Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

The Challenging View of Big Data Mining

[1]Pravin Anil Tak, [2]Dr. S. V. Gumaste, [3]Prof. S. A. Kahate
[1]M.E. II[nd] Year Student, [2]Professor, [3]Assistant Professor,
[1, 2, 3] Computer Engineering Department, Sharadchandra Pawar College of Engineering,
Dumberwadi, Otur, Tal- Junnar, Dist- Pune, Maharashtra, India

Abstract— The "Big data" now a days become a most facing word by many businesses since everyone is interacting with world wide web to fulfil needs since the large amount of data arrives in heterogeneous, unstructured in format evolving a complex relations among data results in an inconsistency and complex in structure of system. Big data mining defines more challenges like heterogeneity, scale, speed, accuracy, privacy. By considering challenges evolved due to big data, an efficient system design can possible, this can be able to mine big data in short time and give real time results.

Keywords—Big data; heterogeneity; scale; trust; privacy.

I. INTRODUCTION

As Big Data handling becomes a largest problem in recent days, since all knows that, now most of businesses using enterprise multitier structure to make it efficient for online selling and purchasing - online shopping. Big data defines vulnerability because of its many factors and properties like huge volume, showing complex in structure. Observation of big data researchers shows as the amount of data level increments from TB level to PB level and in future also this type of increment will be continuously carried out since here is a requirement of defining the systematic structure for handling such big data with help of existing computer systems.

As the amount data get increases, process of information retrieval becomes very complicated since it requires better evaluation techniques and goes through different phases to complete the process. While "big data" has become a most highlighted buzzword since recent year, "big data mining", i.e., mining from big data, has almost seen as an emerging, interrelated research area. Big data mostly comes in the form of streams of a variety of types called as unstructured format. Time is an effective and important dimension of data streams, which often states that the data must be processed or mined in a real-time manner. Besides, the current major consumers of big data, corporate businesses, are especially interested in "a big data environment that can goes to handle critical business questions that demonstrate business values."

From the base of data mining perspective, mining big data has defined many new challenges and opportunities. Even though big data takes greater value (i.e., hidden knowledge and more valuable insights and decision making results), it brings tremendous and time variant challenges to extract these hidden knowledge and insights from big data therefore the establishment of process of knowledge discovering and data mining from conventional datasets was not designed to and will not work well with big data. Big data is surely playing a critical role in recent days in all walks of our lives around world. For example, governments had started mining the contents of social media networks and blogs, and online-transactions and other sources of information to retrieve the need for government facilities, to acquire the suspicious organizational groups, and to predict future events (threats or promises). Additionally, service providers started to track their customer's purchases made through online and customers' behaviours through recorded streams of online clicks, as well as product reviews and ranking, for improving their marketing efforts, predicting new growth points of profits, and increasing customer satisfaction.

This paper consist of four sections such as section I defines introduction, section II shows background section III states challenges and problems occurs during handling of big data, section IV concludes paper.

II. BACKGROUND

As earlier defined, big data is a biggest problem in recent days faced by different information technology and others industries. There are various methods and strategies have defined by different data scientist to acquire solutions towards management of big data.

WANG Shuliang defines big data as: Big data is complex data set that has following main characteristics: Volume, Variety, Velocity and Veracity. These make it difficult to use the existing tools to manage and manipulate. Big data analysis redefines the relationship among individuals, organizations, businesses, governments and societies through networked thinking and further to improve the human living environment, to enhance the quality of public services, to improve performance, efficiency and productivity through the intelligentized interactive operating. The technological progress and industrial upgrading of big data will create new markets, new business models, and new industry rules and more importantly it demonstrates the collective will of a country that looking for strategic advantage. [2]

Junyu Xuan et all defines web events is also one kind of Big data. On the other hand, there are numerous websites publishing webpages to cover the events occurring in society. The web events data satisfies the well-accepted attributes of big data: Volume, Velocity, Variety and Value. As a great value of web events data, website preferences can help the followers of web events, e.g. people or organizations, to select the proper websites to follow their interested aspects of web events. However, the big volume, fast evolution speed, multisource and unstructured data all together makes the value of website preferences mining very challenging**.** [6]

