학술정보서비스에서 인명검색 고도화 방법

Enhanced Method for Person Name Retrieval in Academic Information Service

한희준, 예용희, 류범종
한국과학기술정보연구원 정보유통본부

Hee-Jun Han(hhj@kisti.re.kr), Yong-Hee Yae(yaeyh@kisti.re.kr),
Beom-Jong You(ybj@kisti.re.kr)

요약

웹이든 웹이 아니든 존재하는 모든 학술정보에는 창작자, 즉 그 정보를 생산한 주체가 존재한다. 그 주체는 개인, 단체, 기관이 될 수 있으며 또는 해당 정보의 성격에 따라 국가가 될 수도 있다. 대부분의 정보는 제목과 저자, 내용으로 구성된다. 학술정보 가운데 논문의 경우 제목, 저자, 키워드, 요약, 발행일, 발행처, ISSN 등의 메타정보로 기술되며, 특허의 경우는 명칭, 출원인, 발명자, 대리인, IPC, 출원번호, 청구항 등의 메타정보로 표현된다. 대부분의 웹 기반의 학술정보 서비스에서는 이들 메타정보를 가공 및 처리하여 사용자들에게 검색기능을 제공하며, 특히 인명에 해당하는 저자필드를 이용한 검색기능은 중요한 요소이다. 본 논문에서는 인명검색을 위한 효율적인 색인운영과 구검색 기반의 부스팅 요소를 적용한 인접연산 결과 랭킹 알고리즘을 이용해 인명검색 결과의 정확성 개선 방법을 제안하며, 인명검색시 공저자 및 관련연구자 검색결과를 제공하는 방법을 설명한다. 이는 학술정보서비스에 있어서 정확하고 부가적인 검색결과를 제공하는데 효과적으로 적용될 수 있다.

■ 중심어 : | 인명검색 | 정보검색 | NDSL |

Abstract

In the web or not, all academic information have the creator which produces that information. The creator can be individual, organization, institution, or country. Most information consist of the title, author and content. The article among academic information is described by title, author, keywords, abstract, publisher, ISSN(International Standard Serial Number) and etc., and the patent information is consisted some metadata such as invention title, applicant, inventors, agents, application number, claim items etc. Most web-based academic information services provide search functions to user by processing and handling these metadata, and the search function using the author field is important. In this paper, we propose an effective indexing management for person name search, and search techniques using boosting factor and near operation based on phrase search to improve precision rate of search result. And we describe person name retrieval result with another expression name, co-authors and persons in same research field. The approach presented in this paper provides accurate data and additional search results to user efficiently.

■ keyword : | Person Name Retrieval | Information Retrieval | NDSL |

접수번호 : #090922-002
접수일자 : 2009년 09월 22일
심사완료일 : 2009년 12월 22일
교신저자 : 한희준, e-mail : hhj@kisti.re.kr

I. 서 론

월드 와이드 웹(World Wide Web) 상에서 인명(person name)은 모든 검색엔진 질의어의 30%에 해당할 만큼 중요한 요소이며, 웹 어플리케이션에서 인명을 통한 검색은 중요한 기능 중 하나이다[1][4].

현존하는 모든 정보는 창작자가 존재하는데 학술정보의 측면에서 논문의 경우 저자로, 특허는 출원인, 발명자 및 대리인으로, 연구보고서는 연구참여자, 동향분석 자료는 조사자 및 분석자 등으로 통용되고 있다. 이렇게 다양한 단어로 이해되는 정보의 창작자는 개인이 될 수도 있고, 기관, 단체, 국가 또는 크롤러와 같은 일종의 컴퓨터 시스템이 될 수도 있다[4].

학술정보에서 창작자는 대부분 사람의 이름, 즉 인명(Person name)이다. 예를 들면 한국과학기술정보연구원에서 서비스 중인 NDSL 과학기술정보통합서비스[12]의 대상인 논문, 특허, 학위논문, 연구보고서, 산업표준, 과학기술인력, 동향분석정보, 사실정보의 경우 전체 데이터의 약 95% 이상이 저자, 출원인, 연구참여자 등의 정보의 창작자 필드가 인명으로 기술되어지며, 그 표기 형식은 [표 1]과 같다.

