An Efficient Mining Algorithm for Closed Frequent Itemsets and its Associated Data

N. Kavitha
Research Scholar
KarpagamUniversity
Coimbatore.

S. Karthikeyan
Assistant Professor
Department of Information
Technology,College of Applied
Sciences,Oman

ABSTRACT
Database is a repository of information. Retrieving automatic patterns from the database provide the requisite information and are in great demand in various domains of science and engineering. The effective pattern mining methods such as pattern discovery and association rule mining have been developed and find its applicability in a wide gamut ranging from science to medical to military and to engineering applications. Contemporary methods of retrieval such as pattern discovery and association rule mining algorithms are useful only for retrieving the data. The limitations of using these techniques are that they are unable to provide a complete association and relationship among the diverse patterns that is retrieved. This paper attempts a solution to the above limitation by designing a new algorithm (CFIM) which generates closed frequent patterns and its associated data concurrently. CFIM makes explicit the relationship between the patterns and its associated data.

Keywords
Association Rule Mining, Frequent Closed Itemsets and Pattern Discovery.

1. INTRODUCTION
Frequent Itemset Mining (FIM) is in huge demand because of its ability to mine the interesting patterns from the database. Association rules, correlations, sequences and Frequent Itemset Mining (FIM) is in huge demand because of its ability to mine the interesting patterns from the database. Association rules, correlations, sequences and pattern discovery are some of the techniques of FIM which are used to find the patterns. Frequent patterns are so named, because those patterns appear frequently. FIM helps in disclosing the relationship and association among items in a voluminous data. The solution lies in generation of closed frequent patterns. Consider the fig.1 portrays the itemset lattice which is obtained from the simple database (Market –Basket). The main strategy adopted here is search space and computing closure. Table1. Shows the transaction database which contains set of itemsets and their support. For best understanding of closed itemset, an itemset is closed if none of its immediate supersets has the same support as the itemset. By seeing the fig 1. , it is possible to generate five closed itemset which is comparatively less than that which is generated by frequent itemsets when the minimum support is 1**.**Another technique which is available to generate patterns is pattern discovery. It is widely used; the greatest disadvantage of using this technique is it produces a large number of patterns which poses great difficulty for the user to interpret the information. Pattern Discovery lacks in systematic and objective way of combining the fragments of information. So the difficulties arise in exploring the data. To overcome these problems associated with pattern discovery and Frequent Itemset Mining, a new algorithm CFIM which simultaneously combines the closed patterns and associated data is found.

Tid	Itemset	Mininimum Support
1	Milk,Butter	1
2	Bread	7
3	Eggs	3
4	Bread, butter ,eggs	2
5	Bread ,butter	6
6	Bread ,butter	6
7	Milk,bread,butter ,eggs	1
8	Milk, butter	5
9	Milk,bread,butter	3
10	Milk,bread,butter	3

Table 1: Transaction of database

2. RELATED WORK
Research works in the area of developing algorithms for FIM are numerous. Every algorithm so developed in the course of research has got its own ups and downs. A brief summary of such algorithms are discussed below. The pioneer in this regard is the AIS [1] algorithm. AIS generate frequent item sets by employing candidate generation. The disadvantage of this algorithm is generating large number of candidate item set generation. Next in the hierarchy is the Apriori[2] algorithm which employs breadth first strategy to count the support of item sets with the help of candidate generation function. This algorithm was developed by Agarwal and Srikanth in the year of 1994. The drawback of this algorithm is scanning the databases more times frequently. Direct Hashing and Pruning (DHP) [3] is modified version of apriori which uses hashing techniques for generating candidate item sets. The downside of this algorithm is suitable only for two item sets. The FP-Growth [4] algorithm was also an

