A Study of Negative Association Rules Mining Algorithm Based on Multi-Database

Xushan Peng, Ping Cheng, Maoji Wang
Information Engineering College
Ninbo Dahongying University
Ningbo China
e-mail:dhypengxushan@126.com
e-mail: chpzmm@yahoo.com.cn
e-mail:mogeiwang@gmail.com

***Abstract*—the mutual exclusion relationships among data items are reflected by negative association rules, whitch is very important on the decision-making analysis. In the last several years, negative association rules are frequently researched, while the study object of it is single mining of database now. With the development of database technology, multi-database mining is more and more important. On the basis of analyzing the related technology, research status and shortage of present negative association rules mining , the selecting rules, weighted synthesis and algorithm are discussed on multi-datobase.**

Keywords-data mining; negative association rules; multi-database

I. INTRODUCTION

The data mining is also defined as the knowledge discovery, data analysis, knowledge extraction or data acquisition from database. It is a process of finding and extracting implicit, unknown and potentially useful information and knowledge. Association rules are extracted by Agrawal et al in 1993 [1], it is one of the most important research field in data mining, it reveals the potential useful relationship in large-scale affairs among each item sets. For each relationship of item set, the positive correlation such as A = > B is mainly researched, namely the appearance of data item A will inevitably lead to the appearance of another data item B, such as 90% customer s who buy bread and butter can also buy milk, the correlation has the characteristics of high frequency and strong correlation, thus the research has been complete. In fact, there are still other relations between data items, such as the appearance of data item A lead to the unappearance of data item B , namely 90% customers who buy coffee won't buy tea or when some factors appeared, which factors will not appear or rarely appear and so on. From the reverse side and exclusive Angle to study the relationship among the project, which is also important in the decision analysis. In recent years, the research is quite frequent, namely negative association rules of the association rules.

Negative association rules describe the mutual exclusion relationships among items, such as the association rules of A = > ¬ B, ¬ A = > B, ¬ A = > ¬ B (each of A and B was frequent itemsets) , which meet the user'appoiting the minimum support degree minsupp and minimum confidence minconf association rules in the database D. Negative association rules is the research center of association rules mining in recent years, and has produced a lot of research achievements. At present, the negative association rules mining algorithm is mainly directed against a single database mining, in which the correlation is considered in the same database in each of the relations between the data items. With the fast development of database technology, we have to face multiple databases in any cases and consider the relations of multiple databases among the data items. Multi-Database system has been applied in reality life, decision-makers face the relationship handling between data items distributed in different database, which involves the mining problem in multiple databases. Multiple databases according to its class will be divided into different type of the database and how to excavate the negative association rules among different databases are important content researched in this paper.

II. NEGATIVE ASSOCIATION RULES MINING AND RELATED TECHNOLOGY

A. *related concept description*

Definition 1: Set I = {i1, i2,... , im} are assembles in which include m different attributes (itemset), D is business database ,in which every transaction T is a child of I, namely T □ I. Every transaction has a unique identifier (tid). Set A is a assemble of items of I, if A □ T, then affairs T contains A. If A contains k itemsets, it names A is k itemsets.

Definition 2: The support degree (support) of rule A = > B in the transaction database D is the ratio of the affairs of A and B and affairs of all recorded as support (A = > B) or support (A ∪ B), namely:

support (A = > B) = support (A □ B) = | {T: A □ B □ T, T ∈ D} | / | D |. The unappearance of item set A is recorded as ¬ A, apparently Support (¬ A) = 1 - Support (A). If the support degree of a itemset is greater than that of user 'appointing the minimum support degree (minsupp), then it is frequent, frequent itemsets whose length is k called frequent k - itemsets.The confidence of rules A = > B in affairs is the ratio of including the affairs of A and B and that of A recorded as confidence (A = > B), namely:

confidence (A = > B) = | {T: A □ B □ T, T ∈ D} | / | {T: A □ T, T ∈ D} |.

Definition 3: for A given item set A and B, A ∪ B = Φ, there are eight kinds of association rules between A and B.

(1) A = > B, (2) A = > ¬ B, (3) ¬ A = > B, (4) ¬ A = > ¬ B, (5) B = > A, (6) B = > ¬ A, (7) ¬ B = > A, (8) ¬ B = > ¬ A.

In which (1) is called positive association rules, (2) - (4) is called negative association rules, in which the project show negative relationship. In addition,(5) - (8) is opposite with (1) - (4), and just make the letters A and B of (1) - (4) exchange, here will not discuss.

An effective negative association rules must satisfy three conditions as follows:

(1) A ∪B = Φ

(2) supp (A) was minsupp and supp (B) was minsupp

(3) supp (A □ ¬ B) ≥ minsupp or supp (¬ A □ B) ≥minsupp or supp (¬ A □ ¬ B) ≥ minsupp.

