Methods for Determining the Optimal Number of Topics in LDA Topic Models for Science and Technology Information Analysis (Postprint)

Authors: Guan Peng, Wang Yuefen

Date: 2017-11-08T00:00:00+00:00

Abstract

[Objective] To effectively determine the optimal number of topics for the LDA topic model in scientific and technological intelligence analysis. [Method] By utilizing topic similarity to measure differences between latent topics and combining it with perplexity, a method is proposed to determine the optimal number of topics for LDA, which considers both the effectiveness of topic extraction and the model' s generalization capability to new documents. [Results] Using scientific and technological literature in the domestic new energy field as the dataset, empirical results demonstrate that the proposed method for determining the optimal number of LDA topics achieves higher precision (91.67%), F-score (86.27%) for topic extraction, and recommendation accuracy (71.25%) for scientific and technological literature compared with using perplexity alone. [Limitations] The validation of the proposed method has not been conducted on other types of datasets, such as Weibo short texts, XML documents, etc. [Conclusion] The proposed method can effectively extract highly distinguishable topics from scientific and technological literature datasets and improve the effectiveness of scientific and technological literature recommendation.

Full Text

Identifying Optimal Topic Numbers for LDA Models in Scientific and Technical Information Analysis

Guan Peng[1],[2], **Wang Yuefen**[1]
[1](School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094, China)
[2](College of Applied Mathematics, Chaohu University, Hefei 238000, China)

Abstract

This study aims to effectively determine the optimal number of topics for Latent Dirichlet Allocation (LDA) models in scientific and technical information analysis. The study employs topic similarity to measure differences among latent topics and proposes a novel method for determining the optimal topic number by combining perplexity with topic variance. This approach simultaneously considers both topic extraction effectiveness and model generalization capability for new documents. Using a dataset of Chinese scientific literature in the new energy field, empirical results demonstrate that the proposed method achieves higher precision (91.67%), F-score (86.27%), and scientific literature recommendation accuracy (71.25%) compared to using perplexity alone. However, the validation was not conducted on other types of datasets, such as microblog short texts or XML documents. The proposed method can effectively extract highly distinguishable topics from scientific literature datasets and improve the performance of scientific literature recommendation.

Keywords: LDA; Topic Model; Similarity; Perplexity; Scientific and Technical Information Analysis
Classification Number: G202

Latent Dirichlet Allocation (LDA) represents a typical statistical language model that has gained widespread application in information analysis, knowledge services, and knowledge discovery in recent years. Its primary applications include scientific literature mining [2-4], detection of research hotspots and emerging topics [5-7], research topic evolution analysis [8-10], and academic evaluation [11]. LDA' s popularity in information science stems from its suitability for modeling massive heterogeneous text data and its ability to substantially reduce text dimensionality, thereby avoiding the curse of dimensionality [12]. Numerous empirical studies in scientific and technical information analysis have demonstrated LDA' s reliability and effectiveness, yet several unresolved issues remain. Compared with general text mining tasks, scientific and technical information analysis imposes higher requirements on LDA in two main aspects:

First, in general text mining tasks such as text clustering, classification, and automatic summarization [13-16], LDA typically serves as an intermediate dimensionality reduction step without requiring explicit presentation of topics. However, in scientific and technical information analysis tasks such as research topic discovery and evolution analysis, LDA must present and analyze the extracted topics, where the quality of topic extraction directly impacts the effectiveness of topic identification and evolution analysis.

Second, determining the optimal number of topics receives greater emphasis in information analysis applications. The inability to determine the optimal topic number is widely recognized as LDA' s primary limitation [17], yet this issue has not received sufficient attention in current applications of LDA for scientific

and technical information analysis.

Extensive empirical research confirms that LDA' s topic extraction effectiveness is directly related to the number of latent topics (K value), with results being highly sensitive to this parameter. Consequently, scholars have conducted related research to determine the optimal topic number through various methods, with three commonly used approaches:

(1) **Perplexity-based method**: Blei et al. employed perplexity as a standard for evaluating model quality, selecting the model with minimal perplexity to determine the optimal topic number [1]. While perplexity can identify optimal predictive capability, the resulting topic number tends to be excessively large, leading to high similarity among extracted topics and poor topic distinguishability, which reduces efficiency in scientific and technical information analysis.

