Query Terms Extraction from Patent Document for Invalidity Search

Kazuya Konishi
Research and Development Headquarters, NTT Data Corporation
Kayabacho Tower Bldg., 1-21-2 Shinkawa, Chuo-ku, Tokyo 104-0033, Japan
konishikzy@nttdata.co.jp

Abstract

This paper describes our patent retrieval system participated in the NTCIR-5 Patent Retrieval Task, Document Retrieval Subtask. The main scope of our method is the appropriate query expansion to improve recall. We extracted query terms from the topic claim, and expanded query terms extracted from sentences explained in the patent document including the topic claim. The explanation sentences were extracted by the method based on pattern matching and by the method based on the longest common subsequence length.

Keywords: *Query term extraction, Pattern matching, Longest common subsequence.*

1. Introduction

The NTCIR-5 Patent Retrieval Task, Document Retrieval Subtask is an invalidity search. In this subtask, the topic claim is the first claim in Japanese patent applications rejected by the Japanese Patent Office. The topic document is a patent including a topic claim. Relevant documents are patents that can invalidate the topic claim.

We developed a patent retrieval system for this subtask. The main scope of our method is the appropriate query expansion. In this task, the original query terms are extracted from the topic claim. However, as the claims are described abstractly in many cases to enlarge the scope of the claim necessary query terms could not be extracted from the topic claim, and the recall of the result is low. To solve this issue, we applied query expansion methods in which the query terms are extracted from sentences in the "detailed description" of the topic documents. Figure 1 shows the structure of patent. A patent consists of an "International Patent Classification (IPC) code," "abstract," "claims," "detailed description," and so on. "Claims" has one or more "claim," and a "claim" has one or more "components of the invention." "Detailed description" has many "sentences," and some of the " sentences " are relevant to a "component of the invention." In this paper, we define a "sentence" that is relevant to the "component of the invention" as the "explanation sentence." The "explanation sentence" explains more specifically about the invention to clarify the claim. We implemented the following method of extracting query terms:

(1) Extracting "components of the invention" by analyzing the "topic claim,"
(2) Extracting "explanation sentences" related to the "component of the invention," from the "detailed description," and
(3) Extracting query terms from the "topic claim" and all "explanation sentences."

Moreover, we applied the re-ranking method using "IPC code" assigned to the topic document. We also evaluated a method of re-ranking results based on "IPC code."

This paper explains in detail how we implemented these methods, and reports the results. Section 2 outlines our patent retrieval system. We describe the method of extracting query terms in Section 3, and the method of re-ranking based on "IPC code" in Section 4. We then report the results of applying these methods in Section 5, and conclude them in Section 6.

