ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 143

2nd International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2018)

Automatic Indexing of Patent Right-claiming Document Based on Deep Learning

Qinghong Zhong, Xiaodong Qiao and Yunliang Zhang*

Institute of Scientific and Technical Information of China, Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, SAPPRFT

*Corresponding author

***Abstract*—In recent years, there have been more and more applications of deep learning in natural language processing, and people have paid more and more attention to the value embodied in patent. In this paper, based on the value of right-claiming document in the patent, a deep learning tool word2vec is used to convert text information into a set of word embeddings. The word embeddings carry semantic information, so that the quantified metrics the relationships between words. Then, the k-means clustering method is used to extract words whose distance between words is closer to the center of the cluster, so as to achieve the purpose of automatic indexing the right-claiming document.**

Keywords—deep learning; word2vec; k-means; automatic indexing; keyword extraction

I. INTRODUCTION

At present, the number of patents continues to rise, and the processing of vast amounts of patent information becomes more and more important. How to quickly retrieve and make full use of patent information becomes a challenge. Not only that, patent as the most effective carrier of technical information, which contains economic value, while patent infringement cases are also increasing. The existing patent infringement judgment is mainly based on the provisions of article 56 of the People's Republic of China Patent Law: "The scope of the protection of the patent for invention or utility model is subject to the content of its rights, and the instructions and the attached drawings can be used to explain the claims." This provision states that in the determination of patent infringement, the scope of patent protection is subject to the terms of the right-claiming document. This text information covers the content required to protect the patent, has the direct legal effect, is the core of the patent application, and is also an important legal document to determine the scope of the patent protection. Therefore, it is very important for the full utilization and fast retrieval of patent right-claiming document.

This paper proposes an automatic indexing method based on deep learning, which aims to help patent examiners and users to narrow the scope of search and find similar patents accurately. At the same time, it also has high application value in automatic classification, classification and theme assisted retrieval, and automatic summarization and so on.

This paper is based on the keyword clustering algorithm of the deep learning tool word2vec, which maps words into a more abstract vector space during the training of the language model. In the environment of big data, it can be considered that the distance between two points in the vector space corresponds to the degree of similarity between two words. To a certain extent, this has solved how to use computer to express words[1]. The K-means clustering algorithm is used to cluster the numerically expressed words. Word2Vec-generated vectors are used to calculate the similarity between words, and the closest word to the clustering center are selected as the text keyword.

II. INTRODUCTION OF RESEARCH STATUS

Automatic indexing technology can be divided into two main methods: automatic extracting indexing and automatic assignment indexing. At present, research at home and abroad mainly focuses on the former one and this article also the same. It is also called automatic keyword extraction, that is, it can extract feature vocabularies or keywords that can represent the full-text subject content from various types of text information such as literature, documents, and web pages, and finally be converted into a controlled vocabulary by the system [2].

At present, the method of automatic indexing can be roughly divided into a probabilistic statistical analysis method, a grammatical semantic analysis method, and an artificial intelligence analysis method. Since 2000, the research on automatic indexing in China has focused on the group of authors with Hou Hanqing as the center and has achieved rich results. He built a web page-based automatic indexing and automatic classification system based on the knowledge base [3], and based on multi-word list, "Chinese Library Classification" knowledge base and other automatic indexing experiments. In addition, many researchers have conducted research on other methods, such as genetic algorithm, conditional random field(CRF), decision tree, ontology-based, exploring complex networks, and rough sets theory, which have made some innovations and contributions to automatic indexing of documents [4].

Zhang Chengzhi did a lot of research on the automatic indexing method based on machine learning. The machine learning method is mainly to model the data and obtain the parameters through the training data, thus completing the automatic indexing work. In [5], the document indexing problem is converted into a sequence labeling problem, and an automatic indexing model based on CRF is proposed. The experimental results show that this method is the best way to solve the sequence labeling problem at that time. In [6], he

regards the automatic extraction of keywords as a classification problem, integrates the strengths of the machine learning model and the ensemble learning, and uses a multi-classification model for comprehensive voting, which improves the precision and recall of indexing results. In [7], a general automatic indexing evaluation model is proposed, which uses the principle of similarity to modify the problem of ignoring semantic relations in the traditional evaluation model and pays attention to the use of semantic resources to make the model have a certain degree of reliability.

