Document Structure Analysis in Associative Patent Retrieval

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 retrieval system participated in the Patent Retrieval Task at the Fourth NTCIR Workshop. The main task was an associative patent retrieval task, in which a patent application including a target claim is used to search documents that can invalidate the demand in the claim. Our system can be characterized by the structure analysis for both target claim and entire application. Based on the rhetorical structure, a claim is segmented into multiple components, each of which is used to produce an initial query. The structure of an 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 documents lists. All passages in each document are ranked according to the relevance to the target claim. We show the effectiveness of our system by means of the formal run evaluation.

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

1 Introduction

In the Patent Retrieval Task at the Fourth NTCIR Workshop, the invalidity search task and the patent map generation task were performed [1]. We participated in the invalidity search task. This paper describes our retrieval system and its evaluation.

The purpose of 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 usually 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 case. Intuitively, the prior arts associated with all (or most of) components have promise for invalidating the demand in the target claim. To automatize this process, we analyze a rhetorical structure of a claim and segment the claim into multiple 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 a searching process, the description fields that associate with the target claim must be identified. For this purpose, a structure analysis for the entire application is required.

In summary, the above-mentioned 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. Our research is a step towards the automatic query processing for the invalidity patent search.

2 System Description

2.1 Overview

Figure 1 depicts the overall design of our patent retrieval system, which consists of seven modules, i.e., component analysis, translation, term extraction, query expansion, document retrieval, integration, and passage retrieval modules.

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 is because our method for the local claim analysis is implemented for the Japanese rhetorical structure.

Although the official document collection consists of Japanese patent applications, we also used Japanese and English parallel patent abstracts, which were provided for the NTCIR-3 Patent Retrieval Task [2], for cross-lingual retrieval purposes.

Given a patent application, in which a target claim is specified, the system retrieves the relevant documents as follows:

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

(2) in the case of 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 also performs the conventional pseudo-relevance feedback,

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

(6) the integration module merges 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) are newly introduced for the patent structure analyses. 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 perform a text analysis on a claim, from which multiple components are derived. However, because claims are written with the patent-specific sublanguage and description styles, we use the following two alternative methods, instead of the conventional natural language processing (NLP):

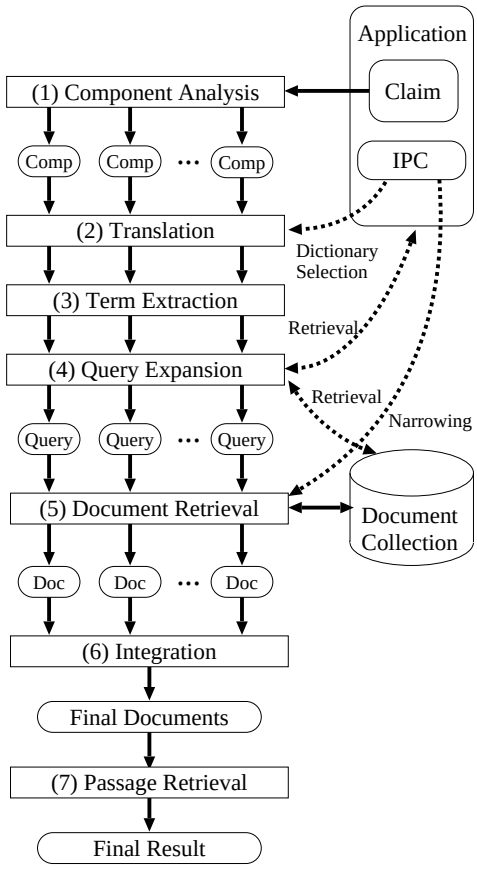

Figure 1. Overview of our patent retrieval system.

- Japanese punctuation (i.e., comma and period) is used as a delimiter to segment a claim into components, because applicants often indicate the components with punctuation,

- and the claim analysis tool proposed by Shinmori et al. [4], which was originally intended to analyze the rhetorical structure (RST), is used.

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.

In practice, we use the subclasses (i.e., the top three codes) in the International Patent Classification (IPC) system, such as, G01R, H01L, and B27N. For this pur-

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

pose, we manually corresponded the IPC subclasses and the domain dictionaries.

2.4 Term Extraction

Although the conventional NLP techniques are not always effective for analyzing patent claims, morphological analysis is a feasible choice to extract content words from each component. 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 used as an initial query.

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 (determined by the official tool) in the input application as independent items and index them as performed in the conventional 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.

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

2.6 Document Retrieval

The document retrieval module is based on an existing probabilistic method [3], which computes 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 fundamentally the same as the baseline system provided for the participants in the Patent Retrieval Task. However, while the baseline system uses only the content words extracted by ChaSen as index terms, we also use character bigrams as index terms for the Japanese documents.

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 numbers 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 other components and thus can potentially 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 users' perspectives.

It should be noted that if we do not perform the query expansion in Section 2.5, the final document list

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.

