Postprint of a Comparative Study on Hypertension Topic Detection and Evolutionary Trends Using SNA and DMR Methods

Authors: Zhou Liqin, Xu Jian, Ba Zhichao, Zhang Bin

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] Detecting the topics and evolution trends of hypertension medical literature is of significant importance for identifying research hotspots and frontiers in the hypertension field, understanding the overall landscape of the domain, and facilitating knowledge exchange among experts. [Method/Process] This study utilized 26,717 bibliographic records of hypertension-related literature retrieved from the PubMed database as the research corpus. High-frequency subject terms were extracted to construct a co-occurrence matrix, while Social Network Analysis (SNA) and Dirichlet Multinomial Regression (DMR) topic models were simultaneously employed to detect the thematic distribution and evolution trends of hypertension medical literature from meso- and micro-level perspectives, and to compare the relationships, similarities, and differences between these two methods. [Results/Conclusion] The findings reveal that hypertension medical literature primarily concentrates on five research themes: risk factors, research methodologies, basic elements, diagnosis and treatment, and animal experiments, with the relative distribution ratios of these themes continuously evolving over time. Subject terms obtained through the SNA method are more specific and explicit, whereas those derived from the DMR method are comparatively broader; however, the DMR method demonstrates greater advantages in exploring the evolution trends of individual themes.

Full Text

Preamble

A Comparative Study of Topic Detection and Evolutionary Trends in Hypertension Research Based on SNA and DMR Methods

Zhou Liqin[1], Xu Jian[1], Ba Zhichao[1], Zhang Bin[2],[3]
[1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2]Center of Traditional Chinese Cultural Studies, Wuhan University, Wuhan 430072
[3]National Institute of Cultural Development, Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] Exploring the topics and evolutionary trends in hypertension medical literature is crucial for identifying research hotspots and frontiers, understanding the overall landscape of the hypertension field, and facilitating knowledge exchange among experts. [**Method/Process**] This study uses 26,717 hypertension-related article records downloaded from PubMed as its research object. High-frequency MeSH terms were extracted to construct a co-occurrence matrix. Social Network Analysis (SNA) and Dirichlet-multinomial Regression (DMR) topic modeling were employed to detect topic distribution and evolution trends in hypertension literature from meso- and micro-level perspectives, and to compare the relationships, similarities, and differences between these two methods. [**Result/Conclusion**] The study found that hypertension literature primarily concentrates on five research themes: risk factors, research methods, basic elements, diagnosis and treatment, and animal experiments. The relative distribution ratios of these themes change continuously over time. The SNA method yields more specific and explicit topic terms, while DMR produces broader terms but offers advantages in exploring the evolutionary trends of each theme.

Keywords: hypertension; community detection; SNA; DMR; topic model; evolution trend

1. Introduction

Hypertension is the most common chronic disease and the most important risk factor for cardiovascular and cerebrovascular diseases. Over the past 50 years, the number of hypertension patients in China has increased year by year, with approximately 200 million patients currently—one in every five adults [**?**]. Hypertension has extremely high rates of disability and mortality; its complications cause about half of all stroke and heart disease deaths, with 9.4 million deaths attributed to hypertension globally each year [**?**]. In China, hypertension accounts for 46% of total deaths, with cardiovascular diseases representing 26.9%, imposing a heavy burden on families, society, and the nation while consuming significant medical and social resources.

Consequently, increasing numbers of scholars, internet giants, and pharmaceutical companies have invested in hypertension research. Biomedical literature related to hypertension serves as a carrier for medical knowledge dissemination and inheritance, containing substantial useful and latent information. However, the explosive growth of biomedical literature makes it difficult for researchers to

obtain an overview of the field and track research hotspots. A PubMed search using "Hypertension[MeSH Terms]" retrieved 284,322 hypertension-related articles as of May 12, 2017.

