров ИТВ можно назвать электронные таблицы, таблицы текстовых процессоров, HTML-таблицы и др. Практически на всех предприятиях накоплены значительные объемы ИТВ и эти предприятия заинтересованы в использовании преимуществ РБД. Представления такого рода информации близки к представлению данных в РБД, и поэтому в принципе процесс преобразования ИТВ в формат РБД можно формализовать и исключить возможные ошибки при проектировании РБД с нуля.

Достаточно большой объем работы в области проектирования РБД на основе использования ИТВ проделал Брешенков А.В. Однако, несмотря на глубокую теоретическую и практическую проработку проблемы, в его работах не рассматриваются важные задачи преобразования ИТВ в формат БД. В частности:

- рассмотрены не все возможные виды подзаголовков в ИТВ;
- не рассмотрены гибридные подзаголовки;
- в качестве атрибутов, которые входят в первичный ключ, анализировалось не более 2-х;
- связи между таблицами рассмотрены для ключевых полей, включающих только один атрибут;

- не проанализировано одно из требований минимальности первичного ключа − никакая часть первичного ключа не должна быть уникальной;
- не проведены детальные исследования по поводу выявления внешних ключей в ИТВ

В диссертации введено понятие ИТВР - расширенная информация табличного вида. При этом под расширением понимается то, что наряду с известными характеристиками ИТВ учитываются и их дополнительные характеристики, которые обусловили необходимость решения задач перечисленных выше.

Существует классическое определение реляционных таблиц (РТ). Но оно не удовлетворяет реальным требованиям к РТ. Поэтому введено понятие и соответствующая расширенная модель реляционных таблиц − РТР. Ее основное отличие от РТ в том, что она отражает свойства ИТВР которые не допустимы в РТ.

К настоящему времени выполнен значительный объем научных исследований, посвященных проектированию реляционных баз данных (РБД). Среди них можно назвать работы Е. Ф. Кодда, К. Дж. Дейта, Гэри Хансена, Джэймса Хансена, Ульмана Дж., Чена Р. Р., Райана Стивенса, Рональда Плю, Дэйва Энсора, Тихомирова Ю.В., Григорьева Ю.А., Баранчикова А.И. и других. Но в этих работах, как правило, методы проектирования РБД основываются на анализе предварительно разработанных схем отношений, когда данных, как таковых, еще нет.

Проблема заключается в отсутствии комплекса методов, алгоритмов, средств и методики, ориентированных на преобразование заполненных ИТВР в РТР.

Предметом исследования являются модели, методы и методика проектирования РБД на основе использования существующей, заполненной ИТВР, а также компоненты математического, лингвистического, информационного и программного обеспечений методики.

Цель и основные задачи исследования. Целью работы является разработка в рамках предложенной автором методики теоретических и практических основ формирования РТР на базе ИТВР, улучшение качественных и количественных характеристик существующих средств и алгоритмов решения задач формирования РБД на основе ИТВР для:

- автоматизированного преобразования заполненных ИТВР, соответствующих современным представлениям о информации табличного вида, в РТР, соответствующих современным представлениям о реляционных таблицах;
- автоматизированного формирования связей между преобразованными таблицами ИТВР;
- автоматизированного назначения первичных ключей в ИТВР;
- автоматизированного назначения внешних ключей в ИТВР;

Методы исследования. При расширении моделей ИТВ и РТ, а также при разработки методов и методики процесса преобразования ИТВР в РТР использована реляционная алгебра, исчисление предикатов, теория множеств, теория алгоритмов, аппарат сетей Петри.

Научную новизну работы определяет концепция и теоретические основы формирования РТР на базе ИТВР, которые воплощены в соответствующую методику проектирования РТ.

Научные результаты, выносимые на защиту:

1. Предложена расширенная модель информации табличного вида, которая отражает ранее не рассмотренные концептуальные особенности этих объектов.
2. Предложена расширенная модель реляционных таблиц, которая отражает ранее не рассмотренные концептуальные особенности этих объектов.
3. Разработан метод автоматизированного преобразования информации табличного вида в реляционные таблицы, который использует адекватные модели и обеспечивает исключение ручных способов и снижение трудоемкости и времени преобразования в десятки раз.
4. Разработан метод автоматизированного назначения ключевых полей в заполненных таблицах, который использует адекватные модели и обеспечивает исключение ручных способов и снижение трудоемкости и времени преобразования в десятки раз.
5. Разработана методика автоматизированного формирования реляционных таблиц на основе использования существующей информации табличного вида, в которой задействованы предложенные модели и методы, и которая сводит к минимуму дефекты преобразования, сокращает в десятки раз трудоемкость и время преобразования.

Достоверность научных положений, рекомендаций и выводов

Обоснованность научных положений, рекомендаций и выводов, изложенных в работе, определена корректным использованием современного математического аппарата. Достоверность положений и выводов диссертации подтверждена положительными результатами внедрения в учебный процесс МГТУ им. Н.Э. Баумана.

Практическая ценность и реализация результатов работы

Научные результаты, полученные в диссертации, доведены до практического использования. Методика, методы, а также программные средства могут быть использованы при решении задач проектирования РБД на основе использования ИТВР.

Содержание отдельных разделов и диссертации в целом было изложено и получило одобрение:

- на Российских НТК и семинарах (2011 - 2014 г.г.);

- на заседании кафедры “Компьютерные системы и сети” МГТУ им. Н.Э. Баумана.

Совокупность научных положений, идей и практических результатов исследований составляет оригинальное направление в области проектирования реляционных баз данных.

По результатам выполненных исследований опубликовано 11 научных работ.

Диссертационная работа состоит из введения, четырех глав и заключения, опубликованных на 161 страницах машинописного текста, содержит 48 рисунков, 20 таблиц, список литературы из 101 наименований и 2-х приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении показана актуальность решаемой проблемы, сформулированы цель и задачи исследования, приведено краткое описание содержания глав диссертации.

В первой главе «Исследование задач построения методики формирования реляционных таблиц на базе заполненных нереляционных таблиц» выполнен аналитический обзор традиционного подхода формирования реляционных таблиц, сформулированы его достоинства и недостатки. Расширено существующее понятие информации табличного вида. Выполнена постановка задачи разработки методики автоматизированного преобразования ИТВР в РТР. Определен состав алгоритмов и средств, разрабатываемых в рамках методики проектирования РТР на основе существующей информации табличного вида.

Проектирование РБД в соответствии с традиционной методологией, включает в себя 4 этапа: формулировка и анализ требований, инфологическое проектирование, даталогическое проектирование, физическое проектирование.

По определению реляционная модель данных (РМД) некоторой предметной области представляет набор отношений, изменяющихся во времени.

Основным понятием РМД является отношение, которое представляет собой подмножество декартового произведения доменов.

$R \subseteq D = D_1 \times D_2 \times \ldots \times D_k$.

Рассмотрены понятия ключевых полей и обеспечения целостности данных. В работах, посвященных теории проектирования РБД, дается определение первичных ключей, обосновывается их необходимость, формулируются требования к ним и определяются свойства внешних ключей. Обосновано утверждение о том, что при наличии только схемы данных непросто безошибочно выбрать атрибуты, удовлетворяющие этим требованиям.

Ряд специалистов в области реляционных баз данных не включают в качестве требований к РТ наличие первичных ключей. Это не оправданно.

Целостность данных в основном рассматривается в двух аспектах – целостность сущностей и целостность согласования. Целостность сущностей − сущности реального мира должны быть различимы. Целостность согласования − ссылаться можно только на те данные, которые существуют. Обеспечение

целостности данных связано с правильным назначением и использованием первичных и внешних ключей.

