NTCIR-5 Patent Retrieval Experiments at Hitachi

Hisao Mase Tadataka Matsubayashi Yuichi Ogawa
Takaaki Yayoi Yusuke Sato Makoto Iwayama
Hitachi, Ltd.
292 Yoshida-cho, Totsuka-ku, Yokohama, Kanagawa, 244-0817, Japan
mase@sdl.hitachi.co.jp
tadamats/y-ogawa/t-yayoi@itg.hitachi.co.jp
y-sato/iwayama@crl.hitachi.co.jp

Abstract

In NTCIR-5, we used five retrieval methods proposed in NTCIR-4: (1) query term weighting using only document frequency, (2) stopword deletion, (3) two-stage patent retrieval, (4) term weighting considering "measurement terms", and (5) related term expansion. In this paper, we compare the retrieval accuracy for two test sets: 34 main queries in NTCIR-4 and 1189 new queries in NTCIR-5. Then, we evaluate the effectiveness of each method from two viewpoints: "ease of retrieval" and "identity of patent applicants". Finally, we introduce our approach to passage retrieval.

Keywords: *Patent Retrieval, Claim Structure Analysis, Term Weighting, Related Term Expansion, Score Merging, Patent Applicants, Passage Retrieval.*

1 Hitachi's approach in NTCIR-5

Our goal in NTCIR-5 is to judge whether the patent retrieval methods we proposed in NTCIR-4 are effective or not. We fixed some program bugs found after NTCIR-4, tuned some processing parameter values using 1256 training queries we collected by hand from IPDL (Industrial Property Digital Library), and enhanced reference term data including a stopword list and a related-term dictionary. To analyze the results, we divided each test set into subsets by focusing on the following two viewpoints:

(1) Ease of retrieval

For some queries, it is very easy to retrieve patents that invalidate the query invention ("relevant patent" hereinafter) with any retrieval method while for some, it is very difficult. We hypothesized that the effectiveness of retrieval methods depends on the ease of retrieval.

(2) Identity of patent applicants

According to our investigation using 210,755 patents, approximately 22% of the patents that examiners at Japan Patent Office quoted as being relevant to reject patent applications have common applicants to the application. We hypothesized that the effectiveness of a retrieval method also depends on the identity of the applicants of a query patent and its relevant patent.

Furthermore in NTCIR-5, we address passage retrieval. We focus on the following two characteristics of passage retrieval: (1) each passage text is so short that a small number of terms can be used and (2) passages in a patent are semantically related to each other. Thus, we use a method that uses sub-strings of kanji-character terms for more flexible term-matching, a method of adding terms in the most relevant claim in a relevant patent to query terms, and a method of excluding common topic terms across passages.

Section 2 briefly reviews the five retrieval methods we proposed in NTCIR-4. Section 3 describes experiments on their effectiveness. Section 4 discusses the affect of the ease of retrieval and the identity of patent applicants on retrieval accuracy. Section 5 describes our approach to passage retrieval.

2 Patent retrieval methods

In this section, we review the retrieval methods in NTCIR-4 and describe their enhancements in NTCIR-5. We introduced Vector Space Model as a retrieval model and TF-IDF method as term weighting. We used GETA[1] as a retrieval engine and Chasen[2] as a morphological analysis tool. Vector Space Model and TF-IDF method. In NTCIR-4, we proposed the following five methods [1][2]. In NTCIR-5, we tuned some processing parameter values and enhanced some reference data.

(1) Query term weighting using only DF

Though the TF-IDF method (Term Frequency Inverted Document Frequency method) is popular in

[1] GETA: http://geta.ex.nii.ac.jp/.
GETA is a research effort in "Innovative Information Technology Incubation Project" promoted by the Information-technology Promotion Agency, Japan (IPA).

