Middle-East Journal of Scientific Research 24 (Techniques and Algorithms in Emerging Technologies): 632-637, 2016
ISSN 1990-9233;

DOI: 10.5829/idosi.mejsr.2016.24.TAET23701

Study on Big Data

N.V. Monisha, A. Nagalakshmi, Namratha Raj, N. Purnashree and K.R. Suneetha

Department of CSE, Bangalore Institute of Technology, Bangalore, India

Abstract: We are living in an era of data flood and as a result, the term "Big Data" is appearing in many contexts. Big Data is the collection of large and complex data set which is difficult to process using traditional data processing application. It has wide scope in the areas of meteorology, genomics, complex physics simulations, biological and environmental research, finance, business and healthcare. This paper reviews the existing work on mining starting from Data Mining to Big Data and discusses each of these methods.

Key words: Data Mining · Web Mining · Big Data · KDD · HDFS · AI

INTRODUCTION

Data Mining is a process of extracting interesting (nontrivial, implicit, previously unknown and potentially useful) information from large data. Data Mining is also known as Knowledge Discovery in the database. There are several Data Mining techniques like Association (pattern discovery based on the relationship between items of the same transaction), Classification (classifying each item from the set of data into predefined groups), Clustering (defines the cluster of objects with similar characteristics) and decision tree, etc. Web Mining is an application of Data Mining technique to discover patterns from World Wide Web. Web Mining is categorized into WUM (Web Usage Mining), WCM (Web Content Mining) and WSM(Web Structured Mining). WUM is used to discover patterns from Web data to understand and give better service for Web-based application. WCM is used to extract and integrate useful data, information and knowledge from Web page content. WSM is used in graph theory to analyze the node and connection structure of a website by Hyperlinks patterns. These days large accumulation of data is difficult to analyze, with the help of traditional methods of mining. This drawback of mining gave rise to the concept of BIG DATA.

Big Data is the collection of large and complex data set which is difficult to process using traditional data processing application. Properties of Big Data are Volume (data sizes over 1014 to 1021), Variety (structured data, unstructured data), Velocity (continuous arrival of high frequencies data, results in incessant high-speed data streams). The challenges in Big Data include analysis, search, sharing, storage, transfer, visualization and privacy violation. Applications of Big Data are in social networking sites, search engines, astronomy, sensor networks, government data, web blocks, mobile phones, scientific research and natural disaster. Tools of Big Data are MapReduce model, iterative MapReduce model, DAG model, Graph model, Collective model, etc. Issues of Big Data are Unknown population representation, Issues of data quality, Privacy and Confidentiality issues, Difficult to access and uncertainty, Management and processing of distributed data, New tools data analysis and visualization.

This paper presents the literature review about the Data Mining, Web Mining and Big Data. The architecture of Big Data is imparted in this paper.

The paper is organized as follows, Section 2 presents related work, Section 3 provides architecture diagrams of Web Mining and Big Data, Section 4 briefs the conclusion and Section 5 presents references.

Literature Survey: In paper [1] contains symbolic classification rules using neural networks. By the experimental results, the effectiveness of the proposed approach is clearly demonstrated to achieve performance efficiency. By this approach, concise symbolic rules can be extracted from the neural network with high accuracy. The network is first trained to achieve the required accuracy rate. The algorithm removes the redundant connections of the network. The activation values of the hidden units in the network are analyzed and classification rules are generated using the result of this analysis. The neural network based approach proposed to

Corresponding Author: N.V. Monisha, Department of ECE, Bangalore Institute of Technology, Bangalore, India.

mine classification rules, from given databases are explained. The work involves Constructing and training a network to classify, Pruning the network while maintaining the classification accuracy and Extracting symbolic rules from the pruned network.

A set of experiments were conducted using a well-defined set of data. The work proposed to generate rules similar to that of decision trees by reducing the training time of neural networks.

The paper [2] discusses on the Knowledge discovery process, Data Mining, various open source tools and improvements in the field of Data Mining from past to the present and explores the future trends. The paper presents the knowledge Discovery Process, advantages, disadvantages and challenges of Data Mining, various open source tools and trends from its beginning to the future. The paper helps to do the research by focusing on the various issues of Data Mining. Data Mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains (finance, marketing, banking, insurance, health care and retailing). Data Mining is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs, enhance research.

