ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Predictive Analysis through Mining in Distributed Databases

Poonam Joshi[1], Anjali Nava[2], Vivek Agarwal[3], Yadnya Nakhwa[4], Maanadh Naik[5]

Assistant Professor, Department of Information Technology, Atharva College of Engineering, Mumbai, India[1]

Student, Department of Information Technology, Atharva College of Engineering, Malad, Mumbai, India[2]

Student, Department of Information Technology, Atharva College of Engineering, Malad, Mumbai, India[3]

Student, Department of Information Technology, Atharva College of Engineering, Malad, Mumbai, India[4]

Student, Department of Information Technology, Atharva College of Engineering, Malad, Mumbai, India[5]

ABSTRACT*:* E-commerce has seen a great spurt in growth in the recent growing years; almost all businesses have their domains online. The current Systems provide the customer with limited searching options. As a result the customer has to search his desired product all by himself. This system can be used as a solution for generating optimized suggestions in Distributed Databases. We make use the concept of Extended Matrix-Based Apriori algorithm.

KEYWORDS*:* Data Mining, Association Rule Mining, distributed database, Extended-Apriori Algorithm, Frequent item-sets.

I. Introduction

Data mining is the practice of examining large pre-existing databases in order to generate new information. Data Mining is primarily used today by companies with a strong consumer focus — retail, financial, communication, and marketing organizations, to "drill down" into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments. There are different kinds of data mining techniques are available. Classification, Clustering, Association Rule and Neural Network Weka are some of the most significant techniques in data mining. In business world today, Online Shopping has proved to be a very successful and popular means of shopping. It first started with just simple book store but now everything is available online at very reasonable prices. Thus with the help of Data Mining it's possible to increase the sales of the products. The overall goal of data mining in this project is to extract information from a data set and analyse the transaction data and transform it into an understandable structure for further use and group the products which are most frequently purchased together. When customer searches for a product, all the products which were frequently purchased together with the searched product will be displayed to him along with the product he searched for. It is obvious that related products are displayed together, a person`s desire increases to purchase both of them. E.g. If we keep an offer on a Smartphone, then a Back-cover and glass-protector are displayed below together. Then it is more probable that customer will purchase all three instead of just a Smartphone .Information extraction methodology separates valuable Subsets of information that can be used for mining.

II. Related work

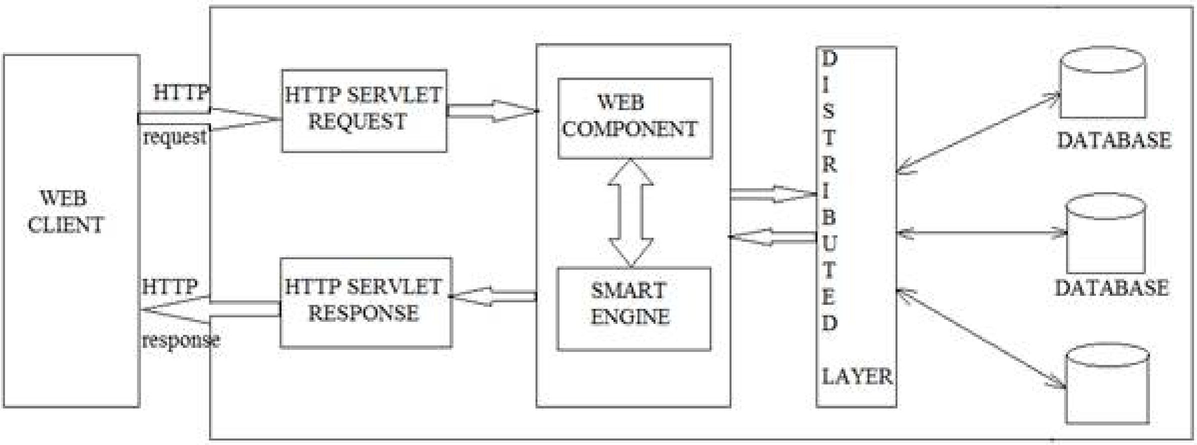
Data mining is the process of extracting hidden patterns from data. Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. The main process of Knowledge Discovery Database (KDD) is the mining process in which different algorithms are applied to produce hidden knowledge from raw data. In post-processing, which, mining result is evaluated according to user's requirements and domain knowledge. A distributed database system consists of loosely coupled sites that share no physical component in

