Research Journal of Applied Sciences, Engineering and Technology 8(8): 1002-1008, 2014
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: May 09, 2014 Accepted: August 03, 2014 Published: August 25, 2014

Mining Association Rules in Big Data for E-healthcare Information System

N. Rajkumar, R. Vimal Karthick, M. Nathiya and K. Silambarasan
Department of CSE, Vel Tech Rangarajan Dr. Sagunthala R and D Institute of Science and Technology, Vel Tech Dr. RR and Dr. SR Technical University, Chennai-62, Tamil Nadu, India

Abstract: Big data related to large volume, multiple ways of growing data sets and autonomous sources. Now the big data is quickly enlarged in many advanced domains, because of rapid growth in networking and data collection. The study is defining the E-Healthcare Information System, which needs to make logical and structural method of approaching the knowledge. And also effectually preparing and controlling the data generated during the diagnosis activities of medical application through sharing information among E-Healthcare Information System devices. The main objective is, A E-Healthcare Information System which is extensive, integrated knowledge system designed to control all the views of a hospital operation, such as medical data's, administrative, financial, legal information's and the corresponding service processing. At last the analysis of result will be generated using Association Mining Techniques which processed from big data of hospital information datasets. Finally mining techniques result could be evaluated in terms of accuracy, precision, recall and positive rate.

Keywords: Association rule mining, autonomous sources, e-healthcare information system, information sharing, medical diagnosis, medical knowledge

INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis that is extracting of interesting patterns or knowledge from huge amount of data. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future actions. Data mining is also known as Knowledge Discovery in Data (KDD).

Data mining has been popularly treated as a synonym of knowledge discovery in databases; a knowledge discovery process consists of Data cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation. A Data Mining system may accomplish one or more of the following tasks such as Classification, Clustering, Association rule and Regression (Koteeswaran and Kannan, 2012).

Research Challenges in Data Mining includes information network analysis, Discovery, usage and understanding of patterns and knowledge, Stream data mining, Mining moving object data, RFID data and data from sensor networks, Spatiotemporal and multimedia data mining, Mining text, Web and other unstructured data, Data cube-oriented multidimensional online analytical mining, Visual data mining, Data mining by integration of sophisticated scientific and engineering domain knowledge.

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. Big data is also refers to data sets or combinations of data sets whose size (volume), complexity (variability) and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. And big data may be as important to business and society as the Internet has become (Xindong *et al*., 2014).

When big data is effectively and efficiently captured, processed and analyzed, companies are able to gain a more complete understanding of their business, customers, products, competitors, etc. which can lead to efficiency improvements, increased sales, lower costs, better customer service and/or improved products and services.

It includes, understanding and utilizing Big Data, new, complex and continuously Emerging Technologies, cloud based solutions, privacy, security and regulatory considerations, archiving and disposal of Big Data, the Need for IT, data analyst and management resources.

LITERATURE REVIEW

The Big data processing is the summarization program of the information, which is from multiple,

Corresponding Author: N. Rajkumar, Department of CSE, Vel Tech Rangarajan Dr. Sagunthala R and D Institute of Science and Technology, Vel Tech Dr. RR and Dr. SR Technical University, Avadi, Chennai-62, Tamil Nadu, India

different, decentralized sources with complex and evolving relationships and keeps growing. An example for big data processing is the summarization process of various types of notions from different media applications or the comments those are receiving in the internet including updated suggestions.

The Big data era has arrived, along with the above example. Nearly 2.5 quintillion bytes of data are created in everyday and in the past two years 90% of the data were produced (IBM, 2012). Since the invention of the Information Technology in the early 19th century, our capability for data generation has never been so powerful and enormous. As another example is the conversation of the first president ship controversy between Barrack Obama and Mitt Romney, which touch off more than 10 million tweets within 2 h. Among these tweets, the specific moments revealed the public interests like discussions about Medicare and Vouchers. These online discussions generate feedback in real time manner and mostly compare it to generic media, such as TV or radio broadcasting (Punnagai *et al*., 2011).

Flicker, a picture sharing site is another example (Michel, 2012). It receives nearly 1.8 million photos per day. Imagine that the size of each photo is 2 megabytes, which results 3.6 terabytes storage every single day. So the pictures on Flicker are a treasure tank to explore the social events, affairs, disasters and so on, only if we have the power to trapping the enormous amount of data.