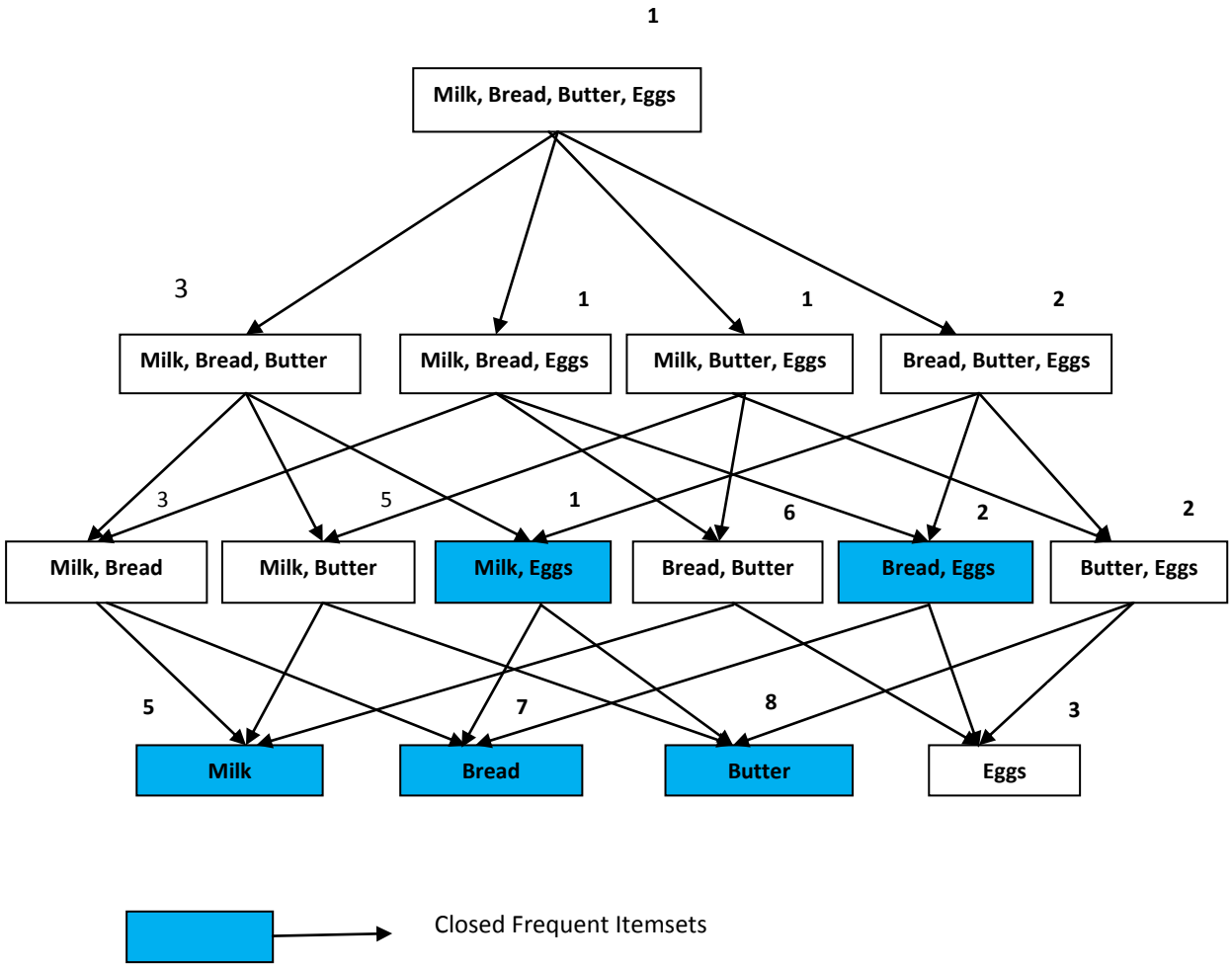

Fig 1. Item set Lattice

improvement of the apriori algorithm which was developed to mine the frequent patterns without generation of candidate item sets. This algorithm adapts to a divide and conquers strategy and used the frequent pattern tree. Brin [5] put forward the idea of using chi-square statistics to detect correlation rules from contingency tables. This method does not provide information regarding the strength of the relationship or its substantive signification in the population. Then the researchers concentrated in generating of maximal frequent itemsets.The algorithm called MAFIA [6] for generating maximal frequent item sets which was efficient for the dense database. This algorithm uses the depth-first traversal of the item set lattice. Max-Miner [7] is another algorithm for generating maximal frequent item sets which efficiently mining long patterns from databases. Later on the research focused on generating closed frequent patterns. A-CLOSE [8] which computes the closures of all minimal generators previously found. Since a single equivalence class may have more than one minimal item sets, redundant closures may be computed. CHARM [9], (Zaki and Hsiao) performs a bottom up depth-first browsing of a prefix tree of frequent item sets built incrementally. As soon as a frequent item set is generated, its tidlist is compared with those of the other item sets having the same parent. If one tidlist includes another one, the associated nodes are merged since both the item sets surely belong to the same equivalence class. Item set tidlists are stored in each node of the tree by using the diff-set technique .Since different paths can, however, lead to the same closed item set, also, in this case, a duplicates detection and pruning strategy is implemented. Another method proposed by Wong and Li [11] helps in simultaneously clustering the discovered patterns and their associated data. Its usefulness lies in its ability to relate the patterns to the set of compound events. The method employs to cluster the patterns is hierarchical agglomerative approach. The Discover * e algorithm is used to generate 'p' patterns. The process employing this technique is highly time consuming. To reduce the time required to cluster patterns, it is suggested by the researchers that to cluster closed frequent item sets which is far less than the number of all Frequents Item sets ies closed frequent item sets and associated data concurrently. This paper discusses the logic behind of using CFIM over the other algorithms such as Apriori CLOSET, CHARM, and Discover * e.

