Research on Chinese Patent Infringement Detection Algorithm Based on Semantic Extended Vector Space Model

Shuangyuan Zou[a], Wenxi Zheng[b] and Min Wu[c]

Center of Modern Educational Technology, University of Science and Technology of China, Hefei 230026, China

[a]zshy@mail.ustc.edu.cn, [b]wxzheng@ustc.edu.cn, [c]minwu@ustc.edu.cn

Keywords: Chinese patent, patent infringement, VSM, Word2Vec

Abstract: As the core-competitiveness of enterprises, patent infringement litigation is becoming more frequent at present. Once companies are involved in patent infringement disputes, they usually spend a lot of time and energy to deal with it. In this paper, the Word2Vec word vector model is trained by some Chinese patent data samples and other Chinese corpus such as Wikipedia, and use this model to calculate the similarity between Chinese words. Then use traditional vector space model (VSM) and Word2Vec model to calculate the similarity of Chinese patent claims. Using similarity to determine whether a patent is infringed. This paper provides an effective method for the retrieval of patent infringement related personnel patent examiners and patent owners, etc.

1. Introduction

With the development of economic globalization and fierce competition in the market, the most important point of enterprise is technological innovation, and high-quality patented technology is the core-competitiveness of enterprises. Due to the promotion of the market, the number of patents in China is exploding. According to the statistics of the China Intellectual Property Office, in 2017, the number of Chinese patent applications was 1.382 million, that is a big number.

With the increase in the number of patents, patent infringement litigation has become more frequent. In order to protect their patents and prevent the risk of patent infringement, the company needs a reliable system for patent infringement retrieval. However, the most of the current patent search systems are based on Boolean search model, such as the State Intellectual Property Office of China and the China Patent Information Network. These systems do not provide infringement analysis.

The basic principles of patent infringement include over all coverage, equivalence and no remorse [1]. The methods for patent similarity determination can be divided into two categories: citation-based method and text-based method. This paper analyzes the latter method. Here is some previous research in this direction. Bergmann [2] used the sentence structure to calculate the patent infringement, and used the DNA chip technology patent infringement case as an example to confirm the feasibility of this method. Lee [3] used the traditional space vector to calculate patent similarity, and got a good result. Taghaboni Dutta [4] used the cosine formula to calculate the similarity between patents. Moehrle [5] combined the methods, processes, and results in the patent to analyze the similarities between patents. Tseng [6] used SAO structure to express patent content. Magerman [7] calculated the similarities between patent documents and scientific publications by mining the semantic information. Yoon [8] used SAO to build a patent map for patent similarity calculation. Indukuri [9] used the parsing and semantic analysis to calculate the similarity of patents. McNamee [10] used patent similarity and distance to classify patents, and used the US patent classification system as an empirical study.

This paper focuses on Chinese patent infringement and proposes a method for judging the similarity of Chinese patents based on vector space model. Different from the traditional vector space model (VSM), this paper uses Word2Vec model to enhance the semantic expression ability of traditional VSM. The new model named semantic extended vector space model (SE-VSM).

2. Calculating Chinese Patent Similarity Based on SE-VSM

The whole idea of this article is shown below in Fig. 1:

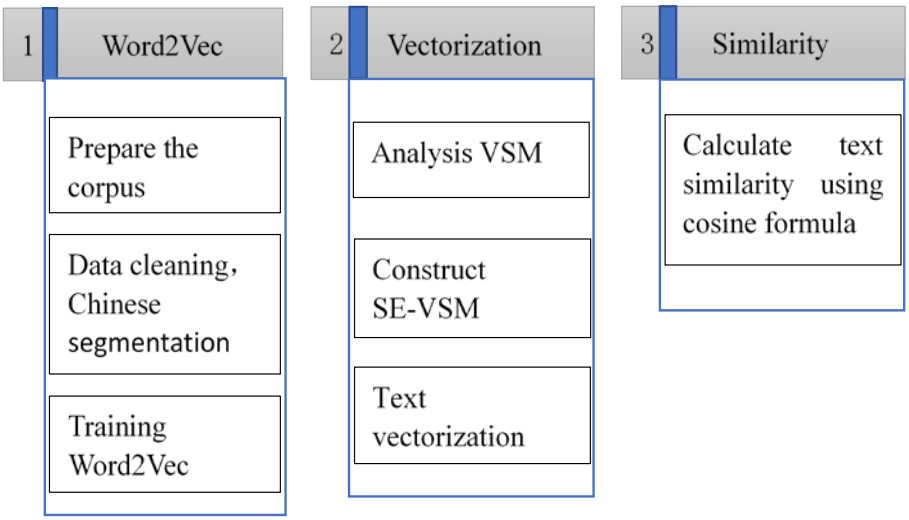

Figure 1. The whole idea of paper

2.1 Text Preprocessing

Before using the vector space model and the Word2Vec model, the text need pre-processing such as word segmentation, removal of stop words, and removal of special symbols. There are many algorithms for Chinese word segmentation such as: dictionary methods, statistical method and combination method. Many researchers have proposed some word segmentation methods for Chinese patent claims. And this paper chooses one method proposed by Jie Zhang [11].

