ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

An Approach for Incremental Frequent Pattern Mining Using Modified Apriori Algorithm.

Harsha Sarah Thomas[1]*, and Nancy Victor[2].

[1]Master of Computer Applications, SITE, VIT University, Tamil Nadu, India.
[2]Assistant Professor, SITE, VIT University, Tamil Nadu, India.

ABSTRACT

Data is getting increased day by day. Usage of social networking sites, hospital administration, banking sectors, etc are some of the reasons for the huge volume of data generated. Data mining is a very effective process which helps to examine the data from various dimensions and it helps to make the data, useful information. This paper is about frequent pattern mining which helps to find out the frequent occurrence of data in a data set. Nowadays the data sets are not static and as a result, dynamic data sets are used to do the frequent pattern mining. Here, an approach of Modified Apriori algorithm is proposed. Different scenarios have been taken to consideration for mining frequent patterns.

Keywords: Data, Data Mining, Frequent Pattern Mining, Modified Apriori Algorithm

Corresponding author

ISSN: 0975-8585

INTRODUCTION

We live in a world where data and information are getting increased day by day. Even though data and information sounds similar, both are having its own differences. Information regarding a particular area, or object or anything can be obtained from the available data. As the data gets increased in a rapid fashion it is necessary to mine these data in order to get useful information. Main theme of this work is mining frequent patterns. As the name implies, mining frequent patterns means extracting the patterns which occurs very frequently. Mostly, a collection of items will be always together like bread and butter or bread and jam etc. Mining frequent patterns helps to find out how the items are associated with each other or how the item is associated with the customer and how they are related.

Association analysis

Suppose the manager of the super market wants to know which items are frequently purchased together. An example of such a rule mined from the super market transaction database is: buys (A, "bread") → buys (A,"butter") [support = 1%, confidence = 50%]

According to the above rule, A represents customer. Support and confidence are the two other terms mentioned along with the rule. Support of 1% means that, 1% of all the particular transactions under the examination tells that bread and butter are bought together. Confidence of 50% means that, if a customer buys bread, and then there is a possibility of 50% that she will be buying butter also.

This paper proposes a methodology, modified Apriori Algorithm, to find the frequent patterns from a dynamic dataset. Dynamic dataset means the dataset which is not a fixed one and it will be changed or getting updated every now and then. The modified Apriori algorithm, which is proposed in this paper, is another way of implementing Apriori algorithm[1]. As the transactions are read from the database, the so called items will be dynamically added and removed. If an item set has to be frequent, all of its subsets also must be frequent. This is an important fact of data mining. The methodology proposed in this paper depends on this fact. Now the examination has to be done in the item sets whose subsets are frequent.

LITERATURE REVIEW

DATA MINING

Data mining is the technique in which the data will be properly examined from all the dimensions and then it will be converted into a very useful piece of information [2]. Data mining is consumer centric and it helps to know, what customers' tastes are and how they are going to buy items.

From a large database, it is very difficult to get the useful information which a user needs or an organization needs just after viewing the database. Data mining helps to extract useful information from large databases with the great ability to help the companies concentrate on the very important information from their large databases or data warehouses. With the help of data mining tools the current fashion and the working in the business will be understood and decisions can be taken accordingly. Data mining is able to tell the important things that are useful and it can even foresee what is going to happen next. Modeling is the name of the process in which the result of a particular situation will be known and this knowledge will be used to create the model for that situation and it will be applied to another situation in which the answer is not known. Data mining works under this technique, modeling.

Mining Frequent Patterns

Mining frequent pattern is one of the data mining functionalities. This is what the paper is about also. From the name itself it is understood that frequent pattern means, the patterns which occurs repeatedly due to a particular similarity [3]. This mainly helps to find the regularity of the behavior of customers and this will be more understood if we take market basket analysis. Here, mostly the patterns are expressed in the form of association rules.

For example if a customer buys bread and jam, then she may also buy butter[4]. The frequent patterns are shown in table 1.

