Hiroshima City University at NTCIR-7 Patent Mining Task

Hidetsugu Nanba
Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima 731-3194 Japan
Phone & FAX: +81-82-830-1584
nanba@hiroshima-cu.ac.jp

Abstract

Our group participated in the Patent Mining Task of the NTCIR-7. We constructed three systems HCU1, HCU2, and HCU3. HCU1 was based on the k-nearest neighbor method using an IR system developed for the NTCIR-6 Patent Retrieval Task. HCU' was the modified version of HCU1 using a method analyzing the structures of titles. HCU3 used automatically created lists of technical terms for each IPC code. We submitted these systems to the Japanese subtask, and obtained 39.13, 39.06, and 14.12 MAP scores, respectively.

Keywords: *k-nearest neighbor method, structure of titles, technical term recognition*

1 Introduction

The need for academic researchers to retrieve patents and research papers is increasing, because applying for patents is now considered an important research activity. However, retrieving patents using keywords is a laborious task for researchers, because the terms used in patents for the purpose of enlarging the scope of the claims are generally more abstract than those used in research papers. Therefore, we have constructed a framework that facilitates patent retrieval for researchers [Kamaya, et al, 2007], and have integrated research papers and patents [Nanba, et al., 2008].

The final goal of our work was to generate technical trend maps from research papers and patents. It is considered that the assigning of research papers to patent classification is a prerequisite for the generation of technical trend maps. We therefore participated in the Patent Mining Task [Nanba et al., 2008] at the NTCIR-7.

This paper is organized as follows. Section 2 describes our approaches. Section 3 contains a system description. To investigate the effectiveness of our method, we conducted experiments, as reported in Section 4. We present our conclusions in Section 5.

2 Our Approaches

We proposed three methods for the classification of research papers.

2.1 The k-Nearest Neighbor-based Approach

The most standard approach for document classification in recent years is applying machine learning [Sebastiani, 2002], such as Support Vector Machine or Naïve Bayes. However, this approach is not easy to apply to the Patent Mining Task, because the number of classes (IPC codes) is large, and it incurs a high calculation cost. We therefore employed the k-nearest neighbor (k-NN) method, which does not require machine learning. This method requires patent documents with manually assigned IPC codes, and an IR system. We used the unexamined Japanese patent applications that were provided by the organizers of the Patent Mining Task, and an IR system [Nanba, 2007], which we constructed for the Patent Retrieval Task [Fujii et al., 2007] in the NTCIR-6. The procedure of our approach was as follows.

1. Retrieve the top k results using the patent retrieval engine for a given query (research paper).
2. Extract the IPC codes with relevance scores for the query from each retrieved patent in step 1.
3. Rank IPC codes using the following equation.

$$\text{Score}(X) = \sum_{r=1}^{n} \text{Relevance score of each patent}$$

Here, X indicates an IPC code and n is the number of patents that X is assigned to within the top k retrieved patents.

2.2 Modification of the k-NN-based Approach Using Analysis of the Structure of Titles

It is generally considered that some terms in a title of a research paper contribute to automatic classification, whereas others do not. For example, within the following title, "サポートベクトルマシンを用いた自動要約 (Automatic Summarization using Support Vector Machine)", "自動要約 (Automatic Summarization)" is useful for classification, whereas "サポートベクトルマシン (Support Vector Machine)" is not, because this term is used in various fields.

Therefore, we analyzed the structures of titles, and detected topic terms that were useful for the classification of research papers, after which we improved the k-NN-based approach, which we described in Section 2.1, using these terms.

Following is an example of the analysis result achieved by our system for the above title.

<METHOD>サポートベクトルマシン (Support Vector Machine) </METHOD>を用いた (using) <HEAD>自動要約 (Automatic Summarization) </HEAD>

In this title, the "METHOD" tag was assigned to "サポートベクトルマシン (Support Vector Machines)", because the cue phrase, "を用いた (based on)" appeared just after it. The "HEAD" tag was assigned to the last noun phrase in the title. We prepared 165 cue phrases to analyze the structure of the titles. Using these cue phrases, we manually made rules to assign 10 types of tags to each word in a title. We show some of these tags and cue phrases in Table 1.

Table 1. Some tags and cue phrases

Tag	Cue phrases
METHOD	を用いた(using), に基づいた(is based on), による(by)
RESTRICT	における(at), に関する(in), の(of)
GOAL	に向けて(towards), のための(for)
CONJ	と(and), や(and), 及び(and)

Among 10 types of tags, we give weight to the "HEAD" tag, because it indicates the main topic of the paper.

2.3 Using Automatically Recognized Domain Specific Terms for Each IPC Code

Generally, the k-NN-based approach is useful for classifying documents into a large number of categories. However, when the number of documents for each category varies much, the k-NN-based approach will be greatly affected by the imbalance, because it tends to classify documents to categories that contain many documents. For the Patent Mining Task, the number of patents for each IPC code varies very much. The number of IPC codes containing only one patent is 1,181, whereas the number of IPC codes containing more than 10,000 patents is 43[1] from among 31,520 IPC codes in the Japanese patent database. To solve this problem, we proposed a new approach.

