ISSN: 2277 – 7962

International Journal of Computer Application and Engineering Technology
Volume 3-Issue 3, July 2014. Pp. 232-236
www.ijcaet.net

APRIORI ALGORITHM FOR MINING FREQUENT ITEMSETS –A REVIEW

Priyanka[1]*, Er. Vinod Kumar Sharma[2]
1.M.Tech (Scholar), 2.Assistant Professor
Department of Computer and Science Engineering,
Guru Kashi University, Talwandi Sabo,
Punjab, India
*Email:Priyankakansal8@gmail.com

Abstract: Data mining is a process of extraction of valuable and unknown information from the large databases. The data mining is a process of analyzing a huge data from different perspectives and summarizing it into useful information. The information can be converted into knowledge about historical patterns. Many Algorithms have been proposed to mine association rule that uses support and confidence as constraint. We are proposing a method that can be combined with Apriori algorithm and reduces storage required to store candidate and the execution time by reducing CPU time. Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules. Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework. In this paper gives an extension to the Apriori algorithm, a classical rule mining algorithm. Apriori finds its application in areas of data mining, finding association between attributes and in prediction systems. To increase the efficiency of the proposed Apriori algorithm a method for incorporating a new correlation factor (threshold) is being introduced.

Keywords: Data Mining, Association Rules, Apriori Algorithm, Frequent Item sets, Support, confidence

1. INTRODUCTION

Data Mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. Data mining functions include clustering, classification, prediction, and associations. One of the most important data mining applications is that of mining association rules. Association rules, first introduced in 1993, are used to identify relationships among a set of items in databases [1, 2]

Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace… etc, to increase their business efficiency. Therefore the needs for a standard data mining process increased dramatically. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

A. Advantages
- Market Basket Analysis
- Fraud Detection

B. Disadvantages
- User privacy/security
- Possible misuse of information
- Great cost at implementation stage

2. OVERVIEW OF ASSOCIATION RULES

Association rules were presented by R.Agarwal and others in 1993. Its main purpose is to find the association relationship among the large number of database items. . It is used to describe the patterns of customers' purchase in the supermarket [1]. Apriori employs an iterative approach known as a level-wise and breadth-first search, which k-item-sets are used to generate (k+l)-item-sets[9]. In terms of the feature of Apriori property, called ant monotone, one can efficiently generate candidate item-sets by discarding unnecessary considered ones.

However, the algorithms based on generated and tested candidate item-sets have two major drawbacks:
1. The database must be scanned multiple times to generate candidate sets. Multiple scans will increase the I/O load and is time-consuming.
2. The generation of huge candidate sets and calculation of their support will consume a lot of CPU time.[4]

Its main aim is to find out the interesting patterns among multiple domains based on a given degree of support and confidence.

Definition1: The association rules can be formally defined as:
- If the support of item-sets *X* is greater than or equal to minimum support threshold, *X* is called frequent item-sets.
- If the support of item-sets *X* is smaller than the minimum support threshold, then *X* is called infrequent item-sets.

Definition2: The support of an item-set is the fraction of the rows of the database that contain all of the items in the item-set. Support indicates the frequencies of the occurring patterns. Sometimes it is called frequency. Support is simply a probability that a randomly chosen transaction t contains both item-sets A and B. Mathematically,

$$\text{Support } (A \Rightarrow B)\ t = P\ (A \subseteq t \wedge B \subseteq t)$$

We will use a simplified notation that

$$\text{Support } (A \Rightarrow B) = P\ (A \wedge B)$$

Definition3: Confidence denotes the strength of implication in the rule. Sometimes it is called Accuracy. Confidence is simply a probability that an item-set B is purchased in a randomly Chosen transaction t given that the item-set A is purchased. Mathematically,

$$\text{Confidence } (A \Rightarrow B)\ t = P\ (B \subseteq t \mid A \Rightarrow t)$$