Many scholars have attempted to study hypertension literature using bibliometric methods. Bibliometrics employs quantitative and statistical analysis to describe publication patterns in a given field or subject [**?**]. Y.S. Oh and Z.S. Galis [**?**] used citation and content analysis to identify and validate key characteristics of the 100 most-cited hypertension research papers over the past century, including citation rankings, publication years, journals, article types, countries, funding sources, and author affiliations. C. Schreiber et al. [**?**] used bibliometrics to study pulmonary arterial hypertension literature published between 2000-2014, aiming to discover research characteristics, influencing factors, and originating countries. M. Götting et al. [**?**] retrieved pulmonary arterial hypertension literature from Web of Science between 1900-2015, analyzing country distribution, temporal distribution, author distribution, citation patterns, and h-indexes.

However, these studies focus on macro-level analysis (major research countries, institutions, authors, etc.) and lack in-depth examination of meso- and micro-level community distribution and topic evolution. Detecting topic distribution and evolution trends in hypertension literature is significant for understanding research hotspots and frontiers. Existing popular topic identification approaches concentrate in two directions: social network analysis (SNA)-based topic and community detection, and topic model-based identification [**?**]. This paper selects 26,717 hypertension-related article records published between 2000-2017 from PubMed, applying both SNA and Dirichlet-multinomial Regression (DMR) topic modeling to detect topic distribution and evolution trends, conduct in-depth analysis of hypertension bio-entity networks and content, examine local changes in topics over time, and explore interactions among significant hypertension bio-entities. Additionally, we compare the topics and evolution trends obtained by both methods to analyze their relationships, similarities, and differences.

2. Literature Review

Co-word analysis [**?**] and citation analysis [**?**] have been widely studied and applied for hotspot topic discovery. D.R. Swanson [**?**] pioneered biomedical literature-based knowledge discovery in 1987 using word co-occurrence methods, introducing ABC theory to mine and infer associations between bio-entities without direct connections, thereby addressing information silos in biomedical literature. However, pure word co-occurrence analysis requires strong professional medical knowledge and extensive manual operation. Subsequent improvements include M.D. Gordon and R.K. Lindsay' s work [**?**], which used word frequency statistics and TF-IDF methods to discover potentially beneficial scientific knowledge with latent connections that simple, standard indexing methods cannot reveal.

Zhu Qingsong and Leng Fuhai [?] employed citation analysis with highly-cited papers for content extraction and topic identification. However, co-word and citation analysis methods have limitations: co-word analysis results are independent of documents, making it impossible to probe topic distribution within a single document, reducing reference value and accuracy; citation analysis suffers from time lags, making it difficult to analyze recent document collections.

Social Network Analysis (SNA) [?] is a research paradigm developed from anthropology, sociology, psychology, and statistics [?], widely applied in data mining, knowledge management, information dissemination, knowledge networks [?], and data visualization. SNA examines sets of social actors and their relationships, measuring, analyzing, and predicting relationship structures and attributes [?]. In academic literature, scholars use similar terms to express domain hotspots, creating massive textual networks where relationships can be linked through subject terms. SNA provides clear, visual representations of these networks, supporting analysis of term importance, network positions, and associated terms [?]. M.L. Wallace et al. [?] demonstrated through case studies that community detection methods are ideal for identifying research directions, revealing more structural details than traditional co-citation analysis. Thus, SNA is a promising approach for mining term relationship networks and detecting topic communities.

Topic models are probabilistic statistical methods for topic discovery in machine learning, widely used in natural language processing [?], information retrieval [?], and text mining [?]. Topic models assume K latent topics in a document collection, where topics are expressed as probability distributions over terms and documents as distributions over topics, representing each document as a bag of words. Topic models originated from Latent Semantic Indexing (LSI) [?], which used Singular Value Decomposition (SVD) for semantic dimensionality reduction. T. Hoffman extended this to Probabilistic Latent Semantic Indexing (pLSI) [?], using probability values to distinguish associations among documents, topics, and terms. D.M. Blei' s Latent Dirichlet Allocation (LDA) [?] brought topic models to maturity. The mainstream topic models in current literature generally refer to LDA and its derivatives.

