Using the K-Nearest Neighbor Method and SMART Weighting in the Patent Document Categorization Subtask at NTCIR-6

Masaki Murata
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
murata@nict.go.jp

Toshiyuki Kanamaru
Kyoto University
Yoshida-Nihonmatsu, Sakyo, Kyoto 606-8501, Japan
kanamaru@hi.h.kyoto-u.ac.jp

Tamotsu Shirado
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
shirado@nict.go.jp

Hitoshi Isahara
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
isahara@nict.go.jp

Abstract

Patent processing is important in industry, business, and law. We participated in the classification subtask (at NTCIR-6 Patent Retrieval Task), in which, we classified patent documents into their F-terms using the k-nearest neighbor method. For document classification, F-term categories are both very precise and useful. We entered five systems in the classification subtask and obtained good results with them. Thus, we confirmed the effectiveness of our method. By comparing various similarity calculation methods, we confirmed that the SMART weighting scheme was the most effective method in our experiments.

Keywords: *Classification, patent documents, k-nearest neighbor method, SMART weighting*

1 Introduction

Patent processing is important in various fields, such as industry, business, and law. We entered our systems into the classification subtask at NTCIR-6 [3], in which we classified patent documents into their F-terms using the k-nearest neighbor method. F-term categories are very precise and thus useful for classifying patent documents. Furthermore, this method makes it possible to classify a large number of documents. This would be difficult using sophisticated machine learning methods, such as the support vector machine [1] and maximum entropy methods [11] because these methods are complicated and require a lot of time and machine resources (memory). In contrast, the k-nearest neighbor method is comparatively easy to use with large amounts of data because it only has to extract a set of data similar to the input data. Yang also pointed out that the support vector machine and the k-nearest neighbor method are the best machine learning methods for document classification [18]. Therefore, we used the k-nearest neighbor method in this study.

2 Problem setting

In this section, we describe the problem addressed in this study.

We participated in the classification subtask during the NTCIR-6 Patent Workshop [10], because F-term categories are very precise and useful for classifying patents. In the subtask, we determined the F-term categories of input Japanese patents when the category theme was given. Our problem was to determine these categories. The subtask details are described on the NTCIR-6 Patent Workshop [10] website.

Each patent can be classified into some theme cate-

Table 1. Japanese patent structure

Section	Tag	Components	Examples
Bibliography	SDO BIJ	Publication number	2000-004182
		Date of publication of application	07.01.2000
		Title of invention	Separation-type portable telephone set
		Application number	10-167909
		Date of filing	16.06.1998
		Applicant	Matsushita Electric Ind. Co. Ltd.
		Inventor ...	Kanazawa Kunihiko
Abstract	SDO ABJ	Problem to be solved	To provide a separation-type portable telephone set that eliminates the ...
		Solution	Voice data are transmitted and received through radio waves, infrared rays, etc.,
Claims	SDO CLJ	Claim 1 ...	A discrete-type cellular phone characterized by transmitting and ...
Description	SDO DEJ	Field of the invention	This invention relates to the discrete-type cellular phone using the ...
		Description of the prior art	Conventionally, cellular phones are used by a method that connects a ...
		Problem(s) to be solved by the invention	However, also in the cellular phone that has the configuration shown in ...
		Means for solving the problem	The discrete-type cellular phone that this invention gets to ...
		Embodiment of the invention	(Gestalt 1 of operation) The discrete-type cellular phone concerning the ...
		Effect of the invention ...	According to the discrete-type cellular phone that this invention gets as ...
Explanation of Drawings	SDO EDJ	Drawing 1 ...	The block diagram of the discrete-type cellular phone concerning the ...
		Description of notations	1 Microphone (body built-in) ...
Drawings	SDO DRJ	Figure 1 ...	

gories and some F-term categories. Theme categories are a higher layer than F-term categories. Both categories were applied to each patent by the Japan Patent Office [5]; there were about 2,600 theme categories, each of which had anywhere from dozens to thousands of F-term categories. Each patent had an average of 1.7 theme categories and 15 F-term categories in the formalrun data.

The subtasks included a dry run and a formal run. For the classification subtask, we were given 760 patent documents to classify and 1,273,757 patent documents for training in the dry run. In the formal run, we were given 21,606 patent documents to classify and 1,273,757 patent documents for training. However, we were able to use documents with given theme categories for training. In the dry run, we were given about 1,920 and 7,314 patent documents with given theme categories, and in the formal run, we were given 1,027 to 35,147 patent documents with given theme categories.

