NTCIR-6 Patent Retrieval Experiments at Hitachi

Hisao Mase Makoto Iwayama
Hitachi, Ltd.
292 Yoshida-cho, Totsuka-ku, Yokohama, Kanagawa, 244-0817, Japan
hisao.mase.qw@hitachi.com
makoto.iwayama.nw@hitachi.com

Abstract

Our goal in NTCIR-6 was to find the accuracy limitations of our current patent retrieval techniques. Thus, we focused only on optional runs, in which not only the claim text used in mandatory runs, but also other texts in a query patent can be used as the input data. We applied six retrieval methods step by step: (1) TF-IDF-based term weighting using TF in a whole query patent text, (2) adding terms extracted from an abstract to the query terms, (3) adding terms with higher weights extracted from a whole patent text, (4) term weight tuning based on the term co-occurrence, (5) document filtering using theme codes, one of the patent classifications, and (6) similarity score tuning using theme codes. Although these methods are simple, we found they are powerful enough to dramatically improve both recall and precision in comparison with a baseline method.

Keywords: *Patent Retrieval, Term Extraction, Term Weighting, Term Co-occurrence, Document Filtering, Score Tuning.*

1 Hitachi's approach in NTCIR-6

We have participated in the NTCIR Patent Retrieval Task since NTCIR-4. In NTCIR-4, we proposed three methods: (1) a two-stage patent retrieval method, (2) term weighting without term frequency (TF), and (3) using measurement terms for term weighting [1]. In NTCIR-5, we evaluated these methods from multiple viewpoints by using even more test sets [2].

In the mandatory runs for NTCIR-4 and 5, the input text was only a key claim text. Our interest in NTCIR-6 (Japanese Retrieval Subtask) was, however, to find the accuracy limitations of our current patent retrieval techniques. Thus, we focused on optional runs, in which not only the claim text used in mandatory runs, but also other texts from a patent can be used as the input data. Since often the whole patent text is used as the input data for the real task of patent retrieval in Patent Offices and/or private companies, this approach is worth trying.

2 Patent retrieval methods in NTCIR-6

In NTCIR-6, we used a whole query patent text as the input data. We applied the following six methods for query term extraction and term weighting:

(1) Term weighting using TF in the whole query patent

In this method, query terms were extracted only from the key claim used in mandatory runs. In the query term weight calculation based on the TF-IDF method, however, TF was calculated using the whole query patent text. Though the key claim has many important terms, the more frequently used terms in the key claim are not always important, because a key claim text is much shorter than the whole patent text.

(2) Adding terms from an abstract

In an abstract, many of the terms characterizing the invention of a query patent are included. Thus, in this method, all the terms extracted from an abstract were added to the query terms extracted from a key claim.

(3) Adding terms from a whole patent text

In some patent texts, important terms appear not only in the key claim and the abstract but also in the body of the text. Thus, in this method, the terms extracted from a whole query patent text were added to the query terms. However, many noise terms are included, because a patent text is very long. In addition, using too many query terms makes the retrieval speed much slower. Thus, we added only the top 30 terms with higher weights to the query terms.

(4) Term weight tuning based on term co-occurrence

We used terms extracted from three areas of a patent text for retrieval: the key claim, the abstract, and the whole patent text. Some of the terms appear only in one area and some in all areas. We had the hypothesis that a term extracted from more areas was more important. Thus, in this method, the weight of a term extracted from only one or two areas was reduced.

(5) Document filtering using patent classification

Patent classification is a powerful bit of information for filtering out the unwanted patents from a retrieved patent set. In this method, we used "theme code", upper classification of F-term, consisting of approximately 2,600 categories. We compared the themes assigned to a query patent with those assigned to each of the retrieved patents and picked out only the patents with one or more common themes. According to our preliminary research, 87% of patents that invalidate the inventions in query patents shared at least one theme code with their corresponding query patents (the other 13% are filtered out though they should not be).

(6) Similarity score tuning using patent classification

This method also used theme codes. If a retrieved patent had one or more common theme codes with those of its corresponding query, the similarity score of the retrieved patent was increased. Since this method does not filter out any retrieved patents, the 13% of retrieved patents mentioned above can be kept in the high retrieval ranks.

Note that the methods (1)-(4) can be combined and that the methods (5) and (6) are exclusive to each other.

