



International Journal of Engineering, Business and Enterprise Applications (IJEBA)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

Association Rule Mining for Distributed Database

Darshana Patel

Gujarat Technological University
Chandkheda, Ahmedabad, Gujarat
INDIA

Abstract: This paper describes details about distributed database and association rule mining algorithms. This paper also focusses on distributed database and, data mining. It also focusses on different types of mining and techniques of mining. Myriad algorithms are available in association rule mining and from that Apriori algorithm has been identified as best suitable algorithm for Distributed Database. Apriori algorithm has been implemented and the results has been generated.

Keywords: Association rule mining; data mining; apriori algorithm; distributed database

I. Introduction

There is an outburst of data from various devices across different geographical locations. The data generated from different heterogeneous sources account for the development of distributed database. Thus to identify important patterns, mining process is performed on this database and the results can be utilized for decision making to enhance the growth of business and industry. Different mining types and techniques are available and from that association rule mining identifies associativity among various data items. In this research paper, apriori algorithm is identified as suitable algorithm for distributed data mining.

II. Distributed Database

Distributed database means single logical database is spread across many locations, where this locations are connected by communication network. [1] The distributed database environment varies based on the communication performed among distributed sites, like different databases co-operate each other, or any master site is there which co-ordinates each other and work in partnership [1]. There are various factors influencing usage of distributed database like Data communications costs and reliability, Data sharing, Distribution and autonomy of business units, multiple application vendor environment, Database recovery, Satisfying both transaction and analytical processing.

III. Data Mining

Data Mining is termed as the process of discovering interesting patterns and knowledge from large amount of data, also known as knowledge discovery. Data mining provides many functionalities like, classification, refinement, mining common pattern, association and clustering. "Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general the task can be classified into two categories: descriptive and predictive. Descriptive mining task characterize properties of the data in a target data set. Predictive mining task perform induction on the current data in order to make predictions". [2] Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. [3] Data mining is successful with three key points. 1. Volume of data 2. Powerful processors and 3. Mining algorithms. With use of data mining many operations are possible like market segmentation, customer churn, fraud detection, direct marketing, interactive marketing, market basket analysis and trend analysis. [4].

A. Types of Data Mining

On the basis of data availability and the access way there are different types of data mining.

Stream mining: "Stream data mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities". [4]

As stream data size is too high so it is not possible to store the complete data at a time in primary memory and new mining algorithms or techniques are required for such kind of data. As data is in huge amount random sampling, sliding windows, histogram, randomized algorithm like methods are used. [5]

Web mining: web mining is another popular approach of data mining which generates important information from web pages. Though there are certain differences between web mining and ordinary data mining. In web mining web pages are generally public so, the information can be fetched. [6]

Multi database data mining: Multi database data mining is one of the popular approach of data mining. Web provides countless information which can be mined and used for any organization. This information can play major role for any organization for decision making. The data of single organization is called internal data and the web data is called external data. External data mining provides very efficient knowledge but it requires multiple database data mining. [7]

Distributed Data Mining: Distributed Data Mining (DDM) is a branch of data mining that offers a framework to mine distributed data closely monitoring the distributed data and computing resources. Distributed data mining originated from the need of mining over decentralized data sources regardless of their physical locations. Distributed data mining is required in two cases, 1) The data is distributed on different locations and mining process is performed and 2) Data mining, sometimes needs lots of resources to perform mining process and to distribute the load, distributed mining process is required. The mining process is carried out at the local sites and final results are aggregated at the global site.

B. Techniques of Data Mining

Information technology has been developed a lot and due to that a huge amount of data available in business, science and many various areas. This large amount of data can be used for decision making and knowledge generation. [10] Data mining is a process of extraction of useful information and pattern identification from existing database. [11] The major techniques of data mining are classification, clustering and association rule mining.

Classification: Classification is used maximum as data mining technique. In classification defined model is available and on the basis of the model the mining is performed so classification is called supervised learning process.

Clustering: Clustering technique provides the facility to group the records to gather. In clustering technique certain classes are created and related data are gathered in individual class.

Association rule mining: Association rule mining is also very popular data mining technique. In association rule mining the pattern is identified based on relationship between different transactional data and so association technique is also called relational technique. [12]

Association rule mining is used to help business to take certain kind of decisions. There are different types of association rules like multi-level, multi-dimensional etc. In this research paper focuses on association rule mining algorithm.

