Research on the incremental updating algorithm of the association rules with the timing decision

Min Zhang[1, a], Yice Zhang[2, b ,*]

[1]School of Dalian University, Dalian 116600, China.

[a]1120834586@qq.com, [b]zhangyice09@sina.com

Keywords: association rule；incremental updating；timing decision；association rule difference degree.

Abstract. In order to deal with the frequently updating of data in practice, association rules incremental updating algorithm combining with the timing decision begun to be widely studied. Ding Hu gave an association rule difference degree judgment method based on the completely frequent itemsets. This method can fully express the changes of association rules before and after the dataset updating, and reduce the times of the updating of association rules, therefore, this method was widely used. However, this method is not very accurate when calculating the new frequent itemsets. DFUP algorithm (Dynamic Fast Update Algorithm) is proposed in this paper, and this algorithm solves the problem of the above algorithm by introducing a dynamic database.

1. Introduction

The generalized definition of data mining is data mining from storage to find interesting knowledge in the process of large amounts of data in database, data warehouse or other information repository[1]. It uses machine learning, statistics, pattern recognition and data visualization technology, knowledge, rules or high-level information from the data, and allows the user to observe or read them from different angles, and access to information and knowledge for business decision-making, financial analysis, process control, business management, information recommendation and the query processing,etc.

Association rules is one of the methods of data mining, the process of discovery association rules is association rule mining. But in the practical application, the database is always in constant change, such as increasing the data into a database or the data update, with the changing of data association rules mining, user requirements can reflect the current state of the database, so as to provide a theoretical basis and useful for market analysis and decision support. Therefore, update the association rules in data change research, because it has very important theoretical value and practical application value, has become an important research direction of association rules mining.

At present, the most committed to updating association rules algorithm incremental updating itself, while ignoring the grasp of the overall situation, people only pursuit algorithm is efficient, did not consider whether can get the new rules, if have not found the new rules, so that people do is useless. As a result, people began to pay attention to choose the right timing to update the association rules, association rules updating time to study to become one of research focuses in updating association rules algorithm.The existing association rules updating timing determination methods, generally uses the data update data set changes before and after the change of the quantity or the association rules to determine the association rules update timing. Including Ding Hu in sampling association rules algorithm based on data warehouse[2] is put forward by giving a difference degree of association rules based on frequent itemsets completely judgement method, the calculation method requires full computing frequent itemsets, can fully express the difference degree of association rules. But this method in determining the update time when need to scan for many times, the new data set that is bad for processing large data sets and long frequent itemsets. This article on the basis of this decision method is improved, and make it more accurate, to further improve efficiency.

2. THE INCREMENTAL UPDATING RELEVANT ALGORITHM

2.1Apriori Algorithm

In 1994, R.Agrawal et al. proposed the famous Apriori algorithm. Apriori algorithm has an important property: either the loophole all the set of frequent itemsets must also is frequent, on the other hand, if a hou anthology of loophole set is not frequently, so the candidate is must not frequently. Pruning algorithm according to the nature, selection of filter hou, so as to reduce the workload.

With minimum support and minimum confidence or transaction data in the database changes, mining the association rules must be updated, can correctly reflect the true relationship between different projects in transaction databases. In order to make the mining of association rules can correctly reflect the current state of the transaction database, a solution is to conduct a association rule mining algorithm, but it has wasted before the invention of association rules. When that happens, the updating association rules algorithm. Incremental updating association rules algorithm in view of the increasing data transaction databases, and on the basis of the original association rules, delete the old association rules does not meet the conditions, adding meet new conditions of association rules, implement updating association rules mining. Efficient incremental updating association rules algorithm is the key to utilize the existing association rules mining results, generate smaller or less candidate itemsets scan the transaction database, its purpose is to try to reduce the amount of calculation.

