JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 2, pp. 373-382, Apr. 2017

국내 과학기술콘텐츠 저자의 소속기관명 식별을 위한 소속기관명 자동 식별 알고리즘에 관한 연구

김 진 영[1] · 이 석 형[1*] · 서 동 준[1] · 김 광 영[1] · 윤 정 선[1]
[1]한국과학기술정보연구원

A Study on the Identification Algorithm for Organization's Name of Author of Korean Science & Technology Contents

Jinyoung Kim[1] · Seok-Hyong Lee[1*] · Dongjun Suh[1] · Kwang-Young Kim[1] · Jungsun Yoon[1]
[1]Korea Institute of Science and Technology Information

[요 약]

과학기술콘텐츠가 증가함에 따라 과학기술콘텐츠의 효율적인 검색을 지원하는 서비스가 요구되고 있다. 저자의 소속기관명을 키워드로 사용할 경우 한 기관에서 생산된 콘텐츠를 확인할 수 있을 뿐만 아니라 저자, 용어를 키워드로 사용한 검색 결과의 식별율을 향상 시킬 수 있다. 검색 키워드로 사용되는 데이터들의 중의성과 모호성으로 인해 검색 결과에 false negative, false positive가 포함될 수 있으므로 데이터의 식별을 통한 통제는 중요하다. 저자의 소속기관명의 식별을 통한 통제 역시 기관의 이명, 약어 검색을 지원가능하게 하므로 매우 중요하지만 기존의 데이터 식별을 통한 통제에 대한 연구는 저자, 용어에 대한 연구가 주를 이루었다. 본 연구에서는 기관명 식별 알고리즘을 제안하고, 한국과학기술정보연구원에서 보유하고 있는 국내 과학기술콘텐츠들에 대한 데이터를 이용한 실험 결과를 보인다.

[Abstract]

As the number of scientific and technical contents increases, services that support efficient search of scientific and technical contents are required. When an author's affiliation is used as a keyword, not only the contents produced by the affiliation can be searched, but also the identification rate of the search result using the author and the term as keyword can be improved. Because of the ambiguity and vagueness of the data used as a search keyword, the search result may include false negative or false positive. However, the previous research on the control through identification of the search keyword is mainly focused on the author data and terminology data. In this paper, we propose the algorithm to identify affiliations and experiment with show the experiment with scientific and technological contents held by the Korea Institute of Science and Technology Information.

색인어 : 과학기술콘텐츠, 전거데이터 구축, 소속기관명 식별, 학술정보 검색
Key word : Science and technology contents, Establishment of authority data, Identification of organization name, Academic information search

http://dx.doi.org/10.9728/dcs.2017.18.2.373

Received 22 April 2017; **Revised** 27 April 2017
Accepted 28 April 2017

***Corresponding Author; Seok-Hyong Lee**

Tel: +82-042-869-1779
E-mail: skyi@kisti.re.kr

Ⅰ. 서 론

과학기술분야 발전 속도가 증가함에 따라 과학기술콘텐츠(논문, 특허, 보고서)의 생산량이 크게 증가하고 있기 때문에 과학기술콘텐츠에 대한 체계적인 관리가 요구된다. 또한, 과학기술분야 연구와 개발을 위해 과학기술콘텐츠들이 참고자료로 많이 활용되고 있기 때문에 효율적인 연구·개발 지원을 위해 다양한 요소를 활용한 검색, 분석 서비스를 제공할 필요가 있다.

과학기술콘텐츠의 관리와 서비스를 위해서 원문과 함께 유지되고 있는 메타데이터가 사용된다. 메타데이터에 저장된 요소들로는 초록, 저자명, 소속기관명, 저자 키워드, 발행연도, 학술지명, 저자명, 출판사명 등이 있다. 이 요소들은 연구자들이 연구·개발을 위해 참고할 자료들을 검색하는데 유용하게 활용된다.

검색 요소들 중 용어, 저자, 소속기관명은 중요한 검색 요소로 사용되지만 여러 의미로 해석되는 중의성(ambiguity)과 여러 의미로 해석되지는 않지만 지시 범위가 불분명한 모호성(vagueness)이 존재한다. 과학기술콘텐츠의 체계적인 관리와 정보서비스를 위해 이 요소들에 대한 식별이 필요하다.

용어, 저자의 식별을 위한 다양한 연구들이 진행되고 있지만 기관 식별을 위한 연구는 활발하지 않다. 용어와 저자는 사용된 문자열이 여러 개체를 의미하기 때문에 용어의 경우는 주변 문자열과의 의미적 연관관계를 고려해 식별해야 하고, 저자의 경우는 주소, 전화번호, 이메일, 사이트, 소속기관 등 다양한 요소들을 활용하여 식별할 수 있다. 그러나 기관의 경우 동일한 개체에 대해 다양한 이명과 약어들, 오탈자가 존재할 수 있고 기관을 식별할 수 있는 추가적인 정보들(주소, 전화번호, 사이트 등)이 없기 때문에 식별하기 어렵다[7]. 예를 들어 한국과학기술원은 과기원, 한국과기원, 카이스트, KAIST 등 다양한 이름을 가지고 과가원, 카아스트, 한과학기술원 등의 오탈자가 발생한다.

과학기술콘텐츠 저자의 소속기관명에 대한 정확한 식별은 소속기관명이 제목, 학술지, 이메일 등에 비해 저자의 식별율과 정확율을 향상[2]시킬 수 있기 때문에 필요하다. 그리고, 소속기관명 식별은 기관명에 대한 이명, 약어를 통제함으로써 검색의 재현율을 향상시킬 수 있다. 예를 들어 한국과학기술원을 검색어로 사용해서 검색했을 때 한국과학기술원 뿐만 아니라 과기원, 한국과기원, 카이스트, KAIST 등의 소속기관 정보가 포함된 과학기술콘텐츠들까지 검색 결과에 포함되어 연구자들이 한국과학기술원의 이명을 사용해 재검색을 하지 않아도 된다. 또한 소속기관명을 다른 검색 요소들(연구 분야, 용어, 저자, 출판연도, 출판사 등)과 함께 사용함으로써 검색결과의 규모를 한정할 수 있기 때문에 연구·개발에 소모되는 노력을 최소화할 수 있다. 마지막으로 특정 기관의 과학기술콘텐츠들에 대한 검색이 가능하게 되어 해당 기관의 연구 분야, 시간에 따른 연구 동향 등을 파악하는데 도움이 된다.

