한국기록관리학회지
23(4), 73-89, 2023.11
https://doi.org/10.14404/JKSARM.2023.23.4.073
pISSN 1598-1487 eISSN 2671-7247

JKSARM
Journal of Korean Society of Archives and Records Management

비정형 텍스트 데이터 분석을 활용한 기록관리 분야 연구동향

Research Trends in Record Management Using Unstructured Text Data Analysis

홍덕용(Deokyong Hong)[1], 허준석(Junseok Heo)[2]

E-mail : igre@korea.kr, sky@atni.kr
[1] 부산광역시 수영구청 기록물관리전문요원
[2] ㈜에이티앤아이 대표이사

OPEN ACCESS

논문접수 2023-10-16
최초심사 2023-10-23
게재확정 2023-11-06

ORCID

Deokyong Hong
https://orcid.org/0000-0002-8052-6963

Junseok Heo
https://orcid.org/0009-0000-3022-2788

초 록

본 연구에서는 텍스트 마이닝 기법을 활용하여 국내 기록관리 연구 분야의 비정형 텍스트 데이터인 국문 초록에서 사용된 키워드 빈도를 분석하여 키워드 간 거리 분석을 통해 국내기록관리 연구 동향을 파악하는 것이 목적이다. 이를 위해 한국학술지인용색인(Korea Citation Index, KCI)의 학술지 기관통계(등재지, 등재후보지)에서 대분류(복합학), 중분류(문헌정보학)으로 검색된 학술지(28종) 중 등재지 7종 1,157편을 추출하여 77,578개의 키워드를 시각화하였다. Word2vec를 활용한 t-SNE, Scattertext 등의 분석을 수행하였다. 분석 결과, 첫째로 1,157편의 논문에서 얻은 77,578개의 키워드를 빈도 분석한 결과, "기록관리"(889회), "분석"(888회), "아카이브"(742회), "기록물"(562회), "활용"(449회) 등의 키워드가 연구자들에 의해 주요 주제로 다뤄지고 있음을 확인하였다. 둘째로, Word2vec 분석을 통해 키워드 간의 벡터 표현을 생성하고 유사도 거리를 조사한 뒤, t-SNE와 Scattertext를 활용하여 시각화하였다. 시각화 결과에서 기록관리 연구 분야는 두 그룹으로 나누어졌는데 첫 번째 그룹(과거)에는 "아카이빙", "국가기록관리", "표준화", "공문서", "기록관리제도" 등의 키워드가 빈도가 높게 나타났으며, 두 번째 그룹(현재)에는 "공동체", "데이터", "기록정보서비스", "온라인", "디지털 아카이브" 등의 키워드가 주요한 관심을 받고 있는 것으로 나타났다.

ABSTRACT

This study aims to analyze the frequency of keywords used in Korean abstracts, which are unstructured text data in the domestic record management research field, using text mining techniques to identify domestic record management research trends through distance analysis between keywords. To this end, 1,157 keywords of 77,578 journals were visualized by extracting 1,157 articles from 7 journal types (28 types) searched by major category (complex study) and middle category (literature informatics) from the institutional statistics (registered site, candidate site) of the Korean Citation Index (KCI). Analysis of t-Distributed Stochastic Neighbor Embedding (t-SNE) and Scattertext using Word2vec was performed. As a result of the analysis, first, it was confirmed that keywords such as “record management” (889 times), “analysis” (888 times), “archive” (742 times), “record” (562 times), and “utilization” (449 times) were treated as significant topics by researchers. Second, Word2vec analysis generated vector representations between keywords, and similarity distances were investigated and visualized using t-SNE and Scattertext. In the visualization results, the research area for record management was divided into two groups, with keywords such as “archiving,” “national record management,” “standardization,” “official documents,” and “record management systems” occurring frequently in the first group (past). On the other hand, keywords such as “community,” “data,” “record information service,” “online,” and “digital archives” in the second group (current) were garnering substantial focus.

Keywords: 기록관리연구동향, 빅데이터, 텍스트마이닝, t-분포확률적임베딩, 산점도
Research Trends in Record Management, Big Data, Text Mining, t-SNE, Scattertext

1. 서론

과거의 기록관리는 종이 문서나 아날로그 형태의 기록물이 주를 이루었지만, 최근 IT기술의 발달과 4차 산업의 발전으로 전자기록물과 행정정보데이터세트 등 다양한 유형의 기록이 등장하였다. 공공기록물 관리에 관한 법과 전자정부법 등 기록관리와 관련된 다양한 법령이 입법됨에 따라 국내기록관리 환경은 지속적으로 발전하고 있다. 대한민국의 발전 역사에서 "기록관리"의 역할은 상당히 중요하며, 기록관리 연구자들의 기여로 인해 학문은 양적으로나 질적으로 큰 발전을 이루었다.

2000년 7월 한국기록관리학회를 시작으로 한국기록학회(2000년 12월) 등 현재 기록관리학 분야는 다양한 학회들을 중심으로 다각적인 연구가 이루어지고 있다. 기록관리의 학술대회는 국내외 다양한 학자들이 기록관리 연구의 성과를 공유하는 장으로, 다양한 국가와 지역에서 개최되었으며 이를 통해 연구 분야가 발전하고 있다. 학회별로 논문 발행 횟수는 연 1회에서 시간이 경과함에 따라 연 4회 이상으로 증가하고 있다.

과거의 연구 동향을 살펴보는 연구는 질적연구와 정형 데이터를 위주로 분석하여 연구가 진행되었으나 빅데이터 시대가 도래함에 따라, 비정형 데이터 특히 텍스트 등을 빅데이터 분석기술을 활용하여 동향 연구가 가능해졌다.

본 연구에서는 텍스트 마이닝 기법을 활용하여 국내 기록관리 연구 분야의 비정형 텍스트 데이터인 국문 초록에서 사용된 키워드를 추출하고, 그 사용 빈도를 분석하여 키워드 간 거리 분석을 통해 특정 주제에 대한 관심도와 추세를 한눈에 파악하는 것이 목적이다. 이를 위하여 한국학술지인용색인(Korea Citation Index)에서 복합학(대분류), 문헌정보(중분류)로 검색된 학술지 7종을 대상으로 하여 2004년에서 2023년까지의 1,157편의 논문들의 77,578개의 키워드를 살펴봄으로써 기록관리연구의 지적구조에 대해 체계적으로 살펴보고자 하였다. 비정형 텍스트 데이터를 분석하여 얻은 특정 키워드의 주기적인 빈도수 확인을 통해, 기록관리 연구자는 어떤 주제나 키워드에 집중해야 할지를 결정할 수 있으며, 이를 기반으로 업무 프로세스를 개선하고 정보자원을 효율적으로 할당할 수 있을 것이다.

2. 이론적 배경

2.1 기록관리연구의 발전

기록관리법 제정 이전의 법령[1]은 관리 자체를 목적으로 하지 않았다. 따라서 행정업무의 능률향상이나 사무처리의 간소화 및 효율화 등이 주된 관심사였다. 또한 행정의 결과물로서 만들어진 문서들은 기록물로서 그것이 정보자원이라는 인식을 하지 않고, 결정된 보존기간에 따라 처리해야 하는 대상물에 지나지 않았다. 따라서 문서를 만들 때 보존기간을 어떻게 결정하도록 할 것인가에 대한 규정이 관심사였다고 할 수 있다.

대한민국의 거버넌스가 발전함에 따라 기록물의 양이 늘어났으며 효율적으로 기록물을 관리하고 보존하기 위한 필요성이 대두되었다. 1999년 「공공기관의 기록물관리에 관한 법률」이 제정되면서 기록정보의 축적을 통해 정보자원으로서 기록의 가치가 새롭게 강조되게 되었다. 따라서 기록정보를 어떻게 축적할 것이며, 축적된 기록정보를 활용하도록 하기 위한 수단과 방법을 강구하며, 서비스를 통해서 정부 내에서 정부와 국민이 상호작용하는 관계를 어떻게 만들어 가는가에 관심을 갖게 되었다. 기록관리법 제정 이전과의 가장 큰 차이점은 이처럼 기록의 생산과 이용의 균형을 이룰 수 있도록 하는 것에 있다. 한국의 기록관리 분야의 발전에 따라 학문적인 연구도 활발하게 수행되어왔다는 사실은 부인할 수 없다.

2.2 기록관리 연구동향 선행연구

기록관리 연구동향에 대한 분석은 기록관리학 분야 학술지와 문헌정보학분야 학술지를 대상으로 진행되었다.

