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

Mao Meixin, Zili Li, Zeng Li, Zhao Zhao and Yang Zhao

September 8, 2018

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

Meixin Mao[1], Zili Li[2], Zhao Zhao[2], Li Zeng[2], Zhao Yang[2]

[1] College of System Engineering, NUDT, Changsha 410073, China
[2] College of Advanced Interdisciplinary Studies, NUDT, Changsha 410073, China
452368828@qq.com zilili@163.com crack521@163.com

Abstract. By using the 2460 papers obtained from the Web of Science database from 1991 to 2018 as the research sample, this paper demonstrates a comprehensive bibliometric analysis of the research status, trends and hotspots in the domain of Semantic Mining. The results indicate that the current global semantic mining research is of great value; Knowledge is mainly distributed in computer science, engineering and linguistics; the international academic communications in semantic mining field are pretty prosperous, which are concentrated on three major region: East Asia, North America and West Europe. In addition, the research hotspots be shown in keywords co-occurring mapping is the research of technology which is represented by text mining, the research of theory which is represented by ontology and semantic network, and the research of application which is represented by knowledge discovery and information extraction. And the current 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. Finally, we discussed to use the mathematical models of logistic curve to predict the number of papers in the future which told us the study is still in the growth stage at present and we need to grasp the golden age of the next five years.

Keywords: Semantic Mining, Mapping of Knowledge Domain, CiteSpace

1 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.

2 Data and methods

2.1 Data Collection

The bibliographic records used for analysis were collected from the Web of Science database, and the specific search strategy is as follows: "TS = (semantic mining) And TI = (overview OR review OR summary OR observe OR assessment OR evaluation OR commentary OR remark OR comment OR current situation OR tendency OR trend) And TS= (bibliometrics OR scientometrics OR mapping knowledge domain OR citespace)". The records retrieved indicates that in the research field of Semantic Mining, few studies were conducted by using the methods such as Bibliometric analysis, scientometrics analysis, 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.

2.2 Methods

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.

3 An analysis of the present situation of semantic mining

3.1 Time and subject distribution analysis

From the historical document volume, semantic mining research has been developing slowly for a long time. Since 2005, with the enhancement of the research strength, the promotion of attention and the breakthrough of technology, the amount of writing has increased exponentially and reached its peak in 2015. The related disciplines have a gradual expansion process. At present, the relevant knowledge is mainly distributed in computer science, engineering and linguistics.

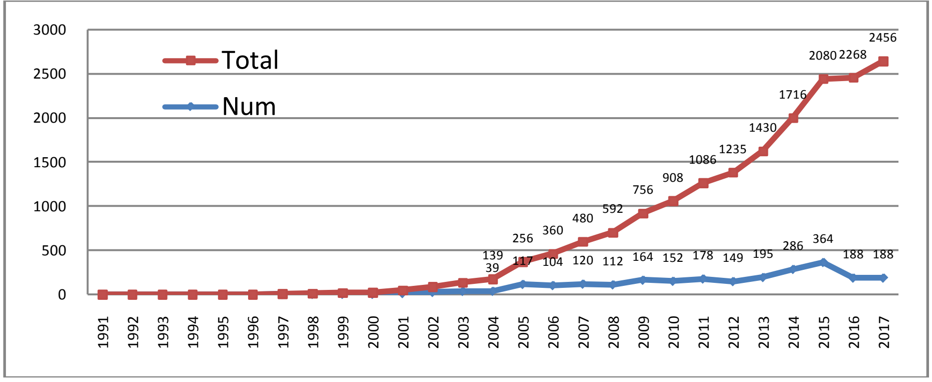


Fig. 1. Number of semantic mining papers (1991-2017)

