Postprint: An Evaluation Model for Patent Keyword Extraction Algorithms Based on Information Gain and Similarity

Authors: Yu Yan, Ju Peng, Shang Mingjie

Date: 2023-04-01T15:51:24+00:00

Abstract

[Purpose/Significance] To address the limitations of current patent keyword extraction algorithm evaluations that primarily rely on matching extracted keywords against manually annotated keywords by experts, this paper proposes an evaluation model based on information gain and similarity. [Method/Process] The proposed model assesses algorithmic accuracy from both internal and external perspectives. The internal evaluation component measures the information gain of each extracted keyword to evaluate its novelty and inventiveness, while the external evaluation component represents patents using the extracted keyword set, computes similarity between related patents, and measures the effectiveness of keywords in describing patent topics. [Result/Conclusion] Validation experiments and empirical application studies demonstrate that the proposed information gain and similarity-based evaluation model is both feasible and effective.

Full Text

Preamble

Volume 66, Issue 6, March 2022

An Evaluation Model for Patent Keyword Extraction Algorithms Based on Information Gain and Similarity

Yu Yan[1],[2], Ju Peng[1], Shang Mingjie[1]

[1]Institute of Information Management and Technology, Nanjing Tech University, Nanjing 210009
[2]Department of Computer Engineering, Chengxian College, Southeast University, Nanjing 211816

Abstract: [Purpose/Significance] To address the limitations of current patent keyword extraction algorithm evaluation methods that primarily rely on matching extracted keywords with manually annotated keywords by experts, this paper proposes an evaluation model for patent keyword extraction algorithms based on information gain and similarity. [Method/Process] The proposed evaluation model assesses the accuracy of patent keyword extraction algorithms from both intrinsic and extrinsic perspectives. The intrinsic evaluation model measures the information gain of each extracted keyword to assess its novelty and creativity, while the extrinsic evaluation model uses the extracted keyword set to represent patents and calculates the similarity between related patents to evaluate the effectiveness of the extracted keywords in describing patent topics. [Result/Conclusion] Through validation experiments and empirical application studies, the results demonstrate the feasibility and effectiveness of the proposed evaluation model based on information gain and similarity.

Keywords: patent; keyword extraction; evaluation; information gain; similarity
Classification Number: G202
DOI: 10.13266/j.issn.0252-3116.2022.06.012

1 Introduction

Patent keywords are terms or phrases that indicate the thematic content of patent documents and are widely used in patent analysis tasks such as patent classification [1], emerging technology monitoring [2], patent retrieval [3], and patent clustering [4]. However, patents typically do not contain keywords and require manual indexing. Due to the lengthy and specialized nature of patent documents and their rapidly increasing volume in recent years, manual indexing has become inadequate for patent analysis needs. Therefore, automatic, efficient, and accurate extraction of patent keywords using computers represents an important research topic.

Current research on patent keyword extraction primarily focuses on algorithmic improvements, with evaluation of these enhanced algorithms typically conducted by matching extracted keywords against manually annotated keywords by experts. However, relying on expert annotation is time-consuming, labor-intensive, and limited in scale, while also suffering from domain limitations, language dependency, and subjectivity issues. These factors hinder effective evaluation of patent keyword extraction algorithms and impede further research advancement.

To address these shortcomings, this paper proposes an evaluation model based on information gain and similarity. The model evaluates algorithm accuracy from both intrinsic and extrinsic levels. The intrinsic evaluation measures the information gain of each extracted keyword to assess novelty and creativity, while the extrinsic evaluation uses extracted keyword sets to represent patents and calculates similarity between related patents to evaluate how well the key-

words capture patent themes. Furthermore, the proposed evaluation model is applicable not only to patent keyword extraction algorithm evaluation but also to keyword extraction evaluation for academic literature and related documents.

Overall, the contributions of this paper are as follows: (1) Proposing an evaluation model for patent keyword extraction algorithms based on information gain and similarity to address current evaluation method limitations (Section 3). (2) Conducting validation experiments to demonstrate the effectiveness of the proposed model (Section 4). (3) Applying the proposed model in empirical studies to evaluate three patent keyword extraction strategies, with results confirming the model's effectiveness and feasibility (Section 5).

