International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

Survey on Big Data Mining Algorithms

Anushree Raj[1], Rio D'Souza[2]
[1, 2]*Department of M.Sc. Big Data Analytics, Department of Computer Science & Engineering*

Abstract: Technology revolution has been facilitating millions of people by generating tremendous data, resulting in big data. It has been a distinct knowledge that massive amount of data have been generated continuously at extraordinary and ever increasing scales. Big Data is a new term used to identify the datasets that due to their large size and complexity. Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This survey paper includes the information about what is big data, big data sources, Data mining, Big data mining and the challenges.
Keywords: Big data, data mining, Big data mining algorithms.

I. INTRODUCTION

Data mining refers to the steps of searching, analyzing and extracting the valuable desired data from a data warehouses to exploit problem-solving and decision-making.
This is known as Knowledge Data Discovery (KDD) [1]. Big data mining is referred to the collective data mining or extraction techniques that are performed on large datasets or volume of data or the big data. Big data mining is primarily done to extract and retrieve desired information or pattern from humongous quantity of data. Big data mining techniques and processes are also used within big data analytics and business intelligence to deliver summarized targeted and relevant information, patterns and relationships between data, systems, processes and more.

II. BIG DATA

Big-data is nothing but a data available at autonomous and heterogeneous sources in extreme large amount which gets updated within a fraction of second.
There exist several definitions for big data based on the characteristics of the data. The well-known definition based on the 3Vs underlines volume, velocity, and variety as the main characteristics of big data.
Volume: Databases include huge amounts of data. Velocity: Data is flowing to the databases in real time: real time streams of data flowing from diverse resources, either from sensors or from internet. Variety: Data is no longer of a single type. Databases include data from a vast range of systems and sensors in different formats and data types.
This may include unstructured text, logs, and videos. O'Reilly [2] defines big data as the data that exceeds the processing capacity of conventional database systems. Further explains that the data is very big, moves very fast, or doesn't fit into traditional database architectures. Big data is voluminous and has large-volume of data coming from heterogeneous, independent sources with dispersed and decentralized control, and attempts to find and explore complex and evolving relationships among them [3].

III. DATA MINING

The explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools. Data mining turns a large collection of data into knowledge. Data mining can be viewed as a result of the natural evolution of information technology.
Data mining can also be termed as "knowledge mining from data," but, knowledge mining may not reflect the emphasis on mining from large amounts of data [4].
The knowledge discovery process is an iterative sequence of the following steps:

1) Data cleaning - to remove noise and inconsistent data.
2) Data integration - where multiple data sources may be combined.
3) Data selection - where data relevant to the analysis task are retrieved from the database.
4) Data transformation - where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.

5) Data mining - an essential process where intelligent methods are applied to extract data patterns.
6) Pattern evaluation - to identify the truly interesting patterns representing knowledge based on interestingness measures.
7) Knowledge presentation - where visualization and knowledge representation techniques are used to present mined knowledge to users.

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [5]. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

IV. BIG DATA SOURCES

The major sources of big data are from the following [6]:

A. Archives are mainly maintained by organizations, to show the function of a particular person or organization functions. Accumulation of archives sometimes does not fit into the traditional storage systems and need systems with high processing capabilities. This voluminous archive contributes to big data.
B. Media Users generate images, videos, audios, live streams, podcasts and so on contributes for big data.
C. Business applications Huge volumes of data are generated from business applications as part of project management, marketing automation, productivity, customer relation management (CRM), enterprise resource planning (ERP) content management systems, procurement, human resource (HR), storage, talent management, Google Docs, intranets, portals and so on. These data contributes to big data
D. Public web Many organizations under government sector, weather, competitive, traffic regulatory compliance, health care services, economic, census, public finance, stock, open source intelligence (OSINT), the world bank, electronic data gathering analysis and retrieval (Edgar), Wikipedia and so on uses web services for communication. These data contributes to big data
E. Social Media Nowadays users rely on social media sites such as twitter, linkedln, facebook, tumblr, blog, slideshare, youtube, google+, instagram, flickr, pinterest, vimeo, wordpress and so on for the creation and exchange of user generated contents. These social networking sites contribute to big data.
F. Data Storage Data storage in SQL, NoSQL, Hadoop, doc repository, file systems and so on also contributes to big data.
G. Sensor Data Accumulation of large quantitative datasets from distributed sensors are now becoming widely available online from medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes or household appliances, assembly lines.

