ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2014

A Survey of Frequent and Infrequent Weighted Itemset Mining Approaches

J.Jaya[1], S.V.Hemalatha[2]

PG Scholar, Dept. of CSE, Kalaignar Karunanidhi institute of Technology, Coimbatore, India[1]

Asst Prof, Dept of CSE, Kalaignar Karunanidhi institute of Technology, Coimbatore, India[2]

ABSTRACT:Itemset mining is a data mining method extensively used for learning important correlations among data. Initially itemsets mining was made on discovering frequent itemsets. Frequent weighted item set characterizes data in which items may weight differently through frequent correlations in data's. But, in some situations, for instance certain cost functions need to be minimized for determining rare data correlations. Determining these types of data is more challenge and interesting research than mining frequent data in items. This paper surveys various methods for frequent itemset and infrequent item set mining of data. This work differentiates various methods with each other during mining of data. Finally, comparative measures of each method are presented which provides the significance and limitations of frequent and infrequent mining of data in itemsets.

KEYWORDS: Clustering, association rules, frequent itemset mining, infrequent itemset mining.

I. INTRODUCTION

Data are any facts, numbers or text that can be processed by computer. The patterns, associations or the relationship among this data can provide information. Information can be converted into knowledge about historical patterns and future trends.

Data Mining is the process of finding correlation or patterns among dozens of fields in large relational databases. It is to extract interesting information or patterns from data in large databases. Data mining is the procedure for discovering data from different viewpoints and summarizing it into valuable information. This information can be used to improve costs and profits of data information or both. Data mining is processed with the great deal of consideration in the information construction and in society recently, because of the extensive preventability of huge amounts of data and the future necessitate for figuring such data into practical information and acquaintance. Data mining finds its application mainly on Market basket analysis, Risk analysis, Fraud Detection, DNA data analysis, Web Mining…etc.

Association rule mining is research topic in data mining and has numerous application. It depicts the implicit relationship among the data attributes. The extraction of interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories is the main objective of Association rule mining. Association rule mining extracts interesting correlation and relation between large volumes of transactions.

This process is divided into two phases. First phase is itemset mining. Second phase is rules construction. Itemset mining was focused on discovering frequent itemset, i.e., patterns whose observed frequency of occurrence in the source data (the support) is above a given threshold. Itemset below the threshold value is referred as Infrequent itemset. Frequent itemsets mining is a central part of data mining and distinctions of association examination, namely association-rule mining and sequential-pattern mining respectively. From large amount of data, frequent itemset are constructed by concerning some rules or association rule mining algorithms to calculate all the frequent itemsets. In many association investigation methods, frequent itemset extraction is considered as a primary step. An itemset is named as frequent if it is available in a large-enough part of the dataset. This frequent occurrence of item is represented by means of the count of support. Consequently, it requires complex techniques for hiding or restructuring users' private information through a data construction process. Furthermore, this technique does not yield the accuracy of mining results. Discovering such frequent pattern is termed as a significant position in mining relations, correlations, and several other relationships among data. In addition, it is used in data clustering, data classification and various other data mining techniques respectively.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2014

However, considerably less consideration has been noticed to mining of infrequent itemsets, even though it has obtained major usage in mining of negative association rules from infrequent itemsets, statistical disclosure risk measurement whereas exceptional patterns in anonymous sample data can direct to statistical disclosure. Then infrequent itemsets is adapted to fraud detection whereas uncommon patterns in financial or tax data might imply unusual action associated with fraudulent behaviour and then applied in the field of bioinformatics where unusual patterns in microarray data could imply genetic disorders. Patterns that are rarely established in database are frequently measured to be irrelevant and are eliminated using the support assessment. Such patterns are named as infrequent patterns. Mining infrequent patterns is a challenging attempt since there is huge number of such patterns that can be incorporated from a well-known data set. Generally, the primary issues in infrequent patterns mining are identification of appropriate infrequent patterns and efficiently discovering such patterns in large data sets.
The following work describes the literature of various methods used for mining frequent and infrequent itemsets respectively

II. RELATED WORK

TECHNIQUES USED FOR FREQUENT ITEMSET MINING

Uniform Distribution of items

In[1]RAgarwal introduces Frequent itemset mining which is widely used data mining technique. Here, the rules are framed based on the itemset mined which is said to be frequent. Those itemset satisfying minimum support and confidence are taken as frequent and is used for framing association rules. Most approaches to association rule mining assume that all items within a dataset have a uniform distribution with respect to support. The main problem with this is items in a transaction are treated equally.

