POSTECH at NTCIR-6 English Patent Retrieval Subtask

Jungi Kim, Yeha Lee, Seung-Hoon Na, Jong-Hyeok Lee
Div. of Electrical and Computer Engineering
Pohang University of Science and Technology (POSTECH)
Advanced Information Technology Research Center (AITrc)
San 31, Hyoja-Dong, Pohang, Republic of Korea, 790-784
{yangpa, sion, nsh1979, jhlee}@postech.ac.kr

Abstract

This paper reports our experimental results at the NTCIR-6 English Patent Retrieval Subtask. Our previous participation at the patent retrieval Subtask revealed that the long length of the patent applications require less smoothing of the document model than general documents such as news paper articles. We setup the initial baseline retrieval system for U.S. patent applications and compare the difference from that of Japanese patent applications.

Keywords: Patent *Retrieval, Language Modeling*

1 Introduction

In our previous participation of patent retrieval Subtask at NTCIR-5, we showed that a patent application has different characteristics from a general document in its structure and size, and has a manually assigned standard taxonomy system, namely International Patent Classification (IPC) code.

We suggested that such characteristics require researchers to reconfirm the well-known previous retrieval techniques such as the application of logarithmic function to term frequencies, document length normalization, pseudo relevance feedback, query expansion, and smoothing in statistical language model, as well as devise new retrieval techniques such as applying patent classification system to cluster-based retrieval model, and etc.

At NTCIR-5, we investigated the effect of smoothing in statistical language model using long and verbose documents such as patent applications.

In the sixth patent retrieval Subtask at NTCIR, we participated in US patent retrieval only. We again verify the effect of smoothing in long and verbose document collections, and, while comparing the Japanese and U.S. patent collection, report the differences and techniques that can apply to U.S. patent collection.

2 U.S. Patent Applications

2.1 Overview of the collection

Year	Size (GB)	Num. of Docs
1993	3.2	98385
1994	3.4	101695
1995	3.6	101431
1996	4.2	109654
1997	4.6	112019
1998	6.1	147577
1999	6.4	153591
2000	6.7	157596
	38	981948

Table 1 Size and Num. of Docs of NTCIR-6 U.S. Patent Collection

NTCIR-6 U.S. patent collection consists of 8 years worth of patent applications submitted to U.S. Patent and Trademark Office (USPTO) from 1993 to 2000. Applications and fields of the applications are marked with XML style tags. Generally required fields of use are: <DOCNO>, <TITLE>, <ABST>, <SPEC>, <CLAIM>. Other fields of interest to researchers are: <APP-NO> and <CITATION> for citation links from application to application or <PRI-IPC> for IPC codes of an application which in many cases are used as clusters.

See [2] for more detailed description of NTCIR-6 patent test collection.

2.2 Differences from Japanese Patent Collection

While carrying out the experiments for U.S. patent retrieval Subtask, we have noticed US patent retrieval have several differences from Japanese patent retrieval.

First, Unlike Japanese patent retrieval system, final retrieved results of U.S. patents cannot be filtered by the filing dates of the applications (<FDATE>). Although U.S. patents have similar field, <APP-DATE>, using this field to filter prior art resulted in poor retrieval performance.

Secondly, while most Japanese patent applications have more that one assigned IPC code [5], U.S. patents are only assigned one IPC code (<PRI-IPC>). Considering an IPC code as a cluster of documents assigned to it, clusters for Japanese patents are overlapping clusters. In [5], we suggested new techniques for cluster-based patent retrieval system using IPC codes as cluster, among which one method employing similarities between clusters using the number of documents in common to redistribute cluster scores in hope of more accurate representation of cluster model. Since IPC clusters in U.S. patent collections do not overlap, such useful information is not available.

Lastly, U.S. patents have direct citations to previously published patent applications by their application number. The citations can be regarded as a direct link between patent applications and popular link analysis methods can be applied to find the link structure of the patent collection.

3 System Description

3.1 Query and Document Processing

We participated in mandatory run of the Subtask only, which requires only the <CLAIM> field of Query be used in retrieval of related documents.

We used word as indexing units. Each token separated by white space in query and documents are processed in same way. First, punctuation marks and special characters are removed, and then they are stemmed with Porter Stemming algorithm. Both stemmed and original tokens are indexed in our system.

3.2 Retrieval Model

Our system is based on Language Modeling framework with Jelinek-Mercer smoothing. Detailed description is found in our report for NTCIR-5 patent retrieval Subtask [4].

