International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 2 Issue 3
March 2015

Apriori Algorithm and its Applications in The Retail Industry for Analyzing Customer Interests

Reddy Sangeetha DM[1], Vanita Saldanha[2], Shoney Sebastian[3]
MS, Post Graduate Department of Computer Science, Christ University, India[1]
MS, Post Graduate Department of Computer Science, Christ University, India[2]
Assistant Professor, Department of Computer Science, Christ University, India[3]

Abstract- Data mining, in the recent times has gained a lot of popularity due to the availability of huge sets of data, which grows day by day. The need to transform this data into useful information has lead to the development of various data mining algorithms. The Apriori algorithm, known to be one of the best algorithms for finding frequent item sets in large transactional databases, is the subject of research in this paper. In Retail Industry, which accumulates a huge amount of sales data on a daily basis, there is a constant need for analyzing data to determine frequently purchased items by a customer, over a period of time. An instance for usefulness of this information can be as follows. This information can be used to generate a shopping cart of items for the customer which is deliverable on demand. The paper discusses also discusses the advantages and disadvantages of Apriori Algorithm. It also explains the implementation of Apriori Algorithm in various domains.

Keywords-Apriori, frequent item sets, Data mining techniques, patterns

1. INTRODUCTION

Today every industry, be it Retail or Banking or Healthcare, has to deal with huge sets of data. The data has to be analyzed and useful information has to be extracted to understand customer's interests and meet their demands in a better way. Data mining, the extraction of predictive information from large databases, is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses. Theoretically, Data mining is a step in the Knowledge Discovery in Databases (KDD) process. Today, Knowledge discovery is useful in many domains like Health Care, Web Analytics, Banking Domain etc. Data Mining is called the core step of the KDD process because it is in this step that data is examined and patterns are generated. Over the years, there have been several algorithms implemented to facilitate data mining in transactional databases. Apriori algorithm, FP Growth, GSP and SPADE (Both improvisations of the Apriori Algorithm) to name a few. Apriori algorithm is termed to be one of the best approaches in data mining. A distinguishing property of Apriori is, if a sequence cannot pass a test (minimum support), all of its super sequences will also fail the test. Use of this property to prune the search space makes the generation of frequent patterns more efficient.

The term 'frequent patterns' basically refers to the item-sets that occur most frequently in a specific transaction/order. Usually all transactions of the customer are together viewed as sequence, where each transaction is represented as an item-set in that sequence. There can be many kinds of frequent patterns like, frequent item-sets, frequent subsequences and frequent substructures. A frequent item-set refers to the set of items that often appear together in a transactional data set- for example, milk and bread, which are frequently bought together in a grocery store. A subsequence, such as buying first a PC, then a digital camera and then a memory card, if it

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 2 Issue 3
March 2015

occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences.

Specifically in the retail industry, frequent patterns are mined from previous transaction records of customers. The retailers can then use such information to analyze the frequent items bought by the customer, to understand their interests, to satisfy their demands and above all to predict their needs. As seen in the past, Retail Stores have attained success by emphasizing on understanding customer-buying-habits from these sets of data. The Apriori Algorithm mentioned earlier can help retail industries to make well informed decisions with respect to frequent item sets.

The paper is structured as follows: Section 2 reviews literature pertaining to applications of Apriori Algorithm in various domains. Section 3 describes the Apriori Algorithm in detail which is meant for finding frequent item sets in transactional databases. Challenges of Implementing Apriori Algorithm in Retail Industry are described in section 4. Section 5 concludes the paper.

2. RELATED STUDY

[1] This paper proposes a way of discovery of association rules between different types of e-banking offered by banks. This paper demonstrates the application of data mining methods to e-banking. Association rules concerning e-banking are discovered using different techniques and prediction models depending on e-banking parameters like the transactions volume conducted through this alternative channel in relation with other crucial parameters like the number of active users.

[2] This paper describes the implementation of web access pattern discovery in a small scale data sources. By using proxy server it explains which kind of data is frequently used by different kinds of web users. This access pattern is helpful in various domains such as cyber crime, search engines pre-fetching concepts. It discovers patterns from the web.

