УДК 004.652.4

МЕТОД И АЛГОРИТМ ФОРМИРОВАНИЯ СЕМАНТИЧЕСКОГО ОПИСАНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ НА ОСНОВЕ СХЕМЫ РЕЛЯЦИОННОЙ БАЗЫ ДАННЫХ

Баранчиков А.И.[1], Костров Б.В. [1], Громов А.Ю. [1]

[1] ФГБОУ ВПО «Рязанский государственный радиотехнический университет», Рязань, Россия (390005, Рязань, ул. Гагарина, д. 59/1), e-mail: rgrtu@rsreu.ru

При построении модели предметной области проектируемой информационной системы часть семантической информации представляется возможным извлечь из актуальных баз данных, которые на настоящий момент времени используются для выполнения задач обработки информации относительно рассматриваемой предметной области. Такие базы данных в достаточной мере адекватно отображают часть семантики предметной области, которая отображена в схеме реляционных баз данных. Предлагаются метод и алгоритм синтаксического анализа схем реляционных баз данных, представленных в виде текстовой информации, с целью их формализации для представления модели предметной области и ее дальнейшей реализации в реляционных базах данных. Исходной информацией для предложенного алгоритма является сценарий создания реляционной базы данных на языке SQL, который может быть получен как стандартными средствами СУБД, так и с помощью широко распространенных CASE-средств.

Ключевые слова: предметная область, сценарий, реляционная база данных, функциональная зависимость, атрибуты

METHOD AND ALGORITHM FOR THE FORMATION OF SEMANTIC DESCRIPTION OF THE SUBJECT AREAS BASED ON THE RELATIONAL DATABASE SCHEMA

Baranchikov A.I. [1], Kostrov B.V. [1], Gromov A.Y. [1]

[1] Ryazan State Radio-Engineering University, Russia, Ryazan (390005, Ryazan, ul. Gagarina, 59/1), e-mail: rgrtu@rsreu.ru

Building a domain model designed information system of the semantic information can be extracted from relevant databases, which at this point in time are used to perform information processing tasks to the subject area. Such databases are adequately represent the semantic content of the subject area that is displayed in the relational database schema. Method and algorithm for parsing schema relational database, presented in the form of text information, with a view to their formalization to represent the domain model and its further implementation in relational databases is proposed. Input information for the proposed algorithm is the scenario of creating a relational database in SQL language which may be obtained by standard database tools, and using widespread CASE-tools.

Keywords: subject area, script, relational database, functional dependence, attributes.

Цель исследования

Получение набора функциональных зависимостей в виде удобном для дальнейшего использования при анализе инфологии предметной области на основе их извлечения из структуры базы данных.

Введение

Перед разработчиками информационных систем часто возникает задача построения формального описания предметной области. Существуют различные варианты получения этого описания, отличающиеся принципами получения ключевых семантических связей между сущностями предметной области [1].

Одним из способов получения семантического описания предметной области является

проведение ее семантического анализа, который чаще всего проводится аналитиками и экспертами в области решаемой задачи автоматизации. В результате получается набор характеристик будущей системы, в которые входят требуемая функциональность, ограничения, накладываемые на данные предметной области, группы пользователей, параметры информационного окружения и другие не менее важные свойства.

Материал и методы исследования

Для предметных областей, которые уже подвергались анализу с целью автоматизации информационных процессов, целесообразнее проводить построение формального описания на основе уже имеющихся моделей данных. Во-первых, это значительно ускоряет процесс анализа ограничений, накладываемых на данные из предметной области. Во-вторых, в процессе эксплуатации информационной системы, реализованной на основе реляционной системы управления базами данных, могут быть выявлены семантические зависимости высоких порядков (например, многозначные зависимости или зависимости соединения [3]), в результате чего схема данных уточняется и дополняется ограничениями.

