УДК 004.932.2:681.3.06

В.О. Радченко, С.С. Танянский

Харьковский национальный университет радиоэлектроники, Харьков

ВЫЯВЛЕНИЕ СКРЫТЫХ ЗАВИСИМОСТЕЙ МЕЖДУ ДАННЫМИ В ЗАДАЧАХ РЕИНЖИНИРИНГА ИНФОРМАЦИОННЫХ СИСТЕМ

Предлагается подход к выявлению новых, ранее неизвестных функциональных зависимостей (ФЗ), основываясь на множестве данных реляционной базы данных. Методы интеллектуального анализа данных (Data Mining) применяются для получения множества строгих ФЗ, удовлетворяющих состоянию базы данных на момент проведения обработки, а также проводится проверка членства отдельно взятой ФЗ в замыкании исходного множества для исключения зависимостей, которые могут быть получены из других с помощью правил вывода.

Ключевые слова: *реинжиниринг, реляционная база данных, функциональная зависимость, выявление зависимостей, универсальное отношение, замыкание функциональных зависимостей.*

Введение

Современные информационные системы (ИС), построенные на основе автоматизированных систем БД, призваны выполнять положенные задачи в тяжелых условиях непрерывного потока требований предметной области (например, усовершенствование бизнес-процессов, расширение круга интересов организации, различные требования со стороны пользователей). Следствием этого является адаптация отдельных компонентов ИС, в частности базы данных, под новые требования. На сегодняшний день наиболее широкое распространение получили реляционные базы данных (РБД), которые обеспечивают наилучшее сочетание простоты, надежности и производительности для решения широкого класса задач.

Одним из основных направлений развития в области баз данных на данный момент является реинжиниринг, позволяющий проводить перепроектирование существующих БД, используя максимум полезной информации, которую можно получить в результате анализа исходной структуры и данных БД. Такой подход позволяет существенно уменьшить затраты средств и времени на проведение перепроектирования.

В данной работе рассматривается задача выявления информации о взаимосвязях между данными, которые могли установиться в процессе функционирования БД. Взаимосвязи представляются в виде зависимостей различных типов, которые затем можно использовать в качестве исходных данных для методов повторного проектирования РБД.

Анализ исследований и публикаций. В научных работах большое внимание уделяется реинжинирингу устаревших БД, то есть таких, которые уже не отвечают текущим требованиям организации; их поддержка и обслуживание крайне затруднительны [1]. С этой целью разрабатываются методики восстановления структуры данных, обусловленной их взаимосвязями, для последующего анализа и переноса данных на современную платформу, как правило, реляционную. Данные имеют наибольшую ценность, поэтому комплекс вышеописанных мер направлен на минимизацию потерь значащей информации в процессе переноса. Другой областью интересов является поддержка и обслуживание современных РБД.

Методы выявления взаимосвязей между данными преимущественно используют ФЗ как средство представления таких связей [2]. Это обусловлено тем, что функциональные зависимости позволяют наиболее простым образом представить связи между объектами рассматриваемой предметной области. Другими типами зависимостей, которые принимаются во внимание, являются зависимости включения и многозначные зависимости, но их использование и методы выявления не рассматриваются в данной работе. Следует заметить, что рассмотренные методы направлены, в основном, на использование в системах интеллектуального анализа данных (Data Mining) и ориентированы на выявление приближенных функциональных зависимостей, которые позволяют представить предполагаемые связи, имеющие некоторую погрешность. В рамках данной работы использование таких методов позволяет также получать множество строгих ФЗ, то есть, справедливых для всего набора входных данных на момент времени проведения обработки.

Постановка задачи. В данной работе рассматривается решение задачи выявления ранее не известных функциональных зависимостей из множества данных целевой РБД, которые будут гарантированно корректными на момент проведения обработки. Задача выявления скрытых зависимостей является составной частью задачи реинжиниринга и относится к этапу предварительного сбора информации об исследуемой РБД. Описанный способ яв-

ляется вариантом для построения автоматизированного решения, непосредственно ориентированного на выявление новых зависимостей в данных, порождаемых предметной областью.

Выявление скрытых зависимостей

Исходными данными для решения поставленной задачи являются: логическая схема реляционной БД $\Sigma = \{\sigma_i\}, i = \overline{1,n}$, где σ_i – схема одного отношения, входящего в БД, n – количество отношений; схема отношения $\sigma_i =< R_i, F_i >$, где R_i – носитель отношения (множество атрибутов), а F_i – множество функциональных зависимостей (ФЗ), удовлетворяющих данному отношению. $P = \{\rho_i, i = \overline{1,n}\}$ – множество отношений рассматриваемой БД.

