Postprint: Research on Literature Topic Novelty Detection Based on Natural Language Word-Pair Method

Authors: Xu Dan, Xu Shuang, Chen Sisi, Han Shuang, Yang Ying, Guo Jijun

Date: 2023-08-26T00:00:00+00:00

Abstract

[目的/意义] This study proposes a novel quantitative metric—Document Topic Novelty—to detect and investigate the novelty of literature thematic content through a natural language word-pair approach, and explores its feasibility, advantages and disadvantages, as well as the relationship between novelty and F1000 recommended literature and citation indicators. [方法/过程] Based on F1000, we selected literature recommended within the recent month under the hematology subject, retrieved from PubMed closely related articles published within six months prior to each recommended publication to constitute the complete document set. We defined the concept and calculation formula for novelty based on the natural language method, implemented the algorithm using Oracle database PL/SQL programming language, and extracted natural language vocabulary through MetaMap software to compute literature topic novelty. [结果/结论] The natural language method demonstrates certain feasibility in the computational detection of literature topic novelty; Document Topic Novelty is not equivalent to F1000 recommendations or citation performance, as they belong to different dimensions and categories of scientific paper evaluation and should not be generalized. This new metric of Document Topic Novelty should be combined with other relevant evaluation indicators such as peer review status and bibliometrics to conduct comprehensive evaluation and analysis of literature, thereby selecting high-quality articles for recommendation.

Full Text

Preamble

Volume 62, Issue 8, April 2018
ChinaXiv Collaborative Journal

Research on Document Theme Novelty Detection Based on Natural Language Word Pair Methods

Xu Dan, Xu Shuang, Chen Sisi, Han Shuang, Yang Ying, Guo Jijun
Library of China Medical University, Shenyang 110122

Abstract

[**Purpose/Significance**] This study proposes a new quantitative indicator—document theme novelty—to investigate the feasibility of detecting literature theme novelty through natural language word pair methods. It explores the relationship between this novelty measure and both F1000 recommended literature and citation metrics, discussing the method's advantages and disadvantages.

[**Method/Process**] Based on the F1000 database, we selected hematology-themed literature recommended within the past month and retrieved closely related publications from the preceding six months in PubMed to constitute the complete document set. We defined the concept and calculation formula for natural language method novelty and implemented the computation using Oracle database PL/SQL programming. Natural language terms were extracted via MetaMap software for calculating document theme novelty values.

[**Result/Conclusion**] The natural language method demonstrates certain feasibility for literature theme novelty detection. Document theme novelty is not equivalent to F1000 recommendations or citation counts; these belong to different dimensions and categories of scientific paper evaluation and should not be conflated. The document theme novelty indicator should be combined with peer review, bibliometrics, and other relevant evaluation metrics for comprehensive literature analysis and quality recommendations.

Classification Number: G250
Keywords: literature theme novelty detection; natural language word pairs; MetaMap; F1000; citation index
DOI: 10.13266/j.issn.0252-3116.2018.08.017

1. Introduction

With the rapid development of science and technology, research activities have become increasingly active, producing substantial numbers of scientific papers daily. Researchers must analyze large volumes of literature to understand disciplinary development trends and grasp cutting-edge information. While this abundance of information provides rich resources, it also creates redundancy problems, forcing researchers to expend considerable time and effort on literature review.

Research on novelty detection can be traced to September 1996, when DARPA (Defense Advanced Research Projects Agency) initiated a sub-project on topic detection and tracking—first story detection or new event detection—within the Text Retrieval Conference (TREC), the most authoritative international evaluation forum in information retrieval. Since TREC 2002, text content novelty tracking and detection projects have been added. Subsequently, experts and scholars from various fields have conducted numerous studies on analyzing and detecting the novelty and innovation of scientific literature themes.