Research in recent years has mainly focused on the research of automatic indexing systems, and it tends to focus on a specific area, such as the fields of medical information and film labels. Wang Jing and Jiang Peng summarized the flow of the four automated indexing systems in [8] and found that the methods in these four systems mostly used TF / IDF word frequency statistical characteristics and word length, location and other characteristics. The other research is to study the function and application of automatic indexing in knowledge organization, knowledge service and thesaurus construction.

The relevant foreign studies are mainly the automatic text indexing of using the SKOS vocabulary by Bueno-de-la-Fuente; Automatic indexing of web pages; Golub, Koraljka proposes a framework for evaluating automatic indexing and classification in search environments.

The deep learning method used in this paper is an artificial intelligence analysis method. This method currently uses machine learning methods. In recent years, deep learning has achieved great success in image recognition and speech recognition. This paper focuses on the application of deep learning technology in natural language processing(NLP). In 2008, the American NEC Institute, which first introduced deep learning to NLP, constructed a network model by mapping words to one-dimensional vector space and multi-layer one-dimensional convolutional structures to solve four typical NLP problems, including Named Entity Recognition(NER), participle, Part-of-Speech tagging and Semantic Role Labeling are all fairly accurate [9]. Mikolov has been making various improvements on the Recurrent Neural Network Language Model (RNNLM). In 2012, he optimized and summarized RNNLM in its doctoral thesis [10]. After entered Google to continue research, based on this research experience, it opened source word2vec in 2013. The rise of Word2vec has led to the study of distributed characteristics of words [11]. Subsequent papers and experiments by Mikolov [12], LEVY O [13] and SCHNABEL T [14] have proved the high availability and excellent performance of word2vec. However, word2vec also has some disadvantages. On the one hand, the sequence of Chinese or English sentences is very important for prediction, while the word2vec model is mainly based on the size of window T to prediction, so the sampling length of data is limited. On the other hand, the word2vec model has no memory and cannot selectively retain important information in the sentence [15]. Of course, this does not prevent word2vec from being widely used in NLP. Word2vec is mostly used in sentiment analysis, text feature extraction, similarity analysis, event identification, and auxiliary information retrieval.

There is less research on patent indexing at home and abroad, and the indexing of right-claiming document is even more limited. The automatic indexing of Chinese patent right-claiming document mainly focuses on the study of its segmentation algorithms and the identification of similar patents. For example, the group of authors with Zhai Dongsheng as the center proposed a Chinese word segmentation model based on a combination of domain dictionaries and rules for Chinese patent claims [16]; Zhang Jie's Chinese similar patent recognition algorithm based on SAO structure. Zhang Jing's [17] analyzes the content elements of patents to find out that the development direction of the indexing work in the information age is automatic indexing, but the current level of technology determines that the indexing of certain key information must be added to human intervention. Research abroad on patent is mainly about legal claims, use of patent information to forecast the market, valuation of patents, and identification of changes in technology topics.

The main result is that some intellectual property companies, and the latest is that in 2017, Guangxi dawning Intellectual Property Service Company Limited announced the "patent automatic indexing tool software V1.0". The tool can automatically index patent, automatically extract patent information, and conduct statistical analysis and visual display of indexed data.

By analyzing the domestic and foreign research status in the fields of automatic indexing, and deep learning, patent indexing, it can be seen that the technologies in these fields tend to mature, but the automatic indexing of patent right-claiming document is mostly based on statistics and based on grammatical semantic analysis methods. There are relatively little research on automatic indexing using deep learning.

III. RESEARCH CONTENT

A. Word2Vec Word Embedding Model Training

The word representation in the traditional language model is string-oriented and does not contain semantic information. Neural network language models are represented by vectors and can store semantic information. The vector representation of semantically similar words is similar in the angle value and distance. Word2vec is the input of neural network language model and is the intermediate result of neural network learning and training two language models. There are Continuous Bag-Of-Words Model (CBOW) and Skip-gram Model (Skip-gram). The CBOW is called a continuous bag-of-words model, which does not consider the word order problem, and uses the context of a word as input to predict the probability of occurrence of the word itself. The Skip-gram takes a word as input to predict the probability of occurrence of its contextual word. The CBOW is more useful for smaller datasets, and the Skip-gram is suitable for larger datasets, and now more commonly used is the Skip-gram. In the process of training the model, two algorithms of Hierarchical Softmax and Negative Sampling are used to perform word embedding training [18]. These two algorithms are word2vec training techniques used to improve computational efficiency and improve model quality. The Python version only uses the skip-gram and the Hierarchical Softmax method to train the corpus. It takes the text set as input

and establishes the semantic relationship of the word based on the contextual relationship of the word. The larger the data set, the closer the information reflected by the training generated word embedding is to the semantics of the data set.

