Regular paper

Journal of the KIECS. pp. 141-146, vol. 13, no. 1, Feb. 28 2018, t. 87, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2018.13.1.141

논문의 저자 키워드를 이용한 실시간 연구동향 분석시스템 설계 및 구현

김영찬* · 진병삼* · 배영철*

Design and Implementation of Real-Time Research Trend Analysis System Using Author Keyword of Articles

Young-Chan Kim* · Byoung-Sam Jin* · Young-Chul Bae*

요 약

저자가 작성하는 논문의 저자 키워드는 논문 내용의 특징을 나타내는 가장 중요한 요소이며, 이를 실시간으로 분석하여 사용자에게 제공하게 함으로써, 연구동향을 파악하는 것이 가능하다. 서지로 작성된 논문의 비정형 데이터를 데이터베이스로 구축 하고, 이를 활용하여 실시간 탐색이 가능하도록 색인데이터 구조를 만든다. 특정 키워드가 포함된 논문을 색인데이터 구조에서 탐색하고, 저자키워드를 대상으로 추출, 클러스터링 하여 가중치에 따라 크기별로 나타낼 수 있는 워드클라우드로 사용자에게 제공하여, 연구동향을 가시화 하는 방법을 설계하였다. 또한, 구현된 시스템에서 “바이러스”와 “홍채인식” 키워드를 통하여 연구동향 분석 결과를 제시하였다.

ABSTRACT

The authors' author keywords are the most important elements that characterize the contents of the paper, By analyzing this in real time and providing it to users, It is possible to grasp research trends. Unstructured data of a journal created in a paper is constructed as a database, make use of this to make index data structure that can search in real time. In the index data structure, a thesis containing a specific keyword is searched, By extracting and clustering the author keywords, By presenting to the user a word cloud that can be displayed by size according to the weight, designed a method to visualize research trends. We also present the results of the research trend analysis of the keywords ″virus″ and ″iris recognition″ in the implemented system.

키워드

Author Keywords, The Research Trend, Visualization, Word Cloud, Index Structure
저자 키워드, 연구 동향, 시각화, 워드 클라우드, 색인 구조

Ⅰ. 서 론

논문의 저자키워드는 논문을 구성하는 필수적인 요소 중에 하나이며, 저자가 연구 작성한 논문 내용의 특징을 나타내기 위한 가장 핵심적이고 중요한 용어를 선정하고 표현한다. 이처럼 저자키워드는 저자뿐 아니라, 논문의 이용자 또한 저자키워드가 중요하다는 것에 대해서는 누구라도 동의할 것이다[1].

* 전남대학교 전기및반도체공학과
* 교신저자 : 배영철(ycbae@jnu.ac.kr)
· 접 수 일 : 2017. 12. 07
· 수정완료일 : 2018. 01. 11
· 게재확정일 : 2018. 02. 15
· Received : Dec. 07, 2017, Revised : Jan. 11, 2018, Accepted : Feb. 15, 2018
· Corresponding Author : Young-Chan Kim
Dept. Electrical and Semiconductor Engineering, Chonnam University,
Email : chanible@naver.com

그동안 논문의 저자키워드와 관련한 연구들을 살펴보면, 정보검색을 위한 색인어로써의 활용과 관련한 연구들이 있으며[2], 저자키워드의 비통제어휘적 특징에 대한 연구를 찾아볼 수 있다[3]. 또한, 저자키워드와 MeSH용어와의 비교를 통해 저자키워드 비통제어휘로써의 특징을 살펴본 연구가 있다[4]. 최근에는 논문의 저자키워드에 대한 네트워크 분석방법을 적용한 연구들이 진행되고 있다. 이러한 연구들은 논문에서 저자키워드를 추출하고 클러스터링 하여 논문이 속한 주제분야의 지적구조 및 연구 동향을 알아내고자 한 연구들이다[5-12].

본 논문은 선행 연구된 결과를 바탕으로, 연구자가 선정한 핵심 용어인 저자키워드를 활용하여 연구동향을 파악할 수 있다는 관점으로, 분석자가 원하는 키워드가 포함된 논문을 탐색하고, 탐색된 논문의 저자키워드를 분석하여 워드클라우드로 제시하고자 한다.

