ISSN: 2321-7782 (Online)
Volume 3, Issue 3, March 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

# *Improving Association Rule Mining in Text Datasets by Prior Knowledge and Iterative Approach – Apriori Algorithm*

**Naziya K. Pathan**[1]
Author
M.Tech. Department of Computer Science & Engineering
J.D. College of Engineering, (M.S)
Nagpur, India

**Mirza Moiz Baig**[2]
Department of Computer Science & Engineering
J.D. College of Engineering, (M.S)
Nagpur, India

***Abstract:*** ***The frequent pattern mining has become an important data mining task and a focused theme in Data mining research. The key problem is how to find useful hidden patterns and to discover association rules between items in a large database of sales transactions, for better business applications. For the solution of these problems, the Apriori algorithm is one of the most popular data mining approaches for finding frequent item sets from a transaction dataset and derives association rules by prior knowledge and iterative approach. Finding such frequent patterns plays essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks as well. Rules are the discovered knowledge from the data base. Finding frequent item set (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. The paper illustrating apriori algorithm on simulated database and finds the association rules on different confidence value. Scale-up experiments show that Apriori scales linearly with the number of transactions. Apriori also has excellent scale-up properties with respect to the transaction size and the number of items in the database.***

***Keywords: Apriori, association rules, support value, confidence value, frequent pattern mining, itemsets.***

## I. INTRODUCTION

Frequent pattern mining is indeed rich. It is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses. Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored through the large amount of customer daily transaction details referred as basket data, due to progress in bar-code technology, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customer shopping behaviour analysis. A record in such data typically consists of the transaction date and the items bought in the transaction. Successful organizations view such databases as important pieces of the marketing infrastructure. Data mining is the extraction of hidden predictive information from very large databases. Data mining tools predict future trends and behaviours, helps organizations to make proactive knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools have the answer of this question. Today retailer is facing dynamic and competitive environment on global platform. How can we take advantage of user preferences or constraints to speed up the mining process? Retail industry is looking strategy where they can target right customers who may be profitable to their business. Market Basket Analysis is the process which analyses customer buying habits by finding associations between

different items that customers place in their "shopping baskets". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.

Those traditionally methods were lot of time consuming to resolve the problems or decision making for profitable business. Data mining prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The field of data mining have been prospered and posed into new areas such as manufacturing, insurance, medicine, etc. Hence, this paper reviews the various trends of data mining and its relative applications from past to present and discusses how effectively can be used for targeting profitable customers in information driven and better marketing campaigns that enable marketers to develop and implement customized marketing programs and strategies.

## II. Related Work

If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently *associated* or purchased together. These patterns can be represented in the form of association rules. For example, the information that customer who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule (1) below:

*computer* ➔ *antivirus_software* [ *support=2%, confidence=60%* ]

Association rule mining is interested in finding frequent rules that define relations between related frequent items in databases, and it has two main measurements of rule interestingness: support and confidence values. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule (1) means that 2% of all the transactions under analysis show that computer and antivirus software purchased together. A confidence of 60% means that 60% of customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.

The *frequent itemsets* is defined as the itemset that have support value greater than or equal to a minimum threshold support value, and *frequent rules* as the rules that have confidence value greater than or equal to minimum threshold confidence value. Association Rule Mining is all about finding all rules whose support and confidence exceed the threshold, minimum support and minimum confidence values. Itemsets with minimum support are called large itemsets, and all others small itemsets.

Association rule mining proceeds on two main steps. The first step is to find all itemsets with adequate supports and the second step is to generate association rules by combining these frequent or large itemsets. In the traditional association rules mining, minimum support threshold and minimum confidence threshold values are assumed to be available for mining frequent itemsets, which is hard to be set without specific knowledge; users have difficulties in setting the support threshold to obtain their required results. Setting the support threshold too large, would produce only a small number of rules or even no rules to conclude. In that case, a smaller threshold value should be guessed (imposed) to do the mining again, which may or may not give a better result, as by setting the threshold too small, too many results would be produced for the users, too many results would require not only very long time for computation but also for screening these rules. That would explain the need to develop an algorithm to generate a minimum support, and minimum confidence values depending on the datasets in the databases.

## III. PROPOSED SYSTEM

This paper proposes the IND-OCPA-P model to analyze the security of the proposed EOB and the encryption schemes supporting an efficient range query over encrypted data.

***Data mining concepts:***

Associations and item-sets: It as denoted as A→B

If A is true then B will also true. Example: People buying two-wheeler also buys helmet.

