A KNN Research Paper Classification Method Based on Shared Nearest Neighbor

Yun-lei Cai, Duo Ji ,Dong-feng Cai
Natural Language Processing Research Laboratory,
Shenyang Institute of Aeronautical Engineering,
Shenyang, China, 110034
jiduo_1@163.com
cyl_315@126.com

Abstract

The patents cover almost all the latest, the most active innovative technical information in technical fields, therefore patent classification has great application value in the patent research domain. This paper presents a KNN text categorization method based on shared nearest neighbor, effectively combining the BM25 similarity calculation method and the Neighborhood Information of samples. The effectiveness of this method has been fully verified in the NTCIR-8 Patent Classification evaluation.

Categories and Subject Descriptors

H.2.8 [Database Applications]: Data mining

General Terms

Performance

Keywords

patent classification, BM25, KNN, SNN

1. Introduction

The NTCIR-8 patent mining task is to classify research papers written in Japanese or English into the International Patent Classification (IPC) at subclass, maingroup, and subgroup levels.[7] In the evaluation, millions of training sets in about 400 subclasses,6000 maingroups,30000 subgroups are used. The large training data and large class space make many classifiers unsuitable for this task. For example, support vector machine (SVM) is a two-category classification model. While confronting multiple-category problem, it is necessary to train many classifiers. Classifier based on Naive Bayes can not be calculated with large classes space and the hardware used. In contrast, KNN is an algorithm based on machine learning, there are not many training parameters, the computational complexity is not high, and the performance is satisfactory, so we chose KNN as our system framework.

Similarity calculation among samples is a key part of KNN algorithm. Traditional methods such as inner product, cosine and Euclidean distance are all based on the vector space model, which did not fully consider the length of the sample. Thus, in the evaluation, each sample is composed of patent title and summary, the length of each sample is generally shorter, and the data sparse problem is serious. In order to integrate all kinds of useful information, our system uses the BM25 similarity calculation method, [5] It is a bag-of-words retrieval function, combines the word frequency and document frequency, balances the length of the document, and is a highly efficient similarity calculation method.

According to international patent classification standards, the patent is defined as a multi-level tree classification structure, as shown in Figure 1, different sub-categories derived from the same parent node have many common attributes, similarity alone is difficult to objectively measure the degree of similarity between samples, therefore, retrieval will introduce noise inevitably, [8] which to some extent affects the performance of the classification system. To solve this problem, this paper introduced the idea of shared nearest neighbors, using the shared Neighborhood Information[4] between samples to amend the similarity again, and ultimately ensure that every search results in a rational and just weight,

NTCIR-8 Patent Classification evaluation demonstrated the correctness of this method , and the details of the nearest neighbor will be described below.

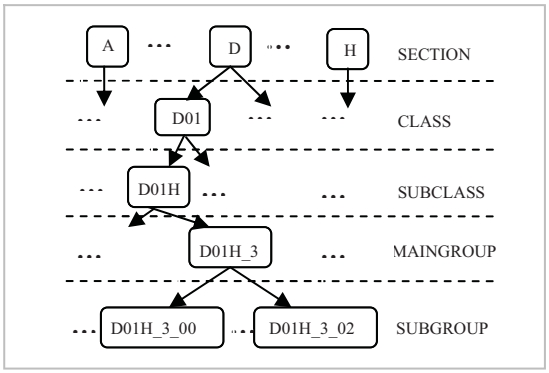

Figure 1 patented multi-level tree

The organization of this paper is as follows: The second part describes the BM25 similarity calculation method, the ideas of shared nearest neighbor is introduced in the third part, the fourth part introduces our experimental results, the last part is the conclusion of this evaluation.

2. BM25 similarity calculation method

BM25 is widely adopted in information retrieval which is used to retrieve documents related to the query keywords, [2] There are many variants of this function, we chose the following form,

$$score(D,Q)=\sum_{i=1}^{n} IDF(q_i)\cdot\frac{tf(q_i,D)\cdot(k_1+1)}{tf(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{avgdl})}$$

Where Q is the query vector containing a series of terms $\{q_1,...,q_n\}$, n is the number of key words in Q, D is the training sample vector $D=\{w_1,...w_M\}$, M represents the number of key words in D, $tf(q_i,D)$ is the frequency of term q_i in D. $|D|$ is the length of document D, $avgdl$ is the average document length in the training data set, k_1 and b are two parameters, namely term-frequency influence parameter and document normalization influence parameter respectively. In our system, k_1 and b are set to 1.5 and 1.0 by default. $IDF(q_i)$ is the inverse document frequency (idf) of term q_i ,which is calculated as follows,

$$IDF(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$$

Where N is the total number of documents in the training data set, $n(q_i)$ is the number of documents containing q_i , it should be noted that $IDF(q_i)$ can be negative in some cases, which may lead to negative $score(D,Q)$, in our system, $score(D,Q)$ is set to 0 if $score(D,Q)$ <0.