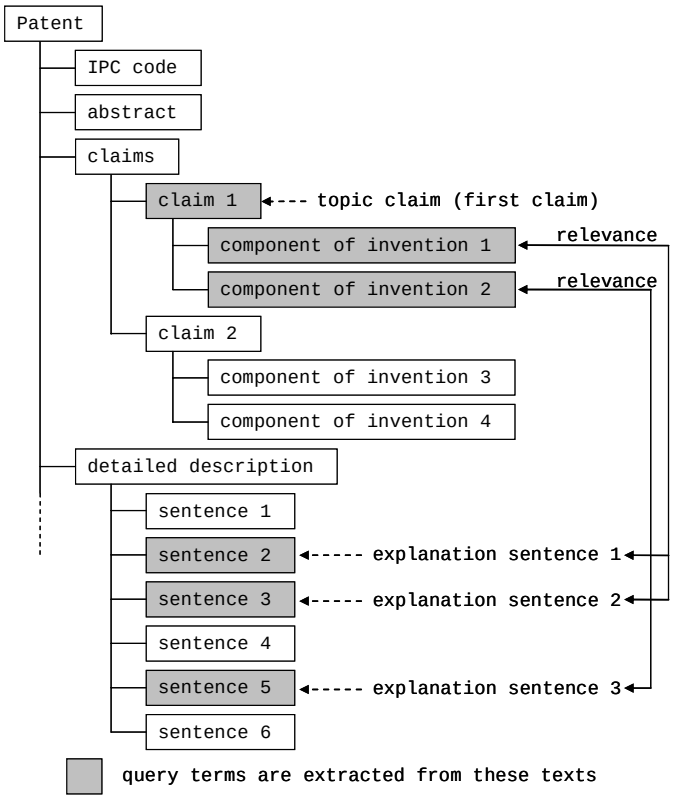


Fig.1 Structure of Patent

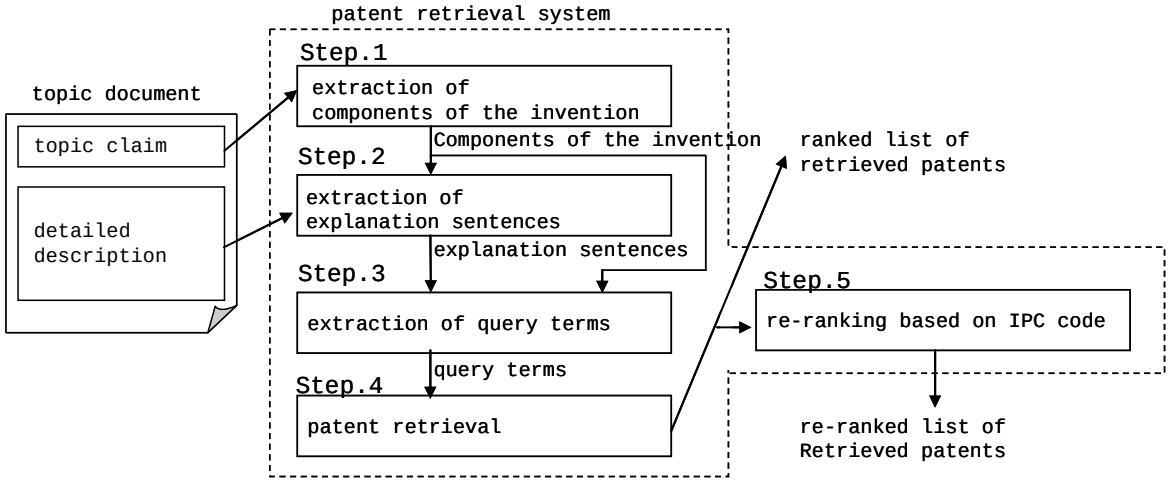


Fig.2 Patent Retrieval System

2. System Description

This section describes our patent retrieval system. The input to this system is a single topic document, and the output is a ranked list of retrieved patents. Here is a summary of each step of the retrieval process. Figure 2 shows the overview of our patent retrieval system.

(1) Extraction of "components of invention":
"Components of the invention" are extracted from the "topic claim" by applying morphological analysis and the pattern matching process. The method of extracting "components of the invention" is discussed in detail in Section 3.1.

(2) Extraction of "explanation sentences":
"Explanation sentences" are extracted from the "detailed description" in the topic document, by applying morphological analysis and the pattern matching process. The method of extracting "explanation sentences" is discussed in detail in Section 3.2.

(3) Extraction of query terms:
We performed morphological analysis on the "topic claim" and each "explanation sentence" to extract words (mainly nouns) as candidate query terms. Sequences of content words were also extracted as compound candidate query terms. We used 73 stop-words that frequently appeared in the existing patents. Moreover, we selected query terms that had a smaller Document Frequency (DF) than the threshold from the candidate query terms.

(4) Patent retrieval:
We retrieved patents that contained query terms and that had a publication date earlier than the filing date of the topic document. We used the Okapi BM25 formula by Robertson et al. [1] for the ranking process in this retrieval. This formula is the conventional ranking model used in many retrieval systems. The relevant score was given to each retrieved patent.

(5) Re-ranking based on IPC code:
We increased the score of the retrieved patents, whose assigned IPC codes were similar to the topic document, and re-ranked the retrieval results.

3. Extraction of Components of the Invention and Explanation Sentences

We extracted "components of the invention" from the "topic claim" and "explanation sentences" from the "detailed description." This section describes our method of extracting the "components of the invention" and "explanation sentences" in detail.

The following are examples of "topic claims" and claim in the relevant document. We used these examples for our explanation in this section.

