Postprint: Identification Methods for Research Hotspots in Multi-source Data Environments

Authors: Qiu Huilin, Shao Bo

Date: 2023-04-01T16:15:50+00:00

Abstract

[目的/意义] In scientific research, identifying and mining research hotspots from scientific and technical literature from diverse sources is of guiding significance for research activities. This study aims to rapidly and accurately identify hotspot topics embedded in multi-source texts through the proposed model methodology, thereby providing supportive services for scientific research innovation.

[方法/过程] We propose a method for identifying research hotspots in multi-source texts based on the LDA2vec model and construct a model specifically for research hotspot identification. This method integrates the advantage of the LDA topic model in mining implicit semantics with the advantage of the Word2Vec word vector model in capturing contextual relationships. Taking scientific literature in the field of machine learning as an example, we verify the feasibility and effectiveness of LDA2vec's application in this domain using two metrics—model perplexity and topic coherence—and compare it with the topic extraction performance of LDA.

[结果/结论] Experimental results demonstrate that the proposed method is feasible for identifying and mining research hotspots when dealing with multi-source data, and achieves a certain degree of performance improvement. It compensates for the shortcomings of using single data sources for topic analysis and enriches the practical application of multi-source data fusion.

Full Text

Preamble

Abstract: [Purpose/Significance] In scientific research, identifying and mining research hotspots from different sources of scientific and technical literature is of guiding significance for conducting research work. This study aims to rapidly and accurately identify hot topics embedded in multi-source texts through the

proposed model and method, providing support services for scientific research innovation. [Method/Process] This paper proposes a method for identifying research hotspots in multi-source texts based on the LDA2vec model and constructs a corresponding model. This method combines the advantages of the LDA topic model in mining implicit semantics and the Word2Vec word vector model in capturing contextual relationships. Taking scientific literature in the machine learning field as an example, the feasibility and effectiveness of LDA2vec in this domain were validated using two metrics—model perplexity and topic coherence—and compared with the topic extraction effects of LDA. [Result/Conclusion] Experimental results demonstrate that the proposed method is feasible for identifying and mining research hotspots under multi-source data conditions, with a certain degree of effectiveness improvement. It compensates for the shortcomings of using single data sources for topic analysis and enriches the practical application of multi-source data fusion.

Keywords: topic model; LDA2vec; research hotspot; LDA; Word2vec; multi-source data fusion
Classification Number: G251.2
DOI: 10.13266/j.issn.0252-3116.2020.05.009

The trend of explosive information growth is intensifying with technological and temporal developments. When retrieving and collecting information on the Internet, besides effective information, one is also disturbed by large amounts of useless and irrelevant information. The same situation exists in scientific research work. Reading and studying existing research results and journal papers within a discipline is a primary means for researchers to quickly grasp the current state of research in the field and form a more comprehensive understanding of the discipline. Therefore, to timely grasp the research status and follow major research hotspots and directions, hotspot identification and mining is an effective and feasible approach.

However, most current research on research hotspot identification primarily focuses on single-source data. When facing multi-source data, where texts from different sources are heterogeneous and likely lack citation relationships, traditional bibliometric analysis methods and keyword/thesaurus-based analytical methods within library and information science cannot effectively produce results. To avoid the relatively macroscopic and rough results brought by traditional methods and to address the issues of insufficient comprehensiveness, objectivity, and depth caused by analysis limited to external document features, an increasing number of studies now employ text mining methods. Content-based text mining enables more effective and objective identification and analysis of documents' internal features, improving research granularity and depth. For identifying and analyzing multi-source journal papers, analysis results from single data sources cannot comprehensively reflect the overall research status of the discipline. Combining papers and patents, which respectively reflect basic research and technological innovation achievements, for topic hotspot analysis is more advantageous in terms of information comprehensiveness and accuracy

of scientific structure division compared to single-source literature, greatly benefiting accurate positioning of research priorities, hotspots, and prediction of research trends. Although these two document types differ in structure and textual expression, belonging to heterogeneous literature, their content can be effectively integrated to form new technical information. Understanding the mutual influence and penetration relationship between science and technology, identifying technological opportunities, and discovering potential commercialization opportunities are significantly meaningful. Therefore, the application and optimization of topic models in the field of research hotspot identification are worth exploring.

1 Related Research

1.1 Research Hotspot Identification Methods

Research hotspot discovery in library and information science depends on discovering and reasoning about hidden relationships between different entities, representing an important extension and practice of scientometrics and intelligence analysis. When identifying and analyzing research hotspots, researchers primarily use methods that can be categorized into those based on external document features and those based on internal document features.

