Comparative Study of Doc2Vec-Based Patent Document Similarity Detection Methods: Post-print

Authors: Cao Qi, Zhao Wei, Zhang Yingjie, Shujun Zhao, Chen Liang

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

Abstract

[Purpose/Significance] Patent similarity measurement (Similarity Measurement) can, at the macro level, assist in formulating national innovation strategic planning, identifying domestic and international hotspots, and countering patent trolls from other countries, while at the micro level providing auxiliary support for patent inventors, patent examiners, and patentees. [Method/Process] This study proposes a deep learning-based Doc2Vec patent similarity analysis method. Based on an uncleaned patent corpus, employing the deep learning Doc2Vec model, and randomly selecting patents, we investigate the patent similarity detection problem and conduct comparative studies with traditional similarity detection models. [Results/Conclusion] Experimental results demonstrate that the deep learning-based Doc2Vec model and the TF-IDF model yield similar results when processing uncleaned patent corpora. This method imposes lower requirements on analysts' patent domain knowledge, eliminating the need for patent domain knowledge-based data cleaning of patent data. Simultaneously, it can provide new intelligent tool support for patent infringement and patent novelty search, reducing research thresholds and workload while enhancing research efficiency.

Full Text

Preamble

A Comparative Study of Patent Document Similarity Detection Methods Based on Doc2Vec

Cao Qi[1], Zhao Wei[1], Zhang Yingjie[2], Zhao Shujun[3], Chen Liang[4]
[1],[2],[4]Institute of Scientific and Technical Information of China, Beijing 100038
[3]Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] Patent similarity detection assists in formulating national innovation strategy planning at the macro level by identifying hotspots both domestically and internationally and addressing patent trolls from other countries. At the micro level, it provides support for patent inventors, examiners, and patentees. [**Method/Process**] This paper proposes a deep learning-based Doc2Vec patent similarity analysis method that operates on an uncleaned patent corpus. Using the deep learning Doc2Vec model, we randomly selected patents to study patent similarity detection problems and conducted comparative research with traditional similarity detection models. [**Result/Conclusion**] Experimental results demonstrate that the deep learning-based Doc2Vec model and the TF-IDF model produce similar results when processing uncleaned patent corpora. This method requires less patent domain knowledge from analysts and eliminates the need for data cleaning based on patent domain expertise. Simultaneously, it provides new intelligent tool support for patent infringement detection and patent novelty searches, reducing research thresholds and workload while improving research efficiency.

Patent similarity detection holds significant research importance at both macro and micro levels. Macroscopically, analyzing patent similarity can assist in national innovation strategy planning and identify research hotspots. In this regard, S. Mukherjee et al. [1] analyzed 28,426,345 papers from 1945 to 2013 and 5,382,833 U.S. patents from 1950 to 2010 to propose definitions of research hotspots and compare their evolution across different stages. Additionally, developed countries frequently exploit intellectual property rights to file unjustified lawsuits against developing nations. S. Padmanabhan et al. [2] analyzed the litigation processes related to HPV patents in India. As a developing country, China also needs relevant countermeasures. Wang Yuefen et al. [3] examined the current state of patent literature similarity research for patent early warning purposes. Microscopically, patent inventors are often technical experts in specific sub-disciplines who aim to protect new product methods in their fields, making it difficult to write cross-disciplinary invention patents. Patent examiners, also technical experts in specific sub-disciplines, often struggle to quickly review cross-disciplinary patents and must spend considerable time learning knowledge from different fields. Current patent text similarity detection primarily relies on patent examiners manually defining lexicons for term frequency detection, requiring high professional backgrounds or substantial time investment to understand relevant patent domain knowledge.

2 Research Status

In recent years, deep learning has been widely applied in text similarity detection and other fields. Patents are a special type of text that also lends itself to deep learning approaches. This paper employs deep learning techniques to study U.S. patents as examples and investigate patent similarity detection problems.

2.1 Patent Document Structure and Similarity Analysis Process

The research object of this study is patent documents. Taking U.S. patent documents as examples, the structure primarily consists of the patent title, abstract, claims, specification, and citations, as shown in Figure 1 [Figure 1: see original paper] [4]. Based on this structure tree, the current patent similarity analysis process is illustrated in Figure 2 [Figure 2: see original paper] [5].

