Research on Automatic Selection of Domain-Specific Stop Words in Patent Text Topic Modeling (Postprint)

Authors: Yú Yǎn, Zhao Naixuan

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

Abstract

[目的/意义] To address the issue that automatic selection of domain-specific stop words in patent text topic modeling has not been adequately studied, this paper proposes a novel automatic method for selecting domain stop words for patent text topic model analysis, thereby improving the discriminability and modeling quality of patent topic models.

[方法/过程] Domain stop words are essentially terms with relatively low information content and poor discriminative power across different categories of patent texts. Therefore, an auxiliary patent text corpus is introduced, category entropy is utilized to measure the distribution of terms, and terms are then ranked according to their category entropy. Those with the highest category entropy are selected as domain stop words.

[结果/结论] Experiments on patent text data validate the feasibility and effectiveness of the proposed method, which can effectively enhance the discriminability of patent topic models.

Full Text

Preamble

Vol. 62 No. 11 June 2018

Research on Automatic Selection of Domain-Specific Stop Words in Patent Text Topic Modeling

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**] To address the insufficient research on automatic selection of domain-specific stop words in patent text topic modeling, this paper proposes a novel method for automatically selecting domain-specific stop words for patent text topic model analysis, aiming to improve the differentiation and modeling quality of patent topic models.

[**Method/Process**] Domain-specific stop words are essentially words with low information content and low discriminative power across different categories of patent texts. Therefore, this study introduces an auxiliary patent text set and uses category entropy to measure word distribution, then selects words with the maximum category entropy as domain-specific stop words based on this ranking.

[**Result/Conclusion**] Experiments on patent text data verify the feasibility and effectiveness of the proposed method, which can effectively improve the differentiation of patent topic models.

Keywords: patent text, topic modeling, domain-specific stop words, automatic selection
Classification Number: G202
DOI: 10.13266/j.issn.0252-3116.2018.11.014

---

1 Introduction

Effective patent text analysis can identify technological hotspots, recognize core technologies, and predict technological development trends, helping researchers gain inspiration and thereby shortening innovation cycles and reducing costs. Consequently, patent text analysis holds significant research importance. Traditional patent text analysis methods typically use words directly from patent texts as topics or concepts for modeling and analyzing technological domain status [**?**]. However, as protected documents, patent applicants often employ vague or abstract expressions to broaden protection scope and increase grant probability. Therefore, understanding patent texts at the latent semantic level yields better analysis results.

Unlike traditional text analysis methods, topic models analyze the probability distribution of word co-occurrence in text collections to mine latent semantic information and have achieved good results in text analysis. As topic models matured, researchers began applying them to patent text analysis to reveal deep-level knowledge structures [**?**]. Although topic models effectively mine latent semantic information in patent texts, the learned topic distributions tend to skew toward high-frequency words during training. These are typically stop words with high frequency but little meaning, such as Chinese particles "的" (de) and "是" (shi), or English words "of" and "the". During iterative topic model generation, these words frequently appear across multiple topics, causing topic distributions to fluctuate, increasing similarity between topics, slowing

convergence, and negatively impacting results [?]. To avoid this, stop word lists are typically used to pre-remove stop words before building patent text topic models. However, this method cannot completely filter out semantically poor words.

Stop words include both general stop words and domain-specific stop words. The former are standard common-domain stop words, while the latter are words with minimal discriminative power in specific domains. In patent text analysis, words like "方法" (method), "包含" (include), and "发明" (invention) frequently appear but contain little information and have low discriminative power, poorly representing semantic information. Some patent text topic model studies use manual methods or word frequency and document frequency to select domain stop words [?, ?]. However, frequency-based selection may include useful domain terms. For example, in 3D printing patent analysis, the word "打印机" (printer) appears frequently but has research value and should not be simply removed as a domain stop word.

Domain-specific stop words in Chinese patent texts have unique characteristics. Domain-general words typically are topic-independent and uniformly appear across multiple patent categories. For instance, "方法", "包含", and "发明" repeatedly appear across chemical, mechanical, physical, and electrical patent categories. Conversely, patent terms containing core domain knowledge appear frequently only in specific categories and rarely or never in others. Therefore, this paper introduces an auxiliary patent text set containing multiple categories, proposes the concept of category entropy to measure word distribution, and automatically selects domain stop words to improve differentiation and modeling quality of patent text topic models.

