Test Collections for Patent Retrieval and Patent Classification in the Fifth NTCIR Workshop

Atsushi Fujii*, Makoto Iwayama†, Noriko Kando‡

*Institute of Library and Information Science
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

†Hitachi, Ltd.
1-280 Higashi-Kougakubo, Kokubunji, 185-8601, Japan
iwayama@crl.hitachi.co.jp
/ Tokyo Institute of Technology

‡ National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, 101-8430, Japan
kando@nii.ac.jp

Abstract

This paper describes the test collections produced for the Patent Retrieval Task in the Fifth NTCIR Workshop. We performed the invalidity search task, in which each participant group searches a patent collection for the patents that can invalidate the demand in an existing claim. For this purpose, we performed both document and passage retrieval tasks. We also performed the automatic patent classification task using the F-term classification system. The test collections will be available to the public for research purposes.

1. Introduction

In the Third NTCIR Workshop (NTCIR-3), the authors of this paper organized the Patent Retrieval Task (Iwayama et al., 2006). The process of patent retrieval differs depending on the purpose of retrieval. In NTCIR-3, the "technology survey" task was performed, in which patents were used as technical publications.

The authors also performed the Patent Retrieval Task in the Fourth NTCIR Workshop (NTCIR-4) focusing on the "invalidity search" and "patent map generation" subtasks (Fujii et al., 2004).

In NTCIR-4, a number of issues remained open questions. First, in the invalidity search subtask, the number of relevant documents per search topic was small and the evaluation result was perhaps less reliable compared with the conventional ad-hoc retrieval tasks. Second, although a subtask for passage retrieval was planned, the evaluation was not performed due to schedule problems. Third, in the patent map generation subtask, a method for quantitative evaluation was not established.

In view of the above problems, we organized the Patent Retrieval Task in the Fifth NTCIR Workshop (NTCIR-5) and performed the following three subtasks:

- Document Retrieval Subtask

  The invalidity search as in NTCIR-4 was performed, but the numbers of search topics and target documents were increased.

- Passage Retrieval Subtask

  In a document retrieved by a topic for invalidity purposes, the passages were sorted according to the relevance to the topic.

- Classification Subtask

  Classifying patent applications has promise to improve the quality of the patent map generation. Additionally, the document classification can automatically be evaluated using the patent classification system. We used a multi-dimensional classification system called "F-term (File Forming Term)".

In this paper, we describe the test collections produced during the above subtasks. Details of evaluation results for participating systems, which are described in the NTCIR-5 Proceedings (Fujii et al., 2005; Iwayama et al., 2005), are beyond the scope of this paper. In Sections 2.–4., we elaborate on the test collection for each subtask, respectively.

2. Document Retrieval Subtask

2.1. Overview

The purpose of the invalidity search is to find one or more patents that can invalidate the demand in an existing claim. This is a patent-to-patent associative retrieval task. In real world, the invalidity search is usually performed by examiners in a government patent office and searchers of the intellectual property division in private companies.

Document Retrieval Subtask was performed as follows. First, the task organizers (i.e., the authors of this paper) provided each participating group with a document collection and search topics.

Second, each group submitted retrieval results obtained by the topics. Each group was allowed to use more than one retrieval method and submit multiple retrieval results. In a single retrieval result, up to the top 1000 retrieved documents must be sorted by the relevance score. Finally, the organizers evaluated the submitted results using relevant documents. The evaluation results were sent to each group.

2.2. Document Sets

In NTCIR-4, the document collection consisted of five years of unexamined Japanese patent applications published in 1993–1997. However, in NTCIR-5, the document collection consists of ten years of unexamined Japanese patent applications published in 1993–2002. The number of documents in the collection is approximately 3.5 M.
The English patent abstracts, which are human translations of the Japanese Patent Abstracts published in 1993–2002, were also provided to train English-to-Japanese cross-language IR (CLIR) systems. We initially planned a CLIR patent retrieval subtask. Because search topics were not completed before the formal run, the CLIR subtask was not performed. However, users of our test collection can evaluate the effectiveness of their CLIR systems.

