Design and Postprint of an Intelligent Search Term Extraction System for Scientific and Technical Novelty Search

Authors: Wang Peixia, Yu Hai, Chen Li, Wang Yongji

Date: 2017-11-08T00:00:00+00:00

Abstract

Abstract

Purpose: To address issues of strong subjectivity, heavy manual workload, lack of standardization, and time-consuming, labor-intensive processes in search term selection for scientific and technological novelty searches.

Application Background: To achieve automation, intelligence, and standardization of the search term extraction process, this paper proposes utilizing real-time relevant corpora identified during the novelty search process as a source of domain knowledge and discusses the relationship between corpus composition types and keyword extraction effectiveness.

Method: Intelligent extraction of search terms in the scientific and technological novelty search domain is achieved through a progressive iterative extraction approach that combines keyword extraction with domain feature expansion.

Results: Through comparison with search terms employed in actual novelty search cases, it was found that extracting 10 search terms using this method after two iterations achieved a recall rate of 80%.

Conclusion: Iterative extraction of search terms based on dynamic relevant corpora composed of literature identified during the novelty search process facilitates rapid and accurate identification of the vast majority of search terms, thereby improving retrieval efficiency and effectiveness.

Full Text

Design and Implementation of an Intelligent Search Term Extraction System for Sci-Tech Novelty Retrieval

Wang Peixia[12], Yu Hai[12], Chen Li[12], Wang Yongji[1]
[1]Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: [Objective] To achieve automation, intelligence, and standardization in the search term extraction process, this paper proposes utilizing real-time relevant corpora identified during sci-tech novelty retrieval as the source of domain knowledge, and discusses the relationship between corpus composition types and keyword extraction effectiveness. [Methods] We implement intelligent search term extraction for sci-tech novelty retrieval through a progressive iterative approach that combines keyword extraction with domain feature expansion. [Results] Compared with search terms used in actual novelty retrieval cases, our method achieves a recall rate of 80% when extracting 10 search terms after two iterations. [Conclusions] Iterative extraction of search terms based on dynamic relevant corpora composed of retrieved documents during the novelty search process helps quickly and accurately identify the majority of search terms, thereby improving retrieval efficiency.

Keywords: Sci-tech novelty retrieval; Search terms; Keyword extraction; Web crawler
Classification Number: TP391

Sci-tech novelty retrieval is an information consulting service that verifies the novelty of research projects through literature search, comparison, and analysis. Based on retrieval scope, it can be categorized as domestic, foreign, or international novelty retrieval. Literature retrieval forms the foundation of sci-tech novelty retrieval and involves constructing search queries, where the selection of search terms plays a critical role and represents one of three key factors affecting retrieval quality. Particularly in international novelty retrieval, the accuracy of English search terms directly impacts the novelty, comprehensiveness, and accuracy of retrieval results.

Search terms refer to substantive words that characterize the thematic content of a novelty retrieval project—essential keywords for revealing and describing the project' s subject matter. Currently, domestic sci-tech novelty consulting agencies typically use both controlled terms (subject terms) and free terms (keywords) for literature retrieval. Controlled terms, also known as descriptors, represent standardized retrieval language derived from thesauri that normalize synonyms and near-synonyms for a given concept. However, literature databases often suffer from inconsistent indexing, and novelty search professionals must rely on specialized thesauri or controlled vocabulary lists that remain relatively static due to long maintenance cycles. These cannot promptly incorporate terms representing new technologies, methods, or theories, which is problematic since

sci-tech novelty retrieval must reflect innovation in technology, methods, and theory. Consequently, using controlled terms may lead to missed relevant literature. Keywords (or free terms), by contrast, are indexed and retrieved from document titles, abstracts, keywords, and even full texts, offering greater flexibility without concerns about indexing standardization or thesaurus update timeliness. They represent a crucial retrieval pathway for electronic information resources and constitute the focus of this research.

Currently, novelty search professionals manually complete the discovery, screening, supplementation, expansion, and final selection of search terms. Based on project materials provided by clients, they use client-supplied keywords as references, combining scientific and technical points, novelty claims, and supplementary materials to initially identify keywords matching the retrieval theme. When necessary, they perform keyword expansion using professional thesauri, dictionaries, terminology standards, and other reference tools, as well as retrieved literature, to obtain standardized names and synonyms. Additionally, professionals conduct trial searches during retrieval, analyzing retrieved documents to evaluate search term appropriateness and make adjustments. This process typically repeats multiple times to achieve satisfactory results.