Additionally, researchers have attempted other methods for automatic domain stop word selection. T.W. Lo et al. [?] proposed a random sampling method for information retrieval, suggesting the most effective stop word list combines classic lists with automatically extracted words. L. Hao et al. [?] proposed an x^2-statistic method to generate domain stop words for furniture queries in e-commerce. M.P. Sinka and D.W. Corne [?] proposed word entropy using clustering and random retrieval algorithms. M. Jungiewicz and M. Lopuszynski [?] developed an unsupervised method based on the observation that stop word distributions typically follow random variable models (e.g., Poisson distribution). M. Makrehchi and M.S. Kamel [?] proposed backward filtering performance and data sparsity index concepts. Gu Yijun et al. [?] calculated joint entropy based on word probabilities within sentences and across the corpus. Gong Zheng and Guan Gaowa [?] used joint entropy for Mongolian stop words. Zhu Jie and Li Tianrui [?] studied Tibetan stop word selection. However, these methods are not suitable for patent text processing due to its unique characteristics.

2 Related Research

Stop words are considered words without semantic information or discriminative power. They constitute most of text data and create significant interference in text analysis, carrying little information while suppressing other words and greatly affecting processing efficiency and accuracy. Stop word removal is widely used in text analysis. W. Frakes et al. [**?**] found that eliminating high-frequency words early in automatic indexing could improve retrieval speed and reduce storage without decreasing accuracy. C. Silva [**?**] verified that SVM-based text classifiers improved accuracy after stop word removal. Guan Qin [**?**] evaluated different Chinese stop word lists and found they significantly impact clustering accuracy, making appropriate list selection crucial. In summary, stop word selection critically affects text analysis results, and removal is an important preprocessing step.

Stop words can be divided into general and domain-specific categories. Domain stop words vary by domain and dataset. For example, "学习" (learning) may be a domain stop word in education but not in computer science. Domain stop words have been applied in human resources [**?**], bioinformatics and gene ontology [**?**], information retrieval [**?**], and e-commerce [**?**]. Selection is typically manual, but automatic selection is more flexible and promising. However, designing efficient automatic selection algorithms is challenging. Common approaches use term frequency (TF) or document frequency (DF). TF-based selection assumes words appearing frequently in a collection are domain stop words. DF calculates how many documents contain a word, assuming words appearing in many documents lack discriminative power.

3 LDA Topic Model

LDA (Latent Dirichlet Allocation) [**?**] is a commonly used topic model that has become a research focus due to its simple parameters and lack of overfitting. This paper uses LDA for patent text modeling. LDA is a three-layer Bayesian probability model consisting of words, topics, and documents. It assumes each document contains several latent topics, and each topic contains specific words. The relationship between documents and words is reflected through latent topics, which are independent and shared across all documents, with each document having a specific topic distribution.

The model typically uses Gibbs sampling to estimate topic posterior distributions, calculated as:

$$p(z_{ij} = k | z_{-ij}, w, \alpha, \beta) \propto \frac{n^k_{i(\cdot)} + \beta}{n^k_{(\cdot)(\cdot)} + V\beta} \cdot \frac{n^k_{(\cdot)j} + \alpha}{n^{(\cdot)}_{(\cdot)j} + K\alpha}$$

where z_{ij} represents the topic variable for word w_i in document d_j; $-ij$ indicates excluding word w_i from document d_j; n_{ijk} counts how many times word w_i in

document d_j is assigned to topic k; $(\cdot)$ denotes summation across corresponding dimensions (word, document, topic); β is the Dirichlet prior for words; α is the Dirichlet prior for topics; K is the number of topics; and V is the total vocabulary size.

Once topic assignments for each word are obtained, posterior estimates for θ and ϕ in LDA are calculated as:

$$\theta_{jk} = \frac{n^k_{(\cdot)j} + \alpha}{n^{(\cdot)}_{(\cdot)j} + K\alpha}$$

$$\phi_{ki} = \frac{n^k_{i(\cdot)} + \beta}{n^k_{(\cdot)(\cdot)} + V\beta}$$

where θ_{jk} is the probability that document d_j contains topic k, and ϕ_{ki} is the probability of word w_i in topic k.

