Cluster-Based Patent Retrieval Using International Patent Classification System

Jungi Kim[1], In-Su Kang[2], and Jong-Hyeok Lee[1]

[1] Division of Electrical and Computer Engineering
Pohang University of Science and Technology (POSTECH)
Advanced Information Technology Research Center (AITrc)
{yangpa, jhlee}@postech.ac.kr
[2] Information System Research Laboratory
Korea Institute of Science and Technology Information (KISTI)
dbaisk@kisti.re.kr

Abstract. A patent collection provides a great test-bed for cluster-based information retrieval. International Patent Classification (IPC) system provides a hierarchical taxonomy with 5 levels of specificity. We regard IPC codes of patent applications as cluster information, manually assigned by patent officers according to their subjects. Such manual cluster provides advantages over automatically built clusters using document term similarities. There are previous researches that successfully apply cluster-based retrieval models using language modeling. We develop cluster-based language models that employ advantages of having manually clustered documents.

Keywords: cluster-based retrieval, patent retrieval, invalidity search, international patent classification.

1 Introduction

Patent applications have different characteristics from other document types such as newspapers or web documents. Patents are generally very long and verbose, and their sizes are much more variable (Iwayama et al., 2003). Patent applications are well-structured and size of the collection is enormous; there are about 5 million U.S. patents, or 100-200 gigabytes of text, which are made up of hundred fields of textual or non-textual information (Larkey, 1998). One can take advantage of its structure for better retrieval or use it as a realistic-sized test collection. Also, patents provide a great test bed for exploring new ideas for manual clusters. Patent applications have one or more manually assigned International Patent Classification (IPC). Potential usefulness of using clusters for information retrieval has long been suggested and explored with no conclusive findings of its benefits (Liu and Croft, 2004). It is only recent that some promising results of cluster-based retrieval using statistical language modeling are reported (Kurland and Lee, 2004; Liu and Croft, 2004).

Kang et al. (2006) is the first to use IPC system as manual clusters for searching patent documents. They define two roles of cluster model: smoothing-oriented and topic-oriented. Their cluster model based on statistical language modeling is used either to smooth document language model or as an independent topic model which is

Y. Matsumoto et al. (Eds.): ICCPOL 2006, LNAI 4285, pp. 205 – 212, 2006.

interpolated with document model for final scores of retrieved documents. They report smoothing document model with cluster model does increase the retrieval performance marginally, and more gain in performance is obtained by interpolating document and cluster model.

Our work extends Kang et al. (2006)'s topic-oriented model with consideration of the characteristics of manually clustered documents. We show and discuss the cluster characteristics obtained by statistically analyzing the corpus. Then, we propose new models that incorporate the information of cluster size and similarity of manually built clusters into a cluster-based retrieval model. Finally, we present and discuss our experimental results and our conclusion.

2 International Patent Classification

2.1 International Patent Classification System

International Patent Classification System (IPC) is a 5-level hierarchical taxonomy for sorting patent applications administered by World Intellectual Property Organization (WIPO). The 5 levels of IPC are: section, class, subclass, main group, and subgroup. A patent application is hand-assigned to one or more appropriate IPC codes by human examiners. There are many overlapping categories in IPC taxonomy, and it is possible that an application has many IPC codes of different subgroups, groups, subclasses, classes, or sections. We consider IPC at each level as a cluster of documents, although for our experiments we only use IPC Code at level 5.

2.2 Statistics of IPC Clusters

To better view the characteristics of manual cluster of the patent applications and to compare with that of automatically built clusters, statistics of IPC cluster size and cluster memberships of patent documents are collected from the NTCIR-4 patent document collection.

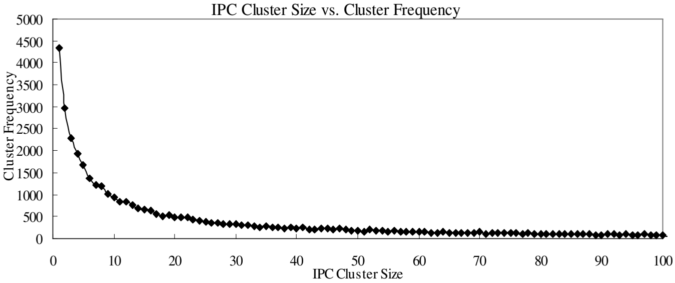

Fig. 1. IPC Cluster Size vs. Cluster Frequency at IPC level 5

The size and the document memberships of IPC clusters are very different from that of automatic clusters over which we generally have controlled sizes or a fixed number of clusters a document can belong.

