Integrating Content and Citation Information for the NTCIR-6 Patent Retrieval Task

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

Abstract

This paper describes our system participated in the Japanese and English Retrieval Subtasks at the NTCIR-6 Patent Retrieval Task. The purpose of these subtasks is the invalidity search, in which a patent application including a target claim is used to search documents that can invalidate the demand in the claim. Although we use a regular text-based retrieval method for the Japanese Retrieval Subtask, we combine text and citation information to improve the retrieval accuracy for the English Retrieval Subtask.

Keywords: *Patent retrieval, Invalidity search, Citation Analysis*

1 Introduction

In the Patent Retrieval Task at the Sixth NTCIR Workshop, three subtasks were performed; Japanese Retrieval Subtask, English Retrieval Subtask, and Classification Subtask [5]. We participated in the Japanese and English Retrieval Subtasks, both of which are intended for invalidity search. This paper describes our retrieval system and its evaluation results in those tasks.

The purpose of the invalidity search is to find the patents that can invalidate the demand in an existing claim. This is an associative patent (patent-to-patent) retrieval task, because the patent application including a target claim is used as a search topic, instead of short keywords and phrases.

For the Japanese Retrieval Subtask, we used the same system for the NTCIR-5 Patent Retrieval Task [3]. For the English Retrieval Subtask, we propose a new method that combines text and citation information.

Section 2 outlines our patent retrieval system. Sections 3 and 4 describe retrieval methods for retrieving Japanese and English patents, respectively. Section 5 describes the evaluation results for our system.

2 System Description

Figure 1 depicts the overall design of our patent retrieval system, which consists of seven modules; component analysis, translation, term extraction, query expansion, document retrieval, integration, and passage retrieval. Although the design in Figure 2 is the same as the system participated in the NTCIR-5 Patent Retrieval Task [3], we enhanced the document retrieval module for English patents.

This system performs monolingual and cross-lingual or multi-lingual retrieval. Although the basis of our method is language-independent, the current system uses a patent application in Japanese to search for documents in Japanese and English.

This system performs monolingual and cross-lingual or multi-lingual retrieval. Although the basis of our method is language-independent, the current system uses a patent application in Japanese to search for documents in Japanese and English.

Given a patent application, in which a target claim is specified, our system retrieves the relevant documents in the following steps:

(1) the component analysis module performs the local structure analysis and segments the target claim into components,

(2) in cross-lingual retrieval, the translation module machine translates the claim into English on a component-by-component basis, for which the patent classification codes associated with the input application are used to select the translation dictionaries,

(3) the term extraction module selects query terms in the claim on a component-by-component basis,

(4) the query expansion module extracts additional query terms from the description field related to the claim by the global structure analysis and performs pseudo-relevance feedback,

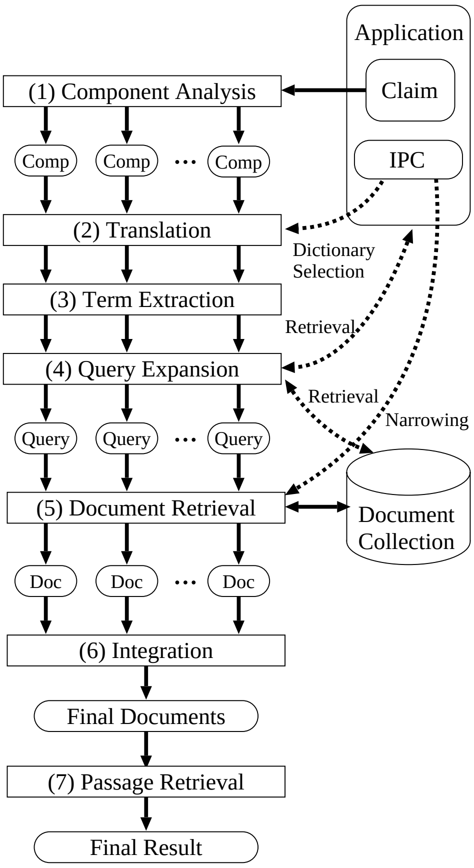

Figure 1. Overview of our patent retrieval system.

(5) the document retrieval module searches a document collection for candidates of relevant documents and produces a document list on a component-by-component basis,

(6) the integration module combines the document lists for each component and re-ranks the documents according to a new relevance score.

Here, (1), (4), and (6) were introduced for patent structure analysis purposes [2, 3]. While the obligatory modules are (3) and (5), any of the remaining modules can be omitted depending on the application.

3 Japanese Retrieval

In the Japanese Retrieval Subtask, mandatory runs must be generated using only the CLAIM and FDATE fields in each topic patent. Thus, we used only steps (3) and (5) in Section 2.

In step (3), we use the ChaSen morphological analyzer[1] to extract nouns. However, nouns in a predefined stopword list are discarded. For topics translated into English, morphological analysis is not performed and we simply discard words in the stopword list. In either language, all remaining words are collected in an unordered list and are used as an initial query.

In step (5), we use Okapi BM25 [7] to compute the relevance score between a query and each document in a collection. To invalidate an invention in a topic patent, relevant documents must be the “prior art”, which had been open to the public before the topic patent was filed. Thus, the date of filing is used to constrain the retrieved documents and only the documents published before the topic was filed can potentially be relevant.

In addition, we use the pseudo-relevance feedback to enhance the query, which enhances a query with two-stage retrieval. In practice, from the top ten documents retrieved in the first stage, the top ten terms are extracted and used in the query for the second stage. Here, the score of each term is determined according to a variant of the TF.IDF term weight.