Evidently, search term selection involves an iterative process of literature retrieval, document browsing, analysis, synthesis, and adjustment, requiring multiple search requests to various databases and repeated trial searches. After each trial, retrieved documents must be analyzed to adjust search terms, with multiple cycles needed for final determination. This process is labor-intensive, time-consuming, and tests the patience of search professionals. Moreover, it heavily depends on professionals' expertise, experience, and knowledge structure, introducing strong subjectivity that is difficult to standardize, thereby directly affecting retrieval effectiveness and novelty report quality.

To overcome these challenges—subjectivity, heavy manual workload, lack of standardization, and time consumption—this study introduces dynamic literature corpora related to sci-tech novelty projects. Grounded in project information, we use real-time, dynamically acquired corpora related to the novelty project as the domain knowledge source, adopting a progressive iterative approach that combines keyword extraction with domain feature expansion.

2.1 Automatic Keyword Extraction

Automatic keyword extraction technology finds extensive applications in literature retrieval, automatic abstracting, text clustering, and classification. In information retrieval, effective keywords can supplement full-text indexing and help users discover relevant documents. The automatic keyword extraction process typically involves two steps: (1) candidate keyword identification using heuristic rules (removing stop words; retaining only nouns, adjectives, and verbs; using external resources like Wikipedia; N-gram methods, etc.) to extract words or phrases as candidates; and (2) candidate keyword selection using supervised or

unsupervised methods to determine correct keywords.

Most research focuses on candidate selection, divided into supervised and unsupervised approaches. Early supervised methods treated keyword extraction as a classification problem, using annotated corpora to train models that determine whether words belong to keyword categories, employing algorithms such as Naive Bayes, decision trees, maximum entropy, multi-layer perceptrons, and vector space models. However, supervised methods require costly manual annotation using author-provided or expert-annotated keywords. Additionally, classifiers evaluate words independently, while research shows keyword selection is not independent—previously selected keywords influence subsequent selections.

Keyword extraction techniques fall into three categories: statistical feature-based, topic model-based, and graph model-based methods. Statistical feature-based methods calculate word features (term frequency, N-gram, TF-IDF, information entropy) combined with positional markers (title, paragraph beginning, first occurrence) to assign weights for keyword extraction. For example, some researchers used TF-IDF scores and first occurrence position, while others incorporated phrase length, title appearance, distribution patterns, and frequency extremes. Adding linguistic knowledge like noun phrase chunking and part-of-speech tags significantly improves accuracy.

Topic model-based methods predominantly use LDA, which infers "document-topic"and "topic-word"distributions from known "word-document"matrices. This approach assumes words from dominant topics are more likely to be identified as keywords. However, effectiveness depends heavily on training data topic distributions.

Graph model-based methods, exemplified by TextRank, draw inspiration from Google' s PageRank. They construct word graphs where nodes represent candidate keywords and edges represent relationships. When two words appear within an observation window, they establish a connection. Each connecting edge represents a "vote,"with importance determined by connected nodes through iterative computation, ultimately selecting keywords by importance scores.

Beyond document features, keyword extraction incorporates domain knowledge from two sources: (1) lexical resources like thesauri, internet dictionaries, terminology databases, or Wikipedia, where Wikipedia entries serve as independent concepts; and (2) corpora including domain-specific, general, and comparative corpora. Experiments show that documents from the same domain significantly improve extraction effectiveness. However, domain-related corpora require manual acquisition of large volumes, consuming substantial time and resources. Moreover, these static corpora become outdated over time, lacking timeliness. In scientific literature keyword extraction, some approaches use titles and keywords as seed words, employing Word2Vec on open-domain corpora to find similar candidates, though open-domain corpora lack domain relevance advantages. Others use comparative corpora for domain relevance calculation, but corpus selection and size directly affect extraction results, and manual ac-

quisition remains challenging.

2.2 Search Term Recommendation

Related research includes data mining-based search term recommendation techniques, typically using rule-based, content-based, collaborative filtering, or hybrid methods. These systems rely on historical behavior records like search logs, modeling user behavior to discover patterns. However, such recommendation techniques build upon existing system operations, focusing on static data mining, whereas our research emphasizes dynamic extraction from retrieved relevant literature, representing a different focus.