1) 정부처무규정(1949), 공문서규정(1950), 정부공문서규정(1961) 등

https://doi.org/10.14404/JKSARM.2023.23.4.073

기록관리 연구동향의 게재순으로 살펴보면 다음과 같다. 이재윤, 문주영, 김희정(2007)은 5년간(2001년 ~ 2006년) 145편의 논문을 대상으로 문헌 클러스터링을 활용하여 연구동향을 파악하였다. 김규환, 장보성, 이현정(2009)은 10년간(1999년 ~ 2009년 9월) 374편의 논문을 대상으로 배경 정보와 주제영역을 활용하여 연구자 특성에 따른 주제 영역 상관성을 분석하였다. 남태우, 이진영(2009)은 13년간(1997년 ~ 2009년) 399편을 대상으로 주제별, 간행시기, 학회지, 연구자 등의 분포를 분석하였다. 김판준, 서혜란(2012)은 연구제목, 저자키워드, 목차, 초록 등의 키워드를 통해 프로파일링 기법을 이용하여 국내 전자기록분야 지적 구조를 분석하였다. 최이랑(2015)은 10년간(2004년 ~ 2013년) 479편의 논문을 대상으로 내용 및 네트워크 분석을 통해 기록관리연구에서 최빈도 키워드와 참여 기관 등을 파악하였다. 손혜인, 남영준(2016)은 16년간(2000년 ~ 2015년) 681편을 대상으로 빈도와 네트워크 분석을 실시하여 연구자 배경과 주제 변화 추이 등을 파악하였다. 박준형, 오효정(2017)은 20년간(1997년 ~ 2016년) 1,027편을 대상으로 전처리 후 LDA토픽모형과 HDP토픽모형을 수행하였는데 LDA는 국내 기록관리학 내에 거시적으로 대표되는 주제들을, HDP는 세부 주제별 미시적인 핵심 키워드를 도출하는데 효과적임을 알 수 있었다. 박준형, 류법모, 오효정(2018)은 1997년부터 2016년까지 기록관리학 관련 논문을 자동 수집하여 토픽모델링을 수행하여 시기별로 상관관계가 있는지를 파악하여 전자기록, 기록정보서비스, 다양한 기록물의 기록화 방안에 관한 연구가 점차 활성화 되는 것을 파악하였다.

2.3 텍스트 마이닝 기법을 활용한 선행연구

텍스트 마이닝(Text mining)은 대량의 비정형 데이터에서 유용한 정보를 추출하고, 이를 분석하여 인사이트를 도출하는 기술이다. 텍스트 데이터는 우리가 생활하면서 남기는 채팅 기록, 이메일, 웹사이트 게시글, 뉴스 기사, 소셜 미디어 게시글 등 다양한 형태로 존재한다. 이러한 데이터를 텍스트 마이닝 기법을 사용하여 분석하면, 대량의 문서에서 키워드를 추출하고, 이를 분석하여 문서 집합의 특성을 파악하는 토픽 모델링(Topic Modeling), 문서 내에서 긍정, 부정, 중립과 같은 감성 정보를 추출하는 감성 분석(Sentiment Analysis), 문서 내에서 사람, 장소, 조직, 날짜 등과 같은 중요한 개체를 인식하고 분석하는 엔티티 인식(Entity Recognition), 문서들의 유사한 특성을 가지는 그룹으로 나누어 분석하는 군집화(Clustering), 문서 내에서 특정 정보를 추출하는 정보 추출(Information Extraction) 등을 할 수 있다.

이러한 텍스트 마이닝 기법은 비즈니스, 마케팅, 경영 전략, 제품 개발과 의료, 법률, 정치 등 다양한 분야에서 활용되고 있다. 다른 학문 분야들에서도 문헌 자료와 논문의 양이 계속해서 증가하기 때문에 대용량 텍스트를 분석하여 인사이트를 도출하는 것이 중요한 이슈로 대두되고 있다. 텍스트 마이닝 기법을 활용하여 연구동향을 분석한 선행연구에는 김규하, 박철용(2015)의 한국데이터정보과학회지의 영문초록을 토픽 모형과 사회연결망기법을 활용하여 연구동향을 분석하였다. 조수곤, 김성범(2012)은 산업공학 학술지의 주제어간 연관관계를 텍스트 마이닝 기법을 활용하여 분석하였으며 배규용 외(2013)는 기후변화관련 식품 분야의 논문초록을 텍스트 마이닝기법을 활용하여 분석하였다. 정용복, 박의섭(2015)은 암반공학분야 SCI논문의 주제어를 텍스트 마이닝 기법으로 분석하였다. 이상복(2019)은 품질경영학회지 연구동향을 텍스트 마이닝 기법으로 분석하였으며 김지영, 김은혜, 이지영(2019)은 무용경영 연구동향과 지식구조 분석, 윤희영, 곽일엽(2020)은 한국무역연구분야 키워드와 초록의 데이터를 텍스트 마이닝 기법을 활용하여 연구동향을 분석하였다.

2.4 단어의 표현과 시각화 이론

. Word2vec

Word2Vec[2]은 자연어 처리 분야에서 널리 사용되는 기술 중 하나로, 단어의 분산 표현을 학습하는 방법이다. 단어의 분산 표현이란, 단어를 고차원 벡터로 표현하는 것을 의미하는데 Word2Vec은 특정 문맥에서 단어가 얼마

2) 2013년에 발표된 자연어처리(Natural Language Processing, NLP)기술

https://doi.org/10.14404/JKSARM.2023.23.4.073

나 자주 등장하는지를 분석하여, 각 단어를 고차원 벡터로 표현한다. 이러한 벡터 표현은 단어의 의미와 관련된 정보를 포착하는 데 도움이 된다. Word2Vec의 거리구하는 공식은 유클라디안, 맨하튼, 코사인 거리 등이 있는데 본 연구에서는 벡터의 크기와 방향성을 고려하는 코사인 거리를 사용하였다. 다음 코사인 거리의 공식은 아래와 같다.(Ethayarajh, Duvenaud & Hirst, 2019)

$$Cosine Distance(A, B) = 1 - \frac{A \cdot B}{||A|| \cdot ||B||}$$

A와 B는 두 개의 벡터이고 벡터의 내적(inner product)을 나타낸다. ||A||는 A의 크기(norm)를 나타내며, ||B||는 B의 크기(norm)를 나타낸다. 예를 들어, 'King'을 나타내는 벡터를 K, 'male'을 나타내는 벡터를 M, 'female'을 나타내는 벡터를 F라고 두고 'King'-'male'과 'King'-'female' 간의 코사인 거리를 아래와 같이 계산하면 다음과 같다.

$$Cosine\ Distance(K, M) = 1 - \frac{K \cdot M}{| |K| | \cdot | |M| |}$$

$$Cosine Distance(K, F) = 1 - \frac{K \cdot F}{| |K| | \cdot | |F| |}$$

||K||는 K의 크기(norm)를 나타낸다. 'King'-'male'벡터의 크기(norm)가 'King'-'female'벡터의 크기(norm)보다 크면, 'King'-'male'과 'King'-'female' 간의 코사인 거리는 'King'-'male'과 'King'-'female'간의 거리보다 작아진다. 이는 'King'과 'male'사이의 유사성이 'King'과 'female'사이의 유사성보다 더 높다는 것을 나타낸다. 이와 같은 방식으로 Word2Vec의 임베딩 기법을 사용하여 단어 간의 유사성을 계산할 수 있으며 이를 통해 단어 간의 관계를 파악하고, 유추 및 비유 관계를 모델링할 수 있다.

Word2Vec은 크게 CBOW(Continuous Bag of Words)와 Skip-gram 두 가지 모델로 구성되는 데 CBOW 모델은 주변에 있는 단어들을 이용해 중심 단어를 예측하는 방식으로 학습을 진행하며, Skip-gram 모델은 중심 단어를 이용해 주변 단어를 예측하는 방식으로 학습을 진행한다. 본 연구에서는 학습하고자 하는 텍스트 데이터를 수집한 후, 단어 단위로 토큰화하고 디폴트값을 사용한 하이퍼파라미터인 벡터 차원(Dimension), 학습 속도(Learning Rate), 그리고 에폭(Epochs)을 설정하여 CBOW모델로 학습시켰다.