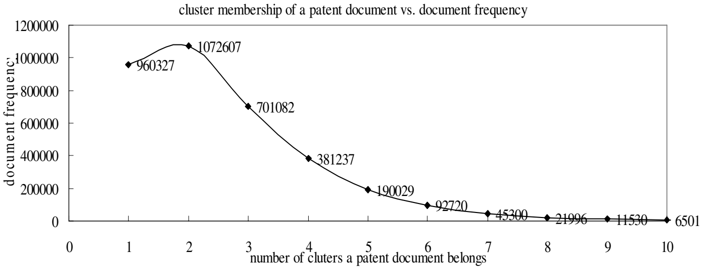

Fig. 2. Cluster Membership of Document vs. Document Frequency at IPC level 5

Figure 1 shows frequencies of different IPC cluster sizes at IPC level 5. Cluster size varies from 1 to 85355 documents, and the average number of documents per cluster is about 172 with the standard deviation of 859. Statistics and distribution of cluster sizes at different IPC level of a different patent collection are also available in the literature written by Fall et al. (2003).

Figure 2 shows how many IPC codes a document belongs and frequencies of such documents. A patent document has at least 1 IPC code and at most 91 IPC codes. On average, a patent document belongs to 2.6 IPC clusters at level 5, and the standard deviation is about 3.

Distributions of cluster size and cluster membership of documents are similar to that of Zipf's: few occur very often, while others occur rarely.

3 Cluster Characteristics

3.1 Automatic Clustering

Many researchers have tried different clustering methods to create automatic clusters of documents, among which are agglomerative algorithms that belong in a hierarchical clustering (van Rijsbergen and Croft, 1975), partitional clustering such as k-means (Liu and Croft, 2004), or different methods that uses common classification algorithms such as k-Nearest Neighbors algorithms (Kurland and Lee, 2004).

There has not been any research on the characteristics of automatically built clusters. However, from the algorithmic and the probabilistic-point of view, we can expect the properties for some of the methods.

For k-Means algorithm, if we assume a document has an equal chance to be assigned to any of the clusters, size of a cluster should follow a binomial distribution. Each document has probability of 1/k and there are n independent trials where n is the number of documents, hence the binomial distribution B (n, p=1/k) describes the distribution of cluster sizes and gives the expected mean of n * 1/k, or n/k, with the variance of n * 1/k * (n–1)/k, or $(n^2-n)/k^2$.

k-Nearest Neighbors algorithms has a fixed size of k + 1 for a cluster. If a cluster is built for every document, there are n numbers of overlapping clusters.

Similarity measure plays an important role in the quality of automatically built clusters. Automatic clustering uses measures of inter-document similarity such as

Dice's coefficient, Jaccard's coefficient, Cosine coefficient, and Kullback-Liebler (KL) divergence. However, such measures solely depend on the presence of terms of the documents in consideration.

3.2 Manual Clustering

In section 2.2, we show that IPC cluster size and cluster membership of documents have Zipf-like distributions. Cluster size noticeably affects the cluster-based retrieval performance according to Kurland and Lee (2004). For their experiments, while small sizes of clusters (4 or 5) do better than baseline cluster-less model, cluster size 40 gives superior retrieval performance.

We argue that cluster size is also important in manual cluster-based retrievals and propose methods to normalize cluster size. One way is to make an artificial cluster with a pre-determined size; if a cluster is too small, expand the cluster by merging with neighbor clusters, in case of IPC clusters, merging into a higher level, and if the cluster is too big, in result of sizing up or in its original state, use automatic clustering techniques to separate the target number of most relevant documents. Doing so can take advantages of manual clustering and automatic clustering: an accurate relevance measure and the control over cluster sizes. Another simple but crude method is to use only clusters of wanted sizes and ignore the rest.

Unlike automatically built clusters, IPC clusters are clustered based on topic, not terms. Each document is hand-assigned to its appropriate clusters based on its subject. Although we may not directly know how the inter-document similarity is measured, we can infer the similarities of clusters using cluster overlap size: the number of documents clusters share. Clusters of more similar topics should have more documents in common. For normalization purposes on cluster size, we use Dice's coefficient for similarity score.