Dirichlet-multinomial Regression (DMR) [?] extends LDA by incorporating a log-linear prior in the document-topic distribution that can be adjusted based on observed document features (author, venue, publication date, etc.) to obtain topic distributions under different conditions. M. Song et al. [?] used DMR to detect topics and evolution trends in Alzheimer' s disease research with excellent results. Compared to co-word analysis, topic models can reveal topic probability distributions for any document; compared to citation analysis, topic models are not time-lagged and better reflect "term-topic-document" relationships, offering significant advantages for hotspot topic discovery.

In summary, SNA and DMR offer major advantages over traditional co-word and citation analysis for hotspot detection. This paper aims to apply both methods from meso- and micro-levels to detect topic distribution and evolution trends in

hypertension literature and compare their relationships and differences.

3. Research Design

The research framework consists of five steps: (1) Data collection and processing: retrieving hypertension literature titles, abstracts, years, and journals from PubMed, then performing word segmentation and stop-word removal; (2) Basic bibliometric analysis: importing processed data into BICOMB [**?**] for analysis of year distribution, journal distribution, and subject term distribution; (3) Hypertension literature topic community detection: constructing a co-occurrence matrix from MeSH terms, importing into Gephi [**?**] to calculate PageRank values and centrality, identify key nodes, detect and visualize topic communities, and describe evolution trends; (4) Apply DMR topic modeling to detect topic distribution and temporal evolution; (5) Compare and validate results from steps 3 and 4. The research framework is shown in Figure 1 [Figure 1: see original paper].

3.1 Data Collection and Processing

Using the search strategy "Hypertension[MeSH Terms] AND ('2000/1/1' [PDat]: '2017/5/1')" in PubMed, we retrieved 99,252 hypertension-related articles published since 2000. We selected records containing both abstracts and full texts, yielding 26,717 articles (as of May 2017), saved in XML format as our research object.

To determine high-frequency MeSH terms, we used the high/low frequency boundary formula from Y. Yang et al. [**?**]:

$$T = \frac{-1 + \sqrt{1 + 8 \cdot I_1}}{2}$$

where I_1 is the number of terms appearing only once, and T is the minimum frequency threshold for high-frequency terms. Based on this formula, the frequency threshold was 77. For better visualization, we removed points below this threshold and isolated nodes, resulting in 632 vertices. Constructing a co-occurrence matrix from the top 100 vertices yielded 4,950 edges. Vertices represent bio-entities derived from articles, edges represent relationships between entities, and edge weights represent co-occurrence frequencies within specific sentences.

3.2 Methods and Models

(1) High-frequency term extraction and co-occurrence matrix construction. The bibliographic co-occurrence matrix builder BICOMB [**?**] was used to extract high-frequency MeSH terms from abstracts and construct the co-occurrence matrix. BICOMB can quickly read, accurately extract fields, categorize, and generate co-occurrence matrices from PubMed, SCI, CNKI, and

other databases. Analysis of the 26,717 articles revealed they were published in 1,701 journals, involved 171,637 authors, and contained 9,978 subject terms.

(2) Community detection based on optimized network modularity. Communities are groups of highly aggregated, tightly connected nodes representing an intermediate network characteristic between macro- and micro-levels [**?**]. Nodes in the same community are more likely to have similar functions, and community structure helps understand network structure-function relationships. The most representative algorithm is M.E.J. Newman's modularity optimization method [**?**], where modularity (Q-value) measures network partition quality:

$$Q = \sum_i (e_{ij} - a_i^2)$$

where e_{ij} represents the fraction of edges between community i and community j, and $a_i = \sum_j e_{ij}$ is the fraction of edges with one endpoint in community i. Essentially, modularity-based algorithms identify communities by analyzing edge betweenness and modularity changes. In co-word networks where nodes are subject terms, identifying communities representing research topics becomes a process of finding core nodes, as a few core nodes represent the scientific theme of each community. Node importance metrics include traditional centrality, prestige, and PageRank values, all considering global network properties like edge counts, centrality, and connectivity to identify core nodes.

(3) Dirichlet-multinomial Regression topic model. DMR [**?**] extends LDA by including a log-linear prior in the document-topic distribution that can be adjusted based on observed document features. This study uses publication time as a variable to explore temporal topic trends. The DMR model is illustrated in Figure 2 [Figure 2: see original paper] [**?**].

