A PATENT RETRIEVAL METHOD USING SEMANTIC ANNOTATIONS

Youngho Kim, Jihee Ryu, Sung-Hyon Myaeng
KAIST, 335 Gwahak-ro Yuseong-gu, Daejeon, South Korea
bruceyhkim@kaist.ac.kr, zzihee5@kaist.ac.kr, myaeng@kaist.ac.kr

Keywords: patent retrieval, invalidity search, semantic annotation, cluster-based retrieval

Abstract: Automatic annotation of key phrases for their semantic categories can help improving effectiveness of a variety of text-based systems including information retrieval, summarization, question answering, etc. In this paper, we exploit semantic annotations for patent retrieval (i.e., patent invalidity search). We first annotated key phrases for two semantic categories, PROBLEM (e.g. "pattern matching") and SOLUTION (e.g. "dynamic programming") in a patent document, which constitute a particular technology. Semantic clusters are formed by grouping patent documents with the same PROBLEM or SOLUTION tag. A language modelling approach to information retrieval is extended to consider the semantically oriented clusters as well as document models. Our retrieval evaluation of the proposed approach using a collection of United States patent documents shows a 22% improvement over the baseline, a smoothed language modelling approach without using the semantic annotations.

1 INTRODUCTION

Patent texts are a rich resource for semantic knowledge discovery as they contain many technological concepts and their semantic relations. Some previous studies attempted to analyze patent texts to discover meaningful information (Shinmori et al., 2003; Yoon and Park, 2004; Fujii et al, 2007a; Kim et al., 2009). The results are often targeted at developing automatic or semi-automatic analytical tools that help patent analysis experts identify past technical progresses and estimate future directions.

On the other hand, patent retrieval, searching past relevant patents to a target technology, can also assist examiners (i.e., experts) in patent offices in the context of *invalidity search* (Fujii et al, 2007b). The main purpose of retrieving patent documents is to validate the genuineness of the technology in a patent application. This task has newly emerged as a focus of recent workshops including the NTCIR workshops (Kando, 2004, 2005, 2007) and the CIKM workshop on patent information retrieval (Tait, 2008). Examining the invalidity (i.e., finding prior patents that contain some conflicting claims) through search is critical for a newly applied patent to be granted.

Our research focuses on the intersection of the two areas: semantic annotation and patent retrieval. Among a variety of possible types of semantic annotations, we opted for two key aspects of patent documents, namely, PROBLEM and SOLUTION that constitute a technology (Kim & Myaeng, 2009). We feel that especially for patent invalidity search, identifying patents with the same PROBLEM and SOLUTION as those in the query patent would be a critical task.

Unlike ad hoc retrieval, mostly targeting at news articles, patent retrieval is unique in that it has to deal with lengthy and structured documents. A patent document usually consists of title, abstract, claim, and description sections. As observed previously (Iwayama et al., 2003), a patent document is 24 times longer than a news document on average, and the variance of the lengths in patent documents is about 20 times larger than that of new articles. These characteristics need to be considered when patent documents are semantically annotated and processed for retrieval and when various language models are constructed.

Another unique aspect of patent retrieval is that every patent has a classification code called IPC (International Patent Classification) manually assigned to it. Since the IPC codes are semantic in

nature and organized in taxonomy, they can be utilized effectively to group similar patents for retrieval (Kang et al., 2007). For example, all the patents with the IPC code G10L 17 have something to do with "speech recognition", the description of the IPC code. If a query patent belongs to the cluster of patents sharing the same IPC code, it is highly probable to find a conflicting patent in the cluster. The IPC codes make cluster-based retrieval since they can be the basis for semantic clustering (Kang et al., 2007).

