International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

DATA MINING ALGORITHM IN DISTRIBUTED DATABASE

Vivek Gohil[1], Ketan Gharge[2], Ritesh Ghorui[3], Silviya Dmonte[4]

[1,2,3]Student, Dept of Computer Engineering, Universal College, Maharashtra, India
[4]Assistant Professor, Dept of Computer Engineering, Universal College, Maharashtra, India

---

Abstract – Data mining is a form of business intelligence and data analysis. It is the process of analysing data to draw useful conclusions or predictions from it. It's a technique frequently adopted by large-scale ecommerce businesses to aid with marketing and product development. Because of the nature of the internet, ecommerce businesses will obtain a lot of data about their customers, or their prospective customers. Data is obtained whenever a purchase is made, an account is created, or a page view is made. This raw data is obtained from the Distributed database. A utilization of a calculations to dispersed databases isn't successful, as it requires a lot of correspondence overhead. In our investigation, a proficient calculation, EDMA (Effective Data Mining Algorithm), is proposed. It minimizes the number of candidate sets and exchange messages by local and global pruning.

***Key Words*: Apriori algorithm, Association rules, parallel and distributed data mining.**

1. INTRODUCTION

Information mining is the act of looking at huge previous databases keeping in mind the end goal to produce new data 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 become very 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. In this project data mining is used to analyze the transaction data 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 natural when related products are displayed together, a person`s desire increases to purchase both of them. E.g. If we keep an offer on Mobile, it`s Screen protector and a back cover together then it is more probable that customer will purchase all three instead of just mobile.

Information mining varies from other exploration system in that it is proposed to deal with information without beginning from a specific theory, presumption or even a specific inquiry. Basically, it inverts the experimental system, beginning from information and moving towards speculations as opposed to taking after the conventional request (Berry, et al. 2000).

While data mining and learning disclosure in databases (KDD) are as regularly as conceivable viewed as proportionate words, information mining is really piece of the learning disclosure process (Zaiane 1999). The accompanying shows information mining as a venture in learning disclosure process.

Information extraction methodology separates valuable subsets of information for mining. Objectives of the extraction procedure are, recognizing concerned data in the database and transforming the database into some suitable structure to investigation by the information mining calculations.

Information arrangement is a standout amongst the most essential ventures in the information mining procedure. Vast database frameworks by and large contain mistakes in the put away information. Inspect the information for blunders, frameworks and missing qualities to the nature of the information. This is the most drawn out and most critical venture 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 can be connected 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 is 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.

2. Literature Review

The following research articles are selected for review, keeping in mind the traditional and conventional approach of an efficient association rule mining algorithm in distributed database.

Neha Saxena, Rakhi Arora, Ranjana Sikarwar and Ashika Gupta in their study of An Efficient Approach of Association Rule Mining on Distributed Database Algorithm journal Applications requiring huge data processing have two main problems, one a massive storage and its supervision and next processing time, when the quantity of data increases. Distributed databases determine the first trouble to a huge amount but second problem increase. Since, current stage is of networking and communication and community are involved in maintenance huge data on networks, therefore, researchers are propose a range of novel algorithms to raise the throughput of resulted data over distributed databases. Within our research, we are proposing an novel algorithm to process large quantity of data at the a variety of servers and collect the processed data on customer machine as much as necessary [1]

Mustapha Ismail, Mohammed Mansur Ibrahim, Zayyan Mahmoud Sanusi, Muesser Nat in their Information Mining in Electronic Business: The fundamental point of this paper is to audit the utilization of information mining in web based business by concentrating on organized and unstructured information gathered exhaustive different assets and distributed computing administrations keeping in mind the end goal to legitimize the significance of information mining. Also, this examination assesses certain difficulties of information mining like creepy crawly recognizable proof, information changes and influencing information to demonstrate fathomable to business clients. Other challenges which are supporting the slow changing dimensions of data, making the data transformation and model building accessible to business users are also evaluated An unmistakable manual for web based business organizations sitting on colossal volume of information to effortlessly control the information for business change which consequently will put them exceedingly focused among their rivals is additionally given in this paper. [2]

A. Prodromidis, P. Chan, and S. Stolfo. In their work on Parallal distributed data mining systems: Issues and approaches in AAAI/MIT Press, 2001.Association Rule Learning is a general technique used to discover associations amongst numerous variables. It is often used by grocery stores, retailers, and anyone with a bulky transactional database [7].

