Invalidity Patent Search System of NTT DATA

Kazuya KONISHI **Akira KITAUCHI** **Toru TAKAKI**
Research and Development Headquarters, NTT DATA CORPORATION
Kayabacho Tower Bldg., 1-21-2 Shinkawa, Chuo-ku, Tokyo 104-0033, Japan
{konishikzy, kitauchia, takakit}@nttdata.co.jp

Abstract

In this paper, we give an overview of our invalidity patent search system for NTCIR-4 PATENT. The system is based on the document retrieval technique and the new methods that are suitable for the invalidity search; the query term extraction based on characteristics of invention, the retrieval model using components of invention, the ranking using the term weighting based on category information, and so on. This paper describes these methods, and evaluates the search results given by our methods.

Keywords: *Patent retrieval, Invalidity search, Query term extraction.*

1. Introduction

The NTCIR-4 Patent Retrieval Task is an invalidity search. In this search, the examiners have to find the existing patent specifications that describe the same invention of the topic claim. However, it is often difficult to retrieve such specifications by using the common type of document retrieval system based on term matching. The reasons for this problem are listed below.

1. Since the terms included in a claim are often abstract or creative in order to expand the claim's scope, different specifications tend to comprise different terms even if these terms explain the same things.
2. It is possible that a subset of terms in a claim match with an invention component that is different from any of the invention components in the topic claim. This happens because a subset of the terms does not necessarily specify the invention component.
3. The degree of distinguishing one invention from another depends on the level of the specialization of the patent classification of the invention. Since patent classifications are highly specialized and independent from each other, the interpretation of the term varies from field to field.

Through consideration of these reasons, we have developed and implemented document retrieval methods that are suitable for invalidity searches. This paper describes these methods from the perspective of the first reason above. Additionally, it evaluates the search results given by our methods.

2. System Description

First of all, we provide a description of the invalidity search system as background information before describing our retrieval methods. The input to this system is a single patent specification. The specification in turn has a single topic claim. The system output is a list of existing specifications that describe the same invention of the topic claim. The system conducts the search after producing queries corresponding to the invalidity search based on the terms included in the topic claim. Here is a summary of each step of the process.

(1) Query term extraction:
We perform morphological analysis to extract the word (mainly nouns) from the topic claim as query terms. We use ChaSen [1] as the morphological analyzer. Additionally, sequences of content words are extracted as compound query terms. We use 73 stopwords that appear frequently in the existing specifications.

(2) Existing patent specification retrieval:
We retrieve the existing patent specifications that describe inventions that might be identical to the one of the topic claim. We use the BM25 formula of Okapi [2] for the ranking process of this retrieval. This formula is a conventional ranking model used in many retrieval systems.

3. Retrieval Methods

3.1. Query Term Extraction based on Characteristics of Invention

In this section, we explain how to extract the query terms focused on the characteristics of the invention [3]. This method solves the problem in which different terms that suggest the same things are described in various different ways in different claim.

By referring to the terms of the topic claim, we extract descriptions of the invention's characteristics from the "detailed description of the invention" in the specification. The terms included in the description are set as additional query terms. Since additional query terms are related to the terms listed in the topic claim about the invention, we refer to them as "related terms" from here on.

