Method of Improving Personal Name Search in Academic Information Service

Heejun Han*, Seok-Hyoung Lee**

ARTICLE INFO

ABSTRACT

All academic information on the web or elsewhere 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 information is composed of a title, an author, and contents. An essay 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 personal names. It also describes a method for providing the results of searching co-authors and related researchers in searching personal names. This method can be effectively applied to providing accurate and additional search results in the academic information services.

1. Introduction

Personal names on the web (World Wide 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 (Guha & Garg, 2004).

All current information is produced by a creator, who is referred to as an author for an essay, 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

* Senior researcher, NTIS Center, Korea Institute of Science and Technology Information, Korea (hhj@kisti.re.kr)
** Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information, Korea (skyi@kisti.re.kr) (Corresponding Author)

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 (Christen, 2006).

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 essays, patents, theses, research reports, industrial standards, science and technology work forces, trend analysis information, and factual information for the science and technology information integration service of the NDSL (National Discovery for Science Leaders) provided by the KISTI (Korea Institute of Science and Technology Information).

Table 1. Exemplary metadata for personal name in academic information

Category	Classification	Format
Korean essays	author (in Korean)	박성준 ; 김주연 ; 김영국
Korean essays	author (in English)	Park, Sung-Joon ; Kim, Ju-Youn ; Kim, Young-Kuk
Overseas essay	author	Zhou, Fushan ; Yang, Deng-Ke ; Molitor, R.J.
Patent	inventor	Sakurada, Masahiro ; Yamanaka, Hideki ; Ohta, Tomohiko
Research report	participant (in Korean)	류범종 ; 김진숙 ; 진두석 ; 이석형
Research report	participant (in English)	Ryu, Beom-Jong ; Kim, Jin-Sook ; Jin, Doo-Seok ; Lee, Seok-Hyung
Trend analysis	analyzer	박로학 ; James Lee

The academic information search service on the web undergoes the process of indexing data required for search, and all search engines for processing data in Korean carry out the step of morphological analysis. This is done by delimiting white spaces, and separating compound words and postpositional words. However, because personal names are data which cannot be decided with compound words and postpositional words, most academic information search services as well as the NDSL carry out delimiting white spaces or separation in delimiters in indexing the personal name field. 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 tokening data in white spaces or delimiters as done for personal names in Korean.

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 name field for essay 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.

This study describes an efficient method of indexing for searching personal names focusing on the NDSL service, and refining the process of search. Section 2 describes related studies; Section 3 describes issues associated with searching personal names; Secction 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 search, including academic information services, news, and other knowledge information, an issue involved is that indexing and searching technology based on simple strings can provide information including inaccurate personal names.

   If data is 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 (Winkler, 2006) for connecting different records that show the same personal name, and a process of author identification (Culotta et al., 2007) for forming a group of entities represented from identification properties, for example, language analysis, essay titles, e-mail addresses, journal names, and institutions to which authors belong (Kanani, McCallum, & Pal, 2007). 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 identification of author information from the enormous amount of available information is still not in an optimal state, and does not completely ensure that the identified author is the same person. Therefore, there is a need for providing more accurate data by changing search models without processing information, and removing unnecessary search results.

   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 personal names is written.

3. Issues involved in personal name search

3.1. Inaccurate search results

   In NDSL essay data, a plurality of personal names is listed with delimiters as shown in Table 1. Indexing the essay author name field is carried out by tokening white spaces and possible special characters (; - , .) to produce indexes. Table 2 shows an exemplary indexing result for an author name field.

Table 2. Result of indexing essay author name

Original data	박성준 ; 김영국 ; 송병수 ; Park, Sung-Joon ; Kim, Young Kuk ; Song, Byung-Soo
Indexes	박성준 김영국 park sung joon kim young kuk song byung soo

All search systems convert application user's queries to those that can be processed by a search engine. In the NDSL, AND operation is used to process white spaces in user queries. Assuming that the index field of the author name is AU if the user query is '김영국 김주연', the search language 'AU:김영국 and AU:김주연' is transferred to the search engine. If personal names in Korean are used as a query, search accuracy does not matter, but unwanted search results are obtained if a query is in English. The words 'park', 'sung', and 'joon' are unwanted search results because they exist regardless of the order and the position of the author field of the relevant essay if the query is 'Park Sung Joon', as shown in Table 3.

