Experiments on Patent Retrieval at NTCIR-5 Workshop

Hironori Takeuchi **Naohiko Uramoto** **Koichi Takeda**

Tokyo Research Laboratory, IBM Research

1623-14, Shimotsuruma, Yamato-shi Kanagawa 242-8502 Japan
{hironori,uramoto,takedasu}@jp.ibm.com

Abstract

In the Document Retrieval Subtask in the NTCIR-5 Patent Retrieval Task, the search topic was the claim in a patent document and the search results were the patents that invalidate the claim in the query. Therefore we can calculate the similarities between the IPCs in the search topic and those of each patent document in the collection and use them for the patent-to-patent search. In this task, we extracted the relevance information from the search results based on the similarities between hierarchical structures of the IPCs, modified the term weighting in the query processing, and got the final retrieved documents using the updated query. As a result, it was found that there are no significant improvements from term re-weighting when considering the relevance information from the search results using hierarchical information of IPCs.

Keywords:similarity metric, hierarchical structural information, relevance feedback, patent retrieval

1 Introduction

The notion of similarity is used in many contexts such as search engines, collaborative filtering, and clustering. In many cases, the objects being compared are treated as sets or bags of elements drawn from a flat domain, and this model is called a ”vector space model”. For example, a document is treated as a bag of words in the vector space model. For similarity calculations, the objects are treated as vectors in an n-dimensional space, where n is the cardinality of the element domain and the cosine of the angle between two objects is used as a metric of their similarity. This cosine metric is mainly used for similarity computations in information retrieval systems based on vector space models [4].

There are objects that have hierarchical structures. For example, several IPCs (International Patent Codes) that represent the information for the patent claims are assigned to each patent document. For such objects that represent hierarchical structural information, there are some similarity metrics that exploit the hierarchical domain structure and that are obtained as natural generalizations of the traditional metrics [5]. We have examined the effectiveness of similarity calculations between two IPC sets in the patent collection considering the hierarchical information in the IPCs and it was found that search result are slightly improved by considering not just the text in the search topic but also the hierarchical structural information of the IPCs [9].

In the Document Retrieval Subtask in the NTCIR-5 Patent Retrieval Task, the search topic was the claim in a patent document and the search results were the patents that invalidate the claim in the query. Therefore we can calculate the similarity calculations between the IPCs in the search topic and those of each patent document in the collection. We examined the effective of relevance feedback using the search results based on the similarity metric of IPCs. From the search result using the similarity metric between IPCs in the search topic and those of each patent document in the collection, we extracted keywords and calculated their frequencies. After that, we modified the weights of the keywords in the text queries (e.g. the target claims) considering the keywords that appeared relatively frequently in the retrieved documents.

The rest of this paper is organized as follows. In Section 2, we describe the metrics to calculate the similarities between objects with hierarchical information. In Section 3, we present the query processing and the relevance feedback using the search results based on the hierarchical information of the IPCs. In Section 4, we describe the outline of our search system and experiments, and cover the results in Section 5. Finally, we will discuss the results and offer conclusions regarding our experimental study.

2 Similarity Metrics for Hierarchical Structure

In this section, we describe the similarity metrics for objects with hierarchical structures that are evaluated in our experiments [9].

First, we introduce a similarity metric based on the generalized vector space model [5]. Let U be a rooted tree, with all nodes carrying a distinct label. Each node can have an arbitrary fan-out, and the leaves of U can be at different levels. Let L_U be the set of all labels in U and LL_U be the set of all labels on the leaves of U. We define the $depth$ of a node in the hierarchy by the number of edges on the path from the root of U to that node. Given any two leaves $\overrightarrow{l_1}$ and $\overrightarrow{l_2}$ in U, we define the *Lowest Common Ancestor* $LCA(\overrightarrow{l_1}, \overrightarrow{l_2})$ to be the node of greatest depth that is an ancestor of both $\overrightarrow{l_1}$ and $\overrightarrow{l_2}$. Let the set of leaf labels LL_U be $\{\overrightarrow{l_1}, \overrightarrow{l_2}, \ldots, \overrightarrow{l_n}\}$. Then collection A (here the IPCs in a patent document) is represented by the vector $\overrightarrow{A} = \sum_{i=0}^{n} a_i \overrightarrow{l_{Ai}}$, where a_i is the weight of $\overrightarrow{l_{Ai}}$. For any two leaves l_i and l_j, we define

$$G(\overrightarrow{l_i}, \overrightarrow{l_j}) = \frac{2 \times depth(LCA(\overrightarrow{l_i}, \overrightarrow{l_j}))}{depth(\overrightarrow{l_i}) + depth(\overrightarrow{l_j})}. \qquad (1)$$