- Topic claim:
  An electronic medical chart input and reference system characterized by IC card that has the record of the user ID and password and the IC card reading system that executes login and logout procedures.
- Claim in the relevant document:
  A medical support system characterized by function that authenticates users based on an authenticated IC card and controls data on patient information so that this can only be accessed by authenticated users.

3.1. Extraction of Components of the Invention

"Components of the invention" are extracted from the "topic claim." Claims are usually described using typical expressions that are common to a large number of existing patents. Therefore, "components of the invention" can easily be extracted through a pattern matching process [2].

First, we separated the "topic claim" per morpheme. Next, we assigned meaning types such as the "name of

the component" or "action" to each morpheme through the pattern matching process. In addition, we specified consecutive morphemes as "components of the invention" with the pattern matching process. We used ChaSen [3] as the morphological analyzer, and the Erie [4] as the pattern-matching engine. We also manually coded 241 patterns based on expressions specific to patents to assign meaning types to each morpheme and specify consecutive morphemes as "components of the invention."

The following two "components of the invention" were extracted from the "topic claim" as exemplified above by this method.

- Component of the invention:
  (A) *IC card that has the record of the user ID and password*
  (B) *IC card reading system that executes login and logout procedures*

3.2. Extraction of Explanation Sentences

In this section we describe the two methods of "explanation sentence" extraction, one method based on pattern matching, and the other method based on long common subsequence (LCS) length. We implemented the method based on LCS length to extract "explanation sentences" that are not extracted by the method based on pattern matching.

The method to expand query terms extracted from "explanation sentences" has been adopted by other systems, and the effect has been reported. On the system, the method of extraction sentences uses the conventional ranking model used in many retrieval systems [5]. We studied the method to extract more pertinent sentences.

3.2.1. Method based on Pattern Matching

This paragraph explains how "explanation sentences" were extracted that were relevant to the "component of the invention" using pattern matching.

We implemented this method for the NTCIR-4 Patent Retrieval Task [6]. It was based on the hypothesis that many "explanation sentences" are expressed using typical sentence structures in the "detailed description." The following is an example of the "component of the invention" and "explanation sentence."

- Component of the invention:
  (A) *IC card that has the record of the user ID and password*
- Explanation sentence:
  The patient information that the user can input and refer to is strictly controlled through user authentication, using the IC card that has the record of the user ID and password.

We can retrieve relevant documents as exemplified above, using terms like *"authentication," "patient,"* and *"control"* as query terms, extracted from this "explanation sentence." The following are typical sentence structures in an "explanation sentence."

- it is possible (….) by (component of the invention)
- (component of the invention) <verb> (….)
- (component of the invention) <be-verb> (….)

We manually coded 104 patterns of sentence structures in "explanation sentences," and extracted "explanation sentences" from the "detailed description" by applying Erie's pattern matching process.

We confirmed from the results of an experiment that Mean Average Precision (MAP) of the retrieval result was higher when extracting query terms both from the "topic claim" and from "explanation sentences" than when query terms were extracted from the "topic claim" only or from the whole topic document [7]. Further addition of patterns may increase the MAP of the retrieval result. However, it is a challenging problem to prevent omissions of "explanation sentences," because it is difficult to define all necessary patterns.

3.2.2. Method using Longest Common Subsequence (LCS) length

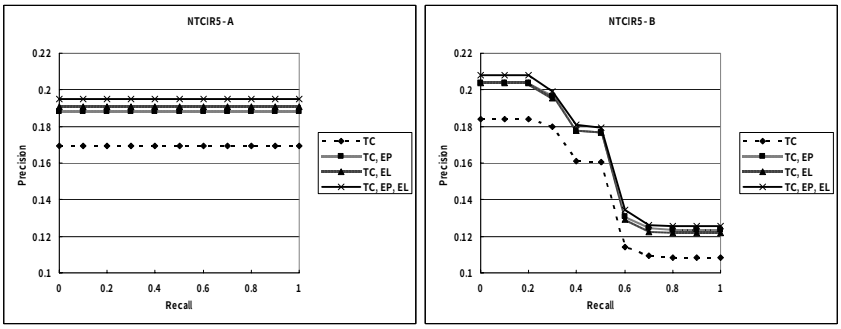
This paragraph explains how "explanation sentences" were extracted using the LCS length. The purpose of this method is to extract "explanation sentences" which could not be extracted by the method based on pattern matching because of the difficulty in defining all necessary patterns.