B. Negative association rules mining technology

Negative association rules contain itemsets (such as ¬ A, ¬ B) which are not existed, and it is difficult to calculate their support and confidence degree directly, while it can be calculated by using positive association rules support and confidence degree of itemsets A and B, literature [2] has the corresponding conclusion:

Theorem 1 set A, B □ I, A □ B = Φ, then

(1) supp (¬ A) = 1 - supp (A);

(2) supp (A □ ¬ B) = supp (A) - supp (A ∪ B)

(3) supp (¬ A □ B) = supp (B) - supp (A ∪ B)

(4) supp (¬ A □ ¬ B) = 1 - supp (A) - supp (B) + supp (A ∪ B)

Inferential 1: set A, B □ I, A ∩ B = Φ, then

(1) conf (A = > ¬ B) = $\frac{\sup p(A) - \sup p(A \cup B)}{\sup p(A)}$ = 1 - conf (A = > B)

(2) conf (¬ A = > B) = $\frac{\sup p(B) - \sup p(A \cup B)}{1 - \sup p(A)}$

(3) conf (¬ A = > ¬ B) = $\frac{1 - \sup p(A) - \sup p(B) + \sup p(A \cup B)}{1 - \sup p(A)}$ = 1 - conf (¬ A = > B)

For the mining problems of negative association rule, solving a contradiction of the rules are the first job, namely the rules conf (A = > B)≥ minconf and conf (A = > ¬ B) ≥minconf establish at the same time, which is obviously a contradiction. For the appearance of contradictions in positive and negative association rules ,it can be avoided by correlation in association rules. The correlation of association rules is definited in literature [3], itemsets A and B can be calculationed by formula :

$$\text{CorrA, B} = \frac{\sup p(A \cup B)}{\sup p(A) \sup p(B)}$$

Through the CorrA, B ,we can judge the correlation of A, B:

If CorrA, B > 1, then A and B are positive correlation;

If CorrA, B = 1, then A and B are mutual independence;

If CorrA, B < 1, then A and B are negative correlation.

The correlation of itemsets A and B have such relationship: if CorrA, B > 1, then CorrA, ¬ B < 1; Corr ¬ A, B < 1; Corr ¬ A, ¬ B > 1; opposition is also opposite. So contradiction rule can be avoided just by judging the correlation among items when excavating positive negative association rules, that is when CorrA, B > 1 only excavating A = > B and ¬ A = > ¬ B; When CorrA, B < 1 only excavating ¬ A = > B and A = > ¬ B; When CorrA, B = 1 don't excavate.

C. the present situation of negative association rules mining and its shortage

For the research of negative association rules mining, Brin et al mentioned negative correlation between two frequent itemsets [3] firstly in 1997, Savasere et al expounds the strong association rules in literature [4]. literature [5] proposed a negative association rules mining algorithm based on matrix, literature [6] proposed an positive and negative association rules mining algorithm depended on interest, literature [7] proposed the positive and negative association rules mining algorithm based on degree of support, confidence and correlation coefficient and literature [8] proposed positive and negative association rules based on sequence model. A kind of PNARC model is extracted in literature [9] ,which not only can excavate positive and negative association rules of the frequent itemsets but also can detect and delete mutual contradictory rules. The above studies about negative association rules algorithm is from different angles and different ways, but the common problems is that database mining is based on a single database, that is to say the considerate connection is in the same database in each of the relations between the data items. In fact, with the fast development of database technology, multi-database system has been applied in real ity life, therefore we must consider researching multi-database mining problem.

III. NEGATIVE ASSOCIATION RULES MINING TECHNOLOGY BASED ON MULTI-DATABASE

Multi-Database mining are based on decision problem of knowledge discovery under global enterprise distribution data and discover novel useful model process from different databases. It is more difficult excavating negative association rules in multiple databases than that in single database. But we can use a single database mining knowledge to excavate negative association rules in the multiple database, the idea is: Firstly, the multiple databases can be classified according to a certain rules, eliminating ambiguity caused by different database [10]; Secondly, the similar data in each databade can be pretreatmented, such as removing meaningless、redundant noise rules and making database become cleaner; Lastly, making a knowledge synthesis excavated from each same type of databases. Therefore, the multi-database mining generally is divided into three steps: First, classify the database. Second, extract knowledge from the same database. Third, make knowledge synthesis from the same database mining, which generally adopt weighted method for synthesis of all the information in the database.

A. database and the weighted value of its correlation rules

In scientific research and application, we usually use weighted method to analyse and synthetise different database information [10], literature [11] adopts the weighted value to synthetise multi-database association rule, synthetised with the weighted value of database and association rule .

