A Survey on Closed Frequent Pattern Mining

Prabha S
Associate Professor,
Department of IT,
K.S.Rangasamy College of Technology, Tiruchengode

Shanmugapriya S
PG Scholar,
Department of IT,
K.S.Rangasamy College of Technology, Tiruchengode

Duraiswamy K, PhD.
Dean,
K.S.Rangasamy College of Technology, Tiruchengode,
Tamilnadu, India

ABSTRACT
The identification of association rule mining has attracted many researchers. Several algorithms for effective discovery of association rule have been proposed. With the vast literature of closed frequent itemset discovery and association rule mining, still we are not able to say that we have found solution for most of the problems. This is the inspiration for my study towards the closed frequent itemsets and association rule mining. In this paper we reviewed few algorithms for closed frequent itemset and presented a comparison.

General Terms
Closed frequent itemset, Patterns, Algorithm, Candidate Generation, Projected Database, Maximal Frequent itemset, and Vertical Data Format.

Keywords
Closed Frequent Itemset, Association rule mining and Pattern

1. INTRODUCTION
A frequent pattern is defined as the itemset, subsequences or substructures that occur in a document with the frequency equal or above the user specified threshold. Frequent itemset[19, 20] are the itemset which occurred together in the given document. If the subsequences appears frequently in a document is known as frequent sequential pattern. A substructure can refer to various structural forms such as subgraphs, subtrees or sublattices. A frequent structural pattern is defined as a substructure appears frequently in a graph database. An essential need of the frequent pattern is mining associations, correlations, classification, clustering and many other relationships among data. The interesting research area in data mining is the frequent pattern mining among the other different area.

An itemset is closed if none of its immediate supersets has the same support as the itemset. The set of closed frequent patterns contains the complete information regarding to its corresponding frequent patterns.

In this survey, we perform the overview of the closed frequent pattern mining algorithm, features and problems. In section 2 it give the details about the various algorithm with the comparisons. Finally it concludes with the remarks.

2. CLOSED FREQUENT ITEMSET ALGORITHMS

2.1 A- CLOSE
The A-Close(Apriori based closed frequent itemset) [18] algorithm is works based on the apriori algorithm along with the concept of the closed itemset lattices concept. It constructs the set of candidate frequent closed patternset in a single pass over the database in a repeated iteration. Then compute the support count for the frequent closed patternset using the minimum support threshold. Finally, closed small patternset can used to construct the set of candidate frequent closed patternset for the next pass.

The Key features of the A- Close algorithm are: It uses closed itemset lattices, it reduce search space and memory consumption.

2.2 CHARM
CHARM [6, 15, 16, 21, 22, 25] stands for Closed Association Rule Mining algorithm is used to mine the closed frequent patterns. It explores patternset and didset(Document idset) space simultaneously which skips many levels quickly to identify the closed frequent patterns. It uses two pruning strategies, Candidate pruning are not only based on the subset infrequency but also branches are prunes based on non-closure property. The fundamental operation used is union of two patternset and an intersection of their document ids.

The key features of the CHARM algorithm are: It explores both itemset and didsets for quick mining of closed frequent patterns and it uses pure bottom up approaches.

2.3 CLOSET
The Closet [13, 15, 21, 22] algorithm used to mine the closed frequent patterns with the help of three techniques such as compression frequent pattern tree structure without candidate generation, Single Path compression technique and partition based projection mechanism. The algorithm initially introduces the divide and conquers method for the mining of the frequent closed patterns by arranging the patterns in a support decreasing order and then divides the search space. The subsets of the frequent closed patterns are mined by constructing corresponding conditional database and mine recursively. Once all the subsets are found, the complete set of frequent closed patterns is obtained.

The key features of the closet algorithm are: It uses FP-Tree structure for mining patterns without candidate generations, single prefix path compression technique for quick identification of patterns and partition techniques for scalable mining of the patterns.

2.4 CLOSET+
The CLOSET+ [5, 14] algorithm is used to mine closed frequent pattern. Initially, it scans the database only once to find the global frequent patterns and sort the database in support descending order and forms the frequent pattern list, scans the document and builds the FP-Tree using the pattern list, using divide and conquer technique and depth first searching paradigm it finds the closed frequent patterns. Finally, stop the process until all the patterns in the global header are mined. The frequent closed patterns are obtained either from result tree or from the output file.

The key features of the CLOSET+ algorithm are: It uses hybrid tree projection method for the conditional projected database and it uses horizontal data format.