3. Category decision-making based on shared nearest neighbor

KNN algorithm is firstly to select pre-K samples when the similarity values are sorted in descending order, then to determine the categories of test sample with class mapping method. Common category decision-making methods are voting and similarity summing, In NTCIR-7 Tong Xiao presented an improved sorting method - the similarity summing algorithm based on position, [2] and achieved good classification results. On the basis of this, we present a new method, namely the similarity summing algorithm based on shared nearest neighbor, specific details are introduced below.

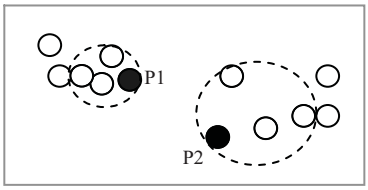

Figure 2 neighborhoods

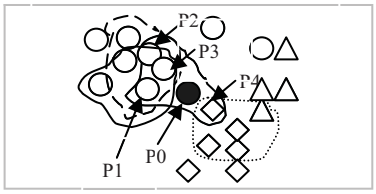

Figure 3 shared nearest neighbors

3.1 Neighbor

$sim(p_i, p_j)$ represents the similarity between the sample p_i and p_j, which can be any kind of similarity calculation method. We take the BM25 method to calculate the similarity between p_i and the other sample p_j, and sort them in descending order. Given a threshold w, if any sample meets formula 3.1, we claim that the set S_i is p_i's Neighbors, [4] As in Figure 2, consider the two sample point p1 and p2, if the threshold value is set to 3, then the nearest three sample points are called their neighbors.

$$S_i = \{p_j \mid rank(sim(p_i, p_j)) \geq w, j = 1,2..n\} \quad (3.1)$$

w is a user-specified parameter which can be trained to acquire, and we define it as the neighborhood radius, in this paper, w is set to 100, according to application requirements, as well as the distribution of samples, the user can select the appropriate w value.

3.2 Shared Nearest Neighbor

$link(p_i, p_j)$ is the number of shared neighbors between the sample p_i and p_j, which is a shared nearest neighbor concept, [4] as shown in formula 3.2,

$$link(p_i, p_j) = |\{x \mid S_i \cap S_j\}| \quad (3.2)$$

In a certain neighborhood radius, the greater $link(p_i, p_j)$, the more similar between the sample p_i and p_j, [4] and the more likely they belong to the same class. As in Figure 3, the graphics of different shapes represent samples of different categories, if the neighborhood radius is set to 4, the neighborhoods of p1, p2, p3, and p4 are roped above, the number of shared neighbors between p0 and p1, p2, p3 are all 2, p0 and p4 shared 0 neighbors, obviously p0 and p1, p2, p3 has a certain similarity, the introduction of $link$ to measure the similarity between samples can fully reflect the neighborhood information. Larger $link$ means that samples contain more similar properties. Through neighborhood information of samples, similarity can be more objectively evaluated.

3.3 similarity weight adding based on shared nearest neighbor

Classifier usually assigns higher weights to the higher ranked samples, Section 3.2 gives a detailed analysis of the importance of neighborhood information. On the basis of this, A new method, namely the similarity weight summing algorithm based on shared nearest neighbor, is presented. In the method penalty factor is added to small shared samples, as shown in formula below,

$$score(x) = \sum_{i=1}^{k} (k_t)^i \times \alpha^{\log(abs(link(d_i, q_j) - \beta) + 0.1)} \times score_{d_i} \times role(x, i)$$

$$role(x, i) = \begin{cases} 1, \text{if document } d_i \text{ has a x ipc classification} \\ 0, otherwise \end{cases}$$

Where k_t, α and β are three parameters, by default, k_t is set to 0.95, α and β are trained by dryrun corpus. $(k_t)^i$ can be regarded as a penalty that punishes the sample that have low ranks. q_j is the test sample, $score_{d_i}$ is the similarity between q_j and d_i.

4 . Experiment

4.1 Corpus

The corpus used by the Subtask of Research Papers Classification in NTCIR-8 is the English patents and Japanese patents provided by the National Institute of Informatics from 1993 to 2002, according to IPC list, we have extracted 3316197 documents from the Unexamined Japanese patent applications data sets, and extracted 3,496,139 documents from Patent Abstracts of Japan for category. Eventually we

extracted IPC number, title and summary from each patent as a training corpus, and conducted pre-processing on this basis such as removing stop words and so on.

Test set is divided into dry-run and the formal-run, each topic only consists of the title and summary of research papers. The official result is on the formal-run test set.

4.2 The evaluation results

The evaluation criteria of Subtask of Research Papers Classification in NTCIR-8 is A-Precision, which differs from NTCIR-7 in that each participant group submits one or more ranked lists of IPC codes at subclass, main group, and subgroup levels. In order to evaluate our system, we have conducted a comparison experiment on Japanese corpus and English corpus, using the KNN system for baseline, and KNN based on shared nearest neighbors (KNN + SNN) is our final results presented.

The results on English corpus and Japanese corpus are shown in table 1, and the bold ones are the official evaluation results, corresponding RUN-IDs match "KECIR_(.*)_A_OR".

sufficient, thereby increased the sample's neighborhood information, through which the similarity between samples can be more accurately measured.