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

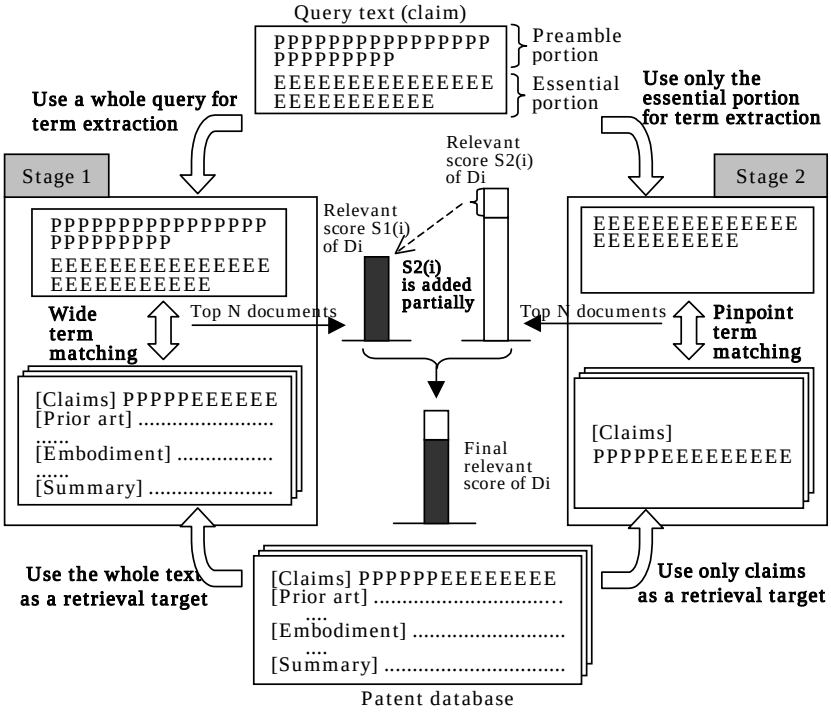

Figure 1. Overview of two-stage patent retrieval method [2].

query term weighting, we do not use TF because most query claims are so short that the term frequency is not proportional to the importance of the term. Thus, we set TF to 1 and use only IDF to assign query term weights.

(2) Stopword deletion

Our system uses nouns, verbs, adjectives, and alphabetical strings in a query claim text to retrieve relevant patents. In this method, unimportant terms are deleted from the terms with either of above part-of-speeches extracted from the query.

Though 2910 stopwords were collected by hand in advance in NTCIR-4, stopword deletion was not as effective as we expected in NTCIR-4. Thus, in NTCIR-5, we collected only 31 words that appeared in claims of more than 20% of patent documents in the patent database to keep the retrieval accuracy high.

(3) Two-stage patent retrieval

An overview of our two-stage patent retrieval is shown in Figure 1 [2]. Stage 1 is a recall-oriented retrieval to include as many relevant patents as possible within the top N retrieved documents. We use macro-level query text analysis and retrieval methods, which are common to generic text retrieval methods. That is, a whole query text is used for query term extraction and weighting in a query text analysis, and a whole text in a patent database is used as a retrieval target in text retrieval.

Stage 2 is precision-oriented retrieval to improve the rank of retrieved relevant documents. In Stage 2, only the top N documents retrieved in Stage 1 are the target of processing. We use micro-level query text analysis and retrieval methods that consider the patent structure, especially the claim structure. That is, in a query text analysis, only the essential portion of a claim is used for term extraction and weighting; the preamble portion is ignored. In text retrieval, only the claims are used as a retrieval target.

In our two-stage retrieval method, the relevant scores calculated in each stage are finally merged document by document into the final relevant score. The final relevant score S(i) of document i is calculated using the following formula:

$$\left.\begin{aligned} S(i) &= S1(i) + S2'(i) * P \\ S2'(i) &= \frac{\text{Av. of S1}}{\text{Av. of S2}} * S2(i) \end{aligned}\right\} \quad \text{[Eq. 1]}$$

where S1(i) is the relevant score of document i in Stage 1, S2(i) is the relevant score of document i in Stage 2, S2'(i) is the normalized score of document i in Stage 2, "Av. of S1" is the average value of the scores of the top N documents retrieved in Stage 1, "Av. of S2" is the average value of the scores of the top N documents retrieved in Stage 2, and P is a weight tuning parameter. According to our preliminary experiment, the optimal value of P is around 0.1. It is necessary to calculate S2'(i) to fix the value of P because the difference of S1(i) and S2(i) is big due to the difference of terms and their weights used in retrieval.

(4) Term weighting considering "measurement terms"

In claims, some terms are accompanied by

Table 1. Patterns of retrieval methods and comparison of their evaluation results.