2.5 CARPENTER

CARPENTER [7, 9, 24] stands for Closed Pattern Discovery by Transposing Tables that are Extremely Long used to mine long biological dataset. It consists of two main steps; Transpose the data into table and row enumeration tree search. In the First step, it transpose the patterns into the table named as transpose table , in that each tuple lists the feature along with the row ides feature occurs in the original table. In the second step, according to the transpose table, construct the row enumeration tree which enumerates row ids with predefined order and search the tree in depth first order without any pruning strategies. It consists of three pruning strategies, in the prune 1 method, it prunes the branch which are not having enough depth, in prune 2 method, if rj has 100% support in project table of ri , prune the branch rj where support is the depth of the node .and in prune 3 method, At any node in the enumeration tree, if the corresponding itemset of the node has been found before it prunes the branch rooted at this node.

The Key features of the CARPENTER algorithm are: It uses row enumeration search for the optimized search and it use depth first approach.

2.6 COBBLER

The COBBLER [8] algorithm abbreviated as ComBining row and coLumn enumeration used to mine the closed frequent patterns. In this algorithm, it iteratively calculates the conditional tables and conditional transposed tables in dynamic enumeration tree for performing a traversal in depth first manner. Here, each conditional table represents a feature enumerated node and each conditional transposed table represents as row enumeration node. Initially, The Frequent closed patterns will be set as empty. Based upon the switching condition it performs row enumeration or feature enumeration is performed. In the row enumeration it has three parameters such as conditional transposed table, set of rows consider for the row enumeration according to the imposed order on the rows and frequent closed patterns founded so far. In the feature enumeration it has three parameters such as conditional original table; set of features for the feature enumerations according to the imposing order of the features and frequent closed patterns founded so far. The iterative calculation stops when in the row enumeration process, the set of row enumeration will become empty or in the feature enumeration process, the set of column enumeration will become empty.

The Key Feature of the COBBLER algorithm is: It combines both the row and column enumeration and it use depth first search methods.

2.7 TD-CLOSE

The TD-Close [10] algorithm uses the Top-Down strategy and closeness checking method to mine the frequent closed patterns .Initially, it performs the transposition operation to transform original table to the transposed table and initialize the frequent closed patterns as an empty set and size of the rowset as zero. Finally, the TopDownMine is used to find the frequent closed patterns. In the TopDownMine, it takes the parameter as x-excluded transposed table, cMinsup and excludedsize where x-excluded transposed table is a table in which each tuple contains rids less than any of rids in x, and at the same time contains all of the rids greater than any of rids in x, cMinsup is a dynamically changing minimum support threshold and excludedsize is the size of the rowset.

The key features of the TD-Close algorithm are: It uses closeness checking method and uses of the Top-Down strategy.

2.8 PGMiner

The PGMiner [17, 23] is the Prefix Graph Miner which mines the frequent closed patterns. This algorithm integrates two methods such as projected database and bit vectors. Initially it projecting the document containing into nodes of a graph as similar to FP-Tree but different in the cost of traversing multiple branches of the tree to collect frequency information are low compared to the FP-Tree .Then, the projection of the nodes are encode into the bit vectors with the shorter length compare to the existing approaches. The efficiency of mining algorithm is improved by using two phase such as intra node itemset mining and inter node pruning mechanism. In the intra node itemset mining, it finds the frequent closed patterns for each node and form local closed patterns. In the inter node pruning mechanism, it checks whether the local closed patterns are also globally closed or not and finally obtain the frequent closed patterns.

The key features of the PG Miner algorithm are: It uses the projection database to collect the frequency information; it uses bit vectors format leads to fast frequency counting of patterns via intersection operations and the inter and inter node strategies used to reduces the search space.

2.9 PTclose

The PTclose [12] stands for the Patricia Tree used to mine the closed frequent patterns. In this algorithm, it uses the PTArray Technique to reduce the scanning of the patricia tree. It has two inputs such as Patricia tree and Closed Frequent Pattern tree. The Patricia tree is a compact tree, used to characterize all relevant frequency information within the document, each branch in patricia tree represents a frequent patterns and the nodes along the branches are in frequency decreasing order from root through leaves. Initially it verifies whether the patricia tree is a single path tree, if so all candidate closed Frequent patterns (CFP) are obtained from the patricia tree and candidate patterns is then compared with all the CFPs within patricia tree. If it is closed patterns it will be inserted into patricia and all CFP – tree existing in memory will be updated until Frequent closed pattern are obtained.

The key features of the PTclose algorithm are: It uses patricia tree and PTArray which reduces the time and memory consumption.