Set $D_1, D_2,\dots, D_m$ for m different databases, $S_1, S_2,\dots, S_m$ respectively for the similar data association rule sets, $S = \{S_1, S_2,\dots, S_m\}$ for total association rule sets, $R_1, R_2, R_j,\dots, R_n$ for specific rule set in total association rule sets S, including j = 1, 2, 3,... , n, Num (D_i) for number of transaction of database D_i, then the weighted value of database D_i is:

$$\omega_{D_i} = \frac{Num(D_i)}{\sum_{i=1}^{m} Num(D_i)}$$

The weights of rules Rj contain the sum of weights with the rules of the database, standardized weights of the rules is:

$$\omega_{R_j} = \frac{\sum_{i=1, R_j \subset S_i}^{m} \omega_{D_i}}{\sum_{j=1}^{n} \sum_{i=1, R_j \subset S_i}^{m} \omega_{D_i}}$$

B. Simplification of association rules before Synthesis

When the negative association rules in databases are excavated, there are conflict rules from one database rules to another, such as D1 have rules A = > B, there are rules in D2 A = > ¬ B, then the conflict come into being. In order to obtain the correct association rules, we can use the correlation to judge, we have used correlation to detect conflict rules with a single database ahead, here this method will be further applied to multi-database.

(1) A = > B, (2) A = > ¬ B, (3) ¬ A = > B, (4) ¬ A = > ¬ B, apparently (1) (4) and (2) (3) is contradictory, if there are this kind of contradictory rules ,then the correlation of itemsets A and B will be judged after synthesis:

$$corr_{\omega(A,B)} = \frac{\sup p_w(AB)}{\sup p_\omega(A) \sup p_\omega(B)}$$

Among $\sup p_\omega(AB)$, $\sup p_\omega(A)$ and $\sup p_\omega(B)$ is the degree of support after frequent item sets sythetised, which is different from the front database is that every itemsets have added the weights.

(1) If the corr $_{\omega(A, B)} > 1$, only mining rules A = > B and ¬ A = > ¬ B;

(2) If the corr $_{\omega(A, B)} < 1$, mining rules A = > ¬ B and ¬ A = > B;

(3) If the corr $_{\omega(A, B)} = 1$, no mining rules.

After this judgment, the contradictory rules of the database will be got rid of.

C. Synthetised negative association rules

Set $D_1, D_2,\dots, D_m$ for m different databases, rule sets Si is association rules of database D_i (I = 1, 2,..., m). $\omega_{D_1}, \omega_{D_2}, \dots, \omega_{D_m}$ respectively is weights of database $D_1, D_2,\dots, D_m$, for specific association rules A = > B, A = > ¬ B (or ¬ A = > B, ¬ A = > ¬ B) the degree of support and confidence after synthesis is:

$$\sup p_\omega(A \Rightarrow B) = \omega_{D_1} \times \sup p_1(A \Rightarrow B) + \omega_{D_2} \times \sup p_2(A \Rightarrow B) + \dots + \omega_{Dm} \times \sup p_m(A \Rightarrow B)$$

$$\sup p_\omega(A \Rightarrow \neg B) = \omega_{D_1} \times \sup p_1(A \Rightarrow \neg B) + \omega_{D_2} \times \sup p_2(A \Rightarrow \neg B) + \dots + \omega_{Dm} \times \sup p_m(A \Rightarrow \neg B)$$

$$conf_\omega(A \Rightarrow B) = \omega_{D_1} \times conf_1(A \Rightarrow B) + \omega_{D_2} \times conf_2(A \Rightarrow B) + \dots + \omega_{Dm} \times conf_m(A \Rightarrow B)$$

$$conf_\omega(A \Rightarrow \neg B) = \omega_{D_1} \times conf_1(A \Rightarrow \neg B) + \omega_{D_2} \times conf_2(A \Rightarrow \neg B) + \dots + \omega_{Dm} \times conf_m(A \Rightarrow \neg B)$$

Confidence degree can be showed through the support degree:

$$conf_\omega(A \Rightarrow B) = \frac{\sup p_\omega(A \cup B)}{\sup p_\omega(A)}$$

$$conf_\omega(A \Rightarrow \neg B) = \frac{\sup p_\omega(A \cup \neg B)}{\sup p_\omega(A)}$$

D. Algorithm design

Before synthesis in the mode, the data needs to be pretreated, namely useful rules is selected. After deleting redundancy and noise,the data we get will be more refined, finally synthetised algorithm of the mode will be more effective and the results will be more accurate.

