Detecting the Landscapes and Hotspots of Scientometrics: A Full-Text Citation Analysis based on Semantic Technology

Li Zeng, Zili Li, Zhao Zhao, Meixin Mao

College of Advanced Interdisciplinary Studies,
National University of Defense Technology Changsha 410073, China
crack521@163.com zilili@163.com z_costa@163.com 452368828@qq.com

Abstract. With the rapid development of information technology, the era of Big Data has come. Big Data technology has brought great opportunities for the research of technology mining, while the "data dizzy" and "data redundancy" effects brought by it cannot be ignored. As one of the basic methods of technology mining, the research of scientometrics also faces the same opportunities and challenges. In order to meet the challenges, an in-depth analysis of scientometrics was conducted. By using the papers of Scientometrics in SpringLinker Database from 1978 to 2017, a Full-Text citation analysis based on semantic technology is used to quantitatively assess the basic status, landscapes, hotspots and future development trends of the "Scientometrics" research area. Besides traditional methods such as co-word analysis, main path analysis and sleeping beauty paper recognization, novel methods such as dynamic topic model and word vectors models are used, furthermore a three-dimensional visualization technology was proposed. It shows that these methods can provide a dynamic view of the evolution of scientometrics research landscapes, hotspots and trends from various perspectives which may serve as a potential guide for future research.

Keywords: Technology Mining; Scientometrics; Visualization; Topic Model

1 Introduction

Scientometrics is the study of measuring and analyzing science, technology and innovation. Major research issues include the measurement of impact, reference sets of articles to investigate the impact of journals and institutes, understanding of scientific citations, mapping scientific fields and the production of indicators for use in policy and management contexts [1]. With the arrival of Big Data Era, the research of scientometrics also faces the same opportunities and challenges with scientometrics. In fact, scientometrics is an old and yet hot field of research, which has gained huge popularity in these days. Modern scientometrics is mostly based on the work of Derek J. de Solla Price and Eugene Garfield. The latter created the Science Citation Index [2].Schubert, A (1989) [3] did a comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981-1985. CALLON, M (1991) [4] introduced a co-word analysis as a tool for describing the network of interactions between basic and technological research using the case of polymer chemistry. Egghe, L (2005) [5] studied the power laws in the information production process. Leydesdorff, Loet (2009) [6] proposed a global map of science based on the isi subject categories. In this paper, a Full-Text citation analysis based on Semantic Technology is used to quantitatively assess the basic status, landscapes, hotspots and future development trends of the "Scientometrics" research area. Novel methods such as dynamic topic model and word vectors models are used, furthermore a three-dimensional visualization technology was proposed with the aim to offer a dynamic view of the evolution of social network analysis research hotpots and trends from various perspectives.

2 Data and methods

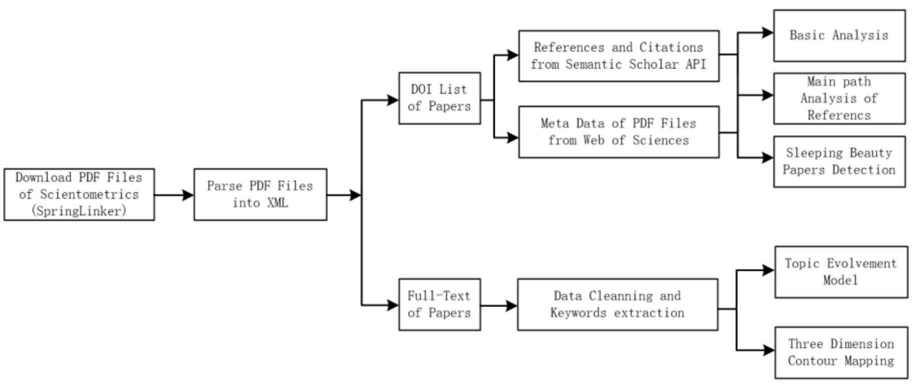


Fig. 1. The research framework of this paper

Figure 1 shows the research framework of this paper. Firstly, Samples are collected from SpringLinker Database on December 31, 2018 and the time span is limited to 1978 and 2017. A python script was used to download full-texts of Scientometric magazine, and 4708 records were downloaded from SpringLinker database and parsed by a machine learning software grobid [7] for extracting, parsing and re-structuring raw PDF files into structured TEI-encoded documents [8] with a particular focus on technical and scientific publications. After getting the DOI [9] list of these papers, references and citation information were expanded by the open api supplied by Semantic Scholar [10], additional meta data such as institutes, citation numbers for these files were supplemented by the web of science database. Secondly, a basic analysis of these files were conducted including maturity forecasting, geographic distribution and internationally collaboration. With the aim to trace the trajectory of scientometrics and the delayed recognized papers, main path analysis and sleeping beauty paper detection was conducted too. Thirdly, the full-text of these files were further processed including steps such as stop-word removal, keyword extraction and so on, then these files were used for the topic model and the three-dimensional visualization technology to vividly show the development of this filed.

