Invalidity Search for USPTO Patent Documents Using Different Patent Surrogates

Yuen-Hsien Tseng, Chen-Yang Tsai*, and Da-Wei Juang**
National Taiwan Normal University, Taipei, Taiwan, R.O.C., 106
samtseng@ntnu.edu.tw
*Fu Jen Catholic University, Taipei, Taiwan, R.O.C., 242
otto4321a@yahoo.com.tw
**WebGenie Information LTD., Taipei, Taiwan, R.O.C., 231
david@webgenie.com.tw

Abstract

This paper describes our work at the sixth NTCIR workshop on the subtask of invalidity search for patent retrieval. We compared different patent surrogates for their effectiveness on invalidity search. Our preliminary results show that the query by the Claims field plus PRF (pseudo relevance feedback) leads to the best results in terms of relevance degree A while the query by all free-text fields yields highest performance under relevance degree B.

Keywords:
Probabilistic retrieval model, relevance feedback, effectiveness, patent surrogates.

1. Introduction

Patent documents contain important research results that are valuable to the industry, business, law, and policy-making communities. The claims in the patent documents protect the intellectual property rights of the inventors and the assignees. A violation of other patent's claims can lead to the invalidity of the applicant's or a lawsuit involving large expenditure of compensation. Therefore, it is important to check the validity of the claims of a patent application for patent examiners. It is also an inevitable task to check the invalidity of existing patents for those whose are involved in a patent-related lawsuit.

From the statistics of WIPO (World Intellectual Property Organization), there were over 1 million patents world-wide by 1995. In 2001, this number had increased to over 1.5 million [1]. The large volume and the increasing speed of patent documents have made the validity, or equivalently invalidity, checking a challenging task for humans and machines alike.

2. The Invalidity Search Task

Since NTCIR-4, the invalidity search, searching existed patents to invalidate the claims of a specified patent or patent application, has been one of the focuses of the patent retrieval task. This year we participated in this subtask for English patents to learn more lessons in dealing with patent documents based on our past experiences [2-4].

Like the other information retrieval evaluation tasks, the English invalidity search subtask [5] consists of a collection of patents documents, a set of query topics, and the relevance judgment files or guidelines for each pair of search topic and target document. The document collection consists of 981,948 patents from USPTO [6] ranging from 1993 to 2000. The queries include 2,221 search topics, each come from a patent‘s claims. Besides the claims, its corresponding patent information such as inventors, title, abstract, etc, can be used in the query, as long as this information does not directly lead to the answers (relevant target patents) of the query. The answers for a search topic are assumed to be those citations provided by the applicant in the search topic. This choice decided by the task organizers seems to aim at reducing the difficulty of invalidity checking. Under this relevance guideline, it can be seen that if a topic patent and its relevant document are assigned to the same IPC (International Patent Classification) class, the document can be retrieved with less difficulty. Thus, the degree of the relevance of each citation is classified into two ranks: If the IPC subclasses assigned to the topic patent and the target document are identical, the relevance is labeled B; if they are not identical, it is labeled A.

Participants are expected to return a ranked list of retrieved patents for each topic. Mean average precision (MAP) reported by the trec_eval program [7] is used as one of the evaluation measures.

3. System Description

Since this is basically a patent-to-patent retrieval task, to be fair, only the patent identifiers and free-text fields tagged by <TITLE>, <ABST>, <SPEC>, and <CLAIM> can be used for indexing. Information from the other fields is not allowed in the mandatory runs. Our system indexed all these free-text fields for the 981,948 patents (37.5 GB). We used a list of stop words to filter non-semantic bearing words and a key-phrase extraction algorithm [8] to extract frequent multi-words phrases. All the single words, extracted multi-word phrases, and bi-words (like bi-grams, any consecutive word pairs) stemmed by the Porter's algorithm are indexed. The bi-grams are helpful for arbitrary phrases searching (and thus are used in our system for Chinese and English text retrieval). However, they lead to far more indexing terms than a cheap PC can handle due to the large text size a normal patent document can have, especially in its <SPEC> field where full details of the invention are mentioned. Therefore, we removed the texts in the <SPEC> tag in our final indexing process. In the end, the index took 369 minutes to build and contained over nine million index terms.

To compare the query document with the indexed documents, we use the BM25 probabilistic retrieval model, which can be described in the following formula:

$$BM25(d_i, q_j) = \sum_{k=1}^{T} \frac{(k_3+1)tf_{j,k}}{k_3 + tf_{j,k}} \log\left(\frac{n - df_k + 0.5}{df_k + 0.5}\right)\left(\frac{(k_1+1)tf_{i,k}}{tf_{i,k} + k_1((1-b) + b\frac{dl_i}{Avgdl})}\right)$$

where $tf_{j,k}$ is the term frequency of index term k in document j, df_k is the document frequency of term k, dl_i is the length of document i, $avgdl$ is the average document length of all indexed documents, and k_1, k_3 and b are parameters. They are set as k_1=1.2, k_3=1000, and b=0.75 in our experiment.

Since pseudo relevance feedback (PRF) is shown to be a robust and effective way to improve initial retrieval [9-10], some of our experiments use this technique. Specifically, fifteen best terms from six top-ranked documents retrieved by the initial query were used. These six documents were first concatenated into one text string and then the keyword extraction algorithm [8] was applied to extract maximally repeated patterns. The extracted patterns were filtered by some stop words and then sorted in decreasing order of occurrence. The first 15 terms were added to the initial query for the second run of document retrieval.