Y. Zhang [3] proposed a novelty detection method based on vector space models, with the novelty score formula: Novelty Score(d_t) = 1 - max_{1≤i≤t-1} cos(d_t, d_i), arguing that greater similarity between current and previous texts results in lower novelty. G. Kumaran et al. [4] enhanced novelty detection precision by combining text classification and named entity techniques to improve document representation. K. Rajaraman et al. [5] developed three computational methods for topic novelty detection, tracking, and trend analysis using adaptive resonance theory neural networks. M. Zhang et al. [6] proposed an overlap-based novelty determination method with the formula Overlap_{BA} = |A B|/|B|, applying a threshold to judge text novelty—higher overlap indicating lower novelty. H.P. Zhang et al. [7] employed semantic distance calculation methods for TREC 2003 specific topics to detect novelty, retrieving new information while filtering redundancy. S. Flora et al. [8] applied a document-to-sentence annotation framework (D2S) to novelty detection for TREC 2004 and 2003 data, converting documents to sentences to identify sentence-level novelty before calculating document-level novelty scores. Their experiments demonstrated D2S's superior ability to detect redundant information based on document novelty percentages compared to standard precision and recall metrics.

Domestic scholars have primarily proposed innovation quantification indicators based on keywords and term frequency. Shen Lü [9] introduced a general equilibrium theory for scientific innovation, defining scientific achievement repetition rate and citation rate through keyword frequency to quantify innovation degree. Shen Yang [10] proposed an innovation evaluation method based on keyword frequency, considering keyword frequency in documents and search expressions, time span, and user evaluations as calculation bases. Hu Shuli and Zhang Jinghui [11] created a mathematical formula to quantitatively calculate literature novelty based on keywords and their ordering, describing novelty from three levels: topic novelty, main argument/conclusion novelty, and evidence novelty. Qian Lingfei et al. [12] defined three innovation evaluation indicators—keyword crossover rate, co-word life index, and effective new word emergence rate—for comparing disciplinary innovation capabilities, where higher effective new word rates indicate stronger innovation sustainability. Subsequently, Yang Jianlin et al. [13] proposed a theme novelty measurement method based on keyword pair inverse document frequency, defining relevant concepts and providing document novelty calculation formulas. Their empirical research concluded that average theme novelty in important core journals within the same discipline is

higher than in ordinary journals.

However, keyword-based methods have limitations: keywords are often non-standardized, subjective, and limited in number, making them insensitive to emerging research and new technological vocabulary. This study draws inspiration from S. Flora's document-to-sentence-to-document approach and Yang Jianlin's co-word analysis, time point, term frequency, and inverse document frequency concepts. Leveraging MetaMap's ability to automatically extract medical natural language vocabulary with high sensitivity to emerging concepts, we propose a new quantitative indicator—document theme novelty—to investigate literature theme novelty through natural language word pairs.

The core idea is that within a document set, searching for information not present in previous documents reveals that the earlier a co-occurring natural language word pair appears, the higher its novelty. In other words, when a word pair first emerges in the document set, it best represents novelty and emerging concepts, with its representational power for document theme novelty gradually weakening as more documents containing it are published over time.

This study uses F1000 as a baseline, selecting natural language word pairs to detect and analyze literature theme novelty, and compares these results with citation data from Web of Science and F1000 recommendation scores (FFa) to explore potential relationships and evaluate the feasibility and 优缺点 of the natural language word pair method.

---

2. Experimental Materials and Methods

2.1 Research Subject

We selected 38 hematology-themed literature items recommended by F1000 within the past month (downloaded on July 30, 2014) as our baseline. In PubMed, we searched for literature closely related to these 38 items (related citations) published within the six months preceding each recommended article's publication, yielding 523 related documents. After limiting document types to journal articles, historical articles, clinical trials (phases I-IV), controlled clinical trials, randomized controlled trials, comparative studies, multicenter studies, evaluation studies, and in vitro studies—the publication types that best represent a discipline's latest frontiers—we excluded case reports, reviews, letters, comments, news, meta-analyses, consensus development conferences, editorials, and non-English or abstract-less documents. The final dataset comprised 401 documents, including 33 F1000-recommended articles. All 401 documents were saved in MEDLINE format for natural language term extraction and novelty calculation.

