International Journal of Engineering Research and General Science Volume 3, Issue 1, January-February, 2015
ISSN 2091-2730

Review paper on finding Association rule using Apriori Algorithm in Data mining for finding frequent pattern

Krutika. K .Jain, Anjali . B. Raut

ME Student, Computer science & Enginering Department, HVPM College,India
Head of the Computer science & Enginering Department, HVPM College, Amravati, India
Email: krutikajain20@gmail.com

Abstract— Because of the rapid growth in worldwide information, efficiency of association rules mining (ARM) has been concerned for several years. Association rule mining plays vital part in knowledge mining. The difficult task is discovering knowledge or useful rules from the large number of rules generated for reduced support In this paper, based on the Apriori algorithm association rules is based on interestingness measures such as support, confidence and so on. Confidence value is a measure of rule's strength, while support value corresponds to statistical significance. Traditional association rule mining techniques employ predefined support and confidence values. However, specifying minimum support value of the mined rules in advance often leads to either too many or too few rules, which negatively impacts the performance of the overall System. In this algorithm, we will create association rules depending upon the dataset available in the database. The algorithm majorly works on finding the minimal confidence and 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..

Keywords— Data Mining, e-Commerce, Apriori algorithm, association rules, support, confidence, retail sector.

INTRODUCTION

Today retailer is facing dynamic and competitive environment on global platform and competitiveness retailers are seeking better market campaign [I][2]. Retailer are collecting large amount of customer daily transaction details. This data collection requires proper mechanisms to convert it into knowledge, using this knowledge retailer can make better business decision. Retail industry is looking strategy where they can target right customers who may be profitable to their business . Data mining is the extraction of hidden predictive information from very large databases. It is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [4][6]. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions [7] [5]. 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.

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. From the last decade data mining have got a rich focus due to its significance in decision making and it has become an essential component in various industries [7][5].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 campaigns.

LITERATURE REVIEW

Algorithms for mining association rules from relational data have been done since long before. Association rule mining was first presented at 1993 by R. Agrawal, T. Imielinski, and A. Swami [3]. Association rule mining is interested in finding frequent rules that

define relations between emulated frequent items in databases, and it has two main measurements: support and confidence values. The frequent item sets is defined as the item set 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. These threshold values are traditionally assumed to be available for mining frequent item sets. Association Rule Mining is all about finding all rules whose support and confidence exceed the threshold, minimum support and minimum confidence values.

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][lO]. In the traditional association rules mining [11][12], minimum support threshold and minimum confidence threshold values are assumed to be available for mining frequent item sets, which is hard 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 [13][14][15][16], another constraint such as similarity or confidence pruning is usually introduced.

Association Rule Mining is all about finding all rules whose support and confidence exceed the threshold, minimum support and minimum confidence values. In the traditional association rules mining with FPtrees and reduction technique[11][12],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.

DESIGN

Mining for association rules:

Association rules are the form

A->B

This implies that if a customer purchase item A then he also purchase item B. For the association rule mining two threshold values are required. As given in the design part.

• Minimum support

• Minimum confidence

The ordering of the items is not important. a customer can purchase item in any order means if he purchased Milk first then Butter and after purchasing both he can buy Bread or after buying Bread he can purchase Milk and Butter. But in the association rules the direction is important.

If A->B is different from B->A

There is general procedure for defining the mining association rules using Apriori algorithm.

• Use apriori to generate frequent item-sets of different Sizes

• At each iteration divide each frequent item-set X into two parts antecedent (LHS) and consequent (RHS) this represents a rule of the form LHS->RHS.

• The confidence of such a rule is support(X) /support (LHS)

• Discard all rules whose confidence is less than minimum confidence

The support of an association pattern is the percentage of task - relevant data transactions for which the pattern is true.

Support (A -> B) = P (A U B)

Support (A -> B) = # # Tuple containing both A & B

Total # of Tuples

If the percentage of the population in which the antecedent is satisfied is s, then the confidence is that percentage in which the consequent is also satisfied.

The confidence of a rule A -> B, is the ratio of the number of occurrences of B given A, among all other occurrences given A. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern A -> B

Confidence (A -> B) = P (B |A) means the probability of B that all know A

CO NFIDENCE (A-> B) = # Tuple containing both A & B

# of Tuples containning A

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.

TABLE II. TRANSACTION DATABASE

TXl	Bread	Butter	Milk
TX2	Ice-cream	Bread	Butter
TX3	Bread	Butter	Noodles
TX4	Bread	Noodles	Ice-cream
TX5	Butter	Milk	Bread
TX6	Bread	Noodles	Ice-cream
TX7	Milk	Butter	Bread

TX8	Ice-cream	Milk	Bread
TX9	Butter	Milk	Noodles
TX10	Noodles	Butter	Ice-cream

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 Bread and noodles means a customer who purchased bread also purchased the noodles is given below.

The support for ten transactions where bread and noodles occur together is two.

Support for {Bread, Noodles} = 2/10= 0.20.

This means the association of data set or item set, the bread and butter brought together with 20 percent support.

Confidence for Bread ----> Noodles = 2/8 = 0.25

This means that a customer who buy bread then there is a confidence of 25 percent that it also buy butter.

