A Survey on Big Data Mining Platforms, Algorithms and Handling Techniques

Rajkumar.D[1], Usha.S[2]

[1]Rajkumar.D, Department of Computer Applications/ SRM University, Ramapuram, Chennai, India

[1]rajkumar.d@rmp.srmuniv.ac.in

[2]Usha.S, Department of Computer Applications/ SRM University, Ramapuram, Chennai, India

[2]usha.s@rmp.srmuniv.ac.in

ABSTRACT

Big data is an magical term that describes a collection of data sets which are large and complex, growing data sets with multiple, autonomous sources , it contain structured and unstructured both type of data. Hence, there seems to be a need for an analytical review of recent developments in the big data technology. This extremely large-scaled data called Big data are in terms of quantity, complexity, semantics, distribution, and processing costs in computer science, cognitive informatics, web-based computing, cloud computing, and computational intelligence. The size of the collected data about the Web and mobile device users is even greater. To provide the ability to make sense and maximize utilization of such vast amounts of web data for knowledge discovery and decision-making is crucial to scientific advancement; we need new tools for such a big web data mining. A review on various big data mining algorithms and the methods employed to handle such a vast data is also discussed in this paper

Keywords — Big data, Big data mining, Big data mining algorithms, Visual web mining, Apache Hadoop.

1. INTRODUCTION

Big data is a term encompassing different types of complicated and large datasets that is hard to process with the conventional data processing systems. Numerous challenges are in place with big data like storage, transition, visualization, searching, analysis, security and privacy violations and sharing. The exponential growth of data in all fields demands the revolutionary measures required for managing and accessing such data. In [1], the authors have highlighted the need for the research in big data, in order to manage the online bio-logical data avenue. They have foreseen the importance of big data in the biological and biomedical research. It has exploded in such a way that it has marginalized a regulatory schema for personally identifiable information [2]. This is possible by analyzing the meta data and by using the predictive, aggregated findings thereby combining the previous discrete data sets. The significance of big data analytics comes when enterprises choose a technical stack, which dictates the type of data to store and to process. Relational Data Base management Systems are doing fine with structured data and continue to be the choice for many requirements. But for the exponential growth of unstructured data in terabytes or even peta bytes, derived from social networks, sensor networks and other federated data with replications, big data is the answer for handling such data. 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 [3]. Thus making big data mining or knowledge discovery of large datasets a difficult process.

2. BIG DATA

Big data technologies defines a new generation of technologies and architectures, designed solely to economically extract useful information's from very large volumes of a wide variety of data, by permitting high velocity capture, discovery, and analysis[4]. O'Reilly [5] defines big data is the data that exceeds the processing capacity of conventional database systems. He also explains that the data is very big, moves very fast, or doesn't fit into traditional database architectures. Further he has extended that to gain value from this data, one has to choose an alternative way to process it. There are mainly 3 types of big data sets- structured, semi structured and unstructured [6]. In structured data, we can group the data to form a relational schema and represent it using rows and columns within a standard database. Based on an organization's parameters and operational needs, structured data responds to simple queries and provides usable information due to its configuration and consistency. Semi structured data [7] does not conform to an explicit and fixed schema. The data is inherently self-describing and contains tags or other markers to enforce hierarchies of records and fields within the data. Certain examples for semi-structured data include weblogs and social media feeds. The formats of unstructured data cannot be easily indexed into relational tables for analysis or querying. Certain examples for unstructured data's are image files, audio files, video files, and health records and so on.

3. FIVE V'S OF BIG DATA

There are many properties associated with bigdata. The prominent aspects are Volume, Velocity Variety, Vercacity and Value.

3.1 Volume: The volume of big data is exploding exponentially day to day. The data accumulated through social websites and sensor networks going to cross from petabytes to Zetabytes.

3.2 Velocity: This is a concept which indicates the speed at which the data generated and become historical. Big data is able to handle the incoming and outgoing data rapidly.

3.3 Variety: Data produced are from different categories, consists of unstructured, standard, semi structured and raw data which are very difficult to be handled by traditional systems.

3.4 Vercacity: It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set.

3.5 Value: All enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services. For that, study on customer attitudes and trends in the market are to be analyzed. Moreover, users can also query the data store to find business trends and accordingly they can change their strategies. By making big data open to all, it creates transparency on functional analysis. Supporting real time decisions and experimental analysis in different locations datasets can do wonderful things for enterprises.

