Power patent classification method based on deep neural network

HUA Min *
Shanghai Municipal Electric Power Research Institute, Shanghai, 200437, China
*Corresponding author’s e-mail: 4200826@qq.com

Abstract. Aiming at the problems in power patent classification, a power patent classification method based on deep neural network is proposed. Firstly, the power patent is analyzed for preprocessing. Then, the keywords in the patent abstract are extracted to construct a power vocabulary dictionary. Then, the power patent classification model is built according to the deep neural network. Finally, the power patent classification model based on deep neural network is used to classify the power patents, and the accuracy is close to 98%. It can be seen that the deep neural network method can not only realize the classification of power patent, but also has good classification effect.

1. Introduction
Patent documents have great research value. If we can accurately analyze patent documents, we will reveal important technical details and relationships, and illustrate business trends. Novel industrial solutions will also be inspired and proposed, so as to make crucial investment decisions. Therefore, we must carefully analyze patent documents and make use of the value of patents. Effective patent classification can improve the quality and efficiency of patent daily management, and deeply excavate and utilize useful patent information, such as patent strategy formulation or improvement, technical subject research and patent value analysis. With the rapid development of intelligent technology, scholars and practitioners have done some research on automatic patent classification, and achieved some results. However, the existing patent automatic classification method is not accurate enough for the power industry patent classification. The effective collection, collation and analysis of power patent data is becoming more and more important for the development of the power industry. Therefore, the research on the patent classification method of the power industry has important social benefits and economic value for the realization of the practical application of intelligent technology in the power industry.

From the perspective of classification methods, patent automatic classification can be divided into three categories: classification methods based on specific rules, citation relationship and text content mining. Based on specific rule classification, such as C. He [1] Based on association rule mining method to identify category rules, and then build an automatic classifier; based on Citation relationship, such as S. Chang [2] Based on patent citation relationship to cluster patents and interpret the technologies involved in the cluster to build a classification system, K Lai et al. [3] established a classification system based on the co citation relationship of basic patents by using the method of factor analysis; there are a large number of researches on text content mining classification, which continue to receive attention, and are discussed in detail below.
Automatic classification based on patent text content mining belongs to the task of text classification in natural language processing (NLP). The classic method of this process is to use machine learning method to determine the potential basis features of patent classification by means of Feature Engineering, and then use machine learning algorithms such as Bayesian classifier, SVM and logical regression to carry out automatic classification. The common feature of this kind of method is bag of words feature, that is, the bag of words model is used to represent the patent text as the word frequency vector of the words contained [4]. However, due to the high-frequency noise problem of invalid words (such as function words, conjunctions and other functional words) caused by simple word frequency representation, the method of using TF-IDF to replace the word frequency in the original vector is widely used, such as Jia Shan Shan et al. [5] used TF-IDF features extracted from patent applications to train naive Bayes, support vector machine and AdaBoost classifiers to predict IPC classification number. In addition, some new features are constantly introduced to improve the classification effect. For example, S. Verberne et al. [6] added the semantic triplet information of feature words on the basis of patent feature words to improve the classification accuracy. Stutzki et al. [7] introduced the geographic data features of patent applicants, used KNN and SVM classifier with one versus rest strategy for patent classification. S. Lim et al. [8] simultaneously extracted features from title, abstract, claim, technical field and background technical information to improve the effect of patent text classification. Patent automatic classification based on classical machine learning method relies on researchers to construct features manually to achieve better classification effect [9]. However, the feature representation represented by bag of words model loses semantic information such as word meaning information and word order information in patent text. For example, two documents of the same category may not be accurately classified due to different word description methods. In recent years, with the rise of deep learning technology and its continuous application in patent information research, a series of research results have been produced in the research scene of patent automatic classification based on patent text content mining. For example, Ma Shuanggang [10] designed an automatic classification method based on the combination of de-noising automatic encoder (DAE) and SVM algorithm based on deep learning theory, and selected six IPC categories in the computer field to verify the classification effect.

