Document Structure Analysis for the NTCIR-5 Patent Retrieval Task

Atsushi Fujii and Tetsuya Ishikawa
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 Document and Passage Retrieval Subtasks at the NTCIR-5 Patent Retrieval Task. The purpose of these subtasks was 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. Our system is characterized by the structure analysis for both target claim and entire application. The target claim is segmented into components, each of which is used to produce an initial query. The structure of the application is used to enhance each query. The candidates of relevant documents are retrieved and ranked on a component-by-component basis. The final document list is obtained by integrating these document lists. All passages in each document are ranked according to the relevance to the target claim. We show the effectiveness of our system experimentally.

Keywords: *Patent retrieval, Invalidity search, Document structure analysis, Passage retrieval*

1 Introduction

In the Patent Retrieval Task at the Fifth NTCIR Workshop, three subtasks were performed; Document Retrieval Subtask, Passage Retrieval Subtask, and Classification Subtask [2, 3]. We participated in the Document and Passage 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.

The conventional method for query processing extracts index terms from a search topic and formulates an unordered list of terms as a query.

However, because a search topic is a patent application, which is structured from a number of perspectives, a different approach is desired in the invalidity search. We introduce two structure analysis methods in a patent retrieval system.

First, because a claim often consists of multiple components (e.g., parts of a machine and substances of a chemical compound), relevance judgment is performed on a component-by-component basis in real world. The prior arts associated with all or most of the components have promise for invalidating the demand in the target claim. To automatize this process, we analyze the structure of a claim and segment the claim into components.

Second, while a claim includes general words and vague descriptions, a different field in the same application, which is usually termed “detailed description”, elaborates on the same content in detail. To utilize effective and concrete index terms in retrieval purposes, the description fields that associate with the target claim must be identified. For this purpose, a structure analysis for the entire application is required.

The above first and second methods correspond to local and global analyses for a patent application, respectively. These analyses have manually been performed by examiners in a government patent office and searchers of the intellectual property division in private companies.

2 System Description

2.1 Overview

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. We used the same system to participate in the NTCIR-4 Patent Retrieval Task [1].

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,

(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,

(7) the passage retrieval module sorts the passages in each document, for which the official tool was used to standardize the passages in the document collection.

Here, (1), (4), and (6) were introduced for patent structure analysis purposes [1]. While the obligatory modules are (3), (5), and (7), any of the remaining modules can be omitted depending on the application. In the following sections, we elaborate on each module, respectively.

2.2 Component Analysis

We analyze the structure of a claim and segments the claim into components. However, because claims are written with the patent-specific sub-language and description styles, we use Japanese punctuation (i.e., comma and period) as a delimiter to segment a claim into components, because applicants often indicate the components with punctuation.

2.3 Translation

We use PAT-Transer/je[1], which is a machine translation (MT) system for patents, to translate Japanese claims into English. Out of 22 domain dictionaries (e.g., chemistry and mechanics), the MT engine can use up to 10 dictionaries simultaneously. Because the translation quality is dependent on the dictionary used, we select the domain dictionaries based on the classification codes assigned to the input application.

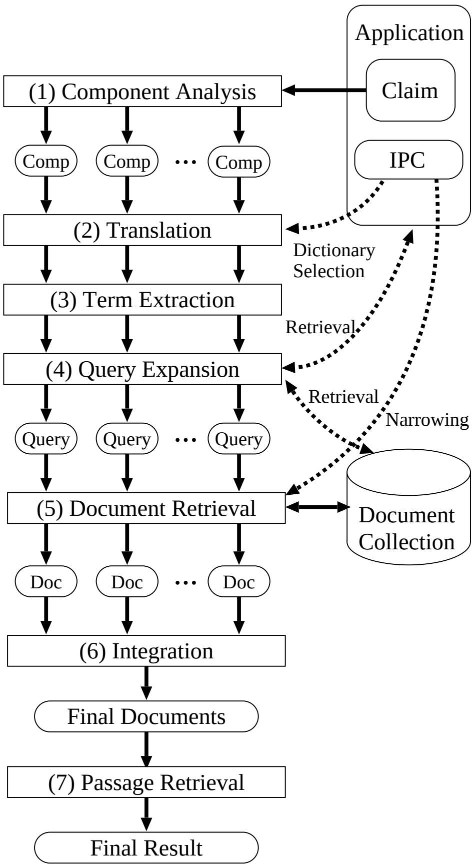

Figure 1. Overview of our patent retrieval system.

We use the subclasses (i.e., the top three codes) in the International Patent Classification (IPC) system, such as, G01R, H01L, and B27N. We manually corresponded the IPC subclasses and the domain dictionaries.

2.4 Term Extraction

For Japanese claims, we use the ChaSen morphological analyzer[2] 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.

[1] http://www.crosslanguage.co.jp

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

2.5 Query Expansion

We use two methods for query expansion purposes.

