Performance Evaluation for Frequent Pattern mining Algorithm

Mr.Rahul Shukla, Prof(Dr.) Anil kumar Solanki

Mewar University,Chittorgarh(India), Rsele2003@gmail.com

Abstract— frequent pattern mining is an essential data mining task, with a goal of discovering knowledge in the form of repeated patterns. Many efficient pattern mining algorithms have been discovered in the last two decades to enhance the performance of Apriori Algorithm, for the purpose of determining the frequent pattern. The main issue for any algorithm is to reduce the Execution time. In this paper we compare the performance of Apriori and ECLAT Algorith on the medical data and find out the interesting pattern in medical data.

Keywords— Apriori, Frequent Patterns, Support , Eclat , Association rule ,confidence, Itemset

introduction

frequent pattern mining has applications ranging from intrusion detection and Market basket analysis, to credit card fraud prevention and drug discovery it is the analysis of dataset to find unsuspected relationship and to summarize the data in new ways that are both understandable and useful. Evolutionary progress in digital data acquisition and storage technology has resulted in huge and voluminous databases. Data is often noisy and incomplete, and therefore it is likely that many interesting patterns will be missed and reliability of detected patterns will be low. This is where, knowledge Discovery in databases (KDD) and Data Mining (DM) helps to extract useful information from raw data. Frequent patterns are those that occur at least a user-given number of times (referred as minimum support threshold) in the dataset. Frequent item sets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters. Frequent pattern mining is one of the most important and well researched techniques of data mining. The mining of association rules is one of the most popular problems. The original motivation for searching association rules came from the need to analyze so called supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. Association rules describe how often items are purchased together. Such rules can be useful for decisions concerning product pricing, promotions, store layout and many others. Being given a set of transactions of the clients, the purpose of the association rules is to find correlations between the sold articles. Knowing the associations between the offered Products and services help those who have to take decisions to implement successful marketing techniques. Based on the obtained results and comparative statistical interpretations, we Issued hypotheses referring to performance, precision and accuracy of the two processes Apriori and Eclat Frequent pattern Approach.

A. Association Rule

Association rule is used to find out the items which are frequently used together. The Presence of one set of items in a Transaction implies other set of items. The terms used in these Rules are

Support***:*** The support of an association rule X implies Y is the Percentage of transaction in the database that consists of X U Y.

Confidence: The confidence for an association rule X implies Y is the ratio of the number of transaction that contains X U Y

To the number of transaction that contains X.

Large Item Set: A large item set is an item set whose number of occurrences is above a threshold or support. The task of association rule mining is to find correlation relationships among different data attributes in a large set of data items, and this has gained lot of attention since its introduction. Such relationships observed between data attributes are called association rules. A typical example of association rule mining is the market basket analysis [2].

Limitation of Apriori

One is to find those item sets whose occurrences exceed a predefined Support in the database; those item sets are called frequent Pattern The Apriori Algorithm can be further divided into two sub-part candidate large item sets generation process and frequent item sets generation process. Frequent item set or large item set are those item sets whose support count exceeds the value of support threshold. Due to Number of passes apriori takes the more time. It scan the Database many time for Frequent pattern Discovery.

Apriori Algorithm for Frequent Pattern Mining

It searches for large item sets during its initial database pass and uses its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent item sets and those below are called small item sets. The algorithm is based on the large item set property which states: Any subset of a large item set is large and any subset of frequent item set must be frequent. The first algorithm for mining all frequent item sets and strong Association rules were the AIS algorithm by [3]. Shortly after that, the algorithm was improved and renamed Apriori. Apriori algorithm is, the most classical and important algorithm for mining frequent itemsets.The Apriori algorithm performs a breadth-first search in the search space by generating candidate k+1-itemsets from frequent k item sets. The frequency of an item set is computed by counting its occurrence in each transaction. Apriori is an influential algorithm for mining frequent item sets For Boolean association rules. Since the Algorithm uses prior knowledge of frequent item set it has been given the name Apriori. It is an iterative level wise search Algorithm, where k Item sets are used to explore (k+1)-item sets. First, the set of frequents 1- item sets is found. This set is denoted by L1. L1 is Used to find L2, the set of frequent 2-itemsets, which is used to Find L3 and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of Database
There are two steps for understanding that how Lk-1 is used to find Lk:-

A. The join step:-

To find Lk, a set of candidate k-item sets is generated by joining Lk-1 with itself. This set of candidates is denoted Ck.

B. The prune step:-

Ck is a superset of Lk, that is, its members may or may not be Frequent, but all of the frequent k-item sets are included in Ck.

A scan of the database to determine the count of each candidate in Ck would result in the determination of Lk.Ck, however, can be huge, and so this could involve heavy computation to reduce the size of Ck.