To sum up, domestic and foreign research on the improvement and application of patent classification methods based on deep learning technology has made some achievements, but at present, most of the patent automatic classification or retrieval are combined with IPC (International Patent) With the development of technology, in many fields, such as smart phones, drugs, finance, etc., automatic patent classification and retrieval are facing challenges such as cross domain and complex technical system, which makes the effect of traditional classification and retrieval methods based on IPC number not ideal [11]. As a cross domain power industry, the research on automatic patent classification of power industry is very important There are two difficulties in the automatic classification of patents:

1) The power industry covers a wide range, and the IPC retrieval method is difficult to completely cover all areas of the power industry, and it is difficult to directly classify according to the traditional IPC number.

2) With the rapid development of power technology, the number of innovation achievements and patents in related fields continues to grow. It is necessary to build an appropriate classification system to accurately identify different types of power technology innovation.

Therefore, this paper focuses on the shortcomings of traditional methods in power patent classification.

2. Patent classification pre-processing

The natural language domain where the text classification problem is located has its unique feature processing logic, and most of the work of traditional text classification task is also here. Since the core idea of a patent abstract is concentrated in the patent abstract, this topic carries out feature engineering related work based on the patent abstract, including three parts: text pre-processing, feature extraction and vocabulary construction. The ultimate goal is to transform the text into a computer understandable format.
2.1. Text preprocessing

Power patent text pre-processing mainly includes four steps, which are described as follows.

The first step is to clean the mastered power patent data. The main purpose of this process is to de-duplicate the mastered data and manually interpret and classify them to obtain higher classification accuracy.

The second step is word segmentation. The reason for word segmentation is that many studies show that the feature granularity is word granularity, which is much better than word granularity. In fact, it is very easy to understand, because most classification algorithms do not consider word order information. Based on word granularity, it obviously loses too much "n-gram" information. This topic intends to use Jieba for word segmentation of patent abstracts.

The third step is part of speech tagging. The main purpose of this process is to eliminate word ambiguity and strengthen word features.

The fourth step is to stop words. This process aims to filter out invalid words [De], [Ba], [But] by establishing a stop word list.

2.2. Feature extraction

TF-IDF is a commonly used text feature vectorization method to evaluate the importance of words to a document in a corpus. The advantages of this method are as follows:

First, the influence of high frequency words is considered.

Second, the inverse text frequency is used to weight the word frequency, that is, the more a word appears in the document, the more important it is. However, if it appears in more documents, the less important it is in the classification process, which reflects the ability of a word to distinguish the current document from other documents.

The specific calculation formula of TF-IDF is as follows:

\[
TF_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}
\]

Where \(n_{i,j}\) is the word in document \(d_j\), and the denominator is the total number of times all the words appeared in file \(d_j\).

\[
IDF = \log \frac{|D|}{1+|\{j: t_i \in d_j\}|}
\]

Where \(|D|\) is the total number of documents in the corpus, \(\{j: t_i \in d_j\}\) denotes the number of documents inclusive word \(t_i\).

\[
TF-IDF(w) = TF(w) \times IDF(w)
\]

TF-IDF is an unsupervised feature extraction algorithm. In this paper, the combination of TF-IDF and supervised method is used for feature extraction, that is, the keywords in the patent abstract are extracted by TF-IDF, and the patents are classified and labelled manually to add labels for each patent. Then, the key words and label are coded by dictionary function and one hot coding method respectively to complete the transformation from text to vector. Thus, the keywords extracted from the power patent abstract are constructed into a vocabulary, which is convenient for the expansion of the following words and the training of the classifier.

3. Deep neural network

Convolutional neural network (CNN) is a kind of artificial neural network, which is used in handwritten digit recognition by Fukushima. With the in-depth research and continuous improvement of convolutional neural network, convolutional neural network has been successfully applied in image processing, pattern recognition and other fields. It not only achieves good results in the field of computer vision, but also can be applied to text classification.

