Acta Technica 62 (2017), No. 4A, 705–712 © 2017 Institute of Thermomechanics CAS, v.v.i.

Research on relational database fusion method for data mining

Xiangqin Li[1], Chuanjun Luo[2]

Abstract. With the development of data mining technology, more and more problems can be solved by data mining technology. In the development of data mining, the scale of data is also growing, so we need to use the appropriate way to store and optimize the data. The relational database is the most commonly used database, however, how to integrate a large number of relational databases in data mining is an urgent problem to be solved. In this paper, we first analyze the existing relational database integration scheme. Then, based on this scheme, an Aprior fusion algorithm for data mining and its optimization algorithm are proposed. Finally, we give an example of data mining to prove the validity of the data fusion scheme proposed in this paper. It can greatly improve the speed of data reading in data mining and finally improve the performance of data mining system. The experiments demonstrate the effectiveness of the proposed relational database fusion method for data mining.

Key words. Relationship database, fusion method, data mining.

1. Introduction

In the real world, a large amount of data are present in government and enterprises: 1) A large amount of data are unstructured, except that some special or specialized data is stored in a structured database (eg relational database) [1]. The data is stored in various forms, such as Word, pictures, video files and so on. However, these data are not uniform standards; 2) These binary images or video files are huge, often large to several tens of megabytes, which is structured with a structured database; 3) These data is decentralized, and a large number of data are not centralized storage, while data are scattered in different departments and personal computers [2]; 4) These data also involve security, sharing, update and other issues.

Relational database develops in the early 20th century, puts forward in the early 70s. After decades of database experts' efforts, theory and practice have achieved remarkable results, which marks the increasingly mature database technology [3].

[1] Workshop 1 - School of Computer Engineering, Jingchu University of Technology, Jingmen, China

[2] Workshop 2 - E-government Information Center, Jingmen city, Hubei province,China

But it is still difficult to achieve the analysis of the data in the relational database, which cannot support the decision very well. Thus, in the 1980s, the idea of data warehouse and the basic principles of data warehouse have been identified with architecture and use principles. The main technologies include data access in the database, network, C / S structure and graphical interface, and some large companies have begun to build data warehouse. For the rapid growth of data in the rapid accumulation of data collection, storage, the human has been unable to solve these problems, while the data warehouse can achieve the useful knowledge of the need for data mining [4]. Data mining and statistical sub-domain "tentative data analysis" and artificial intelligence sub-domain "knowledge discovery" and machine science, are comprehensive technical disciplines. It is necessary and important to understand the differences and relationships between relational databases, data warehouses, and data mining to make better use of these three technologies [5]. As shown in figure 1, the classical data mining contains seven steps.

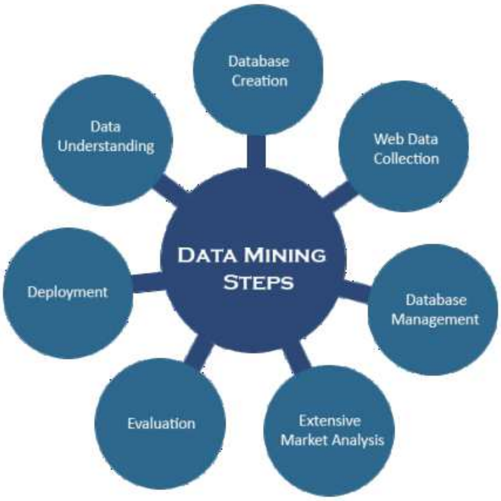

Fig. 1. The classical data mining contains seven steps

There are several differences between relational database and the data warehouse. Relational database is a transaction-oriented design, while the data warehouse is a thematic design. Relational database stores online transaction data, while data warehouse usually stores historical data [6]. In addition, the relational database design will try to avoid redundancy, but the data warehouse is inclined to introduce redundancy. Relational databases are designed to capture data, and data warehouses are designed to analyze data. The traditional relational database is oriented to transaction-oriented system applications, so it cannot meet the requirements of decision support system analysis. Transaction processing and analysis processing have very different properties, and they have different demand data [7]. Data warehouse is the early step of data mining. Through the construction of data warehouse, the efficiency and ability of data mining are improved, which ensures the breadth and completeness of data in data mining [8].