Example for frequent item sets

Transaction database:

1 : {f,i,j}
2 : {g,h,i}
3 : {f,h,j}
4 : {f,h,i,j}
5 : {f,j}
6 : {f,h,i}
7 : {g,h}
8 : {f,h,i,j}
9 : {g,h,j}
10 : {f,i,j}

1 item	2 items	3 items
{f}: 7	{f,h}: 4	{f,h,i}: 3
{g}: 3	{f,i}: 5	{f,h,j}: 3
{h}: 7	{f,j}: 6	{f,i,j}: 4
{i}: 6	{g,h}: 3	
{j}: 7	{h,i}: 4	
	{h,j}: 4	
	{i,j}: 4	

Table 1: Frequent Patterns

APRIORI ALGORITHM

Apriori algorithm is the fundamental algorithm used for finding out the frequent patterns [5]. Apriori helps a lot in learning the association rule also. Since Apriori is the fundamental algorithm for finding out the frequent patterns, it works on the transaction database which is having the details regarding the items purchased by the customer. The main feature of the algorithm is to get the very useful information from the large database. For example, the information that a customer who bought mobile phone will also be likely to buy memory card can be attained from the following association rule:

Support (Mobile phone → memory card)

$$\frac{\text{No. of transactions containing both mobile phone and memory card}}{\text{No. of total transactions}}$$

Confidence (Mobile phone → memory card):

$$\frac{\text{No. of transactions containing both mobile phone and memory card}}{\text{No. of transactions containing (Mobile phone)}}$$

The goal of the algorithm is to find the rules which satisfy the minimum support and minimum confidence.

Working of the Apriori

All the frequent item set should be found out. Items whose existence in the database are more than or equal to the minimum support should be considered and through this the frequent items are acquired. Then the frequent item sets have to be found out. For this, from the frequent item sets attained, candidates have to

be generated. Now the results should be pruned to attain the frequent item sets. Now this frequent item sets will be used to generate the strong association rules which should satisfy the minimum support and confidence threshold.

INCIDENCE MATRIX

Incidence matrix is the matrix which contains 1's and 0's [6]. Incidence matrix shows the connection between two classes of objects. In this paper, incidence matrix is a very important component which is used in the working of the modified Apriori algorithm. Here, incidence matrix is created according to the corresponding transaction database. So the 1s and 0s in the incidence matrix explains the purchase status of the customer. If the customer has bought a particular item, there will be a corresponding entry of '1' in the incidence matrix, else it will be 0. This can be explained by the following example in table 2.

Conversion of transaction database into incidence matrix

Customer	Item 1	Item 2	Item 3
Priya	Bread	Butter	Milk
Ram	Bread	Butter	-
Ann	Bread	-	Milk
Riah	-	Butter	Milk
Meera	Bread	Butter	

Table 2: Transaction database

According to the table, incidence matrix will be:

1	1	1
1	1	0
1	0	1
0	1	1
1	1	0

PROPOSED ARCHITECTURE

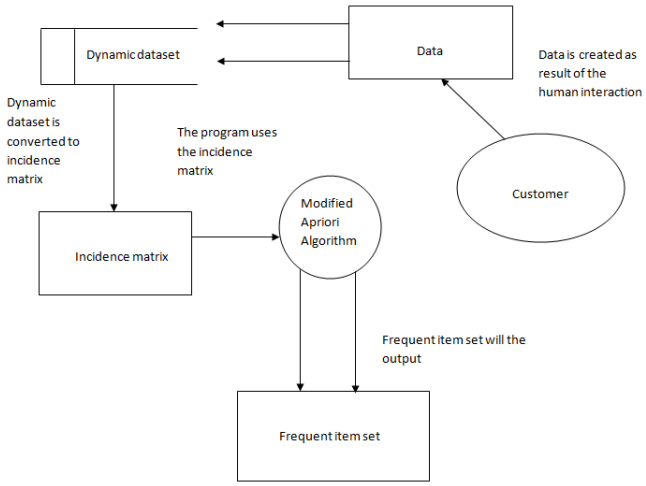

Fig 1: Proposed Architecture

The itemset that is to be considered in the architecture is of dynamic nature. Dynamic itemset means the itemset which will get updated and thus the values of the itemset are not static. For finding the frequent patterns according to the latest dataset, Apriori algorithm is the algorithm from which the references have been taken. This Apriori algorithm which is the fundamental algorithm for finding the frequent patterns is modified and used in this system. The itemset is not directly used as the input to the program. Firstly, the

dataset will be converted to the corresponding incidence matrix. This incidence matrix is then taken as the input for the modified Apriori algorithm. The program works on the basis of the incidence matrix and the frequent patterns are generated according to recent dataset. The architecture is shown in figure 1. Various algorithms have been proposed for incremental pattern mining [7-12].