V. BIG DATA MINING

Big data mining is the capability of extracting useful information from these large datasets or streams of data, which was not possible before due to data's volume, variability, and velocity [7]. Big data is a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques. Big data technologies have great impacts on scientific discoveries and value creation [8, 9, 10]. Structured (numerical) and unstructured (textual) are two main types of data forms in big data.

VI. BIG DATA MINING ALGORITHMS

A. Decision Tree Induction Classification Algorithms

In the initial stage different Decision Tree Learning was used to analyze the big data. In decision tree induction algorithms, tree structure has been widely used to represent classification models. Most of these algorithms follow a greedy top down recursive partition strategy for the growth of the tree. Decision tree classifiers break a complex decision into collection of simpler decision. Hall. et al. [11] proposed learning rules for a large set of training data. The work proposed by Hall et al generated a single decision system from a large and independent subset of data. An efficient decision tree algorithm based on rainforest frame work was developed for classifying large data set [12].

B. Evolutionary Based Classification Algorithms

Evolutionary algorithms use domain independent technique to explore large spaces finding consistently good optimization solutions. There are different types of evolutionary algorithms such as genetic algorithms, genetic programming, evolution strategies, evolutionary programming and so on. Among these, genetic algorithms were mostly used for mining classification rules in large data sets [13]. Patil et al. [14] proposed a hybrid technique combining both genetic algorithm and decision tree to generate an optimized decision tree thus improving the efficiency and performance of computation. An effective feature and instance selection for supervised classification based on genetic algorithm was developed for high dimensional data [15].

C. Partitioning Based Clustering Algorithms

In partitioning based algorithms, the large data sets are divided into a number of partitions, where each partition represents a cluster. K-means is one such partitioning based method to divide large data sets into number of clusters. Fuzzy- CMeans is a partition based clustering algorithm based on Kmeans to divide big data into several clusters[16]

D. Hierarchical Based Clustering Algorithms

In hierarchical based algorithms large data are organized in a hierarchical manner based on the medium of proximity. The initial or root cluster gradually divides into several clusters. It follows a top down or bottom up strategy to represent the clusters. Birch algorithm is one such algorithm based on hierarchical clustering [17]. To handle streaming data in real time, a novel algorithm for extracting semantic content were defined in Hierarchical clustering for concept mining [18].This algorithm was designed to be implemented in hardware, to handle data at very high rates. After that the techniques of self-organizing feature map (SOM) networks and learning vector quantization (LVQ) networks were discussed in Hierarchical Artificial Neural Networks for Recognizing High Similar Large Data Sets [19]. SOM consumes input in an unsupervised manner whereas LVQ in supervised manner. It subdivides large data sets into smaller ones thus improving the overall computation time needed to process the large data set.

E. Density Based Clustering Algorithms

In density based algorithms clusters are formed based on the data objects regions of density, connectivity and boundary. A cluster grows in any direction based on the density growth. DENCLUE is one such algorithm based on density based clustering [20].

F. Grid Based Clustering Algorithms

In grid base algorithms space of data objects are divided into number of grids for fast processing. OptiGrid algorithm is one such algorithm based on optimal grid partitioning [21].

G. Model Based Clustering Algorithms

In model based clustering algorithms clustering is mainly performed by probability distribution. Expectation Maximization is one such model based algorithm to estimate the maximum likelihood parameters of statistical models [22].

