Using the Multi-level Classification Method in the Patent Mining Task at NTCIR-7

Duo Ji, Huan-yu Zhao, Dong-feng Cai
Natural Language Processing Research Laboratory,
Shenyang Institute of Aeronautical Engineering,
Shenyang, China, 110034
Jiduo_1@163.com

Abstract

A patent includes a great deal of practical technical information, and plays an important role in promoting scientific development. The research on patent classification and retrieval has significant application value. A patent is a special technical text with strict hierarchical classification system and normalized structure, and there are a number of relations between patents and their constituents. Based on these relations, this paper proposes a weight calculation method using the patent title, and adopts a hierarchical classification method to complete the English and Japanese classification tasks at the NTCIR-7 patent mining. The effectiveness of the method is proved by corresponding evaluation.

Keywords: *patent classification, multi-level classifier, term weighting*

1. Introduction

The patent mining task at the NTCIR-7 is to classify research papers (abstracts) written in Japanese or English into the International Patent Classification (IPC).[2] In the evaluation, millions of training sets in about 31,000 classes are used in PAJ, and the number of patent in each class is quite different, at most 73895 and at least 1. The large class space and the uneven distribution of classes increase the training complexity of a classifier so that the difficulty of the patent mining task is increased.[5] Thus, this paper adopts the example based KNN algorithm[4] to construct a basic text classification system for different types of text. Compared with other scientific and technical documents, a patent has wide content range, integrates technology, law and economic information, reflects new technological information, and is written in unified format and normalized form, therefore in the patent mining task there are many features for us to use.

First, the hierarchy feature of the IPC classification of a patent is considered so as to propose the hierarchical classification method based on the naive Bayes model[6][9] and the KNN method. Second, in the feature selection of the patent, owing to the limited length of the patent document, the patent document (PAJ) has 195.6 words and 86 features averagely, and the frequency of each feature is 2.25 averagely; many key features such as technical terms may be features with low frequency and ultra-low TF value will cause its weight to decrease, therefore the importance of the features can not be incarnated. Based on the detailed analysis of the title and the summary of the patent, this paper combines the concept of mutual information theory to propose a feature weight calculation method by investigating the correlation of the feature and the topic to cause the vector representation of the text. Experimental results show that the method effectively improves the performance of patent classification.

The organization of this paper is as follows: the feature weight calculating method based on patent title is introduced in the second part, the third part describes the construction of the classification system used in the evaluation, experiment results and conclusion are given in the fourth part, and the last part is the summary of this evaluation.

2. Title based weight calculating method

In the correlation study of natural language processing, point-point mutual information[10] is often used as the measurement for describing the correlation degree between two words, which is called mutual information for short. A title is the topic of a document, so the correlation between the feature and the topic is the correlation between the feature and the title, as shown in formula (1)

$$I(w,T) = \log\frac{P(w,T)}{P(w)p(T)} = \log\frac{P(T\mid w)}{P(T)} \qquad (1)$$

Where w represents the feature; T represents the title. Because the probability of the title cannot be easily obtained, this paper makes a further assumption that the occurrence of words in a title is independent. In this way, formula (2) can be obtained by formula (1), wherein t_i represents the i-th feature in the title T, and K represents the feature number contained in the title T:

$$I(w,T) = \sum_{i=1}^{K}\log\frac{P(w,t_i)}{P(w)p(t_i)} = \sum_{i=1}^{K}\log\frac{P(w\mid t_i)}{P(w)} \qquad (2)$$

Where $P(w)$ represents the probability of containing the feature w in the document; $P(w\mid t_i)$ represents the probability of containing the feature w under the condition of containing the feature t_i in the concentrated document. The calculation formula of the feature weight is the formula (3):

$$TW(i,j)=\begin{cases} I(w_{i,j},T_j) & I(w_{i,j},T_j)\geq\alpha \\ 0 & I(w_{i,j},T_j)<\alpha \end{cases} \tag{3}$$

Where $w_{i,j}$ is the i-th feature in the document j; T_j is the title of the document j; when the mutual information of the feature and the title is smaller than the threshold α, the feature is considered as noise and can be removed from the feature set.

