Postprint: Patent Terminology Extraction Method Incorporating Paper Keyword Knowledge

Authors: Yu Yan, Chen Lei, Jiang Jinde, Zhao Naixuan

Date: 2023-04-01T16:15:57+00:00

Abstract

[Purpose/Significance] To remedy the deficiency that limited information in patent corpora constrains patent term extraction effectiveness, this paper proposes utilizing rich knowledge from paper keywords to obtain effective features beyond patent text, thereby improving patent term extraction accuracy.

[Method/Process] Based on keyword knowledge from relevant papers, two features—domain relevance and head-tail degree—are proposed to measure the likelihood of candidate terms becoming actual terms, which are then integrated into traditional patent term extraction methods.

[Results/Conclusion] Experimental results demonstrate that leveraging domain relevance and head-tail degree information of candidate terms obtained from paper keywords enables methods incorporating such knowledge to achieve significant accuracy improvements over traditional term extraction methods.

Full Text

Patent Term Extraction by Integrating Keyword Knowledge from Papers

Yu Yan[1],[2], **Chen Lei**[1], **Jiang Jinde**[3], **Zhao Naixuan**[1] [1]Information Service Department, Nanjing Tech University, Nanjing 210009 [2]Computer Engineering Department, Southeast University Chengxian College, Nanjing 211816 [3]School of Business, Nanjing Xiaozhuang University, Nanjing 211171

Abstract: [Purpose/Significance] This paper proposes leveraging the rich keyword knowledge from academic papers to obtain effective features beyond patent text, thereby addressing the limitation of insufficient information in patent corpora that constrains patent term extraction performance and improving extraction accuracy. [Method/Process] Based on keywords from relevant papers, we

propose two features—domain relevance and head-tail degree—to measure the likelihood of candidate terms being genuine terms, and integrate these features into traditional patent term extraction methods. [Result/Conclusion] Experimental results demonstrate that utilizing domain relevance and head-tail degree information of candidate terms derived from paper keywords significantly improves the accuracy of patent term extraction compared to traditional methods.

Keywords: patent term extraction; paper; keyword
Classification Number: G202
DOI: 10.13266/j.issn.0252-3116.2020.14.011

Patent literature represents a crucial source of technical information, and effective patent document analysis plays a vital role in national economic, scientific, and social development. Terms in patent documents provide structured knowledge units for analysis, embodying and carrying technical information that becomes a key component of various patent analyses. Therefore, automatically extracting patent terms with minimal or no human intervention constitutes an important research topic.

The C-value method [**?**] is a commonly used statistical approach for term extraction that performs well in extracting long terms. However, C-value primarily relies on term frequency calculations, leading to two main problems: low-frequency terms cannot be identified (e.g., the string "functionalized graphene" was not correctly extracted due to its low frequency in the patent corpus), and some boundaries are incorrectly identified (e.g., the string "introducing inert gas" was incorrectly extracted because the boundary word "introducing" appeared frequently in the patent corpus) [**?**]. Consequently, there remains substantial room for improving extraction accuracy.

High-quality papers constitute the primary output of scientific research and serve as the main theoretical source and knowledge foundation for patents [**?**]. Correspondingly, patents represent the embodiment of technological innovation, inspiring scientific research by revealing problems, expanding research space, and stimulating innovative ideas. In recent years, the increasingly active interaction between scientific research and technological innovation has strengthened their relationship, making papers and patents highly correlated. Papers typically contain author-indexed keywords that describe the main topics. Keyword indexing is not arbitrary; keywords are generally mature terms or phrases in specific domains [**?**, **?**]. To compensate for the limitation of insufficient information in patent corpora that constrains term extraction effectiveness, this paper proposes for the first time utilizing rich paper keyword knowledge to obtain effective features beyond patent text, thereby improving patent term extraction performance. Our approach uses paper keyword knowledge to propose two types of features that measure the likelihood of candidate terms being genuine terms, and integrates these features into the C-value method to enhance patent term extraction accuracy.