(2) **Non-parametric approaches**: Hierarchical Dirichlet Processes (HDP) represent a typical non-parametric Bayesian model that can automatically learn the most appropriate topic number K from document collections [18]. HDP addresses LDA' s topic number selection problem through the non-parametric characteristics of the Dirichlet process, with experiments confirming that HDP' s optimal topic number aligns with perplexity-based selection. However, this method requires building both an HDP model and an LDA model for the same collection, suffers from high computational complexity, and exhibits low efficiency in scientific and technical information analysis applications.

(3) **Bayesian model selection**: Griffiths et al. proposed using Bayesian models to determine the optimal topic number [19]. This method relies on the Gibbs sampling process, exhibits high computational complexity, and can only determine topic numbers without characterizing model generalization capability.

Additionally, scholars have explored the relationship between topic similarity and optimal topic number. Arun et al. conceptualized LDA as a matrix factorization process where topic extraction effectiveness depends on K value selection, experimentally demonstrating that KL divergence measures of topic similarity yield smaller values when topic numbers approach the optimal value and larger values when deviating from it [20]. Cao et al. theoretically and experimentally established the relationship between optimal topic number and topic similarity, using this as a constraint to unify optimal K value selection with LDA parameter estimation within a single framework [21]. Their experiments proved that the optimal K value relates to both the quantity of texts in the collection and their inter-document correlation. Comprehensive analysis reveals that existing methods for determining LDA' s optimal topic number suffer from either high model complexity or low topic distinguishability. Therefore, this paper proposes a new method for determining topic number based on topic similarity.

Methodology: A Perplexity-Topic Similarity Combined Approach

Topic distinguishability correlates closely with inter-topic similarity: smaller similarity yields greater distinguishability. Balancing model generalization capability with topic extraction effectiveness, this paper proposes a combined metric called **Perplexity-Var** to determine the optimal topic number.

Perplexity

In probabilistic language models, perplexity evaluates language model quality based on the principle that better models assign higher probabilities to test sets [22]. Smaller perplexity indicates better predictive performance for new texts, and perplexity generally decreases as the number of latent topics increases.

In LDA topic models, perplexity is calculated as follows [1]:

$$\text{Perplexity}(D) = \exp\left(-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right)$$

where D represents the test set in the corpus containing M documents, N_d denotes the number of words in each document d, w_d represents the words in document d, and $p(w_d)$ is the probability of generating words w_d in the document.

Perplexity-Var

Common methods for calculating topic similarity include Kullback-Leibler (KL) divergence [23] and Jensen-Shannon (JS) divergence [24]. Since KL divergence does not satisfy symmetry or the triangle inequality, this study adopts JS divergence to measure inter-topic similarity.

By introducing the concept of variance from random variables into the latent topic space, we can measure the overall dispersion of the topic space. Topic variance, denoted as $\text{Var}(T)$, represents the average squared distance between each topic and their mean, measuring the deviation of topics from their mean and assessing the overall dispersion and stability of the latent topic space. The calculation proceeds as follows:

1. Compute the mean $\bar{\phi}$ of the topic-word probability distributions ϕ
2. Calculate topic variance using JS divergence:

$$\text{Var}(T) = \frac{1}{K}\sum_{i=1}^{K} D_{JS}(T_i \parallel \bar{\phi})$$

where T represents topics extracted by LDA, K is the topic number, and D_{JS} denotes JS divergence. $\text{Var}(T)$ measures stability and dispersion among topics:

larger variance indicates greater inter-topic differences and better distinguishability, yielding more stable topic structures.

