Research on Big Data and its Usability

Qiong Ren [a], Junming Chang

School of Mathematics and Computer Science, Jianghan University, Wuhan, China

[a] qren@163.com

Abstract

With the rapid development of information technology, especially the great progresses of Internet, cyber physical system Internet of things, cloud computing and social network, big data becomes ubiquitous. Big data brings not only great benefits but also crucial challenges. Improving the data usability is one of the most significant challenges. Dirty data accompanies the tremendous increase of data volume, degrades the data quality and data usability, and brings serious harm to the information societies. Fortunately, there has been widespread concern about the data usability in both industrial and academic communities, and the recent research efforts on data usability have yielded some impressive results. In this paper, the concepts of big data and its characteristics are introduced, and then the works related to the data usability are surveyed.

Keywords

Big data, characteristics, data usability, standardization.

1. Introduction

Nowadays, society is the one of informatization and digitization. The rapid development of the Internet, Internet of things and cloud computing technology, makes the data tremendous in the whole world. at the same time, the data also become a new kind of natural resources. [1] The development of social informatization and network data grow explosively. The ultra large amounts of data need to be used reasonably, and highly efficiently, which can bring more efficiency and value to people's life.

Ubiquitous zaw, mobile devices, RFID and wireless sensors produce data every minute. Hundreds of millions of Internet users generate a huge amount of information interaction. More and more data need to be processed. Furthermore, the business requirements and competitive pressure on the real time and validity of the data processing put forward higher request, the traditional regular data processing technology has been outdate. Big data has brought a lot of practical problems. In order to solve these problems, it's needed to break through the traditional technology according to the characteristics of the large data for new technological change. Big data technology is a gather of collection, storage, management, processing, analysis, and sharing and visualization technology.

2. Definition and characteristics of big data

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of "Big Data," including the recent announcement from the White House about new funding initiatives across different agencies, that target research for Big Data. While the promise of Big Data is real, there is no clear consensus on what is Big Data. In fact, there have been many controversial statements about Big Data, such as "Size is the only thing that matters."

2.1 Definition of big data

Common view is that big data refers to large scale and complex dataset, which is difficult to be handled with the existing database management tools or data processing application. [2] McKinsey defines big data as: datasets that can't be fetched, managed or handled with traditional database software tools in a certain period of time. [3] According to different sources, the big data can be roughly divided into the following categories: [4]

(1) From the people. People produced i all kinds of data, including text, images, video and other information, in the Internet activities and the process of mobile Internet.

(2) From the computer. All kinds of computer information system produce data, which exist in the form of files, database, multimedia, including auditing, automatically generated information etc.

(3) From the object. All data collected by digital equipment, such as cameras, medical Internet, and astronomical telescope, etc.

2.2 Characteristics of big data

It is hard to avoid mention of Big Data anywhere we turn today. There is broad recognition of the value of data, and products obtained through analyzing it. There was general consensus at the workshop that a big data benchmarking activity should begin at the end-application level, by attempting to characterize the end-to-end needs and requirements of big data applications. While isolating individual steps of such an application such as, say, sorting, is also of interest, this should still be done in the context of the broader application scenarios. The characteristics of big data can be summed up in 5 Vs, namely the Volume, the Variety, the Velocity, the Value and the Veracity.

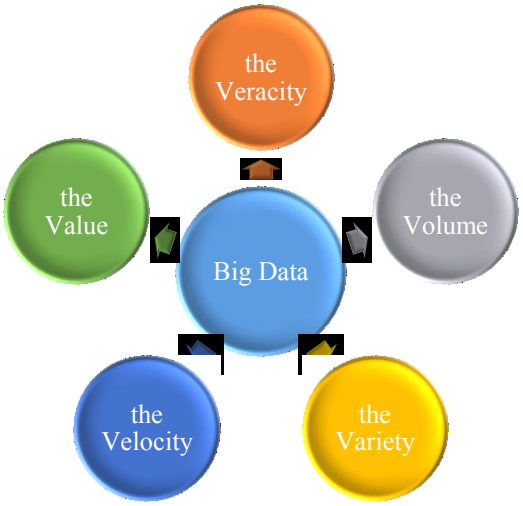

Fig. 1 Five Vs of big data

(1) The volume. The size of the data gathered for the analysis is very large. Google executive chairman said, now the data the world created every two days is equal to the size of data that produced from the human civilization to 2003. The concept of "big" is relative. For search engines, EB level belongs to the larger size, but for all kinds of database or data analysis software, the scale level will have a larger difference.

(2) The Variety. from the generated types, The diversity of data can be divided into trading data, interactive data, sensor data; From the data source, it can be divided into social media, the sensor data, system data; From the data format, it can be divided into text, images, audio, video, spectrum, etc.; From the relationship between data into structured, it can be divided into semi-structured, and unstructured data; Data from the data owner, it can be divided into company, government, social data, etc.

(3) The Velocity. on the one hand, growth data speed is explosive. on the other hand, the speed of data access, processing and delivery should be greatly fast. The amount of Digital data will be doubled every 3 years. The information storing speed is four times faster than the growth rate of the world economy.

(4) The Value. Although we have a large amount of data, only very small part of the data is valuable. The huge value hides behind big data. For instance, the social networking site Facebook has 1 billion users. Through analysis of the information of the users, advertisers can do advertisement accurately

according to the analysis results. For advertiser, one billion user data is as worth as hundreds of billions dollars.

