Интеллектуальные системы

DOI 10.24412/2221-2574-2023-2-40-45 УДК 004.8

ПРИМЕНЕНИЕ МЕТОДОВ DATA MINING ДЛЯ АНАЛИЗА И ВЫЯВЛЕНИЯ ЗАКОНОМЕРНОСТЕЙ В РЕЛЯЦИОННЫХ БАЗАХ ДАННЫХ

Баранчиков Алексей Иванович
доктор технических наук, доцент, профессор кафедры электронных вычислительных машин, ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина».
E-mail: alexib@inbox.ru

Федосова Елена Борисовна
аспирант кафедры электронных вычислительных машин, ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина».
E-mail: lena.fedosova2019@mail.ru

Адрес: 390005, Российская Федерация, г. Рязань, ул. Гагарина, д. 59/1.

Аннотация: В данной статье приводится решение одной из задач извлечения информации о структуре баз данных — выявление общих для нескольких реляционных баз данных атрибутов. Демонстрируются разработанные авторами алгоритмы нахождения общих для нескольких реляционных баз данных атрибутов с помощью применения методов Data Mining. Применение разработанных алгоритмов позволит на основании имеющихся нескольких реляционных БД оптимальным способом синтезировать новую схему единой БД, адекватно отражающую предметную область предприятия

Ключевые слова: Data Mining, классификация текстовых данных, реинжиниринг баз данных, реляционная база данных, структура базы данных.

Введение

Современные предприятия и организации в своей работе всё чаще предпочитают использовать некую единую информационную среду, позволяющую обрабатывать очень большой объем разнородных данных. Но зачастую на крупных предприятиях уже существует набор больших баз данных (БД), имеющих концептуальные связи и содержащих общую информацию, но никак не связанных между собой структурно (технически). Такие БД содержат огромное количество данных, утеря которых может привести к нежелательным последствиям (принятие неэффективных решений, аварии, сбои и т.д.). В связи с этим часто возникает необходимость объединения уже имеющихся БД в единую БД с целью последующей разработки единой информационной среды предприятия, что обусловливает высокую актуальность данной работы.

Для разработки единой БД можно использовать подход, при котором общая БД разрабатывается «с нуля». Однако в силу многих причин (возможная утеря проверенной длительным использованием семантики имеющихся БД, увеличение материальных и временных затрат, снижение достоверности данных и т.д.) такой подход считается недостаточно эффективным [1]. По этой причине на практике часто встречается задача интеграции нескольких реляционных БД в одну.

На данный момент существуют различные методы и алгоритмы, позволяющие получить информацию о структурах БД: алгоритмы определения структурных закономерностей в структурах данных (алгоритм поиска функциональных зависимостей Find_FD, алгоритм поиска многозначных зависимостей Find_MD, алгоритм поиска зависимостей соединения Find_JD) [2], алгоритм определения семантических зависимостей в информационных

структурах данных [3], алгоритм сравнительного анализа схем реляционных баз данных на основе изучения семантики предметной области [4], алгоритм коррекции схемы реляционной базы данных [5], алгоритм Tane [6], и т.д.

Однако эти алгоритмы разработаны, прежде всего, для модификации одной БД в соответствии с требованиями и семантикой, выявленными в ходе эксплуатации, и не предлагают решения задачи интеграции нескольких БД в одну. Кроме того, эти алгоритмы требуют глубокой экспертной оценки и не всегда приносят достаточно высокие результаты в части достоверности данных.

С целью устранения этих пробелов для объединения нескольких БД в одну в данной работе предлагается использовать методы Data Mining. Как известно Data Mining — исследование и обнаружение "машиной" (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком [7].

1. Теоретическая часть

Одним из этапов интеграции нескольких реляционных БД в одну является определение одинаковых для имеющихся БД атрибутов. Поэтому в данной работе была поставлена задача определить с помощью Data Mining, какие атрибуты уже существующих реляционных баз данных являются общими. Подобная информация о структурах баз данных позволит исключить избыточность данных и повысить эффективность их хранения, что необходимо для дальнейшего синтеза новой схемы БД, адекватно отражающей выбранную предметную область.