Related Work

2.1 Term Extraction

Current term extraction methods fall into two categories: statistical methods and machine learning methods. Statistical methods evaluate the likelihood of word strings becoming terms by calculating statistical measures, offering advantages such as minimal manual intervention, strong adaptability, and portability. These methods typically use termhood and unithood to measure candidate term potential. Termhood assesses the degree to which a candidate term belongs to a specific domain, measuring its relevance to that field. Common termhood statistics include word frequency [**?**], TF-IDF [**?**], and the C-value method [**?**]. Such methods primarily rely on term frequency calculations, resulting in problems like failure to identify low-frequency terms and incorrect boundary identification. Current improvements [**?**] typically introduce mutual information and adjacent entropy [**?**] as statistical measures to reconstruct objective functions. However, research shows these improvements still offer limited enhancement.

Unithood measures the structural stability of candidate terms, i.e., the binding strength between internal components. Mutual information is a commonly used unithood indicator [**?**], measuring the dependency between word components in candidate terms by calculating their co-occurrence frequency. While it effectively reflects the binding strength between character strings, it overestimates the strength between low-frequency strings that always appear adjacent to each other. Some studies have attempted to address this issue [**?**, **?**], but results still show considerable room for improvement.

Machine learning methods construct models to extract terms by learning features from training texts. These methods can compensate for statistical methods' inability to identify low-frequency terms by using data-driven models to assess term likelihood. Common machine learning approaches include maximum entropy models [**?**] and conditional random fields [**?**, **?**]. However, machine learning methods require large-scale manually annotated corpora for training, demanding high quantity and quality of training data. Moreover, these methods remain immature and require further experimentation and validation [**?**]. Currently, no targeted, comprehensive, large-scale annotated corpus exists for patent documents. Statistical methods can extract terms with minimal manual intervention, offering an effective solution to the training data acquisition challenge in machine learning approaches. Therefore, this paper focuses on statistical methods for patent term extraction.

2.2 Correlation Between Papers and Patents

Scientific research and technological invention interact and develop spirally through knowledge transfer and feedback [**?**]. Recent studies demonstrate strong correlations between papers and patents.

Internationally, F. Narin et al. [**?**] selected biomedical journals and patents from

the USPTO database related to biotechnology, analyzing citation relationships between papers, between patents, and between patents and papers to reveal the close relationship between high technology and science, demonstrating strong paper-patent correlations. They also found that the knowledge linkage between science and technology doubles every six years [**?**]. T. Magerman et al. [**?**] used LSA text mining to examine patents and papers through inventor-author dual identities and collaborative relationships, discovering high similarity between patent documents and scientific publications. Y. Qi et al. [**?**] collected large-scale nanoscience patents and papers, using thematic keyword extraction to reveal semantic-level topics and demonstrate paper-patent correlations. H. Huang et al. [**?**] analyzed cross-citations in the fuel cell domain, showing increasing convergence in science-technology linkages.

Domestically, Wu Feifei et al. [**?**] used social network analysis on paper-patent citation relationships to discover interaction patterns between scientific and technological fields. Notably, over the past decade, technology has significantly influenced science in chemistry, communications, computer science, medical devices, and measurement, while scientific research outcomes in chemistry, physics, biology, and medicine have universally impacted patent formation. Peng Yanqi et al. [**?**] used citation and patentometric analysis on graphene patents and papers to reveal science-technology correlations. Huang Lucheng et al. [**?**] employed text mining methods with improved SAO structures to identify similarities between papers and patents in perovskite solar cells.

Patent Term Extraction Integrating Paper Keyword Knowledge

To address current term extraction problems, this paper proposes a patent term extraction method that integrates paper keyword knowledge. The method flowchart is shown in Figure 1 [Figure 1: see original paper], comprising five main steps: preprocessing (Section 3.1), candidate term selection (Section 3.2), C-value calculation (Section 3.3), keyword-based feature statistics (Section 3.4), and C-value updating (Section 3.5).

3.1 Preprocessing

First, we preprocess the collected patent and paper text corpora. Preprocessing 主要包括分词、词性标注、去除停用词等工作。Since Chinese text lacks clear word boundaries, segmentation is required to divide sentences into meaningful words. Part-of-speech tagging assigns appropriate POS tags to each word after segmentation. Stop word removal eliminates high-frequency but low-information words using general stop word lists and manual filtering, such as “的” (de), “了” (le), and “发明” (invention). Additional preprocessing includes English case conversion and special symbol removal.