As previously noted, using LDA for scientific literature topic extraction often results in excessively large topic numbers when using perplexity alone, leading to high topic similarity and poor distinguishability. While perplexity reflects model predictive capability, pursuing predictive performance exclusively inevitably yields overly large topic numbers. Combining both metrics effectively addresses the distinguishability problem.

The Perplexity-Var metric is calculated as:

$$\text{Perplexity-Var}(D) = \frac{\text{Perplexity}(D)}{\text{Var}(T)}$$

where Perplexity(D) is the test set perplexity and Var(T) is the test set topic variance.

Interpretation: First, considering model generalization capability, smaller perplexity indicates better LDA generalization. Second, considering topic extraction effectiveness, the optimal LDA model corresponds to minimal average similarity among topic structures [21], which translates to larger topic structure variance. Therefore, larger topic variance indicates better topic extraction effectiveness and smaller Perplexity-Var values. In summary, the LDA model is optimal when Perplexity-Var is minimized.

Experiments

Data and Preprocessing

Data Collection: Experimental data were retrieved from CNKI (China National Knowledge Infrastructure). After deduplication and removal of incomplete records, 1,018 documents from China' s new energy field (1994-2000) were obtained, including title, author, institution, abstract, and keywords (full text not included). Ten percent of the corpus served as the test set for model evaluation, with the remainder used for LDA model training.

Through analysis of titles, keywords, and abstracts from 1,018 documents, we identified 27 valid topics containing 955 documents, with 63 documents having unclear topics, as shown in .

Preprocessing: 1. Domain dictionary extraction and segmentation: Python scripts extracted keywords from 1,018 documents, calculated term frequency, and built a domain dictionary. The jieba segmentation package [25] segmented abstracts using the domain dictionary as a custom user dictionary to improve segmentation quality. 2. LDA implementation: Topic extraction used the gensim machine learning package [26] for Python, with Perplexity-Var calculation and document similarity computation also implemented in Python.

The experimental environment consisted of a Windows 7 Ultimate system with an Intel(R) Core(TM) i5-4570 CPU @ 3.2GHz and 4GB RAM.

Evaluation Metrics and Comparative Analysis

Among the three methods for determining LDA' s optimal topic number, the HDP-based approach suffers from high computational complexity, while the Bayesian method based on Gibbs sampling cannot characterize new document prediction capability. Therefore, this study selected the popular perplexity-based method as the baseline for comparison.

Evaluation was conducted from two perspectives: scientific literature topic extraction effectiveness and similarity-based recommendation performance.

(1) Topic Extraction Effectiveness: Measured using precision (P), recall (R), and F-score. Precision evaluates the proportion of correctly extracted topics among all valid topics extracted by LDA. Recall assesses the proportion of correctly extracted topics relative to domain research topics identified by expert judgment. F-score is the harmonic mean of precision and recall:

$$P = \frac{\text{correct}}{\text{extract}}, \quad R = \frac{\text{correct}}{\text{standard}}, \quad F = \frac{2PR}{P+R}$$

where *extract* is the number of valid topics extracted by LDA, *correct* is the number of correctly extracted topics (those contained within expert-judged domain topics), and *standard* is the number of domain topics identified through literature review and expert evaluation.

(2) Document Similarity Recommendation: High-quality scientific information services must address user needs for finding literature similar to their readings. Recommendation quality directly correlates with extracted topic quality, making it a crucial evaluation criterion.

After LDA topic extraction on the training corpus, documents are represented in a topic vector space with substantially reduced dimensionality compared to word vector space. For new documents in the test set, the trained LDA model extracts topics and maps documents to the topic space, where JS divergence measures similarity between new and training documents to complete recommendation.

The similarity-based recommendation process: 1. Train LDA model on the training corpus with topic number K 2. Extract topics for test set documents using the trained model 3. Calculate JS divergence between each test document and all training documents; smaller JS divergence indicates greater similarity. Rank all documents by similarity, with top-ranked documents being most similar.