4 Automatic Selection of Domain-Specific Stop Words

Compared with patent terms, domain stop words in patents typically have category independence and appear uniformly across categories. Conversely, patent terms containing core domain knowledge appear frequently only in specific categories. Therefore, this paper introduces an auxiliary patent text set containing multiple categories to identify domain stop words. Domain stop words typically appear uniformly both between and within categories, and this distribution can be measured using information entropy—a key concept in information theory that measures uncertainty. When words are unevenly distributed across texts, they provide more information and have stronger discriminative ability. This unevenness can be measured by word entropy.

Specifically, this paper introduces an auxiliary patent text set with categories $c_1, c_2, \ldots, c_m$, each containing relevant patent documents. The distribution of word w_k across different categories is called between-category entropy (EBC), calculated as:

$$EBC(w_i) = -\sum_{t=1}^{m} \frac{df(w_i, c_t)}{df(w_i)} \log \frac{df(w_i, c_t)}{df(w_i)}$$

where $EBC(w_i)$ is the between-category entropy of word w_i; $df(w_i, c_t)$ is the document frequency of w_i in category c_t; and $df(w_i) = \sum_{t=1}^{m} df(w_i, c_t)$ is the total document frequency of w_i in the auxiliary set. When a word appears only in documents of a single category, EBC is minimized; when uniformly distributed across all categories, EBC reaches its maximum.

Similarly, within-category entropy (EIC) measures word distribution within category c_t:

$$EIC(w_i, c_t) = -\sum_{d_j \in c_t} \frac{tf(w_i, d_j)}{tf(w_i, c_t)} \log \frac{tf(w_i, d_j)}{tf(w_i, c_t)}$$

where $EIC(w_i, c_t)$ is the within-category entropy; $tf(w_i, d_j)$ is term frequency of w_i in document d_j; $tf(w_i, c_t) = \sum_{d_j \in c_t} tf(w_i, d_j)$ is total term frequency in category c_t; and $|c_t|$ is the number of documents in c_t. More uniform distribution within a category yields higher EIC values.

Combining EBC and EIC, category entropy (EC) measures overall word distribution across categories:

$$EC(w_i) = EBC(w_i) \times \sum_{t=1}^{m} EIC(w_i, c_t)$$

Higher EC values indicate more uniform distribution both within and between categories, making the word more likely to be a domain-specific stop word in patent texts.

5 Experiments

5.1 Dataset and Experimental Settings

To validate the proposed method, experiments were conducted on 3D printing and intelligent speech-related patent texts. 3D printing is an emerging manufacturing technology that has attracted widespread attention for revolutionizing traditional manufacturing. Intelligent speech represents a new human-computer interaction mode that has developed rapidly with mobile internet, deep learning, and big data. Patents were retrieved from the China National Intellectual Property Administration database using search queries: “3D printing OR rapid prototyping OR additive manufacturing OR three-dimensional printing OR incremental manufacturing OR digital manufacturing OR intelligent manufacturing” and “intelligent speech OR speech recognition OR speech synthesis OR natural language understanding OR speech interaction OR speech technology OR speech control” for the period 2013-2017 (retrieved August 1, 2017). After cleaning and deduplication, 7,790 3D printing patents and 5,272 intelligent speech patents were collected as target sets. Additionally, 2,000 patents were randomly sampled from each IPC class (A-H) as an auxiliary set. Dataset statistics are shown in Table 1 .

The ICTCLAS segmentation system was used for word segmentation, with the Harbin Institute of Technology stop word list removing general stop words. LDA parameters were set as $\alpha = 50/K$, $\beta = 0.01$, with 2,000 Gibbs sampling iterations (saving every 1,000). Optimal topic numbers K were selected via

perplexity with five-fold cross-validation, yielding $K = 15$ for 3D printing and $K = 10$ for intelligent speech.

5.2 Evaluation Metrics

Average KL distance measures topic differentiation. The symmetric Jensen-Shannon distance between topics is:

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

where $KL(\phi_i||\phi_j) = \sum_{v=1}^{V} \phi_{iv} \log \frac{\phi_{iv}}{\phi_{jv}}$. Average KL distance is:

$$avg\_KL = \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} JS(\phi_i, \phi_j)}{K(K-1)}$$

To compare across different vocabularies after stop word removal, normalized average KL distance is used:

$$avg\_KL' = \log_b(avg\_KL/V)$$

where V is vocabulary size. Larger avg_KL' indicates better topic distinguishability.

