IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Association Rules Mining in Cloud Computing Environments using Improved Apriori Algorithm

Avinash Sharma[1*], Sarvottam dixit[2], N. K. Tiwari[3]

[1]Dept. of Computer Science Engineering, Mewar University Chittorgarh Rajasthan India
[2]Dept. of Computer Science Engineering Professor Mewar University Chittorgarh Rajasthan India
Director Patel Group of Institution Bhopal M.P. India

[*]*Corresponding Author: avinashvtp@gmail.com Tel.: +91-7999978185*

Available online at: www.ijcseonline.org

Accepted: 15/15/2018, Published: 31/Dec/2018

Abstract— This paper describes how data mining is used in cloud computing. Data Mining used for extracting potentially useful information from raw data. The integration of data mining techniques into normal day-to-day activities has become commonplace. Every day people confronted with targeted advertising, and data mining techniques help businesses to become more efficient by reducing costs. Cloud computing provides a powerful, scalable and flexible infrastructure into which one can integrate, previously known, techniques and methods of Data Mining. Data security and access control are the most challenging in cloud computing because users send their sensitive data to the cloud service providers. The service providers must have a suitable way to protect their client's sensitive data. Association rules are dependency rules, which predict occurrence of an item based on occurrences of other items. Apriori is the best-known algorithm to mine association rules. In this paper, we use Modified Apriori algorithm to mine the data from the cloud using sector/sphere framework with association rules.

Keywords—Data mining, Cloud Computing Association rules

I. INTRODUCTION

Data mining is the process of extracting meaningful information from data. It help in guessing a trend or value, classifying, categorizing the data, and in finding correlations, patterns from the data set Mining can be done irrespective of the storage format. Data stored in flat files, spreadsheet, word files. Data mining involves association rule mining process. In which done to extract interesting Correlation rule from the items. For example, association rule bread, butter generated from the transaction database of a grocery store can help in formulating marketing strategy around the rule. Association rules are widely used in various areas such as telecommunication networks, marketing and risk management, and inventory control etc. Many companies and firms keep large quantities of their day-to-day transaction data. These data could be analysed to learn the purchasing trend of the customer. Such valuable insight can be used to support variety of business-related applications such as marketing and promotion of the products, inventory management etc. Besides market based data analysis, association rules can be mined for the field of bioinformatics, medical diagnosis, web mining and scientific data analysis. All the above fields deal with the huge amount of input data, whose locations could be distributed.

Processing such huge data requires lots of resources and time. Data Mining algorithm generates extremely large number of association rules in many cases and sometimes the association rules are very large. It becomes almost impossible for the users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Thus we reach at the concept of generating only interesting rules, generating only non-redundant rules, or generating only those rules which satisfies certain criteria. The first is data parallelism in which the input data set could be divided among the participating node to generate the rules. The second method is of dividing the task among the nodes so that each node will access the whole input data set for generating the rules. The input data size is usually quite large and distributed in nature for association rule mining so cloud computing could be used for generating rules. Cloud computing allows consumers and businesses to use applications without any installation and access their files at any computer with internet access. It lets us do efficient computing by centralizing storage, memory, processing and bandwidth. Further, in cloud you pay have to pay only for services which you use and ac-cording to the duration of usage, thus making it a very good and inexpensive option for using it for association rule mining.

Association rule mining is an important research topic of data mining; its task is to find all subsets of items which frequently occur, and the relationship between them. Association rule mining has two main steps: the establishment of frequent item sets and the establishment of association rules. Apriori algorithm [3] is the most classic and most widely used algorithm for mining frequent item sets which generate Boolean association rules. The algorithm uses an iterative method called layer search to generate (k + l) item sets from the k item sets. In this paper we describe a new algorithm which provides the way for data mining or data mining association on cloud environment so that we can achieve a better way to handle a large amount of data.

II. LITERATURE SURVEY

In this section, we briefly review the most related studies including frequent pattern mining algorithms and parallel and distributed algorithms for frequent pattern mining.
In 2015, **Iyer Chandrasekharan P.K. Baruah [1]** Cloud computing has shown a new interest in a paradigm called Data mining is treated as service. This idea aimed at organizations that lack the technical expertise or the computational resources enabling them to outsource their data mining tasks to a third party service provider.