We evaluated it based on the following hypothesis; the same terms specifically appear in the same order in the "component of the invention" and many "explanation sentences." This is because "components of the invention" and "explanation sentences" are written by the same author. An example of the "component of the invention" and "explanation sentence" is indicated below.

- Component of invention:
  (B) *IC card reading system that executes login and logout procedures*
- Explanation sentence:
  Establishing an IC card reading system that can easily execute login and logout procedures from the user's desk prevents her/him from forgetting to log out when she/he leaves her/his desk.

We can extract terms like *"desk," "leaving,"* and *"forgetting"* from this "explanation sentence." We expected to see a similar effect with this "explanation sentence" in the relevant document because a similar invention to that in the topic document is described in the relevant document.

However, the same terms with "component of the invention" appear in the following sentence but in a different order.



Extracting Query Term from	MAP	
	NTCIR5-A	NTCIR5-B
TC	0.1695	0.1424
TC, EP	0.1884	0.1590
TC, EL	0.1908	0.1579
TC, EP, EL	0.1949	0.1619

Fig. 3 MAP by each Extracting Query Term Method

- Sentence:

  When the user removes his/her IC card from the IC card reading system, the logout procedure is executed automatically, and the screen returns to the waiting status for login again.

This sentence is not relevant to any "component of the invention" directly. In fact, terms like *"waiting"* and *"screen"* are relevant to one application of the invention, and do not appear in the relevant document including explanation of another application.

Given two sequences X and Y, the LCS length of X and Y is the maximum length of common subsequences of X and Y. Consequently, if the LCS length of the "component of the invention" and the sentence is longer than the threshold, we can extract the sentence as an "explanation sentence" that is relevant to the "component of the invention."

We assigned the "component of the invention" to X, and each sentence in "detailed description" to Y. We then found LCS length of X and Y for all Y, and computed the similarity of Y to X as the LCS length divided by the length of X. If the similarity of Y was longer than the threshold, we determined that Y was an "explanation sentence" related to the "component of the invention."

4. Re-ranking based on IPC code

This section describes how the results of retrieval were re-ranked based on the "IPC code" assigned to the topic document.

The topic document and relevant document tend to be assigned to similar "IPC code". So we implemented the method of multiplying the score of retrieved patent by constant when the "IPC code" assigned to the topic document and the retrieved patent are the same. In this method, we evaluated the identity of the IPC class, and multiplied the score of the retrieved patent when one or more IPC classes were the same for topic document and the retrieved patent. Therefore, the patent documents that have assigned the same IPC classes as the topic document will be re-ranked high up on the list.

5. Evaluation Results

In this section, we compare the MAP of the retrieval result, using query terms extracted only from the "topic claim," and expanding the query terms extracted from "explanation sentences," to evaluate the effect of our query expansion. The results showed an improvement on MAP by query term extraction from the "explanation sentences."

First, we implemented patent retrieval using the following four types of query term that were extracted from each topic document.

(1) Query terms that were extracted only from "topic claim" (TC),
(2) Query terms that were extracted from TC and from "explanation sentences" that had been extracted by method based on the pattern matching (EP),
(3) Query terms that were extracted from TC and from "explanation sentences" that had been extracted by method based on the LCS length (EL), and
(4) Query terms that were extracted from TC and from both of EP and EL.

Figure 3 shows each MAP. The lowest MAP was scored when extracting query terms only from TC, while highest MAP was scored when extracting query terms from both of EP and EL. This result shows that extracting query terms from "explanation sentences" improves the MAP. Moreover, the higher MAP was scored by extracting query terms from both of EP and EL, compared with the case of using either EP or EL only. This result shows that each method of extracting "explanation sentences" complements one another for improving the MAP.

