International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in

Vol.4., Issue.3., 2016
(May-June)

RESEARCH ARTICLE

ISSN: 2321-7758

A COMPARATIVE ANALYSIS OF APRIORI ALGORITHM AND MATRIX APRIORI ALGORITHM FOR PROFITABLE BUSINESS LOGIC IN RETAIL SECTORS

RASHEEDHA Z KHAN[1], G C DIVYA[2]

[1] HOD of information science and engineering, [2]Student, Department of CSE
Shree Devi Institute of Engineering & Technology, Kenjar, Mangluru, Karnataka, India

ABSTRACT

At present Data mining is one of the effective area for better e-commerce application. Retail sector is facing key problem to find useful hidden patterns for maximum retail strategy. To find frequent item set and to derive the association rules from them the apriori algorithm is vogue in data mining. A transaction database is used for mining association rules from frequent item set. Retail holders need a profitable business logic in a while time to gain more profit. In this study, frequent item set mining algorithms called apriori algorithm and matrix apriori algorithm is used and comparative study is made based on time performance. Evaluation is given to prove the study.

Keywords - datamining; ecommerce; apriorialgorithm; Matrix apriori; retail sector; supourt; confidenc; tabularized

INTRODUCTION

In a very large database extraction of hidden predictive information is done by data mining. Data mining is also a tool to find hidden pattern of frequent item sets generated by association rules. The changing and the focused environment in the global platform are making retail sector holder's face a lot of problem for better market campaign [1].

Collection of Daily transaction details of a customer by the retailer do not have a proper format to hold on details. The data collection requires proper information using which the retail holder can make better business decision for maximum profit. The doctrinal tack was lot of time consuming to steadfast the problem or a decision making for turnout winnings business. Data mining prepares database for finding hidden patterns finding sequential deep informations that a dexterous may miss because it lies outside the surmise. From the lag decades data mining has got rich seat in decision making and it has become essential component in various populous commercial enterprise [3][7][5].The data mining has been formulated and stiffed into areas such as tooling, indemmity ,drug and retail etc.,[3][2][4][6].Hence, this papers reflects various trends of data mining and allied applications from over and on and discusses how effectively can be used to spot profitable customers.

Indent of study

Mining association rules is useful in picking up frequent item sets purchased by customer that define a relation in database between the unrelated item set.It has two met age support value and confidence. Frequent Item set that have support value greater than or equal to minimum threshold value and rules that have confidence value greater

than or equal to minimum threshold confidence value are said to be frequent item set and frequent item rules respectively.

The apriori algorithm is primo algorithm in data mining for extracting association rules [7].

Design

Data mining perceptions:

Item sets and Associations:

Represented as:

X->Y

If X is true then B will also true. Example: Sale of ice-cream and cold beverages are likely to be high in summer than compared to winter.

X=sale of ice-cream and cold beverage.

Y=High in summer than winter.

Using this, association rule can predict that if X is true then B is also true.

A rule X->Y->Y->X, then X and Y are called an "interesting item set"

Example:

Patron purchasing bread also buy jam.

Patron purchasing jam also buys bread.

Sales transaction table:

From market basket analysis of set of products in a single transaction. Considering example, patron purchasing bread also buys jam.

Bread->jam

Transactional database:

The set of all sales transaction is called population. Table 1 is a population table.The representation of one record per transaction. Data tuple is used to represent transaction.

Table 1 population table

Transactions	Item sets
T001	l1,l2,l5
T002	11,12,14
T003	12,13
T004	11,12,14
T005	11,13
T006	12,13
T007	11,13
T008	11,12,13,15
T009	11,12,13

Support and confidence:

Support and confidence threshold level is mandatory in association rule.

Support (X->Y)= p(X u Y)

$$\text{Confidence}(X\to Y)=\frac{\#\_\text{TUPLES\_CONTAINING\_BOTH\_X\&Y}}{\#\_\text{OF\_TUPLES\_CONTAINING\_X}}$$

(# indicates number of)

The percentage of number of occurrence of rule and occurrence of all rules is said to be support.

