Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1173–1182
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Cluster-based patent retrieval

In-Su Kang [a,*], Seung-Hoon Na [b], Jungi Kim [b], Jong-Hyeok Lee [b]

[a] *Korea Institute of Science and Technology Information, Pohang University of Science and Technology (POSTECH), Advanced Information Technology Research Center (AITrc), Republic of Korea*
[b] *Division of Electrical and Computer Engineering, Pohang University of Science and Technology (POSTECH), Advanced Information Technology Research Center (AITrc), Republic of Korea*

Received 1 September 2005; accepted 29 May 2006
Available online 16 January 2007

Abstract

Through the recent NTCIR workshops, patent retrieval casts many challenging issues to information retrieval community. Unlike newspaper articles, patent documents are very long and well structured. These characteristics raise the necessity to reassess existing retrieval techniques that have been mainly developed for structure-less and short documents such as newspapers. This study investigates cluster-based retrieval in the context of invalidity search task of patent retrieval. Cluster-based retrieval assumes that clusters would provide additional evidence to match user's information need. Thus far, cluster-based retrieval approaches have relied on automatically-created clusters. Fortunately, all patents have manually-assigned cluster information, international patent classification codes. International patent classification is a standard taxonomy for classifying patents, and has currently about 69,000 nodes which are organized into a five-level hierarchical system. Thus, patent documents could provide the best test bed to develop and evaluate cluster-based retrieval techniques. Experiments using the NTCIR-4 patent collection showed that the cluster-based language model could be helpful to improving the cluster-less baseline language model.

Keywords: Cluster-based retrieval; Patent retrieval; Invalidity search; International patent classification

1. Introduction

Through the recent NTCIR workshops (Fujii, Iwayama, & Kando, 2004, 2005; Iwayama, Fujii, Kando, & Takano, 2002) which has started the evaluation of retrieving patent documents, *patent retrieval* casts new challenging issues to information retrieval community. *Invalidity search*, which has newly emerged in patent retrieval, is to retrieve published or registered patents that contain some conflicting claim parts that are enough to

* Corresponding author. Address: PIRL 323, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang 790-784, Republic of Korea. Tel.: +82 54 279 5656; fax: +82 54 279 5699.
E-mail addresses: dbaisk@postech.ac.kr (I.-S. Kang), nsh1979@postech.ac.kr (S.-H. Na), yangpa@postech.ac.kr (J. Kim), jhlee@postech.ac.kr (J.-H. Lee).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.11.006

invalidate (or reject) a new patent application. Unlike traditional document retrieval, invalidity search has the problem of modeling patentability, which is very difficult to quantitatively formalize like modeling document relevance to a query. Actually, this search task is manually performed by patent examiners in national patent offices to find a prior art of a patent application. Considering the today's explosively growing number of patent applications, the attempt to automate invalidity search is crucial.

Unlike newspaper articles which account for most of the modern large-scale test collections for information retrieval, patent documents have many different characteristics. They are structurally well formed. A typical patent document consists of a title, an abstract, a claim, a detailed description as well as bibliographic information. In addition, according to (Iwayama, Fujii, Kando, & Marukawa, 2003), the average length of patent documents is about 24 times larger than that of newspaper articles, and the standard variance of the length of patent documents is approximately 20 times larger than that of newspapers. Moreover, all patents have manually-assigned cluster information, international patent classification (IPC) codes. IPC is a standard taxonomy for classifying patents, and has currently about 69,000 nodes which are organized into a five-level hierarchical system. The above characteristics raise the necessity to reassess existing retrieval techniques on patent documents. Among these, this study focuses on the fact that patent documents are already manually clustered according to IPC, and we attempts to apply the cluster-based retrieval techniques to the invalidity search task of patent documents to improve retrieval effectiveness.

