International Journal of Computer Trends and Technology (IJCTT) – volume 23 Number 1 – May 2015

Combining and Analysing Apriori and K-Means Algorithms for Efficient Data Mining on the Web

Nisha Rani[1],Yamini Chouhan[2]

[1]*M.E , Computer Science & Engg. Dept.Shankaracharya Group of Institutions, Bhilai (C.G.), India*
[2]*Asst.Professor, Computer Science & Engg. Dept.Shankaracharya Group of Institutions, Bhilai (C.G.), India*

***Abstract*— Web mining is the combination of data assembled by combining information mining techniques and procedures with data accumulated over the World Wide Web. Mining means extricating something helpful or important from a large no of datasets .Web mining is utilized to comprehend client conduct, assess the adequacy of a specific Web website, and help evaluate the accomplishment of a specified task. The proposed work is aimed to find a solution for generating different frequent item sets at each site in a distributed network. Apriori algorithm is a very popular algorithm for data mining that is dependent upon reducing infrequent item from item sets for mining useful data. Apriori algorithm can be very slow because of no of transactions. In order to increase the efficiency of the algorithm the initial item set is further clustered using K-Means algorithm. Cloud computing and data mining are emerging technologies dealing with major issues such as security and scalability and efficiency. The proposed work aims to increase efficiency of both the technologies.**

***Keywords*— Cloud Computing, Apriori Algorithm ,K-Means Algorithm.**

I. Introduction

Data mining, the extraction of concealed prescient data from large databases, is a compelling new innovation with incredible potential to help organizations concentrate on the most essential data in their information stores. Most organizations effectively gather and process enormous amounts of information for decision making purposes.

Information mining strategies can be actualized quickly on existing programming and equipment environments to improve the benefit of existing data stores, and can be combined with new items and frameworks as they are brought on line. With the incorporation of data sources on line, it is needed that clients use improvised techniques to discover the improvised data stores, and to track and dissect their use designs.

These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge.

A. *Web Mining*

Web mining can be extensively characterized as the disclosure and investigation of helpful data from the World Wide Web. This portrays the programmed searching of data that is available through web based techniques, There are around three information disclosure spaces that relate to web mining: Web Based Content Mining, Web Based Structure Mining, and Web Based Usage Mining.

Web based content mining: The methodology of separating learning from the substance of records or from their existing appearance. Web report content mining, asset disclosure taking into account ideas indexing or specialists based innovation might likewise fall in this classification.

Web based structure mining: The procedure of inducing learning from the World Wide Web association and connections in the middle of references and referents in the Web.

Web usage mining: Involves logging the data usage info on web, is the procedure of separating fascinating examples in web access.

B. *Cloud Computing*

Cloud computing is an recently evolved technology and involves utilization of processing assets.The concept involves collection of remote servers and programming systems that permit concentrated information stockpiling and online access to PC administrations or assets. Cloud computing, or in less complex shorthand simply "the cloud", involves in increasing the availability of the imparted assets. Cloud assets are normally distributed by different clients . This can work for distributing assets to clients. Distributed computing platform displays the involved key attributes:

Agility: Enhances with clients' capacity to re-procurement innovative base assets.

Application programming interface (API) : Availability to programming that empowers machines to interface with cloud programming in the same way that a customary client interface encourages association among people and PCs

Cost : Decreases cost . An open cloud conveyance model believers capital consumption to operational usage. This brings problems to existing passages, as required info is commonly given by an outsider and does not have to be obtained for one-time or not commonly registered assignments. Evaluating on an utility processing is a process involving use based alternatives and less IT aptitudes are needed for execution (in-house).

Device and location independence : Empower clients to get to frameworks utilizing a web program paying little mind to their area or what gadget they utilize (e.g., PC, cellular telephone).

Maintenance : Distributed computing applications can be easily maintained, because they don't have to be introduced on each client's PC and can be gotten to from better places.

Performance : Execution is observed, and predictable and approximately coupled architectures are developed utilizing web benefits as the framework interface.

Productivity: Gainfulness expanded when different clients can deal with the same information all the while easily instead of being communicated.

Reliability: Enhances the utilization of different repetitive destinations that is suitable of business coherence and calamity recuperation.

II. LITREATURE REVIEW

- Authors Loraine Charlet Annie M.C , Ashok Kumar D in their work "Market Basket Analysis for a Supermarket based on Frequent Itemset Mining ", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 stated that Market basket analysis is an important component of analytical system in retail organizations to determine the placement of goods, designing sales promotions for different segments of customers to improve customer satisfaction and hence the profit of the supermarket.

