The Patent Mining Task in the Seventh NTCIR Workshop

Hidetsugu NANBA
Graduate School of Information Sciences
Hiroshima City University
3-4-1 Ozuka-higashi, Hiroshima, 731-3194, Japan
nanba@hiroshima-cu.ac.jp

Makoto IWAYAMA
Hitachi, Ltd.
1-280 Hitashi-Kougakubo, Kokubunji, 185-8601, Japan
makoto.iwayama.nw@hitachi.com

Atsushi FUJII
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

Taiichi HASHIMOTO
Integrated Research Institute, Tokyo Institute of Technology
4259 Nagatsuta, Yokohama, 226-8503, Japan
hashimoto@iri.titech.ac.jp

ABSTRACT
This paper introduces the Patent Mining Task of the Seventh NTCIR Workshop and the test collections produced in this task. The task’s goal was the classification of research papers written in either Japanese or English in terms of the International Patent Classification (IPC) system, which is a global standard. For this task, 12 participant groups submitted 49 runs. In this paper, we also report the evaluation results of the task.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Search process
H.3.4 [Systems and Software]: Performance evaluation
H.3.5 [Online Information Services]: Data sharing

General Terms
Measurement, Performance, Experimentation

Keywords
test collection, classification of research papers, patent, International Patent Classification (IPC)

1. INTRODUCTION
The Patent Mining Task in the Seventh NTCIR Workshop (NTCIR-7) investigated the effective retrieval of necessary information from research papers and patent databases. In this paper, we introduce the task and report the evaluation results.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PaIR’08, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558-256-6/08/10...$5.00.

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 these fields are bioscience, medical science, computer science, and materials science. In fact, the development of an information retrieval system of research papers and patents for academic researchers is central to the Intellectual Property Strategic Programs for 20061 and 20072 of the Intellectual Property Strategy Headquarters in the Cabinet Office, Japan.

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, which 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, the Patent Mining Task aims to develop fundamental techniques for retrieving and classifying both research papers and patents.

In previous NTCIR Workshops, Patent Classification Subtasks were conducted [4][5]. In these subtasks, participants were asked to classify Japanese patent applications in terms of the File Forming Term (F-term) system, which is a classification system for Japanese patent documents. Here, we are focusing on the classification of research papers in addition to patents. The aim of the Patent Mining Task in NTCIR-7 was the classification of research papers written in either Japanese or English in terms of the International Patent Classification (IPC) system, an alternative patent classification system.

The remainder of this paper is organized as follows. In Section 2, we describe some related works. In Section 3, we explain the task description. In Section 4, we describe the participants of the task.

In Section 5, we report the evaluation results. Finally, we conclude in Section 6.

2. RELATED WORKS
The task of research paper classification into the IPC system is considered a cross-genre text classification study. Although patent classification tasks were conducted in the Fifth and Sixth NTCIR Workshops [4][5], and in Falls' study [1], they did not focus on cross-genre text classification. In other evaluation workshops, such as the Text RETrieval Conference (TREC)\(^3\) and the Cross-Language Evaluation Forum (CLEF)\(^4\), no tasks that focused on cross-genre information access have been conducted.

Although the goal of the task was not text classification, a related subtask was conducted in the Patent Retrieval Task in the Third NTCIR workshop [6]. This subtask aimed to retrieve patents relevant to a given newspaper article. In this task, Iloh et al. focused on “Term Distillation” [3]. The distribution of the frequency of the occurrence of words was considered to differ between heterogeneous databases. For example, the word "president" often appears in newspaper articles but seldom appears in patents. Therefore, unimportant words were assigned high scores when using the Term Frequency Inverse Document Frequency (TFIDF) method to weight words. Term Distillation is a technique that can prevent such cases by filtering out words that could be assigned incorrect weights.

There is another approach for cross-genre information retrieval. 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 [11]. 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.

To solve this problem, Kamaya et al. proposed a method to paraphrase scholarly terms into patent terms (e.g., paraphrase "floppy disc" into "magnetic recording medium") [7]. They used hypernym-hyponymy relations between terms for the paraphrasing [10]. Some patent terms (e.g., "magnetic recording medium") are the hyponyms of scholarly terms (e.g. "floppy disc"). Therefore, a paraphrase of some scholarly terms can be realized by finding their hypernyms. However, some scholarly terms do not have such a relationship with their corresponding patent terms. For example, a hyponym of the scholarly term "machine translation" is "natural language processing," but the terms "automatic translation" or "language translation" are used in patents instead of "natural language processing". Therefore, they also used citation relationships between research papers and patents for paraphrasing. Generally, a research paper and a patent that have citation relationships with each other, tend to be in the same research field. Using this idea, a paraphrase of a scholarly term can be realized by using the following procedure:

1. Retrieve research papers that contain a given scholarly term in their titles.
2. Collect patents that have citation relationships with the papers retrieved in Step 1.
3. Extract patent terms from patents collected in Step 2.
4. Output patent terms extracted in Step 3.