In document collection D, for each document d, x_d represents metadata feature vectors, α is a function of observable document features representing topic prior distribution. Given prior distribution $N(0, \Sigma)$, hyperparameter β, the document and word generation process is:

For each topic t, draw $\phi_t \sim Dir(\beta)$. $Dir(\beta)$ is the topic-word distribution;
For each document d, draw $\theta_d \sim Dir(\alpha_d) = Dir(exp(\tau_d))$, $\tau_d \in \tau$. For each document' s α_d, Dirichlet parameters and τ_d are covariance functions $f(y_d, x_k)$, where y_d is document d' s observed attribute vector and x_k is metadata vector;
For each word w, draw $z_{d,w} \sim Multi(\theta_d)$. $z_{d,w}$ is the topic assignment for word w in document d, θ_d is document d' s topic proportion; draw $T_{d,w} \sim Multi(\phi_{z_{d,w}})$. $T_{d,w}$ is document d' s w-th word, ϕ_t is topic t' s preference, $\sum_n \phi_{t,n} = 1$.

In the DMR model, we set three fixed parameters: σ^2 (variance of prior distribution parameters), β (Dirichlet topic-word distribution component), and $|T|$ (number of topics).

4. Analysis and Results

4.1 SNA-Based Detection of Hot Topics and Evolution Trends in Hypertension Literature

Using boundary formula (1), the frequency threshold was 77. After removing terms below this threshold and isolated nodes, we obtained 632 vertices. Selecting the top 100 vertices yielded 4,950 edges. Importing into Gephi and applying community detection algorithms [**?**] with Fruchterman-Reingold layout produced the visualization shown in Figure 3 [Figure 3: see original paper].

(1) Key node identification. To identify the most central bio-entities, we analyzed the top 10 nodes using four centrality metrics: PageRank, weighted centrality, closeness centrality, and betweenness centrality (Table 1), following methods from Wasserman [**?**] and Brin [**?**]. PageRank estimates node importance based on incoming connections [**?**]. The average PageRank was 0.01, with top-ranked entities similar to those from weighted centrality (8 of 10 identical, though rankings differed). Unique entities like "rates" appeared only in PageRank top 10.

Weighted degree centrality extends degree centrality by calculating frequencies of node pairs. Betweenness centrality counts shortest paths through a node; closeness centrality represents the reciprocal of total distance from a node to all others, indicating network reachability [**?**]. Table 1 describes top 10 bio-entities for each metric. Average weighted degree was 23.427; average betweenness centrality was 78.17; average closeness centrality was 0.4. Seven nodes overlapped between weighted and betweenness centrality top 10. "Risk factor," "Prospective Studies," and "Follow-up Studies" appeared only in weighted centrality, while "Blood Pressure," "Animals," and "Body Weight" appeared only in betweenness centrality. Closeness centrality' s top 10 matched weighted centrality exactly. Overall, key network nodes included Hypertension, Male, Female, Age, Adult, Human, Risk Factors, Body Weight, Blood Pressure, Animals, Prospective Studies, and Follow-up Studies—all occupying central positions in the community distribution.

(2) Topic community detection. Using the modularity optimization formula (2) and Blondel et al.' s algorithm [**?**] with resolution set to 1 [**?**], we identified five modules (three major ones shown in red, blue, and green in Figure 3 [Figure 3: see original paper]). Modularity Q was 0.187; average weighted degree was 23.43; graph density was 0.282 (28.2% visibility); average clustering coefficient was 0.808; iterations were 100; eigenvector centrality was 0.00239; average path length was 2.645; diameter was 6.