The rise of big data applications are illustrated in the above examples, that the data collection has grown tremendously and which is beyond the ability of commonly used software tools to manage and process within an endurable expired time. The basic challenge is to examine the high volume of data and extracting the useful knowledge for future actions. The knowledge extraction process has to be efficient and tight to real-time because storing all observed data is closely infeasible (Machanavajjhala and Reiter, 2012).

Integrating and mining of bio-data from various sources to decipher and make use of the structure of biological networks is illustrated in Integrating and Mining Bio-data from Multiple Sources in Biological Networks by NSF, 2013. This addresses the theoretical underpinnings and enables technologies for mining and integrating the biological networks. Thus the information acquisition and processing for information are achieved.

Real-time Classification of big data stream by Zhu *et al*. (2010) builds a big data analytic process framework for high fast response and decision making in real-time manner is proposed in Real time Classification of big data stream. It includes the issues are reducing big data volumes, constructing prediction models and creates framework to ensure data monitoring and classification.

Meanwhile the examination of the NP-complexity of the mining problems, various patter matching with wildcards and personalized information processing is performed in Pattern matching and Mining with wildcards and Length Constraints is examined by Patter Matching and Mining with Wildcards and Length Constraints by NSFC in January 2013.

Chu *et al*. (2006) in Map reduce for machine learning on multicore, describes the parallel programming in the multi core environment and multi processor systems supported by a Map Reduce based programming interface and it also realized data mining algorithms including k-means. Parallelization with Multiplicative Algorithms for big data mining by Birney (2012) suggests that applications of big data mainly focused on the privacy preserving approaches.

The evaluation of performance in single-pass learning, query based and iterative based learning (Koteeswaran and Kannan, 2013) in the framework is achieved by the map reduce implementation in hadoop, which describes how to share data between nodes which involved in parallel learning algorithms, how to react with the distributed storage information and finally shows that this mechanisms (Map reduce) is suitable for mining large scale data from the medium size clusters are explains in Map reduce based application programming interface Phoenix (Ranger *et al*., 2007).

In Challenges and Opportunities with Big Data by Labrinidis and Jagadish (2012) describes that the data which are stored in large amount do not have any security, privacy and also data's will be incomplete. The basic challenges in the big data applications is to examine the large volumes of data closely and gathering the useful knowledge for future works.

Reed *et al*. (2011) proposes a method called Square Kilometre Array which can store the information or data in particular limit and it is also incapable of handling these large volumes of data. So because of this, the novel data volumes need an effective analyzing of data and it requires platform which makes prediction to achieve fast response for such large volumes of data.

Chang *et al*. (2009) expanded the methods of data mining in different ways, which includes the efficient improvement for single source knowledge discovery and Wu and Zhang (2003) and Wu *et al*. (2005, 2013) had designed the data mining mechanism which is from a multi source perspective, the main objective of knowledge discovery is improving the efficiency of data. In a foundation for global knowledge discovery in multi source data mining, Su *et al*. (2006) proposed a local pattern analysis theory for finding global models.

CHALLENGES

Existing methods are incapable of handling the Big Data, such as SKA Square kilometre array which can store the data's to particular limit as a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast-

response and real-time classification for such Big Data. This is because different information collectors use their own schemata for data recording and the nature of different applications also results in diverse representations of the data. Another challenge is that the autonomous data sources with distributed and decentralized controls, being autonomous, each data sources is able to generate and collect information without involving (or relying on) any centralized control it leads to major confusion in existing system. And also the other challenge in complex and evolving relationships underneath the data, while the volume of the big data increases. An existing method are having some drawbacks such as Data security is not in complete form, Data can be stored in limited form only, Information cannot be shared, Knowledge sharing is also less and Data are stored in complex.

METHODOLOGY

In the proposed system, huge data's can be collected and stored in a big data in which for a small data can be formed in different scenario but it will be stored in the big data. It also performs heterogeneous and diverse dimensionality. This is because different information collectors use their own schemata for data recording and the nature of different applications also results in diverse representations of the data the heterogeneous features refer to the different types of representations for the same individuals and the diverse features refer to the variety of the features involved to represent each single observation it becomes a major challenge.

Next is Autonomous Sources with Distributed and Decentralized Control in this process large data collected will be stored in one big data. Data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data and may be used in further analysis.

