Review paper on Mining Association rule and frequent patterns using Apriori Algorithm

Peeyush Kumar Shukla4
Department of Computer Science and Engineering, SRCEM College, Palwal, Affiliated to MD University, Rohtak (Haryana), India
Email:shukla.piyush143@gmail.com

ABSTRACT

Because of speedy development at worldwide information several mining algorithms have been developed over the years. Apriori Algorithm is one of the most productive algorithm which is used to excerpt frequent patterns from huge database likely tera and penta bytes of data and find out the appropriate association rule for distinguish the knowledge. It basically needs two important things: minimum support and minimum confidence. Firstly, we check whether the frequent item are greater than or equal to the minimum support threshold value and we find the frequent item sets respectively. Secondly, the minimum confidence constraint is used to generate association rules according to the minimum confidence threshold value. In this paper we propose an algorithm (Apriori) used to mine the frequent patterns and association rules. The Apriori algorithm generates candidate set during each step. It abbreviates the item sets by dispose the infrequent item sets that exactly not match the minimum threshold from the candidate sets. To avoid the propagation of candidate set which is expensive the FP Growth algorithm is used to mine the item set. The FP Growth does not generate the candidate set instead it generates an optimized data set that is FP tree from the dataset.

Keywords: Data mining, Association rules, frequent item sets, Apriori algorithm, minimum support, minimum confidence.

IJCIR Reference Format:
Peeyush Kumar Shukla. Review paper on Mining Association rule and frequent patterns using Apriori Algorithm. Vol. 10, Issue.1pp 32 - 40. http://www.ijcir.org/volume 10-issue 2/article 4.pdf.

4 Author's Address: Peeyush Kumar Shukla, Department of Computer Science and Engineering, SRCEM College, Palwal, Affiliated to MD University, Rohtak (Haryana), India *shukla.piyush143@gmail.com*

International Journal of Computing and ICT Research, ISSN 1818-1139 (Print), ISSN 1996-1065 (Online), Vol.10, Issue 1, pp. 32-40, June 2016.

1.INTRODUCTION

Frequent patterns are patterns like as item sets, subsequences or substructures that come along in a data set subsequently. Behalf of the transactional database, we can suppose the behavior of the products purchased by the customers. For example a set of items Mobile and Sim card that appear frequently as well as together in a transaction set is a frequent item set. Subsequences means if a customer buys a Mobile he must also buy a Sim card and then head phone etc. From the overall structure of the database these transactions are occurs sequentially is called sequential patterns. The Substructure is concerned to different structural forms such as sub graphs, sub trees which may be manipulate along with item sets or sequences.

Data mining is manipulated to work with amount of data stored in the database, to take out the required information and knowledge [1]. Data mining has various strategies to perform data extraction. Association proficiency is the most effective data mining technique among them. It encounters concealed or craved pattern among the huge amount of data. It is also responsible for finding co-relationships among different data attributes in a large set of items in a database. Since its introduction, this method has acquired lot of attention. Author of [1] has examined that an association analysis [5] is the research of hidden pattern or clause that occur repeatedly common in an applied dataset. Association rule acquire relations and interconnection among data and datasets given. Such association's rules are calculated from the data with help of the concept of probability.

1.1. Basic Concept: The main approach is to finding association rules to interrupt the problem in two parts:

a. Find all frequent item sets

b. Generate strong association rules from frequent items.

Finding all frequent item sets is a difficult task where generating strong association rules are not too much costly.

The problem of association rule mining from frequent items is defined as:

Let $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots\ldots\ldots\ldots \boldsymbol{x_n}\}$ be a set of n binary attributes called *items*.

Let $\boldsymbol{Y} = \{\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots\ldots\ldots\ldots \boldsymbol{y_n}\}$ be a set of transactions called the *database*.

Each transaction in $\boldsymbol{Y}$ has a specific transaction ID and contains a subset of the items in $\boldsymbol{X}$.A rule is defined as an implication of the form:

$$A \rightarrow B$$

An example rule for the supermarket could be **{tea ,sugar→ milk}** meaning that if tea and sugar are bought then customer also buy milk.

2. APRIORI ALGORITHM

The algorithm [5] is designed to find associations in sets of data in a database. Apriori is a definitive algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length $k - 1$. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all $2^{|S|} - 1$ of its proper subsets.

2.1. ITEMSET

Item set is collection of items in a database which is denoted by D= {x1, x2,……..…,xn},Here 'n' is the number of items.

