A Patent Topic Discovery Method Incorporating Terminology Knowledge: Postprint

Authors: Yu Yan, Zhao Naixuan

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

Abstract

[Purpose/Significance] To address the challenge in patent topic analysis wherein word-level basic units impede the recognition of multi-word terms and compromise topic model efficacy, this study proposes a patent topic discovery model that incorporates terminology. [Method/Process] The model first employs class entropy to effectively identify terminology within patent documents; it then utilizes a generalized Polya urn model to enhance the probability of assigning semantically similar terms to identical topics, thereby alleviating data sparsity issues inherent in term-based topic modeling. [Results/Conclusion] Experimental results indicate that the term information integrated into the proposed model improves topic generation quality, conferring enhanced readability and discriminability upon topic representations.

Full Text

A Patent Topic Discovery Method Integrated with Term Knowledge

Yu Yan[1],[2], **Zhao Naixuan**[1]
[1] Information Service Department, Nanjing Tech University, Nanjing 210009
[2] School of Electronics and Computer Engineering, Southeast University Chengxian College, Nanjing 211816

Abstract

[Purpose/Significance] In patent topic analysis, the word-based analytical unit causes multi-word patent terms to be fragmented, leading to topics that are difficult to interpret. To address this issue, this paper proposes a patent topic discovery model integrated with term knowledge. **[Method/Process]** The model first introduces category entropy to effectively identify terms in patent

literature. It then utilizes the Generalized Pólya Urn model to increase the probability of semantically similar terms being assigned to the same topic, thereby alleviating the data sparsity problem caused by using terms as the basic unit of topic model analysis. [**Result/Conclusion**] Experimental results demonstrate that the term information contained in the proposed model improves the quality of topic generation, making topic representations more readable and discriminative.

Keywords: patent analysis; topic discovery; term
Classification Number: G202
DOI: 10.13266/j.issn.0252-3116.2018.21.015

1 Introduction

Patent documents contain rich solutions to problems across various domains. Effective patent document analysis can identify technological hotspots, recognize core technologies, predict technological development trends, and help researchers gain inspiration, thereby shortening innovation cycles and reducing R&D costs. Consequently, patent document analysis holds significant research importance. Unlike traditional literature analysis methods, topic models can mine hidden semantic information in documents by analyzing the probability distribution of word co-occurrence patterns. With the widespread application of topic models, researchers have attempted to apply them to patent literature analysis to reveal the deep-level knowledge structure of patents [**?**]. Numerous studies have demonstrated the practical value of patent analysis based on topic models.

However, patent topic modeling uses words as the basic analytical unit, making it difficult to recognize multi-word terms in patents, which results in unclear and incomprehensible topics. For instance, Chinese patent analysis requires word segmentation beforehand. While current Chinese word segmentation technology can generally meet research needs with a relatively complete segmentation dictionary, patent literature contains numerous unregistered multi-word terms not recorded in general dictionaries. These terms, which embody core domain knowledge, are often fragmented during segmentation and thus fail to reveal their essential meaning. For example, the term "boronizing agent" is segmented into "渗硼剂"; "hot-dip plating" into "热浸镀"; and "duplex stainless steel" into "双相不锈钢". This fragmentation not only prevents recognition of core knowledge but also introduces spurious co-occurrences, causing generated topics to include irrelevant words and yielding poor model performance.

One approach to improving topic semantics is to focus on higher-order semantic units beyond individual words, typically by replacing the traditional word distribution with a distribution of higher-order semantic units in the topic model. However, term-based patent topic analysis has not been thoroughly investigated, and related research in other domains suffers from issues such as excessive

model complexity, poor scalability, low term recognition accuracy, and data co-occurrence sparsity.

To address these problems in patent topic modeling, this paper proposes a patent topic discovery model integrated with term knowledge. This model uses terms as the basic unit of analysis. First, based on the characteristics of patent literature, it introduces category entropy to effectively identify terms in patent documents. Then, it employs the Generalized Pólya Urn model to increase the probability of semantically similar terms being assigned to the same topic, thereby alleviating the data sparsity problem arising from using terms as the basic analytical unit. Experimental results show that compared with traditional patent topic models, the term knowledge-integrated model contains richer semantic information and exhibits stronger readability and topic discriminability.

