Experiments on Patent Retrieval at NTCIR-4 Workshop

Hironori Takeuchi[*] **Naohiko Uramoto**[*†] **Koichi Takeda**[*]

[*]Tokyo Research Laboratory, IBM Research [†] National Institute of Informatics

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

Abstract

In the Patent Retrieval Task in NTCIR-4 Workshop, the search topic is the claim in a patent document, so we use the claim text and the IPC information for the similarity calculations between the search topic and each patent document in the collection. We examined the effectiveness of the similarity measure between IPCs and the term weighting for the occurrence positions of the keyword attributes in the search topic. As a result, it was found that the search results are slightly improved by considering not just the text in the search topic but also the hierarchical structural information of the IPCs. In contrast, the term frequencies for the occurrence position of the attribute did not improve the retrieval result.

Keywords:similarity measure, hierarchical structural information, 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 measure of their similarity. This cosine measure is mainly used for similarity computations in the vector space model based information retrieval systems (Frakes and Baeza-Yates, 1992).

There are objects that have hierarchical structures. For example, some IPCs (International Patent Codes) that represent the information for the patent claims are assigned to each patent document. For these objects with hierarchical structural information, there are some similarity measures that exploit the hierarchical domain structure and that are obtained as natural generalizations of the traditional measures (Ganesan et al., 2003).

The purpose of this paper is to examine the effectiveness of similarity calculations between two IPC sets in the patent collection considering the hierarchy information in the IPCs by using the patent test collection of NTCIR-4. In the Patent Retrieval Task in the NTCIR-4 Workshop, we searched the patent collection for the patents that can invalidate the requirements in an existing claim. Because both the search topic and the document collection are patent documents and include IPCs in this invalidity search, we can calculate the similarity between the search topic and each document in the collection by using the IPC hierarchical information. We use a metric based on the generalized vector space model (Ganesan et al., 2003) and an extended metric for the similarity calculation between two IPC sets, and evaluate the effectiveness of these metrics. In the main task, the text query (e.g. target claim) is divided into some components, so we can also examine the effectiveness of the term weighting considering the occurrence positions of the keyword attributes in the search topic.

The rest of this paper is organized as follows. In Section 2, we describe the metrics to calculate the similarity between objects with hierarchical infor-

mation. In Section 3, we present the query processing and the search systems. In Section 4, we describe the outline of our search experiments and cover the results in Section 5. Finally, we will discuss the results and offer conclusions regarding our experimental study.

2 Similarity Measures for Hierarchical Structure

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

First, we introduce a similarity measure based on the generalized vector space model (Ganesan et al., 2003). Let U be a rooted tree, with all nodes carrying a distinct label. Each node can have 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 to 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 (such as 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

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

This metric defines the similarity between the two leaves $\overrightarrow{l_i}$ and $\overrightarrow{l_j}$. We continue to measure similarity by using the cosine-similarity measure. 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

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

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

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

This measure is called the *Generalized Cosine-Similarity Measure* (GCSM) (Ganesan et al., 2003). Now we will show an example for the calculation of GCSM. Figure 1 shows the two set of IPCs, $A = \{\mathrm{F02M61/14\ 310},\ \mathrm{F02M61/18\ 360}\}$ and $B = \{\mathrm{F02M61/14\ 320},\ \mathrm{F02M65/00\ 302}\}$.

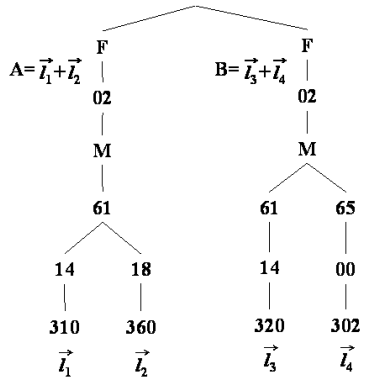

Figure 1: IPC Examples

From (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 get $sim_{GCSM}(A, B) = 0.6847$.

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 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

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

where a and b are the parameters of the sigmoid function. 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).