The rest of the article is organized as follows. ''Result and discussion'' displays the results containing five parts: "Basic analysis" subsection mainly maturity forecast, geographic distribution and internationally collaborated. "Main path analysis" subsection mainly display the trajectory of scientometrics. "Sleeping beauty paper detection" subsection displays the valuable paper that are underestimated. "Topic model" subsection mainly show the result of full-text analysis. "3D keyword contour mapping" subsection displays 3D keyword semantic mapping we proposed. Conclusions and shortcomings of our research are drawn and discussed in ''Discussion'' section.

3 Result and discussion

3.1 Basic analysis of scientometrics research

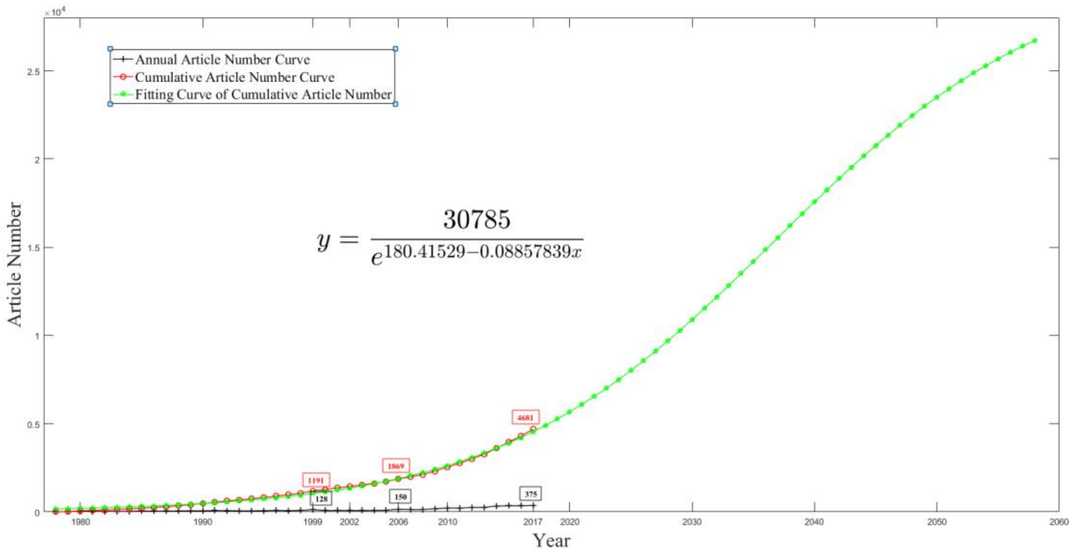


Fig. 2. The geographic distribution of Scientometrics Research

Figure 1 shows the number of papers and maturity forecast between 1978 and 2017 in the field of Virtual Reality. The black curve is the annual number of articles, according to the curve, a substantial interest in scientometrics research did not emerge until 1999, although a few articles related to scientometrics were published previously. The highest number of articles arrived at 2017, with 375 articles, accounting for 7.78% of the total number and the average number of articles was 120.4 per year. The red curve is the cumulative number of papers. According to the theory of technology maturity, the cumulative number of documents could be fitted by the Logistic Growth Model [11]. The least squares method for curve fitting is used to get the parameters in the equation, where the blue curve is the result which is described by (1).

$$y = \frac{30785}{e^{180.41529-0.08857839x}} \tag{1}$$

Here x and y indicate the year and the corresponding article number. According to (1), we can divide the development of Scientometrics into four stages: infant stage (before 2011), growth stage (2012-2061), mature stage (2062-2080) and stable stage (after 2080). According to the above stage division, the research of Scientometrics in 2018 was in the growth stage with a maturity of 15.92%.