Test set documents (102 documents) were manually annotated to identify the top 10 most relevant documents from the training set. For each test document, the top 10 recommended documents were retrieved, and recommendation precision was calculated as:

$$\text{Recommend Precision} = \frac{\sum_{i=1}^{M} |T_i \cap R_i|}{10 \times M}$$

where M is the number of test documents, T_i is the manually annotated top-10 relevant document set, and R_i is the algorithm' s top-10 recommendation set.

Experimental Results and Analysis

(1) Optimal Topic Number Determination: The experiment varied topic number K from 10 to 200 in increments of 10, performing LDA extraction and calculating both Perplexity and Perplexity-Var metrics on the test set.

Perplexity Metric: As shown in [Figure 1: see original paper], perplexity reached its minimum at K=70, identifying 70 as the optimal topic number.

Perplexity-Var Metric: Topic variance across different K values was calculated using JS divergence, as shown in [Figure 2: see original paper]. Variance decreases as topic number increases because larger numbers introduce interfering and semantically redundant topics, increasing inter-topic similarity and reducing topic structure stability.

Applying the Perplexity-Var metric, [Figure 3: see original paper] shows the minimum occurs at K=30, identifying 30 as the optimal topic number.

Comparing both metrics, the perplexity-based result (K=70) deviates significantly from the human-judged topic number (27), while the Perplexity-Var result (K=30) closely aligns with expert evaluation.

(2) Comparative Performance Analysis

Topic Extraction Effectiveness: Using K=70 (perplexity) and K=30 (Perplexity-Var), LDA extracted topics from the new energy literature dataset. Partial results are shown in and (displaying top 10 topics, probability values omitted).

Topic meaning emerges from the comprehensive semantics of topic terms. Comparison with human-judged topics () reveals that the Perplexity-Var model correctly extracted 22 topics with 6 interfering topics among 30 total topics, while the perplexity model correctly extracted 23 topics but included 29 interfering topics among 70 total topics. Performance comparison is shown in :

demonstrates that the perplexity-based method extracts many valid topics, but most are redundant with numerous interfering topics, resulting in lower precision and F-score. The Perplexity-Var method produces fewer interfering topics with higher overall performance. Scientific literature mining requires both accuracy and efficiency; excessive interfering topics severely impact mining effectiveness.

Document Similarity Recommendation: The training set was processed through LDA to obtain the topic space, with test documents represented as vectors in this space. Similar documents were recommended using the proposed

method, retrieving top-10 recommendations. shows the similarity recommendation precision for both metrics.

The Perplexity-Var metric achieves higher recommendation precision because it considers both predictive capability and topic similarity, enhancing inter-topic differences and topic distinguishability. When documents are mapped to the topic space, topics can accurately represent document semantic information. and illustrate recommendation results for sample documents on tidal power and wind power topics, demonstrating that K=30 produces more relevant recommendations than K=70, with more semantically similar documents ranked higher.

Discussion and Conclusion

In the big data era, increasing demand for intelligent information analysis requires algorithms capable of processing massive text data. This paper analyzed LDA' s characteristics and identified key differences between information analysis and general text mining applications, emphasizing that topic extraction effectiveness and topic number determination require greater attention in information analysis work.

By combining topic similarity with perplexity, this study proposed a method for determining optimal topic numbers, empirically demonstrating its effectiveness in scientific literature knowledge mining. The method helps analysts extract salient topics from massive literature and improves similarity-based recommendation performance.

Limitations include validation only on scientific literature datasets without testing on other data types (e.g., microblog posts, XML documents). Additionally, evaluation was limited to topic extraction effectiveness and recommendation performance; further validation across broader dimensions is needed to establish general effectiveness. Expanding validation scope and evaluation metrics represents future work.

References

[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] Wang Ping. Literature Knowledge Mining Based on Probabilistic Topic Model [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(6): 583-590.

[3] Hassan S U, Haddawy P. Analyzing Knowledge Flows of Scientific Literature Through Semantic Links: A Case Study in the Field of Energy [J]. Scientometrics, 2015, 103(1): 1-23.