Methods based on external document features include citation analysis and knowledge unit analysis. Citation analysis takes citation frequency and patterns between documents as research objects, revealing document attributes through citation patterns—links from one document to another. Beyond direct citation, co-citation analysis and bibliographic coupling are widely applied in research hotspot identification with numerous developments, including document co-citation, word co-citation, topic co-citation, author co-citation, and category co-citation, as well as document coupling, author coupling, keyword coupling, and journal coupling. Knowledge units, as the most basic units constituting knowledge systems, can be narrowly understood as non-decomposable words in scientometric research. Therefore, word frequency-based analysis methods and thesaurus co-occurrence analysis methods are categorized as knowledge unit analysis methods. These methods' main characteristic is analyzing external features of the most basic units (words) in literature, revealing entity relationships within disciplinary structures at a more micro level than citation analysis. While widely applied in research hotspot discovery, some scholars recognize that citation analysis based on citation frequency cannot directly reveal literature content. For instance, Zhu Qingsong and Leng Fuhai argue that topic identification based on citation content analysis better reveals why highly cited papers are cited and aligns with overall paper content.

Methods based on internal document features can be understood as hotspot identification methods based on text content mining. Semantic-level mining to identify document set themes and connotations can solve problems of meaningless or context-deviating results from external feature analysis. For exam-

ple, Yang Chao built a topic model by extracting SAO structures from patent texts, solving issues of unclear patent topic semantics and mismatched problem-solution identification. Ruan Guangce used Doc2Vec methods for vector and similarity calculations to generate hot topic paper collections, then applied topic models and clustering algorithms for topic identification and mining, achieving superior semantic feature recognition. Zhao Yifang introduced paragraph information gain for policy texts, addressing the problem that existing topic models cannot effectively allocate contribution differences of specific feature words to similar policy topics, balancing contribution differences between different topics.

Overall, a relatively mature methodological system has initially formed in the research hotspot identification field, including not only discipline-specific methods like bibliometrics but also methods introduced from other disciplines and emerging technologies. However, three main problems exist: (1) insufficient semantic understanding—traditional scientometrics-based methods primarily count thesaurus terms (frequency, co-occurrence, citation counts) without deep text and semantic research (synonyms, near-synonyms, different expression habits); (2) single data source—using one data source to identify research fronts has limitations and cannot comprehensively represent all scientific research frontier information; and (3) time lag—the process from paper writing and review to publication and citation formation is generally lengthy, causing temporal lag in paper data.

1.2 Topic Model Research Status

A topic consists of a core event or activity and all directly related events and activities. Topic models can analyze literature content and extract topics to obtain hotspot knowledge and development trends within a field.

Literature review shows that since the LDA model itself is the most widely used and relatively successful model with good performance in identifying implicit semantics in large-scale document sets, and since the core of research hotspot identification is mining and reasoning implicit knowledge from large-scale disciplinary literature, topic models applied to research hotspot identification are primarily based on the LDA model. Regarding issues when applying LDA to research hotspot and topic identification, many domestic scholars have optimized it or combined LDA with other models to achieve research objectives, including combinations with ontology, SNA social network analysis, citation analysis, co-word analysis, tags, clustering algorithms, and relevant special indicators.

From the current domestic research situation, the overall direction of applying topic models to research hotspot identification is relatively single, with most improvements based on the LDA model and few explorations of the feasibility and effectiveness of new models and methods. In contrast, other fields like public opinion hotspot identification and microblog hotspot identification have more diverse and rich exploration methods. Many new methods may have better efficiency and superior effects in text semantic understanding and mining, making

them worth exploring for research hotspot identification and mining.

2 Research Hotspot Identification Method Based on LDA2vec

2.1 Model Foundation

2.1.1 LDA Topic Model Topic models are unsupervised machine learning methods that differ from traditional library and information science methods based on external document features. While traditional methods only focus on surface relationships between documents or word frequency, topic models can extract deep semantic relationships between words and documents—so-called "latent topic information"—effectively extracting hidden topics from large-scale document sets and corpora. They have been widely applied in text sentiment classification and information extraction, providing excellent opportunities for in-depth text analysis and research topic mining with broad application prospects and practical significance.

Since the earliest topic model LSI was proposed in 1998, many optimized model algorithms have been developed that can explore implicit semantic structures in document collections through calculation and learning of large numbers of documents, sentences, and words. The core purpose of topic models is text dimensionality reduction. Text dimensionality reduction technology has evolved from TF-IDF matrices, unigram mixture models, and pLSA models to the most classic LDA model, which can be understood as Bayesianizing pLSA. LDA is a three-layer Bayesian network model composed of words, topics, and documents, with the core idea that each document can be viewed as a mixture of various topics, where each document is considered to have a set of topics assigned to it through LDA. LDA obtains word clusters by calculating P(word|topic) and P(topic|document). The two most critical steps are: (1) which topic a word belongs to across all documents, and (2) which topic the document containing the word belongs to.