Рассмотрены понятия нормализация и семантическое моделирование. Нормализация отношений – аппарат ограничений на формирование отношений, который позволяет устранить избыточность БД, обеспечивает непротиворечивость хранимых данных в РБД.

Семантическое моделирование. Разработчики БД обычно обладают весьма ограниченными сведениями о смысле хранящихся в них данных. Поэтому широкое распространение получил метод семантического моделирования “сущность-связь” или ER-модель. Связи в модели “сущность-связь” могут иметь тип “один к одному”, “один ко многим”, “многие к одному” и “многие ко многим”.

Определено понятие ИТВР. Представление информации в табличном виде настолько удобно, что это определило появление целого класса систем, ориентированных на работу с табличной информацией − электронных таблиц. Форма представления ИТВР может быть самой различной: на бумаге, в виде файла текстового редактора, в виде электронных таблиц и др.

Всесторонний анализ ИТВР показал, что в общем случае она обладает следующим свойствами.

1. ИТВР − это информация, которая воспринимается ее потребителями, как таблицы.
2. В ИТВ могут отсутствовать разделители строк и разделители столбцов.
3. Элементы данных таблицы могут размещаться в нескольких строках.
4. Типы элементов данных одного столбца могут не совпадать.
5. Заголовки ИТВ могут включать в себя подзаголовки нескольких уровней.
6. В ИТВР могут быть задействованы внутренние подзаголовки.
7. В ИТВР имена заголовков могут совпадать.
8. Любой столбец ИТВР может быть задействован как многоуровневый заголовок.
9. В ИТВР возможны пустые строки.
10. В ИТВР могут отсутствовать заголовки столбцов.

В качестве примера на рис.1 приведен фрагмент реальной ИТВР.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		Продажа модернизации - месяц						Продажа модернизации - YTD						Коммерческие предложения - месяц						
2		Объем		Единицы		Цели		Объем		Единицы		Цели		План		Факт		Цели		
3	Region/branch					Объем	ед.					Объем	ед.	ед.	Объем	ед.	Объем	ед.	Объем	ед.
4		План	Факт	План	Факт	000 руб.		План	Факт	План	Факт	000 руб.			000 руб.		000 руб.		000 руб.	
5	MSR	12500,0	52,6	66,7	1,0	14750,0	77,2	1375,0	52,6	6,0	1,0	3150,0	15,0	75,8	16666,7	35,0	8379,0	178,0	39166,7	62,5
6	РУ-1	250,0	0,0	1,0	0,0	600,0	2,8	750,0	0,0	3,0	0,0	1800,0	8,5	11,4	2500,0	26,0	7309,0	27,3	6000,0	34,1
7	*Б.Кочуров*	0,0		0,0		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0			0,0	0,0	0,0
8	Д.Сапожников	83,3		0,3		158,3	0,8	250,0	0,0	1,0	0,0	475,0	2,3	3,8	833,3			7,2	1583,3	11,4
9	А.Трофимов	83,3		0,3		150,0	0,8	250,0	0,0	1,0	0,0	450,0	2,3	3,8	833,3	7,0	1973,0	6,8	1500,0	11,4
10	О.Базыкин (new)	41,7		0,2		145,8	0,7	125,0	0,0	0,5	0,0	437,5	2,0	1,9	416,7	11,0	3554,0	6,6	1458,3	5,7
11	А.Москвитин (new)	41,7		0,2		145,8	0,7	125,0	0,0	0,5	0,0	437,5	2,0	1,9	416,7	8,0	1782,0	6,6	1458,3	5,7
12	РУ-2	208,3	52,6	1,0	1,0	450,0	2,2	625,0	52,6	3,0	1,0	1350,0	6,5	9,5	2083,3	9,0	1070,0	20,5	4500,0	28,4
13	*А.Моисеев*	0,0		0,0		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0			0,0	0,0	0,0
14	С.Санникова	104,2		0,5		154,2	0,8	312,5	0,0	1,5	0,0	462,5	2,3	4,7	1041,7			7,0	1541,7	14,2
15	В.Левин	104,2	52,6	0,5	1,0	154,2	0,8	312,5	52,6	1,5	1,0	462,5	2,3	4,7	1041,7	9,0	1070,0	7,0	1541,7	14,2
16	Вакансия-1	0,0		0,0		141,7	0,7	0,0	0,0	0,0	0,0	425,0	2,0	0,0	0,0			6,4	1416,7	0,0
17	РУ-3	250,0	0,0	1,0	0,0	600,0	2,8	0,0	0,0	0,0	0,0	0,0	0,0	11,4	2500,0	0,0	0,0	27,3	6000,0	0,0
18	*А.Ямашкин*	0,0		0,0		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0			0,0	0,0	0,0

Рис.1. Фрагмент реальной ИТВР.

Анализ проблемы автоматизированного преобразования ИТВР в РТ показал, что для ее решения необходимы следующие взаимосвязанные алгоритмы (рис. 2):

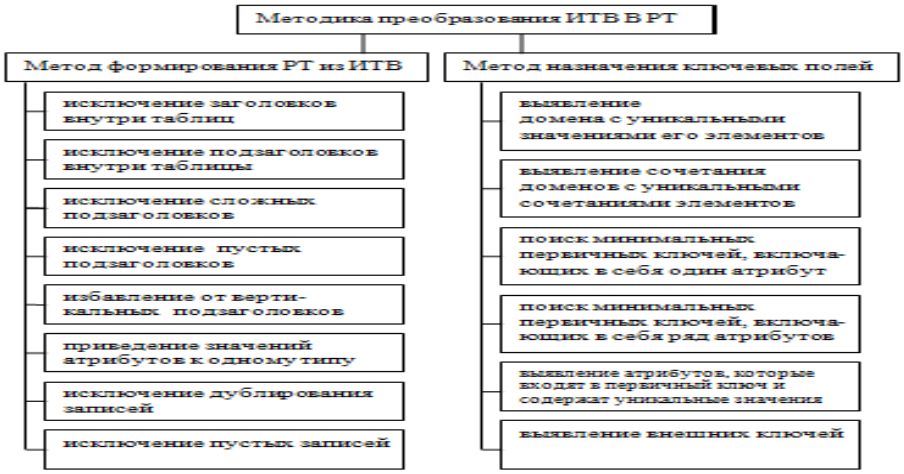

Рис. 2. Схема иерархии разрабатываемых алгоритмов

Во второй главе «Метод преобразования заполненных ИТВ в РТ» разработаны модели объектов исследования (РТР и ИТВР), метод преобразования таблиц ИТВР к реляционному виду. В рамках метода предложены алгоритмы приведения атрибутов ИТВР к единому типу, исключения дублирования записей в ИТВР, избавления от сложных атрибутов и исключения простых и сложных подзаголовков.

Модель РТР в известной литературе в полном объеме не рассматривается.

$RTP=\{Z, D\}$, где Z − множество заголовков, D − множество данных.

$z = \{z_1, \ldots, z_i, \ldots, z_n\}$, $i = 1, n$; $n >= 1$,где n − степень множества заголовков.

Должно быть обеспечено условие $z_i \neq z_m$, $i = 1, n$; $m = 1, n$; $i \neq m$,

где n − степень множества заголовков, т.е. недопустимо совпадение заголовков.

$D = \{SD\}$, где SD − множество строк данных.

$SD = \{SD_1, \ldots, SD_i, \ldots, SD_n\}$, $i = 1, n$; $n >> 1$,

где n − мощность множества строк данных.