Next, we classified all topics into the following four categories based on the amount of improvement in MAP by extracting query terms from "explanation sentences."

(A) The MAP scored lowest when extracting only from TC, but in addition to the TC, extracting query terms from both of EP and EL scored the highest MAP,
(B) The MAP scored highest when extracting only from TC, but in addition to the TC, extracting

categorization	NTCIR5-A	NTCIR5-B
(A) MAP is raised by extracting query terms from EP and EL	290 topics (47%)	609 topics (51%)
(B) MAP is reduced by extracting query terms from EP and EL	167 topics (27%)	356 topics (30%)
(C) MAP is still 0 even if query terms are extracted from EP and EL	91 topics (15%)	134 topics (11%)
(D) others	71 topics (11%)	90 topics (8%)

Fig. 4 The number of topics by categorization of MAP improvement

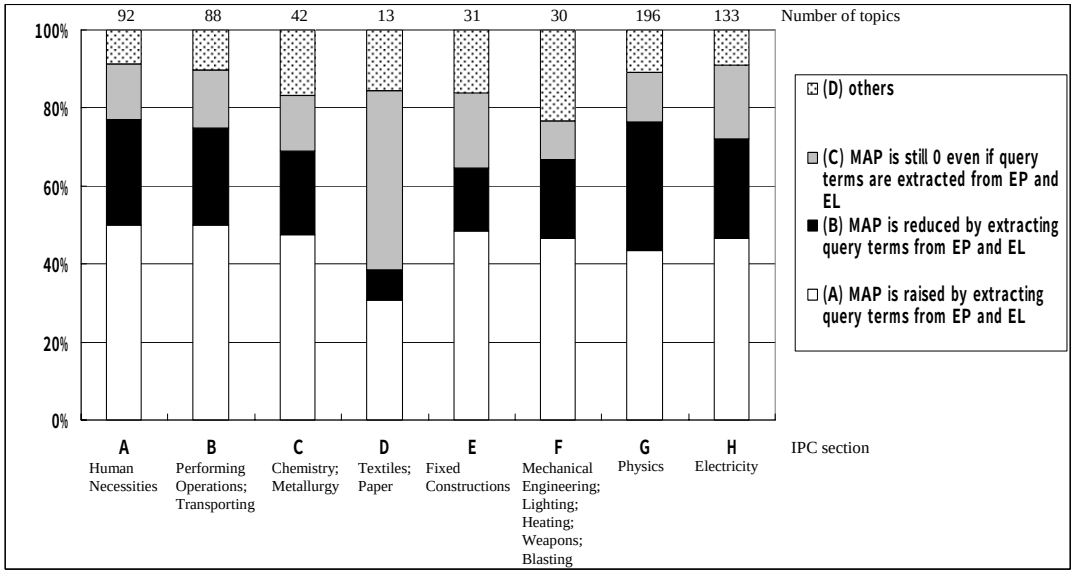


Fig. 5 Topics classification according to IPC

query terms from both of EP and EL scored the lowest MAP,

(C) Both of the cases scored zero as the MAP; one is extracting query terms only from TC, and other is extracting query terms from both of EP and EL, in addition to the TC,

(D) Others.

Figure 4 shows the four categories. (A) indicates the topics of which the extraction of query terms from "explanation sentences" improved the MAP, but (B) shows the topics of which the extraction of query terms from "explanation sentences" lowered the MAP. The extraction of query terms from explanation sentences improved the MAP for more than 40 percent of topics.

Figure 5 shows the classification of all topics based on the amount of improvement in MAP which was scored by extraction of query terms from "explanation sentences" according to the head IPC section assigned to the topic document. In most of the IPC section, extraction of query terms from "explanation sentences" improved MAP for more than 40 percent of the topics. However, for 50 percent of the topics of the D section (Textiles; Paper), the extraction of query terms from "topic claim" and "explanation sentences" would not be able to retrieve relevant documents. Only 1 percent of the terms appear in the whole topic document and the whole relevant document in common, for D section topics that scored 0 as MAP when query terms extracted from "explanation sentences" are expanded. In consequence, it is necessary to expand query terms extracted from some external information sources other than the topic document for these topics.

