IMPLEMENTING HACE THEOREM FOR BIG DATA PROCESSING – REVIEW

Prema Gadling*

Prof. Mahip Bartere**

Prof. Jayant Mehare

ABSTRACT— The aim is propose to elaborate a HACE theorem that states the characteristics of the Big Data revolution, and proposes a Big Data processing model from the data mining view. Here, Data comes from everywhere like sensors, media sites and social media etc. In this useful data can be extracted from this big data using data mining technique for discovering interesting patterns. As enhancement we propose Detection of emerging topics from social networks of big data. Specifically, we focus on mentions of user links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and retweets. In this paper, we are going to talk how effectively analysis is done on the data which is collected from the Twitter using Flume. Twitter is an online web application which contains rich amount of data that can be a structured, semi-structured and un-structured data. We can collect the data from the twitter by using BIGDATA eco-system using online streaming tool Flume. And doing analysis on Twitter is also difficult due to language that is used for comments. And, coming to analysis there are different types of analysis that can be done on the collected data. And this paper provides a way of analysis data using hadoop which will process the huge amount of data on a hadoop cluster faster in real time.

Keywords - Big Data, data mining, HACE, Spectral Clustering, Hadoop.

* **M.E (CSE) 3nd Semester, GHRCEM, Amravati University**

** **Department of computer science, GHRCEM, Amravati University**

Introduction

Data Mining is the technology to extract the knowledge from the data. It is used to explore and analyze the same. The data to be mined varies from a small data set to a large data set i.e. big data. Data Mining has also been termed as data dredging, data archeology, information discovery or information harvesting depending upon the area where it is being used. The data Mining environment produces a large volume of the data. The information retrieved in the data Mining step is transformed into the structure that is easily understood by its user.[8]

Big Data are the large amount of data being processed by the Data Mining environment. In other words, it is the collection of data sets large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications, so data mining tools were used. Big Data are about turning unstructured, invaluable, imperfect, complex data into usable information. [1][2]

Data have hidden information in them and to extract this new information; interrelationship among the data has to be achieved. Information may be retrieved from a hidden or a complex data set. It seeks to explore more complex and evolving relationships among data. Huge data with heterogeneous and diverse dimensionality has autonomous sources with distributed and decentralized control in complex and evolving relationships. So the noise is one of the problems. In this paper, we group the objects by using clustering which is similar to one cluster and is dissimilar group in to other cluster. To overcome this problem, we use spectral algorithm by implementing using java programming language.

HACE Theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

The characteristics of HACE make it an extreme challenge for discovering useful knowledge from the Big Data. The HACE theorem suggests that the key characteristics of the Big Data are

1. Huge with heterogeneous and diverse data sources:-

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This huge volume of data comes from various sites like Twitter, MySpace, Orkut and LinkedIn etc.

2. Decentralized control:-

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers.

3. Complex data and knowledge associations:-

Multistructure, multisource data is complex data, Examples of complex data types are bills of materials, word processing documents, maps, time-series, images and video. Such combined characteristics suggest that Big Data require a ―big mind to consolidate data for maximum values.[9]

The following paper reviews different aspects of the big data. The paper has been described as follows, in section I Introduction about data mining, big data and HACE Theorem. In section II we discuss the works related to data mining with big data. In section III the proposed system is discussed. In section IV describes system architecture of big data. In section V discuss the system implementation.

I. Related Work

Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE presents a HACE theorem that describes the features of the Big Data revolution, and suggests a Big Data processing model, from the data mining perspective. [2] The demand-driven, mining and analysis, user interest modeling, and security and privacy considerations are involved in the demand-driven aggregation of data sources.

Yanfeng Zhang et al [3], proposed a novel incremental method which is an extension of MapReduce known as I2MapReduce which reduces I/O overhead for accessing preserved fine-grain computation states. It performs key-value pair level incremental processing rather than task level re-computation which makes the novel technique improve performance and give optimized result for mining the data. It merges K-means algorithm with MapReduce for better results.

Jie Xu et al [4], proposed a technique for mining the information over the web for prediction of user's behavior. When information is large enough to store, it is difficult to know the user behavior over the web, however the proposed framework reduces the implementation complexity over the real world data.

Srividya K. Bansal et al [5], uses Extraction, Transformation and Loading (ETL) process to integrate data from heterogeneous sources into meaningful semantic model and then can apply various mining algorithms. The extract transform load (ETL) framework is used for integrating data from multiple sources or applications, possibly even from different domains.

Liu Jun et al [6], proposed a Zipf-like model which analyses the behavior of user over the mobile internet era and helps to navigate easily through pages. It uses diurnal pattern clustering algorithm to cluster the data into k-cluster centers. These clustering patterns reveal various usage patterns of the user which has helped to find out hidden patterns and understand user patterns.

II. Existing System

The rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.

The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.[2]

III. System Architecture

The propose system is built on Windows XP/7 operating system and require big data processing framework to discovering useful knowledge from the big data.

In System Architecture, the user wants to gather data from the any data source then preprocess the data. Then it stores in the dataset which here act as MySQL database. It uses the technique called clustering with spectral algorithm and gets the extract data or knowledge for the future analysis of task. Therefore this algorithm is mostly useful for the user to get the extract information while having the large amount of data.