The largest green community (42% of network) contained terms like Risk Factors, Aged (80 and over), Prevalence, Sex Factors, Prospective Studies, Follow-up Studies, Cross-Sectional Studies, and Cohort Studies. The second-largest red community (36%) contained Hypertension, Humans, Female, Middle Aged, Treatment Outcome, Antihypertensive Agents, and Blood Pressure Determina-

tion. The third blue community (20%) contained Animals, Rats (Inbred SHR), RNA (Messenger), and Mice. Manual and expert judgment categorized these into five research themes:

- **Theme 1 (Risk Factors):** Age factors, pregnancy, sex, smoking, obesity, myocardial infarction, etc.
- **Theme 2 (Research Methods):** Prospective studies, follow-up studies, cross-sectional studies, cohort studies, retrospective studies, plus indicators like incidence, disease severity index, and expected values.
- **Theme 3 (Basic Elements):** Sex, age, blood pressure, heart rate, renin, glomerular filtration rate, etc.
- **Theme 4 (Diagnosis/Treatment):** Treatment outcomes, antihypertensive drugs, blood pressure measurement, dose-response relationships.
- **Theme 5 (Animal Experiments):** Animals, rats, RNA, Inbred SHR for validating hypertension indicators.

Table 2 shows the community and topic distribution.

(3) Evolution trend analysis. Dividing the literature into three periods (2000-2005, 2006-2010, 2011-2017) and visualizing each stage's community distribution (Figure 4 [Figure 4: see original paper]) revealed relatively balanced community proportions across periods: 38%/33%/29% (2000-2005), 40%/38%/22% (2006-2010), and 42%/37%/21% (2011-2017). While high-frequency terms like "hypertension," "male," and "female" appeared across all stages, community distributions varied slightly. Table 3 shows community distribution parameters. Node average degree, graph density, and modularity increased over time, indicating growing numbers of MeSH terms and evolving community structures. However, this method is labor-intensive and cannot accurately detect each theme' s proportion per period or temporal evolution paths.

4.2 DMR-Based Detection of Hot Topics and Evolution Trends

(1) Hot topic distribution. The 26,717 records were processed by: (1) stemming; (2) removing stop words, single-character words, and terms appearing fewer than 5 times; (3) removing hypertension superordinate terms. Each record became a text file for DMR modeling. Using the MALLET toolkit [**?**] and setting topic count $|T| = 5$ (to match SNA results), with tuned σ^2 and β values, we identified five topics. To enhance interpretability, we examined top 10, 20, and 30 terms per topic, selecting the most frequent and meaningful terms (Table 4):

- **Topic 1:** mice, angiotensin, renin, vascular, response, effects, receptor, rats—describing animal experiments.
- **Topic 2:** risk, factors, age, obesity, gene, women—describing risk factors (age, diabetes, obesity, sex, genetics).
- **Topic 3:** systolic, diastolic, group, compare, rate, invalid, significant—describing research methods.

- **Topic 4:** patient, blood, gene, treatment, results—describing basic hypertension elements.
- **Topic 5:** treatment, antihypertensive, coronary, mortality, medication, clinical, therapy—describing diagnosis and treatment.

(2) Topic evolution trends. Using publication time as a variable, we examined relative topic distributions from 2000-2017 (Figure 5 [Figure 5: see original paper]). Overall, all themes evolved over time. In 2000, Topic 1 (animal experiments) and Topic 4 (basic elements) dominated, while Topic 5 (diagnosis/treatment) was relatively weak. Over time, Topic 1 declined, Topic 4 decreased then increased but remained important, Topic 5 grew steadily to become relatively significant by 2017, Topic 2 (risk factors) remained stable and important, and Topic 3 (research methods) fluctuated slightly but increased from 2007 onward.

5. Comparison and Discussion

Results show that SNA and DMR detect essentially the same themes: risk factors, research methods, basic elements, diagnosis/treatment, and animal experiments. However, the MeSH terms differ. SNA yields more specific, concrete terms with clear meanings within each community/theme. For example, in risk factors, SNA identified specific terms like "age factors," "Diabetes Mellitus," "sex factors," "smoking," "cardiovascular disease," "obesity," and "lifestyle." DMR produced broader terms like "age," "Diabetes," "obesity," and "gene" —only representing major categories. Similarly, for research methods, SNA identified specific terms like "prospective studies," "logistic models," "surveys and questionnaires," "cross-sectional studies," plus metrics like "risk assessment" and "severity of illness index," while DMR yielded broader terms like "group," "compare," "rate," and "significant."