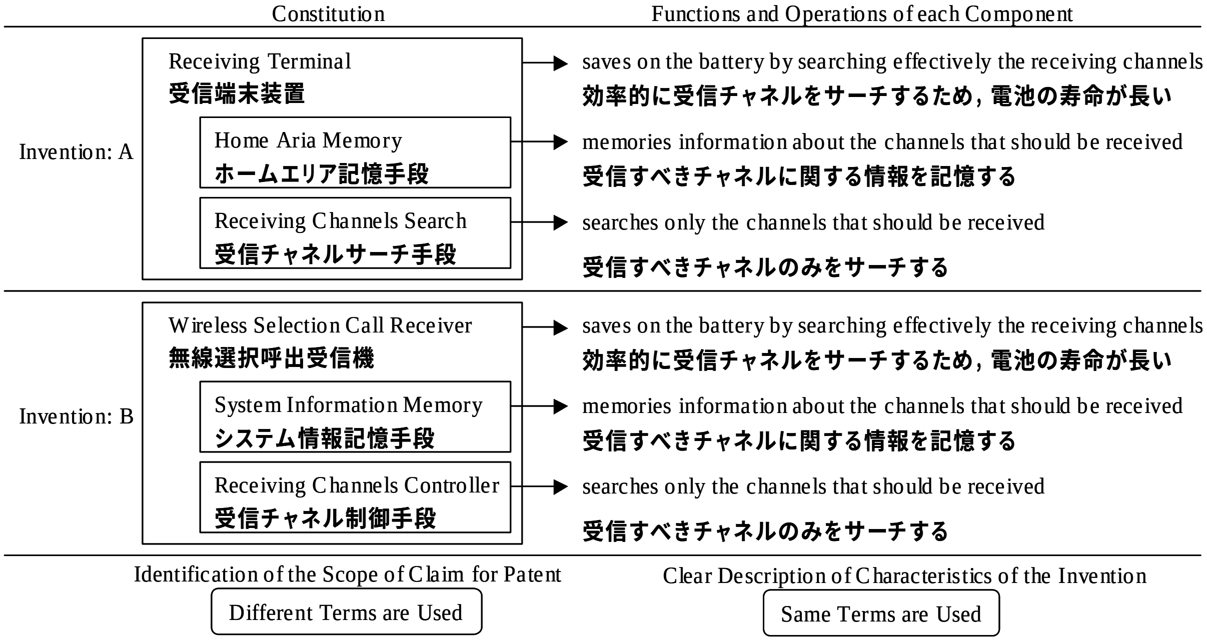

Constitution
Functions and Operations of each Component
Invention: A
Receiving Terminal
受信端末装置
saves on the battery by searching effectively the receiving channels
効率的に受信チャネルをサーチするため，電池の寿命が長い
Home Aria Memory
ホームエリア記憶手段
memories information about the channels that should be received
受信すべきチャネルに関する情報を記憶する
Receiving Channels Search
受信チャネルサーチ手段
searches only the channels that should be received
受信すべきチャネルのみをサーチする
Invention: B
Wireless Selection Call Receiver
無線選択呼出受信機
saves on the battery by searching effectively the receiving channels
効率的に受信チャネルをサーチするため，電池の寿命が長い
System Information Memory
システム情報記憶手段
memories information about the channels that should be received
受信すべきチャネルに関する情報を記憶する
Receiving Channels Controller
受信チャネル制御手段
searches only the channels that should be received
受信すべきチャネルのみをサーチする
Identification of the Scope of Claim for Patent
Different Terms are Used
Clear Description of Characteristics of the Invention
Same Terms are Used

Figure 1. Hypothesis about descriptions of specifications derived from the same invention

This method is based on the hypothesis that the descriptions are common to each specification when they are derived from the same technical ideas of the invention. The description of the characteristics of the invention can itself be characterized as follows; it explains the functions of the invention targeted by the terms of the claim as well as the operations that affect the invention.

Figure 1 shows an example of the above hypothesis. Authors may have to describe the scope of a claim for everyone to get the same interpretation since the patent specification is a technical document. However, they describe their claimed invention using abstract or creative terms which have various meanings to enlarge the scope of the claim. The terms in the claim are not well suited for the query term. Then, we consider that the functions and operations of the invention components of the topic claim are clearly indicated with concrete and general terms in the "detailed description of the invention" of the specification. Those terms are common to many specifications that describe the same invention of the topic claim, and effective as query.

Because the patent specification is a technical document, we assume that there are limitations of the expression types of the description, which explain the thing described by the claim term; the function of the used for the invention, and the operations influencing the invention, about the thing. We developed methods of extracting these descriptions by conducting pattern match. For the pattern match, the expression patterns were defined as continuous morphemes patterns. The below is a summary of the three kinds of expression patterns we developed. The underlined parts indicate the claim terms.