표 1. 학술정보에서의 인명 메타정보의 예

종류	구분	표기형식
국내 논문	저자(한글)	박성준 ; 김주연 ; 김영국
	저자(영문)	Park, Sung-Joon ; Kim, Ju-Youn ; Kim, Young-Kuk
해외 논문	저자	Zhou, Fushan ; Yang, Deng-Ke ; Molitor, R.J.
특허	발명자	Sakurada, Masahiro ; Yamanaka, Hideki ; Ohta, Tomohiko
연구 보고서	참여자 (한글)	류범종 ; 김진숙 ; 진두석 ; 이석형
	참여자 (영문)	Ryu, Beom-Jong ; Kim, Jin-Sook ; Jin, Doo-Seok ; Lee, Seok-Hyung
동향 분석	분석자	박로학 ; James Lee

웹 상에서 학술정보 검색서비스를 하기 위해서 검색에 필요한 데이터를 색인하는 과정을 거치는데, 한글 데이터를 처리하는 모든 검색엔진은 형태소분석의 단계를 수행한다. 이는 공백문자 구분, 복합어분리, 조사분리 등으로 설명된다. 그러나 인명의 경우 복합어 및 조사로 판단할 수 없는 데이터이므로, NDSL 뿐만 아니라, 대부분의 학술정보검색서비스에서 인명필드를 색인할 때 공백문자 구분 또는 구분자(delimiter) 단위의 분리만을 수행하고 있다. 또한 영문 데이터의 색인 과정에는 스테밍(stemming), 단복수 처리 등의 단계를 거치지만, 영문 인명의 경우 한글 인명과 마찬가지로, 공백문자나 구분자 단위로만 데이터를 토큰화해서 색인어를 추출한다.

Web of Secience, Scopus, NDSL 과학기술정보통합서비스와 대부분의 학술정보서비스에서는 논문검색의 경우 저자명 필드를 통한 검색 기능을 기본적으로 제공한다. 특히 Scopus 나 NDSL 의 경우, 저자명 DB를 별도로 구축 및 색인한 후 저자찾기 기능을 통해 저자명을 찾기 위한 기능을 별도로 제공한다[13].

본 논문에서는 NDSL 서비스를 중심으로 인명검색을 위한 색인운영의 효율성과 검색기능의 고도화에 대해 논한다. 2장에서는 관련연구를 소개한다. 3장에서 기존 인명검색의 문제점에 대해 논하고 4장에서는 제안하는 방법을 기술하며, 5장에서 결론을 맺는다.

II. 관련연구

인명 검색은 학술정보 서비스뿐만 아니라, 뉴스 및 기타 지식정보 등 모든 웹문서의 검색에서 그 비율이 커져가고 있지만, 단순 문자열에 기반한 색인 및 검색 기술이 부정확한 인명을 포함한 정보를 사용자에게 제공한다는 문제점을 지닌다.

학술정보에 나타난 저자의 이름과 실세계의 저자를 식별한 데이터가 존재하고, 인명 표기법이 일정하다면 정확한 검색결과를 제시할 수 있다. 하지만 동일 인명을 나타내는 서로 다른 레코드를 연결하는 레코드 링키지 기법[9] 혹은 언어분석 및 논문 제목, 이메일, 게제지명, 소속기관 등 식별 자질로부터 표현된 개체들 간의 군집화를 형성하는 저자식별[10][11]의 과정이 선행되어야 하고, 다양한 인명 표기법이 정형화되어야 한다. 하지만 현재 인명검색을 위해 저자식별 데이터를 구축

활용하거나, 인명 질의어를 다양한 표기형식으로 확장하여 검색해주는 시스템은 없다.

방대한 정보로부터 저자 정보를 식별하는 것은 엄청난 노력이 필요하며, 식별된 저자가 동일인이라는 것을 완벽히 보장하지는 못한다. 하지만 정보를 전혀 가공하지 않고 검색 모델의 변화를 통해 보다 정확한 데이터를 제공하고, 불필요한 검색결과를 제거할 필요성이 있다.