The main objective, A Hospital information system is a comprehensive, integrated information system designed to manage all the aspects of a hospital operation, such as medical, administrative, financial and legal and the corresponding service processing and analysis of result will be generated using Association Mining Techniques which processed from big data of hospital information dataset. Finally mining techniques result could be evaluated in terms of accuracy, precision, recall, positive rate and etc.

Association rule mining: Association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness.

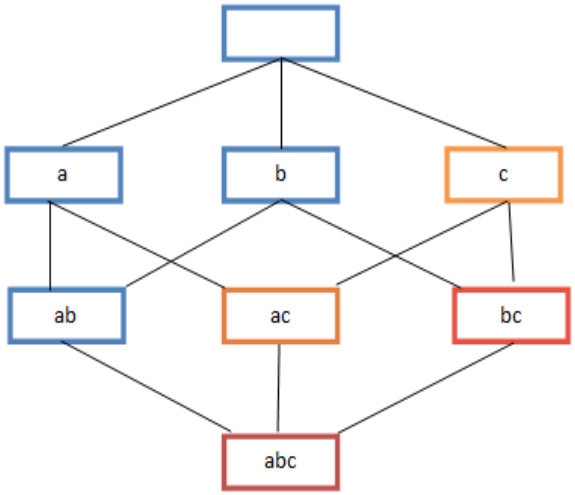

Fig. 1: Association rule mining analysis

Following the original definition, the problem of association rule mining is defined as: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short item sets) X and Y are called antecedent (Left-Hand-Side or LHS) and consequent (Right-Hand-Side or RHS) of the rule, respectively.

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

- First, minimum support is applied to find all frequent item sets in a database.
- Second, these frequent item sets and the minimum confidence constraint are used to form rules.

While the second step is straightforward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations) as shown in Fig. 1. The set of possible item sets is the power set over I and has size 2^n-1 (excluding the empty set which is not a valid item set). Although the size of the power set grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent item set, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent. Exploiting this property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent item sets as shown in Fig. 1.

EXPERIMENTAL EVALUATION

It includes to the overall structure of the system and the ways in which that structure provides conceptual integrity for a system. In a broader sense however components can be generalized to represent major system elements and their interaction.

A modular design reduces complexity, facilitates changes (a critical aspect of software maintainability) and results in easier implementation by encouraging parallel development of different part of system. Software with effective modularity is easier to develop because function may be compartmentalized and interfaces are simplified. Software architecture embodies modularity i.e., software is divided into separately named and addressable components called modules that are integrated to satisfy problem requirements as shown in Fig. 2.

Modularity is the single attribute of software that allows a program to be intellectually manageable. The five important criteria that enable us to evaluate a design method with respect to its ability to define an effective modular design are: Modular Decomposability, Modular Compensability, Modular Understandability, Modular Continuity and Modular Protection.

The following are the basic concepts which are planned in aid to complete the project with respect to the proposed system, while overcoming existing system and also providing the support for the future enhancement system.

Medical Information System (MIS): The MIS is a key component for the overall management of CG clinics and sickbays. MIS is a dynamic tool, which will provide a comprehensive electronic solution for tracking operational medical readiness, health systems management and patient access to care. Patient can access their medicine report in online itself which is developed by organization helping to promptly meet customers.

Distributed decentralization system: A distributed system is a software system in which components located on networked computers communicate and coordinate their actions by passing messages. Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data sources is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function.

Distributed computing also refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers.

Information sharing: Information sharing describes the exchange of data between various organizations, people and technologies. Information shared by individuals (such as medical information shared in Distributed server) Using information sharing intelligently has been shown to be a more effective way to manage any organization Information sharing is crucial to many businesses, helping to promptly meet customer and client needs through customer relationship systems which share information about medical report and services and improve access to their customers. Information sharing has also allowed easy availability of medical history details which helps consumers access more services. Customer can access their medical related information in distributed server which is shared by organization for customer's satisfaction.