Задача определения общих атрибутов может быть рассмотрена как одна из задач Data Mining — классификация. Классификация — это определение категории (класса) объекта по набору его признаков [8, 9]. В нашем случае, объектами являются активные домены атрибутов реляционных баз данных.

Пусть $D = \{d_1,...,d_i,...,d_m\}$ — множество реляционных баз данных d_i. Каждая из d_i состоит из совокупности отношений $R_i = \{r_i^1,...,r_i^k,...,r_i^s\}$. Схемой отношения $r_i^k \in R_i$ является конечное множество атрибутов $A_i^k = \{a_{i,k}^1,...,a_{i,k}^j,...,a_{i,k}^l\}$, причем каждому $a_{i,k}^j \in A_i^k$ поставлено в соответствие непустое конечное множество $Dom_{i,k}^j$, которое является доменом атрибута $a_{i,k}^j$.

Для выделения общих для нескольких реляционных БД атрибутов необходимо выполнить следующие действия (рис. 1):

1) выбрать одну БД $d_i \in D$ в качестве основной;

2) разделить выбранную основную БД d_i на классы так, чтобы каждый атрибут, принадлежащий d_i, представлял собой отдельный класс (за исключением внешних ключей);

3) для каждого объекта обучающей выборки определить характеристики, на основании которых можно идентифицировать принадлежность объектов обучающей выборки соответствующим классам;

4) классифицировать каждый из атрибутов оставшихся баз данных, т.е. для каждого атрибута определить значения характеристик, на основании которых можно идентифицировать принадлежность атрибута одному из классов [10].

Для классификации можно использовать любой алгоритм классификации, например, метод опорных векторов (SVM), наивный байесовский классификатор (Naive Bayes), логистическая регрессия (Logistic Regression), алгоритм *k* ближайших соседей (*k*NN) и т.д.

Перечисленные действия реализуются алгоритмами Find_Training_Set и Find_Joint_Attributes, разработанными авторами статьи.