2.3 Main Contributions

This study extracts search terms based on novelty project titles, keywords, scientific points, novelty claims, and knowledge from retrieved literature, automatically extracting domain feature words related to input search terms as candidates. Key contributions include: (1) Using dynamic literature corpora related to novelty projects for extraction, unlike traditional keyword extraction targeting single documents or static corpora. Our extraction objects consist of multiple relevant documents from novelty applications and retrieval processes, characterized by domain relevance, large volume, rich content, and authority, obtained through web crawlers for synchronization with data sources, ensuring dynamic and real-time updates. (2) Emphasizing domain knowledge introduction. Traditional keyword extraction for indexing limits itself to titles, abstracts, and full texts, while we additionally utilize author-assigned keywords, which represent basic domain concept elements with strong indicative and discriminative power, making them crucial search term sources. (3) Our extracted candidate search terms assist professionals in quickly identifying relevant terms, leveraging the domain expertise inherent in author-assigned keywords to prevent missed detection and improve recall and precision in international novelty retrieval—distinctly different from indexing-oriented extraction. (4) The extraction process is dynamic and progressive. Traditional extraction completes in one pass, whereas novelty retrieval term extraction is interactive and dynamic, with professionals iteratively adjusting terms through repeated searches to achieve satisfactory results.

Based on these characteristics, we first conduct statistical sampling of author-assigned keywords' distribution in titles and abstracts to analyze their relationship. Then, using retrieved corpora, we construct word graphs based on co-occurrence relationships among titles, keywords, and abstracts, using term frequency as association strength for candidate extraction experiments, comparing contributions of the three components. Finally, we generate usable search terms from dynamic corpora through the process shown in [Figure 2: see original paper], implementing intelligent extraction via progressive iteration combining keyword extraction and domain feature expansion, demonstrated through actual cases.

The intelligent search term extraction system for sci-tech novelty retrieval comprises two components: web crawler-based online literature retrieval and intelligent search term extraction. Since extraction depends on novelty projects and dynamic corpora, corpus acquisition forms a crucial system component. The system uses Spring Web MVC and Hibernate frameworks—lightweight web and object-relational mapping frameworks respectively—with MySQL storing collected literature information, achieving separation of data, business, and presentation layers.

3.1 Corpus Acquisition

Corpus acquisition primarily involves using web crawlers to online retrieve scientific literature information identified by various databases. The process is shown in [Figure 1: see original paper]. Web crawlers are programs that fetch web content by analyzing webpage structures using format characteristics. In our system, data preprocessing analyzes webpage tag structures to extract titles, abstracts, keywords, authors, and other information for local database storage. The system maintains search logs to address time consumption from repeated popular queries, displaying previous crawl results for duplicate searches, and implements deduplication to prevent redundant data collection.

3.2 Intelligent Search Term Extraction

The intelligent search term extraction process, shown in [Figure 2: see original paper], consists of: (1) Corpus acquisition: Automatically generating search terms from project titles, constructing queries for cross-database retrieval, and storing retrieved literature locally; (2) Candidate extraction: Automatically extracting 10 candidate search terms from the dynamic corpus; (3) Term expansion: Performing domain feature expansion on candidates to generate a search term list; (4) Evaluation: Manually verifying if the term list meets requirements, and if not, selecting appropriate terms to manually construct queries for repeating steps 1-3; (5) Merging: Combining terms by importance to generate a final search term list.

4. Search Term Extraction Method

Important terms typically appear with higher probability in domain-specific scientific literature corpora. Scientific documents feature rich structures including titles, abstracts, and author-assigned keywords, with professional domain-specific language. Search terms closely relate to novelty projects and often constitute professional terminology. Leveraging domain relevance between retrieved corpora and novelty projects, we initially obtain domain-related corpora using word groups from project titles to construct queries. Based on retrieved corpora, we extract indicative keywords and perform domain-specific expansion to generate candidate search terms. This method assumes that project titles provide concise thematic descriptions while scientific points offer detailed descriptions.

Since titles cannot reflect all domain features, we employ multiple iterative retrieval rounds. In subsequent iterations, professionals select terms from generated lists to construct queries for acquiring domain-related corpora. The extraction process comprises three steps: candidate term extraction from retrieved literature, domain-specific expansion, and merging.

