Post-print: Semantic Retrieval Technology for Academic Resources Based on Word Vector Expansion

Authors: Wang Renwu, Chen Chuanbao, Meng Xianru

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

Abstract

[Purpose/Significance] Guided by statistical methods, this study explores semantic retrieval techniques based on word vector expansion to enhance the semantic retrieval capability of academic resources.

[Method/Process] Utilizing natural language processing and text mining technologies, the metadata of collected academic resources (primarily scholarly papers) is preprocessed. A semantic retrieval system is constructed by integrating the word2vec word vector generation tool with the elasticsearch full-text search engine, enabling exploratory research on semantic retrieval of academic resources.

[Results/Conclusion] The proposed method can effectively improve the retrieval performance of academic information, partially achieve semantic retrieval of academic resources, and provide a reference for further research on semantic retrieval.

Full Text

Preamble

Semantic Retrieval Technology of Academic Resources Based on Word Vector Expansion

Wang Renwu, Chen Chuanbao, Meng Xianru
Department of Information Management, Faculty of Economics and Management, East China Normal University, Shanghai 200241

Abstract

[Purpose/Significance] This study adopts a statistical methodology to explore semantic retrieval technology based on word vector expansion to enhance

the semantic retrieval capabilities of academic resources. [**Method/Process**] Using natural language processing and text mining techniques, we preprocessed the collected metadata of academic resources (primarily academic papers), and constructed a semantic retrieval system by combining the word2vec embedding generation tool with the Elasticsearch full-text search engine to conduct exploratory research on semantic retrieval of academic resources. [**Result/Conclusion**] The proposed method effectively improves academic information retrieval performance, achieves semantic retrieval of academic resources to a certain extent, and provides valuable insights for future semantic retrieval research.

Keywords: word2vec; Elasticsearch; semantic retrieval; academic resources
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2018.19.014

Introduction

Academic users' information needs are typically characterized by professionalism, knowledge intensity, personalization, diversity, convenience, and human-centered design [1]. To meet these specialized needs, numerous academic resource databases have emerged, including Web of Science, ScienceDirect, CNKI, and Wanfang Academic. While these databases provide massive academic resources and high-quality information services, their proliferation also creates challenges: users must search across multiple platforms to obtain comprehensive and accurate information; complex interfaces and numerous search syntaxes create poor user experiences, especially for novices; and most importantly, these databases primarily rely on keyword matching, leading to the "vocabulary problem" [2] where identical search intents produce vastly different results due to variations in user-provided keywords, ultimately compromising retrieval quality.

If academic search systems could offer a simple search interface while transcending keyword limitations to understand users' true retrieval intentions and achieve semantic-level information retrieval, academic information retrieval effectiveness would be substantially enhanced. This study investigates the use of deep learning-based word vector technology for semantic expansion to improve academic information retrieval, thereby achieving semantic retrieval of academic resources to a certain degree.

Related Research

2.1 Semantic Retrieval Research

Current approaches to semantic retrieval can be broadly categorized into rule-based and statistical methods. Rule-based semantic retrieval constructs semantic knowledge networks through manually crafted rule bases for semantic reasoning. Statistical methods, grounded in mathematical statistics, do not require pre-built knowledge rules but instead employ algorithms to extract statistical

patterns from large corpora, computing semantic similarity between words for retrieval applications.

Semantic knowledge bases represent an early implementation of rule-based retrieval. Notable examples include WordNet and HowNet. D.I. Moldovan and R. Mihalcea [3] processed query statements and used WordNet vocabulary to expand query terms, defining synonym sets for application in the AltaVista retrieval system. Gao Xuexia and Yan Shitao [4] proposed a word sense disambiguation method based on Jaccard coefficients, using the WordNet lexicon to disambiguate query terms, achieving a 10% improvement in retrieval precision. Wang Lidong and Zhang Huixi [5] studied Sina Weibo texts, using HowNet to match Chinese subject terms with Weibo vocabulary based on semantic relatedness to satisfy user query intentions, exploring semantic retrieval for short texts.