Data mining data sources are not necessarily data warehouses, which can also be a relational database of data. However, in advance to pre-data processing, it can be used for data mining. Data pre-processing is a key step in data mining and is the main part of the data mining process. Therefore, the data warehouse and data mining is not necessarily linked, some people simply believe that the data warehouse is the preparation of data mining, which is not comprehensive. You can also use the relational database data as data mining data source.

2. Relational Database Fusion Algorithm for Data Mining

2.1. Relational database convergence scheme

Specific implementation should complete the following tasks:

1) A reasonable metadata structure should be defined for the relational database;

2) For metadata and data sets, it should be a complete increase, delete, update and query functions;

3) The new metadata or data set should also set the user and user group permissions distribution, so that both them can achieve the full sharing of multiple information, but also to ensure that these raw data security and integrity;

4) They should be implemented to define the view and other functions.

Relational database table structure definition is generally two cases: 1) Pure text type, that is, the record does not contain the relevant pictures and text[9]; 2) Records may be with one or more pictures, documents, video information and so on. The database designer may use a field to store the path of these pictures, documents, or video data, or it may be possible to save the path information in a single table. Taking into account the above two cases, for the IIMS platform, the authors propose the metadata definition method, in which the structure of the metadata and relational database table structure is corresponding to each metadata project data type and table structure in the definition of consistent. As shown in figure 2, the framework of relational database integration is given in details.

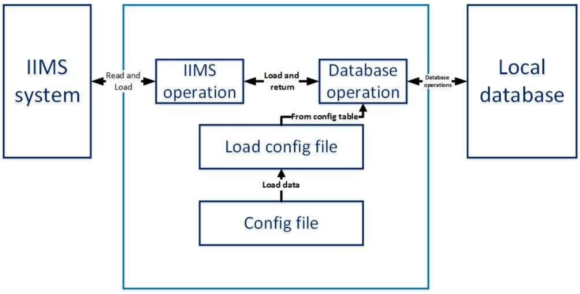

Fig. 2. The framework of relational database integration

According to the above method, we can use the secondary development interface provided by IIMS to develop the implementation framework of the relational database fusion scheme as shown in Figure 2.

2.2. Aprior algorithm and its optimization algorithm

Faced with increasingly fierce market competition, customers are increasingly demanding the ability to respond quickly to various business problems, and increasing the demand for timely processing of excess data. The challenge is that large-scale, complex data systems allow users to feel endless; on the other hand, these large amounts of data behind a lot of meaningful and valuable decision-making information. Such as the computer industry are familiar with the "beer and diapers" story, which is the retail giant "Wal-Mart" from a large number of sales data in the analysis of the law. The United States men go to the supermarket to buy baby diapers, and they will buy beer. "Wal-Mart" put these two "irrelevant" goods placed in the shelves, and also placed some under the side dishes, so that the sales of these goods are increased. So the application of data mining from a large number of data found with specific guidance in the law.

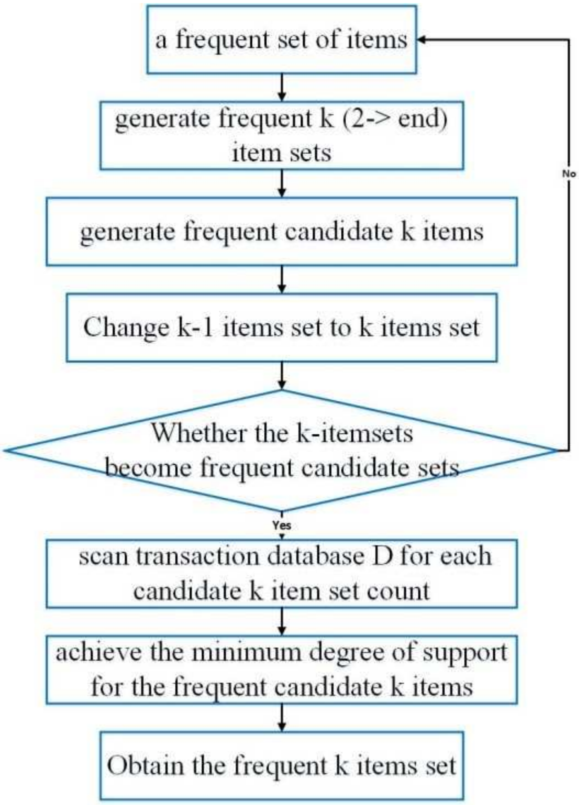
a frequent set of items
generate frequent k (2-> end) item sets
generate frequent candidate k items
Change k-1 items set to k items set
Whether the k-itemsets become frequent candidate sets
No
Yes
scan transaction database D for each candidate k item set count
achieve the minimum degree of support for the frequent candidate k items
Obtain the frequent k items set