2.3. Search Topics

Each search topic is a Japanese patent application rejected by the Japanese Patent Office (JPO). For each topic, one or more citations (i.e., prior art) were identified by examiners of the JPO to invalidate the demand.
To increase the number of topics, we minimized the cost required for producing search topics and relevance judgements. We automatically extracted patent applications rejected by the JPO and the citations used for the rejection. For this purpose, we used the citation information in the "*Seirihyoujunka* (Standardized)" Data, which was extracted from the master database in the JPO. We used only the citations as relevant or partially relevant documents and did not perform relevance judgement by human assessors.
We selected 1,200 applications as topics. The process of selecting these topics is described in the NTCIR-5 Proceedings (Fujii et al., 2005).
The citation information we used did not include the information as to which claim was the target of the rejection. Thus, for each application in the pool we systematically extracted the first claim, which is usually the target.
Each search topic file includes a number of additional SGML-style tags. The claim used as the target of invalidation is specified by <CLAIM>. The date of filing is specified by <FDATE> and only the patents published before this date can potentially be relevant. Thus, target documents must be indexed by both text content and publication date. Other patent fields, such as IPC (International Patent Classification) and applicant, can also optionally be used for indexing purposes. Figure 1 shows an example topic claim translated into English.
During the formal run, we found 11 inappropriate topics. For most of these topics, the automatic method failed to extract the first claim correctly, because the layout of applications is not strictly standardized and can vary depending the applicant. The remaining 1,189 topics were used for evaluation purposes and are included in our test collection.

2.4. Evaluation Method

The relevance degree of the citation with respect to a topic was determined based on the following two ranks:

- the citation used to reject an application was regarded as a "relevant document (A)" because the decision was made confidently,

```
<TOPIC>
<NUM>1048</NUM>
<LANG>EN</LANG>
<FDATE>19950629</FDATE>
<CLAIM>A milk-derived calcium-containing composition comprising an inorganic salt mainly composed of calcium obtained by baking a milk-derived prepared matter containing milk casein-bonding calcium and/or colloidal calcium. </CLAIM>
</TOPIC>
```

Figure 1: Fragment of search topic.

- a citation used to reject an application with another citation was regarded as a "partially relevant document (B)", because each citation is partially related to the claim in the application.

We used Mean Average Precision (MAP), which has commonly been used in past IR literature, to evaluate the submitted runs for Document Retrieval Subtask.

3. Passage Retrieval Subtask

3.1. Task Description

In Document Retrieval Subtask, we performed the invalidity search task, in which the first claim in a patent application was used to search for similar patent documents. However, because patent documents are long, it is effective to indicate important fragments (i.e., passages) in a relevant document.
The purpose of Passage Retrieval Subtask was to sort all passages in a relevant document according to the degree to which a passage provides grounds to judge whether the document is relevant.
We used the 41 search topics and 378 relevant documents produced for the dry run and the formal run of the NTCIR-4 Patent Retrieval Task. We call those relevant documents "target documents". The search topics for NTCIR-4 were used to determine criteria as to how the passages in a target document should be sorted.
The passages in each target document were standardized by the official tool provided by the organizers. In Japanese patent applications, paragraphs are identified and annotated with the specific tags by applicants. Because we used these paragraphs as passages, the passage identification process was fully automated.
Figure 2 depicts a fragment of a Japanese patent application (publication number is 1997-051209) segmented into passages, in which each passage and its ID are specified by <PASSAGE> and <PNUM>, respectively.
A high rank should be given to the passages that provide sufficient grounds to judge whether a target document is relevant with respect to the search topic. In other words, using a target document as a collection consisting of multiple passages, a search topic was used to search the collection for relevant passages and sort these passages.
Of the 378 target documents, 356 documents were used for evaluation purposes and are included in our test collection. For the remaining documents, passages judged as grounds