This metric defines the similarity between the two leaves $\overrightarrow{l_i}$ and $\overrightarrow{l_j}$. We continue to metric similarity by using the cosine-similarity metric. If collection A is represented by the vector $\overrightarrow{A} = \sum_{i=0}^{n} a_i \overrightarrow{l_{Ai}}$ and B by the vector $\overrightarrow{B} = \sum_{i=0}^{n} b_i \overrightarrow{l_{Bi}}$, then

$$G(\overrightarrow{A}, \overrightarrow{B}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j G(\overrightarrow{l_{Ai}}, \overrightarrow{l_{Bj}}). \qquad (2)$$

The cosine similarity between A and B is given by the following formula:

$$sim_{GCSM}(A, B) = \frac{G(\overrightarrow{A}, \overrightarrow{B})}{\sqrt{G(\overrightarrow{A}, \overrightarrow{A})}\sqrt{G(\overrightarrow{B}, \overrightarrow{B})}}. \qquad (3)$$

This metric is called the *Generalized Cosine-Similarity Measure* (GCSM) [5]. Here is an example of the calculation of GCSM. Figure 1 shows two set of IPCs, $A = \{\mathrm{F02M61/14\ 310,\ F02M61/18\ 360}\}$ and $B = \{\mathrm{F02M61/14\ 320,\ F02M65/00\ 302}\}$.

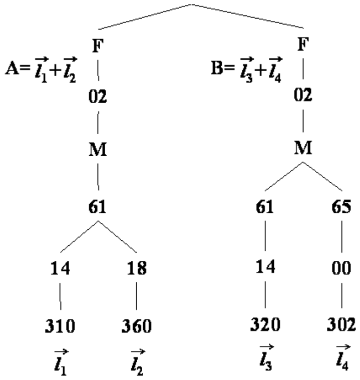

Figure 1: IPC Examples

From Equation (1), the intersections between each pair of leaves are $G(\overrightarrow{l_1}, \overrightarrow{l_2}) = \frac{2}{3}$, $G(\overrightarrow{l_1}, \overrightarrow{l_3}) = \frac{5}{6}$, $G(\overrightarrow{l_1}, \overrightarrow{l_4}) = \frac{1}{2}$, $G(\overrightarrow{l_2}, \overrightarrow{l_3}) = \frac{2}{3}$, $G(\overrightarrow{l_2}, \overrightarrow{l_4}) = \frac{1}{2}$ and $G(\overrightarrow{l_3}, \overrightarrow{l_4}) = \frac{1}{2}$. From these intersections, we can calculate the GCSM between A and B by using Equation (3), and find out $sim_{GCSM}(A, B) = 0.6847$.

Second, we introduce a similarity metric that extends the GCSM. The $depth(LCA(l_i, l_j))$ is monotonously increased in accordance with $LCA(l_i, l_j)$ in the GCSM. In the calculation of the similarity between two leaves l_i and l_j, we introduce a sigmoid function and modify it as follows:

$$EG(\overrightarrow{l_i}, \overrightarrow{l_j}) = \frac{1}{1 + \exp\{-a(G(\overrightarrow{l_i}, \overrightarrow{l_j}) - b)\}}, \qquad (4)$$

where a and b are the parameters of the sigmoid function. By using the sigmoid function, the differences of some pairs of nodes are amplified. By analogy to the GCSM, we introduce the cosine similarity between A and B as follows:

$$sim_{EGCSM}(A,B) = \frac{EG(\overrightarrow{A},\overrightarrow{B})}{\sqrt{EG(\overrightarrow{A},\overrightarrow{A})}\sqrt{EG(\overrightarrow{B},\overrightarrow{B})}}. \quad (5)$$

In this paper, we call this metric the *Extended Generalized Cosine-Similarity Measure* (EGCSM).

In the NTCIR-4 Patent task, we found that the search results using EGCSM were better than those using GCSM. Therefore, we used EGCSM for this task and set the parameters in the sigmoid function so that $a = 25$ and $b = 0.5$.

3 Term Weighting considering Relevance Information

3.1 Relevance Information

In our experiment, we use the search results using the hierarchical information in the IPCs for the relevance information. In the vector space model, the modified query vector Q_1 considering the relevance information is defined as [4],

$$Q_1 = Q_0 + \frac{1}{n_1}\sum_{i=1}^{n_1} R_i - \frac{1}{n_2}\sum_{i=1}^{n_2} S_i, \quad (6)$$

where Q_0 is the vector for the initial query, R_i is the vector for relevant document i, S_i is the vector of the nonrelevant document i, n_1 is the number of relevant documents, and n_2 is the number of nonrelevant documents. The search results using the hierarchy information are not document vectors, and we cannot use the results directly. Therefore we extract the relevance terms from the initial search results.