. t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding)은 고차원 데이터를 저차원으로 압축하고 시각화하기 위한 차원 축소 알고리즘이다.(Maaten & Hinton, 2008) t-SNE는 고차원 벡터를 2D 또는 3D로 축소하여 시각적으로 표현할 수 있는 데이터를 생성하지만 t-SNE 자체만으로는 시각화를 수행할 수 없으며, 축소된 값을 기반으로 한 다른 시각화 라이브러리를 사용해야 한다(Parra et al., 2020). t-SNE는 데이터의 유사성을 보존하면서 차원 축소를 수행한다. 즉, 비슷한 데이터들은 저차원에서도 가깝게, 멀리 떨어진 데이터들은 저차원에서도 멀게 배치된다. 이를 통해 시각화 결과로 비슷한 특성을 가진 데이터끼리 클러스터링이 가능하며, 이를 통해 데이터에 대한 인사이트를 얻을 수 있다. t-SNE는 대규모 데이터셋에서는 계산 비용이 매우 높아지는 단점이 있으며, 학습 파라미터에 따라 결과가 달라질 수 있다는 점도 주의해야 한다.[3)]

3) https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

https://doi.org/10.14404/JKSARM.2023.23.4.073

. Scattertext

Scattertext는 텍스트 데이터를 시각화하고 분석하기 위한 도구이다. Scattertext 플롯은 두 가지 범주의 텍스트가 서로 다른 방식으로 사용되는 좌표 기반 방식을 상호작용적으로 시각화할 수 있다(Opacich, 2021). Scattertext는 텍스트 데이터를 시각화하기 위해 많은 단계를 거쳐 전처리해야 하는데, 이 때 null 값, 이상값, 중복값 등을 제거하고, 기호, 숫자, 공백, 불용어를 제거하는 등의 처리를 수행하여야 한다. Scattertext는 텍스트 데이터의 두 가지 범주 간의 차이를 시각화하며, 각 단어의 사용 빈도와 범주에 따라 단어들을 플롯에 나타내는데 플롯에서 단어의 Y축 위치는 해당 단어가 첫 번째 범주에서 얼마나 많이 사용되었는지를, X축 위치는 두 번째 범주에서 얼마나 많이 사용되었는지를 나타낸다. 즉, 자주 사용되는 일반적인 용어들을 오른쪽 상단에 표시하고 간혹 사용되는 용어들은 오른쪽 하단에 표시한다. Scattertext는 클라이언트 측에서 점에 레이블을 부여하며, 사용자가 해당 점 위로 마우스를 올리면 해당 용어에 대한 정보를 표시한다. 이를 통해 텍스트 데이터의 패턴과 특징을 시각적으로 파악할 수 있다(Kessler, 2020).

. F1-Score와 Scaled F1-Score

데이터 분석의 분류모델 평가지표로 정확도가 있지만 데이터 세트 내에 클래스가 균일하게 분포되어있는 경우에 유용한 평가지표이다. 즉 클래스가 불균형한 데이터 세트에는 좋지 않은 평가지표이기 때문에 혼동행렬(Confusion matrix)에서 정밀도(Precision)와 재현율(Recall)을 조화평균으로 구한 값을 F1-Score라 하며 점수가 높을수록 모델의 분류 성능이 뛰어나다는 것을 의미한다.

$$F1-Score = \frac{1}{\frac{1}{\text{정밀도}(Precision)} + \frac{1}{\text{재현율}(Recall)}}$$

텍스트 데이터는 카테고리 간의 불균형과 단어의 빈도 차이로 인해 F1-Score를 사용하기 어렵다. 이를 조정하여 Scattertext에서는 텍스트 데이터의 특성을 고려하여 Scaled F1-Score 지표를 사용하는데 카테고리 간의 차이와 단어의 중요성을 정확하게 이해하는 데 도움을 준다.

2.5 이전연구와의 구별되는 특징

기존 기록관리연구에 대한 동향 분석은 학술지나 연구자 동향의 분석이 주를 이뤘으며, 논문의 초록과 키워드를 SNA(Social Network Analysis), LDA(Latent Dirichlet Allocation) 등으로 분석하였다. 데이터의 선형 구조나 확률적관계를 다루는 SNA, LDA과는 달리 과거와 현재의 주요 단어 등을 동시에 비선형 구조와 군집상태를 직관적으로 시각적으로 나타내는 t-SNE분석과 고차원적인 해석이 가능한 Scattertext Plot을 사용하였다.

본 연구에서는 텍스트 마이닝 분석방법을 활용하여 국내 기록관리연구의 20년간 연구 동향과 특징을 제시하고, 비정형 텍스트 데이터인 국문 초록 전체를 사용하여 기록관리 분야 키워드 현황을 살펴보았다. 전처리 과정 중 형태소 분석은 Phython-Konlpy 중 Kkma를 활용하였다. 이를 통하여 국내 기록관리 연구에 사용된 키워드 Top100을 도출하였으며 Word2vec 기법을 활용하여, 국문 초록의 주요키워드를 벡터로 나타냈다. Word2vec은 대용량의 텍스트 말뭉치에서 단어 간의 연관성을 학습하기 위해 신경망 모델을 사용하는 데 각각의 고유한 단어를 특정 숫자 목록인 벡터로 나타낸 후 수학함수(코사인 유사도)를 사용하여 단어 간의 의미적 유사도 수준을 실수로 표현한다. 그 결과를 t-SNE 플롯 기법을 통해 초록의 키워드 간 연구적 연결성을 시각적으로 제시하였으며, 최종적으로 Scattertext를 통해 기록관리연구의 20년을 10년 단위로 구분하여 초록의 키워드 빈도값을 시각화하였다. 해당 연구는 한국학술지인용색인(KCI)에서 복합학(대분류) - 문헌정보학(중분류)의 7종의 기록관리학 국내 학술지

https://doi.org/10.14404/JKSARM.2023.23.4.073

를 주제로 하는 논문들의 국문 초록 키워드 데이터를 활용하여 기록관리 연구의 변화와 흐름을 한눈에 파악하는 것이 목적이다.

3. 자료 수집 및 연구방법

국내 기록관리 동향을 파악하기 위한 연구 목적을 달성하기 위해 비정형 텍스트 데이터인 국문 초록을 수집하였다. 이후 각 초록에서 도출되는 키워드 및 단어의 연관성을 확인하기 위해 Word2vec 알고리즘을 활용하였다. 이어서, 시각적인 측면에서 각 단어 간의 의미적 거리를 도출하기 위해 t-SNE 알고리즘을 사용하였다. t-SNE는 단어 시각화에 사용되며, 고차원 데이터를 저차원으로 축소하여 각 단어 간의 의미적 거리를 표현하는 것이다. 마지막으로, Scattertext plot을 활용하여 과거와 현재의 연구에서 자주 사용되는 키워드를 한눈에 알아볼 수 있도록 하였다.

3.1 자료수집

국내 기록관리연구를 대표 할 수 있는 학술지를 선정하기 위해 한국학술지인용색인(Korea Citation Index, KCI)에서 2023년 3월 22일 기준으로 대분류(복합학), 중분류(문헌정보학)으로 검색된 학술지(28종) 중 등재지 7종 즉, 한국도서관.정보학회지, 한국기록관리학회지, 정보관리학회지, 한국문헌정보학회지, 기록학연구, 한국비블리아학회지, 기록과 정보 문화 연구를 선정하였다. KCI에서 제공하는 2004년부터 2023년(총 20년간)까지 해당 논문들에 대한 학술지명, 발행년월, 논문명, 키워드, 국문초록 등의 자료를 수집하였다. 그 중 기록관리학 관련 논문만을 추출하기 위하여 논문명과 키워드에 “기록”, “아카이브”, “아카이빙”, “행정정보”, “정보공개”와 같은 단어가 포함된 논문을 추출하였다. 정확도를 높이기 위해서 “문헌정보”, “도서관”, “사서”, “독서”, “서지”, “책”, “어린이” 키워드는 제외하도록 추출하였다. 총 6,197편의 논문 중 1,157편의 논문들을 활용하여 word2vec 작업후 t-SNE 플롯과 Scattertext 분석을 수행하였다.

3.2 연구방법

기록관리연구 분야의 최빈도 키워드를 추출하고 분석하기 위해 사용한 주요 패키지로는 Pandas, Numpy를 사용하였으며 Word2vec 모델을 사용하기 위하여 Gensim, 시각화를 위하여 Matplotlib, Seaborn를 사용하였다. 데이터 전처리를 위하여 Python의 SpaCy[4] 모듈과 Kkma를 사용하였다. Spacy는 머신러닝(Machine Learning) 기반의 자연어 처리 라이브러리로 기록관리 분야 연구의 국문초록의 키워드별 빈도수, 연도별로 키워드 빈도를 분석할 수 있었다. 자연어 처리를 위하여 형태소 분석기 자체를 개조하여 SpaCy와 Kkma를 병행하여 사용하였다. 데이터 정제 작업은 Kkma는 명사를 추출 한 뒤 ‘것’, ‘수’ 등 분석에 필요 없는 불용어를 삭제하는 과정을 반복적으로 진행한 후 Scatter Text Plot을 제작하였다. 또한 ‘등’, ‘이’ 와 같은 분석에 불필요한 요소들을 삭제하여 최대한 단어 집단 간 동질성을 높이고 이질성을 줄이도록 하였다.

t-SNE와 Scattertext의 시각화 방법을 함께 사용한 이유는 다양한 시각화 결과를 제시하고자 했으며, 두 방법이 서로 보완적인 분석을 가능하게 해주기 때문이다. t-SNE는 주로 단어 간의 의미적 유사성을 시각화하는 데 유용하며, 주로 단어 간의 유사성을 강조한다. 반면 Scattertext는 주로 단어와 카테고리 간의 관계를 시각화하는 데 특화되어 있다. 따라서 두 가지 방법을 함께 사용하면 텍스트 데이터를 다양하게 탐색할 수 있다. 또한, t-SNE로 시각화한 결과를 통해 단어 간의 군집 구조와 유사성을 이해한 후, Scattertext를 사용하여 이러한 단어의 카테고리 간 역할을 확인할 수 있다. 다시 말해, t-SNE 시각화는 데이터의 전반적인 구조를 보여주고, Scattertext 시각화는

4) Industrial-strength Natural Language Processing in Python, 10.5281/zenodo.1212303, https://github.com/explosion/spaCy/blob/master/CITATION.cff

https://doi.org/10.14404/JKSARM.2023.23.4.073

단어의 역할과 중요도를 더 자세히 분석하는 데 도움이 된다. 마지막으로 두 시각화 방법은 서로 보완적이며, 데이터를 다양한 관점에서 분석하는 데 도움이 된다.