$SD_i = \{ED_{i1}, \ldots, ED_{ij}, \ldots, ED_{ik}\}$, $j = 1, k$; $k >= 1$,

где k − степень множества i-ой строки данных, ED_{ij} − элемент данных.

В РТР не должно быть пустых заголовков: $Zi \neq Ø$.

В РТР не должно быть пустых строк: $SDi \neq Ø$.

В РТР содержимое домена не может быть подзаголовком: $SDi \neq zn$.

В РТР не должно быть внутренних подзаголовков: $SDi \neq z$.

В РТР не должно быть сочетаний различного типа подзаголовков: $\neg((Sdi = zn) \& (SDi = z))$.

Модель ИТВР в известной литературе в полном объеме не рассматривается. Рассмотрим эту модель. ИТВР представляются множеством $DT = \{Z, D\}$, где Z − множество заголовков ИТВР, D − множество данных, соответствующих заголовкам.

$Z = \{Z_1, ..., Z_i, ..., Z_n\}$, $i = 1, n$; $n >= 1$, где n − где n число заголовков.

Допустима ситуация, когда $Z_i = Z_m$, $i = 1, n$; $m = 1, n$; $i \neq m$, где n − степень множества заголовков, т.е. возможно полное совпадение заголовков.

$D = \{SD, Z\}$, где SD − множество строк данных.

Такого рода представление D допускает наличие нескольких заголовков и подзаголовков 1-го и 2-го уровней, расположенных в области данных.

$SD = \{SD_1, ..., SD_i, ..., SD_n\}$, $i = 1, n$; $n >> 1$, *где* n − мощность множества строк данных.

$SD_i = \{ED_{i1}, ...,ED_{ij}, ..., ED_{ik}\}$, $j = 1, k$; $k >= 1$, *где* k − степень множества элементов данных i-ой строки данных; ED_{ij} − элемент данных. В ИТВ могут быть пустые заголовки: $Z_i = \varnothing$. В ИТВ могут быть пустые строки: $SD_i = \varnothing$.

В РТ основные типы полей: числовой, текстовый, дата-время. В нереляционных таблицах в одном и том же столбце могут храниться данные различных типов. Это недопустимо в РТ. При преобразовании ИТВР в РТР значения атрибутов заполненных таблиц ИТВР в каждом столбце необходимо привести к одному типу.

В РТР недопустимо дублирование записей, а в ИТВР это возможно. В связи с этим необходимы средства исключения дублирования записей в ИТВР.

Ниже приведены результаты сравнительного анализа моделей ИТВР и РТР.

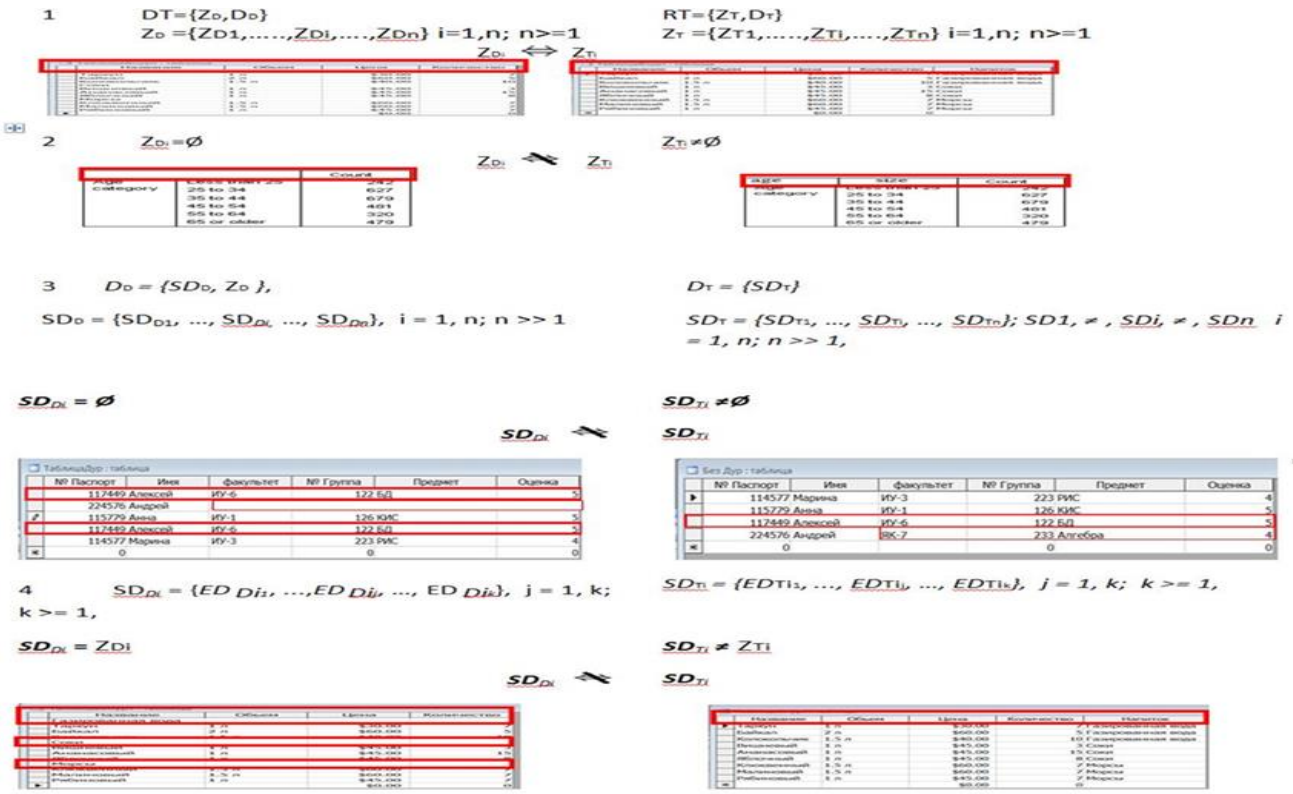

Рис.3. Результаты сравнительного анализа модели ИТВ и модели РТ.

Здесь: Z_D − множество заголовков ИТВ, Z_T − множество заголовков РТ, SD_D − множество строк данных ИТВ, SD_T − множество строк данных ИТВ.

Так как основной задачей работы является разработка алгоритмов, методов, методики преобразования ИТВР в РТР, то каждая позиция сравнительного анализа позволяет определить какие задачи необходимо решать в ходе разработки методики преобразования ИТВР в РТР. Упрощенно разработанная методика представляет собой процесс преобразования объектов, соответствующих модели ИТВР к объекту, соответствующему модели РТР. А совокупность связанных между собой РТР и является РБД. Следует обратить внимание на то, что методика проектирования систем управления базами данных в полном объеме в работе не рассматривается, а лишь представлена основными характерными запросами, которые формируются в ходе преобразования.

