Two-Stage Patent Retrieval Method Considering Claim Structure

Hisao Mase[*] Tadataka Matsubayashi[**] Yuichi Ogawa[**] Makoto Iwayama[***] Tadaaki Oshio[†]

[*]Systems Development Laboratory, Hitachi, Ltd.
292 Yoshida-cho, Totsuka-ku, Yokohama-shi, Kanagawa 244-0817, Japan
mase@sdl.hitachi.co.jp

[**]Software Division, Hitachi, Ltd.
890 Kashimada, Saiwai-ku, Kawasaki-shi, Kanagawa 212-8567, Japan
tadamats/y-ogawa@itg.hitachi.co.jp

[***]Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
iwayama@crl.hitachi.co.jp

[†]Japan Patent Information Organization
4-1-7 Toyo, Koto-ku, Tokyo 135-0016, Japan
t_oshio@japio.or.jp

Abstract

This paper proposes a patent retrieval method that consists of two processing stages. In Stage 1, analysis and retrieval methods to improve recall are applied. In Stage 2, the top N documents retrieved in Stage 1 are re-arranged by applying analysis and retrieval methods that consider the claim structure to improve precision. This paper gives an overview of this retrieval method and evaluates its performance at the NTCIR4 Patent Retrieval Task.

Keywords: *Patent Retrieval, Claim Structure Analysis, Keyword Extraction, Allomorph Expansion, Related Term Expansion, Document Filtering, Score Merging.*

1. Introduction

Text retrieval methods using a natural language text as an input are becoming popular. These methods focus on a keyword set extracted from the input text and calculate the similarity between this keyword set and that extracted from each of the retrieval target documents.

Keyword-based document retrieval methods have three technical issues:

(a) How to extract appropriate keywords
(b) How to assign weights to the keywords
(c) How to treat allomorphs and synonyms

This paper proposes a patent retrieval method to solve these problems and to improve retrieval accuracy. This method consists of two processing stages: in Stage 1, analysis and retrieval methods to improve recall are applied, and in Stage 2, the top N documents retrieved in Stage 1 are re-arranged by applying analysis and retrieval methods that consider the claim structure to improve precision.

Section 2 overviews our two-stage retrieval method. Section 3 describes the analysis and retrieval methods used in each stage. Section 4 evaluates the feasibility of our method by using test data of the NTCIR4 Patent Retrieval Task.

2. Overview of two-stage patent retrieval method

The processing flow of our two-stage patent retrieval method is shown in Figure 1. The following methods are applied in both stages.

(1) Morphological Analysis

Claim text as a query is divided into terms and the part-of-speech is assigned to each term. Before this processing, the hyphens occurring just after a KATAKANA letter in a query text are replaced with a CHOUON letter.

(2) Stopword Deletion

The unimportant terms are deleted from the terms extracted from the query. Approximately 2900 stopwords are collected by hand in advance. They