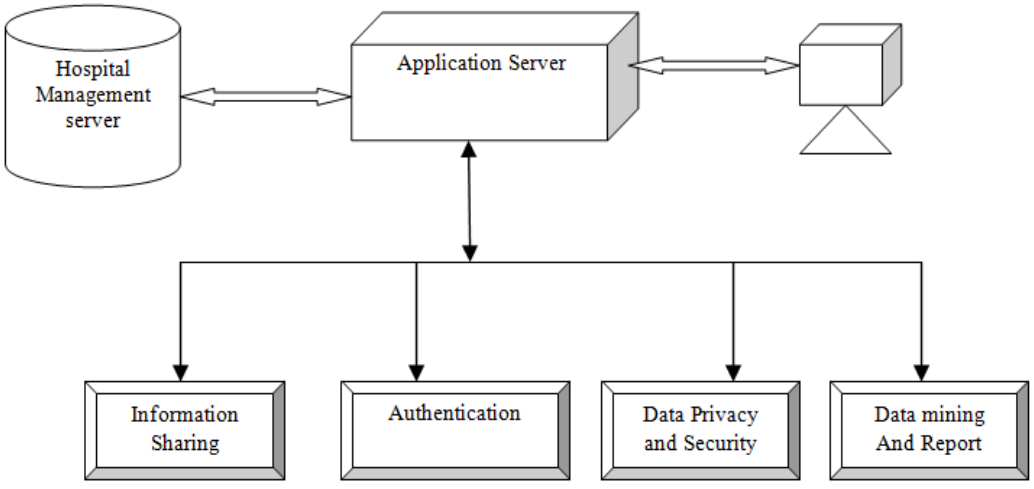

Fig. 2: Block diagram of the e-healthcare information system

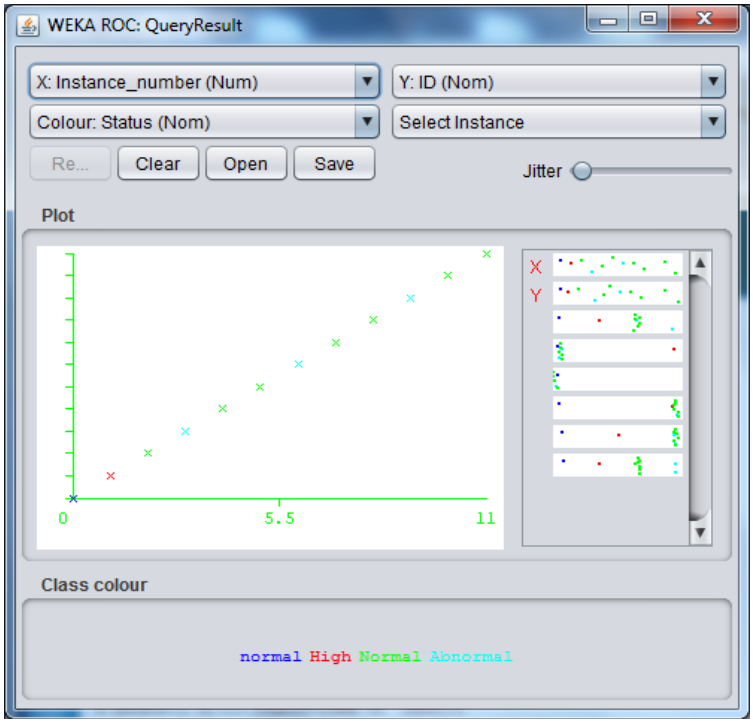

Fig. 3: Result analysis graph

Data privacy: Data privacy is to share data while protecting personally identifiable information. In this system medical related information could be shared in protection manner. Key management system is followed in medical information which is performed by implementing RSA algorithm and it is provide protection for original information. So customer can access their medical information from server in secured manner.

A person may not wish for their medical records to be revealed to others. This may be because they have concern that it might affect their insurance coverage's or employment. Or it may be because they would not wish for others to know about medical or psychological conditions or treatments which would be embarrassing. Revealing medical data could also reveal other details about one's personal life. Privacy Breach There are three major categories of medical privacy: informational (the degree of control over personal information), physical (the degree of physical inaccessibility to others) and psychological (the extent to which the doctor respects patients' cultural beliefs, inner thoughts, values, feelings and religious practices and allows them to make personal decisions). Physicians and psychiatrists in many cultures and countries have standards for doctor-patient relationships which include maintaining confidentiality. In some cases, the physician is legally protected.

Mining technique implementation: Data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data and may be used in further analysis or, for example, in machine learning and predictive analytics as shown in Fig. 3. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. The data collection, data preparation, or result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

RESULTS AND PERFORMANCE EVALUATION

Based on the instances or datasets which are classified as either correct or incorrect instances, from that the mean absolute error, root mean squared error, relative absolute and relative squared error are calculated. Finally by applying the algorithm, detailed accuracy of the instances by class is measured (Table 1).