3.2 Candidate Term Selection

Terms generally do not contain conjunctions, prepositions, auxiliary verbs, adverbs, or punctuation. Therefore, during candidate term selection, we use manually crafted grammatical rules to extract candidate terms from the corpus. The POS pattern matching method selects noun phrases as candidate terms based on specific POS arrangement patterns. This paper adopts POS pattern matching rules from literature [**?**] to select candidate terms. The POS pattern matching rules are shown in Table 1 , where 'a' represents adjective, 'b' distinguishing word, 'c' conjunction, 'd' adverb, 'k' suffix component, 'l' idiom, 'm' numeral, 'n' noun, 'u' auxiliary, 'v' verb, and 'vn' indicates either verb or noun at that position. The plus sign indicates multi-word terms composed of corresponding POS combinations.

3.3 C-value Calculation

The C-value method calculates termhood for each candidate term. C-value relates to term frequency in the corpus—higher frequency indicates greater termhood. It also considers candidate term length, assuming long strings are more meaningful and more likely to be terms than short strings. The C-value formula is:

$$C\text{-value}(x) = \begin{cases} \log|x| \cdot f(x) & \text{if } x \text{ is not nested} \\ \log|x| \cdot \left(f(x) - \frac{\sum_{y \in T_x} f(y)}{|T_x|}\right) & \text{if } x \text{ is nested} \end{cases}$$

where x represents a candidate term, $|x|$ denotes its length, $f(x)$ is its frequency in the patent corpus, T_x is the set of candidate terms containing x, and $|T_x|$ is the number of elements in T_x.

3.4 Keyword-Based Feature Statistics

As mentioned in the introduction, the C-value method primarily considers candidate term frequency, causing problems with low-frequency term identification and incorrect boundary recognition. Since papers and patents are strongly correlated and paper keywords are generally mature domain terms or phrases, this paper leverages paper keyword knowledge to propose two statistical features addressing C-value's limitations: domain relevance (Section 3.4.1) and head-tail degree (Section 3.4.2), thereby improving extraction accuracy.

3.4.1 Domain Relevance To address C-value's inability to identify low-frequency terms, we propose using candidate term frequency as keywords in the paper corpus to measure domain relevance. For example, although "functionalized graphene" appears infrequently in the patent corpus, its frequent occurrence as a keyword in papers indicates high domain relevance, suggesting

it may still be a valid term. This mitigates C-value's low-frequency term problem and improves accuracy. For a given candidate term x, its domain relevance $D(x)$ is:

$$D(x) = N(x)$$

where $N(x)$ represents the frequency of x appearing as a keyword in the paper corpus.

However, due to terminological flexibility, particularly patent applicants' use of vague terms to broaden protection scope and improve grant probability, exact keyword matches in paper corpora are limited. For instance, if "chemical vapor deposition" does not appear as an exact keyword, its domain relevance would be 0. Therefore, we relax exact matching to fuzzy matching, using keywords with similar surface forms to calculate domain relevance. For example, although no exact keyword matches "chemical vapor deposition," similar keywords like "chemical vapor deposition method" or "atmospheric pressure chemical vapor deposition" can be used. Thus, given candidate term x and keyword k, we update x's domain relevance $D(x)$ as:

$$D(x) = \sum sim(x,k) \times N(k)$$

where $N(k)$ is keyword k's frequency in the paper corpus, and $sim(x,k)$ is the similarity between candidate term x and keyword k, measured using the classic Dice coefficient [**?**]:

$$sim(x,k) = \frac{2 \times |x \cap k|}{|x| + |k|}$$

where $|x \cap k|$ is the number of shared words between candidate term x and segmented keyword k, $|x|$ is the word count in x, and $|k|$ is the word count in k. For example, the similarity between "chemical vapor deposition" and "chemical vapor deposition method" is $sim = 2 \times \frac{3}{3+4} = 0.86$. Table 2 shows domain relevance calculation examples for "functionalized graphene" and "chemical vapor deposition."

As shown in Table 2, candidate terms use similar keyword frequencies from the paper corpus to compute domain relevance, mitigating low-frequency term identification problems. To avoid interference from dissimilar keywords, we only consider keywords with similarity above threshold δ when calculating domain relevance.