Previous studies on cluster-based retrieval relied on automatically-generated clusters rather than manually-assigned clusters as in the case of patent documents. The fact that the long history of cluster-based retrieval thus far has not yet obtained conclusive results on whether it affects the retrieval effectiveness positively or not can be partially contributed to the structure, granularity, and correctness of the automatically-generated clusters. IPC is consistently maintained by an authorized official organization, and IPC codes are assigned to patent documents manually by technical specialists such as patent examiners. Thus, a collection of patent documents could provide the ideal test bed to develop and evaluate cluster-based retrieval techniques. This study explores the influence of manual clusters provided by IPC labels inherent in patent documents on the retrieval effectiveness. In particular, we rely on the language modeling approach to incorporate the cluster information into a retrieval function.

The remainder of the paper is organized as follows. Section 2 gives related works about cluster-based retrieval. Section 3 describes the international patent classification (IPC) system. Section 4 explains the cluster-based language models for searching patent documents. Experimental evaluations are given in Section 5. Concluding remarks are given in Section 6.

2. Related Work

2.1. Cluster-less patent retrieval

There have been some attempts (Larkey, 1999; Osborn, Strzalkowski, & Marinescu, 1997) to search patent documents before handling patent documents in NTCIR workshops. Osborn et al. (1997) evaluated the retrieval effectiveness of searching patent documents using the SMART system. However, the evaluation of Osborn et al. (1997) was different from that of the invalidity search task, since Osborn et al. considered the citations of a patent document as its relevance judgment. Larkey (1999) divided a whole collection of patent documents into n sub-collections according to their IPC codes. All documents in each sub-collection were merged into a single large virtual document. Then, a query was matched to each of n virtual documents to select the best m ($\ll n$) collections from which documents were retrieved using the same query. This type of staged search was borrowed from the collection selection technique of distributed IR (Callan, Ku, & Croft, 1995; Xu & Callan, 1998) in order to avoid prohibitively expensive search time for a large collection of patent documents. Unfortunately, the idea of Larkey (1999) was not experimentally evaluated.

IPC codes assigned to a patent document can be considered as topic labels about the content of the patent document. In other words, IPC codes can be used as terms representing a document. Konishi, Kitauchi, and Takaki (2004) viewed each IPC code assigned to a patent document as a category representing the document, and incorporated TF/IDF of categories as well as TF/IDF of terms into a single retrieval function. Mase, Matsubayashi, Ogawa, Iwayama, and Oshio (2004) utilized IPC codes to filter relevant documents.

First, IPC codes of the top n retrieved documents are gathered to create a set S of topically related IPC codes. Then, the set S was used to filter out the retrieved documents of which IPC codes never overlap with the set S.

2.2. Cluster-based retrieval

Cluster-based approaches to information retrieval utilize document clusters on the basis of the assumption that related documents are grouped into the same cluster, in order to improve efficiency or effectiveness. As for *efficiency*, searching and browsing on clusters rather than the individual documents could drastically reduce the retrieval time of the system and information seeking time of users, respectively. As for *effectiveness*, the studies of cluster-based retrieval starts from the *cluster hypothesis* (van Rijsbergen, 1979) that related documents would help to satisfy the same information need. This study focuses on the effectiveness of the cluster-based retrieval. See Willett (1988) for the excellent and exhaustive review of cluster-based approaches to retrieval.

In a typical cluster-based retrieval, documents are automatically grouped into topically related clusters based on the inter-document similarity which is based on the terms two documents have in common. Then, the score of each retrieved document can be determined *partially* or *exclusively* by the strength of the membership that the document belongs to each cluster. In partial scoring, a document score is determined by the combination of a query-document similarity and a query-cluster similarity, while in exclusive scoring a document score is exclusively determined by a query-cluster similarity. This study relies on partial scoring.

The methods of automatically creating clusters from a set of documents are divided into *agglomerative* and *partitioning* approaches (Rasmussen, 1992; van Rijsbergen, 1979; Willett, 1988). An agglomerative method starts to group two most similar documents into a cluster, and then agglomeratively bind two most similar clusters into a larger cluster until a single largest cluster is obtained. A partitioning method divides the entire documents into a set of predefined number of clusters. Agglomerative methods generate a tree-structured hierarchy of clusters, and partitioning methods a flat-structured list of clusters. While partitioning methods are computationally more tractable for large document collections than agglomerative ones, they have a limitation of having to fix the number of clusters in advance (Willett, 1988). In the case of a hierarchy of clusters, there exist *top-down* and *bottom-up* search techniques and their variants in order to find the best-matching cluster for a query (Croft, 1980; Jardine & van Rijsbergen, 1971; van Rijsbergen, 1974, 1975; Voorhees, 1985).

