Agriculture Big Data: Research Status, Challenges and Countermeasures

Haoran Zhang, Xuyang Wei, Tengfei Zou, Zhongliang Li, Guocai Yang

Agriculture Big Data: Research status , challenges and countermeasures

Haoran Zhang[a], Xuyang Wei[b], Tengfei Zou[c], Zhongliang Li, Guocai Yang[d1]

School of Computer and Information Science, Southwest University, Chongqing 400715, China
[a]haoranzhang0715@163.com, [b]weixuyang321@163.com, [c]670543900@qq.com,
[d]paul.g.yang@gmail.com

Abstract. Agriculture data type and amount in human society is growing in an amazing speed which is caused by the emerging of agricultural Internet of things. The growth of agricultural data volume brought many difficulties in storage and analysis. The realization of the big data and cloud computing technologies provides the solution of these problems. The cloud computing and big data technologies can be applied in agriculture to solve the problems in storage and analysis. This paper elaborates the related concept of agricultural big data briefly, and emphasizes some current researches, the challenges that agricultural big data should face in the future and its countermeasures.

Keywords: agricultural big data, cloud computing, challenge, countermeasure

1 Introduction

Until now, the agricultural big data has not been clearly defined. We can reach the definition with two parts — big data and agricultural informationization.

The accurate definition of big data also has some disputes. The well-known consulting company McKinsey argues that big data cannot be obtained, stored and managed in a certain time using a traditional database software tools[1]. The definition given by wikipedia shows big data is a blanket term for any collection of data sets which are so large and complex and it becomes difficult to process using on-hand database management tools or traditional data processing applications[2]. there is another view that big data should meet three terms which are Variety, Velocity and Volume[3]. Some people has a different view. They think that three terms can't describe the characteristics of big data, so they come up with four terms. But there is no unified statement on the fouth characteristic. The International Data Corporation IDC believes big data should also has a term of Value[4]. The IBM corporation argues that big data should be considered with Veracity[5]. I think that the term of value makes the definition of big data more accurate. Although the value of big data is sparse, there is great potential inherent value.

Agricultural informationization is a dynamic concept. It improves the comprehensive productivity of agriculture and the overall productivity of agriculture using modern information technology and information systems[6]. In brief, agricultural

[1] Corresponding Author Email: paul.g.yang@gmail.com

informationization makes full use of information technology methods, means and process to achieve the goals.

Agricultural big data uses concepts, techniques and methods of big data and cloud computing to handle a mount of agricultural data. We can obtain the information which is useful to guide agriculture. Wen Fujiang shows that agricultural big data is related to agricultural production in all aspects. It is a multi-professional data mining process[8]. The implementation of agricultural big data technology is a very important component of agricultural informationization. And it also provides new methods and ideas for agricultural research and agricultural business development. Professor Wang refers that we have many datasets of agriculture. As a result, there is a good deal of agricultural data. But the data standards are not uniform and standardized. How to give a reliable and professional decision is an urgent task[9]. In order to solve the problem of the development of agricultural informationization, we should establish a national big data center and make efforts to develop cloud computing and to search agricultural big data mining techniques.

2 Developing Status of Agricultural Big Data

With the rapid rise of the Internet of things and social networks, the amount of data is growing at an unprecedented rate. According to Winter Corporation's survey, the current amount of data is growing at three-time increases every two years[10]. Its growth rate exceeds the growth rate of Moore's Law far. The era of big data is coming. "NATURE" magazine published a special issue named "Big Data: Science in the Petabyte Era"[11]. This paper described the challenges of big data in many aspects,such as Internet technologies, network economics and so on. "SCIENCE" magazine also launched a special issue named "Dealing with Data"[12] in 2011. It illustrated the important scientific value of big data. European Information Society Science and Mathematics Research journal ERCIM News discussed the management and innovative technology issues in their special issue named "Big Data"[13] in 2012. Gartner Company showed that the development of big data has entered the peak period in 2013 Hype Cycle for Emerging Technologies. We showed it in Fig.2-1. Many transnational corporations, such as IBM,ORACLE and FACEBOOK, are making their efforts to develop big data technologies[14]. Academician Professor Li Guojie noted that we should take the redundant data out of raw data in the data explosion era[15]. In December 2013, China Big Data Technology Conference was held and it mainly discussed the technical means and the commercial value of big data applications. Big data has been a intersection of information science, social science, network science, and many other emerging interdisciplinary field. It has became a hot research point.