In the evaluation, we used average precision (A-Precision), R-precision, and F-measures. Average precision is the average of the precision when each category relevant to the input document is extracted. R-precision indicates the precision when extracting R categories, where R is the number of relevant categories. The F-measure is the harmonic mean of the recall and precision. The recall is the ratio of the correct outputs to all the correct categories. Precision is the ratio of the correct outputs to all the outputs.

2.1 Patent structure and F-terms

In this section, we explain the Japanese patent structure and the F-terms used in this study.

Each Japanese patent document has a sequence of normative sections: the bibliography, abstract, claims, description, explanation of drawings, and drawings, as indicated in Table 1. In the patent data given at the NTCIR-6 Patent Workshop [10], tags for these sections such as “SDO ABJ” were inserted. The bibliography of a patent includes the publication number,

Table 2. Example of F-term classification system

5K067	Mobile radio communication systems					
AA	AA00	AA01	AA02	AA03	AA04	...
	Purpose and Effects	Measures to overcome radio or transmission problems	Measures relating to phasing or multi-pass	Measures to prevent interference or jamming	Prevention of unwanted transmission	...
BB	BB00	BB01	BB02	BB03	BB04	...
	Applications	Telephones	Wireless telephones	Car phones	Cellular phones	...
CC	CC00	CC01	CC02		CC04	...
	Transmission systems	Multiplex systems	Frequency multiplexing		Time-division multiplexing	...
DD	DD00	DD01	DD02	DD03	DD04	...
	Transmission signals	Signal types	Frequency signals	Serial or parallel tones	Binary signals (i.e., binary code)	...
EE	EE00	EE01	EE02	EE03	EE04	...
	System configuration	Station configuration	Mobile stations	Variants that have secondary stations	Use as multiple stations	...
...	...	...	...	...	...	...

Table 3. Theme and F-term categories for published patent 2000-004182

Published patent number	Theme	F-terms
2000-004182	5K011	AA04, BA00, BA10, DA17, JA01, KA12
	5K027	AA13, CC08, DD11, DD14, EE03, HH03, HH16, HH20
	5K067	AA34, AA42, BB04, FF38

date of publication, title, inventors, etc. The abstract contains the abstract, and the claims section contains the claims. The description gives information such as the field and embodiments of the invention.

Next we explain the F-terms. The Japan Patent Office provides a multi-dimensional classification structure called an F-term classification system [6, 13]. An example is shown in Table 2.

In an F-term classification system, each technological field is defined as a theme corresponding to a set of "FI" codes (an extension of IPC). For example, the theme denoted by "5K067" represents the technological field of "Mobile radio communication systems,', and this theme corresponds to the FI codes "H04B7/24-7/26,113@Z;H04Q7/00-7/04@Z." A theme is expressed by a sequence consisting of a digit, a letter, and three digits. There are over 2,500 themes.

Each theme has a collection of viewpoints for specifying the possible aspects of the inventions under the theme. For example, 5K067 has "Purpose and Effect", "Applications", and "Transmission Systems" as viewpoints. The collection of viewpoints varies from theme to theme. A viewpoint is denoted by two letters. For example, "AA" represents the viewpoint "Purpose and Effect". Note that the naming of viewpoints is not uniform across themes, meaning that "AA" may not represent "Purpose and Effect" in other themes.

Each viewpoint has a list of possible elements. For example, "Purpose and Effect" in this theme might be "Measures to overcome radio or transmission problems", and "Applications" in this theme might be "Telephones". The collection of elements varies from viewpoint to viewpoint. An element is represented as two digits. For example, "Telephones" for "Applications" corresponds to "01". As an exception, "00" sometimes represents the elements not enumerated in the list of possible elements. The "00" element is also used to designate the viewpoint as a whole, as shown in Table 2.

A pair comprising a viewpoint and an element is called an F-term. For example, "BB01" is the F-term for mobile radio communication systems whose applications are telephones.

All patents have various theme and F-term categories. To explain these, we use the published patent 2000-004182 described in Table 1. The patent has three theme categories: 5K011, 5K027, and 5K067.