The set of documents to be clustered can be determined in a *query-independent* or *query-specific* manner. In other words, the entire documents may be statically clustered independent of a particular query, while a subset of retrieved documents to a query may be dynamically grouped into clusters. Several researchers (Hearst & Pedersen, 1996; Lee, Park, & Choi, 2001; Liu & Croft, 2004; Tombros, Villa, & van Rijsbergen, 2002) have witnessed the improvement of the retrieval effectiveness over the cluster-less retrieval through query-specific clustering.

Numerous studies have compared the cluster-less retrieval with the cluster-based retrieval, but unfortunately showing inconsistent findings. Some (Croft, 1980; Hearst & Pedersen, 1996; Jardine & van Rijsbergen, 1971; Lee et al., 2001; Liu & Croft, 2004; Tombros et al., 2002) have reported that the use of clusters helped to obtain better retrieval effectiveness, while others (El-Hamdouchi & Willett, 1989; Voorhees, 1985) have found that the cluster-based retrieval was not successful to improve the cluster-less retrieval.

Recently, Liu and Croft (2004) have demonstrated that the cluster-based retrieval outperformed the cluster-less retrieval for a variety of large document collections using the modern statistical language model. Considering that most past studies have used ad hoc formulas or staged approach to integrate query-cluster similarities into retrieval models, and that many of them have experimented on relatively small document collections, the finding of Liu and Croft (2004) is encouraging enough to foster further exploration of cluster-based language models.

Instead of automatically-generated clusters employed in most cluster-based studies thus far, this study uses IPC clusters within a collection of patent documents which depend on manually-assigned cluster labels (or IPC codes) of patent documents. To the best of our knowledge, no cluster-based retrieval studies have been conducted using a hierarchy of manually-assigned clusters for large document collections.

Section	B	Performing Operations; Transporting
Class	B 64	Aircraft; Aviation; Cosmonautics
Subclass	B 64 C	Aeroplanes; Helicopters (air-cushion vehicles B60V)
Main group	B 64 C 25/00	Alighting gear (air-cushion alighting gear B60V 3/08)
Subgroup	B 64 C 25/02	. undercarriages
Subgroup	B 64 C 25/04	.. arrangement or disposition on aircraft
Subgroup	B 64 C 25/10	... retractable, foldable, or the like

Fig. 1. Example of international patent classification.

3. International patent classification

International patent classification (IPC) is an internationally accepted standard taxonomy for sorting, organizing, disseminating, and searching patents. It is officially administered by World Intellectual Property Organization (WIPO[1]). IPC is a five-level hierarchical taxonomy, covering all technical areas. Top-level nodes consist of eight sections such as human necessities, performing operations, chemistry, textiles, fixed constructions, mechanical engineering, physics, and electricity. A section is divided into classes which are subdivided into subclasses. Subclass is divided into main groups which are further subdivided into subgroups. In total, the current IPC has 8 sections, 120 classes, 630 subclasses, and 69,000 groups (main groups or subgroups).

Fig. 1 shows a part of IPC. Section symbols use uppercase letters A through H. A class symbol consists of a section symbol followed by two-digit numbers such as B00, B64, and B99. A subclass symbol is composed of a class symbol followed by an uppercase letter like B64C. A main group symbol consists of a subclass symbol followed by one to three-digit numbers followed by a slash followed by 00 such as B64C1/00, B64C25/00, and B64C789/00. A subgroup symbol replaces the last 00 in a main group symbol with two-digit numbers except for 00 such as B64C1/83, B64C25/09, and B64C789/01. Each IPC node is attached with a noun phrase description which specifies some technical fields relevant to that IPC code. Note that a subgroup may have more refined subgroups. Hierarchies among subgroups are indicated not by subgroup symbols but by the number of dot symbols in node descriptions as shown in Fig. 1.

