ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2018. No. 1 (3)*

UDC 004.932.2:681.3.06 DOI: https://doi.org/10.30837/2522-9818.2018.3.054

V. FILATOV, S. DOSKALENKO

ON THE APPROACH TO SEARCHING FOR FUNCTIONAL DEPENDENCES OF DATA IN RELATIONAL SYSTEMS

The **subject** matter of the study is information systems built on the basis of relational databases. The **goal** of the article is to develop a method for re-engineering relational databases that takes into account implicit interrelated functionally dependent data that affect the structure of the logical model. The following **results** are obtained: the approach to identify previously unknown functional dependencies based on the analysis of a set of relational database data is suggested; the classes of tasks of reengineering relational databases are specified; the stage of developing the target logic diagram which is common for the problems of adaptation and refactoring was studied; the sub-task of checking if the logic diagram of the relational database corresponds to the third normal form within this stage is considered using the synthesis method; it is shown that the solution of this task involves a number of difficulties, in particular, it is necessary to find a set of functional dependencies that are performed on the current instance of the data of a relational database; the approach for finding a set of functional dependencies from an instance of the data of a relational structure is suggested. The direction of further research can be the support of empty values at the stage of identifying functional dependencies as well as the issues of data transfer without any loss from the initial structure of the database to the target data obtained as a result of applying the methods of re-engineering. **Conclusions.** The approach is suggested to identify previously unknown functional dependencies which are based on the analysis of a set of relational database data. The first step is to get a set of functional dependencies for each relationship. The similar operation for the universal relation of the target database is performed at the second step. At this step, functional dependencies among the attributes of different relationships, that is the interrelationships among the data that were established during the information system operation, can be identified. The method for determining their information novelty is suggested; this method consists in verifying the membership of the functional dependencies of the universal relation while discovering the union of sets of dependencies of individual relations. A promising direction for further research is the development of methods to implement the technology for verifying if the obtained dependencies correspond to the logical model of the domain.

Keywords: reengineering, relational database, functional dependence, discovering dependences, universal relation, closing functional dependencies.

Introduction

The rapid development of information technology has led to major changes in all stages of the life cycle of information systems (IS), in particular, at the stage of support and maintenance. Constantly changing requirements in the condition of shortened terms of development adversely affect both the quality of the product and the further opportunities for its development and maintenance. Situations often happen when further support for individual components or the whole system becomes extremely difficult because of the significant complexity of the internal structure while the design of a new system that can meet current requirements is inappropriate for a number of reasons (economic or time factors and so on). One way to solve such problems is to carry out re-engineering, whose goal is to improve the characteristics of the initial IS based on a preliminary analysis of the IS current state and its individual components [1].

The basic problems of information system design

The development and implementation of many modern large information projects are usually long, their cost exceeds the planned one and the final product can be unreliable and difficult to maintain. All this leads to the situation that is known as the “software crisis”. Although the crisis was first mentioned in the late 1980s, even after 30 years it still cannot be bridged over.

Some of the reasons for the general problem of designing complex information systems are as follows:

- the development of about 40% of systems fails to be successful or stops before the work is completed;

- business interests are rationally integrated and the developed information technology is used only in 25% of systems;

- only 20-30% of information systems meet all the criteria for achieving success.

The major failures in developing software have been caused by the lack of the complete specification of all requirements within the design phase, the shortage of acceptable development methodology or by the fact that a general global project is not sufficiently divided into separate components that can be effectively controlled and managed.

In the case of the partial implementation of the requirements of users of the information system or a change in the business process to such an extent that the system ceases does not meet the requirements of users any more, several options of the development are possible:

- developing a new system;

- modifying (developing) the existing system (legacy system);

- reengineering the existing system (legacy system).

The first option is the simplest and preferable for the developer but does not satisfy the requirements of users because additional time and financial resources are required; besides, there are development risks and the risks of loss of the accumulated information within the life cycle of the used IS. The re-engineering of legacy information systems requires that experts in the field of information systems and technologies should be engaged, which makes the work overcomplicated. There is an opinion that in most cases it is easier to develop a new system than to re-engineer it. This is connected with the qualification of specialists who need to be involved in work. Their skills should be sufficiently high level to solve a set of design problems and create a modified information system [1].

Re-engineering of information systems in general and databases, in particular, has become an object of close attention and active studies. To expand the functionality and improve the performance of available systems, the process of re-engineering requires determining and understanding all the components of such systems. The database, which is a part of the system, is the most important component of the information system. Nowadays, relational databases (RDBs) are dominant and the vast majority of available software applications and services use them. This is the reason for selecting the RDB as the object of the research.