3. THE CLOSED FREQUENT ITEMSETS

The item set is presented by I= $\{i_1, i_2, \ldots\ldots i_m\}$. D stands for the database which contains a set of database transactions in which T represents a set of item sets which in turn a subset of I. There is an identifier TID, which is associated with each transaction. Then the Transaction database TDB is a set of transactions. Given a transaction database(TDB),the support of an item set X is denoted as sup(x).An association rule R is defined as X=>Y is an implication between two items X and Y where X,Y $\subset$ I and X$\cap$ Y=ϕ. The support of the rule is

denoted as sup(X=>Y) is defined as sup(X U Y).The incidence of the rule is denoted as conf(X=>Y), is defined as sup(XUY)/sup(X). The problem of association rule mining is to generate large number of frequent item set and generate large number of association rules in the database. CFIM, aimed at saving time in computing item set closures and their supports. Closed item sets mining and their corresponding rules which has the same power as association mining but substantially reduces the number of rules to be presented. Closed item set is defined as "An item set X is closed if none of the proper supersets of X have the same support."

4. METHODOLOGY

Closed Item sets are generated using the proposed algorithm (CFIM). The methodology of the proposed work is given in fig 2.

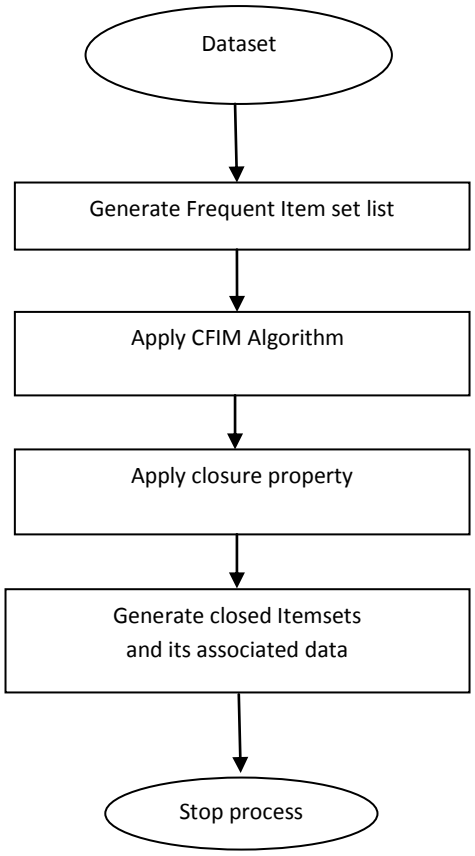

Fig 2: Closed Item set Generation

'For given data input, the proposed algorithm generates the frequent item list and applying closure property to generate closed frequent item sets.

Algorithm: Closed frequent item set mining using CFIM
Input: Transactional Database, Minimum support value.
Output: Closed frequent Item sets and its associated data.
Method:

1. Scan the transaction database to find the frequent item sets using minimum thresh old value.
2. Generate Frequent Item sets for the given datasets.
3. Mine the closed frequent item sets from the generated frequent item sets using the function CFIM (X, MIN_SUPP, and FREQ_LIST).
   CFIM (X, MIN_SUPP, FREQ_LIST).

Parameters
X- Is the item set of the transactional database?
FREQ_LIST– Frequent Item set List
Min_supp- Minimum Support count
Method:

1. X is the set of item sets in the FREQ_LIST which appear in the every transaction of Data Base.
2. Apply the closure property to find the closed frequent Item sets from the generated frequent Item sets.
3. Generates closed frequent item sets (CFI).

5. RESULT AND DISCUSSION

A data set is a set of items.It is represented in tabular form. It is roughly equivalent to a two dimensional spread sheet or data base table. The rows of a table represent the members of a data set. The columns of a table represent the features or attributes of the data items**.**We had taken the synthetic data set T1014D100K from the IBM Data set generator. It consists of 100,000 Transactions with the average length of 10 items. We found the frequent, closed item sets for CFIM algorithm and it is implemented using java programming language with weka tool. The result is performed using Intel(R) Core(TM) 2 Duo CPU with the speed of 2.80GHZ and 1.99 GB of RAM.

