A Method for the Recommendation of Similar Documents and Related Researchers in National R&D Information Collections

Heejun Han¹, Heeseok Choi² and Jaesoo Kim³

¹,²,³ Korea Institute of Science and Technology Information, South Korea
¹hhj@kisti.re.kr, ²choihs@kisti.re.kr, ³jaesoo@kisti.re.kr

Abstract

The users of academic databases and R&D information often use search services to obtain necessary data for their studies. Most web users create various search queries and request these needed information from the system, and as a final destination of searching selected information lists and obtaining information, users are moved to the detailed page of the corresponding information. Similarly, in cases of academic information services providing journal and patent information, the final destination of the user is the specific metadata information page or the full article page, and in this case, providing other similar types of academic information and the names of researchers in other related fields is essential for satisfying the information requirements of the users. In case of the NTIS (National Science and Technology Information Service), it provides search services on national R&D information (tasks, participating personnel, research products, facilities and equipment, etc.), but lacks provision of similar documents within the same DB or among different DBs. In this article, the authors explain user queries and search items provided by the search engine, serviced by the NTIS for data categorized as research products including journals, patents, research reports, and trend analyses, and search service methods for listing similar documents within the same or different contents, and names of researchers in related fields using boosting technologies. This way, the R&D information desired by the user may be efficiently provided in the final service screen, which can reduce repeated efforts on searching.

Keywords: Similar Document Retrieval, Related Researcher Recommendation, NTIS, R&D Information

1. Introduction

Efficiency of search systems is determined not only by the speed and stability of the system, but also by the accuracy of the search results provided to the user from the vast amounts of typical or atypical documents [1, 2]. The purpose of the creation of search queries by those using web based information search services lies in reaching documents that are pertinent to the interest of the users. The user creates search queries and requests results from the search system, and the search system undergoes unique search processes including similarity calculations and alignment algorithms to return refined search lists. The user then moves to the final detailed page by choosing the documents most conforming to their search intentions from the proposed list, and achieves the primary objective of their search. But the users also want other documents similar to the one chosen, or a list of names of researchers in related fields, to be automatically proposed by the system. For example, users using journal search results in an academic information service will want to see other related patents or research reports, or the names of researchers in the related field. Most search engines satisfy these requirements by using document vectors, which are terms representative of the documents, to extract similar other documents [3, 4]. In case of the NTIS, the FAST search engine uses document vectors to provide similar documents within the databases of the same type, at the screen
showing the detailed information of one assignment. As this method calculates similarity using only the TF and IDF values of the document vectors, there is a possibility of decreased similarity compared to the original document, which will not satisfy requirements to use lists on similar documents between different databases and related researchers.

In this paper, the authors explain methods to efficiently service the similar documents within the same DB or among different DBs of tasks and research products (journals, patents, research reports, trend analyses, etc.) of NTIS, and other related researchers. The authors also combine the user queries and document vectors extracted by the search engine to create a candidate group, and also provide similar documents by applying boosting methods for the search results by using theme field codes and author name field values. This paper will also attempt to categorize related researchers by using the grouping functions of the search engine. In chapter 2, the authors’ introduce the NTIS and the FAST search engine, and then review the similar document search functions of the academic information portal NDSL (National Discovery for Science Leaders) operated by the KISTI (Korea Institute of Science and Technology Information). In chapter 3, the authors’ explain these proposed methods and in chapter 4, a conclusion is made.

2. Related Works

2.1. NTIS (National Science & Technology Information Service)

Many research organizations, universities, and enterprises produce and manage many research results while carrying out their R&D projects. The research information project, human resources, research results, technology reports, and equipment are important factors working as reference knowledge on R&D fields, and some plans had to be drawn up by systematic architecture for the management, spread, and application of research results. NTIS is the world’s first national R&D information portal service, which provides government funded R&D information on topics as programs, projects, human resources, equipment and facilities, and outcomes in real time from 2004[5].

