Automatic Translation of Scholarly Terms into Patent Terms

Hidetsugu Nanba
Hiroshima City University
nanba@hiroshima-cu.ac.jp

Hideaki Kamaya
Hitachi Systems & Services
kamaya@ls.info.hiroshima-cu.ac.jp

Toshiyuki Takezawa
Hiroshima City University
takezawa@hiroshima-cu.ac.jp

Manabu Okumura
Tokyo Institute of Technology
oku@pi.titech.ac.jp

Akihiro Shinmori
INTEC Systems Institute
shinmori_akihiro@intec-si.co.jp

Hidekazu Tanigawa
IRD Patent Office
htanigawa@ird-pat.com

ABSTRACT

For a researcher in a field with high industrial relevance, retrieving research papers and patents has become an important aspect of assessing the scope of the field. However, retrieving patents using keywords is a laborious task for researchers, because the terms used in patents are often more abstract than those used in research papers, to try to widen the scope of the claims. We propose two methods for translating scholarly terms into patent terms: the "citation-based method" and the "thesaurus-based method". We also propose a method combining these two with the existing "Mase's method". To confirm the effectiveness of our methods, we conducted some examinations, and found that the combined method performed the best.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Search process

General Terms

Experimentation, Theory

Keywords

cross-genre IR, patent, research paper, citation, thesaurus

1. INTRODUCTION

We propose a method for translating scholarly into patent terms. For example, our method translates a scholarly term "floppy disc" into patent terms, such as "magnetic recording device" or "removable recording media". This translation technology can support users when retrieving both research papers and patents.

For a researcher in a field with high industrial relevance, retrieving research papers and patents has become an important aspect of assessing the scope of the field. Examples of such fields are bioscience, medical science, computer science, and materials science. In addition, research paper searches and patent searches are required by examiners in government patent offices, and by the intellectual property divisions of private companies. An example is the execution of an invalidity search among existing patents or research papers that could invalidate a rival company's patents or patents under application in a Patent Office. However, the terms used in patents are often more abstract or creative than those used in research papers, to try to widen the scope of the claims. Therefore, a technology for translating scholarly terms into patent terms is required.

This technology is also useful in the following situation. When inventors or patent attorneys write patents, they are often confused about which patent terms they should use, because there may be several choices of patent terms for a scholarly term. For example, the scholarly term "floppy disc" can be expressed as "removable recording medium", if the inventors or patent attorneys focus on the floppy disc's feature of removablility. On the other hand, "floppy disc" can also be expressed as "magnetic recording medium", if they focus on the feature of recording information using magnetic force. In such a situation, if it can generate a list of candidate patent terms for a given scholarly term, this technology would support the inventors and the patent attorneys while writing patents.

The remainder of this paper is organized as follows. Section 2 describes some related work. Section 3 proposes our method for translating scholarly terms into patent terms. Section 4 discusses how we investigated the effectiveness of our method by conducting some examinations, and discusses our experimental results. Finally, we provide our conclusions in Section 5.

2. RELATED WORK

There has been much research in the field of cross-genre information retrieval, such as that discussed in the technical survey task of the Patent Retrieval Task of the Third NII Test Collection for Information Retrieval (NTCIR) work-

PaIR'09, November 6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-809-4/09/11 ...$5.00.

shop [2]. This task aimed to retrieve patents relevant to a given newspaper article. In this task, Itoh et al. focused on "Term Distillation" [1]. The distribution of the frequency of the occurrence of words was considered to be different between heterogeneous databases. Therefore, unimportant words were assigned high scores when using TFIDF to weight words. Term Distillation is a technique that can prevent such cases by filtering out words that can be assigned incorrect weights. However, some patent terms, such as "magnetic recording device", appear only in a patent database, and "Term Distillation" can not be applied in such cases.

The Patent Mining Task in the Seventh NTCIR workshop is another research project related to cross-genre information access [6]. The aim of this task was the classification of research papers written in either Japanese or English in terms of the International Patent Classification (IPC) system. Although, we did not examined our method using this data set, our method can also be applied to this task.

Nanba et al. proposed a method to integrate a research paper database and a patent database by analysing citation relations between research papers and patents [5]. For the integration, they extracted bibliographic information of cited literatures in "prior art" fields in Japanese patent applications. Using this integrated database, users can retrieve patents that relate to a particular research paper by tracing citation relations between research papers and patents. However, the number of cited papers among patent applications is not enough to retrieve related papers or patents, even though the number of opportunities for citing papers in patents or for citing patents in papers has been increasing recently. We therefore have studied automatic translation of scholarly terms into patent terms.

3. AUTOMATIC TRANSLATION OF SCHOLARLY TERMS INTO PATENT TERMS

We propose three translation methods: the "citation-based method", the "thesaurus-based method", and "Mase's method". We describe these methods in the following subsections. We then describe a method that combines the three methods.