대표적인 학술정보서비스인 Web of Science, Scopus, NDSL 사이트에서 인명 또는 저자필드 검색은 기본적으로 제공하는 기능이며, 제목, 내용, 출처 필드 등과 함께 기본검색 필드에 포함되어 있다. 사용자 질의어에 존재하는 공백문자는 모두 AND 연산자로 처리되는데, 인명이 질의어일 때는 정확하지 않은 검색결과가 포함된다. 이를 보완하고자 NDSL과 Scopus 의 경우는 인명질의어를 구(phrase)검색 처리하기 위하여 큰따옴표(double quotation marks) 사용이 가능하지만, 다중 인명으로 표기된 저자필드에서는 여전히 부정확한 결과를 초래한다.

III. 인명검색의 문제점

1. 검색결과의 비정확성

NDSL 논문 데이터의 경우 [표 1]에서 보는바와 같이 다수의 인명이 구분자와 함께 나열된 형태이다. 논문 저자명 필드의 색인은 존재할 수 있는 공백문자(white space)와 존재할 수 있는 특수문자(; - , .) 단위로 토큰화하여 색인어를 생성한다. [표 2]는 저자명 필드의 색인결과의 예이다.

표 2. 논문 저자명 색인 결과

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

모든 검색시스템은 어플리케이션의 사용자 질의어를 검색엔진이 처리 가능한 질의어로 변환한다. NDSL 에서는 사용자 질의어에 존재하는 공백문자를 AND 연산으로 처리하는데, 만약 사용자 질의어가 '김영국 김주연' 일 경우에, 저자명의 색인필드가 AU 라고 가정하면 검색엔진에는 'AU:김영국 and AU:김주연' 으로 검색어가 전달된다. 한글 인명을 질의어로 사용한 경우 검색 정확성이 문제되지는 않으나, 질의어가 영문일 경우에는 원하지 않는 검색결과가 나타나는 문제점이 발생한다. 표 3에서와 같이 질의어가 'Park Sung Joon' 일 경우 'park' 과 'sung' 과 'joon' 이라는 단어가 논문의 저자필드에 순서와 위치에 상관없이 존재한다는 이유로 원하지 않는 검색결과가 된다.

표 3. 검색결과의 비정확성의 예

사용자질의어	Park Sung Joon
검색엔진 질의어	(AU:Park) and (AU:Sung) and (AU:Joon)
원하는 검색결과	박성준 ; 김주연 ; 김영국 ; Park, Sung-Joon ; Kim, Ju-Youn ; Kim, Young Kuk.
원하지 않는 검색결과	Park, Dong In ; Kim Sung-Joon ; Oh, Seung Wan
	박승철 ; 이영준 ; 최민기 ; Park, Sung-Chul ; Lee, Young-Joon ; Choi, Min-Ki
	김승해 ; 이준 ; 박명수 ; Kim, Sung-Hae ; Lee Joon ; Park, Myoung-Soo

NDSL 서비스에서는 [표 3]의 원하지 않는 검색결과를 배제하기 위해 저자명 필드에 대한 질의어 처리를 구검색(phrase search)으로 보완하고는 있지만 여전히 사용자 의도와 맞지 않는 검색결과가 존재한다. 예를 들면 사용자가 영문성명 'Seo, Joung Min' 이라는 사람의 논문을 찾고자 검색을 했을 때, 아래와 같이 질의어 단어가 구분자를 사이에 두고 인접한 경우의 저자 리스트를 가진 논문이 검색결과로 제시되며 이는 사용자가 원하지 않는 결과가 분명하다.

- 박서정 ; 민병철 ; Park, Seo Joung ; Min, Byoung Churl

2. 색인의 비효율성

NDSL에서 논문검색을 위한 여러 기능 중 저자찾기가 존재한다. 이는 정확한 저자명을 알지 못하거나, 저자명의 일부분만 알고 있을 경우를 위해 제공하는 기능

이다. 즉, 저자찾기 기능을 이용하면 약 5,200만 건에 해당하는 논문의 모든 저자리스트에 존재하는 인명을 검색할 수 있다. 이는 아래와 같은 선처리 과정을 거친다.