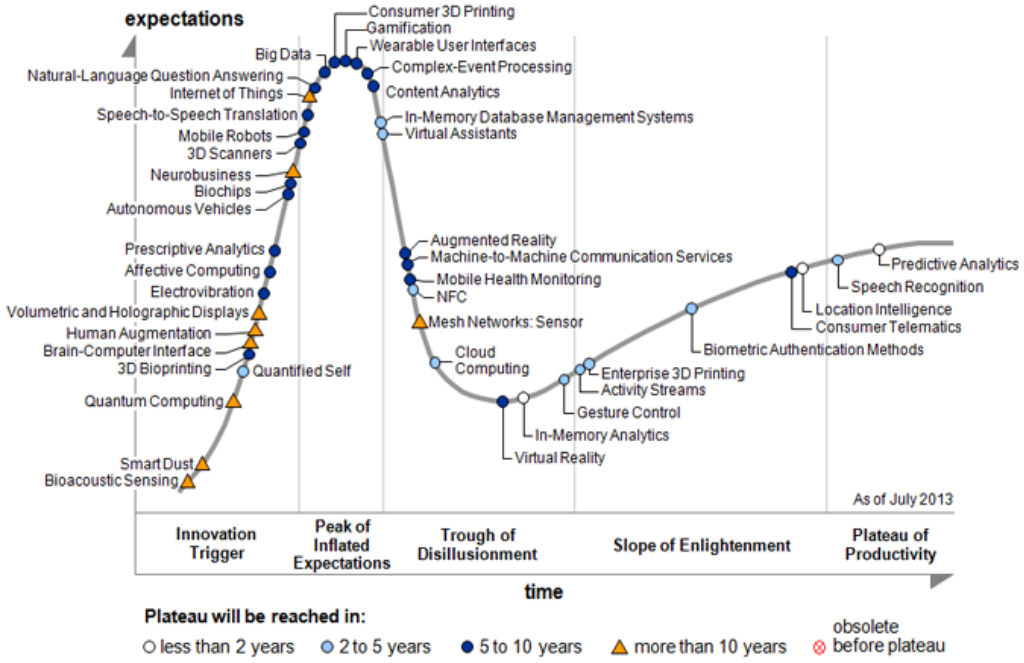

Fig. 1. Emerging Technologies Hype Circle,2013 [16]

Recently, big data has been applied to different points of healthcare, manufacturing, transportation, and financial sector in China. With the use of agricultural informationization and Internet of things in agriculture, agricultural big data will become another focus of big data research. Agricultural big data is a application of big data in agriculture. Agricultural big data which is related to all aspects of agricultural production is a issue of multi-disciplinary, cross-sectoral analysis and data mining. The first domestic agricultural big data industry technology innovation strategic alliance was established in Shandong Agricultural University in June 2013. Shandong Agricultural University President Wen Fujiang pointed out that in China, the current big data research is just beginning, but agricultural big data research is leading research[8]. Many companies are targeting agricultural big data opportunities. The New York Times reported that in 2012 the Solum company received 17 million dollars fund form Andreessen Horowitz and other companies. Solum company commits to use data analysis techniques to determine the amount of fertilizer in cultivation. It helps farmers improve productivity and reduce costs through the analysis of agricultural big data. Multinational agricultural biotechnology company Monsanto spent 930 million to takeover the insurance accident weather company Climate Corporation [18]. Climate Corporation predicts the weather which may damage agricultural production through analyzing their massive weather data. According to the forecast, farmers can select appropriate agricultural insurance, to reduce the impact of bad weather in agricultural process. The analysis of agricultural big data plays an indispensable role in agriculture.

While agricultural big data is still a fresh vocabularies, but the number of data which is produced in the process of agricultural production is far more than small data. Agricultural data generation method has been changing dramatically. With the wisdom agriculture proposed, widely used sensors promote the development of agricultural Internet of things greatly. Agricultural Internet of things generates a lot of agricultural data. We should use big data technologies to analysis the data. The emergence of cloud computing provide a reliable way to deal with these massive data. At this stage, there is few research on agricultural big data. But large number of researchers have realized the value of agricultural big data, and have also been put into analyzing agricultural big data and using agricultural informations. If we can make good use of agricultural big data, it will be not only a great innovation in the history of human agriculture, but also a pioneering work in human history.