Confidence (X->Y)= p(Y|X) the probability of Y that all known Y.

$$\text{Support}(X\to Y)=\frac{\#\_\text{TUPLES\_CONTANING\_BOTH\_X \& Y}}{\text{TOTAL\_\#\_OF\_TUPLES}}$$

The percentage of number of occurrences of Y given X, among all other occurrences given X

The above transaction table 1 is collection of item set that a patron buys in a while span(items are given id as l1..l2...). Table 2 is the detailed description of item sets. List of the support for item set pen and notebook means a patron who purchased pen also purchased notebook is shown below.

The support for 9 transaction where pen and notebook occur together is 4.

Support for{pen, notebook}=4/10=0.40

This summarizes the association of data set ,pen and notebook are bought together with 40 percent support.

Confidence for pen->notebook=4/9=0.44

This summarizes that patron who buy pen then there is a confidence of 44 percent that he also buys notebook.

Table 2 Detailed population table

T001	Pen, notebook, Pencil
T002	Notebook, Eraser
T003	Notebook, Crayons
T004	Pen, notebook, eraser
T005	Pen, crayons
T006	Notebook, crayons
T007	Pen, crayons
T008	Pen, notebook , crayons, gum
T009	Pen, notebook, crayons

Methodology

Apriori algorithm:

Minimum support s and confidence criterion known:

1. Inspect for all single elements that have a minimum support of s.
2. Repeat
   a. Search for all i+1 element item set for a previous i-element from the previous search results.
   b. Resulting is a interesting set of all frequent (i+1) itemset.
3. Down to maximum item set size.

A very large database extraction of hidden predictive information is done by data mining.ata mining is also a tool to find hidden pattern of frequent item sets generated by association rules. The changing and the focused environment in the global platform are making retail sector holder's face a lot of problem for better market campaign [1]. Considering the support count as 2 and confidence value 7,referring to table 1 following transaction occur. Initial support count of item set that have occurred in the transaction table.

Step 1:

C1 → L1

Item set	Support
I1	6
I2	7
I3	3
I4	6
I5	2

Step 2:

C2 → L2

I2,I3	**4**
I1,I2	**4**
I1,I3	**4**
I2,I5	**2**
I2,I4	**2**
I1,I5	**2**

Items set with value less than the minimum support values are discarded and above set of transaction is obtained.

Step 3:

C3 → L3

I2,I3,I4	0
I2,I5,I4	0
I1,I2,I5	2
I1,I2,I3	2
I1,I3,I5	1
I2,I3,I5	1

In the final iteration item set with minimum support are discarded and finally filtered transaction is obtained below.

I1,I2,I5	2
I1,I2,I3	2

Here only two item-sets which satisfy the minimum support value are obtained at the end of three iterations.

Item-sets show that a customer buys pen -> notebook-> pencil, Pen->notebook->crayons frequently. Data is taken only from i iterations not from the whole data this makes apriori more advantageous. In every iterations item sets which does not satisfy the minimum support are just discarded away. New item-set which follow minimum support is obtained at end of every iteration. Algorithm gives the maximum number of item sets which is repeated maximum time. In the example taken to study here a customer purchases pen-notebook- pencil, pen-notebook -crayons together.

Association rule mining generally defined by apriori algorithm:

1. Utilize apriori to generate frequent item-set of different size.

2. A rule of the form LHS->RHS is represented at each iteration. Frequent item-set x of this each iteration is divided into two parts ante cent and consequent.
3. A rules confidence is support(X)/support (LHS)
4. Rules with confidence less than the minimum confidence is discarded.

The frequent item set are {pen, notebook, pencil} {pen, notebook, crayons} together with item set size three. Below shows that item sets divided into some rules.

