6th International Conference on Recent Development in Engineering Science, Humanities and Management
National Institute of Technical Teachers Training & Research, Chandigarh, India (ESHM-17)
14th May 2017, www.conferenceworld.in ISBN: 978-93-86171-36-8

A COMPARATIVE STUDY OF CONVENTIONAL DATA MINING ALGORITHMS AGAINST MAP-REDUCE ALGORITHM

Bhawna[1],Dr. Gagandeep Jagdev[2]

[1]Research Scholar, M.Tech. (CE), Yadavindra College of Engineering, Talwandi Sabo (PB).

[2]Faculty, Dept. of Comp. Science, Punjabi University Guru Kashi College, Damdama Sahib (PB).

ABSTRACT

Data Mining has always remained a prominent aspect of filtering small or huge databases. It is actually data which has gone through many changes in terms of volume, variety and veracity. The past three years have witnessed enormous growth in the size of data. The major sources responsible for thisare social networking, RFID, sensors generated data, banking transactions, hospitals, educational institutes, major organizations in retail sector, stock market and many more. Gone are the days when data was used to be measured in kilobytes, megabytes, gigabytes, terabytes or petabytes. Today new units have evolved for measuring data – exabytes, zettabytes, yottabytes. The traditional data mining algorithms were capable of handling limited data. But data which grows beyond limit is something more demanding and is referred as Big Data. The popular technique, originally proposed by google, capable of handling Big Data is Map-Reduce. This research paper is primarily concerned with studying different conventional data mining algorithms and comparing Apriori Algorithm with Map-Reduce algorithm. The paper also focuses on elaborating Big Data and its characteristics.

Index Terms – Apriori algorithm, Big Data,Data mining, Map-Reduce algorithm.

I. INTRODUCTION

Data mining is all about processing data and identifying patterns in the information to reach to any specific decision. The principles of data mining have become more predominant with the advent of Big Data. It is not possible to obtain relatively simple and easy statistics out of Big Data using conventional data mining techniques.The process of data analysis, discovery, and model-building is often iterative. One must also comprehend how to relate, map, associate, and cluster it with other data to produce fruitful result. The flowchart in Fig.1 show the entire procedure involved in data mining [13

(ESHM-17)
ISBN: 978-93-86171-36-8

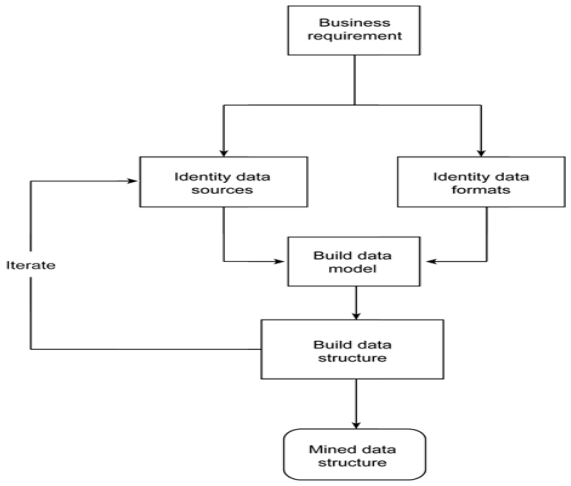

Fig. 1 Figure illustrates flowchart depicting the Data Mining process

Different techniques involved in data mining process are mentioned as under.

- Association - Association is the most popular among available data mining techniques. Based on relationship between items in the same transaction, a unique pattern. Association technique is frequently used in retail sector to identify what customers frequently purchase together.
- Classification - Classification is based on machine learning. Classification classifies each item in a set of data into one of the predefined set of groups. It makes use of mathematical techniques like neural networks, statistics, decision trees and linear programming.
- Clustering - Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.
- Prediction - The prediction, as its name implied, is one of a data mining techniques that discovers the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.
- Sequential Patterns - Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.In sales, with

ISBN: 978-93-86171-36-8

historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

- Decision trees - The A decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

II. DIFFERENT ALGORITHMS CONCERNED WITH DATA MINING

Conventional Data Mining Algorithm

Some of the popular conventional algorithms involved in data mining process are mentioned as under.

