International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 5 (May 2014), PP.25-31

Efficient mining of Positive and Negative Association Rules with weighted FP - Growth

[1]Ms. Varsha N Kavi, [2]Mr. Divyesh Joshi

[1,2]*Department of computer science and Engineering PIET, Gujarat technological University Gujarat, India*

Abstract:- In recent years, data mining has become one of the most popular techniques for data owners to determine their strategies. Association rule mining is a data mining approach that is used widely in traditional databases and usually to find the positive association rules. However, there are some other challenging rule mining topics like negative association rule mining. Assumptions made by classical association rule mining model that all items have same significance without assigning their weight within a transaction or record. But importance for the items and transactions is given by calculating weight on various items have been represented by the proposed method. This proposed system uses weighted FP Growth algorithm using coefficient of correlation for mining both positive and negative association rules. Experiments are carried on synthetic as well as real data sets to provide precise and important positive and negative association rules which benefits in consumption of memory and compaction of search area for negative rule mining which in turn reduces execution time.

Keywords:- Data Mining, Association rule mining, Data Processing, correlation, Market basket analysis.

I. INTRODUCTION

The importance of data mining has been increased rapidly for business domains like marketing, financing and telecommunications. In recent decade the development of economic is violent and swift. Information enhances unceasingly in a highest level. So the organizations and agencies have collected the massive business data. The business organizations urgent need to discover the valuable information and knowledge from the magnanimous data. The typical example of mining a frequent item set is market basket analysis through discovering the interactions between the different merchandise that the customer puts in "the basket". We can analyse a customer's purchasing habit or their interests of buying items.[3] This kind of discovery of may help the retail merchant to understand that which commodities are also purchased by the customers frequently and which are the other items purchased in combinations which helps in developing the better marketing strategies. Analysing data from different perspectives and summarizing it into useful information is what the process of data mining which can be used to increase revenue, cuts costs, or both. Association rule mining is a data mining technique that finds frequent patterns or associations in large data sets. Association rule mining is recognised as positive association rule mining. However, with the increasing usage of data mining technology, researchers have recently focused on finding unique patterns like negative associations. Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other. They are also very convenient for associative classifiers, classifiers that build their classification model based on association rules. It is an expensive process to discover Negative rules. For mining such rules, it has to examine a large search space. Depending on the probability of such an association, marketing personnel can develop better planning of the shelf space in the store or can base their discount strategies on such associations/correlations found in the data. All the traditional association rule mining algorithms were developed to find positive associations between items. By positive associations we refer to associations between items existing in transactions (i.e. items bought). What about associations of the type: customers that buy Pepsi do not buy Coke" or "customers that buy bottled water do not buy juice"? In addition to the positive associations, the negative association can provide valuable information, in devising marketing strategies [6,10,11]. This kind of information help the retail merchant to decide what kind of commodity will be selected to sell and the arrangement of commodity space and quantity. This determination can increase the volume of sales.

II. BASIC CONCEPTS

The association rules are an important research content in data mining which finds frequent patterns or associations in large data sets. An association rule is an implication of the form $A \Rightarrow B$, where A and B are frequent item sets in a transaction database which are called as positive association rules and $A \cap B = \varphi$. In practical applications, the rule $A \Rightarrow B$ can be used to predict that 'If a occurs in a transaction, then B will likely also occur in the same transaction, and we can apply this association rule to recommend who purchase B or

placing B close to A in the store's layout, such application are expected to provide more convenience for customers, and increasing product sales. [1] Recently much work is focused on finding alternative patterns, including unexpected patterns, which are also known as surprising patterns for example while 'bird (X) ⇒ flies (X)' is the well known fact, an exceptional rule is 'bird (X), penguin(X) ⇒ ¬ flies(X)' which indicates negative term and can be treated as a special case of negative rules. In paper [1] extends traditional associations to include association rules of the form ($A \Rightarrow \neg B$), ($\neg A \Rightarrow B$) and ($\neg A \Rightarrow \neg B$) which indicates negative associations between itemsets, and are called negative association rules. Negative association rules assist in decision making which also help the companies to hunt more business chances through infrequent itemsets of interests. Negative associations provide vital information to data owners.