(1) Enumeration expression patterns:
These enumerate the things that contain the same functions or operations of the thing described by the claim term.
(ex) "a memory storage such as a flash memory and ROM"
「フラッシュメモリや、ROM 等の記憶手段」

(2) Defining expression patterns:
These define the functions used for the invention, about the thing described by the claim term.
(ex) "a receiving terminal that achieves a battery saving"
「バッテリーセービングが可能となる受信端末装置」

(3) Explaining expression patterns:
These explain the operations influencing the invention, about the thing described by the claim term.
(ex) "since the search measure starts the search, the receiving terminal can quickly find the channel which should be received next"
「サーチを開始するので、次のサーチにおいて受信すべきチャネルを速やかに発見することができる」

Figure 2 shows the flow of this method. First, the administrator of the search system prepares templates of these continuous morphemes patterns. Second, the system completes the expression patterns by applying the claim terms for the templates. Third, the system extracts the character strings that match the expression patterns from “the detailed description of the invention” part of the specification. We use the Erie system [4] as the character strings extractor. After that, the system extracts the terms included in the character strings and assigns

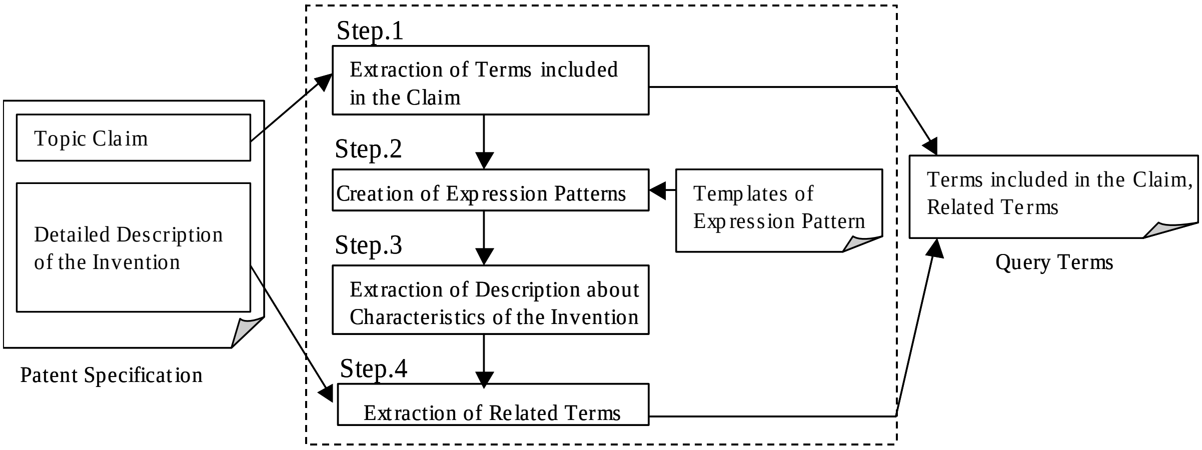

Step.1
Extraction of Terms included in the Claim
Step.2
Creation of Expression Patterns
Step.3
Extraction of Description about Characteristics of the Invention
Step.4
Extraction of Related Terms
Topic Claim
Detailed Description of the Invention
Patent Specification
Templates of Expression Pattern
Terms included in the Claim, Related Terms
Query Terms

Figure 2. Processes of query term extraction based on the characteristics of the invention

them as related terms. The search system performs the search using the claim terms and the related terms as the query terms.

3.2. Other Methods

3.2.1. Retrieval Model using Components of Invention

The invention claimed in a patent application usually includes multiple components. In case of the invalidity search, an examiner intends to find one or more similar patents that include all or the majority of the components in the topic claim. Moreover, it is effective to indicate which component is described or not in the retrieved specification. Although a specification likely contains multiple components, the importance of each component is different. As a query term having its weight in the IR model, the weighting method for each component is needed.

