# Research on WI-LDA Model for Patent Technology Topic Analysis (Postprint)

**Authors:** Wu Hong, Yi Huifang, Ma Yongxin, Li Chang

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

## Abstract

[Purpose/Significance] Improving the low identifiability, weak interpretability, and ambiguous boundary delineation in existing LDA-based patent technology topic analysis is of great significance for grasping technology hotspots and tracking technology frontiers.

[Method/Process] International Patent Classification (IPC) codes are introduced into LDA-based patent topic analysis as context for technical terms. A WI-LDA model is constructed by training on the WI (Word-IPC) structure of <word/phrase, classification code> binary tuples to identify and analyze topics in patent documents.

[Results/Conclusion] Empirical research in China's graphene field and comparative studies with traditional LDA models demonstrate that the WI-LDA model possesses strong generalization capability. It effectively reduces topic identification difficulty, enhances topic interpretability, and produces clearer text topic partitioning in patent technology topic analysis.

## Full Text

## Preamble

**Vol. 62 No. 17 September 2018**

**WI?LDA: A Model for Patent Technology Topic Analysis**

Wu Hong, Yi Huifang, Ma Yongxin, Li Chang

Institute of Scientific and Technical Information, Shandong University of Technology, Zibo 255049

## Abstract

[**Purpose/Significance**] Improving the existing problems of low recognition, weak interpretability, and fuzzy boundary division in LDA-based patent technology topic analysis is of great significance for grasping technology hotspots and tracking technological frontiers. [**Method/Process**] This study introduces the International Patent Classification (IPC) into LDA-based patent topic analysis as the context for technical terms, training on the WI (WordIPC) structure of <word/phrase, classification number> binary tuples to construct the WI?LDA model for identifying and analyzing patent document themes. [**Result/Conclusion**] Through an empirical study of Chinese graphene patents and comparative analysis with traditional LDA models, the WI?LDA model demonstrates strong generalization ability, effectively reducing the difficulty of topic identification, increasing topic interpretability, and making text topic division clearer in patent technology topic analysis.

**Keywords:** WI?LDA; topic model; technical topic in patents; graphene
**Classification Number:** G250
**DOI:** 10.13266/j.issn.0252-3116.2018.17.009

## Introduction

Patent technology topics, as the subject and core of technical content disclosed in patent documents, are highly representative and summarizing [1]. Mining and analyzing these topics provides scientific support for relevant personnel to understand research content in technical fields, grasp technology development opportunities, conduct effective technological innovation, build competitive advantages, and make R&D decisions. Numerous studies have utilized text mining techniques for patent technology topic analysis, among which the LDA (Latent Dirichlet Allocation) model proposed by D.M. Blei et al. [2] is particularly prominent. Representative studies include: Liao Liefa et al. [3] introduced IPC classification numbers to measure technology topic intensity on the basis of LDA modeling, achieving evolutionary research on three aspects: topic intensity, topic content, and technology topic intensity. G.J. Kim et al. [4] used k-means clustering to perform LDA topic extraction on each cluster to describe the main technologies involved. Wu Feifei et al. [6] conducted three-dimensional correlation analysis of technology topics, patentees, and time based on the AToT model to analyze the multi-dimensional dynamic changes of enterprise technology topics. Chen Liang et al. [7] employed the hLDA model to extract hierarchical topics from patent corpora to describe the technical structure hidden in patent texts and revealed temporal changes in topics based on the hierarchical model. In summary, topic models applied in patent technology topic analysis mainly fall into two categories: one directly applies traditional LDA models to corpora composed of patent documents, and the other improves or extends LDA models according to analysis purposes or structural characteristics of patent information. As research on LDA models in patent topic analysis deepens, improvements and extensions of LDA models have gradually become