2.2 Research Method

This study employs a document theme novelty detection method based on MetaMap natural language word pairs, hereafter called the natural language method. The approach follows three principles:

1. **Co-occurrence Principle**: Co-term analysis, proposed by French bibliometricians in the mid-to-late 1970s, originates from citation coupling and co-citation concepts. It statistically counts keyword co-occurrences in documents for cluster analysis to reveal disciplinary or thematic structural changes [14]. Our natural language method extends this to same-document-same-sentence co-occurrence, where two terms appearing in the same sentence have stronger latent connections and greater persuasive power in revealing latest research, connotations, and themes than terms merely co-occurring in the same document.

2. **Time Point Principle**: Within a document set, the earlier a document containing a natural language word pair is published, the higher its novelty degree [13].

3. **Natural Language Word Pair Inverse Document Frequency Principle**: The value of co-occurring natural language word pairs in quantifying a document's theme novelty decreases as the number of previously published documents containing that pair increases [13].

Based on these principles, we define:

Definition 1: Natural Language Term Time Inverse Document Frequency
If t is a natural language term in document D, and N documents published before D contain term t, then N+1 is called the document frequency of term t relative to D, denoted as NLT_{IDF}(D, t). The reciprocal of N+1 is called the time inverse document frequency of term t relative to D, denoted as NLTIDF(D, t).

Definition 2: Natural Language Word Pair Time Inverse Document Frequency
If t1 and t2 are two natural language terms co-occurring in the same sentence of document D, and N documents published before D contain both terms t1 and t2 in the same sentence, then N+1 is called the document frequency of word pair (t1, t2) relative to D, denoted as NLPT_{IDF}(D, t1, t2). The reciprocal of N+1 is called the time inverse document frequency of word pair (t1, t2) relative to D, denoted as NLPTIDF(D, t1, t2). Clearly, NLPTIDF(D, t1, t2) $\geq$ max(NLTIDF(D, t1), NLTIDF(D, t2)).

Definition 3: Document Sentence Novelty
The average value of all natural language word pair time inverse document frequencies in sentence S of document D is called the novelty of sentence S in document D, denoted as NOV(D, S):

$$NOV(D,S) = \frac{\sum_{1 \le i \le j \le n} NLPTIDF(D, t_i, t_j)}{n(n-1) \times 0.5}$$

where t_i and t_j are the ith and jth natural language terms in sentence S of document D. Clearly, NOV(D, S) (0, 1].

Definition 4: Document Theme Novelty
If document D contains K sentences, the average of all K sentence novelty values is the document's theme novelty, denoted as NOV(D, N):

$$NOV(D,N) = \frac{1}{K} \sum NOV(D, S_K)$$

where S_K is the Kth sentence's novelty in the document. NOV(D, N) (0, 1].

2.3 Research Tools

We selected MetaMap, a mapping tool from free text to UMLS super-thesaurus developed by the U.S. National Library of Medicine [15-16], for its ability to automatically extract natural language vocabulary with high sensitivity to emerging concepts and novel terminology—crucial for novelty calculation. MetaMap directly processes MEDLINE-format data to extract natural language terms. We used "MetaMap Result Processing Software" independently developed by China Medical University's Medical Informatics Department to process MetaMap's batch results, generating term frequency counts and co-occurrence matrices. This software extracts each term's occurrence in every sentence, creating records with: ID (program sequence number), article (PubMed PMID), word (best matching term from MetaMap mapping), classes (semantic type), part (location in abstract: ti for title, ab for abstract), and sentence (sentence number). The natural language method's same-sentence co-occurrence refers to word pairs with identical article, part, and sentence values, saved in Excel format for novelty calculation. From 401 documents, we extracted approximately 50,000 term records, forming about 170,000 word pairs.