Исключение сложных атрибутов и подзаголовков. В ИТВР могут встречаться подзаголовки трех типов: внешние подзаголовки, внутренние подзаголовки и подзаголовки-столбцы, а также могут быть различные сочетания подзаголовков. В РТР подзаголовки недопустимы. На основе сравнительного анализа предложенных моделей РТР и ИТВР в диссертации разработаны алгоритмы исключения подзаголовков всех типов. Из соображений ограниченного объема автореферата приводится самый простой алгоритм − алгоритм исключения внутренних подзаголовков.

```
CNT = 0
DO s = 1 то m
   CNT1 = 0
    DO  f =1 то к
        IF a_sk= NULL THEN CNT1 = CNT1 + 1
    EBD  f
    IF CNT1 = k-1 THEN CNT= CNT+1
END s
IF CNT< 2 THEN EXIT
Формирование двух отношений
R0 = R(A_1,....,A_i,.....,A_k)+R(KR)
CNT = 0
DO s = 1 то m
   CNT1 = 0
```

```
DO s = 1 то m
   CNT1 = 0
    DO f =1 то к
        IF a_sk= NULL THEN CNT1 = CNT1 + 1
    END  f
    IF CNT1 = k-1 THEN
       CNT = CNT + 1
       C(R2_CNT,1)=CNT
       C(R2_CNT,2)= a_sk
       DELETE * FROM R0 WHERE (A_1= a_sk)
    ELSE
        C(R0_s,1)=CNT
    END IF
END s
```

Здесь m – мощность R; k – степень R;
$R0 = R(A_{1,\ldots,}A_{i,\ldots,}A_k)+R(KR)$ означает добавление к R атрибута с именем KR;
$C(R2_{CNT,1})$ означает значение элемента R2 в строке CNT и 1-ом столбце;
$C(R0_{s,1})$ означает значение элемента R0 в строке s и 1-ом столбце.

По сути алгоритм организует поиск записей, в которых все значения кроме 1-го − пустые (Null). Найденные подзаголовки удаляются, но предварительно преобразуются в новый атрибут таблицы.

Проблема приведения значений атрибутов заполненных таблиц к одному типу решена двумя способами − с использованием существующих инструментальных средств, в частности Access, для таблиц малой размерности (менее 200 записей) и с использованиеем разработанных алгоритмов и соответствующих процедур для таблиц большой размерности.

Исключение повторяющихся строк в ИТВ в основном, как показано в работе, хорошо обеспечивается существующими инструментальными средствами, которые органично интегрируются в разработанное приложение.

Суть метода преобразования заполненных нереляционных таблиц в реляционные таблицы состоит в преобразовании объектов соответствующих ИТВР в объекты соответствующие РТР на основе использования предложенного комплекта алгоритмов и процедур, методики их использования.

В третьей главе «Метод назначения ключей в ИТВ» разработан метод назначения ключевых полей в ИТВ, который по сравнению с известным методом позволяет учитывать требования к введенным моделям ИТВР м РТР . В частности, в отличии от известных работ, в модели РТР предусмотрено наличие первичных и внешних ключей, полностью учтено требование к минимальности ключей.

В работах Брешенкова А.В. предложен метод назначения ключевых полей в ИТВ, который вполне приемлем для использования. Однако в нем учтены не все особенности ключевых полей. В частности:

− рассматривается возможность включения в первичный ключ только 2-х атрибутов;

− не полностью учитывается требование минимальности первичного ключа;

− не до конца прояснены вопросы формирования первичных ключей из нескольких атрибутов;

− мало освещены вопросы назначения внешних ключей;

− назначение первичных ключей не рассматривается как неотъемлемая задача преобразования ИТВ в РТ.

Ниже сформулированы требования к ключевым полям.

Уникальность. Пусть имеется отношение R:

$R=(A_1, \ldots, A_i, \ldots, A_m, \ldots, A_k)$, $i=\overline{1,k}$, где κ − степень отношения; A_i − атрибут отношения.

$A_i=\{e_{i_1},...,e_{i_j},...,e_{i_n}\}$ $j=\overline{1,n}$, где n − мощность отношения, e_{i_j} − j - й элемент атрибута A_i;

$\mathrm{A_m}=\{e_{m_1},...,e_{m_j},...,e_{mn}\}$ $j=\overline{1,n}$, где n − мощность отношения, e_{m_j} − j - й элемент атрибута A_m.

Необходимо найти такие атрибуты $A_i,\ldots,A_m$ чтобы обеспечилась истинность выражения: $\mathrm{concat}(e_{i_1},\ldots,e_{m_1})\neq,...,\neq\mathrm{concat}(e_{i_j},\ldots,e_{m_j})\neq,...,\neq\mathrm{concat}(e_{i_n},\ldots,e_{mn})$

Необходимо найти такое сочетание атрибутов, чтобы конкатенация их значений была уникальна. При этом:

− проверяемый кортеж атрибутов может включать несколько атрибутов;

− число возможных сочетаний атрибутов может быть очень большим − это зависит от степени отношения (общего числа атрибутов в отношении);

− ключевой атрибут может быть только один;

− может не найтись таких атрибутов, в этом случае назначают суррогатный ключ.

Минимальность. Минимальность ключевого поля рассматривается в двух аспектах.

В первой части требования во главу угла ставится объем памяти, который необходим для хранения значений атрибутов, входящих в первичный ключ. Поэтому самая очевидная целевая функция – минимальное число атрибутов, входящих в первичный ключ: $\min|\mathrm{A1},\ldots,\mathrm{Ai},\ldots,\mathrm{Ak}|$, $i=\overline{1,k}$, где κ − число атрибутов, входящих в ключ; $\mathrm{A_i}$ − атрибут отношения, входящий в ключ.

Строго говоря, целевая функция следующая: $\min(\mathrm{Length}(\mathrm{A_i})+\ldots+\mathrm{Length}(\mathrm{A_к}))$.

Во второй части требования под минимальностью первичного ключа подразумевается отсутствие в составе ключа атрибута, значения которого уникальны. Пусть первичный ключ К представлен множеством атрибутов:

$K = (A_1,..., A_i,..., A_j,..., A_k)$, $i = \overline{1,k}$, где k – число атрибутов, входящих в первичный ключ; $A_i - i$ - й атрибут отношения, входящий в первичный ключ. Тогда должно быть истинно выражение $\neg (e_{i_1}, \ldots, \neq, ..., e_{i_j}, \ldots, \neq, ..., e_{i_n})$.

Алгоритмы назначения первичных ключей в ИТВ. Предложены неформальные (описательные) и формальные алгоритмы назначения первичных ключей. Из соображений объема приводится только один неформальный алгоритм – алгоритм поиска ключа, включающего пару атрибутов:

$MKA^2 = \varnothing$, где MKA^2 – множество пар атрибутов, которые претендуют на роль первичного ключа.

Пусть имеется отношение R: R=(A1, …, Ai, …, Ak), $i = \overline{1,k}$, где к – степень отношения; Ai –атрибут отношения. $MPA = \varnothing$.

Ищутся все возможные сочетания пар атрибутов и запоминаются в MPA:

```
Cnt=0                                        S = Concat (Ai , Aj)
DO  i = 1 to k-1                             MPA (Cnt) = MPA + S
   DO  j = i+1 to k                        END  j
      Cnt = Cnt + 1                     END  i
```

Таким образом, в массиве MPA сформируются все возможные пары атрибутов, а в счетчике Cnt хранится их количество.