2.2. CANDIDATE ITEMSET

Candidate item sets are items which are only to be considered for the processing. Candidate item set are all the possible combination of item set. It is usually denoted by 'Ci' where 'i' indicates the i-item set.

2.3. TRANSACTION

Transaction is a database entry which contains collection of items. Transaction is denoted by and **T**⊆D.
A transaction contains set of items T= {x1, x2, ……., xn}.

2.4. MINIMUM SUPPORT

Minimum support is basically condition which should be satisfied by the given items so that further processing of that item can be completed. Minimum support can be considered as a condition which helps in removal of the in-frequent items in any database. Usually the Minimum support is given in terms of percentage.

2.5. FREQUENT ITEMSET

Frequent item set is commonly large item set i.e. the item sets which satisfies the minimum support threshold value are known as frequent item sets. It is usually denoted by 'Li' where 'i' indicates the i-item set.

2.6. CONFIDENCE

Confidence indicates the certainty of the rule. This argument lets us to count how often a

transaction's item set couple with the left side of the implication with the right side. The item set which does not satisfies the above condition can be discarded. Consider two items X and Y. To calculate confidence of X->Y the following formula is used, Conf(X->Y) = (number of transactions containing both X& Y) (Transactions containing only).

3. LITRATURE REVIEW

Various algorithms for mining association rules and frequent patterns from relational database have been done since long before. Association rule mining was first presented at 1993 by Rakesh Agrawal[3], T. Imielinski, and A. Swami [3].After sometime the Boolean Association rule of mining frequent item set is proposed by Srikant in 1994.The core principles of this theory are the subsets of frequent item sets are frequent item sets and the supersets of infrequent item sets are infrequent item sets. This theory is regarded as the most typical data. A new implementation of mining of frequent closed item set is introduced by Pasquier in 1991. Further a new approach is published for mining of maximum frequent item set by Bayarado,1998.At last Srikant again evaluate some improvements in mining Fuzzy association rules in1996.
Association rule mining proceeds on two main steps. The first step is to find all item sets with adequate supports and the second step is to generate association rules by combining these frequent or large item-sets [8][9][10]. In the traditional association rules mining [2][4], minimum support threshold and minimum confidence threshold values are assumed to be available for mining frequent item sets, which is difficult to be set without specific knowledge; users have difficulties in setting the support threshold to obtain their required results. To use association rule mining without support threshold another constraint such as similarity or confidence pruning is usually introduced.

4. PROPOSED ALGORITHM

Association rule mining is an important task in data mining. Association rules are frequently used by retail stores to assist in marketing, advertising, floor placement and inventory control. In analyzing market basket analysis, people often use Apriori Algorithm, but Apriori generates large number of frequent item sets.

Algorithm for Apriori algorithm:

Apriori (T, ε)

$L_1 \leftarrow \{$*large 1-itemsets that appear in more than* ε*transactions* $\}$

$k \leftarrow 2$

while $L_{k-1} \neq \varnothing$

$C_k \leftarrow$ *Generate*(L_{k-1})

for transactions $t \in T$

$C_t \leftarrow$ Subset(C_k,t)

for candidates $c \in C_t$

$\text{count}[c] \leftarrow \text{count}[c] + 1$

$L_k \leftarrow \{c \in C_k | \ \text{count}[c] \geq \varepsilon\}$

$k \leftarrow k + 1$

return $\bigcup_k L_k$

Let me give you an example to explain it. Suppose you have records of large number of transactions at a shopping center, Wall-mart, showrooms etc. as follows: The Transaction id and sales item details are following:

Transaction ID	**Items Purchased**
T1	{Mango, Orang, Namkeen, Kit Kat, Eggs, Burger}
T2	{Dahi, Orange, Namkeen, Kit Kat, Eggs, Burger}
T3	{Mango, Apple, Kit Kat, Eggs}
T4	{Mango, Ugli, Corn, Kit Kat, Burger}
T5	{Corn, Orange, Onion, Kit Kat, Ice-cream, Eggs}

Now, we follow a simple golden rule: we say an item set is frequently bought if it is purchased at least 60% of times. So for here it should be bought at least 3 times.

Transaction ID	**Items Purchased**
T1	{M, O, N, K, E, B }
T2	{D, O, N, K, E, B}

T3	{M, A, K, E}
T4	{M, U, C, K, B }
T5	{C, O, O, K, I, E}

Step 1: Count the number of transactions in which each item occurs, Note "O=Orange" and "O=Onion" is bought 4 times in total, but, it occurs in just 3 transactions.