However, IPC-based clusters may present a problem when searches are performed within each of them. Since the documents in a cluster are similar to each other, they share many terms, making it difficult to discriminate among each other. Since the goal of invalidity search is to pinpoint the patent documents claiming the same technology, retrieving many grossly similar documents with ordinary index terms would not be very helpful, especially when the size of a cluster is large. Since identifying discriminating features would be difficult but critical for patent invalidity search, we need semantically annotated terms that would help making a fine distinction between the patents claiming the same technology or method from those that are grossly similar to each other based on all the index terms.

The main thrust of this paper, therefore, is to link problem/solution-based semantic annotations, clustering, and patent retrieval. We describe a patent retrieval model based on semantic clusters. The system proposed in this paper consists of two parts: semantic annotation for the PROBLEM and SOLUTION categories and cluster-based retrieval based on extracted semantic key phrases. For the retrieval part, we attempt to distinguish patent documents in a cluster for the same PROBLEM or SOLUTION from those in other clusters, assuming that documents belonging to the same semantic cluster are more likely to be similar and hence conflicting among each other.

The rest of this paper is organized as follows. In Section 2, we present the related work in patent retrieval and cluster-based retrieval. In Section 3, we describe the semantic clustering method based on the problem and solution annotations and a semantic patent retrieval model. We illustrate and interpret the experimental results in Section 4 and finally present our conclusion in Section 5.

2 RELATED WORK

A cluster-based model for Information Retrieval (IR) takes advantages of document clusters by assuming that relevant documents would be grouped within the same cluster. In general, documents are automatically grouped by their topical relatedness and relevant clusters are chosen with respect to a given query (Croft, 1980; Voorhees, 1985), so that the query terms in the relevant cluster are heavily weighted in the retrieval model. In order to verify the superiority of cluster-based retrieval model, Liu and Croft (2004) compared with the cluster-less model in a large test collection, using the language modeling approach.

Prior to the series of workshops related to patent retrieval, Larkey (1999) utilized IPC codes to divide an entire corpus of patents into sub-corpora. The patents in each sub-corpus compose a large virtual document, and a query patent was mapped to each virtual document to select n-best sub-collections. In this approach, the search techniques in distributed IR (Callan et al., 1995) were applied in order to reduce long search time in several sub-collections. The work is considered an important attempt to use a unique aspect of patent documents.

Chen et al. (2003a) proposed a patent document retrieval system concerning semantic and syntactic properties. They utilized Latent Semantic Index to recognize synonymous expressions. The system first finds the patent documents whose vectors lie in the neighbourhood of the query vector. It then uses the template matching algorithm developed by Chen & Tokuda (2003b) to calculate the similarity of the document and the query. Takaki (2004) proposed an associative document retrieval method. They extracted sub-topics from each query and weighted them by a term frequency-based entropy model. They applied this method in patent invalid search by using a query patent claim.

Many previous studies were presented in the series of the NTCIR workshops (Kando, 2004, 2005, 2007). Among the work related to this paper is the one by Konishi et al. (2004) that used an IPC code as a category for each patent and combined TF/ICF (term frequency and inverse category frequency) with a general TF/IDF scoring formula. Fujii (2007) integrated content and citation information to identify an authoritative page by citation information (i.e., a patent is cited by a large number of other patents – foundation patent) like the PageRank method, which was combined with the Okapi BM25 model. His system performed the best among all the participants in the task of patent retrieval in NTCIR-6 (Fujii et al., 2007b).

The work by Kang et al. (2007) seems to be most relevant to our research. They proposed a cluster-based retrieval model utilizing IPC classes. Since the same IPC class would be assigned to somewhat relevant patents, this approach is quite effective to enhance the baseline model (i.e., language model based approach). Our work is different from this approach in that we utilize semantic annotation results in constructing semantic clusters that are simply borrowed as a vehicle to incorporate such semantic annotations for patent retrieval.