Statistical methods have consistently been preferred for semantic retrieval due to their rigorous and scientific nature. With advances in chip technology and machine learning algorithms providing greater computational power and semantic understanding capabilities, statistical methods have experienced a renaissance. In 2003, D.M. Blei et al. [6-7] proposed the LDA topic model, enabling semantic retrieval from a probabilistic perspective. Liu Qihua [8] subsequently designed PMM and TBS models based on LDA, demonstrating effective improvements in semantic retrieval performance. In 2013, Google introduced word2vec [9][10][11], enabling large-scale training of high-quality word vectors for natural language processing tasks. Fan Qiaoqing and Fang Yu [12] used Reuters-21578 and 120ask corpora to train word2vec vectors for semantic similarity computation, combining them with the Axiomatic optimal retrieval model for health Q&A community semantic retrieval. Liu Menglan et al. [13] similarly employed word2vec, proposing a query expansion method tailored to patent literature characteristics that effectively improved patent retrieval. Xu Wentang [14] trained word vectors on Weibo texts using word2vec's skip-gram model, obtaining document and query representations through weighted averaging to design a 微博 semantic retrieval system based on the MRA-E algorithm. Following word2vec, Stanford released GloVe [15], a global word vector training tool. Chen Guohua et al. [16] trained vectors with GloVe, using random mapping for rapid vector localization in large spaces and proposing an academic document vectorization scheme that achieved good results in ScholarNet retrieval. In 2018, Google launched the AI search engine semantic experiences [17], enabling natural language dialogue where the system responds to user questions rather than keywords.

2.2 word2vec

This study employs the word2vec tool to train word vectors for academic texts. Word2vec, released by Google in 2013 and developed by T. Mikolov's research group, takes text corpora as input and uses neural network language models to learn vector representations of vocabulary, outputting word vectors for downstream tasks.

Word2vec offers two training frameworks [Figure 1: see original paper]: (1) CBOW (continuous bag-of-words), which predicts the probability of word w given its context; and (2) Skip-gram, which predicts the probability of context given word w. During training, CBOW excels in speed, while Skip-gram, though slower, performs better for low-frequency words.

System Architecture

We designed an academic resource semantic retrieval system based on word vector technology, as illustrated in [Figure 2: see original paper]. The design integrates deep learning-based word vector text processing with the open-source Elasticsearch full-text search engine to establish a semantic retrieval model, which is then applied to academic resource retrieval services with subsequent analysis and evaluation.

The system comprises five main components: data collection and processing, word vector module, query expansion module, full-text retrieval module, and data analysis module. The detailed structure is shown in [Figure 2: see original paper].

The **data collection and processing module** gathers academic literature through manual importation, web crawling, or API calls from personal repositories, professional databases, and the internet. It performs quality checks and data cleaning before preparing document objects and normalized data for indexing and word vector training.

The **word vector module** trains word embeddings for semantic query expansion. The vector library updates regularly with new datasets to ensure timeliness and coverage.

The **query expansion module** processes user queries through segmentation and stop-word removal, then expands queries using the word vector library. Terms meeting semantic thresholds form an expanded query set for full-text retrieval.

The **full-text retrieval module** indexes processed documents using Elasticsearch, retrieves documents matching query requests, and improves relevance scoring through customized algorithms before presenting sorted results to users.

The **data analysis module** performs analytical and visualization tasks on retrieval results, helping users discover data associations, grasp overall result characteristics, and refine their information needs for secondary retrieval.

Key Technology Research

4.1 Domain Dictionary Construction Technology

A domain dictionary catalogs specialized vocabulary for specific research fields. General-purpose dictionaries in segmentation tools often inadequately cover

domain-specific terms, causing erroneous splits like segmenting "support vector machine" into "support / vector / machine" or "latent semantic analysis" into "latent / semantic / analysis."

We first utilize standardized metadata fields—"Author," "Keyword," and "Organization"—as direct domain word sources. To maximize coverage, we additionally extract domain terms from corpora using pointwise mutual information (PMI) and term frequency as statistical criteria.

PMI measures association strength between random variables, quantifying character cohesiveness for term recognition. The formula is:

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

where $P(x)$ and $P(y)$ are individual probabilities and $P(x, y)$ is co-occurrence probability. PMI > 0 indicates strong association, with larger values showing stronger relationships.

We define each character combination after segmentation, regardless of length, as a "word unit." For example, string $s_1s_2s_3s_4s_5s_6$ segmented as $s_1s_2|s_3|s_4s_5s_6$ yields word units s_1s_2, s_3, and $s_4s_5s_6$. Consecutive word units form patterns: an n-unit pattern is called n-dimensional. For candidate term $s_1s_2|s_3|s_4s_5s_6$, both $s_1s_2|s_3$ and $s_3|s_4s_5s_6$ are sub-patterns.

Library and information science terminology typically contains fewer than 10 characters, usually covered by combinations of 1-4 word units. Zhang Rong's research [18] also shows that 71.723% of terms consist of 2-4 word combinations. Therefore, we only consider 2, 3, and 4-unit combinations (2D, 3D, and 4D patterns).

The domain term extraction process involves: (1) corpus preprocessing (removing garbled text, replacing punctuation with spaces, and removing stopwords); (2) segmentation and word unit frequency counting using jieba; (3) candidate term identification by traversing segmented corpora to generate 2D, 3D, and 4D candidate lists; (4) initial filtering using frequency and PMI thresholds (empirically set at 5 and 20), where valid 2D terms filter 3D terms, and valid 3D terms filter 4D terms.