3 Challenges of Agricultural Big Data

In its infancy, big data technology is facing many challenges, such as wide range of heterogeneous data, problem of real-time, data incompleteness, lack of priori knowledge, private issues and so on. Issues which agricultural big data is facing is consistent with the big data technology. But compared with big data, agricultural big data is not so sensitive in security or privacy issues. Agricultural data mining is targeted at using the result to guide agricultural practices. Therefore, agricultural big data is facing the following issues in sum.

3.1 Problem of Agricultural Big Data Storage

Agricultural big data has a very different modality. From the source point of view, the data comes from Internet of things of the radio equipment, the agricultural information websites, and a variety of advanced mobile terminals. From the content point of view, it includes not only statistical data, but also basic information on agriculture-related economic entities, investment information, import and export information and GIS coordinate information. Data types also include Structured data, semi-structured data and unstructured data. It will be a problem worth studying to store heterogeneous data and the ability to read and write because of the different treatment of different storage hardware devices.

The heterogeneous hardware is also a problem of storing agricultural big data. There will be a very significant performance differences between different machines in the data center. Different hardware devices have different literacy and processing capabilities. Handing equipment will waste a lot of time waitting the slower storage devices. The linear growth of storage devices and servers will not necessarily bring the linear growth of computing power in this case. The "Barrel Effect" restricts the performance of the entire cluster.

3.2 Problem of Agricultural Big Data Analysis

Data analysis is the core of the whole process of agricultural big data, because the value of agricultural big data is produced in the big data analysis process. Currently, there are many problems in agriculture such as food security, soil management, pest forecasting and prevention, soil management and agricultural consumption. We can use the analysis of agricultural big data to solve these problems. Raw data comes from the extraction and integration of agricultural information. We can select all or part of these data to do the research. Conventional analytical techniques are not applicable on processing agricultural big data, such as data mining, machine learning, statistical analysis and other techniques. It is mainly discussed in the following parts.

(1)Traditional data mining algorithms, such as machine learning and other areas, are no longer suitable for agricultural big data. On one hand, algorithms which mine a small amount of data can not be directly applied to big data. On the other hand, agricultural big data has its particularity. The accuracy of the algorithm is no longer the main standard. Algorithm needs to strike a balance between timeliness and accuracy of processing in many cases.

(2)The metrics of the quality of the analysis result is also a major challenge. Big data has complex types. It leads to many problems in designing the indicators to algorithm.

3.3 Problem of Agricultural Big Data Timeliness

As the time elapsing, the inherent value of data keeps attenuating. Therefore, timeliness must be considered during the analysis of agricultural big data. Untimely data analysis may result in the production of agricultural disasters, Especially in the weather, environmental conditions associated with data analysis. For example, the occurrence of "Low grain price hurts farmers" event is the result of managing the cost of production and other information not in time. Therefore, the characteristic of timeliness is particularly important in agricultural big data.

4 Countermeasures of Agricultural Big Data

To solve the problem of agricultural big data, we will give the corresponding countermeasures which will give some guidances for the future work.

4.1 Storage of Agricultural Big Data

Agricultural big data storage affects not only the efficiency of data analysis, but also the cost of data storage. Therefore, we need to come up with high-efficiency and low-cost data storage. One solution is using distributed storage mode. Distributed file system is the physical storage resources in the file system and is not necessarily directly connected to the local node. But the node is connected via a computer network. Many companies which have to deal with huge amount of data have their own distributed file system, such as Google's GFS(Google File System)[19-20], Taobao File System (TFS)[21]. There are some open source distributed file

system,such as HDFS[22]，NFS[23-24]. There is another solution. The research can be concerned in searching the technologies about the acquisition and integration of multi-source ans multi-modal data. In addition, high-speed storage and creating index are also important aspects.
The general solution to solve the problem of heterogeneous hardware is using different storage devices in different aspects in heterogeneous hardware environments. The problem will become very complicated when the scale of heterogeneous environment extends to thousands of clusters.

4.2 Analysis of Agricultural Big Data

After years of researches and developments, data mining, machine learning, statistical analysis and other information analysis have been proved to have significant effects for small data. These algorithms can be adjusted to accommodate cloud computing system. But it must be noted that we must consider the characteristics of agricultural big data in the adjustment process of these algorithms. Real-time and predictable characteristics must be considered. It will be a hot spot in the coming period of scientific research.
It is important and difficult to evaluate the results of agricultural big data algorithms. According to the characteristics of agricultural big data, we can use timeliness as a measure standard. Agricultural big data is massive, so we can use the prior knowledge to test the algorithms. It can measure the quality of the algorithms to a certain extent. It can also measure the reliability of the data results.

