Journal of Mathematical and Computational Science
Vol. 1, Issue 1, June 2016

The Complexity of Extracting Knowledge in Big Data

Rozianiwati Bt Yusof[1*], Maznie Bt Manaf[2], Nor Asma Bt Mohd Zin[3], Marhainis Jamaludin[4]

[1,2,3,4] Faculty of Computer Science and Mathematic, Universiti Teknologi MARA Kelantan, Bukit Ilmu, Machang, Kelantan, Malaysia
rozian696@kelantan.uitm.edu.my
[*]Corresponding author

Abstract: Big data is extremely large data set data in the range of exabytes and the volume of data cannot be processed efficiently with the traditional technology in term of storing, manage and process. With the technologies appear, big data has attracted many researchers to extract knowledge from it. The knowledge can be produced in making a decision. However, there are some issues existed in extracting the important knowledge such as in storing the data, managing the data and processing the data in extracting useful information since its relate with volume, velocity and variety. Therefore, this paper is attempting to list all the possible knowledge that can be extracted from big data as well as discuss the previous researches in knowledge extraction from the huge amount of data. The problem in extracting important knowledge will be examined thoroughly and by identifying the significant problems in knowledge extraction the best knowledge from big data could be revealed. The techniques in analysing big data will be also discussed in the next section.

Keywords: big data, , extraction, knowledge, velocity, , variable, variety

1 Introduction

An issue of Big data have been recently triggered since the tremendous growth of data from technologies such as internet traffic (e.g., clickstreams), mobile transactions, user- generated content, and social media as well as purposefully captured content through sensor networks, business transactions, and many other operational domains such as bioinformatics, healthcare, and finance [1]. However, without an interpretation, a big data is nonsense. With the technologies appear, a big date has attracted many researchers to analyze and extract useful information in making a decision and solve real life problems.

Big data is an extremely large data set data in the range of exabytes and the volume of data cannot be processed efficiently with the current technology in term of storing, manage and process [2]. Its mean that the data set is too large for traditional data-processing systems and it required new technologies. As with the traditional technologies, big data technologies are used for any tasks, including data engineering. Occasionally, big data technologies are actually used for implementing data mining techniques, but more often the well-known big data technologies are used for data processing in support of the data-mining techniques [3].

The definition of big data consists of volume, velocity and variety where the main one is volume [4]. Many factors contribute to the increase in data volume such as transaction-based data stored through the years, data streaming in from social media, increasing amounts of sensor and machine-to-machine data being collected [5]. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

Meanwhile velocity is related to the speed of data is streaming, creation and aggregation and must be dealt with in a timely manner [6]. For example is e-commerce, RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. It has increased the speed and the loading of data in transactions. Reacting quickly enough to deal with data velocity is a challenge for most organizations [2].

Variety is representing the type of data. Data comes in many ways and formats such as image, text, video, audio, document, email, transaction and others. So the diversity of format gives an obstacle to merge and store, manage and process the big data in efficient ways [2].

In a big data, there are three common issues existed such in storing, managing and processing. Since the data are very big and can increase the size and become really big data, cause of technology evolution, the storing of big data becomes a crucial problem. Based on [2], current disc can store about four terabytes per disc. However, big data is in Exabyte where one exabyte need about 2500 discs. Furthermore, it can cause a problem to process the data in single computer and the transferring process can take longer time. Besides that, the storage used to store big data should be more flexible to grow since the data scaling growing rapidly [7]. The variety of data can give more complexity in storing the data because some of the data are structured and some of them are unstructured [5]. So that big data is stored in heterogeneous and different-in-nature data sources for examples are legacy systems, Web, scientific data repositories, sensor and stream databases and social networks into a structured, hence well interpretable format [3].

Besides storing problem, managing the big data also one of the interest issue in the past research. Management is the organization, administration and governance of large volumes of data whether structured or unstructured format [8]. This process is purposely to ensure the data collected are very high quality and accessibility for analytics process. Most of big data storing environment are beyond relational databases and traditional data warehouse where it aggregates with the new technologies. So, it is important in managing big data to find the quality data from the huge amount of data [2]. Furthermore, it must have a good technique to choose what data must be kept and what data can be disposed. However, from the volume of data, it is impractical to validate the data all the times since the data rapidly growing.

The next problem should be considered by researchers is processing issues. The processing issue should consider about data volatile where how long the data can be stored in RAM for accessing the data in doing some analyses towards data. Besides that, the efficient processing of exabytes of data will require a high technology in parallel processing and new analytic processing algorithms to make sure processing time is very effective and actionable information [9].