3 SEMANTIC PATENT RETRIEVAL

We begin with an explanation about the PROBLEM and SOLUTION annotation method (Kim et al., 2009) because it is the basis for making the proposed retrieval method semantic in nature. The semantic clustering method should be considered just one way of incorporating the semantics for patent retrieval. We first extract key phrases in each document, which become the candidates for PROBLEM and SOLUTION annotations and generate semantic clusters: PROBLEM and SOLUTION clusters starting from each patent document. Semantically clustered documents now allow us to measure the probabilistic relatedness between a patent document and a query patent.

Assuming that a patent makes claims for a unique method (solution) to perform a specific task (problem), it would be useful to identify the technology described in a patent in terms of *problem* and *solution* as proposed previously (Kim et al., 2009). For example, when "a dynamic programming method for speech pattern recognition" is the new technology in a patent, the problem and solution parts are "speech recognition" and "dynamic programming method", respectively. This type of semantic annotation of key phrases would make them more unique and help the task of discriminating the patents in a cluster.

For the semantic annotation task, we employ the method proposed recently (Kim et al., 2009), which combines a probabilistic language modeling approach and linguistic clues. The features including linguistic clues are integrated within a statistical classifier framework (i.e., Support Vector Machine).

As a result of the semantic annotation, all the problems and solutions of each patent are identified. Patents sharing the same key-phrases with the same semantic annotation (PROBLEM or SOLUTION) can be clustered, and such semantic clusters can be exploited for patent retrieval. Unlike the patent retrieval model using a language modelling approach and citation links (Fujii et al., 2007a), ours is to use semantic links that connect patents sharing the same problems and/or solutions. Since the key phrases are semantically represented with the annotations, our retrieval method is also considered semantically based.

In the context of patent invalidity search, in addition, identifying PROBLEM and SOLUTION key phrases can facilitate finding the conflicts that would invalidate the target patent. If we can identify and group the patents that share the same PROBLEM as in the query patent, the invalidity search in the group would be a matter of determining whether the SOLUTIONS are sufficiently different. Similarly, if the patents are clustered with the same solution, the remaining task would be just to ensure the problems are the same.

3.1 Semantic Annotation

The task of semantic annotation is basically to identify PROBLEM and SOLUTION key terms in each patent document. In other words, specific semantic knowledge for a patent are first discovered before it is used to index the document. In this work, a technology is viewed as an association of PROBLEM and SOLUTION key terms, e.g., "recognize signal pattern" (SOLUTION) for "noise reduction" (PROBLEM). The annotation task is accomplished in three steps as follows (Kim et al., 2009).

Step 1. All the key phrases from each patent document are extracted as candidates for semantic annotation. A key phrase is recognized as an atomic noun phrase (i.e., the smallest noun phrase, tagged as *NP*) from the result of parsing a sentence in a patent document by a statistical parser (Klein and Manning, 2003). A noun phrase is then expanded, if possible, to a verb phrase by adding a related verb that has a syntactic dependency with the noun phrase. The result of this step is a list of candidate phrases for annotation.

Step 2. PROBLEM phrases are identified based on language model probabilities and linguistic patterns that signal the existence of a problem phrase (e.g. in a pattern consisting of "system for" + [noun phrase], [noun phrase] is annotated as PROBLEM). The linguistic patterns are generalized and integrated into a machine learning framework (i.e. SVM). Since many cited patents share the same PROBLEM key phrase (i.e., patents providing a solution to the

same problem), in addition, the language model including the cited patents is constructed.

Step 3. SOLUTION key phrases are identified from the rest of the key phrase candidates, mostly based on linguistic clues. Unlike the PROBLEM annotation step, SOLUTIONS are revealed not only by pure linguistic patterns but also by PROBLEM tags added in the previous step because a SOLUTION key phrase occurs frequently together with a PROBELM key phrase, often a PROBLEM followed by a SOLUTION as in "speech recognition system using speaker language model." Note that no language modelling is used for SOLUTIONS as they do not occur frequently across different patent documents. The extracted SOLTUTION key phrases are also integrated into the SVM classifier. Finally, each patent document is associated with the set of PROBLEM and SOLUTION key phrases. The annotation methods were evaluated to obtain 76% and 75% in accuracy for PROBLEM and SOLUTION annotations, respectively (Kim et al., 2009).

