F-term Classification System Using K-Nearest Neighbor Method

Kazuya KONISHI
Research and Development Headquarters, NTT DATA CORPORATION
Toyosu Center Bldg. Annex, 3-3-9, Toyosu, Koto-ku, Tokyo 135-8671, Japan
konishikzy@nttdata.co.jp

Toru TAKAKI
Fourth Public Administration Systems Sector, NTT DATA CORPORATION
Akasakacenter Bldg., 1-11-30, Akasaka, Minato-ku, Tokyo 107-0052, Japan
takakit@nttdata.co.jp

Abstract

In the Classification subtask of the NTCIR-6 Patent Retrieval Task, we implemented an F-term classification system using the k-nearest neighbor method. This system is based on the hypothesis that an F-term assigned to many patent documents that are similar to the topic patent document should also be assigned to the topic patent document. In implementing this systenm, we considered and applied methods for calculating similarity between patent documents, extracting patent documents from the training data set, and ranking F-terms to the F-term classification system. In this paper, we report the result of F-term classification.

Keywords: *Patent classification, F-term, K-nearest neighbor method.*

1. Introduction

In this paper, we report F-term classification using the k-nearest neighbor method, implemented in the classification subtask of the NTCIR-6 Patent Retrieval Task. In addition, we analyze the results and consider technological fields where the F-term classification is suitable.

The purpose of the classification subtask of the NTCIR-6 Patent Retrieval Task was to assign F-terms (File Forming Terms) to patent documents automatically [1]. The F-term list was developed by the Japan Patent Office. The Japan Patent Office uses the list to classify patent documents. A technological field is called a theme, and a theme code assigned to each theme. In addition, patent documents in a theme are categorized on the basis of purpose, function, and so on. An F-term is assigned to each category. There are 2,600 theme codes, and ten to thousands of F-terms are defined for 1,800 of them. The classification subtask of the NTCIR-6 Patent Retrieval Task was to assign F-terms to the topic patent documents to which a theme code has been assigned. The same classification task was implemented for the NTCIR-5 Patent Retrieval Task. However, as the number of topic patent documents for evaluating the F-term classification was increased, the evaluation of results is more reliable in the NTCIR-6 Patent Retrieval Task. In this task, the training data set comprised patent documents published from 1993 to 1997 that had been assigned F-terms. The test data set included 21,606 topic patent documents published from 1998 to 1999. The number of theme codes assigned to all topic patent documents was 108.

The result for the NTCIR-5 Patent Retrieval Task showed the possibility of using the k-nearest neighbor method for highly precise F-term classification [2]. When we execute this F-term classification, we can apply some methods for calculating the similarity between patent documents, extracting patent documents from the training data set, and ranking F-terms to the F-term classification system. When we implemented an F-term classification, we attempted several ways to execute each method for applying them to F-term

