The Application of Big Data Analysis Techniques and Tools in Intelligence Research

Mengru Li[1,*], Hong Fu[1], Ruodan Sun[1] and Che Che[4]
[1]Beijing Institute of Science and Technology Information, Beijing, China
[4] Central University of Finance and Economics, Beijing, China
[*]Corresponding author

***Abstract*—The advent of big data era has brought opportunities and challenges to intelligence research. This paper analyzes the emerging techniques of intelligence research under the big data environment, like data mining, visualization, semantic processing, etc. Meanwhile it also summarizes some new tools, such as Weka, Sitespace, etc. In order to promote the development of intelligence theory research and practice, it is vital and useful to explore the updating of intelligence research techniques and tools, and to discover the new model of intelligence analysis.**

Keywords-big data; intelligence research; big data techniques; big data tools

I. Introduction

Under the background of big data, technically, computer technology like visualization and data mining provides a powerful technical perspective for intelligence research, and intelligence knowledge, in turn, given guidance to the development of other techniques. Correspondingly, a lot of big data analysis tools are now widely used in fields of military intelligence, science and technology intelligence, also in the cognitive process of intelligence analysis, in order to guide the development of intelligence analysis tools.

In the face of huge amounts of data, automation technology is indispensable. Through a variety of technical means and different tools, revealing the information content and its relationship all-round is necessary, so as to avoid the misreading of information. [1]

II. The Application of Big Data Techniques in Intelligence Research

The development of intelligence research, determined it could not remain on the qualitative analysis or simple statistics. [2] Therefore, the research on intelligence technology puts forward new requirements. The McKinsey Global Institute has released its research report in May 2011: Big data: The next frontier for innovation, competition, and productivity. This report is divided into six parts, including the second part, which discussed the techniques of big data in three aspects: big data analysis techniques, big data techniques and visualization. In big data analysis techniques, 26 analysis techniques that suitable for numerous industries are enumerated, including clustering analysis, crowdsourcing, data mining, natural language processing, network analysis, predictive modeling, regression, visualization, etc. Most of those techniques are the existing techniques; also some of them are developed based on the original techniques with the development of Internet and the demand for large-scale data mining. These techniques can be roughly divided into big data storage and processing techniques, big data query and analysis techniques, big data analysis and visualization techniques three categories. Among them, the first two techniques are the foundation of big data, and the last one is the most frequently used in intelligence analysis domain currently and should be paid more attention and in-depth study. Big data advanced analysis and visualization techniques mainly include the analysis of data mining and advanced analysis, visual analytics and knowledge representation, and semantic analysis.

A. Data Mining and Advanced Analysis

Data mining, generally refers to the hidden information searching process from a large number of data by algorithm. Data mining is often associated with computer science, and through many methods like statistics, online analytical processing, information retrieval, machine learning, pattern recognition and expert system, to implement data analysis and database knowledge discovery. [3] The task of data mining is to find models from large amounts of data. According to the task, data mining can be divided into many types; the most typical examples are correlation analysis, classification analysis based on decision tree or neural network, clustering analysis and sequence analysis, etc.

The core of big data analysis is the data mining algorithm. Based on diverse data type and format, each data mining algorithm reveals the properties of data scientifically. In addition, Because of handling with big data, many common used data mining algorithms, widely accepted by many statisticians, can touch the inner part of data and mine the true worth.

From the prospect of data mining concept, it has the natural connection with information theory. From the aspect of data mining method, it contains the special properties and implementation process, which can be used to solve the problems from information research. However, so many data mining algorithms are only used for simple applications such as counting statistics, common words count, based on current information research results. In the process of knowledge discovery, these simple applications are only data preprocessing, the deep mining is in need. Therefore, data mining is able to apply in information research domain, which is not only the effect of data mining exploration, but also the result of development of information research.

B. *Visual Analytics and Knowledge Representation*

Visual analytics is the technique for relation analysis through interactive visualization with the purpose of facilitating users to make decision and to draw the perfect analysis figures and tables, depending on the large scale data set with distributed information and complicated data structure. Figures and tables show the information for all related events, data analysis process and the trend of data stream. Visual analytics is different from information visualization, which focuses on the figure representation for automatically generated information and the representation design, development and application. Visual analytics develops based on information visualization and focuses on the choice of analysis methods and the combination between analysis methods with visualization techniques in order to obtain the goal of decision making.