3.2 Semantic Retrieval Model

In general, the IR task is defined to be a ranking method for a set of documents by topical relatedness to a given query. Among many approaches to this task, we focus on cluster-based approaches based on the cluster hypothesis (van Rijsbergen, 1979), which states that closely associated documents tend to be relevant to the same requests. In addition, we were encouraged by a recent result that showed the value of a cluster-based model for patent retrieval using IPC (International Patent Classification) code (Kang et al., 2007). However, the cluster-based method proposed in this paper is just one way of using semantic annotations for IR.

While previous research on using clustering for information retrieval are primarily based on document similarities or classification results based on a classification scheme as in (Kang et al., 2007), we posit that relevant patent documents would share the same PROBLEM and/or SOLUTION key phrases. Our approach coincides with the idea of using citation links, which is motivated by the tendency that cited documents often contain solutions to a similar problem. Our work differs from others in that we explicitly utilize semantic annotations discovered automatically for problems and solutions in patent documents.

Our retrieval model is based on the language modeling approach (Ponte & Croft, 1998) where the probability of generating the query from a document language model is estimated to rank documents. Instead of taking a document alone for language model construction, however, we make use of a semantic cluster that includes the document as in the work by Kang et al. (2007).

$$p(Q|D) \cong (1-\pi)\, p_{mle}(Q|D) + \pi\, p_{mle}(Q|S)$$
$$\cong \sum_{q \in Q} \left\{ (1-\pi) \underbrace{p_{mle}(q|D)}_{\text{term 1}} + \pi \underbrace{p_{mle}(q|S)}_{\text{term 2}} \right\}$$

where *mle* indicates maximum-likelihood estimation, q is a query term in query Q, and S is the semantic cluster constructed for the document D using PROBLEM and SOLUTION annotations, and π is a mixing weight.

By assuming that a query is generated from both of the document and the semantic cluster containing it, we can divide the query generation probability into two parts: document model (term 1) and semantic cluster model (term 2). The mixture of two models is almost the same as the cluster-based model proposed by Kang et al. (2007), but the actual estimation process is different. Instead of using IPC class based clusters, we assume the semantically driven clusters should play a key role in generating a query. This mixture model is estimated as:

$$p_{mle}(q|D) = \frac{cnt(q:D)}{\sum_{t \in V_D} cnt(t:D)}$$

$$p_{mle}(q|S) = \frac{cnt(q:S)}{\sum_{t \in V_S} cnt(t:S)}$$

where V_S is a vocabulary set of S, V_D is a vocabulary set of D, and $cnt(q:S)$ is the frequency of q in S.

We assume that the language model follows a unigram word distribution, and the term in our model would be a unigram word. As such, we count a unigram frequency for each model, which is assumed to express the topicality of a document and a semantic cluster. In addition, we use a weight to mix the two models. A different mixing weight is assigned to each query term as some terms are more related to the cluster than others. The mixing bias (π_q) is estimated by:

$$\pi_q = \frac{Dfreq(q:S)}{Dfreq(q:C)}$$

where *Dfreq* (*q* : *C*) is the number of documents containing a query term *q* in the collection C. Its inverse (IDF) is multiplied by *Dfreq* (*q*: *S*), document frequency of *q* in the semantic cluster *S*, so that a query term appearing in the semantic cluster *S* is considered more important. With the mixing bias for each query, the retrieval model is rewritten as:

$$p(Q|D) \cong \sum_{q \in Q} (1-\pi_q)\, p_{mle}(q|D) + \pi_q\, p(q|S)$$

The next step is to determine which cluster is more applicable to the retrieval task since many clusters can be formed. For an extracted PROBLEM key phrase $p \in D$ and a SOLUTION key phrase $s \in D$, clusters can be formed as follows:

$$CL \in \{CL_p(D),\ CL_s(D),\ CL_p(D) \cap CL_s(D),\ CL_p(D) \cup CL_s(D)\}$$

where $CL_p(D) = \{d_1, d_2, ...\}$ s.t. $p \in d_i$

$CL_s(D) = \{d_1, d_2, ...\}$ s.t. $s \in d_i$

We basically generate a PROBLEM cluster, $CL_p(D)$, and a SOLUTION cluster, $CL_s(D)$, and from these, we further define their intersection and union. One of the four clusters for a particular key phrase can be used as the cluster in the language model defined above.

However, the cluster-based retrieval model defined above can encounter a data sparseness problem. As in other unigram language models (Ponte & Croft, 1998), we use the Jelinek-Mercer smoothing method (Zhai & Lafferty 2001) as follows to alleviate the problem:

$$p(q|D) \cong (1\text{-}\lambda)\, p(q|D) + \lambda\, p(q|C)$$

where C is a collection

When we apply the smoothing method to our model, we have two options: smoothing only the document model or smoothing both the semantic cluster and document models, resulting in the following:

$$p_{mle}(q|D) \cong \pi_q\, p_{mle}(q|S) + (1-\pi_q)\left\{(1-\lambda)\, p_{mle}(q|D) + \lambda\, p(q|C)\right\} \quad \textbf{(1)}$$

$$p'_{mle}(q|D) \cong \pi_q \left\{(1-\alpha)\, p_{mle}(q|S) + \alpha\, p(q|C)\right\} + (1-\pi_q)\left\{(1-\beta)\, p_{mle}(q|D) + \beta\, p(q|C)\right\} \quad \textbf{(2)}$$

In equation **(1)**, only the document language model is smoothed with an assumption that the cluster model would be less vulnerable from the word paucity. In equation **(2)**, we smooth both of the models without the assumption. Also, the parameters $\alpha, \beta,$ and λ are mixing weights for the smoothing scheme.

4 EXPERIMENT

For evaluation of the proposed patent retrieval method using the semantic annotations, we conducted a set of experiments for patent retrieval tasks on an English patent corpus from USPTO (United States Patent and Trademark Office). The collection includes a set of queries that are the claim section of a patent document to mimic invalidity search. For each query, a set of relevant patent documents were pre-determined. As in the ad hoc retrieval task, different retrieval models were evaluated in terms of precision and recall. To build semantic clusters, we utilized the semantic annotation system in (Kim et al., 2009) whose effectiveness was 76% and 75% in accuracy for PROBLEM and SOLUTION annotations, respectively.

4.1 Experimental Setup

The evaluation of the proposed semantic patent retrieval method was done with a test set extended from the NTCIR-6 patent retrieval test set, which consists of 981,948 USPTO patent documents published from 1993 to 2000, covering 3,221 topics that are basically the granted USPTO patents during the period from 2000 to 2001. Among these, we randomly selected 100 topic patents, and each topic patent contains the sections of title, claim, citation, date, and other patent components. Among those components, we used only claim parts as actual queries, i.e., query terms are only from the claim parts, and this is the same condition as that of the NTCIR-6 patent retrieval participants did (Fujii et al., 2007b).

In the NTCIR collection, there are two different relevance judgment sets. *Type A* means that relevant patent documents must be cited among each other but their IPC subclasses are not necessarily identical. *Type B* means that relevant patent documents are not only cited among each other but also under the same

IPC subclass. In our experiments, we used *Type A* because the answers for *Type B* were too sparse – only two or three patent documents are marked as relevant per query. Further details of the gold standard for the experiments are specified in (Fujii et al., 2007b).