DOI: 10.15680/IJIRCCE.2017. 0503366

Database systems that run on each site are independent of each other. Transactions may access data at one or more sites. A DDBMS is the software that manages the DDB and provides an access mechanism that makes this distribution transparent to the users [1]. Association rule mining is an active data mining research area and most ARM algorithms cater to a centralized environment. However, adapting centralized data mining to discover useful patterns in distributed database is not always feasible because merging datasets from different sites incurs huge network communication costs. Most existing parallel and distributed ARM algorithm are based on a kernel that employs the well-known Apriori algorithm. Directly adapting an Apriori algorithm will not significantly improve performance over frequent item-sets generation or overall distributed ARM performance. The main challenges include work-loading balancing, which is especially important for DARM. The distributed database in our model is a horizontally partitioned database, which means the database schema of all the partitions are the same [2]. E-commerce application generate huge amount of operational and behavioural data. Applying association rule mining in E-commerce application can unearth the hidden knowledge from these data. In this paper a survey of association rule mining and its various applications in E-commerce environment are made. In the E-commerce environment all the actions of customers visiting a shop from entry to exit are recorded. So customer navigation pattern and their purchasing behaviour are available in the E-commerce data. Finding association rules from these data helps to make right business decisions in right time. It also helps to improve cross selling website design. Also association rule mining in E-commerce data provide navigation and purchasing suggestions to customers [3]. A novel approach is proposed to improve the Apriori algorithm through the creation of Matrix-File, where the database transactions are saved. Thus repeated scanning is avoided and particular rows & columns are extracted and perform a function on that rather than scanning entire database. The main intention is to determine correlations among large set of items in a database, Apriori algorithm is the first proposed algorithm used to extract association rules from large database. It consists of two procedures: First, finding the frequent item-set in the database using a minimum support and constructing the association rule from the frequent item-sets with specified confidence. The limitations of the algorithm summarized by the generation of a lot of candidate item-set and scans database every time. In other words if database contains huge number of transactions then scanning the database for finding the frequent item-sets will be time costly. These limitations give the opportunities for the researchers to find efficient algorithm with a motive of minimizing the time and number of database scans for Knowledge Discovery [4].An improved algorithm using matrix data structures with simply counting rows and columns and transaction reduction strategies using top down approach for finding out largest regular item-set to smallest regular item-set. Mining association rules is important process in data mining which shows relationship among the variable or affairs stored in data warehouse, database and other information repositories. Association rule mining is two-step process. First it generates regular/frequent item-set set of item having count equal or greater than user specified parameter i.e., minimum support and second it discovers association rules from these frequent item-sets. This paper puts forward an improved algorithm using matrix data structure with simply counting rows and columns and transaction reduction strategies using top down approach for finding out largest regular item-set to smallest regular item-set. In this way, it can greatly reduce complexity and increases the efficiency of improved algorithm [5]

III.PROPOSED ALGORITHM

I. PROPOSED SYSTEM:



Our System consists of 7 modules:
Web Client: It acts as an interface between a user and servlets. A client-tier component may be an application or Web client. A Web client contains two parts: dynamic Web pages and the Web browser.

Servlets:Servlet technology is used to create web application (resides at server side and generates dynamic web page).Servlet technology is robust and scalable because of java language. Before Servlet, CGI (Common Gateway Interface) scripting language was popular as a server-side programming language.

Http Servlet Request: The http servlet request units are responsible for forwarding the request to the Web Component from Web Client.