We will use a simplified notation that

$$\text{Confidence } (A \Rightarrow B) = P\ (B|A)$$

Definition4: Minimum-Support= $\frac{\text{No.Of transactions containing both A\& } B}{\text{Total no.of transactions}}$

Definition5: Minimum-Confidence= $\frac{\text{No.of transactions containing A\& } B}{\text{Transactions containing only A}}$

In general, a set of items (such as the antecedent or the consequent of a rule) is called an Item-set. The number of items in an item-set is called the length of an item-set. Item-sets of some length k are referred to as k-item-sets. Generally, an association rules mining algorithm contains the following steps:

- The set of candidate k-item-sets is generated by 1-extensions of the large (k -1)- item-sets generated in the previous iteration.
- Supports for the candidate k-itemsets are generated by a pass over the database.
- Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

3. APRIORI ALGORITHM

Apriori Algorithm is the most popular & classical algorithm proposed by R. Agarwal in 1994 for mining frequent item-sets. Apriori is used to find all frequent itemsets in a given database. The key idea of Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-itemsets are used to explore (k+1)-itemset

A. Apriori Algorithm [2]

Find frequent item-sets using an iterative level-wise approach based on candidate generation.
Input: *D*, a database of transactions;
min- sup, the minimum support count threshold.
Output: *L*, frequent item-sets in *D*.

Method:
(1) L_1 = find_ frequent_ 1-itemsets (D);
(2) for ($k = 2; L_k\text{-}1 \neq \phi; k{+}{+}$) do begin
(3) Ck = apriori _gen (L_k-1);
(4) for each transaction $t \varepsilon\ D$ do begin// scan *D* for counts
(5) C_t= subset (C_k, *t*); // get the subsets of *t* that are candidates
(6) for each candidate $c \varepsilon\ Ct\ do$
(7) c.count++;
(8) end
(9) $Lk = \{c\varepsilon\ C_k | c.count \geq min\ \_sup\}$
(10) end
(11) return $L = U_k L_K$;

4. LITERATURE SURVEY

4.1. An Efficient Algorithm for Mining Association Rules using Confident Frequent Itemsets

B. Al-Maqeleh et.al**,** introduced a problem with such a process is that the solution of interesting patterns has to be performed only on frequent item sets. An efficient algorithm is proposed to integrate confidence measure during the process of mining frequent item sets, may substantially improve the performance of association rules mining by reducing the search space. The experimental results show the effectiveness of the proposed algorithm in reducing the number of discovered rules comparing with the Apriori algorithm. [3]

TABLE OF COMPARISON

Varaiations	Methodology	Input Parameters	Problems to Overcome	Results
A[3]	Frquent itemsets	Confidence measure	More time of Association rules	Reduce Search space
B[4]	Checkpoint and CPU time	Support and confidence	Calculation of Large candidate sets in more time	Reduces the no of candidate generated and removed candidate at checkpoint
C[5]	Corelation Threshold	Same	Scanning and generation of candidate set more time	Reduce the time complexity
D[6]	Types of Apriori algorithm	Weka Tool	No of cycle generate in frequent item set	Reduce the minimum support and find the resuired no of rules
E[7]	Combination of Data Division and dynamic itemsets counting	Same	Scanning of Database	The Whole database needs twice to be scanned but it reduces the quality of candidate sets generated
F[8]	Hashing	Same	Scanning and geeration of candidate sets	Requires only one scan but still large no of candidate set generated

4.2 Apriori: A Modified Apriori Algorithm Based on Checkpoint

M. Patel et.al, a proposed of many algorithms to mine association rule that uses support and confidence as constraint. We proposed a method based on support value that increase the performance of Apriori algorithm and minimizes the number of candidate generated and removed candidate at checkpoint which is infrequent which interns reduces storage and time required to calculate support of candidate.[4]

