A SURVEY ON LATTICE STRUCTURE WITH CHARM-L TO IMPROVE EFFICIENCY OF GENERATING FREQUENTLY CLOSED ITEM SETS

[1] Bhumi N Patel,[2]Mr. Ishan K Rajani
[1]PG Student,[2]Assistant Prof.,
[1]Computer Department,
[1]Silver Oak College of Engineering and Technology,Ahmedabad,India

Abstract—Efficient algorithms [1] for mining frequent item-sets are crucial for mining association rules as well as for many other data mining tasks. It is well known that countTable is one of the most important facility to employ subsets property for compressing the transaction database to new lower representation of occurrences items. One of the biggest problem in this technique is the cost of candidate generation and test processing which are the two most important steps to find association rules. In this paper, we have developed this method to avoid the costly candidate-generation-and-test processing completely. Moreover, the proposed methods also compress crucial information about all itemsets, maximal length frequent itemsets, minimal length frequent itemsets, avoid expensive, and repeated database scans. The proposed named CountTableFI andBinaryCountTableF are presented, the algorithm has significant difference from the Apriori and all other algorithms extended from Apriori. The idea behind this algorithm is in the representation of the transactions,where, we represent all transactions in binary number and decimal number, so it is simple and fast to use subset and identical set properties. A comprehensive performance study shows that our techniques are efficient and scalable comparing with other methods.

Index Terms—Association rule mining, Frequent, Apriori,Count table, Efficient

I. INTRODUCTION

Data stored in computer increases very fast so we have to try to extract useful knowledge from this data, this process is called knowledge discovery in Database (KDD),or data mining.Knowledge discovery in databases(DBs) is important for many technical, social, and economic problems. Modern DBs contain such a huge quantity of information that it is practically impossible to analyze this information manually for acquiring valuable knowledge for decisions making. Its main algorithm, APRIORI, not only influenced the association rule mining community, but it affected other data mining fields as well.The depth first algorithm is a simple algorithm that proceeds as follows. After some preprocessing, which in-volves reading the database and a sorting of the single items with respect to their support,builds a trie in memory,where every path from the root downwards corresponds to a unique frequent itemset; in consecutive steps items are added to this trie one at a time. Both the database and the trie are kept in main memory, which might cause memory problems: both are usually very large, and in particular the trie gets much larger as the support threshold decreases. Finally the algorithm outputs all paths in the trie, i.e., all frequent itemsets.An extensive set of experiments confirms that CHARM provides orders of magnitude improvement over existing methods for mining closed itemsets, even over methods like AClose,that are specifically designed to mine closed itemsets. It makes a lot fewer database scans than the longest closed frequent set found, and it scales linearly in the number of transactions and also is also linear in the number of closed itemsets found.

II. LITERATURE SURVEY

Marghny H. Mohamed & Mohammed M. Darwieesh developed Efficient algorithms[1] for mining frequent item-sets are crucial for mining association rules as well as for many other data mining tasks.In this paper, we have developed this method to avoid the costly candidate-generation-and-test processing completely. Moreover, the proposed methods also compress crucial information about all itemsets, maximal length frequent itemsets, minimal length frequent itemsets, avoid expensive, and repeated database scans. The proposed named CountTableFI and BinaryCountTableFI are presented, the algorithm has significant difference from the Apriori and all other algorithms extended from Apriori. The idea behind this algorithm is in the representation of the transactions,where, we represent all transactions in binary number and decimal number, so it is simple and fast to use subset and identical set properties.It construct a highly compact count table,which is usually substantially smaller than the original database and discover frequent itemsets with Intersection And operation is faster than the traditional item comparing method used in many Apriori-like algorithm.

Ferenc Bodon said that in his research the efficiency of frequent itemset mining algorithms[2] is determined mainly by three factors: the way candidates are generated, the data structure that is used and the implementation details. Most papers focus on the first factor, some describe the underlying data structures, but implementation details are almost always neglected. In this paper we show that the effect of implementation can be more important than the selection of the algorithm. In this paper we describe an implementation of APRIORI that outperforms all implementations known to us. We analyze, the oretically and experimentally, the principal data structure of our solution. This data structure is the main factor in the efficiency of our implementation. Moreover, we present a simple modification of APRIORI that appears to be faster than the original algorithm.In this paper, we showed that different implementation results in different running time, and the differences can exceed differences between algorithms. We presented an implementation that solved frequent itemset mining problem in most cases faster than other well-known implementations.

Walter A. Kosters and Wim Pijls have discussed DF[3],the depth first implementation of APRIORI as devised in 1999.Given a database,this algorithm builds a trie in memory that contains All frequent itemsets, i.e., all sets that are contained in at least minsup transactions from the

original database. Here min-sup is a threshold value given in advance. In the trie, that is constructed by adding one item at a time, every path corresponds to a unique frequent itemset. We describe the algorithm in detail, derive theoretical formulas, and provide experiments.In this paper, we addressed DF, a depth first implementation of APRIORI.To our experience,DF competes with any other well-known algorithm, especially when applied to large databases with transactions.

Claudio Lucchese and Salvatore Orlando have said One of the main problems raising up in the Frequent closed itemsets mining problem[4] is the duplicate detection.In this paper we propose a general technique for promptly detecting and discarding duplicate closed itemsets, without the need of keeping in the main memory the whole set of closed patterns. Our approach can be exploited with substantial performance benefits by any algorithm that adopts a vertical representation of the dataset. We implemented our technique with in a new depth-first closed itemsets mining algorithm.In this paper we provide a deep study on the problem of mining frequent closed itemsets, formalizing a general framework fitting every mining algorithm. Use such framework we were able to analyse the problem of duplicates rising in this new mining problem.