Существование функциональной зависимости вида $A \to B$ означает, что для любых двух кортежей u, v некоторого отношения ρ_k имеет место следствие: $u(A) = v(A) \Rightarrow u(B) = v(B)$. В качестве сопутствующего примера рассмотрим логическую схему $\Sigma = \{\sigma_1, \sigma_2\}$, состоящую из двух схем отношений:

$$\sigma_1 =< R_1 = \{A, B, C\}, F_1 >, \sigma_2 =< R_2 = \{C, D\}, F_2 > .$$

Предположим, что информация о F_1, F_2 отсутствует или утеряна. Получить множество ФЗ, удовлетворяющих данным отношениям, возможно с помощью метода выявления ФЗ из экземпляров рассматриваемых отношений, в частности, метода Tane, принцип и реализация которого подробно описаны в [3]. В результате его применения будет получено множество минимальных ФЗ, удовлетворяющих набору данных в отношении на момент проведения обработки. Минимальная ФЗ – это такая ФЗ вида $X \to Y, X = \{A_1, \ldots, A_n\}, Y = \{B_1, \ldots, B_m\}$, в которой не существует множества $Z \subset X$, для которого бы соблюдалось $Z \to Y$. Тривиальные ФЗ вида $A_1 \to A_1$ игнорируются используемым методом, поскольку не являются значимыми.

Рассмотрим пример; отношения ρ_1, ρ_2 приведены ниже в табл. 1, 2.

:

Таблица 1

Отношение ρ_1

A	B	C
1	1	1
2	2	1
1	2	2
1	3	3
3	1	1
4	2	2
2	2	4

Таблица 2

Отношение ρ_2

C	D
1	1
2	1
3	1
4	1

Используя метод Tane, получены такие множества ФЗ для приведенных отношений:

$$F_1 = \{AC \to B\}, F_2 = \{C \to D\} .$$

Принимая во внимание множества ФЗ для F_1 и F_2, множество ФЗ схемы Σ будет иметь вид:

$$F_\Sigma = \bigcup_{i=1}^{2} F_i = \{AC \to B, C \to D\};$$

Носитель универсального отношения $R = \{R_1 \cup \ldots \cup R_n\}$ для рассматриваемого примера выглядит следующим образом:

$$R = R_1 \cup R_2 = \{A, B, C\} \cup \{C, D\} = \{A, B, C, D\} .$$

Универсальное отношение может быть получено через естественное соединение всех отношений, входящих в схему. Результат соединения для примера приведен ниже (табл. 3).

Таблица 3

Универсальное отношение

A	B	C	D
1	1	1	1
2	2	1	1
1	2	2	1
1	3	3	1
3	1	1	1
4	2	2	1
2	2	4	1

Применяя метод Tane для полученного универсального отношения, будут выявлены следующие ФЗ: $\overline{F_\Sigma} = \{AC \to B, C \to D, A \to D, B \to D\}$. Данное множество содержит все минимальные ФЗ, принадлежащие F_Σ, а также дополнительные, ранее не известные ФЗ: $F'_\Sigma = \{A \to D, B \to D\}$. Следовательно, множество ФЗ универсального отношения для логической схемы Σ можно выразить таким образом:

$$\overline{F_\Sigma} = F_\Sigma \cup F'_\Sigma ,$$

где F_Σ – множество ФЗ, удовлетворяющих исходным отношениям Σ, а F'_Σ – множество дополнительных ФЗ.

Необходимо установить, являются ли ФЗ из множества F'_Σ выводимыми из F_Σ, или же они являются новой информацией. Для этого предлагается использовать метод проверки принадлежности ФЗ к замыканию $(F_\Sigma)^+$, предложенный Мейером в [4] – решение проблемы членства. Принцип состоит в следующем: так как построение F^+ связано с перебором всех подмножеств множества атрибутов, принадлежащих F, и имеет экспоненциальную сложность, то предлагается строить F-замыкание на множестве атрибутов. F-замыканием множества X называется такое множество атрибутов X^+, что $X \to X^+ \in F^+$ и не существует ни одного атрибута в

R, который бы зависел от X и не принадлежал X^+ [3]. Реализация метода построения F-замыкания имеет линейную сложность. Таким образом, метод проверки принадлежности ФЗ $X \to Y$ к замыканию F^+ заключается в построении F-замыкания X^+ и определения истинности выражения $Y \subseteq X^+$. Если выражение истинно, то $X \to Y \in F^+$.

