УДК 004.358:681.518

В.А. Филатов, Н.В. Касаткина

ОБ ОДНОМ ПОДХОДЕ К ИНТЕЛЛЕКТУАЛЬНОМУ АНАЛИЗУ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

Введение

Современный этап развития общества поднял индустрию информационных технологий до стратегического направления, в котором сосредоточены огромные интеллектуальные и финансовые ресурсы. Информация и инструменты управления информацией − программные продукты различного функционального назначения, приобрели статус информационных ресурсов.

Управление информацией претерпевает изменения в ряде его аспектов: в моделях хранения и доступа, в масштабах проектируемых систем – от баз данных (БД) до систем поддержки принятия решений. Исследованию вопросов, связанных с современным представлением информации, информационных ресурсов и информационных технологий посвящено ряд работ [1,2].

Рассматриваемая статья посвящена исследованию методов и моделей построения интегрированных информационных систем, объединяющих в себе реляционные базы данных как источник первичных данных и средства интеллектуального анализа, позволяющие прогнозировать, планировать развитие предметной области.

Целью работы является исследование и разработка эффективных моделей и методов проектирования информационных систем, объединяющих в одно целое БД и средства оперативного интеллектуального анализа данных.

1. Анализ состояния проблемы и постановка задачи исследования

В основе всех современных информационных как локальных, так и распределенных систем находится база данных. Опыт разработки, внедрения и эксплуатации таких систем показал, что наиболее эффективной структурой, удовлетворяющей требованиям, как разработчика, так и пользователя, является реляционная модель данных. Теоретические основы проектирования реляционных баз данных (РБД), предложенные Коддом [3], с 1970 года по настоящее время практически не изменились.

Неоднократно предпринимались попытки ревизии этой модели, замены ее на различные вариации с объектно-ориентированным подходом, однако опыт разработчиков и время показали – по совокупности выразительных средств и математической основы РМД еще долго будет основной моделью при проектировании БД. При этом, для программной реализации и поддержки РБД за последние 10 лет были вложены большие финансовые ресурсы: созданы и неоднократно модифицировались системы управления реляционными базами данных (СУБД) различного класса от FoxPro и MS ACCESS до ORACLE-8i.

Таким образом, к настоящему времени сложилась следующая ситуация. Базы данных стали неотъемлемой частью информационной поддержки от локальных систем в офисе, до территориально распределенных систем поддержки принятия решений, в которых к настоящему времени накоплены терабайты информации. Этот факт не остался без внимания специалистов в области информационных технологий. Абсолютно логично в такой ситуации возникла проблема эффективного использования годами накапливаемых и хорошо структурированных данных в целях получения качественно новой информации – знаний: закономерностей развития предметной области выраженных в виде законов, тенденций, таблиц, графиков и т.д., позволяющих планировать, прогнозировать ее развитие. Из таких предпосылок возникло современное направление – извлечение знаний из данных (data mining).

Эта область исследований, динамично развиваясь, выделилась в отдельное направление обширного спектра информационных технологий, однако еще раз следует отметить важный факт: data mining исследует методы и технологии обработки баз данных, в которых хранится накопленная в результате длительной эксплуатации информация.

Из известных в настоящее время методов анализа данных следует также выделить методы оперативного анализа баз данных – OLAP-системы. Такие системы основаны на многомерных структурах хранения данных. По типу используемой базы данных OLAP-системы могут разделяться на несколько классов в зависимости от структуры хранения данных:

– системы оперативной аналитической обработки многомерных баз данных (или MOLAP-системы), в которых данные организованы в виде упорядоченных многомерных массивов гиперкубов или поликубов;

– системы оперативной аналитической обработки реляционных баз данных (или ROLAP-системы), которые позволяют представлять данные в многомерной форме, обеспечивая преобразование информации в многомерную модель через промежуточный слой метаданных.

– гибридные системы оперативной аналитической обработки данных (HOLAP-системы) разработаны с целью совмещения достоинств и минимизации недостатков предыдущих систем. Они сочетают аналитическую гибкость и скорость ответа MOLAP-систем с постоянным доступом к реальным данным, свойственным ROLAP-системам [4].

К недостаткам рассмотренных выше методов оперативного анализа данных следует отнести главный – на практике многомерность данных, как правило, заканчивается использованием функций выделения из атрибутов типа «дата/время» расширений: год, полугодие, квартал, месяц, неделя. Все дальнейшие действия сводятся к последовательности выполнения специализированных запросов: «выборка» и «группировка» с использованием функций «CONT», «SUM», «AVG» или других аналогичных.