In our system, parameters are trained by the dryrun data set , which are shown in table 2. In order to verify the stability of the shared nearest neighbors algorithm, we have conducted a multi-group experiments on English corpus and Japanese corpus through adjusting the neighborhood radius. Results are shown in Figure 4, which shows that the method based on shared nearest neighbors is generally superior to the baseline method. Besides, when the neighborhood radius is within 100 or less, the system's performance is nearly the same, which shows that our system is stable to a certain extent. In our experiment, the neighborhood radius is limited to 100. How to determine the optimal neighborhood radius according to the distribution of training data is our next step to research.

corpus	α	β
English	0.6	19
Japanese	0.8	5

Table 2. System parameters

corpus	Method	Subclass(A-precision)	Maingroup(A-precision)	Subgroup(A-precision)
English corpus	KNN	0.6892	0.4969	0.3452
	KNN +SNN	**0.7212**	**0.5474**	**0.3693**
Japanese corpus	KNN	0.6933	0.5111	0.3161
	KNN +SNN	**0.7215**	**0.5138**	**0.3184**

Table 1.Experimental results on English corpus and Japanese

From the results, we can see that KNN + SNN performs best on both corpus. On English corpus, compared to KNN method , KNN + SNN method increases by about 0.02 at subgroup level, but at subclass and maingroup levels, it has increased by about 0.03 and 0.05, respectively. The A-Precision in Japanese corpus is about 0.03 higher at subclass level. With category space narrowing, the performance of KNN + SNN method has improved, which shows that KNN based on shared nearest neighbors is closely related with the density of categories. The number of training samples near the upper class is relatively

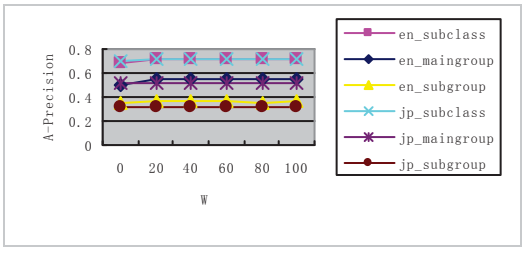

Figure 4 A-precision at different neighborhood radius

5. Conclusion and Future Work

In the Subtask of Research Papers Classification in NTCIR-8, we have built a KNN patent classification system based on shared nearest neighbors. Although

we obtained the best score in the English patent classification task, the results are still not ideal. Firstly, corpus processing is too rough without careful feature selection, the large feature space confused the topic information of patent, consequently weakened performance of our system to a certain extent; secondly, we haven't considered the problem of uneven density of corpus, resulting in wrong category decision-making for topics whose training data are inadequate.

However, in our system, we considered samples' neighborhood information to amend the weight of each search result so as to objectively assign higher weight to the sample which is more similar with the topic, and at the same time avoided the phenomenon that the similar samples with lower ranks are severely punished because of the location. This strategy is more reasonable for category decision-making, and it is worth further investigating for patent classification.

References

[1] Duo Ji, Huan-yu Zhao, Dong-feng Cai. Using the Multi-level Classification Method in the Patent Mining Task at NTCIR-7. In Proceedings of the 7th NTCIR Workshop Meeting. 2008.12. P362-364

[2] Tong Xiao, Feifei Cao, Tianning Li ,Guolong Song ,Ke Zhou,Jingbo Zhu,Huizhen Wang. KNN and Re-ranking Models for English Patent Mining Task at NTCIR-7. In Proceedings of the 7th NTCIR Workshop Meeting. 2008.12. P333-338

[3] Hisao Mase, Makoto Iwayama, Hitachi Ltd.NTCIR-7 Patent Mining Experiments at Hitachi. In Proceedings of the 7th NTCIR Workshop Meeting. 2008.12. P365-368

[4] Levent Ertoz, Michael Steinbach,Vipin Kumar. Finding Clusters of Different Sizes,Shapes ,and Densities in Noisy,High Dimentional Data.Proceedings of the third SIAM International conference on Data Mining 2003.1.P47-58

[5] ACL-2003 Workshop on Patent Corpus Processing[EB/OL]. http://www.slis.tsukuba.ac.jp/~fujii/acl2003ws.html

[6] Nanba H., Fujii A., Iwayama M. Overview of the Patent Mining Task at the NTCIR-7 Workshop[A]. Proceedings of the 7th NTCIR Workshop Meeting[C]. Tokyo, 2008:325-332

[7] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, Taiichi Hashimoto.Overview of the Patent Mining Task at the NTCIR-8 Workshop.Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2010.

[8] Larkey L.S. Some Issues in the Automatic Classification of U.S. Patents[A]. AAAI-98 Workshop on Learning for Text Categorization[C]. Menlo Park, 1998:87-90

[9] Kando N., Leong M.K. Workshop on patent retrieval SIGIR 2000 workshop report[J]. ACM SIGIR Forum, 2000, V34(1):28-30

[10] Peters C., Koster C.H.A. Uncertainty-Based Noise Reduction and Term Selection in Text Categorization[A]. Advances in Information Retrieval: 24th BCS-IRSG European Colloquium on IR Research[C]. Glasgow, 2002:25-27