Study on Key Technologies of Agricultural Information Hotspots based on Big Data Analysis

Ji-chun Zhao[1,*], Guo-jie Wang[2]

1 .Beijing Academy of Agriculture and Forestry Sciences,

The Research Center of Beijing Engineering technology for Rural Remote Information Services, Beijing, China

2 .Handan Polytechnic College，Handan, Hebei, China

Zhaojc@agri.ac.cn

Keywords: Agricultural Information; Hotspots; Big data analysis

Abstract. Agricultural information is growing rapidly, how to quickly get valuable agricultural information is very important to agricultural policy makers and producers. The paper discusses key technologies of agricultural information hotspots based on big data analysis. The method of information access and analysis technology is given, and the agricultural information hotspots system is developed and discussed. The test result shows that the agricultural information hotspots system can provide value information for agricultural policy makers and producers.

1 Introduction

At present, the growth speed of online agricultural information resources is rapid with development of computer and communication technology. Agricultural policy makers and producers facing large information encounter the problems of information loss and resource overload, the main reason is that users can't get useful information because of large amount of disordered information. Agricultural information hotspot is important for agricultural policy makers and producers with agriculture information growing rapidly.

In the past, due to the limitations of technology, a large number of data processing is slow, but with the development of large data analysis technology and progress, there is new technology method for network information process. Growth of and digitization of global information storage capacity is shown in Figure 1.[1]

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk. Analysis of data sets can find new correlations to "spot business trends, prevent diseases, and combat crime and so on."[2] Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics,[3] connectomics, complex physics simulations, biology and environmental research.[4]

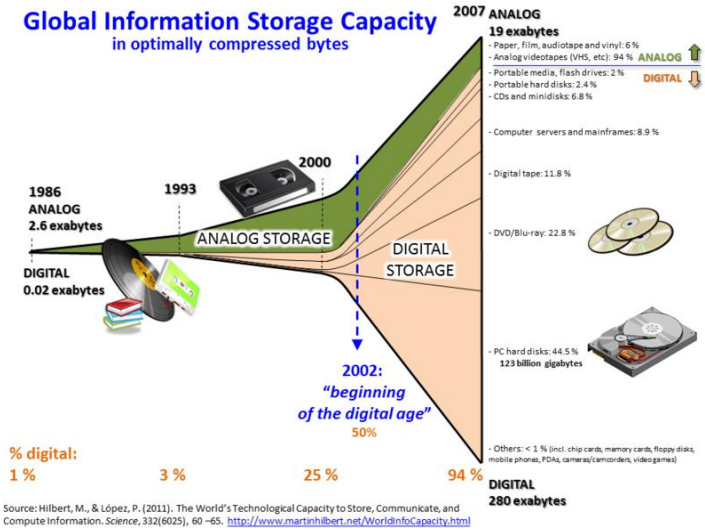

Figure 1 Growth of and digitization of global information storage capacity

2 Information hotspots key technology

- **Big Data Analysis technology**

Techniques for analyzing data, such as A/B testing, machine learning and natural language processingBig Data technologies, like business intelligence, cloud computing and databases

Visualization, such as charts, graphs and other displays of the data, Multidimensional big data can also be represented as tensors, which can be more efficiently handled by tensor-based computation, such as multilinear subspace learning.[5] Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data mining, distributed file systems, distributed databases, cloud-based infrastructure (applications, storage and computing resources) and the Internet.

DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets and in 2008 the technology went public with the launch of a company called Ayasdi.

The practitioners of big data analytics processes are generally hostile to slower shared storage, preferring direct-attached storage (DAS) in its various forms from solid state drive (Ssd) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—Storage area network (SAN) and Network-attached storage (NAS) —is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.

Real or near-real time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in memory is good—data on spinning disk at the other end of a FC SAN connection is not. The cost of a SAN at the scale needed for analytics applications is very much higher than other storage techniques.

There are advantages as well as disadvantages to shared storage in big data analytics, but big data analytics practitioners as of 2011 did not favour it.

- **Cluster analysis technology**

Cluster analysis technology or clustering technology is used for public opinion analysis. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The

appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results.

- **Web crawlers**

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are ants, automatic indexers, bots, Web spiders, Web robots, or—especially in the FOAF community—Web scutters.

This process is called Web crawling or spidering. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

- **Text segmentation**

Text segmentation is the process of dividing written text into meaningful units, such as words, sentences, or topics. The term applies both to mental processes used by humans when reading text, and to artificial processes implemented in computers, which are the subject of natural language processing. The problem is non-trivial, because while some written languages have explicit word boundary markers, such as the word spaces of written English and the distinctive initial, medial and final letter shapes of Arabic, such signals are sometimes ambiguous and not present in all written languages.