Nowadays firms outsource their software's and databases which are called software as a service or database as a service to concentrate on own business [4, 5]. Database as a service have advantage of reliable storage of large volumes of data and saving database administration cost and efficient query processing. Firms outsourcing their XML databases to run trusted parties started looking for new ways of security like W3C, encryption standard, Crypto indexing to securely store data and efficiently query them [4].

III. PRELIMINARIES

Research on enhancing the speed of data processing of positive and negative association rule mining started with survey of different algorithms for market basket analysis. In that we found some different association rule mining algorithms as explained below.

A. Basic association rule mining algorithm

The association rules are an important research content in data mining which finds frequent patterns or associations in large data sets. Widely used example of association rule is market basket analysis. It can also be applied to other domains like marketing, financing, and telecommunication. Association rule mining is an important technique or mechanism in data mining. Association rule is an implication expression of the form X→Y where X is antecedent and Y is consequent. The antecedent and consequent are set of item from item domain I. The antecedent and consequent are a set of items from the domain I. Thus X∩Y=Φ. The support of an item set is defined as the ratio of number of transactions containing the item set to the total number of transactions. The confidence of the association rule X→Y is the probability of Y that exists in a transaction that contains X. First association rule mining algorithm is Apriory algorithm. It was first introduced by Agrawal, Imielinski and Swami (1993). Agrawal and Srikant (1994) have developed most popular association rule mining alogorithm called Aprioi and AprioryTid [2] . They have also shown in [2] that how the best features of Aprioi and AprioyTid algorithms can be combined in to a hybrid algorithm called AprioriHybrid. This hybrid algorithm has excellent scale-up properties. This Apriori algorithm is easy to implement but slow due to many passes over the data set. Therefore another fast rule mining algorithm, FP-Growth is proposed by Han, Pei and Yin (2000). There are two main improvements in FP-Growth algorithm uses FP-tree data structure, which is the compressed form of the data base helps in memory savings.[3] Secondly there is no candidate set generation in FP-Growth which makes overall algorithm fast [3].

B. Positive and Negative association rule mining

Samet and Taflan (2012) proposed a algorithm name PNRMXS (positive negative rule mining on XML stream in database as a service concept) which is based on FP-Growth approach. The processes in PNRMXS take place at two sides, client and server sides. At data owner (client) side, some pre-processing is done on the data set. At the service provider side, the mining takes place [1].The strength of association rule is measured with its support and confidence value. The support value if an item set is the proportion of transactions in the data set which contain the item set. The confidence value of a rule indicates its reliability. The support and confidence is given by Eq. 1 and Eq. 2 respectively

Negative association is like customer who buy product X, but not product Y. The search space is bigger in negative rule mining as compared to positive rule mining. Therefore negative rule with form ($X \Rightarrow \neg Y$) is given by Eq. 3 and Eq. 4 respectively

$$\text{supp}\ (X \Rightarrow \neg Y) = \text{supp}(X) - \text{supp}(XUY) \qquad (3)$$

$$\text{conf}\ (X \Rightarrow \neg Y) = \text{supp}(X) - \text{supp}(X\ UY)\ /\text{supp}(X) \qquad (4)$$

The negative rules are mined from the existing positive rules [1, 6]. Finding valid and sufficiently large number of negative associations is as important as saving memory in this approach Here construction of FP-