Typical practices differ mainly in data source selection strategies and data source analysis algorithms. Such similarity analysis requires patent domain experts to possess knowledge across multiple patent fields. Regarding data source selection, the two main approaches are cross-database comparison between patent databases and other databases, and analysis within the patent database itself. In terms of data source algorithms, these are primarily divided into patent data cleaning algorithms and data search/comparison algorithms. Analyzing patents through their own data mainly involves finding various relationships between patents, such as citation clustering relationships in patent citation networks. Cross-business database comparison between patents and other business databases facilitates manual analysis by patent domain experts, as shown in Figure 3 [Figure 3: see original paper].

2.2 Research Objects of Patent Similarity

Patent similarity research is divided into studies based on linking patents to other business databases and studies based on the patent document database itself.

(1) Patent Linkage with Other Business Databases: In this area, S. Mukherjee et al. [1] studied the consistency and popularity of technology between patents and papers by comparing patent databases with paper databases. J. A. Smith et al. [6] compared patent quality by analyzing the distribution of papers by patent inventors. Li Li et al. [7] studied patent term disambiguation by comparing Chinese and English patents. Lou Yan et al. [8] researched patent alternative solutions by comparing patent data with business data. The advantage of cross-database similarity research lies in its simple principles and verifiable conclusions, but the difficulty lies in processing large amounts of different databases and requiring relevant domain knowledge across different industries.

(2) Patent Self-Database: Based on the patent structure tree in Figure 1, the main data selections are shown in Table 1 [5]. Such research generally analyzes the same type of text in specified research fields. For example, Chen Yunwei et al. [9] studied patent inventor collaboration by analyzing patent citation networks. Others analyze different types of data in the patent structure, such as Wang Xin et al. [10] who measured patent similarity based on classification numbers and citations. Zhu Lei et al. [11] compared patent text data with image data for appearance design patent retrieval based on shape semantics. Methods using the patent self-database rely more on researchers' experience in the patent industry but involve smaller and more manageable data volumes compared to

cross-business database research.

2.3 Domain Knowledge-Based Research Methods

Common similarity detection algorithms require data cleaning based on patent domain knowledge. The main purpose of data cleaning is noise removal, synonym disambiguation, reduction of computational dimensions, and generation of corresponding entity representations. For example, Wang Jin et al. [12] used maximum entropy models to generate entities from patent texts. However, data cleaning itself is not simply deleting invalid word lists. Polysemy and synonymy in natural language processing are inherently complex. From an application perspective, patent examiners may not be industry experts but have a clear understanding of patent rules, requiring patent inventors to supplement their patents with relevant explanations to facilitate examiner comprehension, such as Chen Liang et al. [13] who used KnowledgeGraph to complete implicit entities in patents. For word sense disambiguation, traditional natural language processing methods are generally employed, such as using WordNet to analyze hyponymy relationships for disambiguation or using lexical and semantic analysis to generate lexicons for disambiguation, as Jiang Lixue et al. [14] used semantic roles to generate patent term lexicons. Additionally, due to differences in word formation and grammar across languages, patent sentences need to be simplified to facilitate machine processing and further manual analysis. The main method employed is SAO extraction, such as Xu Haiyun et al. [15] who used SAO (Subject-Action-Object) extraction algorithms to analyze patent sentence structures for comparison. Rao Qi et al. [16] improved upon SAO by incorporating syntactic analysis combined with SPT (the Shortest Path enclosed Tree) structure.

The main purpose of data analysis is to perform similarity analysis using relevant algorithms on the basis of data cleaning, generate patent entities to establish a patent map database, and compare new patents with the established database to judge their similarity and novelty. Currently, domestic patent examiners primarily use the patent search and service system (S system) of the National Intellectual Property Administration, with retrieval mainly based on the VSM (Vector Space Model) and subsequent manual judgment of term position and literal similarity to assess technical solution innovation [17]. Although this approach is simple in principle, it requires substantial manual intervention from patent examiners. In addition to the VSM model, commonly used models include LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation) models. The principles behind these models for similarity comparison are roughly the same: they adopt a bag-of-words model, score each term in a patent to calculate term weights using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm, establish a VSM model for a patent based on term frequency and TF-IDF scores, and then compare the vector angles between different patent documents in the VSM model. Due to the overly sparse vector distribution in the VSM model, LSA models employ SVD (Singular Value

Decomposition) for dimensionality reduction. If topic factors are considered, LDA models use topics for dimensionality reduction. For example, Chen Liang et al. [18] used LDA models to study topic similarity in patent evolution by examining patent entities. Liao Liefa et al. [19] compared the accuracy and recall rates of LDA and VSM models in patent topic similarity analysis.

Although VSM, LSA, and LDA models are widely used and relatively mature, secondary analysis of results still relies heavily on patent analysts' domain knowledge and consumes substantial manual effort.