2.2 Word2Vec Model

The Word2Vec [12] model is a language model proposed by Mikolov. It can quickly and effectively express a word into a vector, and the core architectures are CBOW model And the Skip-gram model, as shown in Fig. 2.

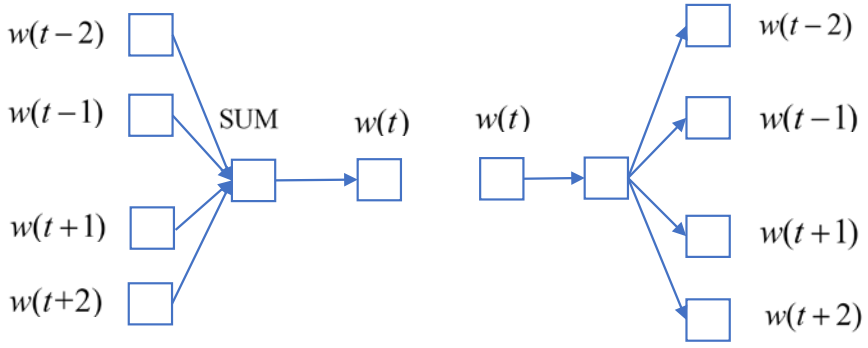

Figure 2. CBOW and Skip-gram

The CBOW model use the context to calculate the probability of the current word $w(t)$. The Skip-gram model is just the opposite. It uses the word $w(t)$ to predict the words before and after $w(t)$.

2.3 Vector Space Model

The vector space model was proposed by G. Salton in 1960 and is a very important mathematical model. The model is based on a hypothesis that the text content is only related to feature items. The meaning of the feature item is the word after the Chinese word segmentation. The model abstracts the text into a feature space vector.

Assume the existing Chinese text library D, and D contains N Chinese docs. First, the Chinese word segmentation method is used to segment the text, and then the stop words are removed to obtain the feature item set $T=\{t_1,t_2,t_3,...,t_n\}$. The feature items of all text are in set T. VSM maps text to n-dimensional vector, for example text D_i can be expressed as one vector V_{D_i}:

$$V_{D_i}=\{w_{i1},w_{i2},w_{i3},...,w_{in}\} \tag{1}$$

$w_{ik}(1\le k\le n)$ is weight of the feature item t_k in text D_i. Generally we use statistical correlation methods to calculate weights w_{ik}, TF-IDF method is a widely used weight calculation method used by VSM. The formula is:

$$w_{ik}=TF_{ik}*IDF_k \tag{2}$$

$$TF_{ik}=\frac{tf_{ik}}{Len_i} \tag{3}$$

$$IDF_k=\log\frac{|D|}{|Num_{t_k}|+\alpha} \tag{4}$$

tf_{ik} is the number of times that t_k occurs in D_i, Len_i the number of feature items in D_i, $|D|$ is the number of texts in D, $|Num_{t_k}|$ is the number of texts containing t_k. α is an empirical coefficient and generally set it to 0.01.

For two texts D_i and D_j, use TF-IDF to calculate the text vector $V_{D_i}=\{w_{i1},w_{i2},w_{i3},...,w_{in}\}$ and $V_{D_j}=\{w_{j1},w_{j2},w_{j3},...,w_{jn}\}$. There are mainly the following methods for calculating the similarity of vectors.

(1) Inner Product:

$$Sim(D_i,D_j)=V_{D_i}*V_{D_j}=\sum_{k=1}^{n}w_{ik}w_{jk} \tag{5}$$

(2) Jaccard Coefficient:

$$Sim(D_i,D_j)=\frac{\sum_{k=1}^{n}w_{ik}w_{jk}}{\sum_{k=1}^{n}{w_{ik}}^2+\sum_{k=1}^{n}{w_{jk}}^2-\sum_{k=1}^{n}w_{ik}w_{jk}} \tag{6}$$

(3) Dice Coefficient:

$$Sim(D_i, D_j) = \frac{2\sum_{k=1}^{n} w_{ik} w_{jk}}{\sum_{k=1}^{n} {w_{ik}}^2 + \sum_{k=1}^{n} {w_{jk}}^2} \tag{7}$$

(4) Cosine Coefficient:

$$Sim(D_i, D_j) = \cos\theta = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{(\sum_{k=1}^{n} {w_{ik}}^2)(\sum_{k=1}^{n} {w_{jk}}^2)}} \tag{8}$$

The most commonly used algorithm is the cosine coefficient shown in formula (8), which uses the cosine of the two vectors to measure the text similarity. The larger $Sim(D_i, D_j)$, the higher the similarity between the two texts.