본 연구에서는 국내 과학기술콘텐츠 저자의 소속기관명 식별을 위해 한국과학기술정보연구원에서 보유하고 있는 국내 과학기술콘텐츠(논문, 특허, 보고서)의 메타데이터에 포함된 소속기관 정보를 분석하여 식별 대상을 선정했고, 과학기술콘텐츠를 생산할 수 있는 기관들에 대한 정보를 수집하여 기관 정보 데이터베이스를 구축했다. 또한 본 연구에서는 구축된 기관 정보 데이터베이스를 사용하여 메타데이터에 포함된 저자의 소속기관명을 식별하는 알고리즘을 개발했다.

본 논문의 구성은 다음과 같다. 2장에서는 국내외 과학기술콘텐츠 정보서비스 사례들과 관련연구들을 소개하고, 3장에서는 소속기관명 식별 대상을 설명한다. 4장에서는 본 연구의 소속기관 식별 시스템에 대해서 설명하고, 구축한 기관 정보 데이터베이스에 대해서 설명한다. 그리고 5장에서는 구축한 기관 정보 데이터베이스에 포함된 기관명들을 사용하여 메타데이터에 포함된 소속기관명을 식별하는 알고리즘을 설명한다. 6장에서는 본 연구의 결론을 기술한다.

Ⅱ. 관련연구

소속기관명 식별결과를 활용한 정보서비스로 해외에서는 Thomson Reuters의 Web of Science와 Elsevier의 Science Direct 등이 저자, 기관, 용어에 대한 식별데이터를 구축하여 서비스하고 있다. 국내에서는 한국과학기술정보연구원의 NDSL(National Digital Science Library)에서 저자와 기관에 대한 식별데이터를 자체 구축하여 서비스하고 있다. 2016년 12월에 새롭게 서비스되고 있는 NDSL에서 본 연구의 기관 식별 결과물을 활용하고 있다. 현재 소속기관명 식별에 사용되는 기관 정보 데이터를 정제, 구축하고 있고 본 연구보다 고도화된 소속기관명 식별 시스템을 개발하고 있다. NDSL에서 서비스되고 있는 기관 식별 결과는 지속적으로 업데이트될 예정이다. 국내의 다른 서비스로는 DBpia에서 저자에 대한 식별데이터를 구축하여 서비스하고 있다. 국가기록원에서는 정부기관 등에서 생산된 기록물들의 식별을 위해 계층구조, 연혁관계 등이 고려된 약 130,000건의 기관 전거 데이터를 자체 구축하여 서비스하고 있다. 네이버에서는 기관을 교육 분야, 정치행정 분야, 경제 분야 등으로 세분화하여 각 기관에 대한 정보를 구축하여 서비스(기관단체사전)하고 있다.

소속기관명 식별을 위한 학술연구는 [1-8]이 있다. [1]에서는 국가 R&D 사업과 관련된 기관명(기업체, 연구원 등)들을 수집하여 패턴을 분석한 후 기관 대표명의 설정 방안과 대표명의 이형관계 설정 기준을 제시했다. 이를 바탕으로 기관전거데이터 구축 방안을 제시했다. 강인수

외[2]에서는 과학기술콘텐츠 메타데이터에 포함되어 있는 항목들 중 저자 식별을 위해 활용될 수 있는 요소들(소속기관명, 제목, 이메일 등) 중 소속기관명이 저자 식별율을 가장 향상시킬 수 있다고 언급했다. 이석형 외 [3]에서는 [2]에서와 같이 저자 식별을 위해 소속기관명 식별이 필요하고 이전기관명, 이후기관명 등이 고려된 기관 전거데이터를 구축이 필요하다고 했다. 이석형[5]에서는 저자의 소속기관의 상태변화(변경, 병합, 분리 등)가 빈번하게 발생하고 특정 기관의 명칭이 다양하거나 저자가 표기한 소속기관명에 오류가 있기 때문에 기관 중심으로 연구 결과물을 관리하기 어렵다고 했다. 그렇기 때문에 국내 학술논문의 소속기관들에 대한 다양한 분석을 통해 기관 식별 데이터를 구축할 수 있는 방안을 제시했다. Emiel Caron 외[7]에서는 과학기술콘텐츠를 생산한 저자의 소속기관명 식별을 위한 기관 정보 데이터 구축이 매우 어렵기 때문에 소속기관들 간 군집(clustering)을 위한 항목(국가명, 도시명, 우편번호, 기관타입, 문자열 유사도 등)을 정하고 많이 일치할수록 높은 기관 간 유사도가 높다고 판정하여 같은 군집으로 분류했다. Emiel Caron 외[7]는 기관 정보 데이터를 구축하지 않고 과학기술콘텐츠 저자의 소속기관명 목록만을 가지고 식별을 할 수 있지만, 식별된 결과가 정확하지 않기 때문에 결국 사람이 검증해야 하는 문제가 있다. 김진영 외[8]에서는 국내 과학기술콘텐츠 저자의 소속기관명들을 분석하고 식별을 위한 시스템을 제안했으며 시스템의 각 분야 별 기능과 상세를 소개했다. 본 연구는 [8]의 후속 연구로서 소속기관명 식별을 위해 필요한 기관 사전 데이터베이스 구축 방법과 실제 예를 제시하고 소속기관명 식별 알고리즘을 제시했으며 실행 예와 결과를 제시했다.

Ⅲ. 소속기관명 식별 필요성 및 대상

본 장에서는 과학기술콘텐츠에 대한 소속기관명 식별의 필요성에 대해 설명하고, 과학기술콘텐츠 메타데이터에 포함된 실제 소속기관명들의 분석을 통해 정의한 식별 대상들을 설명한다.

3-1 과학기술콘텐츠 소속기관명 식별 필요성

과학기술분야 콘텐츠들의 생산이 폭발적으로 증가하고 있고 연구·개발을 위해 논문, 특허, 보고서와 같은 과학기술콘텐츠들이 참고자료로 많이 활용되고 있기 때문에 과학기술콘텐츠들에 대한 체계적인 관리가 필요하고 효율적인 검색과 다양한 분석을 위한 서비스들을 개발/제공할 필요가 있다.