They combined the hypernym-hyponym-based and the citation-based methods, and finally obtained the patent terms as the paraphrase of the given scholarly term.

These related works are a part of solutions for cross-genre information access. The organizers of the Patent Mining Task expected participants to propose other fundamental techniques for cross-genre information access.

3. THE PATENT MINING TASK

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

The sixth edition of the IPC system contains more than 50,000 classes at the most detailed level. The goal of this task was to assign one or more of these 50,000 classes to a given topic, as expressed in terms of the title and abstract of a research paper. An example of a topic is shown in Figure 1. Here, <TOPIC-ID> specifies the topic identification number, and <TITLE> and <ABSTRACT> specify the title and abstract of the research paper to be classified.

![Figure 1. An example of a topic in the English subtask](http://trec.nist.gov/)

Within the overall task, the following subtasks were conducted.

- English subtask: classification of English research papers using patent data written in English.

In addition to these subtasks, we also conducted the following more challenging subtasks, which require both cross-genre and cross-lingual information access techniques.

\(^3\) http://trec.nist.gov/
\(^4\) http://www.clef-campaign.org/
• Cross-lingual subtask (J2E): classification of Japanese research papers using patent data written in English.


These four subtasks are summarized in Figure 2.

3.2. Patent Data

An overview of the patent data used in each subtask is shown in Table 1. In the following, we describe details of the data.

Table 1. Document Sets

<table>
<thead>
<tr>
<th>Data</th>
<th>Year</th>
<th>Size</th>
<th>Number</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Unexamined Japanese patent applications</td>
<td>1993–2002</td>
<td>100 GB</td>
<td>3.50M</td>
<td>Japanese</td>
</tr>
<tr>
<td>(2) USPTO patent data</td>
<td>1993–2000</td>
<td>33 GB</td>
<td>0.99M</td>
<td>English</td>
</tr>
<tr>
<td>(3) Patent Abstracts of Japan (translated into English)</td>
<td>1993–2002</td>
<td>4.2 GB</td>
<td>3.50M</td>
<td>English</td>
</tr>
<tr>
<td>(4) NTCIR-1 and NTCIR-2 CLIR Task test collection (Abstracts of research papers)</td>
<td>1988–1999</td>
<td>1.4 GB</td>
<td>0.26M</td>
<td>Japanese/English</td>
</tr>
</tbody>
</table>

(1) Unexamined Japanese Patent Applications

The data comprised the full texts of Japanese unexamined patents. Each patent was text only, without images or formulas. One or more IPC codes were manually assigned to each patent. These data were distributed to the teams participating in the Japanese subtask and the cross-lingual subtask (E2J).

(2) United States Patent and Trademark Office (USPTO) Patent Data

The data comprised the full texts of USPTO patents, without images or formulas. A primary IPC code was manually assigned to each patent. These data were distributed to the groups participating in the English subtask and the cross-lingual subtask (J2E).

(3) Patent Abstracts of Japan (PAJs)

The data comprised English abstracts translated from Japan Patent Information Organization ( JAPIO ) Patent Abstracts. One or more IPC codes were manually assigned to each abstract. These data were distributed to the groups participating in the English and J2E subtasks.

(4) NTCIR-1 and NTCIR-2 CLIR Task Test Collection

This database was distributed to all participant groups, and they were allowed to use it for any purposes. The database was originally used in the Cross-lingual Information Retrieval (CLIR) tasks in the first and second NTCIR Workshops (NTCIR-1 and NTCIR-2) [8][9]. It contains 255,960 records of Japanese-English paired documents, with each record comprising a title, the author(s), an abstract, keywords, a publication year, and a conference name.

3.3. Relevance Judgements

Sets of topics with manually assigned IPC codes are necessary for the evaluation. However, it is very costly and time consuming to create such data sets. Therefore, we have produced the data sets using the following idea.

Essentially, an invention is not patentable if it was already known before the date of filing. However, Article 30 in the Japanese patent law provides a six-month grace period for disclosures made via a publication or a presentation at a conference or exhibition. In this case, the applicants must mention the proceedings’ title (or the conference name) and the date it was published in an “Indication of exceptions to lack of novelty” field (or exception field) in the patent. Figure 3 gives an example of an exception field.

(English translation)

【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

We can assume that 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 codes that should be assigned to the research paper mentioned in the
exception field, it becomes possible to create a large-scale data set at low cost. In fact, there are more than 9,000 applications with exception fields in the 3,496,253 Japanese patent applications published in the 10-year period 1993–2002.