Current studies in the field of this topic were analyzed. The main areas of development are re-engineering of the logic diagram with the help of an intermediate representation such as an ER-model or own metamodels and applying of a set of pre-set rules for translating model objects in RDB designs, re-engineering of obsolete databases; extracting the structure of both obsolete and relational databases and presenting it as a conceptual data model, in particular, an ER-model [2, 3].

This article examines the task of identifying information about the relationships among data that could be established during the operation of the database. Relationships are represented as dependencies of various types, which can then be used as input data for the methods of re-designing the RDB. To achieve this, methods for restoring the data structure conditioned by their interrelations are being developed for later analyzing and transferring data to a modern platform, usually relational one. The data are of the greatest value, therefore a range of measures described above is aimed at minimizing the loss of meaningful information during the transfer process. Another area of interest is the support and maintenance of modern RDBs.

Methods to identify the relationships among the data preferentially use functional dependencies (FD) as a means of representing such relationships [4]. This happens due to the fact that functional dependencies enable representing in the simplest possible manner the relationships among the objects of the target subject domain. Other types of dependencies that are taken into account are inclusion dependencies and multi-valued dependencies but their use and methods of detection are not considered in this paper. It should be noted that the mentioned methods are directed primarily at using in data-mining systems and focus on identifying approximate functional dependencies that enable presenting probable links that have a certain error [5]. In this paper, the use of such methods also makes it possible to obtain a set of strict FDs, that is, valid for the entire set of input data at the time of processing.

Problem statement

This article deals with the methods for solving the task to identify previously unknown functional dependencies from a set of data of the target RDB, which will be definitely correct at the time of processing. The task of discovering hidden dependencies is an integral part of the task of re-engineering and refers to the stage of preliminary collection of information about the target RDB. The described method is a variant for making an automated decision directly oriented to the detection of new dependences in the data generated by the target area [6].

Discovering hidden dependences

Initial data for solving the set task are – the logic diagram of a relational database $\Sigma = \{\sigma_i\}, i = \overline{1,n}$, where σ_i is a diagram of one relation that is a *DB* part, n is a number of relations; the diagram of relations $\sigma_i =< R_i, F_i >$, where R_i is the medium of relations (many attributes), F_i is a set of functional dependences (*FD*) that meet this relation. $\mathrm{P} = \{\rho_i, i = \overline{1,n}\}$ is a set of relations of the target *DB*.

The functional dependence of $A \to B$ type indicates that for any two tuples u, v of a certain relation ρ_k there is the conclusion $u(A) = v(A) \Rightarrow u(B) = v(B)$. As an associated example, the logic diagram $\Sigma = \{\sigma_1, \sigma_2\}$ is considered, this diagram comprises two diagrams of relations: $\sigma_1 =< R_1 = \{A,B,C\},\ F_1 >, \sigma_2 =< R_2 = \{C,D\},\ F_2 >$. Let us assume that there is no information about F_1, F_2 or it has been lost. A set of FD that meet these relations can be obtained using the method of FD detection from instances of target relations, in particular, Tane method whose principles and methods of implementation are detailed in [7, 8]. When this method is used, a set of minimal FDs that meet a set of data in the relation at the moment of processing can be obtained. A minimal FD is the FD of the $X \to Y, X = \{A_1, \dots, A_n\}, Y = \{B_1, \dots, B_m\}$ type that does not contain a set of $Z \subset X$ where $Z \to Y$ is true. Trivial FDs of the $A_1 \to A_1$ type are neglected by the used method as they are not significant. Let us consider the example; the relations ρ_1, ρ_2 are presented below:

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

ρ_2

C	D
1	1
2	1
3	1
4	1

Using Tane method, the following sets of FD for the given relations $F_1 = \{AC \to B\}, F_2 = \{C \to D\}$ are obtained.

Taking into considerations sets of FD for F_1 and F_2, a set of FD diagram Σ is as follows: $F_\Sigma = \bigcup_{i=1}^{2} F_i = \{AC \to B, C \to D\}$; the medium of universal relation $R = \{R_1 \cup ... \cup R_n\}$ for the target example is as follows: $R = R_1 \cup R_2 = \{A, B, C\} \cup \{C, D\} = \{A, B, C, D\}$;

The universal relation can be obtained through the natural combination of all relations that are included in the diagram. The result of this combination for the example is presented below.