II. 본론

2.1 논문의 DB 설계

논문은 비정형 데이터인 학술지 형태로 발간되며, 이를 정형화된 데이터베이스로 구축하는 것이 필요하다. 이는 어플리케이션 측면에서 데이터를 취급하기 위함이며, 또한 동적인 처리를 위하여 데이터베이스 자원을 참조하기 위함이다[13].

정보처리 방법에 있어 데이터베이스화 하는 것은 데이터의 관리 및 가공, 분석을 원활하게 하기 위한 데이터 설계에 해당한다.

'표 1' 은 논문의 정형화를 위한 데이터베이스 설계의 기초 데이터 구분을 나타낸 것이며, 이러한 기초데이터를 활용하여 실시간 분석이 가능하다.

표 1. 논문의 데이터 속성
Table 1. Data attribute of article

No	Data attribute	Explanation
1	Language	Paper writing language
2	Title - Korean	Korean paper's title
3	Title - English	English paper's title
4	Authors - Korean	Authors - Korean
5	Authors - English	Authors - English
6	Affiliation - Korean	Affiliation - Korean
7	Affiliation -English	Affiliation - Korean
8	e-mail address	e-mail address
9	Corresponding author	Corresponding author
10	Beginning page	Number of Beginning page of paper
11	Ending page	Number of Ending page of paper
12	Abstraction	Abstraction
13	Content	Content
14	Author's keyword	Author's keyword
15	Funding information	Funding information
16	Reference	Reference information

2.2 실시간 탐색을 위한 색인구조 설계

정형화된 논문 데이터베이스를 기반으로 실시간 탐색을 하기 위하여 형태소분석을 통한 색인기법을 활용한다[14]. 색인구조 설계는 대용량 탐색에서 빠른 결과를 도출할 수 있는 장점이 있다. 논문의 데이터 속성별로 탐색을 할 경우에는 많은 시간이 소요되므로 분석결과에 대한 빠른 탐색을 보장한다.

'그림 1'은 특정 키워드를 통해서 문서의 위치정보를 탐색하고, 그 문서의 저자키워드를 클러스터링하여 반환하는 분석시스템의 흐름도를 나타낸 것이다.

흐름도는 아래 6단계 절차를 거쳐 최종적으로 사용자에게 연구동향을 제시한다.

step-1. 탐색을 수행하는 사용자는 연구동향을 파악하기 위하여 특정 키워드가 포함하는 논문을 대상으로 탐색을 요청한다.

step-2. 분석시스템은 색인구조에서 요청한 키워드가 포함하는 논문의 DOC-ID를 탐색한다.

step-3. 탐색된 DOC-ID가 가지고 있는 저자키워드를 데이터베이스로부터 획득한다.

step-4. 획득한 저자키워드를 클러스터링 하여 '그림 1'과 같이 키워드별 가중치로 구조화된 데이터를 만든다.

step-5. 구조화된 데이터를 탐색을 수행한 사용자에게 제공한다.

step-6. 마지막으로 클러스터링된 저자키워드를 통하여 가중치에 따라 키워드별 중요도를 워드클라우드로 시각화하여 나타낸다.