Tree is similar to original FP-Growth algorithm. They do not make any pruning at the beginning in FP Tree. Therefore, in PNRMXS [1], there are more nodes in the FP-Tree than that in the original FP-Growth and this leads to an increase in tree generation time. Here authors have used extra support and confidence thresholds as there is a need of sufficient pruning capability of negative rule mining. So in [1] support thresholds are used only in pruning the frequent item sets and confidence values are used only in the rule generation phase. There are five threshold values "MS", "MSP", "MCP", "MSN", "MCN", which are minimum support, minimum support for positive, minimum confidence for positive, minimum support for negative and minimum confidence for negative. They have adopted a pruning strategy of "correlation coefficient" value. This value is nothing but a covariance of the two variables. COV(X, Y) is divided by their product of standard deviations. This correlation value ranges from -1 to +1. If the value is equal to 0 it indicates that these two variables are independent or else there is a strong correlation between the variables [10,11]. If the item set has the correlation coefficient value greater than 0 it has positive correlation, and if less than 0 then it has negative correlation. The positive co relational frequent itemsets are used to mine positive association rules using MSP and MCP thresholds. Then the consequence part of the rule is to be negated to get negative rules. Then the support and confidence values of the candidate rule are compared with MSN and MCN thresholds to decide if it is a valid negative rule or not. After analysing the figure it can be concluded that in negative rule mining problem, the search space is 6–20 times bigger than that of positive mining[11]. Thus the execution time is longer in negative rule mining problem than that of positive rule mining as expected.

C. Weighted FP-Growth Algorithms

Some weighted FP-Growth algorithms like WARM (2013), WIP, WFIM, and WSFI [7,8,9] are proposed and proved their efficiency over FP Growth. In [7] paper, authors present an efficient algorithm WSFI (Weighted Support Frequent Itemsets)-Mine with normalized weight over data streams. Moreover, they have propose a novel tree structure, called the Weighted Support FP-Tree (WSFP-Tree), that stores compressed crucial information about frequent itemsets. WFIM algorithm is based on the importance of the items. In [8] the approach used to push the weight (Non negative real number) constraints in to the pattern growth algorithm while maintain the downward closure property. Here minimum weight and weight range are defined. WFIM generates more concise and important weighted frequent itemsets in large databases.

IV. RELATED WORK

Association rule mining is a data mining approach that is used often in traditional databases and usually to find the positive association rules. As compared to positive associations, negative association take over lots of search space. In turn needs more execution time as well as consumes more memory. So for faster mining process there is need of precise and special patterns with accuracy. For fast and accurate decision we need to have some most significant patterns and prioritize the selection of target item sets according to their significance in the data set. The main motivation of this project is to provide the data processing accurately and efficiently with significant or prioritized data items from large data base. Prior association rule model assumes that items have the same significance without taking account of their attribute within a transaction or within the whole item space. On the other side, in real world applications, specific patterns and items within the patterns have more importance or priority than other patterns. Implementing a weighted frequent pattern mining functions to retrieve the hidden knowledge from data sets is done. We have proposed a rule mining methodology that mines significant positive and negative association rules. Some information by Negative association rules also provided through which accuracy can be guaranteed and efficiency can be increased. Data owner are benefited by this concept which provides lots of profit to business. Companies are assisted in decision making by Negative association rules which also help to hunt more business opportunities through non frequent item sets of interests. Significant positive as well as negative associations provide vital information to data owners reducing execution time and space occupancy. The main objective of this project is to implement the better system for data processing with excellent time and space management. The goal of using weighted support is to make use of the weight in the mining process and prioritize the selection of target item sets according to their significance in the data set rather than their frequency alone. The idea behind is that, instead of producing huge number of meaningless rules of frequent items it is worth to produce significant precise rules. To propose a rule mining approach which provides significant positive and negative association rule computation on large data set, providing lots of benefits to data owner for faster and correct decision. This determination can increase the volume of sales. Significant and precise information from large data set is for faster decision and execution time. This existing system provides information of positive as well as negative association rules for better business strategies. This kind of information help the retail merchant to decide what kind of commodity will be selected to sell and the arrangement of commodity space and quantity. But PNRMXS algorithm generates lots of negative rule and using no technique to pruning of in frequent item sets for getting all negative rules from existing

positive rules. Because of this it consumes more memory and as well as more 'cpu' time for mining. Instead of generating lot many meaningless rules it is better to have most significant practised special rules for implementing easily and quickly for better strategy which in turn reducing memory consumption and execution time even better faster decision making process. In many applications, some items appear very frequently in the data, while others rarely appear.