B. K-Means Clustering Method

The K-means clustering method is a simple representation learning algorithm that belongs to the unsupervised learning method. K-means clustering initializes k different center points $\{\mu^{(1)}, \cdots, \mu^{(k)}\}$, then iteratively exchanges step one and two until the function converges. Step one, according to the similarity distance formula, each training sample is assigned to the cluster i represented by the nearest center point $\mu^{(i)}$. Step two, each center point $\mu^{(i)}$ is updated to the mean of all training samples $x^{(j)}$ in cluster i [19]. Each of the center points μ is a word obtained in the previous step, and then the average value of the training sample $x^{(j)}$ is calculated using the value of the word embedding, ie, the value of the word embedding is used to calculate the cosine angle to measure the degree of similarity between words. The similarity of the objects in the same cluster is higher; but the similarity of objects in different clusters is small. After repeated iterations, we get k nearest words from the cluster center, which is K keywords [20].

IV. EXPERIMENT PROCESS

A. Experimental Data

A total of 200049 pieces of right-claiming data for patents in the field of nanotechnology were extracted, which is the first item in the right-claiming of each patent. In addition, in order to enhance the experimental results, 50 test data including complete claims and abstracts were added. The file size is 136MB for subsequent word segmentation and word2vec word embedding model training.

B. Experimental Evaluation Index

At present, the evaluation criteria of keyword extraction algorithm effectiveness include Precision (P), Recall(R), and F value. P and R indicators sometimes appear contradictory situations, and F value can perform weighted reconciliation of P and R indicators. The larger the values of P, R, and F, the better the experimental results. Calculated as follows:

$$P=A/B \tag{1}$$

$$R=A/C \tag{2}$$

$$F\text{-measure}=2PR/(P+R) \tag{3}$$

$$A=\text{Number of correct labeled keywords} \tag{4}$$

$$B=\text{Total number of experimentally extracted keywords} \tag{5}$$

$$C=\text{Total number of manually indexed keywords} \tag{6}$$

C. Experiment Process

The experimental process is mainly the following three steps.

1) Text preprocessing: Because the experiment selects the data in the nanometer field, the content contains many chemical symbols, chemical molecular formulas, and chemical nouns. In order to ensure the effectiveness of word segmentation, there is a need for a chemical domain vocabulary. After searching, there was no authoritative chemical domain vocabulary, so I defined a vocabulary that contains 14985 words in the chemical field. Considering that some words with high frequency appearing but little practical significance may interfere with the experimental results, and the characteristics of patent data are more obvious, so stop word list is added. The selected stop word list is the "stop word dictionary of Harbin Institute of technology". In the course of the experiment, the punctuation and special characters were added. After removing and adding some words that had an effect on the experimental results, a stopword vocabulary suitable for this experiment was sorted out. There were a total of 1364 stop words. The word segmentation tool selects the Jieba, which is a Chinese word segmentation module developed by domestic programmers using Python. After about four hours of operation, a series of words separated by spaces were finally obtained. The segmented text size is 114MB.

2) Training word embedding: Use the word2vec tool in the Gensim theme model package in Python 3.6.3 to train the corpora to calculate the word embedding, set the dimension to be 400, the window value to be 5, and min_count to be 5. After about fifteen minutes, the files containing 39578 word embeddings are obtained.

3) Automatic indexing: The two groups of experimental data are preprocessed first, word embeddings are trained, then k-means clustering is performed to obtain the keywords of the clustering center. In this experiment, k=5, 7, 10, 12 are set, that is,index 5, 7, 10, 12 key words respectively. However, there are a few shorter data clusters less than the number of settings.

V. EXPERIMENTAL RESULTS AND ANALYSIS

Since the patent does not include keywords, 12 keywords are manually indexed for the right-claiming document of the 50 patents randomly selected. However, manual indexing is subjective and the evaluation is inconsistent. Learned that the Chinese military-civilian integration platform (SOOIP) provides keyword association analysis, a considerable number of patent have indexed keywords, and the SOOIP platform has the advantages of authoritative data resources and a wide range of data coverage. After selecting keywords and retrieving them in SOOIP, it is found that the keyword source on the platform is the abstract of patent. In order to ensure the consistency of data sources, a set of control groups with patent abstracts was

set up to further test the experimental effect of this method. Two groups of experiments were conducted, and 50 patent data were indexed by 5, 7, 10 and 12 keywords respectively. The first group compares the results of manual indexing with the results of the right-claiming document indexing; the second group compares the SOOIP platform keywords with the abstract indexing results. The P, R, and F value of the two groups of experimental results were obtained, and the results are shown in Tables 1 and 2.