We analyzed the actually extracted "explanation sentences" for topics of which the extraction of query terms from the "explanation sentences" lowered the MAP. Actually, the inappropriate sentences for "explanation sentence" had been extracted from topic documents for those topics. The possible reasons of the extraction of these sentences which excludes query terms for preventing omission are as follows:

(a) <u>It is difficult to specify threshold N of LCS length that distinguishes explanation sentences by constant.</u>

The incorrectly extracted sentence as an "explanation sentence" has the same N terms appearing in the same order with the "component of the invention." However, the sentence that should be extracted as the "explanation sentence" also may have same N terms appearing in the same order with the "components of the invention." Therefore, it is too difficult to specify the threshold N of LCS length simply by constant.

(b) <u>It is difficult to evaluate the identity of the terms without examination of modification relation of the terms.</u>

There are terms that appear in the both "component of the invention" and inappropriate sentence for "explanation sentence" in common. However, each term may modify different term

Extracting Query Term from	MAP	
	NTCIR5-A	
	do not apply re-ranking based on IPC	apply re-ranking based on IPC
TC	0.1695	0.1787
TC, EP	0.1884	0.1939
TC, EL	0.1908	0.1947
TC, EP, EL	0.1949	0.1999

Fig. 6 MAP by Re-ranking based on IPC

respectively. Especially when the LCS length is small, the terms appear in the "component of the invention" and the sentence in common, but the modification relation of the terms tend to be different from each other.

We need to examine specifying a dynamic threshold and evaluation of identity of the terms considering the modification relation in order to apply the method of extracting "explanation sentences" based on LCS length effectively.

Figure 6 plots the results of the application of the re-ranking method based on "IPC code" for the results of patent retrieval. Re-ranking based on "IPC code" consistently improved MAP. It may be possible to assign a similar "IPC code" to the topic document and the relevant document. Therefore, it is essential to consider "IPC code" in invalid searches.

6. Conclusion

We evaluated a system for retrieving patents that were similar to the topic by extracting query terms from "explanation sentences" in the "detailed description" that related to each "component of the invention" in the "topic claim," and from the "topic claim" itself.

In extraction of the "explanation sentences," we evaluated a method based on pattern matching and a method based on LCS length. We found that the method based on LCS complements the method based on pattern matching by applying it to the task. Moreover, we evaluated a re-ranking method based on "IPC code". We confirmed it improved the MAP of retrieval.

The specifying dynamic threshold and evaluation of identity of the terms considering the modification relation remain to be solved regarding the "explanation sentence" extraction method based on LCS length. A future issue is to retrieve relevant documents that have similar content to the "topic claim" using query terms excluded from the topic document.

References

[1] S.E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: Automatic ad hoc filtering, VLC and interactive. Proceedings of the 7th Text REtrieval Conference(TREC-7), NIST Special Publication 500-242, pp.253-264, 1999.

[2] T. Takaki, A. Fujii, and T. Ishikawa. Associative Document Retrieval by Query Subtopic Analysis and its Application to Invalidity Patent Search. Proceedings of the 13th Conference on Information and Knowledge Management (CIKM 2004), pp. 399-405, 2004.

[3] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. Technical Report NAIST-IS-TR99009, NAIST, 1999.

[4] Y. Eriguchi and T. Kitani. NTT Data Description of the Erie System Used for MUC-6. Proceedings of Tipster Text Program (Phase II), pp. 469-470, 1996.

[5] A. Fujii and T. Ishikawa. Document Structure Analysis in Associative Patent Retrieval. NTCIR Workshop 4 Proceedings. http://research.nii.ac.jp /ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-PATENT-FujiiA.pdf

[6] K. Konishi, A. Kitauchi, and T. Takaki. Invalidity Patent Search System of NTT DATA. NTCIR Workshop 4 Proceedings. http://research.nii.ac.jp /ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-PATENT-KonishiK.pdf

[7] K. Konishi, A. Kitauchi, and T. Takaki. Patent Retrieval by Query Terms Extraction based on Characteristics of Invention, Proceedings of Data Engineering Work Shop, DEWS2004, 3-b-1, 2004 (in Japanese).