Таким образом, при построении модели предметной области, для которой уже разработаны схемы баз данных, целесообразно их использовать для получения семантической информации о предметной области. Это обусловлено тем, что одним из основных требований, которые предъявляются к базе данных, является ее адекватность предметной области в пределах ее функциональности. Адекватность обеспечивается соответствием семантики предметной области, которая может быть представлена в виде множества функциональных зависимостей (F-зависимостей), а также многозначных зависимостей (MV-зависимостей) и зависимостей соединения (J-зависимостей), и схемы базы данных $\boldsymbol{R} = \{R_1, R_2, ..., R_n\}$. Следовательно, если схема базы $\boldsymbol{R}$ данных построена верно, то реализованные в ней ограничения должны отражать семантику предметной области. При этом можно утверждать, что правильным является и обратное утверждение: если база данных адекватна предметной области, то по ее схеме можно судить о модели предметной области, которая в этом случае будет представлена в виде множества семантических зависимостей:

$\boldsymbol{U} = \boldsymbol{F} \cup \boldsymbol{MV} \cup \boldsymbol{J}$.

Однако утверждать с полной уверенностью, что схема базы данных $\boldsymbol{R}$ адекватна предметной области, нельзя, поскольку здесь играет важную роль элемент субъективизма. Это связано с тем, что разработчик базы данных мог не учесть определенных особенностей предметной области в силу тех или иных причин, или, наоборот, использовать ограничения, основанные на бизнес-правилах информационной среды, не связанные с нормализацией отдельных схем отношений R. Ограничения, которые не укладываются в классическую теорию нормализации реляционных отношений, представляют особый интерес, так как

реализуют логику информационного обмена и должны быть учтены [4,5]. Разработчики реализуют эту бизнес-логику на уровне клиентских приложений и/или с помощью дополнительных объектов баз данных.

Другой причиной неадекватности может служить значительное расширение функционала разрабатываемой информационной системы, который, естественно, не был учтен в базе данных, которая взята за основу для получения модели предметной области.

Таким образом, следует выделить три основных этапа для построения модели на основе реляционной базы данных:

- синтаксический анализ схем реляционных баз данных, представленных в виде сценариев создания физических баз данных на языке SQL;
- выявление неучтенных семантических зависимостей посредством анализа содержащихся данных;
- объединение результатов двух предыдущих этапов.

В работе остановимся на первом этапе, поскольку именно на нем можно получить большую часть семантических зависимостей.

Метод, основанный на анализе существующих схем баз данных, предназначен для извлечения ограничений предметной области из формализованной записи сценария создания базы данных. Сценарии подобного рода генерируются большинством современных систем управления базами данных.

Не вся информация, содержащаяся в сценарии создания базы данных, может быть использована для построения набора семантических зависимостей. Как правило, требуется проанализировать только блоки создания отношений (create table) и ограничений (constraint). Ограничения могут содержаться как внутри конструкций создания таблиц (задание первичных и внешних ключей), так и отдельно от них. Блоки, содержащие информацию о создании процедур, функций, представлений, правил, умолчаний, последовательностей и других объектов реляционных баз данных, не представляют интереса в контексте решаемой задачи.

Входные данные: сценарий создания схемы базы данных, записанный в стандарте SQL.

Выходные данные: множество семантических зависимостей $\boldsymbol{U}$, отражающих ограничения, накладываемые на данные предметной области.

Алгоритм ConstraintA:

Шаг 1. Преобразование входного сценария в удобный для анализа вид.

Шаг 2. Удаление избыточной информации, содержащейся в сценарии.

Шаг 3. Обработка блоков создания отношений.

Шаг 4. Обработка блоков отдельных ограничений и построение на их основе семантических зависимостей.

На первом шаге происходит конвертация сценария в унифицированный формат. Данный шаг требуется для преобразования файлов сценариев в случае их специфических особенностей хранения, так как не все современные системы управления базами данных хранят файлы сценариев в едином формате.

Исходный файл сценария преобразуется в текстовый файл, в котором каждая строка хранит описание одного объекта базы данных.

