Computer Technology and Application 6 (2015) 57-63
doi: 10.17265/1934-7332/2015.02.001

A Study on Indexing Efficiency and Retrieval Accuracy for Author Name Search of Academic Papers

Heejun Han, Heeseok Choi, and Jaesoo Kim

NTIS Management Department, NTIS Center, Korea Institute of Science and Technology Information, Daejeon 305-806, Korea

Abstract: Most academic information has its creator, that is, a subject who has created the information. The subject can be an individual, a group, or an institution, and can be a nation depending on the nature of the relevant information. Most web data are composed of a title, an author, and contents. A paper which is under the academic information category has metadata including a title, an author, keyword, abstract, data about publication, place of publication, ISSN, and the like. A patent has metadata including the title, an applicant, an inventor, an attorney, IPC, number of application, and claims of the invention. Most web-based academic information services enable users to search the information by processing the meta-information. An important element is to search information by using the author field which corresponds to a personal name. This study suggests a method of efficient indexing and using the adjacent operation result ranking algorithm to which phrase search-based boosting elements are applied, and thus improving the accuracy of the search results of author name. This method can be effectively applied to providing accurate search results in the academic information services.

Key words: Author name search, information retrieval, indexing, search algorithm, boosting.

1. Introduction

Personal names on the web are an important element which accounts for 30% of all search engine queries, and search by means of personal names is an important function in web applications [1, 2].

All current information is produced by a creator, who is referred to as an author for a paper, an applicant, an inventor, and an attorney for a patent, a research participant for research reports, an inspector and an analyzer for trend analysis data, in terms of academic information. The creator of information referred to as such various names can be an individual, an institution, a group, a nation, or a computer system, for example, a crawler [2].

The creator of academic information is mostly a personal name. For example, in more than approximately 95% of all data, the creator field, including authors, applicants, and research participants, is described by personal names, the format of which is shown in Table 1, as can be found in papers, patents, theses, research reports, industrial standards, science and technology work forces, trend analysis information, and factual information for the National Discovery for Science Leaders service provided by the KISTI (Korea Institute of Science and Technology Information).

The academic information search service on the web undergoes the process of indexing data required for search. Person names are data which cannot be decided with compound words and postpositional words, so most academic information search services carry out delimiting white spaces or separation in delimiters in indexing the author name field. Also the process of indexing data in English undergoes stemming, singular and plural number processing, etc. However, for personal names in English, indexes are extracted by tokenizing data in white spaces or delimiters. Web of Science, Scopus, the NDSL science and technology information integration service, and most academic information services enable information to be searched through the author

Corresponding author: Heejun Han, M.S. in Electrical Engineering, research field: future R&D strategy research. E-mail: hhj@kisti.re.kr.

Table 1 Exemplary metadata for personal name in academic information.

Category	Classification	Format
Korean paper	author	Park, Sung-Joon; Kim, Ju-Youn; Kim, Young-Kuk
Overseas paper	author	Zhou, Fushan; Yang, Deng-Ke; Molitor, R. J.
Patent	inventor	Sakurada, Masahiro; Yamanaka, Hideki; Ohta, Tomohiko
Research report	reporter	Ryu, Beom-Jong; Kim, Jin-Sook; Jin, Doo-Seok
Trend analysis	analyzer	Rohak Park; James Lee

Table 2 Result of indexing author name.

Original data 1	Park, Sung-Joon; Kim, Young Kuk; Song, Byung-Soo
Indexes 1	park sung joon kim young kuk song byung soo
Original data 2	Aron Culotta; Pallika Kanani; Robert Hall; Michael Wick; Andrew McCallum
Indexes 2	aron culotta pallika kanani robert hall michael wick andrew mccallum

name field for paper search. In particular, Scopus or the NDSL provides an independent function for finding author names through “Find Author” after building and indexing an author name DB [3].

This study describes an efficient method of indexing for searching author names focusing on the NDSL or NTIS service, and refining the process of search. Section 2 describes related studies; Section 3 describes issues associated with searching author names; Section 4 describes a method suggested in this study; and Section 5 gives a conclusion.

2. Related Studies

While searching personal names is considered increasingly important in all web document searches, including academic information services, news, and other knowledge information, an issue involved is that indexing and searching technology based on simple strings and tokenizing by white space can provide information including inaccurate personal names. If data are provided for identifying the author name included within the academic information, and a given personal name notation is provided, accurate search results can be achieved. However, a prerequisite is the record linkage method for connecting different records that show the same personal name, and a process of author identification for forming a group of entities represented from identification properties, for example, language analysis, paper titles, e-mail addresses, journal names, and institutions to which authors belong [4-6]. Various personal name notations should also be standardized. However, there is currently no system for building and using author identification data for searching personal names, or for extending personal name queries into various notation formats for searching.

