# Transforming Patents into Prior-Art Queries

Xiaobing Xue
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst, MA, 01003, USA
xuexb@cs.umass.edu

W. Bruce Croft
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst, MA, 01003, USA
croft@cs.umass.edu

## ABSTRACT

Searching for prior-art patents is an essential step for the patent examiner to validate or invalidate a patent application. In this paper, we consider the whole patent as the query, which reduces the burden on the user, and also makes many more potential search features available. We explore how to automatically transform the query patent into an effective search query, especially focusing on the effect of different patent fields. Experiments show that the background summary of a patent is the most useful source of terms for generating a query, even though most previous work used the patent claims.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Algorithms, Experimentation, Performance

## Keywords

Prior-art Search, Patent Retrieval, Information Retrieval

## 1. INTRODUCTION

The goal of searching a patent database for prior-art is to find previously published patents on a given topic. It is a common task when the patent examiner needs to decide whether a patent application is novel. As a legal document for protecting the invention, a patent has complex structures and technical content, which can create significant challenges for the retrieval system. Furthermore, some additional factors can make the problem even worse. In order to extend the coverage of a patent, the writers often intentionally use vague words and expressions in the claim, which increases the difficulty of capturing the real content of a patent. Also, in order to pass the patent examination, writers tend to develop their own terminologies, which can cause serious word mismatch problems. The combination of these factors make prior-art search significantly different with other search tasks, such as web search.

Currently, patent retrieval systems use a typical keyword search approach, where the success of the prior-art search relies on the quality of the keywords posed by the user. In this paper, we consider a novel approach to patent search, where the user submits the whole patent application as the query. In this approach, the burden on the user is entirely shifted to the system. Given a patent query, the system will have full access to the abundant information contained in the patent, which can potentially make many more useful search features available. Larkey [1] studied how to transform a patent into a query for patent classification, however this approach has not been fully explored for prior-art search.

Specifically, we focus here on a study of the search features extracted from different fields of a patent. A typical patent contains the following fields: the title (ttl), the abstract (abst), the background summary (bsum), the description of the figures (drwd), the detailed description (detd) and the claims (clms). Most previous work on prior-art search [2, 3, 4] has used the words from the claim field as the query without examining other alternatives. Our results indicate that the words from the summary field are better than those from the claim field when used as queries for prior-art search.

## 2. FEATURES FOR PRIOR-ART SEARCH

In this paper, we consider words extracted from different fields with different weights as search features. With the query patent available, many other features can be extracted, which will be explored in the future. Fig. 1 shows the general algorithm for extracting query words, where the parameters *Field*, *Num* and *Weight* indicate where to extract query words, how many query words should be kept, and what weighting methods are used, respectively. Table 1 list the possible values of these parameters. For the parameter *Weight*, *bool* denotes assigning the weight 1 to all words. For the parameter *Field*, besides the fields mentioned above, we also consider the primary claim (pclm), which is the most important claim among all claims and the whole patent (all), where the query words are selected from the whole patent without considering the structure information. With different configurations of the parameter values, we obtain different search features.

**ALGORITHM**: Extracting query words
**INPUT**: *Patent*, *Field*, *Num*, *Weight*
**OUTPUT**:*Query*
**PROCESS**: Rank words in *Field* according to their *tfidf* scores and then select *Num* top ranked words as the query words. Assign *Weight* to each query word to get *Query*.

**Figure 1: Algorithm for extracting query words.**


*SIGIR'09,* July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

**Table 1: Possible values for the input parameters**

| Name | Parameter Values |
|---|---|
| *Num* | integer |
| *Weight* | *tf*,*tfidf*,*bool* |
| *Field* | ttl,abst,bsum,drwd,detd,clms,pclm,all |

**Table 2: Influence of *Field* and *Weight* on retrieval performance.**

| | MAP | | | P@10 | | |
|---|---|---|---|---|---|---|
| Field | *bool* | *tfidf* | *tf* | *bool* | *tfidf* | *tf* |
| ttl | 0.042 | 0.039 | 0.043 | 0.108 | 0.098 | 0.109 |
| drwd | 0.044 | 0.048 | 0.047 | 0.117 | 0.116 | 0.120 |
| detd | 0.055 | 0.057 | 0.066* | 0.139 | 0.144 | 0.157* |
| pclms | 0.059 | 0.062 | 0.055 | 0.149 | 0.146 | 0.139 |
| clms | 0.066 | 0.066 | 0.064 | 0.155 | 0.160 | 0.157 |
| abst | 0.066 | 0.070 | 0.074* | 0.156 | 0.161 | 0.170 |
| all | 0.067 | 0.068 | 0.078* | 0.165 | 0.164 | 0.182* |
| bsum† | 0.078 | 0.082 | 0.094* | 0.181 | 0.182 | 0.199* |