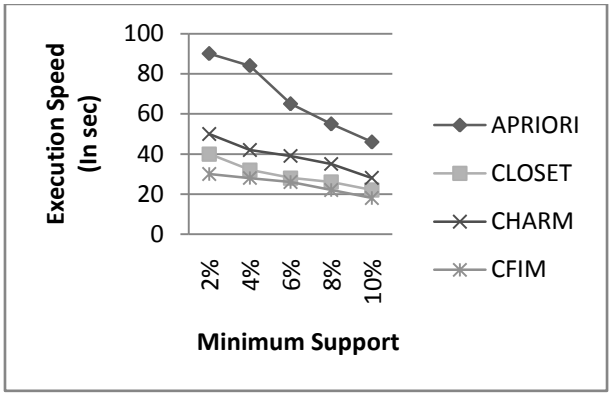

Fig 3 .Performance Analysis of algorithms

The performance analyses of algorithms are shown in the fig3. The computation speed of CFIM with other algorithms Apriori, closet and Charm algorithms are shown in the above chart. A look at the charts reveals the higher execution speed of CFIM stands stark contrast with other algorithms. And also CFIM generates the closed item sets and its associated data which makes the explicit relationship between the item sets and the data. Even if the difference in minimum is maximized, shown in fig 4.CFIM stands out as the best of other algorithms.

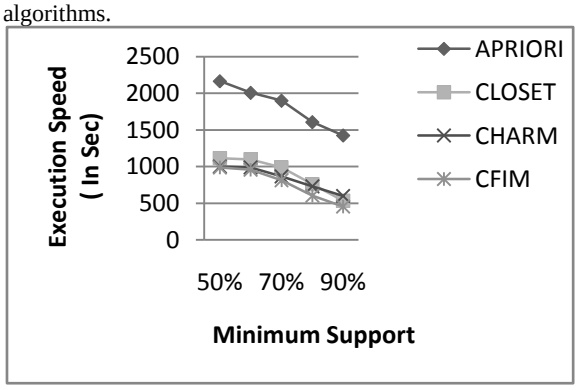

Fig 4.Performance Analysis of algorithms

6. CONCLUSION

We presented a new algorithm named as CFIM Algorithm which mines the closed frequent item sets and its associated data parallely.CFIM Algorithm which is novel algorithm, well suited for transactional databases. This algorithm is aimed at saving time in computing item set closures and their supports. CFIM produced the closed frequent item sets and its associated data is much fewer than the other algorithms which are generated. CFIM makes explicit the relationship between the item sets and its associated data. The CFIM Algorithm saving the memory space while computing closure by eliminating many levels.

7. REFERENCES

[1] Rakesh Agrawal, Tomaz Lmielinski and Arun Swami," Mining association rules between sets of items in large databases", Proc of ACM SIGMOD Conference on Management of Data, Washington, 1993.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. 1995 Int'l Conf. Data Eng. (ICDE '95), pp. 3-14, Mar. 1995.

[3] J.S.Park, M.Chen, P.S.Yu," An effective hash based algorithm for mining association rules", Proc. of ACM SIGMOD International Conference on Management of Data, May 1995.

[4] Jiawei Han, Ian Pei and Yiwen Yin,"Mining Patterns without Candidate Generation", Proc.of 2000, International Conference on Management of Data, May 16-18, 2000 Dallas, Texas, USA.

[5] S. Brin, R. Motwani, and R. Silverstein, "Beyond Market Basket: Generalizing Association Rules to Correlations," Proc. ACM-1997.

[6] "MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases", Proc. 2001 International Conference on Data Engineering (ICDE'01), pp: 443-452.

[7] R.J.Bayardo," Efficiently Mining Long Patterns from Databases", Proc. of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, Washington, United states.

[8] Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal," Efficient mining of association rules using closed itemset lattices", Information System, Vol 24, 1999.

[9] M.J. Zaki and C.-J. Hsiao, "Charm: An Efficient Algorithm for Closed Itemsets Mining," Proc. Second SIAM Int'l Conf. Data Mining, Apr. 2002.

[10] Jiawei Han and Micheline Kamber, "DATA MINING Concepts and Techniques" Elsevier Publishers, 2001 edition.

[11] A.K.C. Wong, Fellow, IEEE, and Gary C.L.Li "Simultaneous pattern and data clustering for pattern cluster analysis" IEEE Trans.Knowledge and Data Eng., vol.20, no. 7, pp. 911-923, JULY 2008.

[12] M.J. Zaki, "Mining Non-Redundant Association Rules," Data Mining and Knowledge Discovery, vol. 9, no. 3, pp. 223-248, 2004.