TABLE I. RIGHT-CLAIMING DOCUMENT OF AUTOMATIC INDEXING RESULTS CALCULATION

	Number of keywords: K			
	5	***7***	***10***	***12***
P	0.552	0.4971	0.4809	0.4824
R	0.23	0.29	0.3983	0.4783
F	0.3247	0.3663	0.4357	0.4803

TABLE II. ABSTRACT OF AUTOMATIC INDEXING RESULTS CALCULATION

	Number of keywords: K			
	5	***7***	***10***	***12***
P	0.4737	0.449	0.4454	0.4437
R	0.1422	0.1871	0.2625	0.3062
F	0.2187	0.2642	0.3303	0.3624

As can be seen from Table 1, the P of a group of claim items can reach about 50%. As the number of index words increases, the P decreases. When k=12, it rises slightly. The R keeps the value of the denominator constant, and its value increases with the number k. In the other group, as the index number increases, the P decreases. The P of the first group is slightly higher than that of the second group. The reason for the analysis is that the keywords of the SOOIP platform should also be the result of machine automatic indexing, which the difference between the indexing results is caused by different indexing methods. Analyzing the results of the two groups of experiments can be considered that this method is more accurate when the number of indexes is small. Comparing the indexing results from the same patent further analyzes the features of the right-claiming document indexing. The results are shown in Table 3 and Table 4.

TABLE III. COMPARISON OF AUTOMATIC INDEXING RESULTS OF RIGHT-CLAIMING DOCUMENT

Manual indexing	ligand,composition,polycarboxylic acid,long chain,acetic acid,alkyl,alkenyl,alkynyl,photoluminescence,quantum dot, valeri cacid, quantum yield
System indexing	govern,alkynyl,carbene,quantumyield,photoluminescence, branching,former,pentene,embrace,flock,long chain, valeric acid

TABLE IV. COMPARISON OF AUTOMATIC INDEXING RESULTS OF ABSTRACT

SOOIP Keywords	nano structure,shine,ligand,compound,size distribution, polycarboxylic acid, carboxylate ligan,dicarboxylic acid,shell formation,quantum dot, productivity,enrichment,quantum,preparation,launch, photoluminescence, wave length
System indexing	Photoluminescence,narrow,quantum,enrichment,altitude,size,construction,offer,ligand,indium,include,quantum yield

Comparing the two sets of indexing results, it can be found that the indexing of the right-claiming document can better index some field of proper nouns, such as "alkynyl, valeric acid, pentene", etc. This is of great benefit for the accurate retrieval of patents and can help the users to make more accurate inquiries of the knowledge points and have a more independent retrieval meaning for nouns. In addition, it can be seen that word segmentation results directly affect the indexing effect, such as "size" and "size distribution", "quantum yield" and "quantum; yield". So, in a particular field of indexing, it is better to have an authoritative domain vocabulary, which can be more accurate in identifying proper nouns and NER, thereby improving the indexing effect.

VI. CONCLUSION

This paper introduces the principles and applications of the word2vec tool and the k-means clustering algorithm and conduct experiment. Through this experiment, it is found that the open source tools and methods of deep learning in NLP are relatively mature. This method of generating word embedding can eliminate the process of artificial discovery and determination of word characteristics. By placing the word in a more abstract space and turning the text into numerical information, the clustering can quickly extract the nearest key words from the distance clustering center by continuous iterative calculation, so as to achieve the function of automatic indexing of the target literature.

From the process of textual preprocessing, it is found that when doing patent automatic indexing, the professional vocabulary in a specific field is very important, which directly affects the effect of word segmentation and the word embedding generated. Therefore, in the future experiment, we can choose to experiment in the field of professional domain vocabularies in order to reduce the experimental error caused by the incompleteness of the custom dictionary. The patent literature is different from the paper literature. The patent does not have the keywords indexed by the author himself. Many of the existing patent with keywords are indexing based on the abstract or the instruction, and there are hardly any patent that specifically indexing right-claiming document. Through experiments, it has been found that the indexing of claim can better reflect the characteristics of a particular field and can identify more domain specific nouns. In turn, it can also be considered that the indexing of claims can not only assist in retrieval, but also can play a role in the construction of domain vocabularies and can help identify proper nouns in the domain and the NER.