Table 1. *Universal relation*

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

Using Tane method for the obtained universal relation, the following FDs are discovered: $\overline{F_\Sigma} = \{AC \to B, C \to D, A \to D, B \to D\}$. This set contains all minimal FDs that are included in F_Σ as well as additional earlier unknown FDs $F_\Sigma^{'} = \{A \to D, B \to D\}$. Consequently, a set of *FD* of the universal relation for the logic diagram Σ can be expressed as follows: $\overline{F_\Sigma} = F_\Sigma \cup F_\Sigma^{'}$, where F_Σ is a set of FDs that meet the initial relations Σ, and $F_\Sigma^{'}$ is a set of additional FDs.

It is necessary to determine whether FDs from a set of $F_\Sigma^{'}$ can be derived from F_Σ or they are new information. To do this, it is suggested that the method of checking if FD corresponds to the closure $(F_\Sigma)^+$ should be used; this method was offered by D. Meyer in [9] – solving the problem of membership. The principle is as follows: as F^+ building is connected with sorting out all the subsets of attributes that belong to F and has an exponential complexity, it is suggested that F should be built – the closure on a set of attributes. F - closure of a set of X is such a set of attributes X^+ where $X \to X^+ \in F^+$ and there are no attributes in R that would depend on X but would not belong to X^+ [10]. The implementation of the method of building F - closure has a linear complexity. Thus, the method of checking if FD $X \to Y$ belongs to the closure F^+ consists of building F - closure of X^+ and determining if the expression $Y \subseteq X^+$ is true. If the expression is true, $X \to Y \in F^+$.

Let us consider the example: to check if $A \to D$ belong to F_Σ^+, A^+ should be built. According to [8], $A^+ = \{A\}$, as F_Σ does not contain FDs where A is the only attribute on the left side of FD. $D \notin A^+$, therefore, $A \to D \notin F_\Sigma^+$. It is shown that $B \to D \notin F_\Sigma^+$ in a similar way. Thus, a set of discovered dependences $F_\Sigma^{'}$ is non-derivable and therefore if new information [11,12].

It should be noted that this approach does not ensure that new dependences fully correspond to the target area. As it is based on a set of data which are included in RDB at the moment of processing and does not take into consideration their semantics, there is a major risk that random FDs can be obtained. A random FD is a FD which is not correct for a particular subject area (for example, the date of a person's birth determines the date of the person's child) and can be removed at any moment by changing or adding the tuples of dates that contradict the discovered dependence in the process of RDB operation [13].

Thus, there is another task – to check if new dependences are correct for the subject area which is modelled by the considered RDB. This work does not consider the solution of this task, this is the area of further studies. An expert assessment can be used as a way for solving this task. Or a numerical criterion for each single FD can be used, which enables establishing a threshold value, and dependences that are below this value are considered as random ones.

Conclusions

This paper suggests the approach to identify previously unknown functional dependencies, which is based on the analysis of a set of relational database data. The first step is to obtain a set of FD for each relationship. The similar operation is performed for the universal relation of the target RDB at the second step. At this step, the FD among the attributes of different relationships – the relationship among the data established in the process of the RDB operation can be discovered. The method for determining their information novelty is suggested; this method consists in checking if the FZ of the universal relation participates in closing the union of FD sets of individual relations.

The area for further studies is the development of methods and means for checking if the obtained dependencies are correct for the subject area, which is modelled by the considered RDB.

References

1. Rossiter, N. (2011), Re-engineering relational databases: the way forward: ISWSA '11, ACM New York, NY, USA, 17 p.

2. Konstantinov, S. M., Ponomarenko, Yu. L., Filatov, V. O. (2016), "Chastkovo vidobrazhennya modeley Danykh pry intehratsiyi informatsiynykh system", *Ekonomiko-matematychne modelyuvannya sotsialno-ekonomichnykh system. Zb. nauk. prats*, Kyiv, P. 140–158.

3. Kosenko, V. (2017), "Principles and structure of the methodology of risk-adaptive management of parameters of information and telecommunication networks of critical application systems", *Innovative Technologies and Scientific Solutions for Industries*, Kharkiv, No. 1 (1), P. 46–52. DOI: https://doi.org/10.30837/2522-9818.2017.1.046.

4. Filatov, V. A., Chaplanova, E. B. (2012), "Development of Information Technology of Object-relational Databases Design" ["Rozrobka informatsiynoyi tekhnolohiyi proektuvannya ob'yektno-relyatsiynykh baz danykh"], *European Researcher*, Vol. (36), No. 12, P. 2095–2101.