After extraction, we apply filtering rules: remove temporal and quantitative expressions; merge sub-patterns; delete strings starting/ending with auxiliary words like "该" (this) or "应" (should); and remove degree adverbs like "非常" (very) or "十分" (extremely).

4.2 Word Vector Semantic Expansion Technology

Semantic concept expansion sources come from academic resource metadata. After preprocessing (filtering, punctuation removal, segmentation, stop-

word removal), we train word2vec on the corpus to obtain vector representations and semantic distances between words.

Semantic concept expansion criteria involve two aspects. First, **semantic distance**: word2vec returns semantically similar words and similarity scores. We use a threshold of 0.85—terms exceeding this value are considered strongly associated and added as expansion terms. Second, **maximum expansion term count**: while query expansion provides additional information to clarify user intent, excessive expansion introduces noise. Following word2vec's default, we return at most 10 similar terms per original query term.

The semantic query expansion process involves: (1) user query input; (2) preprocessing using domain dictionary-enhanced jieba segmentation and stopword removal; (3) expansion using trained word vectors, adding terms meeting semantic thresholds to form an expanded query set; (4) submitting the expanded set to the retrieval system.

4.3 Personalized Scoring Scheme for Academic Literature

We selected Elasticsearch [19] as our search engine, which by default uses Lucene's TF/IDF scoring mechanism:

$$score(q,d) = queryNorm(q) \cdot coord(q,d) \cdot \sum_{t \in d} (tf(t \in d) \cdot idf(t)^2 \cdot t.getBoost() \cdot norm(t,d))$$

Component explanations are provided in .

While effective generally, this algorithm has limitations for academic literature: (1) it ignores positional relevance of query terms; (2) it doesn't account for semantic differences between original and expanded terms; (3) it neglects citation impact as a quality indicator.

Regarding the first two limitations, Zhang Xiaofei and Kong Minxiu [20] suggest assigning different weights to document sections (title, keywords, abstract, full text) and expanded terms. For the third, PageRank [21] offers inspiration: highly cited papers typically indicate higher quality or significance and deserve scoring boosts.

We propose a personalized scoring strategy for academic literature. After extensive testing, we assign boost values of 1.2, 1.1, and 1.0 to title, keyword, and abstract fields respectively, establishing the importance hierarchy: title > keywords > abstract. Original query terms receive weight 1.0, while expanded terms receive weights equal to their semantic similarity scores. For citation impact, we use Elasticsearch's `function_{score}` with `field_{{{value}}_{{factor}}}` to incorporate citations, applying logarithmic smoothing: $\log(1 + citation)$ to avoid distortion from extreme values. The final relevance score is:

$$\text{Final Score} = \text{Base Relevance Score} + \log(1 + citation)$$

System Implementation and Evaluation

5.1 Experimental Data Source

Our experimental dataset comprises 122,519 metadata records of library, information, and archival science (LIS) core journal papers from CNKI (2002-2017). We additionally crawled citation counts for these papers. The 28 core journals were selected based on the 2017 Peking University Core Journal Directory [22] under categories G25 (Library Science) and G27 (Archival Science). Metadata fields are described in .

5.2 Domain Dictionary Segmentation Evaluation

Our domain dictionary construction algorithm initially extracted 432,457 2D candidates, 335,062 3D candidates, and 157,853 4D candidates. After PMI, frequency filtering, and rule-based cleaning, we obtained 4,393 valid domain terms: 3,590 2D terms, 681 3D terms, and 122 4D terms. Combined with 223,831 directly imported terms from metadata fields, we deduplicated with jieba's built-in dictionary to create a final domain lexicon of 205,826 terms.

Examples of 2D, 3D, and 4D terms are shown in , , and respectively.

We integrated the domain dictionary into jieba's user dictionary and compared segmentation performance. Without the domain dictionary, jieba achieved 87.42% average accuracy; with the dictionary, accuracy rose to 97.44%—a 10% improvement. The enhanced segmentation correctly identifies domain-specific terms like "support vector machine," "chaotic time series," and institutional names like "Hainan Vocational and Technical College." Segmentation examples are shown in [Figure 3: see original paper] and [Figure 4: see original paper].

5.3 Academic Resource Word Vector Training

We used Python's gensim library to train word vectors. First, we segmented the LIS corpus using our domain dictionary and removed stopwords. Then we trained word2vec with parameters specified in : vector dimension 200, window size 5, minimum count 5, iterations 5, using the skip-gram model. The resulting binary vector library is periodically updated with new corpora.

