Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.343 – 345

Information System on Market Basket Analysis

Shruti Kawale[1], Rajni Tetwar[2], Nirmala Suryavanshi[3], Prachi Wankhede[4], Prof. Neha Titarmare[5]

[1,2,3,4,5] Department of Computer Science and Engineering, RTMNU, India.

[1] shruti.kawale05@gmail.com; [2] rajnitetwar@gmail.com; [3]nsuryavanshi.10r@gmail.com; [4]prachiwankhede2012@gmail.com; [5] nehatitarmare@gmail.com

Abstract— This paper gives description about market basket analysis (MBA) that is exploring the consumer purchase behaviour. This is a very useful analytical process of discovering associations and correlation among the sets of items. The method used here has iterative procedure which gives us significant association between the items present in the data set. Using these results we can know about the patterns in consumer purchasing behaviour which would help the retailers to improve their marketing strategies.

Keywords— Market Basket Analysis (MBA); transactional dataset; association rules

INTRODUCTION

Market Basket Analysis (MBA) is a very prominent technique of data mining which is widely used in identifying products and contents that go well together [1], [8]. Now a day, there is a boom in online shopping of products due to rapid increase in e-commerce websites, and the ease in using them. This has made an enormous increase in transactional datasets. The availability of such datasets has promoted the researches to analyse the data and use it for productivity of the retailers. The increasing number of e-commerce websites has lead to development of competition between the retailers, thus this analysis of consumer purchase behaviour has become a point of prime importance for them. This kind of analysis has helped them to gain the competitive advantage. To adapt to the needs of consumers retailers need to know the demands and expectations, which can be very well known by performing affinity analysis. This also helps us to know who the consumers are, understand why they make certain purchases, and gain insight about its merchandise. There are various algorithm and methods for these analysis the most commonly used is Apriori algorithm. This paper presents the process of the algorithm and the empirical results found out by them.

This algorithm identifies the frequent patterns present and gives the results based on them. This is an iterative process where the results are given in a step by step process. The frequent itemsets are used for associating items together. The infrequent items are completely pruned out from the frequent sets, thus we have a very reliable technique to be sure of frequent itemset results.

APRIORI ALGORITHM

This algorithm is widely used in association rule mining technique [2]. It was proposed by R. Agarwal and R. Srikant. It deals with the transactional data; this data has entry of every transaction one by one. This algorithm has basic steps of finding frequent itemsets, association, and pruning. The most basic principles for apriori to follow are: [7]

1) All subset of a frequent itemset must be frequent.
2) No super set of any infrequent itemset should be tested or generated.
3) If an itemset is infrequent, then all of its super sets must also be infrequent.

This algorithm works iteratively step by step, ever current state is generated using the previous step of it. It makes multiple pass over the database to reach the definite frequent set, which is used in the further step. The first step is to find out frequent sets in gradual increment form, and then we can associate them together using association rule. Before associating the items together we have to perform the step of pruning which is used to identify the item which is infrequent, if found that subset is pruned from the set.

A. Terminologies used in apriori algorithm are as follows:

1) Itemsets- This is the collective set of items which are present in the dataset, example I= $\{i_1,i_2,i_3 \ldots i_n\}$.this contains every item purchased by the consumer which is irrespective of it frequency.
2) Threshold value- This is the minimum value which is used to limit the frequency item. Frequent itemset- This is a collective item set which is calculated iteratively and there every item present has the support count, above or equal to minimum support count. Eventually, we get a set of frequently occurring itemset. Example, F= $\{i_1,i_2,i_4\}$ This is very important that the frequent itemset obtain must be the subset of itemset (F subset I) [4].
3) Support- This is the total count of occurrence of an individual item present in the dataset. In the case of apriori algorithm we also calculate support for collective items together, example we can calculate support for i_1 and i_2 together. Another alternative is to calculate support relatively by considering the ratio of total number of occurrences of the item to total number of transactions. (Frequency (i_1,i_2) / n) where 'n' is total number of transactions [6].
4) Confidence- This is the probability of an item (i_1) to be purchased with another item(i_2). This is very important that the subsets used for calculate confidence should be present in frequent itemset. If the calculated support and confidence is above threshold then set is accepted to be beneficial. The formula for confidence is the ratio of total number of occurrences of items together to total number of base item. (frequency (i_1,i_2) / frequency(i_1)) [6].