Then formula (3) and the TF*IDF term weight are combined using the linear interpolation as formula (4) to calculate the feature weight.

$$TWTI(i,j)=(1-\lambda)TW(i,j)+\lambda\bullet W(i,j) \tag{4}$$

Where λ is an adjusting parameter, *W(i,j)* is the TF*IDF term weight which is calculated using formula (4).

$$W(i,j)=(1+\log tf_{ij})\log(\frac{N}{df_i}) \tag{5}$$

tf_{ij} is the frequency of the i-th feature in document j. N is the number of the document in the training corpus, and df_i is document frequency of feature i.

3. Multilevel classification method

A patent is a technical document with a certain domain background, and technical terms in different domains express different meanings. For example, in the domain of machinery, a *crane* means a machine for hoisting and moving heavy objects, but in the domain of biology, a *crane* means a large wading bird. The patent mining task provides 889,116 American patents and 2,382,595 Japanese patents applied from 1993 to 2002 respectively, the contents cover many domains. In order to make the patent feature description based on words more accurate, the training sets are divided into 8 (A-H) independent domains to respectively carry out model training according to the IPC section classification. And for the document to be classified, the KNN classification method is used to give N (N=1000) classified candidate results in each domain.

Though the cosine similarity[8] of the documents varies between 0 and 1, the same value can not represent the same similarity degree in different feature space. Therefore, in order to obtain a reasonable class ranking list based on the overall situation, this paper distributes a correlation probability to each document to be classified based on the overall situation by the Bayes model, and finally merging algorithm is adopted to re-rank the 8*N results to obtain the final classification structure, wherein the calculation method is as follows:

$$Weight(S)=(\alpha_i+\beta)S_i \tag{6}$$

Where S_i represents a class in the i-th domain's ranking value by the KNN method; α_i is the result of the Bayes classifier and β is adjustment parameters.

The architecture of the system is shown in Figure 1. The system is mainly divided into text preprocessing, feature selection, Bayes classifier, KNN classifier and result combination. Text preprocessing mainly comprises word stemming of English, word segmentation of Japanese, etc., and the existing tool Chasen is used for Japanese word segmentation; feature selection mainly adopts the traditional DF method to filter the features whose document frequency is higher than a certain threshold T for the corpus of each domain respectively; in the KNN classifier, the cosine similarity calculation method is adopted, and the text is normalized and reserved in an inverted index mode so as to enhance operating speed.

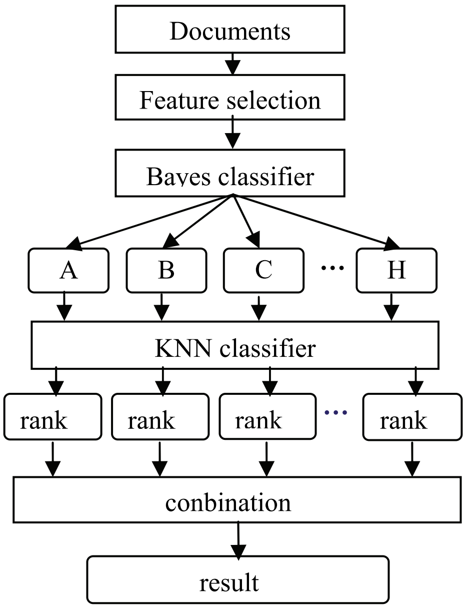

Figure 1. The architecture of the system

4. Experiment

4.1. Test Collections

The corpus used by the NTCIR-7 patent mining task[1] is the patent corpus provided by the National Institute of Informatics and is mainly from Japanese unexamined patent applications from 1993 to 2002, USPTO English patents from 1993 to 2002 and the English translation of Japanese patent summaries, wherein 889,116 patents are extracted from the patent data of the USPTO to be used as training samples, and each sample provides a single IPC classification label which relates to 38,491 IPC classes in subgroups. The text of the USPTO provides the whole text of a patent, and the available fields in the NTCIR-7 evaluation are patent number, patent title, patent summary, detailed patent presentation and patent declaration which are labeled <DOC>, <DOCNO>, <TITLE>, <ABST>, <SPEC> and <CLAIM> by XML labels. Training samples, of 2,382,595 Japanese patents are extracted from the patent applications from 1993 to

2002, and each sample provides multiple or a single IPC class which relates to 30,885 IPC classes in the subgroups. The Japanese patents provide multiple fields among which patent number, patent title, summary, declaration, etc, can be used for patent mining task.