① 논문에서 저자 메타정보 추출
② 문자열 레벨의 저자명 중복체크
③ 정렬(sorting)을 위한 저자명 생성
④ 초성검색을 위한 첫 문자 추출
⑤ 오라클 저자정보 테이블에 적재
⑥ 저자정보를 색인 후 검색기능 제공

NDSL에서 5,200여만 건의 논문 레코드에서 추출되어 문자열 처리에 의해 중복제거 된 저자 레코드 수는 약 2,200만 건이다. 논문 정보의 경우 일주일 단위로 약 5만 여건이 입수되는데 40% 비율로 별도의 저자정보가 계속해서 생성되고 있는 것이다. 저자찾기 기능을 제공하기 위해서 이미 논문 색인정보에 존재하는 저자명을 추가로 색인하는 것은 색인 생성 및 관리 측면에서 낭비이며, 이는 하드웨어적으로 검색속도 및 디스크 부하에 부정적인 영향이 된다.

IV. 제안하는 방법

1. 검색결과의 정확성 개선

[표 3]에서의 원하지 않는 검색결과를 제외시키기 위해서는 나열된 인명을 구분하는 기호인 세미콜론(;) 단위의 검색처리가 필요하다. 구분자를 지정해서 해당 구분자 내에서만 검색 연산자가 동작하도록 설계하였다. 즉, 논문, 특허, 연구보고서, 분석동향 정보 등 모든 학술정보에서 인명으로 기술된 모든 메타필드에 대해서는 구분자(separator) 속성을 세미콜론으로 지정하여 색인처리 하였다. 이 때 구검색은 [표 4]와 같이 세미콜론 내에서 순서에 맞는 연속된 색인어가 존재하는 경우에만 검색결과를 가져온다.

표 4. 색인 속성변경과 구검색을 통한 결과 정확성

사용자질의어	Won-Jae Lee
검색엔진 질의어	(AU: "Won-Jae Lee ", mode="PHRASE")
제안 구검색 검색결과	Sung-Jae Chung ; Won-Jae Lee ; Keun-Shik Lee ; Moon-Sun Chang

동일인명을 표기할 때 한글의 경우 일관적이지만 로마자 표기의 경우에는 그렇지 않다. 성과 이름의 순서, 이름의 표기 양상, 철자의 세 항목에 따라 일관성 없이 표기되는 경우가 다양하다[5][6]. 아래의 예와 같이 '김철수'라는 사람의 이름을 표기하는 방식은 인명의 로마자 표기법이 정해놓은 규칙이 있음에도 불구하고 개인의 취향, 또는 논문작성의 요구방식에 따라 다를 수 있다.

- 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

약 1억명의 사람이 9만개의 인명을 공유할 만큼 개인마다 인명이 유일하지 않은 상황[1-3]에서 저자식별의 단계를 거치지 않은 채로, 또는 동일 인명의 서로 다른 표기를 검색 시스템적으로 보완하기 위해 인명의 철자 및 유형 확장[5][7], 문자열과 음성학적인 유사도 측정과 매칭처리를[8] 수행하지 않고서는 'Kim Chul-Soo'라는 검색어로 'Kim Chul-Su' 또는 'Kim C. S.' 라는 검색결과를 얻는 것은 불가능하다. 즉, 검색 사용자가 '김철수'라는 사람의 논문을 모두 찾기 위해서 질의어로 위 표기형식 중 어느 하나를 사용할 때 다른 표기형식의 결과까지 모두 찾아주기 위해서는 저자식별의 문제가 관여되어야 하지만 본 논문에서는 저자식별의 문제를 논하지는 않는다. 분석 및 처리하지 않은 메타데이터 자체와 검색기법 만으로 다른 표기형식의 저자명을 제시하고, 질의어의 문자열은 같지만 순서가 바뀐 표기의 결과까지 제시해주고자 한다.

색인 속성에 구분자를 지정하고, 구검색을 수행할 경우 표 4와 같이 정확한 검색결과를 보장하지만, 한글성

명 성(last name)과 이름(first name)의 로마자 표기 순서가 바뀌는 경우에는 동일 인명이라 할지라도 검색결과에서 제외된다. 이 점을 보완하기 위해 구검색에 대부분의 검색엔진이 지원하는 연산자 중 인접 검색(near search)을 조합하고, 부스팅 요소(boosting factor)를 적용하여 검색 기법을 개선하였다.