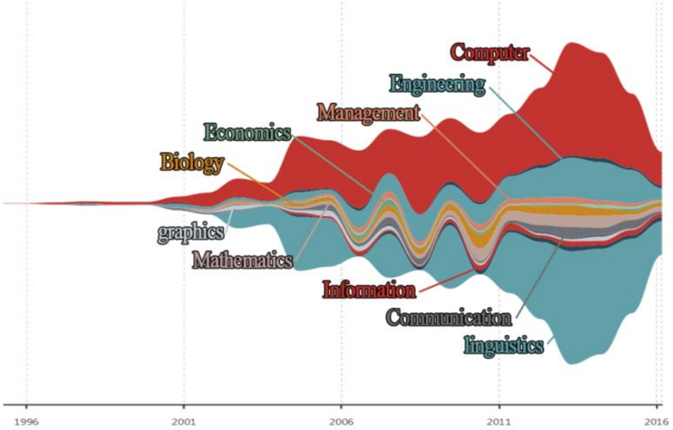


Fig. 2. River chart of subject distribution (1991-2017)

3.2 Knowledge Flow Analysis

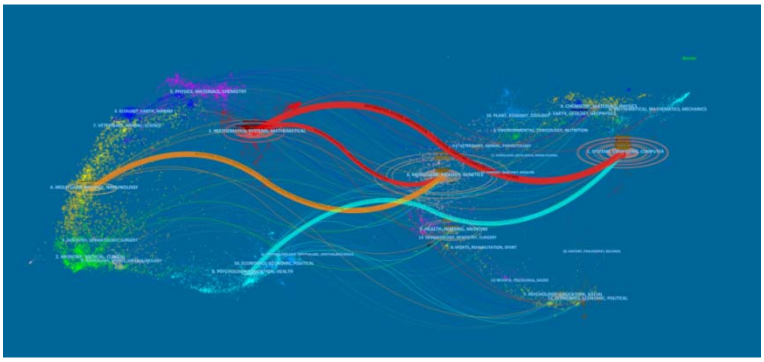


Fig. 3. Subject Categories Dual-map

Fig.3 shows the dual-map overlay of publications in Semantic Mining. The three major subject of knowledge source are Mathematics, Biology and Psychology. Most of those source are flowing to the subjects such as Computer, Systems, Biology and other sources include Medicine, Psychology and so on.

3.3 High Influential Authors Analysis

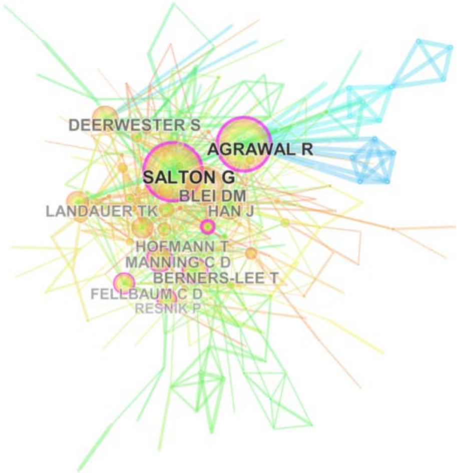


Fig. 4. Authors Co-cited Network

According to the Fig.4 and Table 1, some researchers such as Salton G, Agrawal R are the most influential authors in the field of semantic mining, with both high influence and centrality, which can be considered as the key experts on semantic mining. Among them, Professor Salton G is the founder of modern information retrieval, and one of the founders of computer science department of Cornell University. He created the vector space model (SVM) in the field of information retrieval, and presided over the establishment of the first fully automated text processing and retrieval system

(SMART) in the world. Agrawal R has proposed a Apriori algorithm for mining association rules quickly and has been widely cited. In addition, Blei DM proposes a thematic model (Latent dirichlet allocation, LDA) for mining hidden themes in texts; Deerwester S proposed LSA (latent semantic analysis) for indexing and retrieval; Han J is an authoritative expert in the field of data mining, and compiled Data Mining: Concepts and Techniques, a classic textbook for data mining. Through these, we find that the highly influential authors in the field of semantic mining research are almost come from information retrieval and data mining research.