the research focus, mainly including the following five categories: Integration of temporal information, such as the dynamic topic model DTM [8] that models by time intervals, and the continuous time model TOT [9] that jointly models co-occurring words and document timestamps; Integration of document metadata, such as the LDA institution-topic model [5] that jointly models patent knowledge subjects and objects, and the ICT model [10] that integrates three types of information including patent text, inventors, and patentees; Consideration of complex semantics, such as the bigram topic model BGTM [11] that considers word order, and the N-gram topic model TNG [12] that models with phrases; Consideration of lexical context, such as the LDA model [13] that uses SAO structure as the basic unit to improve the model from the relationship between subjects and objects; Integration of text classification numbers, such as the SSHLDA model [14] and PatentClassificationLDA [15] that combine text indexing information with the patent classification system as a predefined set of technical topics.

However, in terms of the combination of patent information characteristics and topic models, both traditional and extended LDA models still have certain defects in patent technology topic analysis:

Topic models - treat the training corpus as independent words/phrases, ignoring their context. This tends to treat identical keywords from different texts as equivalent, resulting in assimilated topic descriptions [13] and increasing the difficulty of topic identification, leading to low topic recognition. Moreover, isolated words/phrases contain limited semantic information and insufficient representation capacity, making it difficult to clearly express topic concepts and depth. Even increasing the number of descriptive topic words, without an understandable topic context, researchers cannot accurately interpret topic information based on separated words/phrases under topic distributions, especially when topic directions are unclear or ambiguous. This makes topic summarization difficult and results in subjective speculation. Furthermore, text topic division is unclear, particularly when a patent involves multiple topics with very close distribution proportions, where forced division may result in documents belonging to no topic or most topics, which does not match reality.

Topic model replaces noun words/phrases in traditional LDA with SAO structure, increasing the breadth and depth of topic information and improving the above problems to some extent. However, as patent texts are legal writing with strict syntactic structures and obscure language, a single sentence often requires convoluted expression to be comprehensive and leave no loopholes. As a sentence-level structure, SAO extraction efficiency is limited by the development of dependency syntax itself, resulting in an overly sparse "document-SAO matrix" and too few word co-occurrence pairs between documents, missing substantial relevant information and directly affecting the accuracy of SAO-based LDA models.

Topic model mainly combines classification number-indexed document information for topic extraction, using each node in the classification system as a

topic to infer the vocabulary probability distribution corresponding to the topic based on document content classification numbers (or presetting that a document belongs to only one classification number or equal probability extraction of document-classification numbers). Although this method can avoid treating identical keywords from different classification texts as equivalent to some extent and help improve topic recognition, the training unit in the topic extraction process remains unigram (single word) structure, and depicting topic content with word probability distribution still brings interpretation difficulties [15].

To address the problems of low topic recognition, weak interpretability, and fuzzy text topic boundary division, this paper introduces the International Patent Classification (IPC) of the text where technical words reside as their context. Using the <word/phrase, classification number> WI (WordIPC) binary tuple structure for LDA training, the IPC is directly introduced into the topic extraction process to construct the WordIPC?LDA topic model (WI?LDA for short), aiming to reduce the impact of external machine learning and achieve effective identification and analysis of patent technology topics.

## WI?LDA Topic Model

WI?LDA introduces IPC as the context for words/phrases in the corpus, using WI vocabulary as the basic unit for corpus training, and discovers the implicit Topic structure in patent texts through unsupervised learning using a Bayesian probability model. WI vocabulary refers to the binary tuple basic structure formed by combining word/phrase Word with its distributed text technical context IPC. The WI?LDA model identifies topics based on this structured vocabulary, grounded in the principle that patent texts, as documents describing technical solutions, have each technical word distributed in an independent technical environment, while IPC reflects the technology and function that provides context for each word/phrase. Its advantage lies in providing richer and more accurate information for each basic training unit, achieving precise description of each topic and its vocabulary. Taking “material” as an example, in traditional LDA model results, only the meaning of “material” can be seen, with thin effective information, requiring interpretation based on final clustering results and causing great confusion in topic determination. WI?LDA introduces IPC as context into LDA modeling beforehand, transforming the basic structural unit of training vocabulary from unary to binary, further expanding its breadth and depth, and enriching information content. For instance, “material” combined with H01M emphasizes materials used for battery electrodes, while “material” combined with C08L more often refers to materials for preparing composites. Thus, even the same word has different meanings in topic representation, and the model clustering principle no longer relies solely on the latent semantic features of words/phrases but also considers the technical context of words/phrases. Only topic words with similar technical contexts and strong co-occurrence semantic features will maximally belong to the same topic.