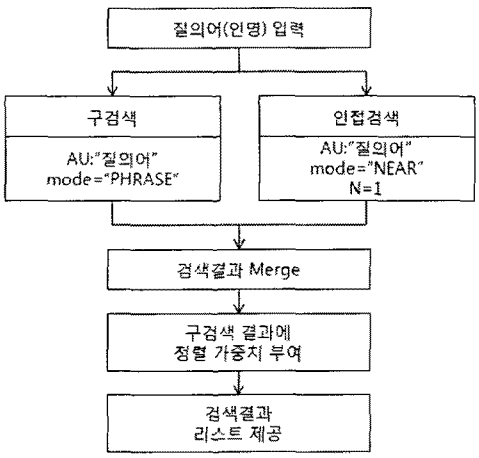

그림 1. 검색결과 개선 알고리즘

[그림 1]에서의 검색 알고리즘을 이용하면 질의어가 'Kim, Chul-Soo' 일 경우 'Chul-Soo Kim' 의 검색 결과를 얻을 수 있다. FAST 엔진에서 제공하는 구검색은 질의어 단어의 나열 순서에 종속적이기 때문에, 먼저 사용자 질의어를 사용한 구검색 결과와 구분자 내에서 동작하는 인접연산 검색 결과를 합친 후에 랭킹 알고리즘을 적용한다. 사용자의 의도는 'Kim Chul-Soo' 가 정확히 단어 순서에 맞게 매칭된 결과를 원하였기 때문에 랭킹을 통해 구검색 결과를 상위로 부스팅 한다.

제시하는 알고리즘에 의해 사용자의 인명 질의어와 정확히 일치하는 저자명이 포함된 논문리스트가 먼저 제시되며, 또한 로마자 표기에 의해서 성과 이름의 표기 순서가 바뀐 저자명이 포함된 논문도 제시된다. NDSL 시스템의 검색엔진 FAST 에서의 검색 알고리즘은 수식 1과 같이 표현된다. 사용되는 연산자는 FAST 뿐만 아니라 IDOL K2, 독크루저, KRISTAL 등 거의 모든 검색엔진에서 지원하고 있는 것이므로, 인명을 통한 검색 기능에는 효과적으로 적용될 수 있다.

$$
\begin{aligned}
&XRANK(\\
&OR((AU{:}\ query,\ \mathrm{mode} = PHRASE), \\
&(AU{:}\ query, \mathrm{mode} = NEAR,\ N=1)), \\
&(AU{:}\ query,\ \mathrm{mode} = PHRASE), \\
&boostall = yes)
\end{aligned}
\tag{1}
$$

내국인이든 외국인이든 인명 표기에 있어서 성과 이름의 표기 순서가 바뀌었으나 동일 인명을 나타내는 경우, 해당 논문을 결과로 제시해 주는 것은 이용자 측면에서 상당히 효율적인 정보가 된다. 제안된 검색결과 개선 알고리즘을 적용한 논문 검색 결과의 유용성을 증명하기 위해 아래와 같은 조건에서 실험을 수행하였다. 검색대상은 NDSL 논문 전체이며, 논문 메타정보로부터 미리 추출한 로마자 표기 인명 질의어 만 건을 검색에 사용하였다. 비교하기 위한 검색은 세 분류로 수행하였는데, 검색방법 1은 공백문자를 AND 로 처리하는 기존의 인명 검색 방법이고, 검색방법 2는 [표 4]에서 설명한 구분자 내에서만 구검색이 수행되는 방법이며, 검색방법 3은 [그림 1]에서 설명하는 제안된 알고리즘에 기반한 방법이다.

- 검색대상 : NDSL 서비스대상 논문 52,110,679 건
- 인명(영어표기) 질의어 테스트셋 10,000 건
- 검색방법 1 : 공백문자를 AND 로 처리하여 검색
- 검색방법 2 : 구분자내에서 구검색
- 검색방법 3 : 구분자내에서 구검색과 인접검색 조합

표 5. 검색결과 개선 실험데이터의 예

질의어	결과건수			(A)-(C)	(C)-(B)
	검색 방법1 (A)	검색 방법2 (B)	검색 방법3 (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**