A= Buying two-wheeler B= Buying helmet

Using this association rule can predict that if A is true then B also true. For any rule if A→B and B→A, then A and B are called an "interesting item-set"; Example: People buying shampoo also buy conditioner. People buying conditioner also buy shampoo.

Sales Transaction Table: From the market basket analysis of the set of products in a single transaction. Discovering for example, that a customer who buys shoes is likely to buy socks Shampoo→ Conditioner

Transactional Database: The set of all sales transactions is called the population. The representation of the transactions in one record per transaction. The transaction is represented by a data Tuple.

TABLE1: TRANSACTION DATABASE

| TID | List of item_IDs and items |
|---|---|
| T100 | I1- Shampoo, I2 - Conditioner, I5 – Washing Powder |
| T101 | I2 - Conditioner , I4 – Bathing Soap |
| T102 | I2 - Conditioner, I3 – Toothpaste |
| T103 | I1 - Shampoo, I2 - Conditioner, I4 – Bathing Soap |
| T104 | I1 - Shampoo, I3 - Toothpaste |
| T105 | I2 - Conditioner, I3 - Toothpaste |
| T106 | I1 - Shampoo, I3 - Toothpaste |
| T107 | I1 - Shampoo, I2 - Conditioner, I3 – Toothpaste, I5 – Washing Powder |
| T108 | I1 - Shampoo, I2 - Conditioner, I3 - Toothpaste |

Support and Confidence: Any given association rule has a support level and a confidence level. The rule A→B holds in the transaction set D with support s, where *s* is the percentage of transactions in D that contains (A U B), this is taken to be the propability, P(A U B) i.e. $support\ (A \rightarrow B) = P\ (A \cup B)$. (1)

The rule A→B has confidence *c* in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability,

P(B|A) i.e. $confidence(A \rightarrow B) = P(B|A)$. (2)

The transaction table given above is showing the item sets purchased by the customer in a period of time. The support for the item sets shampoo and conditioner means a customer who purchased Shampoo also purchased the conditioner is given below. The support for nine transactions where shampoo and conditioner occur together is two. Support for {shampoo,

conditioner} = 2/9 = 22%. This means the association of data set or item set, the shampoo and conditioner brought together with 22% support. Confidence for shampoo→conditioner = 4/6 = 0.66. This means that a customer who buy shampoo then there is a confidence of 66% that it also buy conditioner.

***Mining for frequent item-sets: The Apriori Algorithm:***

Apriori is a seminal algorithm for finding frequent itemsets for Boolean association rules. The name of the algorithm based on the fact that the algorithm uses *prior knowledge* of frequent itemset properties. Apriori employs an *iterative approach* known as a level-wise search, where *k*-itemsets are used to explore (*k*+1)-itemsets.

Given minimum required support *s* as interestingness criterion:

1. Search for all individual elements (*k*-element item-set) that have a minimum support of s.
2. Repeat
    A. From the results of the previous search for *k*element itemset, search for all *k*+1 element itemsets that have a minimum support of item-set.
    B. This becomes the set of all frequent (*k*+ 1) itemsets that are interesting.
    C. Until item-set size reaches maximum.

Apriori property: *All nonempty subsets of a frequent itemset must also be frequent.* To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, is used to reduce the search space. This property belongs to a special category of properties called *antimonotone* in the sense that *if a set cannot pass a test, all of its supersets will fail the same test as well.*

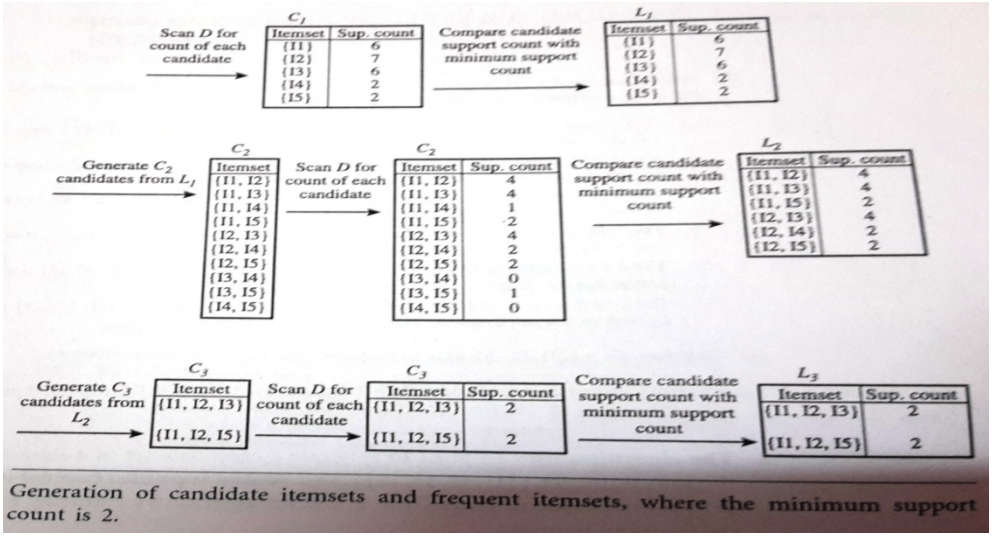
A two-step process is followed, consisting of *join* and *prune* actions.