Algorithm: Calculating Candidate Term Domain Relevance

Input: Candidate term x, paper keyword set K, paper corpus $Docs$
Output: Domain relevance $D(x)$ of candidate term x

1. $D(x) = 0$ // Initialize domain relevance of x to 0
2. **FOR** each keyword k in K **DO**
3. $sim(x,k) = 2 \times \frac{|x \cap k|}{|x|+|k|}$ // Calculate similarity using formula (4)
4. **IF** $sim(x,k) \geq \delta$ **THEN** // Check if similarity exceeds threshold δ
5. `$N(k) = COUNT(k, Docs)$ // Count keyword $k$ occurrences in paper corpus`
6. `$D(x) = D(x) + sim(x, k) \times N(k)$ // Accumulate domain relevance using formula (3)`
7. **END IF**
8. **END FOR**

3.4.2 Head-Tail Degree C-value's second major problem is incorrect boundary term identification. For example, "introducing inert gas" is incorrectly extracted because "introducing" frequently appears in the patent corpus. Using paper keyword information, we observe that "introducing" rarely appears as the first word (head) of keywords, suggesting the candidate term may have an incorrect head word and is unlikely to be a valid term. Similarly, using keyword statistics on the last word (tail) assesses tail word correctness and estimates term likelihood. Therefore, we propose head degree, tail degree, and head-tail degree statistical features using keyword information to evaluate head and tail word correctness, alleviating C-value's boundary identification problem and improving accuracy.

Specifically, given candidate term $x = \{w_1, w_2, \ldots, w_n\}$, head degree H, tail degree T, and head-tail degree HT are defined as:

$$H(x) = N(w_1, *)T(x) = N(*, w_n)HT(x) = \min(H(x), T(x))$$

where $N(w_1, *)$ is the frequency of keywords with w_1 as the head word, $N(*, w_n)$ is the frequency of keywords with w_n as the tail word, and $\min(H(x), T(x))$ selects the smaller value, indicating that if either head or tail word is likely incorrect, the candidate term is probably not a valid term. Table 3 shows head-tail degree calculation examples for "fluorescent nanoparticles" and "introducing inert gas."

As shown in Table 3, "fluorescent nanoparticles" has a high head-tail degree because both "fluorescent" (head) and "particles" (tail) frequently appear in keyword positions, indicating high term likelihood. Conversely, "introducing inert gas" has head-tail degree 0 because "introducing" never appears as a keyword head, suggesting low term likelihood.

Algorithm: Calculating Candidate Term Head-Tail Degree

Input: Candidate term x, paper keyword set K, paper corpus $Docs$
Output: Head-tail degree $HT(x)$ of candidate term x

1. $w_1, w_n = CUT(x)$ // Segment x, set first word as w_1, last word as w_n
2. $H(x) = 0$ // Initialize head degree
3. $T(x) = 0$ // Initialize tail degree
4. $HT(x) = 0$ // Initialize head-tail degree
5. **FOR** each keyword k in K **DO**
6. $k_1, k_m = CUT(k)$ // Segment k, set first word as k_1, last word as k_m
7. **IF** $w_1 == k_1$ **THEN** // If head words match, accumulate head degree
8. `$N(k) = COUNT(k, Docs)$ // Count keyword occurrences`
9. `$H(x) = H(x) + N(k)$`
10. **END IF**
11. **IF** $w_n == k_m$ **THEN** // If tail words match, accumulate tail degree
12. `$N(k) = COUNT(k, Docs)$ // Count keyword occurrences`
13. `$T(x) = T(x) + N(k)$`
14. **END IF**
15. **END FOR**
16. $HT(x) = \min(H(x), T(x))$ // Select minimum as head-tail degree using formula (7)

3.5 C-value Updating

We integrate keyword-based statistical features into C-value to improve patent term extraction accuracy. Specifically:

Combining domain relevance D forms *D-C-value*:

$$D\text{-}C\text{-}value(x) = (1 + D(x)) \times C\text{-}value(x)$$

This definition indicates that higher C-value and higher domain relevance increase term likelihood, mitigating C-value's low-frequency term problem. When $D(x) = 0$, *D-C-value* degrades to C-value.

Combining head-tail degree HT forms *HT-C-value*:

$$HT\text{-}C\text{-}value(x) = (1 + HT(x)) \times C\text{-}value(x)$$

This indicates that higher C-value and higher head-tail degree increase term likelihood, mitigating boundary identification problems. When $HT(x) = 0$, *HT-C-value* degrades to C-value.