- C4.5
- k-means
- Support vector machines
- Apriori

A. C4.5 - C4.5 constructs a classifier in the form of a decision tree. In order to do this, C4.5 is given a set of data representing things that are already classified. A classifier is a tool in data mining that takes a bunch of data representing things we want to classify and attempts to predict which class the new data belongs to. Arguably, the best-selling point of decision trees is their ease of interpretation and explanation. They are also quite fast, quite popular and the output is human readable.

B. k-means - k-means creates *k* groups from a set of objects so that the members of a group are more similar (Fig. 2). It's a popular cluster analysis technique for exploring a dataset. Cluster analysis is a family of algorithms designed to form groups such that the group members are more similar versus non-group members. Clusters and groups are synonymous in the world of cluster analysis.

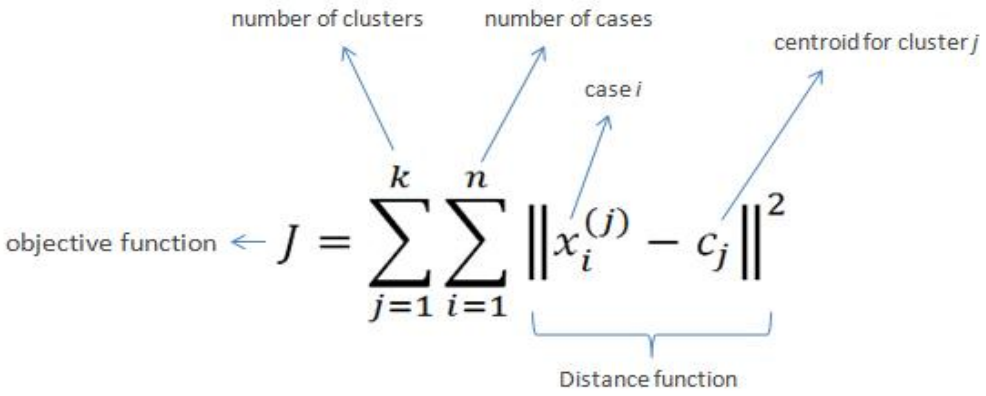

Fig. 2 Formula for K-Means clustering

ISBN: 978-93-86171-36-8

Algorithm for k-means

1. Clusters the data into *k* groups where *k* is predefined.
2. Select *k* points at random as cluster centers.
3. Assign objects to their closest cluster center according to the *Euclidean distance* function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

C. Support Vector Machine (SVM) - Support vector machine (SVM) learns a hyperplane to classify data into 2 classes. At a high-level, SVM performs a similar task like C4.5 except SVM doesn't use decision trees at all. A hyperplane is a function like the equation for a line, y = mx + b. In fact, for a simple classification task with just 2 features, the hyperplane can be a line.SVM can perform a trick to project your data into higher dimensions. Once projected into higher dimension, SVM figures out the best hyperplane which separates your data into the 2 classes. For example - Consider a bunch of red and blue balls on a table. If the balls aren't too mixed together, you could take a stick and without moving the balls, separate them with the stick.When a new ball is added on the table, by knowing which side of the stick the ball is on, you can predict its color. The balls represent data points, and the red and blue color represent 2 classes. The stick represents the simplest hyperplane which is a line.

D. K - Nearest Neighbor (kNN) - KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. Consider the example mentioned as under to place KNN in the scale. Fig. 3 shows the spread of red circles (RC) and green squares (GS):