Проверяются все пары на уникальность.

```
MUP = ∅ /* Массив пар атрибутов, представляющих собой атрибуты, все соответствующие пары значений которых уникальны */
Cnt1 = 0
DO  i = 1 to Cnt
   S = MPA(Cnt)
```

/ По сути S представляет собой пару атрибутов (Ai, Aj) $A_i = (e_{i_1} .... e_{i_m})$, где e_{i_1} – 1-й элемент домена с атрибутом A_i, e_{i_m} – m-й элемент домена с атрибутом A_i.*

*$A_j = (e_{j_1} .... e_{j_m})$, где e_{j_1} – 1-й элемент домена с атрибутом A_j, e_{j_m} – m-й элемент домена с атрибутом A_j, m – мощность отношения. */*

```
   DO  n = 1 to m
```

/ Для каждой пары атрибутов (Ai, Aj) выполняется проверка условия Concat(e_{i_1}, e_{j_1}) $\neq .... \neq$ Concat((e_{j_1}, e_{j_m}) */*

```
   END  n
```

/ Если текущая пара атрибутов имеет все соответствующие пары значений уникальными, то эта пара добавляется к массиву пар с уникальными значениями: */*

```
Cnt1 = Cnt1 + 1
MUP( Cnt1) = S
```

END i

Если претенденты на ключевой атрибут найдены, т.е. MUP $\neq \varnothing$*, то для проверки второго требования минимальности выполняется переход к алгоритму поиска первичного ключа на основе 3-х атрибутов и далее по аналогии. Разработчик на любом шаге поиска может принять решение о назначении суррогатного ключа.*

Алгоритмы назначения внешних ключей в ИТВ. В диссертации предложены неформальные и формальные алгоритмы назначения внешних ключей в заполненных таблицах. Ниже приведен формальный алгоритм.

П1. Осуществляется поиск всех возможных сочетаний пар таблиц анализируемой БД.

Пусть имеется табличное пространство T:

```
/* T=(T1, ..., Ti, ..., Tm, ..., Tk), i = 1,k , где к – число прикладных таблиц БД */
MPT = ∅ _ _ Массив всех пар таблиц БД
/* Ищутся все возможные сочетания пар таблиц */
Cnt=0
DO i = 1 to k-1 _ _ к – число таблиц БД
   DO j = i+1 to k
      Cnt = Cnt + 1_ _ счетчик числа пар таблиц
      S = Concat (Ti , TJ)
      MPT (Cnt) = S
   END j
END i
```

П2. Для каждой пары таблиц из массива MPT выполняется поиск внешнего ключа

```
F = 0 _ _ флажок наличия связанных ключей
Cnt1 = 0 _ _ счетчик совпадений
DO i = 1 to Cnt _ _ – число пар таблиц БД
   DO l = 1 to C1_ _ – степень 1-й таблицы из i-й пары
      DO m= 1 to C2 _ _ – степень 2-й таблицы из i-й пары
         DO j= 1 to M1_ _ –мощность 1-й таблицы из i-й пары
            DO k= 1 to M2_ _ –мощность 2-й таблицы из i-й пары
               If e_lj = e_mk Then Coun1 = Coun1 + 1
            END k
         END j
         If Coun1 >= Max(M1, M2) Then
         Begin
            Print (“Для пары таблиц ”, i, “столбцы”, l, “ и ”, m)
            Print (“ претендуют на первичный и внешние ключи”)
            Cnt1 = 0
            F = 1
         End
      END m
   END l
```

END i

If F = 0 Then Print (“ внешних ключей не обнаружено не для одной из таблиц”)

Следует обратить внимание на то, что атрибуты, предложенные алгоритмом в качестве связанных атрибутов, могут не удовлетворять разработчика, поэтому необходимо предоставить ему возможность окончательного решения.

Суть метода назначения ключевых полей в ИТВ состоит в преобразовании объектов соответствующих ИТВР в объекты соответствующие РТР на основе использования требований к ключевым полям, предложенного комплекта алгоритмов и процедур, методики взаимодействия разработчика и подсистемы.

В четвертой главе «Методика формирования РТ на основе использования заполненных ИТВ» решена задача формализации методики формирования РТР на основе использования заполненных ИТВР. Предложена модель методики в операторной форме. Предложена модель методики с использованием аппарата сетей Петри. Проведено исследование разработанной методики с использованием аппарата сетей Петри. В результате исследований и модификаций модели исключены концептуальные ошибки в ее описании и функционировании.

Формулировка проблемы формализации методики. Целью работы является разработка методики формирования реляционных таблиц на основе существующей информации табличного вида.

Модель методики в операторной форме. На начальном уровне абстрагирования от большинства компонент человеко-машинной системы схема процесса преобразования ИТВР в РТР может быть проиллюстрирована рис.4.

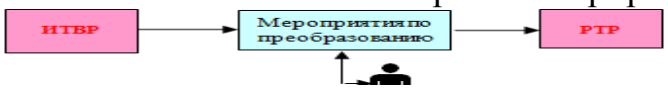

Рис. 4. Укрупненная модель процесса формирования РТР на основе использования заполненных ИТВР

Введем оператор преобразования ИТВР в РТР. Результатом выполнения этого оператора является объект, соответствующий предложенной ранее модели РТР. Операндами оператора *ОПП,* как видно из рисунка, являются модель ИТВР и модель РТР. Укрупненная операторная модель формирования РТР на основе использования заполненных ИТВР приведена на рисунке 5.

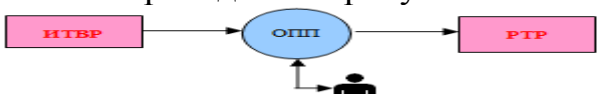

Рис. 5. Укрупненная операторная модель формирования РТР на основе использования заполненных ИТВР

Оператор операторной модели (элипс) соответствует одной или более человеко-машинной процедуре. Состояние операторной модели (прямоугольник) соответствует начальному, промежуточному и окончательному состоянию ИТВР.

В диссертации выполнено 11 итераций построения операторной модели методики преобразования ИТВР в РТР. Модель анализировались и уточнялась. В результате получена результирующая операторная модель процесса преобразования ИТВР в РТР, представленная на рис. 6.

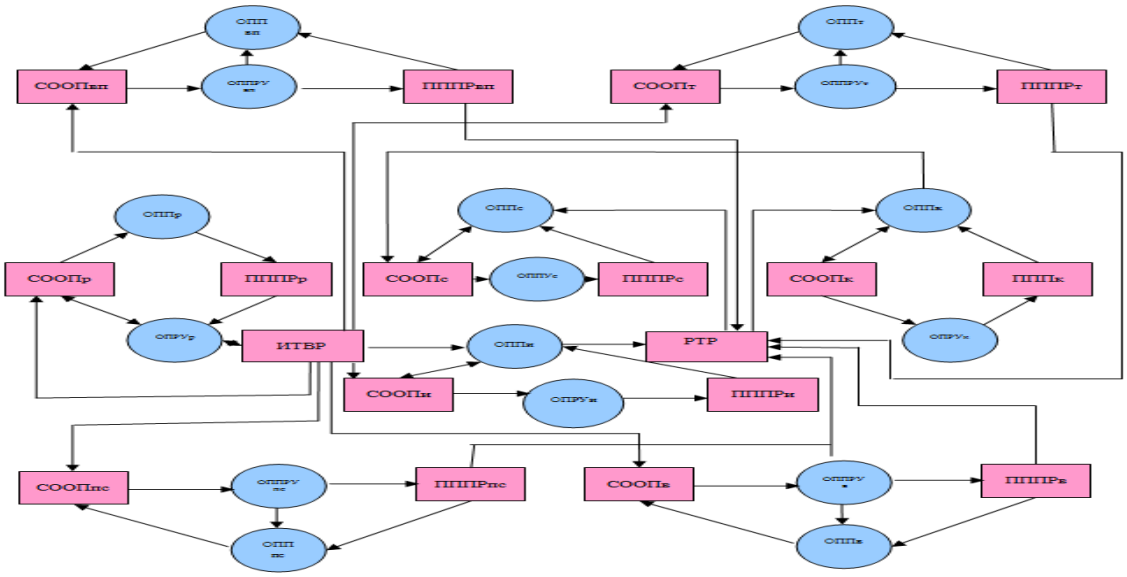

Рис. 6. Операторная модель процесса преобразования ИТВР в РТР

Здесь: ОППс − оператор формирования связей между таблицами ИТВ; СООПс − состояние модели, соответствующее автоматической генерации системы оценок результатов импорта; ОПРУс − оператор принятия решений относительно управляющих воздействий после анализа системы оценок результатов формирования связей; ППППРс − состояние модели, соответствующее формированию системы оценок результатов формирования связей разработчиком РТ.

