NTCIR-4 Patent Retrieval Experiments at RICOH

Hideo ITOH
Software R&D group, RICOH Co., Ltd.
1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, JAPAN
hideo@src.ricoh.co.jp

Abstract

Focusing on the document structure of patent, two search databases AC and WH were built and compared in retrieval performance, where AC includes only abstract and claim sections in patent and WH includes the whole patent texts. Moreover, we attempted to combine search results for the two databases to improve retrieval performance. Another point of our experiments is cross-lingual patent retrieval using a large and high-quality parallel corpus. The query submitted against the parallel database was partially translated and expanded using the same mechanism for pseudo-relevance feedback.

Keywords: *NTCIR, patent retrieval, document structure, cross-lingual retrieval*

1 Introduction

In our experiments, we focused on the following two points. One is the document structure of patent. A patent usually consists of several sections such as abstract, claims and so on. Two search databases AC and WH were built and compared in retrieval performance, where AC includes only abstract and claim sections and WH includes the whole patent texts. Assuming that these two databases are complementary to each other in precision and recall, we attempted to combine search results for each database.

Another point is cross-lingual patent retrieval using parallel corpus. The parallel corpus was constructed with PAJ data which is a set of English abstracts of each patent. The English query submitted against the parallel database was partially translated and expanded using the same mechanism for pseudo-relevance feedback.

We submitted four runs LAPIN1–LAPIN4 for mono-lingual retrieval, and LAPIN5 for cross-lingual retrieval. All of the runs were produced in full automatic manner. In search topic, we used CLAIM and FDATE fields but COMP tags were not used.

2 System Description

In this section, we detail the process of claim-to-patent retrieval. The framework is the same as that of NTCIR-3 [1] and NTCIR-2 [2].

2.1 Indexing

As a retrieval target, the Japanese patents published in 1993-1997 were automatically indexed to build a search database, where the indexing unit is a character n-gram and the index data structure is inverted file.

A Japanese patent consists of several sections such as abstract, claims and detailed description. Since each section is annotated with SGML tags, we can automatically extract a specific section for indexing. Two search databases were built which are denoted by AC and WH in this paper. Abstract and claim sections in each patent were stored in the database AC. On the other hand, the whole patent texts were stored in the database WH.

Apart from the search database, we recorded in RDB the published date of each patent. The date was identified using INID 43 code in the patent text.

2.2 Query Processing

For each search topic, a claim part was automatically extracted using CLAIM tag and used as a query string. The COMP tags were not used and eliminated from the extracted query string. The query string above mentioned was fed to our search engine. Using a Japanese morphological analyzer with a function of word form normalization, the search engine divides an input query string into words. After eliminating invalid words using a stopword dictionary, query terms are extracted from a sequence of words by pattern matching against both word form and part-of-speech tag. All of the extracted query terms were used for the retrieval, namely any term selection was not performed. We used phrasal terms (word bi-grams) in addition to single terms.

2.3 Document Retrieval

In the search engine, each query term is submitted to the ranking search module, which calculates a relevance score of the documents including the term.

The relevance score of the document d for the term t is defined by the following formula, which is based on the OKAPI/BM25 [4] with modified term weighting formula [6].

$$score_{d,t} = \frac{tf_{d,t}}{tf_{d,t}+k_1((1-b)+b\frac{l_d}{l_{ave}})} \cdot weight_t$$

$$weight_t = \log(k_4 \cdot \frac{N}{n_t} + 1) / \log(k_4 \cdot N + 1)$$

where N is the number of documents in the target collection, n_t is the document frequency of the term t, $f_{d,t}$ is the within-document frequency of the term t in the document d, l_d is the document length and l_{ave} is the average document length.

In the above formulae, k_1, k_4, b are tuning parameters and we set the values of the parameters through a preliminary experiments using Search Report Data (2001, 2002, 2003) which was provided by the task organizer. This is a collection of the search reports prepared by professional patent search intermediaries, and the reports were used by patent examiners at the Japanese Patent Office as reference data for patent examination.

Retrieved patents were ranked on the sum of the score and the patents published after the search topic was filed were eliminated. This elimination was performed using the FDATE of the search topic and the INID 34 code of the patent. After the elimination, the top-1000 patents in the ranking were submitted for the official run.

2.4 Combining Search Results

Through a preliminary experiments using the Search Report Data, we observed that the search results for the database AC and WH were complementary to each other in precision and recall. Roughly speaking, it seemed that WH induced a high-recall search and AC gave high-precision at the middle of recall point.