If minsup is set too high, those rules that involve rare items will not be found. To find rules that involve both frequent and rare items, minsup has to be set very low. The key element that makes association rule mining practical is the minsup. It is used to prune the search space and to limit the number of rules generated. In paper [12] argues that using a single minsup for the whole database is inadequate because it cannot capture the inherent natures and/or frequency differences of the items in the database. By the natures of items we mean that some items, by nature, appear more frequently than others[12]. They extended the existing association rule model to allow the user to specify multiple minimum supports to reflect different natures and/or frequencies of items. Specifically, the user can specify a different *minimum item support* for each item. Minimum item supports or the weight of items thus enable us to achieve the goal of having higher minimum supports for rules that only involve the frequent items, and having lower minimum supports or the weight of item for rules that involve less frequent items. For our experiments, to find significant items we need a method to assign weight of an item in the data set. We use the actual frequencies (or the supports) of the items in the data base as the basis for significance of the weighted items. We use the following formulas:

$$Weig(i) = \begin{cases} W(i) & W(i) \geq ms \\ ms & Otherwise \end{cases}$$

$W(i) = f(i) / No.of\ transactions\ in\ database.$

$f(i)$ is the actual frequency (or the support expressed in percentage of the data set size) of item in the database. *ms* is the user-specified minimum item support allowed. Thus Weight value for items should be related to their frequencies in database. Thus, to calculate weight values for items we use two parameters, frequency of the item and number of transactions of the database.

V. PROPOSED ALGORITHM

Input: A Transaction data base TDB , Minimum support threshold : min_sup, Weight of the items : Wi, weight range, Minimum weight threshold : min_weight, Minimum positive support : MSP , Minimum positive confidence: MCP, Minimum negative support : MSN, Minimum negative confidence : MCN

Output: The complete set of weighted frequent item sets, Counts of positive and negative frequent itemsets, Execution time for Weighted FP Tree construction. Total time consumed for the mining only positive association rules and both positive and negative association rules.

Method:

1. Scan TDB once to find the global weighted frequent items. Weight(i)=frequency of item (in percentage) / No.of transactions in a data dase

Weight range = minW to maxW

MaxW=support * maxW and MinW=suppor * minW

Item X is frequent if it does not satisfies these two conditions:

1.1 : (support < min_sup && weight < min_weight)

1.2 : (support * MaxW < min_sup)

2. Sort items in weight ascending order. The sorted weighted frequent item list forms the weight_order and header of the FP Tree
3. Scan the TDB again and build a global FP Tree using weight_order (weighted FP Tree).
4. Mine weighted FP Tree for weighted frequent itemset mining in a bottom up manner. Form conditional data bases for all remaining items using conditions :

4.1 (support < min_sup && weight < min_weight)

4.2 (support * MinW < min_sup

5. Mine all the items in the global header table generate significant frequent items. Find the execution time for constructing weighed FP Tree.
6. Finding Correlation Coefficient of all the frequent items.If corr(A,B) >1 add to positive association rule else If corr(A,B) < 1 add to negative association rule.

After getting frequent items from WFP Growth algorithm WPNRM(proposed algorithm) uses correlation

coefficient for finding the relation between the data items. So this pruning technique mines both positive and negative association rules at the same time [1]. Correlation Coefficient provides dependency of two items. The correlation coefficient (denoted as $corr_{A,B}$) can show the relevance of the two itemsets [14]. It is calculated as follows:

$$Corr(A,B) = SUP\ P(A\ U\ B) \div (SUP\ P(A) * SUP\ P(B))$$

The value of Correlation Coefficient exists in following three situations :

1. Corr(A,B) = 1 then A and B are independent.
2. Corr(A,B) > 1 then A and B are positive correlation.
3. Corr(A,B) < 1 then A and B are negative correlation

VI. EXPERIMENTAL RESULTS

The section evaluates the extended model of PNRMXS. The workstation used for the experiments has Intel core I5 2.50 GHz processor and 4 GB ram. The Operating system is Windows home Basic and installed Java version is JDK (1.7). In the experiments all results on execution time are in milliseconds, and memory usage is in Megabytes and all tests are made on the mentioned workstation without any extra load and with only one user. We showed the proposed model working on different data sets like real and as well as synthetic data base. Table 1 gives the details of experimenting datasets.