Significance of item

In[2]W.Wang introduces the concept of weight to be assigned for item in each transaction which reflects the intensity or the importance of the item within the transaction. The main problem with this is that weights are introduced only during the rule generation step not used for the mining purposes.

Weighted Association Rule Mining

In[3]Feng Tao et.al presents Weighted Association Rule Mining for frequent itemset mining. In this work the limitation of the conventional Association Rule Mining model is avoided specifically its inability for treating units differently. The presented method uses weights which can be incorporated in the mining process to resolve this difficulty. Then the challenge is solved when doing enhancement towards using weight, especially the invalidation of downward closure property. In order to adapt weighting in the new setting, a set of new concepts are used. With this weighted downward closure term is used as a substitute of the unique downward closure property. At last this method is confirmed as suitable and gives reason for the efficient mining scheme in the new construction of weighted support. By learning the simulation of the lattice building, solution is suggested that weight can be utilized to guide the mining focus to those significant itemsets with high degree of consequence. *Transaction weight* is a type of itemset weight. It is a value attached to each of the transactions. Usually the higher a transaction weight, the more it contributes to the mining result. However weights are to be priorly assigned which is difficult in real life cases.

Data trimming framework

In[4]data trimming framework is presented for mining frequent itemsets from uncertain data under a probabilistic framework. This method uses the U-Apriori algorithm, which is a customized part of the Apriori algorithm, to process on various datasets. Then the computational problem of U-Apriori is identified by using a data mining technique. Then LGS-Trimming method is used under the framework and confirmed, by widespread experiments, that it attains very high performance gain by means of Input /output cost and computational cost. In contrast to U-Apriori, LGS-Trimming process well on datasets with increased percentage of low probability items.

W-support mechanism

In[5]Ke Sun and FengshanBai presented novel framework of w-support mechanism in association rule mining. Initially, the HITS model and algorithm are utilized to obtain the weights of transactions from a database record with simply binary attributes. By derived from these weights, a novel assessment of w-support is described to provide the consequence of item sets. However the presented method differs from the conventional support in taking the quality of transactions into account. Then, the w-confidence and w-support of association rules are described in similarity to the description of confidence and support. Then an Apriori-like algorithm is presented to extract association rules whereas

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2014

w-confidence and w-support are resulted above fixed thresholds in nature.The analyzed data set is represented by means of Bipartite graph in order to automate item weight assignment.

Probabilistic Frequent Itemset Mining

In [6]C.K.Chui addresses the issue of relating the weight to probability of occurrence but in most cases both of them are uncorrelated. For example an item which is very likely to occur in a transaction may seem to be of least importance. In[7]Thomas Bernecker et.al presented Probabilistic Frequent Itemset Mining for mining uncertain transactional databases. This probabilistic method brings new probabilistic mechanism of frequent itemset which is based on probable world semantics. In this probabilistic circumstance, an itemset is said to be frequent if the probability that itemset happens in at least minSup transactions is higher than a given threshold. Considerably this is the said to be first method deals with the problem under probable world's semantics. In addition to the probabilistic mechanisms, a framework is presented in which it has the capability to solve the Probabilistic Frequent Itemset Mining (PFIM) problem proficiently.