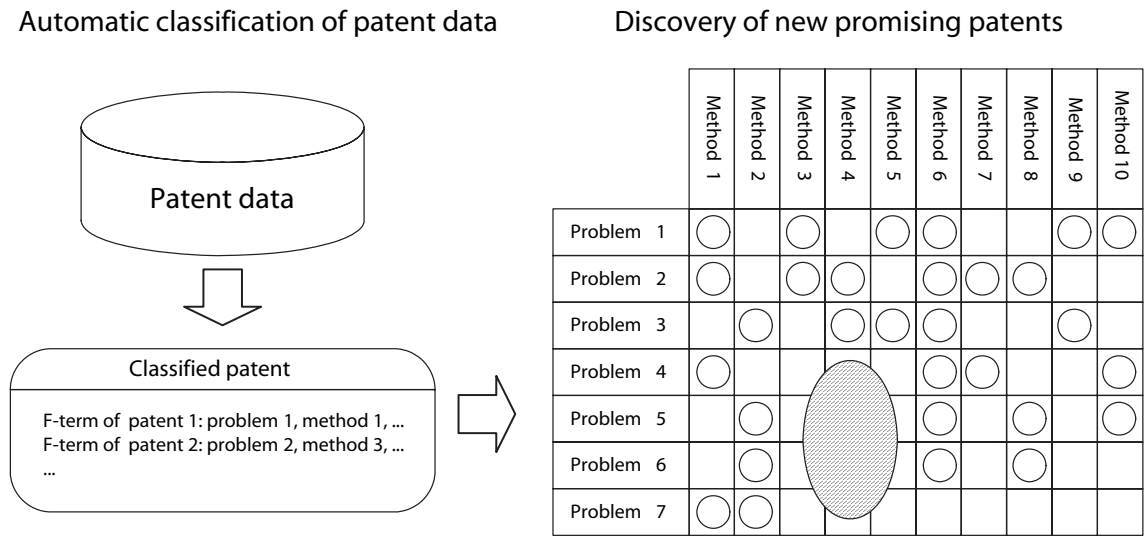

Figure 1. Example of F-term use

Its F-term categories for these theme categories are listed in Table 3. This patent has the F-terms AA34, AA42, BB04, and FF38 for the theme 5K067. The fact that the patent has the F-term BB04 for the theme 5K067 means that the patent relates to mobile radio communication systems, and its application is "cellular phones", as listed in Table 2. The patent does not have F-terms for CC and DD viewpoints in the theme 5K067. A patent does not generally always have F-terms for all viewpoints.

3 Background and motivation

The F-term categories are both precise and useful for categorizing patents. For example, the "radio transmission" theme had many F-term categories, including" purpose", "application", "transmission system", "transmission signal", "system architecture", and "function". These were then further broken down, and "purpose" contained the F-terms of "failure prevention", "service improvement", and "efficiency improvement"; "application" contained the F-terms of "car phone", "cellular phone", and "train radio system"; and "function" contained the F-terms of "memorization", "display", etc. If we arrange the radio transmission patent documents into a two-dimensional table, where the columns are the "purpose" F-terms, and the rows are the "application" F-terms, we can better understand the purpose and application situations in radio transmission patent documents. Figure 1 shows another simple example that demonstrates the usefulness of F-terms. In the example, each patent was given F-terms on problems and methods by automatically classifying the patent data. The patent information containing F-terms was transformed into the table at the right side of the figure. The circles in the table show that there are patents that contain information covering the corresponding problems and methods. The area denoted by the gray circle did not have any patents, which indicates the discovery of promising new patents, such as the patents covering problems 4 to 7 using methods 3 to 5. The F-terms are useful for discovering such patents. Thus, F-term categories can be very useful for categorizing patents. (The patent task organizer also illustrated the importance of F-term categorization for similar reasons. Automatic construction of patent maps were handled in NTCIR-4 Patent Retrieval Task [2].)

This study is therefore useful for the following reasons:

- Our method can help annotators determine the F-term categories of each patent document.
- Our method can be used for documents from outside the patent office that do not contain F-term categories, and can assign F-term categories to these documents.

4 Modification of the k-nearest neighbor method

We used the following modified version of the k-nearest neighbor method.

1. Method 1

   The system first extracts the k patent documents with the highest similarities to an input patent document for all patent documents with the same given input theme in a training data

set. We used the ruby-ir toolkit [16, 17] to extract the documents and experimentally determine the constant k.