First, we search the input application for the fragments that describe the same or similar content in a claim component, because general words in the component are usually expressed by concrete or specific words in those fragments.

In practice, we regard all paragraphs, which are determined by the official tool, in the input application as independent items and index them as performed in the document retrieval. Thus, the corresponding paragraphs can efficiently be retrieved in response to an initial query produced in Section 2.4. For this purpose, we use the same retrieval module in Section 2.6. Consequently, for general words, such as "moving objects", we can be add more concrete words, such as "vehicles" and "trains", in the query.

Second, we use the conventional pseudo-relevance feedback (PFR) to further 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.

Note that while PRF is an inter-document expansion method, the above-mentioned first method is an intra-document expansion method, which can be combined with the first stage in PRF.

It should also be noted that because the effectiveness of PRF is dependent of the accuracy of the first stage retrieval, a combination of the intra- and inter-document expansion methods has promise for improving the entire accuracy of our retrieval system.

Konishi et al. [4] also performed query expansion using the "detailed description" field. While their method relies on a number of hand-crafted rules for structure analysis purposes, our method uses the conventional retrieval method and thus is more robust. However, quantitative comparisons are needed to determine the relative superiority between these methods.

2.6 Document Retrieval

For the document retrieval module, we use Okapi BM25 [5] to compute the relevance score between a (translated) query and each document in a collection. In addition, non-textual constraints, such as the IPC code and date, can be used to reduce the number of retrieved documents.

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.

The document retrieval module is the same as the baseline system provided by the organizers of the Patent Retrieval Task, which uses ChaSen to extract words as index terms and also uses character bigrams as index terms [2].

2.7 Integration

When we perform document retrieval and produce document lists on a component-by-component basis, a number of documents are included in more than one list. Thus, the retrieval documents can be organized in a two-dimension matrix as Figure 2, in which the x/y-axes correspond to the retrieved documents and components, respectively. The values in each cell are the relevance scores determined by the document retrieval module in Section 2.6.

Intuitively, document A, which was retrieved for a large number of components with high scores, can potentially be relevant. Although document B was retrieved for component #1 with a higher score than that for document A, document B has little association with the other components and thus can possibly be irrelevant.

In principle, the final score of a document is computed as a weighted average of the score for each component. However, because currently we do not have a method to determine the weight of a component, we experimentally use the average of the score for each component as the final score. In the final document list, the documents are re-sorted according to the new score.

The component analysis is effective for interactive retrieval purposes, because given a matrix like Figure 2, a user can grasp which document is retrieved by which component. In addition, if an interface allows users to modify the weight of each component manually, the final results can be changed depending on the user's perspectives.

Note that if we do not perform the query expansion in Section 2.5, the final document list does not change whether we use each component as an independent query or we use the entire claim as a single query, because the document retrieval module in Section 2.6 uses each query term independently. In other words, the component analysis does not affect the final result.

However, when combined with the query expansion methods, the additional query terms can be different depending on the component analysis and consequently the final result can be different.

2.8 Passage Retrieval

The passage retrieval module sorts all the passages in a retrieved document. We regard all paragraphs in

Component		Candidate docs		
ID	Text	A	B	C
1	映像を処理してパソコン画面上に動画像を表示させるパソコン用動画像処理装置において、	400	600	200
2	映像入力チャンネルからの NTSC 信号を色相別デジタル輝度信号 . . . NTSC 信号変換部と、	100	0	100
. . .	. . .	. . .	. . .	. . .
8	ことを特徴とするパソコン用動画像処理装置。	300	0	50

Figure 2. Example matrix of components and candidate documents.

a document as independent items and index them as performed in the document retrieval module. Once all items are indexed, the retrieval process is the same as in Sections 2.2–2.7. However, the IPC code and date are not used to reduce the number of passages retrieved. Additionally, we do not use character bigrams for index purposes.

3 Evaluation

3.1 Evaluating Document Retrieval

For the formal run of the Document Retrieval Subtask, we submitted six results using the 1223 Japanese topics. We evaluated the effectiveness of the following methods:

- LSA: component analysis (the local structure analysis)
- GSA: intra-document expansion (the global structure analysis)
- CBI: character bigram indexing
- PRF: pseudo-relevance feedback
- IPC: International Patent Classification

For each method, we compared the cases of “used” and “not used”. For method C, in case of “used” character bigrams were used as index terms in addition to word index terms. As explained in Section 2.7, the use of LSA has no effects on the result when GSA is not used.

In the Document Retrieval Subtask, the evaluation was performed using alternative conditions for “topics” and “relevance level” as follows.

- topics:
  - the search topics for the NTCIR-4 main task were used.
  - new topics for NTCIR-5 were used.
- relevance level:
  - only relevant documents were used as correct answers (A).
  - both relevant and partially relevant documents were used as correct answers (AB).