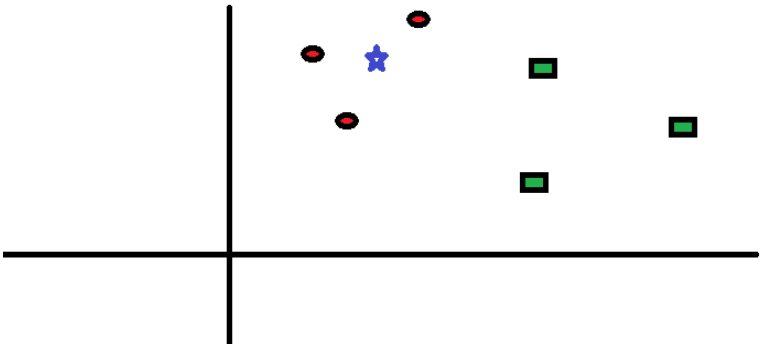

Fig. 3 Figure depicts the spread of red circles and green squares

We intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else. The "K" is KNN algorithm is the nearest neighbors we wish to take vote from. Let's say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three data points on the plane. Refer to Fig. 4 below.

ISBN: 978-93-86171-36-8

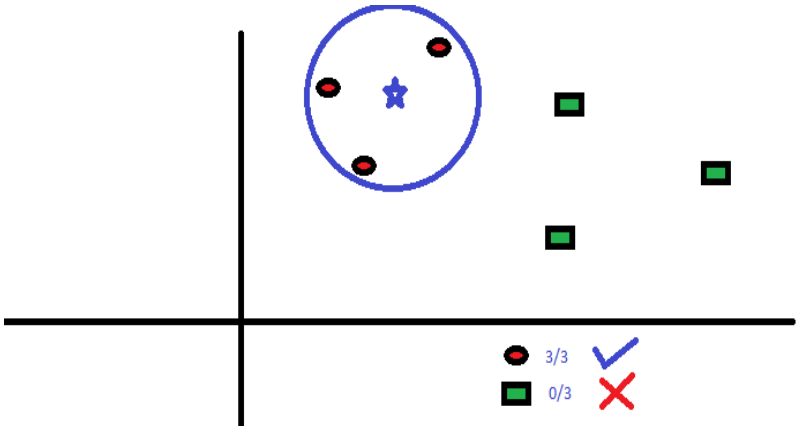

Fig. 4 Figure depicts three similar data points enclosed on the plane.

The three closest points to BS is all RC. Hence, with good confidence level we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm. Next, we will understand what are the factors to be considered to conclude the best K.

E. Apriori Algorithm - It is a classic algorithm used in data mining for learning association rules. It is nowhere as complex as it sounds, on the contrary it is very simple. Apriori [2] is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database.Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Recently, researchers are applying the association rules to a wide variety of application domains such as Relational Databases, Data Warehouses, Transactional Databases, and Advanced Database Systems like Object-Relational, Spatial and Temporal, Time-Series, Multimedia, Text, Heterogeneous, Legacy, Distributed, and web data [1]. Since data generated day by day activities, the volume of data is increasing dramatically. Massive amount of data is available in the data warehouses. Therefore, mining association rules helps in many business decision making processes. Some examples are cross-marketing, Basket data analysis and promotion assortment etc. In the area of association rules mining, a lot of studies have been done [1, 2].

The problem of discovering association rules is decomposed into two stages:

- Discovering all frequent patterns represented by large item sets in the database, and generating the association rules from those frequent item sets.
- The second sub problem is a straightforward problem, and can be managed in polynomial time. On the other hand, the first task is difficult especially for large databases.

ISBN: 978-93-86171-36-8

The Apriori is the first efficient algorithm for solving the association rule mining, and many of the forthcoming algorithms are based on this algorithm. Confidence denotes the strength of implication and support indicates the frequencies of the occurring patterns in the rule. It is often desirable to pay attention to only that rule which may have reasonably large support. Such rules with high confidence and strong support are referred to as strong rules. The prime objective of mining association rules is to discover strong association rules in large databases.

Fig. 5 illustrates an example for understanding the working of Apriori algorithm. The four item sets under study are {1 3 4}, {2 3 5}, {1 2 3 5}, {2 5}. The golden rule set for example is to only consider the items having 2 or more than 2 occurrences. The highlighted rows in tables indicates the items to be rejected.