The Web of Science, Scopus, the NDSL site, and other exemplary academic information service providers provide the function of personal name or author search, which is included in the basic search field in addition to title, contents, and source fields. All white spaces included in a user query are processed with the AND operator, and include inaccurate search results if the relevant query is a personal name. In the NDSL and Scopus, the double quotation marks can be used in order to process phrase search for personal name queries. However, this method still results in inaccurate information in the author field in which a plurality of person names is written.

3. Issues Involved in Author Name Search

3.1 Inaccurate Search Result

In the paper, author and co-author names are listed with delimiters as shown in Table 1. Indexing the author name field is carried out by tokenizing white spaces and possible special delimiters (; - , .) to produce indexes. Table 2 shows an exemplary indexing result for an author name field.

All search systems convert application user’s queries to those that can be processed by a search engine. Most search applications in the NDSL or NTIS translate a white space of user query to AND operation for search engine. If personal names in Korean are used as a query, search accuracy does not matter because the white space is not included at

person name in Korea, but unwanted search results are obtained if a query is in English. The words “parkz”, “sung”, and “joon” are unwanted search results because they exist regardless of the order and the position of the author field of the relevant paper if the query is “Park Sung Joon”, as shown in Table 3.

The phrase search may be applied at this case in order to exclude the unwanted search results shown in Table 3, but search results not wanted by users still exist. For example, if a user intends to search papers by a person called “Kanni Robert”, papers with an author list are presented as a search result if the query words are adjacent with a delimiter (,) as shown below, and it is obviously a result not wanted by the user.

ex) Pallika Kanani, Robert Hall, Michael Wick, Andrew McCallum

3.2 Inefficient Indexing in NDSL Service

The NDSL includes “Find Author” function which is one of the paper search functions. ‘Find Author’ contributes to searching author names in all author lists of about 56 million papers. This is subject to the following pre-processes:

(1) Extract metadata from a paper;

(2) Check redundant author name at the character level;

(3) Create an author name for sorting;

(4) Extract the first character for searching an initial sound (the alphabet in case of English);

(5) Load author information onto the Oracle DB table;

(6) Index the author information and then provide searching function.

The number of author records of which the redundancy was extracted from 56 million paper records in the NDSL and then removed by string processing is approximately 22 million. Approximately 50,000 pieces of paper information are acquired weekly, which means that independent author information continues to be created at the rate of 40%. It is a waste in terms of indexing and management to additionally index author names which already exist in the paper index information in order to implement “Find Author”. This has a negative impact on search speed and disk load in hardware.

4. Proposed Method

4.1 Improving Retrieval Accuracy

Searching an author name is required within units of semicolons (;) which are used to divide listed personal names in order to exclude unwanted search results. A limiter is specified to enable the search operator to operate only in the relevant delimiter. That is, the delimiter (separator) property is specified with a semicolon for all metadata fields described as personal names in all academic information, such as papers, patents, research report, and analysis trend information for indexing. In this case, phrase search brings search results only when there are successive indexes matching the sequence in the semicolon as shown in Table 4.

The same personal name can always be written the same way in Korean, but the notation thereof in Roman characters is not always the same. In many cases, the order of the surname and the first name,

Table 3 Exemplary inaccurate search result.

User query	Park Sung Joon
Search engine query in NDSL or NTIS service	(AU: Park) and (AU: Sung) and (AU: Joon) where “AU” is author name search field
Wanted search result	Park, Sung-Joon; Kim, Ju-Youn; Kim, Young Kuk
Unwanted search result	Park, Dong In; Kim Sung-Joon; Oh, Seung Wan
	Park, Sung-Chul; Lee, Young-Joon; Choi, Min-Ki
	Kim, Sung-Hae; Lee Joon; Park, Myoung-Soo

Table 4 Accuracy of result through index property change and phrase search.

User query	Won-Jae Lee
Search engine query	(AU: “Won-Jae Lee”, mode = “PHRASE”)
Result of suggested phrase search	Sung-Jae Chung; Won-Jae Lee; Keun-Shik Lee; Moon-Sun Chang

characters of the names and detailed alphabets do not always match. For writing the name of “Michael Wick” and “김철수”, although there is a rule of notation in Roman characters for writing personal names, it may be different depending on personal tastes or the requirements of writing paper as shown in Table 5.