The system next calculates $Score(x)$ using the following equation for each F-term category x in the extracted documents.

$$Score(x) = \sum_{i=1}^{k}((k_r)^i \times score_{doc}(i) \times role(x,i)), \quad (1)$$

where

$$role(x,i) = 1 \text{ (if the } i\text{-th document has a F-term } x) \\ = 0 \text{ (otherwise).}$$

Here, $score_{doc}(i)$ is the value of the similarity of the selected document that has the ith highest value of the similarity between the input patent document and the selected document, and k_r is a constant determined using experiments.

The system finally extracts the F-term categories with higher $Score(x)$s than the highest $Score$ multiplied by k_p. We experimentally determined the constant k_p. The extracted F-term categories are output as the desired categories.

5 Method of calculating similarity

We used the following four methods to calculate the similarity between an input patent document and each patent document in a training data set.

1. SMART The System for Manipulating and Retrieving Text (SMART) is a term weighting method in information retrieval [14, 15, 4]. The system first extracts terms[1] for each input patent document. The system then selects documents containing at least one of the terms from all the patent documents with a given input theme in the training data set. It uses the following equation to calculate Sim_{SMART} for each selected document. We used Sim_{SMART} as the similarity between an input patent document and each patent document in the training data.

$$Sim_{SMART} = \sum_{t \in T}(W_d \times W_q), \quad (2)$$

$$W_d = \frac{1 + log(tf)}{1 + log(avtf)} \times \frac{1}{0.8 + 0.2\frac{utf}{pivot}}, \quad (3)$$

$$W_q = (1 + log(qtf)) \times log\frac{N+1}{n} \quad (4)$$

In these equations, T is the set of terms appearing in both the input document and the selected document, tf is the number of occurrences of a term t in the selected document, $avtf$ is the average number of occurrences of each term in the set in the selected document, qtf is the number of occurrences of term t in the query document, utf is the number of unique terms in the selected document, $pivot$ is the average number of unique terms per document in the training documents, N is the total number of patent documents with a given input theme in the training data set, and n is the number of documents in which term t appears.

2. BM25 BM25 is a term weighting method in information retrieval [12, 9, 4]. Using this method, the system first extracts terms for each input patent document. Next, the system selects documents containing at least one of the terms. It uses the following equation to calculate Sim_{BM25} for each selected document. We used Sim_{BM25} as the similarity between an input patent document and each patent document in the training data.

$$Sim_{BM25} = \sum_{t \in T}(W_d \times W_q), \quad (5)$$

$$W_d = \frac{(k_1+1)tf}{k_1((1-b)+b\frac{dl}{avdl})+tf}, \quad (6)$$

$$W_q = \frac{(k_3+1)qtf}{k_3+qtf} log\frac{N}{n} \quad (7)$$

In these equations, T, tf, qtf, N, and n are the same as in SMART, dl is the length of the selected document, $avdl$ is the average length of the documents, and k_1, k_3, and b are constants determined using experiments. We used the default values described in the ruby-ir toolkit as k_1, k_3, and b ($k_1 = 1$, $k_3 = 1000$, and $b = 1$). We used $log\frac{N}{n}$ instead of $log\frac{N-n+0.5}{n+0.5}$ in the original equations of BM25 because Sim_{BM25} sometimes produced negative scores. We confirmed that higher F-measures were obtained when we made this revision in the experiments.

3. Tfidf With this method, the system first extracts terms for each input patent document. It then selects documents containing at least one of the terms. The system uses the following equation to calculate Sim_{Tfidf} for each selected document. We used Sim_{Tfidf} as the similarity between an input patent document and each patent document in the training data set.

[1] We used only nouns as terms, And we used ChaSen[7] to identify the nouns.

Table 4. Experimental results from the F-term categorization dry run

Similarity method	Parameters	A-Precision	R-Precision	F-measure
SMART	$k = 101, k_r = 0.99, k_p = 0.3$	0.4253	0.4064	0.3878
BM25'	$k = 101, k_r = 0.99, k_p = 0.3$	0.4138	0.3916	0.3745
BM25	$k = 101, k_r = 0.99, k_p = 0.3$	0.3690	0.3608	0.3367
Overlap	$k = 101, k_r = 0.99, k_p = 0.3$	0.3590	0.3532	0.3344
Tfidf	$k = 101, k_r = 0.99, k_p = 0.3$	0.3201	0.3204	0.3133

Table 5. Experimental results from the F-term categorization formal run

Similarity method	Parameters	A-Precision	R-Precision	F-measure
SMART	$k = 101, k_r = 0.99, k_p = 0.3$	0.4518	0.4024	0.3840
BM25'	$k = 101, k_r = 0.99, k_p = 0.3$	0.4445	0.3973	0.3783
BM25	$k = 101, k_r = 0.99, k_p = 0.3$	0.4174	0.3755	0.3521
Overlap	$k = 101, k_r = 0.99, k_p = 0.3$	0.3872	0.3473	0.3327
Tfidf	$k = 101, k_r = 0.99, k_p = 0.3$	0.3579	0.3243	0.3096

$$Sim_{Tfidf} = \sum_{t \in T} tf \times log\frac{N}{n}, \qquad (8)$$

In this equation, T, tf, N, and n are the same as in SMART.