VII. CHALLENGES

A. Evaluating the interestingness of mined patterns big data mining or simply mining allows the discovery of knowledge which is potentially useful and unknown. If the knowledge extracted is new, useful or interesting, it is very subjective and mainly depends upon the application and the user who uses the data. It is firm that data mining can create, or discover, a large number of patterns or rules. Identifying and measuring the interestingness of patterns and rules discovered, or to be discovered is a must for the evaluation of knowledge discovery process. Assessing how interesting a mined pattern is still an important research issue [23]

B. Building a global unifying theory of big data mining Many techniques are designed for performing classification or clustering individually, but there is no theoretical framework that unifies different tasks such as classification, clustering and association rules and so on. Therefore building a global unifying theory for mining big data is an active research area [24].

C. Scaling up to meet the growing needs of large data sets In order to meet the growing demands of data, we need to scale up both in terms of capacity and performance measures effortlessly. Big data needs more capacity, scalability, and efficient processing capabilities without increasing the resource demands. In traditional systems, storage architectures were designed in such a way to scale up with the growing needs of data. But it really affects the performance capacity of the storage systems. So organizations dealing with big data should design an optimal storage architecture which offers the features such as scalability, high performance, high efficiency, operational simplicity, interoperability and so on to manage growth.

D. Building efficient big data mining platform to handle big data and its characteristics, an efficient Big data Processing and computing framework is needed. Traditional data mining algorithms only needed all the data to be loaded into the main memory and perform the operation of data mining. In medium scale data processing, parallel computing is used with limited number of processors. As big data applications are characterized by autonomous sources and decentralized controls, consolidating distributed data sources to a centralized node for mining is systematically discouraging due to the potential transmission cost and privacy issues. So building an efficient platform to mine big data is essential.

E. Building efficient mining algorithms/models for big data with the exponential growth of data, traditional data mining algorithms have been unable to meet large data processing needs. In order to deal with big data, an efficient model that deals

with cost effective computation of huge, heterogeneous, sparse, incomplete, complex data are needed. The main drawbacks of big data mining algorithms are lack of user-friendly interaction support, quality and performance. Data mining algorithms usually needs scanning of entire data for obtaining perfect statistics and they may require intensive computing. Therefore it is essential to improve the efficiency and performance of data mining algorithms to handle big data.

F. Maintaining security, trust and data integrity. Security is a major concern with big data. In order to ensure security, organizations need to establish policies, which are self-configurable. Another major issue in the field is trust of data sources which are not well – known and not at all verifiable. Data Integrity should be maintained by adopting the best practices in the industry [25].

G. Data privacy issues Data privacy has been always a serious issue right from the beginning of Data mining applications. The concern has become extremely vigorous with big data mining that often needs personal information in order to give relevant and accurate outputs [26]. Also, the massive volume of big data such as in social media sites that contains tremendous amount of highly interconnected personal information can be easily mined out and when all pieces of the information about a person are mined out and put together, any privacy about that individual rapidly disappears

VIII. CONCLUSION

It is known that big data mining is an emerging trend in all science and engineering domains and also a promising research area. In spite of the limited work done on big data mining so far, it is believed that much work is required to overcome its challenges related to the above mentioned issues. From the perspective of data mining problem, this paper gives a brief introduction to the big data, big data sources and big data mining algorithms and challenges.