Целью проводимых научных исследований является разработка методов и моделей интеллектуального анализа реляционных баз данных. На первом этапе рассматривается задача: на основе формального анализа схемы базы данных представить подмножество всех зависимых атрибутов отношений модели данных типа «многие» ко «многим» N:M, и таким образом, ответить на вопрос: какие атрибуты могут потенциально хранить знания о предметной области, данные о которой хранит РБД.

2. Реляционная модель данных в задачах интеллектуального анализа

Рассмотрим общий подход к проектированию реляционной базы данных.

Для построения структурной схемы баз данных используются традиционные средства спецификации реляционной модели данных. Основной структурной единицей данных в реляционной модели является *n*-арное отношение, представляющее собой конечное подмножество декартова произведения доменов, т.е множеств атомарных значений элементов данных – атрибутов отношения.

Пусть R – конечное множество имен отношений базы данных;

D – $D_1,...,D_i$ – множество доменов, где всякий домен D_i есть именованное множество атомарных значений элементов данных;

A – конечное множество имен атрибутов отношений;

dom – отображение из A в D, определяющее из какого домена выбираются значения атрибутов.

Пару $\langle A_i, domA_i \rangle$, где $A_i \in A$ называют атрибутом.

Структурную схему S_i отношения R_i $R_i \in R$ можно представить в виде $R_i(A_1,...,A_n)$, в котором все A_i различны. Отношение r_i можно определить как расширение схемы $S_i : r_i \subseteq domA_1 \times ... \times domA_n$.

Перестановка атрибутов в схеме не порождает нового расширения и множество $A_1,...A_n$ атрибутов отношения R_i задает тип отношения. Для спецификации состава носителя используется выражение $R_i = A_1...A_n$. Структурная схема U реляционной базы данных – это спецификация вида $(R_1,..., R_p)$, где $R_i \in R$ и все R_i различны [5].

Концептуально реляционная база является информационно-логической моделью некоторой предметной области, такой, что каждое расширений соответствует некоторому состоянию данной области в определенный момент дискретно текущего времени. Каждое состояние моделируется упорядоченной совокупностью значений элементов данных, соответствующих значениям свойств объектов предметной области.

Объекту определенного типа соответствует кортеж отношения определенного типа. Заметим, что реляционная модель данных предполагает сильную типизацию объектов, использование вполне определенных категорий, таких как тип объекта, атрибут (свойство) объекта, домен и отнесение каждого значения и упорядоченной совокупности значений к одной из этих категорий. Объекты определенного типа обладают определенным набором свойств, что задается в реляционной модели схемой отношения.

На этапе проектирования разработчик выполняет обязательную процедуру – нормализацию схемы реляционной базы данных до одной из нормальных форм, как правило, это третья нормальная форма (3НФ). Приведем алгоритм нормализации универсального отношения.

Этап 1. На первом шаге задается одно или несколько универсальных отношений, отображающих понятия предметной области. В соответствии с предметной областью выделяются функциональные зависимости. Если отношения удовлетворяют условиям: в отношении нет одинаковых кортежей, кортежи не упорядочены и все значения атрибутов являются атомарными, то все отношения автоматически находятся в 1НФ.

Этап 2. Если в некоторых отношениях обнаружена зависимость атрибутов от части сложного ключа, то проводится декомпозицию этих отношений на несколько отношений следующим образом: те атрибуты, которые зависят от части сложного ключа выносятся в отдельное отношение вместе с этой частью ключа. В исходном отношении остаются все ключевые атрибуты:

Исходное отношение: $R(K_1, K_2, A_1, ..., A_n, B_1, ..., B_m)$.

Ключ: K_1, K_2 – сложный.

Функциональные зависимости:

$K_1, K_2 \rightarrow A_1, ..., A_n, B_1, ..., B_m$ – зависимость всех атрибутов от ключа отношения;

$K_1 \rightarrow A_1, ..., A_n$ – зависимость некоторых атрибутов от части сложного ключа.

Декомпозированные отношения:

$R_1(K_1, K_2, B_1, ..., B_m)$ – остаток от исходного отношения. Ключ K_1, K_2 .

$R_2(K_1, A_1, ..., A_n)$ – атрибуты, вынесенные из исходного отношения вместе с частью сложного ключа. Ключ K_1 .