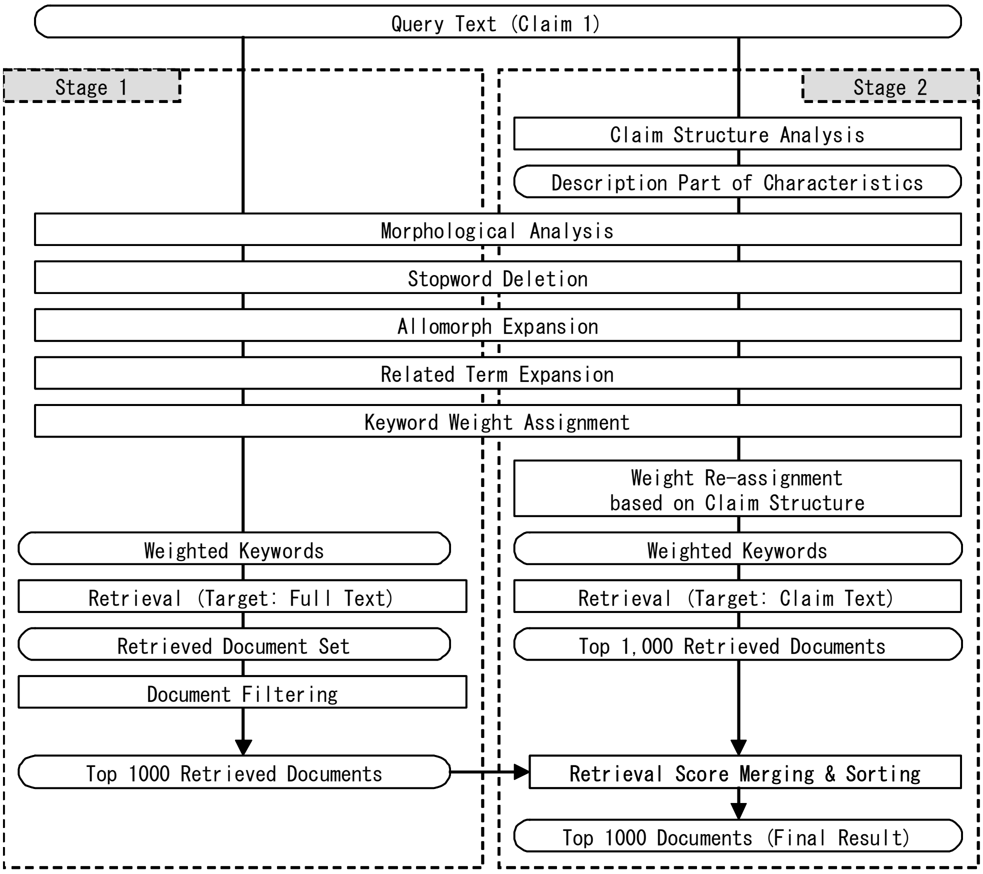

Figure 1. Overview of two-stage patent retrieval method.

include the terms that occur frequently, independent of the document format and the technical field of the invention (e.g., "こと(thing)" and "する(do)") and the terms frequently used in patent documents (e.g., "発明 (invention)", "前記 (said)", and "装置 (apparatus)"). Since stopword deletion processing is effective to improve both recall and precision according to the results of a preliminary experiment, this processing is applied in both stages.

(3) Allomorph Expansion

An allomorph dictionary, which includes a set of terms with the same meaning but with slightly different letter strings, is used to handle allomorphs as the same terms.

Allomorph expansion is executed with a bootstrap approach using an existing allomorph dictionary. First, simple allomorph expansion rules are generated from an allomorph dictionary for machine translation (e.g., "ター" is translated into "タ"). Then, the expanded term candidates for the original term are obtained by applying the expansion rules to the terms in the documents. Finally, only the terms included in the target document set are selected from the candidates. Because there are many KATAKANA terms in patent documents, allomorph expansion is applied to only KATAKANA terms. Since allomorph expansion processing is effective for improving both recall and precision according to the results of a preliminary experiment, this processing is applied in both stages.

(4) Related Term Expansion

The semantic similarity between two arbitrary terms is calculated by analyzing a lot of patent documents in order to generate a dictionary of related terms. The keywords extracted from a query are expanded to related terms by use of the related-term dictionary. Keywords expanded by this processing help to improve the retrieval accuracy.

The related-term dictionary is automatically generated using either of two clues:

(a) Term Co-occurrence

Terms used in the same context are usually related to each other. Based on this hypothesis, each noun is expressed as a vector of verbs that modify the noun, and the nouns whose verb vectors are similar to each other are extracted as related terms. For example, the noun "学校(school)" is expressed with a verb vector which includes "行く(go)" , "入学する(enter)" and "卒業する(graduate)". A term whose verb vector is similar to that of "学校(school)", such as "大学 (university)", is extracted as a related term of "学校 (school)". The relationship between nouns and verb is inferred from the morphological analysis results using normalized expressions.