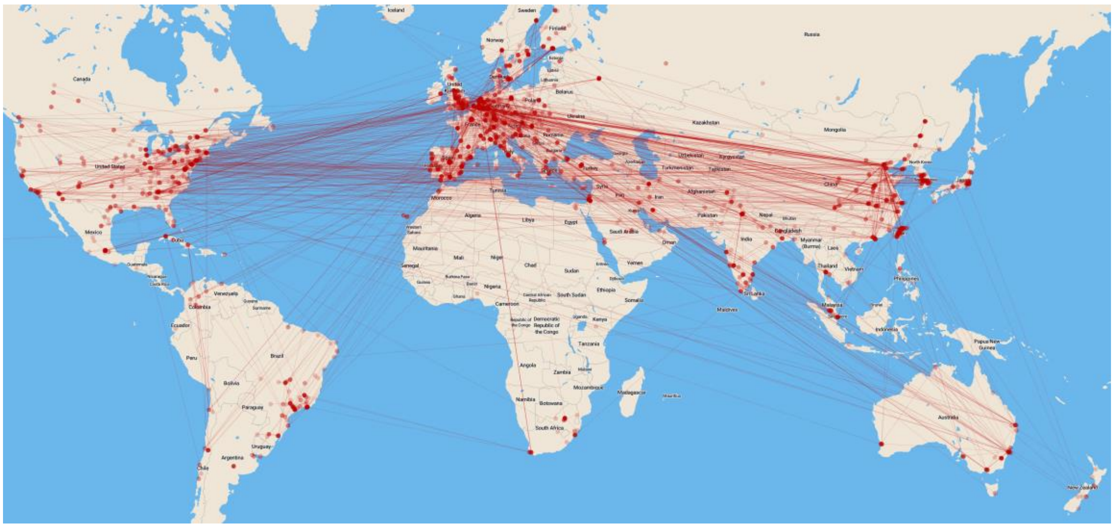


Fig. 3. The geographic distribution of Scientometrics Research

TABLE 1. TOP Ten COUNTRIES/ Territory in Scientometrics

No.	C/T	TP	IP	CP	TC	HI	TI
1	CHINA	564	350	214	5447	31	294
2	USA	525	299	226	8285	41	164
3	Spain	383	269	114	4924	35	287
4	England	262	115	147	5298	41	131
5	Belgium	259	95	164	6620	42	307
6	Germany	255	134	121	4742	33	93
7	Netherlands	252	135	117	8017	49	200
8	Taiwan (China)	191	141	50	3495	28	168
9	India	181	154	27	1970	23	109
10	Italy	162	129	33	1815	21	110

No., Rank By TP; C/T, Country/Territory; TP, Total papers; IP, independent papers; CP, Internationally collaborated articles; TC, Total citations counts; HI, H Index; TI, Total Institutes numbers; (CHINA refers to mainland China).

Figure 3 shows the geographic distribution and cooperation networks of countries/ territories in the field of Scientometrics which was generated from researchers' affiliations. On the whole, these research institutes are mainly located in Europe, Southeast Asia and North America. Table I lists the top ten most productive countries/ territories in this filed. Over all, China is the first most productive and the fourth most influential country in this field, with a total amount of 564 papers (350 independent papers, 214 internationally collaborated papers), 294 institutes and 5447 citations, its top five most productive institutes are Wuhan University (87 papers), Dalian University of Technology (71 papers), Chinese Acad Sci (60 papers), Harbin Institute of Technology (65 papers) and Peking University (38 papers), and CHINA's H-Index is 31. USA is the second most productive but the first most influential country in this field, with a total amount of 525 papers (299 independent papers, 226 internationally collaborated papers), 164 institutes and 8285 citations, its top five most productive institutes are Georgia Institute of Technology (49 papers), Indiana University (48 papers), Drexel University (36 papers), University of Wisconsin (15 papers) and Arizona State University (14 papers), and USA's H-Index is 41. Spain is the third most productive and the sixth influential country in this filed, with a total amount of 383 papers (269 independent papers, 114 internationally collaborated papers), 287 institutes and

4924 citations, its top five most productive institutes are University of Granada (103 papers), Spanish National Research Council (92 papers), Univ Politecn Valencia (40 papers), University de Barcelona (30 papers) and University of Valencia (22 papers), and its H-Index is 35. Netherlands is the seventh most productive but the second most influential country in this field with a total amount of 252 papers (135 independent papers, 117 internationally collaborated papers), 200 institutes and 8017 citations, its top five most productive institutes are Leiden University (101 papers), University of Amsterdam (57 papers), Vrije University of Amsterdam (23 papers), Maastricht University (11 papers) and Erasmus University Rotterdam (8 papers), and Netherlands' H-Index is 49. Other countries/ territories such as Belgium, Taiwan (China), India and Italy also make outstanding contributions in this field.