Overall, LDA has two major advantages: (1) it can handle polysemy or different contexts of the same word because LDA considers the entire document's topic tendency when dividing topics; and (2) it can find words to describe each topic, which better guides comprehensive and profound understanding of a topic's meaning, greatly benefiting scientific research. However, LDA's biggest disadvantage is that it is a typical bag-of-words model treating a document as a set of words without considering word order and sequence relationships.

2.1.2 Word2vec Word Vector Model Although words in LDA can roughly correspond to topics, this is usually not the case for word vectors. To deeply understand text semantics and content, contextual considerations are crucial. The LDA model fails to incorporate word relationships into its calculations, while a major feature of word vector models is describing relationships between

words. Word vectors are unrelated to word content but related to semantics, focusing more on contextual logic.

Word2vec mainly has two models: (1) in the Continuous Bag-of-Words (CBOW) structure, a set of context words predicts the pivot word; and (2) in the Skip-gram architecture, the pivot word predicts surrounding context words. In other words, CBOW inputs the sum of vectors of n words around word w_i and outputs the vector of word w_i itself; Skip-gram inputs the vector of word w_i itself and outputs vectors of n words around w_i.

From the model perspective, LDA's foundation is latent topics, while Word2vec's foundation is context. LDA focuses on document-word co-occurrence, while Word2vec focuses on context-word co-occurrence. The two are complementary for semantic analysis and form the basis of this study's model construction.

2.2 Model Construction

The LDA2vec model proposed by C.E. Moody et al. is a model that jointly learns dense word vectors mixed with Dirichlet-distributed latent document-level topic vectors. It simultaneously leverages LDA's advantage in topic capture and Word2vec's advantage in capturing relationships between words. Built on Word2vec's Skip-gram model, it transforms from using a pivot word to predict context words to using context vectors to predict context words. Specifically, it extends the Skip-gram model by incorporating topic and document vectors, combining ideas from word embedding and topic models. Inspired by Latent Dirichlet Allocation (LDA), the model is extended to simultaneously learn word, document, and topic vectors.

Therefore, referencing C.E. Moody's LDA2vec model, this study hopes to more efficiently predict surrounding words through more data and features to more effectively extract topics implicit in literature. Based on the hybrid LDA2vec model, this study borrows its approach of integrating LDA's global characteristics and Word2vec's local relationships to explore a hotspot topic identification method that mixes sparse document representations with dense word and topic vectors, constructing the model structure and process shown in Figure 1 [Figure 1: see original paper].

C.E. Moody's implementation algorithm for the LDA2vec model has overly high GPU requirements, is suitable only for ultra-large-scale data, and has low efficiency. Based on experiments with their model from GitHub, results show little difference compared to traditional LDA. Therefore, considering hotspot identification needs and the relatively small scale of original data, this study made some improvements to the model implementation: first using mature Word2vec and LDA models to train the corpus, then using LDA2vec's core algorithm for iterative calculation to obtain better results while improving efficiency.

2.3 Model Parsing

The core algorithm in the constructed model mainly includes two parts of calculation and training: one part trains to obtain information about the proportion of different topics in an article; the other part learns context vector representation based on the Skip-gram method when pivot and target words are determined.

2.3.1 Word Vector Representation Word vector learning superficially includes two parts: first obtaining word vector representation through Skip-gram, then introducing context vectors and using Skip-gram negative sampling ideas to learn target word vector representation. However, the second part's word vectors remain unchanged in this model. The actual goal is to learn content vector representation by minimizing the loss function of (pivot word + document, target word) pairs versus (pivot word + document, random word) pairs, borrowing word vector training methods.

Consistent with previous Word2vec methods, negative sampling determines sampling probability based on word frequency. The probability of sampling a certain word is as follows, with parameter set to 3/4:

$$len(w) = \frac{[counter(w)]^{3/4}}{\sum_{u \in D}[counter(u)]^{3/4}} \quad \text{Formula (1)}$$

Like the Word2vec model, when input word and target word pair (j, i) co-occur in a moving window across the corpus, they are extracted. For each (input word-target word) pair, the input word predicts nearby target words. Each word is represented by a fixed-length dense distributed representation vector. Unlike Word2vec, this model uses the same word vectors in input and target representations. The word drawing distribution is u^{β}, where u represents overall word frequency normalized by corpus size. Unless otherwise specified, the sampling power β is set to 3/4, and the number of negative samples is fixed at $n = 15$. Compared with the unigram distribution, this choice emphasizes selecting uncommon words for negative samples. Instead of optimizing softmax cross-entropy, negative sampling studies learning single words conditioned on context by drawing negative samples from marginal popularity in the corpus.

2.3.2 Document Vector Representation This part's significance lies in obtaining document vector representation for corresponding documents, then adding them to word vectors as initial context vector values. The model processes documents by decomposing document vectors into document weight vectors and topic matrices. Document weight vectors represent percentages of different topics, while topic matrices consist of different topic vectors. Therefore, context vectors are constructed by combining different topic vectors appearing in documents.