# 3. EXPERIMENTS

## 3.1 Corpus

The USPTO corpus we used consists of 1,604,386 patents published from 1980 to 1997. The query set consists of patents published in 1997 that have at least 20 citations and all field types mentioned above. The size of the query set is 3,736 and we randomly split it into a training set of 3,361 patents and a test set with 373 patents. Since it is extremely difficult, perhaps even impossible, to get relevance judgments from experts for those thousands of queries, we use a patent's citation field <UREF> as a substitute, which is the same as the strategy adopted by NTCIR5-6[1]. Indri is used to index the full text of patents in the collection. The Krovetz stemmer is used to stem each word and the stopword is not removed. The standard mean average precision (MAP) and precision at 10 (P@10) are used to measure the retrieval performance. Two-tailed t-tests are conducted to decide statistical significance.

## 3.2 Effect of Field and Weight

In this section, we report on the retrieval performance of different search features extracted in Section 2. Specifically, the effect of different configurations of the parameters *Num*, *Weight* and *Field* is compared. For the parameter *Num*, we tested different values for *Num* from 10 to 50 for different fields, where *Weight* is set to *bool*. Results show that 10 words for the title field and 20 words for the other fields are enough to capture the most information, thus *Num* is set to 10 for the title field and 20 for the other fields.

The combinations of three *Weight* values and eight *Field* values were further explored. The results are shown in Table 2[2]. * denotes significantly different with both *bool* and *tfidf*. † denotes all values of 'bsum' are significantly different with the corresponding values of all other fields.

Table 2 shows that for most fields, *tf* is the best weighting method. *tfidf* is also better than *bool*, but not as clearly as *tf*. With respect to fields, it is clear that the words extracted from the background summary field achieve the best performance (much better than the claim field) when used for generating queries. This observation can be explained by the properties of these two fields. The claim field is used for legal purpose, which defines the protection scope for this patent. Thus, authors tend to use language that extends the scope, such as 'mobile unit' instead of 'vehicle'. In contrast, the summary field is mainly written for technical use. The authors typically review previous work and briefly describe their patent relative to related work. It is therefore not unreasonable that the vocabulary used in this field is more effective for prior-art search. The abstract field is also a slightly better source for query words than the claim field. Using only the primary claim is not as effective as using all claims. Although the detailed description (detd) contains much more content than other fields, it seems not as useful as the claim field. The title field and the figure description field (drwd) are among the least effective. The performance of 'all' is also a good choice that is slightly better than the abstract field, but still worse than the summary field.

[1]http://research.nii.ac.jp/ntcir/publication1-en.html

[2]Here, due to the space limit, only the results on the test set are provided. The results on the training set are quite similar.

# 4. CONCLUSION

Prior-art search is an important task in the patent process. In contrast to a typical keyword search, in this paper we consider a search scenario where the user submits a whole patent (or application) as the query instead of selecting keywords. The abundant information contained in the query patent makes it possible to extract different search features. On the USPTO collection, we explored the effect of fields and weighting methods on prior-art search and showed the background summary field is a better source than the widely used claim field for generating query words. In the future, obtaining more reliable relevance judgments is the most important issue. Other techniques for extracting concepts and entities from patent text should be further explored.

## Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0711348 and the Information Retrieval Facility (IRF). Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

# 5. REFERENCES

[1] Larkey, L.: A patent search and classification system. In: Proceedings of the 4th ACM Conference on Digtal Libary, Berkeley, CA (1999) 179–187

[2] T. Takaki, A.F., Ishikawa, T.: Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In: Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC (2004) 399–405

[3] H. Mase, T.M., Ogawa, Y.: Proposal of two-stage patent retrieval method considering the claim structure. ACM Transactions on Asian Language Information Processing **4**(2) (2005) 186–202

[4] Fujii, A.: Enhancing patent retrieval by citation analysis. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands (2007) 599–606