2.4 Deep Learning-Based Research Methods

To address the problem that patent similarity analysis requires numerous professionals with patent domain knowledge, this paper proposes an alternative research approach: a method not based on domain knowledge but on deep learning. This new method is primarily attributed to the maturation of neural network and deep learning technologies in recent years. Neural network methods differ from traditional natural language processing methods by simulating the dendrite and axon functions of the human brain, where information is transmitted through activation functions after judgment. Deep learning methods, building upon traditional neural networks, simulate the functional non-differentiation of human brain neurons when processing information, enabling segmentation of training data, such as the convolutional neural networks widely used in image recognition. Additionally, different classifiers are employed to learn different classifications from the same data segment, with results from different segments merged later.

In the patent similarity analysis field, scholars have introduced neural network algorithms for similarity research, particularly unsupervised learning. Wu Yuying et al. [20] proposed using Self-Organization Maps (SOM) for training, generating matrices of keywords and patent weights through training for patent similarity calculation and infringement detection. Xu Kan et al. [21] used Word2Vec frameworks in deep learning to train patent texts and determine similarity of patent domain terms through training rates. Others have used deep learning algorithms for disambiguation, such as Wang Yanyan et al. [22] who used Word2Vec frameworks for word sense disambiguation. Neural network research methods can relatively reduce manual intervention, and due to the absence of traditional natural language processing constraints during training, computational efficiency is significantly improved.

Traditional natural language processing methods are more akin to manual classification and using mathematical models to eliminate noise and improve accuracy, whereas neural network methods, especially deep learning, are more similar to not pre-defining business models but using neural networks to generate models and discover patterns. Although further optimization of such methods requires researchers to combine domain knowledge for processing [5], the innovation of this paper lies in adopting a research method not based on domain

knowledge but on deep learning. This paper is based on the Doc2Vec model, which is an extension of the Word2Vec model but adds a paragraph matrix to the input layer during word vector training [23]. Doc2Vec is more suitable for processing patent texts because it works at the paragraph level, such as processing patent abstracts.

3 Experimental Design and Results Analysis

3.1 Experimental Technical Route

This experiment adopts the principles of hypothesis and hypothesis verification, statistical principles, and control and experimental control principles. The hypothesis premise of this paper is to not adopt data cleaning based on patent domain knowledge but to use deep learning-based Doc2Vec models, combined with traditional TF-IDF, LSA, and LDA models, to compare patent similarity detection results and find correlations. Specifically, we first generate model files and retrieval files through training, then use deep learning models as experimental control groups for traditional models, conducting comparative analysis using TF-IDF, LSA, LDA, and Doc2Vec models. We randomly select a group of patents for separate comparative analysis, draw hypothetical conclusions, then randomly select another group of patents to repeat the experiment for validation. Since more comparisons are needed when verifying patterns, we select 100 items for statistical analysis to verify patterns.

3.2 Experimental Environment and Preparation

This study primarily uses the Gensim framework [24] for experiments. The experimental code is based on Python 2.7.12, with MariaDB as the database (version 10.1.21). The entire experimental code is based on Gensim framework 3.0.0, with the main development environment being Ubuntu Linux v16, 64-bit operating system, Intel 16-core processor, and 64GB of RAM.

The corpus for this study was downloaded from the United States Patent and Trademark Office (USPTO), covering patent data from January 1, 2015, to August 1, 2017, totaling 3,044,956 patent entries, which were imported into the database. The patent numbers and abstracts were retained to generate CSV text corpus files consisting of patent numbers and corresponding patent abstracts, with the data structure shown in Figure 4 [Figure 4: see original paper].

3.3 Experimental Design and Process

Based on the experimental environment described in Section 3.2, we designed the entire experimental process, which mainly consists of two parts: training library generation (Train) and inference testing (Infer). The training library generation flowchart is shown in Figure 5 [Figure 5: see original paper] (Note: The method for generating LSA indexes is called LSI). After model generation, the trained models need to be used to compare with new patents for testing and

obtain their patent numbers. The specific inference testing process is shown in Figure 6 [Figure 6: see original paper].

3.4 Experimental Results Analysis

According to the experimental process in Figures 5 and 6 from Section 3.3, we conducted training. For the TF-IDF model (Group E), we used the dictionary length of File Group D as the feature number. For the LDA model (Group F) and LSA model (Group G), the feature number was defined as 10. For the proposed deep learning-based Doc2Vec method, considering performance optimization, the 3,044,956 patent documents were stored in blocks, with each block containing 327,680 patents for calculation. During Doc2Vec model training, each patent document was trained 100 times. Specifically, the status of the trained models is shown in Table 2 .

