2011 International Conference on Computer and Automation Engineering (ICCAE 2011)
IPCSIT vol. 44 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V44.25

Research of Data Mining of Association Pattern Pairs in Multidimensional Structured Database

Zhang Jing, Huang Xiantong and Yang Xinfeng[+]

Computer Science and Technology Department,Nanyang Institute of Technology,Nanyang,473000, China

Abstract. User preference model first proposed using an intelligent method to set the minimum support. Secondly, we propose a new data mining problem, the structure of the database to find frequent patterns associated pair. To effectively address these problems, we developed a series of cutting ability with a strong algorithm. The new algorithm is also discussed in the one-dimensional and multi-dimensional structure of the database found on the applicability of the model and to assess the efficiency of the new algorithm.

Keywords*:* Pattern Pairs; Multidimensional Structured Database; Data Mining

1. Introduction

For Mining Association Rules Apriori algorithm for frequent item sets first proposed by Agrawal and Srikant. It is a very influential algorithm is based on the nature of "frequent itemset is also frequent non-empty set." A frequent itemset is depending on the user specified minimum support. Overall, the association mining to meet the minimum support to first find the frequent itemsets, and then those who meet the minimum confidence based on frequent item sets derived association rules. The problem is the minimum support and minimum confidence set no clear standards. Setting different minimum support will find a different number of frequent itemsets, which affects the rules to be excavated. Fayyad et al described the data mining in order to repeat the search process, it implies that the efficiency of mining process most appropriate setting depends on mining query, this set contains the minimum support and confidence setting.

When the association rule mining was first introduced, it only focused on the discovery product purchase analysis of the association between dimensions. Later, in [1-3], the proposed multi-dimensional data from a data warehouse cube or multidimensional association mining. Allows for multiple dimensions, and through the optional filter set of data granularity, the user can specify more precise objectives mining data. Therefore, multidimensional association mining query is automatic, the rules were excavated close to the user want. However, for inexperienced users, to construct an effective query is difficult, especially in the setting appropriate threshold values.

User preference model to maintain logs extracted from the excavation are frequently used queries. It describes a variety of users to mine the experience, thereby providing useful information for the users benefit users, especially beneficial to inexperienced users. With the model of user preferences specified in the query and the similarity between the user's query terms, identifying the most similar to the intensity of mining user queries. Queries from those calculated from the threshold of support, export the appropriate minimum support and provide to the user.

Recently, in many applications, structured data is becoming increasingly common. Structured data through these interrelated or integrated, we will have easy access to more informative and valuable but complex structured data. As such a typical example of complex databases, we are concerned about the structure of multi-dimensional data, ie, attribute value table with the attributes for structured data.

Multi-dimensional structure of the database from the situation in the mining, the discovery of knowledge between attributes is an important concern is also significant. For example, on multi-dimensional structure of

[+] Corresponding author. Tel.: +86-377-62076332
E-mail address: ywind2005@163.com

proteins in the database, if the amino acid and protein structures found some cause for concern among the links, then it may give us some of the amino acid sequence to maintain a useful mechanism for knowledge.

2. Model Based on User Preference Support Threshold Setting Method of Intelligent

From the association rule mining was first Agrawal et al [4] introduced the beginning, a number of scholars dedicated to research on it. In their proposal, the need to specify the user's minimum support to the pattern recognized as interest. Other studies on the measurement of interest was carried out. The main purpose is to provide a better test to be truly useful knowledge. In [5] and cultural, the present and discuss a variety of tests. Lenca et al to collect a lot of measurements on their property, the use of standard adjuvant setting for the specified multi-user needs a good measurement. To provide users with better testing, Zhang et al gives a polynomial strategy will specify the user's fuzzy support threshold drawn into the actual minimum support. Liu et al proposed a fuzzy matching technology, it will set a specific user mode (user feedback) into to help users identify those of interest. We provide users with appropriate support threshold value is present in the user preference model by the additional knowledge of mining it from the previous experience of the user to collect information.

A multi-dimensional association rules is an involving one or more dimensions (attributes) of the association rules. Mining multidimensional association mining query element model is defined as follows:

$$MP:\langle t_G, t_M, [wc], ms, mc\rangle$$

Where t_G, t_M, wc， ms and mc are an integral part of the inquiry, described as follows:

t_G : Transaction ID number set (data interval size);

t_M : Set of attributes of interest to mining;

wc : The optional "where" conditions;

ms : Minimum support;

mc : The minimum confidence.