They defined procedure as website preference is formally defined at first. Then, according to the hierarchical attribute of web events data, proposed a hierarchical network model to organize big data of a web event from different service providers called as organizations, different areas and different nations at a given time stamp. With this hierarchical network structure in hand, two strategies are defined to mine the value of websites preferences from web events data. The first one is a straightforward strategy that utilizes the communities of keyword level network and the mapping relations between websites and keywords to unveil the Value in them. By taking the whole hierarchical network structure into consideration, an iterative algorithm is proposed in second strategy to refine the keyword communities like the first strategy. And secondly, an evaluation criterion of website preferences is designed to compare the performances of two proposed strategies. [6]

Xindong Wu et all proposed HACE Theorem for mining Big data. Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. They represented the conceptual view of big data processing frame work as shown in Fig-1. Huge data comes in heterogeneous manners means, data is in various schemata same way arrives in a network. A schema defines how the data get stored in a database and quantity shows the hugeness as these changes according to time [1] [3]. In a dynamic network, the structure of network changes at every moment means, hosts in the network added newly or may get reduced from existing one. It is very hard to mined huge data that comes from various autonomous sources across distributed decentralized dynamic network. By following different schemata from the distributed database the complex and evolving relations set are evaluated to define a static schema across the network. Evolving relationship evaluated from dynamic network by capturing snapshots of entire network [10]. Through these snapshots Evolving Induced Relational State (EIRS) evaluated to define active hosts from network and flow of transfers of data identified for the better management of big data arrives in a network [10].

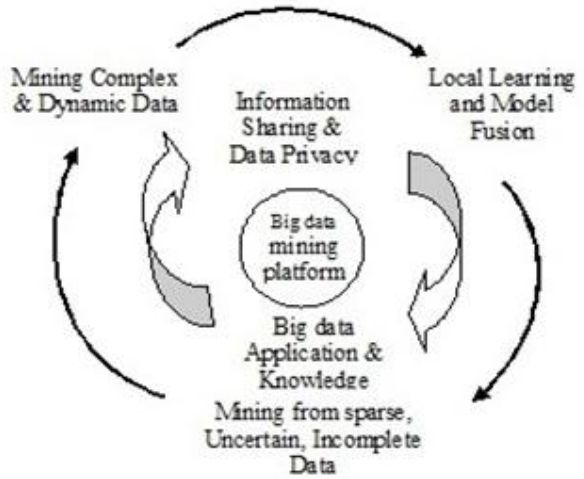

Fig. 1: Big Data Processing Framework

III. CHALLENGING VIEW REGARDING BIG DATA

Before begin to work towards big data mining, consideration of problems and different issues are necessary since the evaluation of method for retrieving efficient information from largest heterogeneous data set can becomes easily possible. As discussed earlier, in background section that mainly big data characterized by features such as Volume, Velocity, Variety and Value individually defines a complex structure of big data and due to this strong methods are required to mine data.

The difference between requirements of the big data management and abilities that current database management systems can provide has reached the historically high peak. [2][3] The three V's (volume, variety and velocity) of big data each define one separate and distinct aspect of critical deficiencies of current database management systems. Gigantic volume acquires equally great scalability and massive parallelism that are beyond the capability of current database management systems; the various data types of big data particularly not suitable to restriction of the closed processing architecture of current database systems.