REFERENCES

[1] Saed Sayad, Data Mining Map, An Introduction to Data Mining, http://www.saedsayad.com/ (2012). (Last seen 05–April–2015)
[2] O'Reilly Radar, What is bigdata? http://radar.oreilly.com/2012/01/what-is-big-data.html. January 11,2012
[3] Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE Transactions on Knoweledge and Data Engineering Vol 26 No1 Jan 2014
[4] D. W.-L. Cheung, V. Ng, A. W.-C. Fu, and Y. Fu. E□cient mining of association rules in distributed databases. Transactions on Knowledge and Data Engineering, 8(6):911–922, Dec. 1996.
[5] C. Clifton and D. Marks. Security and privacy implications of data mining. In A CM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pages 15-19, May 1996
[6] U. Fayyad Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. http:// big-datamining.org/keynotes/#fayyad 2012
[7] W. Fan, A. Bifet, "Mining Big Data: Current Status, and Forecast to the Future," ACM SIGKDD Explorations, Vol. 14, No. 2, pp. 1-5, December 2012.
[8] Demchenko, P. Grosso, C. D. Laat, P. Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure," 2013 International Conference on Collaboration Technologies and Systems (CTS), 20-24 May 2013, San Diego, CA, USA, pp. 48-55, 2013.
[9] D.E. O'Leary, "'Big Data', the 'Internet of Things' and the 'Internet of Signs'," Intelligent Systems in Accounting, Finance and Management, Vol. 20, pp. 53-65, 2013.
[10] H.V. Jagadish, A. Labrinidis, Y. Papakonstantinou, et al., "Big Data and Its Technical Challenges," Communications of the ACM, Vol. 57, No. 7, pp. 86-94, 2014.
[11] Lawrence 0. Hall, Nitesh Chawla , Kevin W. Bowyer, "Decision Tree Learning on Very Large Data Sets", IEEE, Oct 1998
[12] Thangaparvathi, B., Anandhavalli, D An improved algorithm of decision tree for classifying large data set based on rainforest framework, Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on Oct. 2010 Page(s):800 – 805
[13] D. L. A Araujo., H. S. Lopes, A. A. Freitas, "A parallel genetic algorithm for rule discovery in large databases" , Proc. IEEE Systems, Man and Cybernetics Conference, Volume 3, Tokyo, 940-945, 1999.
[14] Mr. D. V. Patil, Prof. Dr. R. S. Bichkar,"A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data Sets", IEEE, 2006
[15] Ros, F., Harba, R. ; Pintore, M. Fast dual selection using genetic algorithms for large data sets, ,Intelligent Systems Design and Applications (ISDA), 12th International Conference on Date of Conference:27-29 Nov. 2012 Page(s):815 – 820, 2012.
[16] J. C. Bezdek, R. Ehrlich, and W. Full. Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2):191–203,1984.
[17] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. ACM SIGMOD Record, volume 25, pp. 103–114, 1996
[18] Moshe Looks, Andrew Levine, G. Adam Covington, Ronald P. Loui, John W. Lockwood, Young H. Cho, 2007,"Streaming Hierarchical Clustering for Concept Mining", IEEE, 2007
[19] Yen-ling Lu, chin-shyurng fahn, 2007, "Hierarchical Artificial Neural Networks For Recognizing High Similar Large Data Sets. ", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, August 2007
[20] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. Proc. of the ACM SIGKDD Conference on Knowledge Discovery ad Data Mining (KDD), pp. 58–65, 1998.
[21] A. Hinneburg, D. A. Keim, et al. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. Proc. Very Large Data Bases (VLDB), pp. 506–517, 1999.
[22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), pp. 1–38,1977.

[23] Vashishtha, J. GJUST, Kumar, D. ; Ratnoo, S., Revisiting Interestingness Measures for Knowledge Discovery in Databases, Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on Jan. 2012 Page(s):72 – 78
[24] QIANG YANG, XINDONG WU, 10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH, International Journal of Information Technology & Decision Making,Vol. 5, No. 4 (2006) 597–604, World Scientific Publishing Company.
[25] A. Rubin and D. Greer. A survey of the world wide web security. IEEE Computer, 31(9):34-41, Sept. 1998.
[26] Office of the Information and Privacy Com- missioner, Ontario. Data Mining: Stak- ing a Claim on Your Privacy, January 1998. Available from http ://.ww. ip¢. on. ca/ web.site, eng/mat t ors/sum_pap/papers/ dat amine, him.