1. *The join step:* To find the $L_k$, a set of candidate k-itemsets is generated by joining $L_{k-1}$ with itself. This set of candidates is denoted $C_k$.
2. *The prune step:* $C_k$ is a superset of $L_{k,}$ i.e., its members may or may not be frequent, but all of the frequent *k*-itemsets are included in $C_{k.}$ $C_k$ however, can be huge, and so this could involve heavy computation. To reduce the size of $C_k$, the Apriori property is used.

By using the consumer database given in table1, Let's illustrate the process of apriori with an example for finding frequent itemsets in D, let takes the consumer database which is showing the number of itemsets purchased by the consumers from a retailor. There are nine transactions in database, i.e. |D|=9. Let Minimum support count required is 2, i.e. *min_sup*=2 Single item like shampoo, conditioner etc. in the given database every item occurs two or more time than the minimum support or minimum support threshold value is 2. Now focus on interestingness of the single itemsets.

1. In the first iteration of the algorithm, each item is a member of the set of candidate 1-itemsets, $C_1$. The algorithm simply scans all of the transactions in order to count the number of occurrences of each item.
2. The set of frequent 1-itemsets, $L_1$, can then be determined. It consists of the candidate 1-itemsets satisfying minimum support. In our example, all of the candidates in $C_1$ satisfy minimum support.
3. To discover the set of frequent 2-itemsets, L2, the algorithm uses the join $L_1$ ⍰ $L_1$ to generate a candidate set of 2-itemsets, $C_2$. $C_2$ consists of $L_1$ 2-itemsets. Note that no candidates are removed from $C_2$ during the prune step because each subset of the candidate is also frequent.




Generation of candidate itemsets and frequent itemsets, where the minimum support count is 2.

4. Next, the transactions in D are scanned and the support count of each candidate itemset in $C_2$ is accumulated, as shown in the middle table of the second row in Fig 1.

5. The set of frequent 2-itemsets, $L_2$, is then determined, consisting of those candidate 2-itemsets in $C_2$ having minimum support.

6. The generation of the set of candidate 3-itemsets, $C_3$, is detailed in Fig. 1. From the join step we first get $C_3$= $L_2$⋈$L_2$={{I1,I2,I3},{I1,I2,I5},{I1,I3,I5},{I2,I3,I4},{I2,I3,I5},{I2,I4,I5}} Based on the Apriori property, the four later candidates cannot possibly be frequent. We therefore remove them from $C_3$. The resulting pruned version of $C_3$ is shown in the first table of the bottom row of Fig 1.

7. The transactions in D are scanned in order to determine $L_3$, consisting of those candidate 3-itemsets in $C_3$ having minimum support.

8. The algorithm uses $L_3$⋈$L_3$ to generate a candidate set of 4-itemsets, $C_4$. Although the join results in {{I1,I2,I3,I5}}, this is pruned because its subset {{I2,I3,I5}} is not frequent. Thus $C_4$= null, and the algorithm terminates, having found all of the frequent itemsets.

So association rules which frequently used and follow the minimum confidence . So the research part of this paper is this by changing the value of minimum confidence, gives different association rules. the value of minimum confidence is high then rules filtered more accurately.

## IV. Conclusion

This paper is an attempt to use data mining as a tool used to find the hidden pattern of the frequently used item-sets. An Apriori Algorithm plays an important role for finding these patterns from large databases so that various sectors can make better business decisions especially in the retail sector. Apriori algorithm is useful for the discovery of such associations that can help retailers develop marketing strategies by gaining insight into which items are frequently purchase together by customers. There are wide ranges 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.