4. BIG DATA MINING PLATFORMS

4.1 Google's Map Reduce, Hadoop and Google Big Table

Google's programming model, Map Reduce, and its distributed file system, Google File System (GFS) [8] are the pioneers in the field. Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing has received a significant amount of attention, with Map Reduce parallel programming. So with this concept many companies provide big data processing framework that support Map Reduce. After that Yahoo and related companies developed Hadoop [9] uses the Hadoop Distributed File System (HDFS) – an open source version of the Google's GFS. Later in this field to support the Map Reduce computing model strategy, Google developed the BigTable [10] in 2006– a distributed storage system designed for processing structured data with size in the order of petabytes.

4.2 Dynamo

In 2006 Amazon developed Dynamo [11], which uses a key-value pair storage system. Dynamo is a highly available and scalable distributed data store built for Amazon's platform. Dynamo is used to manage services that need high reliability, availability, consistency, performance and cost effectiveness.

The following models are also developed to support big data management and processing.

4.3 HBase

HBase [12] is an open source, non-relational, distributed database developed after big table. It works on the top of Hadoop Distributed file system and provides big-table like capabilities for Hadoop

4.4 Apache Hive

Apache Hive [13] is a data warehouse infrastructure built on top of Hadoop. It provides data summarization, query, and analysis of big data.

4.5 Berkeley Data Analytics Stack(BDAS)

The Berkeley Data Analytics Stack (BDAS) [14] is an open source data analytics stack that integrates software components being built by the UC Berkeley AMPLab for computing and analyzing big data. Many systems in the stack provide higher performance over other big data analytics tools, such as Hadoop. Nowadays, BDAS components are being used in various organizations.

4.6 ASTERIX

ASTERIX [15] is an Open Source System for big data management and analysis. With the help of ASTERIX Semi structured data can be easily ingested, stored, managed, indexed, retrieved and analyzed. Many of the drawbacks of Hadoop and similar platforms such as single system performance, difficulties of future maintenance, inefficiency in extracting data and awareness of record boundaries etc are easily overcome by ASTERIX.

4.7 SciDB

SciDB [16] is an open-source data management and analytics software system (DMAS) that uses a multidimensional array data model. SciDB is designed to store petabytes of data distributed over a large number of machines and used in scientific, geospatial, financial, and industrial applications.

4.9 Hadoop Map Reduce

These algorithms work on top of Hadoop and make use of Map Reduce programming model.

4.8 NIMBLE

A open source toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on Map Reduce for large datasets. It allows users to compose parallel ML-DM algorithms using reusable (serial and parallel) building blocks that can be efficiently manipulated using almost all parallel

programming models such as Map Reduce. It runs on top of Hadoop.

4.9 Big Cloud-Parallel Data Mining(BC-PDM)

Big Cloud Parallel Data Mining mainly relies on cloud computing and works on top of Hadoop and mainly used in intelligence data analysis.

4.10 Graph Mining tools

Graphs are widely used in data mining application domains for identifying relationship patterns, rules, and anomalies. Certain examples for domains include the web graph, social networks etc. The ever-expanding size of graph-structured data for the above applications needs a scalable system that can process large amounts of data efficiently. Giraph, GraphLab, Bulk Synchronous Parallel Based Graph Mining (BPGM) are the examples for the system to process graph structured data.

5. BIG DATA MINING ALGORITHMS

5.1 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. [17] 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 [18].

5.2 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 [19]. Patil et al. [20] 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 [21].

5.3 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[22]

5.4 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[23].To handle streaming data in real time, a novel algorithm for extracting semantic content were defined in Hierarchical clustering for concept mining[24].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 [25]. 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.

5.5 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 [26].

5.6 Grid based clustering algorithms

In grid base algorithms space of data objects are divided into number of grids for fast processing. Opti Grid algorithm is one such algorithm based on optimal grid partitioning [27].

5.7 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 [28].

5.8 Scalable visual assessment of tendency (sVAT) algorithm

The scalable visual assessment of tendency (sVAT) procedure satisfies the definition of scalability and can be efficiently applied to any sized re- 63 lational data set. It produces a true VAT ordered image for the sample that is representative of the full data im- 65 age, does not involve any sensitive thresholding parameter, and requires the user to only supply choices for two parameters [29].

