Journal of Digital Contents Society Vol. 17 No. 6 Dec. 2016(pp. 555-563)
https://doi.org/10.9728/dcs.2016.17.6.555

A Study on the Method and System for Organization's Name Authorization of Korean Science and Technology Contents

Jinyoung Kim*, Seok-Hyong Lee, Dongjun Suh***, Kwang-Young Kim****

Abstract

Science and technology contents (research papers, patents, reports) are the most common reference material for researchers involved in research and development in the fields of science and technology. Based on various search elements (title, abstract, keyword, year of publication, name of journal, name of author, publisher, etc.), many services are available for users to search science and technology contents and bibliographic information owned by libraries. Authority data on organization name can be useful as an element for author identification and as an element to search for results produced by specific organizations. However, organization name is not taken into account by current search services for domestic academic information and bibliographic records.

This study analyzes organization name data contained in the metadata of science and technology contents, which are the basis of the establishment of authority data, and proposes a method and system based on string containment and exact string matching.

Keywords: Science and technology contents, establishment of authority data, identification of organization name, academic information search

국내 과학기술콘텐츠 전거데이터 구축을 위한 소속기관명 식별 방법과 시스템에 관한 연구

김진영*, 이석형, 서동준***, 김광영****

요 약

과학기술콘텐츠(논문, 특허, 보고서)는 과학기술에 대한 연구와 개발을 위해 연구자들이 가장 많이 활용하는 참고자료이다. 과학기술콘텐츠와 도서관에서 보유 중인 서지 정보 검색을 위해 다양한 검색 요소(제목, 초록, 키워드, 발행 연도, 학술지명, 저자명, 출판사 등)를 활용한 서비스들이 제공되고 있다. 저자의 소속기관명 전거데이터는 저자 식별을 위한 요소, 특정 기관의 연구, 개발 결과물 검색을 위한 요소 등으로 유용하게 활용될 수 있지만 현재 서비스되고 있는 국내 학술 정보와 도서관 서지 검색 서비스들에서는 소속기관명에 대해 고려하지 않고 있다.

이에 따라 본 연구에서는 국내 과학기술콘텐츠의 전거데이터 구축을 위해 식별 대상인 과학기술콘텐츠의 메타데이터에 포함되어 있는 소속기관 데이터를 분석하고 본 연구에서 제안한 문자열 간의 포함관계를 고려한 문자열 완전일치 검색(Exact String Matching) 방법을 활용한 식별 방법과 시스템을 제안한다.

키워드 : 과학기술콘텐츠, 전거데이터 구축, 소속기관명 식별, 학술정보 검색

※ Corresponding Author : Kwang-Young Kim
Received : December 10, 2016
Revised : December 28, 2016
Accepted : Decembe 30, 2016
* Researcher, Korea Institute of Science and Technology Information (KISTI)
** Senior Researcher, Korea Institute of Science and Technology Information (KISTI)
*** Senior Researcher, Korea Institute of Science and Technology Information (KISTI)
**** Senior Researcher, Korea Institute of Science and Technology Information (KISTI)
Tel: +82-42-869-1778 , Fax: +82-42-869-1767
email: glorykim@kisti.re.kr

1. Introduction

Science and technology contents (research papers, patents, reports), produced and registered in vast amounts each day, are the most common reference material for researchers involved in research and development in the fields of science and technology. Based on various search elements, many services(NDSL*, DBpia**, Naver's academic information service***,Web of Scienc e****, etc.) are available for users to search science and technology contents and bibliographic information owned by libraries. The search elements are the metadata of science and technology contents, including title, abstract, keyword, year of publication, name of journal, name of author, and publisher. However, organization name has not been utilized as a search element.

Searching by organization name can further limit search results when combined with other search elements, thus reducing the number of reference materials to be viewed by researchers. By accessing the results of specific organizations, researchers can gain an enhanced understanding of the primary research areas and research trends of such organizations. Organization name is expected to play an important role in academic information and bibliographic information search [6] because it can be utilized as an element in identifying persons, an important factor in the establishment of authority data using science and technology contents.