2 Related Research

Keyword extraction algorithm evaluation primarily examines how accurately extracted keyword sets reflect document thematic content. Evaluation methods can be divided into two categories: intrinsic evaluation and extrinsic evaluation [5].

Intrinsic evaluation methods match extracted keywords against correct keywords, determining correctness and using metrics to evaluate algorithm accuracy. Matching typically employs exact matching, comparing algorithm-extracted keywords with author- or expert-annotated keywords (called gold-standard keywords). However, exact matching is overly strict, ignoring semantic relationships such as synonyms or partial matches, leading to unreliable evaluation results. To address this, some studies supplement exact matching with fuzzy matching methods, such as using edit distance to calculate morphological similarity [6], probability models to compute semantic similarity [7], or comprehensive similarity measures incorporating both morphological and semantic information [8].

The most widely used metrics in intrinsic evaluation are precision (P), recall (R), and F1-score [9-11]. Precision measures the ratio of matched keywords to extracted keywords; recall measures the ratio of matched keywords to gold-standard keywords; and F1-score is the weighted average of precision and recall. However, these metrics do not consider the order of extracted keywords. In practice, if matched keywords have higher rankings, the extraction algorithm is more accurate. Therefore, some studies have improved these metrics by scoring based on keyword ranking order, such as Precision@K [12] which considers the top K extracted keywords (K typically being 1, 3, 5, 10), Mean Reciprocal Rank (MRR) [13-15] which measures the ranking of the first matched keyword, and Binary Preference Measure (Bpref) [16-17] which calculates the ranking of incorrectly extracted words.

Extrinsic evaluation methods apply extracted keywords to specific applications and indirectly evaluate their effectiveness by measuring application performance, such as text classification [18], clustering, or retrieval [19]. Since extrinsic evaluation is task-specific, the quality and scale of corpora used

in these tasks, as well as the algorithms employed, significantly influence evaluation results. Moreover, the computational cost of specific tasks often exceeds that of keyword extraction itself, making evaluation speed impractical. Consequently, extrinsic evaluation methods are rarely used for keyword extraction algorithm evaluation.

Currently, intrinsic evaluation is the most widely used method. However, it has significant limitations. Some documents, particularly patents, lack keywords and require manual annotation for evaluation. Manual annotation is labor-intensive, subjective, and arbitrary. Variations in corpus type, annotation granularity, and annotator expertise lead to substantial differences across datasets, requiring multiple domain experts and consistency measurement using Kappa statistics. While some open-source annotated datasets exist [20-21], their reliability is often poor [5], typically domain-specific English datasets without dedicated patent keyword annotation datasets.

Therefore, researchers have explored keyword extraction evaluation without gold-standard keywords. For example, Zhang Chengzhi [8] used term frequency and position information to extract keywords as gold standards, then compared them with algorithm-extracted keyword sets to evaluate accuracy. However, this approach assumes that keywords extracted using frequency and position information can serve as gold standards, which is questionable since such keywords may not accurately reflect document content. Conversely, if they could accurately represent document themes, keyword extraction algorithms would be unnecessary.

Based on this analysis, this paper proposes a new evaluation model to address current method limitations and advance patent keyword extraction research.

3 Evaluation Model

3.1 Model Principles

The evaluation model assesses algorithm accuracy from intrinsic and extrinsic perspectives, proposing an intrinsic evaluation model based on information gain and an extrinsic evaluation model based on similarity. Intrinsic evaluation measures the effectiveness of each extracted keyword using information gain, while extrinsic evaluation uses extracted keyword sets to represent patent documents and calculates inter-patent similarity to assess how accurately keyword sets capture document themes.

3.1.1 Intrinsic Evaluation Model Principle Based on Information Gain

Patent keywords should indicate patent themes and embody novelty and creativity. Therefore, patent keywords should carry as much information as possible to distinguish existing patents. Information gain (IG) represents the reduction in uncertainty of a random event under certain conditions, thereby indicating the information carried by that condition. When an extracted keyword reduces

more uncertainty in an existing system, it contains greater information. An algorithm's information gain can be represented by the mean information gain of its extracted keywords to measure its ability to capture novelty and creativity.