표 5는 검색 방법에 따른 논문 검색결과 건수를 보여준다. 질의어가 'Lee, Won-Jae' 일 경우에 검색 방법 1의 알고리즘에 따르면, 모든 논문의 저자필드에서 'lee', 'won', 'jae' 가 순서와 위치에 상관없이 모두 존재하는 논문 결과건수가 4,863건이다. 검색 방법 2의 결과에 의하면 4,863건 논문 가운데 저자필드에 'lee won jae' 가 세미콜론 영역 안에 순서대로 존재하는 경우가 321건에 해당한다는 것을 알 수 있다. 이는 사용자 질의어와 비교한다면 정확한 결과라고 할 수 있다. 하지만 제안한 알고리즘에 해당하는 검색 방법 3은 'Won-Jae Lee' 로 표기된 논문 결과건수 11건을 더 포함하여 제공한다. 이는 식별된 실세계의 동일 인물인 것은 보장할 수는 없지만, 문자열 인명 표기 방식을 고려하면 아주 유용한 검색 결과라고 할 수 있다. [표 5]의 (A)-(C)는 사용자 질의어와는 상관없는 검색결과이며, (C)-(B)는 인명표기에 사용한 문자는 같지만 성과 이름의 순서가 바뀐 동일 인명으로 표기된 저자명을 가진 논문 결과건수이다.

[표 6]은 인명 질의어 만 건을 이용한 논문 검색결과 건수의 평균을 보여준다. ((A)-(C))/(A) 에 해당하는 수치는 기존의 저자필드를 이용한 논문 검색이 약 86% 정도의 정확하지 않을 뿐만 아니라 사용자가 원하는 않는 검색결과를 제공한다는 것이다. 제시된 알고리즘은 인명을 통한 논문 검색시에 정확하지 않은 검색결과를 제외시키며, 동시에 ((C)-(B))/(C) 의 수치가 나타내는 것처럼 약 9%에 해당하는 동일인명의 다른 표기까지 검색결과에 포함시킨다.

표 6. 검색결과 개선 성능

결과건수 평균			((A)-(C))/(A)	((C)-(B))/(C)
검색방법1 (A)	검색방법2 (B)	검색방법3 (C)		
7,433	884	975	0.868	**0.093**

2. 색인의 효율화

NDSL 에서는 논문, 특허 등의 학술정보를 찾기 위한 부가기능으로 인명찾기의 기능을 제공하고 있는데, 이는 학술정보 메타정보로부터 다시 인명필드만을 추출하여 색인하므로, 색인측면에서 불필요한 부하가 될 수 있고, 엄청난 색인량 증가로 인해 검색 속도에 부정적인 영향을 미치게 된다. 본 장에서는 이미 논문 및 특허의 메타 정보에 포함되어 색인된 인명필드 색인정보를 이용해 인명찾기 기능을 제공하는 방법을 설명한다. 논문 및 특허 등 학술정보의 인명(저자, 출원인, 발명자 등) 필드에 수평적으로 나열된 데이터를 검색해서 사용자가 원하는 인명 검색결과를 수직적으로 나열해 주기 위해서 검색결과 그룹핑 개념인 FAST 엔진의 네비게이터(navigator) 기능을 사용한다. [그림 2]는 제안하는 인명찾기 구현방법이다.

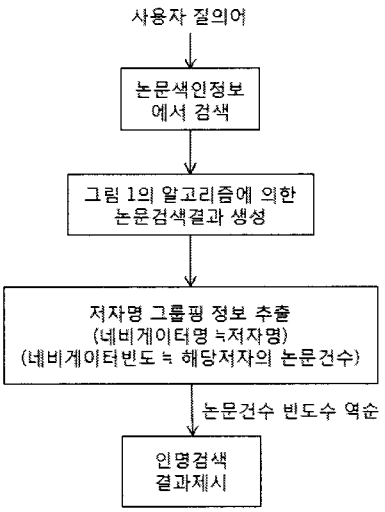

그림 2. 인명찾기 구현방법

별도의 저자리스트에서 검색하지 않고, 논문정보에 포함된 저자필드에서 인명 검색을 수행 후 논문 결과셋으로부터 다시 인명 그룹핑 정보를 생성하여 논문 저작수가 많은 순서로 나열하면 [표 7]에서 보여주는 인명검색 결과를 제시할 수 있다. 이는 인명찾기 기능 한가지만을 위해 약 5,600만 여건의 논문에서 추출한 2,200만 여건의 저자명을 다시 색인하는 과정을 거칠 필요가 없으므로 색인관점에서 약 28%의 색인량을 감소시키고 색인 바이너리를 저장하는 디스크 저장 효율을 약 10% 개선시킨다.