3 Experiments

3.1. Data

We used test sets in the NTCIR-6 Patent Retrieval Task to evaluate the effectiveness of our patent retrieval methods [3]. The relevant patents for a query 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 seven kinds of test sets: NTCIR-4a (31 queries, 158 relevant patents), NTCIR4-ab (34, 342), NTCIR-5a (619, 619), NTCIR5-ab (1189, 2065), SR-a (189, 189), SR-ab (349, 810) and NTCIR6-ab (1685, 9871). Note that the symbols ‘a’ and ‘b’ in the test set “NTCIR6-ab” have the same meaning as those in “NTCIR4-ab” and “NTCIR5-ab”, although in NTCIR-6 they are officially used to show the degree of consistency of the IPC Subclass between a query patent and retrieved patents, because all queries in the NTCIR-6 test set have only partially relevant patents. Also note that there are no queries corresponding to NTCIR-6a. The top 1000 retrieved patents for each query were output automatically as the retrieval result. The retrieval target patent set consisted of approximately 3.5 million patents issued from 1993 to 2002.

3.2. Evaluation metrics

We used “Mean Average Precision (MAP)” for the metrics of the evaluation [3]. We also compared the number of correctly identified patents from the top 300 retrieved patents. This metric was used to evaluate the recall (it is said a patent examiner reads approximately 300 retrieved patent documents per query patent).

3.3. Experiment patterns

As shown in Table 1, we used ten experiment patterns (HTC01-HTC10) combined with ten retrieval methods (#1-#10). HTC01 was a baseline method. We used GETA[1] as the search engine. HTC02 and HTC03 were the mandatory runs. These patterns used the methods we had proposed in NTCIR-4 and 5 [1][2]. HTC02 added the two-stage patent retrieval method and query term weighting using only DF to HTC01. HTC03 added the stopword deletion and the method using the measurement terms to HTC02.

HTC04 through HTC10 were the optional runs. These patterns used a step by step addition of the six methods described in Section 2. HTC04 added the query term weighting using TF in a whole query patent text ((1) in Section 2) to HTC01. HTC05 applied the adding of terms from an abstract method ((2) in Section 2) to HTC04. HTC06 applied the adding 30 terms from a whole patent text method ((3) in Section 2) to HTC05. HTC07 added the query term weight tuning based on term co-occurrence method ((4) in Section 2) to HTC06. HTC08 added the stopword deletion and the method using measurement terms methods to HTC07. HTC09 added the document filtering using theme codes method ((5) in Section 2) to HTC08. HTC10 added the similarity score tuning using theme codes method to HTC08.

3.4. Results and discussion

The result of each experiment pattern is shown in Tables 2 - 5.

Table 2 shows the MAPs by experiment pattern in each test set. The MAPs in almost all the patterns improved in comparison to those in the baseline (HTC01). Table 3 shows the relative values of the MAPs when the MAP in HTC01 is equal to 1. Our six retrieval methods are very effective in improving the precision. As for HTC10, the pattern with the highest MAPs, the MAPs improved by a maximum of 47% (NTCIR-6ab) in comparison to HTC01. Also, we noted that the values in the mandatory runs (HTC02 and HTC03) were not higher than those in the optional runs (HTC04-HTC10). This is because there was much less information extracted from a key claim text in the mandatory runs than in the optional runs. This shows that the information from a key claim is not enough to retrieve similar patents with higher precision.

Table 4 shows the number of correctly identified patents from the top 300 retrieved patents using the experimental pattern in each test set. Table 5 shows the relative values of the number of correctly

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

Table 1. Experimental patterns.

#	Retrieval methods	Experiment patterns									
		HTC01	HTC02	HTC03	HTC04	HTC05	HTC06	HTC07	HTC08	HTC09	HTC10
1	Two-stage patent retrieval	—	applied	applied	—	—	—	—	—	—	—
2	Query term weighting using only DF	—	applied	applied	—	—	—	—	—	—	—
3	Stopword deletion	—	—	applied	—	—	—	—	applied	applied	applied
4	Query term weighting using measurement terms	—	—	applied	—	—	—	—	applied	applied	applied
5	Query term weighting using TF in a whole query patent	—	—	—	applied	applied	applied	applied	applied	applied	applied
6	Adding terms from an abstract	—	—	—	—	applied	applied	applied	applied	applied	applied
7	Adding 30 terms from a whole query patent	—	—	—	—	—	applied	applied	applied	applied	applied
8	Query term weighting based on term co-occurrence	—	—	—	—	—	—	applied	applied	applied	applied
9	Document filtering using theme codes	—	—	—	—	—	—	—	—	applied	—
10	Similarity score tuning using theme codes	—	—	—	—	—	—	—	—	—	applied

Table 2. MAPs in each test set.