Этап 3. Если в отношениях обнаружена зависимость неключевых атрибутов от других неключевых атрибутов, то проводится декомпозиция этих отношений: те неключевые атрибуты, которые зависят других неключевых атрибутов, выносятся в отдельное отношение. В новом отношении ключом становится детерминант функциональной зависимости:

Исходное отношение: $R(K, A_1, ..., A_n, B_1, ..., B_m)$.

Ключ: K .

Функциональные зависимости:

$K \rightarrow A_1, ..., A_n, B_1, ..., B_m$ – зависимость всех атрибутов от ключа отношения;

$A_1, ..., A_n \rightarrow B_1, ..., B_m$ – зависимость неключевых атрибутов других неключевых атрибутов.

Декомпозированные отношения:

$R_1(K, A_1, ..., A_n)$ – остаток от исходного отношения. Ключ K .

$R_2(A_1, ..., A_n, B_1, ..., B_m)$ – атрибуты, вынесенные из исходного отношения вместе с детерминантом функциональной зависимости. Ключ $A_1, ..., A_n$.

3. Разработка и исследование метода формирования функционально-зависимых переменных реляционной модели данных для решения задачи интеллектуального анализа

Технология нормализации позволяет решить несколько задач: избавиться от избыточности данных и уйти от проблем известных как «аномалии» – обновления, удаления и добавления. При этом теория функциональных зависимостей в процессе нормализации используется как критерий степени нормализации. Чем больше функционально-зависимых атрибутов в отношении тем, с точки зрения, теории проектирования реляционных баз данных, «хуже» проект. Однако, с точки зрения интеллектуального анализа данных, сильно нормализованное отношение не представляет ценности как источник знаний. Для доказательства этого утверждения рассмотрим определение функционально-зависимых атрибутов.

Пусть R – отношение. Множество атрибутов Y функционально зависимо от множества атрибутов X (X функционально определяет Y) тогда и только тогда, когда для любого состояния отношения R для любых кортежей $r_1, r_2 \in R$ из того, что $r_1X = r_2X$ следует что $r_1Y = r_2Y$. Символически функциональная зависимость записывается $X \rightarrow Y$.

Из определения следует: во всех кортежах, имеющих одинаковые значения атрибутов X, значения атрибутов Y также совпадают в любом состоянии отношения R . Таким образом, наличие функционально-зависимых атрибутов в отношении (в том числе и транзитивно-зависимых т.е. повторяющихся групп данных) – формальная основа поиска закономерностей групповых свойств данных.

Сформулируем задачу интеллектуального анализа данных применительно к реляционной модели.

Под задачей интеллектуального анализа данных будем понимать процедуру поиска всех пар атрибутов реляционной базы данных, удовлетворяющих условию $X \rightarrow Y$, для которых выполнимы групповые операции.

Решить поставленную задачу можно, по крайней мере, двумя способами:

1. На основе правил теории нормализации, анализируя вторую и третью нормальные формы, выбрать все пары функционально-зависимых атрибутов $X \rightarrow Y$ из отношения $R(K_1, K_2, A_1, ..., A_n, B_1, ..., B_m)$. В список пар также войдут атрибуты, транзитивно-зависимые между собой через отношения, в которых в качестве ключевых атрибутов используется составной ключ.

2. На основе анализа схемы базы данных, нормализованной до третьей нормальной формы.

Этот вариант формирования функционально-зависимых атрибутов для решения задачи интеллектуального анализа данных рассмотрим на примере фрагмента реляционной модели данных информационно-аналитической системы «Вища атестаційна комісія України» (ІАС «ВАК України»).

Пусть дано универсальное отношение **ЭКСПЕРТЫ_СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ**

СПЕЦИАЛИСТЫ_СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ

Код_Специалиста	Фамилия	Место работы	Тел	**Н_Спец совета**	Организация	Специальность
1	Иванов	ХНУРЭ	33-22	Д.О1.001	ХПИ	05.13.06
1	Иванов	ХНУРЭ	33-22	Д.02.011	ХАИ	05.13.06
2	Петров	ХНУРЭ	33-22	Д.О1.001	ХПИ	05.13.23
3	Сидоров	ХПИ	11-44	Д.О1.001	ХПИ	05.13.06
3	Сидоров	ХПИ	11-44	Д.02.011	ХАИ	05.13.23