В соответствии с моделью после каждого выполнения оператора с помощью автоматизированных средств выполняется проверка промежуточного состояния таблицы ИТВР на соответствие модели РТР. Успешным завершением работы в рамках разработанной модели методики считается полное соответствие таблицы, описываемой моделью ИТВР, таблице, описываемой моделью РТР.

Мотивация перехода от операторной модели к модели, построенной на основе аппарата сетей Петри. Операторная модель, построенная на основе анализа отличий моделей ИТВР и РТР, позволила выделить основные компоненты человеко-машинной подсистемы преобразования, определить порядок их использования, сформировать связи между ними. Во избежание ошибок на ранних этапах разработки методики и подсистемы необходимо исследовать следующие ее характеристики:

- обеспечение баланса информационных потоков (устойчивость);
- достижимость всех состояний (живость);

- отсутствие тупиковых ситуаций при работе подсистемы (живость);
- отсутствие ситуаций, когда подсистема приходит в нестационарное состояние и число информационных потоков превышает критическую отметку (безопасность);

Операторная модель не позволяет выполнить эти исследования.

Предложенное операторное представление методики наиболее близко к системному уровню, а на данном уровне обычно применяют модели систем массового обслуживания или сети Петри. В связи с этим для выявления принципиальных ошибок в методике и ее исследования выбран аппарат сетей Петри. После проведения исследований уточнена операторная модель.

Формирование сетевой модели. Сеть Петри состоит из трех элементов: множество мест *S*, множество переходов *T* и отношение инцидентности *F*.

Между набором $N = (S, T, F)$ сети Петри и набором $M = (C, O, L)$ операторной модели установлено соответствие таким образом, чтобы между элементами наборов обеспечивалось взаимно однозначное соответствие: $C \leftrightarrow S$; $O \leftrightarrow T$; $L \leftrightarrow F$. Как следствие этого соответствия: $|C| = |S|$; $|O| = |T|$; $|L| = |F|$.

На рис. 7 приведена начальная модель процесса автоматизированного преобразования ИТВР в РТР в виде сети Петри.

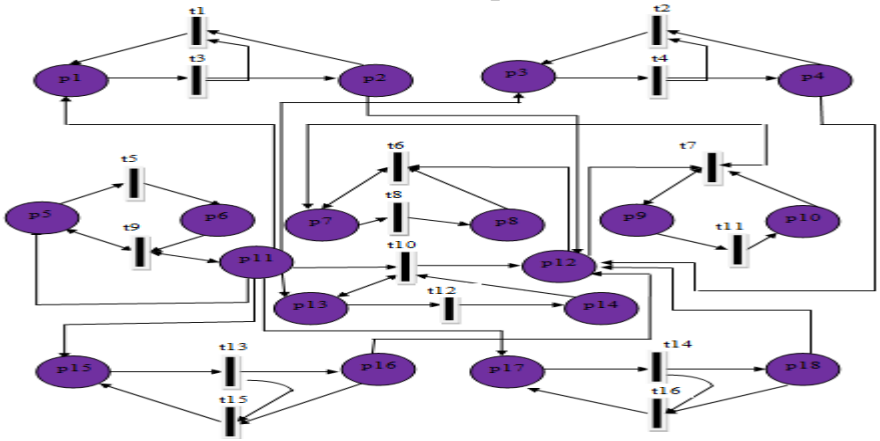

Рис. 7. Начальная сетевая модель процесса автоматизированного преобразования ИТВ в РТ в виде сети Петри

Модели ИТВ, модели РТ, СО, СП поставлены в соответствие положениям сети $\{P\}$. Операторам ОП, ОПР поставлены в соответствие переходы сети $\{t\}$.

Исследование устойчивости сети. Сеть Петри является устойчивой, если она имеет потоковое назначение $\varphi i > 0$ для каждого $ti \in T$.

Для каждого перехода сети рис. 7 назначен поток φi. Для каждого положения Pi запишем уравнения потоков, которые не должны противоречить друг другу, если данная сеть устойчива. В таблице 1 представлена часть уравнений.

Таблица 1

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	t16	
p1	+φ1		-φ3														=0
p2	-φ1		+φ3														=0
p3		+φ2		-φ4													=0
p4		-φ2		+φ4													=0
p5					-φ5				-φ9 +φ9								=0
p6					+φ5			-φ9									=0
p7						+φ6 -φ6	+φ7	-φ8									=0
p8						-φ6		+φ8									=0
p9							+φ7 -φ7				-φ11						=0
p10							-φ7				+ φ11						=0
p11									+φ9 -φ9	-φ10							=0
p12						-φ6	-φ7			+ φ10							=0
p13										+φ10 -φ10 -φ12							=0
p14										-φ10		+ φ12					=0
p15													-φ13		+φ15		=0
p16													+ φ13		-φ15		=0
p17														-φ14		+φ16	=0
p18														+φ14		-φ16	=0

После минимизации уравнений оказалось, что часть потоков имеют нулевые значения, а это противоречит требованию устойчивости сети. Поэтому для сетевой модели выполнен ряд итераций преобразования сети и соответствующей ей операторной модели − введены дополнительные состояния и переходы, отражающие реальное положение дел. Достигнута ситуация, когда в уравнениях потоков нет противоречий. При этом модифицировалась и операторная модель.

Анализ методики для обнаружения дефектов ее функционирования. Динамику функционирования системы можно моделировать перемещением маркеров в сети в соответствии с правилами перехода: M'(P)=M(P)–P(ti)+H(ti), где M(P) = (M(P1), M(P2), …, M(PN)) - разметка сети. На рис. 8 приведен пример перемещения маркеров. Маркеры символизируют состояние системы.

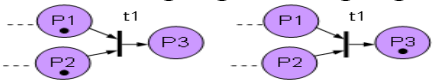

Рис. 8. Пример перемещения маркеров в сети.

Выполнено перемещение маркеров в сети, при этом проанализировано более 50-и состояний. Это позволило сделать заключение о том, что:

- является ли сеть живой, т.е. есть ли потенциальная возможность срабатывания всех переходов;
- является ли сеть достижимой, т.е. есть ли потенциальная возможность достигнуть всех положения сети.

В работе выполнена проверка живости и достижимости сети, выявлены ошибки, исправлена сеть Петри и соответствующая операторная модель.

В результате всех преобразований получена сеть Петри свободная от принципиальных ошибок, что исключает принципиальные ошибки в описываемом ею процессе.

После выполненных преобразований сетевой модели и исключения принципиальных ошибок в методике выполнен обратный переход к операторной модели (рис. 9).

