Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Mining Frequent Item Sets from incremental database: A single pass approach

Sandhya Rani Jetti[1] and r.Sujatha D[2]

[1] JNTUH

Received: 20 October 2011 Accepted: 14 November 2011 Published: 28 November 2011

Abstract

Apriori based Association Rule Mining (ARM) is one of the data mining techniques used to extract hidden knowledge from datasets that can be used by an organization?s decision makers to improve overall profit. Performing Existing association mining algorithms requires repeated passes over the entire database. Obviously, for large database, the role of input/output overhead in scanning the database is very significant. We propose a new algorithm, which would mine frequent item sets with vertical format. The new algorithm would need to scan database one time. And in the follow-up data mining process, it can get new frequent item sets through 'and operation' between item sets. The new algorithm needs less storage space, and can improve the efficiency of data mining.

Index terms— Apriori; vertical format; Association Rule ; data mining.

1 I. INTRODUCTION

ata mining technology is mainly used to process massive amounts of data and information. It can help us find the potential and meaningful knowledge. Association rule mining is one of the most active research methods. It was proposed by Agrawal etc. to analysis market basket question. The purpose is to discover the association rules among different commodities included in trading database..

2 II. APRIORI ALGORITHM a) Brief overview

Apriori [1] algorithm is a classical algorithm to find association rules which was presented by Agrawal and Srikant in 1993. It is the boolean association rules algorithm to mining frequent item sets. The basic idea of the algorithm is to recursively generate frequent item sets. The basic idea of the algorithm is that it starts with the only 1-item sets, recursively generates a frequent 2item set, and then produces a frequent 3-item set, and continues like this, until all frequent item sets are produced. The algorithm will stop till generate all the frequent item sets.

Author ? : Student, CSE, Aurora's Technological and Research Institute, Hyderabad, India. E-mail : Sandya.jetti@gmail.com Author ? : Head Of the Dept, CSE, Aurora's Technological and Research Institute , Hyderabad, India. E-mail : sujatha.dandu@gmail.com b) Properties

The properties of Apriori algorithm is that an item set is frequent and its all non-empty subsets are frequent. In other words, if an item set is not frequent, all of its super-sets will not be frequent. (k-1)-item sets) for each item set p1,Lk-1 for each item set p2,Lk-1 if(p1[1]=p2[1]),(p1[2]=p2[2]),...,(p1[k- 1]=p2[k-1]

) then{ c=p1 connect p2; if has_infrequent_subset(c,Lk-1) then delete c; else add c to Ck ; } return Ck; procedure has_infrequent_subset(c:candidate k-item set; Lk-1:frequent (k-1)-item sets) for each (k-1)-item set s of c If s not in Lk-1 then return TURE; return FALSE; a). This algorithm may generate many candidate item sets in the calculation process. When the number of frequent 1-item set is very big or the frequent patterns are very long, the number of the generated candidate item sets will increase sharply. So the algorithm's efficiency will fall dramatically. For example, if the number of frequent 1-item set is 104, the number of candidate 2-item set we need to generate, will be 107. If the length of frequent mode is 100, we will need to generate 2100 candidate sets. If we search so many candidate sets, the efficiency of the algorithm should be fairly low.

3 III. RELATED WORK a) Brief overview

Apriori is a horizontal format algorithm. It is for mining frequent item sets. It needs to scan the database repeatedly to get support degree of the candidate sets. The time consumption in this process is the key of the algorithm. A variety of improved algorithms proposed to reduce the Comparison times between candidate sets and the Transaction records. Such as the DHP [2], FPgrowth [3] and so on. They all have played a certain role in this regard. Here, we think that if there is a new algorithm can eliminate this comparison process, it will make the performance improved greatly.

This paper presents a new algorithm, which mine frequent item sets with vertical format. The new algorithm only needs to scan database one time to get frequent 1-item set. In the follow-up of the mining process we statistic support degree of candidate sets which is used to generate table. We needn't to scan database again.

4 b) Algorithm Benefits

Benefits of vertical algorithm for mining frequent item sets: It will be able to judge whether the nonfrequent item sets before generating candidate item sets. This can save time. Since each TID set of k-item set to carry the complete information that can calculate the support degree, so we needn't to scan the database to calculate the support degree of (k+1)-item set.

5 c) Description of the algorithm

First, scan the database to generate frequent 1item set.Second, transform horizontal format of frequent 1-item set into vertical format. Then do 'and operation' among each element of frequent item set Lk and record the result. If the result is more than the min_sup, we'll obtain a candidate set Ck+1, else we will do the next 'and operation'. We will stop doing the 'and operation' till the following situations come: there is a request item set left and we have no way to do the 'and operation' or all the results of 'and operation' is less than min_sup.

IV.

6 ALGORITHM EXAMPLES

Mining Frequent Item Sets from incremental database: A single pass approach ©

f. b1?b2=8 FLAG=1<2 delete So the maximal frequent item set is {I1,I2,I3} and {I1,I2,I5}.

7 V. RELATED WORK

The literature "A vertical format algorithm for mining frequent item sets" that considered as motivation for this proposal discussed an effective Apriori approach that avoids the multiple passes to the database. This model named by the author as vertical approach. b). This algorithm needs to scan the database many times and check candidate item sets by matching pattern. If the database is very large and the patterns needed to be matched are also very long, the efficiency of the algorithm will be greatly reduced.

8 VI. MY CONTRIBUTION

The solution discussed in the literature "A vertical format algorithm for mining frequent item sets" given single pass approach to perform Apriori. This solution limited to the stable dataset. I would like to extend this work to handle the multi pass problem in incremental dataset also known as streaming data.

VII. [1] [2]

[1] © 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XXI Version I

[2] December

Figure 1:

.1 ACKNOWLEDGMENT

I would like to thank Prof D.Sujatha(HOD), Sr. Asst.Prof A. Poongodai and Prof D. Sujatha (Aurora's Technological and Research Institute, Hyderabad, India) for proposing the concept of Mining Frequent Item Sets from incremental database: A single pass approach as well as providing their careful reading and valuable suggestions. I would also like to thank the anonymous referees for their helpful comments, correction and suggestions to improve this work.

[Cuiru and Shaohua (2008)] 'An Improved Apriori Algorithm for Association Rules'. Wang Cuiru , Wang Shaohua . *Computer Technology and Applications* February 2008.

[Dunham ()] Margaret H Dunham . *Data Mining Itroductory and Advanced Topics*, 2005. Tsinghua University Press. p. .

[Han and Kamber ()] Jiawei Han , Micheline Kamber . *Data Mining Concepts and Techniques*, 2006. China Machine Press. p. . (2nd ed)

[Song Jingjing Mining Maximal Frequent Patterns in a Unidirectional FP-tree ()] *Song Jingjing Mining Maximal Frequent Patterns in a Unidirectional FP-tree*, May2007. Henan Zhengzhou. Henan University