Experiment ID		HTC01	HTC05	HTC06	HTC07	HTC08	HTC10
Retrieval methods	Term weighting with DF	-	used	used	used	used	used
	Stopword deletion	-	-	used	used	used	used
	Two-stage retrieval	-	-	-	used	used	-
	Measurement term	-	-	-	-	used	-
	Related term expansion	-	-	-	-	-	used
Baseline ID for comparison		-	HTC01	HTC05	HTC06	HTC07	HTC06

#	test set	queries	relevant patents	MAP (degree of improvement in comparison with baseline (%))					
1	4a -ALL	31	158	.2779	.3014(+ 8.5%)	.3030(+0.5%)	.2786(-8.1%)	.2973(+6.7%)	.3048(+0.6%)
2	4ab-ALL	34	342	.2120	.2393(+11.3%)	.2440(+2.0%)	.2397(-1.8%)	.2401(+0.2%)	.2506(+2.3%)
3	5a -ALL	619	619	.1837	.1861(+ 1.3%)	.1848(-0.7%)	.1933(+4.6%)	.1904(-1.5%)	.1841(-0.4%)
4	5ab-ALL	1189	2065	.1496	.1492(- 0.3%)	.1483(-0.6%)	.1555(+4.9%)	.1541(-0.9%)	.1486(+0.2%)
1-1	4a -T10	18	40	.6508	.7519(+15.5%)	.7526(+0.1%)	.6802(-9.6%)	.7145(+5.0%)	.7736(+ 2.8%)
1-2	4a -B10	26	118	.0359	.0502(+39.8%)	.0532(+6.0%)	.0537(+0.9%)	.0534(-0.6%)	.0542(+ 1.9%)
2-1	4ab-T10	23	76	.5998	.6985(+16.5%)	.7065(+1.1%)	.6486(-8.2%)	.6689(+3.1%)	.7237(+ 2.4%)
2-2	4ab-B10	33	266	.0430	.0531(+23.5%)	.0553(+4.1%)	.0588(+6.3%)	.0585(-0.5%)	.0614(+11.0%)
3-1	5a -T10	181	181	.5960	.5873(- 1.5%)	.5806(-1.1%)	.6064(+4.4%)	.5952(-1.8%)	.5699(- 1.8%)
3-2	5a -B10	438	438	.0133	.0203(+52.6%)	.0212(+4.4%)	.0226(+6.6%)	.0232(+2.7%)	.0246(+16.0%)
4-1	5ab-T10	374	427	.5571	.5354(- 3.9%)	.5282(-1.3%)	.5540(+4.9%)	.5499(-0.7%)	.5215(-1.3%)
4-2	5ab-B10	976	1638	.0141	.0191(+35.5%)	.0202(+5.8%)	.0217(+7.4%)	.0218(+0.5%)	.0222(+9.9%)
1-3	4a -SAME	8	40	.4838	.4812(- 0.5%)	.4790(-0.5%)	.4765(- 0.5%)	.4756(- 0.2%)	.4552(-4.3%)
1-4	4a -DIFF	27	118	.2038	.2332(+14.4%)	.2358(+1.1%)	.2115(-10.3%)	.2333(+10.3%)	.2421(+3.8%)
2-3	4ab-SAME	12	76	.3786	.4081(+ 7.8%)	.4068(-0.3%)	.4131(+ 1.5%)	.4145(+ 0.3%)	.4108(-0.9%)
2-4	4ab-DIFF	33	266	.1268	.1507(+18.8%)	.1562(+3.6%)	.1524(- 2.4%)	.1534(+ 0.7%)	.1608(+4.8%)
3-3	5a -SAME	107	107	.4293	.4513(+ 4.9%)	.4408(-2.3%)	.4552(+ 3.3%)	.4513(- 0.9%)	.4445(+0.8%)
3-4	5a -DIFF	512	512	.1323	.1306(- 1.3%)	.1312(+0.5%)	.1385(+ 5.6%)	.1359(- 1.9%)	.1297(-1.1%)
4-3	5ab-SAME	261	341	.3478	.3497(+ 0.5%)	.3462(-1.0%)	.3539(+ 2.2%)	.3543(+ 0.1%)	.3458(-2.4%)
4-4	5ab-DIFF	1051	1724	.1029	.1014(- 1.5%)	.1015(+0.1%)	.1081(+ 6.5%)	.1065(- 1.5%)	.1018(-4.4%)

Note: Hatched MAP value is better than that of baseline for comparison.

numerical values. These terms (called “measurement terms”) are treated as important terms 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.). Not only measurement terms themselves, but also the terms around 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 the 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.

(5) 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 terms extracted from a query are expanded to related terms using the related-term dictionary. Terms expanded by this processing help to improve the retrieval accuracy.

The related-term dictionary is generated automatically using either of two clues: (a) term co-occurrence or (b) expressions in parentheses in a “Description of Symbols” tag in a patent document. We collected approximately 611,098 related-term entries from patent documents covering 10 years.