To verify the reliability of the experiment, we randomly selected two patents from the patent database [25] on September 5, 2017, for hypothesis and inference testing. One data point was used for hypothesis generation, and the other for inference testing, with results shown in Table 3 . The randomly selected patent numbers were 9754858 and 9755578. Comparing the files generated by TF-IDF (E1), LSA (F1), LDA (G1), and Doc2Vec (H1), the experimental results are shown in Table 4 .

The comparison of common items derived from different similarity algorithms in the first 20 groups of data is shown in Table 5 . We found that Group W (TF-IDF model) and Group Z (Doc2Vec model) have 7 common items in the top 20 data points, while Groups W and X, and Groups W and Y have no common items. Both Groups X and Y (LSA and LDA models) are based on the TF-IDF model. This study found that without patent domain knowledge-based data cleaning, Groups X and Y produce no intersection, meaning no similarity. However, the similarity detection effects of TF-IDF and Doc2Vec models are better than those of LSA and LDA models.

Based on this hypothesis, we tested with the second patent (Patent No.: 9755578) to verify the existence of similarity. We increased the experimental data volume. Since Groups X and Y are both based on Group W, but Group Z's model is not based on Group X, we took the top 100 results from Groups W, X, and Y and the top 20 results from Group Z for comparative analysis across different models. Following the process in Tables 4 and 5, we compared any two groups, with the analysis of common items from different similarity algorithms in the first 20 data points shown in Table 6 .

The experimental results in Tables 5 and 6 further validate the hypothesis inferred from Table 4: when comparing patent similarity, the similarity detection of TF-IDF and Doc2Vec models is superior to that of LSA and LDA models. Experimental results demonstrate that without domain knowledge-based data cleaning, TF-IDF and Doc2Vec produce similar results. Currently, the similarity detection methods used in the industry are primarily based on patent

domain knowledge for data selection, followed by TF-IDF detection.

4 Value and Application

4.1 Research Value

This paper proposes a deep learning-based Doc2Vec patent similarity analysis method, conducts case studies using Doc2Vec, and compares results with traditional TF-IDF, LSA, and LDA models through hypothesis testing and control experiments. The innovation lies in using deep learning models and algorithms to study patent similarity problems. The deep learning-based Doc2Vec patent similarity analysis method does not require researchers to have extensive patent domain knowledge. Traditional research approaches classify research objects, but due to massive data volumes, entity definition and data cleaning consume substantial work time and rely on expert systems requiring numerous patent domain experts. The proposed Doc2Vec method using deep networks does not require patent domain knowledge-based data cleaning yet achieves results similar to traditional methods.

Additionally, this study performed minimal data definition and cleaning, instead employing longer machine training periods to prevent information loss. When training TF-IDF, LSA, LDA, and Doc2Vec models, the test patent files were scored against 327,680 patent files, providing an alternative research approach. Although increasing the computational intensity and time of unsupervised machine training, this approach maximizes unsupervised learning mode rather than excessive pre-processing by industry experts, enabling pure computer automation and reducing manual workload.

4.2 Application Value

In application, this paper can support patent map generation using Doc2Vec models. Traditional patent map generation primarily relies on TF-IDF and LSA models, continuously requiring extensive lexicon division by patent domain experts (such as patent semantic classification). The approach proposed here enables patent analysis by researchers without patent domain expertise, where semantic division is based on deep learning-generated models rather than manual work. Although such generated models may produce ambiguous results that are not as clear as pure manual detection, they can significantly conserve the energy of patent agents and analysts.

Furthermore, from the perspective of patent infringement, since the Doc2Vec model learns semantics during training, it can discover infringement cases that TF-IDF and LSA models cannot detect, facilitating supplementary analysis of infringement patents for relevant enterprises in competitive landscape and infringement analysis, thereby improving service efficiency.

References

[1] Mukherjee S, Romero D M, Jones B, et al. The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: the hotspot[J]. Science advances, 2017, 3(4): e1601315.

[2] Padmanabhan S, Amin T, Sampat B, et al. Intellectual property, technology transfer and manufacture of low-cost HPV vaccines in India[J]. Nature biotechnology, 2010, 28(7): 671-678.

[3] Wang Yuefen, Xie Shoufeng, Qiu Yuting. The significance and current status of patent literature similarity research for early warning[J]. Information Studies: Theory & Application, 2014, 37(7): 135-140.

[4] Wang Xiuhong, Yuan Yan, Zhao Zhicheng, et al. A structural tree model of patent literature and its application in similarity calculation[J]. Information Studies: Theory & Application, 2015, 38(3): 107-110.