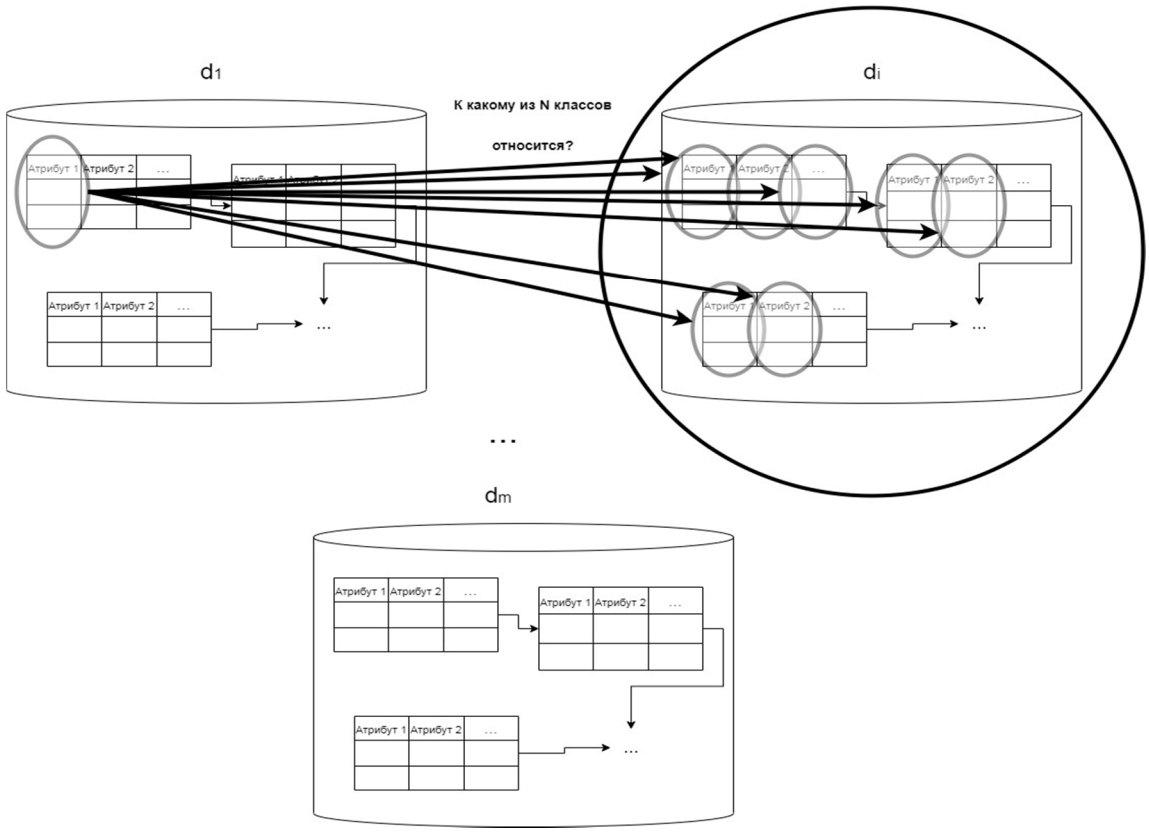

Рис. 1. Определение общих для нескольких БД атрибутов путём применения алгоритма классификации

2. Алгоритм нахождения обучающей выборки

Алгоритм Find_Training_Set:

Вход: множество реляционных баз данных $D = \{d_1,...,d_i,...,d_m\}$.

Выход: обучающая выборка $TS = \{C_1,..., C_p,..., C_n\}$.

begin

Выбрать $d_i \in D$ в качестве основной базы данных;

Подготовить (при необходимости) анализируемые данные к обработке (лемматизация, стемминг и т.д.);

$TS := \varnothing$;

for $k := 1$ to s do

begin

for $j := 1$ to l do

if $a_{i,k}^j \neq внешний\_ключ$ then

begin

$C := \{a_{i,k}^j, Dom_{i,k}^j\}$;

$TS.append\ (C)$;

$j := j + 1$;

end;

else: $j := j + 1$;

$k := k + 1$;

end;

Получили множество $TS = \{C_1,...,C_p,...,C_n\}$, где $C_p = \{a_p, Dom_p\}$

for $p := 1$ to n do

begin

Определить значения элементов множества характеристик $H = \{H_1,...,H_t\}$, на основании которых Dom_p относится к классу a_p;

$C_p.append(H)$;

$p := p + 1$;

end;

$D = \{d_1,...,d_{i-1},d_{i+1},...,d_m\}$;

end.

Конец алгоритма.

В результате выполнения алгоритма Find_Training_Set получаем обучающую выборку $TS = \{C_1,...,C_p,...,C_n\}$, причем C_p имеет вид $\{a_p, Dom_p, H_p\}$, где a_p — класс (имя атрибута); Dom_p — активный домен, соответствующий имени атрибута a_p; H_p — множество характеристик $H = \{H_1,...,H_t\}$, на основании которых можно идентифицировать принадлежность домена Dom_p классу a_p.

3. Алгоритм определения общих для нескольких БД атрибутов

Алгоритм Find_Joint_Attributes:

Вход: множество реляционных баз данных $D = \{d_1,...,d_i,...,d_m\}$.

Выход: множество атрибутов и соответствующих им активных доменов с указанием класса, к которому относится данный атрибут.

begin

FindTrainingSet;

Подготовить (при необходимости) анализируемые данные к обработке (лемматизация, стемминг и т.д.);

for $i := 1$ to $m-1$ do

for $k := 1$ to s do

for $j := 1$ to l do

begin

Определить множество характеристик H для $Dom_{i,k}^{j}$;

Классифицировать каждый атрибут с помощью алгоритма классификации *k*NN;

end;

end.

Конец алгоритма.