2 Related Research

Unlike traditional patent text analysis methods, patent topic models analyze the probability distribution of word co-occurrence in patent text collections to mine implicit semantic information and reveal deep-level knowledge structures. For example, J. Tang et al. [**?**] proposed a topic-driven patent analysis method to assess competitor development status. B. Wang et al. [**?**] extracted technical terms from patent titles and abstracts, then used an extended topic model incorporating institutional information to analyze research hotspots and directions in specific domains. H. Chen et al. [**?**] used topic models to evaluate hidden topics in patent claims. M. Kim et al. [**?**] employed topic models to analyze patent abstracts and claims to generate patent development maps for understanding technology trends. A. Suominen et al. [**?**] used topic models to classify patent text data for predicting future technology trends. Fan Yu et al. [**?**] utilized topic models to transform high-dimensional patent text representations in lexical space into low-dimensional topic space representations, enabling K-nearest neighbor clustering of patents. Wang Bo et al. [**?**] applied topic models to patent text analysis for patent topic classification, addressing issues of excessive generality, poor timeliness, and lack of scientific rigor in previous methods. Wu Feifei et al. [**?**] extracted nouns and noun phrases from patent abstracts to construct topic models, revealing implicit topic evolution in patent texts to mine dynamic changes in corporate strategies. Liao Liefa and Le Fugang [**?**] proposed a patent technology evolution model based on topic models and classification codes. Chen Liang et al. [**?**] employed topic models to extract hierarchical topic models from patent corpora, describing the technical structure hidden in patent texts for patent technology evolution analysis.

Although patent topic models have achieved good analytical results, they use words as the basic unit. Multi-word terms in patents, which embody core domain knowledge, are often fragmented and difficult to recognize, introducing spurious co-occurrences that cause generated topics to include irrelevant words and produce poor results. One approach to improving topic semantics is to focus on higher-order semantic units beyond words, typically by replacing the

traditional word distribution with a distribution of higher-order semantic units in the topic model.

Based on when term recognition is performed during topic discovery, existing methods can be broadly divided into three categories: term preprocessing models, joint models, and term post-processing models. Term preprocessing models first perform term recognition and then discover topics based on the identified terms. Joint models simultaneously identify terms and discover topics [**?**]. Term post-processing models first obtain topics through conventional unigram topic models, then combine unigram words into terms [**?**]. While experiments show that joint models can improve topic model accuracy and produce more meaningful topic words, they are typically complex and face challenges of high computational complexity and poor scalability with large document collections. Term post-processing models separate term recognition from topic discovery, reducing computational complexity and improving scalability, but they cannot guarantee that words within a term share the same topic, which is crucial for term mining research.

Therefore, this paper focuses on term preprocessing models. These models first perform term recognition and then discover topics based on the identified terms, offering simplicity, intuitiveness, and extensibility. However, current methods have limitations: (1) In the term recognition stage, frequency-based methods produce low-quality terms, as high-frequency phrases are not necessarily terms while low-frequency phrases may be valid terms. (2) In the topic discovery stage, they fail to address the sparsity problem of higher-order semantic unit co-occurrence. The core idea of topic models is generally based on element co-occurrence within documents [**?**]. According to the power-law distribution in natural language [**?**], most word co-occurrences are sparse. When documents are treated as term collections rather than word bags, the elements become sparser and co-occurrence decreases further, negatively impacting topic modeling.

The proposed method belongs to the third category of term preprocessing models: first identifying terms, then discovering topics. It is simple, intuitive, and extensible. Based on Chinese patent characteristics, it introduces category entropy to identify terms in Chinese patents and employs the Generalized Pólya Urn model to alleviate sparsity by sampling related terms.

3 Patent Topic Model Integrated with Term Knowledge

This paper proposes a patent topic discovery model integrated with term knowledge, tailored to patent literature characteristics. After preprocessing patent texts (segmentation and data cleaning), the model performs term recognition (detailed in Section 3.1) and then conducts term-based topic modeling (detailed in Section 3.2) to achieve topic discovery.