3 Experiments

3.1. Data

We used 34 main queries in NTCIR-4 and 1189 new queries in NTCIR-5 to evaluate our patent retrieval methods. The relevant patents for the above queries were divided into two ranks: (a) patents that can invalidate a query invention and (b) patents that can invalidate a query invention when combined with other patents. We used four kinds of test sets: NTCIR-4a (31 queries, 158 relevant patents), NTCIR-4ab (34, 342), NTCIR-5a (619, 619), and NTCIR-5ab (1189, 2065). The top 1000 retrieved patent documents for each query were output completely automatically as the retrieval result. The retrieval target document set consisted of approximately 3.5 million patent documents issued from 1993 to 2002.

3.2 Evaluation measurements

NTCIR-5 uses “Mean Average Precision (MAP)”. Average precision 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), and Xi is a value denoting whether the i-th output document is a correct patent or not (the value is 1 if it is a correct patent and 0 otherwise).

3.3 Results and discussion

We used six patterns to evaluate each of our retrieval methods as shown in Table 1. We compared the results for a pattern with its baseline result.

As shown in Table 1 #1-#4, the effectiveness of our methods depended on the test sets. That is, in NTCIR-4a/ab (#1 and #2), four methods other than two-stage retrieval method were effective at improving MAP, while two-stage retrieval was much more effective than the other four methods in NTCIR-5a/ab (#3 and #4).

In NTCIR-4a/ab, "term weighting with DF" was the most effective out of the five methods. In NTCIR-5a/ab, the methods "stopword deletion" and "term weighting considering measurement terms" make MAP worse.

Possible reasons for the above differences are as follows:

(1) Difference in definition of relevant patents

The relevant patents in NTCIR-4a/ab were collected by human experts (called "pooling"). On the other hand, those for 1189 NTCIR-5a/ab were the patents that examiners at Japan Patent Office used to reject the query invention. The average number of relevant patents per query was 10.1 for NTCIR-4ab and 1.7 for NTCIR-5ab. This difference might affect the retrieval accuracy of each retrieval method.

(2) Number of queries

The NTCIR-4 query set is very small (34 queries) and covers narrow technical fields of the invention (47 IPC sub-classes). On the other hand, the NTCIR-5 query set covers much wider technical fields (328 IPC sub-classes). Field-specific factors might affect the retrieval accuracy.

(3) Behavior of MAP

Because MAP is high when the relevant patents are retrieved within top the 10, it depends strongly on the retrieval rank of these "easy-to-retrieve" relevant patents. In the next section, we analyze the retrieval results in more detail from two viewpoints: "ease of retrieval" and "identity of patent applicants".

4 Analysis of results

4.1. Ease of retrieval

We divided each of the four test sets into two subsets: "T10" is a subset including queries whose relevant patent is ranked within the top 10 with our baseline retrieval method ("HTC01" in Table 1), and "B10" is a subset including queries whose relevant document is ranked below rank of 10. Note that one query might be included in both subsets when it has two or more relevant patents.

The comparison of MAP by subset and method is shown in #1-1 to #4-2 of Table 1. In almost all of subsets "B10", MAP values were improved, while those in most of subsets "T10" were worse, especially in NTCIR-5a/ab. These results show that our methods are effective at retrieving relevant patents that are difficult to retrieve.

4.2. Identity of patent applicants

We also evaluated the affect of the identity of patent applicants of a query patent and its relevant patent. Approximately 22.2% of relevant patents in NTCIR-4ab have common applicants to its corresponding query patent and 16.5% in NTCIR-5ab.

We divided each of the four test sets into two subsets: in "SAME" the applicants of a relevant patent included the same applicants as its corresponding query patent and in "DIFF" their applicants were completely different.

The comparison of MAP by subset and method is shown in #1-3 to #4-4 of Table 1. First of all, notice that the MAP values in SAME are much higher than those in DIFF. This means that the MAP depends strongly on the identity of patent applicants. Then, in NTCIR-4a-SAME (#1-3), all of our methods were ineffective, while in NTCIR-4a-DIFF (#1-4), our methods were effective, except for "two-stage retrieval". That is, the identity of patent applicants strongly affected the effectiveness of our methods in NTCIR-4a. In NTCIR-5a/ab, however, the tendency of the results was quite different from that in NTCIR-4a/ab, except for stopword deletion.

In stopword deletion, the degree of improvement was slightly higher in DIFF than in SAME in all test sets (compare MAP values at HTC06 in #1-3 to # 4-4 of Table 1). In other methods the results did not depend on the identity of applicants.

Though we did not use applicant data in the NTCIR-5 Patent Retrieval Task, it is a useful clue for finding relevant patents.