Table 1. TOP 10 Influential Authors

#	Author	Frequency	Centrality	Institution	References
1	Salton G	195	0.27	CORNELL UNIV	[6][7]
2	Agrawal R	190	0.32	IBM CORP	[8][9]
3	Blei DM	133	0.08	Univ Calif Berkeley	[10][11]
4	Deerwester S	124	0.08	BELL COMMUN RES INC	[12][13]
5	Han J	110	0.07	Univ Illinois	[14][15]
6	Berners LEE T	104	0.16	MIT	[16][17]
7	Hofmann T	96	0.07	Brown Univ	[18][19]
8	Manning C D	96	0.13	Stanford Univ	[20][21]
9	Landauer TK	94	0.04	Univ Colorado	[22][23]
10	Fellbaum C D	82	0.12	Princeton University	[24][25]

3.4 International Cooperation Analysis

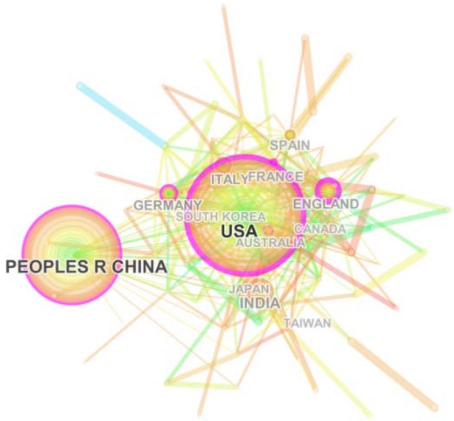


Fig. 5. International Co-occurring Network

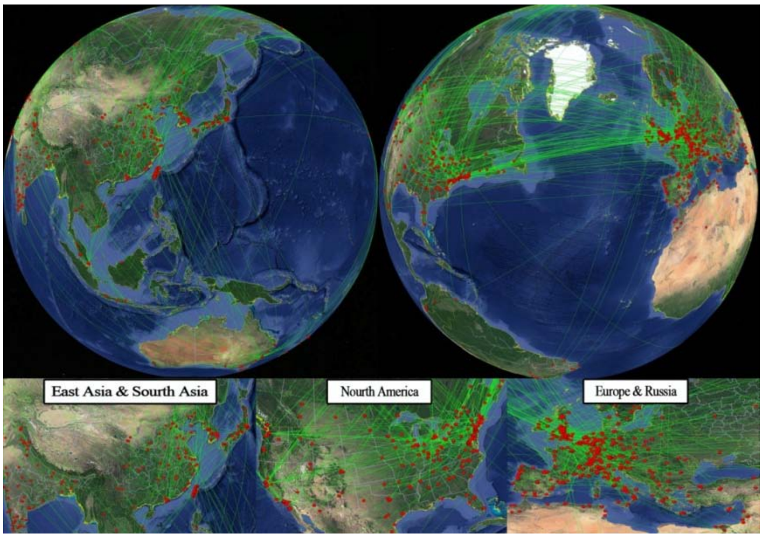


Fig. 6. Geographic Distribution of Countries/Territories

According to the Fig.5 and Fig.6, the international academic communications in the semantic mining field 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. When taking the centrality as measuring index, the U.S. is definitely at the central position of the semantic mining domain, followed by China and those traditional developed countries such as U.K, France, and Germany.