3.1 Term Recognition

Relying solely on phrase frequency cannot distinguish patent terms from common phrases. For example, in patents, "invention relates to" appears far more frequently than actual terms. Compared with general documents, patent literature has distinct characteristics, typically containing general words and terms. General words are topic-independent, appear uniformly across categories, and often introduce terms to connect them with subsequent text. Terms express domain knowledge, exhibit high domain relevance, appear frequently in certain categories, and rarely or never in others. As shown in Figure 1 [Figure 1: see original paper], statements from three different patent categories after segmentation show that "boronizing agent", "electrolytic manganese metal", and "multistage centrifugal pump" are category-specific terms, while "this", "invention", "relates to", "a kind of", "of", "method" are general words appearing across all three categories. In patent literature, the boundary between general words and terms is clearer than in general corpora, and general words are more easily identified. Therefore, this paper first identifies general words to select candidate terms, then ranks these candidates to assess their likelihood of being terms.

3.1.1 Candidate Term Selection General words in patents are typically topic-independent and appear uniformly across and within categories, while terms appear non-uniformly in certain categories. To address this, we introduce an auxiliary patent text set containing multiple categories and use inter-category entropy and intra-category entropy to measure word distribution. Information entropy is an important concept in information theory used to measure information uncertainty.

Specifically, let $c_1, c_2, \ldots, c_m$ be m categories of auxiliary patent text sets, each containing several related patent texts. The distribution of word w across different categories is called inter-category information entropy (EC), calculated as:

$$EC(w) = -\sum_{j=1}^{m} \frac{df(w, c_j)}{df(w)} \log \frac{df(w, c_j)}{df(w)} \quad (1)$$

where $EC(w)$ represents the inter-category information entropy of word w; $df(w, c_j)$ denotes the document frequency of word w in category c_j; and $df(w) = \sum_{j=1}^{m} df(w, c_j)$ represents the total document frequency of word w in the auxiliary patent text set. By definition, when a word appears only in a single category, its inter-category information entropy is minimized; when a word is uniformly distributed across all categories, its inter-category information entropy reaches the maximum. As shown in Equation (1), the larger the category entropy EC, the more uniform the word's distribution across categories, and the more likely it is to be a general word in patents.

Words are sorted in descending order by category entropy, and words exceeding a threshold are selected as general words. The target patent texts are then coarsely segmented based on these general words. Following [**?**, **?**, **?**], word strings with frequency greater than 1 and length greater than 1, along with their substrings, are selected as candidate terms.

3.1.2 Candidate Term Ranking The C-value statistic can calculate the likelihood of each candidate term being a true term [**?**]. It is an improvement over term frequency calculation that enhances extraction of nested multi-word terms and excludes interference from non-term words. When both a substring and its parent string are included in the candidate term set (i.e., nested strings exist), their C-value can be calculated to determine whether they are genuine terms. The C-value method is simple, adaptable, domain-independent, considers the nested nature and length of candidate terms, and performs well in term recognition [**?**].

Specifically, C-value calculation uses four statistical features of candidate strings: total frequency in the corpus, frequency as a nested string, number of parent strings containing the nested string, and candidate string length. The calculation formula is:

$$C\text{-value}(x) = \begin{cases} \log|x| \times tf(x) & \text{if } x \text{ is not nested} \\ \log|x| \times \left(tf(x) - \frac{\sum_{y \in T_x} tf(y)}{|T_x|}\right) & \text{otherwise} \end{cases} \tag{2}$$

where x represents a candidate term; $|x|$ denotes the length of x; $tf(x)$ represents the frequency of x in the target patent text set; T_x denotes the set of candidate terms in the target patent text set that contain x; and $|T_x|$ represents the number of elements in T_x. The formula shows that C-value is related to the candidate term's frequency in the corpus—higher frequency indicates greater termhood. Additionally, it considers candidate term length, assuming that long strings appearing with certain frequency are more meaningful and more likely to be terms than short strings.

3.2 Topic Discovery

Based on conventional topic discovery models, this paper introduces the Generalized Pólya Urn (GPU) model [**?**] to alleviate sparsity by sampling related terms. The GPU model is incorporated into the traditional topic discovery model to mitigate data sparsity issues.