3.2 Main path analysis to trace the trajectory of Scientometrics

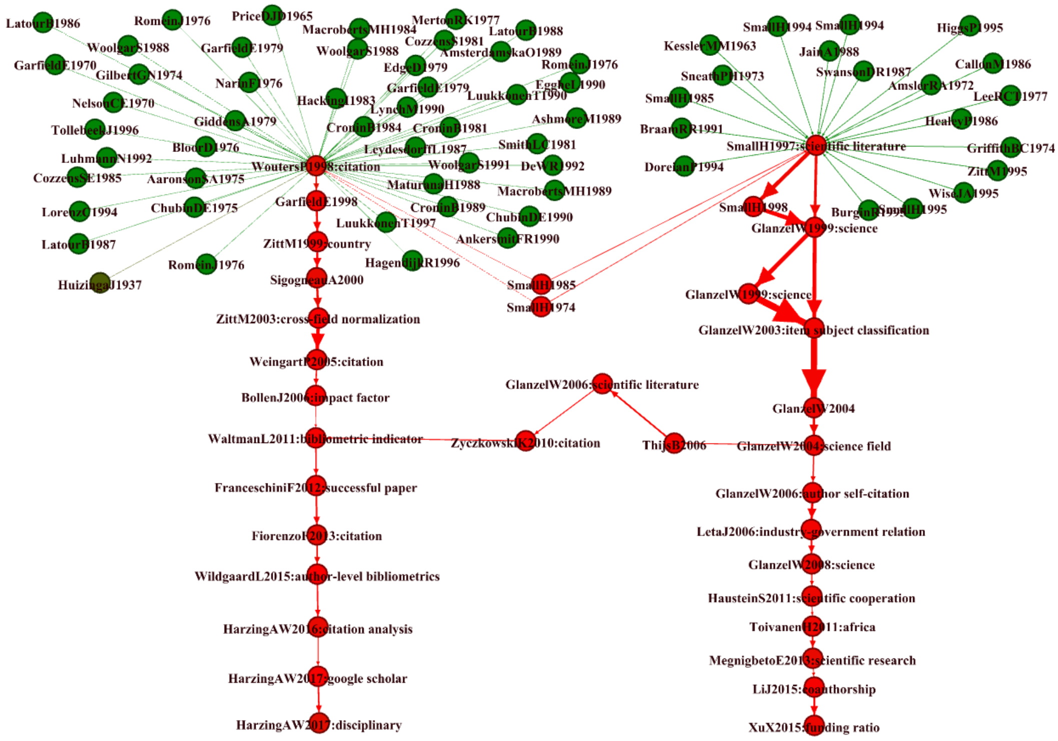


Fig. 4. The Main Path of Scientometrics Research

In order to identify and trace the trajectory and paradigm of Scientometrics, a method called main path analysis was used. Main path analysis examine connectivity in acyclic networks, and are especially interesting when nodes are time dependent, as it selects the most representative nodes at different moments of time [12]. The main path is reconstructed by calculating the connectivity of the links in terms of their degree centrality and outlining the path formed by the nodes with the highest degree. In terms of a citation network, this degree measure considers the number of citations a document receives (in-degree) as well as the number of cited references in the documents (out-degree).The main path is constructed by selecting those connected documents with the highest scores until an end document is reached [13]. By means of Citespace [14] and pajek [15], a reference citation network is generated, after retaining the largest weakly connected sub-graph and decycling the network, the search algorithm based on SPC proposed by Batagelj [13] was used to identify the single main path and Fig.4 shows the result of the algorithm.

From the figure, a conclusion can be concluded that modern scientometric are pioneered by Price DJ [16] who firstly proposed Price Law and used the literary model as a functional simplification of the process of scientific discovery and communication, Garfield, E. [17] who build the citation indexing for studying science. Kessler, M. M [18] who firstly researched the phenomenon of Bibliographic coupling between scientific papers, Glänzel, W., & Schubert [19] who build a new classification scheme of science fields and subfields designed for scientometric evaluation purposes, Small, H.G. [20][21] and Leydesdorff, L [22] who study the scientific mapping domain and so on. Another conclusion can also be drawn that modern scientometrics study exist two main streaming, one mainly focus on citation analysis, impact factor, normalization and so on, the other focus on scientific mapping, subject classification, funding ratio and so forth.

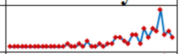
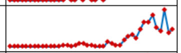
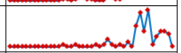
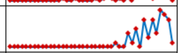
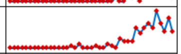
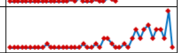
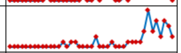
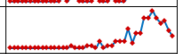
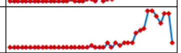
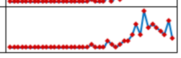
3.3 Sleeping beauty paper detection to mining the valuable

"Sleeping beauty" in science was first proposed by van Raan (2004) [23] in order to describe the phenomenon where papers did not achieve recognition in citations until many years after their original publication. As introduced in [16], a metric to calculate the obsolescence of publications, without examining each citation curve individually to identify shifts, is defined as:

$$\boldsymbol{G}_s = \begin{cases} 1 - \dfrac{2 \times [n \times c_1 + (n-1) \times c_2 + \cdots + c_n] - C}{C \times n} & C > 0 \\ 1 & C = 0 \end{cases} \quad (2)$$

We set the threshold that $0.4 < G_s < 1$ and $C > 40$ to indicates a sleeping beauty, meaning a publication that received recognition after a long period of time.

TABLE 2. TOP Ten Sleeping Beauty Papers in Scientometrics

NO.	DOI	PY	PA	TC	GS	History
1	10.1007/BF02016308 [19]	1979	40	69	0.6291	
2	10.1007/BF02017249 [24]	1986	33	153	0.5854	
3	10.1007/BF02025827 [25]	1984	35	74	0.5684	
4	10.1007/BF02093621 [26]	1996	23	48	0.5587	
5	10.1007/BF02019280 [27]	1991	28	199	0.5496	
6	10.1007/BF02016934 [28]	1985	34	46	0.5455	
7	10.1007/BF02020078 [29]	1988	31	41	0.5366	
8	10.1007/BF02093973 [30]	1991	28	123	0.5086	
9	10.1007/BF02459299 [31]	1997	22	116	0.4959	
10	10.1007/BF02017219 [32]	1994	25	74	0.4775	

No., Rank By GS;DOI, Digital object identifier of paper; PY, publication year; PA, publication age; TC, Total citations counts; GS, obsolescence coefficient of paper