SYSTEM IMPLEMENTATION

Incidence matrix generation

Incidence matrix is the matrix which contains zeros and ones. One is for showing the on state or positive and zero is for showing the off state or negative. Here, in this work, incidence matrix is generated according to the dynamic itemset. This helps to reduce the complexity. The example is shown in table 3.

	Bread	Butter	Jam
Aswin	1	1	0
Harsha	1	1	1
Sanuj	1	0	1

Table 3: Incidence Matrix generation example

This is a sample dataset. The corresponding incidence matrix generated according to the dataset is:

1 1 0
1 1 1
1 0 1

Pseudocode for the generation of incidence matrix:

1. Input the dynamic dataset in .xls format file in which the purchase status of the customer should be shown either as Y or N.
2. Check the cells of the dataset.
3. If Y or y is found, replace it with a 1. Else, replace with 0.

Modified Apriori Algorithm

Apriori algorithm is the fundamental algorithm for finding the frequent patterns. Here, as the frequent patterns are found from a dynamic dataset the Apriori algorithm is modified accordingly. In fact, this is another way of finding frequent patterns. The input to this program will be the generated incidence matrix. The output will be the frequent patterns according to the latest dataset.

Pseudocode for the modified Apriori algorithm:

1. Input the incidence matrix of the corresponding transaction data set. Find the support and confidence for the dataset.
2. Initialize the flag as “0” for all the items in the dataset. Also, maintain the itemset states.
3. Find the corresponding candidate itemsets.
4. Based on the candidate pairs, find the frequent itemset. Mark the flag as “1” and proceed.
5. Add the frequent data to the existing dataset.
6. As new data gets added to the transaction data base, go to step 3 and repeat.

RESULTS AND DISCUSSION

The two results are:

- Incidence matrix

ISSN: 0975-8585

Incidence matrix is the corresponding matrix obtained from the dynamic dataset. Incidence matrix contains ones and zeros. The result is shown in figure 2.

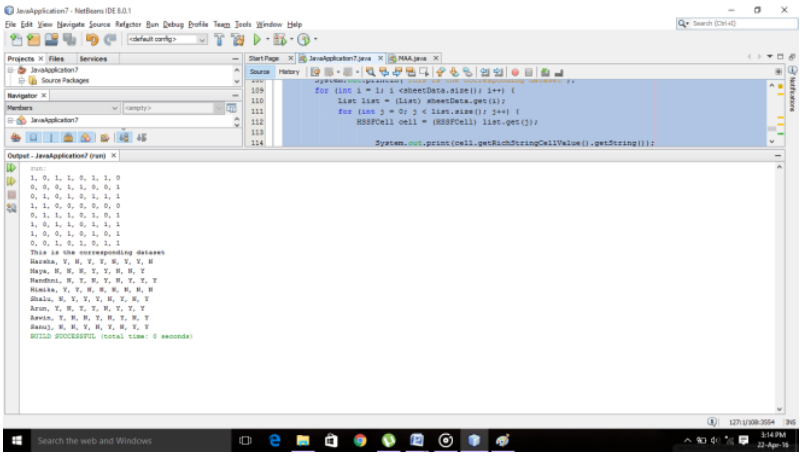

Figure 2: Incidence matrix generation

- Frequent patterns

Frequent patterns are obtained according to the latest dataset. These patterns are obtained from a dynamic dataset. **The frequent patterns are shown in figure 3.**