Simultaneously considering both features yields *D-HT-C-value*:

$$D\text{-}HT\text{-}C\text{-}value(x) = (1 + D(x)) \times (1 + HT(x)) \times C\text{-}value(x)$$

Experiments

4.1 Dataset

To validate the proposed model's feasibility and effectiveness, we conducted experiments on graphene patent literature. Graphene, the thinnest known material, possesses unique structure and exceptional optical, chemical, electrical, and mechanical properties, making it a promising new material with considerable economic benefits and broad industrial application prospects. In recent years, both graphene research papers and patent applications have grown exponentially.

Patent data was collected from the China National Intellectual Property Administration database, searching for "graphene" in Chinese invention patents published between 2014-2018 (retrieved November 15, 2018), yielding 6,445 valid patents. Patent titles and abstracts formed the patent corpus. Paper data was collected from Wanfang Database, searching for "graphene" in 北大核心期刊 (Peking University core journals) from 2014-2018 (retrieved November 15, 2018), yielding 5,236 papers. Paper keywords formed the paper corpus.

4.2 Evaluation Metrics

Given the large patent text dataset, we adopt Precision@N ($P@N$) as the evaluation metric [**?**]:

$$P@N = \frac{r}{N} \times 100\%$$

where N is a constant representing the number of extracted patent terms (ranging from 200 to 2000 in our experiments), and r is the number of correct terms among the N extracted terms. To avoid subjectivity and domain knowledge limitations, we verify extracted terms against Baidu Baike, Wikipedia, and Hudong Baike to determine correctness.

4.3 Results

4.3.1 Impact of Domain Relevance on Patent Term Extraction Accuracy We first investigate domain relevance's impact on extraction accuracy. Similarity threshold δ was set to 0.2, 0.4, 0.6, 0.8, and 1.0, with methods denoted as *D-C-value*-0.2, *D-C-value*-0.4, *D-C-value*-0.6, *D-C-value*-0.8, and *D-C-value*-1.0, compared against traditional C-value. Results are shown in Figure 2 [Figure 2: see original paper].

Figure 2 shows that *D-C-value*-0.2, *D-C-value*-0.4, and *D-C-value*-0.6 achieve lower accuracy than C-value. At $N = 1000$, their accuracies are 20.09%, 14.90%, and 3.61% lower than C-value, respectively. Conversely, *D-C-value*-0.8 and *D-C-value*-1.0 significantly outperform C-value, with *D-C-value*-0.8 achieving the highest accuracy—18.69% and 17.79% higher than C-value at $N = 1000$. This indicates that using keywords with low similarity introduces noise and reduces accuracy, while high-similarity keywords improve accuracy, validating the effectiveness of domain relevance features. Subsequent experiments use $\delta = 0.8$.

4.3.2 Impact of Head-Tail Degree on Patent Term Extraction Accuracy We evaluate head-tail degree's impact by comparing C-value with *HT-C-value*. Results are shown in Figure 3 [Figure 3: see original paper]. At $N = 1000$, *HT-C-value* achieves 14.15% higher accuracy than C-value, demonstrating that head-tail degree features from keyword statistics alleviate C-value's boundary identification problems and improve accuracy.

4.3.3 Combined Features' Impact on Patent Term Extraction Based on previous experiments, we investigate combining domain relevance and head-tail degree. *D-HT-C-value* is compared against single-feature methods (*D-C-value* and *HT-C-value*) and baseline C-value. Results are shown in Figure 4 [Figure 4: see original paper].

At $N = 1000$, *D-HT-C-value* achieves the highest accuracy, outperforming C-value, *D-C-value*, and *HT-C-value* by 27.49%, 8.80%, and 13.34%, respectively. This shows that combining both features yields better accuracy than single features, with domain relevance having greater impact.

4.3.4 Comparison with Other Methods Finally, we compare *D-HT-C-value* with typical C-value improvements:

1. **C-value**: Uses C-value to measure termhood
2. **PMI-C-value**: Integrates mutual information into C-value. Mutual information is a common unithood indicator measuring component binding strength through co-occurrence frequency.
3. **En-C-value**: Integrates adjacent entropy into C-value. Adjacent entropy uses uncertainty of neighboring words to eliminate boundary errors; higher entropy indicates more informative neighbors and higher term probability.
4. **D-HT-C-value**: Our proposed method integrating domain relevance and head-tail degree from paper keywords.