[Figure 1: see original paper] illustrates the principle of the intrinsic evaluation model based on information gain. In the figure, algorithms 1 and 2 each extract 2 keywords from a target patent. By constructing a relevant patent dataset, the information gain of keywords "lithium-ion" and "ternary" extracted by algorithm 1 are calculated as 0.41 and 1.85, respectively, yielding an algorithm information gain of 1.13. For algorithm 2, the information gain values for "invention" and "battery" are 0.01 and 0.11, respectively, resulting in an algorithm information gain of 0.06. This indicates that algorithm 1 extracts keywords carrying more information and better reflecting the target patent's novelty and creativity.

3.1.2 Extrinsic Evaluation Model Principle Based on Similarity Patent keywords reflect patent themes, so the extrinsic evaluation model uses extracted keywords to represent patents and calculates inter-patent similarity to detect how accurately they capture patent themes. Innovation in patents is not isolated; it reflects technological continuity to some extent. Patent examiners compare patents and cite similar patents as references, making citations indicative of similarity between target and cited patents [22-24]. Research shows that target patents have higher similarity with cited patents than with random patents [22-24]. Therefore, this paper uses extracted keyword sets to calculate similarity between target-cited patent pairs and target-random patent pairs to evaluate algorithm accuracy.

[Figure 2: see original paper] illustrates the principle of the extrinsic evaluation model based on similarity. Algorithms 1 and 2 each extract 3 keywords from target, cited, and random patents. For algorithm 1, the target and cited patents share keywords "lithium-ion" and "positive electrode," yielding a similarity of 2, while target and random patents share only "lithium-ion," yielding a similarity of 1. For algorithm 2, target-cited and target-random pairs both share "invention" and "battery," each yielding a similarity of 2. Algorithm 1 produces higher similarity for target-cited pairs than target-random pairs, which aligns with intuition and more accurately captures patent themes, making algorithm 1 superior.

3.2 Detailed Model Description

3.2.1 Intrinsic Evaluation Method Based on Information Gain Information gain represents the reduction in uncertainty of random event X under condition y, indicating the information carried by y:

$$IG(y) = H(X) - H(X|y)$$

where H represents information entropy: $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$, with $p(x_i)$ being the probability of event X being x_i. $H(X)$ measures system uncertainty—greater uncertainty yields higher entropy. $H(X|y) = -\sum_{i=1}^{n} p(x_i|y) \log p(x_i|y)$ represents uncertainty under condition y. Information gain measures the change in entropy before and after condition y appears, indicating how much information is introduced to eliminate uncertainty.

Accordingly, this paper constructs a patent dataset $S = \{C_1, C_2, \dots, C_n\}$, where $C_i (i = 1, 2, \dots, n)$ represents a patent category and $C_i = \{d_{i,1}, d_{i,2}, \dots, d_{i,m}\}$, with $d_{i,j} (j = 1, 2, \dots, m)$ being a patent in category C_i. Patent keywords should carry sufficient information to distinguish between different patent categories and different patents within the same category, eliminating as much uncertainty as possible. This paper defines information gain that eliminates inter-category uncertainty as Class Information Gain (IGC) and information gain that eliminates intra-category uncertainty as Document Information Gain (IGD).

Specifically, for a keyword w extracted from target patent $d_{i,j}$, IGC measures the information w carries to eliminate inter-category uncertainty:

$$IGC(w) = \left(-\sum_{i=1}^{n} p(C_i) \log p(C_i) \right) - \left(-\sum_{i=1}^{n} p(C_i|w) \log p(C_i|w) \right)$$

where $p(C_i)$ is the probability of category C_i and $p(C_i|w)$ is the probability of category C_i given keyword w. Correct patent keywords should carry substantial information to eliminate inter-category uncertainty.

IGD measures w's ability to eliminate uncertainty among different patent documents within the same category C_i:

$$IGD(w) = \left(-\sum_{j=1}^{m} p(d_{i,j}) \log p(d_{i,j}) \right) - \left(-\sum_{j=1}^{m} p(d_{i,j}|w) \log p(d_{i,j}|w) \right)$$

where $p(d_{i,j})$ is the probability of patent $d_{i,j}$ and $p(d_{i,j}|w)$ is the probability of patent $d_{i,j}$ given keyword w. Keywords should carry substantial information to eliminate intra-category uncertainty.

Ultimately, patent keywords should eliminate both inter-category and intra-category uncertainty. Therefore, the information gain of a term $IG(w)$ is calculated as the product of its class and document information gains:

$$IG(w) = (IGC(w) + \alpha) \times (IGD(w) + \alpha)$$

where α is a small real number to prevent IG from being zero.