3.2.1 Traditional Topic Model LDA (Latent Dirichlet Allocation) [**?**] is a commonly used topic model. Due to its simple parameters and lack of overfitting, it has become a research focus in topic modeling. Therefore, this paper uses the LDA model for patent text topic modeling. LDA is a three-layer

Bayesian probability model consisting of words, topics, and documents. The model assumes each document contains several latent topics, and each topic contains specific words. The relationship between documents and words is reflected through latent topics. Latent topics are mutually independent, shared by all documents in the collection, while each document has a specific topic distribution. The model typically uses Gibbs sampling for posterior distribution estimation, calculated as:

$$p(z_n^{(d)} = k \mid \mathbf{z}_{-d,n}, \mathbf{W}, \alpha, \beta) \propto \frac{N_{k|d} + \alpha}{N_d + K\alpha} \cdot \frac{N_{w|k} + \beta}{N_k + V\beta} \quad (3)$$

where $z_n^{(d)} = k$ indicates that the n-th word w in document d is assigned to topic k; $\mathbf{z}_{-d,n}$ denotes topic assignments excluding the n-th word in document d; $\mathbf{W}$ represents all words in the text collection; K denotes the number of topics; and V represents the total vocabulary size. Once each word's topic in each document is obtained, posterior estimates of θ and ϕ in the LDA model can be computed as:

$$\theta_{d,k} = \frac{N_{k|d} + \alpha}{N_d + K\alpha} \quad (4)$$

$$\phi_{k,w} = \frac{N_{w|k} + \beta}{N_k + V\beta} \quad (5)$$

where $\theta_{d,k}$ represents the probability that document d contains topic k; and $\phi_{k,w}$ represents the probability of word w in topic k.

3.2.2 GPU-Based Topic Model The Generalized Pólya Urn (GPU) model is a probabilistic statistical model and an extension of the standard Pólya Urn (PU) model. In topic models, objects of interest are represented as colored balls in an urn. The probability of observing a ball of different colors can be approximated by the proportion of each color in the urn through sampling and replacement operations. In the PU model, a ball is randomly drawn from the urn, its color is observed, then it is returned to the urn along with additional balls of the same color, and the selection process repeats. This process resembles using Gibbs sampling to derive the LDA model. In the GPU model, after randomly drawing a ball, it is returned to the urn along with a ball of the same color. Additionally, some balls of similar colors are also returned to the urn. Therefore, the GPU model ensures that in subsequent sampling, not only can the same color ball be drawn with higher probability, but also other similar-colored balls can be drawn with relatively high probability.

This operation can be analogized to term-based topic models, where the GPU model increases the probability of terms (e.g., "austenitic stainless steel" and "stainless steel") being assigned the same topic, thereby alleviating sparsity

issues in term co-occurrence in the basic model. Recent research has shown that incorporating prior knowledge using the GPU model is a direct and effective method [**?**].

Specifically, in the term-based Chinese patent topic discovery model, the GPU-based Gibbs sampling estimates the posterior topic distribution as:

$$p(z_n^{(d)} = k \mid \mathbf{z}_{-d,n}, \mathbf{W}, \alpha, \beta, \mathbf{A}) \propto \frac{N_{k|d} + \alpha}{N_d + K\alpha} \cdot \frac{\sum_v N_{v|k} A_{v,t_n^{(d)}} + \beta}{\sum_{v'} \sum_v N_{v|k} A_{v,v'} + V'\beta} \quad (6)$$

where $z_n^{(d)} = k$ indicates that the n-th term t in document d is assigned to variable k; $\mathbf{z}_{-d,n}$ denotes topic assignments excluding the n-th term in document d; V' represents the total number of terms; and $\mathbf{A}$ is a $|V'| \times |V'|$ GPU promotion matrix. To increase the probability of semantically similar elements being assigned to the same topic, matrix $\mathbf{A}$ can be assigned values based on term similarity. Element $A_{v,v'} = sim(v, v')$ represents the similarity between terms v and v', defined as:

$$A_{v,v'} = sim(v, v') = \begin{cases} 1, & v = v' \\ sim(v, v'), & sim(v, v') > \sigma \text{ and } v' \neq v \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $|v \cap v'|$ represents the number of identical words between terms v and v', $|v \cup v'|$ represents the number of different words, and σ is a specified similarity threshold. The threshold filters out term pairs with low similarity, considering their similarity insignificant and not adding them to the GPU promotion matrix $\mathbf{A}$. For example, the similarity between "austenitic stainless steel" and "antibacterial stainless steel" is 1/3. This paper selects a similarity threshold of 0.2.