Http Servlet Response: The http servlet response units are responsible for forwarding the response to the Web Client from Smart Engine.

Web Component: Web Components consists of several separate technologies. Web Components can be assumed as reusable user interface widgets that are created using open Web technology. They are part of the browser, and so they do not need external libraries like jQuery. An existing Web Component can be used without writing code, simply by adding an import statement to an HTML page. With a Web Component, you can do almost anything that can be done with HTML, CSS and JSP, and it can be a portable component that can be re-used easily.

Smart Engine: Smart engine is responsible for running the apriori algorithm, and the smart prediction algorithm also the image searching algorithm. It also takes data from the databases, runs the algorithms and combines the result with web component and is forwarded to Response servlet.

Distributed Layer: Distributed layer contains the meta-data and determines as to where the query is to be sent depending on the category. It consists of data about the different categories present in the databases (eg. Electronics, Clothing etc.) This layer then directs the query to the database depending on its type.

Databases: A database is a collection of information that is organized so that it can easily be accessed, managed, and updated. In one view, databases can be classified according to types of content: bibliographic, full-text, numeric, and images .It consists of all the database entries. It consists of many tuples and their attributes.

Apriori Algorithm using Matrix:
Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence. Association rule generation is usually split up into two separate steps:

First, minimum support is applied to find all frequent item sets in a database.

Second, these frequent item sets and the minimum confidence constraint are used to form rules. The general structure of the new approach is shown in Figure 1

From the figure above, the new suggested approach consists of two parts:

First part, find the *frequent item-sets* in the database, this achieves in two steps1. Find the total number of times each item-sets occurs,

2. Find among these item-sets the one that satisfy the condition which is greater or equal to % Minimum Support.

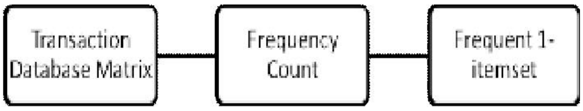


Fig. 1: Generation Frequent 1-Item-sets

*Second part,*prune columns of the Matrix whose frequency count are less than %Minimum Support, a new Matrix areformed with item sets which satisfies the Association rule. The new Matrix consists of frequent item sets only. Hence the size of the Matrix reduces drastically.

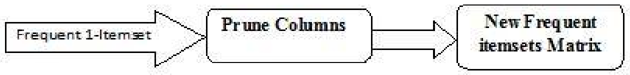


Fig.2: New Matrix Generation

The new Matrix approach is an enhancement to Apriori algorithm in terms of reducing the computation time and memory space, detailed explanation is in the following steps:

Frequent 1-Item-sets:

1. Matrix A contains the Transaction database where each column represents Item Number and row represents transaction of the customer. If the customer has purchased a particular item then it is represented by ‘1’. If the customer has not purchased particular item then is represented by ‘0’.
Frequency of all item sets which is called as Candidates for frequent item sets is found.
Matrix B contains the sum of individual columns, or in other words it counts item frequency, which is frequency of all item sets. As a result, frequency of item is found without scanning the database once again because the matrix already exists.
From Matrix B, a selection is done to frequencies which are greater or equal to the %Minimum Support, and prune the columns which are not frequent. A new Matrix C, is generated which is nothing but Frequent 1-Item-sets Matrix. Simultaneously in another Matrix D, the item number of frequent item sets is stored.

Association rule mining technique:
Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.
Let I= {i1, i2, i3} be a set of n binary attributes called *items*.
Let D = {t1, t2 ...tm} be a set of transactions called the *database*.

Each *transaction* in D has a unique transaction ID and contains a subset of the items in I.

A *rule* is defined as an implication of the form:

X->Y

Where X, Y is a subset of I and X is an intersection of Y which is null.

Every rule is composed by two different sets of items, also known as *item-sets*, X and Y, where X is called *antecedent* or left-hand-side (LHS) and Y *consequent* or right-hand-side (RHS).

IV. Pseudo code

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence. Association rule generation is usually split up into two separate steps:

First, minimum support is applied to find all frequent item sets in a database.