![Figure 1. NTIS Concept Diagram](image-url)
Figure 1 shows the NTIS service concept. The NTIS consists of several services such as national R&D program management service, national R&D participant information service, national R&D outcome information service, equipment and facilities sharing service, science & technology information dissemination infrastructure service, technology and industry information service, and regional R&D information service. NTIS promotes national R&D investment, and also supports decision making for R&D programs management. Here, NTIS uses the FAST search engine for search service in about 6 million R&D information.

2.2. FAST Search Engine

FAST ESP is an integrated software application that provides a platform for searching and filtering services. It is a distributed system that enables information retrieval from any type of information. ESP combines real-time searching, advanced linguistics, and a variety of content access options into a modular, scalable product suite. The FAST ESP system consists of different types of modules that can be categorized according to the purposes of the modules; in other words, what the module does in the system, such as matching and query & result processing. Table 1 shows several modules and their categories of FAST ESP system.

<table>
<thead>
<tr>
<th>Category</th>
<th>Module Name</th>
</tr>
</thead>
</table>
| Data Sources | Web Crawler  
File Traverser |
| Document Processing | Document Processing Engine |
| Matching and Query/Result Processing | Search Engine  
Query & Result Server |
| APIs | Content API  
Search API |
| Administration | Administrator Interface  
License Manager  
Admin Server |
| Relevancy Tuning | Boost Bulk Tool |
| Additional Modules | Content Connectors  
SDK  
Taxonomy Explorer  
Security Access Module |

A web crawler locates and retrieves files on web server. A file traverser traverses and retrieves files from file servers. A document-processing engine performs document-processing tasks for format conversion and document relevancy such as language detection, Asian language tokenization, and lemmatization. A search engine performs the indexing and searching tasks within FAST ESP. It also indexes new documents coming from the FAST document processing engine, matches them against search queries submitted by the query result server, and returns a list of resulting documents and result set navigation options to the query and result server. A query and result server processes search queries and search results to enable relevancy-focused searching and result presentation. It also provides linguistic query processing features like spell checking, and results processing features like result clustering. Content API allows the standard data source modules of FAST ESP, as well as custom applications, to push content to the FAST content distributor. A search API allows external search front end systems to submit their queries and receive result sets in return. In the administration category, an administrator interface provides a browser-based graphical user interface that allows the
system administrator to monitor and configure FAST ESP, while license manager is a license server for all components controlled by the licensing scheme, and an admin server allows system administrators and business users to monitor the end-users’ query behavior and to fine-tune the ranking of individual documents based on the monitoring results. A boost bulk tool can be used to import rank boosting specifications for individual documents into an existing search index. It reads boost records from an XML file and applies the boosts to the specified documents. Additionally various content connectors allow submitted content from databases such as DB2, Oracle, or SQL Server, and other specific applications. SDK, on the other hand, provides additional integration tools for query, content, and document processing integration and search front-end development. A taxonomy explorer is used to create taxonomies for document organization or use concept extraction. Lastly, a security access module provides application level security when integrating FAST ESP with security environments such as active directory.

2.3. Finding Similar Document in NDSL

As a representative academic information service of Korea, the NDSL operated by KISTI can be chosen [5]. The NDSL provides search services for around 100 million cases of academic information including journals, patents, research reports, and trend analyses, all using the FAST search engine. An important characteristic of this service is that it provides similar documents at the detailed metadata information page. In this case, the search engine repeats the search using the document vectors extracted in advance, and the document vectors are composed of top 10 documents with high frequency rates after processing for language, such as separation of compound words and morphologic analyses, and the weighted values of the relevant keywords. The document vectors are extracted from the title, abstract, and keyword fields of the metadata of a document. The document vectors representing a document can be expressed in a formula as shown in equation below (1).

\[
DV_{docid} = docid: [docvec, weight]_1, ..., [docvec, weight]_n \\
\text{where } n = 1, 2, 3, ..., 10
\]

Here, \( DV \) is the assembly of document vectors, \( docid \) is the serial number of the document, \( docvec \) is the document vector and \( weight \) is the weighted value.

The NDSL uses the extracted document vectors to repeat a search on the same DB in the detailed information page to provide similar documents. Figure 2 shows the search processes of the NDSL for similar documents. But as this method does not use the query created by the user, it leads to search results that do not reflect the intentions of the user, and may be improved by refining the search results with grouping and filtering of the search results using theme classification codes (DDC) of journals or IPC categorical values of patents.