4.2 Retrieval Performance

We tested our cluster retrieval model that uses semantic clusters constructed based on PROBLEM and SOLUTION annotations. Our baseline was the Jelinek-Mercer language model as used by Kang et al. (2007). We adopted the model since the general language model approach (Ponte & Croft, 1998) can be problematic with data sparseness. In order to optimize the smoothing parameter λ in the baseline model, we experimented with varying λ values increasing from 0.0 to 0.9. As can be seen in Figure 1, the average precision value is peaked when λ is 0.2 although past research showed that the λ bias over 0.5 would perform the best (Zhai & Lafferty, 2001). We believe this is due to the fact that patent documents are usually long enough to ameliorate some of the sparseness problems. This performance level is quite competitive in comparison with the results reported in (Fujii, 2007b) where most precision values are below 10%. Although the precision values are not directly comparable due to the fact that our queries are only a subset of the queries in the collection, it indicates that the system based on the Jelinek-Mercer language model can be used as a reasonable baseline, in addition to the other evidence in (Kang et al., 2007).

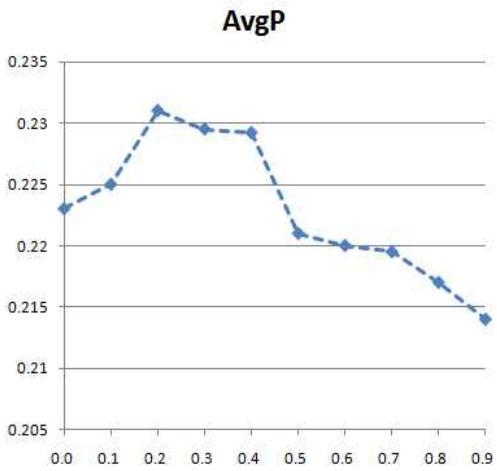

Figure 1. Average precision on each lambda bias

As described in Section 3.2, there are two different smoothing biases, α and β, which were empirically set to 0.1 and 0.3, respectively. Since α is the bias for the semantic cluster that usually has less severe sparseness, it is set to a lower value. We used λ=0.2 as explained above. Table 1 shows the results in MAP (Mean Average Precision) for different cluster types (column) and smoothing methods (rows) (equations *(1)* and *(2)* in Section 3.2) together with the baseline. The four cluster methods correspond to those defined in Section 3.2: PROBLEM based clusters, SOLUTION based clusters, and the union and intersection of both ($CL_p(D)$, $CL_s(D)$, $CL_p(D) \cup CL_s(D)$, and $CL_p(D) \cap CL_s(D)$). Smoothing was applied to both cases: documents only and a combination of documents and clusters.

Table 1: MAP for semantic patent retrieval.

	Prob	Sol	Uni	Inter
Doc $P_{mle}(q\|D)$	**0.281** (*21.7%*)	0.239 (*3.5%*)	0.273 (*18.2%*)	0.232 (*0.4%*)
Doc + Cluster $P'_{mle}(q\|D)$	**0.260** (*14.3%*)	0.238 (*3.0%*)	0.257 (*11.3%*)	0.232 (*0.4%*)
Baseline	0.231 (*0.0%*)			

The best result was obtained when the clusters were constructed based on PROBLEM annotations and smoothing was done for documents only (i.e., equation **(1)**), whereas the improvement made by SOULTION annotations was almost negligible. Combining two sets of clusters with a union or an intersection operation was no better than using the PROBLEM clusters only.

These results coincide with the discussion in (Kim et al., 2009). Many PROBLEM keywords are shared by those patent documents that are linked by citations. In addition, PROBLEM keywords are likely to represent the key theme of the patents containing them. On the other hand, the SOLUTION clusters are far less useful because patent documents rarely share the same SOLUTION keywords for the same PROBLEM. This is due to the fact that in order for a patent to be granted, the solution must be different from those already patented.

It was also found that a key term can rarely be annotated as both PROBLEM and SOLUTION even across patent documents although it is possible. This explains why the intersection case is almost the same as the baseline case. The difference between the two cases in smoothing seems to be due to the fact that sparseness in a unigram language model is less problematic in patent documents as they are

relatively long. In addition, smoothing clusters seems to do more harm than good because they contain enough words.