Fig. 3. The relational dataset fusion algorithm- Aprior algorithm

The association rule is the implication of A B, and rule A B is set up in transaction set D with support degree s, where s is the transaction containing A B (ie, both A and B) percentage, ie probability P (A B). The rule A B has a confidence degree c in the transaction set D, and if the transaction containing A in D also contains the percentage of B, that is, the conditional probability P (B | A), if the set of data items satisfy the minimum support min (s), called the frequent itemsets. We can simplify the mining step into two steps: 1) find all the frequent itemsets, one of the most influential Boolean association rules frequent itemsets algorithm - -Aprior algorithm; 2) generated by the frequent itemsets strong association rules. The relational dataset fusion algorithm- Aprior algorithm is given in details, as shown in figure 3.

Aprior algorithm flow:

1. A frequent set of items;
2. Generate frequent k (2-> end) itemsets;
3. Generate frequent candidate k items;
4. Choose from the frequent k-1 items set as the k items set;
5. Detect whether all k-1 subsets of the k-itemsets are frequent itemsets, and if the k-itemsets become frequent candidate sets;
6. Scan the transaction database D for each candidate k item set;
7. The minimum degree of support for the frequent candidate k items are chosen to become frequent k items set.

To improve the performance of the Aprior algorithm, we improve the Aprior algorithm as follows. (1) In the case of generating frequent candidate k-itemsets, the time cost of filtering non-frequent itemsets by detecting the k-1 subset of k-itemsets is a frequent itemsets, especially in k-1 frequent itemsets. In the case of a higher degree, the k-item set connected by the frequent k-1 itemsets is directly used as the frequent candidate set. In the case, the transaction database can always be stored in memory where the transaction database is not very large, and if the cost of the frequent candidate set is not very high, the k-1 subset of the k-item is not detected item set, directly in the transaction database count, which can better reflect the time efficiency of this method.

3. Experimental application of the relational database fusion method for data mining

3.1. Problem Description

In any university, they often have accumulated the graduates of the students and other characteristics of the college entrance examination results, as well as the course results of university. The relationship between the various factors and the comprehensive evaluation can be used to study the various professional courses and to guide the school enrolment and students to fill volunteer work [5]. For example, if you can find that the "college entrance examination math" "comprehensive performance" results are "excellent", and the frequency of this item sets is high, you can explain that the "college entrance examination" is to determine the professional students as the "comprehensive performance". The enrolment can be appropriate to improve the "math" threshold. Students fill in volunteer, and the same can also be used as a reference, choose their own suitable professional.

3.2. Transformed Relational database tables into a single-dimensional Boolean association rule

The main data involved in the issue is concentrated in a "Student Status Register". It is impossible to get the ideal result of satisfying support and confidence, so we do the corresponding data transformation, merge the similar and related attribute values into one attribute value to meet the support and confidence, from the "attribute - value" into "attribute - the combined value".

3.3. Performance simulation results

The algorithm runs on a computer with a memory of 6G and a CPU of Intel Core i7-6700U. The transaction database consists of 14 items (transactions) and 51 attributes (dimensions), which are randomly extracted. Since the process of generating L1 is the same, the difference between the two algorithms is that the two items produce the 2-> k items, and the time spent in generating the 2-> k items, which can be shown in Table 1.