The authors Anand, Vaibhav, Prithi *et al.* [3] discussed about a large amount of data stored in databases using new techniques and tools. The research field Knowledge Discovery in Databases (KDD) or Data Mining. Attract attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning and data visualization. Data Mining is the process of discovering insightful, interesting and novel patterns, as well as descriptive, understandable and predictive models from large-scale data. The paper overviews different tasks include in Data Mining, anomaly detection, classification, regression, association rule learning, summarization and clustering.

The Wei Fan andreas L. Prodromidis and *et al.* [4] proposed methods for combining multiple learned fraud detectors under a "cost model" demonstrate usefully; The empirical results demonstrated in this work significantly reduces loss due to fraud through distributed Data Mining of fraud models. This process is automated, but it is unavoidable because the desired distribution highly depends on the cost model and the learning algorithm. Unlike a monolithic approach of learning one classifier using incremental learning, the proposed modular multiclassifier approach facilitates adaptation over time and removes out-of-date knowledge. The paper focuses on growing problem of intrusion detection in the network and host-based computer systems. Here sort of task as in the credit card fraud domain and build models to distinguish between bad (intrusions or attacks) and good (normal) connections or processes by applying feature-extraction algorithms, followed by the application of machine-learning algorithms (such as Ripper).

The goal of the paper[5] to discuss about all the possibilities of identifying the way post documents in users' navigation paths and to propose an optimization of navigation structure of a website based on user navigation patterns. Websites may contain numerous documents. Using some web techniques, it's possible to analyze users' data about using resources, contents of documents and structure of web sites.

The paper [6] implements a web usage Mining Intelligent System to provide taxonomy on user information based on transactional data by applying Data Mining algorithm and also offers a public service that is direct access of websites to the third party.

In paper [7] the authors James. Andrews, et al. discussed and apply a lightweight formal method for checking test results. The method assumes that the software under test writes a text log file and is then analyzed by a program to check for failures. They suggested a state-machine-based log file analyzer programs, describe a language and implementation based on that formalism. Report the application of log file analysis to test the random units. They described the results of experiments done by comparing the performance and effectiveness of random unit testing and checking log file analysis to other units testing procedures. Experimental results show LFA that use of analyzer is cost effective and using LFA for complex specifications requiring scores of state machines or hundreds of transactions could be accepted only on safety-critical projects.

In paper [8] the authors Yiyao Lu, et al presented an atomic annotation approach that first aligns the data units on a result page into different groups that is, same group have the same semantic data. Then each group can annotate it from different aspects and different aggregate to different annotations to predict a final annotation label for it. An annotation wrapper for search site is automatically constructed and it is used to annotate new result pages from the same web database. The experimental results show that each of annotators is useful and also capable of generating high quality annotation.

In paper [9] the Thient Thient Shwe, *et al* discusses on the mining of web access logs, web usage data. The framework composes defining the purpose that is a multipurpose analyzer, defining the ontology mapping based on websites, perform preprocessing step, Web usage based mining on frequent item set and proposed the algorithm and Naive Bayesian classifier. The framework is applied to predict for further depending on the current analysis outcomes. The paper also provides details for Website maintenance, Personalization systems, Pre-fetched systems, Resource Systems, Recommender system and web site analysts, etc..

In paper [10] the authors Patrice Buche, *et al*. presented the design of ONDINE system which allows loading and querying of a data warehouse, it is proposed by an Ontological and Terminological Resource (OTR). ONDINE system has been implemented through the development of @web system and development of MIEL++ software. The ONDINE is only software which allows one to simultaneously annotate accurately a data table with an OTR and perform the query process, comparing preferences expressed by the end-user with fuzzy annotations. ONDINE has been tested on three ways of applications are the microbial risk in food, chemical risk in food and aeronautics.

In paper [11] the authors Sheng Di, Member *et al*. proposes a fully distributed, Virtual Machine (VM)-multiplexing resource allocation scheme to manage decentralized resources. The approach presented here not only achieved maximize resource utilization using the Proportional Share Model (PSM), but also delivers adaptively optimal execution efficiency. Self-Organizing Cloud (SOC) architecture presented in this paper acts as both producer and consumer resource. They showed the SOC where optimized algorithms can make an improvement by 15-60% n system throughput than a P2P grid model. The solution also exhibits high adaptability in a dynamic node-churning environment. The novel scheme proposed in this work for virtual resource allocation on a SOC, with three contributions such that Optimization of tasks resource allocation under users budget, with low contention maximized resource utilization based on PSM and Lightweight resource query protocol.