ACKNOWLEDGEMENTS

This work is partially supported by ISTIC Key Project Program (Grant No. ZD2018-07), CKCEST Project Program (Grant No. CKCEST-2018-1-26) and National Digital Composite Publishing System Project (Grant No. XWCB-ZDGC-FHCB/28). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Yuepeng Li, Cui Jin, and Junchuan Ji, "Keyword extraction algorithm based on word2vec(In Chinese)," E-science Technology & application, vol. 6(4), pp:54–59, 2015.

[2] Aiqing Xu, "Research and Improvement of Automatic Indexing Technology of Text Information(In Chinese)," Wuhan University of Technology, 2013.

[3] Hanqing Hou and Pengjun Xue, "Design of Automatic Indexing and Automatic Classification System of Web Pages Based on Knowledge Base(In Chinese)," Journal of Academic Libraries, vol. 22(1), pp. 50–55, 2004.

[4] Fengming Yu, "A Summary of Research on Domestic Automatic Indexing from 2000 to 2009(In Chinese)," Information Research, vol. 5, pp. 28–31, 2011.

[5] Chengzhi Zhang, Xinning Su, "Research on Automatic Indexing Model Based on Conditional Random Field(In Chinese)," Journal of Library Science in China, vol. 34(5), pp. 89–94, 2008.

[6] Chengzhi Zhang, "Research on Automatic Indexing Method Based on Integrated Learning(In Chinese)," Journal of the China Society for Scientific and Technical Information, vol. 29(2), pp. 16–23, 2009.

[7] Chengzhi Zhang, Dongmin Zhou, "Research on General Evaluation Model of Automatic Indexing(In Chinese)," Journal of the China Society for Scientific and Technical Information, vol. 28(1), pp. 40–47, 2009.

[8] Jing Wang, Peng Jiang, "Comparison and Analysis of Automatic Indexing Systems(In Chinese)," Sci-Tech Information Development & Economy, vol. 2(9), pp. 17–21, 2017.

[9] Xianchang Chen, "Research on Deep Learning Algorithm and Application Based on Convolutional Neural Network(In Chinese)," Zhejiang Gongshang University, 2014.

[10] Mikolov T A, "Statistical Language Models Based on Neural Networks," 2012.

[11] Yiou Lin, Hang Lei, Xiaoyu Li and Jia Wu, "Deep Learning in Natural Language Processing: Methods and Applications(In Chinese)," Journal of University of Electronic Science and Technology of China, vol. 46(6), pp. 913–919, 2017.

[12] T. Mikolov, I. Sutskever, K. Chen., G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," International Conference on Neural Information Processing Systems, pp. 3111–3119, 2013.

[13] O. Levy, Y. Goldberg, "Neural word embedding as implicit matrix factorization," Advances in Neural Information Processing Systems, pp. 2177–2185, 2014.

[14] T. Schnabel, I. Labutov, D. Mimno and J. Joachims, "Evaluation methods for unsupervised word embeddings," Conference on Empirical Methods in Natural Language Processing, pp. 298–307, 2015.

[15] Bingya Wu, Miao Wei, "Reviewing natural language processing word embedding method from deep learning(In Chinese)," Computer Knowledge and Technology, vol. 12(36), pp. 184–185, 2016.

[16] Dongsheng Zhai, Wenshan Ma, "Research on Word Segmentation Algorithm in Chinese Patent Claims(In Chinese)," Journal of Intelligence, vol. 30(11), pp. 152–155, 2011.

[17] Jing Zhang, "Research on Quality Control Mechanism of Content Indexing of Patent Analysis Report(In Chinese)," Journal of Modern Information, vol. 37(5), pp. 33–36, 2017.

[18] Lirong Zhang, "Design of Chinese Automatic Indexing Algorithm and Its Application in Internet Public Opinion Monitoring(In Chinese)," Hebei University of Science & Technology, 2017.

[19] Lan Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, 1rd ed., vol. 6. 2017, pp.91–94.

[20] Wenchao Deng, Peng Xu, "Research on Clustering Chinese Words Using word2vec(In Chinese)," Computer engineering & Software, vol. 12, pp. 160–162, 2013.