First, based on the Skip-gram model, pivot and target word pairs appearing in moving windows scanning the corpus are extracted. For each word pair, the pivot word predicts nearby target words. Second, latent vectors are randomly initialized for each document in the corpus. Document weights are softmax-transformed weights producing document proportions. The result is a proportion vector summing to 100%, representing a single document's topic proportions. For example, a document might contain three topics: Topic 0 at 41%, Topic 1 at 26%, and Topic 2 at 34%.

Each topic has a distributed representation in the same space as word vectors. Although each topic is not literally a token in the corpus, it is similar to other tokens. Each document vector is a weighted sum of topic vectors. This analysis produces interpretable topics that help people directly understand document main content without detailed examination.

The initial value setting for the context corresponding to a certain word j is as follows:

$$j = a_{j0} \cdot t_0 + a_{j1} \cdot t_1 \quad \text{Formula (2)}$$

Here, j represents word j's word vector obtained from previous steps; d represents vector representation of all word-context pairs for the word. The specific formula is:

$$j = a_{jk} \cdot t_k \quad \text{Formula (3)}$$

where t_k represents the vector representation of corresponding topic k obtained from the LDA model through matrix decomposition methods consistent with word vector length. a_{jk} represents the probability that document j belongs to topic k, with values between 0 and 1. To ensure interpretability of a_{jk}, softmax is used to guarantee it sums to 1 and is non-negative. After obtaining t_k, relevant topic vocabulary can be obtained based on similarity between word vectors and the topic.

a_{jk} calculation is closely related to L_d, with the specific calculation formula:

$$L_d = \lambda \sum_{jk} (\alpha - 1) \log p_{jk} \quad \text{Formula (4)}$$

When $\alpha < 1$, topic distribution tends to be sparse. Conversely, topic distribution becomes more concentrated. To enhance model interpretability, $\alpha = n - 1$ is used, where n represents the number of topics. Meanwhile, when $\lambda = 200$, the model performs better.

3 Experimental Results and Evaluation—Machine Learning Domain Case Study

3.1 Data Source and Preprocessing

First, a discipline with relatively mature development and clear boundaries was selected as the analysis object. Machine learning research results were chosen as data objects. On February 1, 2019, all Chinese literature published in the CNKI journal database and patent database were retrieved. According to experimental requirements, scientific and technical literature (including academic papers and patent documents) in the machine learning field were searched. The retrieval expression was set as SU='machine learning', with publication time and patent disclosure date limited to the 15-year period from 2004 to 2019. The retrieval results on February 1, 2019, included 5,869 journal articles and 3,865 patent documents. After screening original data to remove irrelevant content and duplicates, a total of 8,928 items were collected, including 5,063 journal papers and 3,865 patent documents.

In the CNKI database, each journal article and patent document has title and abstract indexing, supporting multi-source text fusion. Journal paper data and patent literature data were combined by title and abstract fields to form the preliminary raw dataset.

The content obtained after the above steps was raw data directly crawled from the database. It required segmentation and stop word removal to process into content suitable for subsequent model input and computer recognition. Jieba was used as the tool for word segmentation and stop word removal. Jieba supports stop word filtering; this experiment used jieba to filter out words and punctuation without specific meaning in documents as stop words. Text preprocessing is a repetitive process requiring expansion of custom segmentation dictionaries and feature selection until processing results meet model input requirements.

3.2 Topic Extraction Based on LDA2Vec

The LDA2vec model proposer C.E. Moody open-sourced the core library on GitHub, but experiments based on their model show overly high GPU requirements and little performance difference compared to traditional LDA. Therefore, this study improved the model implementation: first using mature Word2vec and LDA models to train the corpus, then using LDA2vec's core algorithm for iterative calculation to obtain better results while improving efficiency.

3.2.1 Word Vector Representation The preprocessed paper and patent texts were used as corpus to generate word vectors for the fused document set using Word2vec as input for subsequent models. Python's Gensim toolkit encapsulates the Word2vec model. This experiment implemented Skip-gram model word vector training using gensim.models.word2vec in the Gensim pack-

age. According to research requirements, key parameter settings for Word2vec are shown in Table 1 .

Table 1 Word2vec Model Parameter Settings

Parameter	Value	Reason and Purpose
sg	1	Set algorithm to Skip-gram
size	100	Word vector dimension, default 100 for subsequent calculation convenience
window	5	Training window size, generally 5
min_{count}	5	Dictionary truncation minimum frequency, default 5
sample	1e-3	Sampling threshold, higher frequency words are more easily sampled, default 1e-3
negative	3	Do not use HS method, adopt negative sampling method
noise_{words}	15	Number of noise words for negative sampling, generally 3

3.2.2 Document Vector Representation Another LDA2vec model input comes from LDA model output results—the topic-word distribution matrix and document weights. Therefore, this study also used the preprocessed corpus as dataset input for LDA model training. Python includes many packages like gensim and sklearn that encapsulate LDA models. Considering subsequent perplexity evaluation metric calculations, sklearn was chosen for LDA model training.