4.1 Candidate Search Term Extraction Corpus Analysis: Candidate extraction uses literature information returned by databases based on received queries. Crawlers capture titles, keywords, abstracts, and other information as extraction objects. Scientific literature keywords are natural, uncontrolled terms selected for indexing to represent thematic content. In bibliometrics, researchers consider keywords as basic elements representing domain concepts, used to analyze knowledge structure characteristics at macro levels or examine research topic details and relationships at micro levels using "important" words. Thus, keywords essentially reflect domain knowledge structure and thematic characteristics, serving as domain feature words with strong indicative and discriminative power in specific fields—making them crucial sources for high-quality search terms.

To observe keyword distribution in titles and abstracts, we conducted sampling statistics based on CNKI and Wanfang databases, with results shown in [Figure 3: see original paper]. Here, we use average ratio avg(t) = |in(t)| / |t|, where in(t) represents the number of keywords appearing in titles, abstracts, or their combination, and |t| represents total keyword count. The analysis reveals that approximately 50% of keywords appear in titles, while abstracts contain more keywords—especially in CNKI, with average rates exceeding 80%. The combination of titles and abstracts achieves over 80% keyword coverage, making author-assigned keywords a vital candidate source.

To compare extraction effectiveness from titles, keywords, and abstracts, we conducted experiments using each component separately and in combination. **Improved TextRank Candidate Extraction:** Classic TextRank represents a document as an undirected graph where nodes are words and edges connect any two words within a given text window. A node's importance score consists of contributions from neighboring nodes, calculated via PageRank methods, with equal association strength between adjacent words. However, in sci-tech novelty retrieval, retrieved literature volumes are typically large, with documents interconnected through search terms. Words with higher frequency reflect thematic tendencies. Therefore, we improve classic TextRank (denoted MF_{TR}) using term frequency as an importance factor, with the importance transfer matrix also based on term frequency. The importance score calculation formula is:

$$Score(i) = (1-d) + d \times \sum_{j \in set(i)} \frac{tf(i)}{\sum_{k \in set(j)} tf(k)} score(j)$$

where set(i) represents co-occurring words of term i, and tf(i) is term frequency.

Each word's importance score is computed via this formula, sorted in descending order, with the top 10 high-importance words selected as candidate search terms.

4.2 Candidate Search Term Expansion We expand obtained keywords using domain features from retrieved literature keywords or abstracts. Since most Chinese POS tagging systems train on news corpora where scientific terms rarely appear, and segmentation systems may incorrectly split professional terms (for instance, the complete term 'Ixodes persulcatus' might be erroneously segmented as just the partial fragment representing the first part of the full term), we correct such errors by searching the retrieved keyword set for domain feature keywords containing each extracted keyword. If found, they become candidate search terms with calculated importance scores. Considering abstracts as summarizing descriptions, we use candidate term frequency in abstracts as an importance adjustment factor, combined with phrase characteristics (Formula (3)) to calculate final importance scores (Formula (4)):

$$GDC(T) = |T| \times \log \frac{freq(T)}{\sum_{t \in T} freq(t)}$$

$$DGDC(T) = tf(T) \times GDC(T)$$

where T is a candidate search term, |T| represents the number of words (not characters) it contains, N is the number of retrieved documents, and tf(T) is candidate term frequency in descriptive text (abstract or full text). For cases where T appears zero times, we assign a specific initial value of 0.1F, meaning that even if T has high phrase characteristics as a domain keyword, its final domain importance score will be lowered if absent from descriptive text. In practice, tf(T) serves as a weighting factor regulating the final domain importance score.

4.3 Search Term Merging Since retrieved literature significantly influences search terms, the above steps can be repeated. Except for the first automatic query generation, subsequent iterations involve manual query construction using terms from generated lists or project keywords. Finally, expanded term sets are sorted by importance, with the top 10 selected as final search terms.

5.1 Data Sources

We construct queries and send retrieval requests to various literature databases, using the open-source HtmlParser tool to extract titles, keywords, authors, journal names, issues, abstracts, controlled terms, and uncontrolled terms for local database storage. The platform currently supports Chinese (Wanfang, CNKI) and English (Web of Science, EI) retrieval.