В описанном выше алгоритме Find_Joint_Attributes для классификации атрибутов был выбран алгоритм классификации *k*NN (*k Nearest Neighbor*, алгоритм *k* ближайших соседей) [11, 12]. Выбор этого алгоритма для классификации элементов (атрибутов) обусловлен важными для использования и программной реализации алгоритма Find_Joint_Attributes преимуществами *k*NN-алгоритма: простота модификации алгоритма в соответствии с выбранной метрикой, простота программной реализации, возможность обновления обучающей выборки без переобучения классификатора, устойчивость алгоритма к выбросам в исходной выборке данных.

В результате выполнения алгоритма Find_Joint_Attributes получаем набор атрибутов с соответствующими им активными доменами с указанием класса, к которому они относятся, из обучающей выборки TS.

Заключение

В статье представлены разработанные авторами алгоритмы, определяющие общие для нескольких реляционных БД атрибуты. Основной идеей, лежащей в основе алгоритмов, является использование методов Data Mining для решения задачи классификации атрибутов. Применение разработанных алгоритмов позволит на основании имеющихся нескольких реляционных БД оптимальным способом синтезировать новую схему единой БД, адекватно отражающую предметную область предприятия.

Литература

1. *Хомоненко А. Д., Цыганков В. М., Мальцев М. Г.* Базы данных: Учебник для высших учебных заведений / Под ред. проф. А.Д. Хомоненко. 6-е изд., доп. СПб.: КОРОНА-Век, 2009. 736 с.

2. *Нгуен Н.З.* Методики и алгоритмы извлечения знаний из реляционных баз данных на основе семантики предметной области: дис. … канд. техн. наук: 05.13.17. Рязань, 2020. 177 с.

3. *Баранчиков А. И., Нгуен Н. З.* Разработка алгоритма определения семантических зависимостей в информационных структурах данных // Наука, образование, инновации: актуальные вопросы и современные аспекты: сб. статей междунар. науч.-практ. конф. Пенза: МЦНС «Наука и просвещение». 2020. С. 33–36.

4. *Баранчиков А. И., Нгуен Н. З.* Алгоритм сравнения схем реляционных баз данных на основе ана-

лиза семантики предметной области // Вестник РГРТУ. 2019. № 67. - С. 45-49.

5. *Баранчиков А.И., Нгуен Н.З.* Алгоритм коррекции схемы реляционной базы данных // Вестник РГРТУ. 2019. № 69. С. 93–101.

6. *Huhtala Y., Karkkainen J., Porkka P., Toivonen H.* Tane: An Efficient algorithm for discovering functional and approximate dependencies // The Computer Journal. 1999. №2. Vol.42. Pp. 100–111.

7. Анализ данных и процессов: учеб. пособие / *А.А. Барсегян, М.С. Куприянов, И.И. Холод, М. Д. Тесс, С. И. Елизаров*. 3-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с.

8. *Witten I.H., Frank E., Hall M.A.* Data mining: practical machine learning tools and techniques. 3-rd ed. Elsevier, 2011. 629 p.

9. *MacLennan J., Tang Z., Crivat B*. Data Mining with Microsoft SQL Server 2008. Wiley Publishing, Inc, 2009. 636 p.

10. *Баранчиков А.И., Федосова Е.Б.* Применение методов Data Mining для реинжиниринга баз данных // Методы и средства обработки и хранения информации. 2022. С. 129–133.

11. *Батура Т.В*. Методы автоматической классификации текстов // Программные продукты и системы. 2017. Т. 30. № 1. С. 85–99.

12. *Соколова Ю.С. Демидова Л.А*. Классификация данных на основе SVM-алгоритма и алгоритма k-ближайших соседей // Вестник РГРТУ. 2017. № 62. С. 119–132.

Поступила 6 марта 2023 г.