The following is a multidimensional association mining query examples:

t_M： $Prodname, Age$

t_G： $CustID, Date$

wc： $Country = "Japan"$

ms： 65%

mc： 86%

Mining is the strength of the user that he / she wants to understand the different age levels of customer and merchandise daily to purchase the correlation between the set minimum support and minimum confidence 65% 86%. Target mining data sets shown in Table 1. Transaction Ids $t_G = \{CustID, Date\}$, interested in mining properties

$$t_M = \{Age, ProdName\}.$$

Table 1.The target mining data set

tid	tG		tM	
	CustID	Date	Age	*ProdName
1	C001	2009-02-01	20-30	B,E
2	C002	2009-02-03	40-50	A,B,C,E
3	C002	2009-02-10	40-50	A,C,D
4	C003	2009-02-05	30-40	B,C,E
5	C004	2009-02-09	40-50	B,E
6	C004	2009-02-15	40-50	A,E

*A: IBM 60GB B: IBM TP C: RAM 512MB D: Ink Cartridge E: Hard Disk

Multidimensional association mining query constructed by the user started. Upon receiving a user-defined queries submitted, the mining engine's function is triggered. When the user receives the data mining results, he or she subjectively determine whether these results are acceptable. If the answer is no, then will adjust the mining query, run the mining process again until the user get an acceptable result.

3. Multidimensional Association Mining nCP 4.5 Algorithm

Find related items because of the method has been in [6-7] was made, so our goal is to find association patterns right. Digging through the property and is related to some recently proposed by Boolean vectors and numerical models that [8]. On the other hand, also studied the association in the quantitative database, pattern discovery. Multidimensional database as a dig at one of the problems in [9-10] proposed a multi-dimensional sequential pattern mining. The purpose of this problem is to find all item sets and contains a sequence of frequent (closed) mode. In [10] proposed a "related graphics search" problem, it is about the relationship in the graph database mining. The structure and our algorithms are very different, because it considers relevant to a given query subgraph. Carried out in the graphics database is another problem associated with mining mining HSG [11]. In this problem, the use of all - trust as associated measurement [12], all the associated sub-Atlas will be found. However, it did not discuss the mode between the mining property.

Utilization - the correlation coefficient as the correlation test [13], we first consider the two-dimensional structure of the database on the problem of mining association patterns. To effectively solve this problem, developed an algorithm called nCP. By running on a tree data structure, and use - the correlation coefficient of the boundary and joint support of the lower boundary, nCP algorithm can avoid no future for consideration. Also studied two types of top-k mining problem, based on nCP algorithm developed several algorithms for these problems. In addition, we will explore the development to the more general case, that is, not only in the one-dimensional structure of the database mining association pair, also with two or more attributes of the multi-dimensional structure of the database mining association pairs.

4. Setting the Minimum Support Structure of the Intelligent

This section provides minimum support for the mechanism of the smart set. First, we introduce the mining system architecture, and second, to provide minimum support set for the details of the mechanism.

4.1. Smart Set Architecture

Multidimensional association mining data warehouse structures from the user's query. Mining engine is running the user query has been defined, then the rules get sent back to the user. Successful query is kept in the mining log. Described in the previous section by performing the purification process, from mining to generate user preference model log. During the user construct a query, set the minimum support interactive module is responsible to provide a suitable range of minimum support to the user. It calculates the specified user queries and user preferences model similarity between alternative query, to provide users with a reasonable degree of support range setting.

4.2. Set the minimum support

We set the minimum support a reasonable range of strategy is from the user preference model (including the wisdom of experienced users) to extract empirical values. In order to get from the user preference model appropriate minimum support range, the similarity between the two query detection is essential. In particular, constructed by the user, and / or user preference model will be compared with the query to detect similarities. , And the matching degree calculated in different ways, respectively, and then sum them to obtain the degree of similarity queries. Then, for the user's range of minimum support structure from the user query, and / or similar query the top export items. Calculation details are described below:

Match the transaction ID .So constructed by the user, the model from the user preferences. Assumptions. And represent the properties and the domain base, where. And the matching of each and weight, the calculation formula is as follows:

$$T_G(g_{ax}, g_{bw}) = \begin{cases} 0 \text{ if } g_{ax}, g_{bw} \text{ hasn't hierarchy} \\ \dfrac{min(dist(g_{ax}), dist(g_{bw}))}{max(dist(g_{ax}), dist(g_{bw}))} \text{ Otherwise} \end{cases} \quad (1)$$

$$match_G(G_a, G_b) = \frac{\sum_{x=1}^{i}\sum_{w=1}^{j} T_G(g_{ax}, g_{bw})}{|G_a| \times |G_b|} \quad (2)$$

5. Experiment

To assess the effectiveness of the new algorithm, we perform all of the algorithms in Java, and some experiments on a PC machine, PC main unit is configured to: CPU-Intel (R) Core2Quad 2.4GHz, Memory 2GB, Windows XP operating system . For comparison, we also perform some simplification of the algorithm, the algorithm does not use closed pattern tree, and the lower boundary based on the joint support pruning.

