Data Prediction and Optimization in Distributed Databases

Maanadh Naik
Student
Information Technology Department
Atharva College of Engineering
Malad, Mumbai, India

Anjali Nava
Student
Information Technology Department
Atharva College of Engineering
Malad, Mumbai, India

Yadnya Nakhwa
Student
Information Technology Department
Atharva College of Engineering
Malad, Mumbai, India

Vivek Agarwal
Student
Information Technology Department
Atharva College of Engineering
Malad, Mumbai, India

Poonam Joshi
Assistant Professor
Information Technology Department
Atharva College of Engineering
Malad, Mumbai, India

***Abstract*-E-commerce is one of the recent growing trends; almost all businesses have their domains online. The existing 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 have used the concept of Extended Matrix-Based Apriori algorithm which is efficient. This is much more Customer-Oriented Scheme.**

Keywords—Data Mining, Association Rule Mining, distributed database, Extended-Apriori Algorithm

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 analyze 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 *sari*, it`s Earrings and bangles together then it is more probable that customer will purchase all three instead of just *sari.*

Information extraction methodology separates valuable Subsets of information that can be used for mining. Objectives of the extraction procedure are, recognizing concerned data in the database and transforming the database into some suitable manner for carrying out investigation by the information mining calculations.

Arrangement of Information is a standout amongst the most essential ventures in the information mining procedure. Vast database frameworks contain mistakes in the put away information. Inspecting the information for blunders, frameworks, mistakes and missing qualities to the nature of the information. This is the most drawn out and most critical process in the information planning methodology. Strength is an imperative property for the information mining frameworks. Thus, a few procedures are accustomed to figuring out how to information in this methodology. Information cleaning is considered to evacuate commotion and irregularity in the information. There are numerous purposes behind uproarious and deficient information. Some of strategies are accustomed to filling in the missing qualities. The most acclaimed one is the relapse systems for information cleaning. Information coordination combines information from numerous sources. These sources may incorporate numerous databases, information blocks, or level documents.

The principle issue of the incorporation is information confliction. Information change operations utilized for normalizations and conglomeration. Information are changed or incorporated into structure suitable for mining. Information change methodology incorporates smoothing, speculation, standardization and collection methods. Information diminishment operation utilized for lessen the information measure by utilizing one of the information collection, measurement decrease or information examination strategies. Information diminishment techniques can be accustomed to minimizing representation of the information, while diminishing the loss of data substance.

The greater part of the crude information are made and cleaned in the past steps. Hence, information are arranged for the information mining stage. Arranged information may contain numerous credits and we need to choose a subset of the qualities for utilizing as a part of information mining methodology.

A Data mining calculation takes information as data and produces yield as models or examples. In this step a canny strategies are connected keeping in mind the end goal to concentrate information designs. Visualization, order, grouping, relapse or affiliation calculations are utilized for diverse issue. There are numerous algorithmic ways to deal with separating helpful data from information.

II. RELATED WORK

ARM is an important topic of data mining and various developers have been working on different ARM algorithms. This section gives a brief description about the existing solution for generating frequent item sets through ARM.

Nayana Marodkar et al. [1] has proposed a model which makes use of data-mining in distributed databases. 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 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.

A.Anitha et al. [2] has proposed an ARM model for Distributed Databases. 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 distributed database in our model is a horizontally partitioned database, which means the database schema of all the partitions are the same.

Venkateswari S et al. [3] has proposed a model which gives an application of ARM in E-Commerce. E-commerce application generate huge amount of operational and behavioral 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 application 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.

A.Rehab et al. [4] has proposed a model to improve the apriori algorithm through the creation of matrix -file ,where the database transactions are saved. In this paper 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 itemset in the database using a minimum support and constructing the association rule from the frequent itemsets with specified confidence. The limitations of the algorithm summarized by the generation of a lot of candidate itemset and scans database every time. In other words if database contains huge number of transactions then scanning the database for finding the frequent itemsets 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.

