IJCSE International Journal of Computer Science and Engineering Open Access
Review Paper Volume-2, Issue-3 E-ISSN: 2347-2693

# A Survey on Enhancing Data Processing of Positive and Negative Association Rule Mining

Varsha Kavi[1] and Divyesh Joshi[2]
[1*] Computer Engineering Department, PIET, Gujarat Technological University, India, varshaswagat@gmail.com
[2] Computer Science and Engineering Department, PIET, Gujarat Technological University, India, divyeshjoshi89@gamail.com

www.ijcseonline.org

Received: 5 March 2014 Revised: 12 March 2014 Accepted: 22 March 2014 Published: 30 March 2014

***Abstract—*** Importance of data mining has been increased rapidly for business domains like marketing, financing and telecommunications. The business organizations urgent need to discover the valuable information and knowledge from the huge data. We can analyze a customer's purchasing habit or their interests of buying items in market basket analysis. This kind of discovery may help in developing the better marketing strategies. Depending on this strategy this survey paper is based on developing fast processing recommendation engine which suggests the customers for purchasing the needed commodities with the main item. This survey paper focuses on different important association rule mining algorithms like Apriori, FP-Growth and Weighted FP-Growth and on enhancing the processing speed of positive and negative associations rule mining.

***Keywords/Index Term—*** Data Mining, Data Processing, Outsource Services, Market basket analysis, Ajax Technique.

## 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 analyze a customer's purchasing habit or their interests of buying items.[8,9] 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. This paper is based on developing recommendation engine that is suggesting the customers for purchasing the needed commodities with the main item for example, while purchasing DLSR Camera some recommended items are displayed like cover, lens , batteries etc. in the box according to the priorities or frequent item analysis, which is the most better business strategy. 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 quatity. 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 $A \cap B = \varnothing$. 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) $\Rightarrow$ flies (X)' is the well known fact, an exceptional rule is 'bird (X), penguin(X) $\Rightarrow \neg$ 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 [11]. 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 un trusted parties started looking for new ways of security like W3C, encryption standard, Crypto indexing to securely store data and efficiently query them [10].

The rest of the paper is organized as follows. In section III , literature review of the papers is given with their advantages and limitations. Section IV has some possible future research suggestions on the subject.

Corresponding Author: Varsha Kavi

 

## 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 $X \cap Y = \Phi$. 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 that Y exists given that a transaction contain 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 [] 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. Secondly there is no candidate set generation in FP-Growth which makes overall algorithm fast [3,9].

### *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

$$\text{Supp}(X \Rightarrow Y) = (X \cup Y) / N \quad (1)$$

$$\text{Conf}(X \Rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X) \quad (2)$$

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

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

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

In streamed data, algorithms should run fast as possible due to the fast flow of data [12]. It is not possible to store data in memory as it crucial. There are three stream data processing models Landmark, Damped and Sliding window (Zhu and Shasha , 2002). PNRMXS algorithm includes three steps as shown in Fig.1

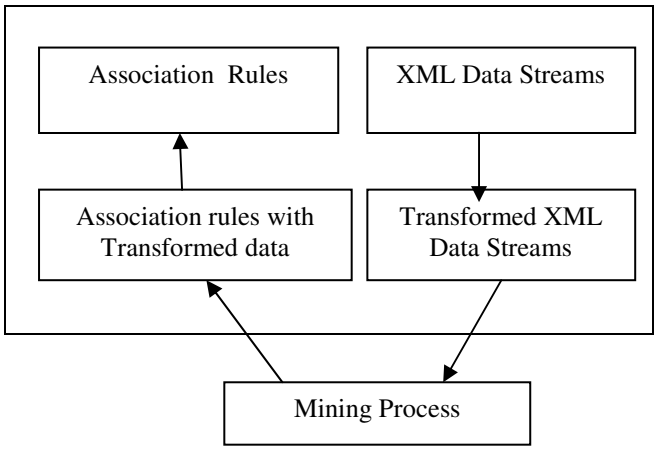


Fig. 1. Flowchart of the PNRMXS

One is data transformation where XML data stream is transformed to flat file for mining, which uses security of one to one mapping which provides privacy for the data the transformed items are with random number generator that makes guessing the original item content difficult. Second step is mining process, where in authors use landmark windows data processing model. The change that they made to the FP Growth algorithm is to scan the data stream only once. Data stream is processed block by block. Each block contains a fixed number of transactions. The negative rules are mined from the existing positive rules. 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. 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. 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