Results are shown in Figure 5 [Figure 5: see original paper]. At $N = 1000$, PMI-C-value, En-C-value, and D-HT-C-value improve accuracy over C-value by 6.58%, 3.89%, and 26.68%, respectively. D-HT-C-value achieves the highest accuracy. While mutual information helps low-frequency terms, it also reduces scores for some high-frequency valid terms, limiting improvement. Some non-term high-frequency strings in patents have many neighbors, limiting adjacent

entropy's effectiveness. In contrast, domain relevance and head-tail degree from paper corpora effectively improve C-value accuracy, validating our approach.

Conclusion

Current patent term extraction methods suffer from failure to identify low-frequency terms and incorrect boundary recognition, leaving substantial room for improvement. Previous studies primarily used features from patent text itself. Papers and patents are strongly correlated, with paper keywords being mature domain terms or phrases that contain rich domain knowledge. To address low external resource utilization and insufficient information in patent corpora, this paper proposes for the first time leveraging rich paper keyword knowledge to obtain effective external features for improving patent term extraction. We propose domain relevance and head-tail degree features based on paper keyword knowledge and integrate them into the C-value method. Experiments show that our approach significantly improves accuracy over traditional methods.

Future research will explore knowledge from Baidu Baike, Wikipedia, and Hudong Baike to further improve patent term extraction accuracy.

References

[1] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International journal on digital libraries, 2000, 3(2): 115-130.

[2] Zhou Shuangshuang, Xu Jin'an, Chen Yufeng, et al. A microblog new word discovery method integrating rules and statistics[J]. Computer Applications, 2017, 37(4): 1044-1050.

[3] HIROYUKI T, TAKAKAYUKI T. A bibliometric analysis of scientific literature cited by influential patents[J]. Journal of informetrics, 2003, 58(2): 369-390.

[4] Chen Hongmei. Keyword selection in scientific papers[J]. Journal of Xi'an Shiyou University (Natural Science Edition), 2011, 26(4): 109-110.

[5] Li Na, Rong Wenhui, Bian Zhiying. How to determine keywords[J]. Clinical Focus, 2003, 18(12): 674-674.

[6] Qin Jiahui, Xu Shuo, Zhang Yunliang, et al. Research and analysis of automatic term extraction technology for scientific literature[J]. New Technology of Library and Information Service, 2014(1): 51-55.

[7] Zeng Wen, Xu Shuo, Zhang Yunliang, et al. Research and analysis of automatic term extraction technology for scientific literature[J]. New Technology of Library and Information Service, 2014(1): 51-55.

[8] SPASIC I, GREENWOOD M, PREEC A, et al. FlexiTerm: a flexible term recognition method[J]. Journal of biomedical semantics, 2013, 27(4): 1-15.

[9] Han Hongqi, Zhu Donghua, Wang Xuefeng. Patent technology term extraction method[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(12): 1280-1285.

[10] Hu Apei, Zhang Jing, Liu Junli. Chinese term extraction based on improved C-value method[J]. New Technology of Library and Information Service, 2013, 230(2): 24-29.

[11] Zhang Leihan, Lü Xueqiang, Li Zhuo, et al. Research on domain ontology term extraction methods[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(2): 167-174.

[12] Zhou Shuangshuang, Xu Jin'an, Chen Yufeng, et al. A microblog new word discovery method integrating rules and statistics[J]. Computer Applications, 2017, 37(4): 1044-1050.

[13] Yu Yan, Zhao Naixuan. Patent term extraction based on general words and term components[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(7): 742-752.

[14] Ding Jie, Lü Xueqiang, Liu Kehui. Patent literature term extraction method based on boundary marker set[J]. Computer Engineering and Science, 2015, 37(8): 1591-1598.

[15] Liu Jian, Tang Huifeng, Liu Wuying. A Chinese term extraction method based on statistical techniques[J]. China Terminology, 2014, 16(5): 10-14.

[16] Du Liping, Li Xiaoge, Yu Gen, et al. New word discovery based on improved mutual information algorithm for Chinese word segmentation system improvement[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 35-40.