4 Research Hotspots and Fronts Analysis

4.1 Research hotspots

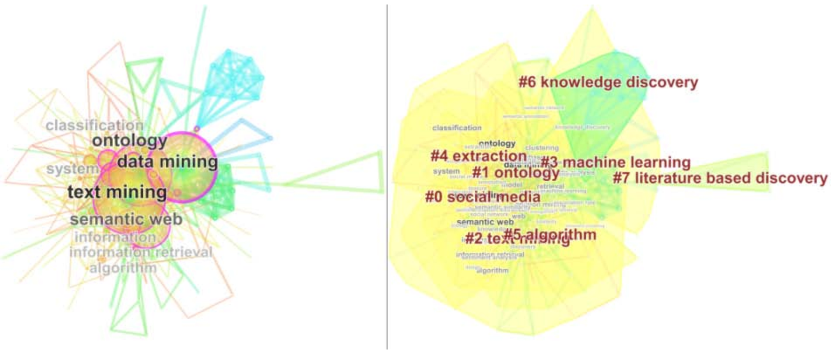


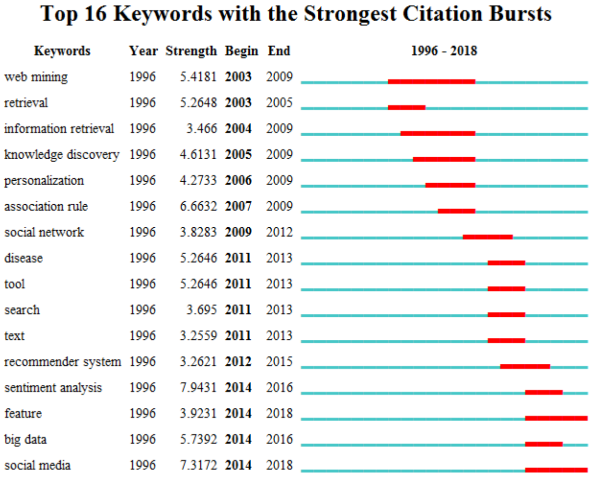
Fig. 7. Keyword co-occurring network(left) and Keyword Clusters(right)

Table 2. TOP 10 Concurrence keywords

#	Keyword	Frequency	Centrality
1	text mining	254	0.15
2	data mining	244	0.24
3	ontology	225	0.16
4	semantic web	178	0.13
5	system	100	0.14
6	classification	97	0.11
7	information	87	0.06
8	algorithm	81	0.15
9	information retrieval	81	0.09
10	web	80	0.04

According to Fig.7 and Table 2, some keywords such as text mining, data mining, ontology, and semantic web, are of high centrality and located in the center of the co-occurring network, which play the important role to extending the research area. Among those keywords, text mining and data mining are the technical basis of semantic mining; Ontology, semantic network and algorithm are the theoretical basis of semantic mining; Classification and information retrieval are the main applications of semantic mining. All of these are the hot topics of research.

Keyword Burst is also an important method to analysis hotspots. Fig.8 shows the top 16 keywords with the strongest citation bursts. From the point of view of time, In 2003-2009, web mining, retrieval, knowledge discovery, personalization, association rule are the keywords with the strongest citation bursts which are the focus of research at that time, reflecting the main function and basic theory of semantic mining; In 2009-2012, the keyword is social network, which reflects that the hotspots at that time was the main application domain, and other keywords such as disease, tool could be considered to be a description of its infectivity and function; Since 2012, recommender system, sentiment analysis, feature, big data, social media are the keywords with the strongest citation bursts, which reflects that the hotspots at that time were new specific applications and new technologies. From the point of view of strength, sentiment analysis, social media, association rule, big data, web mining, retrieval are the keywords with the strongest citation bursts, which reflects that the research of application domain and technology and the research of algorithm rules were the focus of semantic mining research.



Top 16 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	1996 - 2018
web mining	1996	5.4181	**2003**	2009	
retrieval	1996	5.2648	**2003**	2005	
information retrieval	1996	3.466	**2004**	2009	
knowledge discovery	1996	4.6131	**2005**	2009	
personalization	1996	4.2733	**2006**	2009	
association rule	1996	6.6632	**2007**	2009	
social network	1996	3.8283	**2009**	2012	
disease	1996	5.2646	**2011**	2013	
tool	1996	5.2646	**2011**	2013	
search	1996	3.695	**2011**	2013	
text	1996	3.2559	**2011**	2013	
recommender system	1996	3.2621	**2012**	2015	
sentiment analysis	1996	7.9431	**2014**	2016	
feature	1996	3.9231	**2014**	2018	
big data	1996	5.7392	**2014**	2016	
social media	1996	7.3172	**2014**	2018	

Fig. 8. Keyword Bursts

Putting high-frequency keywords and their clusters together, and reading some related articles, we can conclude that the hotspots of semantic mining are the research of technology which is represented by text mining, the research of theory which is represented by ontology and semantic network, and the research of application which is represented by knowledge discovery and information extraction.