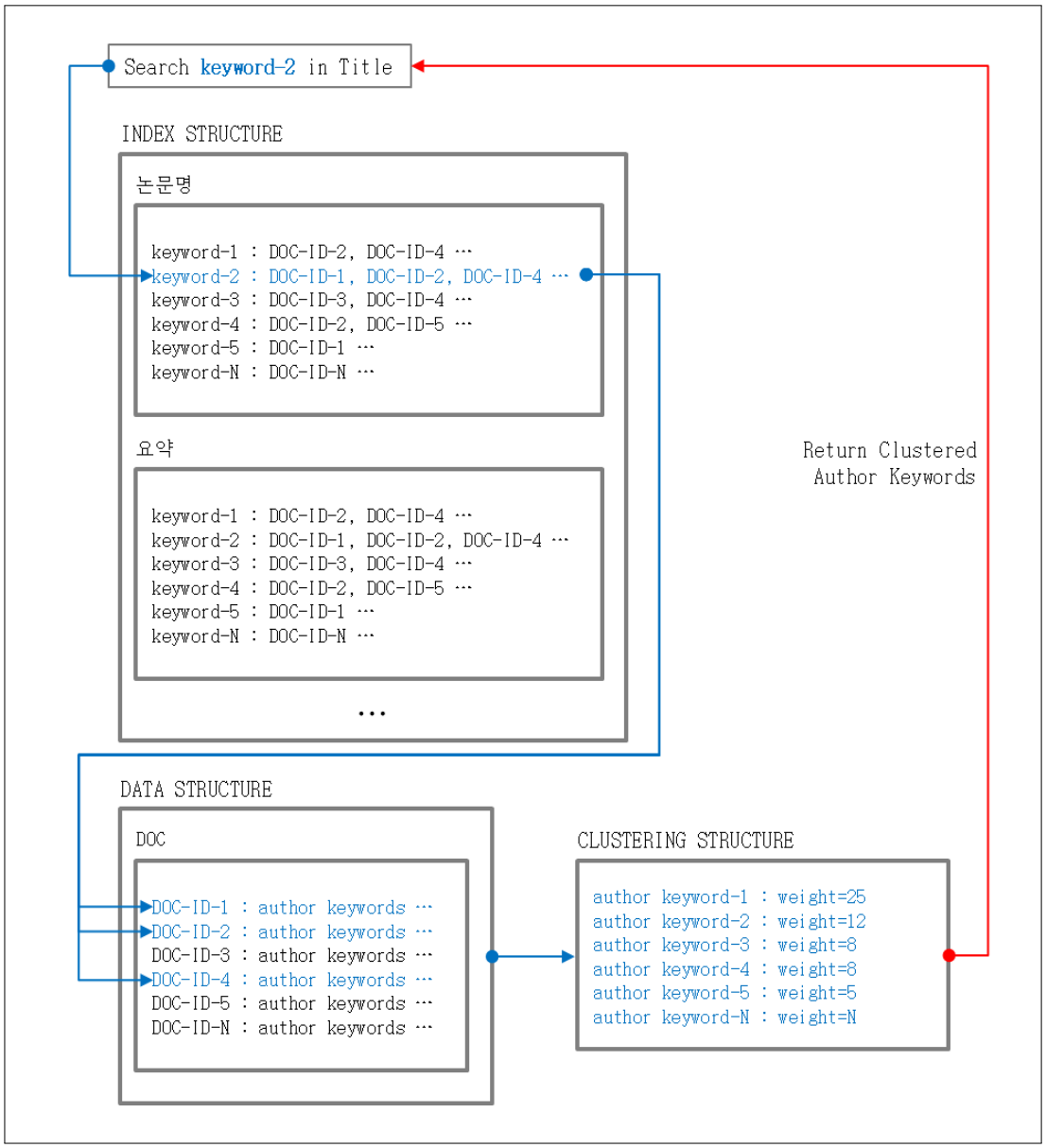

그림 1. 분석을 위한 색인구조 및 검색 흐름도
Fig. 1 Index structure and search flow chart for analysis

2.3 분석결과 시각화

워드클라우드는 문서의 키워드를 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 크기형태로 구분하여 돋보이게 하는 기법이다.

앞서 실시간 탐색방법에 의해 검색된 저자키워드를 클러스터링하고, 가중치를 활용하여 워드클라우드로 가시화 하여 의미있는 분석결과를 제공한다.

‘표 2는 “바이러스” 키워드가 논문명에 포함된

4867건의 논문을 실시간 분석한 클러스터링 결과 목록이다.

표 2. “바이러스”의 클러스터링과 가중치 값
Table 2. Clustering and weight values of “바이러스”

No	Keyword	Weight
1	RT-PCR	59
2	ELISA	31
3	Rotavirus	30
4	PCR	27
5	Adenovirus	21
6	Children	16
7	Hantavirus	16
8	CMV	13
9	“자궁경부암”	13
10	Enterovirus	12
11	Human papillomavirus	12
12	Influenza virus	12
13	Knowledge	12
14	Norovirus	12
15	immunohistochemistry	12
16	“지식”	12
17	CGMMV	11
18	Gastroenteritis	11
19	Japanese encephalitis virus	11
20	Epidemiology	10

이 클러스터링된 목록을 다시 워드클라우드로 표현하면 ‘그림 2’와 같이 표현할 수 있다.