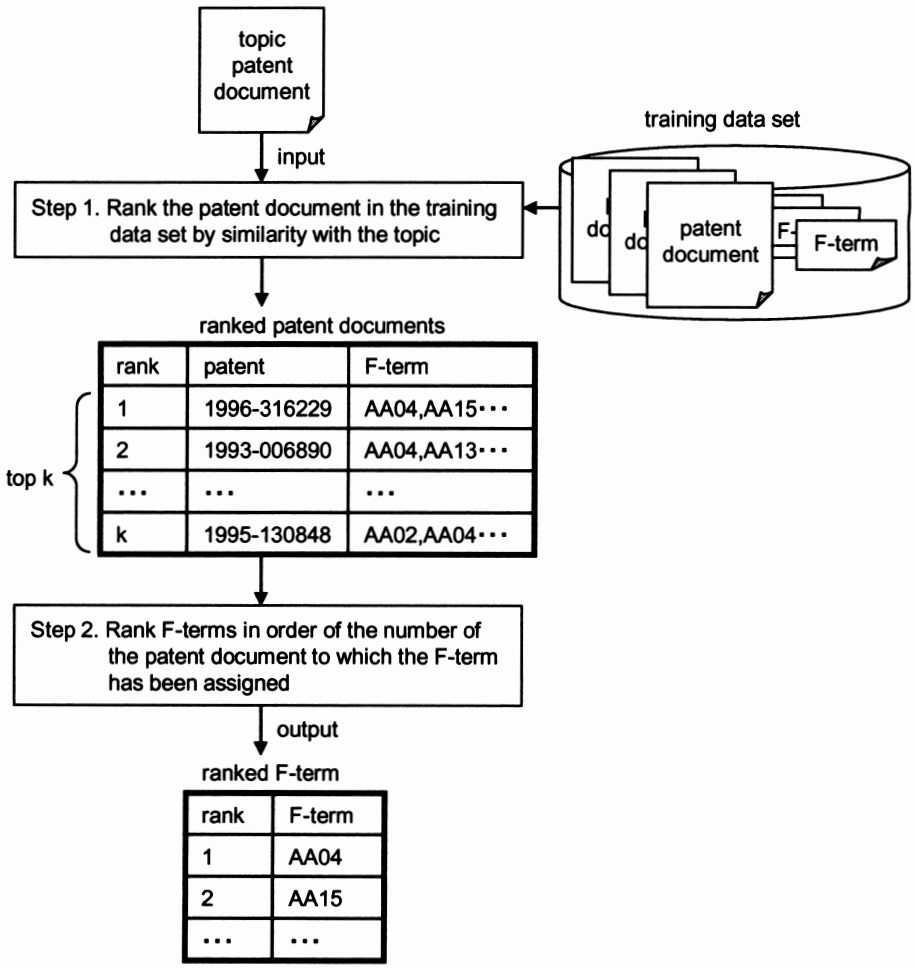

Fig.1. F-term classification using k-nearest neighbor method

classification using the k-nearest neighbor method.

2. F-term Classification System using k-nearest neighbor method

The steps of F-term classification using the k-nearest neighbor method are outlined in Fig. 1. Step 1 extracts patent documents that are similar to the topic patent document from training data set, and step 2 obtains F-terms assigned to many extracted patent documents.

Step 1:

Rank the patent documents in the training data set by similarity with the topic patent document, and extract top-*k* patent documents.

Step 2:

Rank F-terms assigned to extracted *k* patent documents in order of the number of patent documents to which the F-term has been assigned.

In implementing this F-term classification, we can consider some method for calculating the similarity between patent documents, extracting *k* patent documents from training data set, and ranking F-terms.

3. System Description

We implemented the F-term classification system using the k-nearest neighbor method and attempted to apply the following methods to the system, considering the characteristics of patent documents.

3.1. Calculating similarity between patent documents

A patent document is composed of several regions such as the "scope of the claim for a patent", "technical field to which the invention pertains", and "conventional technology". We considered calculating the similarity between patent documents using only regions that include descriptions as the basis for assigning F-terms.

Currently, information about which descriptions in each patent document are the basis for assigning F-term has not been disclosed. As a result of reading of some patent documents, we presume that the following two fields include many descriptions as the basis for assigning F-term, and extract query terms from the fields of the topic document.

1. abstract

   "abstract" field summarizes issues of invention along with solutions and their effectiveness.

2. abstract and first claim

   "first claim" field specifies the main scope of the patent claim. We extracted query terms from both of "abstract" and "first claim" fields.

For each patent document of the training data set, we calculated the similarity with the topic patent documents using query terms extracted from only the abstract as well as from both the abstract and first claim of the topic patent document.

For this purpose, we used Okapi BM25 [3]. Higher similarity was calculated for patent documents containing many terms that appear in the topic patent document, without omission. Because single words or compound words are considered as suitable elements for the similarity calculation [4], we set the following elements as terms and calculated the similarity between patent documents.

1. Single words
2. Single words and compound words

We performed a Japanese language morphological analysis using ChaSen [5]. We set a single noun or unknown word to a single word, and set a sequence of nouns or unknown words to a compound word.

3.2. Extracting patent documents applied to k-nearest neighbor method

The important general issue in the k-nearest neighbor method is how to determine the appropriate value of k. The value of k in the F-term classification is the number of patent documents extracted from the training data set by the system. In order to determine the appropriate value of k, we examined the number of patent documents whose assigned F-term is suitable for assignment to the topic patent document.

We set k to empirical values appropriate for F-term classification, and extracted k patent documents from the training data set.

1. $k = 50$
2. $k = 100$

3.3. F-term ranking

We output the F-term assigned to the k extracted patent document and ranked each F-term according to the number of patent documents assigned it.

Our idea is that the F-term assigned to a patent document with high similarity to the topic patent document should be assigned to the topic patent document.

We set a weight to each F-term by patent document extracted from the training data set, and calculated the sum of the weights of each F-term for F-term ranking. To define the weight, we use the similarity between patent documents, the rank of the similarity, or a constant weight.