5.2 Data Preprocessing

Preprocessing includes format standardization, deduplication, and text processing. **Format Standardization:** Different databases employ different standards, necessitating unified, standardized processing. **Deduplication:** Chinese databases like CNKI, Wanfang, and CQVIP exhibit duplicate indexing (93.6% overlap between Wanfang and CQVIP, 94.1% between CNKI and CQVIP). We classify duplicates into: (1) database duplicate indexing, where identical literature appears multiple times in one or more databases; and (2) duplicate publication, where one research outcome is published in different journals. Our deduplication algorithm uses paper titles, first authors, journals, and years as criteria, categorizing detected duplicates for different handling strategies—removing type 1 duplicates while temporarily retaining type 2.

Text Processing: This step identifies candidate keywords from retrieved literature. Using HanLP's POS tagger, we segment and tag literature, removing punctuation, numerals, discriminators, conjunctions, interjections, onomatopoeia, prepositions, measure words, auxiliaries, modal particles, status words, and pronouns, while retaining other POS types. We also remove stopwords, adding common abstract terms to the general stopword list (e.g., research, has, adopts, conduct, results show, application, method, problem, analysis).

5.3 Candidate Search Term Extraction Experiments

Compared Methods: We evaluate four common keyword extraction methods: Most Frequent (MF), TF-IDF, LDA, and TextRank (TR), selecting the top 10 importance-scored words as candidates. **MF** uses term frequency as importance: score(t) = f(t). **TR** importance scores are calculated as: score(t) = (1-d) + d $\times$ Σ_{t_i,j} (δ{t_i,t_j} $\times$ *score(t_j) / Σ{t_j,k}* e_{t_j,t_k}), where d = 0.85 is the damping factor, and δ indicates co-occurrence. **TF-IDF** calculates importance as: score(t) = TF-IDF(t) = f(t) $\times$ log(N/n(t)), where f(t) is total occurrence count, N is total documents, and n(t) is documents containing term t. **LDA** calculates word probability across t topics as: P(w|D) = Σ_z _{z|D} $\times$ ϕ_{w|z}, where represents document-topic distribution and ϕ represents topic-word distribution, selecting top σ words (σ = 10).

Corpus and Extraction Effectiveness: Using "Graphene in Lithium Batteries: Applications and Prospects" as an example, we control online retrieval time and avoid excessive irrelevant literature by limiting each database to $ $110 records, employing progressive, interactive experimental approaches. Searching "graphene" in Wanfang yields 110 records. Using MF_{TR} (with co-occurrence window w = 5 and iteration termination at >200 iterations or difference $ $0.001f), we extract 10 candidates from titles, keywords, and abstracts separately, shown in . The 16 distinct candidates across three sources share only 5 terms (31.25% overlap). For "lithium battery" , 20 distinct candidates share only 3 terms (15% overlap). This demonstrates varying co-occurrence distributions across different queries. Keyword-based extraction shows better

synonym/hypernym-hyponym relationships than abstract-based methods, with superior domain professionalism and comprehensiveness.

To verify whether combining keywords with titles/abstracts improves quality, we conducted experiments using "graphene" with MF_{TR} on keyword+title, keyword+abstract, and all three combinations, shown in . Results show three new candidates enter the top 10 when combining keywords with titles or abstracts. Keyword+abstract performance approximates the three-component combination, while keyword+title replaces three terms (graphite, structure, nano with preparation, performance, preparation) without significant improvement.

Increasing retrieved documents from 110 to 231 for "graphene" yields results in . Within the observation window (top 10 words), increased document volume only minimally affects extracted terms, limiting changes to one word and thus having negligible impact on final results while consuming more resources and time.

Method Comparison: Searching "graphene and lithium battery" in Wanfang yields 131 documents. Comparative extraction results from titles, abstracts, and keywords using MF, TR, TF-IDF, and LDA are shown in through . **Title-based methods** produce consistent results across algorithms with minor positional variations due to limited title text. **Abstract-based methods** show 70% overlap with title-based methods but different term rankings. **Keyword-based methods** outperform title/abstract combinations because author-assigned keywords, carefully selected to represent thematic content, offer natural advantages in topicality and professionalism. **Combined methods** tend toward title+abstract effects since their larger word counts dominate frequency and co-occurrence metrics, overshadowing the smaller keyword set.

MF_{TR} essentially integrates traditional TextRank with MF methods. Results show that incorporating term frequency as a weighting factor yields outcomes largely consistent with TR but boosts rankings of terms co-occurring with high-frequency words, aligning with our objective of identifying relevant terms closely associated with frequent query terms.