2.10 ICMiner

ICMiner [1] stands for Inter-transaction Closed patterns Miner for mining closed inter transaction patterns. It consists of two phases, in first phase scans the document to find all frequent patterns and in second phase, it constructs a pattern - dataset tree to generate the closed inter – transaction patterns in a depth first search manner. It uses two pruning strategies such as downward closure property and four properties for closed patterns. The downward closure property defines that, if a pattern is frequent then all of its sub-patterns are frequent.

The key features of the ICMiner algorithm are: It uses Depth – First search strategy and effective pruning strategies to avoid costly candidate generation and repeated support counting.

2.11 TTD-CLOSE

TTD- Close [11] algorithm mines the closed frequent pattern based on the trace based top down mining strategy. Initially ,

it generate the transposed table and transformed it into the FR-Tree(Frequent Rowset Tree) with the supporting structure named as IP- List (Itemset Pointer List) to keep information for the recursion node. In IP-List, the explicit rowset is assigned the complete set of r ids and implicit rowset are assigned as empty. During the iterations, any patternset in a certain IP-List must contain all the r ids in the corresponding implicit rowset and may contain some rids in the explicit rowset, excluding any rid outside these rowset for guaranteed the complete traversal of all the subset based on the row enumeration tree.

The key features of the TTD- Close algorithm are: It uses row enumeration strategy and it uses top down strategy for minimize the number of scanning and memory usage.

2.12 PCP-Miner

The PCP-Miner [2] is abbreviated from Pointset Closed Pattern Miner which is used to mine the frequent closed patterns in a pointset database. The algorithm consists of two phases. In first phase, it generates frequent length 2 – patterns and for each frequent pattern it engenders the projected database. In second phase, it generate the frequent (k+1) – patterns by joining k-patterns in a joinable class by depth first manner iteratively. The joinable class is used to localize the support counting, candidate pruning and pattern joining in a smallest of projected databases. During the enumeration process, various pruning strategies are applied to prune impossible candidates and remove the frequent but non-closed patterns.

The key features of the PCP-Miner are: It finds closed patterns using pointset database and it follows depth first search strategy.

2.13 CFIM-P

CFIM-P [3, 4] algorithm stands for Closed Frequent Itemset Mining and Pruning algorithm for mining the closed frequent patterns. The algorithm consists of 3 phases. In the first phase, it traces the null document and filters them for ensuing mining procedures. In the second phase, it mines the closed frequent pattern based on the minimum support count. If the already mined superset exists for the subset of frequent pattern then subset is eliminated by the top down manner. After obtaining the closed frequent itemset, it is added to the list of frequent itemset. In the third phase, the mined closed frequent itemset constitute to form patterns.

The key features of the CFIM-P algorithm are: It uses Top down strategy and it eliminates the null transaction before starts the mining process.

Table 1 Comparison of the Closed Frequent Itemset Algorithms

S.No	Algorithm Name	Author Name	Year	Used For Which Type	Advantage	Disadvantage
1.	A-CLOSE (Apriori based closed frequent itemset)	Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal	1999	Closed Frequent Itemset	• Reduced set of association rules without having to determine all frequent itemsets, thus lowering computation costs.	• Costly when mining long patterns or with low minimum support thresholds in large databases
2.	CHARM (Closed Association Rule Mining)	Mohammed Javeed Zaki, Ching-Jiu Hsiao	1999	Closed Frequent Itemset	• Use a novel search method that skips many levels to quickly identify the closed frequent itemsets.	• It does not work well in higher support of order of magnitude.
3.	CLOSET	Jian Pei, Jiawei Han, Runying Mao	2000	Closed Frequent Itemset	• Computes a much smaller set of candidates and also employs a compact representation of association rules. • Algorithm is based on memorization mechanism which avoids redundant computations.	• It does not work well in lower support of order of magnitude.
4.	CLOSET+	Wang J, Han J, Pei J	2003	Closed Frequent Itemset	• Improvement in terms of runtime, memory usage and scalability	• Work well for datasets with small average row length,
5.	CARPENTER	Pan F, Cong G, Tung AKH, Yang J, Zaki M	2003	Closed Frequent Itemset	• Very efficient in finding frequent closed patterns on datasets with small number of rows and large number of features.	• Encounters problem for datasets that have large number of rows and features
6.	COBBLER (Combining Column and Row	Pan F, Tung AKH, Cong G, Xu X	2004	Closed Frequent Itemset	• Dynamically switch between row and feature enumeration for frequent closed pattern	• It cannot make full use of the minimum support threshold to prune search space. As a result, experiments