1. Similarity between patent documents

   We set the weight *weight_1(f)* as the similarity *sim(p)* between the patent document p and the topic patent document to the F-term f assigned to the patent document p extracted from the

Table 1. Average precision of F-term classification

Calculating similarity between patent documents		F-term ranking		
region for query term extraction	Element as term	Constant weight (*weight 3*)	Similarity between patent document (*weight 1*)	Rank of similarity (*weight 2*)
abstract + 1st claim	single word + compound word	0.2717	0.2705	0.2694
abstract	single word + compound word	0.2711	0.2702	0.2689
abstract	single word	0.2714	0.2700	0.2687
abstract + 1st claim	single word	0.2709	0.2698	0.2684

Table 2. F-measure of F-term classification

Calculating similarity between patent documents		F-term ranking		
region for query term extraction	Element as term	Constant weight (*weight 3*)	Similarity between patent document (*weight 1*)	Rank of similarity (*weight 2*)
abstract	single word + compound word	0.2415	0.2397	0.2401
abstract + 1st claim	single word + compound word	0.2414	0.2402	0.2395
abstract	single word	0.2407	0.2396	0.2391
abstract + 1st claim	single word	0.2402	0.2391	0.2385

training data set:

$$weight\_1(f) = sim(p) \quad (1)$$

2. Rank of similarity

We set weight *weight_2(f)* as the total number k of patent document similar to the topic patent document minus the value of the patent document *rank(p)* minus 1:

$$weight\_2(f) = k - (rank(p) - 1) \quad (2)$$

3. Constant weight

We rank F-terms assigned to extracted k patent documents in order of the number of patent documents to which the F-term has been assigned. We set weight *weight_3(f)* of 1 to the F-term f assigned to the patent document p extracted from the training data set:

$$weight\_3(f) = 1 \quad (3)$$

Here, we call the sum of weights as score *Score(f)* of the F-term f such that

$$Score(f) = \Sigma weight(f) \quad (4)$$

where *weight(f)* means *weight_1(f)*, *weight_2(f)*, or *weight_3(f)*. For *weight_3(f)*, *Score(f)* becomes the number of the patent documents with assigned F-term f among k patent documents, which is exactly the same as the original k-nearest neighbor method.

Another requirement in the NTCIR-6 Patent Retrieval Task is to show the confidence level for each output F-term. We set the *confidence* to 1 when are highly confident in the F-term, and set it to 0 when we are not confident in the F-term. The *confidence* is used to calculate the F-measure for the evaluation of the result of F-term classification.

We calculated the average value c, which is the number of F-terms assigned to one theme code for a patent document in the training data set. We think c represents the appropriate number of F-terms that should be assigned to the topic patent document, and set the confidence to 1 for top-c cases and to 0 for the rest of the F-terms.

Table 3. Average precision of each theme code assigned to the topic patent document

Rank	Theme code	Content	Average precision
1	4J034	polyurethane, polyurea	0.4813
2	5F102	junction FET	0.4642
3	4C055	pyridine compound	0.4634
4	4C063	heterocyclic compounds	0.4466
5	4L045	method of fiber spinning and system	0.4436
6	5B029	character entry	0.4398
7	4H045	peptide or protein	0.4123
8	5J065	code error detecting, correcting	0.4093
9	4F210	stretching and molding as plastic	0.4074
10	4H057	coloring	0.4044
99	3K068	fuel supply and control	0.1319
100	4F070	processing high polymeric substance	0.1295
101	5F051	photovoltaic system	0.1263
102	4J002	polymeric composition	0.1258
103	5F056	electron beam exposure	0.1224
104	4E081	butt welding and welding of particular article	0.1164
105	5B076	stored program control	0.1037
106	2C088	pinball game machine (pachinko etc.)	0.0972
107	3C045	turning	0.0918
108	4L056	yarn spinning and twisted yarn	0.0881

4. Experiment

We applied each method and executed F-term classification using the k-nearest neighbor method. Below, we present the average precision values calculated from the results of the F-term classification.

When we used the same method for calculating the similarity between patent documents and F-term ranking, and executed F-term classification, the average precision was higher for k of 100 than for k of 50. Table 1 shows the average precision of F-term classification for k of 100. The 12 average precision values in Table 1 resulted from combinations of each of the *region* for query term extraction, the type of term used to calculate the similarity between patent documents, and the method of F-term ranking. The difference in each average precision was very small. The highest average precision was obtained when we used query terms extracted from the abstract and first claim of the topic patent document to calculate the similarity between patent documents, used single and compound words as the elements of terms for calculating the similarity between patent documents, and used *weight_3(f)* as the weight for F-term f classification for F-term ranking (In another words, when we did not weight according to the similarity with the topic patent document to each F-term).

Table 2 shows the F-measure obtained from the result of F-term classification calculated using the confidence. The F-measure and average precision show a similar tendency. However, the F-measure was higher when query terms were extracted from only the abstract of the topic patent documents.

5. Analysis

We analyzed the results of F-term classification by each theme code to clarify the effective technological field for F-term classification with the k-nearest neighbor method.

Table 3 shows the average precision obtained for each theme code when we executed F-term classification using the combination that produced the highest average precision.

Table 4 shows part of the F-term list in theme code 4J034 (polyurethane, polyurea), for which the average precision of F-term classification was ranked high. The names of substances appear on the list frequently. When we read the topic patent document assigned theme code 4J034, the same substance names as those in the F-term list appeared in the text.