Visual analytics, one of hot research topics in information research domain, are able to improve the effect of information analysis significantly when applying it into this domain. Information visualization application overcomes the disadvantage of traditional methods for information research and analyzes information from a new prospect. It reveals, explains and analyzes the hidden information that hard to find using previous methods. It is able to generate the valuable conclusion for decision making, which significantly improve the effect and effort of information analysis. [4]

The users for big data analysis contain big data exports as well as normal users. Both of them need visual analytics as the basic requirement. Because visual analytics is able to directly represent the characters of big data as well as be accepted easily by reviewers just like reading figures.

C. *Semantic Analysis*

Semantic is the science about the meaning. Semantic analysis checks whether there is semantic error in source code in order to collect information for coding generation phase. The process of semantic analysis is to check the context, type for the correctly structured source code. Semantic analysis techniques provide a better way for machine to understand the data description and code, integrating the natural language process, information index, database techniques and other methods with purpose of facilitating computer process, integrate, reuse the structured or unstructured information.

The core techniques for semantic analysis include semantic labeling, knowledge sampling, indexing, modeling, inference and so on. Semantic techniques makes a good foundation for deep mining, which recognizes the potential patterns that hide in the information through the semantic process for different types of information and data mining algorithms for the structured data with the extracted semantics.

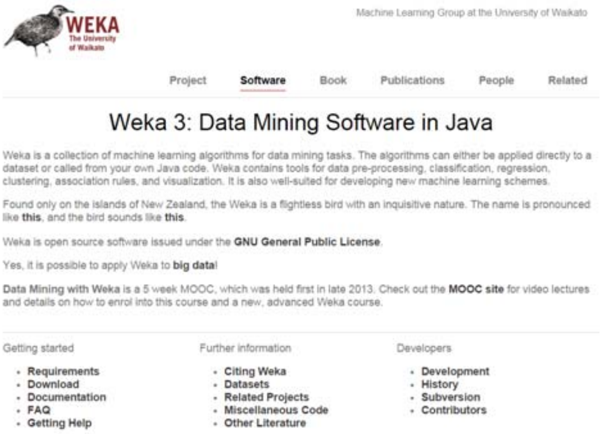
III. THE APPLICATION OF BIG DATA TOOLS IN INTELLIGENCE RESEARCH

Data analysis is the most important part for big data, it helps to make better decisions when finding valuable data from inside. Therefore, based on the above technologies and platforms, application of big data analysis tools in intelligence research can be divided into three types: data mining tools, visual analysis tools and semantic engine tools.

A. *Data Mining Tools*

Due to fewer existing mining tools specifically for intelligence research, the researchers usually use tools from interdisciplinary fields to do intelligence analysis. It leads to redundant work because different tools have different functions. Also, it brings problems that analysis is lack of integrity, which may cause the loss of potential model. Here we mainly introduce two commonly used data mining tools: Weka and RapidMiner.

1) *WEKA, open source software:* WEKA is short for Waikato Environment for Knowledge Analysis, it is a free, noncommercial, open source machine learning and data mining software based on JAVA environment. WEKA works as an open platform for the data mining, brings together a large number of machine learning algorithms that can undertake the task of data mining, including data preprocessing, classification, regression, clustering, association rules, and visualization on new interactive interface.


WEKA
The University of Waikato
Machine Learning Group at the University of Waikato
Project
Software
Book
Publications
People
Related

Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

Yes, it is possible to apply Weka to **big data**!

Data Mining with Weka is a 5 week MOOC, which was held first in late 2013. Check out the **MOOC site** for video lectures and details on how to enrol into this course and a new, advanced Weka course.

FIGURE I. WEKA OPERATOR INTERFACE

In August 2005, on the 11th ACM SIGKDD international conference, WEKA team from University of Waikato won the highest award in the field of data mining and knowledge discovery. WEKA system has been widely recognized, and hailed as a milestone in the history of the data mining and machine learning. It is now one of the most complete data mining tools, and it has been downloaded over ten thousand times per month.

2) *RapidMiner, open source software for data and text mining:* RapidMiner is the world's leading solution for data mining, in a very large extent, owns advanced techniques. Its tasks cover a wide scope, including various data art, which can simplify the design and evaluation of data mining process.

RapidMiner offers free data mining technology and database, 100% Java code, simple process, which is powerful and intuitive. It can process large-scale data with simple scripting language automatically. [5] It shows multi-level data view to ensure the effectiveness of data, and also owns a powerful visualization engine and advanced visualization modeling of high-dimensional data, supporting by over 400 data mining operators.

RapidMiner Studio

Downloadable GUI for machine learning, data mining, text mining, predictive analytics and business analytics.

Easy-to-use visual environment for predictive analytics. No programming required.