All patents are assigned related IPC codes according to their technical fields by human examiners. This means that patent documents are already clustered into a human-organized topic hierarchy. From the viewpoint of cluster-based retrieval, this feature eliminates the overhead of clustering document collections. This study investigates the cluster-based retrieval on patent documents using IPC as a cluster hierarchy.

4. Cluster-based language model

The language modeling approaches to information retrieval assume individual models for documents and views a query as a random sample from each document model (Ponte & Croft, 1998). At retrieval, documents are then generally ranked by the query likelihood that a document model D will generate a given query Q. The simple and common approach of calculating the query likelihood views queries as a sequence of independent terms as shown in Formula (1), where $freq(q)$ is the count of query term q in Q. This multinomial view of document models was chosen by Miller, Leek, and Schwartz (1999), Song and Croft (1999), and Hiemstra (2001).

$$\log P(Q|D) = \log \prod_{q \in Q} P(q|D)^{freq(q)} = \sum_{q \in Q} \log P(q|D)^{freq(q)} \quad (1)$$

Then, the retrieval problem is reduced to estimating a unigram language model for each document. However, the simple maximum likelihood estimation for unigram language models assign zero probabilities to unseen document terms. To avoid this data sparseness problem, the language modeling approach normally employs smoothing techniques among which the simple and popular one is Jelinek–Mercer smoothing (Zhai &

[1] http://www.wipo.int/

Lafferty, 2001) as shown in Formula (2), where *ml* indicates maximum likelihood estimation, and *Coll* means the collection model. See Zhai and Lafferty (2001) for an empirical study of several smoothing methods.

$$P(q|D) = (1-\lambda)P_{ml}(q|D) + \lambda P_{ml}(q|Coll) \tag{2}$$

To incorporate the cluster information obtained from a collection of documents into the above language model, a cluster model is defined for each cluster that corresponds to a particular IPC code in this study. Thus, all documents classified into a specific IPC code are concatenated to create a cluster. Since a single document is associated with two or more IPC codes, a document may belong to several clusters. Intuitively, this means that a document has multiple topics.

Using the cluster models, two types of cluster-based language models can be defined according to how to view cluster models from the viewpoint of statistical language models: *smoothing-oriented* and *topic-oriented*. The smoothing-oriented model considers the cluster models as another component for smoothing document models, while the topic-oriented model regards the cluster models as different topic models about which documents are written. Formulas (3) and (4) show the two cluster-based language models where *cluster*(D) indicates the set of clusters which document D belongs to, and C means the cluster model for a cluster in the set *cluster*(D).

$$P(q|D) = (1-\lambda)\left((1-\alpha)P_{ml}(q|D) + \alpha\frac{\sum_{C\in cluster(D)}P_{ml}(q|C)}{|cluster(D)|}\right) + \lambda P_{ml}(q|Coll) \tag{3}$$

$$\begin{aligned} P(q|D) &= (1-\beta)P(q|D) + \beta\frac{\sum_{C\in cluster(D)}P(q|C)}{|cluster(D)|} \\ &= (1-\beta)((1-\lambda_1)P_{ml}(q|D) + \lambda_1 P_{ml}(q|Coll)) + \beta\frac{\sum_{C\in cluster(D)}((1-\lambda_2)P_{ml}(q|C) + \lambda_2 P_{ml}(q|Coll))}{|cluster(D)|} \end{aligned} \tag{4}$$

In Formula (3) that is the smoothing-oriented model, the document model is first smoothed with the average of cluster models and then further smoothed with the collection model. In Formula (4) that corresponds to the topic-oriented model, the term distribution is determined by a mixture of a document model and the average of its related cluster models. Formula (3) is similar to that of Liu and Croft (2004) except for its generalization to the case of overlapping clusters which means that a document may belong to multiple clusters. Formula (4) can be viewed as a variant of the mixture model of Zhang, Callan, and Minka (2002). The major difference is that in Formula (4) the collection model only plays the role of smoothing while in the work of Zhang et al. (2002) the collection model accounts for a component in the mixture model.