3 Query Processing and IR System

3.1 Term Weighting and Query Processing

In the NTCIR-4 patent task, each search topic is an unexamined patent application. We extract the target claim from each search topic. A claim usually consists of multiple components and relevance judgment is performed on a component-by-component basis in the real patent search. So, for each target claim, the <COMP> tags are inserted by a person who produces search topics and who is in charge of the relevance assessment for these topics. For each component, we define the weight for the terms by using the following function as

$$w_i = \begin{cases} i & 1 \leq i \leq \frac{n}{2} \\ -i+n+1 & \frac{n}{2} < i \end{cases} \quad i = 1, 2, \ldots, n \qquad (6)$$

where n is the number of the <COMP> tags in the target claim. Using w_i, we modify the term frequencies as follows:

$$tf_k = \sum_{i=1}^{n} w_j tf_{ki} \qquad (7)$$

where tf_{ki} is the term frequency of the k-th attribute keyword in the i-th component. We use this modified term weight for the query.

We also extracted the filing date and the applicant name from each search topic to filter the retrieved documents so that the retrieved patents should have been filed prior to the topic patent and should not have the same applicant name.

3.2 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 and their weights from a text (i.e., a claim) in the query patent document. For the query IPCs, we constructed the similarity search systems based on GCSM and EGCSM. For the weighted keywords we used a baseline IR system provided by the task organizer. The baseline IR system uses a word-based indexing by Chasen 2.2.1 and the IPA dictionary 2.4.4. The retrieval model in the baseline system is BM25 (Robertson and Spark-Jones, 1976; Robertson and Spark-Jones, 1994).

The retrieved documents and their similarity scores from the two retrieval systems were merged. In the results from the baseline system, each score was normalized by the maximal score so that the similarity of the first retrieved document should be 1.0. For each retrieved document $\overrightarrow{d}_i$, we calculated the following integrated ranking status value

$$\begin{aligned} IRSV_i &= \alpha ipcsim_i + (1-\alpha) nscore_i, \qquad (8) \\ 0 &\leq \alpha \leq 1, \end{aligned}$$

where $IRSV_i$ is the integrated ranking status value, $ipcsim_i$ is the similarity between the collection of IPCs in the query and that in the $\overrightarrow{d}_i$, and $nscore_i$ is the normalized score from the baseline system. The ranking document list by $IRSV_i$ was filtered by the filing date and the applicant name. This filtered document list was the result of our IR system. Figure 2 shows the overview of our system.

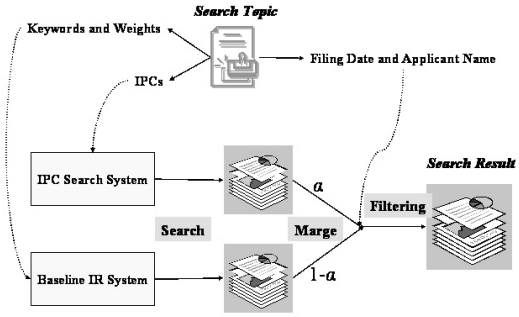

Figure 2: System Overview

4 Outline of Retrieval Experiment

In this section, we describe the outline of our retrieval experiment. We set up 2 systems using GCSM and EGCSM for the searches using IPC information. For both GCSM and EGCSM, we set

the same merge parameter, α as defined in Equation (8). Table 1 shows the parameters and corresponding system IDs. Because the retrieved documents came only from the baseline IR system in for $\alpha = 0.0$, the result of TRL2 was the same as that of TRL4. In EGCSM, we set the parameters so that $a = 25$ and $b = 0.5$. In the additional run, the <COMP> tags were not inserted in the target claim, so the traditional tf model was used for the query processing.

Table 1: Merge Parameter

	GCSM					
α	0.0	0.2	0.4	0.6	0.8	1.0
System ID	TRL2	TRL8	TRL7	TRL6	TRL5	TRL1
	EGCSM					
α	0.0	0.2	0.4	0.6	0.8	1.0
System ID	TRL4	TRL12	TRL11	TRL10	TRL9	TRL3

The relevance judgments for patents were made based on the following two ranks. The documents that could invalidate the demands of all essential components in a target claim were judged as "A". The documents that could invalidate the demands of most of the essential components in a target claim (but not all of the essential components) were judged as "B" (Fujii et al., 2004). These relevant documents were obtained from various the different sources, including the citations maed by the examiner of the patent office, the manuals searched by the evaluator, and the 30 systems participating in the pooling.

5 Results

In this section, we show the result of our retrieval experiment. Tables 2 and 3 show the mean average precision(MAP) of each system by using relevant patent A and A+B, respectively. In this task, there are only a few relevant documents for each search topic (Fujii et al., 2004). In this case, MAP is greatly influenced by the precision in the top ranking (e.g. 10) retrieved documents. So, for each system, tables 4 and 5 show the macro average ranking that is the micro average of rankings of the first relevant documents for each topic.