4.2 Research Front Analysis

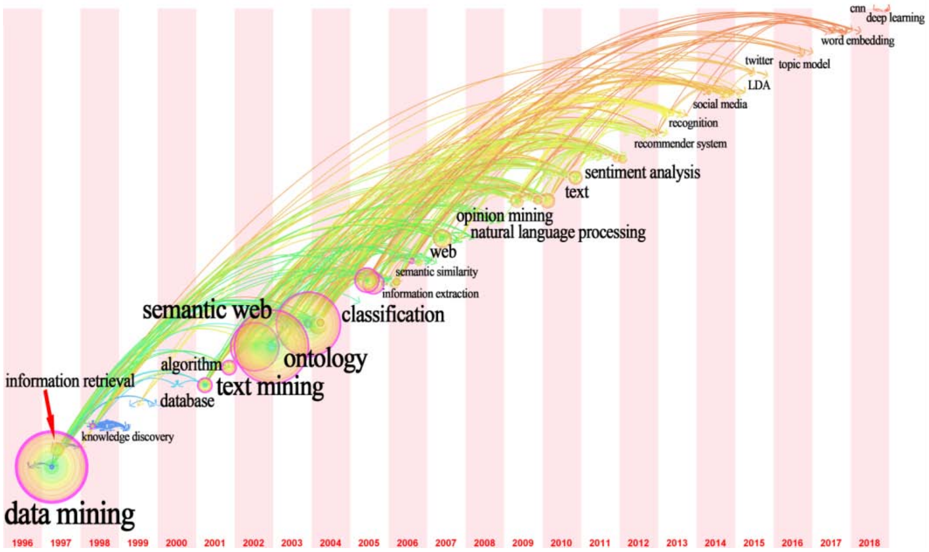


Fig. 9. Research Hotspots Evolution Mapping Knowledge Domain

Fig.9 shows the evolution of research hotspots in the field of semantic mining during the period from 1996 to 2018 and the connections between those hotspots, which take a group of keyword at the up-right corner of the figure as the current research front. Table 3 shows some hotspot keywords of semantic mining key technologies. And we can get the conclusion that the development path of semantic mining research is a 3-step process: first the application requirements, then the theories, and ending up with the key technology researches. So, the current 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.

Table 3. 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)

4.3 References Co-cited Analysis

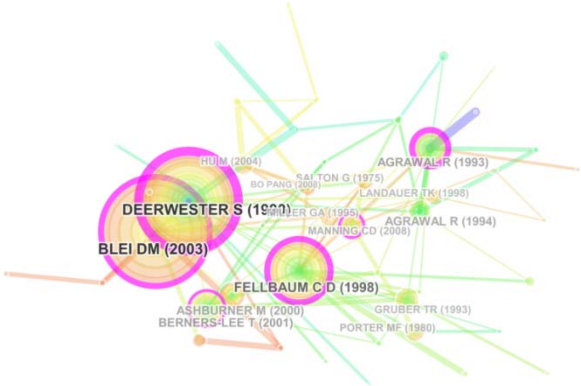


Fig. 10. References Co-cited Network

Highly-cited reference is an important indicator of the research hotspots. According to the analysis of the top 10 highly-cited references, it is found that a majority part of those references are studying some specific algorithm, and the rest are studying some correlation theories or just making some introduction or research of the semantic network or Lexical Database. In addition, among the Top 10 highly-cited references, the first author has a lot of overlap with the highly influential authors, indicating that these scholars are indeed the leading figures in the field of semantic mining. But these papers comes from 9 authors and have no cooperation between each other. It proves that this field of semantic mining research had not produced a group of authors with core influence and dominant position.

Table 4. Top10 Co-cited References

Title	First Author	Frequency	Centrality
Latent dirichlet allocation [10]	Blei D M	6988	0.28
Indexing by latent semantic analysis [12]	Deerwester S	4065	0.46
WordNet: An Electronic Lexical Database [26]	Fellbaum C	3673	0.32
The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities [16]	Berners-Lee T	3180	0.17
Mining association rules between sets of items in large databases [27]	Agrawal R	3867	0.23
Gene Ontology: tool for the unification of biology [28]	Ashburner M	15195	0.04
Fast algorithms for mining assoiciation rules [9]	Agrawal R	3147	0.08
Introduction to Information Retrieval [29]	Manning C D	3182	0.12
An algorithm for suffix stripping [30]	Porter M F	2797	0.02
A translation approach to portable ontology specifications [31]	Gruber T R	4782	0.06

5 Conclusion and Discussion

To sum up, 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 and other sources include Medicine, Psychology and so on.