The context is the name of a person is not unique in that about 100 million people share 90,000 personal names [1, 2]. It is impossible to get a search result of “M. Wick” or “ML Wick” with a search word of “Michael Wick” without extending the characters and type of personal name, measuring and matching string and phonetic similarity in order to supplement different notations for the same personal name in terms of the search system while the step of author identification is not carried out. That is, when a user uses one of the above notations as a query to find all papers by “Michael Wick”, the issue of author identification should be involved in order to find all results of different notations. However, this study does not handle the issue of author identification. This study intends to suggest author names in different notations only with the metadata not analyzed and processed, and with a search method, and to suggest the result of notations of which the query string is the same but in a different sequence.

If a delimiter is specified in an index property and phrase searching is carried out, this process ensures an accurate search result shown in Table 4. However, if the sequence of notations in Roman characters for the last name and the first name of personal names in Korean or an abbreviation is used for last name in Roman expression, even the same personal name is not included in the search result. To address this issue, the near search is combined among the operators supported by most search engines with the phrase search, and the boosting factor is applied to improve the search method.

Using the search algorithm in Fig. 1 results will include a search for “Chul-Soo Kim” if the query is written as “Kim, Chul-Soo”. Because the phrase search depends on the sequence of writing query words, the phrase search result in which the user query was used is combined with the search result for adjacent operation which operates in a delimiter, and the ranking algorithm is then applied. Because the user wants a matching result with “Kim Chul-Soo” which is correct for the sequence of the words, the phrase search result is boosted to the higher layer through ranking.

First, the paper list including the author name exactly matching the personal name query by the suggested algorithm is suggested. The paper including the author name in which the sequence of notations for the surname and the first name changes according to the notations in Roman characters is also suggested. The search algorithm in the search engine FAST of the NDSL system is written as shown in Eq. (1).

Table 5 Many expressions on person name or Romanize Korean.

Michael Wick	M. Wick; M., Wick; ML Wick; M Wick
김철수	Kim Chul Soo; Kim, Chul-Soo; Chul-Soo Kim; Chul Soo Kim; Kim, C. S.; Kim, Chulsoo; Kim Chul Su; Chul-Su Kim; Kim, Chulsu; Kim. Choel-Soo

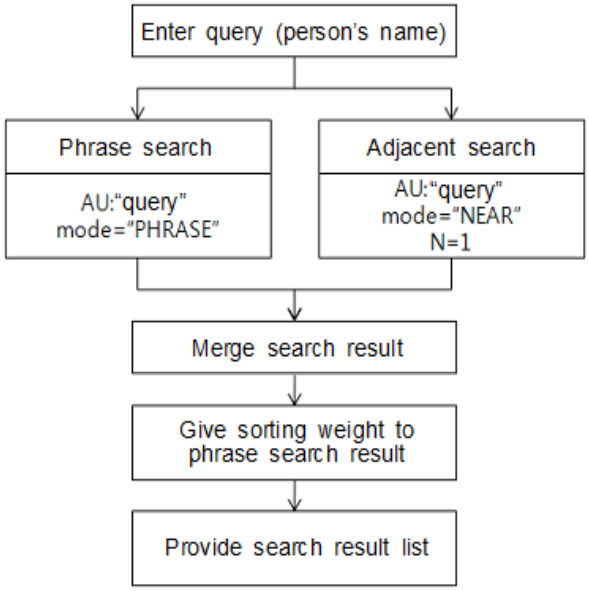

Fig. 1 Algorithm for improving search results.

$$XRANK(OR(AU:query,\mathrm{mod}\,e=PHRASE),\\(AU:query,\mathrm{mod}\,e=NEAR,N=1)),\\(AU:query,\mathrm{mod}\,e=PHRASE),boostall=yes) \quad (1)$$

If the sequence of notations for the surname and the first name changes but they indicate the same personal name in writing personal names of Koreans or foreigners, a result suggesting the relevant papers is significantly more efficient for users. An experiment was carried out in the following condition in order to prove usability of paper search results to which this result improvement algorithm was applied. In our testing, search 1 is an existing method of searching author names in which the white spaces are processed with AND. Search 2 is a method of carrying out the phrase search only in the delimiter described in Table 4. Search 3 is a method of search based on the suggested algorithm described in Fig. 1.