5.9 Distributed ensemble classifier algorithm

Distributed ensemble classifier algorithm was developed in the field based on the popular Random Forests for big data. This proposed algorithm makes use of Map Reduce for improving the efficiency and stochastic aware random forests for reducing randomness[30]

5.10 CulsiVAT Algorithm

ClusiVAT produces true single linkage clusters in compact, separated data. We also show that single linkage fails, while clusiVAT finds high quality partitions that match ground truth labels very well. And clusiVAT is fast: it recovers the preferred c = 3 Gaussian clusters in a mixture of 1 million two

dimensional data points with 100% accuracy in 3.1 seconds[31].

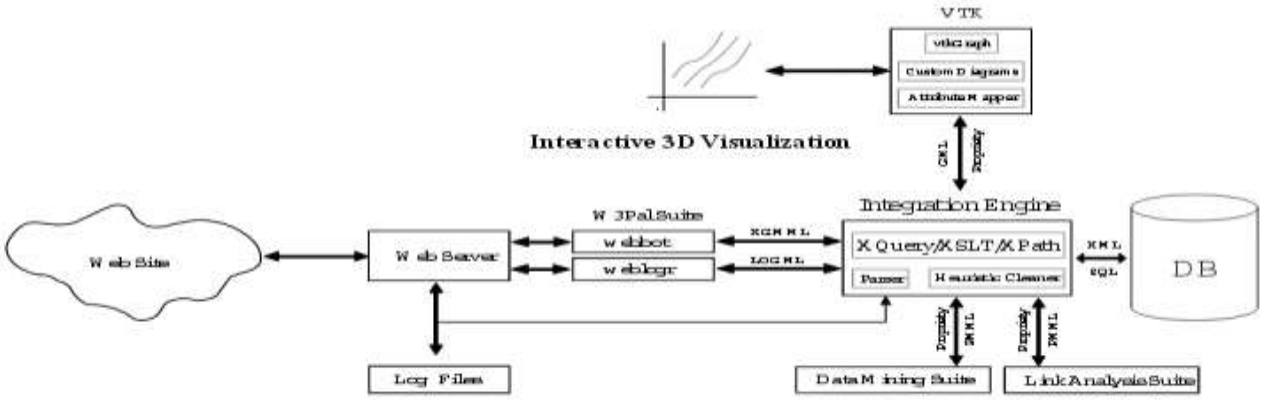

Figure 1: Sample implementation architecture of VWM

6. TECHNIQUES FOR HANDLING BIG WEB DATA

6.1 Visual Web Mining Architecture

The architecture of implementing the visual web mining is shown in below Figure. We target one or a group of websites for analysis. Input of the system consists of web pages and web server log files. Access to web log is done by the local file system, or by downloading it from a remote web server. A web robot (webbot) is used to retrieve the pages of the website [32]. In parallel, Web Server Log files are downloaded and processed through a sessionizer and a LOGML [33] file is generated

6.2 Handling Big Web Data with Hadoop Map reduce

Big Data is an emerging growing dataset beyond the ability of a traditional database tool to handle. As the use of internet and the web is becoming a daily concern of many individuals, the growth of data is becoming so high beyond the imagination of normal internet user. In addition, such a large amount of data leads to the big data problem. Hadoop rides the big data where the massive quantity of information is processed using cluster of commodity hardware. Web server logs are semi-structured files generated by the computer in large volume usually of flat text files. It is utilized efficiently by Map reduce as it process one line at a time.

7. CONCLUSION

We are living in a digital world of big data where massive amounts of heterogeneous, autonomous, complex and evolving data sets are constantly generated at unprecedented scale.The data size in all areas is exploding day to day. he velocity and variety of data growth is increasing due to the proliferation of sensor and mobile devices with internet connection. Data generated by this way, is the greatest asset for enterprises in developing and defining business strategies. 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.

measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

REFERENCES

[1] Howe AD, Costanzo M, Fey P, et al. 2008. Big data: The future of biocuration, Nature. 455(7209): 47-50. Doi: 10.1038/455047a.

[2] Crawford Kate and Jason Schultz. 2014. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, Boston College Law Review. 55(93): 93-128.