In 2015, **Abdur Rahim Mohammad Forkan [2]** Context-aware monitoring is an emerging technology that provides real-time personalised health monitoring services and a rich area of big data application. In this paper, they explain a knowledge discovery-based approach that allows the context-aware system to adapt its behaviour in runtime by analysing large amounts of data generated in ambient assisted living (AAL) systems and stored in cloud repositories

In 2014, **Liao et al.[3],** presented a MRPrePost algorithm based on MapReduce framework. MRPrePost is an improved version of PrePost. Performance of PrePost algorithm is improved by including a prefix pattern. On this basis, MRPrePost algorithm is well suitable for mining large data's association rules. In case of performance MRPrePost algorithm is more superior to PrePost and PFP. The stability and scalability of MRPrePost algorithm is better than PrePost and PFP. The mining result of MRPrePost is approximate which is closer to original result.

In 2014, **Aditi V. Jarsaniya, Shruti B. Yagniket al [4]** proposed A Literature Survey on Frequent Pattern Mining for Biological Sequence. Its limitation is, When the database is large, it is sometimes difficult and unrealistic to construct a main memory based FP-tree.

In 2014, **Dr.Vijayalakshmi M N, S.Anupama Kumar, Kavyashree BN [5]** presents the application of association mining on educational data to understand the knowledge and performance of students. Author implemented Apriori algorithm on student log data to bring out the interesting rules. Those rules can be used to infer the performance of the students and to impart the quality of education in the educational institutions. The algorithm generated frequent item sets using support measure in order to understand the interest of the students in the course. Interesting rules are generated based on frequent item sets using confidence factor of the dataset.

III. CLOUD COMPUTING

Data mining techniques and applications are very much needed in the cloud-computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage. Using data mining through Cloud Computing reduces the barriers that keep small companies from benefiting of the data mining instruments. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users. The implementation of data mining techniques through Cloud Computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage. CDM (Cloud Data Mining) offers tremendous potential for analysing and extracting the (useful) information in various fields of human activities: finance, banking, medicine, genetics, biology, pharmacy, marketing, etc. The application of this technology should enable that with just a few clicks of the mouse one can reach the desired information about customers, their habits, interests, purchasing power, frequency of purchases of certain items, location and so on. Cloud provides technology that can "handle" huge amounts of data, which cannot be processed efficiently and at reasonable cost using standard technologies and techniques. Data mining in Cloud (CDM) is, from a technical point of view, a very tedious process that requires a special infrastructure based on application of new storage technologies, handling and processing. Big Data/Hadoop is the latest type in the field of data processing.

IV. INTEGRATED DATA MINING AND CLOUD COMPUTING

Data mining in Cloud Computing allow the organizations to centralize the management of software and data storage with

assurance of efficient, reliable and secure services for their users. It provides technology that can handle large amounts of data which cannot be processed efficiently at reasonable cost using standard technologies and techniques. It also allows the users to retrieve meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage. We can provide new ways and means to effectively solve the distributed storage of massive data mining and efficient computing through Cloud Computing, mass data storage and distribution of computing, massive data mining environment for cloud computing. Extension of Cloud Computing will drive the Internet and technological achievements in the public service to promote the depth of information resources sharing and sustainable use of new methods and new ways of traditional data mining. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage with assurance of efficient, reliable and secure services for their users.

4.1 Advantages of IDMCC Integration

The following are the advantages of the Integrated Data Mining and Cloud Computing Environment

- Virtual computers that can be started with short notice
- Redundant robust storage
- No query structured data
- Message queue for communication
- The customer only pays for the data mining tools that he needs
- The customer doesn't have to maintain a hardware infrastructure as he can apply data mining through a browser

4.2 Advantages Of Using Data Mining With Cloud Computing

Cloud Computing combined with data mining can provide powerful capacities of management. Due to the explosive data growth and amount of computation involved in data mining, an efficient and high performance computing is an excellent resource necessary for a successful data mining application. Data mining in the cloud computing environment can be considered as the future of data mining because of the advantages of cloud computing paradigm. Cloud Computing provides greater capabilities in data mining and data analytics. The major concern about data mining is that the space required by the operations and item sets is very large.

4.3 Disadvantages Of Using Data Mining With Cloud Computing

There are certain issues associated with data mining in the cloud computing. The major issue of data mining with cloud computing is security as the cloud provider has complete control on the underlying computing infrastructure. Special care has to be taken so as to ensure the security of data under cloud computing environment.