The analyze process in big data is called analytic process. Analytic is a process to extract value or knowledge from data. Accuracy in big data can lead more confident in decision making. However, in a big data, it consists of volume, velocity and variety. So the problems will trigger in extracting knowledge from big data.There are some past research had already discussed about big data in many areas. The researchers had focused on issues of storing, managing and processing. Most of them using new technologies combine with the traditional technologies in handling big data. There are some common techniques widely used in analysing big data such as association rules learning, genetic algorithm, machine learning, regression analysis, sentiment analysis and sosial network analysis. All of these techniques purposely to extract knowledge from the big data.

The remaining part of this article is organized as follows; section 2 discuss about the problem existed in extracting knowledge. In section 3, the past research in big data will be discussed. Section 4 discuss about the techniques widely used in analysing big data and finally brief the conclusions and future work in section 5.

2 Problem In Extracting Knowledge From Big Data

There are many problems in handling big data. However, the complexity is to analyze and find the useful knowledge and hidden information in the big data. The knowledge produce can be used to make a decision. In producing the knowledge, there are some issues existed in extracting knowledge from big data since the processing problem is related with the storing and management issues. The problem produces also come from the way of storing and manage the big data. The process of extracting useful knowledge from big data is called analytic. Analytics is the complex process and

procedures because these processes involve in large scale and enormous size of data repositories [7]. There are many problems existed in extracting knowledge from big data and it happened when to understand the data, accessing the data and to improve the quality of data.

A Understanding the Data

The common problem when the research relates with data is understanding the data. It takes a lot of time to understand the data very well. It is important to understand it in analyzing and to get the right interpretation of data. For example, if the data is from medical industry, then the researchers have to understand how to visualize out of the data. To make sure the process of interpretation of data in the right path, there need some involvement from domain expertise for example is doctor for medical industry. So, the researchers who analyzing the data will understand very well and extract the useful knowledge from the big data. However, since the big data is involving very high volume of data, it is the biggest challenge to consume the valuable information for decision making purpose.

Beside that, to extract the knowledge from the big data, the common problem is there are need a combination of expertise in handling the data, such as a domain expert, database expert, statistical skill, mathematical expert, machine learning expert, data visualization and business and communication skills. Even though, there have cooperation between all expertises, it is difficult for analyzer to explain knowledge produced to the domain expert [7].

B Accessing the Data

Besides understanding the data by involving an expertise in such area, the data must be accessed quickly. However the challenges is the volume of data very high need to acces all the levels needed in high speed. The big data normally heterogenous data set [2]. So the major challenge is to figure out what the data is and how to analyze it. Beside that, its also involve integration problem, mainly coming from active literature on data and schema integration issues, but it also has deep consequences on the kind of analytics to be designed. Beside that, high-flux, streaming data methods are often required because the analyst might have only one shot at accessing the data [7].

There are some of big data in unstructured data sources such as social network data, biological experiment result and transaction. In order to extract knowledge, the dataset should transformed to suitable format. There are some issues existed involve classical ETL processes of data warehousing systems. Based on [7], transformations of data from unstructured to structured format should be performed on the basis of the analytics to be designed, according to a sort of goal-oriented methodology.
Thus, effective processing of Exabyte of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information [2].

C Data Quality

Access a big data should be very quickly to make sure the process to extract useful knowledge does not take longer time. However, the valuable information extracted more important for decision making. To produce a good information, the quality of data should be assured. In order to get high quality of data, the information management phase have to ensure the data is clean. So, the researchers have to find the best method to address quality to make sure the problem will not arise later.
Big data also can caused uncorellated data. Due to the large size of large data repositories, dealing with a large amount of data that is uncorrelated with the type of analysis to be designed happen very often. So it is a big problem in filtering out uncorellated data to improve the quality of data as well as affect the quality of knowledge produced [7].

3 Past Research In Extracting Knowledge From Big Data

George et al.[10] have studied an Australian state emergency service using Big Data to improve the delivery of emergency services. In this study, a research methodological approach that encom passes two phases was adopted. In the first phase, a comprehensive literature review of journal articles dealing with 'big data'- related topics was conducted. In the second phase, an analysis of an in-depth case study of an Australian state emergency service which is currently using 'big data' for improved emergency service delivery is realized. The approach has three key characteristic: (i) the development of a classification framework (ii) conduct the literature review (iii) realize the classification of relevant journal articles.