Authors Ekaterina Olshannikova, Aleksandr Ometov, YevgeniKoucheryavy and Thomas Olsson [12] overviews the research issues and achievements in the fields of Big Data and its visualization tools and techniques. The paper provides a classification of existing data types, analytical methods, visualization tools, impacts of new technologies like, virtual reality display and augmented reality.

The author Remya Panicker [13] gave a short glimpse on Big Data technologies used to build Big Data infrastructure. The paper explores various possibilities to improve decision-making in critical development areas of crime and natural disaster etc. and also discusses various opportunities useful in policy making and decision making with use of Big Data.

Authors Jaseena K.U. and Julie M. David [14] had discussed about the challenges and solution of Big Data. The exponential growth of data due to explosion of social network sites, search and the new resource and gave an area for Big Data. Computational and scalability challenges introduce by data includes- storage bottleneck, noise accumulation, spurious correction and measurement error. The paper also discusses about Big Data tools like map reduces over Hadoop, HDFS(Hadoop Distributed File System).

Authors Raymond Kosala, Hendrik Blockeel [15] point outs some confusion regarding the usage of Web Mining. He also suggested three Web Mining categories and the relationship with related paradigms and research done on it. The work related to Information integration and Web warehouses, such as IR(Information Retrieval), AI(Artificial Intelligence) and machine learning is also discussed in this paper.

Authors Harshna, Navneet Kaur [16] discusses about Data Mining techniques that are being used for intrusion detection, their advantages and disadvantages and also discusses about applications and the issues.

Authors Rohit Pitre, Vijay Kolekar [17] generalizes the meaning of Big Data and discusses about challenging issues, related works and the challenges for future.

Authors Raj Kumar, Dr. Rajesh Verma [18] proposed classification algorithms and explains about classification algorithms like C4.5, k-nearest neighbor classifier, Naive Bayes, SVM(Support Vector Machines), Apriori and AdaBoost, etc.

Authors Fabricio Voznika, Leonardo Viana [19], discusses about hidden patterns and automatic methods used for pattern extraction.

Authors Shakir Khan, Dr. Arun Sharma, Abu Sarwar Zamani, Ali Akhtar [20] has discussed about Data Mining applications in safety measures and also outlines research on confidentiality and Data Mining.

Authors Neelamadhab Padhy, Dr. Pragnyaban Mishra and RasmitaPanigrahi [21] have discussed a variety of techniques, approaches on important fields of Data Mining technologies. They have also discussed about data ware housing and scope of Data Mining. They have even discussed about Data Mining applications, classification algorithms and significance of evolutionary computing.

In this paper, Elisa Bertino [22] describes important research directions about Big Data - Opportunities and Challenges and includes some questions that have been debated by the panel.

[23] Here the authors Aarti Sharma, Rahul Sharma, Vivek Kr. Sharma, VishalShrivatava explains about an Application of Data Mining and its various techniques used in various fields. Based on data, the paper indicates that the decision tree provides details about various symptoms of diseases to decide future trends to find the patterns in the medical field.

The authors Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule [24] describes about Big Data. It requires a set of new techniques for integration to uncover largely hidden values from large datasets. Hadoop is open-source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers. Hadoop Map Reduce is a large scale, open source software framework which breaks up large data into smaller parallelizable chunks and handles scheduling. For incremental processing for Big Data applications, minimum changes are made for original MapReduce framework.

[25] Here the authors Nikita Jain, Vishal Srivastava describes about the key technology, which is surveyed to achieve the Data Mining on Neural Network and Genetic Algorithm. A formal review of the area of rule extraction from ANN (Artificial Neural Network, suitable for solving the problems of Data Mining and its applications) and GA(Genetic Algorithm) is also explained.

In this paper authors, Manjith B.C. and Shijin C.S. [26] explains about Mobility Prediction for Delay Reduction in WLAN(Wireless Local Area Network), which uses Location Tracking and Data Mining. To reduces the handoff delay, a system is proposed called Predictive Mobility Management Scheme (PMMS) which uses mobility prediction and Data Mining to optimize the handoff delays through pre-scanning, pre-authentication and pre-reassociations.