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 considers each query term independent. 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 passages in a retrieved document. We regard all paragraphs (determined by the official tool) in a document as independent items and index them as performed in the conventional document retrieval. Once all items are indexed, the retrieval process is fundamentally 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.

3 Evaluation

For the formal run, we submitted ten results obtained with the Japanese topics. Because our method is implemented for Japanese topics and thus we did not submit the results obtained with the English topics.

We evaluated the effectiveness of the following optional methods:

- A: component analysis (the local structure analysis)
- B: intra-document expansion (the global structure analysis)
- C: character bigram index terms
- D: pseudo-relevance feedback (PRF)
- E: International Patent Classification (IPC)

For method A, we had three choices, i.e., “not used”, “punctuation was used”, and “the RST tool was used”. However, the remaining four options we simply compared the cases of “used” and “not used”. For method C, character bigrams were used as index terms in addition to word index terms.

Table 1 shows the mean average precision (MAP) values averaged over the 101 search topics, for different methods. The column of “Rigid” denotes the case in which the documents judged A were regarded as the correct answers and the column of “Relaxed” denotes the case in which the documents judged B were also regarded as the correct answers. In this table, #1–#10 correspond to the official results in the formal run.

Looking at Table 1, the best MAP values were obtained when we performed the component analysis relying solely on the Japanese punctuation and used all options but the IPC (#6), for both Rigid and Relaxed.

By comparing #1 and #3, one can see that a combination of the component analysis and intra-document expansion improved the MAP values, for both Rigid and Relaxed. By comparing #1 with #4 and #5, the effectiveness of the character bigrams and PRF was observed, respectively.

In addition, the use of the Japanese punctuation was more effective than the RST tool in our experiments. However, note that the RST tool was not primarily developed for the invalidity search task.

By comparing #1 and #6, the use of IPC decreased the MAP value. Thus, we performed additional experiments in which the IPC was not used. Those results are #11–#14 in Table 1.

By comparing #6 with #11–#14, each method (i.e., A–D) was effective to improve the MAP values.

At the same time, the differences among the different methods in MAP were generally marginal. Detailed qualitative analyses are needed.

4 Conclusion

For the Patent Retrieval Task at NTCIR-4, an associative patent retrieval task was performed, in which a patent application including a target claim is used to search documents that can invalidate the demand in the claim.

For this purpose, we proposed a patent retrieval system, which can be characterized by the structure analysis for both target claim and entire application. Based on the rhetorical structure, a claim is segmented into multiple components, each of which is used to produce an initial query. The structure of an application is used to enhance each query. The candidates of relevant documents are retrieved and ranked on a component-by-

Table 1. MAP values for different methods (#1–#10: the results in the formal run, #11–#14: results obtained after the formal run).

	A	B	C	D	E	Rigid	Relaxed
#1	1	1	1	1	1	0.1383	0.1297
#2	0	1	1	1	1	0.1308	0.1233
#3	0	0	1	1	1	0.1370	0.1263
#4	1	1	0	1	1	0.1137	0.1087
#5	1	1	1	0	1	0.1361	0.1255
#6	1	1	1	1	0	**0.1464**	**0.1343**
#7	2	1	1	1	1	0.1320	0.1310
#8	2	1	0	1	1	0.1110	0.1090
#9	2	1	1	0	1	0.1310	0.1290
#10	2	1	1	1	0	0.1370	0.1310
#11	0	1	1	1	0	0.1405	0.1301
#12	0	0	1	1	0	0.1460	0.1323
#13	1	1	0	1	0	0.1078	0.1032
#14	1	1	1	0	0	0.1446	0.1296

A: component analysis (0: not used, 1: punc, 2: RST)
B: intra-document expansion (0: not used, 1: used)
C: character bigram index terms (0: not used, 1: used)
D: PRF (0: not used, 1: used)
E: IPC (0: not used, 1: used)

component basis. The final document list is obtained by integrating these documents lists. In addition, all passages in each document are ranked according to the relevance to the target claim. We showed the effectiveness of our system by means of the formal run evaluation.

Our method can potentially be applied to general associative retrieval tasks, in which an input document is long and thus consists of multiple components or subjects. However, the effectiveness of our method in different document genres remains an open question and needs to be explored.

Acknowledgments

The authors would like to thank Yuka Yamada for her support with experiments. They would also like to thank Akihiro Shinmori for his support with the RST tool.

References

[1] A. Fujii, M. Iwayama, and N. Kando. Overview of patent retrieval task at NTCIR-4. In *Working Notes of the Fourth NTCIR Workshop*, 2004.

[2] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. An empirical study on retrieval models for different document genres: Patents and newspaper articles. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 251–258, 2003.

[3] 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.

[4] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. Patent claim processing for readability: Structure analysis and term explanation. In *Proceedings of the ACL-03 Workshop on Patent Corpus Processing*, pages 56–65, 2003.