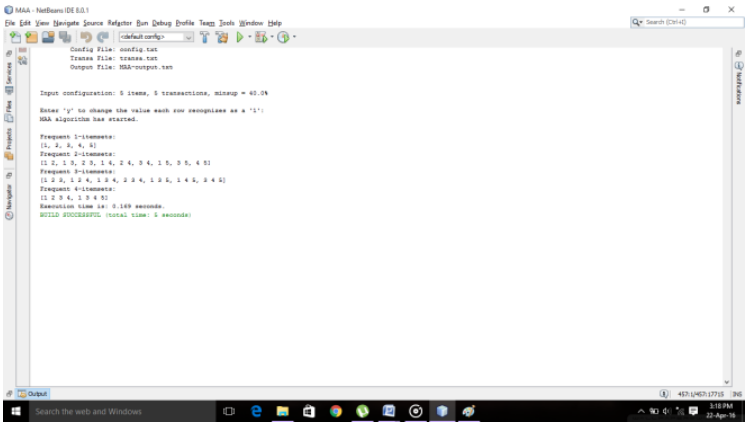

Figure 3: Frequent patterns

CONCLUSIONS

Apriori algorithm is one of the algorithms used for identifying the frequent items. The work proposes a modified Apriori algorithm in which the same will be working on the dynamic item sets. Modified Apriori algorithm is another way of detecting the repeated items. The algorithm will not directly use the dynamic item set. Instead it will be converted to an incidence matrix and then this incidence matrix will be used as the input for obtaining the frequent item sets. As the dataset is dynamic, the obtained result will also be dynamic. The frequent patterns obtained may not be similar always. Because each time, data will be getting updated or removed in the dataset and that is what actually dynamic dataset is. The data inside the dynamic dataset is not static and it can be changed at anytime.

FUTURE RESEARCH RECOMMENDATIONS

In future the research can be extended to work with MapReduce computation model[13]. As the dynamic dataset gets incremented frequently, it may not be possible for the traditional database systems to handle the massive quantity of data which is produced. MapReduce plays a major role in handling big datasets and hence can be incorporated in the future work for working with market basket dataset [14].

REFERENCES

[1] Wasilewska, Anita. "Apriori Algorithm." Lecture Notes, http://www. cs. sunysb. edu/~ cse634/lecture_notes/07apriori. pdf, accessed 10 (2007).

[2] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996 Mar 15;17(3):37.

[3] Aggarwal CC, Han J, editors. Frequent pattern mining. Springer; 2014 Aug 29.

[4] Padhi I, Mishra J, Dash SK. Predicting Missing Items in Shopping Cart using Associative Classification Mining. International Journal of Computer Applications. 2012 Jan 1;50(14).

[5] Agrawal R, Srikant R. Fast algorithms for mining association rules. InProc. 20th int. conf. very large data bases, VLDB 1994 Sep 12 (Vol. 1215, pp. 487-499).

[6] Hahsler M, Grün B, Hornik K. Introduction to arules–mining association rules and frequent item sets. SIGKDD Explor. 2007 Jul 13;2(4).

[7] Wang JL, Xu CF, Pan YH. An incremental algorithm for mining privacy-preserving frequent itemsets. In Proceedings of Fifth International Conference on Machine Learning and Cybernetics, Dalian 2006 Aug 13 (Vol. 13, p. 16).

[8] Xie Y, Xu Z, Zhu X, Xie P. A parallel algorithm PMASK based on privacy-preserving data mining. InInstrumentation & Measurement, Sensor Network and Automation (IMSNA), 2012 International Symposium on 2012 Aug 25 (Vol. 2, pp. 398-402). IEEE.

[9] Shana J, Venkatachalam T. An Improved Method for Counting Frequent Itemsets Using Bloom Filter. Procedia Computer Science. 2015 Dec 31;47:84-91.

[10] Toon Calders , Nele Dexters , Joris J.M. Gillis and Bart Goethals, “Mining frequent itemsets in a stream”, ScienceDirect(2012)

[11] Troiano L, Scibelli G. Mining frequent itemsets in data streams within a time horizon. Data & Knowledge Engineering. 2014 Jan 31;89:21-37.

[12] Tseng FC. Mining frequent itemsets in large databases: The hierarchical partitioning approach. Expert Systems with Applications. 2013 Apr 30;40(5):1654-61.

[13] Farzanyar Z, Cercone N. Accelerating Frequent Itemsets Mining on the Cloud: A MapReduce-Based Approach. In2013 IEEE 13th International Conference on Data Mining Workshops 2013 Dec 7 (pp. 592-598). IEEE.

[14] Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. InACM SIGMOD Record 1997 Jun 1 (Vol. 26, No. 2, pp. 255-264). ACM.