Table 5 shows part of the F-term list of theme code 4L056 (spinning and flammable yarn), for which the average precision of F-term classification was ranked low. Descriptions in the abstract, such as "supply of multiple raw materials" and "by something mechanical", appear on the list. When we read the topic patent document assigned theme code 4L056, descriptions created from the abstract's descriptions appeared in the text.

Many terms having a fixed meaning (in a specific technological field, there are no terms having the same meaning except the term) appear on patent documents with the theme code 4J034, on the other hand they do not appear on patent documents with the theme code 4L056. We also found some other examples in each theme, such as the cases of theme codes with high or low average precision. We considered the

Table 4. Parts of F-term list (theme code : 4J034)

4J034		polyurethane, polyurea									
CA	CA00		CA01	CA02	CA03	CA04	CA05				
	type of active hydrogen substrate from low molecular active hydride compounds		・hydroxy	・・monohydroxy compound	・・polyhydroxy compound	・・・diol	・・・triol				
			CA11	CA12	CA13	CA14	CA15	CA16	CA17		CA19
			・N containing active hydrogen substrate	・・amine	・・・monoamine	・・・polyamine	・・・・diamine	・・・・triamine	・・・except primary amine (secondary, third amine)		・・ammonia, ammonium
			CA21	CA22	CA23	CA24	CA25	CA26			
			・carboxylic acid(derivative)←anhydride、ester)	・・onocarboxylic acid (derivative)	・・polycarboxylic acid (derivative)	・・・dicarboxylic acid (derivative)	・・・tricarboxylic acid(derivative)	・・・tetra karate etra karte (deriative)			

Table 5. Parts of F-term list (theme code : 4L056)

4L056		yarn spinning and twisted yarn									
BA	BA00		BA01		BA03		BA05		BA07			
	supply *		・supply of sliver		・supply of roving		・supply of yarn		・supply of the multiple raw material			
			BA11	BA12		BA14	BA15		BA17			
			・component *	・・coiler can		・・creel	・・・bobbin hanger		・・feed roller			
BB	BB00		BB01	BB02		BB04		BB06		BB08		
	opening *		・by something mechanical	・・by opening roller		・by fluid		・by electrostatic		・remove contaminant		

theme code assigned to the patent documents which include many terms having a fixed meaning may be more suitable for F-term classification using the k-nearest neighbor method. This is because the precision of similarity between patent documents would be high when the terms appearing on the patent document have a fixed meaning. In fact, as table 6 shows, the score of the F-terms calculated from the similarity between patent documents was high with the theme code 4J034, but was low with the theme code 4L056.

Table 6. F-term score

(a). Topic # : F412409
Theme code : 4J034

Rank	F-term	Score(f)
1	HA01	89
2	HA07	88
3	DA01	77
4	HC71	74
5	HC12	72
6	HC61	71
6	HC67	71
8	DB07	70
8	HC64	70
10	CB07	66

(b). Topic # : F414253
Theme code : 4L056

Rank	F-term	Score(f)
1	AA02	22
2	AA01	19
2	BF08	19
4	AA45	16
5	AA19	14
6	BF09	12
7	AA21	11
7	AA32	11
7	FA05	11
10	CB06	10

6. Conclusion

We implemented the F-term classification system using the k-nearest neighbor method in the classification subtask of NTCIR-6 Patent Retrieval Task. In this paper, we reported the results of executing F-term classification with methods for the calculation of the similarity between patent documents, the extraction of patent documents, and F-term ranking.

We also considered the effective technological field for applying the F-term classification using the k-nearest neighbormethod. The F-term classification may be more effective when the term appearing on the patent document has a fixed meaning.

When the information about the basis for F-term assignment is disclosed in the future, we will be able to make large improvements in the method of calculating the similarity between patent documents.

References

[1] M. Iwayama, A. Fujii, N. Kando. Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task. Proceedings of the 6th NTCIR Workshop, 2007.

[2] M. Murata, T. Kanamaru, T. Shirado, and H. Ishihara. Using the K Nearest Neighbor Method and BM25 in the Patent Document Categorization Subtask at NTCIR-5. Proceedings of 5th NTCIR Workshop, 2005.

[3] S.E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: Automatic ad hoc filtering, VLC and interactive. Proceedings of the 7th Text REtrieval Conference(TREC-7), NIST Special Publication 500-242, pp.253-264, 1999.

[4] K. Yamada, T. Mori, H. Nakagawa. Information Retrieval Based on Combination of Japanese Compound Words Matching and Co-occurrence Based Retrieval. IPSJ journal, Vol.41, No.4, pp.1162-1170, 2000 (in Japanese).

[5] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, M. Asahara. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. Technical Report NAIST-IS-TR99009, NAIST, 1999.