The core idea of the WI?LDA topic model is that each WI vocabulary in an

article is generated by "selecting a topic with a certain probability, and this topic selecting a WI vocabulary with a certain probability." The specific model is shown in Figure 1 [Figure 1: see original paper]:

**Figure 1. WI?LDA Topic Model**

Where hollow circles represent latent variables and solid circles represent observable variables (i.e., WI vocabulary). Assuming the patent corpus contains D documents, words/phrases based on individual words and their IPCs from documents are extracted as technical contexts to construct WI vocabulary, ultimately obtaining N WIs and forming a D×N document-WI matrix WI_{D×N}. Each element WI in this matrix represents the occurrence probability of the nth WI in the dth document. The corpus includes K topics Z = {$z_1$, $z_2$, $z_3$, …, z_K}, Φ_z is the multinomial probability distribution of WI vocabulary for the kth topic z_k. For each z_k in Z, the probability of generating different WIs forms a K×N topic-WI probability matrix Φ_{K×N}. Each element Φ_{zn} in this matrix represents the probability of the nth WI in the kth topic, meaning each WI in a document can be considered as generated from a conditional probability distribution p(WI|Φ). Topic generation is determined by p(Z| _d), forming a D×K document-topic matrix _{D×K}, where each element represents the probability of the dth document generating the kth topic z_k. The entire process can be expressed in matrix form as WI_{D×N} = Φ_{K×N} × _{D×K}.

WI?LDA is a probabilistic generative model. The document generation steps are as follows:

a. For each document d D, sample _d ~ Dir($\alpha$) to obtain the multinomial distribution parameters _d for topics on document d;

b. For each topic z K, sample Φ_z ~ Dir($\beta$) to obtain the multinomial distribution parameters Φ_z for WI vocabulary on topic z;

c. For the nth WI_{d,n} in document d, sample the topic z_{d,n} ~ Multi( _d) according to multinomial distribution, then sample the specific WI_{d,n} vocabulary ~ Multi(Φ_z) according to multinomial distribution.

From the above, the joint distribution formula for all variables is:

P(WI_d, Z_d, _d, Φ_z|$\alpha$, $\beta$) = *{n=1}^N p(WI*{d,n}|Φ_{z_{d,n}}) p(Z_{d,n}| _d) p( _d|$\alpha$) p(Φ_z|$\beta$) (1)

In WI?LDA, the WI vocabulary of texts is observable data after construction, while text topics are latent variables. Based on text generation rules and known data, WI?LDA can derive text topic distribution and each topic's WI distribution Φ through probability inference. Common inference methods include EM (expectation maximization), variational Bayesian, and Gibbs sampling [16]. Among them, Gibbs is an MCMC (Markov chain Monte Carlo) sampling method

that is easier to implement than EM, with lower computational complexity and comparable speed and results [17], and has been widely applied in probabilistic generative models. Therefore, this paper adopts the Gibbs sampling method from reference [18] for parameter estimation.

## Empirical Study

Graphene is a two-dimensional crystal composed of carbon atoms only one atomic layer thick, with excellent electrical, mechanical, optical, chemical, and thermal properties, as well as high specific surface area. It is considered a new potential material for the "post-silicon era" with broad application prospects [6]. As a representative new material that can change five major industries in China's future [6], graphene has attracted significant attention in China. Since 2008, patent applications have been growing rapidly, and China now ranks first in the world, far exceeding the United States and other Asian and European countries [19]. In this context, analyzing Chinese graphene patent topics can clarify China's graphene technology distribution, which is important for maintaining national competitive advantages and sustainable development, and also provides a good reference for the entire graphene industry.