5. Experiments

5.1. Experimental setup

We evaluate the retrieval effectiveness of cluster-based language models on the invalidity search task for patent retrieval using the NTCIR-4 patent test set[2] (Fujii et al., 2004). The document collection consists of 1,707,185 patent documents of unexamined Japanese patent applications published in 1993 through 1997. The test set has a total of 101 search topics which correspond to Japanese patent applications rejected by the Japanese Patent Office. Each search topic is composed of a title, a claim, the date of filing, and the other parts. We used claim parts and the date of filing as queries. Relevant documents of the invalidity search task should be among the prior art which had been open to the public before the topic patent was filed. Thus, we used the date of filing to filter out retrieved documents of which filing date is more than the filing date of the topic patent.

[2] We used the test collection produced for the NTCIR-4 workshop, which is not exactly the same as the final collection (i.e., the "research purpose use of test collection").

NTCIR-4 patent retrieval organizers define two types of relevance judgments: *type A* and *type B*. The documents that can invalidate the demands of all essential components in a target claim are considered as type A answer documents, and the documents that can invalidate the demands of most of the essential components in a target claim are considered as type B. Thus, types A and B roughly correspond to rigid and relaxed relevance. We used type A as relevance judgments for our experiments.

A total of 101 search topics are divided into 34 main topics and 67 additional topics. Among 34 main topics, our experiments used 32 main topics as query files since the two topics had no relevant documents in type A relevance judgment data. Topic numbers of the two excluded topics were 013 and 020.

From the claim parts of documents and topic files, character bi-grams were extracted as index terms and query terms. All retrieval results are reported using the non-interpolated mean average precision which is computed by executing the TREC_EVAL program.

5.2. Experimental results

In recent language modeling approaches, it is generally known that document models should be smoothed a lot with the collection model (Kraaij, Nie, & Michel, 2003). In other words, better performance is observed when λ in Formula (2) is more than 0.5. To see whether this observation applies also to patent documents, we have performed the invalidity search over different values of λ using the Jelinek–Mercer language model which corresponds to Formulas (1) and (2). According to Zhai and Lafferty (2001), smoothing in language modeling approaches has two different roles: *estimation role* and *the role of query modeling*. The first role is to estimate document models from their samples that constitute the target document collection. The second is to reflect the importance of query terms. Our intuition is that patent retrieval would not be dominated by the estimation role, since patent documents are long and cover a diverse spectrum of topicality. From the viewpoint of topicality, note that the claim part in patent documents includes highly specific topic words while the detailed description of patent documents provides more general background information.

As shown in Fig. 2, smoothing was helpful when estimating document language models for patent documents better. As λ increase from 0.01 to 0.1, the performance improves, demonstrating the estimation role of smoothing. In terms of the role of its query modeling, however, the result of Fig. 2 violates the general behavior of query-modeling role that Jelinek–Mercer smoothing shows, considering that topic claims are long. According to Zhai and Lafferty (2001), long and verbose queries need more smoothing, since heavy smoothing could discriminate the common and non-informative terms from informative terms by emphasizing IDF factor in terms of term weighting.

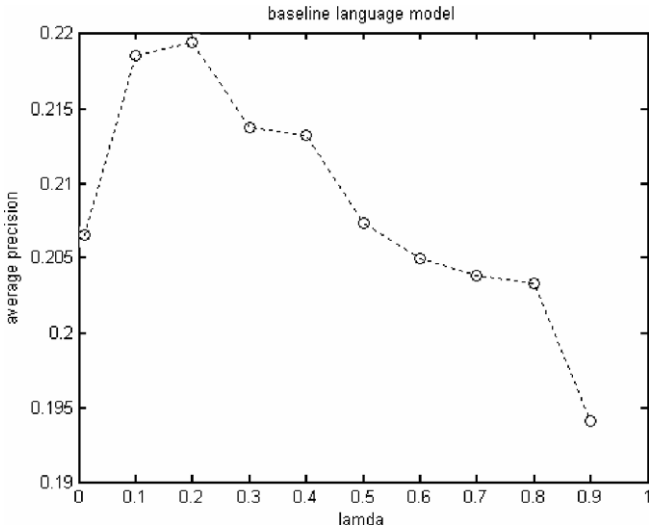

Fig. 2. Effect of smoothing in patent documents (*lamda* indicates λ in Formula (2)).