A scan of the database to determine the count of each candidate in Ck would result in the determination of Lk. Ck,However, can be huge, and so this could involve heavy computation. To reduce the size of Ck , the Apriori property is used as Follows.

i. Any (k-1)-item set that is not frequent cannot be a subset of frequent k-item set.

ii. Hence, if (k-1) subset of a candidate k item set is not in Lk-1 then the candidate cannot be frequent either and so can be removed from C.

Based on the Apriori property that all subsets of a frequent item set must also be frequent, we can determine that four latter Candidates cannot possibly be frequent. How? For example, let’s take {I1, I2, I3}. The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1, I2, I3} are members of L2, We will keep {I1, I2, I3} in C3.

Let’s take another example of {I2, I3, I5} which shows how the pruning is performed. The 2-item subsets are {I2, I3}, {I2, I5} & {I3,I5}.

BUT, {I3, I5} is not a member of L2 and hence it is not

Frequent violating Apriori Property. Thus we will have to

Remove {I2, I3, I5} from C3.

Therefore, C3 = {{I1, I2, I3}, {I1, I2, I5}} after checking for

All members of result of Join operation for Pruning.

C.Apriori algorithm pseudo code:

```
Procedure Apriori (T, minSupport)
{ //T is the database and minSupport is the minimum support
    L1= {frequent items};
    for (k= 2; Lk-1 !=∅; k++)
    {
        Ck= candidates generated from Lk-1
        for each transaction t in database do
        {
            increment the count of all candidates in Ck
            that are contained in t
            Lk = candidates in Ck with minSupport
        }//end for each
    }//end for
    return ;
}
```

ECLAT Based Approach on Apriori Algorithm

Eclat is a vertical database layout algorithm used for mining frequent itemsets. It is based on depth first search algorithm. In the first step the data is represented in a bit matrix form. If the item is bought in a particular transaction the bit is set to one else to zero. After that a prefix tree needs to be constructed. To find the first item for the prefix tree the algorithm uses the intersection of the first row with all other rows, and to create the second child the intersection of the second row is taken with the rows following it [6]. In the similar way all other items are found and the prefix tree get constructed. Infrequent rows are discarded from further calculations. To mine frequent itemsets the depth first search algorithm is applied to prefix tree with backtracking.. Frequent patterns are stored in a bit matrix structure. Eclat is memory efficient because it uses prefix tree. The algorithm has good scalability due to the compact representation.

EXPERIMENTS

In this section, we evaluate the performance of Apriori and Eclat based Algorithm. To make the evaluation, we check the performance of Apriori and Eclat approach on the different support count and fixed support count with different size of dataset.

A. Dataset

In tune with our application, we have taken a database of Hospital transaction of 633 items. In this analysis process we considered 2000 transactions in Horizontal Transaction format to generate the frequent pattern . In the horizontal Transaction each Transaction contain the multiple medicine in the single row.The horizontal Transaction view is Given below in Table 1.

TID	Medicine
1	STERILE WATER 5ML,COLISPAS DROP,PERINORM INJ
2	MEROTEC 250 MG INJ
3	kefragard 0.75 inj
4	combiflam tab
5	MONOCEF SB 1GM INJ
6	RANTAC INJ,DISPO VAN 2ML SYREING,DISPO VAN 5ML SYRIENG,DISPO VAN 1 ML SYRIENG,Dany Ing.,DISPO VAN 50ML SYRINGE,AUGPEN-300 INJ
7	RANTAC INJ,PYRIMOL INJ,Dany Ing.,MONOCEF 500 INJ
8	MIKACIN 250 INJ,MONOTAX SB 750 INJ
9	NIZONIDE-O SYP
10	ALLEGRA 120 TAB,CRITUS-XF SYP,SENSICLAV-625 TAB,DOLO-500 TAB

Table-1.

Time Comparison of Apriori and Eclate Algorithm

As a result of the experimental study,revealed the performance of Apriori and Eclat algorithm.The run time is the time to mine the frequent pattern.we have taken transaction 2063 and experiment is applied on 2000 transaction of hospital data with different support count result of time is shown in the figure-1 reveals that outperformance over Eclat Algoruthm.

A. Dataset

In tune with our application, we have taken a database of Hospital transaction of 633 items. In this analysis process we considered 2000 transactions in Horizontal Transaction format to generate the frequent pattern . In the horizontal Transaction each Transaction contain the multiple medicine in the single row.The horizontal Transaction view is Given below in Table 1.