Отношение **СПЕЦИАЛИСТЫ_СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ** находится в 1НФ, т.к. в отношении нет одинаковых кортежей, кортежи не упорядочены и все значения атрибутов являются атомарными. В качестве потенциального ключа отношения можно выбрать атрибуты **Код_Специалиста**, **Н_Спецсовета**. Рассматриваемое отношение не находится в 2НФ, т.к. существуют атрибуты, зависящие от части сложного ключа: **Код_Специалиста → Фамилия**; **Код_Специалиста → Место_работы; Код_Специалиста → Тел; Н_Спецсовета → Организация.**

Декомпозируем отношение **СПЕЦИАЛИСТЫ_СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ** на три отношения – **СПЕЦИАЛИСТЫ_МЕСТО_РАБОТЫ; СПЕЦСОВЕТЫ; СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ.**

СПЕЦИАЛИСТЫ_МЕСТО_РАБОТЫ

Код_Специалиста	Фамилия	Место работы	Тел
1	Иванов	ХНУРЭ	33-22
2	Петров	ХНУРЭ	33-22
3	Сидоров	ХПИ	11-44

СПЕЦСОВЕТЫ

Н_Спец совета	Организация
Д.О1.001	ХПИ
Д.02.011	ХАИ

СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ

Код_Специалиста	**Н_Спец совета**	Специальность
1	Д.О1.001	05.13.06
1	Д.02.011	05.13.06
2	Д.О1.001	05.13.23
3	Д.О1.001	05.13.06
3	Д.02.011	05.13.23

Отношение **СПЕЦИАЛИСТЫ_МЕСТО_РАБОТЫ** не находится в 3НФ, т.к. имеется функциональная зависимость неключевых атрибутов **Место_работы → Тел.** Декомпозируем его на два отношения – **СПЕЦИАЛИСТЫ**, **ТЕЛЕФОНЫ.**

СПЕЦИАЛИСТЫ

Код_Специалиста	Фамилия	Место работы
1	Иванов	ХНУРЭ
2	Петров	ХНУРЭ
3	Сидоров	ХПИ

ТЕЛЕФОНЫ

Место работы	Тел
ХНУРЭ	11-22-33
ХПИ	33-22-11

Схема реляционной базы данных **СПЕЦИАЛИСТЫ_СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ** представлена на рис.1.

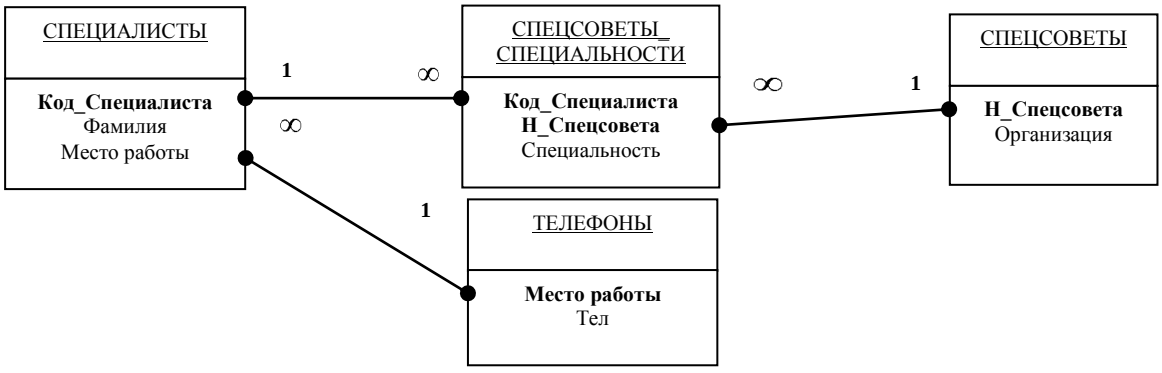

Рис.1. – Схема реляционной базы данных
СПЕЦИАЛИСТЫ_СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ

В результате анализа представленной выше схемы базы данных, можно выделить следующие зависимости и указать их информационное содержание:

Код_Специалиста → Н_Спецсовета – в каком количестве спецсоветов работает каждый специалист;

Код_Специалиста → Место работы – сколько специалистов работают в каждом вузе;

Код_Специалиста → Н_Спецсовета – сколько специалистов в каждом спецсовете;

Код_Специалиста → Специальность – сколько существует специалистов по каждой специальности;

Н_Спецсовета → Специальность – сколько и какие специальности в каждом спецсовете,

а также транзитивные связи через отношение **СПЕЦСОВЕТЫ_СПЕЦИАЛЬНОСТИ**:

Место работы → Н_Спецсовета → Специальность – сколько специалистов каждой специальности имеется в каждом вузе;

Место работы → Н_Спецсовета → Организация – сколько специалистов каждой организации в каждом спецсовете.