Third, on the aspect of the high influential authors, Salton G, Agrawal R are the most influential authors in the field of semantic mining, with both high influence and centrality, which can be considered as the key experts on semantic mining. 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. When taking the centrality as measuring index, the U.S. is definitely at the central position, followed by China and those traditional developed countries such as U.K, France, and Germany.

Fifth, on the aspect of the research hotspots, the research of technology which is represented by text mining, the research of theory and algorithm, and the research of application which is represented by knowledge discovery and information extraction 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.

We also discussed how to use mathematical models to predict the number of papers in the future. In 1981, Little A.D. [32] found that the development of technology and biological evolution had an amazing similarity. Then, S curves were introduced to describe the development of Technology.

The general formula of the Logistic curve model is as follows:

$$y = a / (1 + be^{-cx}) \qquad (1)$$

"a" means the limit value of the number of papers, and the ratio of y to a represents the different stages of technical development.

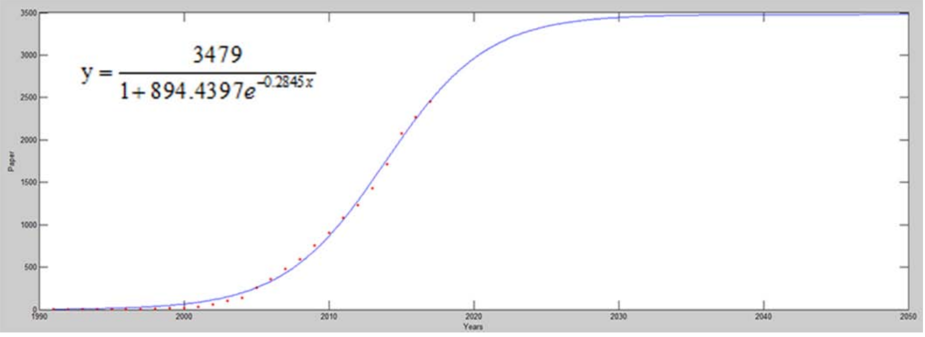


Fig. 11. Logistic curve of semantic mining papers

Fig.11 shows the embryonic stage of semantic mining research is in 1991-2006, then the research will entering a period of growth, 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. Fig.12 and Table 5 shows the Logistic curve of the major countries and the forecast value(Actual value) statistics of their major years.

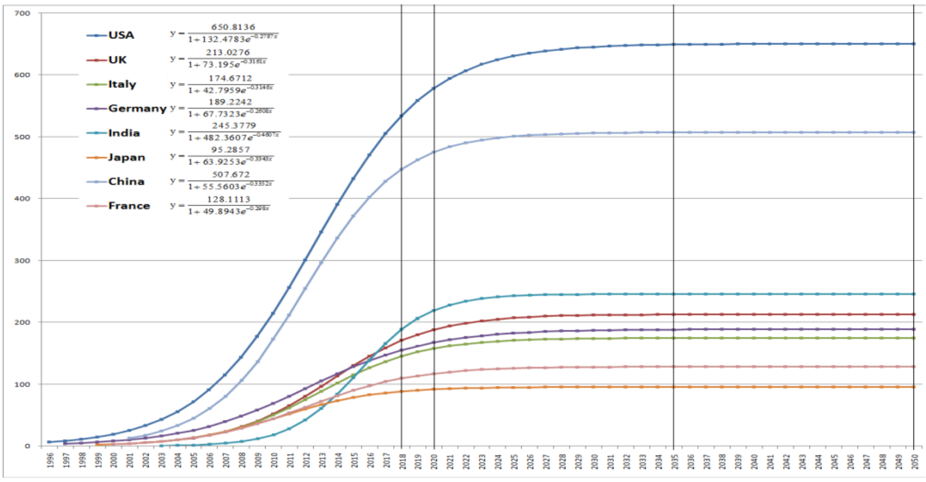


Fig. 12. Logistic curve of semantic mining papers in various major countries

Table 5. A statistical table of the forecast (actual) value of major countries in major years

