Modern Applied Science; Vol. 11, No. 4; 2017
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

The Study of Semantic Analysis on Intelligence Research under the Environment of Big Data

Hong Gu[1] & Hongwei Yuan[1]

[1] Nanjing University of Aeronautics and Astronautic, China

Correspondence: Hong Gu, Library of Nanjing University of Aeronautics and Astronautic, Nanjing, China. Tel:15851863750. E-mail: guhongnuaa@163.com

Received: November 30, 2016 Accepted: December 23, 2016 Online Published: January 24, 2017

doi:10.5539/mas.v11n4p1 URL: http://dx.doi.org/10.5539/mas.v11n4p1

Abstract

Faced with complex, large mass of data, how to find the information we need from these data, then to do intelligence research, it is an issue of concern in the intelligence community. This paper analyzes the significance of research and three technologies to ensure the rigor of intelligence research: visualization, data mining and semantic analysis technology, focuses on the semantic analysis technology in the application of intelligence research, exemplified by the semantic role annotation and semantic-based text orientation analysis of two methods, described the meaning of these two methods, the semantic database, the basic flow of information, their strengths and weaknesses, as well asdevelopment and raised its outlook in information research.

Keywords: bigdata, semantic analysis, intelligence research

1. Introduction

With the development of cloud computing, networking, online social media and other emerging technologies, more and more data appear in our life with the explosive growth. All these massive data mark the arrival of the era of Big Data. In the era of Big Data, various digital information appears broadly. What the problem we should focus more on is how to get more, better, more accurate and more abundant data from so many data. Researches on Information Science on the era of Big Data have also had a lot of changes. The research object, research environment, research methods and tools in traditional Informatics present a qualitative leap under the environment of big data and generate many new areas. Many traditional research methods have been unsuitable for researching in a new context. Therefore the intelligence workers must find new ways to meet research needs under the environment of big data.

2. The Influence of Big Data to Intelligence Research

2.1 The Concept of Big Data

Big data suggests large data sets. The research institution, Gartner, defines Big Data: a kind of information asset with abundant and high growth rates needing new treating mode to have a stronger decision-making power, insight and capabilities of process optimization. The definition of big data in Wikipedia is: the information that the amount of data is too large to acquire, manage, handle and clear up for helping companies to decide by mainly present software tool (Big Data, April 2016). McKinsey defines big data as the data collection that we can not use traditional database software tool to collect, store, manage and analyze in certain time.

During the understanding of big data, it is not only a simple number concept, but also its great complexity, such as data-growing continuously,a wide range of data, data-exchanging frequently and complex relationship.

According to the 2011 World Forum of Big Data, big data held 4V features of huge numbers of data, various types of data, low value density, fast processing speed (Liu, H. X., & Bai, W. H. 2014).

2.2 The Development of Intelligence Research in Big Data Environment

These features of big data give intelligence research new development environment and promote further a variety of Internet technology, data mining technology and cloud computing technology more mature, resulting in increasing the forecasting components of the trends of information in the field of intelligence research and putting data analysis into an unprecedented height. Enhancing cross-border cooperation between the information and data diversity also contribute to the improvement of the usage of data integration. As a result, under the

background of big data, intelligence analysis has become the consensus of the whole society and has been more comprehensive, diverse and complex. All of these points not only provide more requirements for the concepts of intelligence officers, but demand higher improvement of intelligence analysis methods .

The intelligence research activity is a kind of sensemaking (Gary, K. et al. 2006) at the macro level. It depends on the analyst who can build cognitive frame based on existing knowledge and get understanding through constant modification by cognitive framework(Gary, K. et al. 2006). So intelligence research depends on personal judgment of specialist, with great uncertainty to a certain extent. How to ensure the rigor of intelligence activities, reduce uncertainty of intelligence and improve the quality of intelligence activities have been more and more concerned by Intelligence Community. Under the background of big data, the requirement of keeping intelligence research rigor is not only reflected in the concepts, but also reflected in the techniques. While in traditional intelligence research process, using automatic tools is essential and we try our best to avoid the people's initiative to some extent, to , however, when facing the same problem, we can use different tools or hold different ideas to get different results. From the side, we have not yet overcome technological prejudice. So on the era of big data, we need to use a variety of techniques to ensure the rigor of intelligence research.Intelligence analysis under the environment of big data is in the following figure:

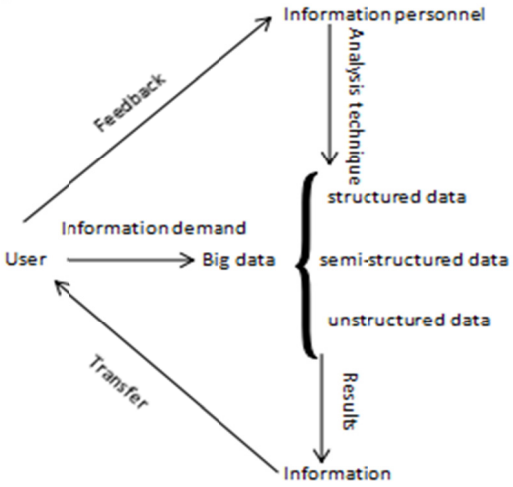

Figure 1. The process of intelligence analysis

Under the background of big data, the concern and deep-researching techniques in the current intelligence area are roughly divided into three types: visual analysis technique、data mining technology、semantic analysis technique.

3. The Study of Semantic Analysis on Intelligence Research under the Environment of Big Data

3.1 The Concept of Semantic, and the Application of Semantic Analysis in the Era of Big Data

Semantics is the science of meaning. Semantics represents not only the nature of things, but also cause and effect of things, hypernym-hyponym, agent and other logical relationship. Semantic analysis is to identify information containing semantics(Qin, Zhu, Zhao & Zhang, 2014). The semantic technology provides better processing or machine-understandable data description, procedures and infrastructure(John. Domingue, et al. 2011). Semantic technology integrates WEB technology with artificial intelligence, machine learning, natural language processing, information retrieval and other computer technology methods. It is designed to allow computers to collect and integrate structured and unstructured data better. Semantic technology can provide better technical support for semantic analysis, and lay the foundation for deep mining of data.

During the traditional intelligence research, information analysis has been always lack of semantics support. In the analysis of some structured data such as papers, reports, patents and other literature, there are some mature technologies for extracting some key words from the summary of the literature to reflect the theme of the literature, but in data processing, the relevance and proximity of words could not be acquired because of no way to distinguishing synonyms, which needs an artificial distinction between literature themes. These data sets in a small environment between words are still acceptable.

But with the advent of Internet technology and big data , the cardinal number of data increases, and data generate around our life all the time. Our behaviors of purchase, bank transactions and other acts are presented in the form

of data and the data are growing exponentially. The development of Internet technology not only promotes the generation of text data, but also presents us unstructured data with images, audio, video and other types. These contents contained in different types of data are different, so are the structures and storage. On the era of big data, data generate very fast, it's difficult for us to find valuable information among our redundant data, like looking for a needle in a haystack, embodying the features of low value of the density of data. In this case, if still relying on traditional semi-automatic techniques for data processing integration, we will feel obviously powerless. And for the new types of data tracking and monitoring, although there has been some technologies, we still inevitably need human intervention. Therefore, semantic analysis on the era of big data has become an inevitable trend and necessity.

Now there are many development of semantic analysis, we will highlight semantic role labeling and Semantic-based Textual Analysis study.

3.2 Semantic Role Labeling

The substance of semantic role labeling is to have shallow semantic analysis to the sentence level, divide words into groups in a sequence according to the current knowledge of grammar and classify them in accordance with the semantic roles, which does not make the whole sentence a detailed semantic analysis, but only mark semantic roles (parameters) of given predicates (verbs, nouns, etc.) in order to have a "shallow" understanding of computer words (Yang & Zhang, 2010). Semantic role labeling is a mixture of natural language processing techniques, such as segmentation, POS tagging, syntactic analysis, so the study of semantic role labeling also provide a research platform for the study of machine learning methods and the underlying technology(Li, Sun & Li, 2011). Definite task-analyzing and convenient evaluation are its advantages. Semantic role labeling in many applications has played a significant role, it can be used to quiz systems, information retrieval, machine translation, automatic abstracting and other natural language processing.

