

NTT DTEC at Patent Retrieval Task

Yohichi Nakatani Koutarou Takada Michihiro Isoda

NTT DATA TECHNOLOGY CORPORATION

2-2-12 Akasaka, Minato-ku, Tokyo 107-0052, Japan

{yohichi.nakatani,takadak,michihiro.isoda}@nttdtec.co.jp

Manabu Okumura Makoto Iwayama Yuzo Marukawa Akihiro Shinmori

Precision and Intelligence Laboratory

Tokyo Institute of Technology

{oku,iwayama,maru}@pi.titech.ac.jp, shinmori@lr.pi.titech.ac.jp

Abstract

Search effectiveness is investigated when a corpus is created by using only “Title,” “Abstract,” and “Claims,” which are expected to briefly express the invention, instead of using the entire document in the search for documents similar to a patent application. In addition, the JAPIO patent abstract that expresses the invention is used to make a comparison with the search effectiveness of “Title + Abstract” for the patent application, and the merits of each are discussed.

1. Introduction

Recently, research into and commercialization of concept search has been steadily achieved, attaining the status of being closely related to our lives. Application of concept search (especially searches for similar documents) is demanded in fields requiring prior technical search, most prominently in the field of patents.

However, as is well-known, the reality is that the application of similar-document searches is not yet practical because of uniqueness in patent

specification expressions. One factor is that patent specification has an extremely wide variation in expressions (including coined words). Another factor is that the quantity of “noise” increases because actual examples not directly related to the invention itself and explanation of conventional techniques may be included in the document.

Other indirect factors preventing progress in application and investigation of similar-document searches in patent literature may include the fact that there is no pure test collection on patent documents.

One of the purposes of this patent retrieval task was “to create a high-quality test collection,” and we thought that investigation of similar-document searches on patent documents would be further progressed when the test collection is completed. Furthermore, we participated in the task since the task would clarify the problems in application of similar-document search on patent documents.

Although the task was officially executed from the viewpoint of “technical survey”, the emphasis of ours was mainly placed on “technical trend investigation by expert patent searchers.”

2. Search System

The similar-document search system used in this study mainly adopts “TF/IDF” and “Document Length Normalization” function [1]. Since it is also equipped with simple Boolean search and a relevance feedback function, it is capable of reducing search results by AND search and selecting the relevant patents from the search results to execute relevance feedback, etc., and this function was used in manual search.

3. Data Set

We used 2 years’ patent specification (years 1998 and 1999) and 2 years’ JAPIO patent abstracts (years 1998 and 1999). As mentioned earlier, we thought that using the entire document would allow inclusion of factors other than the nature of invention and might cause increase in noise. Therefore, we took note of “Title,” “Abstract,” and “Claims” that seem to express the invention clearly and it was decided that only the data for these sections should be used.

Moreover, combination within the above fields was examined, and search was executed in 3 patterns of “Title + Abstract + Claims,” “Title + Abstract,” and “Title + Claims.”

As to JAPIO patent abstracts, it was decided that the description was an overview of the invention and that comparison with “Title + Abstract” of patent application was possible, and they were used as corpus.

4. Search Method

Two patterns were used as search method and submitted. One was “manual search” which referred to only the <article><supplement> of the topic and in which the search query was created manually in order

to execute search, while the other was “automatic search” which used <article><supplement> of the topic and executed search automatically. Correspondence between searched corpus and search method is shown below. "ID" indicates the ID added when the task results were presented.

ID	CORPUS	MODE
F005	Patent Application (T+A+C)	Manual
F006	Patent Application (T+A)	Manual
F007	JAPIO Application Excerpt	Manual
F008	Patent Application (T+C)	Manual
F009	Patent Application (T+A+C)	Auto

(T: Title/A: Abstract/C: Claim)

Manual search was executed by two of the authors, and one person taking charge of ranges “topic-1 - 16” and the other taking charge of range “topic-17 - 31”.

5. Search Results

As stated previously, 5 patterns of results were submitted as search results. They are arranged in the order of higher expected search effectiveness in our judgment. The results are presented below.