Iterative Extraction: To verify whether iterative extraction helps professionals quickly locate relevant terms, we compared extracted candidates with terms used in actual novelty retrieval cases. For a project titled "Discovery and Research of Lyme Disease in China" with scientific points covering epidemiological surveys, tick transmission, spirochete analysis, and monoclonal antibody preparation, the final retrieval report included terms: Lyme disease, epidemiology, Ixodes persulcatus, Borrelia burgdorferi, and monoclonal antibody.

First iteration: Using automatically generated query "China and Lyme disease and discovery" retrieved 85 documents. MF_{TR} on keywords returned 10 candidates: Lyme disease, spirochete, test, genotype, polymorphism, epidemiology, serum, diagnosis, virus, Borrelia. After domain expansion using Formula (4), top terms included Lyme disease spirochete, Lyme disease, Borrelia burgdorferi,

epidemiology, epidemiological survey, spirochete, indirect immunofluorescence assay, diagnosis, enzyme-linked immunosorbent assay, genotype.

Second iteration: Using "Lyme disease spirochete" as query retrieved 130 documents. MF_{TR} extraction yielded: Lyme disease, spirochete, protein, genotype, Borrelia, expression, host, part of Ixodes persulcatus, epidemiology, transmission. After expansion: Lyme disease spirochete, Lyme disease, Borrelia burgdorferi, Ixodes persulcatus, genotype, epidemiological survey, restriction fragment length polymorphism, epidemiology, transovarial transmission, spirochete.

Merging: Iteration produces varying scores for the same term across different corpora, using the higher score for final sorting. After two iterations, the final list contains: Lyme disease spirochete, Lyme disease, Borrelia burgdorferi, epidemiology, epidemiological survey, Ixodes persulcatus, spirochete, genotype, indirect immunofluorescence assay, diagnosis. Comparing with the final report, four terms match exactly (Lyme disease, epidemiology, Borrelia burgdorferi, Ixodes persulcatus), achieving 80% recall.

Conclusion and Outlook

This paper proposes an intelligent search term extraction method based on real-time relevant corpora identified during sci-tech novelty retrieval, using progressive iterative extraction combining keyword extraction and domain feature expansion. Comparison with actual cases demonstrates 80% recall for 10 extracted terms after two iterations. Since corpus acquisition via web crawling requires substantial literature and time, future research will seek a practical balance through real-world novelty retrieval practice. Additionally, extraction effectiveness heavily depends on literature database quality, particularly spelling errors affecting English term extraction, making error self-correction a future research direction.

References

[1] Huang Jiangling. Analysis of Important Factors Affecting the Quality of Science and Technology Novelty Search[J]. Information Research, 2008(8): 67-68.

[2] Cao Huanzeng. Some Measures for Increasing the Recall Ratio of Sci-tech Literatures[J]. Sci-Tech Information Development & Economy, 2008, 18(32): 72-74.

[3] Chen Yulin. Keyword Search Method Application Research on Science and Technology Novelty Check[J]. Journal of Henan Normal University: Natural Science Edition, 2011, 39(3): 171-173.

[4] Zhang Baiqiu, Wu Xiaohuang. Keywords Selection in Science Technology Novelty Retrieval[J]. Information Science, 2008, 26(9): 1344-1348.

[5] Hasan K, Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art[C]. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 1262-1273.

[6] Frank E, Paynter G W, Witten I H, et al. Domain-specific Keyphrase Extraction[C]. In: Proceedings of the 16th International Conference on Artificial Intelligence (IJCAI-99), 1999: 668-673.

[7] Turney P D. Learning Algorithms for Keyphrase Extraction[J]. Information Retrieval, 2002, 2(4): 303-336.

[8] Nguyen T D, Kan M-Y. Keyphrase Extraction in Scientific Publications[C]. In: Proceedings of International Conference on Asian Digital Libraries (ICADL), 2007: 317-326.

[9] Lopez P, Romary L. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID[C]. In: Proceedings of International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010: 248-251.

[10] Krapivin M, Autayeu M, Marchese M, et al. Improving Machine Learning Approaches for Keyphrases Extraction from Scientific Documents with Natural Language Knowledge[C]. In: Proceedings of the Joint JCDL/ICADL International Digital Libraries Conference, 2010: 102-111.