Второй шаг необходим для удаления строк сценария, не содержащих искомую информацию о семантике предметной области. Из сценария удаляются все строки, кроме блоков, описывающих создание отношений (таблиц), и отдельных ограничений.

На третьем шаге каждый блок создания таблицы T_i, содержащий набор ключевых атрибутов $\boldsymbol{X}$, неключевых атрибутов $\boldsymbol{Y}$ и ключевых атрибутов дочерних таблиц, зависимых от подмножества $\boldsymbol{X}$, преобразуется в набор семантических зависимостей $\{\boldsymbol{X \to Y, X' \to Z_1, ..., X'' \to Z_n}\}$:

$$T_i\{\boldsymbol{X,Y,Z}\} \Rightarrow \{\boldsymbol{X \to Y, X' \to Z_1, ..., X'' \to Z_n}\},$$

где $\boldsymbol{X', ..., X''}$ — подмножества ключевых атрибутов родительской таблицы, однозначно определяющих набор ключевых атрибутов дочерних таблиц, причем $\boldsymbol{X \subseteq X', ..., X \subseteq X''}$.

Множество семантических зависимостей $\boldsymbol{U_T}$, полученных на основе анализа таблиц T_i, представляет собой набор множеств $\boldsymbol{U_{Ti}}$:

$$\boldsymbol{U_{Ti} = \{ X \to Y, X' \to Z_1, ..., X'' \to Z_n \}};$$

$$\boldsymbol{U_T = U_{T1}, U_{T2}, ..., U_{Ti}, ..., U_{Tv}}.$$

На четвертом шаге происходит заполнение множества $\boldsymbol{U_C}$ семантических зависимостей, выявленных из блоков отдельных ограничений первичных и внешних ключей.

Ограничения внешнего ключа $\boldsymbol{CF_j}$ содержат набор ключевых атрибутов $\boldsymbol{X}$ и неключевых атрибутов $\boldsymbol{Y}$. Каждое такое ограничение преобразуется в семантическую зависимость вида $\boldsymbol{X \to Y}$, которая включается в множество зависимостей $\boldsymbol{U_{CF}}$:

$$\boldsymbol{CF_j\{ X, Y \} \Rightarrow U_{CFj} = \{ X \to Y \}};$$

$$\boldsymbol{U_{CF} = U_{CF1}, U_{CF2}, ..., U_{CFj}, ..., U_{CFw}}.$$

Ограничения первичного ключа CP_k содержат набор ключевых атрибутов $\boldsymbol{X}$ и ссылку на таблицу T_m, в которой данный ключ используется в качестве первичного. Каждое такое

ограничение преобразуется в зависимость следующего вида:

$T_m => U_{CPm} = \{ \boldsymbol{X} \rightarrow \boldsymbol{Y} \}$;

$CP_k \{ \boldsymbol{X}, T_m \} => U_{CPm} = \{ \boldsymbol{X} \rightarrow \boldsymbol{Y} \}$,

где $\boldsymbol{Y}$ — множество неключевых атрибутов таблицы T_m.

В результате формируется набор семантических зависимостей $\boldsymbol{U_{CP}}$:

$\boldsymbol{U_{CP}} = \boldsymbol{U}_{CP1}, \boldsymbol{U}_{CP2}, ..., \boldsymbol{U}_{CPm}, ..., U_{CPq}$.

Итоговое множество семантических зависимостей, описывающих ограничения, накладываемые на данные предметной области, является объединением полученных множеств зависимостей:

$\boldsymbol{U} = \boldsymbol{U_T} \cup \boldsymbol{U_{CF}} \cup \boldsymbol{U_{CP}}$.