5.4 Retrieval Performance: Precision, Recall, and F1

Precision and recall are fundamental evaluation metrics. **Recall** measures the proportion of relevant results retrieved among all relevant documents:

$$Recall = \frac{\text{Retrieved Relevant Results}}{\text{All Relevant Results}}$$

Precision measures the proportion of relevant results among all retrieved documents:

$$Precision = \frac{\text{Retrieved Relevant Results}}{\text{All Retrieved Results}}$$

F1 score comprehensively evaluates both metrics:

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

We evaluated the system using 300 documents: 50 relevant to each of three keywords ("archive culture," "data mining," "information literacy") and 150 irrelevant documents. Performance metrics are shown in and visualized in [Figure 5: see original paper], [Figure 6: see original paper], and [Figure 7: see original paper].

Compared to keyword matching, word vector-based semantic expansion shows more significant improvement in recall (average +19.33%) than precision (average +2.77%). This occurs because semantic expansion adds more relevant terms, retrieving more relevant documents (higher recall) but also introducing some noise (smaller precision impact). The F1 score improved by an average of 11.56%, demonstrating overall system superiority.

Conclusion

This study constructed a semantic retrieval system using word2vec and Elasticsearch, designing key technologies including automated domain dictionary construction, word vector semantic expansion, and personalized scoring for academic literature. Experiments on 122,000+ LIS domain papers demonstrated significant retrieval improvement, particularly in recall. Future work should address result ranking algorithms and personalized recommendation.

References

[1] Wang Jiehui. Analysis of functional requirements of university researchers for library one-stop resource discovery platforms[J]. Information Studies: Theory & Application, 2014(12): 95-98, 80.
[2] FURNAS G W. The vocabulary problem in human-system communication[J]. Communications of the ACM, 1987, 30(11): 964-971.
[3] MOLDOVAN D I, MIHALCEA R. Using wordnet and lexical operators to improve internet searches[J]. IEEE Internet Computing, 2000, 4(1): 34-43.
[4] Gao Xuexia, Yan Shitao. Research on semantic retrieval based on WordNet word sense disambiguation[J]. Natural Science Journal of Xiangtan University, 2017(2): 118-121.
[5] Wang Lidong, Zhang Huixi. Research on semantic retrieval of microblog

text based on HowNet[J]. Information Science, 2016(9): 134-137.
[6] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.
[7] BLEI D M, LAFFERTY J D. Correction: a correlated topic model of science[J]. Statistics, 2007, 1(1): 17-35.
[8] Liu Qihua. Text semantic retrieval model based on LDA[J]. Information Science, 2014(8): 38-43, 55.
[9] GOOGLE. Word2vec[EB/OL]. [2017-08-26]. https://code.google.com/archive/p/word2vec/.
[10] MIKOLOV T. Word2vec[EB/OL]. [2017-08-26]. https://github.com/tmikolov/word2vec.
[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2018-06-17]. https://arxiv.org/pdf/1301.3781v3.pdf.
[12] Fan Qiaoqing, Fang Yu. Research and analysis of semantic retrieval technology for health Q&A communities[J]. Electronic Technology & Software Engineering, 2017(2): 202-204.
[13] Liu Menglan, Liu Bin, Peng Zhiyong. Research on patent automatic query expansion based on word vectors[J]. Computer Engineering & Science, 2017(12): 2297-2305.
[14] Xu Wentang. Research and implementation of microblog retrieval system based on word vectors[D]. Shanghai: Donghua University, 2017.
[15] STANFORD. GLOVE[EB/OL]. [2017-08-26]. https://nlp.stanford.edu/projects/glove/.
[16] Chen Guohua, Tang Yong, Xu Yuying, et al. Research on academic semantic search based on word vectors[J]. Journal of South China Normal University (Natural Science Edition), 2016, 48(3): 53-58.
[17] GOOGLE[EB/OL]. [2018-06-17]. https://research.google.com/semanticexperiences/.
[18] Zhang Rong. Research on term definition extraction, clustering and term recognition[D]. Beijing: Beijing Language and Culture University, 2006.
[19] ELASTICSEARCH[EB/OL]. [2017-08-26]. https://www.elastic.co/cn/.
[20] Zhang Xiaofei, Kong Fanxiu. Research on scientific literature retrieval based on semantic concept analysis[J]. Information Studies: Theory & Application, 2016, 39(8): 115-118.
[21] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: bringing order to the web[R]. Stanford InfoLab, 1999.
[22] Baidu Baike. 2017 latest edition of "Chinese Core Journals Directory"[EB/OL]. [2017-08-26]. https://wenku.baidu.com/view/15c20df10d2259010202020740be1e650e52eacfa4.ht

Author Contributions

Wang Renwu: Responsible for topic selection, framework design, experimental design, and paper revision.
Chen Chuanbao: Responsible for conducting experiments and drafting the initial manuscript.
Meng Xianru: Responsible for data collection and processing, and conducting experiments.

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