The procedure used to create the data set was as follows. Firstly, we extracted publication years and proceedings titles from the exception fields in the 9,000 applications. Although the title and authors of a paper are not mentioned in the exception field, the authors are usually the same as the inventors of the patent. We therefore extracted and used the inventors of the patent instead of the authors’ names.

Secondly, we compared these extracted data with records in a research paper database using a simple string matching method. From this automatic matching, we obtained, on average, six candidate records for each exception field.

Thirdly, we manually identified the correct match from among the candidate records. This produced 976 pairs of matching patents and research papers.

From these pairs, we created English and Japanese topics (titles and abstracts) and their correct classifications (IPC codes extracted from patents). For each topic, an average of 2.3 IPC codes was assigned.

We then randomly assigned 97 topics to the "dry run" and the remaining 879 topics to the "formal run". The dry run data were provided to the participant teams as training data for the formal run. A list of pairs of a patent ID and one or more IPC codes were also provided as additional training data. These IPC codes were extracted from each patent in the data sets (1), (2), and (3).

Participant teams were asked to submit one or more ranked lists of IPC codes for each topic, to be evaluated using Mean Average Precision (MAP), Recall, and Precision measurements. To calculate these measurements for each submitted run, the organizers produced a Perl program that was compatible with the trec_eval program. The values for MAP, Recall, and Precision can potentially be different depending on the version of trec_eval used.

4. PARTICIPANTS

We had 24 participating systems for the Japanese subtask, 20 for the English subtask, and five for the Cross-lingual subtask. As far as the number of groups is concerned, we had 12 participating groups of universities and companies. Table 2 shows the breakdown of the groups.

<table>
<thead>
<tr>
<th></th>
<th>Japan</th>
<th>Other Asian Countries</th>
<th>Europe</th>
<th>North America</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Company</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The number of runs for each subtask was as follows.

- Japanese subtask: 24 runs from five groups
- English subtask: 20 runs from nine groups
- J2E: five runs from two groups

There were no runs submitted to E2J.

5. RESULTS

5.1. Evaluation Results of the Patent Mining Task

We show the evaluation results for the Japanese, English, and Cross-lingual subtasks in Tables 3, 4, and 5, respectively. We also show the recall-precision curves for each subtask in Figures 4, 5, and 6. The systems "A13", "F1", and "G5" obtained the best scores in Japanese, English, and Cross-lingual subtasks, respectively. These systems employed the k-Nearest Neighbours (k-NN) method in common, and exceeded 0.4 of MAP scores. In addition to these, many other systems, such as C1 and C2 in the Japanese subtask and "G1", "G2", "G3", and "G4" in the English subtask, also employed the k-NN method.

On the other hand, some systems using machine learning approaches also obtained remarkable results. The "B4" system in the Japanese subtask applied logistic regression to the dry run data for tuning of the model, and obtained 39.64 of the MAP score for the formal run data. In the English subtask, "B6" obtained 34.79 of the MAP score using the hybrid model of logistic regression and Naïve Bayes.

5.2. Comparison with the Results of Patent Classification

To investigate the effectiveness of a cross-genre text classification system, we compared it with a patent classification system. For the cross-genre text classification system, we used "C1" for the Japanese subtask. The "C1" system comprises a patent retrieval engine [10] developed for the Patent Retrieval Task in NTCIR-6 [2]. The engine introduced the Vector Space Model as a retrieval model and SMART [12] for term weighting. The "C1" system obtained a list of IPC codes using the follow procedure.

1. Retrieve top 170 results using the patent retrieval engine for a given query (research paper).
2. Extract IPC codes with relevance scores for the query from each retrieved patent in step 1.
3. Rank IPC codes using the following equation.

\[
\text{Rank IPC code } i = \frac{\text{Relevance score of IPC code } i}{\text{Total relevance score}}
\]

From the results of the dry run, we found that detecting patents that are the counterparts of given research papers (topics) and using them for the classification of papers was more effective than using all the data in the list. We therefore removed the data for the 976 patents from the list and distributed the list to the participant teams before the formal run, because the approach of detecting the counterpart patents is not our intended purpose here.

The maximum number of IPC codes for a single topic is 1,000.

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

Here, X and n indicate an IPC code and the number of patents that X is assigned to within the top 170 retrieved patents, respectively.

We also used this system for patent classification. As we described in Section 3.3, each query has its counterpart in a patent database. We regarded these counterparts as queries and conducted patent classification using the "C1" system. We used the full text of patents as queries. Instead of using the top 170 retrieved patents, we used the top 20 for the calculation of each IPC code in Step 3, which was determined using the dry run data.

From these experiments, we obtained 37.06 for the MAP score, which is almost the same as the 39.13 obtained for the "C1" system, as shown in Table 3. This result indicates that the performance of "C1" as a cross-genre text classification reached almost the same level as an intra-genre text classification.