To better understand how semantic annotations help patent retrieval, we compared the best case (PROBLEM-based clustering with the document smoothing) with the baseline for individual queries. For each query *q*, a comparison was made by computing the ratio using MAP as follows:

$$r(q) = \frac{best(q) - base(q)}{base(q)} \times 100$$

where *best* is the best system's MAP for *q*

base is the baseline's MAP for *q*

In Figure 2, the x-axis and y-axis represent individual queries and percent increase values, respectively. It can be seen that except for the two queries below the x-axis and three on the axis, all the queries show improvements, the majority of which range between 20% and 45%. In those exceptional cases with no improvement, most relevant documents were not included in the clusters partly because there are only a very small number of PROBLEM and SOLUTION keywords in the collection that no clusters were formed. In other words, as long as semantic clusters were formed (and in most cases, semantic clusters were indeed formed), the semantic annotations, especially the PROBLEM annotations, helped improving retrieval effectiveness. This is strong evidence that our semantic retrieval model is effective for the task of patent retrieval.

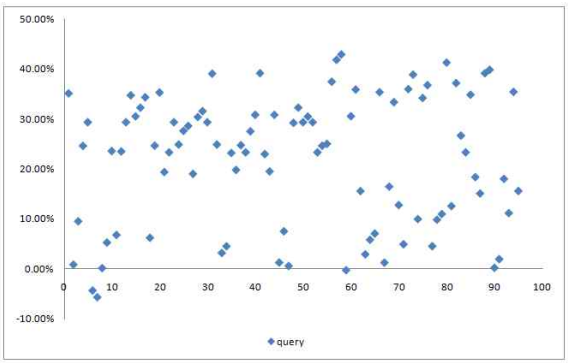

Figure 2. Improvements for individual queries when the PROBLEM clusters were used

5 CONCLUSION

In this paper, we proposed a novel patent retrieval method exploiting two kinds of semantic annotations, which identifies key terms for PROBLEM and SOLUTION categories in patent analysis. The annotation method is borrowed from a recent patent analysis work. Based on the identified PROBLEM and SOLUTION annotations, we investigated on the effectiveness of using such semantic annotations for patent retrieval. We proposed a new semantic clustering method based on the PROBLEM and SOLUTION key phrases whose occurrences make two documents belong to the same cluster. A cluster-based retrieval method was adapted to our situation by adding semantic cluster information to a conventional language model based retrieval method. The experimental result shows that the annotation based semantic retrieval method in fact improves retrieval effectiveness significantly, making it possible to conclude that our semantic retrieval method is desirable for enhancing retrieval performance in patent retrieval.

For future work, we plan to identify additional semantic categories that can help patent retrieval. Once they are identified, the same model can be used to improve patent retrieval further. As a shorter term plan, we consider to compare and combine the semantic clustering with conventional document clustering based on words. We also plan to devise and test different ways of incorporating semantic annotations for retrieval, such as incorporating semantically annotated phases as index terms and conducting two-level retrieval, one for PROBEM-based retrieval and the other for SOLUTION-based matching, especially for an invalidity search task.

ACKNOWLEDGEMENTS

This work was supported by Microsoft Research Asia and the IT R&D program of MKE/IITA [2008-F-047-02, Development of Urban Computing Middleware].

REFERENCES

Ahmad, K., Al-Thubaity, A. 2003. Can text analysis tell us something about technology progress? In *Proceedings of the ACL-03 workshop on patent corpus processing*, pages 41-45.

Callan, J., Ku, Z., Croft, B. 1995. Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '95)*, pages 21-28.

Chen, L., Tokuda, N., Adachi, H. 2003. A patent document retrieval system addressing both semantic and syntactic properties. In *Proceedings of the ACL-03 workshop on patent corpus processing*, pages 1-6.