Second, these frequent item sets and the minimum confidence constraint are used to form rules. The general structure of the new approach is shown in Figure 1

From the figure above, the new suggested approach consists of two parts:

First part, find the *frequent item-sets* in the database, this achieves in two steps

1. Find the total number of times each item-sets occurs,

2. Find among these item-sets the one that satisfy the condition which is greater or equal to % Minimum Support.

Second part, prune columns of the Matrix whose frequency count are less than %Minimum Support, a new Matrix are formed with item sets which satisfies the Association rule. The new Matrix consists of frequent item sets only. Hence the size of the Matrix reduces drastically. The new Matrix approach is an enhancement to Apriori algorithm in terms of reducing the computation time and memory space.

V. SIMULATION RESULTS

The simulation involves displaying related products as per the user’s purchases. This however is done by implementing the Extended-Apriori Algorithm in fig.1. The output is based on client-side and server-side. On the server-side, the company can analyze the pattern of purchases and thereby identify associations between products as shown in fig.2. By using Distributed environment, real time simulation is shown as all the data cannot be centralized. Data is distributed categorically to prevent redundancy and inconsistency problems. As shown in fig.3 we have established a connection with a slave database by sending a ‘ping’ to its IP address. The selection of database to which a request should be sent is determined by look-up on the ‘server-details’ table in the master database. Each database has an id assigned depending on the category as shown in fig.4. Users often do not know the name of a product needed. But they have an image of it. Then they can upload the image on the website and search for the product by using Image-Based Searching as shown in fig.5.

DOI: 10.15680/IJIRCCE.2017. 0503366

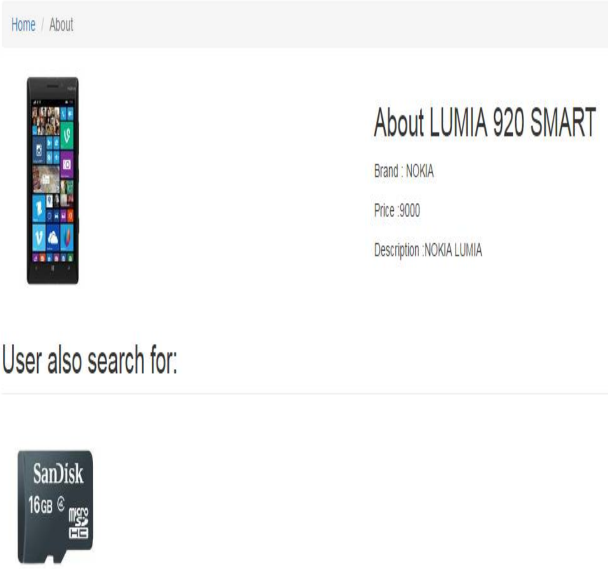
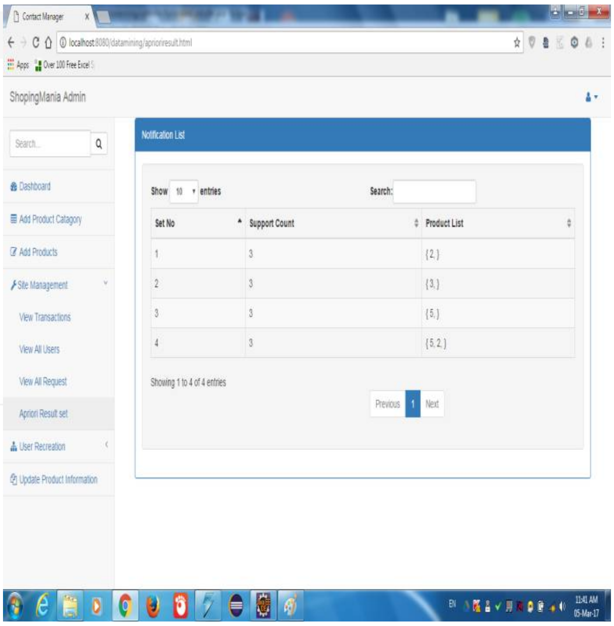