V. ASSOCIATION RULES

Association rule is very popular and well researched method for discovering interesting relations between variables in large databases. Given a set of transactions, where each transaction is a set of items, an association rule is an expression X => Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions in the database which contain the items in X tend to also contain the items in Y. An example of such a rule might be that 98% of customers who purchase tires and auto accessories also buy some automotive services; here 98% is called the confidence of the rule. The support of the rule X => Y is the percentage of transactions that contain both X and Y. The problem of mining association rules is to find all rules that satisfy a user-specified minimum support and minimum confidence. Applications include cross marketing, attached mailing, catalogue design, loss-leader analysis, add-on sales, store layout, and customer segmentation based on buying patterns. The problem of mining association rules can be decomposed into two sub problems:

1. Find all sets of items (item sets) whose support is greater than the user-specified minimum support. Item sets with minimum support are called frequent item sets.

2. Use the frequent item sets to generate the desired rules. The general idea is that if, say, ABCD and AB are frequent item sets, then we can determine if the rule AB => CD holds by computing the ratio conf = support(ABCD)/support(AB). If conf >= minimum confidence, then the rule holds. That is, the rule will have minimum support because ABCD is frequent. Much of the research has been focused on the first sub problem as the database is accessed in this part of the computation and several algorithms have been proposed. In association rule mining algorithm, most of the algorithms are based on Apriori algorithm to calculate and in the mining process they can produce amount of option set which reduce the efficiency of association rule mining.

VI. PROBLEM IDENTIFICATION

Association rule mining is a popular and well-researched area for discovering interesting relations between variables in large databases for Cloud Computing Environment. We have to analyse the colouring process of dyeing unit using association rule mining algorithms using frequent patterns. These frequent patterns have a confidence for different treatments of the dyeing process. These confidences help the dyeing unit expert called dyer to predict better combination or association of treatments.

Various algorithms are used for the colouring process of dyeing unit using association rules. For example. LRM,FP

Growth Method., H-Mine and Aprori algorithm But these algorithm significantly reduces the size of candidate sets . However, it can suffer from three-nontrivial costs:
(1) Generating a huge number of candidate sets, and
(2) Repeatedly scanning the database and checking the candidates by pattern matching.
(3) It take more time for generate frequent item set.
(4) The large databases can not be executed efficiently in H-Mine and LRM algorithms,
We have to proposed such that algorithm that it has a very limited and precisely predictable main memory cost and runs very quickly in memory-based settings. it can be scaled up to very large databases using database partitioning and to identify the better dyeing process of dyeing unit.

VII. PROPOSED ALGORITHM

The Apriori algorithm had a major problem of multiple scans through the entire data. It required a lot of space and time. The modification in our paper suggests that we do not scan the whole database to count the support for every attribute. This is possible by keeping the count of minimum support and then comparing it with the support of every attribute. The support of an attribute is counted only till the time it reaches the minimum support value. Beyond that the support for an attribute need not be known. This provision is possible by using a variable named flag in the algorithm. As soon as flag changes its value, the loop is broken and the value for support is noted. The pseudo code for the proposed algorithm is as follows:

Input: Database, D, of transactions;
Minimum support threshold, min_sup
Output: L, frequent item sets in D
Method:

```
1) L (1) = find_frequent _1-itemsets (D);
2) For each transaction t belongs to D
3) count_items= count_items(t);
4) For (k=2; L(k-1)!=null; k++)
5) {
6) C(k)= apriori_gen(L(k-1, min_sup);
7) flag=1;
8) For each transaction t belonging to D
   Where count_items>=k
9) {
10) If (flag==1)
11) {
12) c=subset(C(k),t);
13) c.count++;
14) if (c.count==min_sup)
15) flag=0;
16) }
17) if (flag==0)
18) Exit from loop
19) }
20) L(k)={c.count=min_sup}
21) }
22) return L=U(k) L(k);
```

VIII. CONCLUSION

Cloud Computing provides storage of data in a server by protecting data by using data mining concept. Actually, we are discussing the cloud computing data mining for the advance use of security in data loss purpose. In Cloud computing, the data is being shifted from one server to another server in a peer-to-peer transaction. Data mining technologies provided through Cloud Computing is an absolutely necessary characteristic for today's businesses to make proactive, knowledge driven decisions as it helps them have future trends and behaviours predicted. In this paper, we have attempted to give a new perspective algorithm with the eye of a modified Apriori algorithm. This algorithm is better than both of the previous methods, i.e., FP Growth tree algorithm and TPFP algorithm.

REFRENCES

[1]. Iyer Chandrasekharan P.K. Baruah "Privacy-Preserving Frequent Itemset Mining in Outsourced Transaction Databases" Sri Sathya Sai Institute of Higher Learning Prashanti Nilayam, A.P., India 2015 IEEE.