4.3 Timeliness of Agricultural Big Data

Timeliness is a core demand of agricultural big data analysis. A lot of research are also expanded around this demand. There are three methods to ensure timeliness.
The first method is using stream processing mode. Although streaming mode is suitable to real-time system, its application field is relatively limited. Streaming application model focuses on real-time statistical system, online monitoring. The second method is batch mode. In recent years, the development of batch model real-time system has became a hot topic and has achieved a lot of achievements. MapReduce programming model which Google company made in 2004 is the most representative batch mode. The third method is using a combination of stream processing and batch mode. The main idea is to use the MapReduce programming model to achieve stream processing.

Acknowledgment

This work was supported by the Key Technologies R & D Program(2012BAD35B08), China.

References

1. James M, Michael C, Brad B. Big data: The next frontier for innovation, competition, and productivity [J]. The McKinsey Global Institute, 2011.
2. Big data [EB/OL]. http://en.wikipedia.org/wiki/Big_data/
3. Grobelnik, Marko. Big Data Tutorial [EB/OL]. http://videolectures.net/eswc2012_grobelnik_big ㄧ data/
4. Hamish Barwick. The" four Vs" of Big Data. Implementing Information Infrastructure Symposium [EB/OL]. http://www.Computerworld.com.au/article/396198/iiis_four_vs_big_data/
5. IBM. What is big data? [EB/OL]. http://www901 .ibm.com/software/data/bigdata/
6. Zeng X J, Ding C Y, Wen Hongxia, et. al. Effective Methods for Agricultural Informationization[J]. Agricultural Library and Information Sciences,2004(2):34-37.
7. Qian X J. Research of Agricultural Informationization in the Process of China’s Agricultural Modernization[D]. Beijing: China Agricultural University, 2005.
8. Wen F J. Strategic significance of data and collaboration mechanisms agricultural research[J]. Higher Agricultural Education, 2013, 11: 002.
9. Wang R J. Bottleneck of Agricultural Informatization Development in China and the Thinking of Coping Strategies[J]. China Academic Journal Electronic, 2013, 28(003): 337-343.
10. WinterCorp. The Large Scale Data Management Experts [EB/OL]. http://www.wintercorp.com/
11. Nature. Big Data [EB/OL]. http://www.nature.com/news/specials/bigdata/index.html/
12. Science. Special Online Collection; Dealing with Data [EB/OL]. http://www.sciencemag.org/site/special/data/
13. ERCIM News. Big Data [EB/OL]. http://ercim-news.ercim.eu/en89/
14. Li G J, Cheng X Q. Research status and scientific thinking of big data[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657.
15. Li G J. The New Focus of Information Technology Big Data[J]. [2013-04-12]. http://www. cas. cn/xw/zjsd/201206/t20120627_ 3605350. Shtml.
16. Jackie Fenn, Hung LeHong. Emerging Technologies Hype Cycle for 2013: Redefining the Relationship [EB/OL]. http://my.gartner.com/portal/server.pt?open=512&objID=202&mode=2&PageID=5553&showOriginalFeature=y&resId=2546719&fnl=search&srcId=1-3478922244
17. Big Data Across the Federal Government [EB/OL]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheetfmall .pdf
18. Monsanto Acquires The Climate Corporation [EB/OL]. http://www.monsanto.com/features/pages/monsanto-acquires-the-climate-corporation.aspx
19. Ghemawat S, Gobioff H, Leung S. The Google file system [C]. the ACM Symposium on Operating Systems Principles, 2003:29-43.
20. McKusick M K, Quinlan S. GFS: Evolution on Fast-forward [J]. ACM Queue, 2009, 7(7):10-20.
21. Chucai. TFS Introduction [EB/OL]. http://rdc.taobao.com/blog/cs/
22. Konstantin S, Hairong K, Sanjay R et al. The Hadoop Distributed File System [C]. The 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010:1-10.
23. Osadzinski A. Network File System(NFS) [J]. Computer Standards and Interfaces, 1988, 8(1):45-48.
24. Anderson T E,Dahlin M D, Neefe J M, et al. Serverless Network File Systems [J]. ACM Transaction on Computer Systems, 1996, 14(1):41-79.