Simply taking the final equation from the previous work, we have

$$P(Q\,|\,D_M)=\prod_{q\in Q}P(q\,|\,D_M)^{freq(q)} \qquad (1)$$

P(Q|D$_M$) is the query likelihood that a document model D$_M$ will generate a given query Q. We assume unigram language model, where each term q is independent of each other. To avoid assigning zero probabilities to unseen words, we smooth document model with collection model as in equation (2).

$$P(q\,|\,D_M)=(1-\lambda)P_{mle}(q\,|\,D_M)+\lambda P_{mle}(q\,|\,Coll) \qquad (2)$$

We do not use all query terms for retrieving but select terms whose document frequency is less than some threshold θ * |D|.

4 Experimental Results

Our official retrieval performance is 0.0282 in MAP for Rigid relevance judgment and 0.0572 in MAP for Relaxed relevance judgment. They are neither good nor bad performances considering all the submissions. However, since we only used the baseline system, there are possibilities for much improvement.

We provide here the unofficial evaluation results of our system using the sample topics provided before submitting the formal result. Model and the parameter settings are same as the officially submitted result. Out of the 1000 sample topics given for training, we decided to use smaller number of topics by selecting topics with 1 at its tenth digit place of topic number, i.e. 0010, 0011, and so on, reducing the number of topics to 100.

There are two parameters we tune for: λ for Jelinek-Mercer smoothing and threshold Θ for query term selection.

As Figure 1 shows, for selected sample topics, best performance in MAP using relevance judgment A is achieved with λ = 0.4 and θ = 0.4.

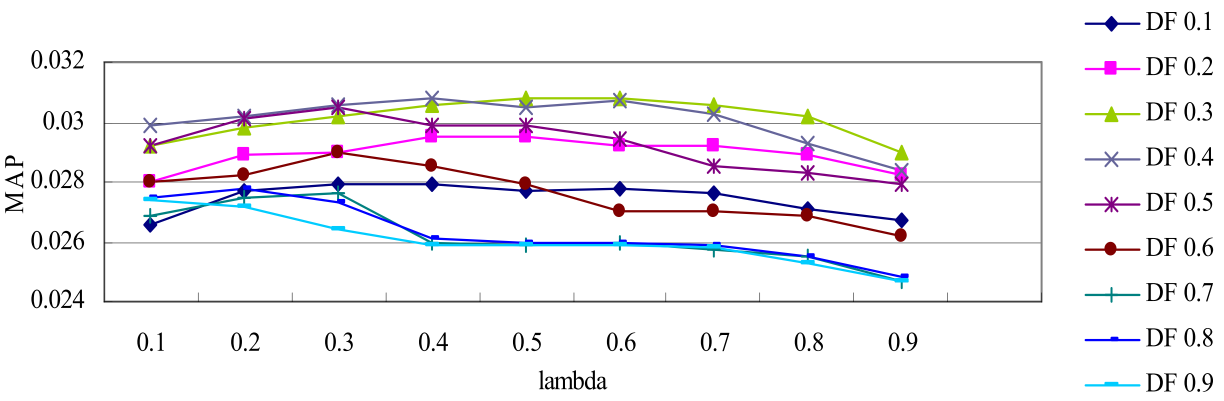

Figure 1 Effect of Smoothing and Query Term Selection in MAP

5 Conclusion

This paper presented the retrieval results of U.S. patent retrieval system using the baseline language model with Jelinek-Mercer smoothing. The performance is nether satisfactory nor disappointing, and it does not clearly verify our conclusion from NTCIR-5 Japanese Patent Retrieval Subtask where heavy smoothing hurts retrieval performance. More investigation into the characteristics of U.S. patent collection is required.

We also plan to further explore various methods using resources available in patent applications such as IPC codes and citations.

Acknowledgements

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

References

[1] Fujii, A., Iwayama, M. and Kando, N. Overview of patent retrieval task at NTCIR-5. In Proceedings of the Fifth NTCIR Workshop, 2005.

[2] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of patent retrieval task at NTCIR-6 Workshop. Proceedings of the Sixth NTCIR Workshop Meeting. 2007.

[3] Hiemstra, D. Using language models for information retrieval. PhD Thesis, University of Twente, 2001.

[4] Kang, I.S., Na, S.H., Kim, J. and Lee, J.H. 2005. POSTECH at NTCIR-5 Patent Retrieval: Smoothing Experiments in a Language Modeling Approach to Patent Retrieval, pages 300-303, 2005.

[5] Kim, J., Kang I.S., and Lee, J.H. Cluster-based Patent Retrieval Using International Patent Classification System, In Proceedings of the 21st ICCPOL, pages 205-212, 2006.

[6] Ponte, J.M. and Croft, W.B. A language modeling approach to information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 275-281, 1998.

[7] Zhai, C. and lafferty, J. 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, pages 334-342, 2001.