The analysis of result will be generated using Association Mining Techniques which processed from big data of hospital information datasets (Fig. 4). Finally mining techniques result could be evaluated in terms of accuracy, precision, recall and positive rate (Fig. 5).

Table 1: Detailed accuracy by class

TP rate	FP rate	Precision	Recall	F-measures	ROC area	Class
1	0	1	1	1	1	Normal
1	0	1	1	1	1	High
1	0.60	0.70	1	0.824	0.614	Normal
0	0	0	0	0	0.352	Abnormal
WtdAvg	0.75	0.35	0.575	0.750	0.647	0.613

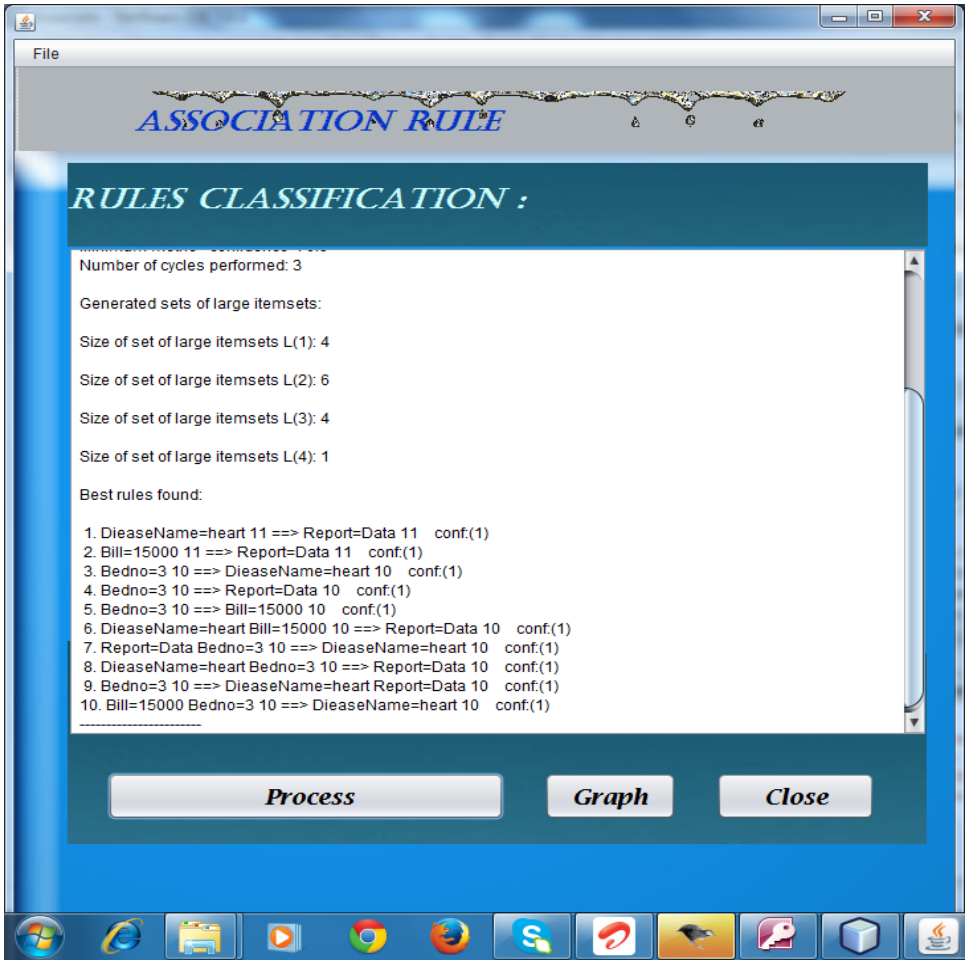

Fig. 4: Association rule classification

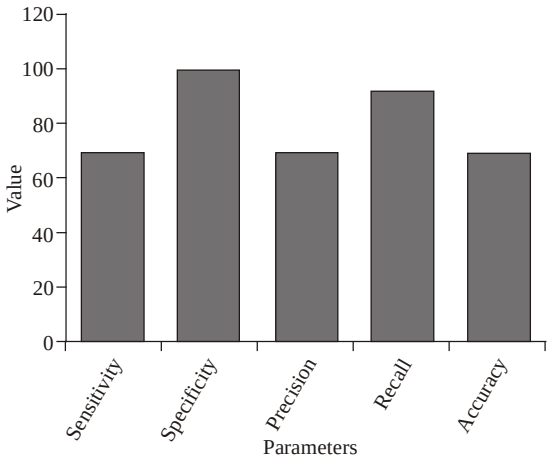

Fig. 5: Performance evaluation

Based on our proposal system, it achieves an effective data analysis for heterogeneous type of data and Fast response and real time classification for big data.