Shalini Dutt et al. [5] has proposed a model which makes use of 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 itemset to smallest regular itemset. 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 itemset 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 itemsets. 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 itemset to smallest regular itemset. In this way, it can greatly reduce the complexity and increases the efficiency of improved algorithm.

Vartika Mohan et al.[6] has pointed out that Apriori Algorithm contains many loopholes such as frequently scanning of database, generating large number of candidate-key, in addition to all these Apriori Algorithm also consumes very large amount of storage space for its processing. Thus to solve limitations of Apriori Algorithm, the author has presented an algorithm i.e. Matrix-Over-Apriori. This algorithm is an improvement over Apriori Algorithm which is supported by matrix and the proposed algorithm is efficient in both ways i.e. space and time. Also it results in decrease of the amount of candidate keys that are produced during entire processing.

III. METHODOLOGY

A. Apriori Algorithm using Matrix

Association rules are usually required to assure a user-specified least amount of support and a user-specified minimum confidence. Association rule generation is usually split up into two separate steps:
First, least 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 itemsets* in the database, this achieves in two steps
1. Find the total amount of times each item sets occurs,
2. Find among these itemsets the one that satisfy the condition which is greater or equal to % Minimum Support.

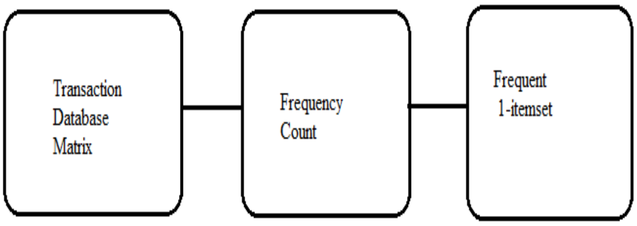

Fig. 1: Generation Frequent 1-Itemsets

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 repeated item sets only. Hence the size of the Matrix reduces significantly.

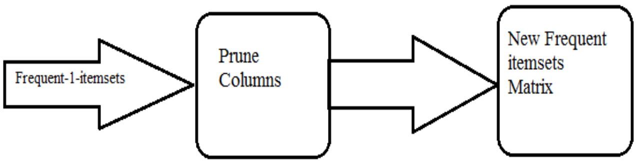

Fig.2 : New Matrix Generation

The new Matrix approach is an improvement to Apriori algorithm in terms of dropping the memory space and computation time. The following steps will explain in detail:

B. Frequent 1-Itemsetss

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 X, contains the sum of individual columns, or in other words it counts item occurrence, which is frequency of all item sets. As a result, occurrence of item is found without scanning the database once again because the matrix already exists.
From Matrix X, a selection is done to frequencies which are greater or equal to the %Minimum Support, and prune the columns which are not frequent. [3] A new Matrix C, is generated which is nothing but Frequent 1-Itemsets Matrix. Simultaneously in another Matrix D, the item number of frequent item sets is stored.

C. 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 *itemsets*, X and Y,where X is called *antecedent* or left-hand-side (LHS) and Y *consequent* or right-hand-side (RHS).

IV. ARCHITECTURE & IMPLEMENTATION

To resolve the problem of existing work, we propose a new approach Data Prediction And Optimization in Distributed Databases. This approach is an extension of the Apriori algorithm used in existing systems. The proposed work will solve problems of efficiency and accuracy. It will also tackle the issues of limited searching options.