- Authors Ritika Agarwal, Dr. Barjesh Kochar, Deepesh Srivastava in their work "A Novel and Efficient KNN using Modified Apriori Algorithm", International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012 stated that In the field of data mining, classification and association set rules are two of very important techniques to find out new patterns. K-nearest neighbor and apriori algorithm are most usable methods of classification and association set rules respectively. However, individually they face few challenges, such as, time utilization and inefficiency for very large databases. The current paper attempts to use both the methods hand in hand.

- Authors Mohammad Ali Farajian & Shahriar Mohammadi "Mining the Banking Customer Behavior Using Clustering and Association Rules Methods", International Journal of Industrial Engineering & Production Research , December 2010, stated that In the field of data mining, classification and association set rules are two of very important techniques to find out new patterns. K-nearest neighbor and apriori algorithm are most usable methods of classification and association set rules respectively. However, individually they face few challenges, such as, time utilization and inefficiency for very large databases. The current paper attempts to use both the methods hand in hand.

- Authors Chad West, Stephanie MacDonald, Pawan Lingras , and Greg Adams in their work titled "Relationship between Product Based Loyalty and Clustering based on Supermarket Visit and Spending Patterns ", International Journal of Computer Science & Applications , 2005 stated that Loyalty of customers to a supermarket can be measured in a variety of ways. If a customer tends to buy from certain categories of products, it is likely that the customer is loyal to the supermarket. Another indication of loyalty is based on the tendency of customers to visit the supermarket over a number of weeks. Regular visitors and spenders are more likely to be loyal to the supermarket. Neither one of these two criteria can provide a complete picture of customers' loyalty.

- Authors Thomas H Beach, Omer F Rana, Yacine Rezgui and Manish Parashar in their work titled "Cloud computing for the architecture, engineering & construction sector: requirements, prototype & experience ", Beach et al. Journal of Cloud Computing: Advances, Systems and Applications 2013 stated that The Architecture, Engineering & Construction (AEC) sector is a highly fragmented, data intensive, project based industry, involving a number of very different professions and organisations. Our efforts in engaging with the industry have shown that Cloud Computing is still an emergent technology within the AEC sector. Technologies such as Google Drive and DropBox are often used informally and in an adhoc way between individuals - but concerns over security and the protection of intellectual property often dissuade major companies from adopting such services.

III. PROBLEM IDENTIFICATION

Apriori algorithm, in spite of being simple and clear, has some limitations. It is costly to handle a large number of candidate sets. For example, if there are 104 frequent 1-item sets, the Apriori algorithm ill need to generate more than 107 length-2 candidates ,accumulate and test their occurrence frequencies. Moreover, to discover a frequent pattern of size 100, such as {a1, a2. . . , a100}, it must generate 2100 -2 ˜ 1010 candidates in total. This is the inherent cost of candidate generation, no matter what implementation technique is applied. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns. Apriori Algorithm Scans the database too many times, When the database storing a large number of data services, the limited memory capacity, the system I/O load, considerable time scanning the database will be a very long time, so efficiency is very low.

Method sto improve the algorithm performance can be:

- **Reducing Transactions:** Reduction of frequent transactions.
- **Itemset Partioning:** Any itemset that is frequent in database must be frequent in at least one of the partitions of database.
- **Creating a subset of data:** Mining on a subset of given data, lower support threshold value a method to determine the completeness,

IV. PROPOSED METHODOLOGY

A. *Apriori Algorithm*

- The Apriori Algorithm is an popular algorithm for mining frequent itemsets obtaining association rules.
- Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time .This step is known as candidate generation, and generated candidates are tested against the data.
- Apriori is useful for mining useful information from database containing transactions like collections of items bought by customers.

Basic steps are:

- Discover all regular itemsets:
- Get frequent items: Things whose event in database is more noteworthy than or equivalent to the min.support limit.
- Get frequent itemsets: Produce candidates from frequent itemsets and Prune the outcomes to discover the continuous itemsets.
- Produce association rules from successive itemsets .Rules which fulfill the min.support and min.confidence edge.

B. *K-Means Algorithm*

The K-Means algorithm is a simple yet effective statistical clustering technique. Basic steps are:

- Choose a value for K, for determing no of clusters.
- Choose K data points) from dataset at random. These are the initial cluster centers.
- Use simple Euclidean distance to assign the remaining instances to their closest cluster center.
- Use the instances in each cluster to calculate a new mean for each cluster.
- If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centers and repeat steps 3-5.