Topic number setting significantly impacts LDA model output results. Too many topics lead to insignificant results; too few topics result in some words corresponding to multiple topics. Perplexity is commonly used as a main evaluation metric for topic models, describing topic division certainty and reflecting model quality to some extent. Although topic number selection affects perplexity calculation and perplexity can only serve as a reference, topic number determination also requires considering subjective needs. As topic numbers vary, model perplexity changes continuously, as shown in Figure 2 [Figure 2: see original paper]. After comprehensive consideration of perplexity values and subjective research needs, the topic number K was set to 15, with hyperparameters taking default values for LDA model training to obtain topic-word matrices and document weights. The extracted results are shown in Figure 3 [Figure 3: see original paper].

Topics were sorted by occurrence probability from high to low, totaling 15 topics. Top 10 probability topic words for each topic were selected to more clearly and accurately understand each topic's implicit semantics.

3.2.3 Topic Extraction Based on LDA2vec LDA2vec model inputs include word vectors obtained in Section 3.2.1 and document vectors calculated in Section 3.2.2, which are input into the LDA2vec model for fusion training.

3.3 Topic Visualization Based on LDA2vec

The pyLDAvis toolkit was used for visual display of topic identification results, enabling more intuitive observation and analysis of hotspot topic results. The visualization results are shown in Figure 4 [Figure 4: see original paper].

The pyLDAvis visualization interface consists of two parts: the left side visually displays all identified topics, with graphic size representing topic occurrence probability and graphic positions indicating relationships between different topics; the right side visualizes topic word probabilities, with light bars representing total occurrence frequency of topic words. When selecting a topic on the left, the corresponding topic words' frequencies in that topic are highlighted on the light bars.

Based on this visualization, hotspot topics, prominent topics, and inter-topic relationships can be more clearly explored. Results show that identified Topics 1-4 account for the vast majority of all literature. Observing characteristic words of topics 0-3, they can be summarized as "Algorithms and Methods," "Text Classification," "Feature Detection," and "Data Analysis." Among these, research on machine learning-related algorithms and methods holds an absolute position in the machine learning field and is associated with other topics. Eight topics in the first quadrant—"Medical Applications," "Predictive Analysis," "Images," "Educational Applications," "Mechanical Applications," "Communication and Signals," "Early Warning Systems," "Genetic Applications," and "Semantics"—though accounting for less proportion with larger gaps compared to the first four topics, have many associations and overlaps, showing that machine learning applications in different fields are closely interrelated and mutually referential.

From both visualization results, topics extracted by the LDA2vec model are substantial, with relatively clear relationships and minimal overlap or crossing. The pyLDAvis tool enables convenient in-depth exploration and analysis of identified topics' relationships, connotations, and meanings, greatly benefiting researchers.

3.4 Experimental Comparison and Evaluation

C.E. Moody validated the LDA2vec model's feasibility using Hacker News comment data and the Twenty Newsgroups text classification/clustering dataset, primarily demonstrating model identification results and calculating topic coherence for a small portion without comparing model performance with traditional models. This experiment conducts comparison and evaluation from two aspects to verify the proposed method's feasibility and effectiveness. First, evaluation based on the widely used metric—perplexity—understood as the uncertainty of the trained model about which topic document d belongs to; lower perplexity

indicates better clustering effects. Second, evaluation based on topic coherence metrics quantifies similarity relationships between characteristic words under identified topics, reflecting which topics are usable and valuable.

3.4.1 Model Perplexity In information theory, perplexity is an important metric for measuring how well a probability model predicts samples. In natural language processing, a language probability model can be viewed as a probability distribution over entire sentences or paragraphs, with the basic idea that better language models assign higher probability values to test set sentences. The formula is:

$$P(\tilde{W}|M) = \exp\left(-\frac{\sum_{m=1}^{M} \log p(w_{m1}, w_{m2}, ..., w_{mN_m})}{\sum_{m=1}^{M} N_m}\right) \quad \text{Formula (5)}$$

From the formula, smaller perplexity means larger sentence probability and better language model performance.

The same data were trained using both LDA and LDA2vec models. Using Python's sklearn package lda_{perplexity} function, perplexity values for both models were calculated. Topic number range was set to [1, 100] with interval 5, and perplexity variation curves for both models were calculated and plotted as topic numbers varied from 1 to 100. The perplexity value distribution curves for LDA and the experimental model are shown in Figure 5 [Figure 5: see original paper].