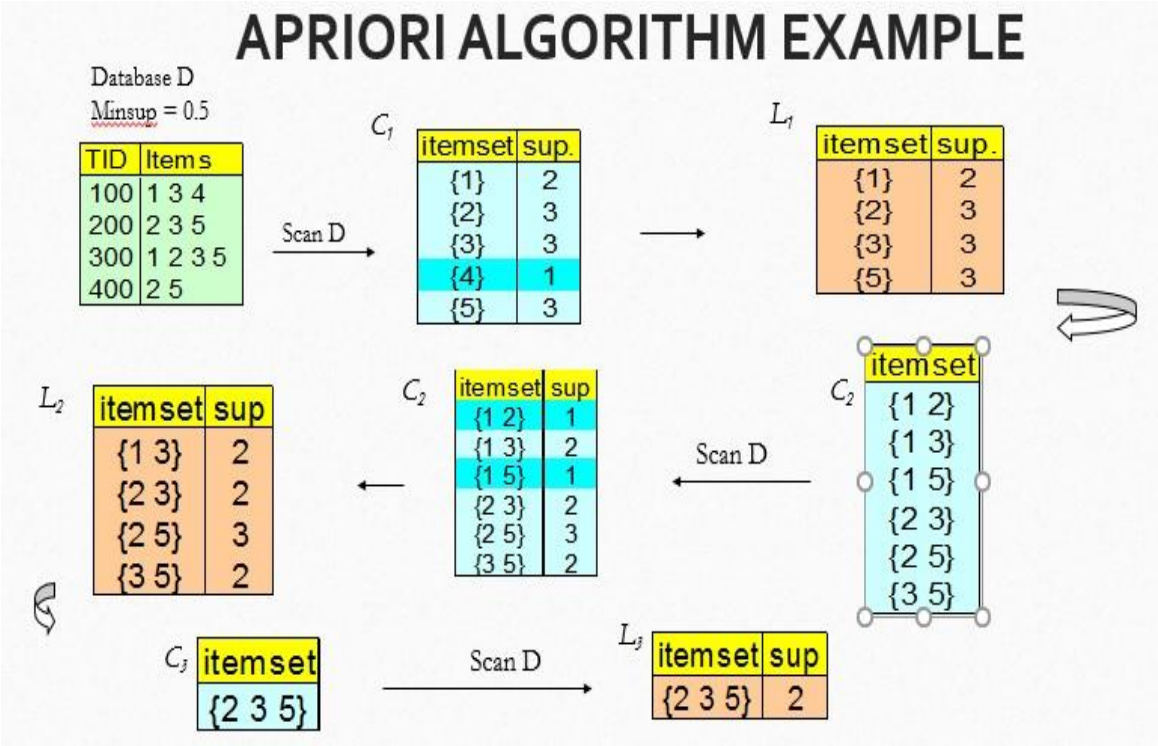

Fig. 5 Figure depicting an example of working of Apriori algorithm

III. MAP-REDUCE ALGORITHM FOR BIG DATA

Hadoop is one such java framework which is capable of handling large data sets in a distributed computing environment. Apache Software Foundation [8, 9] is the authority sponsoring Hadoop. Applications run on systems constructed of thousands of nodes involved in processing thousands of Terabytes of data. Because of distributed nature, Hadoop facilitates very fast data transfer and system continues to operate even if any node failure occurs. This notion minimizes the risk of devastating system failure even when multiple nodes become non-operational. The creator of this technology was Google. It was developed by them during their early days to index all valuable textural and structured information. The primary motive behind all this was to provide

ISBN: 978-93-86171-36-8

meaningful results to the users. Hadoop finds its application in several sectors which comprises of retail, sports, medical science, business, education and of course now in elections.

Map-Reduce algorithm [10, 11] works behind Hadoop and can be written in any language [5, 12]. Fig. 6 depicts the working of Map-Reduce technology.