#	Test set name (# of queries, # of correct patents)	Experiment patterns									
		HTC01	HTC02	HTC03	HTC04	HTC05	HTC06	HTC07	HTC08	HTC09	HTC10
1	NTCIR-6ab (1685, 9871)	0.0783	0.0812	0.0815	0.0843	0.0953	0.1041	0.1071	0.1077	0.1134	0.1151
2	SR-a (189, 189)	0.1510	0.1620	0.1625	0.1591	0.1612	0.1674	0.1719	0.1680	0.1832	0.1825
3	SR-ab (349, 810)	0.1276	0.1405	0.1443	0.1438	0.1599	0.1663	0.1744	0.1717	0.1834	0.1845
4	NTCIR-5a (619, 619)	0.1852	0.1967	0.1917	0.1986	0.2011	0.2165	0.2224	0.2243	0.2391	0.2391
5	NTCIR-5ab (1189, 2065)	0.1503	0.1578	0.1548	0.1630	0.1660	0.1795	0.1847	0.1856	0.1967	0.1973
6	NTCIR-4a (34, 158)	0.2779	0.2993	0.2973	0.3409	0.3014	0.2422	0.2856	0.2812	0.2828	0.2834
7	NTCIR-4ab (34, 342)	0.2123	0.2409	0.2411	0.2596	0.2418	0.2266	0.2485	0.2531	0.2627	0.2636

Table 3. Relative values of MAPs in each test set (HTC01=1).

#	Test set name (# of queries, # of correct patents)	Experiment patterns									
		HTC01	HTC02	HTC03	HTC04	HTC05	HTC06	HTC07	HTC08	HTC09	HTC10
1	NTCIR-6ab (1685, 9871)	1	1.04	1.04	1.08	1.22	1.33	1.37	1.38	1.45	1.47
2	SR-a (189, 189)	1	1.07	1.08	1.05	1.07	1.11	1.14	1.11	1.21	1.21
3	SR-ab (349, 810)	1	1.10	1.13	1.13	1.25	1.30	1.37	1.35	1.44	1.45
4	NTCIR-5a (619, 619)	1	1.06	1.03	1.07	1.09	1.17	1.20	1.21	1.29	1.29
5	NTCIR-5ab (1189, 2065)	1	1.05	1.03	1.08	1.10	1.19	1.23	1.23	1.31	1.31
6	NTCIR-4a (34, 158)	1	1.08	1.07	1.23	1.08	0.87	1.03	1.01	1.02	1.02
7	NTCIR-4ab (34, 342)	1	1.14	1.14	1.22	1.14	1.07	1.17	1.19	1.24	1.24

Table 4. Number of correctly identified patents from the top 300 retrieved patents.

#	Test set name (# of queries, # of correct patents)	Experiment patterns									
		HTC01	HTC02	HTC03	HTC04	HTC05	HTC06	HTC07	HTC08	HTC09	HTC10
1	NTCIR-6ab (1685, 9871)	4525	4538	4502	4731	5044	5397	5524	5527	5814	5916
2	SR-a (189, 189)	109	113	113	118	118	125	134	136	140	143
3	SR-ab (349, 810)	450	449	442	479	486	528	545	547	577	591
4	NTCIR-5a (619, 619)	409	404	401	414	430	455	463	463	478	493
5	NTCIR-5ab (1189, 2065)	1201	1191	1193	1237	1287	1360	1383	1387	1444	1484
6	NTCIR-4a (34, 158)	107	108	111	112	118	124	125	125	121	124
7	NTCIR-4ab (34, 342)	229	230	238	245	261	259	272	272	267	274

Table 5. Relative values of number of correctly identified patents from top 300 retrieved patents (HTC01= 1).