	Enumeration)				discovery. • Automatically select an enumeration method according to the characteristics of the datasets before and during the enumeration.	show that it often cannot run to completion in a reasonable time for large microarray data, and it sometimes runs out of memory before completion.
7.	TD-CLOSE (Top down frequent pattern)	Liu H, Han J, Xin D, Shao Z	2006	Closed Frequent Itemset	• More efficient and uses less memory • Produces accurate closed patterns	• More time is needed when the minsup becomes smaller. • More influenced by the number of tuples increase
8.	PGMiner (Prefix Graph Miner)	H. D. K. Moonesinghe, Samah Fodeh, Pang-Ning Tan	2006	Frequent Closed Itemsets	• Faster frequency counting of itemsets through intersection operations. • Reduces the search space. • Reduces the memory usage especially for low support thresholds.	• Memory consumption gets high when the threshold is gradually lowered.
9.	PTclose (Patricia tree closed frequent itemset)	J. Tahmores Nezhad, M.H.Sadreddini	2007	Frequent Closed Itemsets	• Suits both dense and sparse datasets. • Relatively low response time and memory consumption. • Shorter time to scan the patterns	• Patricia tree condenses one-element nodes into their parent node. Condensed nodes contain keys that are no longer uniquely specified by their search path. • when nodes are made to store pointers to the key they contain, complexity is introduced
10.	ICMiner (Inter-transaction Closed patterns Miner)	Anthony J.T.Lee, Chun-Sheng Wang, Wan-Yu Weng, Yi-An Chen, Huei-Wen Wu	2008	Closed Inter Transaction Itemsets	• Avoids costlier generation of candidate and repeated support counting	• Extra memory space to store candidate inter-transaction patterns
11.	TTD-Close (Trace-Based Top down frequent pattern)	Hongyan Liu , Xiaoyu Wang , Jun He, Jiawei Han , Dong Xin , Zheng Shao	2009	Frequent Closed Pattern Mining	• Reduces search and memory space.	• More time is needed when the minsup becomes smaller. • More influenced by the number of tuples increases.
12.	PCP-Miner (Pointset Closed Pattern Miner)	Anthony J.T.Lee, Wen-KwangTsao, Po-YinChen,Ming-ChihLin,Shih-HuiYang	2010	Frequent Closed Pattern Mining	• It is more efficient and scalable. • Reduce the number of candidate patterns and eliminate non- closed patterns.	• The algorithm is a memory-based algorithm. When the dataset or the number of frequent patterns is getting larger and larger, it may not be able to be loaded into main memory. • If the pattern can be represented as a bit string, the cost of pattern joins and subsumption checks.
13.	CFIM-P (Closed Frequent	Binesh Nair, Amiya Kumar Tripathy	2011	Closed Frequent Itemset	• Eliminates the redundant patterns • Reduces the processing	• More time is needed when the minsup becomes smaller

	Itemset Mining and Pruning)			Mining	time • Minimize the depth of the tree, without losing any important information. • Vertical data representation technique for frequent itemsets, identifies the null transactions and filters them for the subsequent mining process	• Less efficient for small itemsets.

3. CONCLUSION

Over the decade of year a tremendous numbers of the researches are developed in this domain. In this paper, we present a brief outline and analyses the assessment of the various closed frequent pattern mining algorithms. The comparison table shows the advantages and disadvantages of different closed frequent pattern mining algorithms. The complete coverage on the closed frequent pattern mining is not possible with limited space and knowledge. Short overview may give a rough outline of the recent work of the field. However, in-depth research is needed on several critical issues in various data mining applications.

4. REFERENCES

[1]. Anthony J.T. Lee , Chun-Sheng Wang, Wan-Yu Weng, Yi-An Chen and Huei-Wen Wu,2008, "An efficient algorithm for mining closed inter-transaction itemsets", Data & Knowledge Engineering, Vol.66, pp.68–91.

[2]. Anthony J.T.Lee, Wen-KwangTsao, Po-YinChen, Ming-ChihLin and Shih-HuiYang, 2010, "Mining frequent closed patterns in pointset databases", Information Systems, Vol.35, pp.335–351.

[3]. Binesh Nair, Amiya Kumar Tripathy, 2011, "Accelerating Closed Frequent Itemset Mining by Elimination of Null Transactions", Journal of Emerging Trends in Computing and Information Sciences, Vol. 2, No.7.

[4]. Binesh Nair, and Amiya Kumar Tripathy, 2011, "Optimizing Frequent Pattern Mining through Elimination of Null Transactions", Computational Intelligence and Information Technology, vol. 250, pp 518-523.

[5]. Claudio Lucchese, Salvatore Orlando, Raffaele Perego, 2006, "Fast and Memory Efficient Mining of Frequent Closed Itemsets", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, pp. 21-36.