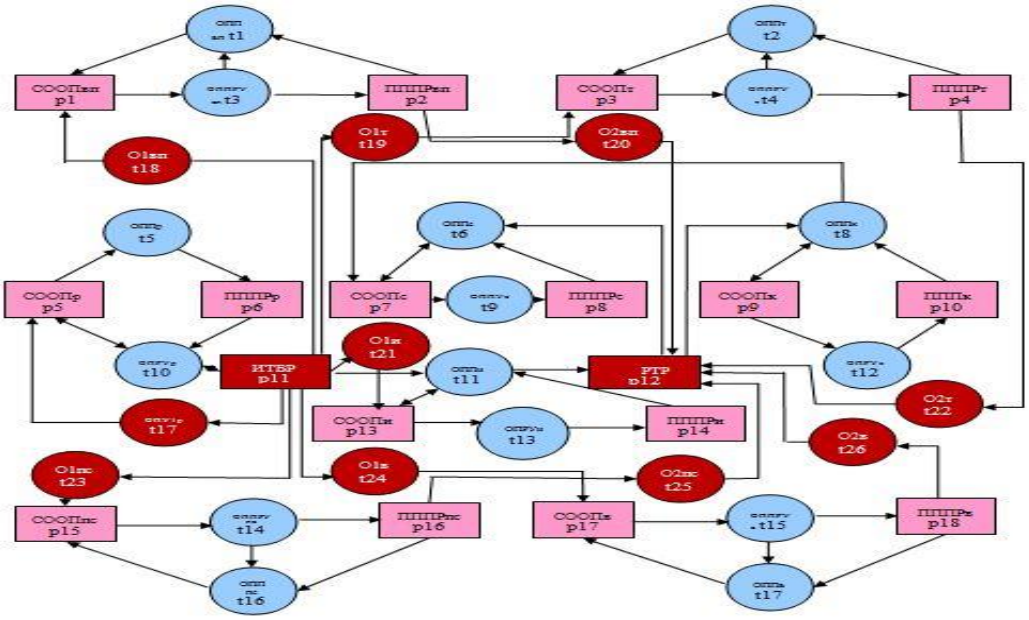

Рис. 9. Окончательная операторная модель процесса преобразования ИТВ в РТ

Разработанная операторная модель методики служит в качестве исходной формализации для разработки процедур подсистемы преобразования ИТВ в РТ , а также как руководство к действиям пользователя подсистемы.

Суть предложенной методики состоит в последовательном использовании разработанных методов, алгоритмов и процедур, в которых задействованы предложенные модели РТР и ИТВР, в соответствии с формализованной моделью процесса преобразования.

В приложении 1 «Программная реализация методики формирования РТ на основе ИТВ» выполнена программная реализация методики формирования реляционных таблиц на основе информации табличного вида. Выполнены экспериментальные исследования временных характеристик процедур подсистемы. На рис. 10. приведено схематичное изображение методики преобразования ИТВ в РТ.

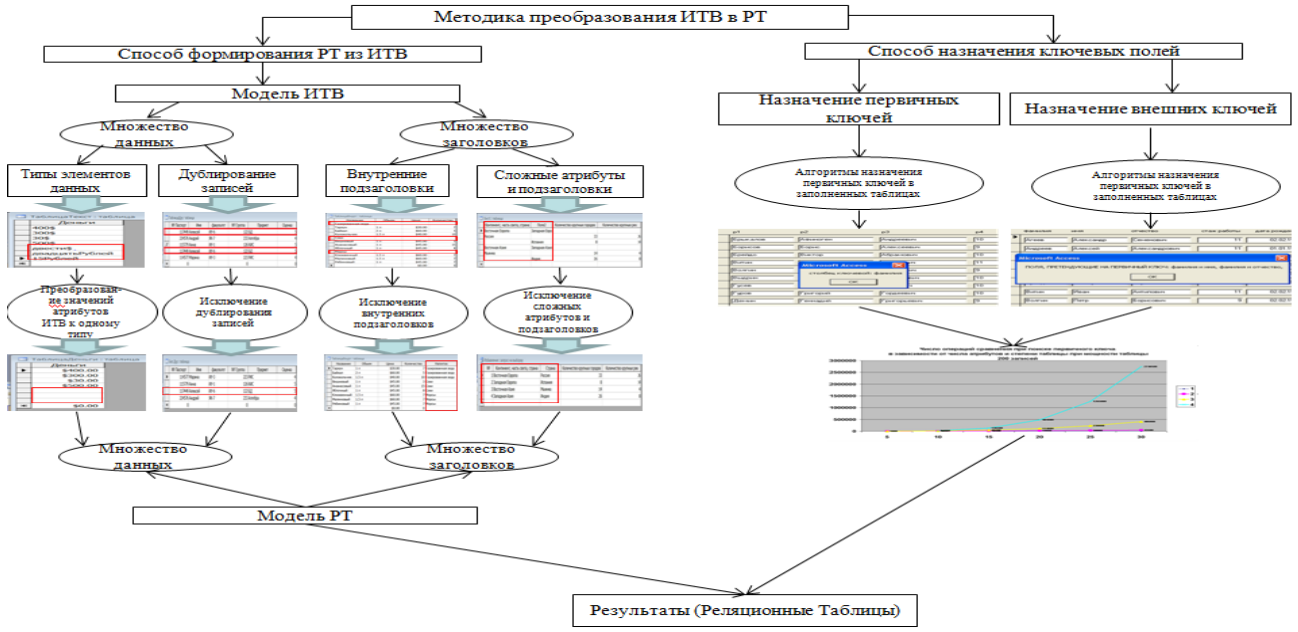

Рис. 10. Схематичное изображение методики преобразования ИТВ в РТ.

Программная реализация назначения ключевых полей выполнена для ключевых полей сформированных на базе одного, двух и трех атрибутов таблицы. На рис. 11 приведен фрагмент соответствующего интерфейса

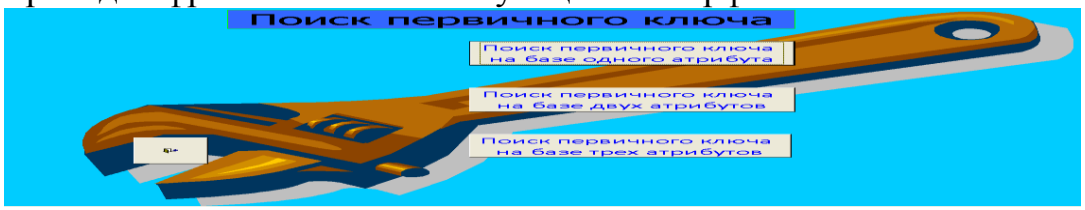

Рис. 11. Фрагмент интерфейса

Ниже приведен фрагмент результата функционирования процедуры, обеспечивающей поиск первичного ключа на базе одного атрибута.

В результате выполнения поиска первичного ключа на базе трех атрибутов может сформироваться сообщение типа рис. 12.

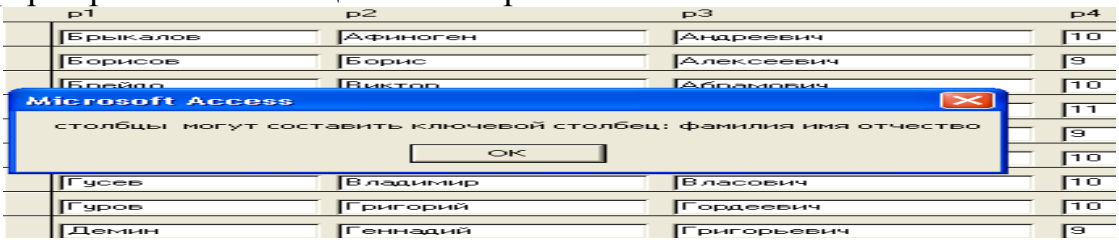

Рис. 12. Сообщение, формируемое при поиска ключа на базе 3-х атрибутов.

Оценка вычислительной сложности процедур назначения первичных ключей. Для анализируемой таблицы необходимо проверить все сочетания полей. Число возможных сочетаний C_n^k = n!/(n-k)!/k!, где n – общее число атрибутов таблицы, а к – количество проверяемых атрибутов на принадлежность к

первичному ключу. Выполнены экспериментальные оценки разработанных средств (рис. 13).