5. Filatov, V., Voloshchuk, O., Spivak, N. (2016), "Implementation and support fuzzy systems by means the relational data model" ["Realizatsiya ta pidtrymka nechitkykh system zasobamy relyatsiynoyi modeli danykh"], *Współpraca Europejska, European Cooperation*, Vol. 4, No. 11, P. 49–61.

6. Huhtala, Ykä (1999), "Tane: An Efficient Algorithm For Discovering Functional and Approximate Dependencies", *The Computer Journ*al, No. 42 (2), P. 100–111.

7. Filatov, V. A., Chaplanova, E. B., Spivak, N. O. (2014), "Komponenta obmezhen tsilisnosti yak element ob'yektno-relyatsiynoyi modeli danykh", *Informatsiyno-keruyuchi systemy na zaliznichnomu transporti*, No. 6 (109), P. 30–34.

8. Radchenko, V. A., Tanyanskyy, S. S. (2012), "Vyyavlennya prykhovanykh zalezhnostey mizh danymy v zadachakh reinzhynirynhu informatsiynykh system", *Information Processing Systems*, Vol. 3 (101), 268 p.

9. Meyer, D. (1987), *Teoriya relyatsiynykh baz danykh*: trans. for English, Moscow : Svit, 609 p.

10. Rudenko, D. A., Filatov, V. A. (2013), "Formalnyy pidkhid do opysu vlastyvostey danykh v informatsiynykh systemakh", Visnyk *Khersonskoho natsionalnoho tekhnichnoho universytetu*, No. 1 (46), P. 146–149.

11. Filatov, V. (2014), "Fuzzy models presentation and realization by means of relational systems", *Econtechmod: an international quarterly journal on economics in technology, new technologies and modelling processes*, Lublin, Rzeszow, Vol. 3, No. 3, P. 99–102.

12. Filatov, V., Radchenko, V. (2015), "Reengineering relational database on analysis functional dependent attribute", *Proceedings of the X Intern. Scient. and Techn. Conf. "Computer Science & Information Technologies"* (CSIT'2015), 14-17 sept. 2015, Lviv, Ukraine, P. 85–88.

13. Radchenko, V. A. (2011), "Modyfikatsiya metodu vyyavlennya funktsionalʹnykh zalezhnostey v relyatsiynykh bazakh danykh", *Informatsiyni tekhnolohiyi v navihatsiyi y upravlinni: stan ta perspektyvy rozvytku. Materialy Druhoyi mizhnarodnoyi naukovo-tekhnichnoyi konferentsyy*, Kyiv : DP «TSNDI NiU», 52 p.

Receive 15.02.2018

Відомості про авторів / Сведения об авторах / About the Authors

Філатов Валентин Олександрович – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, завідувач кафедри штучного інтелекту, м Харків, Україна; e-mail: valentin.filatov@nure.ua; ORCID: 0000-0002-5653-6301.

Филатов Валентин Александрович – доктор технических наук, профессор, Харьковский национальный университет радиоэлектроники, заведующий кафедрой искусственного интеллекта, г. Харьков, Украина; e-mail: valentin.filatov@nure.ua; ORCID: 0000-0002-5653-6301.

Filatov Valentin – Doctor of Sciences (Engineering), Professor, Kharkiv National University of Radio Electronics, Head of the Department of Artificial Intelligence, Kharkiv, Ukraine; e-mail: valentin.filatov@nure.ua; ORCID: 0000-0002-5653-6301.

Доскаленко Станіслав Миколайович –Харківський національний університет радіоелектроніки, аспірант кафедри штучного інтелекту, м Харків, Україна; e-mail: doskalenko.s@gmail.com; ORCID: 0000-0003-4625-210X.

Доскаленко Станислав Николаевич –Харьковский национальный университет радиоэлектроники, аспирант кафедры искусственного интеллекта, г. Харьков, Украина; e-mail: doskalenko.s@gmail.com ORCID: 0000-0003-4625-210X.

Doskalenko Stanislav – Kharkiv National University of Radio Electronics, post graduate student at the Department of Artificial Intelligence, Kharkiv, Ukraine; e-mail: doskalenko.s@gmail.com; ORCID: 0000-0003-4625-210X.