Previous works [1][4][5] on organization name authorization focus on analysis of the various from of organization name in the science & technology contents and methodologies of the constructing identified data of organization name. These studies suggest how to build the data needed for identifying organization information of content's author, but do not suggest a method for identifying organization information automatically. Emiel Caron et al. [7] suggests un-supervised rule-based organization name identification method rather than supervised approach because organization dictionary data set is usually not available. The method can not be sure the results are correct because it is based on rules excluding real organization dictionary data set.

In this paper, to identify the affiliated organizations of authors of science and technology contents, this study analyzed the characteristics of Korean and English organization names, contained in the metadata of science and technology contents. Multiple names may exist in Korean and English when searching by organization name; these names can be written in full or abbreviated, which can result in poor accuracy and recall. Accordingly, this study established a database for real organization name data set (an organization dictionary), and proposed a automatic method of identification and systematization.

Section 2 explains the need for identification by organization name, and introduces past studies conducted by Korean and international researchers. Section 3 contains analysis of the metadata of science and technology contents, including the characteristics of organization name information. Section 4 describes the organization name identification system proposed in this study. Section 5 is the conclusion, and Section 6 presents directions for future research.

2. Related Work

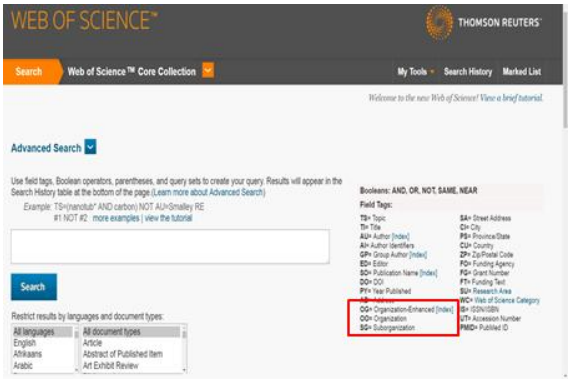
* http://www.ndsl.kr/
** http://www.dbpia.co.kr/
*** http://academic.naver.com/
**** https://apps.webofknowledge.com/

2.1 Need for Identification by Organization Name

Science and technology contents are the most common reference materials for researchers, and are produced and registered in vast amounts each day. Thus, it is essential to develop a method that eliminates redundant efforts in research and development by limiting the number of required reference materials. Identification by organization name is necessary as it can limit search results when combined with other search elements. In addition, researchers can acquire an enhanced understanding of the primary research areas and research trends of specific organizations. Organization name plays an important role in academic information and bibliographic information search because it can be utilized as an important factor in the establishment of authority data using science and technology contents. This method also improves accuracy and recall by enabling researchers to search for organization names using standardized search keywords, preventing inaccuracies caused by names written in different languages or abbreviated forms. In this way, researchers can save time by not having to familiarize themselves with the various names of specific organizations.

2.2 Past Research on Identification by Organization Name

Domestic search services for science and technology contents such as NDSL, DBpia and university library search services do not use organization name as a search element. However, the new version of NDSL launched in December 2016 offers users the option to search by organization name, as proposed in this study. The identification results for organization name will be continuously updated. Naver's academic information service includes title, author, journal, conference, and



(Figure 1) Web of Science

keyword as search elements for academic publications and patents. Its dictionary of organizations* provides information on various organizations. As shown in Fig. 1, Web of Science by Thomson Reuters includes organization name as a field tag.

One study proposed a method of establishing organization authority data based on an analysis of the names of organizations involved in national research and development projects [1]. In-Su Kang et al. [2] mentioned that organization name information is more effective as a feature for author disambiguation than are other elements (titles, academic journals, e-mail addresses, etc.). Seok-Hyoung Lee et al. [3] stated that it is important to construct a name authority database that integrates not only organization name, but also publisher, conference name, previous organization name, and future organization name. Seok-Hyoung Lee et al. [4] designed an author and organization name authority data system based on the Functional Requirements for Authority Data (FRAD). Seok-Hyoung Lee [5] analyzes the various patterns of organization name in metadata of science & technology contents and suggests the construction methodologies of the identified data of author's organization of contents. Emiel Caron et al. [7] suggests un-supervised

* http://terms.naver.com/

rule-based organization name identification method rather than supervised approach because organization dictionary data set is usually not available.