In the second phase, the study draws on an in-depth case study on the use of 'big data' by The New South Wales State Emergency Service (NSWSES),Australia for improved emergency service delivery, so as to draw lessons for the effective use of 'big data'. The better management of emergency \operations required the integration of multiple sources of data (structured and unstructured) across multiple agencies; the combination of these data with historical information for better emergency service delivery. In the case of The NSW SES, the agency has developed a range of IT capabilities over time. For example, the NSW SES has a bi- directional direct communication between its website and the Bureau of Meteorology website during major operations to offer the public a real-time access to accurate weather or emergency information. The same capabilities allow the NSW SES to share resources (humans and assets) with other states during major disaster events. The organization has been aggressively using cutting edge tools and technologies such as paging, telephony, radio, spatial systems, enterprise resource planning (SAP), communications, and mapping tools, in order to provide improved capabilities to its volunteers during emergency response operations. In October 2009, the NSW SES upgraded the corporate IT infrastructure to implement a new collaborative platform based. on Microsoft SharePoint. In December 2009, it completed a successful implementation of the emergency services shared SAP system with other states emergency agencies to enhance their level of information sharing and collaboration at the local and state levels for improved service delivery.

NSW SES has deploy social media tools to expand the service's communication with keystakeholders during emergency events and to assist in a positive profiling of the organization and its members. More recently, it started a project of equipping all staff members and selected volunteers in regions with Smart- phones to support field operations while on the move. Using this improved IT infrastructure, the NSWSES can now integrate information from various databases and flood plans so as to identify the potential risks to which different regions maybe exposed ,and then take preventive actions(e.g.,evacuation, alert messages).For example, by merging the Bureau of Meteorology's external data with its own internal data(e.g.data from flood plan, historical data information from various databases),the NSWSES can now apply predictive analysis and therefore anticipate the impact of a disaster on a given region.

Key insights from the in-depth case study indicate that creating and capturing business value from 'big data' can allow a real-time access and sharing of information across local and national government agencies for improved decision making to enhance emergency service response.Another key benefit realized from 'big data' by the NSWSES is the improvement of intra-and inter-organizational transparency and accountability, which represent major issues in the government environment. Moreover, the ability of the NSWSES to handle and support data from various sources and formats(structured and unstructured),as well as to push 'intelligence' from these data to various channels so as to support emergency operationon the field, was a critical success factor in this process of creating and capturing business value from 'big data'.

In biology, most of the methods used in genome-wide research are based on statistical testing and designed for analyzing a single experimental dataset. The data explosion introduced by modern genomics technologies requires biologists to rethink data analysis strategies and to create powerful new tools to analyze the data. In recent decades, machine learning has been envisaged by life

scientists as a high-performance scalable learning system for data-driven discovery. Therefore, the primary goals of this review are to introduce the basic concepts and procedures of machine learning in biology and to envisage how machine learning could interface with Big Data technology to facilitate basic research and applied biotechnology in plants.

HDLSS data are common in plant and other science studies [11]. Such data contain a large number of attributes and a relatively smaller number of training examples, which tends to cause overfitting [12]. In addition, many of the collected attributes are irrelevant (weakly correlated with the output) or redundant (highly correlated with each other). Their inclusion may make the learning process unstable and yield a model with large variance and poor discriminative power. Therefore, dimension reduction is necessary when using HDLSS data. Feature selection may take place at the data preprocessing or model learning step. When the number of features is too high, correlation analysis is often used to preselect or to screen features prior to model building [13]. Feature extraction is used to create new features by the transformation or function of raw features. One popular feature extraction procedure is principal component analysis (PCA), which extracts a small set of directions (called leading principal components; PCs) to represent the data and achieves great dimension reduction. Although most machine-learning applications were developed for animals, many of them are readily applicable to plants. For example, to improve the assembly quality of the *Drosophila mojavensis* and *Escherichia coli* genomes based on shotgun sequencing reads, machine learning was used to detect assembly errors caused by repetitive DNA sequences [14,6].

4 Technique In Analysing Big Data

A wide variety of techniques and technologies have been developed and adapted to aggregate, manipulate, analyze, and visualize big data.These techniques and technologies draw from different orders including statistics, computer skill, applied mathematics, and political economy.This signifies that an association that means to getvalue from big data has to sweep up a flexible, and multidisciplinary approach.

A few techniques and technologies were produced in a world with access to far little volumes and variety of data, yet have been effectively adapted so they are relevant to very large sets of more different data. Others have been built up more recently, particularly to capture value from big data. Some were developed by academics and others by organizations, especially those with online business models predicated on analyzing big data.

There are six (6) widely used big data analysis techniques that were discussed in this article, there are association rule learning[15], genetic algorithms[16], machine learning[17], regression analysis[18], sentiment analysis[19], and social network analysis[10].

A Association Rule Learning

Association rule learning is a method for finding interesting connections between variables in expansive databases furthermore experienced as if/then statements that help uncover relationships between apparently irrelevant data in a relational database or other data repository.
It was first practiced by real grocery store chains to discover interesting relations between items, utilizing data from market point-of-sale (POS) systems.In data mining, association rules are valuable for analyzing and predicting client conduct. They play an important part in shopping basket data analysis, itemsgrouping, and catalog design and store layout.