A. *Flow Chart*

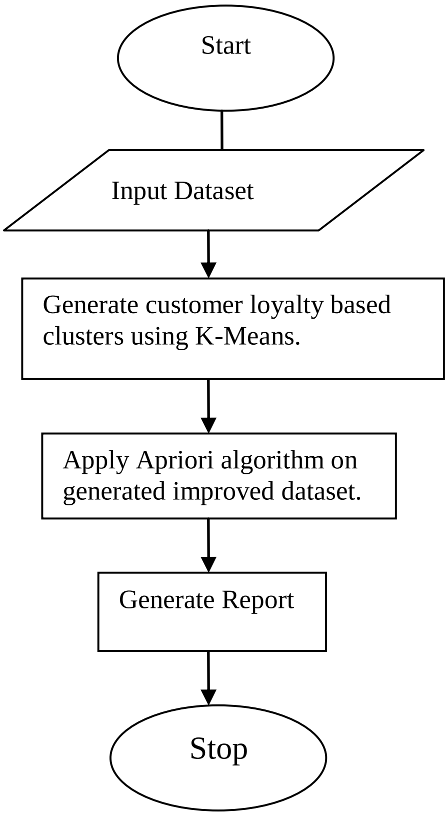
Start
Input Dataset
Generate customer loyalty based clusters using K-Means.
Apply Apriori algorithm on generated improved dataset.
Generate Report
Stop

Fig: Proposed Methodology

V. CONCLUSION

Cloud computing and data mining are emerging technologies dealing with major issues such as security and scalabilty and efficiency. The proposed work aims to increase efficiency of both the technologies. Future work aims to increase the efficiency by using better parameters of other algorithms also.

It is not possible to develop a system that makes all the requirements of the user. User requirements keep changing as the system is being used. Some of the future enhancements that can be done to this system are:

- As the technology emerges, it is possible to upgrade the system and can be adaptable to desired environment.
- Because it is based on object-oriented design, any further changes can be easily adaptable.
- The efficiency of algorithm can be further increased by applying more efficient data mining algorithms in near future.
- More work is possible on security of data in cloud servers.

- Security can be increased by applying efficient encryption/decryption algorithms.

REFERENCES

[1] *[1] Agrawal R, Srikant R (1994) Fast algorithms for mining associationrules.In:Proceedings of the 20th VLDB conference, pp 487–499*

[2] [2] Mining Association Rules between Sets of Items in Large Databases:Rakesh Agrawal ,Tomasz Imielinski,Arun SwamiACM SIGMOD ConferenceWashington DC, USA, May 1993

[3] [3] G.K. Gupta,Introduction to data mining with case studies:Prentics Hall of India, New Delhi, 2006

[4] [4] Han, David, et al. Principles of Data Mining: MIT press. Cambridge, 2001.

[5] [5] Mining Association Rules between Sets of Items in Large Databases:Rakesh Agrawal ,Tomasz Imielinski,Arun SwamiACM SIGMOD ConferenceWashington DC, USA, May 1993

[6] [6] Fast Algorithms for Mining Association Rules: Rakesh Agrawal Ramakrishnan Srikant VLDB Conference Santiago, Chile, 1994

[7] [7] High Performance Data Mining Using the Nearest Neighbor Join Christian Böhm Florian Krebs

[8] [8] A Review of various k-Nearest Neighbor Query Processing Techniques :International Journal of Computer Applications (0975 – 8887) Volume 31–No.7, October 2011

[9] [9] Mining of Meteorological Data Using Modified Apriori Algorithm,European Journal of Scientific Research ISSN 1450-216X Vol.47 No.2 (2010), pp.295-308EuroJournals Publishing, Inc. 201

[10] [10] Fast Algorithms for Mining Association Rules: Rakesh Agrawal Ramakrishnan Srikant VLDB Conference Santiago, Chile, 1994High Performance

[11] [11] Data Mining Using the Nearest Neighbor Join Christian Böhm Florian KrebsA Review of various k-Nearest Neighbor Query Processing

[12] [12] Top 10 algorithms in data mining, Xindong Wu, Springer-2007

[13] [13] Han, Jiawei and Kamber, Micheline, Data Mining Concepts and Techniques. Morgan Kaufman Publishers. San Fransisco 2000.