2.2 The FUP Algorithm

In order to make full use of already mined association rules, implement the incremental updating of association rules, D.W.C heung et al proposed FUP (Fast Update) algorithm [3]. FUP algorithm is based on Apriori algorithm thought, through the iterative search for many times, get the updated frequent itemsets of transaction databases, and thus on the basis of frequent itemsets find association rules. But FUP algorithm can effectively use the result of the frequent itemsets which has been excavated ,generate smaller candidate itemsets, reduce the cost of the incremental updating association rules, improve the efficiency of the incremental updating association rules mining.

For any of a set of k ($k \geqslant 1$) or more, has the following three properties:

Property 2.6: If the item set X is frequent in the transaction database DB and the new data set db, it's frequent in the new transaction database DB $\cup$ db.

Property 2.7: If the item set X is infrequent in the transaction database DB and the new data set db, it's infrequent in the new transaction database DB $\cup$ db.

Property 2.8: If the item set X is frequent in the transaction database DB(db) and is infrequent in the db(DB),you need to use the support count in DB (db) plus the support count in dB (DB) to determine whether or not the frequent item set.

DB is the original database, huge scale; db is a new database, initially empty.

FUP algorithm according to the Properties of the three updated in mining frequent itemsets of transaction databases, pruning of candidate itemsets on a large scale, thus greatly reducing the scanning times of original data and new data, improve the efficiency.

Although the efficiency of the FUP algorithm compared to the directly to the updated database using apriori algorithm is much higher, but there are still shortcomings:

1. Things as the original database scale increasing, the FUP efficiency is lower, mainly because if the itemsets is not at the same time the original database and new frequent itemsets, the algorithm need to repeat scan the transaction database of updated hou selection for pattern matching.
2. For large data sets and long frequent itemsets FUP performance is not high, the number of FUP algorithm scans database is mainly depends on the length of frequent item sets, in the presence of long frequent items data, need to scan the database repeatedly.
3. The efficiency of FUP algorithm is not high on the small size of the new database. this is because as long as there is a frequent itemsets in the new db, but is not frequently in the original DB,

you need to scan the original DB and the new db, and especially when the DB is large, the algorithm efficiency is low.

In view of this, the FUP still need to repeat scanning k times on the DB∪db. but in general, the size of the original database DB is very big, the scanning k times is quite time consuming, so can't easily to DB, then you need to make accurate judgments to the update timing, reduce the number of association rules updating, as far as possible to avoid doing this.

2.3 Timing Decision

Discovered association rules updating time basis mainly include: association rule difference degree based on 1- frequent item sets [4], based on the change of the data set percentage[5], based on the difference degrees of the data sets[6], the difference degree of association rules based on completely frequent itemsets [2]. Among them ,Ding Hu in sampling association rules algorithm based on data warehouse is presented a determine method of the difference degree of association rules based on completely frequent itemsets, it can fully express the changes before and after the update association rules data set, reducing the number of association rules updating, so it has been widely applied. The specific methods for:

$$diff(D,S)=\frac{\sum_{k=1}^{n}(|Lk(D)-Lk(S)|+|Lk(S)-Lk(D)|)}{\sum_{k=1}^{n}(|Lk(D)+Lk(S)|)}$$

Among them, the $L_k(D)$ said frequent k- itemsets in data set D, $L_k(S)$ in the data set S frequent k- itemsets, k = 1, 2, 3... n. Through the complete frequent itemsets is calculated between two sets of data, and then obtain the difference degree of association rules.

This method is mainly through the difference degree of frequent item set between the original DB and the new data db, given a threshold $\bar{d}$ [0, 1],when the difference degree thin than $\bar{d}$ update. With the coming of the new data, if don't need to update the association rules, the accumulation of new data sets are constantly changing and updating. After each association rules updating, at the same time, the accumulation of new data set to empty.In the case of several times in a row to update timing decision does not need to update the association rules, Because of we get the support count of the frequent itemsets in the accumulation of new data set in the process of determining, when new data come again to update the timing decision association rules,we can take advantage of the support count of the frequent itemsets in the accumulation of new data set, don't need to scan again to this part of the data set, to avoid the multiple scan accumulation of new data sets.