Correspondingly, the posterior estimates of θ and ϕ in the GPU-based topic model are calculated as:

$$\theta_{d,k} = \frac{N_{k|d} + \alpha}{N_d + K\alpha} \quad (8)$$

$$\phi_{k,w} = \frac{\sum_v N_{v|k} A_{v,t_n^{(d)}} + \beta}{\sum_{v'} \sum_v N_{v|k} A_{v,v'} + V'\beta} \quad (9)$$

In this model, by introducing term similarity knowledge and using the GPU model to increase the probability of terms being assigned to the same topic, the sparsity problem in term co-occurrence in the basic model is alleviated. When $\mathbf{A}$ is an identity matrix, the model reduces to the Pólya Urn model. Due to

similarity symmetry, matrix **A** is sparse. Therefore, applying the GPU model directly to the sampling process does not increase model complexity or inference difficulty.

4 Experiments

4.1 Dataset

To validate the effectiveness of the proposed model, we selected patent literature from two domains: rare earth steel and electrolytic manganese. Rare earth steel refers to steel with added rare earth elements that exhibits improved transverse performance, wear resistance, and corrosion resistance. With recent industry decapacity and transformation efforts, leveraging rare earth resource advantages to enhance the market position of rare earth steel has become an important issue. Manganese is the second most used alloy element in the steel industry after iron, playing a crucial role in steel smelting, particularly in manganese-substituted nickel stainless steel and advanced special steel smelting. Additionally, manganese is important in non-ferrous metals, magnetic materials, catalysts, and battery materials. Systematic analysis of Chinese electrolytic manganese patents can help explore research status and technology development trends for rational planning and scientific decision-making.

Experiments were based on the China National Intellectual Property Administration patent database, searching for Chinese invention patents published up to December 8, 2017. Using “rare earth steel” and “electrolytic manganese” as keywords, we obtained 1,547 and 1,544 Chinese invention patents, respectively. After data crawling, cleaning, and deduplication, patent titles and abstracts were used as the domain patent text sets for analysis.

To obtain patent general words, 2,000 Chinese invention patent titles and abstracts were randomly selected from each of the A-H IPC classification categories as auxiliary patent text sets. Table 1 shows basic dataset information.

Table 1 Basic Dataset Information

Dataset Type	Target Patent Text Set	Auxiliary Patent Text Set
Category A (Human Necessities)	-	2000
Category B (Operations; Transport)	-	2000
Category C (Chemistry; Metallurgy)	-	2000
Category D (Textiles; Paper)	-	2000
Category E (Fixed Constructions)	-	2000
Category F (Mechanical Engineering)	-	2000
Category G (Physics)	-	2000
Category H (Electricity)	-	2000
Rare Earth Steel	1547	-
Electrolytic Manganese	1544	-

4.2 Evaluation Methods

Experiments primarily evaluate term recognition accuracy and topic model quality.

First, the P@N method assesses patent term recognition accuracy by judging the precision of the top N terms in the final ranked candidate list. The top N terms automatically extracted by the model are manually evaluated. To avoid subjectivity and domain knowledge limitations, obviously correct or incorrect terms are directly labeled, while ambiguous terms are verified using knowledge websites like Baidu Baike, Wikipedia, and Hudong Baike. The calculation formula is:

$$P@N = \frac{\#\text{correct terms in top N}}{\#\text{top N candidate terms}} \times 100\% \quad (10)$$

Second, the average KL (Kullback-Leibler) distance between topics evaluates generated topic quality. KL distance measures the distance between two probability distributions—larger KL values indicate greater inter-topic distance and higher topic quality. The average KL distance is defined as:

$$avg\_KL = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} KL(\phi_i \| \phi_j) \quad (11)$$

where KL distance $KL(\phi_i \| \phi_j) = \sum_{v=1}^{V} \phi_{iv} \log \frac{\phi_{iv}}{\phi_{jv}}$. KL distance is asymmetric, but topic similarity measurement should be symmetric. Therefore, we adjust the formula using symmetric Jensen-Shannon distance to measure the distance between two topic word distributions, replacing KL in Equation (11). The specific calculation is:

$$JS(\phi_i, \phi_j) = \frac{KL(\phi_i, \phi_j) + KL(\phi_j, \phi_i)}{2} \quad (12)$$

4.3 Experimental Procedure and Parameter Settings

First, term recognition is performed. The ICTCLAS segmentation system from the Institute of Computing Technology, Chinese Academy of Sciences (http://ictclas.nlpir.org/) is used for word segmentation on both target and auxiliary patent text sets. This system provides Chinese word segmentation and part-of-speech tagging and is widely used [**?**]. Based on segmentation information, word category entropy in the target set is calculated. Through manual evaluation, the top 500 words with highest category entropy are selected as general words. The target set is coarsely segmented using these general words, candidate terms are selected, and then C-value ranking is applied to select the top candidates as terms.