Mandatory-A

ID	Average-precision	R-precision
F005	0.2144	0.2343
F006 (*)	0.1892 (0.1898)	0.214 (0.2123)
F007 (*)	0.0015 (0.186)	0.0025 (0.2122)
F008 (*)	0.1627 (0.1636)	0.1891 (0.1912)
F009	0.0959	0.1293

Mandatory-A+B

ID	Average-precision	R-precision
F005	0.2014	0.2483
F006 (*)	0.1785 (0.1793)	0.2159 (0.2194)
F007 (*)	0.0013 (0.183)	0.0027 (0.2315)
F008 (*)	0.1646 (0.1642)	0.2034 (0.2017)
F009	0.1083	0.1499

There are items here for which (*) is added in the ID section of results and for which two result values (upper and lower) exist. The upper value is the result officially presented on the Web, while the lower value is the result of re-executed after the official submission. The reason the lower value was re-executed was that the following mistakes were present at submission.

(1) Search was executed through GUI, and the results were made into a text file by copying & pasting. Data was pasted although copying was not done correctly. Therefore, the following sections turned out incorrect.

- “Title + Abstract”: Results of topic2 and topic4 are identical with the results of “Title + Abstract + Claims.”
- “Title + Claims”: Results of topic5 and topic10 are identical with the results of “Title + Abstract.”

(2) Although the document IDs of “F007” should be “-KKH-,” submission was made with “-JSH-” format since a JAPIO patent abstract was used. In addition, similarly with the above case, copying & pasting of the contents of topic10 failed. The results of topic10 were obtained again for correction, and “-JSH-” was changed to “-KKH-” to obtain the corrected value (lower).

6. Discussion

The order of runs turned out to be as we had expected in the case of Judgment A. It is especially interesting that the values of “Title + Abstract” and JAPIO patent abstracts are close. However, in the case of A+B judgments, the order of the JAPIO patent abstract was higher than “Title + Abstract.” In addition, “Title + Claims” was higher in order than

we had expected, although we expected it to be much lower. On the other hand, automatic search was very low in search effectiveness compared to manual search. A discussion of these matters follows.

6.1 “Manual search” vs. “Automatic search”

Manual search and Automatic search are briefly described here. Basically, in manual search, a searcher creates a search query referring to <ARTICLE><SUPPLEMENT> and executes the search. No special restriction is imposed on the search method of referring to the search results and adding/deleting search terms or making relevance feedback by selecting relevant documents from the search results, etc. The only restriction is having to use the same created search query for the runs of the same topic as much as possible. However, changing the query is allowed if identical search query cannot in any way provide satisfaction. Furthermore, there is no restriction obliging the searcher to select the same patents in the runs of the same topic when relevance feedback is made. This is because there is no guarantee that the same patents will be searched in the runs of the same topic.

Automatic search used <ARTICLE> <SUPPLEMENT> as search query, and nothing special was done.

Consequently, it seems that manual search is very effective when the overall results of the 31 topics are considered. However, when comparison is made for each topic, the number of cases where one retrieve more relevant documents than the other was 17 topics for manual and 10 topics for AUTO in Judgment A, and as for Judgment A + B, 16 topics for manual and 14 topics for AUTO. Comparison by each topic does not always lead to clear judgment regards the superiority of manual or auto.

The reason manual is better in the overall results of the 31 topics originates in the difference of topics for which the number of the retrieved relevant documents was superior. That is, topics superior in manual had many relevant documents retrieved compared to auto. On the other hand, topics superior in auto did not have a number of relevant documents retrieved drastically different from manual, having rather a number only a little larger.

However, there are cases in which there is a significant difference between manual search and automatic search (number of relevant documents retrieved for one search is 0). Topic3 and topic22 fall into this case. In the following, these 2 topics are focused on for discussion.

Although the number of A-relevant documents retrieved is 0 in automatic search in topic3, manual search resulted in 21 A-relevant documents retrieved. The method of manual search used AND search with “stepping motor” and “minute angle”, and documents were selected from the search result to do relevance feedback.

Since there is nothing unique to the manual search method, in this case, we assumed that <ARTICLE><SUPPLEMENT> of topic3 had certain distinctive features. Actually the <ARTICLE> format of topic3 was in interview style, and little of the actual invention was described. Therefore, it is assumed that similarity was higher in sections other than the invention when the all of the <ARTICLE> field was used as the search query, leading to a large amount of noise not strongly related to the invention and none of the A-relevant documents was retrieved.

On the other hand, topic22 had 0 cases of A-relevant documents in manual, while the number of A-relevant documents retrieved by automatic search was 8. In manual search, the search query was “NOx, nitrogen oxide, reduction, removal,

purification, infrared radiation, magnetism.” As far as search results are concerned, documents related to NOx reduction, removal and purification were retrieved without problem.