Figure 5 shows that the new model's curve appears below the LDA topic model curve within a certain range, especially when topic number K $\leq$ 40. This meets most research topic identification needs, as identifying research hotspots in a discipline requires dividing large numbers of documents into limited, effective topic classifications to facilitate subsequent scientific research support. Too many topics do not meet our needs. Since larger perplexity values indicate worse model classification effects, while smaller values indicate better classification effects and stronger generalization ability, the experimental model is more suitable for research hotspot identification within a certain range.

3.4.2 Topic Coherence The corpus was experimented with under both LDA and LDA2vec algorithms to identify the same number of hotspot topics and top 10 characteristic words under each topic. Hotspot topic identification results based on LDA and LDA2vec models are shown in Tables 2 and 3 respectively.

Table 2 LDA2vec Hotspot Topic Identification and Induction Results

Topic	Top 10 Topic Words
Algorithms and Methods	algorithm, method, learning, machine, research, data, technology, analysis, algorithm, field, development
Text Classification	model, prediction, algorithm, learning, machine, method, classification, data, feature, regression
Feature Detection	algorithm, feature, method, classification, learning, detection, data, machine, sample, propose
Data Analysis	artificial intelligence, development, technology, intelligent, system, learning, analysis, field, machine, computing
Financial Applications	artificial intelligence, development, technology, data, research, economy, decision, risk, planning, city
User Analysis	data, system, analysis, user, technology, early warning, machine, network, platform, fault
Medical Applications	diagnosis, model, clinical, patient, prediction, learning, method, analysis, machine, imaging
Predictive Analysis	prediction, variable, algorithm, soil, research, machine, model, utilization, information, learning
Video Data	video, machine, research, technology, learning, generation, target, monitoring, utilization, artificial intelligence
Teaching Applications	data, curriculum, teaching, learning, student, major, enterprise, innovation, training, artificial intelligence
Mechanical Applications	system, testing, algorithm, ship, environment, machine, path, learning, tracking, operation
Communication and Signals	network, neural, recognition, generation, fraud, machine, learning, structure, chip, signal
Early Warning Systems	forecast, model, learning, drop, knowledge, construction, effect, relationship, accident
Genetic Applications	effect, hereditary transmission, prediction, constitution, material, DNA, data, research, genome, hierarchy
Semantics	entity, semantics, relationship, classification, research, purpose, meaning, information, extraction, structured

Table 3 LDA Hotspot Topic Identification and Induction Results

Topic	Top 10 Topic Words
Model Prediction	model, prediction, risk, learning, machine, based, data, patient, clinical, method
Adversarial Samples	adversarial, sample, power grid, webpage, swarm, agent, sound, attack, descriptor, content

Topic	Top 10 Topic Words
Diabetes Research	diabetes, research, method, learning, algorithm, model, prediction, machine, detection, classification
Machine Learning	learning, machine, research, data, depth, algorithm, application, method, based, classification
Text Data	based, method, text, data, learning, algorithm, sentiment, information, model, algorithm
Neural Networks	algorithm, learning, prediction, machine, research, information, diagnosis, method, flight, forecast
Fault Diagnosis	neural network, model, abnormal sound, effect, pattern, quality, chip, based, fault, feature
Ship Applications	feature, based, algorithm, prediction, machine, equipment, learning, data, clustering, image
Android Applications	segmentation, research, based, algorithm, prediction, device, learning, machine, data, application
Commodity Effects	commodity, effect, meridian conduction, hash, web, influence, sales, brand, robot, environment
Library Forecasting	environment, library, forecast, commercial bank, customer, identification, city, division, operation
Brand Recognition	brand, robot, environment, library, forecast, commercial bank, customer, identification, city

Preliminary observation of the topic word results in the two tables shows that topics identified by the LDA2vec model have higher comprehensibility. Section 3.3's pyLDAvis visualization can show the number of words contained in each topic and their distances, making clustering effects more interpretable, but cannot provide numerical quality assessment. The topic coherence method quantitatively evaluates model effectiveness with specific values. For further verification, topic coherence was used as an evaluation metric. Since people understand topic models as words belonging to the same topic co-occurring frequently in corpora, topic coherence measures semantic similarity between high-scoring words in topics, helping distinguish interpretable topics from statistically inferred ones.

Gensim 0.13.1 provides several calculation methods, including C_{UCI} and U-Mass, which mainly differ in "co-occurrence" definitions. UCI formula is:

$$score(w_i, w_j, \epsilon) = \log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)} \quad \text{Formula (6)}$$

It calculates word probabilities by computing word co-occurrence frequencies in sliding windows of external corpora (like Chinese Wikipedia). This metric can be considered an external comparison of known semantic evaluation. The U-Mass metric defines scores based on document co-occurrence:

$$score(w_i, w_j, \epsilon) = \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)} \quad \text{Formula (7)}$$

where $D(x, y)$ counts documents containing both words x and y, and $D(x)$ counts documents containing x. The U-Mass metric calculates counts from the original corpus used to train the topic model, making it more intrinsic. For this evaluation, the U-Mass method was adopted for topic coherence measurement, as this evaluation measure has been proven to better match human judgment of topic quality.