These differences arise partly from dataset size: SNA analyzed only the top 100 high-frequency terms, while DMR used the entire document collection. Additionally, SNA community/theme counts are subjectively determined, as is DMR's topic count, potentially introducing error. However, DMR better reflects "term-topic-document" relationships, allowing topic probability distributions for any document.

For evolution trends, SNA can only detect community distributions per time period, making cross-period comparisons difficult. DMR can detect each topic's proportion per period and temporal evolution paths, offering advantages for exploring theme evolution. To facilitate comparison, both methods were set to detect five themes, though this subjective setting may introduce error and requires further investigation.

6. Conclusion

This study applied SNA at the meso-level and DMR at the micro-level to detect hypertension literature topic communities and evolution trends, comparing both methods' relationships and advantages. Key findings:

1. Hypertension research comprises five main themes: risk factors, research methods, basic elements, diagnosis/treatment, and animal experiments.
2. Over time, all themes evolve: basic elements research remains prominent; animal experiments decline; risk factors remain stable and important; research methods fluctuate but increase from 2007; diagnosis/treatment grows steadily.
3. SNA and DMR detect similar themes but with different term specificity. SNA terms are more concrete and explicit; DMR terms are broader but better for evolution analysis. Combined use yields complementary advantages.

These findings help newcomers understand the hypertension field, identify hotspots, predict frontiers, facilitate intra- and cross-domain knowledge exchange among experts, and assist decision-makers in tracking knowledge flow. The community detection and evolution analysis methods can be extended to other chronic diseases like diabetes, cardiovascular disease, and coronary heart disease.

Limitations: DMR requires pre-setting topic numbers; while perplexity [**?**] could determine this, it yields overly large numbers for this context, so we used subjective determination, potentially introducing error. Additionally, this study lacks exploration of internal community/topic structures, representing future work.

References

[**?**] Chinese Hypertension Prevention and Treatment Guidelines Revision Committee. Chinese Hypertension Prevention and Treatment Guidelines 2010. Chinese Journal of Cardiology, 2011, 39(7): 701-708.

[**?**] NATIONAL INSTITUTES OF HEALTH. Report: estimates of funding for various research, condition, and disease categories (RCDC). http://report.nih.gov/categorical_{spending}.aspx.

[**?**] SONG M, KIM S, ZHANG G, DING Y, et al. Productivity and influence in bioinformatics: A bibliometric analysis using PubMed. Journal of the American Society for Information Science and Technology, 2014, 65(2): 352-371.

[**?**] OH Y S, GALIS Z S. Anatomy of success: the top 100 cited scientific reports focused on hypertension research. Hypertension, 2014, 63(4): 641-647.

[**?**] SCHREIBER C, EDLINGER C, EDER S, et al. Global research trends in the medical therapy of pulmonary arterial hypertension in 2000-2014. Pulmonary pharmacology & therapeutics, 2016, 39(8): 21-27.

[?] GÖTTING M, SCHWARZER M, GERBER A, et al. Pulmonary hypertension: scientometric analysis and density-equalizing mapping. PloS one, 2017, 12(1): e0169238.

[?] DING Y. Community detection: Topological vs Topical. Journal of informetrics, 2011, 5(4): 498-514.

[?] KLAVANS R, BOYACK K W. Identifying a better measure of relatedness for mapping science. Journal of the American Society for Information Science and Technology, 2006, 57(2): 251-263.

[?] RONDA-PUPO G A, GUERRAS-MARTIN L A. Dynamics of the evolution of the strategy concept 1962-2008: a co-word analysis. Strategic management journal, 2012, 33(2): 162-188.

[?] CHEN C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.

[?] SWANSON D R. Two medical literatures that are logically but not bibliographically connected. Journal of the Association for Information Science. 1987: 228-233.

[?] LINDSAY R K, GORDON M D. Literature-based discovery by lexical statistics. Journal of the American Society for Information Science and Technology, 1999, 50(7): 574-587.

[?] GORDON M D, LINDSAY R K. Toward discovery support systems: a replication, re-examination, and extension of Swanson' s work on literature-based discovery of a connection between Raynaud' s and fish oil. Journal of the Association for Information Science and Technology, 1996, 47(2): 116-128.