However, since the description “NOx, nitrogen” did not exist in most of the A-relevant documents in the test collection, it is assumed that many A-relevant documents could not be retrieved by this manual search method that concentrated on “nitrogen.” In addition, even when the description of “NOx” existed, it was in the “invention embodiment” section, which was not included in the corpus in our case. This also seems to have led to lower performance in manual search.

Thus the contents of topic22 and correct documents were considered to verify what kind of search method would have been effective. Although NOx was not effective, the word “fuel” was found as another characteristic term. Search was made with “magnetism, infrared radiation, fuel,” and 4 A-relevant documents were retrieved. In addition, “magnet” was found as another useful term, and 7 A-relevant documents were retrieved when “magnet, infrared radiation, fuel” were input.

According to these results, it is evident that search results may become very effective or contain completely nonconforming cases depending on the terms selected with reference to topic, when manual search is made. When similar documents are searched, it may be an appropriate strategy to execute both “manual search” and “automatic search” and compare the results each other.

Moreover, search by the entire patent application document may be a very effective way to search with high recall, although it will increase the quantity of noise, since there were cases in which relevant documents could not be retrieved because the relevant passage did not exist in “Title + Abstract +

Claims.”

6.2 “Title + Abstract” vs. “JAPIO Patent Abstracts”

JAPIO patent abstracts include cases in which the patent application abstract is directly used and cases in which JAPIO modifies the original abstract by making correction or addition, or develops new sentences. For this reason, we assumed that there would be some significant difference between the results of “Title + Abstract” of the patent application (i.e., original abstract) and the “JAPIO patent abstract.” In addition, since “added keywords” exists in JAPIO patent abstracts, these keywords are expected to work effectively.

When data was overviewed based on these expectations, a significant difference did exist. That is, while cases with very short abstracts or very redundant descriptions exist in the original abstracts, most JAPIO patent abstracts have been corrected or added by JAPIO to unify the document length and description level for the entire collection. When several topics were considered, there was tendency for these corrected or added abstracts to be retrieved in higher orders, compared to original abstracts, and actually about 50% of the top 20 documents were JAPIO revised abstracts.

However, when each topic was considered, the number of cases where one retrieved more relevant documents than the other was nearly equal for judgment A (8 topics for “TITLE + Abstract” vs. 12 topics for JAPIO patent abstract). Also, for judgment A+B, there was no significant difference, although JAPIO patent abstracts were somewhat superior (7 topics vs. 16 topics). In addition, there was no significant difference between the number of relevant documents retrieved as in the case of automatic vs.

manual search.

Furthermore, added keywords in JAPIO patent abstracts did not improve the search effectiveness. This is due to the low probability of matching between the terms in the input search query and the added keywords.

According to the results, it seems there is little difference between the original abstracts and JAPIO patent abstracts. It may be possible to find some tendencies if each topic is analyzed more deeply.

We think that large difference may not occur because document length and description level is unified in abstracts, and in such situation, it may be difficult to apply a similar-document search. In addition, as stated in the previous section, there are cases of JAPIO patent abstracts in which relevant passages are lacking. For the purpose of high recall searching, it is assumed that JAPIO patent abstracts were not appropriate for use as corpus, and that using the entire patent application document would be appropriate.

6.3 “Title + Abstract + Claims” vs. “Title + Abstract”

Basically, with addition of claims data, the search effectiveness of “Title + Abstract + Claims” is better. However, there were factors assumed to be the cause for lower effectiveness in some topics in which search efficiency was lower in “Title + Abstract + Claims.” These factors are discussed here.

As stated earlier, “Title + Abstract + Claims” tends to be a little superior in search effectiveness because Claims information has been added. However, when topic6 and topic27 in which the search efficiency of “Title + Abstract” was higher were considered, the following tendencies seemed to exist.

One characteristic of topic6 and topic27 is that both cases used words from the title in the search query. Details of each query are “防止(prevention), フィルム(film), レンズ(lens), 再利用(recycling), 不正(invalid), 不正規(irregular), 不当(unfair), 再使用(reuse)” for topic6, and “薄型電波吸収体(thin radio wave absorber), 薄い(thin), 厚い(thick)” for topic27. “レンズフィルム(Lens film)” for topic6 and “電波吸収体(radio wave absorber)” for topic27 existed in the title.

