International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Survey on Data Mining and Ontology for handling Big Data

[1]Mr. N. Suresh Kumar, [2]Dr. M. Thangamani

[1]Assistant Professor (Selection Grade), Sri Ramakrishna Engineering College, Coimbatore, nsuresh2@gmail.com
[2]Assistant Professor, Kongu Engineering College, Perundurai-638052, manithangamani2@gmail.com

Abstract: *Big data is large volume, heterogeneous, distributed data and now rapidly expanding in all science and engineering domains. Big Data mining is the ability of extracting useful information from large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. With increasing size of data in data warehouse it is expensive to perform data analysis. This survey includes the information about handling big data with data mining Data.*

Keywords: Data Mining, Big data and Clustering.

I. Introduction

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. Big data is the data that exceeds the processing capacity of conventional database systems. Big data is very big, moves very fast, or doesn't fit into traditional database architectures. There are different definitions of big data as it is more often used as an all-encompassing term for everything from actual data sets to big data technology and big data analytics.

There are mainly 3 types of big data sets- structured, semi structured and unstructured. 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 does not conform to an explicit and fixed schema.

Document clustering is the use of cluster analysis of textual documents. It has many applications like organizing large document collection, finding similar documents, recommendation system, duplicate content detection, search optimization. Document clustering has been considered for use in a number of different areas of text mining and information retrieval. There are many search engines that are using for information retrieval, but the main challenge in front of the search engine is to present relevant results of the user. Even though there are many knowledge discovery tools to filter, order, classify or cluster their search results exists, still user make extra effects to find the required document. In order to provide solution, combining the entire web mining based data mining techniques. Now a day's huge data is producing by many social networking websites, e-commerce websites, and many organizations. Analysing this huge data is tedious task for any organization. To analyse the huge data, database management techniques may not be sufficient, so the big data came into existence.

II. Related work

Big data is redefining the data management from extraction, transformation and processing to cleaning and reducing. Currently Big Data processing depends upon parallel programming models like MapReduce, as well as providing computing platform of Big Data services. Data mining algorithms need to scan through the training data for obtaining the statistics for solving or optimizing model parameter [1].

Recent studies done on Big Data Analytics in the field of Health Informatics which sought to answer various clinical questions, using data acquired from the molecular, tissue, and patient levels of Health Informatics [2]. Modern health information systems can generate several Exabyte's of patient data, the so called 'health big data [3]', per year. Many health managers and experts believe that with the data, it is possible to

easily discover useful knowledge to improve health policies, increase patient safety and eliminate redundancies and unnecessary costs. Information Retrieval is a large and growing field, Search engines like Google is used by many people to retrieve data or to return thousands of related web pages. Clustering can be used to group search results into small number of clusters, each of which captures a particular aspect of the query [4]. For unstructured query 20 news group datasets are used for evaluation [5-6]

Document clustering is one of the important areas in data mining. Document Clustering is the task of grouping a set of documents in such a way that objects in the same group are more similar to each other than to those in other groups clusters. Datta et al 2006 proposed an exact local approach for handling a K-means clustering, in addition to an approximate local K-means clustering approach for P2P networks. The P2P K-means approach has its fundamental in a parallel implementation of K-means by Dhillon and Modha [7]. Though, K-means monitoring algorithm does not generate a distributed clustering, it assists a centralized K-means process and recompute the clusters by monitoring the distribution of centroids across peers, and triggering a re-clustering if the data distribution changes over time. Alternatively, the P2P K-means approach works by updating the centroids at each peer depending on information obtained from their immediate neighbors. The algorithm stops when the information obtained does not result in considerable update to the centroids of all peers.

The main issue in the existing clustering approaches is that the efficiency and accuracy of the clustering results is minimized with the increase of the network size. A novel text clustering approach proposed by Qing He et al [8] depends on frequent term sets for P2P networks. It needs relatively lower communication volume while attaining a clustering result whose quality is not affected by the size of the network. Furthermore, it provides a term set describing each cluster, which makes the user to have a clear comprehension for the clustering result, and assists the users to find resource in the network or manage the local documents in accordance with the whole network.

Zhongjun Deng et al [9] examined clustering algorithm in P2P network. Conventional clustering approaches cannot be applied to P2P systems because of its lack of central control and very large size. Jia Zhen and Wang Yong Gui [10] presents a genetic clustering approach depending on dynamic granularity. From the perspective of a parallel, random search, global optimization and diversity features of genetic algorithm (GA), it is integrated with dynamic granularity approach. In the process of granularity changing, suitable granulation can be made by refining the granularity, which can improve the efficiency and the accuracy of the clustering algorithm. From the experimental results, it is observed that the approach enhances the clustering algorithm based on GA local search ability and convergence speed.

Internet made it easy to access huge collection of information across the world. This opportunity has encouraged an increasing demand for understanding how to combine multiple and heterogeneous information sources. This research mainly focuses on recognistion and integration of the appropriate information to provide a better knowledge on a specific domain. The combination is predominantly helpful when it allows the communication between dissimilar sources without affecting their autonomy. The difficulty in combining heterogeneous resources has been dealt in the literature. Only few researchers concentrated on heterogeneity between resources other than databases.