Authors D. Jayalatchumy, Dr. P. Thambidurai [27] K. R. Suneetha, Dr. R. Krishnamoorthi [28] discussed about the existing techniques of Web Mining and its issues. The paper reports the summary of current techniques of Web Mining and its classification. Related issues and drawbacks of existing techniques have been discussed here. Finally the paper concludes saying the importance of Semantic Web Mining in overcoming the cons of Web Mining.

Multiplexer:

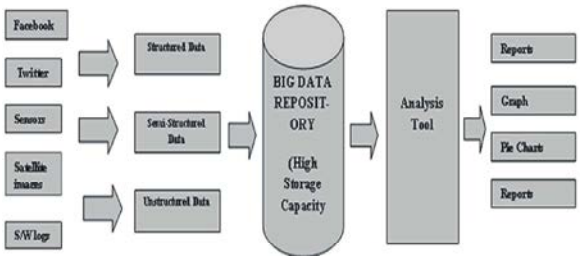

Fig. 1: [13] Big Data Analytics Block Diagram

Big Data analytics refers to tools and methodologies transform massive quantities of raw data into "Data about the Data". Analysis of these data allows us to discover the relation, facts and other important information that lies in this large data set. Data from various resource are collected, refined and stored under uniform schemes. Which is then analyzed into the form of reports, graphs, spatial charts, pie diagrams, etc. Social media can provide information on emerging concern and pattern at the local level which is relevant to development.

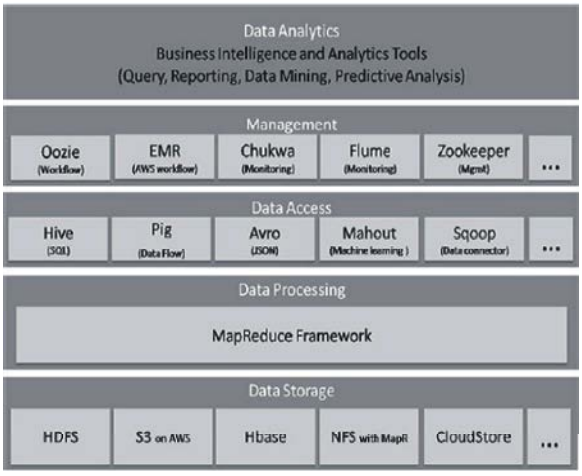

Fig. 2: [14] Big data analysis tools

The above figure is about Big Data analysis tools, where it is divided mainly into Data Analytics, Management, Data Access, Data processes, Data Storage. In Data Analytics, iBusiness, intelligence and analytics tools are developed, which simplifies workflow, monitoring and coordination of the tasks such as Oozie, EMR, Chukwa, etc. Data access tools are developed for defining data types, protocols as data connection, etc. In data processing, large data sets with a parallel, distributed algorithms on a cluster. To store the data many Data Storage Tools such as HDFS, S3, etc. are developed.

CONCLUSIONS

The paper reviews the work related to Data Mining concepts and Big Data. The amounts of data are 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. This paper glimpses the various properties of Big Data, interesting applications, followed by its architecture. Big Data challenge is becoming one of the most exciting opportunities in the upcoming years and has set its stand in this technological era.