4.3. Applying Correlation Threshold on Apriori Algorithm

A.H.S *et.al***,** introduced an Apriori algorithm, a classical rule mining algorithm finds its application in areas of data mining, finding association between attributes and in prediction systems. Performance of the redesigned algorithm is evaluated and is compared with the traditional Apriori algorithm. To increase the efficiency of the Apriori algorithm and reduce the time complexity of the proposed algorithm into O (n). [5]

4.4. Utility of Association Rule Mining: a Case Study using Weka Tool

A.Lekha *et.al* **,** introduced a few case studies pertaining to breast cancer, mushroom, larynx cancer and other datasets are studied to find the utility of association rule mining using Weka tool. They are three association algorithms - Apriori, PredictiveApriori and Tertius Algorithms and comparative study of the three algorithms is also made and also the implementation of the three algorithms gives the strong association rules they have problems with the number of cycles taken to generate the frequent item-sets, minimum support needed, memory utilized and non-numeric data.[6]

4.5. An Improved Apriori Algorithm Based on Association Analysis

Y. Jia *et.al* , proposed an improved algorithm based on a combination of data division and dynamic item-sets counting. The proposed algorithm has improved the two main problems which are faced by classical apriori algorithm. First is the repeatedly scanning of transactional database and second is the generation of large number of candidate sets. In data division, the transactional database is divided into n parts that don't intersect each other. In first scan, all the frequent sets of each division are mined which is called local frequent sets. In second scan, the

whole database is scanned again, getting support degree of all candidate item-sets and then deciding the global frequent item-sets. After data division, dynamic item-sets counting are used to decide candidate item-sets before scanning database every time. So, the whole process needs only twice the entire database scan. [7]

4.6. An Improved Apriori Algorithm

R. Chang et.al**,** proposed an APRIORI-IMPROVE algorithm in which level L2 is directly generated from one scan over the database without generating candidate sets C1, L1 and C2. APRIORI-IMPROVE uses hash table and efficient horizontal data representation. APRIORI-IMPROVE also optimized strategy of storage to save time & space. The performance of APRIORI-IMPROVE is higher as compared to apriori and fpgrowth. [8]

5. CONCLUSION

In this paper, we analyzed and studied various existing improved apriori algorithm to mine frequent itemsets. Mainly common drawbacks are found in various existing apriori algorithm which is improved by using different approaches. It can be applied to many different applications like market basket analysis, telecommunication, network analysis, banking services and many others. In the future work, the problem of large number of candidate sets generated can still be improved and doing my work of improved apriori algorithm in multiple fields like as contact lenses, market basket analysis,voting etc.

6. RERFERENCES

[1]. Arun K Pujari “Data Mining Techniques”, Edition 2001.
[2]. Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”Second Edition.Morgan Kaufmann Publisher, 2006,Pp.123-134.
[3]. Basheer Mohamad Al-Maqaleh and Saleem Khalid Shaab,” An Efficient Algorithm for Mining Association Rules using Confident Frequent Itemsets,” 2012 Third International Conference on Advanced Computing & Communication Technologies.
[4]. Mihir R. Patel, Dipti P. Rana, and Rupa G. Mehta,” FApriori: A Modified Apriori Algorithm Based on Checkpoint,”IEEE 2013.
[5]. Anand H.S. and Vinodchandra S.S.,” Applying Correlation Threshold on Apriori Algorithm,” 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013).
[6]. A. Lekha, Dr. C V Srikrishna and Dr. Viji Vinod,” Utility of Association Rule Mining: a Case Study using Weka Tool,”2013 IEEE.
[7]. Yubo Jia, Guanghu Xia, Hongdan Fan, Qian Zhang and Xu Li, “An Improved Apriori Algorithm Based on Association Analysis,” ICNDC 2012, 3rd IEEE International Conference, pp208-211.
[8]. Rui Chang and Zhiyi Liu , “ An Improved Apriori Algorithm,” ICEOE 2011, IEEE International Conference, vol. 1, pp v1- 476 -v1-478.
[9]. Sanjeev Rao and Priyanka Gupta,” Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm,” IJCST Vol. 3, Issue 1, Jan. - March 2012.