5 Approach to passage retrieval

5.1. Passage retrieval methods

Passage retrieval is to identify relevant passages in a given relevant patent, which describe the basis of why the patent can invalidate a query invention. For passage retrieval we use an n-gram indexing method. Our passage retrieval processing consists of two main steps: (1) query term extraction and (2) scoring of the similarity between query terms and terms in each passage. In this paper we focus on only query term extraction. Passages have the following two characteristics:

(1) Each passage is so short that a small number of terms can be used in passage retrieval.
(2) Passages are semantically related to each other.

Based on (1), we chose (A) a method of using sub-strings of kanji-character terms for more flexible term-matching and (B) a method of adding terms in the most relevant claim in a relevant patent to query terms. Based on (2), we chose (C) a method of excluding common topic terms across passages.

(A) Extended n-gram extraction method

The term "加熱(heating)" and the phrase "熱を加える(apply heat)" are different expressions though their meanings are the same. Those expressions are not matched in a simple term-matching method because the terms used are different. We focused on the fact that kanji-characters are ideograms and extracted all n-grams (sub-strings) from the terms consisting of only kanji-characters. For example, from the kanji-term "加熱", three n-grams "加熱", "加", and "熱" were extracted and used in term-matching. In term weighting, we chose a method of using only DF as described in Section 2 (1) (that is, TF of all n-grams was set to 1).

(B) Query expansion using claims in a relevant patent

In approach (A), it is impossible for our system to identify two terms with completely different strings, such as "モニター(monitor)" and "ディスプレイ(display)". Thus, we hypothesized that a relevant passage is related to claim texts in a relevant patent. We extracted terms from the most relevant claim identified by the system in a given relevant patent and added them to the query terms.

(C) Term exclusion considering claim structure

To identify relevant passages, the terms describing common topics across passages should be excluded before term-matching. Thus, we identified terms used in the "invention target description part" which is often located at the end of a claim text and excluded them from a query term set.

5.2. Results and discussion

The results of evaluating our methods using NTCIR-5 formal run data are shown in Table 2. In NTCIR-5, two evaluation measures are used: expected search length and MAP.

(A) Effectiveness of extended n-gram extraction

For this method, the expected search length was improved (12.1 -> 11.7) while MAP values were worse.

(B) Effectiveness of query expansion using claims

For this method, the expected search length was worse while MAP values were improved. One of the reasons for the bad expected search length is that the relevant passage was selected from the embodiment. The terms in claims are more abstract than those in embodiments, which results in term mismatching between them. More flexible matching of terms in claims and embodiments is one topic for future work.

Table 2. Passage retrieval methods and their accuracy.

Experiment ID			1	2	3	4	5
Method	(A) All n-grams extraction			used	used	used	used
	(B) Use terms in claims				used		used
	(C)Exclusion of common terms					used	used
Result	Expected search length		12.1	11.7	13.2	12.6	12.6
	MAP	x=a y=ab	.526	.504	.544	.485	.540
		x=a y=ab	.482	.476	.504	.456	.502
		x=ab y=a	.512	.497	.555	.472	.553
		x=ab y=ab	.479	.474	.525	.457	.521

x=a: relevant patent
x=b: partially relevant patent
y=a: relevant patent consisting of a single passage
y=b: relevant patent consisting of two or more passages

(C) Effectiveness of term exclusion considering claim structure

This method makes both measures worse. One reason is that term exclusion makes the number of terms very small in some queries. We should discuss how to identify common terms across passages more precisely.

6 Conclusion

We evaluated our patent retrieval methods by using the NTCIR-4 and NTCIR-5 test sets. We also discussed the affect of the ease of retrieval and the identity of patent applicants on retrieval accuracy. We found that our methods were affected by identity of applicants in NTCIR-4 test sets and strongly affected by the ease of retrieval in NTCIR-5 test sets. Though the retrieval accuracy was not stable and depended on test sets overall, our methods were effective at retrieving relevant patents that are difficult to retrieve.

In future work, we should analyze the results more deeply to find the tendency of patent retrieval. We should use other evaluation measures to evaluate the effectiveness of retrieval methods, such as the rate of the number of ranked-up/down relevant patents, to evaluate more completely. We should also consider technical fields. The text analysis and document retrieval algorithm should be adjusted depending on the technical field of the query invention.

References

[1] H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama and T. Oshio: Two-Stage Patent Retrieval Method Consider-

ing Claim Structure, Working Notes of the Fourth NTCIR Workshop Meeting, pp. 256-261, June, 2004.

[2] H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama and T. Oshio: Proposal of Two-Stage Patent Retrieval Method Considering Claim Structure, ACM-TALIP, 2006 (to appear).