Table 1 Details of experimental data sets.

Data set	No. of Transactions	Attributes
Retail	**88161**	**572**
T10I4D100K	**1,00,000**	**1000**

It is experimented for finding positive and negative association rules with different minimum supports, yet without generating a huge number of meaningless rules with frequent items. The proposed method WPNARM tree generation time, memory consumption and also execution on different data sets is compared with existing PNRMXS method. The graphs depict the efficiency of the WPNARM. Weighted frequent items are decreased as the weight range is decreased.WPNARM can adjust the number of weighted frequent item sets by user's feedback in the dense data base with very low minimum support, which helps in reducing memory consumption for the weighted FP Tree. Run time is sharply reduced even with low minimum support as the weight ranges become lower. Table 1 and Table2 shows the comparison of PNRMXS (FP-Growth) and WPNARM (WFP_Growth) experimented on synthetic data set.

Table 3 and Table 4 show the observations of retail dataset.

Table1 Readings of synthetic dataset

Minimum Support	FP tree Time (ms)	WFp tree Time (ms)	FP-memory (MB)	WFp-memory (MB)
0.0025	**179516**	**106690**	**39.741**	**32.34**
0.003	**125583**	**85691**	**37.781**	**31.34**
0.004	**93165**	**66861**	**33.749**	**30.41**
0.005	**58466**	**44584**	**30.74**	**28.49**

Retail data set is very dense so it consumes lot of cpu time at very less minimum support because at this stage there will be more frequent itemsets. Need more time for generating rules for large frequent patterns. On comparing the execution times of FP-Growth(PNRMXS) and WFP-Growth(WPNARM) our proposed system outperforms as it generates only significant frequent items, and it generates Significant Association Rule and thus reducing execution time. Even in comparison of memory occupancy, only significant data consumes memory.

Table 2 Readings of synthetic dataset

Minimum Support	FP Rules	WFp Rules	Total FP-Growth execution time (ms)	Total WFP Growth execution-time (ms)
0.0025	**5809**	**1267**	**428487**	**76987**
0.003	**3883**	**783**	**418641**	**65926**
0.004	**1158**	**546**	**150871**	**67159**
0.005	**846**	**412**	**297728**	**64881**

And hence WFP Growth of proposed system overturns FP Growth of existing system. Important frequent item sets consumes less tree generation time for WFP Growth of proposed work. Significant Rules helps to take quick and accurate decisions for implementation.

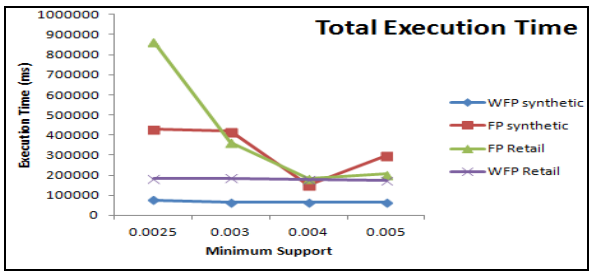

Figure1Graph showing total execution Time

Table 3 Readings of Retail dataset

Minimum Support	FP tree Time (ms)	WFp tree Time (ms)	FP-memory (MB)	WFp-memory (MB)
0.0025	**834352**	**281928**	**45.77**	**39.12**
0.003	**354697**	**184049**	**44.786**	**39**
0.004	**176358**	**178495**	**42.786**	**39.36**
0.005	**198572**	**172849**	**41.78**	**38.12**

Table 4 Readings of retail dataset

Minimum Support	FP Rules	WFp Rules	Total FP-Growth execution time (ms)	Total WFP Growth execution-time (ms)
0.0025	2574	252	863390	182521
0.003	1676	171	364776	184643
0.004	809	78	183785	178948
0.005	589	51	205110	173332