Table 3. Exemplary inaccurate search result

User query	Park Sung Joon
Search engine query	(AU:Park) and (AU:Sung) and (AU:Joon)
Wanted search result	박성준 ; 김주연 ; 김영국 ; Park, Sung-Joon ; Kim, Ju-Youn ; Kim, Young Kuk.
Unwanted search result	Park, Dong In ; Kim Sung-Joon ; Oh, Seung Wan
Unwanted search result	박승철 ; 이영준 ; 최민기 ; Park, Sung-Chul ; Lee, Young-Joon ; Choi, Min-Ki
Unwanted search result	김승해 ; 이준 ; 박명수 ; Kim, Sung-Hae ; Lee Joon ; Park, Myoung-Soo

The NDSL service supplements query processing about the author name field by means of phrase search 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 essays by a person called 'Seo, Joung Min' in English, essays 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) 박서정 ; 민병철 ; Park, Seo Joung ; Min, Byoung Churl

3.2. Inefficient indexing

The NDSL includes Find Author which is one of the essay search functions. This option is provided for cases when a user does not know an accurate author name or knows only some part of an author's name. That is, Find Author contributes to searching personal names in all author lists of about 52 million essays. This is subject to the following pre-processes:

① Extract meta-information in an essay.
② Check redundant author name at the row level.
③ Create an author name for sorting.
④ Extract the first character for searching an initial sound (Chosung).
⑤ Load onto the Oracle author information table.
⑥ Index the author information and then provide searching.

The number of author records of which the redundancy was extracted from 52 million essay records in the NDSL and then removed by string processing is approximately 22 million. Approximately 50,000 pieces of essay 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 essay index information in order to implement Find Author. This has a negative impact on search speed and disk load in hardware.

4. Mapping system for classification systems

4.1. Improving accuracy of search results

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 in Table 3. 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 meta fields described as personal names in all academic information, such as essays, 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.

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

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, characters of the names, and detailed alphabets do not always match. For writing the name of '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 essays.

> ex) 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

The context is the name of a person is not unique in that about 100 million people share 90,000 personal names. It is impossible to get a search result of 'Kim Chul-Su' or 'Kim C. S.' with a search word of 'Kim Chul-Soo' 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 essays by '김철수',

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, 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.
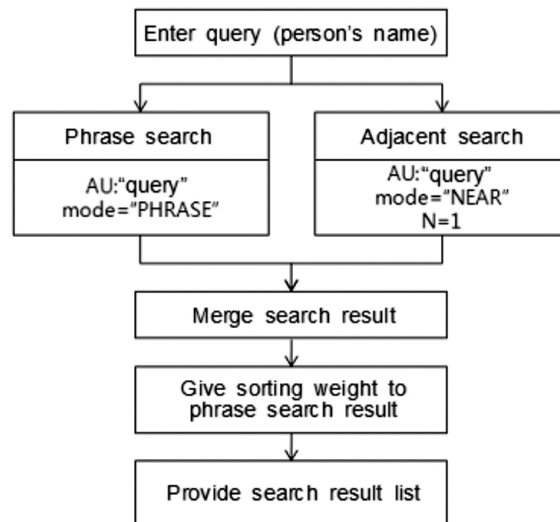
Fig. 1. Algorithm for improving search results

Using the search algorithm in Figure 1 results will include a search for 'Chul-Soo Kim' if the query is written as 'Kim, Chul-Soo'. Because the phrase search provided by the FAST engine 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 wanted 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 essay list including the author name exactly matching the personal name query by the suggested algorithm is suggested. The essays 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 equation 1. Because the used operator is supported by all search engines, such as IDOL K2, DOCCRUZER, and KRISTAL, they can be applied effectively to searching personal names.

$$XRANK(OR((AU:query, mode = PHRASE), (AU:query, mode = NEAR, N = 1)), \\ (AU:query, mode = PHRASE), boost all = yes) \tag{1}$$

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 essays is significantly more efficient for users. An experiment was carried out in the following condition in order to prove usability of essay search results to which this result improvement algorithm was applied. The entirety of NDSL's essays were searched, and 10,000 queries for personal names written in Roman characters extracted from essay metadata in advance were used. Three types of search were carried out for the purpose of comparison. Search 1 is an existing method of searching personal 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 Figure 1.