![Figure 2. Process of Finding Similar Document in NDSL](image-url)
3. Proposed Method for Finding Similar Documents and Related Researchers

For the NTIS national R&D information search services, the FAST search engine extracts 10 document vectors for each document. Out of these, extraction of the candidate group for the similar documents is done by combining the document vectors having a weighted value of more than 0.5, the original query of the user, and the title and keywords of the single detailed metadata information (hereafter written as source document) used by the user. The search results are then optimized by realigning the candidate group using boosting techniques through the author name values and theme classification codes of the source document. The items necessary for performing searches for similar documents between different DBs proposed by the authors are as follows:

- **User’s query**

- **5 document vectors of the source document**

- **Document ID, title, keywords, name of authors, and theme code of the source document**

When the user uses one detailed information page, the principle of providing similar documents between different DBs are as shown in Figure 3.

![Diagram showing the proposed process to recommend similar documents and related researchers at NTIS](image-url)
First of all, a search on the title field of the total index DB using the user's original query is performed to obtain the result set A. After that, the intersection of the results obtained by searching the title field of the total index DB using the title value of the source document, and the results obtained by searching the title and document vector field of the total index DB using combined keywords and document vectors of the source document, which is called the intersection B, is obtained. The co-occurrences of the title, keywords and document vector values of the source document are handled with the OR operator. The union set of the result set A and B is then created, and to exclude the source document from this set, the document ID of the source set is processed with the ANDNOT operator to obtain the final similar document list of the result set C. Afterwards, to cluster the documents that have the same theme field as the source document or documents including the author names of the source document, the boosting method is applied to the result set C. Thus the documents of the result set C are realigned by the boosting factors including the theme classification code and author name values of the source document, and the XRANK operator provided by the FAST engine is applied. Finally, the author name field values of the relevant metadata of the obtained similar documents list are grouped and aligned in the order of the appearance rate, to be provided as the related researcher list.

The result set A is composed of search results of user queries on the title field of the total DB, reflecting the initial user search intentions. The result set B is composed of search results of title, keywords, and document vectors extracted from the case of detailed information presently used by the user, applied to the title and document vector fields of the total DB, and this shows the results similar to the source document. After merging these two sets, the source document is excluded, and to include only the documents similar to the source document in the serviced list from the candidate group result set C, boosting methods are applied. If documents including the author names or theme classification codes of the source document exist within the documents of the result set C, or if documents with identical IPC categories exist for patents, these results are boosted to the higher ranks. Simultaneously, an author grouping information is created from the author fields of the result set C, and the related authors, related researchers or related patent applicants may be provided from the similar documents by aligning the names in the order of appearance rates. The steps for handling queries in the search process of similar documents and related researchers are shown in Table 2 and 3. Table 2 shows an example of metadata of a source document, and Table 3 shows a step of handling process to recommend some documents similar to a source document and finding related researchers’ names.

**Table 2. Example of Metadata of a Source Document**

<table>
<thead>
<tr>
<th>Source document</th>
<th>User Query</th>
<th>Document ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document ID</td>
<td>DNA vaccination</td>
<td>1460002939</td>
</tr>
<tr>
<td>Title</td>
<td>DNA vaccination with using epitope in bronchial asthma</td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td>Gastric cancer; MAGE; vaccination</td>
<td></td>
</tr>
<tr>
<td>Name of authors</td>
<td>Sunseok Kwon, Misuk Lee, Dongin Kim</td>
<td></td>
</tr>
<tr>
<td>Theme classification code (Korea S&amp;T code)</td>
<td>G08</td>
<td></td>
</tr>
<tr>
<td>Document vectors (w&gt;=0.5)</td>
<td>DNA vaccination, bronchial asthma, epitope, T cell epitope, plasmid DNA, Vaccine, therapy, MMPI, TIMP, trichome</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Query Process for Searching Similar Documents and Related Researchers