- Search target: 56,752,186 papers provided by the NDSL service;
- 50,000 personal name (written in English) query test sets;
- Search 1: processes white spaces with AND;
- Search 2: phrase search in a delimiter;
- Search 3: combines phrase search with near search in a delimiter.

Table 6 shows the number of paper search results by each search method. If the query is “Lee, Won-Jae”, 4,863 paper results with all of “lee”, “won” and “jae” are obtained regardless of the sequence and position thereof, according to the algorithm of search 1. According to the result of search 2, there are 321 results in which “lee won jae” is positioned in sequence within semicolons in the author field among 4,863 papers. This is regarded as an accurate result for the user query. However, search 3 corresponding to the suggested algorithm provides 11 more paper results which include “Won-Jae Lee”. This cannot ensure the identified real same personal name, but is considered as a very useful search result in consideration of the method of writing personal name strings. (A)-(C) in Table 5 are search results not related to the user query. (C)-(B) are the paper results with the author name the same for the personal name in which the characters used in writing the personal name are the same but the sequence of the surname and the first name is changed.

Table 7 shows the average of paper search results for 50,000 queries for personal names. The number in ((A)-(C))/(A) means that the existing paper search by using the author field results in about 86% inaccuracy, and provides search results not wanted by the user. The suggested algorithm does not include inaccurate search results in paper search by means of personal names, and even includes different notations for the same personal name which are approximately 9% shown by the number in ((C)-(B))/(C) in the search results.

Table 6 Exemplary experiment data for improving search results.

User query	Number of results			(A)-(C)	(C)-(B)
	Search 1 (A)	Search 2 (B)	Search 3 (C)		
Lee, Won-Jae	4,863	321	332	4,531	11
Hwang, Woo-Suk	174	57	59	115	2
Kim, Young-Jin	13,545	682	712	12,833	30
Choi, Jin Young	4,628	291	299	4,329	8
Lee, Jong Ho	5,475	491	507	4,968	16
Ahn, Kang-Min	375	27	28	347	1
Eun-Hee Kim	3,756	0	199	3,557	199
Kim, Dae-Jung	2,102	107	115	1,987	8
Choi, Jung	13,597	1,226	1,238	12,359	12
Choi, J. H.	34,678	3,032	3,153	31,525	121
Park Jae Hyun	3,459	205	288	3,171	83
Oh, Jae-Eung	126	90	91	34	1
Tom P	1,930	262	267	1,663	5
Alex, S.	1,547	354	361	1,186	7
James K	13,828	3,193	3,201	10,627	8

Table 7 Improved performance for search results.

Average results			((A)-(C))/(A)	((C)-(B))/(C)
Search1 (A)	Search2 (B)	Search3 (C)	((A)-(C))/(A)	((C)-(B))/(C)
7,433	884	975	0.868	0.093

Table 8 Exemplary personal name search results. (A Korean character "김철수" is expressed "Kim Chul Soo" in English).

User query	Kim Chul Soo
Author field of paper search result (7 papers)	김철수; 김주연; Kim, Chul-Soo; Kim, Ju-Youn
	김철수; 박명수; Chul-Soo Kim; Myoung-Soo Park
	김성해; 김철수; Kim, Sung-Hae; Kim, Chul-Soo
	Kim, Chul-Soo; Kim, Sung-Hae
	Lee, Joon; Kim, Chul-Soo; Han, Hee-Jun
	Chul-Soo Kim; Ki-Won Lee
	Chul-Soo Kim; Jong-Suk Lee; Ki-Won Lee
Author name search results through grouping information	Kim, Chul-Soo (4) Chul-Soo Kim (3) 김철수 (3) Ki-Won Lee (2) Kim, Sung-Hae (2) 김성해 (1) Lee, Joon (1) Han, Hee-Jun (1) Kim, Ju-Youn (1) 김주연 (1) Myoung-Soo Park (1) 박명수 (1) Jong-Suk Lee (1)