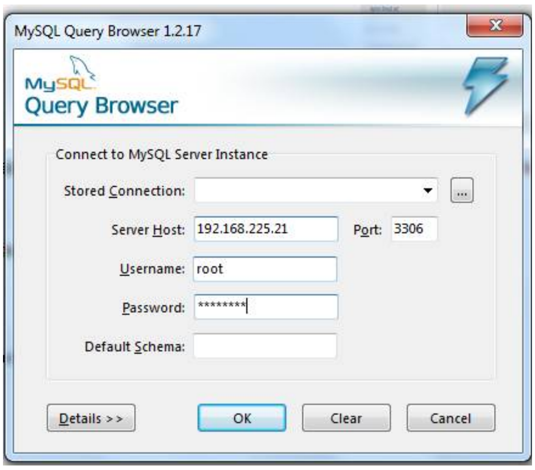
Fig.1. Smart Suggestions



Set No	Support Count	Product List
1	3	{2,}
2	3	{3,}
3	3	{5,}
4	3	{5,2,}

Showing 1 to 4 of 4 entries

Fig. 2.Sales Analysis by result of Apriori Algorithm



Stored Connection:

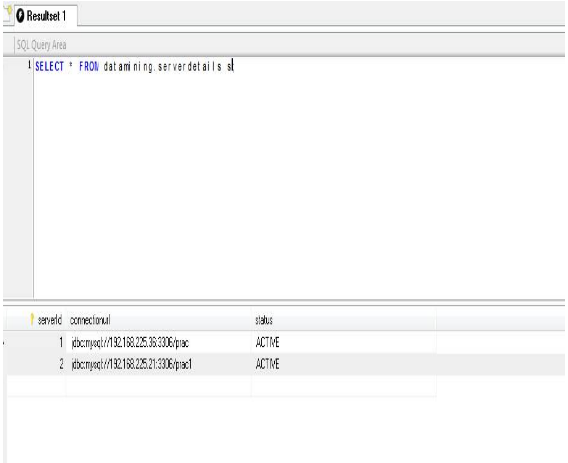
Server Host: 192.168.225.21 Port: 3306

Username: root

Password: ********

Default Schema:

Fig. 3.Connecting to a Slave database



```
SELECT * FROM datamining.serverdetails s
```

serverId	connectionurl	status
1	jdbc:mysql://192.168.225.36:3306/prac	ACTIVE
2	jdbc:mysql://192.168.225.21:3306/prac1	ACTIVE