A term with higher information gain carries more information and is more likely to be a patent keyword. [Figure 4: see original paper] illustrates the intrinsic evaluation method with a target patent $d_{1,1}$ concerning lithium-ion battery cathode materials. shows the information gain values for terms "invention," "battery," and "ternary." The term "invention" appears uniformly across all categories and within the target category, yielding small IGC and IGD values. The term "battery" appears frequently across patents in the target category but rarely in others, resulting in low IGD but high IGC. The term "ternary" appears only in a few patents within the target category, yielding high values for both IGC and IGD and consequently the highest final information gain, demonstrating strong ability to distinguish between and within categories. This shows that the proposed information gain effectively captures the incremental information a term brings to existing patent literature.

3.2.2 Extrinsic Evaluation Method Based on Similarity The dataset construction for the extrinsic evaluation method is shown in [Figure 5: see original paper]. Given a target patent, similar patent pairs are constructed using the target patent and its cited patents, while random patent pairs use the target patent and random patents. Similarity between similar and random patent pairs is calculated using extracted keyword sets. If extracted keywords accurately represent patent themes, similar patent pairs should have higher similarity than random patent pairs.

Specifically, given target patent d, its cited patent d_c, and random patent d_r, patents are represented using extracted keywords, and similarity between two patents d_i and d_j is calculated as:

$$sim(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i|}$$

where $|d_i|$ is the number of keywords in patent d_i and $|d_i \cap d_j|$ is the number of shared keywords. More shared keywords yield higher similarity, indicating more similar patents; fewer shared keywords yield lower similarity, indicating less similar patents.

Finally, the similarity difference (SD) between similar and random patent pairs is calculated:

$$SD = sim(d, d_c) - sim(d, d_r)$$

A higher SD indicates more shared keywords in similar pairs and fewer in random pairs, suggesting more reasonable keyword extraction.

4 Evaluation Model Validation Experiments

This section validates the proposed evaluation model using lithium-ion battery patent data.

4.1 Data Source

Lithium-ion batteries, developed and commercialized by Sony in 1990, have become a research hotspot in electrochemistry due to their high voltage, high energy density, long cycle life, low self-discharge, no memory effect, and environmental friendliness. With technological advances, they are now widely used in mobile phones, portable computers, cameras, power tools, electric vehicles, and energy storage stations. Key materials include cathode materials, anode materials, electrolytes, and separators in the midstream industrial chain.

Based on the China National Intellectual Property Administration database, this study retrieved Chinese invention patents for lithium-ion batteries, extracting 2,000 patent titles and abstracts each for four categories: C1 (cathode materials), C2 (anode materials), C3 (electrolytes), and C4 (separators). Dataset details are shown in .

4.2 Data Processing

First, the collected patent corpus was preprocessed. Since Chinese text lacks word delimiters, segmentation was performed. High-frequency, low-information words like "的" (de) and "是" (shi) were removed using a stopword list. Additional preprocessing included case conversion and special symbol removal.

For intrinsic evaluation validation, candidate keywords were generated for each target patent, and the average information gain of manually annotated keywords was calculated alongside non-keywords, with $\alpha = 0.01$ in equation (4). For extrinsic evaluation, target patents and their cited patents formed similar pairs, while target patents and different-category patents formed random pairs. Expert-annotated keywords were used to calculate similarity differences.

4.3 Results Analysis

4.3.1 Intrinsic Evaluation Method The experiment selected 50 patents with citations from each category C1-C4 as target patents, using cited patents as similar patents (). Three domain experts annotated 8 keywords per patent, with pairwise intersections forming the final annotation set. Kappa scores exceeded 0.8, confirming annotation validity.

[Figure 6: see original paper] shows the intrinsic evaluation results. Across categories C1-C4, the mean information gain for keywords was 10.51, 9.43, 9.38, and 12.61, respectively, while non-keywords averaged 4.04, 4.26, 2.23, and 5.11. Keywords exhibited significantly higher information gain than non-keywords,

demonstrating that patent keywords carry more information to distinguish between and within categories. These results validate using information gain for keyword evaluation.

provides an example from category C1 (cathode materials) for the invention "Carbon-coated Ternary Cathode Material Preparation Method and Carbon-coated Ternary Cathode Material" (Application No. CN201310433513.7). Keywords like "ternary cathode material," "organic carbon source," and "precursor" show high IGC and IGD values, indicating they carry substantial information to distinguish between categories and within categories. Conversely, terms like "compound" and "mixture" appear frequently across categories and patents, yielding low information gain and failing to reflect novelty and creativity. This demonstrates the effectiveness of the intrinsic evaluation method.