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

[4] IDC, Extracting Value from Chaos: http://idcdocserv.com/1142, june 2011

[5] O'Reilly Radar, What is bigdata? http://radar.oreilly.com/2012/01/what-is-big-data.html. January 11,2012

[6] IDC, 2012

[7] Peter Buneman, Semistructured Data http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf, 1997,

[8] Ghemawat, S. Gobioff, H., Leung, S.T, The Google File System. In: 19th ACM Symposium on Operating Systems Principles, pp. 29-

43. Bolton Landing, New York ,2003

[9] 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

[10] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes,

Robert E. Gruber, Bigtable: A Distributed Storage System for Structured Data, 2006

[11] DeCandia, G., Hastorun, D., Jampani, et al : Dynamo: Amazon's Highly Available Key-Value Store. In: 21st ACM SIGOPS

Symposium on Operating Systems Principles, pp.14-17. Stevenson, Washington, USA, 2007

[12] Lizhi Cai,Shidong Huang;Leilei Chen;Yang Zheng,Performance analysis and testing of HBase based on its architecture, Computer and

Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on June 2013

[13] Taoying Liu, Jing Liu ; Hong Liu ; Wei Li , A performance evaluation of Hive for scientific data management. Big Data, IEEE

International Conference on October 2013

[14] Mike FranklinUC Berkeley, USA, The Berkeley Data Analytics Stack: Present and future Big Data, 2013 IEEE International

Conference 9 Oct. 2013

[15] Sattam Alsubaiee, Yasser Altowim, Hotham Altwaijry, Alexander Behm, Vinayak R. Borkar, Yingyi Bu, Michael J. Carey, Raman

Grover, Zachary Heilbron, Young-Seok Kim, Chen Li, Nicola Onose, Pouria Pirzadeh, Rares Vernica, Jian Wen. ASTERIX: An Open

Source System for "Big Data" Management and Analysis, 2012

[16] Stonebraker, M.;Brown, P.;Donghui Zhang;Becla, J., SciDB: A Database Management System for Applications with Complex.

Computing in Science & Engineering (Volume:15 , Issue: 3),july 2013

[17] Lawrence 0. Hall, Nitesh Chawla , Kevin W. Bowyer, "Decision Tree Learning on Very Large Data Sets", IEEE, Oct 1998

[18] 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. 2010Page(s):800 – 805

[19] 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.

[20] Mr. D. V. Patil, Prof. Dr. R. S. Bichkar,"A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data

Sets", IEEE, 2006

[21] 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.

[22] J. C. Bezdek, R. Ehrlich, and W. Full. Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2):191–203,1984.

[23] 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

[24] 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

[25] 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

[26] 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.

[27] 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.

[28] 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.

[29] Richard J. Hathaway, James C. Bezdek, Jacalyn M. Huband: "Scalable visual assessment of cluster tendency for large data sets" February 2006.

[30] Assuncao, J. ; Comput. Sci. Dept., PUCRS Univ., Porto Alegre, Brazil ; Fernandes, P. ; Lopes, L. ; Normey, S, Distributed Stochastic Aware Random Forests -- Efficient Data Mining for Big Data, Big Data (Big Data Congress), 2013 IEEE International Congress on
June -July 2 2013

[31] Dheeraj Kumar,Marimuthu Palaniswami,Sutharshan Rajasegarar, Christopher Leckie
James C. Bezdek, Timothy C. Havens: "clusiVAT: A Mixed Visual/Numerical Clustering
Algorithm for Big Data": October 2013.

[32] J. Punin and M. Krishnamoorthy. Wwwpal system- a system for analysis and synthesis of web pages. In Proc. WebNet, 1998.

[33] J. Punin, M. Krishnamoorthy, and M. J. Zaki. Logml: Log markup language for web usage mining. In WebKDD Workshop, ACM SIGKDD , pages 88–112, 2001.

[34] T. Sathis Kumar: "A Study on Effective Business Logic Approach for Big Data Mining", 2015

[35] SHERIN A, Dr S UMA, SARANYA K, SARANYA VANI M : "SURVEY ON BIG DATA MINING PLATFORMS, ALGORITHMS AND CHALLENGES" , 2015.

[36] S. Justin Samuel, Koundinya RVP, Kotha Sashidhar and C.R. Bharathi : "A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES" :2015

[37] Pranit B. Mohata, **"**Web Data Mining Techniques and Implementation for Handling Big Data" , April 2015.