To exploit the characteristics, we combined the ranking lists of AC and WH. More specifically, we assigned a combination score to each patent in the search results and re-ranked the patents on the new score. Before calculating combination scores, we normalized each score in the ranking list of AC and WH to take a value between 0 and 1. The combination $score_d$ of the document d was given by linear combination as follows :

$$score_d = (1-\lambda) \cdot score_d(AC) + \lambda \cdot score_d(WH)$$

where $score_d(AC)$ and $score_d(WH)$ are the normalized scores in the ranking list of AC and WH respectively, and λ is a tuning parameter. In the preliminary experiments, the combining method improved the mean average precision by 8.4 % and the precision at top-30 by 10.0 %.

2.5 Cross-lingual Retrieval

We performed English-to-Japanese patent retrieval using a parallel corpus. In the cross-lingual retrieval process, the English query is submitted against the English database and top-n documents are obtained. Their counterparts in the Japanese database are exploited as seed documents to extract Japanese query terms. The extraction can be performed using completely same mechanism for query expansion in pseudo-relevance feedback [5].

This well-known strategy seems to be promising in the patent retrieval, because we can build a large and high-quality parallel corpus using PAJ data, which is a set of English abstracts translated by human experts.

The PAJ data corresponding to the retrieval target (patents in 1993-1997) was used for construction of the English database. The indexing unit was a word and stemming was performed. For the Japanese-side database, we used the database AC previously mentioned.

The English query string was extracted using CLAIM tags in the English search topic file. The number of seed documents was set to ten. The number of Japanese query terms selected on Robertson's Selection Values [3] was fixed to twelve. No phrases were used for both English and Japanese terms. The whole process was performed automatically using the same search engine for mono-lingual retrieval explained in the previous sub-sections.

3 Results

Table 1 shows the evaluation results for the monolingual (Japanese-to-Japanese) retrieval. The table includes results from post-submission experiments which have no RunIDs. LAPIN3 differs from LAPIN4 in the value of the linear combination parameter λ. The column "DB" denotes the retrieval target database of the run. The column "Phrase" shows whether phrasal terms were used (yes) or not (no). The column "Expansion" indicates whether pseudo-relevance feedback were performed or not.

34 main topics								
					Mean Average Precision			
ID	RunID	DB	Phrase	Expansion	A	A+B	C(A)	C(A+B)
M1	LAPIN1	AC	yes	no	0.1702	0.1861	0.1644	0.1617
M2	-	AC	yes	yes	0.1990	0.2091	0.1854	0.1962
M3	-	AC	no	no	0.1831	0.1903	0.1617	0.1680
M4	LAPIN2	WH	yes	no	0.2152	0.2174	0.1709	0.1600
M5	-	WH	yes	yes	0.2226	0.2190	0.2102	0.1942
M6	-	WH	no	no	0.2241	0.2062	0.2001	0.1826
M7	LAPIN3	AC+WH	yes	no	0.1996	0.2180	0.1755	0.1745
M8	LAPIN4	AC+WH	yes	no	0.1980	0.2159	0.1661	0.1661

67 additional topics								
					Average Precision			
ID	RunID	DB	Phrase	Expansion	A	A+B	C(A)	C(A+B)
A1	LAPIN1	AC	yes	no	–	–	0.0794	0.0813
A2	-	AC	yes	yes	–	–	0.0616	0.0604
A3	-	AC	no	no	–	–	0.0808	0.0800
A4	LAPIN2	WH	yes	no	–	–	0.1284	0.1265
A5	-	WH	yes	yes	–	–	0.0884	0.0868
A6	-	WH	no	no	–	–	0.1312	0.1287
A7	LAPIN3	AC+WH	yes	no	–	–	0.1054	0.1044
A8	LAPIN4	AC+WH	yes	no	–	–	0.0987	0.0998

Table 1. Evaluation results for mono-lingual retrieval

34 main topics						
			Query Translation		Average Precision	
ID	RunID	nterm	Precision	Recall	A+B	P@10
EJ1	–	6	0.667	0.213	0.1199	0.1265
EJ2	LAPIN5	12	0.522	0.333	0.1656	0.1588
EJ3	–	18	0.376	0.360	0.1852	0.1853
EJ4	–	24	0.308	0.393	0.1883	0.1941
EJ5	–	30	0.274	0.437	0.1885	0.2000
EJ6	–	36	0.241	0.462	0.1903	0.1882
EJ7	–	42	0.215	0.480	0.1890	0.1941
EJ8	–	c12	1.000	0.333	0.1565	0.1676
EJ9	–	c36	1.000	0.462	0.1813	0.1941
EE	–	–	–	–	0.1449	0.1676
JJ	–	–	–	–	0.1903	0.2206

Table 2. Evaluation results for cross-lingual retrieval

The column "A" and "A+B" of Table 1 shows the mean average precision measured with a set of relevant documents judged as A and either A or B respectively. "C(A)" and "C(A+B)" correspond to the measurement with relevant documents cited by the Japanese Patent Office.