Table 2 listed the top 10 sleeping beauty papers in scientometrics. Glänzel, W. et al (1996) [19] [27] proposed a new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Beaver, D.D. et al (1979) [24] studied in scientific collaboration of the professionalization and the natural history of modern scientific co-authorship. Schubert, A. et al (1986) [25] researched the relative indicators and relational charts for comparative assessment of publication output and citation impact. Rip, A. et al (1984) [26] conducted a co-word maps of biotechnology as an example of cognitive scientometrics. Callon M et al (1991) [28] used co-word analysis to describe the network of interaction between basic and technological research of polymer chemistry. Porter, A.L el al (1985) [29] designed an indicator of cross-disciplinary research. Law, J. et al (1988) [30]: conducted a co-word analysis of research into environmental acidification to study the policy and the mapping of scientific change. Narin, F. et al (1991) [31] studied the scientific co-operation in Europe and the citation of multinationally authored papers. Katz, J.S. et al (1997) [32] proposed a calibrated bibliometric model to evaluate the value of a collaboration. Narin, F. (1994) done a bibliometrics analysis of patent.

3.4 Topic model to describe the evolvement of Scientometrics

In order to uncover the research of scientometrics using the full-text, a probabilistic topic modeling approach called Latent Dirichlet allocation (LDA) [34] was used. LDA was introduced by Blei el al (2003) as a generative probabilistic modeling approach to reveal hidden semantic structures in a collection of textual documents. The basic idea is that each document exhibits a mixture of latent topics wherein each topic is characterized by a distribution over the words.

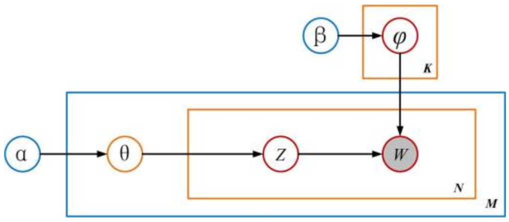


Fig.5. The Latent Dirichlet Allocation For Topic Modelling

Figure 5 shows the whole generative process of LDA, for the convenience of description, the sequence of text was marked as $M = \{m_1, m_2, \cdots m_i, \cdots m_M\}$, where m_i indicates the i-th document in the sequence which is generated by probability distributions of potential topics; the set of topics were marked as $Z = \{z_1, z_2, \cdots z_j, \cdots z_k\}$, where K is the number of topics and z_j is the j-th topic id which was generated by θ_i (the topic distribution over document m_i,) and $W_i = \{w_1, w_2, \cdots w_p, \cdots w_{N_i}\}$ is the words of document m_i which was generated z and φ (the topic distributions over whole corpus).

Table 3. Pseudo algorithm description of LDA

The Generation process of LDA
// Topic level
1、 For each topic index $k \in [1, K]$
Sampling to get mixed parameters of words on each topic $\vec{\varphi}_k \sim Dir(\vec{\beta})$
End For
// Document level
2、 For each document index $m_i \in [1, M]$
Sampling to get the topic mix parameters for each document $\vec{\theta}_m \sim Dir(\alpha)$
Sampling to get the length of each document $N_m \sim Possion(\xi)$;
// Word level calculation
For each word index $n \in [1, N_m]$ in document m_i
Sampling to get the topic index $z_{m,n} \sim Mult(\vec{\theta}_m)$
Sampling to get words $w_{m,n} \sim Mult(\vec{\varphi}_{z_{m,n}})$
End For
End For

Table 4 includes some basic statistics for 4688 full-text of documents we used to generate the topic model of scientometric research. After tokenization of the field contents 22,815 tokens, or individual words, were identified; the number of unique tokens, or distinct words, was 10,515. The topic number is 15, and the hyper parameter α is 3.3333 and β is 0.1.