Our approach used automatically created lists of technical terms for each IPC code. Nakagawa and Mori proposed a method that automatically recognized domain-specific terms from a set of documents in a specific field [Nakagawa and Mori, 2003]. The method assumes that component words in a technical term tend to form other technical terms by combining with other component words. Using Nakagawa's method, we constructed lists of domain specific terms for each IPC code using the following procedure.

1. Select an arbitrary IPC code.
2. Collect all patent documents for the IPC code was manually assigned, from 1993 to 2002 (3.5 million documents).
3. Apply Nakagawa's method, and create a list of terms specified in the IPC.
4. Return to step 1.

Finally, we obtained lists of domain-specific technical terms for each IPC code. Table 2 is part of the list for "G06F17/28" (machine translation). The numerical values shown with each term are the relevance scores automatically calculated by Nakagawa's method. From the results in Table 2, we can find that most of the terms are relevant to the "machine translation" field, but the list also contains inadequate terms, such as "場合 (case)", "構成 (constitution)", and "本発明 (the invention)". These terms are commonly used in patents. TO reduce the scores of these terms, we used the following equation for the calculation of S(T), which indicates a relevance score for the term T.

$$S(T) = \text{Score of term T by Nakagawa's mehotd} \times IIF$$

Here, IIF (Inverse IPC Frequency) was calculated by dividing the number of IPCs by the number of lists that contain the term T. If the term T appeared in many lists, its S(T) score would be lower than the original score calculated by Nakagawa's method.

Table 2. An automatically generated term list for G06F17/28 (machine translation)

Term	Score
言語 (language)	1390612790.07
翻訳 (translation)	1310117190.84
単語 (word)	1224644033.02
訳語 (translated word)	1184913164.93
訳文 (translated sentence)	726003355.74
記憶 (memory)	693395802.62
入力 (input)	634170775.96
表示 (display)	632619205.42
処理 (process)	489133063.49
文書 (document)	479121143.31

Using these lists, we calculated the relevance of each IPC code for a given topic (research paper) using the following procedure.

1. Extract technical terms from a given query (research paper) using Nakagawa's method[2].
2. Calculate the similarities between the list of terms extracted in step 1 and lists of terms preliminarily created for each IPC code[3].

[1] The average patent number for an IPC code is 224.3.

[2] TermExtract: http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

3. Rank IPC codes by their similarities.

3 System Descriptions

We constructed three systems, HCU1, HCU2, and HCU3 for the Patent Mining Task. Of these, we explain the details of the HCU1 and HCU2 systems in the remainder of this section.

Both HCU1 and HCU2 employed the k-NN approach. These systems comprise the following three steps.

(Step 1) Morphological analysis

We introduce the Vector Space Model as a retrieval model and SMART [Salton, 1971] for term weighting. We use GETA[4] as a retrieval engine and MeCab[5] for Japanese morphological analysis tools.

(Step 2) Stopword deletion

Our systems use nouns, verbs, adjectives, and unknown words to retrieve relevant patents. In this method, unimportant words are stripped from terms that are in any of the above parts of speech and are extracted from the query. In this step, HCU2 extracts topic terms from the title using the title analysis method described in Section 2.2, and gives weights[6] to these terms, to which the HEAD tags are assigned.

(Step 3) Retrieval of relevant IPC codes

Our systems retrieved the top k patents, and then ranked relevant IPC codes using the procedure described in Section 2.1. Here, we used the values of 170 and 90 for k for HCU1 and HCU2, respectively. These values were determined using the dry run data, which were provided by the organizers of the Patent Mining Task.

4 Evaluation

4.1 Data and Evaluation

We used 879 topics for the Japanese subtask to evaluate our patent retrieval method [Nanba et al., 2008]. In this data set, relevance judgements were conducted from the following two viewpoints.

- A paper in an exception field and a candidate paper are exactly the same (group A: topics 300-772)
- Authors and research topics of the two papers are almost the same, but the publication years are different (group B: topics:1000-1405)

All the systems were evaluated by mean average precision (MAP).

[3] For calculating the similarity, we used cosine distance. We used scores calculated by Nakagawa's method as weights for each term.
[4] GETA: http://geta.ex.nii.ac.jp/
[5] MeCab: http://mecab.sourceforge.net/
[6] Add one to the tf (term frequency) scores of these terms.

4.2 Results and Discussion

For the formal run, we submitted the three results provided by HCU1, HCU2, and HCU3 to the Japanese subtask. The experimental results are shown in Table 3. As can be seen from the table, MAP scores of the k-NN based systems HCU1 and HCU2 were much better than that of HCU3.