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、光／電気通信分
野、光／電気情報処理分野において使用される高速ＩＣ
の信号を伝達する配線基板および配線基板に用いる誘電
体基板に関するものである。
</PASSAGE>
<PASSAGE>
<PNUM>PATENT-JA-UPA-1997-051209-9</PNUM>
【０００２】
【従来の技術】従来、この種の配線基板は、ほぼ一様な
比誘電率を持つガラスエポキシ、セラミック材料等の誘
電体基板に薄い導電体パターンを形成して、マイクロス
トリップ線路やコプレーナ線路として用いられている。
これらの配線基板は、単層あるいは積層構造をとるが、
厚さはほぼ一様となっている。
</PASSAGE>
<PASSAGE>
<PNUM>PATENT-JA-UPA-1997-051209-10</PNUM>
【０００３】マイクロストリップ線路の場合には、この
特性インピーダンスＺ<SB>0</SB> は誘電体基板の厚さＨと導電体
パターン幅Ｗ<SB>1</SB> および誘電体基板の比誘電率ε<SB>r</SB>
によっ
てほぼ決まる。すなわち、同一の特性インピーダンスＺ
<SB>0</SB> を有する配線基板では、誘電体基板の厚さＨが一定な
らば導電体パターン幅Ｗ<SB>1</SB> も一定となる。定性的には、
誘電体基板の厚さＨが大、導電体パターン幅Ｗ<SB>1</SB> が小、
比誘電率ε<SB>r</SB> が小になるほど特性インピーダンスＺ<SB>0</SB> は
大になる。
</PASSAGE>
<PASSAGE>

Figure 2: Fragment of Japanese patent application segmented into passages.

include figures and can not be used to evaluate text retrieval systems. The number of passages per target document is 47.

3.2. Evaluation Method

Relevant passages were determined based on the following criteria.

- If a single passage can be grounds to judge the target document as relevant or partially relevant, this passage was judged as relevant.
- If a group of passages can be grounds to judge the target document as relevant or partially relevant, this passage group was judged as relevant.

Assessors exhaustively identified all relevant passages and passage groups. During NTCIR-4, we asked 12 members of the Intellectual Property Information Search Committee in the Japan Intellectual Property Association for this task. Each member belongs to the intellectual property division in the company he or she works for, and they are all experts in patent searching.

A relevant passage group is equally informative as a single relevant passage. We introduced the concept of "combinational relevance". This concept provides a salient contrast to the conventional IR evaluation method, in which all relevant items (documents or passages) are independently important and thus combinations of relevant documents are not considered.

We calculate the evaluation score for each run as the rank at which a user obtains sufficient grounds to judge the target document as relevant or partially relevant. To obtain sufficient grounds, a user must read a relevant passage or all the passages in a relevant passage group. To calculate the final score, the ranks are averaged over all target documents. In other words, given a list of passages, we calculate the evaluation score as an expected search length at which a user satisfies their information need. We call this score "Combinational Relevance Score (CRS)".

4. Classification Subtask

4.1. Overview

We evaluated patent classification systems through a multi-dimensional classification structure called "F-term" (Schellner, 2002), which is used in the JPO.

F-term classification system has over 2,500 "themes" covering all the technological fields of patents. Patents under each theme can be classified from several viewpoints, such as purpose, function, and effect. The collection of possible viewpoints varies from theme to theme. Each viewpoint defines a set of its possible elements and a pair of a viewpoint and its element is called "F-term".

F-term classification system serves as an effective tool for narrowing down relevant patents in searching. Experts assign a patent to one or more F-terms in two steps. They determine themes of the patent, and for each theme they assign the patent to F-terms. According to this procedure, we divided our subtask into "Theme Categorization Subtask" and "F-term Categorization Subtask".

4.2. Theme Categorization Subtask

Participants were requested to submit a ranked list of 100 themes for each patent. Unlike the filtering track in TREC4, our subtask is not for binary text classification in which systems only decide for each document whether it should be accepted or rejected as a member of a category.

In a ranked list, participants were also requested to determine the threshold of their confidence on theme assignments. The themes above the threshold were regarded as the submissions with confidence, which were used for calculating the F-measure.

Training documents of this subtask are full texts of Japanese patents published from 1993 to 1997, and test documents were randomly selected from those published from 1998 to 1999. Every Japanese full text has its English abstract and participants were able to use both collections in training and testing.

Submitted results were evaluated based on recall and precision. For a ranked list for each test document, we calculated the 11 point interpolated precision, the MAP (Mean Average Precision), and the F-measure. These values were averaged over all the test documents (macro averaging).

Because most of the test documents are associated with only one or two themes (40% for one theme and 33% for two themes), interpolation of precision did not work effectively to distinguish recall/precision trade-off curves between submitted results.

