Editorial: Breakthroughs in algorithms and applications for big data analysis in modern business world

Rita Yi Man Li

Real Estate and Economics Research Lab,
Hong Kong Shue Yan University,
852, Hong Kong
Email: ymli@hksyu.edu

Biographical notes: Rita Yi Man Li graduated from the University of Hong Kong. She is the Founder and Director of Sustainable Real Estate Research Center and Real Estate and Economics Research Lab. She is an editor of many academic journals. Outside of academia, she is a Chartered Surveyor by profession.

1 Introduction

Data can be ‘big’ in many different ways. Some national and international projects, for example, Europe’s particle physics laboratory near Geneva and the Large Synoptic Survey Telescope in Chile, challenge the best computation approach, systems administration and data stockpiling. However, data can likewise be big by being enduringly important due to the difficulties that may require an experimental setting (Lynch, 2008). Big data can also refer to large volumes of unstructured and structured data from multiple sources with the 3Vs characteristics:

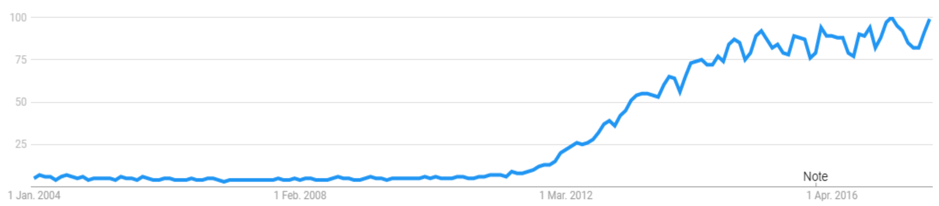
1 Variety: big data includes readings from sensors; messages, updates and images posted to social networks; and GPS signals from cell phones, etc. Many of the sources of big data are quite new. For example, Facebook was launched in 2004, Twitter in 2006 and the iPhone was only unveiled five years ago.

2 Volume: about 2.5 exabytes of data are created each day as of 2012, and this figure is doubled every 40 months or so. More data are now put into the internet every second than two decades ago.

3 Velocity: the velocity of data creation is more crucial than volume; real-time or nearly real-time information makes it possible for a company to be far more active than its competitors (McAfee and Brynjolfsson, 2012).

Since computerised information is so effortlessly shared, duplicated and recombinable, it shows enormous reuse opportunities, quickening the studies under way and exploiting past interests in science (Lynch, 2008). Academically speaking, the advent of big data enables novel research across a wide range of topics, and big data inevitably becomes one of the hot topics (Figure 1). That also leads to the development of various large-data statistical methods (Athey, 2017) and the application of big data in exploring the changes in trends over different times and places. For example, Li et al. (2016) and Li (2018) have studied the changes in the number of searches in smart homes and various automated tools which affect construction safety.

Figure 1 The keyword searches of 'big data' from 2004 to 2017 in Google big data (2017) (see online version for colours)



Whilst the existence of big data offers ample opportunities in research and technological breakthrough, it also imposes challenges to researchers and industry practitioners. 'Big data hubris' suggests that big data is an alternative choice to, rather than a complement to, conventional data series and evaluation. In fact, the amount of data does not imply the possibility of ignoring fundamental issues of construct validity and reliability amongst data. The core venture is that most big data is not an output of instruments designed to produce reliable and valid data for scientific analysis (Lazer et al., 2014).

Other challenges include the development of an appropriate data management and programming capabilities, and; methods which design creative and scalable approaches to summarise, describe and analyse large-scale unstructured data. For example, machine-learning forecast strategies have been productive in applications ranging from pharmaceutical to health applications in urban areas. The gaps between prediction, choice making and hidden axioms need to be comprehended (Athey, 2017).

In view of the abovementioned opportunities and challenges that can be brought by big data, this special issue offers insight into the possible usage of big data's application such as storage points optimisation, name matching, teaching evaluation and data analysis application of the survey. It also sheds light on the various breakthroughs in the algorithm for clustering MapReduce and the multicriteria decision-making approach.

Brojo et al. implement opinion mining via a large dataset in teaching evaluation. Classification technique, GPU architecture via CUDA-C programming model and Hadoop MapReduce programming model are used to evaluate the university faculty's teaching evaluation in a faster way. Srinivasa et al. study the detailed information of about 3.5 million households that participated in the US Census Bureau's American Community Survey. Their study attempts to derive useful insight by using classifiers.

To lower the storage costs, storage sites are usually far away from residential areas. Hongying and Yu propose an integer programming model to minimise the cost in operating online to offline commerce charges by optimising the locations of reverse logistics. An improved genetic algorithm is designed to solve the two-stage heredity under a random circumstance. It builds a multilayer reverse logistics network to recycle and transport the customer returns to the collecting point according to the best solutions suggested by the algorithm and then moves them to the remanufacturing centre in factories. Both the simulation and numerical examples confirm the feasibility and effectiveness of this genetic algorithm.

Paweł presents an eco-innovation implementation risk in enterprises. Eco-innovation solves emerging environmental problems that arise due to strong economic growth and results in a new or significantly improved product or process or a new marketing method. Eco-innovation should be seen as an integral part of innovation efforts across all the economic sectors. European countries impose many barriers on eco-innovation such as high investments risk and limited interest.

Kirubakaran and Aramudhan address the complications in the name matching process due to structure, spelling, phonetic variations, cross language translation, operating system transformation, batch feeds, data migration and so on. This paper discusses the methods to parse and ways to attain name matching with high accuracy. The proposed methods can be applied to the financial sectors' risk intelligence analysis like anti-money laundering, customer due diligence, fraud detection, anti-terrorism and watch list screening.

The traditional k-means clustering algorithm is very sensitive to the cluster centres' initial placement and many initialisation methods are generally used for it. Kazemi and Khodabandehlouie suggest two new non-random initialisation methods for the k-means algorithm. They are tested by data and evaluated by the TOPSIS multicriteria decision-making approach.

Lakshmi implements shuffle as a service component to lower the execution time in MapReduce application, monitor the map phase via skew handling and raise resource utilisation in a cluster. Naveen et al. present an automatic system based on fuzzy logic to predict the understanding time of conceptual data warehouse models. It utilises quality metrics as the inputs and understanding time as the output. The results are validated with the actual data. The predicted results confirm the validity and efficiency of the designed automatic system. Besides, Seelammal and Vimala Devi provide the best selection feature via multicriteria decision making, which is decision tree learning with an emphasis on optimisation of constructing trees to handle large data sets.

References

Athey, S. (2017) 'Beyond prediction: using big data for policy problems', *Science*, Vol. 355, No. 6324, pp.483–485.

Google (2017) [online] http://www.google.com.hk, (accessed 8 October 2017).

Lazer, D. et al. (2014) 'The parable of Google flu: traps in big data analysis', *Science*, Vol. 343, No. 6176, pp.1203–1205.

Li, R.Y.M. (2018) *An Economic Analysis on Automated Construction Safety: Internet of Things, Artificial Intelligence and 3D printing*, Springer, Singapore.

Li, R.Y.M. et al. (2016) 'Sustainable smart home and home automation: big data analytics approach', *International Journal of Smart Home*, Vol. 10, No. 8, pp.177–187.

Lynch, C. (2008) 'Big data: how do your data grow?', *Nature*, Vol. 455, NO. 7209, pp.28–29.

McAfee, A. and Brynjolfsson, E. (2012) 'Big data: the management revolution', *Harvard Business Review*, October, pp.1–9.