[2]. Abdur Rahim Mohammad Forkan, Ibrahim Khalil "BDCaM: Big Data for Context-aware Monitoring - A Personalized Knowledge Discovery Framework for Assisted Healthcare" IEEE Transaction on cloud computing, vol. x, no. x, February 2015.

[3]. Jinggui Liao, Yuelong Zhao, and Saiqin Long,―MRPrePostA Parallel algorithm adapted for mining big data,‖ IEEE Workshop on Electronics,Computer and Applications, 2014

[4]. Aditi V. Jarsaniya, Shruti B. Yagnik. "A Literature Survey on Frequent Pattern Mining for Biological Sequence". © 2014 IJIRT | Volume 1 Issue 6 | ISSN: 2349-6002

[5]. Dr. Vijayalakshmi M N, S.Anupama Kumar, Kavyashree BN. 2014. ― Investigating Interesting Rules Using Association Mining for Educational Data‖, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 2, pp.268-271

[6]. Lingjuan Li , Min Zhang , "The Strategy of Mining Association Rule Based on Cloud Computing", 2011 IEEE.

[7].]T.R. Gopalakrishnan Nair, K.Lakshmi Madhuri , "Data Mining Using Hierarchical Virtual KMeans Approach Integrating Data Fragments In Cloud Computing Environment",2011 IEEE.

[8]. L. J. Li and M. Zhang, "The strategy of mining association rule based on cloud computing," in Proc. 2011 International Conference on Business Computing and Global Informatization.

[9]. F. Marozzo, D. Talia, and P. Trunfio, "A cloud framework for parameter sweeping data mining applications," in Proc. 2011 Third IEEE International Conference on Coud Computing Technology and Science.

[10]. Jiabin Deng, JuanLi Hu, Anthony Chak Ming LIU, Juebo Wu, "Research and Application of Cloud Storage",2010 IEEE.

[11]. Yang Lai , Shi ZhongZhi ," An Efficient Data Mining Framework on Hadoop using Java Persistence API" , 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

[12]. K. W. Lin, Y.-C. Luo, 2009, "A Fast Parallel Algorithm for Discovering Frequent Patterns", GRC '09. IEEE Int. Conf. on Granular Computing, pp. 398 – 403.
[13]. J. Zhou and K.-M. Yu, 2008, "Tidset-based Parallel FP-tree Algorithm for the Frequent Pattern Mining Problem on PC Clusters", Lecture Notes in Computer Science 5036, pp. 18- 28.
[14]. J. Zhou and K.-M. Yu, 2008, "Balanced Tidset-based Parallel FP-tree Algorithm for the Frequent Pattern Mining on Grid System", Fourth Int. Conf. on Semantics, Knowledge and Grid, pp. 103-108.
[15]. A. Javed, and A. Khokhar, 2004, "Frequent Pattern Mining on Message Passing Multiprocessor Systems", Distributed and Parallel Databases, vol. 16, pp. 321–334.
[16]. G. Grahne and J. Zhu, 2003, "Efficiently Using Prefix-trees in Mining Frequent Itemsets", In Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations.
[17]. J. Han, J. Pei, and Y. Yin, 2000, "Mining Frequent Patterns without Candidate Generation", In Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp.1-12
[18]. R. J. Bayardo, Jr., Brute-force mining of high-confidence classification rules. In Proceedings of the 3rd international conference on knowledge discovery and data mining (KDD'97), Newport Beach, California, USA.
[19]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, 1996, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231.
[20]. R. Agrawal, R. Srikant, Mining Sequential Patterns, in: Proc. of the 11th Int'l Conf. on Data Engineering, 1995, pp. 3-14.
[21]. R. Agrawal and R. Srikant. Quest Synthetic Data Generator. IBM Almaden Research Center, San Jose, California, http://www.almaden.ibm.com/cs/quest/syndata.html.
[22]. R. Agrawal, T. Imielinski*, and A. Swami, 1993, "Mining association rules between sets of items in large databases", In Proc. of the 1993 ACM-SIGMOD Int. Conf. on management of data (SIGMOD'93), p

Authors Profile

Mr C H Lin pursed Bachelor of Science and Master of Science from University of New York, USA in year 2009. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Telecommunication, University of New York, USA since 2012. He is a member of IEEE & IEEE computer society since 2013, a life member of the ISROSET since 2013 and ACM since 2011. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 5 years of teaching experience and 4 years of Research Experience.