3.1 Translation Using Citation Relationships Between Research Papers and Patents

A research paper and a patent that have citation relationships with each other, generally tend to be in the same research field. Using this idea, translation of a scholarly term can be realized by using the following procedure.

1. Input a scholarly term,
2. Retrieve research papers that contain the given scholarly term in their titles.
3. Collect patents that have citation relationships with the papers retrieved in Step 2.
4. Extract patent terms from patents collected in Step 3, and output them in order of frequency.

We call this the "citation-based method". To extract patent terms from patents that were collected in Step 3, we used Shinmori's method[7], which focuses on the patent claim.

3.2 Translation Using an Automatically Constructed Thesaurus

To enlarge the scope of the patent, hypernyms of scholarly terms are often used in patents. We therefore propose a method using a thesaurus in addition to the citation-based method. We used a hypernym/hyponym thesaurus, which Nanba[4] automatically constructed using a pattern "A ya B nadono C" (C, such as A (or|and) B). The thesaurus contains 7,031,159 hypernym/hyponym relations, which were extracted from Japanese patents published in the 10 years from 1993 to 2002. This thesaurus also give the frequencies of each hypernym/hyponym relation in patents.

Using this thesaurus, we realize translation of a scholarly term by extracting hypernyms of the given scholarly term from the thesaurus, and by outputting them in order of frequency. We call this the "thesaurus-based method".

3.3 Translation Using Mase's Method

In patent applications, inventors may explicitly describe related terms by using parentheses, as in "floppy disc (magnetic recording medium)". The term preceding the parentheses and the term in parentheses have a broader/narrower relationship. Mase et al.[3] extracted related term from the text in the "description of symbols" fields of Japanese patents. They experimentally confirmed that these terms are effective for query expansion of patent retrieval. This method can also be used in our work.

Using Mase's method, we realize translation of a scholarly term by extracting related terms of a given scholarly term from the "description of symbols" fields, and by outputting them in order of frequency.

3.4 Translation Combining the Three Methods

We propose a method combining the above three methods in the following two steps.

(Step 1) Combining Mase's method with the other two methods

Using Mase's method, we extracted a total of 679,931 pairs of related terms. We translated some scholarly terms into patent terms and found that Mase's method could output correct patent terms at high rates. However, the number of terms obtained by Mase's method is very small and in the worst case, no terms were output[1]. Therefore, we improve the citation-based and the thesaurus-based methods using Mase's method, instead of using Mase's method by itself. Consider an example in which Mase's method obtained two patent terms "magnetic recording device" and "removable storage device" for a given scholarly term "floppy disc". From these results, "floppy disc" can be inferred to be a term related to a "device", because the last word of both patent terms is "device". If there is another patent term for "floppy disc", the last word of the term is probably "device". Therefore, we improve both the citation-based and the thesaurus-based methods by giving a higher priority using Mase's method. The procedure is as follows.

1 Normalize the frequencies of each candidate term in a list given by the citation-based method (or the thesaurus-based method) to a value between 0 and 1 by dividing each score by the score of the term ranked 1.

2 Extract the last word of each candidate term obtained by Mase's method. In this step, we also extract the frequencies of each term in "description of symbols" fields.[2]

[1] We will report this experimental result later.

[2] When Mase's method outputs three candidate terms "magnetic recording device" (freq. 10), "removable storage de-

3 Sum the scores (frequencies) for each last word obtained in Step2, and normalize them to a value between 0 and 1 by dividing each score by the score of the word at rank 1.[3]

4 When the last word of a candidate term by the citation-based method (or the thesaurus-based method) and one of the words obtained in Step 3 match, give the scores of their words to each term, and output in order of score.[4]

(Step 2) Combining the citation-based method and the thesaurus-based method

The terms output by both the citation-based and the thesaurus-based methods, which were improved by Mase's method, seem to be correct patent terms. We therefore combine both methods using the following equation.

$$\begin{aligned}&\text{Score of a candidate patent term by the combined method}\\ &\quad = \lambda * \text{Score by the citation-based method}\\ &\quad +(1-\lambda) * \text{Score by the thesaurus-based method}\end{aligned}$$

Here, λ is a parameter that adjusts the effects of the citation-based and the thesaurus-based methods. We will describe how to determine this parameter in Section 4.1.

4. EXPERIMENTS

To confirm the effectiveness of our methods, we conducted some examinations. We describe the experimental conditions in Section 4.1, report the experimental results in Section 4.2, and discuss the results in Section 4.3.

4.1 Experimental Conditions

Documents

We used Japanese patent applications published in the 10 years from 1993 to 2002. We also used about 85,000 bibliographic records of cited papers in patents, which were automatically created using Nanba's method[5]. We created the correct data set using the following procedure.