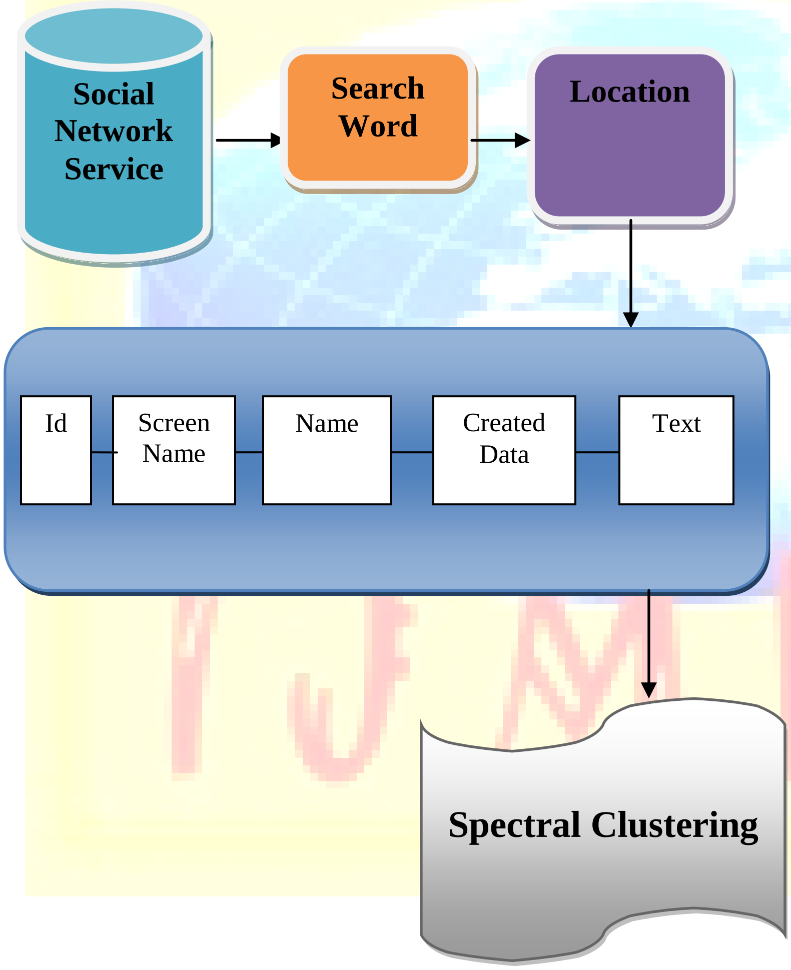

Fig -1: **System Architecture**

IV. System Implementation

If we have a set of n objects x1, x2, x3, …., xn with a pairwise similarity function defined between them which are symmetric and non-negative. Spectral clustering is the set of methods and techniques that partition the set into clusters by using the eigenvectors of matrices. The motivation behind using eigenvectors for clustering is that the change of representation induced by the eigenvectors makes the cluster properties of the initial data set much more evident. In this way, spectral clustering is able to separate data points that could not be resolved by applying directly k-means clustering, for instance, as the latter tends to deliver convex sets of points. Since the introduction of spectral methods in there have been several researches where scientists have tried using different matrices for the calculation of eigenvectors followed by applying clustering on the eigenvectors. Therefore, the final clustering result generated by spectral clustering must be improved from the initial clustering result. That is, our method is regarded as the method to improve the given clustering result. In the experiment, we used document data sets to evaluate our method.

The system is developed using MySQL database has been used as the back end. Eclipse JAVA framework will be needed for development of JAVA Server Programs. The system makes personal recommendations to users based on their purchase history. All the algorithms are written in java. We will conduct a number of experiments to verify effectiveness of the proposed methods. All the experiments will be conducted on PC with an Intel Core i3 and 512 MB (min) RAM.

V. Conclusions

The amounts of data is growing exponentially worldwide due to the explosion of social networking sites, search and retrieval engines, media sharing sites, stock trading sites, news sources and so on. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities.In this paper we discussed about the how to analyze the twitter data. Big Data analysis tools like Map Reduce over Hadoop and HDFS which helps organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data. In addition to that we introduce some big data mining tools and how to extract a significant knowledge from the Big Data. That will help the research scholars to choose the best mining tool for their work.

References

1. “Big Data for Development (2012) -: Challenges and Opportunities”, Global Pulse, May
2. G. Q. Wu, X. Wu, X. Zhu (January 2014) “Data mining with big data,” IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107.
3. Ge Yu, Qiang Wang, Shimin Chen and Yanfeng Zhang (January 2014) ”i2MapReduce: Incremental MapReduce for Mining Evolving Big Data”, IEEE Transactions On Knowledge And Data Engineering, Vol. , No. ,
4. Cyrus Shahabi, Jie Xu, Dingxiong Deng, Mihaela van der Schaar, Ugur Demiryurek, “Mining the Situation: Spatiotemporal Traffic Prediction with Big Data”,DOI 10.1109/JSTSP.2015.2389196, IEEE Transaction, 015.
5. Sebastian Kagemann, Srividya K. Bansal (2015) “Integrating Big Data: A Semantic Extract- Transform-Load Framework”, Cover Feature Big Data Management, IEEE Computer Society.
6. LIU Jun, LI Tingting, CHENG Gang, YU Hua, LEI Zhenming (December 2013) “Mining and Modelling the Dynamic Patterns of Service Providers in Cellular Data Network Based on Big Data Analysis”, China Communications, Ict Industry Convergence.
7. Sagiroglu, S.; Sinanc, D. (2013) "Big data: A review," Collaboration Technologies and Systems (CTS), International Conference on , vol., no., pp.42,47, 20-24 May 2013
8. Albert Bifet, Wei Fan, “Mining Big Data: Current Status and Forecast to the Future”, SIGKDD Explorations, 14 (2), pp1-5
9. Sharayu s. Sangekar1, pranjali p. Deshmukh, “data mining of complex data with multiple, autonomous sources”, international journal of pure and applied research in engineering and technology, issn: 2319-507x, volume 2 (9): 793-799