C. Association rule mining and distributed database

Association Rule Mining (ARM) is one of the most important data mining techniques which find associativity from large set of data items. ARM specially focuses on distributed architecture, which will require integration of knowledge generated from multiple data sites [13]. This mining technique generates rules on the basis of minimum support count. In Association Rule Mining, Association rule express an association between set of data items.

D. Algorithms for Association Rule Mining

There have been many algorithms developed for mining frequent patterns which can be classified into two categories 1) candidate generation with Apriori and its variants as the most popular, and 2) Pattern growth methods with F-P Growth method and its variants.

F-P Growth use two data structures F-P Tree and header table. F-P Growth algorithm is very fast but it requires significant storage space as F-P Trees are repeatedly generated [14]. F-P tree is expensive to build as it requires high amount of memory. According this algorithm first, F-P tree is generated to represent frequent items after this first step it divides the compressed database into set of conditional database and each one associated with one frequent pattern. In last each such database is mined separately. The problem with FP growth is, when database is large FP tree cannot be hold in main memory and to overcome the problem the database must be partitioned into set of smaller database.

E. Apriori Algorithm

The Apriori algorithm iteratively identifies frequent item sets in data by applying the downward closure property of item sets in the generation of candidate item sets where a candidate item set is confirmed as frequent only when all its subsets are identified as frequent in the previous pass. Apriori algorithm is most popular algorithm of Distributed Data Mining [15]. Minimum support count is important in apriori algorithm, which is entered by user. Apriori algorithm identifies item lists that are higher than minimum support count. If minimum support count is less than more itemsets are identified and if it is high then less pair of items are identified.

Typically, association rules are discarded as uninteresting if they do not satisfy a minimum support threshold. Additional analysis can be performed to uncover interesting statistical correlations between associated attribute-

value pairs. A data mining system has the potential to generate thousands or even millions of patterns or rules. A pattern is also interesting, if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge.

Apriori algorithm is association rule mining algorithm proposed by R. Agrawal and R. Srikant in 1994. This algorithm identifies frequency of pair of items in database. Bottom up approach is applied by Apriori algorithm. In Apriori algorithm, during first mining complete transactions are scanned and for individual items support count is identified. Then group of candidates are generated and tested again data. This algorithm is terminated when no further groups can be generated.

The steps of Apriori algorithms are described here:

1. In the first step of the algorithm it will scan all the transactions and generate candidate set C1. It counts occurrence of all the items. In the first iteration all the items are member of the set of candidate 1-itemset.
2. L1 is generated from C1. L1 contains that items only, whose support count is greater than minimum support. L1 consist of candidate 1 itemset from C1 with higher support count then minimum.
3. To generate L2, the join operation of L1 is used. The join of L1 $\bowtie$ L1 is used to generate a candidate set of 2-itemsets C2. C2 is used to identify frequency of 2 itemsets. From C2, no candidates are removed.
4. In the next step the complete transactions in D are scanned to identify the frequency of itemsets available in C2.
5. L2, is generated from C2 where L2 is set of frequent 2-itemsets. The L2 contains items from C2 that are having support count more than minimum support count.
6. Candidate set of 3 itemsets is generated, from L2 $\bowtie$ L2. The frequency will be identified through scanning of transaction set D. Here property of Apriori algorithm "All subsets of frequent itemsets must also be frequent" [2] is followed and non-frequent itemsets are removed.
7. L3 is generated through C3. L3 contains that items only whose support count is higher than minimum support count.
8. The algorithm is further used to generate candidate set of 4 itemsets and so on. This steps are repeat until further candidate sets cannot be generated. The algorithm terminates after generating all frequent itemsets.

The above steps describes Apriori algorithm [2].

Strong association rules can be generated if the frequent itemsets from transactions in database D have been identified. The association rules satisfy both, minimum support and minimum confidence.