5.3 Experimental Results

5.3.1 Determining Domain Stop Word Selection Threshold Using the auxiliary set, words' category entropy was calculated after segmentation and general stop word removal. Table 2 shows the top 20 words with highest category entropy. These words appear across all patent categories, are unrelated to specific patent topics, contain little semantic information, and can serve as domain stop words.

Thresholds of 120, 130, 140, 150, 160, and 170 were tested. Words exceeding the threshold were removed as domain stop words, and topic models were built. Results in Figure 1 [Figure 1: see original paper] show that at threshold 150, both datasets achieve maximum avg_KL' values, indicating optimal topic differentiation. Therefore, threshold 150 was used in subsequent experiments.

5.3.2 Comparison of Different Domain Stop Word Methods Three methods were compared: (1) TF: selecting highest frequency words; (2) DF: selecting highest document frequency words; (3) EC: the proposed category entropy method. Corresponding models are named TF-LDA, DF-LDA, and EC-LDA. Baseline models without any stop word removal (LDA) and with only general stop words (Gen-LDA) were also compared.

Table 3 shows that: (1) Gen-LDA outperforms LDA, confirming general stop word removal improves topic differentiation; (2) TF-LDA outperforms LDA, showing appropriate high-frequency word removal helps; (3) DF-LDA outperforms TF-LDA, indicating document frequency is better than term frequency for patent texts; (4) EC-LDA achieves the highest average KL distance on both datasets, demonstrating that using auxiliary patent sets and category entropy most accurately measures word distribution and information content, yielding best performance.

Table 4 lists top 20 TF and DF words, with domain stop words in bold. Besides low-information words like "发明" (invention) and "包括" (include), TF and DF also include domain terms like "打印" (print), "3D", "语音" (speech), and "模块" (module), which should not be removed. This explains why DF-LDA performs only slightly better than Gen-LDA—some content-bearing words were incorrectly removed.

5.3.3 Patent Topic Word Comparison Table 5 and Table 6 compare top 5 words per topic from Gen-LDA and EC-LDA models. Gen-LDA results contain many low-meaning words like "发明", "方法", "技术", and "包括" (bolded). EC-LDA automatically selects low-information, low-discrimination words as domain stop words, producing more coherent and understandable topics.

6 Conclusion

This paper addresses the lack of research on automatic domain stop word selection for patent text topic modeling. Based on patent literature characteristics, we introduce auxiliary patent sets and propose category entropy to measure word distribution for automatic domain stop word selection, improving patent topic model differentiation and quality. Experiments demonstrate that the proposed category entropy method better captures word distribution characteristics than traditional TF/DF methods, building better patent topic models with increased inter-topic distances and distinguishability.

References

[1] YOON B, PARK Y. A text-mining-based patent network: analytical tool for high-technology trend[J]. Journal of high technology management research, 2004, 15(1): 37-50.

[2] Guo Weiqiang, Dai Tian, Wen Guihua. Patent automatic classification based on domain knowledge[J]. Computer engineering, 2005, 31(23): 52-54.

[3] 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.

[4] Gao Lidan, Xiao Guohua, Zhang Xian, et al. Research on co-occurrence analysis in patent mapping[J]. Modern intelligence, 2009, 29(7): 36-39.

[5] Zhang Jie, Liu Meijia, Zhai Dongsheng. Research on RFID technology topics based on patent co-word analysis[J]. Science and technology management research, 2013, 33(10): 129-132.

[6] 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.

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

[8] 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: IEEE, 2015: 1116-1116.

[9] SUOMINEN A, TOIVANEN H, SEPPÄNEN M. Firms' knowledge profiles: mapping patent data with unsupervised learning[J]. Technological forecasting & social change, 2017, 115(1): 131-142.

[10] Fan Yu, Fu Hongguang, Wen Yi. Patent information clustering technology based on LDA model[J]. Computer applications, 2013, 33(S1): 87-89.

[11] Wang Bo, Liu Shengbo, Ding Kun, et al. Patent content analysis method based on LDA topic model[J]. Science research management, 2015, 36(3): 111-117.

[12] Wu Feifei, Zhang Yaru, Huang Lucheng, et al. Multi-dimensional dynamic evolution analysis of technology topics based on AToT model: a case study of graphene technology[J]. Library and information service, 2017, 61(5): 95-102.