3. Proposed System

Since Mining is the crucial and very important in every sector of business it cannot be ruled out.
To make this system faster & efficient we divide or split larger chuck of operational data on multiple physical database

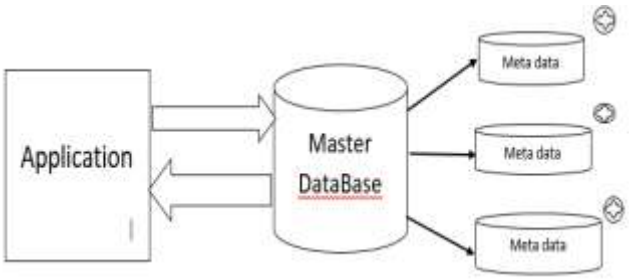

Figure 1. System Architecture

Following are the modules of this system:

3.1. Authentication,Authorization ,Registration Module:

- This will be secured prototype means every user has to sign up for using the prototype. Based on the role they will redirected to their corresponding home page.

3.2. Admin :

- Admin can add the product.
- Admin can view the details of the users.
- Admin can add the new category of the product.

3.3 Users :

- Search for the product.
- User can view own details & can make changes.
- Can add the product to cart.
- Can initiate new Transaction.

3. 4. Mining Module :

- There will be parallel mining performed at each distributed node on operational data.
- Mining algorithm will take past user transaction as input and generated corresponding output.
- This will generate various resultant set containing relative between products.

3.5. Fragmentation :

- It is technique of splitiing data on various distributed node.
- We have chosen Hybrid fragmentation approach.
- We will split the product based on category.This will fragment each product table at every node but with different product data.

3.6. Parallel processing :

- When user search for product .search on separate node takes place simultaneously.
- This improves performance.
- Also mining algorithm on each node will run parallely so that waiting time will be reduced.

3. 7. Query optimization :

- Query is run on subset of data resulting faster retrivation hence waiting time is reduced.

3. 8. Technical feasibility :

- We have chosen Java-8 as implementation language.
- My-SQL 5.6 as Relational Database.
- For powerful front end we have chosen JSP.
- To host application we have chosen Apache Tomcat 7.0

4. **Discussion**

Electronic business, or e-business, is the use of data and correspondence advancements (ICT) in help of the considerable number of exercises of business. Trade constitutes the trading of items and administrations between organizations, gatherings and people and can be viewed as one of the fundamental exercises of any business Electronic trade centers around the utilization of ICT to empower the outside exercises and connections of the business with people, gatherings and different organizations or e business alludes to business with help of web i.e. working with the assistance of web organize.

Mostly, customers are attracted to such software, tools, etc. which are easy to use and user friendly. Thus it would be beneficial to develop a Shopping site which is more user-friendly and with a good User Interface. Following points of improvements can be considered while making such sites.

- Accessibility,
- Quality,
- Availability,

If above mentioned problems are considered, it is possible to improve the sales by increasing customer`s desire of shopping.

- To analyze the transactions.
- To shortlist products.
- To examine the product which customer searches most frequently.
- To give the suggestions on the basis of finding of the products.
- To suggest a good relative product along with the searched product.

With the help of Data Mining it's possible to increase the sales of the products. In this project data mining is used to analyze the transaction data 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 natural when related products are displayed together, a person`s desire increases to purchase both of them. E.g. If we keep an offer on Mobile, it`s Screen protector and a back cover together then it is more probable that customer will purchase all three instead of just mobile.

5. CONCLUSION

Analyzing the previous records in order to boost up the current business with help of identifiable pattern is a good strategy for increase in business which ultimately leads to increased profit. The aim of data mining is to extract knowledge from information stored in database and generate clear and understandable description of patterns. The main aim of this paper is to predict possible combinations of products which are more likely to be purchased together with the help of Apriori algorithm for data mining tool. Then this algorithm was implemented using Association data mining technique. 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 algorithm shortlists the products in to on basis of how frequently they are purchased together and helps us in displaying those grouped products with any one of the searched product so that user gets attracted towards the combos or such offers. This algorithm can also be used in other systems like E Learning. In that same thing can be applied for analyzing most frequently accessed tutorials by a user and group tutorials into a set consisting of tutorials of related subjects. Eg if a searches for Core Java Tutorial, then by analyzing history of other users, we can create a set of tutorials related to java,

i.e. say Applet or AWT and suggest them to other users. This way we can make it user friendly.

REFERENCES

[1] A. W. Jilani," Usage and Effectiveness of E-Commerce Tool amongBusiness-To-Consumer", 2017.

[2] Neha Saxena, Rakhi Arora, Ranjana Sikarwar and Ashika Gupta, "An Efficient Approach of Association Rule Mining on Distributed Database", Volume 3, Number 4 (2013).

[3] Mustapha Ismail, Mohammed Mansur Ibrahim, Zayyan Mahmoud Sanusi, Muesser Nat,"Data Mining in Electronic Commerce", 2015.

[4]Miva, M. and Miva, B.The History of Ecommerce: How Did It All Begin?http://www.miva.com/blog/the-history-Of ecommerce-how -didit-all-begin. 2011

[5] A. Prodromidis, P. Chan, and S. Stolfo,"Parellal,distributed datamining systems",2001.