### 3.1 Data Collection

This study selected Chinese graphene patents from the Derwent Innovation Index (DII) database as the data sample. The processed patent titles and abstracts in the database cover main content, methods, application fields, novelty, and other aspects of original patents, with descriptions tending toward standardization and interpretability, ensuring meaningful patent topic word extraction while improving technical word extraction efficiency [20]. Based on literature review and expert knowledge, "graphene" was determined as the keyword, with the search formula "TI=(graphene or graphenes) and PN=(CN*)" applied for the period 2008-2015 (search date: July 2017). After processing and screening, 9,021 patents were obtained, from which patent numbers, titles, abstracts, international classification numbers, application dates, and other relevant information were extracted to complete data collection.

### 3.2 Data Preprocessing

First, patent abstracts were segmented and part-of-speech tagged to extract nouns and noun phrases from patents. Simultaneously, noise reduction was performed, mainly including: singular/plural unification, synonym merging, hyphen "-" handling, full name and abbreviation processing, and removal of stop words (e.g., a, for), patent description words (e.g., comprise, involves), academic vocabulary (e.g., advantage, method), and some experiment-specific high-frequency but meaningless words (e.g., degree, amount) to ensure result objectivity and scientificity.

When extracting IPC classification numbers, since different IPC levels may produce different clustering effects, this paper separately extracted the main IPC class, subclass, and group for small-scale text experiments. Results showed that topic word division based on class was too coarse with unclear clustering effects, while matrix formation based on group was too sparse and unsuitable for topic training. Subclass-based topic words could ensure clear topic distinction while keeping matrix size manageable. Therefore, the main classification subclass was selected as the limitation for topic word language context. For simplicity and intuitive result presentation, IPC subclasses involved in the graphene field were encoded, with the main technical field IPC subclass-code distribution shown in Table 1 .

After obtaining text IPC subclasses, technical nouns and noun phrases were extracted from patent texts using the NLTK toolkit [21] in Python. IPC subclasses were assigned to nouns/noun phrases in their patent texts to form <noun/noun phrase, main IPC subclass> binary tuples, thereby constructing WI vocabulary and achieving the transformation of each patent document into a feature vector composed of multiple WI binary structures to form domain WI vocabulary training.

### 3.3 Results Analysis

**3.3.1 Model Generalization Ability Analysis** This paper uses the LDA model as a baseline to comparatively analyze the modeling effect of the WI?LDA model. Data preprocessing before modeling is basically similar to the above and will not be repeated. Model generalization ability is evaluated using the standard evaluation function of language models—perplexity [22]. This metric measures corpus modeling capability: the smaller the perplexity, the stronger the model's generalization ability. The expression is:

$$\text{Perplexity}(D\_\{test\}) = \exp(- \{d=1\}\hat{}M \log p(w\_d) / \{d=1\}\hat{}M\ N\_d) \quad (2)$$

The comparison mainly evaluates the generalization ability of the WI?LDA model from two aspects: First, analyzing perplexity changes with increasing topic numbers, mainly by continuously increasing topic numbers to judge the model's predictive ability for uncertain data. Second, analyzing perplexity changes with increasing observed vocabulary, mainly by randomly extracting N vocabularies from a training document using a trained model, continuously adjusting N size, retraining the model, and observing document perplexity changes.

Experimental parameters are set as follows: alpha (document-topic associations) = 5, beta (topic-term associations) = 0.1, iteration times = 5000. The perplexity change with topic numbers is shown in Figure 2 [Figure 2: see original paper]. Under the same topic number, the initial WI?LDA has higher perplexity and weaker generalization ability than the traditional LDA model. However, the model converges quickly, rapidly dropping far below the traditional LDA model's perplexity value, indicating significantly better generalization ability. When topic numbers exceed 40, the WI?LDA model's perplexity stabilizes earliest,

while the LDA model continues to decline, showing better convergence speed and effect of WI?LDA.