2.4 Semantic Extended Vector Space Model

The biggest advantage of the VSM is text representation. It simplifies the processing of text content into vector operations by transforming the text content into a vector, which can greatly improve the operability and computability of natural language text. At the same time, the TF-IDF method can objectively represent text content to a certain extent.

Although the VSM model has many advantages, it also has the following problem: VSM assumes that there is no relationship between feature items, but the actual situation is that the features items of the texts often have rich semantic relations with each other. If you ignore the semantic relationship, VSM will produce inaccurate results.

In order to solve the problem that the traditional VSM contains too little semantic information, this paper use Word2Vec model to extend VSM, and the new model is named Semantic Extended Vector Space Model (SE-VSM). The specific steps are as follows:

1) Training Word2Vec model. Then we will get a Word2Vec model and feature item set $T = \{t_1, t_2, t_3, ..., t_n\}$. Using this model, we can get the similarity between feature items, $sim(t_j, t_k)$ is the similarity between t_j and t_k.

2) For one Chinese patent D_i, First get the vector of D_i according to formula (2).

$V_{D_i} = \{w_{i1}, w_{i2}, w_{i3}, ..., w_{in}\}$, and set $V_{D_i}' = V_{D_i} = \{w_{i1}', w_{i2}', w_{i3}', ..., w_{in}'\}$.

3) For any feature item t_j in D_i, and for any feature item $t_k \in T (1 \le k \le n)$. If $sim(t_j, t_k) \ge 0.5$ then add t_k to the set T_j. For $t_k \in T_j$, use $w_{ik} = w_{ik} + w_{ij}' * sim(t_j, t_k)$ to update V_{D_i}.

4) Finally, normalize V_{D_i} with 0-1 regularization.

3. Evaluation

The training corpus of the Word2Vec model in this paper is mainly composed of Wikipedia, Sogou News and more than 4000 Chinese patent claims. Each word in the trained Word2Vec model is represented by a 128-dimensional vector.

According to the principle of patent infringement, only authorized patents can be infringed. In order to simulate the actual patent infringement retrieval process, this paper selected eight patents,

and these eight patents belong to eight main categories. Their legal status is rejected. The contents of table1 are the patent number of these eight patents.

Table 1. Patents to be detected

Number	Patent number
1	CN200710009513.9
2	CN201010256932.4
3	CN20103125080.7
4	CN201010279077.9
5	CN200910198099.X
6	CN201010546297.3
7	CN20081002304.0
8	CN200780022801.6

For these eight patents, we build eight sets of experimental data. For each patent, we manually design six infringement patents using several methods, such as conceptual replacement, feature replacement, etc. Core-related patents are artificially found patents that are suspected of infringement.

There is the experimental data in Table 2.

Table 2. Experimental datasets

Number	IPC category	Artificial patents	Core-related patents	Subject related patent	Randomly selected patent
1	A41F1/00	6	8	24	34
2	B24C3/00	6	6	21	25
3	C22C1/00	6	7	30	40
4	D06H3/00	6	5	18	39
5	E21B4/00	6	6	35	45
6	F42D1/00	6	8	23	41
7	G05B19/00	6	6	28	28
8	H04W84/00	6	9	17	33

The ultimate goal of patent infringement search is to accurately and completely identify all patents that are suspected of infringement from the patent database. In order to compare VSM and SE-VSM, we use precision and recall as evaluation indicators.

$$precision = \frac{|\{\text{retrieved infringement patents}\}|}{|\{\text{retrieved patents}\}|} \tag{9}$$

$$recall = \frac{|\{\text{retrieved infringement patents}\}|}{|\{\text{all infringement patents}\}|} \tag{10}$$

If the similarity of two patents greater than 0.5, we believe that these two patents have infringement.