Let D be the document collection and N_D be its size, and D_s be the documents retrieved in the initial search and N_{D_s} be the size of the results. For each term t_i extracted from D_s, we defined the following ratio $r(t_i)$.

$$r(t_i) = \frac{f_{D_s}(t_i)/N_{D_s}}{f_D(t_i)/N_D} \quad (7)$$

This $r(t_i)$ is a relative frequency that shows tje degree to which the term t_i occurs more frequently in the retrieved document [6]. We use this ratio for the term weighting as the relevance information.

3.2 Term Weighting

In the query processing, we extract terms in the target claim from each search topic. For each extracted term t_j, we define the term weight w_j as

$$\begin{aligned} w_j &= w_{0j} r(t_j) \qquad (\text{ratio}) \\ &\text{or} \\ w_j &= w_{0j} \log r(t_j) \quad (\log\ \text{ratio}), \end{aligned}$$

where w_0 is the initial term weighting. In our experiment, we used a TF (term frequency) model, an IDF (inverse document frequency) model, and a TF-IDF model for the initial term weighting.

4 Outline of Search System and Retrieval Experiment

In this section, we present an outline of our search system and our retrieval experiment.

4.1 IR System

In our experiment, the search topic (query) was divided into two parts. One of them was a collection of IPCs assigned to the query patent document. The other was a collection of keywords from the text (i.e., the claim) in the query patent document. For the query IPCs, we constructed similarity search systems based on EGCSM. From the search result using the hierarchical information of the IPCs, we extracted the terms in the retrieved documents and their relative frequencies. For each keyword in the query document, we calculated its weight considering its relative frequency. For the weighted keywords we used a baseline IR system provided by the task organizer [1]. The baseline IR system uses a word-based indexing by Chasen v2.2.1 and the IPA dictionary v2.4.4. The retrieval model in the baseline system is BM25 [7, 8].

The ranking document list from the baseline system was filtered by the filing date and the applicant name. Figure 2 shows an overview of our system.

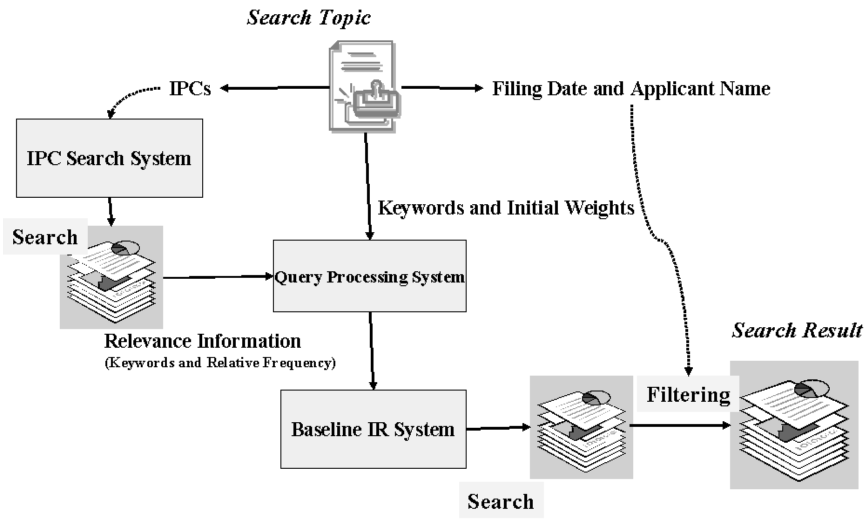

Figure 2: System Overview

4.2 Retrieval Experiment

In our experiment, we set up three initial term weighting models (TF, IDF, and TF-IDF) and three relative frequency models (none, ratio, and log ratio) in the query processing (see Table 1).

Table 1: Term Weighting Models

w_0	TF	IDF	TF-IDF
relative frequency	-	-	-
Run ID	TRL1	TRL2	TRL3
w_0	TF	IDF	TF-IDF
relative frequency	ratio	ratio	ratio
Run ID	TRL4	TRL5	TRL6
w_0	TF	IDF	TF-IDF
relative frequency	log ratio	log ratio	log ratio
Run ID	TRL7	TRL8	TRL9

For the comparison, we tried using the pseudo-relevance feedback provided in the baseline IR system for each initial term weighting model. The IDs using the pseudo-relevance feedback were TRL10 for TF model, TRL11 for IDF model, and TRL12 for TF-IDF model. We also set up a search system using only IPC information based on EGCM (Run ID:TRL13).