The process of selection rules are as follows:

Algorithm: RuleSelection (S)

Input: min., minimum turnout; S: rules set of number for N; weights of database D_i, i = 1, 2,... , m;

Output: S, select simplified rules set;

(1) if there are contradictory rules existed in S do

$$corr_{\omega(A,B)} = \frac{\sup p_w(AB)}{\sup p_\omega(A) \sup p_\omega(B)}$$

If $corr_\omega(A, B) > 1$

S←S-{¬ A=>B,A=>¬ B};

If $corr_\omega(A, B) < 1$

S←S-{A=>B,¬ A=>¬ B};

(2) for each rule R in the rule set S do

$$\omega_{R_j} \leftarrow \frac{\sum_{i=1, R_j \subset S_i}^{m} \omega_{D_i}}{\sum_{j=1}^{n} \sum_{i=1, R_j \subset S_i}^{m} \omega_{D_i}}$$

If ω_{R_j} < min.

S←S-{ R_j };

End for;

(3) output S;

General idea of RuleSlection is: if there are conflicts in the rules, remove such rules firstly; if the turnout of rules Ri are not satisfied with the threshold min., cancell Ri from rules set S, and min. is defined by the specific user or experts, its value may vary in different database. Through the selection rules, the number of rule set is reduced and then reuse database weights to synthetise association rules.

The algorithm of rules sythesis as follows:

Input: S_1, S_2,... S_m rules set; Minsupp: support degree threshold value; Minconf: value reliability threshold value;

Output: synthetised association rules;

(1) $S \leftarrow \{ S_1 \cup S_2 \cup \cdots S_m \}$;

(2) Call RuleSlection (S);

(3) for Rule set S of the rules A = > B do

$\sup p_\omega (A \Rightarrow B) = \omega_{D_1} \times \sup p_1 (A \Rightarrow B) + \omega_{D_2} \times \sup p_2 (A \Rightarrow B) + \ldots + \omega_{Dm} \times \sup p_m (A \Rightarrow B)$;

$$conf_\omega (A \Rightarrow B) = \frac{\sup p_\omega (A \cup B)}{\sup p_\omega (A)};$$

(4) according to the degree of support to arrange the rule R in the rules S;

(5) output the association rules R whose degree of support and confidence are greater than the threshold value in S;

IV. CONCLUSION

In this paper, above algorithm is correct and feasible certificated by the experimental data. the multi-databases contain different kinds of databases, the data mining is discussed about the same kind of database only. How to excavate the negative association mining among different kinds of databas is still a new research direction, which needs to be further studied. In addition, there are many factors for the mining of multi-database to be considered. And there are many methods for the selection of negative association rules, while which is discussed in this paper with only one method .How to select rules based on different factors still needs to be further researched.

REFERENCES

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. Proc of ACM SIGMOD Int Conf Management of Date [C]. Washington D C,

[2] Xushan Peng, Yanyan Wu. The Research and Application of Algorithm for Mining Positive Association Rules. 2011 International Conference on Electronic and Mechanical Engineering and Information Technology, 2011442 9-4431.

[3] Brin S, Motwani R, Silverstein cristiano Beyond market: Generalizing association rules to correlations [A]. Processing of the ACM SIGMOD Conferrence 1997 [m]. New York: ACM Press, 1997.265-276.

[4] Savasere A, Omiecinski E, Navathe S.M ining for strong negative associations in A large database of customer transaction [C] / / Proceeding of the IEEE 14 th Int Conference on Data Engineering losAlamitos: IEEE - CS, 1998, : 494-502.

[5] LuXueYan, wang yong, ZhouYongQuan. A bit matrix based on the negative association rules mining new algorithm [J]. Journal of guangxi university for nationalities (natural science edition), 2007, (4) : 57-60.

[6] DongXiangJun, SongHan Bristol, ginger together, et al. Based on the minimum degree of interest in the positive and negative association rules mining [J]. Computer engineering and application, 2004, 27:24-31.

[7] zhang qian, WangZhi and, ZhangGuoZhi. Based on the correlation coefficient of the positive and negative association rules mining algorithm [J]. Journal of shaanxi institute of science and technology, 2005, (4) : 35-38.

[8] GuoYueBin, ZhaiYanFu, DongXiangJun et al. Based on sequence model of positive and negative association rules [J]. Journal of shandong university (physical edition), September 2007 (9) : 88-95.

[9] DongXiangJun, WangShuJing, SongHan et al. Negative association rules [J]. Journal of Beijing university of science and technology, 2004, (11) : 78-81.

[10] SuYiJuan, YanXiaoWei. An improved algorithm for mining frequent set [J]. Journal of guangxi normal university (natural science edition), 2001, 12 (3) : 22-26.

[11] TangYiFang, cow dint, zhang t ultra. Data mining of association rules algorithm [J]. Journal of guangxi normal university (natural science edition), 2002, 16 (4) : 27-31.