Fig 4. Distributed Layer

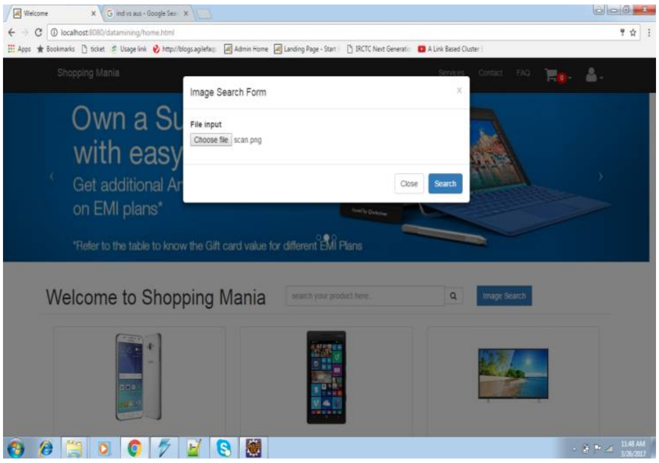


Fig.5.Image-Searching feature

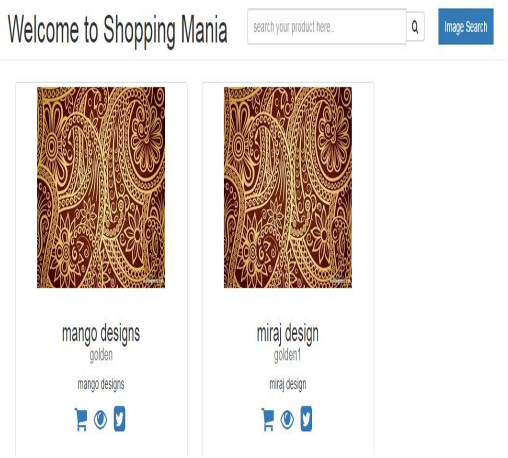


Fig.6. Image Search Results

VI. CONCLUSION AND FUTURE WORK

The main aim of this system is to predict possible combinations of products which are more likely to be purchased together with the help of Extended Matrix based-Apriori algorithm for data mining tool. This algorithm was made to run after scheduled time so that the time taken by this algorithm does not affect the searching time of user as it would have if it was implemented such that it executes at time when user hits ‘Search’ button. This system also enables the user to search a product despite not knowing the name of the product through the image searching option. It will also give every customer smart suggestions based on his previous purchases and suggest latest products suited for the customer. This increases the likelihood that the customer may indeed buy the product thereby increasing sales even more. Unlike the current systems, giving freedom to the customer of choosing his preferred date will add to his convenience while buying the products. The future of this system can be working on a dynamic dataset, integrating various other mining algorithms and further backing up of databases in RAID form and security aspects.

REFERENCES

[1] NayanaMarodkar, ManojChaudhari, ‘Mining of Association Rules in DistributedDatabases’, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, Volume 4 Issue 2, February 2015

[2] David W. Cheung, Member IEEE, Vincent T. Ng, Ada W. Fu, Member IEEE, and YongjianFu ‘Efficient Mining of Association Rules in Distributed Databases’, IEEE transactions on knowledge and data engineering, Volume 8, No. 6

[3] A. Anitha, G. Raja Suhananthan, Dr. N. Krishnan, ‘An Efficient Association Rule Mining Model for Distributed Databases’, VIJCST ISSN : 0976-8491 (Online) Volume 3, Issue 1, Jan. - March 2012

[4] Dr. M. Renuka Devi, Mrs. A. Baby Sarojini ‘Applications Of Association Rule Mining In Different Databases ‘, ISSN: 2229-371X (Online) Volume 3, No. 8, August 2012

[5] Venkateswari S, Suresh.R.M, ‘Association Rule Mining In E-Commerce-A Survey’, ISSN:0975-5462 (Online)Volume 3, No. 4, April 2011

[6] ShikhaBhardwaj, PreetiChhikara, SatenderVinayak, NishantPai, KuldeepMeena ‘Improved Apriori Algorithm for Association Rules’, International Journal of Technical Research and Applications e-ISSN: 2320-8163(Online) Volume 3, Issue 3 (May-June 2015)

[7] VipulMangla, ChandniSarda, SarthakMadra ‘Improving the efficiency of Apriori Algorithm in Data Mining’, International Journal of Engineering and Innovative Technology (IJEIT) ISSN: 2277-3754 (Online) Volume 3, Issue 3, September 2013

BIOGRAPHY

Poonam Joshi is an Assistant Professor at Atharva College of Engineering, Mumbai in the Department of Information Technology; she has completed her B.E. in Computer Science and M.E in Information Technology.

Maanadh Naik is a final year student at Atharva College of Engineering, Mumbai. He is currently pursuing his B.E in Information Technology

Anjali Nava is a final year student at Atharva College of Engineering, Mumbai. She is currently pursuing her B.E in Information Technology

Vivek Agarwal is a final year student at Atharva College of Engineering, Mumbai. He is currently pursuing his B.E in Information Technology

Yadnya Nakhwa is a final year student at Atharva College of Engineering, Mumbai. She is currently pursuing her B.E in Information Technology