In these tables, the "main" and "add" lines show the results for the main search topics and the additional search topics. The "all" lines show the results for the combined search topics. The "(c)" notation indicates that only the citations by the patent office examiner were used as the relevant documents. The underlined values show the best scores for both of the IPC search systems.

Figure 3 shows the recall-precision curve of each system in the main task. Figure 4 shows the recall-precision curves from the GCSM based IPC search system(TRL1) and the EGSCM based IPC search system(TRL3).

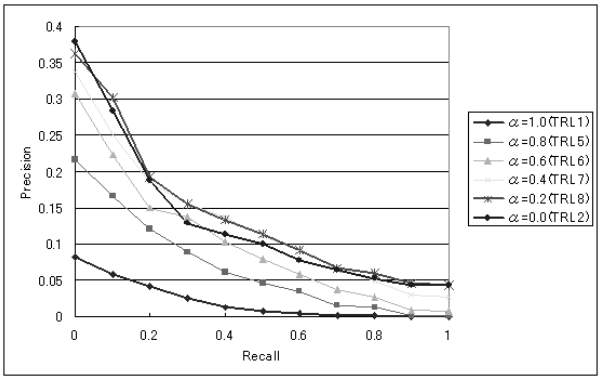

Figure 3: Recall-Precision Curve (Main task A+B(relaxed))

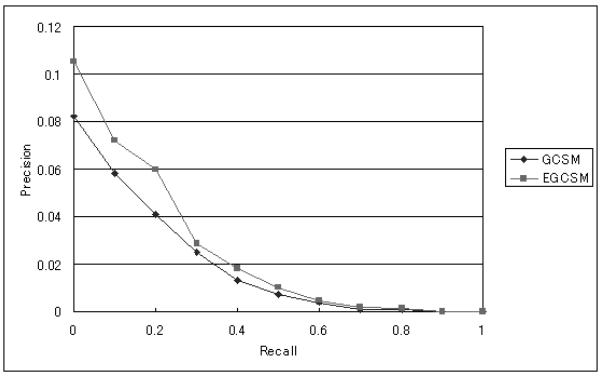

Figure 4: Recall-Precision Curve (Main task A+B(relaxed))

From these tables and figures, it is apparent that the search results of the baseline IR system were slightly improved upon by considering the results of the IPC search system. From the evaluations of both GCSM and EGSM in Figure 4, it is found that the search result of EGSM is better than that of GCSM. Comparing the evaluations of the main runs with those of the additional runs, the term

frequencies for the occurrence positions of the attributes did not improve the retrieval results.

6 Discussion

Here, we will discuss our experimental results. Judging from the result tables, the optimal merge parameter α defined in (8) is about 0.2. This means that we should not rely too much upon the information of the IPC for the patent-by-patent search. In the EGCM, we set the parameters so that $a = 25$ and $b = 0.5$. In that model, the differences between the intermediate nodes were amplified. We need to examine other amplification, the differences between the end nodes and those between the top nodes. If some search topics and their relevant documents are provided, we can estimat the parameters and evaluate their effectiveness.

From the comparison between the results of main run and those of additional run, we can not find the clear effectiveness of the tf model considering the occurrence position of the attribute. However we need to examine it by the same search topics. We also need to examine other term weighting model and the comination of our modified tf model and them.

7 Conclusion

In this paper, we examined the effectiveness of the similarity metrics between IPCs and that of the term weighting for the occurrence positions of the keyword attributes in the search topics by using the patent test collection of NTCIR-4. As a result, it was found that the search results are slightly improved by considering not just the text in the search topic, but also the hierarchical structural information of the IPCs. We need to examine the optimal model parameters for EGCSM. In contrast, the term frequencies for the occurrence positions of the attributes did not improve the retrieval results, though we need to evaluate the differences using the same search topics.

Acknowledgements

We would like to thank the organizers of the Patent Retrieval Task of NTCIR-4 Workshop for preparing a valuable test collection. We also would like to thank the evaluators who created the topics and who assessed the relevance of the results.