To deal with the different information resources, necessitates the formation of few techniques. Hence, Ontologies have been formulated to make information sharing and reuse the various data in all areas and tasks [11]. The key task is to develop the differences in semantics explicit. An ontology is defined as a logical theory accounting for the intended meaning of a formal vocabulary, specifically its ontological commitment to a specific conceptualisation of the world [12]. Ontologies offers a generally agreed understanding of a domain that can be reused and shared across several applications. Among the different techniques for the combination of heterogeneous sources presented in the literature, the main attention is on the structure of multiple shared ontologies which is suggested [13]. This architecture aggregates multiple shared ontologies into clusters, so as to obtain a structure that is able to reconcile different types of heterogeneity and is also intended to be more convenient to implement and give better prospects for maintenance and scaling. Furthermore, such a structure is thought to avoid information loss when performing translations between diverse resources.

Fuzzy ontologies are very much useful in the Web. Ontologies serve as fundamental semantic infrastructure, offering shared understanding of certain domain across different applications, so as to assist machine understanding of Web resources. Moreover, Ontology can also handle fuzzy and imprecise information which is very important to the Web [14]. Semantic Web has been criticized for not addressing uncertainty. In order to provide solution for addressing semantic meaning in an uncertain and inconsistent world, fuzzy ontologies [15] have been proposed. When

using fuzzy logic, reasoning is approximate rather than precise. The main purpose is to avoid the theoretical pitfalls of monolithic ontologies, assist interoperability between different and independent ontologies [16], and offer flexible information retrieval competence [17]

The authors Jiawei Han, Micheline Kamber, they have presented various clustering Techniques for Data Mining. They have published how to use K-means clustering algorithm to cluster large data set in various disjoint clusters [18]. For distributed text clustering, fuzzy ontology is proposed [19] for handling big data.

Hadoop handle the big data. Hadoop is being used by the Yahoo, Google, Face book and Twitter business companies for implementing real time applications. Email, social media blog, movie review comments, books are used for document clustering. Clustering of class labels can be generated automatically, which is much lower quality than labels specified by human. If the class labels for clustering are provided, the clustering is more effective. In classic document clustering based on vector model, documents appear terms frequency without considering the semantic information of each document. The property of vector model may be incorrectly classified documents into different clusters when documents of same cluster lack the shared terms. To overcome this problem are applied by the knowledge based approaches. This approach uses a similarity between the class label terms and term weights to improve the quality of the document clustering. It also can cluster the big data size of document using the distributed parallel processing based on Hadoop [20]. The Hadoop Distributed File System paper focused on the Hadoop implementation using single node implementation as well as multimode implementation details [21, 22].

Big data is getting more attention in today's world. Although MapReduce is successful in processing big data, it has some performance bottlenecks when deployed in cloud. Data locality has an important role among them. Good data locality reduces cross network traffic and hence results in high performance [23]. Big data usually exists in cyberspace as the form of the data stream. It brings great benefits for information society. Meanwhile, it also brings crucial challenges on big data mining in the data stream [24].

References

[1] Shital Suryawanshi1, Prof. V.S.Wadne2 , Big Data Mining using Map Reduce: A Survey Paper , IOSR Journal of Computer Engineering, Volume 16, Issue 6, Ver. VII (Nov – Dec. 2014), PP 37-40.

[2] Herland, M. ; Florida Atlantic Univ., Boca Raton, FL, USA ; Khoshgoftaar, T.M. ; Wald, R., Survey of Clinical Data Mining Applications on Big Data in Health Informatics, IEEE, Machine Learning and Applications (ICMLA), 2013 12th International Conference , vol 2 pp. 465 – 472.

[3] Mu-Hsing Kuo; Tony Sahama; Andre W. Kushniruk; Elizabeth M. Borycki; Daniel K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International journal of big data intelligence, interscience, Vol. 4, issue 8, pp.114 – 126.

[4] A. Ngoumou and M. F. Ndjodo, "A Rewriting System Based Operational Semantics for the feature Oriented Resue Method" , International Journal of Software Engineering and Its Applications, vol. 7, no. 6, (2013), pp. 41-60.

[5] S. Mal and K. Rajnish, New Quality Inheritance Metrics for Object-Oriented Design", International Journal of Software Engineering and Its Applications, vol. 7, no. 6, (2013) , pp.185-200.

[6] S. H. Lee, J. G. Lee and K. I. Moon, "A preprocessing of Rough Sets Based on Attribute Variation Minimization" , International Journal of Software Engineering and Its Applications, vol. 7, no. 6,(2013), pp. 411-424.

[7] Dhillon .I.S and Modha .D.S (2001), "Concept decompositions for large sparse text data using clustering", Machine Learning, Vol. 42, pp. 143-175.