〈표 1〉 연구방법 요약

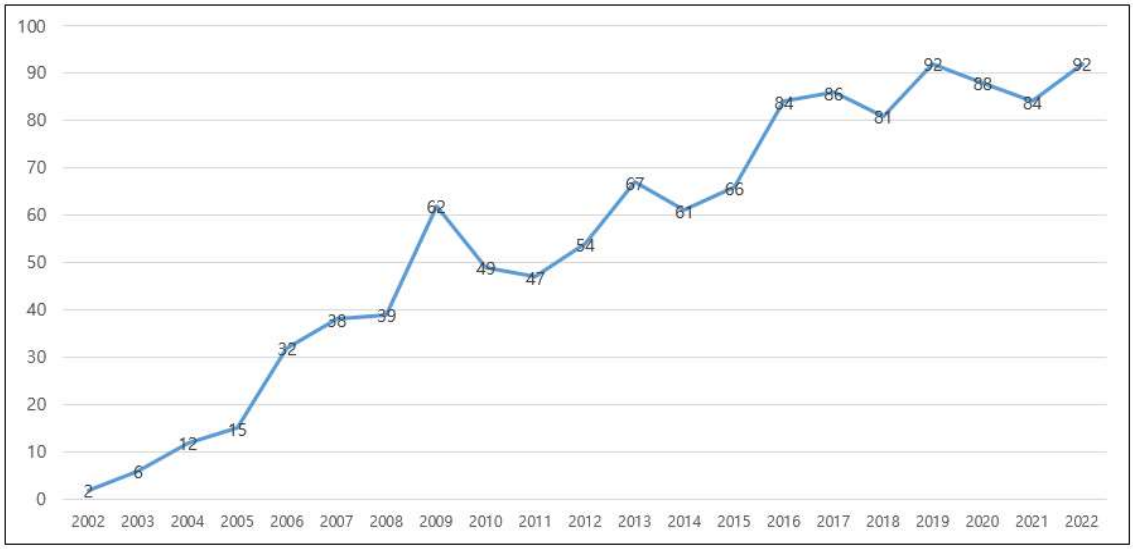
□ 데이터 정의
○ (데이터 수량) 총 6,197개 중 필수·제외키워드, 전처리 작업 등 진행 결과 1,157개 사용
- 필수키워드: 기록, 아카이브, 아카이빙, 아키비스트, 행정정보, 정보공개
※ 기록학, 기록관리라는 키워드는 기록과 같이 나오는 것을 발견하여 필수키워드에 포함
- 제외키워드: 문헌정보, 도서관, 사서, 독서, 서지, 책, 어린이
○ (데이터 속성) 고유명사 출현빈도가 매우 높기에 사전 튜닝 작업이 필요
○ (사용 변수) 학술지명, 발행년월, 논문명, 키워드, 초록
□ 분석 요구 사항
○ 주어진 키워드를 바탕으로 필수값 외 키워드 제거 후 데이터 셋 제작
- 만약 키워드 내에서 존재하지 않는다면 논문명 활용(반영)
○ “초록” 컬럼을 활용하여 분석 진행(반영)
□ 분석 과정
○ (데이터 추출) 전체 데이터에서 사용 변수 추출
○ (데이터 전처리Ⅰ) 형태소 분석기를 활용 명사 추출
○ (데이터 전처리Ⅱ) 연구분야인 ‘기록학’은 고유명사가 많이 등장하기에 형태소 분석기 자체를 개조하여 고유명사 등록 후 명사 재추출
○ (데이터 전처리Ⅲ) ‘등’, ‘이’와 같은 분석에 불필요한 요소 삭제를 통하여 최대한 단어 집단 간 동질성을 높이고 이질성을 줄임
○ (t-SNE) Word2vec 모델을 활용하여 단어를 실수로 표현 후, 시각화
○ (Scattertext) Scattertext module과 형태소 분석기를 결합하여 한국어 맞춤형 Scattertext plot 제작

4. 연구결과

4.1 조사표본과 특성

조사대상 학술논문의 수는 아래의 <그림 1>과 같이 2002년부터 2023년 3월까지 총 1,157편의 한국기록관리학회지, 기록학연구 등에 게재된 논문을 본 연구에서 사용하였으며 기록관리에 관한 연구가 매년 꾸준히 증가하는 것으로 나타났다.

https://doi.org/10.14404/JKSARM.2023.23.4.073



〈그림 1〉 연도별 논문 수

<그림 2>와 같이 분석에 활용된 학술지별 비중을 살펴보면 「한국기록관리학회지」는 441편(38%)로 가장 높은 비중을 차지하였다. 다음으로 「기록학연구」는 388건(33%), 뒤를 이어 「한국비블리아학회지」 92편(7.9%)를 차지하였다.

<표 2>에서는 분석된 학술지의 영향력 지수와 중심성 지수 등을 나타내었다. KCI 인용지수 설명에 따르는 영향력 지수(Impact Factor)[5]는 특정기간동안 한 학술지에 수록된 하나의 논문이 다른 논문에 인용된 평균횟수로 동일분야 저널의 상대적 중요성을 비교 평가하는 방법이다. 중심성지수(SJR)는 엘스비어의 Scopus에서 개발되어 해당 데이터베이스에 근거하여 평가하는 방법이다. IF는 피인용 횟수라는 한 가지 요인만을 고려하는 반면, SJR은 피인용 횟수 외에 인용하는 학술지의 수와 명성도, 그리고 네트워크상의 유기적 인용관계를 고려한다.

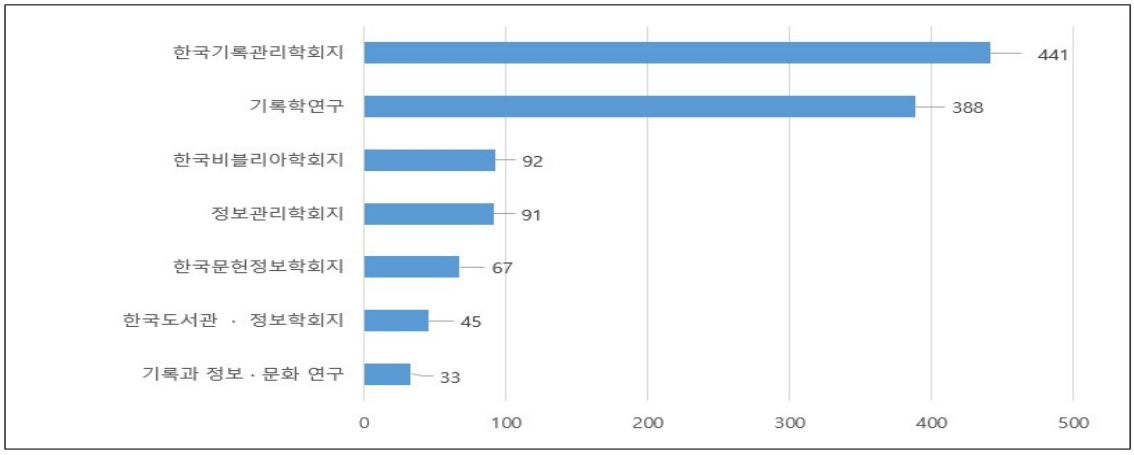
복합학(대분류)-문헌정보학(중분류) 분야에서 분석된 학술지 중 영향력지수(IF)가 높은 학술지는 1.11로 「한국도서관.정보학회지」였으며 「한국기록관리학회지」는 1.05로 두 번째로 나타났으며, 「정보관리학회지」는 0.91로 세 번째로 나타났다. 이 결과, 「한국도서관.정보학회지」가 기록관리연구분야에서 상대적으로 가장 높은 인용빈도를 가진 학술지로 나타났으며, 「한국기록관리학회지」와 「정보관리학회지」가 그 뒤를 이었다.