4 English Retrieval

For the English Retrieval Subtask, we used only steps (3) and (5) in Section 2. However, we enhanced step (5). The target documents in the English Retrieval Subtask are USPTO patents, in which citation information is well-organized compared with patent applications used for the Japanese Retrieval Subtask. Thus, we explored the use of citation information for patent retrieval purposes.

Traditional research in citation analysis can be used in different applications for patents [6]. For example, if a patent is cited by a large number of other patents, this cited patent is possibly a foundation of those citing patents and is, therefore, important.

This idea is similar to identifying authoritative pages by analyzing hyperlink structures on the World Wide Web. For example, Yang [8] combined text-based and link-based methods in the Web retrieval.

Following the above ideas, we combine text and citation information in the invalidity patent search.

For the text-based retrieval, we use the claim(s) in each document to perform word-based indexing. We use Okapi BM25 to compute the text-based score for each document with respect to a query.

For the citation-based retrieval, we use two alternative methods. In either method, we first perform the text-based retrieval and obtain top N documents. We then compute the citation-based score for each of the N documents. Finally, we combine the text-based and citation-based scores and resort the N documents. We compute the final score for document d, $S(d)$, by Equation (1).

$$S(d) = S_T(d) \times S_C(d)^{\alpha} \quad (1)$$

[1] http://chasen.aist-nara.ac.jp/

$S_T(d)$ and $S_C(d)$ denote the text-based and citation-based scores for d, respectively. α is a parametric constant to control the effects of S_C.

As a citation-based method, we use PageRank [1], which estimates the probability that a user surfing on the Web visits a document. We use this probability as the citation-based score for each document. Given a document collection, the value of PageRank for each document is a constant and is independent of the topic.

As an alternative citation-based method, we propose a topic-sensitive method. We use only citations among the top N documents. We currently set $N = 1000$. As in PageRank, the citation-based score of document d is determined by the total votes by other documents. If d is cited by a large number of documents, a high score is given to d. However, if a document cites n documents, the vote for each cited document is $\frac{1}{n}$. We compute $S_C(d)$ by Equation (2).

$$S_C(d) = \sum_{x \in D_{* \rightarrow d}} \frac{1}{|D_{x \rightarrow *}|} \qquad (2)$$

$D_{* \rightarrow d}$ and $D_{d \rightarrow *}$ denote a set of documents citing d and a set of documents cited by d, respectively.

5 Experiments

5.1 Evaluating Japanese Retrieval

For the Japanese Retrieval Subtask, we submitted a single run (AFLAB1), which used only the CLAIM and FDATE fields for query construction purposes. Table 1 shows Mean Average Precision (MAP) for AFLAB1 in different conditions.

In the first column of Table 1, “Def0” and “Def1” denote the different definitions for relevance judgement and “A” and “B” denote the relevance degrees. The column “SR” denote the MAP for the Search Report. In NTCIR-6, there is no “A” relevance documents for Def0. Please see the overview paper [5] for details of the relevance judgement and relevance degree.

Details of the evaluation for each module in our system are described in our papers for NTCIR-4 [2] and NTCIR-5 [3].

Table 1. MAP for Japanese Retrieval.

	NTCIR-4	NTCIR-5	SR	NTCIR-6	Total
Def0 A	21.37	19.16	13.50	N/A	18.38
Def0 B	16.15	15.39	13.27	8.21	11.46
Def1 A	12.59	13.24	11.77	6.69	9.25
Def1 B	16.15	15.39	13.27	8.21	11.46

5.2 Evaluating English Retrieval

For the English Retrieval Subtask, we submitted the following three runs.

- AFLAB1: text-based retrieval
- AFLAB2: text-based retrieval + topic-sensitive citation-based method
- AFLAB3: text-based retrieval + PageRank

While AFLAB1 is mandatory, AFLAB2 and AFLAB3, which used citation information, are optional.

We determined the optimal value of α in Equation (1) through preliminary experiments. The values of α were 0.1 and 0.01 for AFLAB2 and AFLAB3, respectively.

Table 2 shows MAP for different runs. Looking at Table 2, AFLAB2 and AFLAB3 were more effective than AFLAB1, irrespective of the relevance degree. However, AFLAB2 was more effective than AFLAB3, irrespective of the relevance degree.

In summary, a combination of the text-based and citation-based methods improved the effective for the invalidity patent search. The improvement was even greater when we used the topic-sensitive citation-based method.

Table 2. MAP for English Retrieval.

	AFLAB1	AFLAB2	AFLAB3
A	3.65	4.17	3.81
B	7.12	8.11	7.48

6 Summary

We participated in the NTCIR-6 Patent Retrieval Task and evaluated our system in the Japanese and English retrieval Subtasks. Although we have not found any knowledge for retrieving Japanese patent applications, we demonstrated that a combination of text and citation information improved the retrieval accuracy for USPTO patents.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.

[2] A. Fujii and T. Ishikawa. Document structure analysis in associative patent retrieval. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[3] A. Fujii and T. Ishikawa. Document structure analysis for the NTCIR-5 patent retrieval task. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 292–296, 2005.

[4] A. Fujii, M. Iwayama, and N. Kando. Overview of patent retrieval task at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 269–277, 2005.

[5] A. Fujii, M. Iwayama, and N. Kando. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting*, 2007.

[6] M. M. S. Karki. Patent citation analysis: A policy analysis tool. *World Patent Information*, pages 269–272, 1997.

[7] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

[8] K. Yang. Combining text- and link-based retrieval methods for Web IR. In *Proceedings of the 10th Text REtrieval Conference*, pages 609–618, 2001.