4.2. Evaluation Results

The TREC tool is provided by the conference, and the A-Precision and the R-Precision are used in experiment evaluation.[1][2][3] In order to evaluate our system, we designed four experiments. The first (MKNN-TI) uses the multilevel classifier and title based term weight. The second (KNN-TI) uses the KNN classifier with α_i=1 for each domain and the term weight also uses the title based method. And the last two use the same classifier as the first two experiments, but use only TF*IDF term weight, named KNN and MKNN. .

The results on English corpus and Japanese corpus are shown in the two tables below, and the bold ones are the official evaluation results.

Table 1. Experimental results on English corpus

Method	A-Precision	R-Precision
KNN	0.2612	0.2388
KNN-TI	0.2687	0.2492
MKNN	0.2903	0.2216
MKNN-TI	**0.2903**	**0.2708**

Table 2. Experimental results on Japanese corpus

Method	A-Precision	R-Precision
KNN	0.2488	0.2012
KNN-TI	0.2556	0.2302
MKNN	0.2646	0.2107
MKNN-TI	**0.2727**	**0.2548**

From the results, we can see that MKNN-TI is the best method on both corpus: the A-Precision is 0.2903 on English corpus and 0.2727 on Japanese corpus. Besides, we can see that the title based weight calculation method can slightly improve our system. The KNN-TI's A-Precision is close to the KNN in both languages, but the R-Precision is lower than the KNN method. This may be caused that the title is too abstract to represent the topic of the articles. The MKNN experiment result is better than KNN, about 0.02 higher in A-Precision and 0.01 in R-Precision.

4. Conclusion and discussions

In the classification subtask of the NTCIR-7 Patent Mining Task, we implemented a weight calculation method and a multilevel classifier. Although these methods improved our evaluating result, we fail to get ideal results. First, when implementing this system, we adopted cosine method for calculating similarity between documents, which is proved ineffective in dealing with different types of text. Second, infrequent technical terms contained in each patent make the mutual information method slightly improved our system.

But in our system, we considered the domain background and divided the training corpus into several domain corpora based on the IPC section, independently selected features to train models, and finally used a Bayes classifier to re-rank the multi-result from each model. This multilevel process can reduce the calculation complication and get more exact domain feature space. It is worth further investigating for the patent classification.

References

[1]Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, Taiichi Hashimoto. Overview of the Patent Mining Task at the NTCIR-7 Workshop. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

[2]NTCIR committee. NTCIR-7 Patent Retrieval Task. 2008. http://www.nlp.its.hiroshima-cu.ac.jp/~nanba/n tcir-7/cfp-en.html

[3]M. Iwayama, A. Fujii, and N. Kando. Overview of classification subtask at NTCIR-6 patent retrieval task. Proceedings of the 6th NTCIR Workshop, 2007.

[4] M. Murata, T. Kanamaru, T. Shirado, and H. Isahara. Automatic f-term classification of japanese patent documents using the k-nearest neighborhood method and the smart weighting. *Journal of Natural Language Processing*, 14(1), 2007.

[5] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. 1999.

[6]Eibe Frank and Remco R. Bouckaert. Naive bayes for text classification with unbalanced classes. In *Proc 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Germany, pages 503-510. 2006.

[7] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24 (1998), 513-523.

[8]Metzler, D., Dumais, S., and Meek, C. "Similarity Measures for Short Segments of Text," in the Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007), 16-27, 2007.

[9] Mark Hall and Eibe Frank. Combining naive Bayes and decision tables. In *Proc 21st Florida Artificial Intelligence Research Society Conference*, Miami, Florida. AAAI Press, 2008.

[10] Jana Novovicová, Antonín Malík, Pavel Pudil: Feature Selection Using Improved Mutual Information for Text Classification. SSPR/SPR 2004: 1010-1017