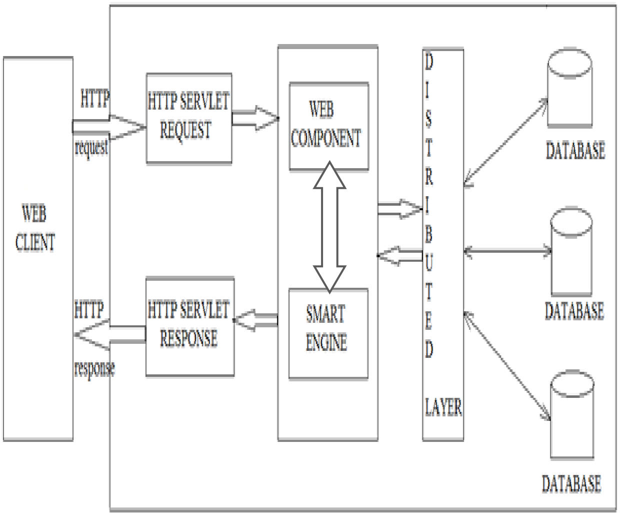

Fig. 3: System Block Diagram

*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 browser and the Web pages. Dynamic Web pages are created by components that run in the Web tier, and a Web browser delivers Web pages received from the server.

*Servlets***:** Servlet technology is used to generate web application (resides at server side and generates dynamic web page).Servlet technology is robust and scalable because of java language. Here are many classes and interfaces in the servlet API such as Servlet, GenericServlet, HttpServlet, ServletRequest, ServletResponse etc.

*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. Web Components use new or still-developing standard browser capabilities.
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..It consists of all the database entries. It consists of many tuples and their attributes.

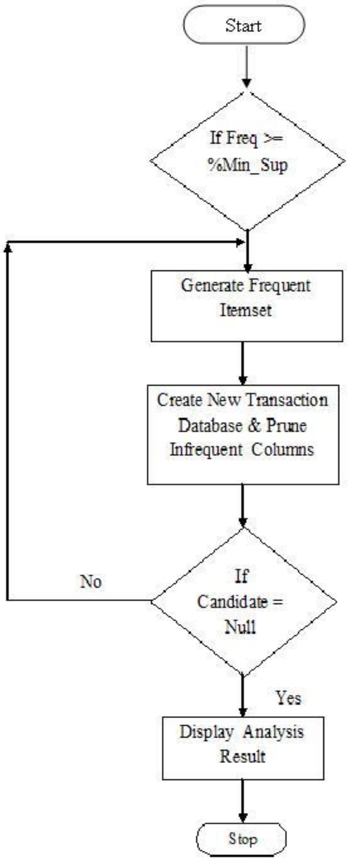
Start
If Freq >= %Min_Sup
Generate Frequent Itemset
Create New Transaction Database & Prune Infrequent Columns
No
If Candidate = Null
Yes
Display Analysis Result
Stop

Fig. 4: Algorithm Flowchart

V.CONCLUSION AND FUTURE RESEARCH 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 do 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 inclusion of voice recognition systems and further backing up of databases in RAID form and security aspects.

ACKNOWLEDGEMENT

We would also like to thank our project guide Prof.Poonam Joshi for her valuable guidance and also express our appreciation and gratitude to the IT department and Atharva College of Engineering authority for helping us to develop this paper.

REFERENCES.

[1] Nayana Marodkar, Manoj Chaudhari, "Mining of Association Rules in Distributed Databases" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064,Volume 4 Issue 2, February 2015

[2] A. Anitha, G. Raja Suhanantham, Dr. N. Krishnan, "An Efficient Association Rule Mining Model forDistributed Databases" IJCST ISSN : 0976-8491 (Online) Volume 3, Issue 1, Jan. - March 2012

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

[4] A.Rehab, H.Alwa & Anasuya V Patil "New Matrix Approach to Improve Apriori Algorithm" December.2013-Volume1.No4 http://www.ijcsns.com ISSN 2345-3397

[5] Shalini Dutt, Naveen Chaudhary & Dharm Singh "An Improved Apriori Algorithm Based on Matrix Data Structure" Global Journal of Computer Science and Technology: C Software & Data Engineering Volume 14 Issue 5 Version 1.0 Year 2014

[6] Vartika Mohan, Dharmveer Singh Rajpoot "Matrix-Over-Apriori: An Improvement Over Apriori Using Matrix". International Journal of Computer Science Engineering (IJCSE) ISSN : 2319-7323 Vol. 5 No.01 Jan 2016