[8] Qing He, Tingting Li, Fuzhen Zhuang and Zhongzhi Shi (2010), "Frequent term based peer-to-peer text clustering", 3rd International Symposium on Knowledge Acquisition and Modeling (KAM), pp. 352–355.

[9] Zhongjun Deng, Wei Song and Xuefeng Zheng (2010), "P2PKMM: A Hybrid Clustering Algorithm over P2P Network", Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), pp. 450–454.

[10] Jia Zhen and Wang Yong Gui (2010), "Genetic Clustering Algorithm Based on Dynamic Granularity", International Conference on Computing, Control and Industrial Engineering (CCIE), Vol. 1, pp. 431-434.

[11] Gomez Perez .A and Benjamins .V.R (1999), "Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods", In Proceedings of the IJCAI'99 Workshop on Ontology and Problem-Solving Methods: Lesson learned and Future Trends, Benjamins, V.R., (Ed.), CEUR Publications, Amsterdam, Vol. 18, pp. 1.1-1.15.

[12] Guarino .N (1998), "Formal Ontology and Information Systems", In Proceedings of Conference on Formal Ontology (FOIS'98), Guarino .N (Ed.), Trento.

[13] Visser .P.R.S and Tamma .V.A.M (1999), "An Experience with Ontology-based Agent Clustering", In Proceedings of the IJCAI'99 Workshop on Ontology and Problem-Solving Methods: Lesson learned and Future Trends, Benjamins .V.R (Ed.),

CEUR Publications, Amsterdam, Vol. 18, pp. 12.1-12.13.

[14] shridharan .B, Tretiakov .A and Kinshuk (2004), "Application of ontology to knowledge management in web based learning", In IEEE International Conference on Advanced Learning Technologies, pp. 663 – 665.

[15] Amy J.C Trappey, Charles V. Trappey, Fu-Chiang Hsu, and David W. Hsiao (2009), "A Fuzzy Ontological Knowledge Document Clustering Methodology", IEEE Transactions on System, Man and Cybernetics-Part B, Vol. 39, No. 3.

[16] Cross .V (2004), "Fuzzy Semantic Distance Measures between Ontological Concepts", IEEE Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2004), Banff, Alberta, Canada, pp. 27–30.

[17] Thomas .C and Sheth .A (2006), "On the Expressiveness of the Languages for the Semantic Web – Making a Case for 'A Little More", Fuzzy Logic and the Semantic Web (Capturing Intelligence), Sanchez .E (Ed.), Elsevier.

[18] Jiawei Han, MichelineKamber,"Data Mining Concepts and Techniques", Second Edition, Morgan Kaufman Publishers, 2006 Elsevier In.

[19] Dr. M. Thangamani, A. Senthil karthick kumar, A. M. J. Mohamed zubair Rahman, Bio-Inspired Fuzzy Expert system for Mining Big data, Mathematical and Computational Methods in Science and Engineering, pp. 223-227, 2014

[20] Yong-Il Kim1, Yoo-Kang Ji2 andSun Park3, Big Text Data Clustering using Class Labels and Semantic Feature Based on Hadoop of Cloud Computing, International Journal of Software Engineering and Its Applications, Vol.8, No. 4 (2014), pp. 1-10.

[21] Tom White, "Hadoop: The Definitive Guide", First Edition, Published by Reilly Media, June 2009, in United States of America.ao,Tao Li , IEEE transaction,2012.

[22] Konstantin Shvachko, HairongKuang, "The Hadoop Distributed File System", Published by IEEE, 2010.

[23] Somesh S Chavadi1, Dr. Asha T2, Text Mining Approach for Big Data Analysis Using Clustering and Classification Methodologies, International Journal of Emerging Technology and Advanced Engineering, Vol. 4, issue 8, 2014.

[24] Yi Wu, Network Big Data: A Literature Survey on Stream Data Mining, Journal of Software, vol. 9, no. 9, 2014, pp-2427-2434

Author Profile

Mr. N.Suresh Kumar is currently working as Assistant Professor in the Department of Information Technology at Sri Ramakrishna Engineering College, Coimbatore. He has a Master's degree in Software Engineering (2010) at Anna University of Technology, Coimbatore. He has nearly 11 years of academic experience and 2 Years of Industry experience. He is proficient in Data Mining, Software Engineering, Database Management System and Object Oriented Analysis and Design. He has published the papers in International and National Journals and presented the papers in National conferences. He is a life member of ISTE, IAENG and IACSIT.

Dr. M. Thangamani is nearly 20 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published 50 articles in International journals and presented over 67 papers in national and international conferences in above field. She has delivered more than 35 Guest Lectures in reputed engineering colleges on various topics. She has organized many self supporting and sponsored national conference and Workshop in the field of data mining, big data and cloud computing. She is on the editorial board and reviewing committee of leading research journals, and on the program committee of top international data mining and soft computing conferences in various countries. She also seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She has organizing chair and keynote speaker in international conferences in India and abroad. She is currently working as Assistant Professor in Engineering at Kongu Engineering College.