Making semantic labeling needs a good support from semantic data bank. At present, the well-known English semantic database FrameNet, PropBank, NomBank. FrameNet is developed by U.C.Berkeley who labels British National Corpus based on the theory of semantic framework. It tries to describe each predicate (verb, noun and adjective partial) and attempts to describe the relationship between these frameworks. PropBank is that UPenn who labels shallow semantic information based on Penn TreeBank syntactic analysis. The difference between FrameNet and PropBank is that PropBank only marks verbs (non-verb), and correspondingly the verb is called the target verb.Semantic Roles in PropBank are divided into two categories: one core semantic role---ARG0-ARG5, the other modifying semantic role---ARGM. Typically ARG0 is the agent of the predicate action, ARG1 is the object of predicate action, ARG2-ARG5 have different meanings in different semantic frameworks. Different from the verbs being as predicates marked in Penn TreeBank by PropBank, NomBank just marks the nouns being as predicates, but categories of parameters and expression are the same as PropBank. It was developed in order to compensate the situation that PropBank only uses verbs as predicates have disadvantages of rough labeling.

The study of Chinese semantic labeling mainly uses three resources: Chinese Proposition Bank（CPB），Chinese Nombank (Xue, 2006), Chinese FrameNet(You & Liu, 2005). These are all resources to have shallow semantic labeling to Chinese.

With semantic database as a foundation, the next step is to learn from these existing semantic database and automatically carry out semantic role labeling to various resources. the basic unit of the automatic labeling of semantic role labeling system is syntactic constituents, phrases, words or dependencies. Generally thinking, each semantic role and a syntactic component is corresponding. It also means that a syntactic component corresponds to a semantic role, but the converse is not all true. Syntactic component mainly uses for the system of semantic role labeling based on the role of phrases, but now the majority of semantic role labeling systems tend to use syntactic component as the basic unit which is better used in English environment.But in other language environment, it is difficult to get the result automatically of this deep syntactic analysis. The current syntactic analysis system is in poor performance in the field of general. For this reason, someone tried to make deep semantic analysis base on shallow syntactic analysis. After all, the applicability of shallow syntactic analysis is better than the deep syntactic analysis (Li, Sun & Li, 2011). Hacioglu used dependency as the basic unit to do semantic role labeling(Hacioglu, 2004) and achieved a similar effect. Some scholars also tried to use words as units on semantic role labeling analysis, but the results are not as good as that in using the phrases and sentence compositions as labeling units.

The first step is to do syntactic analysis in semantic role labeling, and then identify the predicates on the basis of syntactic analysis. The labeling process of semantic role labeling system is divided into four steps: pruning,

identification, classification and post-processing.

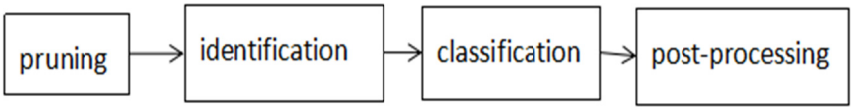

Figure 2. The process of semantic role labeling

Pruning is the basis for semantic role labeling, which is the equivalent of pre-processing of information in the process of systematical labeling. Facing more complicated big data, we should carry out a pre-processing of these data if we want to get the data we need. Pruning is a kind of method that is trying to filter the data which can not be used as some syntactic elements of semantic roles in order to improve efficiency and accuracy. Identification is trying to judge whether these labeling units are semantic roles after the pruning step and keep those labeling units which have been identified as semantic roles. Classification is the step after identification. It can reduce the number of instances that are into the classification and then speed up processing. Post-processing is to process these issues of some inherent constraints of these ingredients which have been classified in semantic role labeling.