CONCLUSION

The main objective is, A Medical Information System which is extensive, integrated knowledge system designed to control all the views of a hospital operation, such as medical data's, administrative, financial, legal information's and the corresponding service processing. At last the analysis of result will be generated using Association Mining Techniques which processed from big data of hospital information datasets. Finally mining techniques result could be

evaluated in terms of accuracy, precision, recall and positive rate.

The term big data which related to data volumes, thus our theorem suggests the important characteristics of big data which needs a big mind for consolidate the data. And also analyzed the many challenges at the data, system and model level and the system is also effectually preparing and controlling the data generated during the diagnosis activities of medical application through sharing information among medical devices.

REFERENCES

Birney, E., 2012. The making of ENCODE: Lessons for big data projects. Nature, 489: 49-51.

Chang, E.Y., H. Bai and K. Zhu, 2009. Parallel algorithms for mining large-scale rich-media data. Proceedings of the 17th ACM International Conference on Multimedia (MM '09). New York, USA, pp: 917-918.

Chu, C.T., S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng and K. Olukotun, 2006. Map-reduce for machine learning on multicore. Proceeding of the 20th Annual Conference on Neural Information Processing Systems (NIPS'06), pp: 281-288.

IBM, 2012. What is Big Data: Bring Big Data to the Enterprise. IBM. Retrieved from: http://www-01.ibm.com/software/data/bigdata/.

Koteeswaran, S. and E. Kannan, 2012. Terrorist intrusion monitoring system using outlier analysis based search knight algorithm. Eur. J. Sci. Res., 74(3): 440-449.

Koteeswaran, S. and E. Kannan, 2013. Analysis of Bilateral Intelligence (ABI) for textual pattern learning. Inform. Technol. J., 12(4): 867-870.

Labrinidis, A. and H. Jagadish, 2012. Challenges and opportunities with big data. Proc. VLDB Endowment, 5(12): 2032-2033.

Machanavajjhala, A. and J.P. Reiter, 2012. Big privacy: Protecting confidentiality in big data. ACM Crossroads, 19(1): 20-23.

Michel, F., 2012. How Many Photos are Uploaded to Flickr Every Day and Month? Retrieved from: http://www.flickr.com/photos/franckmichel/6855169886/.

Punnagai, N., K. Ayarpadi, C. Leena and S. Koteeswaran, 2011. Simulation of broadcasting algorithm using neighbor information in mobile ad hoc networks. Proceeding of the IET International Conference on Sustainable Energy and Intelligent Systems (SEISCON, 2011). Tamilnadu, India, pp: 908-912.

Ranger, C., R. Raghuraman, A. Penmetsa, G. Bradski and C. Kozyrakis, 2007. Evaluating MapReduce for multi-core and multiprocessor systems. Proceeding of the 13th IEEE International Symposium on High Performance Computer Architecture (HPCA'07), pp: 13-24.

Reed, C., D. Thompson, W. Majid and K. Wagstaff, 2011. Real time machine learning to find fast transient radio anomalies: A semi-supervised approach combining detection and RFI excision. Proceeding of the International Astronomical Union Symposium on Time Domain Astronomy, UK.

Su, K., H. Huang, X. Wu and S. Zhang, 2006. A logical framework for identifying quality knowledge from different data sources. Decis. Support Syst., 42(3): 1673-1683.

Wu, X. and S. Zhang, 2003. Synthesizing high-frequency rules from different data sources. IEEE T. Knowl. Data En., 15(2): 353-367.

Wu, X., C. Zhang and S. Zhang, 2005. Database classification for multi-database mining. Inform. Syst., 30(1): 71-88.

Wu X., K. Yu, W. Ding, H. Wang and X. Zhu, 2013. Online feature selection with streaming features. IEEE T. Pattern Anal., 35(5): 1178-1192.

Xindong, W., Z. Xingquan, W. Gong-Qing and D. Wei, 2014. Data mining with big data. IEEE T. Knowl. Data En., 26(1): 97-107.

Zhu, X., P. Zhang, X. Lin and Y. Shi, 2010. Active learning from stream data using optimal weight classifier ensemble. IEEE T. Syst. Man Cy. B, 40(6): 1607-1621.