Table 1 shows Mean Average Precision (MAP) of our submissions, in which “$\surd$” denotes the corresponding method was used. In Table 1, “NTC-X-Y” denotes the case where the topics for NTCIR-X, which corresponds to NTCIR-4 or NTCIR-5, are used and the relevance level is Y, which corresponds to A or AB. For each of the four cases, the best MAP was written in boldface.

For NTC-5-A and NTC-5-AB, IFLAB5 outperformed the other method in MAP. However, for NTC-4-A and NTC-4-AB, IFLAB1 and IFLAB3 outperformed the other method in MAP, respectively. This is mainly due to the method of producing the search topics and relevance judgement. In NTCIR-4, relevance judgement was performed by professional searchers to increase the number of relevance documents. However, in NTCIR-5 relevance judgement was not performed and only the citations that were used by examiners of the Japanese Patent Office to reject the topic patent were used as relevant or partially relevant documents.

3.2 Evaluating Passage Retrieval

In the Passage Retrieval Subtask, search topics and the relevant documents for each topic for NTCIR-4 Patent Retrieval Task were used. We call these 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.

A high rank should be given to the passages that provide sufficient grounds to judge if a document in question is relevant with respect to a search topic. In other words, using a target document as a collection consisting of multiple passages, a search topic is used to search the collection for relevant passages and sort these passages. In this subtask, retrieval results were submitted on a document-by-document basis, instead of on a topic-by-topic basis. The evaluation was performed using alternative conditions for “topics” and “relevant passages” as follows.

- topics: the relevance level of a target document with respect to the NTCIR-4 search topic.
  - relevant (A)
  - partially relevant (B)

Table 1. Evaluation results for Document Retrieval.

RunID	LSA	GSA	CBI	PRF	IPC	NTC-4-A	NTC-4-AB	NTC-5-A	NTC-5-AB
IFLAB1	√	√	√	√		**0.2137**	0.1615	0.1916	0.1539
IFLAB2	√	√	√	√	√	0.2084	0.1665	0.2075	0.1654
IFLAB3	√	√		√	√	0.1986	**0.1823**	0.1850	0.1515
IFLAB4	√	√	√		√	0.2115	0.1666	0.2018	0.1614
IFLAB5		√	√	√	√	0.1983	0.1717	**0.2107**	**0.1684**
IFLAB6			√	√	√	0.2084	0.1665	0.2075	0.1654

Table 2. Evaluation results for Passage Retrieval.

RunID	LSA	GSA	PRF	A/A	A/B	B/A	B/B	Rank
IFLAB1	√	√	√	0.4499	0.4297	0.4310	0.4239	12.01
IFLAB3	√	√		0.4438	0.4205	0.4282	0.4189	12.12
IFLAB4		√	√	0.4747	0.4520	0.4850	0.4614	**10.91**
IFLAB5			√	**0.5072**	**0.4713**	**0.4891**	**0.4636**	11.23
BASE				0.3361	0.3451	0.3700	0.3717	16.23

- relevant passages: the relevance level of passages in the target document.
  - single passage is sufficient grounds (A).
  - a group of passages is sufficient grounds (B).

Table 2, which uses the same notation as Table 1, shows the evaluation results for the Passage Retrieval Subtask. However, we did not used the character bi-gram indexing and IPC. Although we also submitted IFLAB2, this submission was mistakenly the same as IFLAB1. Thus, we do not show the result of IFLAB2 in Table 2.

In Table 2, “X/Y” denotes the MAP corresponding to the case where the target document is judged at X, which corresponds to A or B, and the grounds is a single passage or a group of passages (Y corresponds to either A or B). The column “Rank” denotes the result of precision-oriented evaluation, which is the average rank at which a user obtains sufficient grounds to determine whether the target document is relevant with respect to the NTCIR-4 search topic. Method “BASE” is a control in which all passages in the target document is sorted according to the passage ID.

Looking at Table 2, all of our method outperformed BASE in both MAP and Rank significantly. While IFLAB5 generally outperformed the other methods in MAP, IFLAB4 outperformed the other methods in Rank. Thus, while PRF was effective to improve the MAP, a combination of GSA and PRF was effective to improve the precision-oriented evaluation measure.

4 Summary

We participated in the NTCIR-5 Patent Retrieval Task and evaluated our system in the Document and Passage retrieval Subtasks. Our method can be applied to general associative retrieval tasks, in which an input document is long and consists of multiple components. However, the effectiveness of our method in different document genres remains an open question and needs to be explored.

References

[1] 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.

[2] A. Fujii, M. Iwayama, and N. Kando. Overview of patent retrieval task at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop*, 2005.

[3] M. Iwayama, A. Fujii, and N. Kando. Overview of classification subtask at NTCIR-5 patent retrieval task. In *Proceedings of the Fifth NTCIR Workshop*, 2005.

[4] K. Konishi, A. Kitauchi, and T. Takaki. Invalidity patent search system of NTT DATA. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[5] 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.