Frequent pattern tree (FP-tree) structure

In[8] Jiawei Han et.al presented novel frequent pattern tree (FP-tree) structure, which is an widened prefix-tree construction for storing compressed, critical information about frequent patterns, and expands an effective FP-tree-based mining system, FP-growth, for mining the absolute set of frequent patterns by pattern fragment growth. Effectiveness of mining is attained with three methods:1) a huge database is compressed into a largely reduced. 2)the presented FP-tree-based mining approves a pattern fragment growth process to eliminate the costly generation of a huge number of candidate sets. 3)Finally a partition-based method known as divide-and-conquer system is used to divide the mining job into a set of minor tasks for mining detained patterns in conditional databases, where the search space is reduced appropriately

TECHNIQUES USED FOR INFREQUENT ITEM MINING

Positive and Negative Association rule

In[9]X.wuEfficient mining of both positive and negative association rules . They focus on identifying the associations among frequent itemsets. They designed a new method for efficiently mining both positive and negative association rules in databases. This approach is novel and different from existing research efforts on association analysis. Some infrequent itemsets are of interest in this method but not in existing research efforts. They had also designed constraints for reducing the search space, and had used the increasing degree of the conditional probability relative to the prior probability to estimate the confidence of positive and negative association rules.

Minimal infrequent itemset mining

In[10]David et.al presented a new algorithm of MINIT, for finding minimal τ -infrequent or minimal τ -concurrent item sets. Firstly, a ranking of items is organized by estimating the need of each of the items and then generating a record of items in rising order of support. Minimal τ –infrequent itemsets are determined by using each item in rank order, iteratively calling MINIT on the maintained set of the dataset with regard to items using only those items with superior rank than current items , after that checking each candidate of minimal infrequent items (MII) against the original dataset is performed. A system that can be utilized to judge only superior-ranking items in the iteration is to preserve a "liveness" vector representing which items stay feasible at each level of the iteration

Rare Association Rules generation

In [11]Laszlo et.al presented generation of rare association rules for mining of infrequent itemsets. This work presented a method to taking out rare association rules that stay hidden for traditional frequent itemset mining algorithms. When compared with other method the presented method finds strong but rare associations that are local regularities in the data are found. These rules are said to be "mRI rules" .Apriori computes the support of minimal rare itemsets (mRIs), i.e. rare itemsets such that all proper subsets are frequent. Instead of pruning the mRIs, they are retained. In addition, it is shown that the mRIs form a generator set of rare itemsets, i.e. all rare itemsets can be restored from the set of mRIs which have two merits. Initially, they are highly informative in the case that they have an antecedent which is a producer itemset while adding up the resultant to give ways for a closed itemset. Secondly, the amount of these rules is minimal, that is the mRG rules comprise a dense illustration of all largely confident associations that can be taken from the least rare itemsets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2014

Pattern-Growth Paradigm and Residual Trees

In[12]Ashish Gupta e.al presented pattern-growth paradigm to discover minimally infrequent itemsets.They recommend a new algorithm based on the pattern-growth paradigm to find minimally infrequent itemsets. It has no subset which is also infrequent. This work uses novel algorithm of IFP min for mining minimally infrequent itemsets. Then the residual tree concept has been incorporated by using a variant of the FP-Tree structure which is known as inverse FP-tree. In order to mine the minimally infrequent itemsets, optimization of Apriori algorithm is performed. Finally the presented tree are used for mining of frequent itemset as well

Optimization rule based algorithm

In [13] NikkySuryawanshiRai et.al presented a new algorithm for optimization of association rule mining. This method determines the crisis of negative rule generation and as well as optimized the method of rule generation. This method used a multi-level multiple support of data table as binary values of 0 and 1. The divided process minimizes the examining time of database. The presented method works in the combination of genetic algorithm and MLMS. An algorithm of MIPNAR_GA has been presented for mining interesting negative and positive rule from infrequent and frequent pattern sets. The algorithm is proficient in to three stages: First it extracts frequent and infrequent pattern sets by incorporating apriori approach. Secondly positive and negative rule are generated. And finally prune redundant rule has been applied for interest measurements.

Confabulation-Inspired Association Rule Mining

In [14] AzadehSoltani&Akbarzadeh presented confabulation-inspired association rule mining (CARM) algorithm for mining frequent and infrequent item sets. CARM is motivated by the method of idea in the human brain, and particularly the theory of confabulation for mining association rules. The presented algorithm holds two phases of knowledge attainment and rule extraction. Knowledge attainment holds two modules whereas the axonal association links between these two modules are made to archive all domain knowledge. The second phase of rule extraction is then executed derived from the weight age of these communication links.