METHODOLOGY

The Apriori Algorithm:

Apriori is a seminal algorithm for finding frequent item-sets using candidate generation [18]. Mining for association among items in a large database of sales transaction is an important database mining function. Given minimum required support s as interestingness criterion:

1. Search for all individual elements (I-element item-set) that have a minimum support of s.

2. Repeat

   l. Form the results of the previous search for I element item-set, search for all i+ 1 element item.

   Sets that have a minimum support of item-set.

   2. This becomes the set of all frequent (i+ 1) item Sets that are interesting

   3. Until item-set size reaches maximum.

A large supermarket tracks sales data by Stock-keeping unit (SKU) for each item, and thus is able to know what items are typically purchased together. Apriori is a moderately efficient way to build a list of frequent purchased item pairs from this data.

By using the consumer database given in table no.2

Let's illustrate the process of Apriori with an example, let takes the consumer database which is showing the number of item sets purchased by the consumers from a bakery shop.

The first step of Apriori is to count up the frequencies, called the supports, of each member item separately:

C1 -> L1

item	support	item	support
Bread	0.8	Bread	0.8
Butter	0.7	Butter	0.7
Ice-Cream	0.5	Ice-Cream	0.5
Milk	0.5	Milk	0.5
Noodles	0.5	Noodles	0.5

Now the support for two element item- sets. Interestingness 2-element item-sets

C2

item-sets	support
{Bread, Butter}	0.5
{Bread, Milk}	0.4
{Bread, Noodles}	0.2
{Bread, ice-cream}	0.3
{Butter, Milk}	0.4
{Butter, Noodles}	0.3
{Noodles, ice-cream }	0.3

L2

item-sets	support
{Bread, Butter}	0.5
{Bread, Milk}	0.4

{Bread, ice-cream}	0.3
{Butter, Milk}	0.4
{Butter, noodles}	0.3
{Noodles, ice-cream }	0.3

Here {Bread, Noodles} Item-set thrown away because its support value is less then minimum support

Interestingness 3-element item-sets

C3

item-sets	support
{Bread, Butter, Milk}	0.3
{Bread, Milk, ice-cream}	0.1
{Bread, Butter, ice-cream}	0.0
{Butter, Milk, Noodles}	0.1
{Bread, Milk, Noodles}	0.0
{Noodles, ice-cream, Bread}	0.2

L3

item-sets	support
{Bread, Butter, Milk}	0.3

Here only one item-sets which satisfy the minimum support value. So after three iteration, only one item-set filtered. {Bread, Butter, Milk}

ASSOCIATION RULES FOR FREQUENT ITEM-SETS

Rules	Confidence(percentage)
{Bread}->{Butter, Milk}	37

{Bread, Butter}->{Milk}	60
{Bread, Milk}->{Butter}	75
{Butter}->{Bread, Milk}	42
{Butter, Milk}->{Bread}	75
{Milk}->{Bread, Butter}	75

If the minimum confidence threshold is 70 percentages then discovered rules are

{Bread, Milk}-> {Butter}

{Butter, Milk}-> {Bread}

{Milk}-> {Bread, Butter}

Because the confidence value of these rules are greater than minimum confidence threshold value which is 70 percent. So in the simple language if a customer buy Bread and Milk he is likely to buy Butter. A customer buy Butter and Milk is likely to Bread. A customer buy Milk is likely to buy Bread and Butter.

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 may play 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 may find the tendency of a customer on the basis of frequently purchased item-sets.

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 may play 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 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

REFERENCES:

[1] Jugendra Dongre, Gend Lal Prajapati, S. V. Tokekar “The Role of Apriori Algorithm for Finding the Association Rules in Data Mining”, IEEE conference publication, 2014.
[2] J. Dongre ,G.L. Prajapati, S. V. Tokekar ”Apriori Algorithm in data mining”.
[3] 'The 6 biggest challenges retailer Face today", www.onStepRetail .com. retrieved on June 2011.
[4] R. Agrawal, T. Imielinski, and A. Swami.. Mining association rules between sets of items in larged databases, In Proceedings of the 1993 ACM SIGMODInternational Connference on Management of Data, pages 207-216, Washington, DC, May 26-28-1993.
[5] Fayyad, U. M; Piatetsky-Shapiro, G. ; Smyth, P.; and Uthurusamy, R.1996. Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press.

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

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

[8] Literature Review: Data mining, http://nccur.lib.nccu.edu. twlbitstream OS.pdf, retrieved on June 2012.

[9] 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.

[10]] 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 .

[11] 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.

[12]] P. Tang, M. Turkia "Para llelizing frequent item set mining with FP trees". Technical Report titus.compsci.ualr.edu/-ptang/papers/par-fi.pdf, Department of Computer Science, University of Arkansas at Little Rock,2005.

[13] M. Ashrafi, D. Taniar, and K. Smith "Redundant Association Rules Reduction Techniques". Lecture Notes in Computer Science, Volume 3S09, 2005, pp. 254 -263 .

[14] M. Dimitrijevic, and Z. Bosnjak "Discovering interesting association rules in the web log usage data"Interdisciplinary Journal of Information, Knowledge, and Management, 5, 20 I 0, pp.191 -207.

[15] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo.: Fast discovery of association rules- In Advances in Knowledge Discovery and Data Mining (1 996)