[17] ZHANG W, YOSHIDA T, TANG X, et al. Improving effectiveness of mutual information for substantival multiword expression extraction[J]. Expert systems with applications an international journal, 2009, 36(8): 10919-10930.

[18] Muheyaeti · Niyazibieke, Gulishawuli · Talifu. Research and implementation of Kazakh IT domain term recognition[J]. Journal of Chinese Information Processing, 2016(3): 68-75.

[19] Wang Hao, Wang Miping, Su Xinning. Research on Chinese patent term extraction for ontology learning[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(6): 573-585.

[20] ZENG D, SUN C, LIN L, et al. LSTM-CRF for drug name entity recognition[J]. Entropy, 2017, 19(6): 283-295.

[21] CONRADO M, PARDO T, REZENDE S. A machine learning approach to automatic term extraction using a rich feature set[C]//The 2013 conference of the north American chapter of the association for computational Linguistics: human language technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013: 16-23.

[22] BHATTACHARYA S, KRETSCHMER H, MEYER M. Characterizing intellectual spaces between science and technology[J]. Scientometrics, 2003, 58(2): 369-390.

[23] NARIN F, NOMA E. Is technology becoming science?[J]. Scientometrics, 1985, 7(3): 369-381.

[24] NARIN F, HAMILTON K S, OLIVASTRO D. The increasing linkage between U.S. technology and public science[J]. Research policy, 1997, 26(3): 317-330.

[25] MAGERMAN T, LOOY B V, SONG X. Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications[J]. Scientometrics, 2010, 82(2): 289-306.

[26] QI Y, ZHU N, ZHAI Y, et al. The mutually beneficial relationship of patents and scientific literature: topic evolution in nanoscience[J]. Scientometrics, 2018, 115(1): 893-911.

[27] HUANG M H, YANG H W, CHEN D Z. Increasing science and technology linkage in fuel cells: a cross citation analysis of papers and patents[J]. Journal of informetrics, 2015, 9(2): 237-249.

[28] Wu Feifei, Huang Lucheng, Shi Yuanmiao. Science-technology relationship analysis based on literature and patent mutual citation[J]. Science of Science and Management of S.&T., 2013, 34(10): 13-20.

[29] Peng Yanqi, Qin Jiahui, Ye Ying. Cross-citation analysis of patents and papers in graphene research[J]. Information Studies: Theory & Application, 2018, 41(7): 18-21.

[30] Huang Lucheng, Wang Jingjing, Li Xin, et al. Technology opportunity analysis of perovskite solar cells based on papers and patents[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(7): 686-695.

[31] Chen Erjing, Jiang Enbo. Survey of text similarity calculation methods[J]. Data Analysis and Knowledge Discovery, 2017, 6(6): 1-11.

Author Contributions

Yu Yan: Conceived research idea, designed methodology, conducted experiments, wrote paper
Chen Lei: Data cleaning
Jiang Jinde: Analyzed data, revised paper
Zhao Naixuan: Revised paper

Academic Integrity Statement for *Library and Information Service* Authors

Library and Information Service has consistently upheld the mission of publishing excellent academic research and promoting scholarly exchange, while 致力于净化学术出版环境，创建良好学术生态 (committed to purifying the academic publishing environment and creating a healthy academic ecosystem). In 2013, we led the formulation, release, and implementation of the *Joint Statement on Upholding Academic Ethics and Purifying the Academic Environment by Library Science Journals* (the "Statement") (see: http://www.lis.ac.cn/CN/column/item202.shtml). Subsequently, we led the formulation and release of the *Joint Action Plan for Chinese Library and Information Science Journals to Resist Academic Misconduct* (the "Joint Action Plan") (see: http://www.lis.ac.cn/CN/column/item247.shtml). To implement this philosophy, we hereby declare that, effective immediately, all submitting authors must 承诺 (pledge): Papers submitted to this journal must comply with the above "Statement" and "Joint Action Plan," consciously uphold academic ethics, and resolutely resist academic misconduct. *Library and Information Service* maintains a zero-tolerance policy toward all forms of academic misconduct, including plagiarism and appropriation, and will implement corresponding disciplinary measures.

Library and Information Service Editorial Office

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