Chen, L., Tokuda, N. 2003. Robustness of regional matching scheme over global matching scheme. *Artificial Intelligence*, Vol. 144(1-2), pages 213–232.

Croft, B. 1980. A model of cluster searching based on classification. *Information Systems*, Vol. 5, pages 189–195.

Fujii, A., Iwayama, M., Kando, N. 2004. Overview of patent retrieval task at NTCIR-4. In *Proceedings of NTCIR-4 Workshop Meeting*, pages 225-232.

Fujii, A., Iwayama, M., Kando, N. 2007a. Introduction to the special issue on patent processing. *Information Processing & Management*, Vol .43 (5), pages 1149-1153.

Fujii, A., Iwayama, M., Kando, N. 2007b. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 359-365.

Fujii, A. 2007. Integrating content and citation information for the NTCIR-6 patent retrieval task. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 377-380.

Iwayama, M., Fujii, A., Kando, N., Marukawa, Y. 2003. An empirical study on retrieval models for different document genres: patents and newspaper articles. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '03)*, pages 251–258.

Itoh , H., Mano, H., Ogawa, Y. 2003. Term distillation in patent retrieval. In *Proceedings of the ACL-03 workshop on patent corpus processing*, pages 41-45.

Kando, N. 2004. Overview of the Forth NTCIR Workshop. In *Proceedings of 4th NTCIR Evaluation Workshop*, pages 1-9.

Kando, N. 2005. Overview of the Fifth NTCIR Workshop. In *Proceedings of 5th NTCIR Evaluation Workshop*, pages 1-9.

Kando, N. 2007. Overview of the Sixth NTCIR Workshop. In *Proceedings of 6th NTCIR Evaluation Workshop*, pages 1-9.

Kang, I-S., Na, S-H., Kim, J. 2007. Cluster-based patent retrieval. *Information Processing & Management*, Vol .43 (5) pages 1173-1182.

Kim, Y., Tian, Y., Jeong, Y., Ryu, J., Myaeng, S-H. 2009. Automatic discovery of technology trends from patent text. In *Proceedings of the 24th Symposium on Applied Computing (SAC '09)*. pages. 1480-1487.

Klein, D., Manning, C. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL '03)* pages 423-430.

Konishi, K., Kitauchi, A., Takaki, T. 2004. Invalidity patent search system of NTT data. In *Proceedings of NTCIR-4 Workshop Meeting*, pages 250-255.

Larkey, L. 1999. A patent search and classification system. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 179-187.

Liu, X., Croft, B. 2004. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '04)*, pages 186-193.

Ponte, J. M., Croft, W. B. 1998. A language modelling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '98)* pages 275–281.

Tait, J. 2008. *Proceeding of the 1st ACM workshop on patent information retrieval*, Publication dept. ACM, Inc. Denver, MA, US.

Takaki, T., Fujii, A., Ishikawa, T. 2004. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In *Proceedings of the 13th ACM international conference on Information and Knowledge Management (CIKM '04)* pages 399-406.

Takeuchi, K., Collier, N. 2003. Biomedical entity extraction using Support Vector Machines. In *Proceedings of the ACL-03 workshop on natural language processing in biomedicine* pages 57-64.

van Rijsbergen, C. J. 1979. Information retrieval. Newton, MA: Butterworth-Heinemann.

Shinmori, A., Okumura, M., Marukawa, Y., Iwayama, M. 2003. Patent claim processing for readability: structure analysis and term explanation. In *Proceedings of the ACL-03 workshop on patent corpus processing*, pages 56–65.

Voorhees, E. 1985. The cluster hypothesis revisited. In *Proceedings of the 8th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '85)*, pages 188–196.

Yoon, B., Park, Y. 2004. A text mining-based patent network: analytical tool for high-technology trend. *Journal of High Technology Management Research,* Vol. 15 (1), pages 37–50.

Zhai, C., Lafferty, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '01)* pages 334–342.