Table 4. Basic Statistics of Full-Text of Scientometrics

# of authors	# of documents	# of tokens	# of unique tokens	# topics	alpha	beta
6128	4688	22815	10515	15	3.33333	0.1

Table 5 lists the 15 topics recognized by LDA and their representative keywords and Figure 6 is the topic evolution map over time. Combined with the above information, the specific characteristics of each topic can be summarized as follows. Topic #0 represents the newly proposed methods used in scientometrics such as text-mining, keyword analysis, clustering, social network analysis, visualization and so on. Topic #1 mainly focuses on patents, researchers, journals, citations, articles, scientists and so on. Topic #2 represents the data source that such used in scientometrics such as web of science, scopus, patent, journal citation report, google scholar, sci-expanded and so on. Topic #3 focuses on bibliometrics, citation analysis, scientometrics, research evaluation, co-word analysis and so on. Topic #4 mainly focused on the indexes in scientometrics such as h-index, citation, impact factor, journal impact factor, quantity, quality, indicator, Hirsch index, p-index and so on. Topic #5, Topic #7 and Topic 10 mainly contain the research object in scientometrics such as journals, number, publications, countries, data, term, indicators, journals, field and so on. Topic #6 is an emerging topic which contains keywords such as altmetrics, university ranking, big data, page rank, machine learning, RPYS, social media, data quality, twitter, data mining and so on. Topic #9 is the earliest topic in this filed which contains keyword such as science, impact, research, structure, indicators, measurement, scientific productivity, indicators, scientometrics, price, Bradford distribution and so forth. Topic # 11 mainly focuses on university, ranking, peer review, china, research productivity, higher education, collaboration, self-citation, collaboration pattern, informetrics, g-index and so on. Topic #12 mainly focuses on bibliometrics, altmetrics, peer review, wos, scopus, google scholar, patent analysis, citation rate, emerging technology, science mapping, normalization. Topic #13 mainly contains the patent analysis which include keywords such as webometrics, patent, triple helix, patent citation, innovation, patent analysis, research performance, citation index, patent mining, patent count and so on. Topic #14 mainly focuses on citation, citation analysis, scientometrics, science, scientific collaboration, patent, patent citation, collaboration, china, nano-technology, innovation and so on.

Table 5. 15 Topics Detected by LDA

#	
0	Bibliometrics, SNA, Bibliometric analysis, scientometrics, text mining, innovation, cluster analysis, keyword analysis, visualization, citespace
1	Number, patents, researchers, journals, citations, articles, scientists, sci, us, biotechnology, science, nanotechnology, patterns, institute
2	wos, h-index, scopus, patent, journal citation report, research assessment, google scholar, sci-expanded, sleeping beauty, open access
3	bibliometrics, citation analysis, scientometrics, research evaluation, co-word analysis, webometrics, altmetrics, h-index, network analysis
4	bibliometrics, h-index, citation, impact factor, journal impact factor, quantity, quality, evaluation, indicator, Hirsch index, p-index, evaluation
5	journals, number, publications, science, countries, data, scientists, literature, international collaboration, growth, bradford, indicators
6	scientometrics, altmetrics, university ranking, big data, page rank, machine learning, RPYS, social media, data quality, twitter, data mining
7	science, technology, authors, paper, fields, terms, publications, countries, model, web, patents, scientists, researchers, documents, methods, maps
8	citations, h index, journals, countries, publication, quality, terms, institutions, data, impact factor, researcher, method, time, author, internet
9	science, impact, research, structure, indicators, measurement, scientific productivity, indicators, scientometrics, price, Bradford distribution,
10	evaluation, papers, research, citations, journals, science, field, role, development, indicators, innovation, impact, social sciences, universities, set,
11	university, ranking, peer review, china, research productivity, higher education, collaboration, self-citation, collaboration pattern, informetrics, g-index
12	bibliometrics, altmetrics, peer review, wos, scopus, google scholar, patent analysis, citation rate, emerging technology, science mapping, normalization,
13	webometrics, patent, triple helix, patent citation, innovation, patent analysis, research performance, citation index, patent mining, patent count
14	citation, citation analysis, scientometrics, science, scientific collaboration, patent, patent citation, collaboration, china, nano-technology, innovation