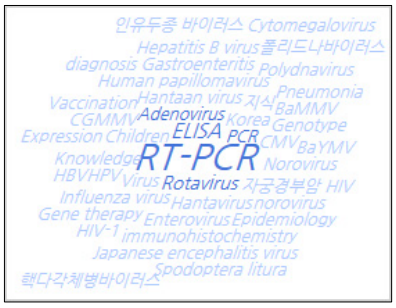

그림 2. “바이러스”의 트랜드 분석 결과
Fig. 2 Trend analysis result of “바이러스”

RT-PCR, ELISA, Rotavirus, PCR, Adenovirus 등의 키워드가 가중치가 높으며 시각적으로 돋보이는 것을 확인할 수 있다.

‘표 3’은 “홍채인식” 키워드가 논문명에 포함된 139건의 논문을 실시간 분석한 클러스터링 결과 목록이다.

표 3. “홍채인식”의 클러스터링과 가중치 값
Table 3. Clustering and weight values of “홍채인식”

No	Keyword	Weight
1	IrisRecognition	11
2	“홍채인식”	8
3	Biometrics	6
4	Iris	6
5	“생체인식”	4
6	“홍채”	4
7	Scale-spacefiltering	3
8	Biometric	2
9	EyelashDetection	2
10	EyelidDetection	2
11	Identification	2
12	Irisrotationinvariance	2
13	Recognition	2
14	ZernikeMoment	2
15	“정보보호”	2
16	Adaboost	1
17	BigData	1
18	Biometries	1
19	BrandEquity	1
20	Cameramodule	1

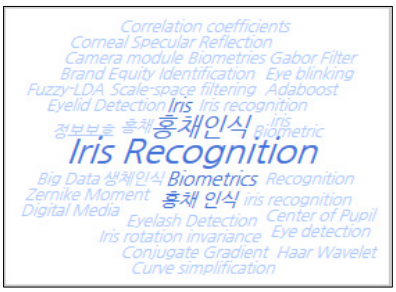

그림 3. “홍채인식”의 트랜드 분석 결과
Fig. 3 Trend analysis result of “홍채인식”

이 클러스터링된 목록을 다시 워드클라우드로 표현하면 '그림 3'과 같이 표현할 수 있다.

Iris Recognition, 홍채인식, Biometrics, Iris, 생체인식 등의 키워드가 가중치가 높으며 시각적으로 돋보이는 것을 확인할 수 있다. 이것은 이 키워드 분야를 중심으로 많은 연구가 진행되어 왔음을 의미한다.

III. 결론 및 향후개선 방향

본 논문에서는 논문의 저자가 작성하는 저자키워드의 중요성을 바탕으로, 탐색에 최적화된 색인기법의 적용과 클러스터링 기법을 활용하여, 저자키워드 가중치에 따른 워드클라우드를 시각적으로 돋보이게 제시함으로써, 연구동향을 가시화 하는 분석시스템 설계 및 구현방법을 제안하였다.

사용자가 질의하는 검색어와 탐색대상이 되는 제목, 요약, 본문 등을 선별하여 논문을 실시간 탐색하였을 때, 탐색결과에 해당하는 논문들의 저자키워드를 클러스터링 하여 노출빈도에 따른 가중치 값을 활용하여 워드클라우드로 사용자에게 실시간으로 분석 결과를 제시한다.

사용자는 질의어에 대한 분석결과에서 특정 저자키워드가 많이 연구되고 있음을 시각적으로 파악할 수 있게 되고, 이를 통하여 연구동향을 알 수 있게 된다.

향후에는 저자키워드들 간의 관계를 분석하고, 특정키워드의 유사키워드들을 필터링하여 단일의미를 지니고 있는 저자키워드들이 노출되도록, 분석결과의 품질을 높이는 연구가 필요하다.

References

[1] S. Kwon, "A Study on the Application to Network Analysis on the Importance of Author Keyword based on the Position of Keyword," *J. of the Korea Society for Information Management*, vol. 31, no. 2, 2014, pp. 121-142.