3. Characteristics of Organization Information

The metadata of science and technology contents consists of author name (Korean, English, Chinese characters), date of birth, organization name (Korean, English), e-mail, co-author, name of conference/journal, year of publication, keyword, publication title/abstract, author profile, and theme. The subject of identification in this study is the field of organization name, stored in either Korean or English. Korean organization data, from the start to end of a character string, is listed in the order of upper-level organization to lower-level organization. English organization data, from the start to end of a character string, is listed in the order of upper-level organization to lower-level organization, or in reverse. When an author is affiliated to more than one organization, a separator character (E.g.: ;, ,) is inserted between organization names.

One important characteristic of organization name data in science and technology contents is that the name of an organization can be expressed in various forms, including abbreviations. For instance, Korea Advanced Institute of Science and Technology is also known as Gwagiwon in Korean and is abbreviated as KAIST. If these names and abbreviations are not taken into account, search results will have poor accuracy and recall. Since the names and abbreviations may be very different from the standardized name, this issue cannot be resolved by string similarity matching. As such, this study established a database of organization names containing standard names, nicknames, and abbreviations, and employed the method of exact string matching. The establishment of the organization name database is described in Section 4.

Another characteristic of organization name data is the character string inclusive relationship between organization names. In other words, the organization name matching the longest character string among exact matches in the database is identified and selected. The method proposed in this study is a character string exact matching comparison with consideration of inclusive relationships.

The next section describes the proposed identification system for organization names based on the aforementioned characteristics and identification method.

4. Organization Name Identification System

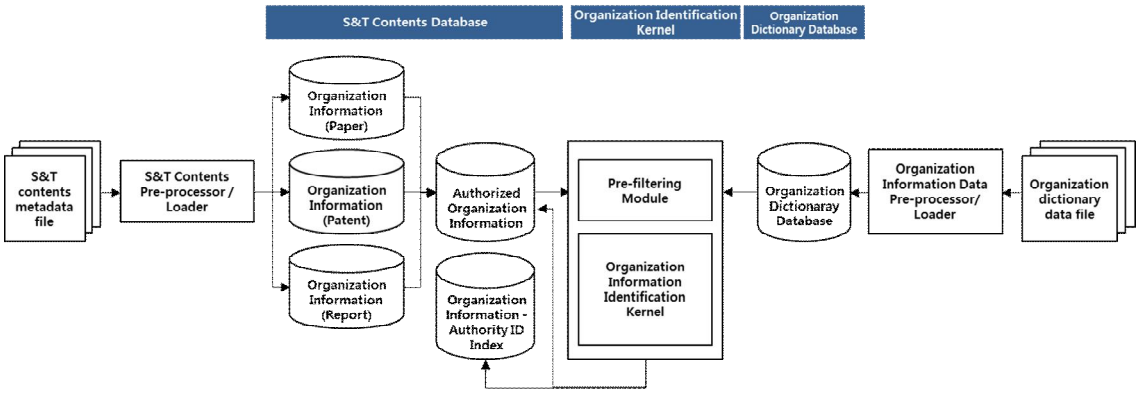
The proposed identification system for organization names automatically identifies representative organization names based on the affiliated organizations contained in the metadata of science and technology contents.

4.1 System Overview

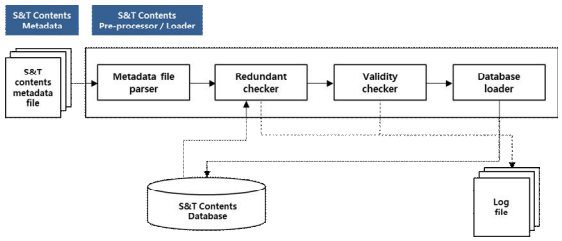
As shown in Fig. 2, the organization identification system consists of the S&T contents pre-processor/loader, organization dictionary database, and organization information identification kernel.