While semantic role labeling analysis can address the needs of intelligence to some extent under the environment of big data, it relies heavily on syntactic analysis to a larger extent. It is generally thought that using syntactic analysis in more mature environment such as in English instead of in Chinese is better, but semantic role labeling is base on database and its capacity of cross-discipline is not very good, so the adaptation should be thought more by us. On the other hand, independent system of semantic role labeling analysis is difficult to obtain satisfactory results. The integration of a variety of semantic role labeling system could improve accuracy but we inevitably face the dilemma of complex system structure and low efficiency. So the usability for semantic role labeling in the field of intelligence research still needs our further study.

3.3 Semantic-Based Textual Tendency Analysis

Since entering the Web2.0 era of the Internet, people are no longer just accept information from the internet passively, they are more likely to tent to express ideas on the platform of the internet. More and more people participate in the process of generating information. Everyone’s views or attitude are not same, thus leading to diverse information. A large number of personal views are covered in the network, mostly semi-structured or unstructured comments, such as the discussion about social problems in news comments and blog sites or the evaluation to goods in electronic commerce. These comments will increase gradually as time goes on. With such rapid increase of comments, how to organize these unstructured or semi-structured information, how to identify subjective attitude, appraise tendency in these comments , how to obtain intelligence information we want are all the problems we need to solve. Text orientation analysis is generated in this background.

Text Orientation analysis, being known as sentiment analysis, is a process of analyzing, summarizing, handling and concluding to subjective text with emotional colors. Text orientation analysis will judge whether people's sentiment is positive or negative by tapping people's comments. These analysis is significant to analyzing people’s purchasing in electronic commerce and public’s attitudes to public opinions. In the field of natural language processing, it pours new vitality into text selection, automatic abstract, subjective and objective classification and other traditional natural language processing technologies.

The basic process of text orientation analysis is divided into four parts: Collection of raw materials, text pre-processing, subjectivity of text recognition and text orientation identification. Collection of raw materials refers using some reptiles machine in a network to collect network information on some social networking sites as well as some e-commerce websites, we need to collect more comprehensive information as possible as we can to be the basis of our text orientation analysis. Text pre-processing is to clear up the information we collect, eliminate information noise (useless information), classify and filter the information, and let a lot of unstructured or semi-structured data be structured. Subjectivity of text identification is to use database to identify the subjective words and weed out word that do not contain the emotional objective text in order to improve the accuracy of the analysis of text orientation. Text orientation identification is targeted at the extracting subjective words combining corpus, and use simple statistical methods or the way based on machine learning or the way based on relativity to judge the tendency of appraise of the text.

Text orientation analysis includes semantic-based text orientation and machine-based text orientation. We mainly

discuss the study of semantic-based text orientation. Because the semantics tendency is for adjectives and embodies an evaluated tendency, this method is less applied to the theme of text that mainly embodying evaluated text classification of views tendency. The research methods of text orientation analysis on the basis of semantics have two kinds: the first one is to extract adjectives in text or phrases reflecting subjective colors, then judge these extracted adjectives or phrases one by one and give them tendency a alignment, and finally add up all of the above alignments to give the overall tendency of the text. The second one is a pre-established tendency semantic model library, sometimes with a tendency dictionary, then match the testing document with reference to the semantic pattern library, and obtain the tendency of the whold text by accumulating all alignments corresponding to matching models(Yang, 2009).

HowNet is a common repository whose describing objects are the representative of Chinese and English words and that reveals the relationship between the concepts and the nature of concepts(HowNet April 2016). It is a Chinese semantic database major in semantic-based analysis of text orientation. In HowNet, the description of Chinese words is based on the concept of "original meaning", which can be considered the most basic Chinese, and can not be subdivided to the smallest semantic units. Since the meaning of the Chinese words is complex, the same word may have different meanings in different circumstances. So in HowNet, the meaning of the Chinese words could be understood as a collection of a number of items. In HowNet's semantic dictionary, each record is made up of a meaning item of a word and its description. Due to the similarity and correlation of semantics, calculation programs obtain the relative value of number. In HowNet, there is a polarity words, it refers to each word giving a metric of the semantic orientation whose size is related to the degree of association of the words and paradigm words. Paradigm word refers to a kind of word whose appraise attitude is very clear, strong, and representative. The closer the relationship with commendatory paradigm words,the stronger the commendatory tendency of the words, the closer the relationship with the derogatory paradigm words, the stronger the derogatory tendency of the words.