[13] Liao Liefa, Le Fugang. Research on patent technology evolution based on LDA model and classification number[J]. Modern intelligence, 2017, 37(5): 13-18.

[14] Chen Liang, Zhang Jing, Zhang Haichao, et al. Application research of hierarchical topic model in technology evolution analysis[J]. Library and information service, 2017, 61(5): 103-108.

[15] FRAKES W B, BAEZA-YATES R. Information retrieval: data structures and algorithms[M]. Prentice-Hall, Inc., 1992.

[16] SILVA C, RIBEIRO B. The importance of stopword removal on recall values in text categorization[C]//International joint conference on neural networks. Piscataway: IEEE, 2003: 1661-1666.

[17] Guan Qin, Deng Sanhong, Wang Hao. Comparative study of common Chinese stop word lists for text clustering[J]. Modern library and information technology, 2017(3): 72-80.

[18] CROW D, DESANTO J. A hybrid approach to concept extraction and recognition-based matching in the domain of human resources[C]//IEEE inter-

national conference on TOOLS with Artificial Intelligence. Piscataway: IEEE, 2004: 535-541.

[19] SEKI K, MOSTAFA J. An application of text categorization methods to gene ontology annotation[C]//International Conference on Research and Development in Information Retrieval. New York: ACM, 2005: 138-145.

[20] TONG S, LERNER U, SINGHAL A, et al. Locating meaningful stopwords or stop-phrases in keyword-based retrieval systems[EB/OL]. [2018-04-06]. http://www.google.com/patents/US8626787.

[21] WHITE B J. Impact of domain-specific stopword lists on ECommerce website search performance[J]. Journal of strategic e-commerce, 2007, 5(2): 83-101.

[22] LO T W, HE B, OUNIS I. Automatically building a stopword list for an information retrieval system[C]//International conference on computer science and software engineering. Piscataway: IEEE Computer Society, 2008: 718-722.

[23] HAO L, HAO L. Automatic identification of stopwords in Chinese for information retrieval[J]. Journal of digital information management, 2005, 3(1): 3-8.

[24] SINKA M P, CORNE D W. Evolving better stoplists for document clustering and web intelligence[C]//Design and application of hybrid intelligent systems. Amsterdam: IOS Press, 2003: 1016-1025.

[25] JUNGIEWICZ M, ŁOPUSZYŃSKI M. Unsupervised keyword extraction from Polish legal texts[C]//International conference on natural language processing. Berlin: Springer International Publishing, 2014: 65-70.

[26] MAKREHCHI M, KAMEL M S. Extracting domain-specific stopwords for text classifiers[J]. Journal of machine learning research, 2017, 21(1): 39-62.

[27] Gu Yijun, Fan Xiaozhong, Wang Jianhua, et al. Automatic selection of Chinese stop word list[J]. Journal of Beijing Institute of Technology, 2005, 25(4): 337-340.

[28] Gong Zheng, Guan Gaowa. Comparative study of Mongolian and English stopwords[J]. Chinese information journal, 2011, 25(4): 35-38.

[29] Zhu Jie, Li Tianrui. Research on Tibetan stopword selection and automatic processing method[J]. Chinese information journal, 2015, 29(2): 125-132.

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

Author Contributions

Yu Yan: Proposed research ideas, designed the study, conducted experiments, and wrote the paper.
Zhao Naixuan: Collected, cleaned, and analyzed data, and revised the paper.

Abstract: [Purpose/significance] Because research on automatic selection of domain-specific stop words in patent text topic model is insufficient, this paper proposes a new method of automatic selection of domain-specific stop words, for patent text topic model analysis, in order to improve the differentiation and modeling quality of patent topic model. [Method/process] In essence, domain-specific stop words are less important words which contain relatively less information, such words are poorly differentiated in different kinds of patent. Therefore, this paper introduced the auxiliary multi-category patent text dataset and measured the distributions of words through the category entropy. Then, according to the category entropy of words. It chose some words that have the maximum category entropy as the domain-specific stopwords. [Result/conclusion] Experimental results show the feasibility and validity of the method proposed in this paper, which can improve the differentiation and quality of topic model for patent text analysis.

Keywords: patent text, topic model, domain-specific stopword, automatic selection

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