Table 2 shows the top 10 words with highest category entropy values. These words typically appear across all patent categories, are unrelated to specific patent topics, contain minimal semantic information, and can serve as general words.

Table 2 Top 10 Words with Maximum Category Entropy Values

Rank	Word	Category Entropy
1	this	High
2	invention	High
3	relates to	High
4	a kind of	High
5	of	High
6	method	High
7	disclose	High
8	provide	High
9	technology	High
10	field	High

For identified terms, the maximum length matching method is used to optimize the segmentation of fragmented phrases in target patent texts, forming a bag-of-terms for topic modeling. In topic modeling, parameters are set based on experience to $\alpha = 50/K$, $\beta = 0.01$, Gibbs sampling iterations to 2000, and saving iterations to 1000. The number of topics K is selected by calculating perplexity of the basic patent topic model (Section 3.2) with optimal values chosen via five-fold cross-validation. Based on calculations, $K = 15$ for the rare earth steel dataset and $K = 20$ for the electrolytic manganese dataset.

4.4 Term Recognition Method Evaluation

To validate the effectiveness of the proposed term recognition method, we compare it with two alternative methods:

1. **Rule-C-value**: Traditional term formation rule matching selects noun phrases, extracts matched noun terms [**?**], then ranks candidates using C-value.
2. **EC-C-value**: Uses the category entropy method proposed in Section 3.1 to select candidate terms, then ranks them by C-value in descending order.

Results are shown in Figures 2 [Figure 2: see original paper] and 3 [Figure 3: see original paper]. In both datasets, EC-C-value significantly outperforms Rule-C-value, indicating that category entropy-based candidate selection combined with C-value ranking is more effective than rule matching.

Table 3 lists the top 10 candidate terms from both methods, with correct terms in bold. The rule-based method cannot effectively filter general words, which

appear with high frequency but are not in stopword lists, generating many false candidate terms and resulting in low extraction accuracy. In contrast, the general word-based method can identify “preparation”, “method”, “invention”, and “production” as general words, excluding these high-frequency words from candidate terms and thereby improving term extraction accuracy.

Table 3 Top 10 Candidate Terms from Both Methods

Rule-C-value	EC-C-value
weight	**duplex stainless steel**
percentage	**mass percentage**
preparation	**manufacturing method**
method	**production method**
invention	**rare earth permanent magnet**
disclose	**rare earth alloy**
rare earth	**alloy element**
permanent magnet	**austenitic stainless steel**
rare earth	**steel-bonded carbide**
alloy	**waste magnetic steel**

4.5 Topic Model Evaluation

To validate the effectiveness of the proposed GPU-based topic model, we use terms selected by the Section 3.1 method and compare two topic models:

1. **EC-PhraseLDA**: This model assumes that co-occurring multi-word terms are not coincidental and must have certain connections. It uses a potential function to represent the interaction between words in terms, where term words share the same latent topic [**?**].
2. **EC-GPU-LDA**: Uses the GPU-based topic discovery model proposed in Section 3.2.2 to model the bag-of-terms.

Results are shown in Figure 4 [Figure 4: see original paper]. EC-GPU-LDA outperforms EC-PhraseLDA. While EC-PhraseLDA replaces fragmented word groups with terms, it also reduces term co-occurrence, affecting topic modeling effectiveness. The proposed EC-GPU-LDA introduces the GPU model to increase sampling probability of similar terms, thereby increasing inter-topic distance.

Table 4 lists the top 10 topic words and terms from both models, with multi-word terms in bold. With EC-PhraseLDA, due to term sparsity, the top 10 topic words contain no multi-word terms. In contrast, EC-GPU-LDA can infer topics for similar terms based on similarity, enabling some sparse multi-word terms to enter the top 10 words, thus solving the term sparsity problem.