- Essays searched : 66,752,186 essays provided by the NDSL service.
- 10,000 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 5. Exemplary experiment data for improving search results

Query	Number of results			(A)-(C)	(C)-(B)
Query	search1 (A)	search2 (B)	search3 (C)	(A)-(C)	(C)-(B)
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 5 shows the number of essay search results by each search method. If the query is 'Lee, Won-Jae', 4,863 essay 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 essays. This is regarded as an accurate result for the user query. However, search 3 corresponding to the suggested algorithm provides 11 more essay 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 essay 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 6 shows the average of essay search results for 10,000 queries for personal names. The number in ((A)-(C))/(A) means that the existing essay 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 essay 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. 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

4.2. Efficient indexing

The NDSL provides Find Author as an additional function for finding academic information, such as essays, patents, etc. This option is used to extract only the personal name field from metadata of the academic information for indexing. Therefore, this process can be an unnecessary load in terms of indexing, and has a negative impact on the search speed due to an enormous increase in indexing volume.

Now, Find Author is described by using the personal name field index information already included in the metadata of essays and patents. The suggested method uses the navigator of the FAST engine which is the concept of search result grouping in order to vertically list the personal name search results wanted by a user by searching the data horizontally listed in the personal name (author, applicant, inventor, etc.) field of academic information. Figure 2 shows the suggested method of implementing Find Author.
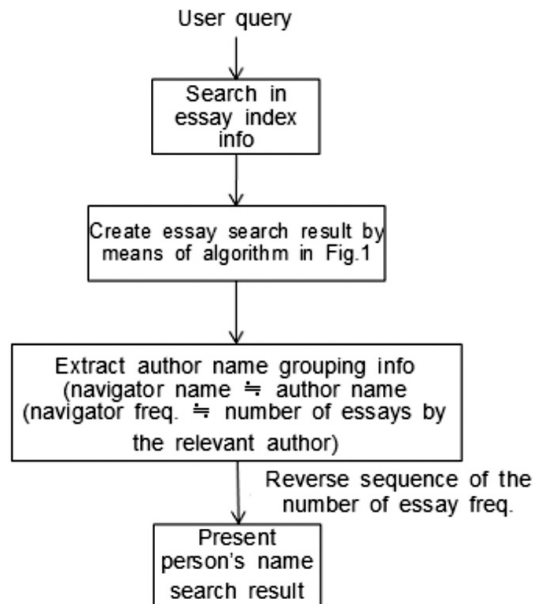
Fig. 2. Implementing Find Author Name

In this method, personal names are first searched in the author field included in the essay information without searching an independent author list. The personal name grouping information is then created from the essay result set to list the essays which have more authors in sequence. Table 7 shows the result of searching personal names. This method does not need a process of indexing 22 million author names extracted from 56 million essays 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 essay and author information indexes : approx. 78 million indexes
- (2) Indexes of essays : approx. 56 million indexes
- Index volume improvement : $((1)-(2))/(1) = 0.282$
- (3) Capacity for storing essay and author information indexes : 840GB
- (4) Capacity for storing essay indexes : 750GB
- Disk capacity improvement : $((3)-(4))/(3) = 0.107$

Table 7. Exemplary personal name search results

User query	Kim Chul Soo
Author field of essay search results (5 essays)	김철수 ; 김주연 ; Kim, Chul-Soo ; Kim, Ju-Youn
Author field of essay search result (5 essays)	김철수 ; 박명수 ; Chul-Soo Kim ; Myoung-Soo Park
Author field of essay search result (5 essays)	김성해 ; 김철수 ; Kim, Sung-Hae ; Kim, Chul-Soo
Author field of essay search result (5 essays)	Kim, Chul-Soo ; Kim, Sung-Hae
Author field of essay search result (5 essays)	Lee, Joon ; Kim, Chul-Soo ; Han, Hee-Jun
Author field of essay search result (5 essays)	Chul-Soo Kim ; Ki-Won Lee
Author field of essay search result (5 essays)	Chul-Soo Kim ; Jong-Suk Lee ; Ki-Won Lee
Personal 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)

4.3. Providing co-authors and involved researchers

   The author metadata of Korean essays is written as a pair of names in Korean and English in most cases. Therefore, if a query is written in English, the name written in Korean can be suggested together with the name written in Roman characters with a changed sequence of writing the surname and the first name. However, the names in Roman characters with different spellings cannot be suggested. That is, if a query is 'Kim Chul Soo', 'Chul Su Kim' is not searched. However, the suggested method always presents the relevant name written in Korean as a search result if a user query is a personal name written in English, and the user selects the presented personal name in Korean or enables access to other notations of the personal name written in English.