Four data sets used in the experiment:

(1) Synthesis of the two attributes graphics database, which is generated by the graphics generator;

(2) MUT: a chemical compound data set, with a compound (graphics) of the structure and hydrophobic compounds (digital) logarithm;

(3) PTE: predictive toxicology evaluation challenges the graphic data set;

(4) DTP: from the DTP AIDS antiviral screening data sets of graphics data sets.

PTE and DTP is only one attribute of the graphic data sets, and the MUT is a two-dimensional data sets. For the MUT in the hydrophobic properties of the algorithm, we apply the format of the interval model. If the variables meet the conditions, then we believe that the establishment.

In the experiment, by gradually changing the top-k in, and value, we tested the new algorithm. In the one-dimensional graph database (PTE and DTP) in the excavation, we add the following constraints: projects must meet the conditions, which is the maximum common subgraph and represents the number of edges, is already used - defined threshold. We are set in the experiment.

The transaction number: Each of vertices and edges of the month: Vertex and edge labels a significant number.

Represent the number of projects available (unit: thousands) and number of projects being tested (unit: thousands). For and on behalf of frequent closed pattern set to obtain the running time (unit: seconds), for it means that the entire running time. The number in square brackets in order to obtain the running time of closed pattern set. In parentheses represent the ratio relative to the simplified algorithm (percentage).

In short, all problems within a reasonable run time to resolve. Observed in the run time has improved to a large extent, and is the number of test items on a large number also decreased. These results are all that strong evidence of the effectiveness of the new algorithm, in particular the joint support from the lower boundary determined, based on the general sort of pruning techniques.

6. Summary

For multidimensional Apriori association mining algorithm used, in order to provide an automated support threshold settings, we propose in this paper set the minimum support an intelligent system architecture, the integration of user preference models. Using this method, the system queries the specified user in the user preferences model to find the most similar to the query, they add up to the scope of access to appropriate support for user reference. According to our proposed method, Apriori algorithm is used to support threshold set is no longer entirely subjective, but also the experience from the additional knowledge of other users. This structure improves the efficiency of the process user queries, access to the rules or mining also tend to close the user's requirements.

7. References

[1] B. Liu, W. Hsu, L.F. Mun, et al. Finding interesting patterns using user expectations. IEEE Trans. Knowledge and Data Engineering, 1999, 11(9): 817-831

[2] G. C. Garriga, H. Heikinheimo, J. K. Sepp¨anen. Cross-mining binary and numerical attributes. In Proc. of the 7th IEEE International Conference on Data Mining, 2007: 481-486

[3] Y. Ke, J. Cheng, W. Ng. Mining quantitative correlated patterns using an information-theoretic approach. In Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006: 227-236

[4] P. Songram, V. Boonjing, S. Intakosum. Closed multidimensional sequential pattern mining. In Proc.of the 3rd International Conference on Information Technology: New Generations, 2006: 512-517

[5] H. Xiong, P.-N. Tan, V. Kumar. Hyperclique pattern discovery. Data Mining and Knowledge Discovery, 2006, 13(2): 219-242

[6] J. Cheng, Y. Ke, W. Ng. Graphgen. A graph synthetic generator. Available from: http://www.cse.ust.hk/graphgen/.

[7] A. Srinivasan, S. Muggleton, R. King, et al. Mutagenesis: Ilp experiments in a nondeterminate biological domain. In Proc. of the 4th International Workshop on Inductive Logic Programming, 1994: 217-232

[8] C. Borgelt, M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In Proc. of the 2002 IEEE International Conference on Data Mining, 2002: 51-58

[9] Wynne Hsu, Mong Li Lee, Ji Zhang. Image mining: Trends and Developments. Journal of Intelligent Information Systems, 2002, 19(1): 7-23

[10]Kitamoto. Data mining for typhoon image collection. In Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), 2001: 68-78

[11]Fayyad U.M., Djorgovski S.G., N Weir. Automating the analysis and cataloging of sky surveys. Advances in Knowledge Discovery and Data Mining, 1996: 471- 493

[12]O.R Zaiane. Mining Multimedia Data. CASCON: Meeting of Minds, 1998: 76-87

[13]Ordonez C., Omiecinski E. Discovering association rules based on image content. In IEEE Advances in Digital Libraries Conference, 1999: 113-125

[14]Gao Cong, Anthony K. H., Tungxin Xu, et al. FARMER: Finding interesting rule groups in microarray datasets. In Proc. 2004 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'04), 2004: 129 -141