Country	Begin	inflection point	saturation	2035	2050
USA	1996, 6/1, 0.99	2013, 347/341, 53.25	2029, 644, 98.99	650, 99.81	651, 100
UK	2001, 4/1, 1.84	2014, 114/114, 53.3	2028, 211, 98.96	212, 99.89	213, 100
Italy	2002, 5/2, 3.1	2013, 88/84, 50.47	2028, 173, 99.13	174, 99.9	175, 100
Germany	1997, 4/1, 1.88	2012, 93/88, 48.93	2030, 187, 99.05	189, 99.74	189, 99.99
India	2003, 1/2, 0.33	2015, 111/116, 45.27	2025, 242, 98.8	245, 99.99	245, 100
Japan	1999, 2/1, 2.14	2010, 44/46, 46.35	2024, 94, 98.94	95, 99.97	95, 100
China	2001, 12/2, 2.45	2012, 254/247, 50.12	2026, 503, 99.1	507, 99.96	507, 100
France	2000, 3/1, 2.63	2012, 63/59, 49.1	2028, 127, 99.13	128, 99.89	128, 100

*The sequence of data in the table: Time, value/Actual, Maturity%, 2035/2050 remove "time".

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. Salton, G., & Buckley, C. (1998). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523.

7. Salton, G. (1975). A vector space model for automatic indexing. Communications of the Acm, 18(11), 613-620.
8. Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: A performance perspective. IEEE Trans. Knowledge and Data Engineering (Vol.5, pp.914 - 925).
9. Agrawal, R. (1994). Fast algorithms for mining assoiciation rules. Proc Vldb.
10. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. J Machine Learning Research Archive, 3, 993-1022.
11. Blei, D. (2011). Probabilistic topic models. ACM SIGKDD International Conference Tutorials (Vol.27, pp.1-1). ACM.
12. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (2010). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.
13. Klein, S. T., Bookstein, A., & Deerwester, S. (1989). Storing text retrieval systems on cd-rom: compression and encryption considerations. Acm Transactions on Information Systems, 7(3), 230-245.
14. Han, J., & Kamber, M. (2011). Data mining: concepts and techniques. Data Mining Concepts Models Methods & Algorithms Second Edition, 5(4), 1 - 18.
15. Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining & Knowledge Discovery, 8(1), 53-87.
16. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 284(5), 34-43.
17. Bizer, C., Heath, T., Berners-Lee, T., et al.. (2009). Linked data: the story so far. INTERNATIONAL JOURNAL ON SEMANTIC WEB AND INFORMATION SYSTEMS, 5(3), 1-22.
18. Hofmann, T. (1999). Probabilistic latent semantic indexing. International ACM SIGIR Conference on Research and Development in Information Retrieval (Vol.42, pp.50-57). ACM.
19. Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1-2), 177-196.
20. Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the, Meeting of the Association for Computational Linguistics (Vol.25, pp.63-70). Association for Computational Linguistics.
21. Manning, C. D. (2002). Probabilistic syntax. Philosophy.
22. Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.
23. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2-3), 259-284.
24. Miller, G. A., & Fellbaum, C. (1991). Semantic networks of english. Cognition, 41(1-3), 197.
25. Fellbaum, C. (2010). Wordnet. Theory & Applications of Ontology Computer Applications, 231-243.
26. Fellbaum, C., & Miller, G. (1998). WordNet:An Electronic Lexical Database. MIT Press.

27. Agrawal, R., Imielinski, T., & Swami, A. N. (1993). Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.
28. Ashburner, M., Ball, C. J., Botstein, D., Butler, H., Cherry, J., & Davis, A., et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genetics, 25(1), 25-9.
29. Manning, C. D., & Raghavan, P. (2008). Introduction to Information Retrieval. Cambridge University Press.
30. Porter, M. F. (1980). An algorithm for suffix stripping. Program Electronic Library & Information Systems, 14(3), 130 - 137.
31. Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199-220.
32. Little, A. D. (1981). The strategic management of technology. Cambridge Harvard Business School Press.