ALGORITHM

The Apriori algorithm is a simple algorithm for mining frequent itemset for Boolean association rules. It is one of the classical algorithms for extracting frequent patterns. It works on transactional data sets that contain information of each transaction row by row respectively; every row has data relevant to transaction of individual customer. It works on multiple pass method which is an iterative process and thus generates step by step results every time a pass is executed. This algorithm is called apriori, because every time it when it generates results it uses the most recent (current) knowledge it has with it we do not provide it with any extra prior knowledge, to generate results. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger itemsets as long as those itemsets appear sufficiently up to the support count, it also makes sure that down closure property is followed along with it [6].

A. Steps of apriori algorithm:

1) Generating frequent item list

Here first we have to generate a candidate set before generation of frequent item list. The database is scanned and support of items is counted from the transaction this is called the candidate set generation. After generation of candidate set we have to make use of it to obtain frequent itemset list. So for this we calculate a k-itemset candidate list and then a k-itemset frequent list, this list contains all the items which are above or equal to the minimum support count and should be present all last frequent itemset [3].Now this frequent k-itemset is used to generate a k+1-itemset candidate list. And then we again use this k+1- candidate list to generate k+1 frequent list. This is the nature of apriori algorithm to generate frequent items from the data set. This goes on till we get an empty set of candidate or frequent items. The most important point here is that the final set as well the set of frequent list which were gathered from the candidate set previously should have support count greater than or equal to minimum support threshold value.

2) Associating frequent items

This step completely deals with the set of items which are frequent, and this step has nothing to deal with the candidate sets used previously as they do not necessarily have items above support count. Here we use the final frequent list to make non-empty subset from individual items. The following example explains this, F= $\{i_2,i_4,i_7\}$ be the set of frequent item set, therefore the non-empty subsets are $\{i_2,i_4\}$, $\{i_2,i_7\}$, $\{i_4,i_7\}$, $\{i_2\}$, $\{i_4\}$, $\{i_7\}$. The next step is to associate all of them together for example relation between $\{i_2,i_4\} \rightarrow \{i_2\}$ "in press" [5]. Using these non-empty subsets we associate them together one by one and then calculate the confidence for them. If this calculated value comes above or equal to the minimum confidence value then the pair of items is considered to be profitable if used together for marketing purpose. So this is how apriori algorithm helps us to know the association between the items or more specifically between the frequent items. But if the calculated confidence is less than the threshold value then the pair of items is neglected. This value of confidence is calculated for every association possible with the subsets irrespective whether the confidence will be above or below threshold. This is how we get the sets of beneficial pairs of items which were purchased together by the consumers. Thus this is a step by step process to be followed to get the final results regarding the motive o f market basket analysis.

CONCLUSIONS

Apriori algorithm is one of the most prominent algorithm which uses association rule mining to generate frequent patterns. It also successfully gives the results of frequent items, which can be used by the retailers to improve their marketing strategy. One can rely on the results generated by the algorithm as it takes into confederation each and every transaction of the transactional database.

REFERENCES

[1] Charu C. Aggarwal, Cecillia Procopiuc and Philip S.Yu, "Finding Localized Associations in Market Basket Data," IEEE, Jan/Feb 2002.
[2] Charu c. Aggarwal and Philip s. Yu, "A New Framework for Itemset Generation," IBM T J Watson Research Center, Aug 1998.
[3] Rakesh Agrawal, Tomaz Imielinski and Arun Swami, "Mining Association Rules between Sets of Items in Large Database," IBM Almaden Research Center.
[4] Soumen Chakrabharti, Sunita Sarawagi and Byron Dorn, "Mining Surprising Patterns using Temporal Description Length," IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.
[5] Charanjeet Kaur, "Association Rule Mining using Apriori Algorithm: A Survey," ISSN: 2278 - 1323 volume 2 Issue 6, June 2013.
[6] Rakesh Agrawal and Ramakrishna Srikant, "Fast Algorithm for Mining Association Rules," IBM Almanden Research Canter 650 Harry Road, San Jose, CA 95120.
[7] Goswami D.N., Chaturvedi Anshu and Raghuvanshi C.S, "An Algoritham for Frequent Pattern Mining Based on Apriori," IJCSE volume 02, No. 04, 2010, 942-947.
[8] Meenakshi Malik, Mamta and R.P.Agarwal, "A survey On Association Rule Mining,'' IJREAS, vol.05, ISSUE 06, June 2015.