<table>
<thead>
<tr>
<th>Q1</th>
<th>TITLE:string(“dna vaccination”, mode=AND)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2</td>
<td>TITLE:string(“DNA vaccination with using epitope in bronchial asthma”, mode=OR)</td>
</tr>
<tr>
<td>Q3</td>
<td>TITLE:string(“dna vaccination, bronchial asthma, epitope, T cell epitope, plasmid DNA, Vaccine, therapy, MMPI, TIMP, trichrome”, mode=OR)</td>
</tr>
<tr>
<td>Q4</td>
<td>DOCVECTOR:string(“dna vaccination, bronchial asthma, epitope, T cell epitope, plasmid DNA, Vaccine, therapy, MMPI, TIMP, trichrome”, mode=OR)</td>
</tr>
<tr>
<td>Q5</td>
<td>AND(Q2, Q3, Q4)</td>
</tr>
<tr>
<td>Q6</td>
<td>OR(Q1, Q5)</td>
</tr>
<tr>
<td>Q7</td>
<td>ANDNOT(Q6, DOCID:filter(“1460002939”))</td>
</tr>
<tr>
<td>Q8</td>
<td>AUTHOR:string(“Sunseok Kwon, Misuk Lee, Dongin Kim”, mode=OR)</td>
</tr>
<tr>
<td>Q9</td>
<td>STCODE:filter(“G08”)</td>
</tr>
<tr>
<td>Q10</td>
<td>OR(Q8, Q9)</td>
</tr>
<tr>
<td>Q11</td>
<td>XRANK(Q7, Q10, boostall=yes)</td>
</tr>
</tbody>
</table>

The search fields used in the proposed methods exist commonly for tasks, journals, patents, research reports, and trend analysis data, so they may be efficiently applied for searching similar documents between different DBs. The user may use patent information, and concomitantly obtain a list of research reports and names of researchers and patent applicants in the related fields, and also obtain search services on other documents using these names. Figure 4 shows an example of the similar documents and related researchers clustering service screen using the proposed method, for academic information services. These are the related journals, related patents, related research reports, and related trend analysis data provided with the detailed information of a journal article, and a list of related researchers such as related authors and related inventors are provided concomitantly.

Figure 4. Service Screen Example for Recommending Similar Documents and Related Researchers in NTIS
4. Conclusions

Some academic information services such as NDSL provide documents similar to the documents used by the user, but in the case of NDSL, the scope is limited to documents within the same DB and the accuracy of similar documents is low, and does not satisfy the desires of the users to obtain data on similar documents and related researchers between different DBs, including journals, patents, research reports, and trend analysis data. Also, the NTIS providing national R&D information provides clustering functions for similar documents on task information, but does not provide information on similar documents between different DBs, which is frequently demanded, and also needs to provide a related researchers list. In this article, the authors proposed an algorithm finding similar documents and related researchers’ name within the same DB and between different DBs. The documents with high similarity are serviced using the combination of the user query and title, keywords, document vectors of the source document, and applying boosting techniques using the author names and theme classification codes. Through these results, a list of related researchers is provided.

Acknowledgments

This research was supported by the NTIS Service Program of the Korea Institute of Science and Technology Information (KISTI) (N-15-NM-CU01-S01).

References


Authors

Heejun Han
- 2002 B.S. in Information and Communication Engineering from Chonbuk National University of Korea
- 2004 M.S., in Electrical Engineering, KAIST of Korea
- 2004 Senior Researcher, KISTI of Korea
- Research area: Content based Retrieval, Semantic Web, Multimedia Search, S&T Strategy

Heeseok Choi
- 2000 M.S., in Software Engineering, Busan National University of Korea
- 2007 Ph.D., in Software Engineering, Busan National University of Korea
- 2005 Senior Researcher, KISTI of Korea
- Research area: Bigdata, Software Architecture
Jaesoo Kim
- 1987 M.S., in Computer Science, Hankuk University of Foreign Studies of Korea
- 2009 Ph.D., in Software Engineering, Hongik University of Korea
- 1991 Department Manager of NTIS Center, KISTI of Korea
- Research area: Software Engineering, Database System, Distribution System