1. Extract all noun phrases from the 85,000 bibliographic records of cited papers in patents, and rank them in order of frequency.
2. Manually select scholarly terms from the noun phrases.
3. Output candidate terms using all our methods and baseline methods, which we will describe later.
4. Manually identify correct patent terms in all candidates obtained in Step 3.

Finally, we obtained 47 scholarly terms (input) with 2.8 patent terms (output) on average for each scholarly term. We show some of these in Table 1.

Table 1: Data for evaluation (example)

scholarly term (input)	patent term (output)
word processor	document processing device, document information processing device, document editing system, document writing support system
TV camera	photographic device, image shooting apparatus, image pickup apparatus

Evaluation Measure

As an evaluation measure, we used ϵ, which is an expansion of MRR, a standard evaluation measure for evaluating question-answering systems. The evaluation score will be close to one when many correct terms are given high ranks.

$$\epsilon = \frac{\sum_{i \in R} \frac{1}{i}}{\sum_{j \in \{1,2,\ldots,n\}} \frac{1}{j}}$$

Here, n indicates the number of correct patent terms for a given scholarly term, R indicates a set of ranks of correct terms in a system output, i is the rank of a correct term in a system output. In addition to the ϵ measure, we also used recall and precision.

$$\text{Recall} = \frac{\textit{The number of correctly extracted patent terms}}{\textit{The number of correct patent terms}}$$

$$\text{Precision} = \frac{\textit{The number of correctly extracted patent terms}}{\textit{The number of candidate terms extracted by a system}}$$

We evaluated only the top 20 terms in each system output.

Alternatives

We conducted experiments using the following methods. Abbreviations for each method are shown in parentheses.

Our methods

(1) Citation-based method (Cite)

(2) (1) + improvement by Mase's method (Cite(M))

(3) Thesaurus-based method (Thes)

(4) (3) + improvement by Mase's method (Thes(M))

(5) (2) + (4) combined method (Cite(M)+Thes(M))

Baseline methods

(6) Mase's method (Mase)

(7) Term co-occurrence-based method (GETA)

(8) Synonyms extraction method (Syn)

(9) JST thesaurus-based method (JST)

Methods (1), (3), (5), and (6) correspond to those mentioned in Sections 3.1, 3.2, 3.3, and 3.4, respectively. Methods (2) and (4) are improved by Mase's method as described in Section 3.3. We will explain the procedures for parameter tuning later.

As one baseline method, we employed the word co-occurrence method (7). In this method, terms co-occurred frequency with a given scholarly term are extracted as candidates using the IR engine GETA.

As another baseline method, we used an automatically constructed synonym dictionary[4]. Nanba constructed a thesaurus using a pattern "A ya B nadono C" (C, such as A (or|and) B). In the expressions, there are several cases in which parentheses were used. He obtained 50,161 pairs of synonyms, and confirmed that the synonyms were useful for

vice" (freq. 5), and "information recording medium" (freq. 3), the three words "device" (freq. 10), "device" (freq. 5), and "medium" (freq. 3) are extracted from the terms.

[3] For the example in Step 2, "device" (score 15) and "medium" (score 3) are obtained. Then, the scores of the words are normalized by dividing by 15, which is the score for "device", resulting in "device" (score 1) and "medium" (score 0.2).

[4] For example, if the citation-based method obtained a term "recording medium" (score 0.5), a score 0.2 $\times m$ for "medium" is added to 0.5. Here, m is a parameter that indicates the influence of Mase's method on the citation-based method. We will describe how to determine m in Section 4.1.

Table 2: Evaluation using ϵ

Our method					Baseline			
(1)Cite	(2)Cite(M)	(3)Thes	(4)Thes(M)	(5)Cite(M)+Thes(M)	(6)Mase	(7)GETA	(8)Syn	(9)JST
0.136	0.173	0.231	0.240	0.298	0.107	0.011	0.058	0.050

query expansion in patent retrieval. We used this synonym dictionary as a baseline method (8).

As the other baseline method, we used a free online thesaurus, which was provided by the Japan Science and Technology Agency (JST). Using the JST thesaurus, we translate scholarly terms into patent terms, in the same way as the thesaurus-based method, which we mentioned in Section 3.2.

Parameters in methods (2), (4), and (5)

We conducted a pilot study to determine a value for parameter m, which indicates the influence of Mase's method for both the citation-based and the thesaurus-based methods, and a value of λ, which adjusts the relative contributions of the citation-based and the thesaurus-based methods. We prepared a data set that consists of 25 scholarly terms and their correct patent terms, and used it for the pilot study. The pilot study was conducted in two steps. In the first step, we changed values of m from 0 to 1 at 0.1 intervals, and calculated ϵ scores of the citation-based method (2) and the thesaurus-based method (4). We found the highest ϵ scores, when m for method (2) was 0.8, and m for method (4) was 0.2. In the second step, we optimized the λ score by changing it from 0 to 1 at 0.1 intervals, and calculating the ϵ scores for each step. We obtained the highest ϵ score, when λ was 0.3. We used this score for the combined method (9).