We then used Oracle 10g database [17-18] with PL/SQL programming [18] to calculate each article's novelty according to our defined algorithm. We compared these novelty values with F1000 scores (FFa) and SCI citation frequencies to analyze feasibility and 优缺点.

---

3. Experimental Results and Conclusions

3.1 Overall Document Theme Novelty and Partition Analysis

Partial results are shown in Table 1 and Table 2 .

The natural language method calculated novelty values for all 401 documents in the dataset, ranging from 1 (maximum) to 0.525 (minimum). Results were divided into six intervals, with an average novelty of 0.8713. Documents exceeding the average numbered 216, accounting for 53.87% of the total. Documents with novelty = 1 represent the earliest published literature in the set, containing entirely new natural language word pairs and serving as the reference standard for identifying novel information in subsequent documents. The novelty distribution differences are not particularly pronounced, with most documents (80.05%) concentrated in the (1, 0.8] interval. This likely results from PubMed's backend similarity calculation during dataset collection, yielding highly relevant documents with minimal differences. We anticipate that datasets obtained through direct topic-based retrieval would show more distinct partitioning and variation.

3.2 Novelty of F1000 Recommended Literature vs. SCI Citations, F1000 Scores, and Journal Impact Factors

Table 3 and Table 4 present the aggregated results and partitioning.

The natural language method calculated novelty for all 33 F1000-recommended documents, with values ranging from 0.525 to 0.992. Results were divided into six intervals, with an average novelty of 0.8155. Twenty documents exceeded the average, representing 60.61% of F1000-recommended literature—a higher proportion than the overall dataset (53.86%).

We anticipated statistical relationships between document theme novelty and citation frequency, F1000 scores, and journal impact factors. However, after applying several statistical methods, we found no statistically significant correlations. Document theme novelty is not equivalent to F1000 scores or citation metrics; these belong to different evaluation dimensions and categories.

F1000 represents expert evaluations of academic content from perspectives of innovation, importance, rationality, and methodology. Most recommended articles score 1, following a long-tail distribution. F1000's initial goal was to provide rapid post-publication expert assessment of expected impact for high-impact paper filtering [20]. However, a paper's true impact requires years to measure, influenced by research field, publication delays, journal accessibility, and citation cycles [21]. In our study, although we retrieved recently recommended literature, some high-quality papers dated back 5-10 years or more.

Journal impact factor represents a journal's two-year citation performance and cannot individually reflect a single paper's novelty or influence. Song Liping et al. [19] found that journal impact factor and single paper influence can diverge, a conclusion our results support.

3.3 Citation Patterns in Overall Dataset vs. F1000-Recommended Literature

Table 5 compares citation patterns.

Among 401 dataset documents, 368 (91.77%) were SCI-indexed, with 33 (8.23%) not indexed. Citation counts ranged from 0 to 4,822. Twenty-nine documents (7.88% of SCI-indexed set) exceeded 100 citations, while 46 (12.5%) had zero citations. The average citation count was 42.98, with 64 documents (17.39%) exceeding this average.

Among 33 F1000-recommended documents, 32 (96.97%) were SCI-indexed. Citation counts ranged from 0 to 4,822, with 12 documents (36.36% of SCI-indexed set) exceeding 100 citations and 4 (12.12%) having zero citations (all published within three months of retrieval date). The average citation count was 243.125, with 5 documents (15.15%) exceeding this average. F1000-recommended literature shows generally higher citation counts than the overall dataset, demonstrating its high value.

Given the large time span, we compared publications by year: In 2002, 6 documents exceeded 100 citations (3 F1000-recommended); in 2003, 2004, and 2006, only F1000-recommended documents exceeded 100 citations; in 2007, 7 documents exceeded 100 citations (5 F1000-recommended); in 2008, among 15 documents, only 2 F1000-recommended documents exceeded 100 citations; in 2011, F1000-recommended documents ranked top two in citation frequency; in 2013 and 2014, differences were less pronounced due to recent publication dates and citation cycles, though the highest-cited 2014 documents were F1000-recommended. Additionally, 8 dataset documents were not yet indexed in SCIE due to database delay.