Following are some important challenges and issues regarding big data mining as heterogeneity, scale, speed, accuracy and trust, privacy crisis, inter activeness, and garbage mining.

a) Heterogeneity

An existing data mining techniques have been used to discover unknown patterns and relationships of interest from structured, homogeneous, and small datasets. Heterogeneity is one of the essential features of big data, comes from the

phenomenon that there exists an unlimited different source that generates or contributes to big data.[1] This phenomenon naturally leads to the highest value in heterogeneity of big data. The data from different and various sources inherently possess many different types and representation forms, and it may formed interconnected, interrelated, and delicately and inconsistently represented. Mining from such a large and heterogeneous dataset, which is typically a tremendous network of interrelated data elements of diverse types, such as an academic social network consisting of authors, papers, conferences, universities, and companies, containing links such as work-at, write, written-by, appear-in, and present, etc. Mining such a dataset, the biggest challenge is perceivable and the degree of complexity is not even imaginable before deeply get there. Heterogeneity in big data also means that it is an obligation to acquire and deal with structured, semi-structured, and even entirely unstructured data simultaneously. [1] [5] [7]

b) Scalability

The unprecedented scale of big data defines commensurately high scalability of its data management and mining tools. Instead of being timid, the extreme scale of big data because more data bears more potential insights and knowledge that have no chance to discover from conventional data of smaller scales. Big data mining straightforwardly implies extremely time-consuming navigation in a gigantic search space, and prompt feedback/interference/guidance from users ideally domain experts must be beneficially exploited to help make early decisions, adjust search/mining strategies.[12][17]

c) Speed-

The capability of fast accessing and mining big data is not just a subjective desire; it is an obligation especially for data streams that are not in structured or semi-structured format since processing results are less valuable or even worthless. The speed of data mining depends on two major factors such as data access time and efficiency of mining algorithms themselves.[8][10] An additional approach to enhance the speed of big data access and mining is through maximally identifying and exploiting the potential parallelism in the access and mining algorithms. To enhance speed of big data mining the concept of parallelism has been utilized. In data parallelism the original data set is divided into number of small sub set and the same program runs on each of the partitions. Results obtained by running the program on each data partitions is later on combined to get the final result.[8][11][15]

d) Accuracy and Trust-

As in earlier data mining results accurate results since they are typically fed to accurate data which is in same format and small in size. However now data comes from different sources and comes in their own schemata since the upcoming values may reliable or not therefore to process this kind of data strong analytical algorithms are required since it gives an accurate and trust worthy results.[13][14] The vast volume of big data attributes additional properties defined are high dynamics and evolution. So an adequate system for big data management and analysis must allow dynamic changing and evolution of the hosted data items. This makes data provenance an integral and essential feature in any system that deals with big data. Provenance relates to the evolution history or the origin that a data item was extracted or collected from.[4] [6]

Provenance directly contributes to accuracy and trust of the source data and the derived (or mined) results. However, provenance information may not be always recorded or available. When the missing provenance of some data becomes a keen interest of the users, data mining can be reversely applied to derive and verify the provenance. Without a great many sources in the past, many provenance mining problems are unsolvable. [13][11]

e) Privacy Crisis

Data privacy has been always a challenge even from the beginning when data mining was applied to real-world data. The concern has become extremely serious with big data mining that often requires personal information in order to produce relevant/accurate results such as location-based and personalized services, e.g., targeted and individualized advertisements. Also, with the huge volume of big data such as social media that contains tremendous amount of highly interconnected personal information, every piece of information about everybody can be mined out, and when all pieces of the information about a person are dug out and put together, any privacy about that individual instantly disappears.[17] [18]

f) Interactiveness

By inter-activeness means the capability of a data mining system that allows fast and adequate user interaction such as feedback/interference/guidance from users. Inter-activeness is relatively an underemphasized issue of data mining in the past. When our society is now confronting the challenges of big data mining, inter-activeness becomes a critical issue. Inter-activeness relates to all the attributes of big data and can help overcome the challenges coming along with each of them. Great inter-activeness boosts the acceptance of a complicated mining system and its mining results by potential users. [7] [5]

g) Garbage Mining

Garbage means having no value. As data comes from different sources in the network and upcoming data is in a unstructured format, collection of such big data at central depository results enormous amount of garbage collection and this results in wastage of storage space and decreases the speed of operations that carried out to mine data.[10] [9]