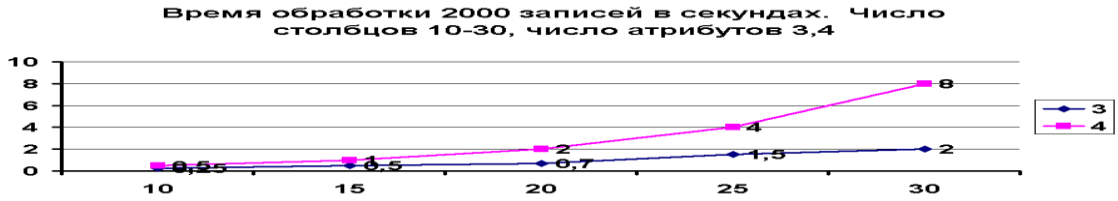

Рис. 13. Графики времени выполнения процедур в зависимости от числа столбцов и числа анализируемых атрибутов.

Анализ типов атрибутов таблицы. Все столбцы ИТВР имеют тип строковый. Поэтому задача состоит в том, чтобы проанализировать в каждом столбце все его элементы и по внешнему виду элементов определить его тип. На рис.14 приведен фрагмент интерфейса для анализа типов ИТВР.

Рис 14. Фрагмент интерфейса для анализа типов ИТВР.

После нажатия на соответствующую кнопку для каждого столбца сформируется сообщение о том, сколько полей не являются датами. Пример одного из сообщений приведен на рис. 14.

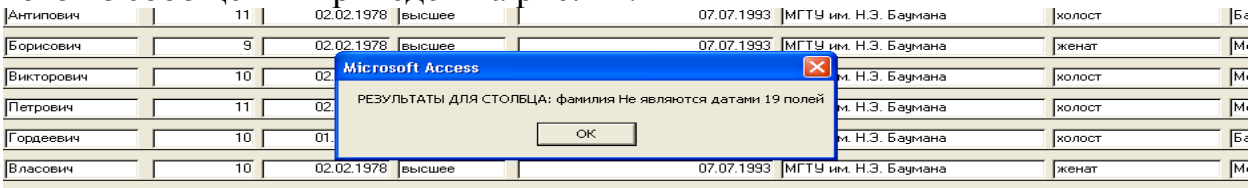

Рис. 14. Сообщение о том, сколько полей не являются датами.

Следующее сообщение (рис. 15) конкретизирует в каких записях таблицы значения полей не являются датами.

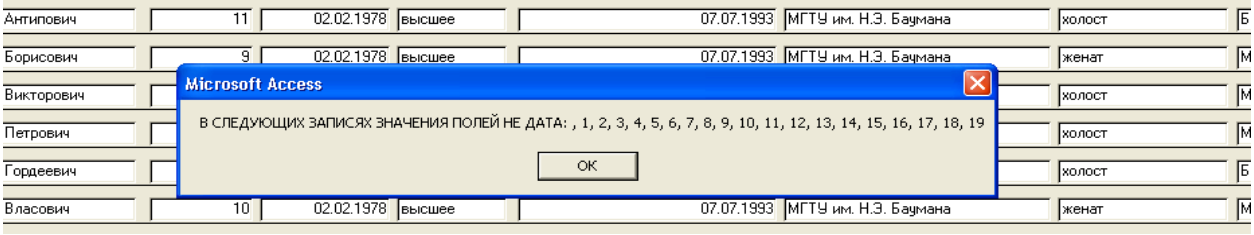

Рис. 15. Сообщение о том, в каких записях таблицы тип значения − не дата.

В соответствии с полученными сообщениями пользователь может принять решение о том, какие поля необходимо исправлять, а какие нет.

Основные результаты диссертации отражены в научных работах

1. Мин Тхет Тин. Электронная цифровая подпись. Современные информационные технологии // Сб. трудов кафедры ИУ-6. – М.: НИИ РЛ МГТУ им. Н.Э. Баумана, 2011. – С. 112–115.

2. Брешенков А.В., Мин Тхет Тин. Исключение внутренних подзаголовков

и избавление от сложных атрибутов при преобразовании нереляционных таблиц к реляционному виду // Современные информационные технологии: Сб. трудов кафедры ИУ-6. – М.: НИИ РЛ МГТУ им. Н.Э. Баумана, 2011. – С. 176–183.

3. Брешенков А.В., Мин Т.Т. Вычислительная сложность процедур назначения первичных ключей в заполненных таблицах // Информатика и системы управления в XXI веке: Сб. трудов МГТУ им. Н.Э. Баумана. – М.: МГТУ им. Н.Э. Баумана, 2012.– №9. – С. 136–142.

4. Брешенков А. В., Мин Т. Т. Мотивы разработки метода преобразования информации табличного вида в реляционное представление. // Инженерное образование, 2012. − №3. − 13 с. (Наука и образование: Эл. науч. издание. Номер гос. регистрации 0421200025.)

5. Брешенков А. В., Мин Т. Т. Аналитический обзор традиционного подхода формирования реляционных таблиц с учетом использования существующей информации табличного вида // Инженерное образование, 2012. − №8. − 16 с. (Наука и образование: Эл. науч. издание. Номер гос. регистрации 0421200025.)

6. Брешенков А. В., Мин Т. Т. Модели реляционных таблиц и информации табличного вида. // Инженерное образование, 2012. − №7. − 10 с. (Наука и образование: Эл. науч. издание. Номер гос. регистрации 0421200025.)

7. Брешенков А. В., Мин Т. Т. Алгоритмы назначения первичных ключей в заполненных таблицах. // Инженерное образование, 2012. − №6. − 14 с. (Наука и образование: Эл. науч. издание. Номер гос. регистрации 0421200025.)

8. Брешенков А.В., Мин Т. Т. Преобразование нереляционных таблиц к реляционному виду без использования сложных атрибутов // Вестник Московского государственного технического университета им. Н.Э.Баумана − М., 2012. − №2. – С. 59− 60.

9. Мин Тхет Тин, Брешенков А. В., Гудзенко Д. Ю. Назначение внешних ключей в заполненных реляционных таблицах // Современные компьютерные системы и технологии: Сб. трудов каф. ИУ−6 МГТУ им. Н.Э. Баумана, фак. “Информатика и системы управления ”. М., − 2012.– С. 128–135.

10. Мин Тхет Тин, Брешенков А.В., Гудзенко Д.Ю. Анализ типов атрибутов информации табличного вида // Современные компьютерные системы и технологии: Сб. трудов каф. ИУ-6 МГТУ им. Н.Э. Баумана, фак. “Информатика и системы управления”. М., − 2012.– С. 15–23.

11. Мин Т. Т. Анализ проблем разработки методики формирования реляционных таблиц на основе использования информации табличного вида. Россия в XXI веке: проблемы, тенденция, перспективы // Материалы XIV Международного симпозиума “Уникальные феномены и универсальные ценности культуры”: Сборник научных статей. – М.: МГТУ им. Н.Э. Баумана, 2012. – С. 270–272.

Автореферат

Диссертация на соискание ученой степени кандидата технических наук

Мин Тхет Тин

Тема диссертационного исследования

Методика формирования реляционных таблиц
на основе информации табличного вида

Научный руководитель
Брешенков Александр Владимирович

Изготовление оригинал-макета

Подписано в печать _____ Тираж _____ экз.

Усл. п. л. ____.

МГТУ им. Н.Э. Баумана

Отпечатано в типографии МГТУ им. Н.Э. Баумана. Заказ № ____