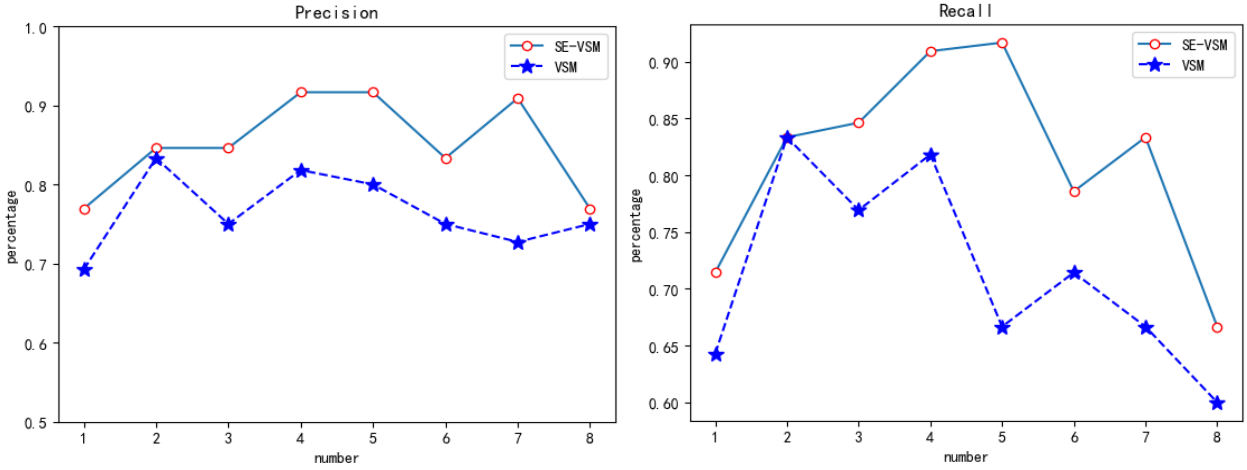

Figure 3. Precision and Recall

It can be seen from Fig. 3 that for the Chinese patent infringement retrieval, SE-VSM can achieve higher accuracy and recall than the traditional VSM. Judging from the results of the search, at least 4 of the top 6 patents with similarity are artificially designed infringement patents, which can explain the feasibility of the method.

4. Conclusion

Patent is the main carrier of scientific and technological innovation, once a company is involved in a patent infringement dispute, it usually takes a lot of time and energy. In serious cases, it may cause huge economic losses. At present, most patent search systems only perform simple search term matching and cannot effectively verify patent infringement.

This paper proposes semantic extended vector space model for the calculation of Chinese patent similarity. This model is based on Word2Vec model and traditional vector space model. Compared with traditional VSM, the model proposed in this paper has stronger semantic expression ability. The validity of the new model was also verified by experiments.

References

[1] YongShun Cheng. Patent infringement judgment pragmatic [M]. BeiJing: Law Publisher, 2009.

[2] Bergmann I, Butzke D, Walrter L, et al. Erdmann. Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis: The Case of DNA Chips [J]. R&D Management, 2008 (38): 550 - 562.

[3] Lee S, Yoon B, and Park Y. An approach to Discovering New Technology Opportunities: Keyword based Patent Map Approach [J]. Tec novation, 2009, 29 (6-7): 481 - 497.

[4] Taghaboni-Dutta F, Trappey AJC, Trappey CV, et al. An Exploratory RFID Patent Analysis [J]. Management Research News, 2009, 32 (32): 1163 - 1176.

[5] Moehrle MG, Gerken JM. Measuring Textual Patent Similarity on the Basis of Combined Concepts: Design Decisions and their Consequences [J]. Scientometrics, 2012, 91 (3): 805 - 826.

[6] Tseng Y H, Lin C J, Lin Y I. Text Mining Techniques for Patent Analysis [J]. Information Processing & Management, 2007, 43 (5): 1216 - 1247.

[7] Magerman T, Looy BV, Song X. Exploring the Feasibility and Accuracy of Latent Semantic Analysis based Text Mining Techniques to Detect Similarity between Patent Documents and Scientific Publications [J]. Scientometrics, 2010, 82 (2): 289 - 306.

[8] Yoon J, Kim K. Generation of Patent Maps using SAO-based Semantic Patent Similarity [J]. Entrue Journal of Information Technology, 2011, 10 (1): 19 - 27.

[9] Indukuri KV, Ambekar A A, Sureka A. Similarity Analysis of Patent Claims using Natural Language Processing Techniques[C]. Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on. IEEE, 2007 (4): 169 - 175.

[10] McNamee RC. Can't see the Forest for the Leaves: Similarity and Distance Measures for Hierarchical Taxonomies with a Patent Classification Example [J]. Research Policy, 2013, 42 (4): 855 - 873.

[11] Jie Zhang, HaiChao Zhang, DongSheng Zhai. Research of the Word Segmentation for Chinese Patent Claims [J]. Data Analysis and Knowledge Discover 2014 (9): 91 – 98.

[12] Le Q V, Mikolov T. Distributed Representations of sentences and Documents [J]. Eprint Arixiv, 2014 (4): 1188 - 1196.