[4] Liang H, Fang L. Topic Discovery and Trend Analysis in Scientific Literature Based on Topic Model [J]. Journal of Chinese Information Processing, 2012, 26(2): 109-115.

[5] Fan Yunman, Ma Jianxia. Detection of Emerging Topics Based on LDA and Feature Analysis of Emerging Topics [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(7): 698-711.

[6] He Q, Chen B, Pei J, et al. Detecting Topic Evolution in Scientific Literature: How Can Citations Help? [C]. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, 2009: 957-966.

[7] AlSumait L, Barbará D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking [C]. In: Proceedings of the 8th IEEE International Conference on Data Mining. 2008.

[8] Liu Tong, Yang Guancan, Jiang Jiya, et al. Research on the Evolution and Dynamic Analysis of Multi-relation Integrated Patent Network: A Case Study on Lithium-ion Battery [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(12): 1288-1301.

[9] He Liang, Li Fang. Topic Evolution in Scientific Literature [J]. New Technology of Library and Information Service, 2012(4): 61-67.

[10] Wu Q Q, Zhang C D, Hong Q Q, et al. Topic Evolution Based on LDA and HMM and Its Application in Stem Cell Research [J]. Journal of Information Science, 2014, 40(5): 611-620.

[11] Gerrish S, Blei D M. A Language-based Approach to Measuring Scholarly Impact [C]. In: Proceedings of the 27th International Conference on Machine Learning. 2010.

[12] Dhillon I S, Modha D S. Concept Decompositions for Large Sparse Text Data Using Clustering [J]. Machine Learning, 2001, 42(1-2): 143-175.

[13] Wang Lidong, Wei Baogang, Yuan Jie. Document Clustering Based on Probabilistic Topic Model [J]. Acta Electronica Sinica, 2012, 40(11): 2346-2350.

[14] Lee H, Kihm J, Choo J, et al. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling [J]. Computer Graphics Forum, 2012, 31(3): 1155-1164.

[15] Kabán A, Girolami M A. A Dynamic Probabilistic Model to Visualise Topic Evolution in Text Streams [J]. Journal of Intelligent Information Systems, 2002, 18(2-3): 107-125.

[16] Chua F C T, Lauw H W, Lim E P. Generative Models for Item Adoptions Using Social Correlation [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(9): 2036-2048.

[17] Zhang Han, Xu Shuo, Qiao Xiaodong, et al. Review on Topic Models Integrating Intra- and Extra-Features of Scientific and Technical Literature [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(10): 1108-1120.

[18] Teh Y, Jordan M, Beal M, et al. Hierarchical Dirichlet Processes [J]. Journal of the American Statistical Association, 2007, 101(476): 1566-1581.

[19] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(S1): 5228-5235.

[20] Arun R, Suresh V, Veni Madhavan C E, et al. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations [A]. //Advances in Knowledge Discovery and Data Mining [M]. Springer Berlin Heidelberg, 2010.

[21] Cao Juan, Zhang Yongdong, Li Jintao, et al. A Method of Adaptively Selecting Best LDA Model Based on Density [J]. Chinese Journal of Computers, 2008, 31(10): 1780-1787.

[22] Grossman D A. Information Retrieval: Algorithms and Heuristics [M]. Springer Science & Business Media, 2004.

[23] Duda R O, Hart P E, Stork D G. Pattern Classification [M]. John Wiley & Sons, 2012.

[24] Lin J. Divergence Measures Based on Shannon Entropy [J]. IEEE Transactions on Information Theory, 1991, 37(1): 145-151.

[25] Sun J Y. jieba0.37 [EB/OL]. [2015-10-08]. https://pypi.python.org/pypi/jieba/.

[26] Rehurek R. gensim 0.10.2 [EB/OL]. [2014-12-11]. https://pypi.python.org/pypi/gensim.

Author Contributions

Guan Peng: Conceived the research design, conducted experiments, drafted and revised the manuscript.
Wang Yuefen: Expanded research ideas, reviewed the manuscript, provided revision suggestions.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the journal's online version at http://www.infotech.ac.cn.

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