Table 4 Topic Representation Examples

EC-PhraseLDA	EC-GPU-LDA
polyacrylamide	polyacrylamide
cladding metal	cladding metal
rare earth ferrosilicon	rare earth ferrosilicon

4.6 Comparison with Traditional Patent Topic Models

Finally, we compare the proposed method with commonly used patent topic analysis methods:

1. **LDA**: Removes stopwords from patent texts and performs topic modeling on word-based preprocessed patent texts [**?**, **?**].
2. **RuleLDA**: Selects noun patent terms based on lexical rules, where words within terms share the same topic for topic modeling [**?**].
3. **EC-GPU-LDA**: Uses the proposed method with EC for general word selection to identify patent terms, then applies the GPU-based topic model.

Results are shown in Figure 5 [Figure 5: see original paper]. LDA performs worst, RuleLDA performs second worst, and the proposed EC-GPU-LDA achieves the best modeling results.

Table 5 lists the top 10 topic words and terms from three models, with multi-word terms in bold. The word-based topic model contains general words like "invention" and "the said", affecting topic quality. The rule-based topic model has low term recognition accuracy, e.g., "production method" is not a genuine term. The term-based representation uses terms as the basic unit, includes correct multi-word terms, and uses GPU to solve multi-word term sparsity, making the model more interpretable.

Table 5 Topic Representation Examples

LDA	RuleLDA	EC-GPU-LDA
invention	production method	**cladding metal**
method	electrolytic manganese slag	**rare earth ferrosilicon**
technology	flux-cored wire	**polyacrylamide**

5 Conclusion

Addressing the problem of fragmented patent terms leading to uninterpretable topics in patent topic discovery, this paper proposes a patent topic model integrated with term knowledge. The model first introduces category entropy based on patent literature characteristics to effectively identify terms in Chinese patent documents. It then uses the Generalized Pólya Urn model to increase the probability of semantically similar terms being assigned to the same topic, alleviating data sparsity caused by using terms as the basic analytical unit. By

using terms as the basic unit of topic model analysis, the model contains richer semantic information and exhibits stronger readability compared to traditional word-based representations. Experimental results show that the term-based representation significantly improves topic readability and effectively reduces the impact of term sparsity from both co-occurrence and semantic relevance perspectives. The model requires no domain knowledge or complex linguistic rules, is data-driven, and is suitable for large-scale Chinese patent literature topic analysis.

Although the term knowledge-integrated patent topic model achieves better results than traditional models, the accuracy of low-frequency term recognition remains low. After C-value ranking, low-frequency candidate terms have small C-values and are removed. Future work will further investigate how to improve low-frequency term recognition accuracy.

References

[1] TANG J, WANG B, YANG Y, et al. PatentMiner: topic-driven patent analysis and mining[C]//ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2012: 1366-1374.

[2] WANG B, LIU S, DING K, et al. Identifying technological topics for patent competitive intelligence analysis: a topic distribution probability approach for patent competitive intelligence analysis: a case study in LTE technology[J]. Scientometrics, 2014, 101(1): 685-704.

[3] CHEN H, ZHANG G, LU J, et al. A fuzzy approach for measuring development of topics in patents using Latent Dirichlet Allocation[C]//IEEE international conference on fuzzy systems. Piscataway, NJ: IEEE, 2015: 1116-1116.

[4] KIM M, PARK Y, YOON J. Generating patent development maps for technology monitoring using semantic patent-topic analysis[J]. Computers & industrial engineering, 2016, 98(1): 289-299.

[5] SUOMINEN A, TOIVANEN H, SEPPANEN M. Firms' knowledge profiles: mapping patent data with unsupervised learning[J]. Technological forecasting & social change, 2016, 115(1): 1-12.

[6] FAN Y, FU H, WEN Y. Patent information clustering technology based on LDA model[J]. Computer Applications, 2013, 33(S1): 87-89.

[7] WANG B, LIU S, DING K, et al. Patent content analysis method based on LDA topic model[J]. Science Research Management, 2015, 36(3): 111-117.

[8] WU F, ZHANG Y, HUANG L, et al. Multi-dimensional dynamic evolution analysis of technology themes based on AToT model: a case study of graphene technology[J]. Library and Information Service, 2017, 1(5): 95-102.

[9] LIAO L, LE F. Research on patent technology evolution based on LDA model and classification codes[J]. Modern Information, 2017, 37(5): 13-18.