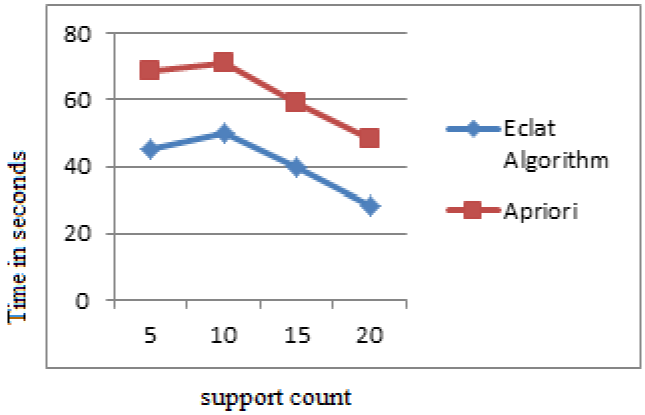

Fig-1:Execution Time Comparison on different support Count

Now The experimental is again applied on different size dataset with fixed support count 10 result of time is shown in the Fig-2 reveals that the Eclat performs better than Apriori

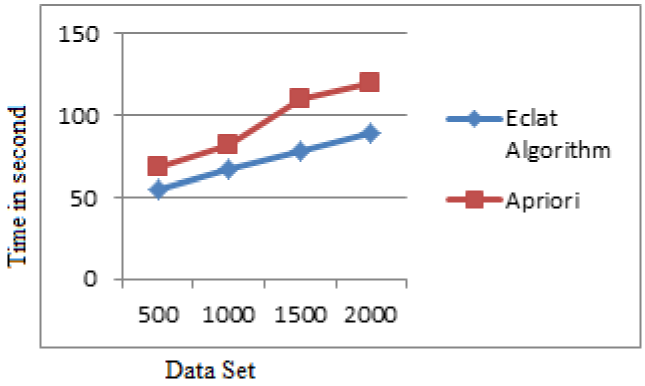

Fig-2:Execution Time comparison on different size of dataset

Conclusion

The association rules play a major role in many data mining applications, trying to find interesting patterns in data bases.Apriori is the simplest algorithm which is used for mining of frequent patterns from the transaction database. The main drawback of Apriori algorithm is that the candidate set generation is costly, especially if a large number of patterns and/or long patterns exist.Apriori algorithm uses large item set property, easy to implement, but it repeatedly scan the database. Apriori takes more time to scan the large Frequent patterns**.** The Eclat approach is used for efficient mining of frequent patterns in large databases. Eclat approach we use the intersection query and it is more efficient than apriori algorithm and also takes lesser time and gives better performance.

REFERENCES:

[1] Rahul Mishra, Abha choubey, Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012

[2]Huiping peng “Discovery of Interesting Association Rules on Web Usage

Mini ng" 2010 International Conference.

[3] Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules",

VLDB. Sep 12-15 1994, Chile, 487-99, pdf, ISBN 1-55860-153-8.

[4] Paresh Tanna, Dr. Yogesh Ghodasara**,** Using Apriori with WEKA for Frequent Pattern Mining, (IJETT) – Volume 12 Number 3 - Jun 2014

[5]Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216

[6] C.Borgelt. "Efficient Implementations of Apriori and Eclat". In Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations, CEUR Workshop Proceedings 90, Aachen, Germany 2003.

[7] D.N. Goswami, Anshu Chaturvedi, C.S. Raghuvanshi. Frequent Pattern Mining Using Record Filter Approach. International Journal of Computer Science Issues. 2010; 7(4): 38–43. Available from: ijcsi.org/papers/7-4-7-38-43.pdf

[8] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. IBM Research Report RJ9839, IBM Almaden Research Center, San Jose, California, June 1994.

[9] A. Amir, R. Feldman, and R. Kashi. A new and versatile method for association generation. Information Systems, 2:333–347, 1997.

[10] J. Park, M. Chen and Philip Yu, "*An Effective Hash-Based*

Algorithm for Mining Association Rules", Proceedings of ACM Special Interest Group of Management of Data, ACM SIGMOD'95, 1995.

[11] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "*New Algorithms for Fast Discovery of Association Rules*", Proc. 3rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'97, Newport Beach, CA), 283-296 AAAI Press, Menlo Park, CA, USA 1997

[12] Shruti Aggarwal, Ranveer Kaur, "*Comparative Study of Various Improved Versions of Apriori Algorithm*", International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue4- April 2013

[13] Adriano Veloso, Matthew Eric Otey, Srinivasan Parthasarathy, Wagner Meira Jr.; Parallel and Distributed Frequent Itemset Mining on Dynamic Datasets; Int'l Conf. on High Performance Computing; 2003.

[14] Haixun Wang, Wei Fan, Philip S. Yu, Jiawei Han; Mining Concept-Drifting Data Streams using Ensemble Classifiers; ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining; August 2003.

[15] Ran Wolff, Assaf Schuster; Association Rule Mining in Peer-to-Peer Systems; IEEE Transactions on Systems, Man and Cybernetics, Part B, Vol. 34, Issue 6; December 2004.

[16] Similarity Search using Concept Graphs, Rakesh Agrawal Sreenivas Gollapudi Anitha Kannan Krishnaram Kenthapadi