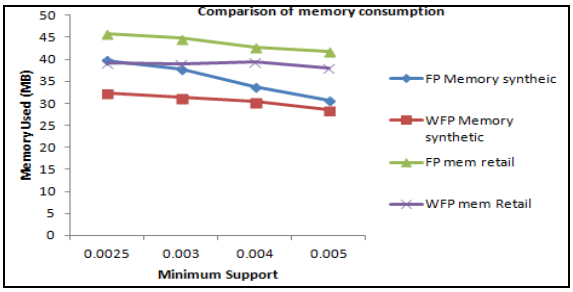

Figure 2 comparison of memory consumption

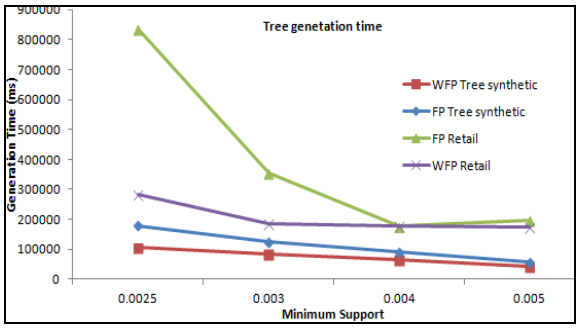

Figure 3 Tree generation Time

VII. CONCLUSIONS

Most of the existing methods concentrate on only positive association rule mining. Traditional association rule mining algorithms can be modified to find even the negative association rule mining which help in performing better business strategies. The FP-Growth algorithm can be extended to weighted FP-Growth for generating more accurate positive and negative rules on different data sets proving the efficiency of WPNARM in memory consumption due to weighted FP Tree and précised significant frequent items. Execution time also improved in mining significant frequent items. The research work can be extended by modifying the technique of finding dependency or correlation of the frequent itemsets.

REFERENCES

[1]. Samet Cokpinar, Taflan Imre Gunden, "Positive and negative rule mining on XML data streams in database as a service concept", Expert Systems with Applications, Vol.39, pp.7503-7511, 2012.

[2]. Agrawal, R., Srikant, R., "Fast algorithms for mining association rules in large Databases", In 20th International Conference on Very Large Data Bases, 1994.

[3]. Liu, Y., & Guan, Y. (2008), "FP-Growth Algorithm for Application in Research of Market Basket Analysis." IEEE International Conference on Computational Cybernetics, 269-272

[4]. Chit Nilar Win, Khin Haymar Saw Hla "Mining frequent patterns from XML data" University of Computer Studies, Yangon

[5]. Unay, O., & Gundem, T. _I. (2008). A survey on querying encrypted XML documents for databases as a service. ACM SIGMOD Record, 37(1), 12–20.

[6]. Wu, X., Zhang, C., & Zhang, S. (2004). "Efficient mining of both positive and negative association rules",ACM Trans. Inf. Syst. (pp. 381–405).

[7]. Younghee Kim, Wonyoung Kim and Ungmo Kim "Mining Frequent Itemsets with Normalized Weight in Continuous Data Streams", Journal of Information Processing Systems, Vol.6, No.1, March 2010 DOI : 10.3745/ JIPS.2010.6.1.079

[8]. Unil Yun and John J. Leggett "WFIM: Weighted Frequent Itemset Mining with a weight range and a minimum weight" ,Texas A&M University College Station, TX 77843, U.S.A

[9]. V.Vidya ,"Mining Weighted Association Rule using FP – tree", IJCSE, 2013

[10]. Maria-Luiza Antonie Osmar R. Za¨ıane ," Mining Positive and Negative Association Rules: An Approach for Confined Rules", 2004 IEEE

[11]. Honglei Zhu, Zhigang Xu," An Effective Algorithm for Mining Positive and Negative Association Rules",2008 IEEE

[12]. Bing Liu, Wynne Hsu and Yiming Ma,"Mining Association Rules with Multiple Minimum Supports",1999 ACM SIGKDD