For example, if a document is associated with only one theme, the precision at the recall 0.0 is always interpolated by the precision at the recall 1.0, which means that the interpolated recall/precision curve becomes a horizontal line. If a document is associated with two themes, the interpolated recall/precision curve becomes a shape of the two-step function (from 0.0 to 0.5 and from 0.5 to 1.0). In this subtask, because 73% of the test documents are associated

Table 1: The themes used in F-term Categorization Subtask.

Theme code	Theme name	#Viewpoints	#F-terms	Example viewpoint
2B022	Cultivation of vegetables	9	95	Target vegetables
3G301	Electrical control of the air and fuel supply to internal combustion	21	369	Engine models
4B064	Manufacture of chemical compounds by using	23	541	Products containing oxygen
5H180	Traffic-control systems	11	215	Means of detection
5J104	Ciphering device, decoding device and privacy communication	14	271	Purpose and effect

with one or two themes, a shape of the macro averaged recall/precision curve over the test documents is similar for every submitted result.
To resolve this problem, we additionally calculated the micro averaged precisions as follows. Assume that there are N test documents. We first collect K top-ranked categories for every test document and pool N×K categories. We then calculate the recall and the precision for this pool. For all values of K, we calculate the corresponding recall and precision values, which are used to interpolate the precisions at the 11 levels of recall.

4.3. F-term Categorization Subtask

Participants were requested to submit a ranked list of 200 possible F-terms for each patent whose theme had been given. Participants were also requested to determine the threshold of their confidence on F-term assignments. We used the five themes listed in Table 1. Although the total number of possible F-terms across all the themes reaches to 337,027, the number of F-terms within each theme is small. In this subtask, the numbers of possible F-terms for the five themes are between 95 and 541.
Training documents are full texts of Japanese patents published from 1993 to 1997 and test documents were randomly selected from those published from 1998 to 1999. English abstracts were allowed to use in training and testing. Evaluation measures are almost the same as those in Theme Categorization Subtask. The only difference is that we did not need to calculate the micro averaged precisions. Interpolation of precision works effectively because the average number of F-terms per test document is 11.4.

4.4. Datasets

Unexamined Japanese patent applications published from 1993 to 2002, which are full texts of Japanese patents written in Japanese, were released. The same years' English abstracts (the Patent Abstracts of Japan) were also released. That is, every Japanese full text has its corresponding English abstract.
At the same time, descriptions of themes and F-terms were released. This collection is called PMGS (Patent Map Guidance System)[1], which is in both Japanese and English. For every patent published from 1993 to 1997, we released the lists of correct themes and correct F-terms as training data. Those themes and F-terms were taken from *Seirihyoujunka* Data, which contains bibliographic information of patents in the SGML format. Although a number of full texts include sections for their themes and F-terms, these themes and F-terms may not be the latest ones. In many times, themes and F-terms are added or deleted after publishing the texts and these revisions are reflected only on the databases in the JPO.
In Theme Categorization Subtask, we randomly selected 2,008 patents from all the patents published from 1998 to 1999 as test data. In F-term Categorization Subtask, we firstly selected five themes which have enough numbers of patents in every year and whose collections of viewpoints are typical ones. The five themes are listed in Table 1. For each theme, we randomly selected about 500 patents from the patents having the theme and published from 1998 to 1999 as test data.

5. Conclusion

This paper described test collections for evaluating retrieval and classification systems targeting patents, which will be available to the public for research purposes[2]. Users of our test collections can evaluate their systems even if they did not participate in the Fifth NTCIR Workshop.

6. References

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2004. Test collections for patent-to-patent retrieval and patent map generation in NTCIR-4 workshop. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1643–1646.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2005. Overview of patent retrieval task at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pages 269–277.

Makoto Iwayama, Atsushi Fujii, and Noriko Kando. 2005. Overview of classification subtask at NTCIR-5 patent retrieval task. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pages 278–286.

Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. 2006. Evaluating patent retrieval in the third NTCIR workshop. *Information Processing & Management*, 42(1):207–221.

Irene Schellner. 2002. Japanese file index classification and f-terms. *World Patent Information*, 24:197–201.

[1] http://www5.ipdl.ncipi.go.jp/pmgs1/pmgs1/pmgs_E

[2] http://www.slis.tsukuba.ac.jp/~fujii/ntcir5/cfp-en.html