NTCIR-5 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 IPC (International Patent Classification) of patent, two retrieval methods were examined. One is to use the IPC code of the query patent and the other is to exploit the codes assigned to top-N retrieved patents in a similar manner of pseudo-relevance feedback. In both methods, the codes were used as constraint on retrieval results. As a result, we found the former is clearly effective and the latter's effect is positive but small. Another point of our experiments is exploitation of synonyms, which were automatically collected from a machine-readable dictionary for each query term.

Keywords: *NTCIR, patent retrieval, IPC, synonym*

1 Introduction

The main focus of our experiments is to clarify the effects of IPC information of patents on the precision improvement in similarity patent retrieval. In fact, searchers have usually used IPC codes for patent investigation in addition to query terms. Although it is questionable whether the IPC can improve the effectiveness in similarity search in general, because the search space may be strictly restricted, in the case of invalidity search, we assumed the search space may be apt to be restricted in the technical field of target query patent. For the purpose, two retrieval methods were examined. One is to use the IPC code of the query patent and the other is to exploit the codes assigned to top-N retrieved patents in a similar manner of pseudo-relevance feedback. In both methods, the codes were used as constraint on retrieval results.

Another point of our experiments is similarity patent retrieval using synonyms. We automatically collected synonyms from a machine-readable dictionary for each query term, and used them in combination with original query terms.

We submitted four runs for "Document Retrieval Subtask" of NTCIR-5 Patent Retrieval Task. All of the runs were produced in full automatic manner. In search topics, we used CLAIM and FDATE fields.

2 System Description

In this section, we detail the process of claim-to-patent retrieval. The framework is almost same as that of NTCIR-4 [1] and NTCIR-3 [2].

2.1 Indexing

As a retrieval target, the Japanese patents published in 1993-2002 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. The whole patent text was used for indexing of each.

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 query string was fed to our search engine. For a query string, the search engine produces a sequence of normalized word forms and part-of-speech tags, using a Japanese morphological analyzer, which had been originally developed and built in the search engine. In order to eliminate so frequent and useless word forms and corresponding tags from the above mentioned sequence, we used a stop word dictionary, which had been developed at the NTCIR-4 Patent Task. The number of entries in the dictionary is about one hundred. Query terms are extracted from the resultant sequence by pattern matching against some rules on word form and tag. The rules had been manually developed depending on our part-of-speech tag system. All of the extracted query terms were used for the retrieval, in other words, any term selection was not performed. We didn't used phrasal terms (word bi-grams).

2.3 Document Retrieval

In the search engine, each query term is submitted to the ranking search module, which calculates relevance scores 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 [5].

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

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

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 the values are the same as them in NTCIR-4 experiments. We had set the values of the parameters through a preliminary experiments using Search Report Data (2001, 2002, 2003) which was provided by the task organizer at the NTCIR-4. 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. The Search Report Data provided at the NTCIR-5 was not used.

The only difference in settings between NTCIR-4 and NTCIR-5 is we didn't use within-query frequencies of the query terms, for the information was not effective for somewhat short queries used at the task.

Retrieved patents were ranked on the sum of the scores and the patents published after the query patent had been 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. However, some of runs include fewer patents as mentioned in the following sections.

2.4 Exploitation of IPC codes

Two retrieval methods were examined. One is to use the IPC code of the query patent and the other is to exploit the codes assigned to top-N retrieved patents in a similar manner of pseudo-relevance feedback. In both methods, the codes were used as constraint on retrieval results. We will show the details of them.

2.4.1 Use of IPC code of query patent

The main IPC code of query patent was used as constraint on retrieval results, where the main IPC code means the one which is positioned at the first place of IPC description in the query patent text area indicated by INID 51. We used the first four characters of the IPC code, for example "G01P" for "G01P 15/09". Using the code as constraint, we eliminated from the result of baseline run the patents which does not include the code in the IPC description area.

2.4.2 Use of IPC codes of retrieved patents

The main IPC codes of retrieved patents were used as constraint on retrieval results, where the main IPC code means the one which is positioned at the first place of IPC description in the retrieved patent text. We used the first six characters of the IPC code, for example "G01P15" for "G01P 15/09". In order to collect the IPC codes above mentioned, we used top-five patents in the baseline run. Using the codes as constraint, we eliminated from the result of baseline run the patents which does not include any of the codes in the IPC description area.