[3] Explains the use of association rule mining in extracting pattern that occurs frequently within a dataset. Association algorithm can be applied to the conversion of quantitative data into qualitative data. The author considers two Association Rule algorithms namely Apriori algorithm and Predictive Apriori algorithm. They implement Apriori algorithm in mining association rules from dataset of crime against women collected from session court and compares the result of both the algorithms using data mining tool called WEKA. M Suman , T Anuradha,K Gowtham, A Ramakrishna

[4] proposed Apriori-Growth algorithm which is based on Apriori algorithm and FP-Growth algorithm. The advantage of the Apriori-Growth algorithm is that it doesn't need to generate conditional pattern bases and sub-conditional pattern tree recursively. The proposed Apriori Growth algorithm overcomes the disadvantages of Apriori algorithm and efficiently mines association rules without generating candidate itemsets, and also the disadvantages of FP-Growth

[5] This paper explains implementation of Apriori Algorithm in Network Cyber Attacks. It mentions about Botnet which is one of the most widespread and serious threats in cyber-attacks. A Botnet is a group of compromised computers which are remotely controlled by hackers to launch various network attacks, such as DDoS attack, spam, click fraud, identity theft and information phishing. Recently malicious botnets evolve into HTTP botnets out of typical IRC botnets. Data mining algorithms can be used to automate detecting characteristics from large amount of data, which the conventional heuristics and

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 2 Issue 3
March 2015

signature based methods could not apply. Here, author presents a new technique for Botnet detection that makes use of Timestamp and frequent pattern set generated by the Apriori algorithm. The main advantage of the proposed technique is that prior knowledge of Botnets like Botnet Signature is not required to detect malicious botnets

[6] The main objective of choosing Apriori is to find frequent item sets. Association Rule Mining can be applied to all fields like Business, Medical and Online Transactions. This paper explains implementation of Apriori Algorithm for detecting crimes against women. Apriori Algorithm in mining association rules from a data set containing crime data concerning woman. A comparative study is made between Apriori and predictive Apriori, the result shows that Apriori Algorithm is better and faster.

[7] This paper proposes a modified Apriori algorithm that is meant for discovering locally frequent patterns from medical data sources. Apriori Algorithm has been modified with pre-processing steps that makes the algorithm even more efficient. The empirical results reveal that this algorithm has plenty of scope to improve the quality of service in the healthcare industry

3. ASSOCIATION RULE MINING ALGORITHMS

An association rule is a rule like "If a customer buys burger, he/she often buys ketch up too". Association rules are of the form X=>Y, each rule has two measurements Support and Confidence. An association rule states that if we pick a customer at random and find out the items she/he selected, we can be assured; indicate the quantity by a percentage, that he/she also selects some other items.

Association rule mining is a two step process:

1. Find all frequent item sets
2. Generate strong association rules from the frequent item sets.

For example, the information that customers who purchase bread also tend to purchase butter at the same time is represented in the following association rule

Bread => Butter [support= 2%, confidence=80%]

Support and confidence are two measures of association rules. A support of 2% means that 2% of all the transactions under analysis show that bread and butter are purchased together. A confidence 80% means that 80% of the customer who purchased bread also bought butter.

Support (A=>B) =P (AUB)

Confidence (A=>B) =P (B/A)

Apriori Algorithm and FP Growth algorithms are the Association Rule Mining algorithms widely used today. The basic implementation of Apriori algorithm which is the subject of this paper is described in detail followed by its comparison with FP Growth Algorithm for time and efficiency.

3.1 APRIORI ALGORITHM

The Apriori Algorithm is described below in brief:-

- Find the frequent itemsets: the sets of items that have minimum support

  1. A subset of a frequent itemset must also be a frequent itemset. • i.e., if {AB} is a frequent itemset, both {A} and { B} should be a frequent itemset.

  2. Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)

- Use the frequent itemsets to generate association rules.

