The Patent Mining Task in the Seventh NTCIR Workshop

The Patent Mining Task Organizers
2008.3.31

ABSTRACT
This paper introduces the Patent Mining Task in the Seventh NTCIR Workshop, which is currently in progress, and the test collections produced in this task. Its goal is the classification of research papers written either in Japanese or in English into the International Patent Classification (IPC) system, which is a global standard patent classification system.

Keywords
test collections, classification of research paper, patent

1. INTRODUCTION
The Patent Mining Task in the Seventh NTCIR Workshop investigates how to retrieve necessary information from both research papers and patents databases easily. To appreciate the scope of a particular research field, it has become important for researchers in research fields with a high industrial relevance, such as bioscience, medical science, computer science, and materials science. Actually, the development of an IR system of research papers and patents for academic researchers is enshrined in Intellectual Property Strategic Program 2006[1] and 2007[2] by Intellectual Property Strategy Headquarters of the Cabinet Office, Japan.

Searching both research papers and patents is also required for examiners in government patent offices, and for searchers in the intellectual property divisions of private companies. Their particular purpose is to carry out an invalidity search on existing patents or research papers that can invalidate the patents of rival companies or patents under application in a patent office.

However, the terms used in patents are generally more abstract and more creative than those used in research papers for the purpose of enlarging the scope of the claims. Thus, the Patent Mining Task aims to develop fundamental techniques for retrieving and classifying both research papers and patents.

In the past NTCIR Workshops, Patent Classification Subtask was conducted [1, 2]. In these subtasks, participants were asked to classify Japanese patent applications into File Forming Term (F-term), which is a classification system for Japanese patent documents. In addition to patents, we focus on classification of research papers. The purpose of the Patent Mining Task in NTCIR-7 is the classification of research papers written either in Japanese or in English into the International Patent Classification (IPC), which is one of the other patent classification systems. In this paper, we describe the details of this task. Currently, nineteen teams are participating in this task, and the dry run is in progress.

2. THE PATENT MINING TASK

2.1 Task Overview
As we described in Section 1, the goal of the Patent Mining Task is the classification of research papers into the IPC system, which is a global standard hierarchical patent classification system, and one or more IPC codes are manually assigned to each patent for the purpose of effective patent retrieval.

The sixth edition of this system contains more than 50,000 classes at the most detailed level. The goal of this task is to assign one or more classes among 50,000 to a given topic (a title and an abstract of a research paper). An example of an English topic is shown in Figure 1. Here, <TOPIC-ID> indicates the topic identification number. <TITLE> and <ABSTRACT> indicate a title and an abstract of a research paper to be classified, respectively.

In the task, the following two subtasks are conducted.

- Japanese subtask: classification of Japanese research papers into the IPC system.
- English subtask: classification of English research papers into the IPC system.

2.2 Relevance Judgments
A number of topics with manually assigned IPC codes are necessary for evaluation. However, it is very costly and time-consuming to create such data sets. Therefore, we created the data based on the following idea.

<TOPIC>
<TOPIC-ID> 100 </TOPIC-ID>
<TITLE> DTMF (Dual Tone Multi-Frequency) transmission method for a mobile communication system </TITLE>
<ABSTRACT> High efficient speech encoding scheme called VSELP, is adopted for Japanese digital mobile communication systems. However, DTMP (Dual Tone Multi-Frequency) signals are distorted by using this encoding scheme. This paper presents a DTMF signal transmission scheme. DTMF signals are transmitted in the form of call control messages from mobile station (MS) to mobile control center (MCC). In addition, necessary control capabilities in MS and MCC is described. </ABSTRACT>
</TOPIC>

Figure 1. An example of an English topic

[1] http://www.kantei.go.jp/jp/singi/titeki2/keikaku2006_e.pdf

[2] http://www.kantei.go.jp/jp/singi/titeki2/keikaku2007_e.pdf

(original)
【新規性喪失の例外の表示】特許法第３０条第１項適用申請有り２０００年３月１４日 社団法人情報処理学会発行の「第６０回（平成１２年前期）全国大会講演論文集（４）」に発表
(translation)
[Indication of exceptions to lack of novelty] The provisions set forth in Article 30, Paragraph 1 in Japanese patent law. Proceedings (Volume 4) of the 60th Annual Meeting of the Information Processing Society of Japan, published in March 14, 2000.

Figure 2. An example of "Indication of exceptions to lack of novelty" field

Essentially, an invention is not patentable if it was already known before the data of filing. However, the article 30 in Japanese patent law provides a six-month grace period for disclosures made through a publication or a presentation at a conference or an exhibition. In this case, the applicants need to mention the proceeding titles (or conference names) and the date it was published in "Indication of exceptions to lack of novelty" field (exception field) in the patent. Figure 2 is an example of the field.

Now, we can consider that the most of the content of the paper mentioned in the exception field overlaps with the patent. Therefore, if we regard the IPC codes that were assigned to the patent as the code that should be assigned to the research paper mentioned in the exception field, it is possible to create a large-scale data set at low cost. In fact, there are totally more than 9,000 applications in 3,496,253 Japanese patent applications published in the 10 years between 1993 and 2002.

The procedure of creating the data is as follows. First, we extracted publication years and the proceeding titles from the exception fields in the 9,000 applications. Though, titles and authors of the papers are not mentioned in this field, the authors are usually the same as the inventors of the patents. We therefore extracted and used the inventors of the patents instead of the author's names. Second, we compared these extracted data with records in a research paper database using a simple string matching method. The database was originally used in Cross-lingual Information Retrieval Task in the first and the second NTCIR Workshop. It contains 255,960 records of Japanese-English paired documents, and each record consists of a title, author(s), an abstract, keywords, a publication year, and a conference name. As a result from the automatic matching, we obtained six candidate records on average for each exception field. Thirdly, we manually identified a correct one from the six candidates. Finally, we obtained 976 patent-research paper pairs. From these pairs, we created English and Japanese topics (titles and abstracts) and their correct answers (IPC codes extracted from patents). For each topic, 2.3 IPC codes are assigned on average. Among them, we use 97 topics for the dry run, and the remainder 879 topics for the formal run. Participant teams are asked to submit one or more ranked lists[3] of IPC codes for each topic, and they are evaluated using the Mean Average Precision.

2.3 Document Sets

The document sets used in the task are shown in Table 1. The data (1) and (4) were distributed to the teams participated in Japanese subtask, while the data (2), (3), and (4) were distributed to those in English subtask.

Table 1. Document sets

Data	Year	Size	Number	Language
(1) Unexamined Japanese patent applications	1993-2002	100GB	3.50M	Japanese
(2) USPTO patent data	1993-2000	33GB	0.99M	English
(3) Patent Abstracts of Japan (English translations for Japio patent abstracts)	1993-2002	4.2GB	3.50M	English
(4) NTCIR-1, 2 CLIR Task Test Collection (Abstracts of research papers)	1988-1999	1.4GB	0.26M	Japanese / English

3. CONCLUSION

We described an overview of the evaluation and design used for the Patent Mining Task in NTCIR-7. We focused on "Indication of exceptions to lack of novelty" field in Japanese patent applications, and created 976 English and Japanese topics and their correct answers (IPC codes). Using this data set, the dry run is in progress.

4. REFERENCES

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

[2] Iwayama, M., Fujii, A., and Kando, N. 2005. Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task. Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access.

[3] The maximum number of IPC codes for a single topic is 1000.