[11] Jiang X, Hu Y, Li H. A Ranking Approach to Keyphrase Extraction[C]. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009: 756-757.

[12] Turney P D. Coherent Keyphrase Extraction via Web Mining[C]. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003: 434-439.

[13] Kumar N, Srinathan K. Automatic Keyphrase Extraction from Scientific Documents Using N-gram Filtration Technique[C]. In: Proceedings of the 8th ACM Symposium on Document Engineering, 2008: 199-208.

[14] Pan Limin, Wu Junhua, Lin Meng, et al. Algorithm of Chinese Keywords Extraction Based on Multi-feature[J]. Netinfo Security, 2014(8): 40-44.

[15] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge[C]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2003: 216-223.

[16] Pasquier C. Task 5: Single Document Keyphrase Extraction Using Sentence Clustering and Latent Dirichlet Allocation[C]. In: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010: 154-157.

[17] Shi Jing, Li Wanlong. Topic Words Extraction Method Based on LDA Model[J]. Computer Engineering, 2010, 36(19): 81-83.

[18] Liu Jun, Zou Dongsheng, Xing Xinlai, et al. Keyphrase Extraction Based on Topic Feature[J]. Application Research of Computers, 2012, 29(11): 4224-4227.

[19] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]. In: Proceedings of EMNLP-04, 2004: 404-411.

[20] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[C]. In: Proceedings of the 7th International World Wide Web Conference, 1998: 161-172.

[21] Han Qichen, Li Dongmei. Semantic Model with Thesaurus for Forestry Information Retrieval[J]. Journal of Frontiers of Computer Science & Technology, 2016, 10(1): 122-129.

[22] Xiong Xia. Domain Information Retrieval Based on Term Relationships of Thesaurus[D]. Beijing: Chinese Academy of Agricultural Sciences, 2011.

[23] Hulth A, Karlgren J, Jonsson A, et al. Automatic Keyword Extraction Using Domain Knowledge[C]. In: Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics, 2001: 472-482.

[24] Coursey K H, Mihalcea R, Moen W E. Automatic Keyword Extraction for Learning Object Repositories[J]. Proceedings of the American Society for Information Science & Technology, 2009, 45(1): 1-10.

[25] Li G, Wang H. Improved Automatic Keyword Extraction Based on TextRank Using Domain Knowledge[C]. In: Proceedings of the 3rd CCF Conference, NLPCC 2014, 2014: 403-413.

[26] Jiang B, Xun E, Qi J. A Domain Independent Approach for Extracting Terms from Research Papers[C]. In: Proceedings of the Australasian Database Conference, 2015: 155-166.

[27] Lopes L, Fernandes P, Vieira R. Estimating Term Domain Relevance Through Term Frequency, Disjoint Corpora Frequency-TF-DCF[J]. Knowledge-Based Systems, 2016, 97: 237-249.

[28] Zhan Hengfei, Yang Yuexiang, Fang Hong. Research and Optimization of Nutch Distributed Crawler[J]. Journal of Frontiers of Computer Science & Technology, 2011, 5(1): 68-74.

[29] Lu Ping, Cai Qun. Keyword Indexing of Chinese Scientific and Technical Paper[J]. Academic Journal of Guangzhou Medical College, 2000, 28(2): 93-94.

[30] Guo C, Lu X. Selecting Publication Keywords for Domain Analysis in Bibliometrics: A Comparison of Three Methods[J]. Journal of Informetrics, 2016, 10(1): 212-223.

[31] Hong Daoguang. Research on Data Integration of Google Scholar[J]. Modern Information, 2010, 30(7): 39-41.

[32] Rossi R G, Maracini R M, Rezende S O. Analysis of Domain Independent Statistical Keyword Extraction Methods for Incremental Clustering[J]. Learning and Nonlinear Models, 2014, 12(1): 17-37.

Author Contributions: Wang Peixia: conceptualization, methodology, experimentation, writing; Yu Hai, Chen Li: paper revision; Wang Yongji: research design modification, final version revision.

Conflict of Interest Statement: All authors declare no conflict of interest.

Supporting Data: Supporting data is self-archived by authors, E-mail: peixia@nfs.iscas.ac.cn. [1] Wang Peixia. Experimental Examples.csv. Literature data information.

Received: 2016-07-28 **Revised:** 2016-09-26

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