(5) The Veracity. Under the virtual network environment, the authenticity and objectivity of the large amount of data should be clear, which is the necessity of big data technology and the urgent needs of the development of business. Furthermore, through the analysis of the big data, we can restore the true color of things and predict their development trend.

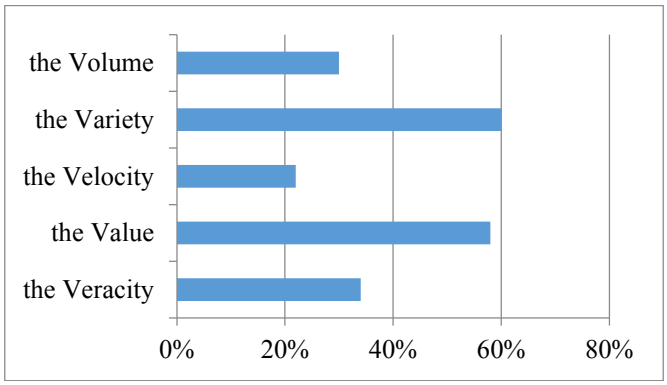

Fig. 2 The attention ratio of the characteristics of big data

Through the social widespread survey, we can see the attention ratio of the big data characteristics in figure 1. It is not hard to see from the graph, "variety" and "value" are paid more attention by people. The reason that the variety is the most popular characteristic is that the variety of data makes changes on the storage and application. While "value" goes without saying that both the value of the data and the implied value are the goals that enterprises, departments and government want. Therefore, it's the problem that the big data should solved that how to convert data of such diversification into value.

3. Big data usability

With the explosive growth of big data, bad data followed, resulting in poor quality of data, which greatly reduce the usability of data. The fact shows that the big data exist serious problems in terms of usability. Data usability problems are inherent in the information society. They not only exist in the western developed countries, but also are common in any place of the information society. Data usability is related to the national economy, people's livelihood and social harmony of the era of large data of a major strategy. Data usability is the important premise of the successful completion of big data management infrastructure construction and giving play to the role of big data effectively. Therefore, it has important strategic significance to research the basic theory o and key technology. [5]

3.1 High quality big data acquisition and integration technology

High quality data acquisition is an important premise to ensure information usability. Therefore, we need to address the following challenges: in the data acquisition phase, we should check the quality of data carefully. We should explore the multiple data sources, such as physical information system, to obtain high quality big data effectively. Then we establish a multimodal data fusion theory and algorithm to achieve high quality precision data acquisition and integration, and find the data evolution law.

The Existing data usability method and system lack a solid theoretical foundation, which cannot achieve automatic error detection and repair. In order to realize automatic detection and reparation of data errors, we need to solve the following challenge on the basis of data usability theory system: putting forward the issue of large data errors automatically detect and repair the computability theory, the big data error automatically detect and fix the problem of computational complexity theory.

When errors in the data cannot be completely repaired, these data are called weakly available data. When calculate on the weakly available data directly, the results are always fine. So it is a meaningful choice. However, the existing theories and algorithms cannot support weak approximate calculation on the available data. Therefore, we need to solve challenging problems as follows: putting forward weakly approximate calculation and undertaking the feasibility analysis of big data.

3.2 The standardization of big data

We know that big data exist in many fields and various processing products. Not only the definition of big data, related terms, classification, structure aspects are lack of unified description, but also all kinds of big data products technical requirements are different. To some extent, this kind of situation hindered the healthy development of industry of big data. Big data is like a large oil tanks, analyzed data are various. We should solve the problem of data explosion and solve the problems of the garbage data. Data standardization and enhancement can complete data resources, prevent garbage data, implement the data capitalization.

When comes to the implementation of standardized work, from the principles of standardization, big data have intersection with lots of existing technologies, such as relational databases, data mining. Big data standardization should focus on the big data under the background of new technology and new applications, such as the relational database and real-time unified. From the scope of standardization, the lifecycle of the big data each link should be taken into consideration, at the same time the new application and new products is also the focus of the standardization work, triggered by big data.

4. Summary

Big data usability research has just been started and in its infancy. The existing research work mainly limited in the relationship between the centralized storage data consistency and the real identity. Only a handful of researches are about the timeliness and completeness of big data. Big data accuracy is neglected. The existing data error detection and restoration algorithm is not suitable for big data system. Usability research put forward the complete theory system of large data usability, the high quality of the theory and methods of data acquisition.

Acknowledgments

2014 Wuhan Municipal colleges and universities teaching and research project (No.2014088); Wuhan City Bureau of Education in teaching and research project (No. 201391).

References

[1] Li Jianzhong, Liu Xianmin: Journal of Computer Research and Development, Vol. 50 (2013) No.6, p.1147.
[2] Information on http://en.wikipedia.org/wiki.Big_data.
[3] Ahlswede R, Cai N, Li S Y R, et al: IEEE, Trans. on Inform Theory, Vol. 46 (2000) No.1, p.1204.
[4] Liguo Jie, Chengxue Qi: Bulletin of Chinese Academy of Sciences, Vol. 27 (2012) No.6, p.647.
[5] Sun Li, Yang Jun, Pan Kunyou: Science and Technology Management Research, (2014) No.19, p.35.