과학기술콘텐츠들에 대한 관리와 다양한 서비스들을 위해 과학기술콘텐츠의 메타데이터들을 활용한다. 과학기술콘텐츠 메타데이터에는 문서 제어번호, 저자명(국문, 영문), 소속기관명(국문, 영문), 제목(국문, 영문), 초록, 저자 키워드(국문, 영문, 주소, 이메일, DOI 등의 정보가 포함되어 있다.

과학기술분야 각종 정보서비스들은 주로 연구 분야, 키워드 등에 대해 검색어를 입력하여 매치되는 결과를 제공했다. 그러나 연구자들은 과거 연구 내용을 확인하거나 연구 이력관리를 위해 본인의 연구 결과물들만을 모아서 보고 싶어 한다. 그리고 본인의 연구 방향을 설정하거나 협력 연구자를 찾기 위해 특정 과학기술분야의 저명한 연구자들의 연구 결과물들만을 모아서 보고 싶어 한다. 또한 연구자 또는 정책입안자의 경우 특정 기관의 연구 동향 파악을 위해 정보서비스를 사용할 수도 있다. 이때 기관은 설립, 폐지, 변경, 병합, 분리될 수 있기 때문에 특정 기관의 연구 동향은 기관의 연혁과 계층 관계에 따라 모두 보여야 한다. 이렇게 다양한 요구사항에 따른 효율적인 연구·개발을 지원하기 위해 과학기술콘텐츠를 생산한 저자들과 저자들의 소속기관명에 대한 정확한 식별이 필요하다.

용어, 저자, 소속기관명은 중요한 식별 요소이다. 각 요소들은 하나의 언어 표현이 둘 이상의 해석을 가능하게 하는 중의성(ambiguity)과 언어 표현이 어떤 대상을 지시하는지의 범위가 명확하지 않은 모호성(vagueness)을 가진다. 용어의 경우 '배'는 배나무의 열매, 물 위로 떠다니는 나무나 쇠로 만든 물건, 어떤 수나 양을 두 번 합한 만큼 등으로 해석될 수 있다. 용어의 경우는 주변 문자열과의 의미적 연관관계와 통계 정보를 고려해 식별해야 한다. 저자의 경우 동명이인(同名異人)이 존재한다. 예를 들어 본 논문의 저자인 '김진영'은 본 연구에서 사용한 국내 학술지 약 1,300,000건 중 1,122건의 학술지의 저자명 정보에 포함되어 있었다. 많은 수의 사람들이 사용하는 저자명 식별은 반드시 필요하며 저자 식별을 위해서는 과학기술콘텐츠 메타데이터에 포함된 저자의 소속기관명, 이메일, 주소, 전화번호, 저자 키워드 등이 사용된다.

기관은 저자와 같이 서로 다른 기관들이 동일한 이름을 가질 수 있다. 예를 들어 본 연구에서 구축한 기관 정보 데이터베이스에 '밝은안과의원'은 71건이 있으며 실제로는 이보다 많을 수 있다. 특정 논문 저자의 소속기관명이 '밝은안과의원'이라고 했을 때 어떤 '밝은안과의원'인지 식별이 필요하다. 또한 기관은 하나의 기관에 대해 여러 가지 이름을 가진다. '한국전기연구원'은 '전기연구원', '전기연', '전기硏', '케리', 'KERI', '전기연구소'와 같은 다양한 이름을 가진다. 그리고 기관 영문명의 경우 축약어를 포함할 수 있는데 대학교(University)는 Univ., 학과(Department)는 Dept., 과학(Science)는 Sci. 등으로 표기되어 있다.

과학기술콘텐츠 소속기관명 식별은 다양한 이름을 가지고 있는 기관명들을 대표 기관명으로 통제하고 같은 이름의 여러 기관들 중 하나를 특정할 수 있다.

소속기관명 식별을 통해 특정 기관의 연구 결과물들만을 열람할 수 있다. 이는 연구자, 정책입안자 등이 기관의 연구 동향을 파악하거나 향후 연구 방향을 설정하는데 유용하게

표 1. 과학기술콘텐츠 저자의 소속기관명 표기의 예

Table. 1. Organization's Name of Author of Korean Science & Technology Contents

국문명	영문명
동신대학교 한의과대학 침구학교실	Dept. of Acupuncture & Moxibustion Oriental Medical College, Dongshin University
고려대학교 전자및정보공학부	Department of Electronics and Info. Engr., Korea University
단국대학교 전자공학과	Dan Kook University, Dept. of Electronics
국립의료원 침구과	Department of Acupuncture &Moxibustion, National Medical Center

활용된다. 기관은 시간의 흐름에 따라 설립, 폐지, 변경, 합병, 분리되는 연혁관계를 가지고 계층구조를 가지기 때문에 기관 간 연혁과 계층구조까지 고려하여 식별하게 되면 더욱 활용도가 향상될 것이다.

소속기관명은 저자 식별 시 식별율과 정확율 향상을 위해 가장 중요한 요소이다[2]. 동일 기관에 대한 다른 이름의 기관명에 대해 대표명으로 식별하여 식별아이디를 부여하여 이용하게 되면 더욱 효율적이고 효과적으로 식별할 수 있다.

과학기술분야 연구자들이 연구·개발 시 필요한 참고자료들을 열람하기 위해 과학기술콘텐츠 정보서비스를 사용하여 검색할 때 연구분야, 용어, 저자 등과 함께 기관명을 함께 사용함으로써 검색 결과물의 범위가 더욱 한정되어 연구·개발의 효율성이 향상된다.

3-2 소속기관명 식별 대상

본 연구에서는 한국과학기술정보연구원이 입수하여 관리하고 있는 국내 과학기술콘텐츠(논문, 특허, 보고서) 약 5,000,000건(논문: 약 1,300,000건, 특허: 약3,300,000건, 보고서: 213,000건)을 대상으로 했다.

논문, 특허, 보고서에 대한 저자 단위 소속기관명의 수는 약 7,500,000건(논문: 약 3,700,000건, 특허: 약 3,600,000건, 보고서: 약 213,000건)이다. 이 통계는 소속기관명 중복을 허용한 수치이다.