Table 3 association rules for frequent item sets.

Rule	Confidence(percentage)
{I5}-{I2,I5}	**100**
{I5}-{I1,I2}	**100**
{I1}-{I1,I5}	**100**
{I1}-{I2,I5}	**100**
{I4}-{I2,I4}	**100**
{I5}-{I1,I5}	**100**

The above table 3 are the rules discovered which are greater than with minimum threshold confidence value 70. Different association rules are obtained by defining different confidence value. As the confidence value is high more is the accurate rules.

Matrix apriori algorithm:

The frequent item set in matrix apriori algorithm are stored in a matrix called Tabularize frequent item and support of the vector stores the patterns of support.

First scan,

1. Scan the database to determine frequent item set.
2. Sort the items in descending support count, trimmed to those that are above the minimum support value for frequent item set list.
3. The sorted item list is the basis for the order of columns of TFI.

Second scan ,

1. TFI first row is left empty.
2. According to following rule zeros and ones is inserted for each new transaction in database which is contructed by following order in frequent items list.
3. If the transaction contains item or not , items in the column of list is added either “0” or “1”.If transaction in TFI is already exists it is not stored in new row but support count is updated by 1.
4. Use the empty row to order up the frequent item set search.
5. In remaining rows of unmodified matrix, considering each column beginning from first row, each cell value is set to number of the row where cell value is equal to “1”.

Table 4.1 a

TID	Items	Item sets	Items
T001	I1,I3,I4	{I2}	3
T002	I2,I3,I5	{I3}	3
T003	I2,I3,I5	{I5}	3
T004	I1,I2,I5	{I1}	2
		{I4}	1

TFI-Tabularize Frequent item & SOP- support of pattern

Table 4.1 b

TFI				SOP
I2	I3	I5	I1	
0	1	0	1	1
1	1	1	0	2
1	0	1	1	1

TFI				SOP
I2	I3	I5	I1	
3	2	3	2	1
0	3	0	4	2
4	1	4	0	1
1	0	1	1	

Itemset	Sup Count
I3,I5	2
I2,I5	3
I2,I3	2
I2,I3,I5	2

Reverse indexing is used in transaction to reach the frequent item set. Table 4.1 a shows the construction and modification of matrix. Table 4.1 b is a frequent item set collection after matrix construction. Item that has least support count. item is compared with items found on its left in order to

find frequent item sets. Sequentially adding the related rows of SOP from top to bottom item set set support count is identified.

Performance evaluation

Considering the time as entity for both apriori and matrix apriori time consumed by matix apriori is minimum when compared with apriori. Graph is shown to state the same.

Conclusion

Matrix apriori scans the database only twice and management of matrix is easy. Matrix apriori uses minimum time compared with apriori. Retailers can use matrix apriori for a better retail strategy in while time for better business logics.

REFERENCES

[1]. ”The 6 biggest challenges retailer Face today", www.onStepRetail .com. retrieved on June 20 II

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

[3]. Andre Bergmann, "Data Mining for Manufacturing: PreventiveMaintenance, Failure Prediction, and Quality Control".

[4]. Fayyad, U. M; Piatetsky-Shapiro, G. ; Smyth, P.; and Uthurusamy, R.1 996. Advances in Knowledge Discovery and Data Mining. Menlo Park,Calif.: AAAI Press.

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

[6]. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts andTechniques, published by Morgan Kauffman, 2nd ed.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 201 2 .

[7]. H. Mahgoub,"Mining association rules from unstructured documents"[II] P. Tang, M. Turkia "Para llelizing frequent itemset mining with FPtrees".Technical Report titus.compsci.ualr.edu/-ptang/papers/par-fi.pdf,Department of Computer Science, University of Arkansas at Little Rock,2005.

[8]. Wu, x., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. 1. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg (2007, December), "Top 10 algorithms in data mining. Know!. Inf. Syst.". 14(1), 1-37.