Using this evaluation standard, the correlation relationships under divided topics are quantitatively manifested. Characteristic words from 15 topics output by both models were selected, and the U-mass coherence function was used to calculate topic relevance, with results shown in Table 4 .

Table 4 Topic Coherence Comparison Between LDA2vec and LDA

Topic #	LDA2vec Coherence	Topic #	LDA Coherence
1	0.679	1	0.623
2	0.729	2	0.425
3	0.668	3	0.635
4	0.678	4	0.691
5	0.687	5	0.653
6	0.612	6	0.624
7	0.712	7	0.564
8	0.675	8	0.635
9	0.472	9	0.653
10	0.615	10	0.635
11	0.574	11	0.624
12	0.645	12	0.564
13	0.612	13	0.635
14	0.713	14	0.653
15	0.479	15	0.635
Average	**0.653**	**Average**	**0.635**

As shown, within most identified topics, the proposed LDA2vec model's topic coherence values are slightly higher than traditional LDA's, with an average of 0.653 slightly greater than LDA's 0.635. Therefore, from quantitative verification based on topic coherence, the internal correlation of topic words under certain topics is higher, making them easier to understand and summarize, providing greater convenience for researchers' next steps in scientific innovation.

Overall, compared with traditional LDA, the proposed model inherits advantages of traditional LDA algorithm and Word2vec word clustering algorithm,

providing reference value for topic research. Moreover, in multi-source text environments, this method demonstrates good performance. This study better introduces the LDA2vec topic model method into library and information science, rapidly and accurately identifying hot topics embedded in multi-source texts to support scientific research innovation.

4 Conclusion and Limitations

4.1 Conclusion

This study's innovation lies in: (1) exploring application scenarios for multi-source data fusion. Current research on multi-source data fusion in library and information science mostly explores macro-level significance and methods, but lacks in-depth discussion from practical application scenarios, solutions, and specific technical implementation details. This research provides one approach. As mentioned, different scientific texts contribute differently to research—some focus on theoretical research, some on methods and technology implementation, and some on frontier exploration. One innovation is fusing multi-source texts as objects for research hotspot topic identification and studying specific implementation methods and technical details. This experiment selected two different data sources—journal paper data and patent literature data—to achieve preliminary integration of theory and practice when identifying research hotspot topics in a discipline. (2) Exploring and verifying application scenarios of the LDA2vec topic model. Currently, research hotspot identification primarily uses traditional LDA models, with some optimizations. Existing LDA2vec model research mainly applies to news recommendation and sentiment tendency analysis. This study innovatively applies this model to library and information science, proposing specific implementation methods for research hotspot topic identification, expanding the model's practical applicability.

In summary, the proposed LDA2vec-based research hotspot identification method achieves relative improvement in topic extraction effectiveness. The model's perplexity is lower than traditional LDA across most ranges, with stronger generalization ability. Higher internal correlation of topic words under certain topics makes them easier to summarize and name, providing more convenience for researchers.

4.2 Limitations

First, regarding experimental data source selection. This study's core content is topic identification facing multi-source texts. However, only two data sources—journal papers and patent documents—were selected for fusion analysis, without involving other data types and sources, and without discussing impacts of different data source types on experimental results.

Second, regarding data acquisition and processing. Due to the enormous size of full-text data, this study mainly focused on titles and abstracts of the two data

sources. Fusion of these two data types relied on scientific literature databases' indexing functions for titles and abstracts, and these two document types have similar, complete functional structures—meaning this study's data foundation is isomorphic. When facing heterogeneous data sources, deeper exploration of data acquisition and processing would be needed.

In conclusion, research hotspot identification is significant for scientific research work. This study conducted some methodological and application explorations, but future research should address more complex multi-source data and more efficient identification effects.

References

[1] Qiu Junping, Wen Fangfang. Visual analysis of research hotspots and frontiers in library and information science in recent five years—A bibliometric study based on 13 high-impact foreign source journals[J]. Journal of Library Science in China, 2011, 37(2): 51-60.

[2] Ren Hongjuan. Research on scientific structure analysis methods based on literature feature fusion[J]. Information Studies: Theory & Application, 2013, 32(7): 97-100.

[3] Morris S A, Yen G, Wu Z, et al. Timeline visualization of research fronts[J]. Journal of the Association for Information Science & Technology, 2003, 54(5): 413-422.

[4] Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents[J]. Journal of the Association for Information Science & Technology, 1973, 24(4): 265-269.

[5] Lei Xiaoqing, Liu Xiaoyan. Statistics and analysis of paper keyword features[J]. Library and Information Service, 1998(5): 19-20, 32.