| Data Set Name | Positive Rule Mining | | | Negative Rule Mining | | | |
|---|---|---|---|---|---|---|---|
| | Frequent set | Positive Rule | Time (ms) | Frequent Set | Positive Rule | Negative Rule | Time (ms) |
| T10N1000I4D100K | MS=0.5%,MSP=0.5%, MCP=1% | | | MS=0.1%,MSP=0.5%,MCP=1%, MSN=0.5%,MCN=1% | | | |
| | 596 | 2 | 9093 | 12490 | 2 | 1202 | 11328 |
| T10N4000I4D100K | MS=0.3%,MSP=0.3%, MCP=1% | | | MS=0.1%,MSP=0.3%,MCP=1%, MSN=0.3%,MCN=1% | | | |
| | 1252 | 18 | 13302 | 8237 | 18 | 0 | 15328 |
| T10N000I4D200K | MS=0.5%,MSP=0.5%, MCP=1% | | | MS=0.1%,MSP=0.5%,MCP=1%, MSN=0.5%,MCN=1% | | | |
| | 599 | 2 | 21875 | 12465 | 2 | 1244 | 25594 |
| T10N4000I4D200K | MS=0.3%,MSP=0.3%, MCP=1% | | | MS=0.1%,MSP=0.3%,MCP=1% ,MSN=0.3%,MCN=1% | | | |
| | 1250 | 18 | 24781 | 8241 | 18 | 0 | 36124 |

Fig. 2 Negative rule mining effect on execution time

negated. 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. The last step is data retransmission. Experiment is done on using different IBM synthetic data sets and execution time for the process at different stages is found in milliseconds and even showed memory usage and proved the efficiency of the algorithm. Various tests have been made with different parameter settings and different synthetic data sets to show the scalability and stability of the proposed methodology. They have shown effect of negative rule mining in Fig 2. 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. 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 [4,5,6,7] are proposed and proved their efficiency over FP Growth. In [4] 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. Empirical results show that their algorithm outperforms comparative algorithms under the windowed streaming model[7]. Use of weight in the mining process and prioritize the selection of target item sets according to their signification in the data set. In [7] paper weighted support of an item is defined as the fraction in the item set occurs. W-support is considered as a common inference of support which employs weights of all transactions into accounts. The w-support of X→Y is given in Eq. 5 and Eq. 6

$$Wsup(X \rightarrow Y) = (X \cup Y) \quad (5)$$

$$Wconf(X \rightarrow Y) = Wsup(X \cup Y) / Wsup(Y) \quad (6)$$

WFIM algorithm is based on the importance of the items. In [5] 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. WFIM algorithm steps are as follows:

1. Scan TDB to find frequent itemset if the item ste does not satisfy the following condition.

   1.1 (support<min_sup && weight < min weight)

   1.2 (support * MAXW < min_sup)

2. Ascending order sorting of item in weight which forms the weight_order and header of FP Tree.

3. Scan the TDB and global FP Tree using weight_order is built.

4. Mine global FP Tree for weighted frequent itemset mining in bottom up manner forming conditional databases using conditions 1.1 and other as (support*MINW<min_sup).

5. When all the items in the global header table have been mined , WFIM is completed.

Researchers have used real data sets and synthetic data sets and showed that WFIM has good scalability against the number of transactions. WFI is decreased as the weight range is decreased. WFIM 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 Show results of WFIM and Table 2 shows the comparisons of all the association rule mining algorithms

TABLE 1 COMPARISON OF FREQUENT ITEMSETS BY WFIM

| Support of Connect Dataset | Number of W.F.I WR: 0.5 –1.5 MW : 1.5 | Number of W.F.I WR : 0.5 – 1.5 MW : 1.0 | Number of W.F.I WR : 0.5 – 1.5 MW : 0.5 | Num of F.C.I | Num of F.I |
|---|---|---|---|---|---|
| 64179 (95%) | 125 | 784 | 1471 | 812 | 2205 |
| 60801 (90%) | 690 | 2346 | 5312 | 3486 | 27127 |
| 54046 (80%) | 2769 | 2989 | 3044 | 15107 | 533975 |
| 47290 (70%) | 3997 | 4089 | 4093 | 35875 | 4129839 |

TABLE 2 COMPARISONS OF ARM ALGORITHMS

| ALGORITHM | ADVANTAGES | DISADVANTAGES |
|---|---|---|
| **APRIORI** | -Easy to implement<br>-New pruning tech.<br>-Avoids wastage of counting candidate which are infrequent | -Too many scans on database high CPU usage |
| **FP GROWTH** | -Only two passes on database<br>-No candidate generation<br>-Faster than Apriori<br>-Computation cost decreased<br>-FP Tree construction | -FP Tree is difficult to use in an interactive mining system<br>-FP Tree is not suitable for incremental mining |
| **WEIGHTED FP GROWTH** | -Good scalability<br>-Generates more concise and important weighted frequent item sets.<br>-Reduction in memory consumption | -Extra burden of defining proper weight ranges |