---

4. Discussion

4.1 Significance of Document Theme Novelty Detection

Innovation is the soul of academic activity. As scientific research outcomes, academic works are characterized by "new content and new knowledge," quantified through knowledge units or information volume [22]. Our proposed document theme novelty provides a content-based literature evaluation method that statistically analyzes term frequency and trend patterns to determine a paper's novelty within a document set. The natural language method reflects theme novelty through same-sentence co-occurring word pair inverse document frequencies, essentially searching for information not present in previous documents.

While novelty is necessary but not sufficient for innovation—novel literature may not necessarily have high impact—it still holds research value for identifying latest developments and disciplinary trends. We can use document theme novelty as a reference indicator for literature recommendation, selecting highly novel and innovative documents (those proposing new viewpoints, methods, or theoretical explorations) to help researchers understand cutting-edge developments and improve reading efficiency.

4.2 Feasibility Analysis and 优缺点 of Natural Language Method

The natural language method operates on same-sentence co-occurrence, which we argue has stronger latent connections and greater significance in revealing concepts, themes, and connotations than same-document co-occurrence. Using natural language vocabulary extracted from titles and abstracts—unstandardized natural terms—the method can reveal thematic meaning to some extent. It operates without time constraints, calculating novelty across entire document sets. As MetaMap's source vocabulary continuously updates, it can extract novel and recently emerged scientific terminology, providing high value for revealing emerging thematic concepts through novelty calculation.

4.3 Relationships Between Document Theme Novelty, F1000, and Citation Metrics

This study confirms low correlation between natural language method novelty and F1000 scores or citation metrics. High novelty doesn't guarantee high F1000 scores or citations, and low novelty doesn't preclude them. Literature novelty, influence, and quality belong to different evaluation dimensions.

Citation relationships demonstrate knowledge inheritance and utilization, marking scientific development [23]. However, citation volume has time lags and Matthew effects, limiting its utility for analyzing emerging topic novelty, and bears no necessary relationship to document novelty.

Thomas Kuhn's [24] scientific paradigm concept identifies two innovation types: cumulative progressive research within existing paradigms, and revolutionary, high-risk, transformative research causing scientific revolutions. Peer review is the primary mechanism for evaluating scientific research [25], representing expert opinions without necessary connection to value, influence, or novelty.

Du Jian, Tang Xiaoli, and Wu Yishan's team [26] studied factors affecting differences between peer review and citation metrics in paper evaluation. They found F1000 experts tag papers with identifiers; those labeled "new finding," "confirmation," "technical advance," "review comment," and "systematic review/meta-analysis" receive relatively high citations but few recommendations (mostly confirmatory and evidence-based research). Papers tagged "interesting hypothesis," "controversial," "refutation/overturn," "new drug target," or "can change clinical practice" receive high recommendations but fewer citations (mostly transformative and translational research). This shows citation behavior reflects knowledge relationships within academic communities, while peer review better evaluates transformative, high-risk research whose potential to overturn existing paradigms and clinical applicability is best judged by practitioners.

A *PNAS* study analyzed scientific peer review effectiveness. K. Siler et al. [27] examined 1,008 manuscripts submitted to *Annals of Internal Medicine*, *BMJ*, and *The Lancet* in 2003-2004, finding these journals rejected many subsequently highly-cited papers, including 14 of the most-cited papers (12 rejected by ed-

itors). This suggests peer review effectively predicts "good" papers but may struggle to identify exceptional or breakthrough research [27].

Song Liping et al. [28] found F1000 factors and Web of Science show significant positive correlation, with bibliometric indicators and expert peer review yielding consistent conclusions to some extent. However, some highly F1000-rated articles lack high citations, and both methods have limitations for single-paper quality evaluation. Bibliometric indicators may miss important papers that experts rate as excellent [13]. Song Liping et al. [19] further noted in digital-era scientific impact evaluation that F1000, Mendeley, and traditional bibliometric indicators represent a multi-dimensional evaluation landscape.