Currently, we believe that this results from the characteristics of patent documents and the invalidity search. The first characteristic is as follows. As mentioned in Section 1, the average length of patent documents is about 24 times larger than that of newspaper articles (Iwayama et al., 2003). Thus, the degree of smoothing of document models for a patent collection is expected to be smaller than that of document models for much shorter newspaper articles. The second characteristic is that although the claim part is long, but it is not verbose. This means that the differentiation of claim terms by heavy smoothing could be harmful to obtaining better query model in the invalidity search of patent retrieval.

Figs. 3 and 4, where *lamda*, *alpha*, and *beta* indicates λ, α, and β in Formulas (3) and (4), show the retrieval results of the two types of cluster-based retrieval models: smoothing-oriented and topic-oriented, respectively. In the experiments of the two figures, the lowest-level IPC codes (that is IPC level 5) were used to create cluster models. The comparison among the cluster-based retrievals based on the uses of different levels of IPC codes is provided later.

Fig. 3 traces the retrieval effectiveness of a smoothing-oriented cluster-based model over varying values of λ and α. As α increases, the performance starts to gradually decrease and then sharply drops independent of a particular value of λ. The best performance was 21.85 for $\lambda = 0.1$ and $\alpha = 0.2$, not outperforming that (21.93) of the baseline language model. This means that the use of cluster models for smoothing does not improve term distribution of document models. In other words, it is hard to find a unique role of cluster models that is different from that of the collection model in terms of smoothing.

Fig. 4 shows the retrieval effectiveness of a topic-oriented cluster-based model over varying values of λ and β. Although Formula (4) defines two smoothing parameters λ_1 and λ_2, we simplified the two into the same parameter λ for simplicity. As with the smoothing-oriented one, the performance of a topic-oriented model declines with the more emphasis on cluster models. However, the performance change was more sensitive to the cluster models than the case of Fig. 3. The best performance of a topic-oriented model was 22.57 for $\lambda = 0.1$ and $\beta = 0.2$, slightly outperforming that (21.93) of the baseline language model.

Table 1 shows the effect of different levels of the IPC hierarchy on retrieval effectiveness using the topic-oriented cluster model of Formula (4). IPC levels 1 through 5 correspond respectively to sections, classes, subclasses, main groups, and subgroups in IPC. Thus, IPC level 5 means the use of bottom-level clusters in the IPC hierarchy. As shown in Table 1, the use of the most specific bottom-level clusters enables better retrieval effectiveness than the baseline language model, although the improvement is marginal. As upper-level cluster models are used, the performance of the cluster-based language model drastically drops. We believe that this is due to the heavy smoothing added by cluster models, since higher-level cluster models resembles the collection model, obscuring the topicality of a topic claim.

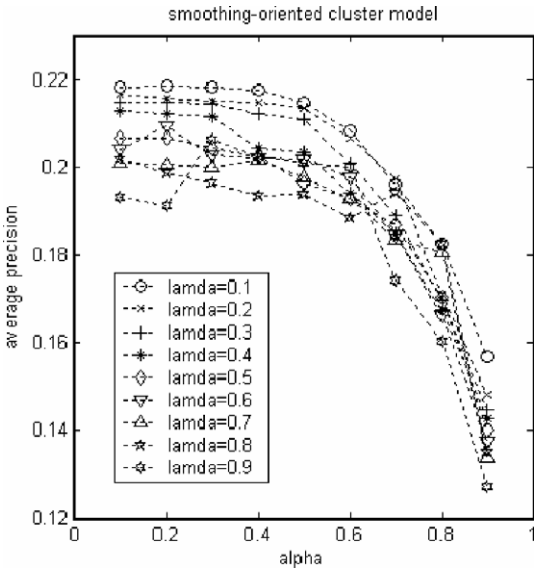

Fig. 3. Evaluation of a smoothing-oriented cluster model (*lamda* and *alpha* indicate λ and α in Formula (3)).