2.5 Exploitation of synonyms

Another point of our experiments is similarity patent retrieval using synonyms. We automatically collected synonyms from a machine-readable dictionary for each query term, and used them in combination with original query terms. We can find such an approach to collect synonyms in [3].

More specifically, we borrowed an English-Japanese word dictionary, which had been originally developed for a machine translation system. The dictionary includes a set of records and each record consists of an English word and its translations in Japanese. For each query term, we got the English words which include the term as a translation. And then, as synonyms of the query term, we collected the Japanese words which were described as translations of the English words. In other words, we exploited transitive relations in a parallel dictionary to get a set of synonyms for the target word. In order to get synonyms only for specific terms, we restricted the target as a term of which document frequency is smaller than 200,000. Finally the collected synonyms were used for query expansion.

3 Results

Table 1 shows the evaluation results for the each submitted run. The column "A" and "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.

4 Analysis

Before giving analysis of each result, we would like to show in Table 2 the ratio of topics of which average

NTCIR-4 topics			
run-id	desc.	A	B
d0045	baseline	0.2405	0.2023
d0046	synonym	0.2424	0.2035
d0047	query IPC	0.2444	0.2019
d0048	retrieved IPC	0.2369	0.2012

NTCIR-5 topics			
run-id	desc.	A	B
d0045	baseline	0.1653	0.1357
d0046	synonym	0.1702	0.1381
d0047	query IPC	0.1766	0.1447
d0048	retrieved IPC	0.1657	0.1366

Table 1. Evaluation results

precisions were increased (better), decreased (worse), and not changed (equal) in comparison with baseline run using judgment B.

4.1 Use of IPC code of query patent

Comparing d0047 with baseline run d0045 in Table 1, the average precisions are improved except for the case of NTCIR-4 topics with B judgments. In the case of NTCIR-5 topics with A and B judgments, the average precisions are improved 7% in both cases. For the case of NTCIR-4 topics with B judgments, we can find d0047 gives more topics of which average precision increased than decreased in Table 2. So we conclude the use of IPC code of query patent as constraint is effective in the invalidity search task.

4.2 Use of IPC codes of retrieved patents

Comparing d0048 with baseline run d0045 in Table 1, we cannot find clear improvement in average precision. In Table 2, however, clearly we can find that d0047 gives more topics of which average precision increased than decreased in almost same degrees of d0047. As a conclusion, we think the use of IPC codes of retrieved patents has positive effect but the amount of improvement is smaller than the one given by the use of query patent IPC code.

NTCIR-4 topics with judgment B				
run-id	desc.	better	worse	equal
d0046	synonym	29 %	44 %	27 %
d0047	query IPC	68 %	15 %	17 %
d0048	retrieved IPC	62 %	24 %	14 %

NTCIR-5 topics with judgment B				
run-id	desc.	better	worse	equal
d0046	synonym	17 %	37 %	46 %
d0047	query IPC	48 %	10 %	42 %
d0048	retrieved IPC	42 %	11 %	47 %

Table 2. Comparison with baseline run

4.3 Exploitation of synonyms

Comparing d0046 with baseline run d0045 in Table 1, we find a little improvement in average precision. However, in Table 2, the number of topics which were hurt in average precision is larger than the one of improved topics. As a conclusion, we think the amount of improvement by exploitation of synonyms is large only if good synonyms are found, but in average cases more harmful synonyms may be obtained by the above mentioned method, because we didn't use any context of query term to get synonyms. Human judgment for use of each synonym must be needed for consistent improvement in average precision.

5 Conclusions

On the IPC of patent, two retrieval methods were examined. One is to use the IPC code of the query patent and the other is to exploit the codes assigned to top-N retrieved patents in a similar manner of pseudo-relevance feedback. In both methods, the codes were used as constraint on retrieval results. As a result, we found the former is clearly effective and the latter's effect is positive but small in comparison with the former. Another point of our experiments is exploitation of synonyms, which were automatically collected from a machine-readable dictionary for each query term. The effect depends on the topic and human judgment for use of each synonym must be required for consistent improvement in average precision.

References

[1] H.Itoh. NTCIR-4 patent retrieval experiments at ricoh. *Proc. of NTCIR Workshop 4 Meeting*, 2004.

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

[3] H.Wu and M.Zhou. Optimizing synonym extraction using monolingual and bilingual resources. *Proc. of the Second International Workshop on Paraphrasing*, pages 72–79, 2003.

[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] 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.