[6]. El Far M. Moumoun L., Gadi, T., and Benslimane R., 2010, "An efficient CHARM algorithm for indexation 2D/3D and selection of characteristic views" , Proceeding of 5th International Symposium on I/V Communications and Mobile Network (ISVC), pp.1-4.

[7]. Feng Pan, Gao Cong, Anthony K. H. Tung, Jiong Yang and Mohammed J. Zaki, 2003, "CARPENTER: Finding Closed Patterns in Long Biological Datasets, in Proceedings of the SIGKDD '03.

[8]. Feng Pan, Gao Cong, Xu Xin and Anthony K. H. Tung, 2004, "COBBLER: Combining Column and Row Enumeration for Closed Pattern Discovery", In Proceeding of International Conference on Scientific and Statistical Database Management.

[9]. Gao Cong, Kian-Lee Tan, Anthony K.H. Tung and Feng Pan, 2004, "Mining Frequent Closed Patterns in Microarray Data", Proceedings of the Fourth IEEE International Conference on Data Mining, pp. 363-366.

[10]. Hongyan Liu, Jiawei Han, Dong Xin and Zheng Shao, 2006, "Mining Frequent Patterns from Very High Dimensional Data: A Top-Down Row Enumeration Approach", pp.280-291.

[11]. Hongyan Liu, Xiaoyu Wang, Jun He, Jiawei Han, Dong Xin and Zheng Shao, 2009, "Top-down mining of frequent closed patterns from very high dimensional data, Information Sciences, Vol.179, pp. 899–924.

[12]. J. Tahmores Nezhad and M.H.Sadreddini, 2007, "PTclose: A novel algorithm for generation of closed frequent itemsets from dense and sparse datasets", in Proceedings of the World Congress on Engineering, Vol.1.

[13]. Jian Pei, Jiawei Han and Runying Mao, 2000, "CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets", in Proceeding of the ACM-SIGMOD, pp. 11–20.

[14]. Jianyong Wang, Jiawei Han and Jian Pei, 2003, "CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets", in proceeding of the SIGKDD '03.

[15]. Jiawei Han, Jianyong Wang, Ying Lu, and Petre Tzvetkov, 2002, "Mining Top-K Frequent Closed Patterns without Minimum Support" , Proceedings of the IEEE International Conference on Data Mining, p. 211.

[16]. Mohammed J. Zaki and Ching-Jui Hsiao, 1999, "CHARM: An Efficient Algorithm for Closed Association Rule Mining", Technical Report 99-10., Computer Science Dept., Rensselaer Polytechnic.

[17]. Moonesinghe H.D.K., 2006, Samah Fodeh and Pang-Ning Tan, "Frequent Closed Itemset Mining Using Prefix Graphs with an Efficient Flow-Based Pruning Strategy".

[18]. Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal, 1999, "Efficient Mining of association rules using closed itemset lattices", Information Systems, vol.24, No.1, pp .25-46.

[19]. Rakesh Agrawal and Ramakrishnan Srikant, 1994, "Fast Algorithms for Mining Association Rules", in Proceedings of the 20th VLDB Conference.

[20]. Rakesh Agrawal, Tomasz Imielinski and Arun Swami, 1993," Mining Association Rules between Sets of Items in Large Databases", in Proceedings of the ACM SIGMOD Conference.

[21]. Wang J., Han J., and Li C., 2005, " Frequent Closed Sequence Mining without Candidate Maintenance", IEEE Transactions on Knowledge and Data Engineering, vol.19, No.8, pp.1042-1056.

[22]. Yan, X., Han, J., and Afshar, R., 2003, "CloSpan: Mining closed sequential patterns in large datasets", In Third SIAM International Conference on Data Mining (SDM), San Fransico, CA, pp. 166–177.

[23]. Yi Pan, and HongYan Du, 2011, "A Novel Prefix Graph Based Closed Frequent Itemsets Mining Algorithm", proceeding of IEEE International Conference on Computational Science and Engineering, pp. 627-631.

[24]. YuQing Miao, GuoLiang Chen, Bin Song, and ZhiHao Wang, 2006, "TP+Close: Mining Frequent Closed Patterns in Gene Expression Datasets", Data Mining and Bioinformatics ,Vol. 4316, pp 120-130 .

[25]. Zaki M.J., and Hsiao C.J., 2005, "Efficient algorithms for mining closed itemsets and their lattice structure", IEEE Transactions on Knowledge and Data Engineering, vol.17, No.4, pp.462 – 478.