[5] Bubela T, Golder R, Graff G D, et al. Patent landscaping for life sciences innovation: toward consistent and transparent practices[J]. Nature biotechnology, 2013, 31(3): 202-206.

[6] Smith J A, Arshad Z, Thomas H, et al. Evidence of insufficient patent quality of reporting in patent landscapes in the life sciences[J]. Nature biotechnology, 2017, 35(3): 210-214.

[7] Li Li, Liu Zhiyuan, Sun Maosong. Research on automatic phrase paraphrase extraction based on Chinese-English parallel patent corpora[J]. Journal of Chinese Information Processing, 2013, 27(6): 151-158.

[8] Lou Yan, Zhang Shang, Huang Lucheng. Research on alternative technology selection based on patent analysis[J]. Science and Technology Management Research, 2015, 35(20): 150-154.

[9] Chen Yunwei, Fang Shu. Research on social network analysis methods for patentee association networks[J]. Library and Information Knowledge, 2011(3): 58-66.

[10] Wang Xin, Zhao Yunhua, Gao Fang. Research on patent similarity measurement method based on classification numbers and citations[J]. Digital Library Forum, 2015(01): 57-62.

[11] Zhu Lei, Jin Hai, Zheng Ran, et al. Appearance design patent retrieval based on shape semantics[J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(3): 372-378.

[12] Wang Jin, Sun Yong, Wang Congwei. Text similarity algorithm based on domain ontology[J]. Journal of Soochow University (Engineering Science Edition), 2011, 31(3): 13-17.

[13] Chen Liang, Zhang Haichao, Yang Guancan, et al. Patent representation method using KnowledgeGraph and its application[J]. Library and Information

Service, 2017, 61(9): 123-129.

[14] Jiang Lixue, Ji Duo, Cai Dongfeng. Patent term similarity calculation method based on semantic roles[J]. Journal of Chinese Information Processing, 2016, 30(4): 37-43.

[15] Xu Haiyun, Wang Zhenmeng, Hu Zhengyin, et al. Review of key technologies for identifying technology themes using patent text analysis[J]. Information Studies: Theory & Application, 2016, 39(11): 131-137.

[16] Rao Qi, Wang Peiyan, Zhang Guiping. Chinese patent SAO structure extraction for text analysis[J]. Journal of Chinese Information Processing, 2015, 29(5): 151-158.

[17] Yang Hongzhang, Fu Jing. Method for constructing intelligent semantic retrieval system for patent information using structured features of patent texts[J]. Information Studies: Theory & Application, 2015, 38(4): 136-138.

[18] Chen Liang, Yang Guancan, Zhang Jing, et al. Research on multi-main path method for technology evolution analysis[J]. Library and Information Service, 2015, 59(10): 124-130, 115.

[19] Liao Liefa, Le Fugang, Zhu Yalan. Application of LDA model in patent text classification[J]. Journal of Modern Information, 2017, 37(3): 35-39.

[20] Wu Yuying, Ma Yuxiang, Zhai Dongsheng. Research on Chinese patent infringement detection based on SOM[J]. Journal of Intelligence, 2014, 33(2): 33-39.

[21] Xu Kan, Lin Yuan, Qu Chen, et al. Research on word vector method for patent query expansion[J]. Computer Science and Exploration, 2017(9): 1-9.

[22] Wang Yanyan, Wang Peiyan. An entity disambiguation method for patent entities[J]. Journal of Shenyang Aerospace University, 2015, 32(1): 77-83.

[23] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//Proceedings of the 31st international conference on machine learning (ICML-14). [EB/OL]. [2014-01-27]. http://proceedings.mlr.press/v32/le14.html.

[24] Řehůřek R, Petr S. Software framework for topic modelling with large corpora. [EB/OL]. [2018-05-22]. https://is.muni.cz/publication/884893/en.

[25] United States Patent Trademark Office. Patent Grant Full Text Data/XML Version 4.5 ICE (JAN 2017 - DEC 2017) [EB/OL]. [2017-12-26]. https://bulkdata.uspto.gov/data/patent/grant/redbook/fulltext/2017/.

Author Contributions

Cao Qi: Responsible for paper concept and framework construction, main content writing, experiment implementation, and result analysis and discussion.
Zhao Wei: Responsible for literature review, participation in paper framework construction and experimental result analysis and discussion.

Zhang Yingjie: Participation in model preprocessing, analysis and discussion.
Zhao Shujun: Participation in experimental model data collection, preprocessing, result analysis, and result refinement.
Chen Liang: Responsible for data proofreading and providing revision suggestions during paper writing.

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