Рассмотрим пример: для того, чтобы проверить $A \to D$ на принадлежность к F_Σ^+, нужно построить A^+. В соответствии с [2], $A^+ = \{A\}$, поскольку F_Σ не содержит ФЗ, для которых A являлась бы единственным атрибутом в левой части ФЗ. $D \notin A^+$, следовательно, $A \to D \notin F_\Sigma^+$. Аналогичным образом показывается, что $B \to D \notin F_\Sigma^+$. Таким образом, множество $F_\Sigma^{'}$ выявленных зависимостей является не выводимым и, тем самым, представляет собой новую информацию.

Следует отметить, что данный подход не гарантирует полное соответствие новых зависимостей рассматриваемой предметной области. Так как он базируется на множестве данных, содержащемся в РБД на момент проведения обработки, и не учитывает их семантику, велика вероятность получения случайных ФЗ. Случайная ФЗ – это такая ФЗ, которая не является корректной для конкретной предметной области (например, дата рождения человека определяет дату рождения его ребенка) и может быть устранена в любой момент путем изменения либо добавления кортежей с данными, противоречащими выявленной зависимости, в процессе функционирования РБД.

Таким образом, ставится еще одна задача – проверка новых зависимостей на предмет корректности для предметной области, которая моделируется анализируемой РБД. В данной работе ее решение не рассматривается и представляет собой направление для дальнейших исследований. В качестве возможных путей решения можно предложить использование экспертной оценки, либо же некоего численного критерия для каждой отдельной ФЗ, что позволит устанавливать пороговое значение, ниже которого зависимости будут считаться случайными.

Выводы

В работе предложен подход к выявлению ранее неизвестных функциональных зависимостей, который основывается на анализе множества данных реляционной БД. Первым шагом является получение множества ФЗ для каждого отношения. На втором шаге проводится аналогичная операция для универсального отношения рассматриваемой РБД. На этом шаге становится возможным выявить ФЗ между атрибутами различных отношений – взаимосвязи между данными, которые установились в процессе функционирования РБД. Предложен способ определения их информационной новизны, который состоит в проверке членства ФЗ универсального отношения в замыкании объединения множеств ФЗ отдельных отношений.

Направлением для дальнейших исследований является разработка методов и способов для осуществления проверки полученных зависимостей на предмет корректности для предметной области, которая моделируется анализируемой РБД.

Список литературы

1. Rossiter Nick. Re-engineering relational databases: the way forward / Nick Rossiter. – ISWSA '11, ACM New York, NY, USA, 2011. – 17 с.

2. Henrard J. Data dependency elicitation in database reverse engineering / J. Henrard // Software Maintenance and Reengineering Conference, 2001. – С. 11-19.

3. Ykä Huhtala. Tane: An Efficient Algorithm For Discovering Functional and Approximate Dependencies / Ykä Huhtala // The Computer Journal 42 (2),1999. – С. 100-111.

4. Мейер Д. Теория реляционных баз данных: Пер. с англ. / Д. Мейер. – М.: Мир, 1987. – 609 с.

Поступила в редколлегию 28.02.2012

Рецензент: д-р техн. наук, проф. О.Г. Руденко, Харьковский национальный университет радиоэлектроники, Харьков.

ВИЯВЛЕННЯ ПРИХОВАНИХ ЗАЛЕЖНОСТЕЙ МІЖ ДАНИМИ В ЗАДАЧАХ РЕІНЖИНІРИНГУ ІНФОРМАЦІЙНИХ СИСТЕМ

В.О. Радченко, С.С. Танянський

Пропонується підхід до виявлення нових, раніш невідомих функціональних залежностей (ФЗ), базуючись на множині даних реляційної бази даних. Методи інтелектуального аналізу даних (Data Mining) використовуються для отримання множини строгих ФЗ, що задовольняють стану бази даних на момент проведення обробки, також проводиться перевірка членства відібраної ФЗ у замиканні початкової множини для виключення залежностей, що можуть бути отримані з інших за допомогою правил виводу.

Ключові слова: *реінжиніринг, реляційна база даних, функціональна залежність, виявлення залежностей, універсальне відношення, замикання функціональних залежностей.*

DISCOVERY OF HIDDEN DATA RELATIONSHIPS IN TASKS OF INFORMATION SYSTEMS REENGINEERING

V.A. Radchenko, S.S. Tanyansky

Approach for the discovery of new, not obtained before functional dependencies (FD) is proposed. It is based on the set of data stored in relational database. Methods of intellectual data analysis (Data Mining) are used to obtain set of strict FDs, that satisfy database state on the moment of processing. Also every single FD goes through initial closure membership check to exclude FDs that can be obtained from others with the rules of inference help.

Keywords: *reengineering, relational database, functional dependency, discovery of dependencies, universal relation, closure of functional dependencies.*