4.3.2 Extrinsic Evaluation Method Extrinsic evaluation results are shown in [Figure 7: see original paper]. Across categories C1-C4, similarity differences were 1.92, 1.73, 1.79, and 2.34, respectively, indicating that correct keywords yield higher similarity for cited patents than random patents. This confirms that proper keywords reflect patent themes, making cited patents more similar than random patents when represented by extracted keywords.

provides an example using the same target patent. Its cited patent "A Preparation Method for Lithium-ion Battery Ternary Cathode Material" (Application No. CN201110314584.6) shares keywords "ternary cathode material" and "lithium-ion battery" with the target patent. The random patent from category C2 (anode materials) shares only "lithium-ion battery." This demonstrates that the extrinsic evaluation method effectively assesses keyword validity.

5 Empirical Application of the Evaluation Model

This section applies the proposed model to compare three patent keyword extraction strategies.

Patent documents include four sections: title, abstract, claims, and specification. Title and abstract provide brief descriptions for retrieval purposes without legal effect. Claims are legal documents describing technical features, containing all essential technical means reflecting novelty and creativity, and defining patent protection scope. Specifications provide detailed support for claims, typically including technical field, background, content, drawings, and embodiments. Current research typically focuses on titles and abstracts for keyword extraction. This section explores extracting candidate keywords from different patent sections using classic TF-IDF and evaluates strategies using the proposed model. The three strategies are shown in .

Experiments used data from Section 4.1, downloading titles, abstracts, claims, and specifications of target and cited patents. After preprocessing and part-of-speech matching, candidate keywords were generated from different sections

based on each strategy. TF-IDF ranked candidates, selecting the top eight as keywords. Results were evaluated using intrinsic and extrinsic methods.

Intrinsic evaluation results ([Figure 8: see original paper]) show that across categories C1-C4, the "claim" strategy achieved the highest information gain values (9.42, 8.55, 8.61, and 10.07), outperforming "abstract" and "all" strategies. details the target patent's keywords and average information gain across strategies, confirming that the claim strategy yields the most accurate keywords.

Extrinsic evaluation results ([Figure 9: see original paper]) show that across categories, the claim strategy achieved the highest similarity differences (1.51, 1.37, 1.25, and 1.49), indicating it best captures patent themes. provides an example showing the claim strategy produces the largest similarity difference between cited and random patents.

In summary, patent documents differ from other literature. Relying solely on titles and abstracts may miss important keywords because abstracts are often too brief for effective TF-IDF application. Specifications, while detailed, contain noise that reduces extraction accuracy. Claims, as both technical and legal documents embodying novelty and creativity, are the core of patents. Empirical results demonstrate that extracting keywords from claims yields better accuracy than from abstracts or specifications.

6 Conclusion

Patent keywords are essential for patent classification, emerging technology monitoring, retrieval, and clustering. Automatic extraction is crucial, but current evaluation methods relying on expert annotation are problematic due to cost, subjectivity, and limitations. This paper proposes intrinsic and extrinsic evaluation models based on information gain and similarity to address these issues.

The intrinsic model evaluates each extracted keyword using class and document information gain to measure novelty and creativity. The extrinsic model uses extracted keyword sets to represent patents and compares similarity between cited and random patent pairs to evaluate thematic accuracy. Validation experiments confirm the model's effectiveness. Empirical application comparing three extraction strategies further demonstrates feasibility and effectiveness. While designed for patent keyword extraction, the model is also applicable to academic literature.

Future improvements include: (1) Better handling of numerical innovations (e.g., "20% capacity improvement"); (2) Investigating the impact of patent set construction on model stability; (3) Validating applicability to academic literature keyword extraction.

References

[1] Hu J, Li S, Yao Y, et al. Patent keyword extraction algorithm based on

distributed representation for patent classification[J]. Entropy, 2018, 20(2): 104-124.