The researcher regularly uses association rules to assemble programs equippedfor machine learning. Machine learning is a kind of artificial intelligence (AI) that tries to establish programs with the ability to suit more efficient without being expressly modified.

B Genetic Algorithms

A technique used for enhancement that is inspired by the procedure of regular advancement. In this technique, potential arrangements are encoded as "chromosomes" that can consolidate and transform.These individual chromosomes are chosen for survival within a modeled "environment" that impact the performance of each person in the population.Frequently depicted as a case of "evolutionary algorithm," these algorithms are appropriate for solving nonlinear issues that need optimization.Samples of applications include improving job scheduling in manufacturing and optimizing the performance of an investment portfolio.

C Machine Learning

A subspecialty of computer science (inside of a field generally called "artificial intelligence") concerned with the outline and improvement of algorithms that permit computers to advancepractices in view of extract data. A major focus of machine learning examination is to consequentlyfigure out how to recognize complex patterns and make intelligent decisions based on data. It gives computers the capacity to learn without being explicitly programmed, and is focused on making expectationsin view of known properties gained from the arrangement of "training data". Natural language processing is an example of machine learning. Machine learning is being connected to separate in the middle of spam and non-spam email messages, learn client preferences and clear recommendations based on this information, and determine the best content for connecting with prospective clients.

5 Conclusion

This paper has outlined several issues that has been rise in Big Data. The main issues discussed here is regarding storing, manage and process the massive amount of data. As the name goes, storing the big data in Exabyte is definitely not an easy task and the data keep growing incrementally each day. Then, managing the data is also another issue because researchers have to deal with several formats of data. Beside, good techniques should be acquired to ensure quality data is chosen. Later, knowledge extraction is performed by employing analytic process. Big data is exploited with current technology to help researchers in diverse domain such as medical, agriculture, oil and gas and so forth to extract knowledge. The valuable extracted knowledge could be used in helping researchers or practitioner making decision.

References

[1] A. Cuzzocrea and K. C. Davis, "Analytics over Large-Scale Multidimensional Data : The Big Data Revolution !," pp. 101–103, 2011.
[2] A. R. Syed, K. Gillela, and C. Venugopal, "The Future Revolution on Big Data," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 6, pp. 2446–2451, 2013.
[3] Bifet, A., Holmes, G., Pfahringer, B., & Gavaldà, R. (2011). Detecting Sentiment Change in Twitter Streaming Data, *17*, 5–11. Retrieved from http://eprints.pascal-network.org/archive/00008650/
[4] Choi, J.H. et al. (2008) A machine-learning approach to combined evidence validation of genome assemblies. Bioinformatics 24, 744–750
[5] Devijver, P.A. and Kittler, J., eds (1982) Pattern Recognition: A Statistical Approach, Prentice Hall
[6] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78. doi:10.1145/2347736.2347755
[7] F. J. Alexander, A. Hoisie, and A. Szalay, "Big Data," *Comput. Sci. Eng.*, vol. 13, no. 6, pp. 10–13, Nov. 2011.
[8] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013.
[9] Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). Foundations of rule learning. *Zhurnal Eksperimental'noi I Teoreticheskoi Fiziki*, (2003). doi:10.1007/978-3-540-75197-7
[10] George, G., Haas, M.R., Pentland, A., 2014. Big data and management. Acad. Manage.J. 57 (2), 321–326.

[11] Hall, P. et al. (2005) Geometric representation of high dimension, low sample size data. J. R. Stat. Soc. B 67, 427–444
[12] Palmer, L.E. et al. (2010) Improving de novo sequence assembly using machine learning and comparative genomics for overlap correction. BMC Bioinformatics 11, 33
[13] R. Kitchin, "Big data and human geography: Opportunities, challenges and risks," *Dialogues Hum. Geogr.*, vol. 3, no. 3, pp. 262–267, Dec. 2013.
[14] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," *2013 46th Hawaii Int. Conf. Syst. Sci.*, pp. 995–1004, Jan. 2013.
[15] S. Sagiroglu and D. Sinanc, "Big data: A review," *2013 Int. Conf. Collab. Technol. Syst.*, pp. 42–47, May 2013.
[16] Saeys, Y. et al. (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517
[17] Varian, H. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, (June 2013), 1–36.
[18] Xu, Y., Zeng, M., Liu, Q., & Wang, X. (2014). A Genetic Algorithm Based Multilevel Association Rules Mining for Big Datasets. *Mathematical Problems in Engineering*, *2014*, 9.
[19] Fosso, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). Int . J . Production Economics How " big data " can make big impact : Findings from a systematic review and a longitudinal case study. *Intern. Journal of Production Economics*, *165*, 234–246. http://doi.org/10.1016/j.ijpe.2014.12.031