[?] Zhu Qingsong, Leng Fuhai. Research on topic identification of highly-cited papers based on citation content analysis. Journal of Library Science in China, 2014, 40(1): 39-49.

[?] WASSERMAN S, FAUST K. Social network analysis: methods and applications. Contemporary sociology, 1994, 91(435): 219-220.

[?] Zhu Qinghua, Li Liang. Social network analysis and its application in information science. Information Studies: Theory & Application, 2008, 31(2): 179-183.

[?] Xu Yuanyuan, Zhu Qinghua. Empirical research on social network analysis in citation analysis. Information Studies: Theory & Application, 2008, 31(2): 184-188.

[?] Li Liang, Zhu Qinghua. Empirical research on social network analysis in co-authorship analysis. Information Science, 2008, 26(4): 549-555.

[?] BUTTS C T. Social network analysis: a methodological introduction. Asian journal of social psychology, 2008, 11(1): 13-41.

[?] Wang Hongwei, Gao Song, Lu Ping. Research on online news hotspot identification based on LDA and SNA. Journal of the China Society for Scientific and Technical Information, 2016, 35(10): 1022-1037.

[?] WALLACE M L, GINGRAS Y, DUHON R. A new approach for detecting scientific specialties from raw cocitation networks. Journal of the American Society for Information Science and Technology, 2009, 60(2): 240-246.

[?] GROSS A, MURTHY D. Modeling virtual organizations with Latent Dirichlet Allocation: a case for natural language processing. Neural networks, 2014, 58: 38-49.

[?] TANG X B, Fang X K. Research on the subject retrieval of Weibo based on the integration of text clustering and LDA. Information studies: theory & application, 2013, 8: 85-90.

[?] ZHANG P J, SONG L. Overview on topic modeling method of micro-blogs text based on LDA. Library and information service, 2012, 24: 120-126.

[?] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 2010, 41(6): 391-401.

[?] HOFFMANN T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 1999: 50-57.

[?] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation. Journal of machine Learning research archive, 2003, 3: 993-1022.

[?] MIMNO D, MCCALLUM A. Topic models conditioned on arbitrary features with Dirichlet-multinomial Regression. University of Massachusetts-Amherst, 2012, 2008: 411-418.

[?] SONG M, HEO G E, LEE D. Identifying the landscape of Alzheimer's disease research with network and content analysis. Scientometrics, 2015, 102(1): 905-927.

[?] CUI L. Development of a text mining system based on the co-occurrence of bibliographic items in literature databases. New technology of library and information service, 2008, 24(8): 70-75.

[?] BASTIAN M, HEYMANN S, JACOMY M. Gephi: an open source software for exploring and manipulating networks. Proceedings of the third international conference on Weblogs and social media. California: ICWSM, 2009.

[?] YANG Y, WU M, CUI L. Integration of three visualization methods based on co-word analysis. Library and Information Service, 2013, 57(8): 91-96.

[?] Cheng Qikai, Wang Xiaoguang. A research theme evolution analysis method based on co-word network communities. Library and Information Service, 2013, 57(8): 91-96.

[?] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks. Physical review E, 2004, 69(2): 026113.

[?] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10): 155-168.

[?] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine. Computer networks, 1998, 30(1-7): 107-117.

[?] LAMBIOTTE R, DELVENNE J C, BARAHONA M. Laplacian dynamics and multiscale modular structure in networks. Physics, 2008, 1(2): 1-29.

[?] WALLACH H M, MIMNO D M, MCCALLUM A. Rethinking LDA: Why priors matter. Advances in neural information processing systems, 2009, 23: 1973-1981.

[?] WANG Y, AGICHTEIN E, BENZI M. TM-LDA: efficient online modeling of latent topic transitions in social media. Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2012: 123-131.

Author Contributions

Zhou Liqin: Data collection and analysis, paper writing;
Xu Jian: Data processing;
Ba Zhichao: Research framework design and guidance;
Zhang Bin: Paper revision suggestions.

Note: Figure translations are in progress. See original paper for figures.