4. Experiment Results

As mentioned, there are a total of 2,221 search topics. Evaluation of them all would take a lot of time. Due to the unexpected events happened to us, we eventually had very little time to complete this task before the submission due. Therefore, not all the search topics were evaluated. Also, if all topics return 1000 ranked documents, as normal NTCIR results submitted, it would take up more than 130 MB, which would cause a long wait while up-loading our result. Finally we decided to return only the first 100 documents for each topic, although we know that this would reduce the MAPs compared to those returning 1000 documents.

Table 1 shows the set of topics, the fields submitted as queries, and the MAP results (for relevance degree A and B). As noted, the MAPs are all small compared to the normal SLIR (Single Language Information Retrieval) results in the NTCIR workshops. The low performance can also be seen from Figure 1, where most topics have MAPs below 0.05.

However, this is due to the difficulty of the invalidity search (far less relevant documents per topics). Also the number of topics is far more than that of the SLIR task (2221 vs 50). Thus a smaller difference may still be significant.

Table 1. NTNU English Invalidity Search Results.

Set	Topics	Fields used	MAP.A	MAP.B
1	1001~1015	INVENTOR, TITLE, ABST, CLAIM, SPEC+PRF	0.010662	0.124486
2	1016~1530	CLAIM+PRF	0.028012	0.053675
3	1531~1930	CLAIM	0.021573	0.036696
4	2031~2326	TITLE, INVENTOR	0.020224	0.036332
5	2523~2800	TITLE	0.019477	0.046047
6	1931~2030	unfinished		
7	2327~2522	unfinished		
8	2801~3221	unfinished		

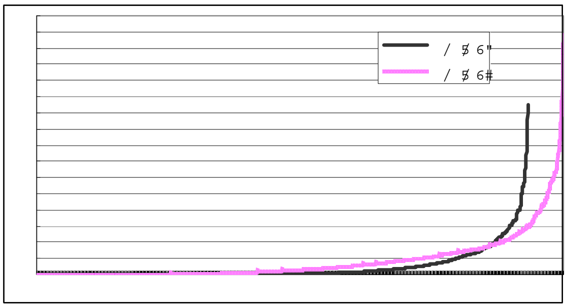

Figure 1. MAPs sorted in increasing order.

As can be seen from Table 1, the most extreme query case has the most extreme results: in Set 1 all the free-text fields plus the names of the inventors are submitted as queries. It leads to the least effectiveness for relevance degree A and highest effectiveness for B. Use of the claims as queries is only marginally better than the use of titles and inventors. However, this is inconclusive if only title is used, since title-only queries lead to slightly worse performance for relevance degree A but far better for B. The pseudo relevance feedback remains the most robust technique which leads to relatively high performance compared to the other sets of queries.

5. Conclusions

Low MAPs are observed among all the invalidity search results from all participants. Although the conclusions from these experiments may be skeptical, the results from a large set of retrieval evaluations should reveal a tendency from which we can learn and apply in the future studies.

Due to unexpected events, this year we have very little time in conducting more experiments and technique comparisons. We expect to devote more efforts in future invalidity search for patent retrieval.

Acknowledgement

This work is supported in part by NSC under the grant numbers: NSC 95-2221-E-003-016- and NSC 95-2524-S-003-012-.

References

[1] WIPO, WIPO Patent Report 2006. http://www.wipo.int/ipstats/en/statistics/patents/

[2] Yuen-Hsien Tseng, Da-Wei Juang, and Chi-Jen Lin "Automatic Categorization of Japanese Patents based on Surrogate Texts," Proceedings of the Fifth NTCIR Workshop on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Dec 6-9, 2005, Tokyo, Japan, pp. 348-354.

[3] Yuen-Hsien Tseng, Yeong-Ming Wang, Yu-I Lin, Chi-Jen Lin and Dai-Wei Juang, "Patent Surrogate Extraction and Evaluation in the Context of Patent Mapping", accepted for publication in Journal of Information Science, 2007.

[4] Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin, "Text Mining Techniques for Patent Analysis", to appear in Information Processing and Management, 2007

[5] Atsushi Fujii, Makoto Iwayama, and Noriko Kando, “Overview of the Patent Retrieval Task at the NTCIR-6 Workshop”, Proceedings of the Sixth NTCIR Workshop Meeting. 2007.

[6] United States Patent and Trademark Office, http://www.uspto.gov/.

[7] Chris Buckley, ftp://ftp.cs.cornell.edu/pub/smart/trec_eval.v3beta.shar

[8] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, Nov. 2002, pp. 1130-1138.

[9] Yuen-Hsien Tseng, Yu-Chin Tsai, and Chi-Jen Lin "Comparison of Global Term Expansion Methods for Text Retrieval," Proceedings of the Fifth NTCIR Workshop on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Dec 6-9, 2005, Tokyo, Japan, pp. 150-155.

[10] Yuen-Hsien Tseng, Chen-Yang Tsai, and Da-Wei Juang, “On the Robustness of Document Re-Ranking Techniques: A Comparison of Label Propagation, KNN, and Relevance Feedback,” Proceedings of the Sixth NTCIR Workshop on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, May 15-18, 2007, Tokyo, Japan.