The Jepson style is a writing form for patent claims. The Jepson claim consists of two description parts. The first part is a preamble portion that describes existing technologies, and the second part is an essential portion that describes the features peculiar to the invention. The components in the essential portion are more important than those in the preamble portion. The invalidity search system should have a function that enables a precise search focusing on the essential points of novelty and the existing technologies.

In the invalidity search, although there are usually many specifications that include the query terms, there are generally few specifications to which the essential contents match almost completely. Thus, it is important to produce queries that reflect the essence of the topic claim. We implemented a method that uses the individual components in the claim. For each component, a query is produced and relevant specification candidates are retrieved based on the relevance score. Then, by integrating each relevance score weighted by the importance of each component, the final relevant s are determined.

3.2.2. Query Term Expansion (LCA)

A method similar to LCA [5] was adopted as the query expansion technique. In our search system, the extended terms were extracted from the top ten ranked passages, although the original LCA method extracts from the top ranked specifications. We restricted the maximum number of extended terms to ten.

3.2.3. Ranking using Term Weighting based on Category Information

The invention of patent specifications is classified in accordance with the International Patent Classification (IPC). We developed an algorithm for term weighting based on the use of category information labeled specifications [6]. Our approach is to weight a term differently for each category only if the term has high relevance to the specific categories.

The basic idea of category-based term weighting is to extend the relationship between terms and documents (specifications) in the tfidf measure to that between terms and categories, which are given by

$$tfidf(d,t) = tf(d,t) \cdot idf(t)\,, \qquad (1)$$

where

$$tf(d,t) = \log(\frac{f_d^t}{f_d} + 1)\,,$$

$$idf(t) = \log\frac{N}{N_t}\,,$$

f_d^t is the term frequency of term *t* in document *d*, f_d is the total frequency of all terms in document *d*, *N* is the total number of documents, and N_t is the document frequency of term *t*, and

$$cdficf(c,t) = cdf(c,t) \cdot icf(t), \qquad (2)$$

where

$$cdf(c,t) = \log(\frac{N_c^t}{N_c} + 1),$$

$$icf(t) = \log \frac{NC}{NC_t},$$

N_c^t is the document frequency of term *t* in category *c*, N_c is the number of documents in category *c*, *NC* is the total number of categories, and NC_t is the category frequency of term *t*.

The criterion for determining whether a term has high relevance to specific categories is defined as below:

$$rel(t) = \frac{\log(N_t + 1)}{\log(NC_t + 1)}. \qquad (3)$$

The term weight considering the relationship between terms and categories is

$$weight_{cat}(c,t) = \begin{cases} cdficf(c,t) & (rel(t) > th_r) \\ \log(\frac{N_t}{N} + 1) \cdot icf(t) & (rel(t) \le th_r) \end{cases} \qquad (4)$$

where th_r is a threshold to judge whether the term *t* should be weighted for each category. We further integrate the term weight with the *tfidf* weighting, which is the measure based on the relationship between terms and documents.

$$weight_{comb}(d,c,t) = \sqrt{weight_{cat}(c,t) \cdot tfidf(d,t)} \qquad (5)$$

To reduce the execution time for ranking documents, a two-step approach is used for retrieval. The first step outputs the top 3,000 documents ranked by a score using the BM25 that is the same weighting scheme based on the relation between terms and documents as tfidf. In the second step, we rerank these documents by a score using our weighting scheme, and take the top 1,000 documents as the final result for the retrieval.

IPC is organized with a five-level hierarchy, and we employ the third level called "subclass" which has 1,233 categories as the set of categories for the term weighting.

3.2.4. Ranking using Passage Retrieval Score

In the ranking processing of our search system, a score is usually given to each specification. A low ranking may be given to long specifications that include the relevant description in a specific portion of the specification because the most often used ranking method uses document length as a ranking feature. We developed the method of calculating the final score with the specification score and the passage score in order to give a higher score to partially relevant specifications.