Mohammed J. Zaki and Ching-Jui Hsiao have said that the task of mining association rules[5] consists of two main steps. The first involves finding the set of all frequent itemsets. The second step involves testing and generating all high confidence rules among itemsets. It is also not necessary to mine the set of all possible rules. We show that any rule between itemsets is equivalent to some rule between closed itemsets. Thus many redundant rules can be eliminated. Furthermore, we present CHARM, an efficient algorithm for mining all closed frequent itemsets. In this paper we presented and evaluated CHARM, an efficient algorithm for mining closed frequent itemsets in large dense databases. CHARM is unique in that it simultaneously explores both the itemset space and tidset space, unlike all previous association mining methods which only exploit the itemset space. The exploration of both the itemset and tidset space allows CHARM to use a novel search method that skips many levels to quickly identify the closed frequent itemsets, instead of having to enumerate many non-closed subsets.

Sujatha Dandu, B.L.Deekshatulu & Priti Chandra have said in research paperFrequent itemset mining plays an important role in association rule mining. The Apriori & FP-growth algorithms are the most famous algorithms [7] which have their own shortcomings such as space complexity of the former and time complexity of the latter. Many existing algorithms are almost improved based on the two algorithms and one such is APFT [11], which combines the Apriori algorithm and FP-tree structure of FP-growth algorithm . The advantage of APFT is that it doesn't generate conditional & sub conditional patterns of the tree recursively and the results of the experiment show that it works fasts than Apriori and almost as fast as FP-growth. We have proposed to go one step further & modify the APFT to include correlated items & trim the non-correlated itemsets. This additional feature optimizes the FP-tree & removes loosely associated items from the frequent itemsets. We choose to call this method as APFTC method which is APFT with correlation. APFTC performs as expected proving to be efficient in time consumed and also in retrieving the most correlated itemsets.

Table 2.1: Comparison table

Sr. no.	Description	Approach	Pros.	Cons.
1	Efficient mining frequent itemsets algorithms	Solving the candidate itemsets generation and support count problems	No candidate set generation	Problem of computationally complexity of frequent itemsets
2	A fast APRIORI implementation	Determining Support with a Trie	Candidate generation becomes easy and fast	Data analysis can be difficult
3	APRIORI, A Depth First Implementation	Depth First Method	straightforward, handy, straight-forward and quick calculation for discovering all regular itemsets	Memory limitations
4	DCI Closed: a Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets	Search spacebrowsing and closure calculation	Requires orders of magnitude less memory	Lengthy calculation
5	CHARM: An Efficient Algorithm for Closed Association Rule Mining	Hash based approach	It stores all hash tables in main memory.	Leads memory inefficient problm because it stores all itemsets in memory if they are closed or nonclosed.

III. CONCLUSION

CHARM- L can give requests of size change over existing techniques for mining shut itemsets.CHARM - L is a cutting edge calculation that creates the continuous shut itemset cross section.These calculations all the while investigate both the itemset space and tidset space utilizing the new IT-tree structure, which permits a novel hunt strategy that skips many levels to rapidly distinguish the shut regular itemsets, rather

than enumerating numerous nonclosed subsets.Moreover, since we pruned nonclosed item sets timely, we reduced the search space. And with some optimized operations that reduce work and time, our algorithm also runs faster.

IV. REFERENCES

[1] Marghny H. Mohamed & Mohammed M. Darwieesh,Int. J. Mach. Learn. & Cyber, DOI 10.1007/s13042-013-0172-6, Springer-Verlag Berlin Heidelberg 2013

[2] Ferenc Bodon∗Informatics Laboratory , Computer and Automation Research Institute ,Hungarian Academy of SciencesH -1111 Budapest, Lagym´anyosi u. 11, Hungary, Research supported in part by OTKA grants T42706, T42481 and the EU-COE Grant of MTA SZTAKI

[3] Walter A. Kosters ,Leiden Institute of Advanced Computer Science ,Universiteit Leiden ,P.O. Box 9512, 2300 RA Leiden,The Netherlands ,kosters@liacs.nl & Wim Pijls, Department of Computer Science, Erasmus University ,P.O. Box 1738, 3000 DR Rotterdam, The Netherlands ,pijls@few.eur.nl

[4] Claudio Lucchese,ISTI "A. Faedo",Consiglio Nazionale delle Ricerche (CNR),Pisa, Italy,email:claudio.lucchese@isti.cnr.it & Salvatore Orlando,Computer Science Dept.,Universit`a Ca' Foscari,Venezia, Italy.email:orlando@dsi.unive.it &Raffaele Perego ISTI "A. Faedo",Consiglio Nazionale delle Ricerche (CNR)Pisa, Italy. email:raffaere.perego@isti.cnr.it

[5] Mohammed J. Zaki and Ching-Jui Hsiao,Computer Science Department, RensselaerPolytechnicInstitute,TroyNY12180,zaki,hsiaoc@cs.rpi.edu,http://www.cs.rpi.edu/zai

[6] Han J, Kamber M (2006) Data mining: concepts and techniques,2nd edn. Morgan Kaufmann, San Francisco

[7] Sujatha Dandu, B.L.Deekshatulu & Priti ChandraAurora's Technological and Research Institute, Hyderabad, India, Volume 13 Issue 2 Version 1.0 Year 2013, Type: Double Blind Peer Reviewed International Research Journal,Publisher: Global Journals Inc. (USA),Online ISSN: 0975-4172& Print ISSN: 0975-4350