Although the patent classification system used full texts in patents, the MAP score was rather lower than that for the cross-genre text classification system, which used titles and abstracts of research papers as queries. One of the reasons for this result could be the way terms are used in claims. As we mentioned in Section 1, more abstract and creative terms are used in patents than those used in research papers, to try to widen the scope of the claims. In particular, the terms in claims are more abstract than those used in specifications. It is therefore considered that using terms in specifications may obtain a better MAP score than using whole terms in patents for patent classification. This may also indicate that using specifications instead of full texts can improve the performance of cross-genre text classification.

### Table 3. MAP for Japanese Subtask

<table>
<thead>
<tr>
<th>Run ID</th>
<th>MAP</th>
<th>Run ID</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A13</td>
<td>44.02</td>
<td>A4</td>
<td>41.65</td>
</tr>
<tr>
<td>A11</td>
<td>43.71</td>
<td>B4</td>
<td>39.64</td>
</tr>
<tr>
<td>A12</td>
<td>43.61</td>
<td>*C1</td>
<td>39.13</td>
</tr>
<tr>
<td>A7</td>
<td>43.60</td>
<td>*C2</td>
<td>39.06</td>
</tr>
<tr>
<td>A1</td>
<td>43.34</td>
<td>A14</td>
<td>38.62</td>
</tr>
<tr>
<td>A6</td>
<td>43.29</td>
<td>B3</td>
<td>35.72</td>
</tr>
<tr>
<td>A5</td>
<td>43.26</td>
<td>B2</td>
<td>34.35</td>
</tr>
<tr>
<td>A8</td>
<td>43.23</td>
<td>B1</td>
<td>33.03</td>
</tr>
<tr>
<td>A10</td>
<td>43.18</td>
<td>D1</td>
<td>27.27</td>
</tr>
<tr>
<td>A3</td>
<td>42.68</td>
<td>*C3</td>
<td>14.12</td>
</tr>
<tr>
<td>A2</td>
<td>42.36</td>
<td>E1</td>
<td>6.98</td>
</tr>
<tr>
<td>A9</td>
<td>42.27</td>
<td>E2</td>
<td>4.06</td>
</tr>
</tbody>
</table>

(C1, C2, and C3 are the task organizer's systems)

### Table 4. MAP for English Subtask

<table>
<thead>
<tr>
<th>Run ID</th>
<th>MAP</th>
<th>Run ID</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>48.86</td>
<td>H2</td>
<td>14.37</td>
</tr>
<tr>
<td>F2</td>
<td>47.21</td>
<td>I1</td>
<td>14.36</td>
</tr>
<tr>
<td>F3</td>
<td>44.53</td>
<td>H1</td>
<td>14.23</td>
</tr>
<tr>
<td>G1</td>
<td>42.45</td>
<td>I2</td>
<td>13.39</td>
</tr>
<tr>
<td>G2</td>
<td>42.09</td>
<td>J2</td>
<td>12.65</td>
</tr>
<tr>
<td>G3</td>
<td>41.83</td>
<td>K4</td>
<td>10.45</td>
</tr>
<tr>
<td>G4</td>
<td>41.49</td>
<td>J4</td>
<td>9.90</td>
</tr>
<tr>
<td>B6</td>
<td>34.79</td>
<td>K1</td>
<td>9.78</td>
</tr>
<tr>
<td>B5</td>
<td>33.74</td>
<td>J3</td>
<td>9.37</td>
</tr>
<tr>
<td>D2</td>
<td>29.03</td>
<td>L1</td>
<td>3.79</td>
</tr>
</tbody>
</table>

### Table 5. MAP for Cross-lingual Subtask (J2E)

<table>
<thead>
<tr>
<th>Run ID</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>G5</td>
<td>43.80</td>
</tr>
<tr>
<td>K5</td>
<td>10.70</td>
</tr>
<tr>
<td>K6</td>
<td>10.41</td>
</tr>
<tr>
<td>K2</td>
<td>9.41</td>
</tr>
<tr>
<td>K3</td>
<td>9.34</td>
</tr>
</tbody>
</table>

### 6. CONCLUSION

We have given an overview of the evaluation and design of the Patent Mining Task in NTCIR-7. We focused on the "Indication of exceptions to lack of novelty" field in Japanese patent applications and thereby created 976 English and Japanese topics and their correct classifications (IPC codes). Forty-nine runs from 12 participant groups were submitted to the formal run, and the systems using the k-NN method obtained the best MAP scores in each subtask. Details of each system will be reported in the NTCIR-7 final meeting to be held in December 2008. After the NTCIR-7 meeting, the test collection described in this paper will be available to the public for research purposes.

### 7. REFERENCES


Figure 4. Recall-precision Curves for All Topics (Japanese Subtask)
Figure 5. Recall-precision Curves for All Topics (English Subtask)

Figure 6. Recall-precision Curves for All Topics (Cross-lingual Subtask: J2E)