Представим алгоритм, базирующийся на описанном методе, в виде псевдокода:

Алгоритм **ConstraintA**;

begin

Шаг 1. Конвертация сценария в унифицированный текстовый формат;

Шаг 2. **for** каждая строка из сценария **do**

if (строка не содержит описания таблицы) **or**

(строка не содержит описания ограничения ключа)

then

Удалить текущую строку;

Шаг 3. **for** каждая таблица $T_i \{ \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \}$ из сценария **do**

begin

$U_{Ti} = \{ \boldsymbol{X} \rightarrow \boldsymbol{Y}, \boldsymbol{X'} \rightarrow \boldsymbol{Z}_1, ..., \boldsymbol{X''} \rightarrow \boldsymbol{Z}_n \}$;

$\boldsymbol{U_T} = \boldsymbol{U_T} + U_{Ti}$

end

Шаг 4. **for** каждое ограничение ключа C_h из сценария **do**

if (C_h — ограничение внешнего ключа)

then **begin**

$U_{CFj} = \{ \boldsymbol{X} \rightarrow \boldsymbol{Y} \}$;

$\boldsymbol{U_{CF}} = \boldsymbol{U_{CF}} + U_{CFj}$

end

else **begin**

$U_{CPm} = \{ \boldsymbol{X} \rightarrow \boldsymbol{Y} \}$;

$$\boldsymbol{U_{CP}} = \boldsymbol{U_{CP}} + U_{CPm}$$

end

$\boldsymbol{U} = \boldsymbol{U_T} \cup \boldsymbol{U_{CF}} \cup \boldsymbol{U_{CP}}$;

end.

Результаты исследования

Следует отметить, что разработчики информационных систем часто отходят от принципов нормализации для достижения определенных целей автоматизации бизнес-процессов. Например, при использовании технологии OLAP часть отношений может быть денормализована до нормальных форм низкого порядка с целью построения многомерных кубов. Другим примером отступления от канонов реляционной модели хранения информации является использование избыточных отношений, хранящих временную или статистическую информацию.

Выходное множество семантических зависимостей $\boldsymbol{U}$ нельзя назвать минимальным, так как избыточные зависимости являются неотъемлемой частью современных информационных систем. В случае необходимости полученное множество семантических зависимостей $\boldsymbol{U}$ может быть минимизировано с помощью поиска классов эквивалентности и построения на их основе кольцевых покрытий с дальнейшей минимизацией [3].

Предложенный алгоритм является сходящимся, поскольку в нем рассматривается конечное число конструкций сценария на языке SQL, который описывает конечное число объектов базы данных, содержащих конечное число атрибутов.

Заключение

Предложены метод и базирующийся на нем алгоритм, который позволяет за счет анализа сценария создания базы данных (сценарий может быть получен стандартными средствами систем управления базами данных) получить математическую модель предметной области. Модель описывается в виде набора семантических зависимостей, которые представлены в специально разработанной структуре базы данных, что дает возможность ее дальнейшего использования в процессе автоматизированного проектирования информационных систем [2].

Список литературы

1. Баранчиков А. И., Костров Б.В. Теория и методы исследования моделей и алгоритмов представления данных для предметных областей с ранжируемыми атрибутами // Вестник

РГРТУ. № 4 (выпуск 46), часть 2. 2013. С. 59–64.

2. Баранчиков А. И., Громов А. Ю. Алгоритм построения схемы реляционной базы данных на основе множества многозначных и функциональных зависимостей, учитывающий атрибуты различной степени секретности // Системы управления и информационные технологии: науч.-техн. журн. № 1(43). Москва; Воронеж, 2011. С. 53–56.

3. Мейер Д. Теория реляционных баз данных: Пер. С англ.: М.: Мир, 1987. — 608 с.

4. Дейт К. Дж. Введение в системы баз данных // 8-е изд. Пер. с англ. М: Вильямс, 2005. - 1328 с.

5. Date C.J. Twelve Rules for Business Rules // (May 1, 2000).

Рецензенты:

Овечкин Г.В., д.т.н., профессор, профессор кафедры ВПМ, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Рязанский государственный радиотехнический университет», Министерство образования и науки РФ, г. Рязань;

Скворцов С.В., д.т.н., профессор, профессор кафедры САПР ВС, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Рязанский государственный радиотехнический университет», Министерство образования и науки РФ, г .Рязань.