ПРО ОДИН ПІДХІД ДО ПОШУКУ ФУНКЦІОНАЛЬНИХ ЗАЛЕЖНОСТЕЙ ДАНИХ У РЕЛЯЦІЙНИХ СИСТЕМАХ

Предметом дослідження є інформаційні системи, побудовані на основі реляційних баз даних. **Метою** статті є розробити метод для реінжинірингу реляційних баз даних, що враховує неявні взаємопов'язані функціонально залежні данні, які впливають на структуру логічної моделі. Отримані такі **результати**: в статті запропоновано підхід до виявлення раніше невідомих функціональних залежностей, який ґрунтується на аналізі безлічі даних реляційної бази даних. Виділено класи завдань реінжинірингу реляційних баз даних; досліджений етап формування цільової логічної схеми, яка є спільною для задач адаптації та рефакторінга. Розглянуто підзавдання перевірки відповідності логічної схеми реляційної бази даних третій нормальній формі в межах даного етапу за допомогою методу синтезу; показано, що її рішення пов’язане з низкою труднощів, зокрема, необхідністю знаходження безлічі функціональних залежностей, що виконуються на поточному екземплярі даних деякої реляційної бази даних. Запропоновано підхід для знаходження безлічі функціональних залежностей з примірника даних реляційної структури. Напрямком для подальших досліджень може стати реалізація підтримки порожніх значень на етапі виявлення функціональних залежностей, а також питання перенесення даних без втрат з вихідної структури бази даних в цільову, отриману в результаті застосування методів реінжинірингу. **Висновки.** В роботі запропоновано підхід до виявлення раніше невідомих функціональних залежностей, який ґрунтується на аналізі безлічі даних реляційної бази даних. Першим кроком є отримання безлічі функціональних залежностей для кожного відношення. На другому кроці проводиться аналогічна операція для універсального відношення даної бази даних. На цьому кроці стає можливим виявлення функціональні залежності між атрибутами різних відносин – взаємозв’язку між даними, які встановилися в процесі функціонування інформаційної системи. Запропоновано спосіб визначення їх інформаційної новизни, який полягає у перевірці членства функціональних залежностей універсального відношення в замиканні об’єднання множин залежностей окремих відносин. Для подальших досліджень перспективним напрямком є розробка методів для реалізації технології перевірки отриманих залежностей на предмет відповідності логічної моделі предметної області.

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

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

ОБ ОДНОМ ПОДХОДЕ К ПОИСКУ ФУНКЦИОНАЛЬНЫХ ЗАВИСИМОСТЕЙ ДАННЫХ В РЕЛЯЦИОННЫХ СИСТЕМАХ

Предметом исследования являются информационные системы, построенные на основе реляционных баз данных. **Целью** статьи является разработать метод для реинжиниринга реляционных баз данных, учитывающий наличие неявных взаимосвязанных функционально зависимых данных, которые влияют на структуру логической модели. Получены следующие **результаты:** в статье предложен подход к выявлению ранее неизвестных функциональных зависимостей, который основывается на анализе множества данных реляционной базы данных. Выделены классы задач реинжиниринга реляционных баз данных; исследован этап формирования целевой логической схемы, которая является общей для задач адаптации и рефакторинга. Рассмотрена подзадача проверки соответствия логической схемы реляционной базы данных третьей нормальной форме в рамках данного этапа с помощью метода синтеза; показано, что ее решение сопряжено с рядом трудностей, в частности, необходимостью нахождения множества функциональных зависимостей, выполняющихся на текущем экземпляре данных некоторой реляционной базы данных. Предложен подход для нахождения множества функциональных зависимостей из экземпляра данных реляционной структуры. В качестве направления для дальнейших исследований можно выделить реализацию поддержки пустых значений на этапе выявления функциональных зависимостей, а также вопросы переноса данных без потерь из исходной структуры базы данных в целевую, полученную в результате применения методов реинжиниринга. **Выводы.** В работе предложен подход к выявлению ранее неизвестных функциональных зависимостей, который основывается на анализе множества данных реляционной базы данных. Первым шагом является получение множества функциональных зависимостей для каждого отношения. На втором шаге проводится аналогичная операция для универсального отношения рассматриваемой базы данных. На этом шаге становится возможным выявить функциональные зависимости между атрибутами различных отношений – взаимосвязи между данными, которые установились в процессе функционирования информационной системы. Предложен способ определения их информационной новизны, который состоит в проверке членства функциональных зависимостей универсального отношения в замыкании объединения множеств зависимостей отдельных отношений. Для дальнейших исследований перспективным направлением является разработка методов для реализации технологии проверки полученных зависимостей на предмет соответствия логической модели предметной области.

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