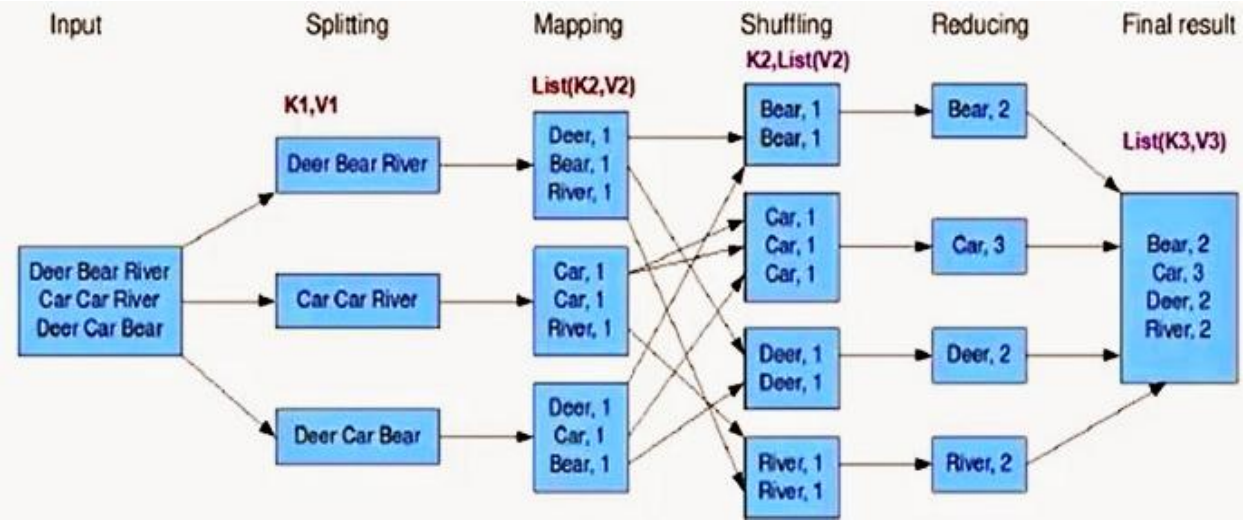

Fig. 6 Flowchart showing working of Map-Reduce technology

The input file comprises of three statements, each having three words in it. Firstly, the input is distributed in three different <key, value> pairs. Thereafter mapping is conducted in which the number of occurrences of each word is written against it. During shuffling, pairs having similar keys are gathered. Then in reduction, the values of similar keys are summed up and placed against the key. Finally, all obtained <key, value> pairs are accumulated [13].

Algorithm for Map-Reduce

- The incoming data can be alienated into n number of modules which depends upon the amount of input data and processing power of the individual unit.
- All these fragmented modules are then passed over to mapper function where these modules undergo simultaneous parallel processing.
- Thereafter, shuffling is conducted in order to gather similar looking patterns.
- Finally, reducer function is called which is responsible for getting the ultimate output in a reduced form.
- Moreover, this technique is scalable and depending upon increase in the data to be processed, the processing units can be further extended.

IV. MAP-REDUCE VS. APRIORI ALGORITHM

The research work is based on mining a huge database constructed and to use best available platform or algorithm or technology for this purpose. Among many prominent data mining algorithms already existing,

(ESHM-17)
ISBN: 978-93-86171-36-8

Apriori algorithm [3, 4, 6, 7] is regarded as the best whenever there's a need to perform mapping or mining database in combinations. So, the target is to compare Apriori algorithm with the Map-Reduce algorithm related to Hadoop platform. For this, a small excel file is taken as input comprising of few numbers. The snapshot of the input file is shown in Fig. 7.

	A	B	C	D	E	F
1	**A1**	**A2**	**A3**	**A4**	**A5**	**Sum**
2	1	5	2	0	0	8
3	2	3	0	1	0	6
4	3	4	0	0	0	7
5	2	1	3	0	0	6
6	1	2	3	0	0	6

Fig. 7 Input data file

A source code is constructed in C++ for implementing working of Apriori algorithm and the above-mentioned data in Fig. 6 is given as input.

On executing the source code of Apriori algorithm, the following output was obtained.

1 2 3 3

The result shows that 1, 2 and 3 appears 3 times together in the input file and it took 17.967033 seconds. The snapshots shown in Fig. 8 shows time consumed as the output obtained on implementing Apriori algorithm.