4.1.1 S&T Contents Pre-processor / Loader

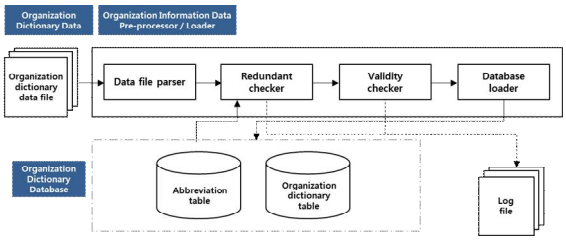
The schematic diagram of the S&T contents pre-processor/loader is given in Fig. 3. S&T contents metadata retrieved from numerous sources (e-Gate, OCEAN, NDSL, NTIS, etc.) are extracted according to the various formats (XML, JSON, CSV). After validity and redundancy checking, they are



(Figure 2) Organization Identification System



(Figure 3) S&T Contents Pre-processor/Loader



(Figure 4) Organization Information Data Pre-processor/Loader

loaded into the database by category (paper, patent, report) and by subject (person, organization, term). Each table containing the subject of identification has an identifier field. The text data file parser extracts data according to the format of the metadata file, and the redundant checker checks whether the extracted data already exists in the database. The validity checker examines the extracted data for errors, and the database loader loads data into the S&T contents database after validity and redundancy checking.

4.1.2 Organization Dictionary Database

The organization dictionary database consists of the abbreviation table and the organization dictionary table. The abbreviation table stores the original forms of nicknames or abbreviations of organization names in Korean or English. It is largely based on the list of abbreviations provided by Thomson Reuters. The organization dictionary table is based on organization information (Korean name, English name, Chinese characters, address, website, etc.) obtained from universities and from the National Medical Center by organization type (education, medical, other). The required fields are added, and the generated data are stored in the database. Korean and English name and Chinese characters are fields to store the representative organization names. The Korean nickname, English nickname, and Chinese nickname are fields to store the different name forms and abbreviations. Additional information such as address, telephone number, and website are used as identification elements in cases of organizations having overlapping names.

To select organizations for the organization information table, hierarchical structures are defined. Organization names for each level and basic information (address, telephone number, website, etc.) are collected. This study

classified organizations into education (university, high school, middle school, elementary school), medical (general hospital, tertiary hospital, hospital, clinic, etc.), and other (government, public, company, etc.). Each categorized organization has lower-level organizations (college, department, medical department, center, research center, major, etc.). Among educational organizations, most high schools, middle schools, and elementary schools do not have lower-level organizations. However, universities can be broken down into lower-level organizations such as campus, college, department, and major. A significant portion of science and technology contents is produced by universities, and the name of a university is usually accompanied by lower-level information. This highlights the importance of identifying the lower-level organizations of universities. In the case of medical organizations, names are accompanied by medical departments. Companies have lower-level organizations such as research institutes and centers, but their organizational structures tend to change more frequently than those of educational or medical organizations. As such, this study did not consider the lower-level organizations of organizations categorized as other.

4.1.3 Organization Name Identification Kernel

The proposed identification kernel for organization names consists of an organization name identification kernel, which employs the exact string method with consideration of character string inclusive relationships as described in Section 3, and a pre-filtering module.

The organization table for the author and organization dictionary database is generated in the S&T contents pre-processor/loader. After identifying organization names for newly acquired science and technology contents using the method described in Section 3, an identification ID is assigned to the identified organization and stored in the identification ID field of the organization dictionary table.

The pre-filtering module minimizes unnecessary calculations performed by the organization identification kernel. Organization names tend to be repeated in more than one language, and there is no need to re-process organization data using the organization name identification kernel if the organization has already been identified. The authority index table, containing Korean and English name pair, the assigned identifier, and the order of authors, is maintained. The pre-filtering module checks the authority index table to determine whether organization names in newly acquired contents have already been identified, and assigns the same identification ID if the organization was previously identified. Otherwise, the organization name identification kernel is used to perform identification, and this is reflected in the organization table in the author and organization data authority search table.

4.1.4 Case study

In this case study, the method is used for the identifying organization name of author's affiliation data in science & technology contents. We collect organization data set through various provider, and build organization dictionary database. Organization dictionary database consists of abbreviation table and organization dictionary table. Tab. 1 is the example of organization dictionary table.