4.2 Experimental Results and Discussion

The ϵ scores for each method are shown in Table 2. We further investigated our methods (2), (4), (5) and baseline methods (6) and (8), all of which obtained better scores among all methods compared. The results are shown in Table 3. In the table, we also show recall, precision, and ϵ scores for an ideal system. Here, precision scores for the ideal system were less than one, because the average number of correct patent terms for each scholarly term is 2.8. Therefore, the scores for the ideal system are an upper bound.

In Table 2, we see that the ϵ score for method (2) is larger by 0.037 points than that by for method (1), which indicates that Mase's method was effective in improving the citation-based method. On the other hand, Mase's method did not improve the thesaurus-based method, because the difference in ϵ scores for methods (3) and (4) is only 0.009. However, the performance by the thesaurus-based method was good enough, and there was little room to improve the thesaurus-based method by Mase's method. The combined method (5) obtained the best ϵ score of all methods. This method also obtained best recall and precision scores in Table 3.

In Table 2, the ϵ score for the JST thesaurus-based method (9) was smaller than those for the thesaurus-based methods (3) and (4), although the JST thesaurus was manually created, while the thesaurus used in methods (3) and (4) was created automatically. This result was caused by the number of terms in the JST thesaurus. The original JST thesaurus contains about 400,000 scholarly terms, but the freely available online version contains only 10% of the original. As a result, there were many cases in which no terms were extracted by the method (9). If we had been able to use the original one, the performance of method (9) would be better.

Table 3: Evaluation using ϵ, Recall, and Precision

	Method	Measure	top 5	top10	top15	top20
Our method	(2) Cite(M)	ϵ	0.151	0.165	0.170	0.173
		Recall	0.169	0.242	0.275	0.311
		Prec.	0.115	0.073	0.056	0.047
	(4) Thes(M)	ϵ	0.213	0.235	0.239	0.240
		Recall	0.274	0.362	0.393	0.399
		Prec.	0.145	0.104	0.078	0.061

	(5)
	Cite(M)
	+Thes(M)
		ϵ	0.261	0.286	0.292	0.298

		Recall	0.309	0.421	0.459	0.533
		Pre.	0.170	0.121	0.092	0.076
Base -line	(6) Mase	ϵ	0.083	0.097	0.106	0.107
		Recall	0.108	0.172	0.246	0.264
		Prec.	0.072	0.061	0.055	0.045
	(8) Syn	ϵ	0.054	0.055	0.057	0.058
		Recall	0.080	0.087	0.101	0.104
		Prec.	0.053	0.038	0.037	0.035
Upper bound		ϵ	1.000	1.000	1.000	1.000
		Recall	1.000	1.000	1.000	1.000
		Prec.	0.587	0.294	0.196	0.147

5. CONCLUSIONS

In this paper, we have proposed three methods: the citation-based method, the thesaurus-based method, and the method combining these two methods. To confirm the effectiveness of our methods, we conducted some examinations. We found that the combined method performed the best in terms of recall, precision, and ϵ, which is an extensional measure of Mean Reciprocal Rank (MRR) widely used for the evaluation of question-answering systems.

6. REFERENCES

[1] H. Itoh, H. Mano, and Y. Ogawa. Term Distillation for Cross-db Retrieval. Working Notes of NTCIR-3 Meeting, Part III, pages 11–14, 2002.

[2] M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of Patent Retrieval Task at NTCIR-3. Working Notes of NTCIR-3 Meeting, Part III, pages 1–10, 2002.

[3] H. Mase, T. Matsubayashi Y. Ogawa, Y. Yayoi, Y. Sato, and M. Iwayama. NTCIR-5 Patent Retrieval Experiments at Hitachi. Proceedings of NTCIR-5 Meeting, pages 318–323, 2005.

[4] H. Nanba. Query Expansion using an Automatically Constructed Thesaurus. Proceedings of NTCIR-6 Meeting, pages 414–419, 2007.

[5] H. Nanba, N. Anzen, and M. Okumura. Automatic Extraction of Citation Information in Japanese Patent Applications. International Journal on Digital Libraries, Vol. 9, No. 2, pages 151–161, 2008.

[6] H. Nanba, A. Fujii, M. Iwayama, and Y. Hashimoto. The Patent Mining Task in the Seventh NTCIR Workshop. Proceedings of PaIR'08, pages 25–31, 2008.

[7] A. Shinmori, M. Okumura,, Y. Marukawa, and M. Iwayama M. Rhetorical Structure Analysis of Japanese Patent Claims using Cue Phrases. Working Notes of NTCIR-3 Meeting, Part III, pages 69–77, 2002.