3 Information Hotspots system development

3.1 The technology of information hotspots system structure. The technology of information hotspots system structure includes fives layers, which is user layer, application service layer, public opinion layer, information processing layer, network information collection layer. The users can be agriculture policy makers and agriculture producers. The application service includes information analysis, information presentation, information special report and information statics. The public opinion analysis layer includes full text retrieval, topic discovery, automatic clustering, negative information judgment and information track. Information processing layer includes information content management platform, automatically remove duplicate information, automatically update and automatically abstract. Network information collection layer includes directional, whole network acquisition, multi thread technology, Internet web, news, blog, and forum. The technology of information hotspots system structure is shown in Figure 2.

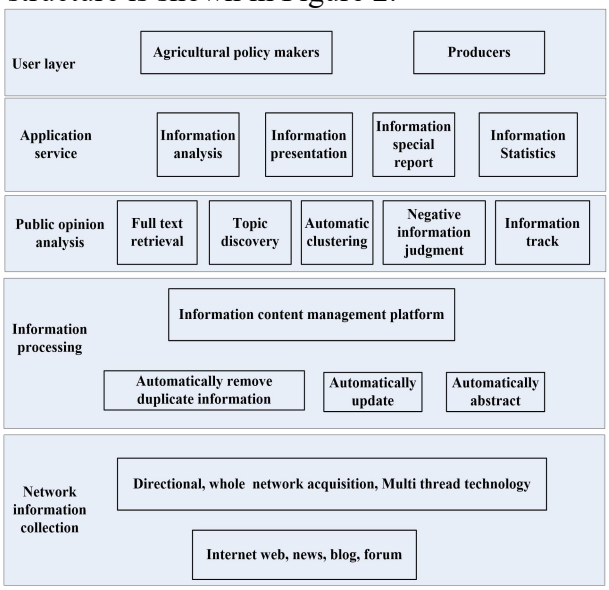

Figure 2 the technology of information hotspots system structure

3.2 Information hotspots system development Environment. System mainly uses the J2EE (Java 2 Platform Enterprise Edition) architecture, java programming language, MyEclipse integrated development environment, and using struts, spring and Hibernate framework. The middleware for the

system is Tomcat. In the development stage, the system uses the Windows platform, combined with Fox Fire Bug Fire plug-in for Bug debugging. The database design and development system is Oracle11. The software version management tool in the development process is SVN. The system integrated development environment MyEclipse page map is shown in Figure 3.

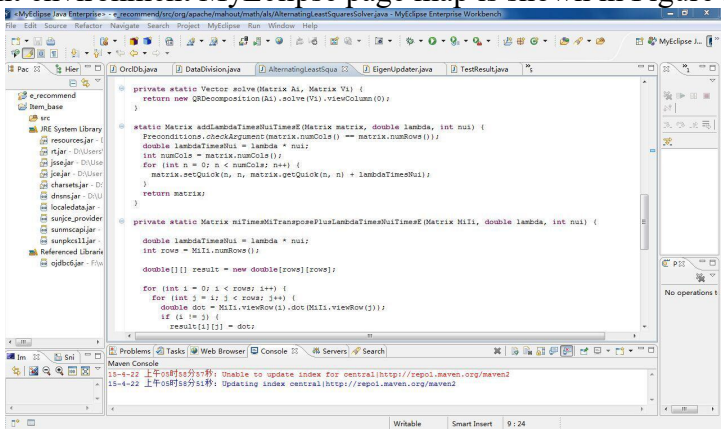

Figure 2 Information hotspots system integrated development environment MyEclipse

4 Test results and conclusion

The paper discusses key technologies of agricultural information hotspots based on big data analysis. The method of information access and analysis technology is given, and the agricultural information hotspots system is developed and discussed. When information hotspots system is finished, we start to test the system. The result shows that the system can get useful information from large amount of disordered information. The agricultural information hotspots system can provide value information for agricultural policy makers and producers.

Acknowledgment

This work was sponsored by Beijing Science and Technology Program Support (Project No. Z151100002115029), which is development of intelligent multi terminal system for modern distance learning in rural areas, and supported by Distance Learning Innovation Team Project of Beijing Academy of Agriculture and Forestry Sciences (Project No. JNICST201612).

References

[1] The World's Technological Capacity to Store, Communicate, and Compute Information. MartinHilbert.net. Retrieved 13 April 2016.

[2] Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). Challenges and Opportunities of Open Data in Ecology. Science 331 (6018): 703–5. doi:10.1126/science.1197962. PMID 21311007.

[3] Manyika, James; Chui, Michael; Bughin, Jaques; Brown, Brad; Dobbs, Richard; Roxburgh, Charles; Byers, Angela Hung (May 2011). Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Retrieved January 16, 2016.

[4] Future Directions in Tensor-Based Computation and Modeling. May 2009.

[5] Lu, Haiping; Plataniotis, K.N.; Venetsanopoulos, A.N. (2011). A Survey of Multilinear Subspace Learning for Tensor Data. Pattern Recognition 44 (7): 1540–1551. doi:10.1016/j.patcog.2011.01.004.