IV. Implementation Results

This dataset has been created to distribute data generated for the NCBI Influenza Virus Sequence Database. The source of the dataset is ftp://ftp.ncbi.nih.gov/genomes/_INFLUENZA/. The NCBI Influenza Virus Sequence Database contains nucleotide sequences, protein sequences and their encoding regions of all influenza viruses in GenBank, including the complete genome sequences from the NIAID Influenza Genome Sequencing Project genomeset.dat - Table with supplementary genomeset data. The genomeset.dat contains information for sequences of viruses with a complete set of segments in full-length (or nearly full-length). Those of the same virus are grouped together and separated by an empty line from those of other viruses. The genomeset.dat file is tab-delimited tables which have the following fields: GenBank accession number Host Genome segment number or protein name Subtype Country Year/month/date Sequence length Virus name Age Gender. The following table describes the results of Apriori algorithm for one lakh, fifty thousand and twenty five thousand records of Genome dataset. The time in milliseconds is calculated for different support counts from 1000 to 100 for both the algorithms for the first mining process. It indicates that as the support count reduces the time taken by Apriori algorithm decreases.

S. Count	Apriori (ms)
1000	25642
500	77956
250	196707
200	267628
150	341380
100	582196

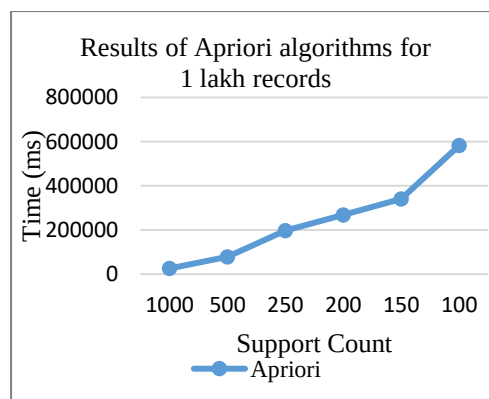


Figure 1: Results of Apriori Algorithm

V. References

- [1] Jeffery A.hoffer, Mery Prescott and Heikki Topi, "Modern Database Management System" 12th edition, Chapter 13 Distributed Database, Prentice Hall publication
- [2] Jaiwei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques", Elsevier, Third Edition, Page 21: chapter 1.4 What kind of Patterns can be mined?, Second Edition.-2006
- [3] "An Introduction to Data Mining", Thearling.com, 2017. [Online]. Available: <http://www.thearling.com/text/dmwhite/dmwhite.htm>. [Accessed: 10- June- 2016].
- [4] "Data Mining", Laits.utexas.edu, 2017. [Online]. Available: <http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/>. [Accessed: 11- July- 2017].
- [5] Twinkle B. Ankleshwari, J.S.Dhobi, "Mining Data Streams: A survey", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 2, February -2014
- [6] Surbhi Singh, Devyani Gupta, Akash Chandhoke, "Web Mining: Taxonomy and Survey", Advances in Computer Science and Information Technology, Volume 1, Number 2, November 2014, pp. 56-60
- [7] Shichao Zhang, Xindong Wu, Chengqi Zhang, "Multi-Database Mining", IEEE Computational Intelligence Bulletin, June 2003, Vol 2, No.1
- [8] Fuggetta, Alfonso Gian Pietro Picco and Giovanni Vigna, Understanding code mobility, IEEE Transactions on Software Engineering Vol 24 No 5, 1998
- [9] Ichiro Satoh, "Mobile Agents", National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
- [10] Yongjain Fu, "Data Mining: Tasks, Techniques and Applications", University of Missouri, Rolla, [Online]. Available: <http://academic.csuohio.edu/fuy/Pub/pot97.pdf>. [Accessed: 10- June- 2016].
- [11] Bharati M. Ramageri, Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305 [Online]. Available: <http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf> [Accessed: 10- June- 2016].
- [12] Martin Brown, Data mining techniques, IBM Developer Networks, December 11, 2012 [Online]. Available: <https://www.ibm.com/developerworks/library/ba-data-mining-techniques/> [Accessed: 14- June- 2016]. H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representatives," Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670. (Article in a conference proceedings)
- [13] O. Folorunso, A. S.Sodiya, G. O. Ogunleye and A. O. Ogunde, "A Review of Some Issues and Challenges in Current Agent Based Distributed Association Rule Mining", Medwell Journals, vol. 10, no. 2, pp.84-95, 2011
- [14] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach", Data Mining and Knowledge Discovery, 8 (2004), pp. 53-87
- [15] Mohammed Al-Maolegi and B. Arkok, "An improved Apriori Algorithm for Association Rules", IJNL C February 2014

VI. Acknowledgments

We would like to express our very great appreciation to significant advisors and contributors for their valuable and constructive suggestions during the planning and development of this research work. Their willingness to give time so generously has been very much appreciated. The authors would like to thank them for their helpful comments to improve this paper.