[2] Taghva. K, Borsack. J, Nartker. T, Condit. A. "The role of manually assigned keywords in query expansion," *Information Processing & Management*, vol. 40, 2004, pp. 441-458.

[3] C. Lee, S. Lee, "A Study on the Analysis of Web Dance Information and Its Users to Build A Web-Based Dance information System," *Korea Society for Information Management*, 2000. 08.

[4] S. Park, K. Park, "Coincidence Analysis of Key Words and MeSH Terms in the Journal of the Korean Society of Occupational Therapy," *J. of the KSOT*, vol. 19, no. 4, 2011, pp. 131-146.

[5] Y. Lee, "An Analysis of the Key Words in Korean Journal of Japanology in Korea," *J. of Japanese Language and Literature*, no. 31, 2006, pp. 107-121.

[6] O. Park, "Knowledge Structures in Knowledge Organization Research: 2000-2011," *J. of the Korean Biblia Society for Library and Information Science*, vol. 22, no. 3, 2011, pp. 247-267.

[7] H. Lee, S. Kwak, "Relation Analysis Among Academic Research Areas Using Subject Terms of Domestic Journal Papers," *J. of the Korean Biblia Society for Library and Information Science*, vol. 22, no. 3, 2011, pp. 353-371.

[8] J. Cho, "A Study for Research Area of Library and Information Science by Network Text Analysis," *J. of the Korean Society for Information Management*, vol. 28, no. 82, 2011, pp. 65-83.

[9] B. Kang, J. Park, "Profiling and Co-word Analysis of Teaching Korean as a Foreign Language Domain," *J. of the Korean Society for Information Management*, vol. 30, no. 4, 2013, pp. 195-213.

[10] S. Seo, E. Chung, "Domain Analysis on the Field of Open Access by Co-Word Analysis," *J. of the Korean Biblia Society for Library and Information Science*, vol. 24, no. 1, 2013, pp. 207-228.

[11] L. Zhang, H. Hong, "Examining the Intellectual Structure of Reading Studies with Co-Word Analysis Based on the Importance of Journals and Sequence of Keywords," *J. of the Korean Biblia Society for Library and Information Science*, vol. 25, no. 1, 2014, pp. 295-318.

[12] Y. Kim, Y. Bae, "A proposal of real time analysis method on research trends of researchers using author keywords in articles," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 10, no. 1, 2016. 06.
[13] Y. Kim, T. Kim, S. Lee, K. Rim, J. Lee, "Database Connection Pool Architecture for User Interconnections Access," *J. of the Korea Contents Association*, vol. 9, no. 1, 2009, pp. 89-97.
[14] M. Park, "A Study on the Extraction and Utilization of Index from Bibliographic MARC Database," *J. of the Korean Library and Information Science Society*, vol. 36, no. 2, 2005, pp. 327-348.

저자 소개

김영찬(Young-Chan Kim)

2007년 평생교육진흥원 컴퓨터공학과(공학사)
2009년 인하대학교 공학대학원 정보공학과(공학석사)
2016년 ~ 현재 전남대학교 대학원 전기및반도체공학과(박사과정)
2009년 ~현재 ㈜미소테크 차장
※ 관심분야 : 데이터베이스, 데이터분석, 강화학습

진병삼(Byoung-Sam Jin)

1998년 한국항공대학교 통신정보공학과(공학사)
2004년 고려대학교 컴퓨터과학기술대학원 소프트웨어공학과(공학석사)
2016년 ~ 현재 전남대학교 대학원 전기및반도체공학과(박사과정)
2009년 ~현재 ㈜미소테크 대표이사
※ 관심분야 : 데이터베이스, 데이터분석, 강화학습

배영철(Young-Chul Bae)

1984년 광운대학교 전기공학과 (공학사)
1986년 광운대학교대학원 전기공학과(공학석사)
1997년 광운대학교대학원 전기공학과(공학박사)
1986년 ~ 1991년 한국전력공사
1991년 ~ 1997년 산업기술정보원 책임연구원
1997년 ~ 현재 전남대학교 전기·전자통신·컴퓨터 공학부 교수
2002년 ~ 2002년 Brigham Young University 방문교수
2011년 ~ 2011년 University of Utah 방문교수
※ 관심분야 : Chaos Control and Chaos Robot, Robot control etc.