4 Cluster-Based Language Models

4.1 Cluster-Less Language Model

We set cluster-less language model as our baseline to compare cluster-less model with cluster-based model. We take the performance of Kang et al.'s Jelinek-Mercer Smoothed unigram maximum likelihood language model:

$$P_{LM}(Q \mid D) = \prod_{q \in Q} \left[(1-\lambda)P_{ml}(q \mid D) + \lambda P_{ml}(q \mid Coll)\right]^{freq(q)} \tag{1}$$

where Q and D, coll are query, document, and collection model respectively, λ is a smoothing parameter for Jelinek-Mercer smoothing, and q is a query term. P_{ml} indicates the probability induced from maximum likelihood and freq(q) is the number of times a term q appears in D. Best retrieval performance, 21.93 in MAP, is obtained when λ =0.2.

4.2 Interpolation Model

Kang et al. (2006) defines two cluster-based models: smoothing-oriented and topic-oriented. Each model uses the cluster model generated by language modeling for different purposes. Smoothing-oriented model uses cluster model for smoothing document language model before smoothing with collection model, while topic-oriented model interpolates document model and cluster model after smoothing each model with collection model.

Our cluster-based models are extended version of the topic-oriented model which performed better than smoothing-oriented model.

Topic-oriented model, which we will refer to as interpolation model is:

$$P_{IM}(Q \mid D) = \prod_{q \in Q} \left[(1-\beta) P_{LM}(q \mid D) + \beta \frac{\sum_{C \in cluster(D)} P_{LM}(q \mid C)}{|cluster(D)|} \right]^{freq(q)} \quad (2)$$

and

$$P_{LM}(q \mid D) = (1-\lambda_1) P_{ml}(q \mid D) + \lambda_1 P_{ml}(q \mid Coll), \quad (3)$$

$$P_{LM}(q \mid C) = (1-\lambda_2) P_{ml}(q \mid C) + \lambda_2 P_{ml}(q \mid Coll), \quad (4)$$

where β is an interpolation parameter, C is a cluster, and cluster(D) is a set of clusters that document D belongs.

For our experiments, both λ1 and λ2 are set to 0.2, and β to 0.1 which give the best performance in Kang et al.'s work for the interpolation model. Following models extends the interpolation model and have the same parameter settings.

4.3 Cluster Size-Limit Model

As Kurland and Lee (2004) have done, we try to prevent too many irrelevant documents from being added by restraining on cluster size. As described in 3.2, the best way to achieve this is to control the degree of relevance, one of which being the number of documents in a cluster. It is possible to normalize the sizes of manual clusters as described in 3.2, however, since implementing and carrying out such method is complicated and takes a long time, we simply add a size-limit parameter so that any clusters having size larger than the parameter are ignored entirely. Relevant documents in large clusters may not be benefited. Nonetheless, for experimental purposes, it should be sufficient enough to show how cluster size affects the retrieval performance.

4.4 Cluster Expansion Model

We infer the cluster similarity information from the corpus as described in section 3.2. We expand initially retrieved clusters by adding similar clusters based on their topics. The score of the added cluster is calculated averaging the scores of clusters that expand it. Since initial retrieval scores about 40,000 IPC clusters out of 42239 clusters, for expansion, we limit the number of clusters we expand. There are two parameters: the number of top clusters from which clusters are expanded, and the number of clusters to expand from each top cluster.

Experimental Setup. For our test collection, we use NTCIR-4 patent which contains 1,707,185 Japanese patent applications. The test collection has 101 search topics which are patent applications rejected by Japanese Patent Office, of which 32 main topics are used for our experiments.

Among various sections a patent application has, claim, date of filing, and detailed description are of our interest. Claim sections and dates of filing of rejected patent applications are used as queries. Claim, date of filing, and detailed description sections are extracted to represent documents.

Relevant judgment set are prior arts that invalidates the topic patent application. Hence, only the documents that are filed before the topic application can appear in relevant document set.

For index and query terms, character bigrams of Japanese, numbers, and English words are used.

5 Experimental Results

5.1 Size-Limit Model

As table 1 shows, the cluster size plays an important role in cluster-based retrieval. Retrieval performance using only clusters with small sizes stayed around that of cluster-less baseline. With the increasing cluster sizes, however, MAP fluctuates quite a bit, indicating, as the number of retrieved relevant documents shows, that some informative clusters are added at some size-limit value, but irrelevant documents are brought in with increased size-limit.