Выбранные зависимости и транзитивные связи позволяют осуществлять оперативный анализ данных и получать новые знания о предметной области.

Выводы.

На основе анализа особенностей проектирования реляционной модели данных рассмотрены основные проблемы построения интегрированных информационных систем, объединяющих в себе реляционные базы данных как источник первичных данных и средства интеллектуального анализа, позволяющие извлекать знания из данных, прогнозировать и планировать развитие предметной области.

В статье исследованы функциональные зависимости атрибутов и предложены варианты выбора атрибутов, информационная значимость которых может использоваться для интеллектуального анализа данных.

Приведен пример формирования функционально-зависимых атрибутов для решения задачи интеллектуального анализа данных.

ЛИТЕРАТУРА:

1. Саймон А.Р. Стратегические технологии баз данных: менеджмент на 2000 год: Пер. с англ. // Под ред. и с предисл. М.Р. Когаловского. – М.: Финансы и статистика, 1999. – 479 с.
2. Конноли Т., Бегг К. Базы данных: проектирование, реализация и сопровождение. Теория и практика, 2-е изд.: Пер с англ. – М.: «Вильямс», 2000. – 1120 с.
3. Мейер Д. Теория реляционных баз данных: Пер. с англ. – М.: Мир, 1987. – 608 с., ил.
4. Базы данных. Интеллектуальная обработка информации // В.В Корнеев, А.Ф. Гареев, С.В. Васютин, В.В. Райх. 2-е изд.М.: Нолидж, 2001. – 496 с.
5. Интеллектуальные технологии в задачах принятия решений технологических комплексов на основе нечеткой интервальной логики / Е. И. Кучеренко, Р. Н. Байдан, И. С. Творошенко, В. А. Филатов. // Восточно-европейский журнал передовых технологий. – 2006. – №2. – С. 92–96.

ФИЛАТОВ Валентин Александрович

Доктор технических наук, профессор кафедры Искусственного интеллекта Харьковского национального университета радиоэлектроники

Научные интересы: базы данных и знаний, агентные технологии, мультиагентные системы, извлечение знаний из данных.

КАСАТКИНА Наталья Викторовна

Главный специалист отдела технических наук ВАК Украины

Научные интересы: модели данных, базы данных и знаний, распределенные информационные системы.

ПРО ОДИН ПІДХІД ДО ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ РЕЛЯЦІЙНИХ БАЗ ДАНИХ

В.О. Філатов, Н.В. Касаткіна

Стаття присвячена дослідженню методів і моделей побудови інтегрованих інформаційних систем, що поєднують реляційні бази даних в якості джерела даних та засоби інтелектуального аналізу, що дозволяють добувати знання з даних. На основі аналізу властивостей реляційної моделі визначено функціональні залежності атрибутів, запропоновано підхід до вибору таких, що можуть бути використані

для аналізу даних, планування розвитку предметної області, тощо. Наведено приклад формування функціонально-залежних атрибутів для вирішення задачі інтелектуального аналізу даних.

ОБ ОДНОМ ПОДХОДЕ К ИНТЕЛЛЕКТУАЛЬНОМУ АНАЛИЗУ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

В.А. Филатов, Н.В. Касаткина

Рассматриваемая статья посвящена исследованию методов и моделей построения интегрированных информационных систем, объединяющих в себе реляционные базы данных как источник первичных данных и средства интеллектуального анализа, позволяющие прогнозировать, планировать развитие предметной области. Исследованы функциональные зависимости атрибутов и предложены варианты выбора атрибутов, информационная значимость которых может использоваться для интеллектуального анализа данных. Приведен пример формирования функционально-зависимых атрибутов для решения задачи интеллектуального анализа данных.

ABOUT ONE APPROACH TO INTELLECTUAL ANALYSIS OF RELATIONAL DATABASES

V. Filatov, N. Kasatkina

This article is devoted to the research of methods and models of integrated information systems. This systems use relational databases as the source of primary data and intellectual analysis tools that allow prediction, planning and developing of the data domain. Functional dependence of attributes was researched. Choices of attributes, whose the significance of the information can be used for data mining, were offered. Example of forming functional-dependant attributes has been made to solve the problem of intellectual data analysis.