Fig. 8 Output obtained after executing Apriori algorithm source code

When the same input was given to Hadoop Framework, the result was obtained in only 1.2 Seconds. The comparative graph of Apriori and Hadoop is shown below in Fig. 9.

ISBN: 978-93-86171-36-8

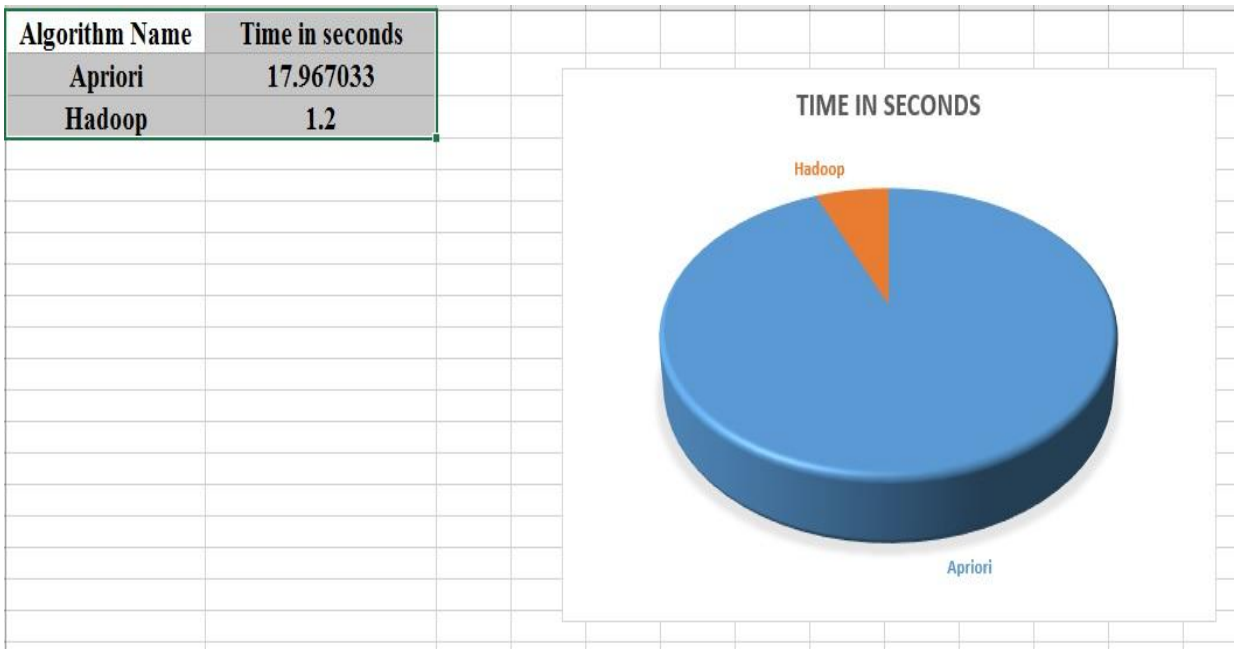

Fig. 9 Comparison of Apriori and Map-Reduce algorithm in terms of time in seconds.

This result shows that Map-Reduce algorithm is much speedier and efficient in mining as compared to Apriori algorithm.

V. CONCLUSION AND FUTURE SCOPE

Most of the researchers have considered association rule mining problems as single objective problem and validated on static database but it is a multi-objective problem because it uses measures like support count, comprehensibility and interestingness for mining the strong association rule. Since the database is being updated periodically due to daily business requirement. Incremental mining deals with generating association rules from updated database.Most of the existing algorithms for incremental mining rescan the entire database again. Cost ofscanning large database is high. The association rules generated on static database are not meaningful, effective and not appropriate for making business strategies and decisions. Hence,it requires to devise a new and efficient algorithm for incremental mining without rescanning of database. Therefore, there is a need to shift the paradigm form single objective to multi-objective association rule mining and also requires consideration of incremental data. Data mining is a new and significant area of research, and soft computing tools itself are extremely appropriate to solve the problems. Soft computing characteristics include high robustness, parallel processing, self-organizing adaptive, high degree of fault tolerance distributed storage etc. are much suitable for data mining applications. It also obtains a greater attention in Artificial Neural Networks, whichoffer qualitative methods for business and economic systems [1, 3].