소속기관명 데이터를 분석한 결과 국문명은 일반적으로 문자열의 시작에서 끝으로 상위기관부터 하위기관으로 나열되는 기관의 계층 구조를 잘 표현하고 있다. <표 1>에서 '동신대학교 한의과대학 침구학교실'은 '대학교-단과대-교실' 순으로 표기되어 있고 '고려대학교 전자및정보공학부'는 '대학교-학부' 순으로 표기되어있으며 '단국대학교 전자공학과'는 '대학교-학과' 순으로 표기되어 있다.

그러나 소속기관의 영문명의 경우 국문명과 같이 문자열들이 계층적으로 나열되어 있지 않은 경우가 있었다. 영문 소속기관명들은 일반적으로 하위기관부터 상위기관 순으로 나열된 경우가 일반적(학과-단과대-대학교, 학부-대학교 등)이지만 반대로 국문표기와 비슷하게 (대학교-단과대-학과, 대학교-학부 등)으로 나열된 경우도 존재했다. <표 1>에서 '단국대학교 전자공학과'의 영문명인 'Dan Kook University, Dept. of Electronics'는 '대학교-학과' 순으로 문자열이 나열되어 있음을 볼 수 있다.

[5]에서는 국내 학술지를 생산한 저자의 소속기관명에 대한 분류 별 출현 빈도를 조사했는데 대학교가 가장 많이 나타났고 일반 기업체 및 부설연구소, 정부출연연구기관, 정부부처 순으로 출현한 것으로 조사되었다. 소속기관명 식별을 위해 사용될 기관 정보 데이터베이스 구축과 관리, 추후 분야 별 연구 결과 분석 서비스 등을 위해 식별 대상 기관 분류를 수행했다.

위와 같은 출현 빈도뿐만 아니라 기관의 특성이 과학기술콘텐츠의 연구 분야 또는 성격을 결정할 수 있음에 착한하여 교육기관, 의료기관, 기타기관으로 분류했다.

교육기관으로는 초·중·고등학교와 대학교로 분리하여 구축, 관리하고, 의료기관으로는 종합병원(상급병원 포함), 병원, 의원, 의료원, 보건소 등을 동일한 계층으로 구축, 관리했다. 기타기관은 교육기관과 의료기관을 제외한 기관을 의미하며 정부기관과 기업체로 분리하여 구축 관리하고 있다.

기관은 <대학교-단과대-학과>, <의료기관-진료과목>, <기업체-부-과-팀> 등 다양한 계층관계를 가진다. 기관은 시간의 흐름에 따라 빈번하게 설립, 폐지, 변경, 합병, 분리되기 때문에 최상위기관인 교육기관, 의료기관, 기타기관이 가지는 모든 하위기관에 대해 기관 정보 데이터베이스를 구축할 수는 없다.

그러므로 본 연구에서는 최상위기관과 함께 식별할 하위 기관들의 한계 계층을 다음과 같이 정의했다.

* 교육기관: 대학교, 초·중·고등학교(하위 없음)
  - 하위1: 캠퍼스
  - 하위2: 대학원
  - 하위3: 단과대 또는 학부
  - 하위4: 학과, 전공, 계열, 기타
* 의료기관: 종합병원(상급병원 포함), 병원, 의원, 의료원 등
  - 하위1: 진료과목
* 기타기관: 정부기관, 기업체
  - 하위 없음

소속기관명 데이터를 분석한 결과 초·중·고등학교는 하위기관이 거의 표기되지 않았다. 그러므로 초·중·고등학교는 하위기관에 대한 데이터를 구축/관리/식별하지 않는다. 대학교의

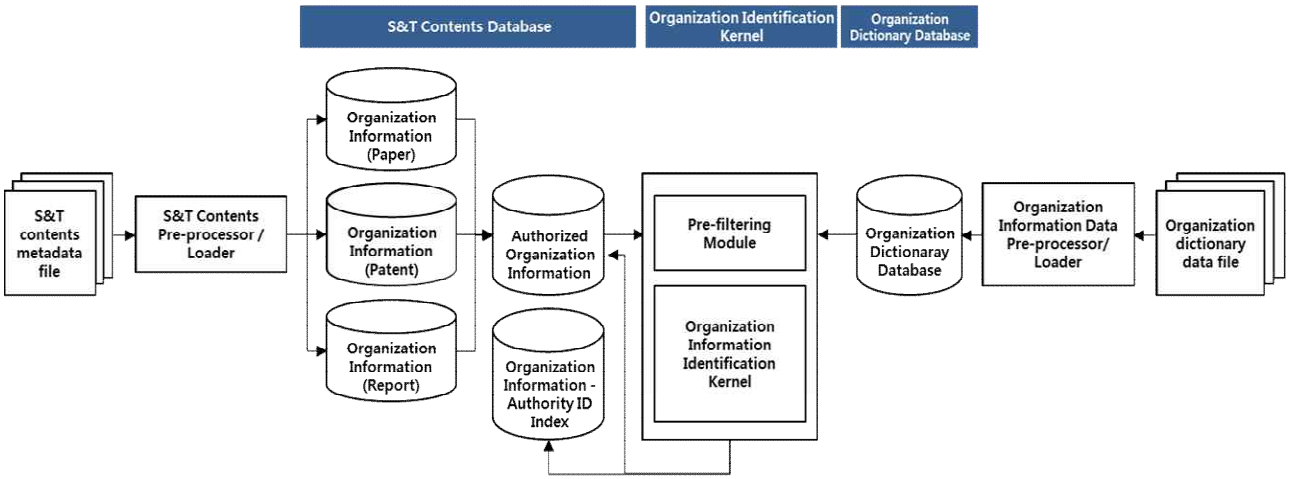

그림 1. 소속기관명 식별 시스템
Fig. 1. Organization Name Identification System

경우 하위 기관으로 캠퍼스, 대학원, 단과대, 학부, 학과, 전공, 계열, 기타(연구소, 연구실, 센터 등)이 표기되었다. 대학교의 최 하위기관인 단과대, 학부, 전공 등은 대학교와 교육부의 정책, 사회의 요구에 따라 자주 바뀌는 경향이 있지만 과학기술콘텐츠를 주로 생산하는 기관이기 때문에 식별 대상으로 포함시켰다.