5.2 Cluster Expansion Model

We expected Cluster Expansion Model to perform well. However, it performed even worse than Cluster-less Model. The poor performance can be explained in several

Table 1. Performance of Size-Limit Model and reported in MAP and number of relevant documents retrieved

Cluster Size-Limit	MAP	Rel. Ret.
10	0.2187	117
50	0.212	117
100	0.2171	117
300	0.2296	120
500	0.2137	121
700	0.2135	123
1000	0.2325	123
1500	0.2278	123
2000	0.2283	122
2500	0.2276	122
3000	0.2264	119
3500	0.2264	119
4000	0.2265	120
4500	0.2266	120
5000	0.2266	120

Table 2. Performance of Cluster Expansion Model and reported in MAP and number of relevant documents retrieved at different number of top IPC clusters and number of expanded clusters

Cluster Size-Limit	Num. Top Clusters	Num. Expanded Clusters	MAP	Rel. Ret.
1000	100	1	0.2161	115
1000	100	5	0.2154	119
1000	100	10	0.209	114
1000	1000	1	0.2126	121
1000	1000	5	**0.2273**	118
1000	1000	10	0.2195	117
1000	10000	1	0.2243	117
1000	10000	5	0.2143	119
1000	10000	10	0.2091	118
1000	∞	1	0.2183	114
1000	∞	5	0.2119	121
1000	∞	10	0.2127	119
∞	100	1	0.2137	116
∞	100	5	0.2148	116
∞	100	10	0.2126	118
∞	1000	1	0.2098	116
∞	1000	5	0.2177	117
∞	1000	10	**0.2181**	117
∞	10000	1	0.2148	120
∞	10000	5	0.2052	116
∞	10000	10	0.2056	116
∞	∞	1	0.2138	116
∞	∞	5	0.2037	117
∞	∞	10	0.2038	116

ways. First, parameters are too coarse. Although the number of top clusters and the number of expanded clusters seemed reasonable from the authors' point of view, the range defined by experimenter may not cover the optimum parameters. Also, the ratio of document model and cluster model of the Interpolation Model was fixed throughout the experiments, but reducing the number of clusters decreased the portion of cluster model in the final score of the Interpolation Model. At different number of clusters, One needs to search the optimal ratio exhaustively.

The effect of changing the number of clustered used and the number of expanded cluster, however, is well demonstrated.

6 Conclusions

We have proposed new models for cluster-based patent retrieval using International Patent Classification system. We first showed manual clusters are statistically different from automatically built clusters. As pointed out by other literatures, cluster size plays an important role in cluster-based retrieval, and we were able to show that it applies to manual cluster as well. With such knowledge, we proposed models more suitable and beneficial for manual cluster-based retrieval and show justifiable results.

Currently we are working on normalizing clusters for a target size and investigating the optimal size for cluster-based retrieval. Also, we are devising a more appropriate and general model for retrieving manually clustered documents. Cluster Expansion Model can also apply to automatically clustered documents and we plan carry out such experimentation, soon.

Acknowledgments. This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc), and also partially by the BK 21 Project in 2006.

References

W. Bruce Croft.1980. *A model of cluster searching based on classification*. Information Systems, 5, 189-195.

Abdelmoula El-Hamdouchi, and Peter Willett. 1989. *Comparison of hierarchic agglomerative clustering methods for document retrieval*. The Computer Journal, 323, 220-227.

C. J. Fall, A. Torcsvari, K. Benzineb, and G. Karetka. 2003. *Automated Categorization in the Internation Patent Classification*. SIGIR Forum 37(1): 10-25.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2004. *Overview of patent retrieval task at NTCIR-4*. Working Notes for the Fourth NTCIR Workshop Meeting pp. 225-232.

Djoerd Hiemstra. 2001. Using language models for information retrieval. PhD Thesis, University of Twente.

Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. 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.

In-Su Kang, Seung-Hoon Na, Jungi Kim, and Jong-Hyeok Lee. 2006. *Cluster-based Patent Retrieval*. Information Processing and Management.

Oren Kurland and Lillian Lee. 2004. Corpus Structure, Language Models, and Ad Hoc Information Retrieval. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Leah S. Larkey. 1998. Some *Issues in the Automatic Classification of U.S. patents*. In Working Notes of the Workshop on Learning for Text Categorization, 15th National Conference on Artificial Intelligence (AAAi-98).

Leah S. Larkey. 1999. *A patent search and classification system*. In Proceedings of the fourth ACM Conference on Digital Libraries pp. 179-187.

Xiaoyong Liu, and W. Bruce Croft. 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.

Jay M. Ponte and W. Bruce Croft. 1998. *A language modeling approach to information retrieval*. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275-281.

C.J. van Rijsbergen, 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA.

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

Chengxiang Zhai and John Lafferty. 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.