ISBN: 978-93-86171-36-8

REFERENCES

[1] Hemant Kumar Soni et al., "ASSOCIATION RULE MINING: A DATA PROFILING AND PROSPECTIVE APPROACH", International Conference on Futuristic Trends in Engineering, Science, Humanities, and Technology (FTESHT-16) ISBN: 978-93-85225-55-0, January 23-24, 2016, Gwalior.

[2] Gianni D'Angelo, Salvatore Rampone et al., "Developing a Trust Model for Pervasive Computing Based on Apriori Association Rules Learning and Bayesian Classification".

[3] JayshreeJha and LeenaRagha, "Educational Data Mining using Improved Apriori Algorithm", International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 5 (2013), pp. 411-418.

[4] Jiao Yabing, "Research of an Improved Apriori Algorithm in Data Mining Association Rules", International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.

[5] Sheila A. Abaya, "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation", International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012 1 ISSN 2229-5518.

[6] Shweta et al., "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms", IJARCSSE, Volume 3, Issue 6, June 2013 ISSN: 2277 128X.

[7] Dean, Jeffery, and Ghemawat Sanjay. 2004. "MapReduce: Simplified Data Processing on Large Clusters." Google.

[8] Dr. Gagandeep Jagdev et. al., "Comparing Conventional Data Mining Algorithms with Hadoop based Map-Reduce Algorithm considering elections perspective", in International Journal of Innovative Research in Science and Engineering (IJIRSE), ISSN: 2454-9665 (O), ISSN: 2455-0663(P), Volume – 3, Issue – 3, March 2017.

[9] Gagandeep Jagdev et. al., "Association of Big Data with Map-Reduce Technology Augments for Economic Growth in Retail", in International Journal of Engineering Technology Science and Research (IJETSR), ISSN: 2394 - 3386, Volume 4, Issue 2, February 2017.

[10] Dr. Gagandeep Jagdev et. al. "Implementation of Big Data concerned with Elections using Map-Reduce as novel mining algorithm" has been accepted for publication at International Conference ICCCN-2017 at NITTTR, Chandigarh on 23rd-24th March 2017.

[11] Dr. Gagandeep Jagdev et. al., "Boosting Revenue growth in retail sector via mining Big Data to utmost potential", at 24th International Conference on Finite or Infinite Dimensional Complex Analysis and Applications (24-ICFIDCAA-2016) held at Anand International College of Engineering, Jaipur, Rajasthan sponsored by IEEE, SIAM, ISAAC, INSERB from 22nd -26th August, 2016.

[12] Dr. Gagandeep Jagdev et. al., "Implementation of Map-Reduce Algorithm for Mining Big Data concerned with Indian Election Scenario", at 24th International Conference on Finite or Infinite Dimensional Complex

ISBN: 978-93-86171-36-8

Analysis and Applications (24-ICFIDCAA-2016) held at Anand International College of Engineering, Jaipur, Rajasthan sponsored by IEEE, SIAM, ISAAC, INSERB from 22nd -26th August, 2016.

[13] Dr. Gagandeep Jagdev et. al., "Analyzing and Mining Big Data In Health Care Sector For Healthy Human Survival Via Competent Technologies", at 1st International Conference on Technology Management held at National Institute of Management, Hamirpur on July 14-15, 2016.

ABOUT THE AUTHOR

Dr. Gagandeep Jagdev is working in the capacity of faculty member in Dept. of Computer Science, Punjabi University Campus, Damdama Sahib (PB). His total teaching experience is above 10 years and has above 105 international and national publications in reputed journals and conferences to his credit. He is also a member of editorial board and advisory board of several reputed international peer reviewed journals and Technical Committee Member of many reputed universities. His areas of interest are Big Data, Data Mining, RFID, Biometrics, Cloud Computing.