2.4 The DFUP Algorithm Based on Dynamic Database

So you say the completely difference degree of frequent item sets of association rules decision method in determining the timing update has two shortcomings:

1. Unable to grasp the opportunity of the new data db scanning, the new data db is relatively large, but not for its frequent scanning.
2. In the case of difference degree of association rules between DB and db is always not big enough ,must to scan db for many times, and in the process of db accumulated many times, is bound to appear this kind of circumstance: in the early stage of the new data db is not frequent itemsets, and then continuously increases, became the frequent itemsets, but as mentioned above, the timing of the original decision and there is no record of frequent itemsets support number, also did not consider this kind of situation, so you can't ensure that the resulting frequent itemsets is accurate.

For these two shortcomings,this paper proposes a DFUP (Dynamic Fast Update Algorithm) algorithm based on dynamic database,the proposed algorithm by introducing a dynamic database DD (Dynamic Database),as a buffer of the new database db and original database DB before the update. This can reduce the amount of data db, convenient for frequent scanning of the entire db. The ideas are described as follows:

Assume that the initial database DB scale is quite large, dynamic database DD and new database db is empty.

1. Using apriori algorithm for mining frequent items from the original database DB and preserved;

2. The increasing data in db, for the first time the data in the db will be added to the DD based on a certain period of time (the time period db has accumulated enough amount), at the same time with the apriori algorithm to a mining of DD, keep its frequent itemsets, by formula (1) to determine whether need to update the entire database. If need, then according to the FUP algorithm to update. If don't need ,to step 3;

3. db continues to join the data continuously，at regular intervals for a db mining (this time is shorter, because the db size is very small, so the db can be frequent mining), then determine the difference degree of frequent itemsets between the db and DD, if not big enough, the db continues to join the data, DD remains unchanged. If big enough ,to step 4;

4. Because the frequent itemsets of both db and DD have been excavated, before adding all the db to DD, the db and DD can be a small range of FUP , mining frequent item sets out after the merger, and judge the difference degree of frequent itemsets between the updated DD and DB, if big enough, according to the FUP algorithm to update. If not big enough, back to step 3.

Repeat the above process, it can ensure that the data could be updated timely and accurately.

3. Summary

The traditional FUP algorithm does not proceed the time to determine before the update,in front of the big data frequently updated, often update many times in a row, but didn't find the new rules, do a lot of busywork. After joining the time to determine, ensures that each update can be found that the new rules, looked from the overall work, work efficiency is greatly increased obviously. Compared with the determine method of difference degree of association rules based on completely frequent itemsets , DFUP algorithm accuracy and efficiency has been improved, which is more efficient and reliable mining.

References

[1] Han Jiawei,M.Kamber.Fan Ming,Meng Xiaofeng,eds. Data Mining: Concepts and Techniques [M]. 2nd ed. Beijing: Mechanical Industry Press, 2007:146-160.

[2] Ding Hu. Sampling association rules algorithm based on data warehouse research[D]. Harbin: Harbin Engineering University,2006.

[3] D．Cheung，J．Han，V．Ng，et a1．Maintenance of Discovered Association Rules in Large Databases：An Incremental Updating Technique[C]． Proceedings of the 1 2th International Conference on Data Engineering， New Orleans，1996：106-114．

[4] C.Bin,H.Peter,S.Peter.A New Two-phase Sampling Based Algorithm for Discovering Association Rules[C].Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,Edmonton,Alberta,Canada,2002:23-26.

[5] N.K.Sharma,N.K.Nagwani.Study and Analysis of Incremental Apriori Algorithm[J].Communications in Computer and Information Science,2011,169(3):470-472.

[6] Zhang Genxiang, Chen Haishan.Large-scale Data Set of Incremental Association Rules Mining [J]. Computer Engineering and Application, 2009, 45(29)：120—124．