- (1) 논문과 저자정보 색인량의 합 : 약 7,800만건
- (2) 논문의 색인량 : 약 5,600만건
- 색인량 개선률 : ((1)-(2))/(1) = 0.282
- (3) 논문과 저자정보 색인 저장용량 : 840GB
- (4) 논문 색인 저장용량 : 750GB
- 디스크 용량 개선률 : ((3)-(4))/(3) = 0.107

표 7. 인명검색결과의 예

사용자질의어	Kim Chul Soo
논문검색결과 (5건)의 저자필드	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
그룹핑 정보를 통한 인명검색결과	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)

3. 공저자 및 관련연구자 제공

국내 논문의 저자 메타정보는 한글성명과 영문성명이 한 쌍으로 존재하는 경우가 대부분이므로 질의어가 영어일 때 해당 인명의 한글표기와 성과 이름의 표기 순서가 바뀐 로마자 표기는 제시 가능하지만, 사용하는 철자(spelling)가 다른 로마자 표기는 제시할 수 없다. 즉, 질의어가 'Kim Chul Soo' 일 경우 'Chul Su Kim' 은 검색 대상에서 제외된다. 하지만, 제안한 방법을 이용하면 사용자 질의어가 영문인명 일 때 항상 해당인명의 한글표기를 검색결과로 제시해 주며, 사용자가 제시된 한글인명을 클릭하거나 재검색 함으로써 영문인명의 다른 표기로의 접근을 얼마든지 가능하게 한다.

검색엔진이 제공하는 그룹핑 기법을 사용하면 논문 결과셋으로부터 저자명 자질을 추출하기 때문에 질의어와 관련된 한글성명, 영문성명과 영문성명의 다른 표기형식까지 제시 가능할 뿐만 아니라, 공저자명도 검색결과에 포함시킨다. [표 4]에서와 같이 사용자는 'Kim Chul Soo'를 질의어로 이용해서 'Chul-Soo Kim' 와 '김철수'를 제시받고 공저자인 'Ki-Won Lee', 'Kim, Sung-Hae', '김성해', 'Lee, Joon' 등의 리스트를 제공받는다. 제공된 인명들은 관련 분야의 연구자라고 예상이 가능하고, 이는 공저자명을 활용하여 관련 연구 분야의 논문을 탐색하는 효과적인 방법이다.

동일 인물에 대한 인명표기의 다양성을 고려하여 검색결과의 정확성을 향상하고 인명찾기 기능을 위한 불필요한 색인을 생성하지 않으며, 동시에 관련 연구자 리스트를 제공하는 기능을 포함한 학술정보서비스는 인명 검색을 통한 학술정보 요구에 유용하다. 이를 검증하기 위해 NDSL 논문 5,200만 여건을 대상으로 인명질의 논문검색시스템을 구현하였다.

본 시스템은 질의어가 인명일 때, [그림 3]과 같이 한글성명, 영문성명, 영문성명의 다른 표기까지 저자명 검색 결과로 제공하며, 동시에 학술논문 리스트를 제공한다. 사용자는 저자검색결과에서 인명을 클릭함으로써 선택한 저자의 논문을 자유롭게 탐색할 수 있다. 본 논문에서 제시한 방법은 첫째 인명 검색시 논문결과의 정확성을 보장하며, 둘째 동일 인명의 다른 표기를 제공하고, 셋째 저자들의 논문건수와 관련 연구분야 종사자라 판단할 수 있는 공저자 리스트를 효과적으로 제공한다.

그림 3. 검색결과 화면

그림 4. 인명질의 논문검색시스템

V. 결론 및 향후계획

모든 정보에는 인명으로 표현되는 창작자(학술정보의 경우 저자, 발명자, 연구자, 출원인, 분석자 등으로 이해된다.) 정보가 존재한다. 웹을 통한 검색서비스에서 인명 검색은 중요한 기능중의 하나이다. 인명 메타정보를 대상으로 한 대부분의 학술정보시스템은 제목, 초록 등 기타 정보와 동일한 색인 및 검색처리를 함으로 인해서, 정확한 검색결과를 제공하지 못하는 문제점이 존재한다.