Table 3. MAP for Japanese Subtask

	MAP
HCU1	39.13
HCU2	39.06
HCU3	14.12

Table 4. Recall for Top n Results (groups A + B)

rank	HCU1	HCU2	HCU3
1	17.1(351)	17.0(348)	6.1(126)
2	27.5(564)	27.4(561)	8.8(181)
3	33.3(682)	33.3(682)	11.0(225)
4	37.9(777)	37.8(776)	12.5(257)
5	41.1(843)	41.6(853)	13.8(284)
10	51.9(1065)	52.4(1075)	19.8(407)
20	62.2(1276)	62.1(1274)	27.0(553)
50	73.4(1506)	73.2(1503)	37.8(776)
100	77.1(1581)	77.4(1587)	46.6(955)
500	77.9(1598)	78.4(1607)	68.0(1394)
1000	77.9(1598)	78.4(1607)	76.0(1559)

Table 5. Recall for Top n Results (group A)

rank	HCU1	HCU2	HCU3
1	19.4(216)	18.9(211)	5.8(65)
2	31.0(346)	31.2(348)	8.3(93)
3	38.3(428)	38.4(428)	10.6(118)
4	43.7(487)	43.7(487)	12.5(139)
5	47.6(531)	47.9(534)	13.9(155)
10	58.6(653)	58.8(656)	20.4(227)
20	69.8(778)	69.5(775)	28.6(319)
50	80.9(902)	80.8(901)	40.4(451)
100	84.3(940)	84.4(941)	49.8(555)
500	84.7(944)	84.8(946)	70.3(784)
1000	84.7(944)	84.8(946)	77.3(863)

We also show recall values for the top n results in Tables 4 and 5. Table 4 shows the results using all topics, whereas Table 5 shows the results using the topics in group A. From the results in Table 5, we could determine that almost 60% of IPC codes were found within top the 10 results and 84% were found within the top 100 for HCU1 and HCU2. For the generation of the technical trend map from research papers and patents, we need to improve recall at top 1, but still we believe that these results are useful for supporting beginners in patent search. It is often necessary for searchers to use patent classification codes for effective patent retrieval, but professional skills and much experience are required for the selection of relevant IPC codes. In such cases, our systems are useful to look for relevant IPC codes.

Error analysis of HCU2

To confirm the effects of the title analysis method, we randomly selected 100 cases from the results of analysis of the titles' structures by HCU2. From 100 cases, HEAD tags were mistakenly assigned in 15 cases. Following are some of the terms for which our title analysis method mistakenly assigned HEAD tags.

新概念(new concept), 防止策(prevention measure), 付与(assignment), 試験データ(test data), キャラクタリゼーション(characterization)

As these terms are often used in many fields, giving weights to these terms leads to impairing the MAP score of HCU2 in comparison to that of HCU1.

5 Conclusions

We participated in the NTCIR-7 Patent Mining Task. We constructed three systems, HCU1, HCU2, and HCU3. HCU1 was based on the k-nearest neighbor method using an IR system developed for the NTCIR-6 Patent Retrieval Task. HCU2 was the modified version of HCU1 using a method of analyzing the structures of titles. HCU3 used automatically created lists of technical terms for each IPC code. We submitted these systems to the Japanese subtask, and obtained 39.13, 39.06, and 14.12 MAP scores, respectively.

References

A. Fujii, M. Iwayama, and N. Kando. Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. Proceedings of the 6th NTCIR Workshop Meeting, 2007.

M.A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–545, 1992.

H. Kamaya, H. Nanba, M. Okumura, A. Shinmori, H. Tanigawa, and T. Suzuki. Paraphrasing Scholarly Terms into Patent Terms using Citation Relations between Research Papers and Patents. IPSJ SIG Notes NL-178, pp.97-102. 2007. (in Japanese)

H. Nakagawa and T. Mori. Automatic Term Recognition based on Statistics of Compound Nouns and their Components. Terminology, Vol.9 No.2, pp. 201-209, 2003.

H. Nanba. Query Expansion using an Automatically Constructed Thesaurus. In Proceedings of the 6th NTCIR Workshop, 414-419, 2007.

H. Nanba, N. Anzen, and M. Okumura. Automatic Extraction of Citation Information in Japanese Patent Applications. International Journal on Digital Libraries, 2008.

H. Nanba, A. Fujii, M. Iwayama, T. Hashimoto. Overview of the Patent Mining Task at the NTCIR-7 Workshop. In Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

G. Salton. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, 1971.

F. Sebastiani. Machine Learning in Automated Text Categorization. ACM Computing Surveys (CSUR), Vol. 34, Issue 1, pp.1-47, 2002.

Y. Taniguchi, and H. Nanba. Identification of Bibliographic Information Written in both Japanese and English. Research and Advanced Technology for Digital Libraries, 12th European Conference, ECDL 2008, Aarhus, Denmark, September 2008 Proceedings, Lecture Notes in Computer Science, Vol. 5713, Springer, pp. 431-433, 2008.