복합학(대분류)-문헌정보학(중분류) 분야에서 분석된 학술지 중 중심성지수(SJR)가 높은 학술지는 1.434로 「기록과 정보.문화 연구」였으며, 「정보관리학회지」는 1.406으로 두 번째로 나타났다. 이 결과, 「기록과 정보.문화 연구」가 기록관리연구분야에서 중요한 위치에 있으며, 「정보관리학회지」가 그 뒤를 이었다. 최근 5년간 피인용된 학술지 중 「한국도서관.정보학회지」가 374회로 인용되었다는 것은 다른 연구자들에게 의해 많이 활용되고 있으며 두 번째로는 「한국문헌정보학회지」가 264회 인용되었으므로 상당히 많이 활용되고 있는 것으로 알 수 있다.

KCI 인용지수 설명에 따르면 학술지가 좋은 평가를 받기 위한 조건은 IF와 중심성지수로 두 가지는 차이가 있다. 중심성지수(SJR)는 세계 양대 학술지 인용 데이터베이스라고 할 수 있는 클래리베이트의 Wos와 엘스비어의 Scopus에서 개발되어 해당 데이터베이스에 근거하여 발표한다. 분석된 학술지 중 「기록과 정보.문화 연구」는 1.434로 복합학(대분류)-문헌정보학(중분류) 분야에서 높은 중심성지수를 보였으며, 「정보관리학회지」는 1.406으로 두 번째로 나타났다. 최근 5년간 피인용된 학술지 중 「한국도서관.정보학회지」가 749회로 가장 많았으며, 두 번째로는 「한국문헌정보학회지」가 634회이다.

5) 영향력지수(IF)=학술지의 논문이 인용된 총 횟수/학술지에 수록된 논문의 수(www.kci.go.kr)

https://doi.org/10.14404/JKSARM.2023.23.4.073



〈그림 2〉 분석에 사용된 학술지별 순

〈표 2〉 분석학술지

자료출처 : KCI(검색일 : 2023년 4월 10일, 영향력지수 기준으로 정렬)

학술지명	발행기관	영향력 지수 (5년 IF)	중심성 지수 (3년치 기준)	자기인용 비율(%) (2년 KCI IF)	총 논문수 (5년)	총 피인용 회수(5년)
한국도서관·정보학회지	한국도서관·정보학회	1.11	1.294	19.28	326	749
한국기록관리학회지	한국기록관리학회	1.05	1.056	23.96	187	384
정보관리학회지	한국정보관리학회	0.91	1.406	14.95	239	514
한국문헌정보학회지	한국문헌정보학회	0.89	1.211	13.24	290	634
기록학연구	한국기록학회	0.81	1.085	35.29	159	434
한국비블리아학회지	한국비블리아학회	0.58	0.665	8.89	259	453
기록과 정보·문화 연구	한국외국어대학교 정보·기록학연구소	-	1.434	30	31	34

4.2 키워드 분석결과

본 연구의 대상인 1,157편의 논문의 저자들이 초록에 제시한 키워드는 총 77,578개로 이들의 빈도를 분석한 결과는 <표 3>이다. 77,578개의 키워드 중 약어나 괄호 등으로 분류가 다르게 된 동일한 키워드들을 재분류한 결과, 중복을 제외한 키워드는 총 9,774개로 나타났다. 이 중 기록관리 연구자들에 의해 가장 활발하게 연구되고 있는 분야를 알아보기 위해 상위 50위까지 키워드 빈도를 조사하였다. “기록관리”키워드를 889회로 가장 많이 사용하였고 그 다음으로 “분석”, “아카이브”, “기록물”, “활용”, “수집”, “서비스”의 키워드를 연구에 가장 많이 사용하고 있는 것으로 나타났다.

https://doi.org/10.14404/JKSARM.2023.23.4.073

〈표 3〉 키워드 빈도값(TOP 50)

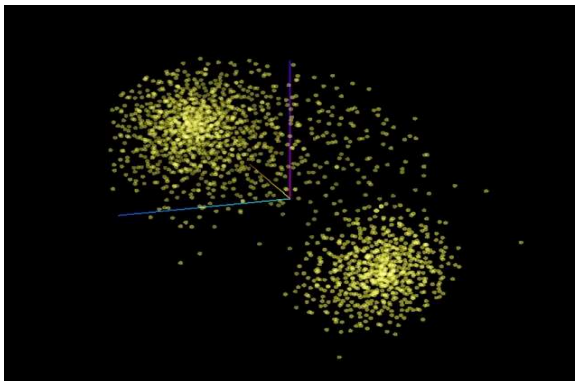
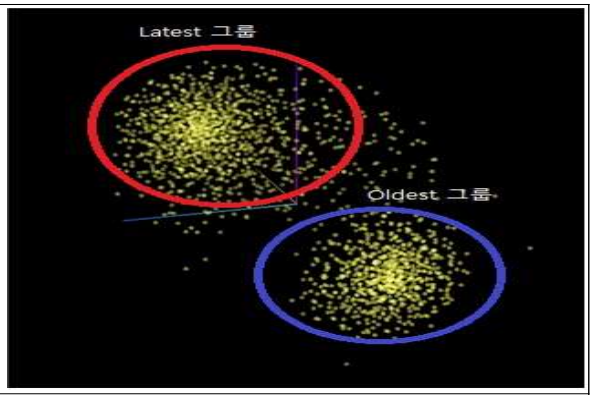
순위	키워드(빈도수)	순위	키워드(빈도수)
1	기록관리(889)	26	적용(235)
2	분석(888)	27	기관(230)
3	아카이브(742)	28	내용(224)
4	기록물(562)	29	이용(223)
5	활용(449)	30	운영(205)
6	수집(389)	31	기록화(198)
7	서비스(376)	32	아카이빙(195)
8	보존(368)	33	메타데이터(194)
9	생산(366)	34	역할(193)
10	방안(362)	35	데이터(186)
11	평가(358)	36	전자기록(181)
12	이용자(331)	37	개념(180)
13	기록관(330)	38	가능(178)
14	목적(330)	39	분야(177)
15	설계(325)	40	개선(177)
16	업무(322)	41	콘텐츠(171)
17	정보(320)	42	국가기록원(169)
18	기술(316)	43	검색(169)
19	개발(309)	44	표준(169)
20	디지털(305)	45	교육(166)
21	가치(270)	46	방향(166)
22	기능(265)	47	우리나라(163)
23	국내(255)	48	환경(158)
24	수행(255)	49	자료(155)
25	제공(249)	50	변화(155)

4.3 word2vec 및 t-SNE 분석 결과

<표 4>는 word2vec을 통해 초록에 나타난 모든 키워드 중 전처리가 된 9,774개의 키워드를 100차원 벡터 공간에 임베딩하여 t-SNE 플롯을 통해 2차원 데이터로 차원 축소하여 시각화하였다. t-SNE를 적용하기 전에 먼저 고차원 데이터에서 각 데이터 간의 거리를 계산하여 최근접 단어를 구하였다. 본 연구에서는 두 벡터 간 각도를 측정하여 유사도를 산출하는 코사인 유사도(cosine similarity)를 활용하여 단어 간 거리를 구하여 최근접 단어를 구하였다. 이를 반복하여 최근접 단어 거리를 구할 수 있었다. 이후 기록관리연구 키워드 간의 연구적 연결성을 시각적으로 나타낼 수 나타냈다. 분석결과 아래의 <표 4>와 같이 기록관리 연구의 주요키워드는 두 군집으로 나누어져 나타났다.

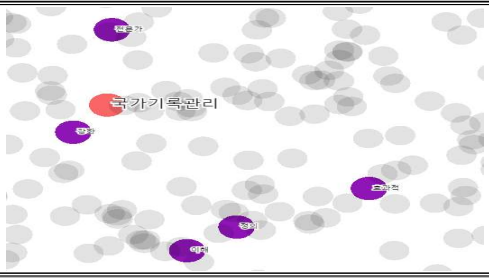
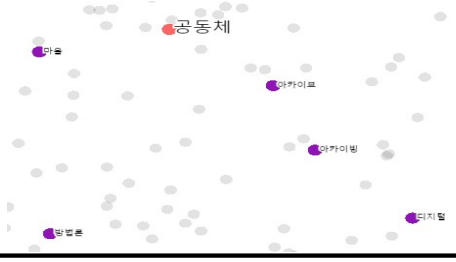
https://doi.org/10.14404/JKSARM.2023.23.4.073

〈표 4〉 t-SNE 분석 시각화 장면

	
100차원 벡터 표현	2개 그룹으로 구분

예를 들어 첫 번째 군집은 연구자가 Oldest그룹으로 명하였는데 이 그룹의 많은 키워드 중 “국가기록관리”키워드를 예로 들면 “효과적”, “정의”, “전문가”, “강화”, “이해” 등이 벡터 공간에서 최근접 하고 있어 서로 간에 연관성을 가지는 것으로 나타났다. 두 번째 군집은 Latest 그룹으로 연구자가 명하였는데 이 그룹의 많은 키워드 중 “공동체”의 키워드는 “아카이브”, “마을”, “아카이빙”, “방법론”, “디지털”등이 벡터 공간에서 최근접 되어 있기 때문에 서로 간에 연관성을 가지는 것으로 나타났다.

