huang_2022_identification_of_topic_evolution_network_analytics_with_piecewise_linear_representation_and_word_embedding

Year

2022

Author(s)

Huang, Lu and Chen, Xiang and Zhang, Yi and Wang, Changtian and Cao, Xiaoli and Liu, Jiarun

Title

Identification of topic evolution: network analytics with piecewise linear representation and word embedding

Venue

Scientometrics

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Network community labeling based on Z score

Topic labeling parameters

Z-score threshold: 2.5

Label generation

Label of community based on Z-Score

After identifying the communities in the network, we need to set a label for each community.
Z-Score index is used to rank the internal nodes of each community
The node with the highest Z-Score value is selected as the label of the community, and the community could be treated as the topic finally. The formula are as follows:

$$z_i = \frac{N_M^i - B/M^o}{\sqrt{Q/M^o - (B/M^o)^2}}$$

$$B = \sum_{(j \in M)} N_M^j$$

$$Q = \sum_{j \in M} \left(N_M^j \right)^2$$

where *z_i* is the Z-Score value of the i *th* node in community *M*, *N_Mi* represents the sum of the weight of the edges between the i*th* node and other nodes in community *M*, *M_o* represents the number of nodes in community *M*. The higher the Z-Score of a node is, the closer the relationship between the node and other nodes in the community is, and the more representative the node is. Referring to the Study of, nodes with Z-Score greater than or equal to 2.5 can serve as the core node of the community.

Table 6 Z-Score distribution of the community in MAY 2012-SEP 2013

ID	Keyword	Z-Score
1	Citation Analysis	7.6157
2	Citations	7.5343
3	H-Index	6.3953
4	Journal Impact Factor	5.0123
5	Research Evaluation	3.7920
6	Scientometrics	3.6293
7	Impact Factor	3.3852
8	Evaluation	3.2717
9	Webometrics	2.8276
10	Peer Review	2.6462
11	Indicators	2.5649
12	Web of Science	2.5022

Table 6 only presents keywords with Z-Scores greater than 2.5 (Guimerà et al., 2007)

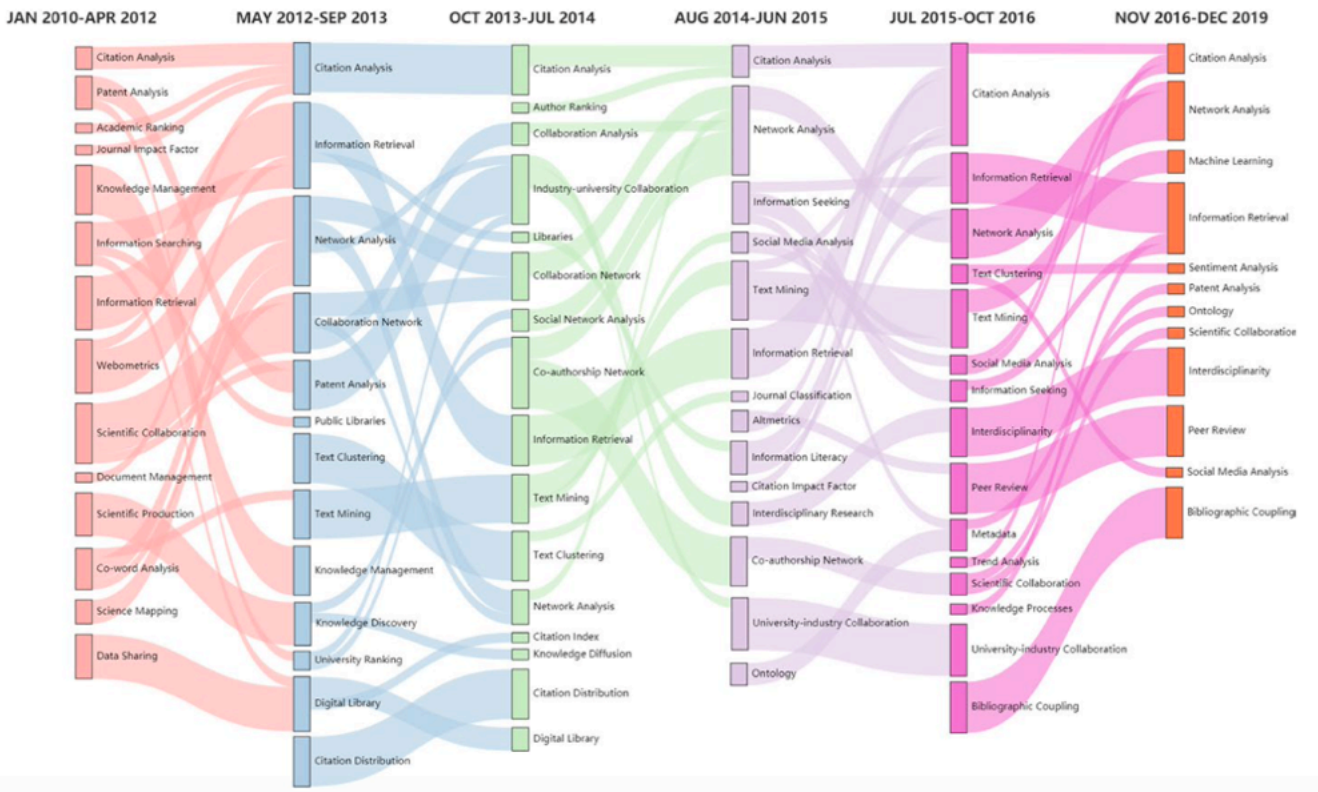

Fig. 6 Topic evolutionary pathways (Piecewise linear representation)