3.2 FP GROWTH ALGORITHM

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 2 Issue 3
March 2015

Unlike Apriori algorithm, FP Growth is a divide-and-conquer approach. The Algorithm:-

- Compresses a large database into a compact, Frequent Pattern tree (FP Tree) structure
- Highly condensed, but complete for frequent pattern mining.
- Avoids costly database scans
- Develops an efficient, FP Growth-based frequent pattern mining method.
- A divide-and-conquer methodology: decompose mining tasks into smaller ones.
- Avoids candidate generation: sub-database test only

3.2 TIME COMPARISON

The main difference between the two approaches is that the Apriori technique is based on bottom-up generation of frequent itemset combinations and the FP-Tree based ones are partition-based, divide-and-conquer methods. An experimental study revealed the comparative performances of Apriori and FP Growth algorithm. The run time is the time to mine the frequent item sets. The experimental result of time is shown in Fig.1 reveals that the FP Growth outperforms the Apriori approach. [8]

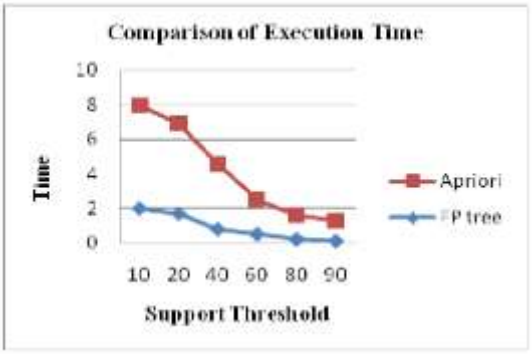

Fig. 1 Execution Time Comparison

3.3 IMPLEMENTATION OF APRIORI ALGORITHM

The goal of retailers (discount stores, department stores, convenience stores, supermarkets, etc) is to increase the gross profit margin through sales and cost reduction. One application of the frequent patterns generated could be, to generate a deliverable shopping cart for the customer consisting of his most frequently purchased items. In this paper, therefore, we propose a method which analyses customer's shopping behavior. The Apriori algorithm generates candidate set during each pass. It reduces the dataset by discarding the infrequent item sets that do not meet the minimum threshold from the candidate sets.

Take an Example

The below table explains the list of monthly transactions and items bought by the customer

Table-1: The Transactions

Monthly Transactions	List of Items
Jan	Potato, Rice, Dal, Rava, Masala powder, Bread
Feb	Avalaki, Rice, Onion, Eggs, Pav
March	Refined Oil, Potato, Rice, Rava
April	Potato, Onion, Bread, Bun,Dal
May	Rice, Dal, Rava, Avalaki, Eggs, Bread, Oil

Now, we follow a simple golden rule: we say an item/itemset is frequently bought if it is bought at least 60% of times. So for here it should be bought at least 3 times.

Lets simplify the table

Potato = P
Rice = R
Dal = D
Rava = Ra
Avalaki = A
Onion = O
Refined Oil = Oi
Eggs =E
Bread = B
Bun = Bu

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 2 Issue 3
March 2015

So the table becomes:

Table-2: Transformed Table

Monthly Transactions	List of Items
Jan	P, R, D, Ra, M, B
Feb	A, R, O, E, Pa
March	Oi, P, R, Ra P
April	P, O, B, Bu,
May	R, D, Ra, A, E, B, Oi

Step 1: Count the number of transactions in which each item occurs.

Table -3 : Candidate 1- Item set

Items	No.Of Transactions
P	4
R	4
D	3
Ra	2
M	1
B	3
Pa	1

Step 2: Now remember we said the item is said frequently bought if it is bought at least 3 times. So in this step we remove all the items that are bought less than 3 times from the above table and we are left with

Table- 4: Frequent 1-Item set

Items	No.of Transactions
P	4
R	4
B	3
D	3

Step 3: We start making pairs from the first item, like PR,PB,PD and then we start with the second item like RB,RD. We did not do BP because we already did PB when we were making pairs with R and buying a Rice and Dal together is same as buying Rice and Dal together. After making all the pairs we get,

Table -5: Frequent 2-Item set

Item Pairs
PR
PB
PD
RB
RD
BD

Step 4: Now we count how many times each pair is bought together.