English

APPLYING DATA MINING METHODS TO ANALYZE AND IDENTIFY PATTERNS IN RELATIONAL DATABASES

Aleksej Ivanovich Baranchikov — Grand Dr. in Engineering, Professor, Department of Electronic Computers, Ryazan State Radio Engineering University named after V.F. Utkin.
E-mail: alexib@inbox.ru

Elena Borisovna Fedosova — Postgraduate student, Assistant, Department of Electronic Computers, Ryazan State Radio Engineering University named after V.F. Utkin.
E-mail: lena.fedosova2019@mail.ru

Address: 390005, Russian Federation, Ryazan, Gagarina str, 59/1.

Abstract: Modern enterprises and organizations in their work increasingly prefer to use shared data environment that enables them to process a very large amount of heterogeneous data. However, often large enterprises already have a set of large databases (DB) with conceptual links and comprise general information but not structurally (technically) related to each other. Such DB have a huge data amount, the loss of which can cause unwanted effect (inefficient decisions, accidents, failures, etc.). Hence, there is often need to aggregate existing databases into a single database aimed at further development of enterprise unified information environment and that makes this research work highly relevant. The article deals with the solution of one of the problems of information retrieval about databases structure, in particular, identification of attributes common to several relational databases. Such information about DB structures will eliminate data redundancy and raise their storage efficiency, which will be required for further synthesis of a new DB schema. Search algorithms developed by the authors are shown for attributes common to several relational databases using Data Mining methods. The task of finding common features is viewed as classification (classification is the definition of a category (class) of an object through the set of its features). Objects are active domains of the relational database in our case. One of the developed algorithms is meant to find a learning sample, the other one is to classify objects (attributes) using kNN algorithm.

Keywords: Data Mining, text data classification, database reengineering, relational database, database structure.

References

1. *Homonenko A.D., Cygankov V.M., Mal'cev M.G.* Databases: Textbook for higher educational institutions. Ed. by *A.D. Homonenko.* Saint Petersburg: KORONA-Vek, 2009. 736 p.

2. *Nguen N.Z.* Methods and algorithms for extracting knowledge from relational databases based on the semantics of the subject area: dis. … cand. tech. Sciences: 05.13.17. Ryazan, 2020. 177 p.

3. *Baranchikov A.I., Nguen N.Z.* Development of an algorithm for determining semantic dependencies in information data structures. Science, education, innovations: topical issues and modern aspects: Iss. articles of the international scientific-practical. conf. Penza: ICNS "Science and Education". 2020. Pp. 33–36.

4. *Baranchikov A.I., Nguen N.Z.* Algorithm for comparing schemas of relational databases based on the analysis of the semantics of the subject area. Vestnik of RSREU. 2019. No. 67. Pp. 45–49.

5. *Baranchikov A.I., Nguen N. Z.* Relational database schema correction algorithm. Vestnik of RSREU. 2019. No. 69. Pp. 93–101.

6. *Huhtala Y., Karkkainen J., Porkka P., Toivonen H.* Tane: An Efficient algorithm for discovering functional and approximate dependencies. The Computer Journal. 1999. No.2. Vol. 42. Pp. 100–111.

7. Analysis of data and processes / *A.A. Barsegjan, M.S. Kuprijanov, I.I. Kholod, M.D. Tess, S.I. Elizarov*. 3-rd ed. Saint Petersburg: BHV-Peterburg, 2009. 512 p.

8. *Witten I.H., Frank E., Hall M.A.* Data mining: practical machine learning tools and techniques. 3-rd ed. Elsevier, 2011. 629 p.

9. *MacLennan J., Tang Z., Crivat B*. Data Mining with Microsoft SQL Server 2008. Wiley Publishing, Inc, 2009. 636 p.

10. *Baranchikov A.I., Fedosova E.B.* Data Mining application for database reengineering. Methods and means of processing and storing information. 2022. Pp. 129–133.

11. *Batura T.V.* Methods for automatic text classification. Software products and systems. 2017. Vol. 30. No. 1. Pp. 85–99.

12. *Sokolova Ju.S. Demidova L.A*. Two-stage data classification method based on SVM-algorithm and the k nearest neighbors algorithm. Vestnik of RSREU. 2017. No. 62. Pp. 119–132.