Table 8 Comparisons of topic labels

Topic	SZ	SF	Supplement Keywords
#1	Patent Analysis, Patent Citation, Innovation, Data Mining, Collaboration	Publication Analysis, Patent Analysis,Library and Information Science, Co-Link Analysis, H-Core	H-Index, Patent Stock, Non-Journal Publications, Co-Term, Network Theory
#2	Citation Analysis, Research Evaluation, University Ranking, Research Collaboration, Publication Analysis	Bibliometrics, Citation Analysis, Scientometrics, Classification, Higher Education	Collaboration Network,Citation Rate, Journal Articles, Citation Behavior,Librarianship
#3	Information Retrieval, Computer Science, Image Retrieval, Clustering, Search Engines	Bibliometrics,Information Retrieval,Information Seeking, Altmetrics, Information Literacy	Metadata, Community Detection, Data Base, Elsevier, Sleeping Beauty
#4	Social Network Analysis, Co-word Analysis, Citation Network, Collaborative Networks, Co-author Networks	Social Network Analysis, Google Scholar, Concept Clustering, Credibility, Critical Thinking	University System, Scientific Ranking, Academic Research Group, Economics Departments, Domain Analysis
#5	Text Mining, Science Mapping, Mapping, Co-citation Analysis, Bibliographic Coupling	Impact Factor, Evaluation, Information Retrieval, Web of Science, Research Performance	Machine Learning, Data Accuracy, Factor Analysis, Data Accuracy, Model
#6	Network Analysis, Scientometrics, Open Access, Webometrics, Open Source	Network Analysis, Incubators, Longitudinal Study, Author Metric, Co-citation	Co-Citation Network, 3-D Computer Graphs, Anomaly Detection,Citation Potential, Liking Networks

SZ is the set of top five keywords with the highest Z-score value; SF is the set of top five keywords with the highest frequency; underlined keywords are the topic's labels selected based on word frequency or Z-Score

Motivation

\

Topic modeling

Community detection based on Fast Unfolding

Topic modeling parameters

After constructing the network, the purpose of this part is to recognize the topics based on community discovery, which includes two sections: (1) Community detection based on Fast Unfolding and (2) Label of community based on Z-Score.

We use Fast Unfolding algorithm to identify communities in each keyword network. Fast Unfolding is a community detection algorithm based on maximization of the modularity.
Modularity is an index, which can measure the tightness of con- nections within communities and the sparsity of connections between communities.
The higher the modularity is, the better the result of community detection is, that is, the internal connection is closer and the connection between communities is sparse.

Nr. of topics

Time period	Pruning threshold δ	Community number	R index
JAN 2010-APR 2012	0.45	31	0.6333
MAY 2012-SEP 2013	0.40	25	0.5224
OCT 2013-JUL 2014	0.30	23	0.4576
AUG 2014-JUN 2015	0.35	23	0.4253
JUL 2015-OCT 2016	0.35	24	0.5347
NOV 2016-DEC 2019	0.30	28	0.4746
Total	–	154	–
Mean Modularity		–	0.5080

Label

Keyword extracted from the community

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Bibliometrics

Dataset: Information Science

Problem statement

This study proposes a framework of identifying topic evolutionary pathways based on network analytics: Firstly, keyword networks are constructed, in which a piecewise linear representation method is used for dividing time periods and a Word2Vec mode is used for capturing semantics from the context of titles and abstracts; Secondly, a community detection algorithm is used to identify topics in networks; Finally, evolutionary relationships between topics are represented by measuring the topic similarity between adjacent time periods, and then topic evolutionary pathways are identified and visualized. An empirical study on information science demonstrates the reliability of the methodology, with subsequent empirical validations.

Corpus

Origin: Web of Science

Nr. of documents: 10,135

Details:

- ten leading journals in IS
- between 2010 and 2019

Document

author keywords, titles, abstracts and published years

Pre-processing

- term clumping process is carried out, which includes removing the garbled code in the title and abstract

---


```
@article{huang_2022_identification_of_topic_evolution_network_analytics_with_piecewise_linear_representation_and_word_embedding,
    abstract = {Understanding the evolutionary relationships among scientific topics and learning the evolutionary process of innovations is a crucial issue for strategic decision makers in governments, firms and funding agencies when they carry out forward-looking research activities. However, traditional co-word network analysis on topic identification cannot effectively excavate semantic
```

relationship from the context, and fixed time window method cannot scientifically reflect the evolution process of topics. This study proposes a framework of identifying topic evolutionary pathways based on network analytics: Firstly, keyword networks are constructed, in which a piecewise linear representation method is used for dividing time periods and a Word2Vec mode is used for capturing semantics from the context of titles and abstracts; Secondly, a community detection algorithm is used to identify topics in networks; Finally, evolutionary relationships between topics are represented by measuring the topic similarity between adjacent time periods, and then topic evolutionary pathways are identified and visualized. An empirical study on information science demonstrates the reliability of the methodology, with subsequent empirical validations.},
    author = {Huang, Lu and Chen, Xiang and Zhang, Yi and Wang, Changtian and Cao, Xiaoli and Liu, Jiarun},
    date-added = {2023-04-28 11:27:30 +0200},
    date-modified = {2023-04-28 11:27:30 +0200},
    day = {01},
    doi = {10.1007/s11192-022-04273-1},
    issn = {1588-2861},
    journal = {Scientometrics},
    month = {Sep},
    number = {9},
    pages = {5353--5383},
    title = {Identification of topic evolution: network analytics with piecewise linear representation and word embedding},
    url = {https://link.springer.com/content/pdf/10.1007/s11192-022-04273-1.pdf},
    volume = {127},
    year = {2022},
    bdsk-url-1 = {https://link.springer.com/content/pdf/10.1007/s11192-022-04273-1.pdf},
    bdsk-url-2 = {https://doi.org/10.1007/s11192-022-04273-1}}

#Thesis/Papers/FS