Table-6: Frequent 3-Item set

Item Pairs	Item Transactions
PR	3
PB	2
PD	4
RB	4
RD	2
BD	1

Step 5: Golden rule to the rescue. Remove all the item pairs with number of transactions less than three and we are left with

Table-7: Frequent 4-Item set

Item Pairs	Item Transactions
PR	3
PD	4
RB	4

Step 6: To make the set of three items we need one more rule (it's termed as self-join),
It simply means, from the Item pairs in the above table, we find two pairs with the same first Alphabet, so we get
PR and PD gives PRD
PR and RB gives PRB
Then we find how many times P,R,D are bought together in the original table and same for P,R,B and we get the following table.

Table-8: Frequent 5-Item set

Item Set	No.of Transaction
PRD	3
PRB	1

Step 7: Thus we can conclude that three items bought frequently by the customer every month are Potato, Rice and Dal.

4. DATA MINING CHALLENGES IN RETAIL INDUSTRY

One of the most significant challenges of data mining in Retail Industry is to obtain high quality and relevant data. It can quite difficult to acquire precise and complete data which is not erroneous as well as gather analytical data.

Retail data is heterogeneous in nature because it is collected from various sources. Before applying Data mining techniques in Retail Industry, it is essential to collect and record the data from different sources into a central data ware house which is actually a costly and time consuming process. Faulty data warehouse design does not contribute to effective data mining.

5. CONCLUSION

In this paper we described how Apriori algorithm is used for discovering locally frequent patterns from retail data sources. Various data mining techniques were used earlier for pattern analysis. However, for finding locally frequent items, Apriori is most suitable especially for transactional databases. This has lead to various improvisations of the core approach. However this technique is useful for only small datasets. For larger datasets, generation of candidate sets and finding frequently occurring items is time consuming and complex. As the size of data increases, the algorithm takes more number of data scans and eventually greater number of iterations for deriving the strong association rules.

References:

[1]Vasilis,Aggelis,DimitrisChristodoulakis,"Association Rules and Predictive Models for e-Banking Services"Proceedings 7th Workshop Research Issues Data Engineering

[2] Sheetalchouhan, Dr.ManishShrivastava "Implementation of web access pattern discovery". International Journal of Emerging Technology and Advanced EngineeringWebsite: www.ijetae.com (ISSN 2250-2459,Volume 2, Issue9,September 2012

[3] DivyaBansal, LekhaBhambhu. 2013. "Usage of Apriori Algorithm of Data Mining as an Application to Grievous Crimes against Women", International Journal of Computer Trends and Technology, Vol.4, Issue.9, pp.3194-3199

[4] M Suman , T Anuradha , K Gowtham, A Ramakrishna. 2012. "A Frequent Pattern Mining Algorithm Based On FP Growth Structure and Apriori Algorithm", International Journal of Engineering Research and Applications, Vol.2, Issue.1, pp.114

[5] S.S.Garasia, D.P.Rana, R.G.Mehta. 2012. "HTTP Botnet Detection Using Frequent PatternsetMining",International Journal of Engineering Science & Advanced Technology Vol.2, Issue.3, pp.619-624.

[6] DivyaBansal, LekhaBhambhu. 2013. Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women", International Journal of Advanced Research in Computer Science and Software

[7] Mohammed Abdul Khaleel, Sateesh Kumar Pradhan "Finding Locally Frequent Diseases Using Modified Apriori Algorithm" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10,October 2013

[8] Rahul Mishra, Abhachoubey "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data." International Journal of Computer Science and Information Technologies, Vol. 3 (4) 2012,4662 - 4665

[9] A Comparative study of Association rule Mining Algorithm – Review Cornelia Győrödi*, Robert Győrödi*,prof. dr. ing. Stefan Holban

[10]R. Agrawal, R. and R. Srikan, "Fast algorithms for mining association rules", Proceedings of the 20th international conference on very large data bases, 478-499, 1994.

[11]W. Lin, S.A. Alvarez and C. Ruiz, Efficient Adaptive-Support Association Rule Mining for Recommender Systems, Data Mining and Knowledge Discovery, 6, Kluwer Academic Publishers, 2002, pp.83-105.

[12]M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall , 2002