(b) Expression using Parentheses

In some patent documents, related terms are explicitly described using parentheses by inventors, such as "プリンタ(印字装置)". The term just before the parenthesis and the term in the parentheses have a broader/narrower relationship to each other. The expression patterns with parentheses are extracted from the text in a "Description of Symbols" in the patent documents for 5 years. Then, the noise terms are deleted. Finally, the related-term dictionary is generated from 35,835 distinct parentheses expressions.

(5) Keyword Weight Assignment

Keywords are identified from the terms extracted from query texts according to their part-of-speech. Then, their weights are assigned using the TF/IDF method (Term Frequency Inverted Document Frequency method). Document frequency (DF) is calculated focusing on the keyword occurrence in claim texts (NOT in the whole patent text). Also, since the length of a query text is short and the really important keywords do not always occur frequently in a claim text, term frequency (TF) might not be a useful clue for keyword weight assignment. Therefore, two weight assignment methods are used: (a) using both TF and IDF and (b) using IDF only (TF is fixed to 1, see Section 4).

(6) Execution of Patent Retrieval

Patent documents in the database are searched to find ones similar to a query text and the top N documents with higher retrieval scores are output. Note that the search scope is whole texts in Stage 1 but claims in Stage 2.

The following section describes the main analysis and retrieval methods in each stage.

3. Description of analysis and retrieval method in each stage

3.1 Stage 1: retrieval for higher recall

In Stage 1, the retrieval is executed focusing on improving recall, by which more correct patent documents are included in the top 1000 retrieved patent documents. In this stage, three kinds of document filtering processing methods are applied as well as stopword deletion, allomorph expansion, and related term expansion: category-based document filtering using "International Patent Classification (IPC)", filtering considering retrieval results using subsets of keywords extracted from a query text, and filtering using "relevant passages".

(1) Document Filtering using IPC

In this processing, IPC Subclasses related to the query text are identified by applying the KNN method (K-Nearest Neighbors method) and the documents to which none of the identified IPC Subclasses are assigned are filtered out. In applying the KNN method, the IPC Subclasses are extracted from the top 15 retrieved patent documents, then the document frequency of each IPC Subclass is calculated, and finally the four most frequent IPC Subclasses are identified as related IPCs (if there are IPCs with the same frequency, all of them remain).

(2) Document Filtering considering Retrieval Results using Keyword Subsets

When the patent retrieval is executed using all keywords extracted from a query text, the retrieval result is sometimes bad even though appropriate keywords are included in the keyword set. This is mainly because the keyword set includes many noisy keywords. We thought that using keyword subsets is effective for retrieving patent documents without any misses. That is, multiple patent retrieval results are derived using keyword subsets extracted from a query text and the top N documents of each result set are collected as the retrieval results.

The processing flow is as follows. After the keywords have been extracted from a query, the patent retrieval is executed using all keywords and the result (Result-1) is derived. Then, the multiple keyword subsets are extracted. Each consists of three keywords neighboring each other in the query (the keyword weights are calculated using TF in the whole query). Then, the patent retrieval is executed using each keyword subset. In this retrieval, document filtering using IPC Subclasses is applied. Then, the top 1500 patent documents in each patent retrieval result are merged into a filtering document list. Finally, the documents included in both this filtering list and Result-1 are output while keeping the order of Result-1. This filtering result is the final output of Stage 1.

In this paper, the method of extracting three keywords neighboring each other is used for keyword subset generation. However, a better method might be to divide a claim text into elements and extract keywords from each element as a keyword subset. Though research efforts on a method to automatically divide a claim into elements have been reported [1], some technical problems remain to be solved. That is why we applied the simpler method mentioned above to evaluate the feasibility.

(3) Document Filtering using "relevant passages"

The patent documents that invalidate the invention described in a query text include "relevant passages" that describe the elements of the invention of the query.