5 Results

In this section, we show the result of our retrieval experiments. Table 2 shows the mean average precision (MAP) of each run by using relevant patents A and B [1], respectively.

In Table 2, NTCIR-4 and NTCIR-5 denote the search topics used in NTCIR-4 and NTCIR-5, respectively. From these results, it can be seen that the term reweighting considering the search results using the information of IPC improved the search results for the search topics used in NTCIR-4 but did not improve the results for the search topics used in NTCIR-5.

6 Discussion

Here, we will discuss our experimental results. It was reported that the TF term weighting model had a negative impact on the search results in the newspaper-to-patent search task [3]. From Table 2, the different tendency was observed in the patent-to-patent search task. However it can be seen that the differences between the MAP for the TF model and those for the DF model were reduced by considering the relevance information from the IPC search results.

In query patents A, the number of relevant document is one, so MAP is greatly influenced by the precision in the top ranking retrieved documents and becomes unstable. We think that the results of the NTCIR-4 and NTCIR-5 search topics using the term re-weighting model considering the IPC search results deffered because the number of search topics in NTCIR-4 is much smaller than

Table 2: Mean Average Precision(MAP)

Run ID	TRL1	TRL2	TRL3	TRL4	TRL5	TRL6	-
NTCIR-5 A	0.0849	0.0655	0.0720	0.0529	0.0453	0.0522	-
NTCIR-5 B	0.0675	0.0539	0.0585	0.0456	0.0398	0.0445	-
NTCIR-4 A	0.1025	0.1023	0.1012	0.1023	0.1034	0.1023	-
NTCIR-4 B	0.0975	0.0961	0.0965	0.0916	0.0962	0.0924	-
Run ID	TRL7	TRL8	TRL9	TRL10	TRL11	TRL12	TRL13
NTCIR-5 A	0.0690	0.0560	0.0668	0.0768	0.0623	0.0696	0.0176
NTCIR-5 B	0.0570	0.0473	0.0547	0.0653	0.0544	0.0577	0.0150
NTCIR-4 A	0.1014	0.0968	0.1060	0.0868	0.0862	0.0989	0.0176
NTCIR-4 B	0.1001	0.0928	0.0981	0.1087	0.1042	0.1113	0.0170

that in NTCIR-5. We think that the term reweighting used in our experiment does not improve the search result significantly. We did not select keywords from the retrieved documents based on the IPC information, so we think we need to examine which terms contribute the relevance information.

In the filtering of the ranking document list, we considered not only the filing date but also the applicant name. We think we also need to investigate the effectiveness of each filtering condition.

7 Conclusion

In this paper, we examined the effectiveness of relevance feedback using search results based on a similarity metric for the IPCs. We introduced a term reweighting model considering the search results based on this hierarchical information of IPCs. The results showed that the search results were not significantly improved by the term reweighting when considering the search results based on the IPC information. For the future work, we need to examine the term selections from the documents retrieved by the IPC information for their relevance information, and the effectiveness of filtering condision.

Acknowledgements

We would like to thank the organizers of the NTCIR-5 Patent Retrieval Task for preparing a valuable test collection.

References

[1] A. Fujii, M. Iwayama and N. Kando. Overview of Patent Retrieval Task at NTCIR-5. *Proc. of the 5th NTCIR Workshop*, 2005.

[2] A. Fujii, M. Iwayama and N. Kando. Overview of Patent Retrieval Task at NTCIR-4. *Proc. of the 4th NTCIR Workshop,* 2004.

[3] M.Iwayama, A.Fujii, N.Kando and A.Takano. Overview of Patent Retrieval Task at NTCIR-3. *Proc. of the 3rd NTCIR Workshop,* 2002.

[4] W.B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structure & Algorithms. Prentice Hall.* 1992.

[5] P. Ganesan, H. Garcia-Molia and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1):64–93, 2003.

[6] T. Nasukawa and T. Nagano. Text analysis and knowledge mining system. *IBM System Journal*, 40(4):967-984, 2001.

[7] S.E. Robertson and K. Spark-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science,* 27(3):129–146, 1976.

[8] S.E. Robertson and K. Spark-Jones. Some simple effective approximation to the 2-poisson model for probabilistic weighted retrieval. *Proc. of the* 17^{th} *SIGIR*, 232-241, 1994.

[9] H. Takeuchi, N. Uramoto and K. Takeda. Experiments on Patent Retrieval at NTCIR-5 Workshop. *Proc. of the 4th NTCIR Workshop,* 2004.