It can be said that the convolutional neural network was first applied to text classification in 2014. In this part, we need to determine the model structure of CNN, including the number of convolution layer, pooling layer and pooling mode. CNN generally uses one-dimensional convolution for text classification,
that is, conv1d. At present, the structure of CNN model used for text classification is relatively simple, as shown in Figure 1. The model consists of only one convolution layer, one pooling layer and two fully connected layers.

![Figure 1. CNN classification model of power patent.](image)

### 4. Patent classification experiment

In order to test the proposed method, we select 290 patents from power patent data. Firstly, the above patents are divided into three categories by manual interpretation. After text pre-processing and feature extraction, a vocabulary which can be used for power patent three classification is constructed. Then, based on CNN neural network, the three-classification model of power patent is constructed. The specific process of model construction is as follows.

1. Key words are extracted from the abstracts of 290 patents.
2. After removing the same keywords, according to the frequency of these keywords, the number of feature words is determined to 703.
3. The electric power vocabulary dictionary is composed of characteristic words vocab.txt file.
4. Ten keywords with the highest frequency are extracted from each patent as features, and the patent category is marked by manual annotation. The results are shown in Figure 2.
5. The `dict` function is further used to encode the words in the dictionary, as shown in Figure 3.

![Figure 2. Patent keywords and classification labels.](image)
(7) 75% of all patent data sets are used as the training set, and the matrix corresponding to keywords is used as \( X \). The corresponding category of patent is label as \( y \). Train for CNN network training. The training process is shown in Figure 4.

(8) Finally, the remaining 25% data is used as the test set to test the patent classification effect of the deep neural network model. The classification accuracy is shown in Figure 5.
As can be seen from the above figure, the accuracy of using CNN to classify power patent intelligently is close to 98%. It can be seen that the method proposed in this paper can achieve the accurate classification of power patents.

5. Conclusion
In this paper, deep learning method is used to classify power patent. In the case of data preprocessing and keyword extraction, convolution neural network is used to build a power patent classification model. Through the classification experiment, we know that the deep neural network can achieve good results in patent classification. In the following work, we will test more patent samples to improve the deep learning classification model.

References
[1] Cong H, Han T L. Pattern-oriented associative rule-based patent classification.(2009) Expert systems with applications, 37(3): 2395-2404.
[2] Shann-bin C, Kuei-kuei L, Shu-min C. Exploring technology diffusion and classification of business methods: Using the patent citation network.(2008) Technological forecasting and social change, 76(1): 107-117.
[3] Lai K K, Wu S J. Using the patent co-citation approach to establish a new patent classification system.(2005) Information processing and management, 41(2): 313-330.
[4] Fall C J, Torcsvari A, Benzineb K, et al.(2003) Automated categorization in the international patent classification. In: Acm Sigir Forum. New York. pp.10-25.
[5] Jia S S, Liu C, Sun L Y, Liu X A, Peng T. Patent Classification Based on Multi-feature and Multi-classifier Integration. (2017) Data Analysis and knowledge Discovery, 1(08): 76-84.
[6] Verberne S, Vogel M, Dhondt E K L. (2011) Patent classification experiments with the linguistic classification system LCS. In: CLEF-IP 2011.Amsterdam.
[7] Stutzki J, Matthias S.(2016) Geodata supported classification of patent applications. In: Proceedings of the third international ACM SIGMOD workshop on managing and mining enriched geo-spatial data. San Francisco.pp.1-6.
[8] Yun J, Geum Y. Automated classification of patents: A topic modeling approach.(2020) Computers and industrial engineering, 147:106636.
[9] Lim S, Kwon Y J. (2016) IPC multi-label classification based on the field functionality of patent documents. In: SIGIR Forum. Gold Coast. pp. 677-691.

[10] Ma S G. (2016) The Study of Automatic Chinese Patent Classification Based on Deep Learning Theory and Method. Jiangsu University, Zhenjiang.

[11] XIE J J, HONG L J. Patent Application Quick Retrieval Skills in the Field of Pharmaceutical. (2018) China Invention & Patent, 15(12): 125-128.