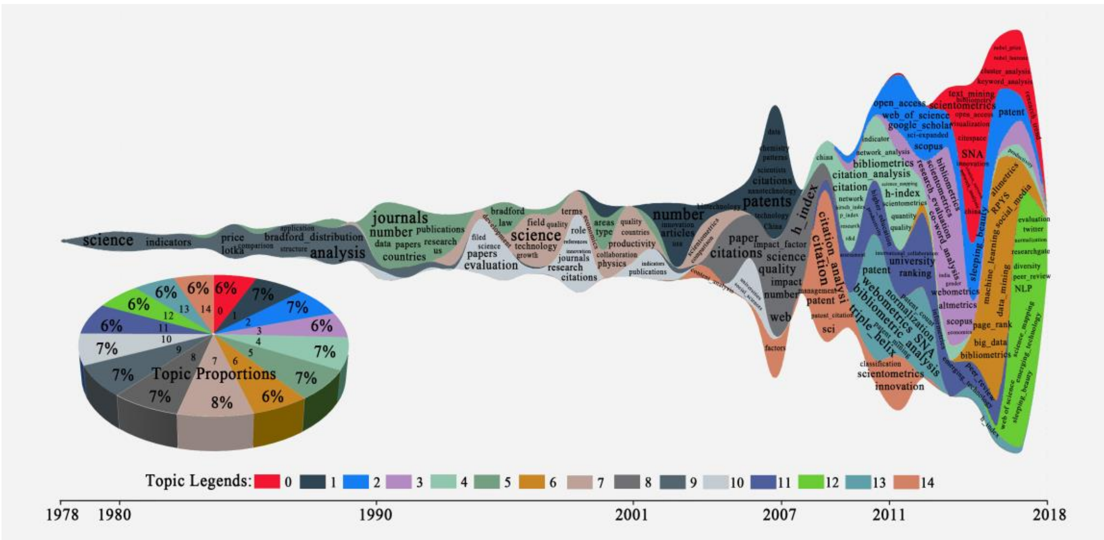


Fig.6. Topic evolution map over time

3.5 3D keyword contour mapping of scientometrics

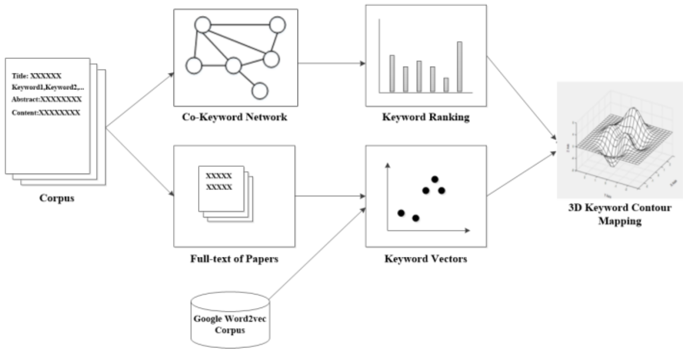


Fig.7. The flow of algorithm to generate 3D keyword contour mapping

In order to more deeply demonstrate the development of scientometrics, we propose a 3D keyword semantic mapping based on word2vec [35]. Fig 7 shows the flow of such algorithm. Firstly, we leverage the NLP tool autophrase [36] to extract high quality phrase which can represented the content of these papers, then we training this corpus by using of word2vec and project the keyword vectors generated by word2vec into 2 dimension by using of TSNE [37]; at the same time, we use the co-occurrence relationship between keywords to calculate the importance of nodes in the network. Finally, we can generate the 3D keyword semantic mapping using the following formula.

$$z_w = \mu \sum_{v \in N} \frac{K(w,v) a_i(v)}{\max(a_i)} \tag{3}$$

Here $K(w,v)$ is the distance mapping kernal function between keyword vector w and v, $a_i(v)$ is the ranking value of keyword v, and $\frac{a_i(v)}{\max(a_i)}$ means the normalized ranking value of keyword v's importance, and μ is the height linear scaling factor.

Figure 8 shows the result of 3D keyword contour mapping of scientometrics. Each peak in the figure represents a keyword or topic in the field. The distance between peaks is determined by the semantic similarity between them, and the height of the peaks indicates the importance of the keywords which can be calculated by indicators such as frequency, betweenness centrality and so on. Here betweenness centrality was chosen as the basic indicator. From the figure, we can clearly concluded that keywords such as biblio_analysis, citation analysis, h_index, scientometric, scientific mapping and so on are the research hotspots in this fields.

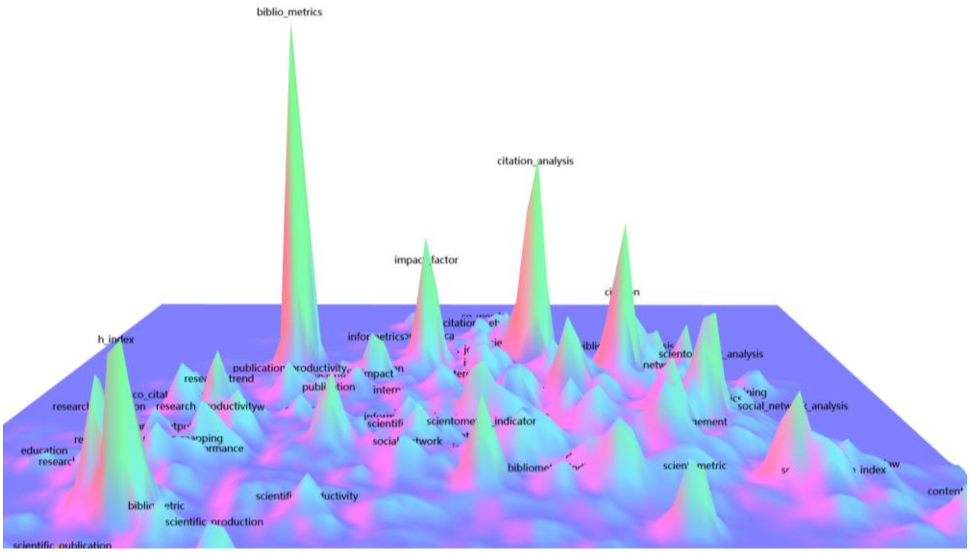


Fig.8. Keyword Landscape Mapping of Scientometrics

4 Conclusion

This paper demonstrates a comprehensive assessment of publication data in the Scientometrics domain. A research framework was proposed to comprehensively assesse the current research hotspots and trends on Scientometrics, using the related Full-text in the SpringLinker database from 1978 to 2017. Analysis about Scientometrics were concentrated on the analysis of technical maturity, scientific outputs, geographic distribution, research trajectory and sleeping beauty paper detection. Moreover, innovative methods such as topic model and keyword 3D semantic mapping were applied which can vividly reveal the landscape and trends from various perspectives. Due to limited data, we conducted the above research based on the data of the scientometric journal itself. In the next phase, we will combine more data to further enrich and improve the research.