#	Test set name (# of queries, # of correct patents)	Experiment patterns									
		HTC01	HTC02	HTC03	HTC04	HTC05	HTC06	HTC07	HTC08	HTC09	HTC10
1	NTCIR-6ab (1685, 9871)	1	1.003	0.995	1.046	1.115	1.193	1.221	1.221	1.285	1.307
2	SR-a (189, 189)	1	1.037	1.037	1.083	1.083	1.147	1.229	1.248	1.284	1.312
3	SR-ab (349, 810)	1	0.998	0.982	1.064	1.080	1.173	1.211	1.216	1.282	1.313
4	NTCIR-5a (619, 619)	1	0.988	0.980	1.012	1.051	1.112	1.132	1.132	1.169	1.205
5	NTCIR-5ab (1189, 2065)	1	0.992	0.993	1.030	1.072	1.132	1.152	1.155	1.202	1.236
6	NTCIR-4a (34, 158)	1	1.009	1.037	1.047	1.103	1.159	1.168	1.168	1.131	1.159
7	NTCIR-4ab (34, 342)	1	1.004	1.039	1.070	1.140	1.131	1.188	1.188	1.166	1.197

Table 6. Rate of query patents improving retrieval accuracy by applying proposed methods. (1685 queries in the NTCIR-6 test set)

Evaluation metrics		Experiment patterns								
		HTC02	HTC03	HTC04	HTC05	HTC06	HTC07	HTC08	HTC09	HTC10
# of relevant patents in top 300 retrieved patents	better	17.2%	18.3%	18.6%	33.3%	43.1%	44.6%	44.9%	51.2%	52.2%
	worse	16.4%	19.4%	9.2%	17.2%	16.5%	11.1%	11.1%	11.6%	8.8%
	no change	66.4%	62.3%	72.2%	49.6%	40.4%	44.3%	44.0%	37.3%	38.9%
Average Precision	better	51.0%	50.5%	56.4%	63.1%	66.6%	70.9%	70.6%	75.8%	76.8%
	worse	43.9%	44.4%	38.5%	34.1%	31.8%	27.4%	27.8%	23.3%	22.5%
	no change	5.1%	5.1%	5.0%	2.8%	1.7%	1.8%	1.6%	0.9%	0.7%

identified patents. The tendency of the results is very similar to that in MAP: almost all the methods were good enough to increase the number of correctly identified patents. As for HTC10, the number of correctly identified patents increased by approximately 30% (NTCIR-6ab).

The tendency of the results in NTCIR-6ab, SR-a, SR-ab, NTCIR-5a and NTCIR-5ab are very similar, while those in NTCIR-4a and NTCIR-4ab are different. In the NTCIR-4 test sets, HTC04 was very effective but HTC05 and HTC06 were not, while in other test sets, all the methods were effective. In the NTCIR-4 test sets, the correctly identified patents were collected by the "pooling method". As a result, the average number of correct documents per query is much greater than the other test sets. Furthermore, the technical fields covered in the NTCIR-4 test sets are relatively narrow. These factors might be the cause of the tendency differences.

Table 6 shows the rate of query patents improving retrieval accuracy by applying proposed methods. We used a test set of NTCIR-6 only. In HTC02 and HTC03 in which only a claim text is used as input, the number of query patents improving the value of evaluation metrics is almost the same as that making the value worse. However, in HTC04-HTC10 in which the whole patent text and patent classification are used, the number of query patents improving the value of evaluation metrics is much higher than that making the value worse. The results in Table 6 shows that proposed methods are effective to improve retrieval accuracy.

4 Conclusion

We proposed and evaluated six patent retrieval methods using a whole patent text as the input. These methods were simple but powerful for improving both the recall and precision. This result shows that a key claim text alone is not sufficient enough input to collect similar patents with a higher accuracy. When our methods are applied in actual retrieval systems, however, we need to consider the processing speed, because our methods use a lot more terms for the retrieval, which could be the cause of the slower retrieval processing.

In future work, we will analyze the results in more detail to find the patent retrieval tendency. We need to tune the term weighting using the term co-occurrence. We should use other patent tags to identify the important terms. We should also consider technical fields. The text analysis and 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 Considering Claim Structure, Working Notes of the Fourth NTCIR Workshop Meeting, pp. 256-261 (2004).

[2] H. Mase, T. Matsubayashi, Y. Ogawa, T. Yayoi, Y. Sato, and M. Iwayama: NTCIR5 Patent Retrieval Experiments at Hitachi, NTCIR Workshop 5 Meeting, pp. 318-323 (2005).

[3] A. Fujii, M. Iwayama, and N. Kando: Overview of the Patent Retrieval Task at the NTCIR-6 Workshop, Proceedings of the Sixth NTCIR Workshop Meeting (2007).