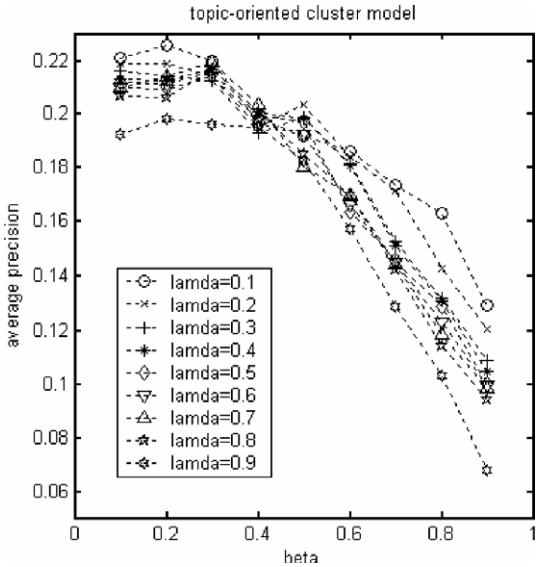

Fig. 4. Evaluation of a topic-oriented cluster model (*lamda* and *beta* indicate λ_1 (or λ_2) and α in Formula (4)).

Table 1
Effect of different levels of the IPC hierarchy on retrieval effectiveness

Retrieval model	Mean average precision	Number of clusters
Baseline language model	21.93	
Topic-oriented cluster model ($\lambda = 0.1$, $\beta = 0.2$)		
IPC level 1	0.25	8
IPC level 2	2.38	118
IPC level 3	5.80	613
IPC level 4	21.91	6133
IPC level 5	22.57	42,239

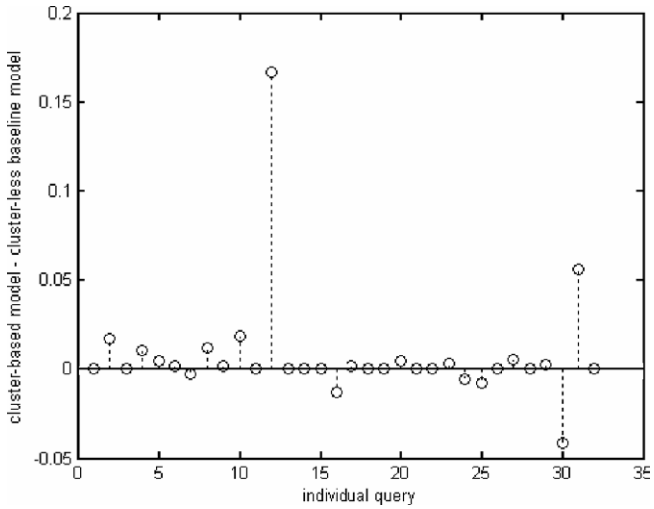

Fig. 5. A query-by-query comparison between a topic-oriented cluster-based model and the cluster-less baseline model (each circle indicates the difference between MAP values of Formula (4) and (2) for each individual query).

Fig. 5 compares a topic-oriented cluster-based model with the cluster-less baseline model by showing the difference between their MAP values for each individual query. In the figure, 1 to 32 query numbers on the *x*-axis correspond to the ascending order of the main topic numbers excluding two topic numbers 013 and

020 which had no relevant documents in type A relevance judgment. The figure shows that the use of cluster information improves the retrieval effectiveness of the cluster-less model for many queries although the difference is not large for most queries. In other words, the cluster-based model outperformed the baseline model for 50% of queries.

In summary, although cluster information (that is, IPC codes) within patent documents was human-supplied correct one, its exploitation in the form of cluster models in the language modeling approach did not bring a substantial success. The reason is partially related to the fact that the invalidity search task has relatively the small number of relevant documents. In the invalidity search task in reality, a few patent documents which invalidate the claim part of a patent application are sufficient. Actually, most patent examiners stop searching the prior art when they obtain at most two or three invalidating patents. Actually, the average number of relevant document was 4.97 per a query for our 32 main topics of the NTCIR-4 patent test set. The success of the cluster-based retrieval depends on the possibility that the relevant documents to a query are grouped into some clusters. Such an odd could increase a lot if a query has more relevant documents.