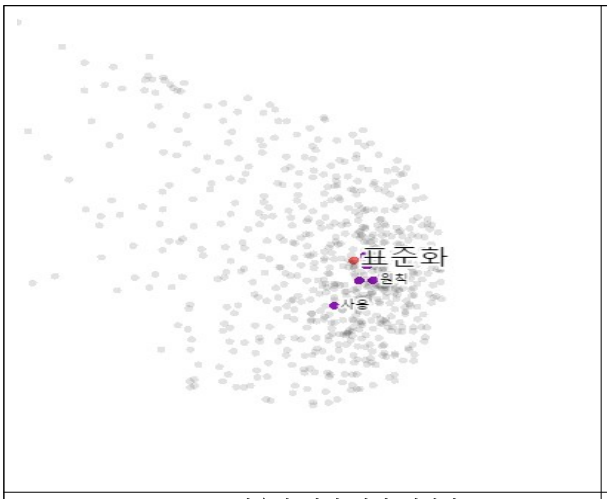
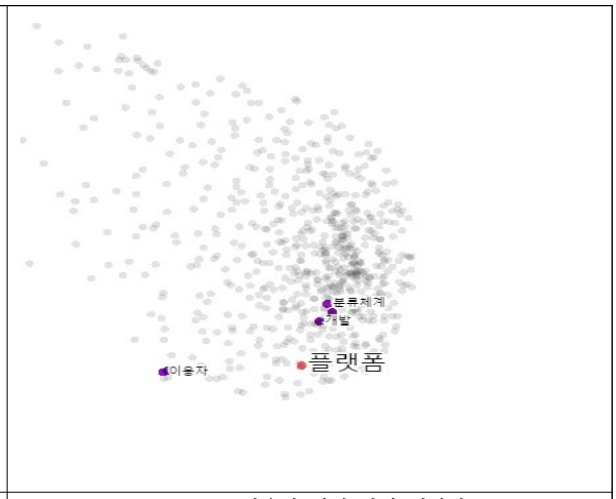
〈표 5〉 t-SNE 분석 그룹별 예시

Oldest 그룹 예시	최근접 단어	
	효과적	0.001
	정의	0.001
	전문가	0.001
	강화	0.001
	이해	0.001
Latest 그룹 예시	**최근접 단어**	
	아카이브	0.001
	마을	0.002
	아카이빙	0.002
	방법론	0.002
	디지털	0.002

<표 6>에서 확인할 수 있듯이, Oldest 그룹에서 자주 등장한 단어인 ‘표준화’의 주변 단어 상위 5개는 ‘기준’, ‘프로세스’, ‘원칙’, ‘사용’, ‘결과’이며 Latest 그룹에서 자주 등장한 ‘플랫폼’의 주변 단어 상위 5개는 ‘개발’, ‘이용자’, ‘제공’, ‘방법’, ‘분류체계’이다. 이러한 결과는 시간의 변화에 따라 사용된 주요키워드가 다르다는 것으로 알 수 있다. t-SNE 분석결과는 키워드간 거리측정 학습장면과 시각화 구현의 영상이 중요하기 때문에 동영상 파일을 연구자 블로그에 업로드를 해두었다.(https://blog.naver.com/igre/223184802479)

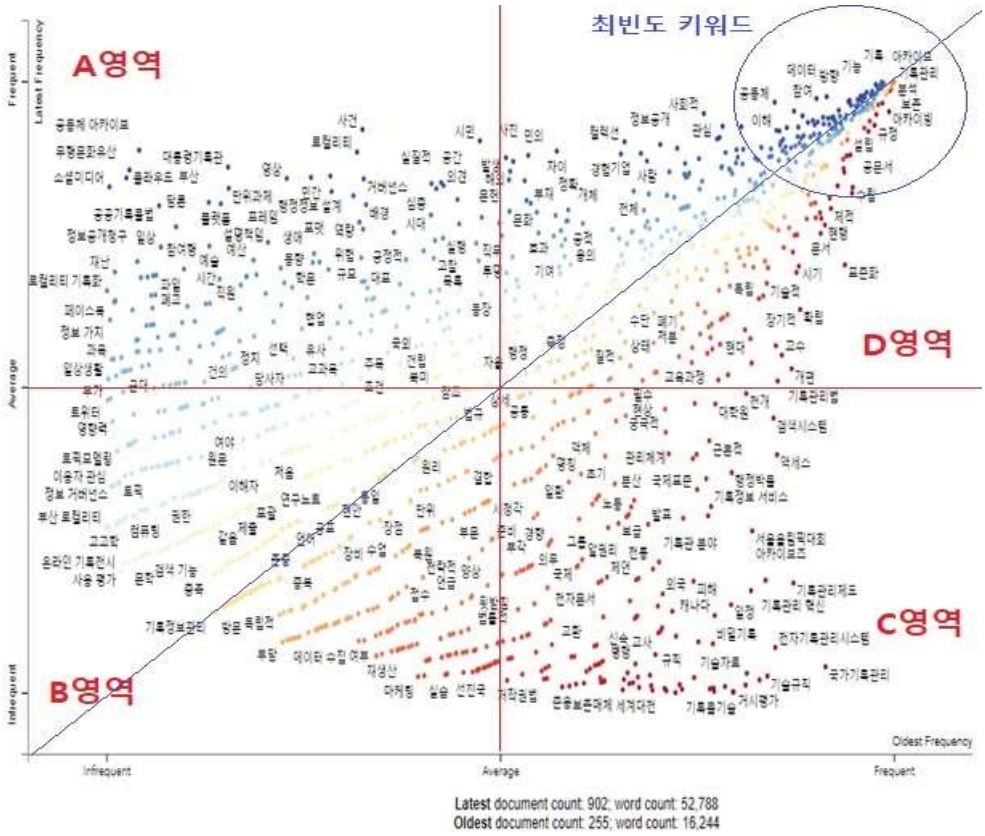
https://doi.org/10.14404/JKSARM.2023.23.4.073

〈표 6〉 그룹별 빈도가 높은 단어의 연관키워드

	
t-SNE를 활용한 과거 단어 시각화	t-SNE를 활용한 최신 단어 시각화

4.4 Scattertext 분석결과

본 연구의 분석에 사용된 1,157편의 기록관리 동향 연구 논문 중 2002년 ~ 2011년에 작성된 논문은 255편의 16,244개의 키워드였으며, 2012년 ~ 2023년에 작성된 논문은 902편의 52,788개의 키워드였다. ScatterText 분석 결과는 <그림 3>와 같다.



〈그림 3〉 Scattertext 분석의 세부 구분

https://doi.org/10.14404/JKSARM.2023.23.4.073

Scattertext 분석 결과를 설명하기 쉽게 연구자가 대각선으로 나누었고 세부적으로 A, B, C, D 총 4개 영역으로 나누었다. 첫 번째 그룹은 Y축의 윗부분(A영역)으로 2012-2023년(현재)에 작성된 논문들의 최빈도 키워드이며 과거에는 등장하지 않은 키워드들의 영역이다. Y축의 아랫부분(B영역)은 현재와 과거에 빈도가 낮게 등장하는 키워드들의 영역이다. 두 번째 그룹은 X축의 아랫부분(C영역)으로 2002-2011년(과거)에 작성된 논문들의 최빈도 키워드이며 현재에는 등장하지 않은 키워드들의 영역이다. X축의 윗부분(D영역)은 기록관리 연구에서 현재와 과거에 계속 사용되는 하는 최빈도 키워드들의 그룹이다.

Y축의 A영역에서 최빈도 키워드는 “플랫폼”, “프레임”, “클라우드”, “정보공개”, “단위과제” 등을 확인 할 수 있고 B영역에서는 최빈도 키워드는 “부산 로컬리티”, “트위터”, “정보 거버넌스”등이 있다. X축의 C영역에서 최빈도 키워드는 “아카이빙”, “기술자료”, “표준화”, “공문서”, “수립”, “기록관리제도”, “기술규칙”, “거시평가”, “건축문화재”, “기록물관리법” 등의 단어들을 확인 할 수 있으며 D영역에서는 최빈도 키워드는 “공동체 아카이브”, “국가기록원”, “데이터”, “사회적”, “ 사건”, “기록정보서비스”, “온라인”, “시민”, “참여”, “디지털 아카이브” 등을 확인 할 수 있다. 이 결과는 html으로 작성된 웹페이지에서 구현되도록 만들었다.(https://blog.naver.com/igre/223184802479)

. 첫 번째 그룹(대각선 위)에서 자주 사용된 키워드

<표 7>에서 Scaled F1-score 기준으로 의미 있는 단어들을 12개 선별하여 각 단어의 2002년 - 2011년의 빈도수와 2012년 - 2023년의 빈도수를 비교하였다.