Semantic-based text orientation analysis is applied to many fields of intelligence research. Under the environment of big data, people not only want to acquire some wanted intelligence, but also need some information about the performance of attitudes so that they can make certain decisions by the attitudes. But semantic-based text orientation analysis also has some drawbacks. For example, although available semantic emotional library can be found in English, mature spread of more widely emotion library is comparatively fewer. However in Chinese, since the tendency of word sentiment is too complex, there are many problems to establish Chinese semantic emotional library. Therefore, semantic-based text orientation analysis in Chinese should be developed more mature and the analysis of Chinese emotional tendency should be study further.

4. Conclusion

Under the environment of big data, intelligence research faces many opportunities and challenges. Although the large amount of data bring more information for intelligence research, the good and bad information give intelligence research a major challenge. The development of semantic analysis technology is a good direction for intelligence research. Semantic role labeling and semantic-based text orientation analysis have yet to be developed. However, with the development of words and semantic technologies, these technologies will become increasingly favored by intelligence officers. Bear in mind that we must study the universal of semantic database and some analysis tools, so that it can be applied to more areas and disciplines.

References

Big Data. (April 2016). Retrieved from http://en.wikipedia.org/wiki/Big_Data

Gary, K. et al. (2006). *Making sense of sensemaking1: Allernative Perspectives.* Intelligent System, 21(4), 70-73. http://doi.ieeecomputersociety.org/10.1109/MIS.2006.75

Gary, K. et al. (2006). *Making sense of sensemaking2: A Macrocongnitive Mode.* Intelligent System, 21(5), 88-92. http://doi.ieeecomputersociety.org/10.1109/MIS.2006.100

Hacioglu, K. (2004). *Semantic role labeling using dependency trees*. COLING 04 Proceedings of the 20th international conference on computational. Linguistics. Geneva: Association for computational Linguistics, pp. 1273-1281. http://dx.doi.org/10.3115/1220355.1220541

HowNet (April 2016). *HowNet's Home Page.* http://www.keenage.com

John, D. et al. (2011). *Handbook of Semantic Web Technologies.* Springer Publishing Company. http://dx.doi.org/10.1007/978-3-540-92913-0

Li, Y. G., Sun, F. Z., & Li, J. B. (2011). *Summary of Semantic Role Labeling*. Journal of Shandong University of

Technology (Natural Sciences Edition), 6, 19-24. http://dx.doi.org/10.3969/j.issn.1672-6197.2011.06.004

Liu, H. X., & Bai, W. H. (2014). *Applied Information Science Researchunder the Background of Big Data.* vol.01, pp.27-30. http://dx.doi.org/10.3969/j.issn.1002-0314.2014.01.004

Qin, C. X., Zhu, T., Zhao, P. W., & Zhang, Y. (2014). *Research Process of Semantic Analysis in Natural Language.* Library and Information Service, 22, pp.130-137. http://dx.doi.org/10.3969/j.issn.1672-6197.2011.06.004

Xue, N. (2006). A Chinese semantic lexicon of senses and roles. *Language Resources and Evaluation, 40*(3), 395-403. http://dx.doi.org/10.1007/s10579-007-9025-9

Yang, T. M. (2009). *Analysis and Applied Research of Text Orientati on Based on Semantics*. Jiangsu University. http://dx.doi.org/10.7666/d.y1604280

Yang, X. X., & Zhang, L. (2010). *Information extraction based on semantic role and concept graph*. Computer Applications, 2, 411-414.

You, L. P., & Liu, K. Y. (2005). *Building chinese framenet database.* Mrques k. Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE). NewYork: IEEE, pp. 301-306. http://dx.doi.org/10.1109/NLPKE.2005.1598752