References

1. Leydesdorff, L. and Milojevic, S., "Scientometrics" arXiv:1208.4566 (2013), forthcoming in: Lynch, M. (editor), International Encyclopedia of Social and Behavioral Sciences subsection 85030. (2015).
2. De Solla Price, D., editorial statement. Scientometrics Volume 1, Issue 1 (1978).
3. SCHUBERT, A,GLANZEL, W,BRAUN, T.SCIENTOMETRIC DATAFILES - A COMPREHENSIVE SET OF INDICATORS ON 2649 JOURNALS AND 96 COUNTRIES IN ALL MAJOR SCIENCE FIELDS AND SUBFIELDS 1981-1985[J].SCIENTOMETRICS,1989,(1-6):3-&.
4. CALLON, M,COURTIAL, JP,LAVILLE, F.CO-WORD ANALYSIS AS A TOOL FOR DESCRIBING THE NETWORK OF INTERACTIONS BETWEEN BASIC AND TECHNOLOGICAL RESEARCH - THE CASE OF POLYMER CHEMISTRY[J].SCIENTOMETRICS,1991,(1):155-205.
5. Egghe, L.Power Laws in the Information Production Process: Lotkaian Informetrics[S].POWER LAWS IN THE INFORMATION PRODUCTION PROCESS: LOTKAIAN INFORMETRICS,2005,():1-427.
6. Leydesdorff, L,Rafols, I.A Global Map of Science Based on the ISI Subject Categories[J].JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,2009,(2):348-362.
7. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. P. Lopez. Proceedings of the 13th European Conference on Digital Library (ECDL), Corfu, Greece, 2009.
8. Sperberg-McQueen, C.M., & Burnard, L. (1995). The design of the TEI encoding scheme. Computers and the Humanities, 29, 17-39.
9. Paskin, N. (2005). Digital Object Identifiers for scientific data. Data Science Journal, 4, 12-20.
10. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H., Peters, M.E., Power, J., Skjonsberg, S., Wang, L.L., Wilhelm, C., Yuan, Z., Zuylen, M.V., & Etzioni, O. (2018). Construction of the Literature Graph in Semantic Scholar. NAACL-HTL.
11. D. Rogosa, D. Brandt, and M. Zimowski, “A growth curve approach to the measurement of change.”, Psychological bulletin, vol. 92(3), 1982, pp. 726.
12. Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite(TM)-based historiograms. JASIST, 59, 1948-1962.
13. Batagelj, V. (2002). Efficient Algorithms for Citation Network Analysis. CoRR, cs.DL/0309023.
14. Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. JASIST, 57, 359-377.
15. Batagelj, V., & Mrvar, A. (1998). Pajek – Program for Large Network Analysis.
16. Price, D.J. (1965). Networks of Scientific Papers. Science, 149 3683, 510-5.
17. Garfield, E. (1970). Citation Indexing for Studying Science. Nature, 227, 669-671.
18. Kessler, M. M. (1963). Bibliographic coupling between scientific papers. Journal of the Association for Information Science & Technology, 14(1), 10-25.
19. Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics, 56, 357-367.
20. Small, H.G. (1997). Update on science mapping: Creating large document spaces. Scientometrics, 38, 275-293.
21. Small, H.G. (1998). A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system. Scientometrics, 41, 125-133.
22. Leydesdorff, L. (1987). Various methods for the mapping of science. Scientometrics, 11, 295-324.
23. Sun, J., Min, C., & Li, J. (2015). A vector for measuring obsolescence of scientific articles. Scientometrics, 107, 745-757.
24. Beaver, D.D., & Rosen, R. (1979). Studies in scientific collaboration Part III. Professionalization and the natural history of modern scientific co-authorship. Scientometrics, 1, 231-245.

25. Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. Scientometrics, 9, 281-291.
26. Rip, A., & Courtial, J.P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. Scientometrics, 6, 381-400.
27. Glänzel, W., & Czerwon, H. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. Scientometrics, 37, 195-221.
28. Callon M, Courtial JP, Laville F (1991) Co-word analysis as a tool for describing the network of interaction between basic and technological research: The case of polymer chemistry. Scientometric 22: 155–205.
29. Porter, A.L., & Chubin, D.E. (1985). An indicator of cross-disciplinary research. Scientometrics, 8, 161-176.
30. Law, J., Bauin, S., Courtial, J.P., & Whittaker, J. (1988). Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. Scientometrics, 14, 251-264.
31. Narin, F., Stevens, K., & Whitlow, E.S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. Scientometrics, 21, 313-323.
32. Katz, J.S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. Scientometrics, 40, 541-554.
33. Narin, F. (1994). Patent bibliometrics. Scientometrics, 30, 147-155.
34. Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.
35. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. NIPS.
36. Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C.R., & Han, J. (2017). Automated Phrase Mining from Massive Text Corpora. CoRR, abs/1702.04457.
37. Maaten, L.V., & Hinton, G.E. (2008). Visualizing Data using t-SNE.