Improving the Performance of Apriori Algorithm by Combining with Clustering Techniques

Nisha Rani
Research Scholar, Computer Science & Engg. Dept.Shankaracharya Group of Institutions, Bhilai (C.G.), India.

Yamini Chouhan
Asst.Professor, Computer Science & Engg. Dept.Shankaracharya Group of Institutions, Bhilai (C.G.), India.

Abstract – Data mining on World Wide Web is utilized to comprehend client conduct, assess the adequacy of a specific Web website, and help evaluate the accomplishment of a specified task. It is the combination of data assembled by customary information mining philosophies and procedures with data accumulated over the World Wide Web. Data Mining means extricating something helpful or important from a baser substance. This paper is aimed to find a solution for generating different frequent item sets at each site in a cloud based 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.

Index Terms – World Wide Web , Web Mining , Cloud Computing , Apriori Algorithm , K-Means Algorithm.

1. INTRODUCTION

Information mining strategies can be actualized quickly on existing programming and equipment stages to improve the benefit of existing data assets, and can be incorporated with new items and frameworks as they are brought on the web. With the dangerous development of data sources accessible on the World Wide Web, it has gotten to be progressively vital for clients to use mechanized instruments in discover the coveted data assets, and to track and dissect their use designs. The extraction of data from expansive databases is a compelling new innovation with incredible potential to help organizations concentrate on the most essential data in their information stockrooms. Most organizations effectively gather and refine enormous amounts of information. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge.

Web Data Mining: Web mining can be extensively characterized as the disclosure and investigation of helpful data from the World Wide Web. This portrays the programmed hunt of data assets accessible online, i.e. Web substance mining, and the revelation of client access designs from Web servers, there are around three information disclosure spaces that relate to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining.

Web based content mining: It is the process of separating learning from the substance of records or their portrayals. Web report content mining, asset revelation taking into account ideas indexing or specialists based innovation might likewise fall in this classification.

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

Web based usage mining: It is the process of separating fascinating examples in web access.

2. RELATED WORK

Market basket analysis [1] 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. In the field of data mining, [2] 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. In the field of data mining, classification [3] 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. Loyalty of customers [4] 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. The Architecture, Engineering & Construction (AEC) [5] sector is a highly fragmented, data intensive, project based industry, involving a number of very different professions and organizations. 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 ad hoc way between individuals - but concerns over security and the protection of intellectual property often dissuade major companies from adopting such services.

3. PROBLEM IDENTIFICATION

Apriori algorithm, in spite of being simple and clear, has some limitation. It is costly to handle a huge number of candidate sets. For example, if there are104 frequent 1-item sets, the Apriori algorithm ill need to generate more than107 length-2 candidates and 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.

Methods to improve the algorithm performance can be:

- **Transaction reduction:** Reduction of frequent transactions.
- **Partitioning:** Any item set that is frequent in database must be frequent in at least one of the partitions of database.
- **Sampling:** Mining on a subset of given data, lower support threshold value a method to determine the completeness,

4. PROPOSED METHODOLOGY

Apriori Algorithm:The Apriori Algorithm is an popular algorithm for mining frequent item sets 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 item sets.
- 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 fulfil the min.support and min.confidence edge.

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.

4.1. Flow Chart

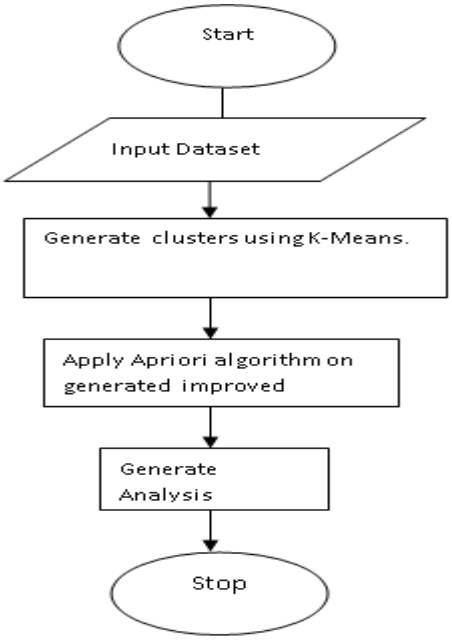

Figure 1 Proposed Methodology

5. CONCLUSION

Data mining is emerging technology dealing with major issues such as security and scalability and efficiency. The proposed work aims to increase efficiency of exiting algorithms namely Apriori and K-Means technique. 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] Agrawal R, Srikant R (1994) Fast algorithms for mining associationrules.In:Proceedings of the 20th VLDB conference, pp 487–499.
[2] Mining Association Rules between Sets of Items in Large Databases:Rakesh Agrawal ,Tomasz Imielinski,Arun SwamiACM SIGMOD ConferenceWashington DC, USA, May 1993.
[3] G.K. Gupta,Introduction to data mining with case studies:Prentics Hall of India, New Delhi, 2006.
[4] Han, David, et al. Principles of Data Mining: MIT press. Cambridge, 2001.
[5] Mining Association Rules between Sets of Items in Large Databases:Rakesh Agrawal ,Tomasz Imielinski,Arun SwamiACM SIGMOD ConferenceWashington DC, USA, May 1993
[6] Fast Algorithms for Mining Association Rules: Rakesh Agrawal Ramakrishnan Srikant VLDB Conference Santiago, Chile, 1994.
[7] High Performance Data Mining Using the Nearest Neighbor Join Christian Böhm Florian Krebs.
[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] 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] Fast Algorithms for Mining Association Rules: Rakesh Agrawal Ramakrishnan Srikant VLDB Conference Santiago, Chile, 1994High Performance.
[11] Data Mining Using the Nearest Neighbor Join Christian Böhm Florian KrebsA Review of various k-Nearest Neighbor Query Processing.
[12] Top 10 algorithms in data mining, Xindong Wu, Springer-2007