References

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

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

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

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

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

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

Table 2: Mean Average Precision(MAP) (A)

	GCSM					
α(System ID)	0.0(TRL2)	0.2(TRL8)	0.4(TRL7)	0.6(TRL6)	0.8(TRL5)	1.0(TRL1)
main	0.0644	0.0645	0.0481	0.0261	0.0119	0.0029
main(c)	0.0784	0.0910	0.0808	0.0607	0.0460	0.0057
add	0.0752	0.0836	0.0866	0.0449	0.0398	0.0186
all	0.0716	0.0773	0.0739	0.0387	0.0306	0.0134
all(c)	0.0868	0.0962	0.0987	0.0620	0.0488	0.0176
	EGCSM					
α(System ID)	0.0(TRL4)	0.2(TRL12)	0.4(TRL11)	0.6(TRL10)	0.8(TRL9)	1.0(TRL3)
main	0.0644	0.0641	0.0482	0.0323	0.0301	0.0038
main(c)	0.0784	0.0958	0.0850	0.0602	0.0550	0.0084
add	0.0752	0.0790	0.0705	0.0492	0.0335	0.0147
all	0.0716	0.0741	0.0631	0.0436	0.0324	0.0111
all(c)	0.0869	0.0949	0.0868	0.0675	0.0473	0.0157

Table 3: Mean Average Precision(MAP) (A+B)

	GCSM					
α(System ID)	0.0(TRL2)	0.2(TRL8)	0.4(TRL7)	0.6(TRL6)	0.8(TRL5)	1.0(TRL1)
main	0.1218	0.1310	0.1179	0.0908	0.0584	0.0174
main(c)	0.0839	0.0996	0.0941	0.0778	0.0574	0.0084
add	0.0892	0.0975	0.1071	0.0650	0.0510	0.0232
all	0.1001	0.1088	0.1107	0.0737	0.0535	0.0212
all(c)	0.0874	0.0982	0.1027	0.0693	0.0531	0.0182
	EGCSM					
α(System ID)	0.0(TRL4)	0.2(TRL12)	0.4(TRL11)	0.6(TRL10)	0.8(TRL9)	1.0(TRL3)
main	0.1218	0.1300	0.1096	0.0854	0.0622	0.0217
main(c)	0.0839	0.0999	0.0870	0.0703	0.0561	0.0102
add	0.0892	0.0945	0.0896	0.0770	0.0491	0.0196
all	0.1001	0.1065	0.0964	0.0798	0.0535	0.0203
all(c)	0.0874	0.0963	0.0887	0.0747	0.0515	0.0165

Table 4: Macro Average Ranking (A)

	GCSM					
α(System ID)	0.0(TRL2)	0.2(TRL8)	0.4(TRL7)	0.6(TRL6)	0.8(TRL5)	1.0(TRL1)
main	436.59	414.56	401.64	526.29	616.84	756.03
main(c)	422.23	387.58	396.41	498.72	570.19	740.44
add	454.27	428.52	439.42	499.73	641.57	714.93
all	448.44	423.91	426.96	508.49	633.41	728.49
all(c)	444.38	415.89	426.15	499.42	619.55	722.80
	EGCSM					
α(System ID)	0.0(TRL4)	0.2(TRL12)	0.4(TRL11)	0.6(TRL10)	0.8(TRL9)	1.0(TRL3)
main	436.59	403.89	416.93	512.02	606.35	750.20
main(c)	422.23	386.95	418.51	480.75	575.79	751.88
add	454.27	428.64	461.99	532.95	636.84	752.68
all	448.44	420.47	447.12	526.04	626.78	751.86
all(c)	444.38	415.78	448.58	516.84	618.00	752.43

Table 5: Macro Average Ranking (A+B)

	GCSM					
α(System ID)	0.0(TRL2)	0.2(TRL8)	0.4(TRL7)	0.6(TRL6)	0.8(TRL5)	1.0(TRL1)
main	492.21	465.45	443.42	526.61	642.31	765.03
main(c)	439.18	411.90	398.79	477.62	575.82	725.16
add	441.84	402.39	415.99	479.17	621.15	691.80
all	458.79	423.62	425.23	495.14	628.28	716.45
all(c)	440.94	405.60	510.20	478.65	605.89	703.03
	EGCSM					
α(System ID)	0.0(TRL4)	0.2(TRL12)	0.4(TRL11)	0.6(TRL10)	0.8(TRL9)	1.0(TRL3)
main	492.21	463.79	461.70	540.85	652.16	763.25
main(c)	439.18	412.76	417.22	515.15	606.93	750.71
add	441.84	403.99	437.39	511.86	612.85	725.80
all	459.79	424.12	445.57	521.62	626.08	738.40
all(c)	440.94	406.94	430.60	512.96	610.85	734.18