의료기관의 경우 의학 분야의 과학기술콘텐츠들을 생산하지만 각 진료과목 별로 분야가 현저하게 차별되기 때문에 식별 대상으로 포함시켰다. 현재 약 170개의 진료과목이 기관 정보 데이터베이스 구축되어 있고 각 의료기관 별 진료과목들이 하위기관으로 구축되어 있다.

기타 기관의 경우 정부기관, 기업체를 포함한다. 정부기관은 [5]에 따르면 국내 학술지 생산 빈도가 저조하고 계층 구조가 비교적 깊고 복잡하기 때문에 4.2절에서 설명한대로 국립중앙도서관에서 제공하는 한국정부기관 부호표에 포함된 기관명을 최상위 기관으로 지정하고 각 기관들의 계층관계를 구축하고, 연혁관계를 고려하여 구축된 이전 기관들과 이후 기관들까지 식별 대상으로 포함시켰다. 포함된 각 기관들의 하위기관(국, 과, 계, 팀 등)들은 식별 대상에서 제외했다.

기타기관의 기업체의 경우 과학기술콘텐츠를 많이 생산하지만 폐지, 병합, 분리, 변경이 다른 기관 분류에 비해 더욱 빈번히 발생한다. 또한 기업체의 하위기관의 경우도 상태 변화가 자주 발생하기 때문에 기업체의 최상위 기관만을 식별 대상에 포함시켰고, 하위기관(센터, 부, 실, 팀, 연구소 등)은 식별 대상에서 제외했다.

Ⅳ. 소속기관 식별 시스템/기관 정보 데이터베이스

본 장에서는 본 연구의 선행 연구인 [8]에서 제안한 소속기관 식별 시스템(Organization Name Identification System)을 소개하고 3장에서 제안한 소속기관명 식별 대상에 따라 구축한 기관 정보 데이터베이스를 설명한다.

4-1 소속기관 식별 시스템

과학기술콘텐츠를 생산한 저자의 소속기관명 식별 시스템 <그림 1>은 과학기술콘텐츠 데이터베이스(S&T Contents Database), 소속기관명 식별 커널(Organization Identification Kernel), 기관 사전 데이터베이스(Organization Dictionary Database)로 구성된다.

과학기술콘텐츠 데이터베이스에는 한국과학기술정보연구원이 수집한 다양한 형태(XML, CSV 등)의 국내 과학기술콘텐츠(논문, 특허, 보고서)들에 대한 메타데이터들이 데이터베이스 스키마에 맞게 가공되어 저장되어 있다. 논문, 특허, 보고서 각각에 대해 식별 대상(인물, 기관, 용어) 별로 식별을 요구하는 과학기술콘텐츠 메타데이터들을 가공하여 저장하고 있다.

본 연구의 소속기관 식별 시스템에서는 과학기술콘텐츠의 저자들이 다수가 될 수 있기 때문에 콘텐츠 당 저자/소속기관으로 분리하여 소속기관 중심으로 재구성하여 사용한다. 이 데이터베이스에는 소속기관명(국문, 영문), 저자명(국문, 영문), 제어번호, 학술지(국문, 영문), 제목(국문, 영문), 출판연도 등에 대한 정보가 포함되어 있다.

기관 사전 데이터베이스는 축약어 데이터베이스와 기관 정보 데이터베이스로 구성된다. 축약어 데이터베이스는 소속기관명에 포함된 약어(예: Univ.=University, Dept.=Department)를 처리하기 위한 약어 사전 데이터가 저장되어 있다. 기관 정보 데이터베이스는 기관의 이명, 약어 목록과 기본 정보(주소, 전화번호, 사이트, 설립일, 폐지일 등)과 기관 간 각종 관계 정보(상위 관계, 하위 관계, 이전 관계, 이후 관계 등)가 저장되어 있다. 기관 사전 데이터베이스는 4-2절에서 다시 설명한다.

표 2. 기관 정보 데이터베이스에 포함된 레코드들의 예

Table. 2. Records in the organization information database

기관 분류	국문명	영문명	한문명	국문이명	영문이명	한문이명	국문주소	영문주소	전화번호	사이트
대학교	고려대학교	Korea University	高麗大學校	[고려대학, 고려대]	[KU]	[高麗大學,高麗大]	서울특별시 성북구 안암로 145	145 Anam-ro, Seongbuk-gu, Seoul	02-3290-1114	www.korea.ac.kr
학과	도시토목환경학과	Dept. of Urban, Civil Engineering Environment			[department of urban, civil engineering environment]					
병원	원광대학교 의과대학산본병원	Wonkwang University Sanbon Hospital		[원광대학교 의과대학 산본병원]			경기도 군포시 산본로 321, (산본동)		031-390-2300	
회사	한국가스기술공사	Korea Gas Technology Corporation		[가스기술공사]	[KOGAS-TECH]				02-2657-1301	

표 3. 축약형 소속기관명의 예

Table. 3. Abbreviations of Organization Name

소속기관 국문명	소속기관 영문명
동아대학교 기계공학과	Dept. of Mech. Eng. Dong-A Univ.
연세대학교 의과대학 의용공학교실	Dept. of Medical Eng., College of Medicine, Yonsei Univ.
(주)바이오시스	Biosys Co. Ltd.

소속기관명 식별 커널에서는 과학기술콘텐츠 데이터베이스에 저장된 소속기관 중심으로 재구성된 기관 데이터베이스와 축약어와 과학기술콘텐츠를 생산하는 각종 기관 정보가 저장된 기관 사전 데이터베이스를 사용하여 실제 식별이 수행된다. 식별을 위한 알고리즘은 5장에서 설명한다.

4-2 기관 사전 데이터베이스

소속기관 식별 시스템의 구성요소 중 하나인 기관 사전 데이터베이스는 축약어 사전 데이터베이스와 기관 정보 데이터베이스로 구성된다.

축약어 사전 데이터베이스는 주로 기관의 영문명에 포함된 축약어들을 원형으로 복원하는데 필요한 <축약어-원형>들에 대한 정보가 저장되어 있다.