Here, when claims are considered, there is a tendency for the words in the title to be re-used in claims. Therefore, patents with many claim sections inevitably had repeated title words, making them higher in the retrieved ranking.

Since topic6 and topic27 used the words included in the title, patents with many claim sections were searched to the orders of top 20 - 40 for “Title + Abstract + Claims,” compared to “Title + Abstract.” Although words other than the title are included in the search query, they lose importance and have little effect on relevance ranking. That is, cases with many claim sections with short descriptions emphasize the title words, and thus it is assumed that they are higher in the ranking.

It is thought that this problem occurred because all cases of patent specifications were searched. That is, when the words used in the title are considered within all cases of patent specifications, they become terms that are extremely characteristic in the corpus. Therefore, creation of corpus that may reduce the importance of terms used in title may be necessary to solve this problem.

6.4 “Title + Abstract + Claims” vs. “Title + Claims”

Basically “Title + Abstract + Claims” is superior.

This is due to the fact that the abstract itself describes the invention and claims did not supply sufficient information. However, there are topics in which “Title + Claims” became superior, and they are discussed here.

In topic3, “Title + Claims” was superior (this may be a very interesting topic in this task because in the comparison of automatic and manual search, manual is completely superior for this topic). When compared by the number of relevant documents retrieved, judgment A gave 12 cases for “TITLE + Abstract + Claims” and 26 cases for “Title + Claims”, while judgment A + B gave 43 cases and 58 cases where “Title + Claims” being superior.

It is thought that this superiority may be due to the large effect of relevance feedback. The search method of topic3 used AND search with “stepping motor” and “minute angle” and then documents supposed to be relevant were selected from the search results to perform relevance feedback. Here, 10 documents were retrieved as search results in AND search of “Title + Abstract + Claims”, and 7 of them were selected for feedback. Out of the 7 documents, 4 documents were officially A, 1 document was officially B, and 2 documents were officially non-relevant. On the other hand, AND search of “Title + Claims” retrieved 4 documents as search results, and 3 of them were selected for feedback. The selected 3 documents were all of judgment A officially.

When only official A documents (4 documents) were actually selected with “Title + Abstract + Claims” to perform relevance feedback, it worked very effectively with the number of 30 documents with judgment A, and 63 documents with judgment A + B.

On the basis of the above results, it is said that search effectiveness deteriorated with inclusion of 2

non-relevant documents which increased noise when relevance feedback was performed in “Title + Abstract + Claims. On the other hand, “Title + Claims” selected only the relevant documents, and this selection seems to have directly taken effect by the feedback. That is, it became evident that selecting only relevant documents is very important in relevance feedback since it largely depends on the selected documents.

Selection of non-relevant documents in this case seems to be ascribable to the fact that the searchers did not understand the contents of the applicable invention, and to the fact that the entire document could not be referred to because only the contents of “Title + Abstract + Claims” were displayed. It is expected that the data should be prepared so that the entire documents are available for reference even when “Title + Abstract + Claims” is used for indexing.

7. Conclusions

It was very surprising that the results of manual and auto search did not produce a large difference for each topic compared to the difference between the total values. Search effectiveness becomes higher in manual search if effective words can be selected. Conversely, search effectiveness may increase in automatic search due to the effective word overlooked in manual. Since searchers cannot understand which word is the effective word, it may be most effective if both manual search and automatic search are executed and the results are examined.

Furthermore, it was very beneficial result that the selected documents in relevance feedback greatly affects the search effectiveness, because it renewed our awareness of the danger in the method which

feeds back a vague decision of relevance on the grounds that the contents seemed somewhat relevant in executing manual search.

In this study, a rather irregular treatment was used with “Title + Abstract + Claims,” “Title + Abstract,” “Title + Claims”, instead of using the entire patent application documents. Consequently, it is judged that “Title + Abstract + Claims” is most efficient in searching and is effective. Since such search efficiency is obtained even without using the entire patent application documents, we consider that the idea of regarding “Title + Abstract + Claims” as briefly expressing the invention was not necessarily wrong.

However, there were problems such as lack of relevant passage in “Title + Abstract + Claims”. It is expected that the advantage of each search method will be clarified if search is made by using the entire patent application documents in the future.

References

- [1] A. Singhal, C. Buckley and M. Mitra.
Pivoted document length normalization.
In proceedings of the Annual International ACM
SIGIR Conference on Research and Development in
Information Retrieval, pages 21-29, 1996.