Table 1. Performances of different algorithms

Support degree	**7**	**6**	**5**	**4**
Without Aprior algorithm	20s	2min12s	54min23s	10h35min12s
With Aprior algorithm	5s	22s	19min46s	4h28min59s
With improved Aprior algorithm	4s	14s	1min54s	52min34s

From the operation of the situation, it can be seen that the improved algorithm in the speed of operation has been significantly improved, especially against the mining object has a large number of frequent mode and long mode, when user gives the minimum support threshold low. The transaction database records increase to 148, and then they will find similar results.

In order to further analyze the performance of the proposed algorithm, we compare the running time of three different methods under different project set size.

From the above figure 3, we can see that as the number of items increases, the computational time of the three methods increases accordingly. This is because that the increase in the number of items will increase the storage complexity of the relational database whose data is taken at data mining, then the time cost of the operation will also increase accordingly. In addition, the two methods of using the relational database fusion algorithm Aprior proposed in this paper are far superior to those that do not use this method. And with the increase in the number of projects, the performance advantages of Aprior algorithm will be more obvious. This is because with the increase in the number of projects, Aprior algorithm will adopt the relational database integration, thus it can greatly reduce the data between the database operands. So you can reduce the data mining operation time. At the same

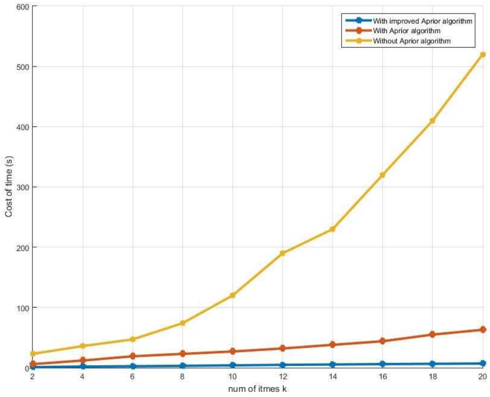

Fig. 4. The cost of time vs the number of items

time, the performance of the proposed Aprior algorithm is better than that of the Aprior algorithm. Because in the enhanced Aprior algorithm, we subtly design the relational database to further reduce the operand, so the same number of items of the improved Aprior algorithm has the best performance.

4. Conclusion

More and more problems can be solved by data mining technology with the development of data mining technology. In the development of data mining, the scale of data is also growing, so we need to use the appropriate way to store and optimize the data. The relational database is the most commonly used database, however, how to integrate a large number of relational databases in data mining is an urgent problem to be solved. In this paper, we first analyze the existing relational database integration scheme. Then, based on this scheme, an Aprior fusion algorithm for data mining and its optimization algorithm are proposed. Finally, we give an example of data mining to prove the validity of the data fusion scheme proposed in this paper. It can greatly improve the speed of data reading in data mining and finally improve the performance of data mining system. The experiments demonstrate the effectiveness of the proposed relational database fusion method for data mining.

References

[1] O. Maimon, A. Browarnik: *NHECD-Nano health and environmental commented database.* Data mining and knowledge discovery handbook. Springer US (2009) 1221–1241.

[2] X. Wu, X. Zhu, G. Q. Wu: *Data mining with big data.* IEEE transactions on knowledge and data engineering *26* (2014), No. 1, 97–107.

[3] A. H. Doan, A. Y. Halevy: *Semantic integration research in the database community: A brief survey.* AI magazine *26* (2005), No. 1, 83.

[4] S. H. Liao, P. H. Chu, P. Y. Hsiao: *Data mining techniques and applications–A decade review from 2000 to 2011.* Expert systems with applications *39* (2012), No. 12, 11303–11311.

[5] M. Saeed, C. Lieu, G. Raber: *MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring.* Computers in Cardiology (2002) 641–644.

[6] J. Zhang, W. Hsu, M. L. Lee: *Image mining: Issues, frameworks and techniques Proceedings of the Second International Conference on Multimedia Data Mining.* Springer-Verlag (2001) 13–20.

Received November 16, 2016