Thus, in this paper, we hypothesize that the basis is described by the passage and that the retrieved documents that include the passage describing the elements of the invention of the query should be treated as important documents.

In this filtering processing, after keywords have been extracted from the query (called "query keywords”), the distinct number N of query keywords included in each retrieved document is counted. Then, the distinct number P of query keywords included in each passage of the retrieved document is counted, and the ratio R (=P/N) is calculated. The retrieved document that includes at least one passage whose value of R is higher than a threshold is selected as the retrieval document in Stage 1.

In this filtering method also, it might be better to consider the claim structure. In this paper, however, the above-mentioned method is applied as a first step to evaluate the feasibility.

3.2 Stage 2: retrieval for higher precision

In Stage 2, retrieval is executed focusing on improving precision. In this stage, only the 1000 retrieved documents in Stage 1 are targets of processing. Therefore, if the correct document is included in the 1000 documents, the result document set in Stage 2 includes it. One key point is that more correct documents are included in the document set in Stage 1. Stage 2 uses more detailed analysis and retrieval methods to re-calculate retrieval scores and re-arrange the document rank in Stage 1.

An overview of the analysis and retrieval processing in Stage 2 is given below (see Figure 1 again).

(1) Keyword Re-extraction Considering Claim Structure

A patent claim consists of three description parts:

(A) “premise description part”

This part describes the premise of the invention using the expressions such as “～において、” and “～であって、”.

(B) “characteristics description part”

This part describes the newness of the invention.

(C) “invention target description part”

This part describes the target to which the invention is applied.

Though the premise description part is helpful to identify its technical field, it is not necessary to judge the similarity between two inventions. Thus in Stage 2, the premise description part is deleted from a query text and the other two description parts are analyzed for keywords.

(2) Keyword Weight Re-assignment based on Claim Structure Analysis

Since the target scope of the query text is changed as mentioned above, the keyword weight should be re-assigned. Also, in Stage 2, the following two kinds of keywords are focused on and additional weight is assigned to them:

(a) Keywords in Invention Target Description Part

The invention target description part includes many important keywords that identify the invention target. These keywords, however, appear less frequently than other keywords. Thus, the keywords appearing in the invention target description part are assigned additional weights (in reality, the TF value is changed).

(b) Keywords Accompanied by Measurements

In claims, some terms are accompanied by numerical values. In this paper, these terms (called “measurement terms”) are treated as important keywords in the query and additional weight is assigned to them.

In this weighting method, a measurement term dictionary (consisting of 361 words) is prepared by hand (e.g., “速度(speed)”, “温度(temperature)”, “pH”, etc.). Also, not only measurement terms themselves, but also the terms neighboring them and the terms modifying them are the targets of additional weight assignment. For example, in the phrase “/用紙/の/搬送/速度/を/制御/する (control paper feed speed) ”, the word “速度(speed)” is a measurement term, and its neighboring words “搬送 (feed)” and “用紙 (paper)” are also given additional weight.

(3) Execution of Patent Retrieval

In Stage 1, the retrieval is applied to full texts as a target scope to keep higher recall. On the other hand in Stage 2, the retrieval target is a claim text. Since the number of target documents is reduced to 1000, higher precision is expected.

(4) Retrieval Score Merging

The final retrieval result is derived by merging the retrieval result in Stage 2 with that in Stage 1. In this paper, N% of the retrieval score of each document in Stage 2 is added to the score of the same document in Stage 1.

The processing flow is as follows. First the top 1000 documents are selected with their retrieval scores. Then, the average score of the top 1000 documents in each stage is calculated, and the ratio X of the two values (X=[average score in Stage1] ÷ [that in Stage 2]). Then, each score in Stage 2 is adjusted by multiplying the score by X. Finally, each adjusted value is further multiplied by coefficient Y and the resulting value is added to the score in Stage 1. According to the preliminary experiment to identify the best value of coefficient Y, the best value is Y=0.1.

