A TECHNIQUE OF MINING DATA WITH LARGE DATASET-SURVEY

Surya.k

P.G scholar, department of CSE,
Karpagam academy of higher education, Coimbatore-21,
Tamil Nadu, India.

Abstract— **Big data and data mining are the vast emerging technique used for mining or extracting the data from large dataset. Generally data mining refers to extracting the knowledge or data from the dataset, where the data may be relevant or irrelevant, in case of searching a particular data from the large records. we are using data mining domain to extract a particular important hidden. Whereas Big data refers to large data where it contains structure or unstructured data .The main motive of this survey deals with, how the mining process deals with large dataset. Many authors were in process of doing a research dealing with mining a data or knowledge from large data or records. one such survey is here with a technique of mining a data with large data set.**

Index Terms—**About four key words or phrases in alphabetical order, separated by commas.**

INTRODUCTION:

Data mining is different from database and the evaluation of data mining is from data collect in which data collection can be done and are accessed by data access and next come data warehouse and decision support and finally come data mining where the extraction process takes place. Here this paper deals with the technique of mining data with large data set. Consider the database in which data are present and the data mining deals with solving problems by analyzing data present in the database. Generally Big data deals with large dataset with large size and complexity. Consider lot of data to be stored in the specific database compared to google, so in order to reduce the spare of data, we have many technologies, one such technology is Map Reduce technology.

KNOWLEDGE ABOUT DATA MINING AND BIGDATA:

Data mining:

Major elements of load transition data and can also have the capacity to store and manage the data and also providing data access to all aspects such as business intelligence because now a days there is a necessary to store and process large amount of structure and unstructured data based on cluster classification technique. And next well suited for web search that is web is the extremely a popular medium for one who wishes to publish and share their own business or useful information in which user at the other side make use of these data and this process of getting a related things with key words is done through data mining and next is the bio informatics, finance, digital libraries and digital governments etc..,

Big data:

Big data deals with large growing complex data set, is now recently growing or expanding in all

domain expect, especially in bio informatics because now a days that is day by day we can found a genomic information getting expanding and there is need of storing a large data dealing with sensitive and non sensitive information. So here the big data plays the major role and next widely field using big data is digital government and digital library and mainly in medical field. Consider the example face book which is a famous social media in which we can share our own ideas and send or share a picture or videos. Generally speaking it is a picture or status sharing site which approximately receiving more than million photos per day. Focusing on size of a single picture it may ranges approximately from 1 MB then on considering the whole there may leads to million or trillion space thus results for the large space for the storage, thus there is the rise for the big data.

On doing the mining process with large data the first thing we can save the time and for easy access where the records are large. One such method is clearly explained by flowchart below which shows how the system gets the data as input and how clustering and classification done and how the fitness is calculated and how memory is evaluated is shown clearly.

ROLE OF MAP REDUCE:

Instead of doing a work in serial that everyone knows that if we done in parallel for then the work will be done sooner. The same concept is also available for one such model is Map Reduce which are used for the analysis and mining step.

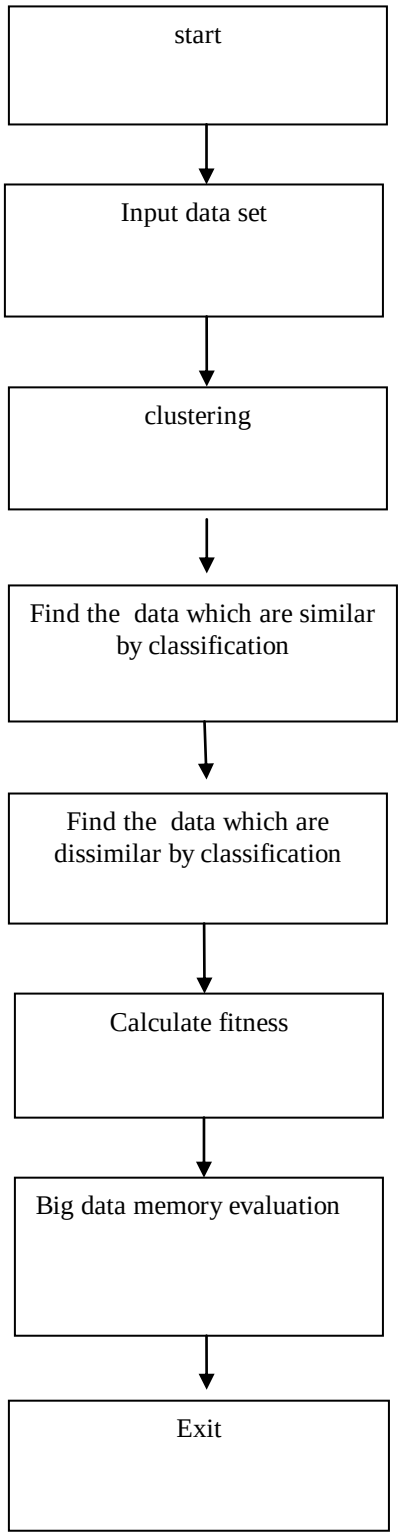
Map Reduce Programming generally deals with the two steps the first task is "Map" and second task is "Reduce" by providing key values pairs.

MODULE DESCRIPTION:

Consider there are totally five modules. There all about how the data collected and how the clustering and classifying is done with the large data set and finally how the result is obtained.

DATA FLOW DIAGRAM:

Data mining with big data:



Inputting the data set:

The initial step is to mine the large data set is to providing the large input to the database otherwise the data are collected or gathered and inputting relevant data set

clustering:

After inputting the data Set, the next step is to cluster the data usually the data are clustered based on similarities This clustering can also be done for both homogeneous clustering and heterogeneous clustering

Heterogeneous classification

Classifying technique usually takes place,when there are large dataset. Hetero meaning different. In this step, the data which are different are classified as a single classification phase.

Homogeneous classification:

Homo meaning same, in this step the data which are similar are grouped forming different classification.

Results and reports:

Results and reports are the final step or phase which the Big Data Module evaluation and fitness classified is done here.

CONCLUSION:

This survey deals with and mainly focus on how the mining process could takes place when the dataset size is large. It could be very helpful in business intelligence, web search,bio informatics, health informatics finance and digital governments etc.., now a days mining with large dataset is necessary in all aspects.

REFERNCE:

[1] J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman, "Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center," Proc. 11th USENIX Conf. File and Storage Technologies (FAST '13), 2013.

[2] D. Luo, C. Ding, and H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining," Proc. IEEE 12th Int'l Conf. Data Mining, pp. 489-498, 2012.

[3] T. Mitchell, "Mining our Reality," Science, vol. 326, pp. 1644-1645,2009.

[4] S. Papadimitriou and J. Sun, "Disco: Distributed Co-Clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-to- End Mining," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08),pp. 512-521, 2008.

[5] A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press, 2011.

[6] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel
Classifier for Data Mining," Proc. 22nd VLDB Conf., 1996.

[7] A. da Silva, R. Chiky, and G. He´brail, "A Clustering Approach for Sampling Data Streams in Sensor Networks," Knowledge and Information Systems, vol. 32, no. 1, pp. 1-23, July 2012.

[8] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09),pp. 296-301, 2009.

[9] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems, vol. 30, no. 1, pp. 71-88, 2005.