본 논문에서는 학술정보서비스에서 인명 검색의 고도화를 위하여 다양한 인명표기 형식과 특징을 고려한 검색 기법과 검색 정확성 개선을 위한 방법에 대해 설명하였다. 또한 인명찾기를 위한 효율적인 색인 설계에 대해 논하였으며, 이는 NDSL 과학기술정보통합서비스의 경우 약 28%의 색인용량 절감과 10%의 디스크 용량 절감을 증명하였다. 또한 사용자에게 인명 검색시 공저자 및 관련연구자 리스트 등 유용한 정보를 효과적으로 제공함을 설명하였다.

향후 실세계에서 식별되는 인물과 관련된 학술정보를 제공하기 위해서는 인명, 이메일, 소속기관, 주소 등의 메타정보를 분석하여 다중 자질간의 조합 기법을 연구하고 저자식별의 개념을 적용해야 한다. 검색엔진을 활용한 메타검색의 정확성과 저자식별의 결과를 결합할 때 더 유용한 학술정보서비스의 구현이 가능하다.

참 고 문 헌

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

[2] J. Artiles, J. Gonzalo, and F. Verdejo, "A testbed for people searching stategies in the www," In Proc. of SIGIR'05, pp.569-570, 2005.

[3] P. Jakub, W. Karol, and S. Marcin, "On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages," Information retrieval, Vol.12, No.3, pp.275-299, 2009.

[4] C. Peter, "A Comparison of Personal Name Matching: Techniques and Practical Issues," Technical report, TR-CS-06-02, Computer Science Laboratory, The Australian National University, Canberra, Australia. 2006.

[5] 이준호, "로마자로 표기된 한글 인명의 검색 방법", 논문집: 이학편·공학편, 제31호, pp.181-189, 2002.

[6] 김혜숙, "한국인의 로마자 인명 표기의 통일성과 일관성: ≪영어영문학≫게재자를 중심으로", 영어학, 한국영어학회, 제1권, 제3호, pp.417-435, 2001.

[7] 송재용, 조영화, 류근호, "로마자표기 한글 인명을 위한 검색 모듈 설계와 인명 질의 확장기 구현", 제25권, 제1호, pp.196-198, 1998.

[8] U. Pfeifer, T. Poersch, and N. Fuhr, "Retrieval effectiveness of proper name search methods," Information Processing and Management, Vol.32, No.6, pp.667-679, 1996.

[9] W. Winkler, "Overview of record linkage and current research directions," Research Report Series #2006-2, Statistical Research Division, U.S. Census Bureau., 2006.

[10] A. Culotta, P. Kanani, R. Hall, M. Wick, and A. McCallum, "Author disambiguation using error-driven machine learning with a ranking loss function," IIWeb-2007, 2007.

[11] P. Kanani, A. McCallum, and C. Pal, "Improving author coreference by resource-bounded information gathering from the Web," IJCAI-2007, 2007.

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

[13] http://scholar.ndsl.kr/artsrch.do

저 자 소 개

한 희 준(Hee-Jun Han) 정회원

- 2002년 2월 : 전북대학교 정보통신공학과(공학사)
- 2004년 2월 : 한국과학기술원 멀티미디어공학(공학석사)
- 2004년 3월 ~ 현재 : 한국과학기술정보연구원 연구원

<관심분야> : 시멘틱웹, 대용량 검색, 내용기반 멀티미디어검색

예 용 희(Yong-Hee Yae) 정회원

- 1978년 2월 : 경북대학교 전자공학과(공학사)
- 1991년 2월 : 한양대학교 전자계산학과(공학석사)
- 1991년 1월 ~ 2000년 12월 : 산업기술정보원
- 2001년 1월 ~ 현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 정보검색, 정보유통, 디지털라이브러리, 시멘틱 기술

류 범 종(Beon-Jong You) 정회원

- 1984년 2월 : 서강대학교 전자공학과(공학사)
- 1987년 5월 ~ 1993년 3월 : 시스템공학연구소
- 1993년 4월 ~ 2000년 12월 : 연구개발정보센터
- 2000년 2월 : 충남대학교 문헌정보학(이학석사)
- 2005년 8월 : 충남대학교 문헌정보학(이학박사)
- 2001년 1월 ~ 현재 : 한국과학기술정보연구원 책임연구원
- 2007년 3월 ~ 현재 : 충남대학교 문헌정보학과 겸임교수

<관심분야> : 시멘틱웹, 지식베이스, 테크마이닝