   Because using the grouping method provided by a search engine contributes to extracting author name properties from an essay result set, it is possible to suggest the name in Korean, English, and other notations of the name in English related to the query, and the search result also includes co-author names. The user is suggested both 'Chul-Soo Kim' and '김철수' by using a query of 'Kim Chul Soo', and receives the list of 'Ki-Won Lee', 'Kim, Sung-Hae', '김성해', and 'Lee, Joon', co-authors. It is expected that the suggested personal names are researchers in a related field, and this is an effective method of searching essays in related research fields by using co-author names.

   The academic information service is very useful for academic information request through searching personal names, which can improve the accuracy of search results, does not create unnecessary

indexes for Find Author in consideration of various personal names written for the same person, and can provide a related researcher list. To verify the service, an essay search system for personal name queries is implemented for 52 million NDSL essays.

This system provides results for searching author names including the names in Korean and English and other notions in different spellings for the name in English, as shown in Figure 3, if the query is for a personal name, and also provides an academic essay list. Users can search the essays of the author selected by clicking the personal name in the author search result. The method suggested in this study ensures accurate essay results in searching personal names, and effectively provides the number of an author's essays and a co-author list which can be considered as people involved in related research.
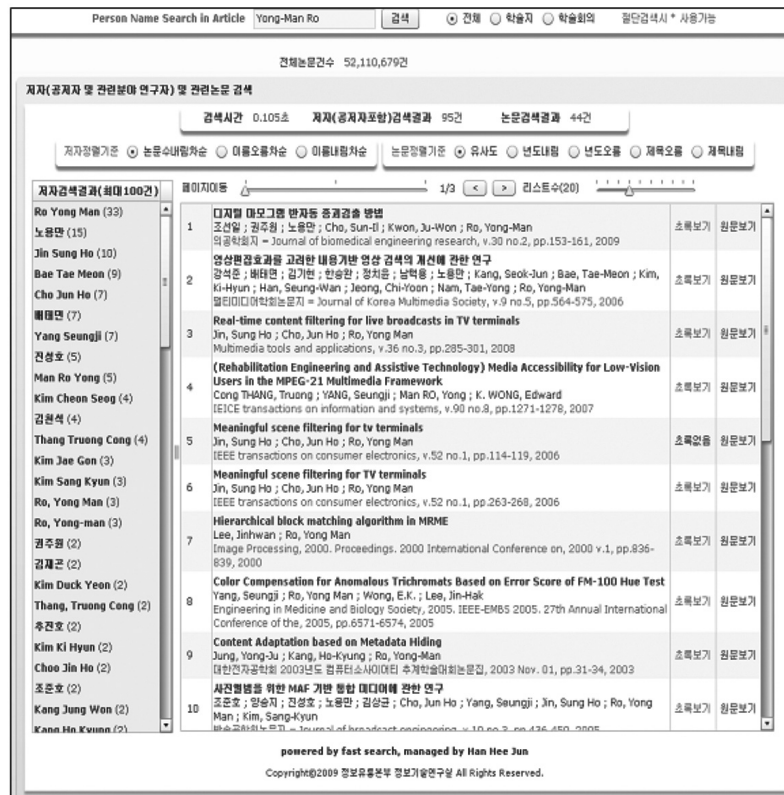

Fig. 3. Search results

Fig. 4. Essay search system for personal name queries

5. Conclusion

All information has creator data in the form of personal 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.

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

Christen, P. (2006). A comparison of personal name matching: techniques and practical issues. *IEEE International Conference on Data Mining - ICDM*, 2006, 290-294. doi:10.1109/ICDMW.2006.2

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. *Workshop on Information Integration on the Web - WIIW*, 2006, 32-37.

Guha, R. V., & Garg, A. (2004). Disambiguating people in search. *World Wide Web Conference Series - WWW*, 2004.

Kanani, P., McCallum, A., & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the web. *International Joint Conference on Artificial Intelligence*, 2007, 429-434.

Noh, Y., Choi, M. J., Choi, Y. W., Jeong, S. W., Jung, E. J., Kang, M. S., ··· & Park, S. Y. (2011). An analysis of user satisfaction of K university's library service. *International Journal of Knowledge Content Development & Technology*, *1*(1), 61-79.

Winkler, W. E. (2006). *Overview of Record Linkage and Current Research Directions*. Washington, DC 20233, U.S : Statistical Research Division, U.S. Census Bureau.