---

5. Limitations of This Study

5.1 Natural Language Word Pair Extraction Constrained by MetaMap

Natural language word pair extraction depends on MetaMap, whose performance critically constrains our calculations. As its vocabulary source continuously updates, MetaMap's effectiveness in extracting emerging scientific terminology directly affects calculated novelty values.

5.2 Literature Collection Method

Our dataset collection aimed to calculate novelty across literature with certain relevance within a discipline, expecting successful partitioning with targeted scope. Direct free-term retrieval methods may yield different results with more pronounced distribution differences.

5.3 MetaMap Result Processing Software and Operational Applicability

Since MetaMap data sources constantly update while our processing software (developed by China Medical University in 2010) may have compatibility issues with newer data formats, objective errors could affect results. Manual removal of stop words, numerals, pronouns, prepositions, and symbols without substantive meaning involves subjectivity that may influence outcomes. Additionally, sentences yielding only one natural language term cannot form word pairs and were excluded, potentially affecting results.

5.4 Lack of Weighting

Our method treats natural language word pairs from titles and abstracts equally. However, given titles' importance, different weight assignments for title versus abstract terms should be considered in future research.

5.5 Lack of Evaluation Method for Calculated Results

We lack an effective evaluation method for novelty results. We considered expert evaluation but abandoned it due to experts' busy schedules, limited availability, and subjectivity. Using F1000 scores also proved unsuitable due to subjectivity and dimensional differences. TREC evaluation metrics (recall, precision, F-value) [1] apply to prediction results with established answers and are unsuitable here. Future research should select appropriate scientific paper evaluation indicators [29] for validation.

---

6. Conclusion

1. This study confirms the natural language method's feasibility for document theme novelty calculation.
2. Document theme novelty is not equivalent to F1000 recommendations or citation metrics; they belong to different evaluation dimensions and should not be conflated.
3. The document theme novelty indicator should be combined with peer review, bibliometrics, and other evaluation metrics for comprehensive literature analysis and quality recommendations.

Future research will compare medical subject heading word pair methods with natural language methods on the same dataset, explore different data collection approaches, and carefully select multiple appropriate evaluation indicators for validation.

References

[1] Xing Meifeng, Guo Shiming. Review of text content novelty detection research [J]. Information Science, 2011, 29(7): 1098-1103.
[2] HARMAN D. Overview of the TREC 2002 novelty track [EB/OL]. [2017-01-08]. http://trec.nist.gov/pubs/trec11/papers/NOVELTY.OVER.pdf?origin=publication_{detail}.
[3] ZHANG Y, TSAI FS. Chinese novelty mining [EB/OL]. [2017-01-08]. http://www.aclweb.org/anthology/D09-1162.
[4] KUMARAN G, ALLAN J. Text classification and named entities for new event detection [EB/OL]. [2017-01-08]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.9552&rep
[5] RAJARAMAN K, TANA H. Topic detection, tracking, and trend analysis using self-organizing neural networks [EB/OL]. [2017-01-08]. http://www3.ntu.edu.sg/home/asahtan/papers/trac_{kdd01}.pdf.
[6] Expansion-based technologies in finding relevant and new information: the TREC 2002 novelty track experiments [EB/OL]. [2017-01-09]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.8780&rep=rep1&type=pdf.
[7] ZHANG HP, SUN J, WANG B, et al. Computation on sentence semantic distance for novelty detection [J]. Journal of Computer Science and Technology,