[10] CHEN L, ZHANG J, ZHANG H, et al. Application research of hierarchical topic model in technology evolution analysis[J]. Library and Information Service, 2017, 1(5): 103-108.

[11] WALLACH H M. Topic modeling: beyond bag-of-words[C]//International conference on machine learning. New York: ACM, 2006: 977-984.

[12] WANG X, MCCALLUM A, WEI X. Topical N-grams: phrase and topic discovery, with an application to information retrieval[C]//Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA: ACL, 2007: 697-702.

[13] LINDSEY R V, HEADDEN III W P, STIPICEVIC M J. A phrase-discovering topic model using hierarchical Pitman-Yor processes[C]//Joint conference on empirical methods in natural language processing and computational natural language learning. Stroudsburg, PA: ACL, 2012: 214-222.

[14] DANILEVSKY M, WANG C, DESAI N, et al. Automatic construction and ranking of topical keyphrases on collections of short documents[C]//Proceedings of the 2014 SIAM international conference on data mining. Philadelphia, PA: SIAM, 2014: 398-406.

[15] EL-KISHKY A, SONG Y, VOSS C R, et al. Scalable topical phrase mining from text corpora[J]. Proceedings of the VLDB endowment, 2014, 8(3): 305-316.

[16] ZHANG Q, ZHANG Z. Research on topic phrase mining method based on PhraseLDA model[J]. Library and Information Service, 2017, 61(8): 120-125.

[17] HEINRICH G. A generic approach to topic model[M]//Machine learning and knowledge discovery in databases. Berlin: Springer, 2009: 517-532.

[18] ZIPF G K. Selected studies of the principle of relative frequency in language[J]. Language, 1933, 9(1): 89-92.

[19] HAN H, ZHU D, WANG X. Method for extracting patent technology terms[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(12): 1280-1285.

[20] XU C, SHI S, FANG X, et al. Chinese patent document term extraction[J]. Computer Engineering and Design, 2013, 34(6): 2175-2179.

[21] 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.

[22] SPASIC I, GREENWOOD M, PREECA A, et al. FlexiTerm: a flexible term recognition method[J]. Journal of biomedical semantics, 2013, 4(1): 27-42.

[23] MAYNARD D, ANANIADOU S. Identifying terms by their family and friends[C]//Conference on computational linguistics. Stroudsburg, PA: ACL, 2000: 530-536.

[24] LI C, WANG H, ZHU M, et al. Automatic extraction of multi-word strings based on domain category information C-value[J]. Journal of Chinese Information Processing, 2010, 24(1): 94-99.

[25] LIU L, LIU X. Domain phenomenon term extraction based on delimiters and context terms[J]. Journal of South China University of Technology (Natural Science Edition), 2011, 39(7): 146-152.

[26] HU A, ZHANG J, LIU J. Chinese term extraction based on improved C-value method[J]. New Technology of Library and Information Service, 2013, 29(2): 24-29.

[27] ZHANG J, ZHANG H, ZHAI D. Research on word segmentation method for Chinese patent claims[J]. New Technology of Library and Information Service, 2014, 30(9): 91-98.

[28] MAHMOUD H. Pólya urn models[M]. New York: Chapman & Hall/CRC, 2009.

[29] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993-1022.

[30] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the national academy of sciences, 2004, 101(1): 5228-5230.

[31] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing semantic coherence in topic models[C]//Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA: ACL, 2011: 262-272.

[32] CHEN Z, MUKHERJEE A, LIU B, et al. Leveraging multi-domain prior knowledge in topic models[C]//International joint conference on artificial intelligence. Menlo Park, CA: AAAI, 2013: 2071-2077.

[33] CHEN Z, MUKHERJEE A, LIU B, et al. Discovering coherent topics using general knowledge[C]//ACM international conference on information & knowledge management. New York: ACM, 2013: 209-218.

[34] CHEN Z, LIU B. Mining topics in documents: standing on the shoulders of big data[C]//ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2014: 1116-1125.

[35] SUN R, GUO S, JI D. Topic representation method incorporating event knowledge[J]. Chinese Journal of Computers, 2017, 40(4): 791-804.

Author Contributions: Yu Yan: Conceived research idea, designed research plan, conducted experiments, wrote paper.
Zhao Naixuan: Analyzed data, revised paper.

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