III.COMPARATIVE TABLE

S.No	Author and year	Category	Techniques	Merits	De-merits
1	W.Wang,J.yang,P.S.Yu 2000	Frequent itemset mining	Significance of item	Reflects the significance of item	Weight not used for mining process
2	Feng Tao, FionnMurtagh, Mohsen Farid-2003	Frequent itemset mining	Weighted Association Rule Mining	Scalable and efficient in discovering significant relationship in weighted terms	Difficult to find a generic mechanism to determine the relaxation factor
3	Jiawei Han, Jian Pei, and Yiwen Yin-2000	Frequent itemset mining	Frequent pattern tree (FP-tree) structure	Efficient and scalable result is achieved	High computational cost is required
4	David J. Haglin and Anna M. Manning-2007	Infrequent itemset mining	Minimal infrequent itemset mining	Better performance is obtained	Improved running time is not observed
5	Ashish Gupta, Akshay Mittal, Arnab Bhattacharya-2011	Infrequent itemset mining	Pattern-Growth Paradigm and Residual Trees	Improved performance is obtained with less computational time	Better scalability is not achieved for mining maximally frequent itemsets

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2014

IV. CONCLUSION

The present work surveys various methods for mining frequent and infrequent item sets of data. The presented work surveys different viewpoint on different types of interesting frequent and infrequent patterns. The related concepts of positive and negative correlated pattern and its association rules are mined. The current survey makes review on several papers related to frequent and infrequent patterns and in addition rare item sets and also offers the knowledge on different algorithms presented for mining infrequent patterns. The major advantage for mining infrequent itemset was to advance the profit of rarely originated datasets in the transactions. The first effort is to discover the frequent item set mining and then determine the infrequent weighted item sets. Merits and demerits of each method are described in comparative table to efficiently differentiate the each methods functionality. As per the analysis of all the existing algorithms, infrequent itemset mining frequent pattern growth uses algorithms that works out in very less computing time and the efficiency of performance has been improved when the large databases has been accounted.

.

REFERENCES.

1. R. Agrawal, T. Imielinski, and Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '93), pp. 207-216, 1993.
2. W. Wang, J. Yang, and P.S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and data Mining (KDD '00), pp. 270-274, 2000.
3. F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework," Proc. nineth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '03), pp. 661-66, 2003..
4. C.-K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets from Uncertain Data," Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD '07), pp. 47-58, 2007.
5. C.-K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets from Uncertain Data," Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD '07), pp. 47-58, 2007
6. C.-K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets from Uncertain Data," Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD '07), pp. 47-58, 2007.
7. T. Bernecker, H.-P.Kriegel, M. Renz, F. Verhein, and A. Zuefle,"Probabilistic Frequent Itemset Mining in Uncertain Databases,"Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 119-128, 2009
8 J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 1-12, 2000
9 X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules," ACM Trans. Information Systems,vol. 22, no. 3, pp. 381-405, 2004.
10] D.J. Haglin and A.M. Manning, "On Minimal Infrequent Itemset Mining," Proc. Int'l Conf. Data Mining (DMIN '07), pp. 141-147,2007.
11 Laszlo Szathmary, PetkoValtchev, and Amedeo Napoli," Finding Minimal Rare Itemsets and Rare Association Rules" Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (KSEM 2010)
12 A. Gupta, A. Mittal, and A. Bhattacharya, "Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees," Proc. Int'l Conf. Management of Data (COMAD), pp. 57-68,2011
13 NikkySuryawanshiRai, Susheel Jain, Anurag Jain," Mining Interesting Positive and Negative Association Rule Based on Improved Genetic Algorithm" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1, 2014
14 AzadehSoltani and M.-R.Akbarzadeh-T," Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets", IEEE transactions on neural networks and learning systems, 2014