[2] Joung J, Kim K. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data[J]. Technological Forecasting and Social Change, 2017, 114: 281-292.

[3] Zhou S. Application of keywords in patent literature retrieval[J]. Information Theory and Practice, 2018(7): 73-79.

[4] Wang K, Wang J, Tang Y, et al. Technology opportunity identification based on patents and scientific papers[J]. Science and Technology Management Research, 2020, 33(5): 67-70.

[5] Firoozeh N, Nazarenko A, Alizon F, et al. Keyword extraction: Issues and methods[J]. Natural Language Engineering, 2020, 26(3): 259-291.

[6] Ristad E S, Yianilos P N. Learning string-edit distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522-532.

[7] Dagan I, Pereira F. Similarity-based estimation of word cooccurrence probabilities[EB/OL]. [2021-10-28]. https://arxiv.org/pdf/cmp-lg/9405001.pdf.

[8] Zhang C, Zhou D. Research on general evaluation model for automatic indexing[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(1): 40-47.

[9] Yu Y, Shang M, Zhao N. Patent keyword extraction method driven by claim features[J]. Journal of the China Society for Scientific and Technical Information, 2021, 40(6): 610-620.

[10] Ma H, Liu F, Xia Q, et al. Literature keyword extraction algorithm based on weighted hypergraph random walk[J]. Acta Electronica Sinica, 2018, 46(6): 1410-1414.

[11] Wang Z, Guo Y. Research on automatic Chinese patent keyword extraction based on word and sentence importance[J]. Information Studies: Theory & Application, 2018, 41(9): 123-129.

[12] Singhal A, Kasturi R, Srivastava J, et al. Leveraging web resources for keyword assignment to short text documents[EB/OL]. [2021-10-28]. https://arxiv.org/ftp/arxiv/papers/1706/1706.05985.pdf.

[13] Voorhees E M. The TREC-8 question answering track report[EB/OL]. [2021-10-28]. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.6392&rep=rep1&type=pdf.

[14] Florescu C, Caragea C. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents[C]//Proceedings of the 55th annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 1105-1115.

[15] Zhang Y, Chang Y, Liu X, et al. Mike: keyphrase extraction by integrating multidimensional information[C]//Proceedings of the 2017 ACM on conference

on information and knowledge management. New York: ACM, 2017: 1349-1358.

[16] Buckley C, Voorhees E M. Retrieval evaluation with incomplete information[C]//Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2004: 25-32.

[17] Liu Z, Huang W, Zheng Y, et al. Automatic keyphrase extraction via topic decomposition[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. Stroudsburg: ACL, 2010: 366-376.

[18] Zhang K, Xu H, Tang J, et al. Keyword extraction using support vector machine[C]//Proceedings of the 6th international conference on advances in Web-age information management conference. Berlin: Springer-Verlag, 2006: 85-96.

[19] Turney P D. Mining the Web for lexical knowledge to improve keyphrase extraction: learning from labeled and unlabeled data[EB/OL]. [2021-10-31]. https://arxiv.org/ftp/cs/papers/0212/0212011.pdf.

[20] Kim S N, Medelyan O, Kan M-Y, et al. SemEval-2010 task 5: automatic keyphrase extraction from scientific articles[EB/OL]. [2021-10-31]. https://aclanthology.org/S10-1004.pdf.

[21] Augenstein I, Das M, Riedel S, et al. SemEval2017 task 10: ScienceIE-Extracting keyphrases and relations from scientific publications[EB/OL]. [2021-10-31]. https://arxiv.org/pdf/1704.02853.pdf.

[22] Rodriguez A, Kim B, Turkoz M, et al. New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network[J]. Scientometrics, 2015, 103(2): 565-591.

[23] Li R, Zhang L, Guo S. Comparative analysis of patent co-citation clustering and patent bibliographic coupling clustering[J]. Library and Information Service, 2012, 56(8): 91-95.

[24] Lu Y, Xiong X, Zhang W, et al. Research on classification and similarity of patent citation based on deep learning[J]. Scientometrics, 2020, 123(1): 813-839.

Author Contributions

Yu Yan: Conceptualized the research, designed the study, conducted experiments, and wrote the manuscript.
Ju Peng: Analyzed data and revised the manuscript.
Shang Mingjie: Collected and cleaned data.

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