## V. FUTURE WORK

The major extensions of frequent pattern mining includes the following:

(1) Incorporating taxonomy in items: Use of Taxonomy makes it possible to extract frequent itemsets that are expressed by higher concepts even when use of the base level concepts produces only infrequent itemsets.

(2) Incremental mining: In this setting, it is assumed that the database is not stationary and a new instance of transaction keeps added. The algorithm updates the frequent itemsets without restarting from scratch.

(3) Using numeric valuable for item: When the item corresponds to a continuous numeric value, current frequent itemset mining algorithm is not applicable unless the values are discretized. A method of subspace clustering can be used to obtain an optimal value interval for each item in each itemset.

(4) Using other measures than frequency, such as information gain or $\chi 2$ value: These measures are useful in finding discriminative patterns but unfortunately do not satisfy anti-monotonicity property. However, these measures have a nice property of being convex with respect to their arguments and it is possible to estimate their upper bound for supersets of a pattern and thus prune unpromising patterns efficiently.

(5) Using richer expressions than itemset: Many algorithms have been proposed for sequences, tree and graphs to enable mining from more complex data structure.

(6) Closed itemsets: A frequent itemset is closed if it is not included in any other frequent itemsets. Thus, once the closed itemsets are found, all the frequent itemsets can be derived from them.

## ACKNOWLEDGMENT

I express my deep sense of gratitude to my Institution, J D College of Engineering and Management, NAGPUR for providing an opportunity in fulfilling the most cherished desire for reaching my goal. I extend the immense gratitude to the Head of the Department Prof. Shrikant V. Sonekar for his motivation, inspiration, and encouragement for the completion for my project. The valuable and unflinching requital support in this Endeavor of Prof. Mirza Moiz Baig, M.Tech. Coordinator, whose support & guidance was immeasurable to the completion of this project. I take this opportunity to express my hearty thanks to Prof. Mirza Moiz Baig, His extreme energy, creativity and excellent coding skills have always been a constant source of motivation for me in the completion of my project work.

## References

1 Jugendra Dongre and Gend Lal Prajapati, 'The Role of Apriori Algorithm for Finding the Association Rules in Data Mining', ICICT 2014.

2 Agrawal R, Srikant R (1 994), 'Fast algorithms for mining association rules.' In: Proceedings of the 20thVLDB conference, pp 4S7-499

3 Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

4 Huangshan, P. R. China, August 21-23, 2009, 'A Fast Incremental Clustering Algorithm', pp. 175-178, Proceedings of the 2009 International Symposium on Information Processing (ISIP'09).

5 Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman 'An Efficient k-Means Clustering Algorithm: Analysis and Implementation', IEEE Transactions on pattern analysis and machine intelligence, Vol. 24, No.7, July 2002.

6 Sergio M. Savaresi and Daniel L. Boley, 'A comparative analysis on the bisecting K-means and the PDDP clustering algorithms.', Intelligent Data Analysis 6(2002) 1-18 IDA174.

7 Yuepeng Cheng, Tong Li and Song Zhu, 'A Document Clustering Technique Based on Term Clustering and Association Rules', ©2010 IEEE

8 Vipin Kumar and Xindong Wu, 'Top 10 algorithms in data mining', Published online 4 December 2007 © Springer-Verlag London Limited 2007

9 Literature Review: Data mining, http://nccur.lib.nccu.edu. twlbitstream/1 40.1 I 9/3523 I/S/35603 I OS.pdf, retrieved on June 2012.

10 'The 6 biggest challenges retailer Face today", www.onStepRetail .com. retrieved on June 2011.

11 Berry, M. J. A. and Linoff, G. Data mining techniques for marketing, sales and customer support, USA: John Wiley and Sons,1997

12 Fayyad, U. M; Piatetsky-Shapiro, G. ; Smyth, P.; and Uthurusamy, R.1996. Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press.

13 Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

14 H. Mahgoub,"Mining association rules from unstructured documents" in Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic, Aug. 25- 27, 2006, pp. 1 67-1 72.

15 S. Kannan, and R. Bhaskaran "Association rule pruning based on interestingness measures with clustering". International Journal of Computer Science Issues, IJCSI, 6(1 ), 2009, pp. 35-43 .

16 M. Ashrafi, D. Taniar, and K. Smith "A New Approach of Eliminating Redundant Association Rules". Lecture Notes in Computer Science, Volume 31 S0, 2004, pp. 465 -474.

## AUTHOR(S) PROFILE

**Ms. Naziya Pathan,** received the B.E. degree in Computer Technology and MBA in Finance and Human Resource.