메타데이터에 포함된 소속기관명들을 분석한 결과 국문명의 경우 대학교를 대학 또는 대(예: 서울대학교, 서울대학, 서울대)로 표기한 경우가 대부분이었다. 영문명의 경우 <표 3>과 같이 Univ., Dept., Mech., Eng., Co., Ltd. 등의 축약어들이 출현한다. 축약어 사전 데이터베이스에는 <Univ. - University>, <Dept. - Department>, <Mech. - [Mechanical, Mechanics, Mechanism]>, <Eng. - Engineering>, <Co. - Company>, <Ltd. - Limited> 등에 대한 정보가 저장되어 있다. 5장에서 설명할 식별 알고리즘에서는 축약어가 포함된 소속기관 영문명을 이 데이터베이스를 사용하여 원형명으로 변환한다. 예를 들어 <표 3>의 '동아대학교 기계공학과'의 영문명은 ['Department of Mechanics Engineering, Dong-A University', 'Department of Mechanical Engineering, Dong-A University', 'Department of Mechnism Engineering, Dong-A University']로 변환된다. '연세대학교 의과대학 의용공학교실'의 영문명은 'Department of Medical Engineering, College of Medicine, Yonsei University로 변환되고, '㈜바이오시스'는 'Biosys Company Limited'로 변환된다.

기관 정보 데이터베이스에는 과학기술콘텐츠를 생산하는 기관들에 대한 각종 정보가 저장되어 있다. 여기에는 <표 2>와 같이 3장에서 설명한 식별 대상 분류에 따른 코드, 언어별 대표 명칭(국문명, 영문명, 한문명), 언어별 이명과 약어 명칭 목록과 각종 기본 정보인 국문 주소, 영문 주소, 전화번호, 사이트 등이 포함된다. 추가적으로 각 기관의 제어번호(식별아이디를 의미함)와 기관 간 관계 정보인 상위관계, 하위관계, 이전관계, 이후관계 등이 포함된다.

현재 초·중·고등학교에 대한 약 12,000건, 폐지된 대학교를 포함하여 대학교에 대해 약 1,700여 건, 각종 의료기관 약 64,000건, 정부기관 2,500건 등을 구축했고 분류 별 하위 기관과 연혁 관계를 고려하여 약 500,000건의 기관 정보들이 구축되어 있다. 기관은 신규 설립, 폐지, 서로 다른 기관들 간 병합, 한 기관이 여러 기관들로 분리, 하나의 기관이 다른 기관들로 변경되는 등 다양한 변화가 발생하기 때문에 추후 지속적으로 업데이트를 할 예정이다. 이때 기존 입수되거나 신규 입수된 과학기술콘텐츠의 소속기관들 중 구축된 기관 정보 데이터베이스에 없는 기관들 위주로 업데이트할 것이다.

기관 정보 데이터베이스 구축을 위해 식별 대상 분류 별로 다양한 경로를 통해 기반 데이터를 입수했다.

교육기관(초중등학교, 대학교) 기반 데이터 구축을 위해 교육학술정보원에서 총괄관리하고 있는 '학교알리미'와 한국대학교육협의회에서 운영하고 있는 '대학알리미'에 공식 요청하여 입수한 데이터를 사용했다.

의료기관 기반 데이터 구축을 위해 공공데이터포털에서 구할 수 있는 국립중앙의료원의 OpenAPI 등을 활용하여 수

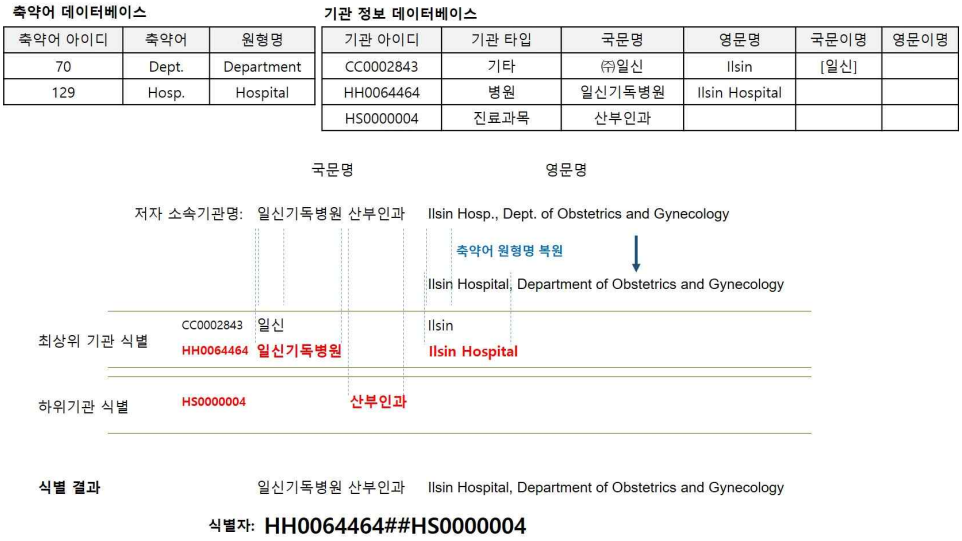

그림 2. 소속기관명 식별 알고리즘의 실행 예
Fig. 2. The execution example of identification algorithm for organization's name of author

집했다. 정부기관은 국립중앙도서관에서 제공하는 한국정부기관 부호표와 국가기록원에서 운영하고 기록물 생산기관 변천정보를 활용했다. 기업에 대해서는 국내 상장기업 목록, 전문기업목록을 입수하여 활용했다.

이와 같은 기반 데이터에 대하여 명칭, 기본 정보 등에 대한 오류 여부를 검증하여 정제하여 한국과학기술정보연구원에서 기 구축한 기관 전거 데이터와 병합했다. 기반 데이터에는 기관 간 계층 관계와 연혁관계가 없기 때문에 각 기관의 공식 홈페이지와 국가기록원과 같은 공신력있는 홈페이지들에 대한 검색을 통하여 구축했다.

기관 정보 데이터베이스 구축은 많은 시간과 인적, 물적 비용이 소요되기 때문에 저자와 용어 식별 결과, 메타데이터에 포함된 다른 요소들과 연계하여 구축한 후 검증하여 기관 정보 데이터베이스에 적재하는 반자동 구축 시스템이 필요하다.