Step 2: Now remember we said the item is said frequently bought if it is bought at least 3 times. So in this step we remove all the items that are bought less than 3 times from the above table.

Item	Number of transactions
M,O,B	3
E	4
K	5

This is the single items that are bought frequently. Now let's say we want to find a pair of items that are bought frequently. We continue from the above table (Table in step 2)

Step 3: We start making pairs from the first item, like MO,MK,ME,MB and then we start with the second item like OK,OE,OB. We did not do OM because we already did MO when we were making pairs with M and buying a Mango and Onion together is same as buying Onion and Mango together. After making all the pairs we get,

Item Sets
MO,MK,ME,MB,OK,OE,OE,OB,KE,KB,EB

Step 4: Now we count how many times each pair is bought together. For example M and O is just bought together in {M,O,N,K,E,B}.While M and K is bought together 3 times in {M,O,N,K,E,B}, {M,A,K,E} And {M,U,C, K, B}

After doing that for all the pairs we get-

Item Sets	Number of transactions

MO	1
ME,MB,OB,EB	2
MK,OK,OE,KB	3
KE	4

Step 5: Golden rule to the rescue. Remove all the item pairs with number of transactions less than three and we are left with.

Item Pairs	Number of transactions
MK,OK,OE,KB	3
KE	4

These are the pairs of items frequently purchased together. Now let's say we want to find a set of three items that are brought together. We use the above table (table in step 5) and make a set of 3 items.

Step 6: To make the set of three items we need one more rule (it's termed as self-join).It simply means, from the Item pairs in the above table, we find two pairs with the same first Alphabet, so we get

· OK and OB, this gives OKB

· KE and KB, this gives KEB

Then we find how many times O,K,E are bought together in the original table and same for K,E,B and we get the following table

Item Set	Number of transactions
OKE	3
KEB	2

While we are on this, suppose you have sets of 3 items say PQR, PQS, PRS, PRT, QRS and you want to generate item sets of 4 items you look for two sets having the same first two alphabets.

· PQR and PQS -> PQRS

· PRS and PRT -> PRST

And so on … In general you have to look for sets having just the last alphabet/item different.

Step 7: So we again apply the golden rule i.e. the item set must be purchased together at least 3 times which leaves us with just OKB, Since KEB are bought together just two times .Thus the set of three items that are bought together most frequently are O,K,E.

5. CONCLUSION

This paper is an attempt to use data mining as a tool used to find the frequent pattern and its association rule of different item sets. An Apriori Algorithm may play an vital role for finding these patterns from huge database so that various sectors can make better business decisions especially in the retail sector. Apriori algorithm may find the tendency of a customer on the basis of frequently purchased item-sets. There are wide range of industries have deployed successful applications of data mining. Data mining in retail industry can be deployed for market campaigns, to target profitable customers using reward based points. The retail industry will gain, sustain and will be more successful in this competitive market if adopted data mining technology for market campaigns.

6. REFERENCES

[1] Pranay Bhandari, K. Rajeswari, Swati Tonge, Mahadev Shindalkar improved Apriori Algorithms.

[2] M. Ashrafi, D. Taniar, and K. Smith "A New Approach of Eliminating Redundant Association Rules". Lecture Notes in Computer Science.

[3] Agrawal R, Imielinski T, Swami A, ―Mining Association Rules between Sets of Items in Large Databases, ‖In: Proc of the ACM SIGMOD International conference on Management of Data.

[4] P. Tang, M. Turkia "Para llelizing frequent item set mining with FP trees".

[5] International Institute for Infrastructural, Hydraulic, and Environmental Engineering, ’’Predictive Data Mining.

[6] R Agrawal and R Srikant ― Fast Algorithm for Mining Association Rules.

[7] Data Mining Introductory and advanced topics by Margaret H.Dunham and Data Mining concepts and Techniques by jiawei Han and Michelin kamber second edition.

[8] Literature Review: Data mining, http://nccur.lib.nccu.

[9] H. Mahgoub,"Mining association rules from unstructured documents" in Proc. 3rd Int. Conf. on Knowledge Mining .

[10]] S. Kannan, and R. Bhaskaran "Association rule pruning based on interestingness measures with clustering". International Journal of Computer .