*D. A web Asynchronous communication mechanism based on AJAX.*

Data exchange between a client and a server through the asynchronous communication mechanism is implemented by AJAX [13,14]. Ajax belongs to a series of web 2.0 technology. Ajax is the application made combing of (Javascript + XML). It is not a new language but a new technology with advantages like small amount of data transfer, lightening the burden on the server and the bandwidth, with great experience of speed to customers. When the client operates browser Ajax engine makes communication between browser and server asynchronously [13,14]. Not all the operations of the user are passed to server, some verification and processing will be completed by Ajax. While only new data return, the client page is refreshed. There is not the reloading of the whole page, only changed part is sent to client. So Ajax eliminates the problem of traditional web information exchange "process-wait-process". XMLHttpRequest object is the key technology to achieve Ajax asynchronous communication. IE, Mozilla, Opera, Safari and other browsers contain this object. Ajax provides improved user interaction with web based application. But this new technique too have some limitations like, lack of formal semantics make Ajax application difficult to build, debug, and to understand and validation. Technical features of Ajax include XHTML + CSS pages, DOM(Document object Model) for dynamic and interactive, XML and XSLT for data exchange, XMLHttpRequest object to asynchronous data query and operations together.

## IV. SUGGESTIONS FOR FUTURE WORK

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. This literature survey information can be used for developing an recommendation engine for commodities whose data processing can be enhanced by using Ajax data set, instead of XML data stream. The FP-Growth algorithm can be extended to weighted FP-Growth for generating more accurate positive and negative rules on different data sets. Even to reduce memory consumption due to weighted FP Tree.

## REFERENCES


[1]. Samet Cokpınar, Taflan _Imre Gundem "Positive and negative association rule mining on XML data streams in database as a service concept ", ELESVIER 39 ,page no(7503–7511) ,Jan 2012.

[2]. Rakesh Agrawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules", 20th VLDB Conference Santigo Chile ,page no(487-499), Sep 1994.

[3]. Jiawei Han, Jian Pei, Yiwen Yin ,"Mining Frequent Patterns without Candidate Generation" ACM SIGMOD International onference on Management of Data, Dallas, Texas, United States.

[4]. 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, page no (79-90) , March 2010.

[5]. Unil Yun and John J. Leggett, "WFIM: Weighted Frequent Itemset Mining with a weight range and a minimum weight" ,SIAM, page no (636-640).

[6]. V.Vidya "Mining Weighted Association Rule using FP – tree", IJCSE , Vol. 5 No. 08,page no (741-752), Aug 2013

[7]. Raorane A.A. Kulkarni R.V. and Jitkar B.D, "Association Rule – Extracting Knowledge Using Market Basket Analysis" , Research Journal of Recent Sciences,Vol. 1(2), page no(19-27), Feb. 2012.

[8]. Yongmei Liu and Yong Guan, "FP-Growth Algorithm for Application in Research of Market Basket Analysis ",IEEE International Conference on Computational Cybernetics, page no(269–272), Nov 2008.

[9]. Chit Nilar Win, Khin Haymar Saw Hla ," Mining frequent patterns from XML data", APSITT, page no (208–212), 2005.

[10]. Ozan Ünay Taflan !.Gündem "A Survey on Querying Encrypted XML Documents for Databases as a Service" SIGMOD Record, Vol. 37, No. 1,page no(12-20),March 2008 .

[11]. Xindong WU,Chengqi Zhang and Shichao Zhang," Efficient Mining of both Positive and Negative Association Rules", ACM Transaction on Intenational System,Vol 22 No 3, page no(381-405), July 2004.

[12]. Sun Zhaoyun,Zhang Xiaobo,Zhao Li , "The Web Asynchronous Communication Mechanism Research Based


  

on Ajax" , 2nd international Conference on Education Technology and Computer (ICETC), Vol 3, page no(370-372), 201 0.

[13]. Supratik Mukhopadhyay, Ramesh Bharadwaj, Hasan Davulcu, "Functional "AJAX" in Secure Synchronous Programming", Proceedings of the 44th Hawaii International Conference on System Sciences, IEEE , 2011.

[14]. Wang Jing and XU Feng , "The Research of Ajax Technique Application Based on the J2EE", Foundation project: The Design of the System for Zhanjiang Meteorological Forecast (0910262), IEEE, 2010.

AUTHORS PROFILE

*Mrs. Varsha Kavi*

She has received her B.E (CST) in 1988 from BAS Engineering college Bagalkot, Karnataka University. She is currently a master student in Computer Engineering at PIET Gujarat Technological University. Her research interests include data mining , Data stream Mining.

*Mr. Divyesh Joshi*

He is working as a Assistant Professor at PIET, Gujarat Technological University. He has received his B.E (I.T) degree from S.P.B. Patel College of Engineering-Hemchandracharya North Gujarat University(HNGU) in 2011 and M.tech(CE) from U.V. Patel College of Engg-Ganpat University-Mehsana-2013. His area of interest include Data Mining , Data stream Mining.

 