2005, 20(3): 331-337.
[8] TSAI FS, ZHANG Y. D2S: document-to-sentence framework for novelty detection [J]. Knowledge & Information Systems, 2011, 29(2): 419-433.
[9] Shen Lü. General equilibrium theory of scientific innovation—scientometric analysis of scientific achievement innovation evaluation [J]. Science Research, 2003, 21(2): 205-209.
[10] Shen Yang. A keyword-based innovation evaluation method [J]. Information Theory and Practice, 2007, 30(1): 125-127.
[11] Hu Shuli, Zhang Jinghui. A mathematical model for measuring scientific literature novelty [J]. Information Theory and Practice, 1995(5): 23-24.
[12] Qian Lingfei, Yang Jianlin, Zhang Li. Disciplinary innovation comparison based on keyword analysis—taking information and library science as examples [J]. Information Theory and Practice, 2011, 34(1): 117-120.
[13] Yang Jianlin, Qian Lingfei. Theme novelty measurement method based on keyword pair inverse document frequency [J]. Information Theory and Practice, 2013, 36(3): 99-102.
[14] Xue Chen. Bibliometric analysis of international big data research papers [J]. Modern Information, 2013, 33(9): 129-139.
[15] MetaMap: a tool for recognizing UMLS concepts in text [EB/OL]. [2015-03-10]. http://metamap.nlm.nih.gov/.
[16] Zhang Yunqiu, Leng Fuhai. Study on MetaMap's text mapping principle and its impact on retrieval effectiveness [J]. Journal of Information Science, 2007, 26(3): 344-349.
[17] Baidu Baike. Oracle database [EB/OL]. [2015-03-20]. http://baike.baidu.com/view/1685727.htm.
[18] Baidu Baike. PL/SQL [EB/OL]. [2015-03-20]. http://baike.baidu.com/link?url=TOjaqL199OyPA1Gk0UK
[19] Song Liping, Wang Jianfang, Wang Shuyi. Comparison of F1000, Mendeley and traditional bibliometric indicators from scientific evaluation perspective [J]. Journal of Library Science in China, 2014, 40(7): 48-54.
[20] Liu Chunli. F1000 factor: a scientific paper impact evaluation method based on soft peer review [J]. Chinese Journal of Scientific and Technical Periodicals, 2012, 23(2): 383-386.
[21] Brody T, Harnad S, Carr L. Earlier web usage statistics as predictors of later citation impact [J]. Journal of the American Association for Information Science and Technology, 2006, 57(8): 1060-1072.
[22] Ren Quan'e. Innovation evaluation of humanities and social science research based on informatics [J]. Information and Documentation Services, 2009(2): 20-23.
[23] Qiu Junping. Literature citation patterns and citation analysis methods [J]. Information Theory and Practice, 2001, 24(3): 236-240.
[24] KUHN TS. The structure of scientific revolutions [M]. Chicago: University of Chicago Press, 2012.
[25] Scientists analyze peer review effectiveness [EB/OL]. [2015-03-10]. http://paper.sciencenet.cn/htmlpaper/201511219413977135306.shtm.
[26] DU J, TANG XL, WU YS. The effects of research level and article type on the differences between citation metrics and F1000 recommendations [J]. Journal of the Association for Information Science and Technology, 2016,

67(12): 3008-3021.
[27] SILER K, LEE K, BERO L. Measuring the effectiveness of scientific gatekeeping [J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(2): 360-365.
[28] Song Liping, Wang Jianfang. Correlation study between peer review and bibliometrics based on F1000 and Web of Science [J]. Journal of Library Science in China, 2012, 38(2): 62-69.
[29] Wang Wenxia, Liu Chunli. Research on differences in paper impact evaluation indicator models across disciplines [J]. Library and Information Service, 2017, 61(13): 108-116.

Author Contributions

Xu Dan: Constructed the paper framework, wrote and revised the manuscript
Xu Shuang: Supplemented and revised portions of the paper
Chen Sisi: Provided guidance and suggestions for research topic selection
Han Shuang: Collected and cleaned data
Yang Ying: Collected and cleaned data
Guo Jijun: Proposed research topic and 思路, revised the manuscript

---

Xu Dan, Xu Shuang, Chen Sisi, Han Shuang, Yang Ying, Guo Jijun. Document Theme Novelty Detection Research Based on Natural Language Pairs [J]. Library and Information Service, 2018, 62(8): 130-138.

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