For identifying organization name in authorized organization information table, we execute proposed organization identification kernel with organization dictionary database. Fig. 5 is the example of identification kernel execution result. Organization information(eg. "경북대 자연과학대학 화학과", "Dept. of Chemistry, College of Nat. Sciences, Kyungpook National Univ.") is identified three organizations.

Field name	Description	Records		
		1	2	3
OCN	Control number	UU0000034	UC0000022	UD0000187
OCD	Category type	UU	UC	UD
OCDN	Category typename	대학교	단과대	학과
OKR	Korean name	경북대학교	자연과학대학	화학과
OEN	English name	Kyungpook National University	College of Natural Sciences	Chemistry
OCH	Chiness name	慶北大學校	-	-
AKR	Korean name variations	[경북대학## 경북대]	-	-
AEN	English name variations	[KNU]	-	-
ACH	Chiness name variations	[慶北大學## 慶北大]	-	-
KADDR	Address in korean	대구광역시 북구 대학로 80 경북대학교	-	-
EADDR	Address in english	-	-	-
TEL	Phone number	053-950-6072	-	-
SITE	Website URL	www.knu.ac.kr	-	-

<Table 1> Example of organization dictionary data

```
Korean: 경북대 자연과학대학 화학과
English: Dept. of Chemistry, College of Nat. Sciences, Kyungpook National Univ.

ORG # 1
-- dic information --
--- OCN, OCD, OCDN: UU0000034      UU      대학교
--- OKR, OEN, OCH: 경북대학교      Kyungpook National University      慶北大學校
--- AKR, AEN, ACH: [경북대학, 경북대, 慶北大學, 慶北大]      [KNU]      null
--- KADDR, EADDR: 대구광역시 북구 대학로 80 경북대학교
--- TEL, SITE: 053-950-6072      www.knu.ac.kr

-- dic information --
--- OCN, OCD, OCDN: UC0000022      UC      단과대
--- OKR, OEN, OCH: 자연과학대학      College of Natural Sciences
--- AKR, AEN, ACH: null      null      null
--- KADDR, EADDR:
--- TEL, SITE:

-- dic information --
--- OCN, OCD, OCDN: UD0000187      UD      학과
--- OKR, OEN, OCH: 화학과      Chemistry
--- AKR, AEN, ACH: null      null      null
--- KADDR, EADDR:
--- TEL, SITE:
```

(Figure 5) Example of organization identification kernel result

5. Conclusion

Science and technology contents (research papers, patents, reports) are the most common reference materials for researchers, and developing a method that limits the scope and amount of such contents will improve the efficiency of the research process.

Based on various search elements (title, abstract, keyword, year of publication, name of journal, name of author, publisher, etc.), many services are available for users to search science and technology contents and bibliographic information owned by libraries. Authority data on organization name can be useful as an element for author identification and as an element to search for results produced by specific organizations. However, organization name is not taken into account by domestic search services.

This study proposed a method and system for the identification of names of organizations affiliated to the authors of science and technology contents. Since this system links nicknames and abbreviations to standard organization names, researchers can achieve higher accuracy and recall in their search results.

6. Future Work

As it is based on exact matching of character strings to names stored in the database, the system proposed in this study cannot perform identification for unsaved organization names. The organization name database must be regularly updated to keep up with new openings, abolishments, mergers, and name changes.

The authors of science and technology contents may enter organization names with spelling errors. Because the proposed identification system does not take typos into account, a manual spell check is required to improve accuracy.

A system capable of manual checking and regular database update is being developed. However, such functions involve higher costs, and the system should be further improved to ensure economic feasibility. String similarity matching is being studied as a method of minimizing the cost of manual checking.

While organization history was not

considered in the proposed system, it will be taken into account in the upgraded version. The upgraded system will allow users to analyze results produced by specific organizations and make policy decisions based on organization history and trends over time.

References

[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.

김 진 영

2009년 : 서강대학교 대학원 (공학석사)
2009년~현재 : 한국과학기술원 전산학과 박사과정(수료)
2015년~현재: 한국과학기술정보연구원(KISTI) 연구원
관심분야 : 빅데이터, 데이터베이스 시스템, 그래프 데이터베이스, 정보 검색(IR), 개체식별기술, 접근제어 등

이 석 형

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)

서 동 준

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

김 광 영

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

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