4. Experiments

The effectiveness of our patent retrieval method above described was evaluated using the “formal run query set of the NTCIR4 Patent Retrieval Task”. The query set consisted of “Main data (34 queries)” and “Additional data (69 queries)”. The top 1000 retrieved patent documents for each query were output full-automatically as the retrieval result. The retrieval target patent document set consisted of 1.7 million documents issued from 1993 to 1997.

In this evaluation, Chasen[2] and ANIMA were used as morphological analysis engines, and GETA (morphological base, [3]) and a commercial document retrieval system (valuable-length n-gram base, [4]) were used as document retrieval engines.

4.1 Evaluation measurements

The NTCIR4 Patent Retrieval Task uses "Average Precision (AP)" and the number of correct documents in the top 1000 output documents as the main evaluation measurements. AP is calculated using the following formula:

$$\text{Average Precision} = \frac{1}{\sum_{i=1}^{N} X_i} \sum_{i=1}^{N} \left[\frac{X_i}{i} \left(1+\sum_{k=1}^{i-1} X_k \right) \right]$$

where, N is the total number of output documents (N=1000 in this experiment), Xi is a value denoting whether output i-th document is a correct document or not (the value is 1 if it is a correct document and 0 otherwise).

4.2 Experiment patterns

In our proposed method, many kinds of analysis and retrieval methods are used. It is necessary to evaluate the effectiveness of each processing. The following are key alternatives for constructing experiment patterns:

(1) Treatment of TF in keyword weight assignment
(2) Stopword deletion
(3) Allomorph expansion
(4) Related term expansion
(5) Document filtering
(6) Keyword weight re-assignment for measurement terms and invention target terms
(7) Score merging

Since it is difficult to execute all combinations of these alternatives, we used the patterns shown in Table 1.

4.3 Results and discussion

Experiment results are also shown in Table 1. In "Main" queries, similar documents to the query were selected from the submitted retrieval results by human experts and added to the correct documents (the average number of correct documents per query was 10.4). In "Additional" queries, only the documents selected by examiners to invalidate the query invention were used as correct documents (the average number of correct documents per query was only 1.8). Also, correct document set A was a set that completely invalidated the invention of the query and correct document set B was one that partly invalidated it.

In Table 1, the experiment pattern with the highest AP is dispersed by query set and correct document set. The following results were also derived for the effectiveness of each processing:

(1) Treatment of TF

The APs in the experiment patterns where TF was fixed to 1 were higher than those in any experiment patterns where a real TF value was used. There are two reasons for this: (a) the claim text as a query is too short to use TF values as a useful clue for keyword weight assignment and (b) in the claim text, the same word is repeated, which makes the value of TF higher than it should be.

(2) Score Merging

Score merging processing greatly improves APs in Additional queries, but not those in Main queries.

(3) Document Filtering

Though this processing is used to improve recall in Stage 1, its effectiveness is very low. In filtering using IPC, a total of six correct documents in queries #014, #015, and #025 were filtered out of the retrieved documents.

In filtering by "relevant passages", 161 (59%) correct documents out of 274 in Main queries improve their ranks, but in 70 (26%) correct documents, their ranks were worse.

Almost all of the above 70 queries included many noise keywords that were not related to the essential nature of the invention, especially in queries #012, #033, and #036. In query #036, for example, many keywords extracted from the description part on chemical formula were included in the correct document.

From these results, we conclude that only the keywords related to the essential nature of the invention should be extracted to identify "relevant passages". The approach of the claim structure analysis and technical fields specified search is useful for identifying these keywords, which helps to improve retrieval accuracy.

(4) Keyword Weight Re-assignment

Approximately half of the query texts included measurement terms. In Main queries, keyword re-assignment helped to improve AP, but not in Additional queries. In the current algorithm, all measurement terms and their neighbor terms were given additional weight. The context before and after the measurement terms should be considered for higher retrieval accuracy (for example, giving additional weight only to the measurement terms that the numerical value follows).