6. Conclusion

In this paper, we have applied a cluster-based language model to the invalidity search task of patent retrieval. From the experimental evaluations using the NTCIR-4 patent document collection, our findings are as follows. First, in terms of smoothing of the baseline language model, it can be harmful to smooth document models heavily with the collection model in the case of patent retrieval, because topic claims are long but not verbose. This observation leads us to the further study on the effect of other smoothing techniques for patent documents. Second, the cluster-based language model can be helpful to improve the retrieval effectiveness of the baseline language model. Unfortunately, however, the finding was not conclusive since the improvement was marginal. Our recent analysis has shown that the size of IPC level 5 clusters is highly variable, ranging from 1 to 26,361 documents. The size of a cluster means the number of documents which belong to the same cluster. In the future, we plan to explore the techniques to balance the size of clusters such as pruning the large clusters of more than a certain threshold, or automatically partitioning the large clusters into smaller clusters.

Acknowledgements

This work was supported by the KOSEF through the Advanced Information Technology Research Center (AITrc) and by the BK21 project.

References

Callan, J., Ku, Z., & Croft, W. 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* (pp. 21–28).

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

El-Hamdouchi, A., & Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal, 32*(3), 220–227.

Fujii, A., Iwayama, M., & Kando, N. (2004). Overview of patent retrieval task at NTCIR-4. In *Working notes of the fourth NTCIR workshop meeting* (pp. 225–232).

Fujii, A., Iwayama, M., & Kando, N. (2005). Overview of patent retrieval task at NTCIR-5. In *Proceedings of the fifth NTCIR workshop*.

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 76–84).

Hiemstra, D. (2001). Using language models for information retrieval. *PhD thesis*, University of Twente.

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* (pp. 251–258).

Iwayama, M., Fujii, A., Kando, N., & Takano, A. (2002). Overview of patent retrieval task at NTCIR-3. In *Working notes of the third NTCIR workshop meeting* (pp. 1–10).

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval, 7*, 217–240.

Konishi, K., Kitauchi, A., & Takaki, T. (2004). Invalidity patent search system of NTT data. In *Working notes of the fourth NTCIR workshop meeting* (pp. 250–255).

Kraaij, W., Nie, J. Y., & Michel, S. (2003). Embedding Web-based statistical translation models in cross-language information retrieval. *Computational Linguistics, 29*, 1–37.

Larkey, L. S. (1999). A patent search and classification system. In *Proceedings of the fourth ACM conference on digital libraries* (pp. 179–187).

Lee, K. S., Park, Y. C., & Choi, K. S. (2001). Re-ranking model based on document clusters. *Information Processing and Management, 37*, 1–14.

Liu, X., & Croft, W.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* (pp. 186–193).

Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., & Oshio, T. (2004). Two-stage patent retrieval method considering claim structure. In *Working notes of the fourth NTCIR workshop meeting* (pp. 256–261).

Miller, D., Leek, T., & Schwartz, R. (1999). A hidden Markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 214–221).

Osborn, M., Strzalkowski, T., & Marinescu, M. (1997). Evaluating document retrieval in patent database: a preliminary report. In *Proceedings of the conference on information and knowledge management* (pp. 216–221).

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

Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms*. Englewood Cliffs, NJ: Prentice-Hall.

Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 279–280).

Tombros, A., Villa, R., & van Rijsbergen, C. J. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management, 38*, 559–582.

van Rijsbergen, C. J. (1974). Further experiments with hierarchic clustering in document retrieval. *Information Storage and Retrieval, 10*, 1–14.

van Rijsbergen, C. J. (1975). Document clustering: an evaluation of some experiments with the Cranfield 1400 collection. *Information Processing and Management, 11*, 171–182.

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

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

Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management, 24*(5), 577–597.

Xu, J., & Callan, J. (1998). Effective retrieval with distributed collections. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 112–120).

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* (pp. 334–342).

Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 81–88).