Since, for the 67 additional topics, the relevant documents were provided by only the citation, the values of the column "A" and "A+B" are not available (denoted by '–').

Table 2 shows the evaluation results for the cross-lingual (English-to-Japanese) retrieval. The table includes results from post-submission experiments which have no RunIDs.

The mono-lingual retrieval run EE and JJ are included in the table in order to compare with cross-lingual cases EJ1–EJ9. The run EE corresponds to the first stage of our cross-lingual retrieval (English-to-English retrieval). The run JJ corresponds to the run M3 in Table 1. Because phrasal terms were not used in the second stage of our cross-lingual retrieval, we selected M3 (with no phrasal terms and no query expansion) for comparison.

The column "nterm" denotes the number of Japanese terms extracted and used for each retrieval. The column "A+B" denotes the mean average precision measured with a set of relevant documents judged as either A or B. The column "P@10" shows the retrieval precision at top-10.

The column "precision" and "recall" indicate the performance of query translation and are defined as follows:

$$precision = \frac{n}{N} \qquad recall = \frac{n}{M}$$

where n is the number of Japanese terms "correctly" extracted. The correctness is judged by whether the term occurs in the original Japanese query or not. N is the number of Japanese terms extracted. M is the number of Japanese terms in the original Japanese query.

In the run EJ8 and EJ9, the cross-lingual retrieval was performed using only Japanese terms correctly extracted. "c12" and "c36" in the column "nterm" mean the number of Japanese terms extracted for each retrieval are 12 and 36 respectively.

4 Analysis

- Comparison between AC and WH
  Comparing LAPIN2 with LAPIN1 in Table 1, we can conclude that the database WH gives better performance for claim-to-patent retrieval than AC.

- Effects of combining search results
  In the evaluation using C(A+B), LAPIN3 gives better performance for the main topics than LAPIN2. In this case, the mean average precision is improved by 9 %. This observation is consistent with the result of the preliminary experiments using the Search Report Data. However this strategy hurt retrieval performance in the other cases.

- Cross-lingual retrieval
  In LAPIN5, the precision of the query translation was 52.2 % and the rest of terms consists of related terms newly introduced from seed documents. Among EJ runs (EJ1–EJ7), the best mean average precision was given by EJ6 which has lower precision of query translation than LAPIN5. This suggests that the related-terms had a good effect on the performance. It can be ascertained by comparing the performance of EJ6 and EJ9. Since the performance of EJ6 is comparable with that of JJ and is better than that of EE, the cross-lingual retrieval using PAJ data can be regarded as promising.

- Effects of query expansion and phrasal term
  Comparing M1 with M2 and M4 with M5 respectively, query expansion by pseudo-relevance feedback has a positive effect for main topics consistently. On the contrary, for additional topics, query expansion seems to hurt the performance at least in the evaluation using only citation. The effectiveness of phrasal term is unstable, comparing M1, M4, A1, A4 with M3, M6, A3, A6 respectively.

5 Conclusions

In claim-to-patent retrieval, we conclude the database WH gives better performance than AC. The effects of combining ranking lists from AC and WH is unstable. The English-to-Japanese patent retrieval using PAJ data is promising. More experiments are needed as well as careful observation on the effect of phrasal term and query expansion.

References

[1] H.Itoh, H.Mano, and Y. Ogawa. Term distillation for cross–db retrieval. *Proc. of NTCIR Workshop 3 Meeting*, 2003.

[2] Y. Ogawa and H. Mano. RICOH at NTCIR-2. *Proc. of NTCIR Workshop 2 Meeting*, pages 121–123, 2001.

[3] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[4] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *Proc. of 17th ACM SIGIR Conf.*, pages 232–241, 1994.

[5] S. E. Robertson and S. Walker. On relevance weights with little relevance information. *Proc. of 20th ACM SIGIR Conf.*, pages 16–24, 1997.

[6] M. N. Y. Ogawa, H. Mano and S. Honma. Structuring and expanding queries in the probabilistic model. *The Eighth Text REtrieval Conference (TREC-8)*, pages 541–548, 2000.