Ⅴ. 소속기관명 식별 알고리즘

본 장에서는 소속기관명 식별 시스템에서 사용하는 소속기관명 식별 알고리즘을 설명한다. 이 알고리즘은 먼저 4-1절에서 설명한 과학기술콘텐츠 데이터베이스에 저장된 소속기관명 목록에 대해 축약어 데이터베이스를 참조하여 소속기관명에 포함된 축약형을 원형명으로 복원한다. 복원 전, 후 소속기관명에 대해 기관 정보 데이터베이스에 구축되어 있는 각 언어 별 대표명과 이명 목록을 문자열 포함관계를 고려하여 비교한 후 가장 넓은 범위를 포함하고 있는 기관의 아이디들의 목록을 소속기관명에 대한 식별자로 출력한다.

5-1 축약형 소속기관명 원형 변경

식별 대상 소속기관명에 대하여 영문명에 축약어가 포함되어 있는 경우 축약어 데이터베이스에 구축되어 있는 <축약어-원형> 사전을 참조하여 원형명으로 변경한다. <그림 2>에서 식별이 요구되는 소속기관 영문명(Ilsin Hosp., Dept. of Obstetrics and Gynecology)에 축약어인 Hosp.와 Dept.가 포함되어 있기 때문에 축약어 데이터베이스를 참조하여 Ilsin Hospital, Department of Obstetrics and Gynecology로 변환한다.

5-2 문자열 포함관계를 고려한 소속기관명 식별 알고리즘

축약형 소속기관명을 원형 명으로 변경하고 변경 전 소속기관명과 변경 후 소속기관명에 대해 기관 정보 데이터베이스에 저장된 대표기관명(국문, 영문), 이명목록(국문, 영문)들을 비교한다.

이 때, 2.2절에서 정의한 기관 분류 와 계층 별 식별 대상에 따라 각 분류 별 최상위 기관에 해당하는 분류 코드(기관 타입)들과 일치하는 기관 목록을 구성한 다음 가장 넓은 범위의 문자열을 포함하는 기관을 최상위기관으로 선택한다. 선택된 기관명 문자열을 제외한 나머지 문자열에 대해 선택된 기관명 문자열 계층의 하위에 속하는 기관 분류코드에 해당하는 기관명들과의 비교를 통해 다시 가장 넓은 범위의 문자열을 포함하는 기관을 선택한다. 2.2절에서 정의한 기관 분류 별로 더 이상 찾아지는 기관명이 존재하지 않을 경우 다음 소속기관명을 처리한다.

<그림 2>의 경우 저자 소속기관 국문명인 '일신기독병원'에 기타 분류(정부기관 또는 기업체)인 '(주)일신'의 이명인 '일신'

과 일치하는 부분문자열이 존재하고, 의료기관 분류(종합병원, 병원 등)인 '일신기독병원'과 일치하는 부분문자열도 존재한다. 이 때 '일신기독병원'이 '(주)일신'의 이명인 '일신'에 비해 더 넓은 범위에 해당하기 때문에 HH0064464, '일신기독병원'이 최상위기관으로 선택된다. 선택된 HH0064464가 의료기관에 해당하기 때문에 기관 정보 데이터베이스에 저장된 진료과목명들에 대해 최상위기관의 문자열 매칭 방법을 사용하여 HS0000004, '산부인과'가 선택되었다. 영문명에서도 국문명과 동일한 방법으로 문자열 포함관계를 고려하여 기관명들을 식별한 후 국문명의 식별 결과와 병합한다. <그림 2>의 예에서는 결국 'HH0064464##HS0000004'가 식별자로 출력된다.

2.2절에서 언급한 논문, 특허, 보고서에 대한 저자 단위 소속기관명 약 7,500,000건 중 약 6,400,000건(85.3%)이 최소 1개 이상의 최상위기관이 식별되었다.

Ⅵ. 결 론

과학기술분야 콘텐츠들의 생산이 폭발적으로 증가하고 있고 연구·개발을 위해 논문, 특허, 보고서와 같은 과학기술콘텐츠들이 참고자료로 많이 활용되고 있기 때문에 과학기술콘텐츠들에 대한 체계적인 관리가 필요하고 효율적인 검색, 다양한 분석 서비스를 개발/제공할 필요가 있다.

과학기술콘텐츠의 관리와 수요자들의 효율적이고 효과적인 연구·개발을 위하여 기관, 용어, 저자의 중의성과 모호성을 해소한 식별에 대한 연구가 진행되고 있다. 용어와 저자의 식별에 비해 소속기관에 대한 식별 연구가 부족하다.

본 연구는 국내 과학기술콘텐츠들을 생산한 저자들의 소속기관명을 식별하는 시스템을 제안한 [8]의 후속 연구로서 실제 구축한 데이터베이스를 소개하고 시스템에서 사용한 소속기관명 식별 알고리즘을 제안했다.

소속기관명 식별은 기관명이 가질 수 있는 중의성과 모호성을 해소하여 대표 기관명으로 식별해주는 것을 의미한다. 이러한 소속기관명 식별은 저자 식별 시 식별율과 정확율을 향상시키고, 특정 기관에 대한 연구 결과물들 검색 시 재현율이 높으며 한 기관의 과거 연구 결과물들을 한데 모아서 볼 수 있다. 또한 과학기술콘텐츠 수요자들이 연구분야, 용어, 저자명 등으로 검색할 때 검색 결과를 한정시킬 수 있어 연구·개발의 효율성을 향상시킨다.

Ⅶ. 향후연구

본 연구에서 제안한 소속기관명 식별 알고리즘은 기관명 사전 데이터베이스에 저장되어있는 기관명과 기관 분류 코드만을 고려하여 기관명과 기관의 일부 계층들을 식별할 수 있다. 그러므로 최상위 기관과 최하위 기관이 동일하게 식별되었더라도 중간 기관들이 다르다면 동일한 기관을 의미하는지 알 수 없는 모호함이 존재한다. 중간계층이 다르지만 최하위 기관 이름이 동일한 경우가 존재하기 때문이다.

이를 보완하기 위해 현재 4-2에서 설명한 기관 정보 데이터베이스에 기관의 계층관계(상위기관, 하위기관), 연혁관계(이전관계, 이후관계) 등에 대한 정보를 구축하고 있고 현재까지 구축된 데이터는 새롭게 설계된 기관 정보 데이터베이스에 적제 되어 있다.