Forget sifting through code: RapidMiner is the most powerful, easy to use and intuitive graphical user interface for the design of analytic processes. Let the Wisdom of Crowds and recommendations from the RapidMiner community guide your way.

FIGURE II. RAPIDMINER OPERATOR INTERFACE

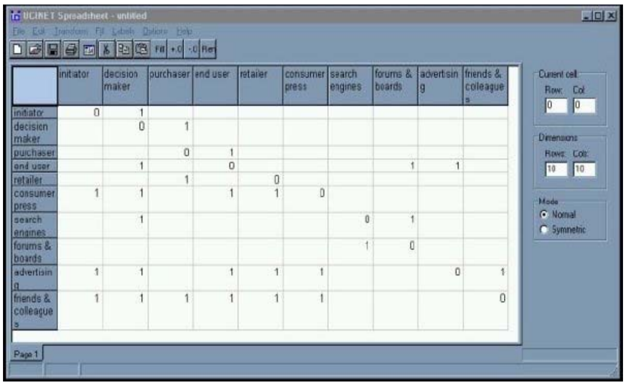
Now, it has been successfully used in many different applications by Yale University, including text mining, multimedia mining, function design, data stream mining, integrated development and distributed data mining.

B. Visual Analysis Tools

Although current intelligence analysis system offers a variety of views to reveal information, it's more of a presentation of analysis results; analyst will never know the analysis method, the limitation and effectiveness of the results. At the same time, the existing analysis tools need to enter all sorts of multifarious parameters to work, while analyst are lack of support for the intelligence analysis of cognitive process, which increasing the difficulty of analysis. However, visual analytics can better solve this problem, it integrates methods in multiple areas including information analysis, geography space analysis and scientific analysis, applies the outcomes in the field of data management and knowledge representation, statistical analysis, knowledge discovery to carry on automatic analysis. It coordinates the linkup between man and machine to better spread, understand and analysis results. Here we introduce four kinds of visualization software: Pajek, UCINET, Jigsaw and Citespace.

*1) Pajek, information visualization software:*Pajek is a specially designed network analysis and visualization program for processing large data sets, it is a helpful tool used to study all kinds of complicated nonlinear network. [6] Pajek runs under Windows environment, it is used for large network systems analysis and visualization operation with thousands or even millions of node. Pajek can handle multiple networks at the same time, 2- mode network and time events network as well, it also provides a longitudinal network analysis tool. [7]

2) UCINET, information visualization software: UCINET is by far the most popular social network analysis software. Social network analysis methods including centrality analysis, subgroup analysis, character analysis, and statistical analysis based on displacement, etc. [8]



	initiator	decision maker	purchaser	end user	retailer	consumer press	search engines	forums & boards	advertisin g	friends & colleague s
initiator	0	1								
decision maker		0	1							
purchaser			0	1						
end user		1		0				1	1	
retailer			1		0					
consumer press	1	1		1	1	0				
search engines		1					0	1		
forums & boards							1	0		
advertisin g	1	1		1	1	1			0	1
friends & colleague s	1	1	1	1	1	1				0

Current cell
Row: 0 Col: 0

Dimensions
Rows: 10 Cols: 10

Mode
● Normal
○ Symmetric

Page 1

FIGURE III. UCINET OPERATOR INTERFACE

UCINET is written by a group of network analysts from University of California, Irvine. UCINET network analysis integration software is including one and two-dimensional data analysis software NetDraw, and three-dimensional display analysis software Mage, which is on the march, and also integrates the Free application software for large-scale network analysis. [7] In addition, the package has a strong function of matrix analysis, such as matrix algebra and multivariate statistical analysis. It is by far the most popular social network analysis software, also it's easy to use and suitable for the beginners.

*3) Jigsaw, document visualization analysis system:*John Stasko, et al. from Georgia Institute of Technology, set up a visualization analysis system called Jigsaw, based on the intelligence analysis conceptual model put forward by Pirolli, and applied it to the field of academic and network research, it also illustrates the feasibility of visualization analysis technology applied into intelligence research.

Jigsaw is a software establishing visual views for all kinds of document by using text mining algorithm. It can generate document clustering figures, timelines, word tree graphs, etc. Modeling is auto-completed according to specific tasks, and it will generate visual chart relatively. Also users are free to adjust the properties and appearance of the chart, it is a typical visualization analysis system.

*4) Citespace, visualization tools:*Citespace and its updated version are drawing tools in knowledge mapping domains, which are widely used by researchers both domestic and abroad. [9] The tool is developed by Dr Chaomei Chen, and it's free to use. Users can get their mapping knowledge domains of a certain direction based on data set, and the knowledge mapping domains is stable, informative and readable.