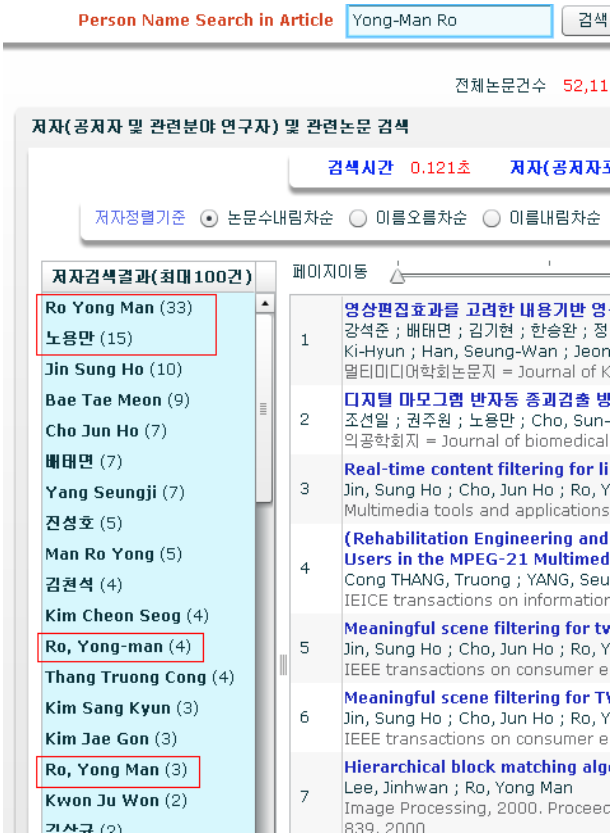

Fig. 2 Paper Search based on proposed method.

4.2 Indexing Efficiency

Author names are first searched in the author field included in the paper information without searching an independent author list. The personal name grouping information is then created from the paper result set to list the papers which have more authors in sequence. Table 8 shows the result of searching personal names. This method does not need a process of indexing 22 million author names extracted from 56 million papers only for finding authors. Therefore, this method reduces approximately 28% of the volume of indexing, and improves approximately 10% of disk storage efficiency for storing index binaries.

(1) Sum of paper and author information indexes: approx. 78 million indexes;

(2) Indexes of papers: approx. 56 million indexes

Index volume improvement: ((1)-(2))/(1) = 0.282

(3) Capacity for storing paper and author information indexes: 840 GB;

(4) Capacity for storing paper indexes: 750 GB

Disk capacity improvement: ((3)-(4))/(3) = 0.107

5. Conclusions

All information has creator data in the form of author names (authors for academic information, inventors, researchers, applicants, or analyzers). Searching personal names is one of the important functions in search services through the web. Most academic information systems for metadata of personal names cannot often provide accurate search results because of the same indexing and searching as other information, for example, titles, abstract, etc.

This study described a method of searching information and improving the accuracy of search in consideration of various personal name notations and properties in order to refine searching personal names in the academic information services. An efficient index design was also described for finding personal names, which was proved to reduce approximately 28% of index capacity and 10% of disk capacity in the NDSL science and technology information integration

service. A method of effectively providing users with useful information, for example, co-authors and involved researcher lists in searching personal names was described. Fig. 2 shows our paper search system based on our proposed personal name search method.

It is necessary to further study a method of analyzing metadata including personal names, e-mail addresses, and institutions for which the relevant person works to combine a plurality of properties and to apply the concept of author identification, in order to provide academic information related to the relevant person identified in the real world. More useful academic information services can be implemented by combining accurate metadata search by using a search engine with the result of author identification.

References

[1] Guha, R. V., & Garg, A. 2004. "Disambiguating People in Search." In *Proceedings of the 13th World Wide Web Conference*, ACM Press.

[2] Christen, P. 2006. "A Comparison of Personal Name Matching: Techniques and Practical Issues." In *Data Mining Workshops, ICDM Workshops, Sixth IEEE International Conference on IEEE*, 290-4.

[3] http://www.ndsl.kr.

[4] Winkler, W. 2006. "Overview of Record Linkage and Current Research Directions." Research Report Series #2006-2, Statistical Research Division, U.S. Census Bureau.

[5] Culotta, A., Kanani, P., Hall, R., Wick, M., and McCallum, A. 2007. "Author Disambiguation Using Error-Driven Machine Learning with a Ranking Loss Function." IIWeb-2007.

[6] Kanani, P., McCallum, A., and Pal, C. 2007. "Improving Author Coreference by Resource-Bounded Information Gathering from the Web." IJCAI-2007.

[7] Vu, Q. M., Masada, T., Takasu, A., and Adachi, J. 2007. "Disambiguation of People in Web Search Using a Knowledge Base." In *Research, Innovation and Vision for the Future, 2007 IEEE International Conference on IEEE*, 185-91.

[8] Artiles, J., Gonzalo, J., and Verdejo, F. 2005. "A testbed for People Searching Stategies in the www." In *Proc. of SIGIR'05*, 569-70.