Bibliometric Analysis of the Semantic Mining Research Status with the Data from Web of Science

Mao Meixin 452368828@qq.com
National University of Defense Technology
Li Zili zilili@163.com
National University of Defense Technology
Zeng Licrack521@163.com
National University of Defense Technology
Zhao Zhao z_costa@163.com
National University of Defense Technology
Zhao Yang zyhellon@163.com
College of Advanced Interdisciplinary Studies, NUDT

Introduction

According to the data from Web of Science database, the first paper on semantic mining topic was published in 1991, but Shanon B[1] thought that computers could not display the main characteristics of human consciousness. Because psychological theory couched in terms of semantic representations and the computational operations associated with them is bound to be inadequate. The phenomenology of consciousness is a specific case marking this inadequacy.

With the rapid development of Internet technology, the amount of interactive resources and information on the network is increasing exponentially, but the expansion of information brings people the lack of resources. Because the amount of information is growing, it is even more difficult to find valuable information for users in the huge amount of information. This leads to data mining based on network. The useful information will be automatically extracted from the web document.

Data mining is an advanced process that extracts potential, effective and understandable patterns from massive data according to established goals. The process usually includes problem definition, data extraction, data preprocessing, knowledge extraction, knowledge assessment and so on (2001) [2].

Semantic mining is a new data mining technology that accurately extracts useful information and knowledge from unstructured data. It uses intelligent computing based on semantics to realize the collection of unstructured information and to dig valuable information from it (2008) [3]. The main task of semantic mining is knowledge discovery, exploring potential and interesting knowledge from the semantic database that has described concepts, attributes, and attribute values(2011) [4].

To analysis the research status of semantic mining must be an interesting but important thing.

Data and methods

The bibliographic records used for analysis were collected from the Web of Science database. The records retrieved indicates that in the research field of Semantic Mining, few studies were conducted by using the methods such as Bibliometric, scientometrics, mapping knowledge, and so on, nor by using the visualization analyzing tools such as CiteSpace. So, some novelty could be gained in this paper by analyzing the research status in the semantic mining domain with CiteSpace, which may help those semantic mining researchers clarify the developing trends, explore the research hotspots and fronts, and determine their future research orientation.

After data collection, deduplication and other operations, an analysis as regard to geographic distribution of scholars was mapped by Google Earth, network analysis of different type entities such as countries/territories, institutes, categories, highly cited references, highly cited authors and keywords was conducted by the scientometric software CiteSpace which created by Chaomei Chen [5]. Finally, we try to predict the number of papers using the logistic curve model.

Conclusion and Discussion

According to the analysis of the present situation, the research hotspots and fronts of semantic mining, we can get six conclusions as follows:

First, from the historical document volume, since 2005, the amount of writing has increased exponentially and reached its peak in 2015. At present, the relevant knowledge is mainly distributed in computer science, engineering and linguistics. Second, on the aspect of the knowledge flow, the major sources of semantic mining knowledge are Mathematics, Biology and Psychology. And most of those source are flowing to the subjects such as Computer, Systems, Biology. Third, on the aspect of the high influential authors, Salton G, Agrawal R are the most influential authors, which can be considered as the key experts. In addition, the highly influential authors are almost come from information retrieval and data mining research. Fourth, on aspect of the international cooperation, the academic communications are pretty prosperous, which are concentrated on three major region: East Asia, North America, and West Europe, and the academic cooperation between the United States and Europe are much more intense. Fifth, on the aspect of the research hotspots, the research of technology, theory and algorithm, and application are the key points of the semantic mining research. At last, the current semantic mining research fronts can be categorized into two layers: the model research by using deep learning technology for semantic mining, the application research such as applying semantic mining to social media.

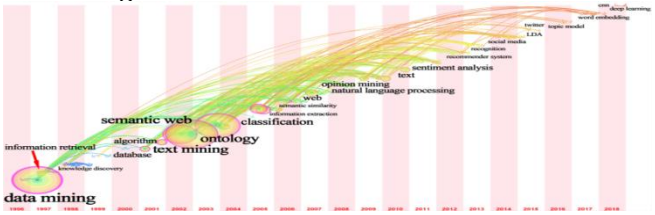

Fig. 1. Research Hotspots Evolution Mapping Knowledge Domain

Table 1. Time Sequence of Research Hotspots — Key Technologies

Category	Keyword(Time, Frequency)
Theory	ontology (2003, 225), gene ontology (2008, 4), domain ontology (2014, 6)
Method	machine learning (2003, 33), association rule (2000, 28), ontology learning (2014, 9), topic modeling (2016, 5), uml (2013, 5), owl(2011, 4)
Algorithm	latent semantic analysis (2003, 52), latent dirichlet allocation(2015, 6)
Mining	text mining(2002, 254), data mining(1996, 244), opinion mining (2009, 55), web mining (2003, 53)
Analysis	sentiment analysis (2011, 46), semantic analysis (2005, 23), formal concept analysis (2011, 4)
Classification	clustering(2005, 58), document clustering (2006, 10), categorization(2007, 5)
Extraction	information extraction (2005, 51), feature extraction (1997, 9), knowledge extraction (2014, 8)
Processing	natural language processing (2009, 50), semantic annotation(2009, 20), integration (2011, 13), relevance feedback (2005, 6)

We also discussed how to use mathematical models to predict the number of papers in the future. In 1981, Little A.D. [6] found that the development of technology and biological evolution had an amazing similarity. Then, S curves were introduced to describe the development of Technology.

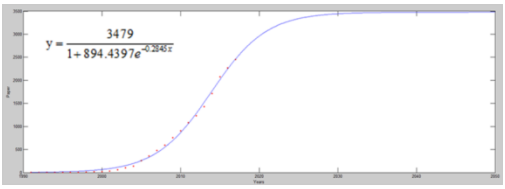

Fig. 2. Logistic curve of semantic mining papers

Fig.2 shows the embryonic stage of semantic mining research is in 1991-2006, the inflection point is in 2014; after 2022, with the technique of iterative upgrade, research will be entering the mature stage, and will enter the saturation period in 2030. At present, the study is still in the growth stage. It has great research value and needs to grasp the golden age of the next five years.

References

1. Shanon, B. (1991). Consciousness and the computer: a reply to henley. Journal of Mind & Behavior, 14(1), 48-55.
2. Zhong, Xiao., Ma, Shaoping., Zhang, Bo., Yu, Ruizhao. (2001). Data Mining: A Survey. Pattern Recognition and Artificial Intelligence.
3. Wang, W. (2008). Research on semantic mining-based intelligent competitive intelligence system. Information Studies Theory & Application.
4. Yang, Jie. (2011). Research on Semantic Web Mining Based on Ontology and Algorithm Apriori. (Doctoral dissertation, Taiyuan University of Technology).
5. Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. John Wiley & Sons, Inc.
6. Little, A. D. (1981). The strategic management of technology. Cambridge Harvard Business School Press.