〈표 7〉 첫 번째 그룹(C영역)의 키워드 빈도분석

	Words	연 도			Words	연 도	
		2002-2011	2012-2023			2002-2011	2012-2023
1	아카이빙	186	65	7	기술규칙	29	3
2	국가기록관리	45	3	8	거시평가	22	0
3	표준화	59	15	9	건축문화재	22	0
4	공문서	67	24	10	기록물관리법	27	2
5	수립	70	25	11	전자기록 관리시스템	29	4
6	기록관리제도	33	5	12	규정	108	49

첫 번째 그룹(2002년 - 2011년, 과거)에서 자주 사용된 키워드를 분석한 결과, 해당 기간 동안 국내의 기록관리 제도와 법령이 새로 제정되었고, 공공기관의 새로운 기록관리체계가 구축되는 과정에 대한 연구 및 표준화와 기술규칙을 포함한 기록관리체계 수립과 관련된 연구가 활발하게 진행되었음을 확인할 수 있었다. “아카이빙” 키워드는 주로 기관의 장기 보존 시스템, 주제, 보존, 메타데이터 등 다양한 관련 주제에 대한 연구가 이루어진 것으로 파악되며, “국가기록관리” 키워드는 대한민국의 기록관리 인프라와 시스템, 정책, 발전 등 다양한 주제로 연구가 진행되었다고 나타났다. 또한, “전자기록관리시스템” 키워드는 전자기록물 검색 기반 환경, 관리 절차 등과 관련한 다양한 주제의 연구가 진행되었다. <표 7>에 나타난 각 단어들은 2002년 - 2011년의 초록 데이터에서 많이 사용되었지만 2012년 - 2023년에는 사용 빈도가 줄어드는 경향을 보이고 있다.

https://doi.org/10.14404/JKSARM.2023.23.4.073

. **두 번째 그룹(대각선 아래)에서 자주 사용된 키워드**

첫 번째 그룹(2002년 - 2011년, 과거)과 비교하여 두 번째 그룹(2012년 - 2023년, 현재)에서 자주 등장하는 키워드를 Scaled F1-score 값에 기반하여 10개의 단어로 <표 8>에 나타냈다. “공동체” 키워드의 경우, 민간 공동체, 마을 공동체 및 지역 로컬리티 연구에서 미래에 정체성을 전승하기 위한 시민들과 공공기관 간의 관계 형성에 대한 관심이 증가하고 있는 것으로 나타났다. 또한, “국가기록원” 키워드는 국가의 기록관리 제도를 운영하는 중요한 국가기관으로, 책임과 성과에 관한 연구가 활발히 이루어지고 있는 것으로 분석되었다. “데이터” 키워드의 경우, 최근에는 전자기록물 뿐만 아니라 해당 레코드의 메타데이터 관리 및 공공기관의 데이터 세트 관리 등이 기록관리의 중요한 부분으로 포함되었다(시행령 제34조의 3, 2020년 제정). 이에 따라, 이러한 데이터들을 효과적으로 보존하고 관리하기 위한 연구가 활발히 진행되고 있다고 볼 수 있다. “시민”과 “참여” 키워드는 공공기관의 기록물관리 제도에 많은 시민 참여를 허용하기 위한 민간 기록물 수집과 관리 방법에 관한 연구가 진행 중이며, “사회적”과 “콘텐츠” 키워드는 기관이 보유한 기록물을 역사적 및 문화적 가치에 따라 교육 콘텐츠로 활용하기 위한 연구가 활발히 진행되고 있음을 나타냈다. 요약하면, 새로운 기술과 형태의 기록물을 효율적으로 관리하고 시민들과 공유하기 위한 디지털 아카이빙 방법과 관련한 연구가 주로 진행되고 있음을 확인할 수 있었다.

〈표 8〉 두 번째 그룹(A영역)의 키워드 빈도분석

	Words	연 도			Words	연 도	
		2002-2011	2012-2023			2002-2011	2012-2023
1	공동체	29	81	7	공동체아카이브	0	34
2	국가기록원	43	92	8	온라인	22	48
3	데이터	47	101	9	시민	9	37
4	사회적	20	62	10	참여	37	59
5	사건	6	47	11	디지털아카이브	25	48
6	기록정보서비스	35	65	12	콘텐츠	67	86

5. 결론 및 시사점

본 연구는 최근 20년간의 기록관리 연구의 키워드들을 텍스트 마이닝 기법과 데이터 시각화를 통하여 제시하고자 하였다. 기존의 선행연구는 질적연구와 정형 데이터를 위주로 기록관리동향에 관한 연구가 진행되었으나 본 연구에서는 빅데이터 분석 기술 중 하나인 텍스트 마이닝 기법을 활용하여 국내 기록관리연구에서의 비정형 텍스트 데이터인 국문초록을 분석하여 시각화하고자 하였다. 한국학술지인용색인(Korea Citation Index)에서 복합학(대분류)에서 문헌정보학(중분류)의 학술지 7종을 분석대상으로 하여 2002년부터 2023년까지 기록관리연구를 연구한 국내 학술지 1,157편에 대하여 국문초록의 키워드 77,578개를 추출하여 특정 문맥에서 단어가 얼마나 자주 등장하는지, 각 단어를 고차원 벡터로 표현하는 Word2vec를 사용하였다.

분석 결과는 다음과 같다. 첫째, 기록관리 연구는 2002년부터 2023년까지 1,157편으로 계속 증가하였으며, 이 중에서 한국기록관리학회에서 발행한 『한국기록관리학회지』가 가장 많은 441편으로 나타났다. 이어서 발행된 『기록학 연구』가 388편으로 그 뒤를 이었다. 총 7개의 학회 중에서 최근 5년간 영향 인자(IF)가 가장 높은 저널은

https://doi.org/10.14404/JKSARM.2023.23.4.073

『한국도서관·정보학회지』(1.11), 『한국기록관리학회지』(1.05), 『정보관리학회지』(0.91)이다. 둘째, 1,157편의 논문에서 77,578개의 키워드에 대한 빈도 분석 결과, "기록관리" (889회), "분석" (888회), "아카이브" (742회), "기록" (562회), "활용" (449회), "수집" (389회), "서비스" (376회), "보존" (368회), "생산" (366회), "측정" (362회)이 가장 빈도가 높았다. 셋째, 기록관리 연구의 키워드의 빈도를 분석하여 각 단어를 고차원 벡터로 표현한 뒤 t-SNE 플롯을 통해 알아보았는데 시간이 지남에 따라 확연하게 두 개의 그룹으로 나누어지는 것을 알 수 있었다. 또한 Scattertext분석을 시행하여 시간에 따라 사용되는 키워드의 빈도가 크게 두 개의 그룹으로 나누어지는 것을 알 수 있었다.

첫 번째 그룹에는 과거의 논문들의 최빈도 키워드들의 현황이 분포되고 있는데 기록관리제도와 법령이 제정됨에 따라 공공기관의 새로운 제도변화에 대한 연구나 표준화 또는 기술규칙 등 기관의 기록관리체계를 수립하는 연구 등이 다양하게 실시되었다. “아카이빙”, “국가기록관리”, “전자기록관리시스템”등의 키워드들의 빈도가 높았다. 또한 첫 번째 그룹(과거)의 키워드들은 두 번째 그룹(최신)에서 줄어드는 양상을 보였다. 두 번째 그룹은 최근 현재의 논문들의 최빈도 키워드의 현황이 분포되고 있는데 이를 살펴보면 “공동체”, “사회적”, “참여”의 키워드가 자주 등장하였다. 최근 현대에 들어 아카이브의 최종 목적인 시민들과 공공기관간의 관계 형성과 정체성을 미래에 전승하기 위한 연구나 마을공동체 및 지역 로컬리티 연구 등의 관심이 증가하고 있는 것으로 보인다. 또한 “데이터”, “디지털아카이브”, “콘텐츠”의 키워드는 최근 전자기록물의 메타데이터의 관리방안과 행정정보시스템의 데이터세트까지 기록관리 대상에 포함(시행령 제34조의 3, 2020년 신설)되었기 때문에 효율적인 보존과 관리 등에 관한 연구가 활발히 진행 중인 것을 파악할 수 있다. 두 번째 그룹의 최빈도 키워드들은 첫 번째 그룹에서도 등장은 하지만 빈도수가 적은 것으로 나타났다.