[6] Zeng Qian, Yang Siluo. Comparative study on knowledge exchange in foreign library and information science—From the perspective of journal citation analysis[J]. Information Studies: Theory & Application, 2013, 36(10): 114-119.

[7] Think Tank Encyclopedia[EB/OL]. [2019-02-20]. https://wiki.mbalib.com/wiki/知识单元. 2018-12-02-2019-02-28.

[8] Wang Xiaoguang. Formation and evolution of scientific knowledge networks (I): Proposal of co-word network method[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(4): 599-605.

[9] Zhu Qingsong, Leng Fuhai. Research on topic identification of highly cited papers based on citation content analysis[J]. Journal of Library Science in China, 2014, 40(1): 39-49.

[10] 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(3): 86-96.

[11] Ruan Guangce, Xia Lei. Journal paper hot topic selection identification based on Doc2Vec[J]. Information Studies: Theory & Application, 2019, 42(4): 110-115.

[12] Zhao Yifang, Pei Lei, Kang Lele. Policy text topic identification based on paragraph information gain[J]. Digital Library Forum, 2018(11): 2-10.

[13] Lowe S A. The beta-binomial mixture model for word frequencies in documents with applications to information retrieval[C]//Proceedings of the sixth European conference on speech communication and technology. Budapest: Dragon System Inc, 1999.

[14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3: 993-1022.

[15] Feng Jia, Zhang Yunqiu. Research on ontology-based research topic semantic analysis method[J]. Library and Information Service, 2018, 62(7): 96-103.

[16] Wang Hongwei, Gao Song, Lu Pin. Online news hotspot identification research based on LDA and SNA[J]. Journal of the China Society for Scientific and Technical Information, 2016(10): 1022-1037.

[17] Li Yongzhong, Cai Jia. Evolution and visual analysis of domestic e-government research topics based on LDA[J]. Modern Information, 2017, 37(4): 158-164.

[18] Ye Chunlei, Leng Fuhai. Research on scientific literature topic identification method based on citation-topic probability model[J]. Information Studies: Theory & Application, 2013, 36(9): 100-103.

[19] Wang Lianxi. Analysis of domestic microblog research hotspots and topic mining—Taking computer and library information disciplines as research objects[J]. Journal of Intelligence, 2015(4): 127-132.

[20] Ma Hong, Cai Yongming. Chinese text topic analysis of co-word network LDA model: Taking traffic law literature (2000-2016) as an example[J]. Data Analysis and Knowledge Discovery, 2017, 32(12): 17-26.

[21] Shen Si, Xu Fei, Wu Peng. Analysis and mining of implicit time information in literature for scientific research topics[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(4): 370-381.

[22] Pu Shanshan. Research on expert recommendation model for scientific research collaboration based on knowledge complementarity[J]. Information Studies: Theory & Application, 2018, 41(8): 100-105.

[23] Zhou Na, Li Xiuxia, Gao Dan, et al. Research on knowledge combination analysis based on latent topics—Taking communication as an example[J]. Agricultural Library and Information Science Journal, 2018(9): 85-.

[24] Liu Yuwen, Wu Xuangou, Guo Qiang. Network hotspot news focus identification and evolution tracking[J]. Small Microcomputer System, 2017(4): 738-

743.

[25] Zhang Cong, Yi Xiushuang, Zhu Minghao, et al. A Spark-based academic research hotspot mining method[J/OL]. Computer Engineering, 2019. [2019-02-20]. http://kns.cnki.net/kcms/detail/31.1289.TP.20190129.1332.005.html.

[26] Guan Peng, Wang Yuefen. Analysis of author research interest evolution in disciplinary field life cycle[J]. Library and Information Service, 2016, 60(10): 116-124.

[27] Hofmann T. Probabilistic latent semantic analysis[C]//Fifteenth conference on uncertainty in artificial intelligence. Berkeley: Morgan Kaufmann Publishers Inc., 1999: 289-296.

[28] Moody C E. Mixing dirichlet topic models and word embeddings to make LDA2vec[EB/OL]. [2019-02-20]. http://arxiv.org/abs/1605.02019.

[29] Hua Bolin, Li Guangjian. Research on multi-source fusion competitive intelligence in big data environment[J]. Information Studies: Theory & Application, 2015, 38(4): 1-5.

[30] Hua Bolin, Li Guangjian. Discussion on theory and application of multi-source information fusion in big data environment[J]. Library and Information Service, 2015, 59(16): 5-10.

[31] Hua Bolin. Research on multi-source information fusion methods[J]. Information Studies: Theory & Application, 2013, 36(11): 16-19.

Author Contributions: Qiu Huilin: Designed overall paper structure and wrote the paper; Shao Bo: Proposed research ideas and framework, guided paper revision and finalization.

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