3.2.5. Hybrid Method

We implemented a module that changes the patent retrieval method according to the features of the topic claim. The features are the importance of the query terms in the claim and the existence of the preamble portion.

The former feature was used to judge whether the query term extraction based on the characteristics of the invention should be used, and the latter was to judge whether the ranking method that uses the individual components in the topic claim should be used.

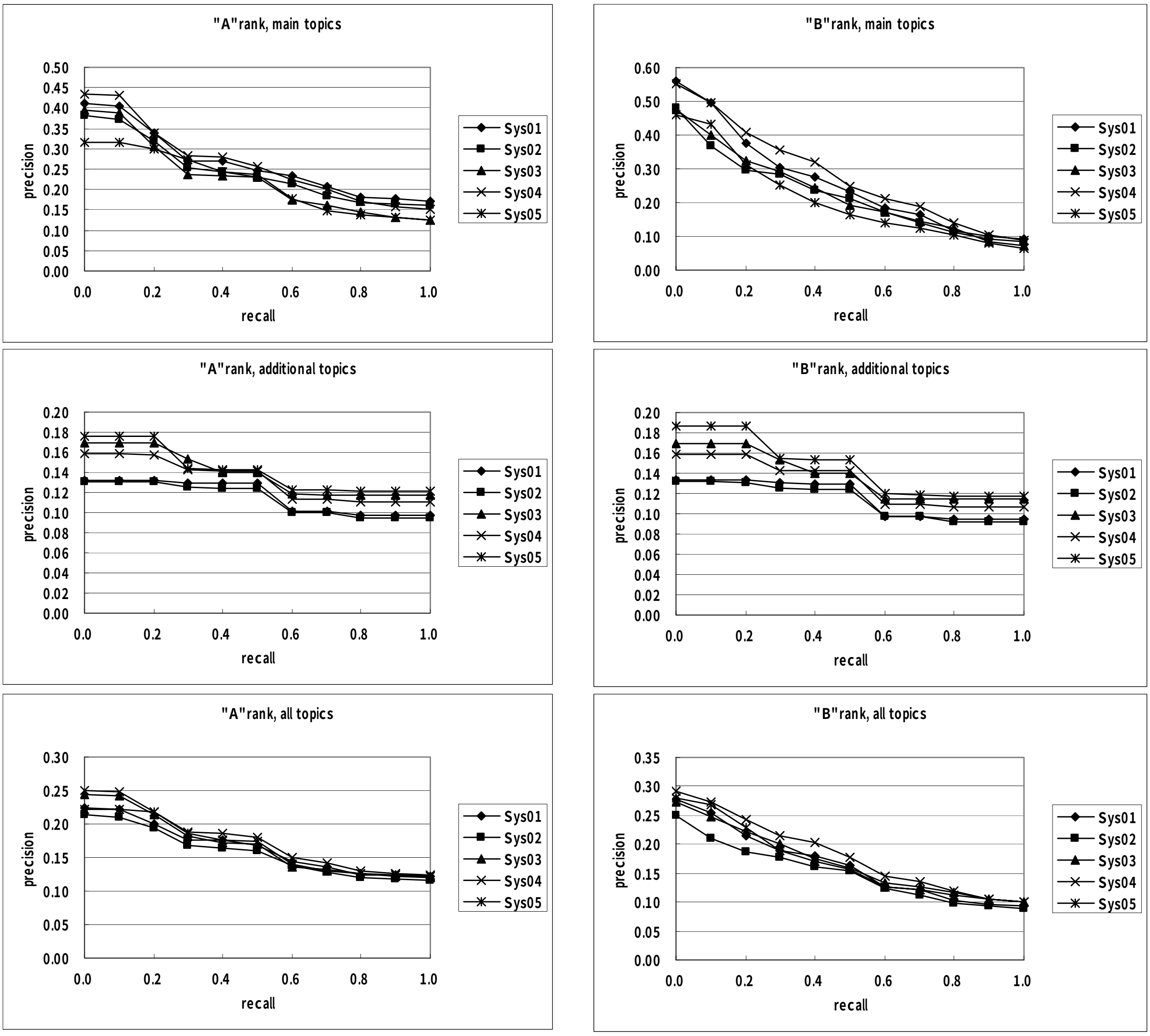
4. Search Result

We submitted a number of systems for the NTCIR-4 Patent Retrieval Task. For all systems, the collection was a publication of unexamined patent applications in 1993-1997. The index consisted of morphemes. All systems were produced using the base system described in section 2 and combinations of methods described in section 3. Figure 3 summarizes the results of the evaluation about some of the systems relating to the extraction of the query terms based on the characteristics of the invention.

- **Sys01**: the base system
- **Sys02**: the base system using the query term expansion (LCA)
- **Sys03**: the base system using the query term extraction based on characteristics of invention
- **Sys04**: the base system using the query term extraction based on characteristics of invention and the hybrid method
- **Sys05**: the base system using the query term extraction based on characteristics of invention, the retrieval model using components, and the ranking using term weighting based on category information

5. Discussion

There was not much difference in the precision of the retrieval between the "A" rank patents (These patents can invalidate a topic claim by itself) and the "B" rank patents. (These patents can invalidate a topic claim when it is used with other patents.) On the other hand, we could confirm that there were differences in the precision



System	MAP					
	A, main	A, additional	A, all	B, main	B, additional	B, all
Sys01	0.2476	0.1144	0.1583	0.2465	0.1135	0.1583
Sys02	0.2303	0.1127	0.1515	0.2101	0.1111	0.1444
Sys03	0.2143	0.1375	0.1629	0.2151	0.1354	0.1622
Sys04	0.2475	0.1308	0.1693	0.2666	0.1293	0.1755
Sys05	0.2108	0.1404	0.1636	0.1961	0.1444	0.1618

Figure 3. Results of the evaluation about some of the systems