In the big data era, the volume of data generated and populated on the World Wide Web keeps increasing at an amazingly fast pace.[7] [9] In such an environment, data can become out dated, corrupted, and useless as time passes. As in daily life professional communication can be carried out via mail during this enormous data arrives, which is not used at all and consuming a large space called as junk since such useless information must be removed time to time or may be avoid results in enhancing reliable communication. Thus Cleaning a garbage collection from an existing data structure is becomes necessary to mine data efficiently and fast. [15][16]

IV. CONCLUSION

In a current age, the big data era where enormous amounts of heterogeneous, semi-structured and unstructured data are continually generated at unprecedented scale. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area, still in its infancy. In spite of the limited work done on big data mining so far, can believe that much work is required to overcome its challenges related to heterogeneity, scalability, speed, accuracy, trust, provenance, privacy, and inter-activeness.

ACKNOWLEDGMENT

All faith and honours to Lord Shri Ganesh for his grace and inspiration. I wish to express my sincere thanks to all the departmental staff members for their support. Last but not the least; I would like to thank all my Friends and Family members who have always been there to support and helped me to complete this paper work in time.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE *Data Mining with Big Data* IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 1, JANUARY 2014.

[2] WANG Shuliang, DING Gangyi, ZHONG Ming *Big Spatial Data Mining* IEEE International conference on Big Data 2013.

[3] U Kang and Christos Faloutsos *Big Graph Mining: Algorithms and Discoveries* SIGKDD Explorations Volume 14, Issue 2.

[4] Tak Pravin, Kahate Sandip *An Efficient Approach for Big Data Mining* International Journal of Informative and Futuristic Research 2014.

[5] Hui Chen, Tsau Young Lin, Zhibing Zhang and Jie Zhong *Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce* IEEE International conference on Granular Computing 2013.

[6] Junyu Xuan, Xiangfeng Luo, Jie Lu *Mining Websites Preferences on Web Events in Big Data Environment* 2013 IEEE 16th International Conference on Computational Science and Engineering.

[7] Shuliang WANG Hanning YUAN *Spatial Data Mining in the Context of Big Data* 2013 IEEE International Conference on Parallel and Distributed System.

[8] Jinggui Liao, Yuelong Zhao, Saiqin Long *MRPrePost-A parallel algorithm adapted for mining big data* 2014 IEEE Workshop on Electronics, Computer and Applications.

[9] Carson Kai-Sang Leung, Richard Kyle MacKinnon, Fan Jiang *Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data* IEEE International conference on Big Data 2014.

[10] Rezwan Ahmed, George Karypis *Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks.*

[11] Johan Bollen1,Huina Mao,Xiao,Jun Zeng *Twitter mood predicts the stock market*.

[12] Yang Song, Gabriel Alatorre, Nagapramod Mandagere, and Aameek Singh *Storage Mining: Where IT Management Meets Big Data Analytics* IEEE International conference on Big Data 2013.

[13] Abdul-Aziz Rashid Al-Azmi *Data, Text, and Web Mining for Business Intelligence: A Survey* International Journal of Data Mining and Knowledge Management Process 2013.

[14] Li Liu,Murat Kantarcioglu and Bhavani Thuraisingham *Privacy Preserving Decision Tree Mining from Perturbed Data* International Conference on System Sciences 2009.

[15] Shanqing Li, Lirong Song, Hui Zhao *A Discriminant Framework for Detecting Similar Scientific Research Projects Based on Big Data Mining* 2014 IEEE International Congress on Big Data.

[16] Leila Ismail, sMohammad M. Masud ,Latifur Khan *FSBD: A Framework for Scheduling of Big Data Mining in Cloud Computing* 2014 IEEE International Congress on Big Data.

[17] Feng Ye, Zhijian Wang,Fachao Zhou, Yapu Wang,Yuanchao Zhou *Cloud-based Big Data Mining and Analyzing Services Platform integrating R* 2013 International Conference on Advanced Cloud and Big Data.

[18] Jianjun Yu, Fuchun Jiang, Tongyu Zhu *RTIC-C: A Big Data System for Massive Traffic Information Mining* 2013 International Conference on Cloud Computing and Big Data.