(5) Allomorph Expansion

In this experiment, allomorph expansion contributed little to AP. According to additional experiments, performed after submission, which were based on experiment pattern #16 in Table 1 without stopword deletion and allomorph expansion, the AP for correct document A was 0.1680 (0.2675 in Main queries) and that for correct document A+B was 0.1538 (0.2218 in Main). These values are higher than that of original pattern #16.

The cause is that too many allomorphs were expanded to improve recall, which resulted in many

Table 1. Combination of analysis and retrieval methods and their evaluation results.

Experiment ID				16	08	12	04	18	06	02	14	10	20
Method	Target Scope of IDF Calculation			Claim Text									
	Target Scope of Retrieval			Full Text									
	Stopword Deletion			Applied									
	Allomorph Expansion			Applied									
	Value of Term Frequency			Use a real TF value				TF is 1					
	Document Filtering			-	Used	-	Used	Used	-	Used	-	Used	Used
	Weight Re-assignment			-	-	Used	Used	-	-	-	Used	Used	Used
	Score Merging			-	Used	Used	Used	-	Used	Used	Used	Used	Used
	Related Term Expansion			-	-	-	-	-	-	-	-	-	Used
Results	Main 34 queries	Correct A 159 docs	AP	.2416	.2430	.2397	.2411	.2697	.2658	.2666	.2705	**.2714**	.2673
			NoD	119	117	119	117	114	**120**	114	**120**	114	112
		Correct A+B 344 docs	AP	.2142	.2251	.2201	.2241	.2384	.2430	.2441	.2436	.2433	**.2465**
			NoD	267	270	267	270	**273**	**273**	**273**	**273**	**273**	267
	Add. 69 queries	Correct A 97 docs	AP	.1100	.0897	.1034	.0876	.0942	**.1124**	.0979	.1051	.0919	.0899
			NoD	76	**78**	76	**78**	76	76	76	76	76	76
		Correct A+B 115 docs	AP	.1107	.0908	.1045	.0886	.0956	**.1134**	.0995	.1063	.0936	.0912
			NoD	89	**90**	89	**90**	89	89	89	89	89	87
	Total 103 queries	Correct A 256 docs	AP	.1534	.1403	.1484	.1383	.1521	**.1630**	.1536	.1597	.1511	.1484
			NoD	195	195	195	195	190	**196**	190	**196**	190	188
		Correct A+B 459 docs	AP	.1455	.1360	.1434	.1342	.1437	**.1570**	.1482	.1526	.1440	.1435
			NoD	356	360	356	360	**362**	**362**	**362**	**362**	**362**	354

Note 1: NoD "number of retrieved correct documents" Note 2: bold shows the best results

noise allomorphs being included.

(6) Related Term Expansion

Overall, the effectiveness of related term expansion is unknown[1]. In the related term expansion based on parentheses expression, the extracted related terms included many compound nouns such as "発光ダイオード/発光素子", which caused mismatches between keywords and related terms.

5. Conclusion

A two-stage patent retrieval method considering claim structure was proposed. Its effectiveness was demonstrated using the NTCIR4 Patent Retrieval Task.

As future work, the claim structure should be used more to improve the retrieval accuracy. The difference of technical fields should be also considered. Furthermore, it is essential to enhance the word dictionary for more correct patent document analysis.

Acknowledgment

GETA is a research effort in "Innovative Information Technology Incubation Project" promoted by the Information-technology Promotion Agency, Japan (IPA).

References

[1] A. Shinmori, et al.: Rhetorical Structure Analysis of Japanese Patent Claims using Cue Phrases, 149th IPSJ SIGNL, 2002-NL-149, pp. 65-72, 2002.

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

[3] GETA: http://geta.ex.nii.ac.jp/.

[4] HiRDB: http://www.hitachi.co.jp/Prod/comp/soft1/textsearch/index.html.

[1] A minor bug is found in a related term expansion program (morphological analysis base) after submission of the result. The experiment is being done again.