of retrieval among the main topic, the additional topic, and all the topics. The main topic was that the assessors identify the relevant specifications in addition to the citations provided by the examiners of the Japanese Patent Office (JPO). The additional topics used only the citations provided by the JPO examiners as the relevant specifications.

As for the main topic, **Sys04** retrieved the relevant specifications with high precision. In other words, the query term extraction based on the characteristics of the invention could extract terms that were common to many specifications that describe the same invention and the terms found by the examiners.

For the additional topics, **Sys05** retrieved the relevant specifications with high precision. The precision of **Sys04** for the additional topics was not bad, and **Sys04** was the best for all topics with high precision.

However, overall, the precision of **Sys01** was good; therefore, our method leave to be improved. Note that by and large, the precision of **Sys02** was bad. Consequently, we can expect that the relevant terms on the claim terms selected on the basis of the common sense are not suited for identifying inventions.

6. Conclusion

We have analyzed the characteristics of patent specifications and examined methods of retrieving the specifications of inventions identical to the one described in the topic claim. The results of the NTCIR-4 Patent Retrieval Task showed that our methods had a beneficial effect on the invalidity search. It is considered that focusing on the extraction of the common terms in the specifications particularly describing identical inventions was the reason for this result. However, the increase in precision by applying our methods was modest at best. Further examinations of our methods are planned in the future.

References

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

[2] S.E. Robertson, S. Walker, 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.

[3] K. Konishi, A. Kitauchi, 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)

[4] Y. Eriguchi, 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] J. Xu and W.B. Croft. Query expansion using local and global document analysis. In Proc. of the 19th annual international ACM SIGIR conference on research and development in information retrieval, pp. 4-11, 1996.

[6] A. Kitauchi, K. Konishi, T. Takaki, Term Weighting Using Category Information for Information Retrieval, Proceedings of Data Engineering Work Shop, DEWS2004, 2-b-5, 2004. (In Japanese)