When analyzing perplexity changes with observed vocabulary increase, the topic model with 40 topics was used as the initial training model. Statistics on observable vocabulary in single training documents showed a maximum vocabulary count of 156, so N was set in the interval [1:156]. To ensure result stability, document perplexity was calculated for all training documents, using the mean as the perplexity under that N value. The final perplexity curve with N values is shown in Figure 3 [Figure 3: see original paper]. When initial N is small, the perplexity values of both models are similar with no obvious difference in vocabulary topic distribution effects. As N increases, the WI?LDA model's perplexity value is lower than LDA's under the same observed data, indicating superior vocabulary topic assignment.

**3.3.2 Patent Technology Topic Effect Analysis** To analyze the WI?LDA model's effect on technology topic analysis, it is compared with traditional LDA technology topic analysis. Figure 2 shows that when topic numbers are between (5, 10), both models have equal perplexity. To compare the two models' topic analysis effects, topic models were defined in this range. Research found that when both have 8 topics, technical nouns in each topic have good differentiation and relatively high probability [6], making results more representative. Therefore, the graphene field was divided into 8 main topics. The top 10 topic words by probability in each topic were selected to determine technical topic content. Topic distributions under both models are shown in Table 2 .

Comparison of topic results reveals that WI?LDA model results have obvious topic contexts, enabling rough understanding of domain topic distribution and quickly grasping topic content and technology direction. Secondly, from a local relationship perspective, in LDA model results based on single words, some topic words have significant overlap. For example, words like "substrate" and "electrode" appear in more than half of the topics, adversely affecting topic classification and interpretation. The WI?LDA topic model avoids such problems, enabling deeper understanding of topic word connotations. For instance, "material200" under H01M refers to materials for battery electrodes, while "material98" under C08L more often represents materials for preparing composites.

In terms of topic interpretability, the WI?LDA model outperforms single-word LDA. Taking Topic 5 content under both methods as an example, traditional LDA results contain both composite-related words (nano, composite) and battery electrode-related words (battery, lithium, ion, cathode), creating ambiguous scenarios: the topic might be composites with battery electrode as one application direction; battery electrode might be the main topic content with composite used to modify electrodes; or it might be defined as graphene composites and their battery applications. Without reading patent texts, forced definition would affect result objectivity. In contrast, WI?LDA's Topic 5, although containing similar words, defines them under the H01M premise, ef-

fectively avoiding ambiguity and enabling clearer category labeling: graphene application in battery electrodes.

Regarding text topic division, since WI?LDA introduces technical word context, only texts with the same context and belonging to the topic correspond to higher text-topic probabilities. Compared with traditional LDA's relatively average text-topic distribution probabilities, text-topic probability distances should be larger and division levels clearer. To effectively evaluate text-topic probability distinction under both models, this paper defines a text-topic probability average distance metric:

Dis = ( _{i=1}^N |P_i - (1/K)|) / N (3)

Where Dis represents the average distance of text-topic probabilities under the model, indicating probability value differences. N is text quantity, K is topic number, and P_i is the probability of each topic belonging to the text. The formula shows that larger Dis values indicate clearer text-topic division and better topic model effects, while smaller values indicate fuzzier division.

Calculations show that traditional LDA's text-topic probability average distance Dis is 0.0117, while WI?LDA's Dis is 0.0224, about twice that of traditional LDA, demonstrating WI?LDA's superior and more significant text-topic division effect.