REFERENCES

1. Hongjun Lu, Member, IEEE Computer Society, RudySetiono and Huan Liu, 1996. “Effective Data Mining Using Neural Networks”, IEEE Transactions on Knowledge and Data Engineering, 8(6): 957-961.
2. Hameetha Begum, S., 2013.. “Data Mining Tools and Trends - An Overview”, International Journal of Emerging Research in Management & Technology ISSN: 2278-9359, pp: 6-12, I.S. Jacobs and C.P. Bean, 1963. “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, pp: 271-350.
3. Anand V. Saurkar, Vaibhav Bhujade and Priti Bhagat Amit Khaparde, 2014. “A Review Paper on Various Data Mining Techniques”, 4(4) ISSN: 2277 128X International Journal of Advanced Research in computer Science and Software Engineering, pp: 98-101.
4. Philip K. Chan, Florida Institute of Technology, Wei Fan Andreas L. Prodromidis and Salvatore J. Stolfo, 1999. Columbia University, ” Distributed Data Mining in Credit Card Fraud Detection”, 1094-7167/99/$10.00 © 1999 IEEE, pp: 67-73.
5. Zeljko Eremic, Dragical Radosav and Branko Markoski, 2010. “Mining user Logs to Optimize Navigational Structure of Adaptive Web Sites”, 11th IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary, pp: 271-275, 18-20 November, 2010.
6. Naveena Devi, B., Y. Rama Devi, B. Padmaja Rani, R. Rajeshwarrao, 2012. “Design and Implementation of Web Usage Mining Intelligence System in the field of e-commerce”, International Conference on Communication technology and System Design, pp: 20-27.
7. James H. Andrew and Yingjum Zhang, 2003. “General Test Result Checking with log File Analysis”, IEEE Transaction on Software Engineering, 29: 634-648.
8. Yiyo Lu, Hai He, Hongkun Zhao, Weiyi Meng and Clement Yu, 2013. “Annotating Search Results from Web databases”, IEEE Transaction on Knowledge and Data Engineering, 25(3): 514-527.
9. Thient ThientShwe, Thient Myint, Thient Thient Aye, Su Su Htay, Swe Swe Nyein and Mie Mie Su Thwin, 2010. “FrameWork for Multi-Purpose Web log access Analyzer”, IEEE 2nd International Conference on Computer Engineering and Technology, pp: v3.289-v3.293.
10. Patrice Buche, Julette Dibie-Barthelemy, Liliana Ibanescu and Lydie Soler, 2013. “Fuzzy Web data tables Integration Guided by an Ontological and Terminological resource”, IEEE Transaction on Knowledge and Data Engineering, 25(4): 805-814.
11. Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng and Clement Yu, 2013. “Dynamic Optimization of Multiattribute Resource Allocation in Self-Organizing Clouds”, IEEE Transaction on Parallel and Distributed Systems, 24(3): 464-476.
12. Ekaterina Olshannikova, Aleksandr Ometov, Yevgeni Koucheryavy and Thomas Olsson, 2015. "Visualizing Big Data with augmented and virtual reality: challenges and research agenda "Olshannikova *et al*. Journal of Big Data, 2: 22.
13. Remya Panicker, 2013. "Adoption of Big Data Technology for the Development of Developing Countries" Proceedings of National Conference on New Horizons in IT - NCNHIT.
14. Jaseena, K.U. and Julie M. David, "Issues, Challenges and Solutions:Big Data Mining".
15. Raymond Kosala, Hendrik Blockeel, " Web Mining Research: A Survey".
16. Harshna, Navneet Kaur, 2013. "Survey paper on Data Mining techniques of Intrusion Detection "International Journal of Science, Engineering and Technology Research (IJSETR), 2(4).
17. Rohit Pitre, Vijay Kolekar, 2014. "A Survey Paper on Data Mining With Big Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1(1).
18. Raj Kumar, Dr. Rajesh Verma, "Classification Algorithms for Data Mining: A Survey", International Journal ofInnovations in Engineering and Technology (IJIET), pp: 1.

19. Fabricio Voznika, Leonardo Viana, "Data Mining Classification".
20. Shakir Khan, Dr. Arun Sharma, Abu Sarwar Zamani and Ali Akhtar, 2012. " Data Mining For Security Purpose & Its Solitude Suggestions"international Journal of Scientific & Technology Research, 1(7).
21. Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, 2012. " The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(3).
22. Elisa Bertino, 2013. "Big Data - Opportunities and Challenges", IEEE 37th Annual Computer Software and Applications Conference.
23. Aarti Sharma, Rahul Sharma, Vivek Kr. Sharma, Vishal Shrivatava, 2014. "Application of Data Mining - A Survey Paper" (IJCSIT) International Journal of Computer Science and Information Technologies, 5(2): 2023-2025.
24. Ms. Vibhavari Chavan, Prof. Rajesh N. Phursule, 2014. "Survey Paper On Big Data" (IJCSIT) International Journal of Computer Science and Information Technologies, 5(6): 7932-7939.
25. Nikita Jain, Vishal Srivastava, " Data Mining Techniques: A Survey Paper"IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.
26. Manjith, B.C. and C.S. Shijin, 2012. "Mobility Prediction for Delay Reduction in WLAN using Location Tracking and Data Mining" International Journal of Computer Applications (0975-8887), 52(21).
27. Jayalatchumy, D. and Dr. P. Thambidurai, 2013. "Web Mining Research Issues and Future Directions - A Survey"IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727, 14(3): 20-27.
28. Suneetha, K.R. and Dr. R. Krishnamoorthi, 2009. "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, 9(4).