본 연구에서 제안한 소속기관명 식별 알고리즘에 추가적으로 계층관계와 연혁관계, 설립일, 폐지일, 과학기술콘텐츠의 출판일 등을 고려한 새로운 소속기관명 식별 시스템을 개발 중에 있다.

본 연구는 구축되어 있는 기관 정보 데이터베이스를 사용하여 식별하기 때문에 해당 데이터베이스 구축되어 있지 않은 기관명을 식별하기 어렵다. 이 경우는 저자의 소속기관명에 오자, 탈자가 존재하는 경우와 소속기관명이 잘 표기되어 있으나 기관 정보 데이터베이스에 구축되어 있지 않은 경우에 해당한다.

잘 표기된 소속기관명에 대한 정보가 데이터베이스에 구축되어있지 않은 경우 사람이 직접 식별을 하고 관련 정보를 데이터베이스에 추가해야하는 시간, 인적/물적 비용이 소요된다. 그러므로 식별되지 않은 소속기관명들에 대해 신규 구축 대상 여부를 알리거나, 자동 구축할 수 있는 연구가 필요하다.

오자, 탈자가 존재하는 소속기관명들에 대해서는 기관 정보 데이터베이스에 구축된 기관명들과의 문자열 유사도 비교를 통해 가장 높은 유사도를 가지는 기관으로 추천한 후 검증을 통해 전거하는 방법이 필요하다. 이 때 국문은 영문과 다르게 음소 단위의 문자열 유사도 방법이 필요하고 이에 대한 연구를 진행할 예정이다.

식별 연구는 모든 식별 대상을 정확하게 자동으로 식별할 수 없다는 한계를 가진다. 본 연구의 최종 목표는 시간, 인적/물적 비용을 최소화하면서 기관들의 계층관계, 연혁관계 등을 고려한 과학기술콘텐츠 저자의 소속기관 식별 시스템을 개발하는 것이 목표이다.

참고문헌

[1] Sung Ho Shin, "An Approach of Organization's Name Authority Control for Improving Data Searching Results," *Fall Conference*, *Korean Society for Internet Information*, pp. 403-407, November 2008.

[2] In-Su Kang, Seungwoo Lee, Hanmin Jung, Pyung Kim, Heekwan Koo, Mi-Kyung Lee, Won-Kyung Sung, and Dong-In Park, "Features for Author Disambiguation," *Journal of the Korea Contents Association*, Vol. 8, No. 2, pp. 41-47, 2008.

[3] Seok-Hyong Lee and Seung-Jin Kwak, "A Study on the Construction for Name Authority Data of the Korean

Academic Papers," *Journal of the Korean Biblia Society for Library and Information Science*, Vol. 21, No. 1, pp. 105-118, March 2010.

[4] Seok-Hyong Lee and Seung-Jin Kwak, "Development and Evaluation of Authority Data based Academic Paper Retrieval System," *Journal of the Society for Library and Information Science*, Vol. 46, No. 2, pp. 133-156, May 2012.

[5] Seok-Hyoung Lee, "A Study on the Construction of Identified Data of Author's Affiliation in Academic Papers," *Journal of the Institute for Social Sciences*, Vol. 25, No. 4, pp. 391-410, 2014.

[6] Anderson A. Ferreira, Marcos André Gonçalves and Alberto H. F. Laender, "A Brief Survey of Automatic Methods for Author Name Disambiguation," *ACM SIGMOD Record*, Vol. 41, No. 2, pp. 15-26, June, 2012.

[7] Emiel Caron and Hennie Daniels, "Identification of Organization Name Variants in Large Databases using Rule-based Scoring and Clustering With a Case Study on the Web of Science Database," In *Proc. of the 18th International Conference on Enterprise Information Systems(ICEIS 2016)*, Vol. 1, pp. 182-187, 2016.

[8] Jinyoung Kim, Seok-Hyong Lee, Dongjun Suh, and Kwang-Young Kim, "A Study on the Method and System for Organization's Name Authorization of Korean Science and Technology Contents," *Journal of Digital Contents Society*, Vol. 17, No. 6, pp. 555-563, Dec. 2016.

김진영(Jin-Young Kim)

2009년 : 서강대학교 대학원 (공학석사)
2009년~현재 : 한국과학기술원 전산학과 박사과정(수료)

2015년~현재: 한국과학기술정보연구원(KISTI) 정보융합연구실 연구원
※관심분야 : 빅데이터, 데이터베이스 시스템, 그래프 데이터베이스, 정보 검색(IR), 개체식별기술, 접근제어 등

이석형(Seok-Hyoung Lee)

2001년 : 충남대학교 대학원(공학석사)
2012년 : 충남대학교 대학원(정보학박사)

2001년~현재: 한국과학기술정보연구원(KISTI) 정보융합연구실 선임연구원
※관심분야 : 정보처리(information on processing), 정보분석(information analysis),
빅데이터 분석(bigdata analysis) 등

서동준(Dongjun Suh)

2007년 :한국과학기술원 디지털 미디어프로그램 (공학석사)
2014년 :한국과학기술원 건설 및 환경공학과 건설IT융합프로그램 (공학박사)

2007년~2010년: HUMAX 소프트웨어 연구원
2014년~2015년: KAIST IT 융합연구소 연구조교수
2015년~현재: 한국과학기술정보연구원(KISTI) 선임연구원
2017년~현 재: 과학기술연합대학원대학교(UST) 과학기술정보과학과 부교수
※관심분야 : 빅데이터 분석, 딥러닝, 기계학습, 건설 ICT 융합 등

김광영(Kwang-Young Kim)

2001년 : 부산대 대학원 (공학석사) -한글형태소분석기
2011년 : 충남대 대학원 (문헌정보박사-개인화검색시스템)

2001년~현재: 한국과학기술정보연구원
※관심분야 : 정보검색(IR), 딥러닝기반 개체명인식기, 개인화 검색시스템, PLOT기반 식별기술

윤정선(Kwang-Young Kim)

1993년 : 한국과학기술원 대학원 (공학석사)

1993년~2000년: 한국표준과학연구원
2000년~현재: 한국과학기술정보연구원
※관심분야 : 과학기술 데이터 서비스, 빅데이터, 인공지능 등