## Conclusion

Addressing the deficiencies of low topic recognition, weak interpretability, and fuzzy text division in most current topic models for patent technology topic analysis, this paper proposes a WI vocabulary-based LDA topic model according to patent text characteristics by introducing IPC as the context for technical words, using <word/phrase, classification number> WI (WordIPC) binary tuple structure to identify topic content. The case study proves the effectiveness of the WI?LDA model. Compared with traditional word-based LDA topic models, this model has stronger generalization ability, reduces difficulty in identifying assimilated topics, increases topic readability and interpretability, improves clustering effects, facilitates clear topic direction for subsequent analysis and decision-making, and enables clearer text-topic division.

This paper also notes potential limitations, such as the matrix size expansion and model space dimension explosion caused by introducing IPC. How to better balance these issues while improving clustering effects remains a direction for future improvement.

## References

[1] Hu Apei, Zhang Jing, Lei Xiaoping, et al. A Review of Patent Technology Topic Analysis Based on Text Mining [J]. Journal of Intelligence, 2013, 32(12): 88-92.

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3(4/5): 993-1022.
[3] Liao Liefa, Le Fugang. Research on Patent Technology Evolution Based on LDA Model and Classification Number [J]. Modern Information, 2017, 37(5): 13-18.
[4] Kim G J, Sang S P, Jang D S. Technology Forecasting Using Topic-Based Patent Analysis [J]. Journal of Scientific & Industrial Research, 2015, 74(5): 265-270.
[5] 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.
[6] Wu Feifei, Zhang Yaru, Huang Lucheng, et al. Multi-Dimensional Dynamic Evolution Analysis of Technology Topics Based on AToT Model—Taking Graphene Technology as an Example [J]. Library and Information Service, 2017, 61(5): 95-102.
[7] 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.
[8] Blei D M, Lafferty J D. Dynamic Topic Models [C]//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh: ACM, 2006: 113-120.
[9] Wang X, McCallum A. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]//Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2006: 424-433.
[10] Tang J, Wang B, Yang Y, et al. PatentMiner: Topic-Driven Patent Analysis and Mining [C]//Proceedings of the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing: ACM, 2012: 1366-1374.
[11] Wallach H M. Topic Modeling: Beyond Bag-of-Words [C]//Proceedings of the 23rd 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 Seventh IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2007: 697-702.
[13] Yang Chao, Zhu Donghua, Wang Xuefeng, et al. Patent Technology Topic Analysis: LDA Topic Model Method Based on SAO Structure [J]. Library and Information Service, 2017, 61(3): 86-92.
[14] Mao X L, Ming Z Y, Chua T S, et al. SSHLDA: A Semi-Supervised Hierarchical Topic Model [C]//2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2012: 800-809.
[15] Chen Liang. PatentClassificationLDA Model for Patent Analysis [J]. Library and Information Service, 2014, 58(5): 58-63.

[16] Zhang Chenyi, Sun Jianling, Ding Yiqun. Microblog Topic Mining Based on MB-LDA Model [J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.
[17] Tang Xiaobo, Xiang Kun. Hotspot Mining Based on LDA Model and Microblog Popularity [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(8): 864-874.
[18] Sugimoto C R, Li D, Russell T G, et al. The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science Dissertations Using Latent Dirichlet Allocation [J]. Journal of the Association for Information Science & Technology, 2011, 62(1): 185-204.
[19] Zhao Zhenxia, Chen Hong. Analysis of Current Status and Trends of Graphene Technology Development in China—Based on Patent Data [J]. China Textile Leader, 2016(9): 40-43.
[20] 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.
[21] Liu Xu. Application of Python Natural Language Processing Toolkit in Corpus Research [J]. Journal of Kunming Metallurgy College, 2015, 31(5): 65-69.
[22] Li Baoli, Yang Xing. Research Topic Evolution Analysis Based on LDA Model and Topic Filtering [J]. Journal of Chinese Computer Systems, 2012, 33(12): 2738-2743.

## Author Contributions

Wu Hong: Topic selection and design, guiding paper writing, revision, and finalization;
Yi Huifang: Responsible for framework design, completing data collection, processing, and paper writing;
Ma Yongxin: Paper framework design and content revision;
Li Chang: Paper data cleaning.

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