CiteSpace combined text mining, information visualization and scientometrics creatively, formed visualization techniques suitable for multiple, time sharing, dynamic analysis, and promoted science and technology intelligence research to a new stage on the basis of mapping knowledge domains and knowledge visualization.

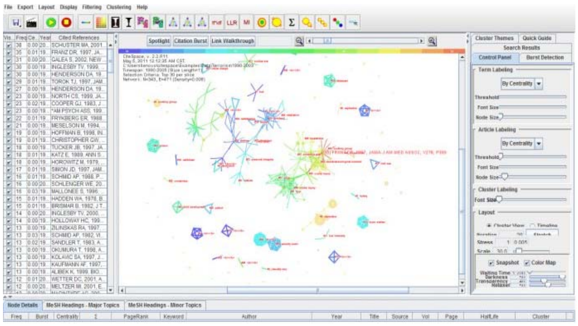


FIGURE IV. CITESPACE OPERATOR INTERFACE

C. *Semantic Engine Tools*

Lack of semantic support is a common problem for existing intelligence research practices and tools. For traditional intelligence research objects, such as scientific papers, patents, etc., many of their analysis tools are lack of deep semantic support. For emerging intelligence research object, such as network news, blogs, etc., are still at the initial state on the whole. Currently, it still relies on analysts to figure out the required information, and collate them into structured data. The existences of above problems make semantic technologies become inevitable.

1) Annotation.cn, large-scale semantic computing platform: Annotation.cn is a multi-domain annotation service platform system, mainly contains three parts: vocabulary management module, tagging module, user management module. Among them, vocabulary management includes four functions: vocabulary maintenance, type management, concept management, concept view; tagging module contains two functions: document management and document annotation; system management includes user management and role management functions.

IV. CONCLUSION

The development of big data, is a positive impetus to the intelligence research and its organization and service work. It not only promote the further changes in field of intelligence research and practice, but also push ahead a higher duty requirement for intelligence work in data management, data analysis, data use and data services. Innovative new techniques and tools are needed now in aid of intelligence research in the era of big data. Therefore, big data techniques and tools have brought both opportunities and challenges to intelligence research in the field of intelligence theory and practice. However, big data intelligence techniques and tools, specifically for storage, processing and analyzing intelligence data, remain inadequate. We should take this opportunity to explore the application of new techniques and methods actively, to integrating, processing, organizing and using big data, to better improve the quality of data service, promote intelligence effect in the field of knowledge management and application.

ACKNOWLEDGMENT

This research was financially supported by Beijing Municipal Social Science Foundation Under Grant No. 16XCC017. “Research on Network Public Opinion Ecology in Wechat Field” and Beijing Municipal Financial Project Under Grant No. PXM2016_178214_000006. “Capacity Building on Intelligence Processing and Analysis Faced to Text Oriented Information”

REFERENCES

[1]. Li Guangjian, Yang Lin. Intelligence Analysis and Intelligence Technology in View of Big Data. Library & Information, 2012 (6): 1-8

[2]. Gu Tao. Research on Collaboration Analysis of Competitive Intelligence Based on Big Data. Information Science. 2013, 31 (12): 114-118

[3]. Tang Zhixiong, Xian Donglai. The Implement Method of Analyzing Customer Retention by Data Mining Technology. Information & Communications. 2011 (2): 99-100

[4]. Tang Tianbo, Gao Feng. Case Study of Visual Analytics in Intelligence Research. Information Studies: Theory & Application. 2009,8 (32): 63-67

[5]. Liu Zhilong. Data Analysis and Data Mining Application in Statistics Industry. Statistics and consultation. 2014 (1): 36-38

[6]. Zhou Qingshan, Zhao Xue, Zhao Xuyao, Zhou Gefei. Knowledge Mapping Analysis in Digital Content Industry in China. Information Studies: Theory & Application. 2014, 35 (4): 56-61

[7]. Fei Zhonglin, Wang Jingan. Social Network Analysis: Method and Perspective of Management Research. Science and Technology Management Research. 2010 (24): 216-219

[8]. Li Gang, Li Ang. Study on Coauthorship Based on Social Network Analysis. Journal of Information Resources Management. 2011 (3): 43-47

[9]. Liao Shengjiao. The Comparative Study on the Scientific Knowledge Mapping Tools: VOSviewer and Citespace. Sci-Tech Information Development & Economy. 2011, 21(7): 137-139