4. Overlap The system first selects terms for each input patent document. Next, it selects documents containing at least one of the terms. The system uses the following equation to calculate $Sim_{Overlap}$ for each selected document. We used $Sim_{Overlap}$ as the similarity between an input patent document and each patent document in the training data set.

$$Sim_{Overlap} = \sum_{t \in T} 1, \qquad (9)$$

In this equation, T is the same as in SMART.

6 Sections used to extract terms

We extracted terms from the following two sections of the patent document.

1. Abstract
2. Claims

7 Experiment

7.1 Experiments in the classification subtask

We conducted experiments during the classification subtask. We used our modified version of the k-nearest neighbor method, along with various other methods to calculate similarity. We extracted terms from two sections (the abstract and the claims section).

The results are presented in Tables 4 and 5. BM25' indicates our modified method of BM25 using $log\frac{N}{n}$ instead of $log\frac{N-n+0.5}{n+0.5}$. BM25 indicates the original BM25 using $log\frac{N-n+0.5}{n+0.5}$.

Tables 4 and 5 indicate the following.

- When we compared the similarity calculation methods, SMART had the best score. We confirmed that SMART was effective.
- Comparing BM25 and BM25', we confirmed that our modification of BM25 (BM25') was effective.

In a previous paper, we described many experiments we conducted, including a comparison of variations of the k-nearest neighbor method and statistical tests [8]. Please refer to that paper for more details.

8 Conclusion

Patent processing is important in fields such as industry, business, and law. In an classification subtask we participated in at NTCIR-6, we classified patent documents into their F-terms using the k-nearest neighbor method. F-term categories are precise and useful for classifying patent documents. We used five systems in the classification subtask and obtained good results. This indicates that our method was effective. By comparing various similarity calculation methods, we confirmed that SMART was the most effective method in our experiments.

In the future, we would like to construct application systems that show users the results of classifying patent documents by applying the automatic F-term classification technique used in this study.

References

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[2] M. Iwayama, A. Fujii, and N. Kando. Overview of patent retrieval task at NTCIR-4. *Proceedings of the 4th NTCIR Workshop*, 2004.

[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. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. Evaluating patent retrieval in the third ntcir workshop. *Information Processing and Management*, 42:207–221, 2006.

[5] JPO. Japan patent office. 2005. www.jpo.go.jp/ index.html.

[6] JPO. Japan patent office, 2005. http://www.jpo.go.jp/ index.html.

[7] 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.

[8] 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.

[9] M. Murata, Q. Ma, K. Uchimoto, H. Ozaku, M. Utiyama, and H. Isahara. Information retrieval using location and category information. *Journal of the Association for Natural Language Processing*, (2):141–160, 2000. (in Japanese).

[10] NTCIR committee. NTCIR-6 Patent Retrieval Task. 2006. http://if-lab.slis.tsukuba.ac.jp/fujii/ntcpat/index-en.html.

[11] E. S. Ristad. Maximum Entropy Modeling for Natural Language. ACL/EACL Tutorial Program, Madrid, 1997.

[12] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the third Text REtrieval Conference (TREC-3)*, pages 109–126, 1994.

[13] I. Schellner. Japanese file index classification and F-terms. *World Patent Information*, 24:197–201, 2002.

[14] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96)*, pages 21–29, 1996.

[15] A. Singhal, J. Choi, D. Hindle, and F. Pereira. At&t at trec-6. In *SDR Track in NIST Special Publication 500-226: The 6th Text REtrieval Conference (TREC6)*, pages 227–232, 1997.

[16] M. Utiyama. Information retrieval module for ruby, 2005. www2.nict.go.jp/jt/a132/members/mutiyama/software.

[17] M. Utiyama and H. Isahara. Large scale text classification. *9th Annual Meeting of the Association for Natural Language Processing*, 2003. (in Japanese).

[18] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.