본 연구의 시사점은 다음과 같다. 첫째, 기록관리 연구 동향에 텍스트 마이닝, 키워드 추출, 용어간 거리측정, 시각화 등 다양한 데이터 분석기술을 활용하여 추세변화를 확인 할 수 있었다. 신기술을 적극 수용하여 다양한 유형의 기록을 후대에 남길 수 있도록 효율적인 방법을 시민들과 공유해야 한다. 둘째, 빅데이터 분석기술을 통해 기록관리 연구 분야의 인사이트를 도출하기 위해 기관별 메타데이터를 공유하여 기록관리 분류 자동화, 검색지원도구, 컨텐츠 개발 등 다양한 영역으로 발전시킬 수 있는 참고 사례로 활용 될 수 있다. 셋째, 국내외 연구동향을 비교하고 국외와 협력하는 것이 필요하다. 국제 학술 교류를 통해 해외 기록관리 연구 동향 데이터를 수집하여 다양한 관점과 아이디어를 수렴하고 국내외 표준 및 규제와의 조화를 이끌어내야 한다. 이 연구는 특정 주제의 변화에 대한 논의는 하지 않았으며, 주로 키워드의 변화를 중점으로 분석하였다. 최근 개발되고 있는 다양한 데이터 분석 기술을 활용하여 기록관리 분야의 주요 키워드가 시간에 따라 어떻게 연구되고 있는지 분석하는 연구는 향후 기록관리 연구의 중요한 주제가 될 것이다.

참고문헌

김규하, 박철용 (2015). 토픽 모형 및 사회연결망 분석을 이용한 한국데이터정보과학회지 영문초록 분석. 한국데이터정보과학회지, 26(1), 151-159. https://doi.org/10.7465/jkdi.2015.26.1.151

김규환, 장보성, 이현정 (2009). 우리나라 기록관리학 분야의 연구영역 분석-논문제목의 구문 및 의미 구조를 중심으로. 한국문헌정보학회지, 43(3), 417-439. https://doi.org/10.4275/KSLIS.2009.43.3.417

김지영, 김은혜, 이지영 (2019). 텍스트 마이닝 및 의미연결망 분석을 활용한 무용경영 연구동향과 지식구조 분석. 한국스포츠산업경영학회지, 24(3), 85-103. https://doi.org/10.31308/KSSM.24.3.6

김판준, 서혜란 (2012). 프로파일링 기법을 이용한 국내 전자기록 분야 지적구조 분석. 한국기록관리학회지, 12(2), 29-50. https://doi.org/10.14404/JKSARM.2012.12.2.029

남태우, 이진영 (2009). 우리나라 기록관리학 연구 동향 분석. 한국도서관.정보학회지, 40(2), 451-472. https://doi.org/10.16981/kliss.40.2.200906.451

박준형, 류법모, 오효정(2018). 시계열 기반 국내 기록관리학 토픽 트렌드 분석. 한국기록관리학회지, 18(1), 29-47. https://doi.org/10.14404/JKSARM.2018.18.1.029

박준형, 오효정 (2017). 국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교 - LDA와 HDP를 중심으로. 한국도서관.정보학회지, 48(4), 235-258. https://doi.org/10.16981/kliss.48.4.201712.235

배규용, 박주현, 김정선, 이영섭 (2013). 텍스트 마이닝 기법을 활용한 기후변화관련 식품분야 논문초록 분석. 한국데이터정보과학회지, 24(6), 1429-1437. https://doi.org/10.7465/jkdi.2013.24.6.1429

손혜인, 남영준 (2016). 기록관리학 분야 국내 학술지의 연구동향에 관한 연구-『한국기록관리학회지』와 『기록학연구』를 중심으로. 정보관리학회지, 33(1), 85-110. https://doi.org/10.3743/KOSIM.2016.33.1.085

윤희영, 곽일엽 (2020). 초록 데이터를 이용한 한국무역연구분야 키워드, 문서간 관계성 모델링. 국제상학, 35(2), 45-64. https://doi.org/10.18104/kaic.2020.35.2.45

이상복 (2019). 텍스트 마이닝 처리로 품질경영학회지 연구동향 분석. 품질경영학회지, 47(3), 597-613. https://doi.org/10.7469/JKSQM.2019.47.3.597

이재윤, 문주영, 김희정 (2007). 텍스트 마이닝을 이용한 국내 기록관리학 분야 지적구조 분석. 한국문헌정보학회지, 41(1), 345-372. https://doi.org/10.4275/KSLIS.2007.41.1.345

정용복, 박의섭 (2015). 텍스트 마이닝을 이용한 암반공학분야 SCI논문의 주제어 분석. 터널과 지하공간, 25(4), 303-319. https://doi.org/10.7474/TUS.2015.25.4.303

조수곤, 김성범 (2012). 텍스트 마이닝을 활용한 산업공학 학술지의 논문 주제어간 연관관계 연구. 대한산업공학회지, 38(1), 67-73. https://doi.org/10.7232/JKIIE.2012.38.1.067

최이랑 (2015). 국내 기록관리학 연구동향에 관한 연구-최근 10년간(2004-2013) 학술논문을 중심으로. 기록학연구, 43, 147-177. https://doi.org/10.20923/kjas.2015.43.147

Ethayarajh, K., Duvenaud, D. & Hirst, G. (2019). Towards Understanding Linear Word Analogies. The Association for Computational Linguistics, 57, 3253-3262. https://doi.org/10.18653/v1/P19-1315

Kessler, J. (2020). Visualizing thousands of phrases with Scattertext, PyTextRank and Phrasemachine. Analytics Vidhya. Available: https://medium.com/analytics-vidhya/visualizing-phrase-prominence-and-category-association-with-scattertext-and-pytextrank-f7a5f036d4d2

Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605.

Opacich, J. (2021). Interpreting Scattertext: A seductive tool for plotting text. Towards Data Science. Available: https://towardsdatascience.com/interpreting-scattertext-a-seductive-tool-for-plotting-text-2e94e5824858

Parra, C., Cebollada, S., Payá, L., Holloway, M. & Reinoso, O. (2020). A Novel Method to Estimate the Position of a Mobile Robot in Underfloor Environments Using RGB-D Point Clouds. IEEE Access, 8, 9084-9101. https://doi.org/10.1109/ACCESS.2020.2964317

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Bae, Kyu-Yong, Park, Ju-Hyun, Kim, JeongSeon & Lee, Yung-Seop (2013). Analysis of the abstracts of research articles in food related to climate change using a text-mining algorithm. Journal of the Korean Data And Information Science Sociaty, 24(6), 1429-1437. https://doi.org/10.7465/jkdi.2013.24.6.1429

Cho, Su-Gon & Kim, Seoung-Bum (2012). Finding Meaningful Pattern of Key Words in IIE Transactions Using

Text Mining. Journal of the Korean Institute of Industrial Engineers, 38(1), 67-73. https://doi.org/10.7232/JKIIE.2012.38.1.067

Choi, Yilang (2015). A Study on the Research Trends of Archival Studies in Korea : Focused on Research Papers between 2004 and 2013. The Korean Journal of Archival Studies, 43, 147-177. https://doi.org/10.20923/kjas.2015.43.147

Jung, Yong-Bok & Park, Eui-Seob (2015). Keyword Analysis of Two SCI Journals on Rock Engineering by using Text Mining. Tunnel and Underground Space, 25(4), 303-319. https://doi.org/10.7474/TUS.2015.25.4.303

Kim, Gyuha & Park, Cheolyong (2015). Analysis of English abstracts in Journal of the Korean Data & Information Science Society using topic models and social network analysis. Journal of the Korean Data And Information Science Sociaty, 26(1), 151-159. https://doi.org/10.7465/jkdi.2015.26.1.151

Kim, Gyuhwan, Jang, BoSeong & Yi, Hyunjung (2009). A Study on Intellectual Structure of Records Management and Archives in Korea: Based on Syntactic and Semantic Structure of Article Titles. Journal of the Korean Society for Library and Information Science, 43(3), 417-439. https://doi.org/10.4275/KSLIS.2009.43.3.417

Kim, Ji Young, Kim, Eun Hye & Lee, Ji Young (2015). A Study on the Research Trends and Knowledge Structure of Dance Management Using Text-Mining and Semantic Network Analysis. Korean Journal of Sport Management, 24(3), 85-103. https://doi.org/10.31308/KSSM.24.3.6

Kim, Pan Jun & Suh, hye-Ran (2012). A Study on the Analysis of Intellectual Structure of Electronic Records Research in Korea Using Profiling. Journal of Korean Society of Archives and Records Management, 12(2), 29-50. https://doi.org/10.14404/JKSARM.2012.12.2.029

Lee, Jae-Yun, Moon, Ju-Young & Kim, Hee-Jung (2007). Examining the Intellectual Structure of Records Management & Archival Science in Korea with Text Mining. Journal of the Korean Society for Library and Information Science, 41(1), 345-372. https://doi.org/10.4275/KSLIS.2007.41.1.345

Nam, TeaWoo & Lee, Jin-Young (2009). A Study on the Research Trends of Records and Archives Management in Korea. Journal of Korean Library and Information Science Society, 40(2), 451-472. https://doi.org/10.16981/kliss.40.2.200906.451

Park, JunHyeong & Oh, Hyo-Jung (2017). Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea: focused on LDA and HDP. Journal of Korean Library and Information Science Society, 48(4), 235-258. https://doi.org/10.16981/kliss.48.4.201712.235

Park, JunHyeong, Ryu, Pum-Mo & Oh Hyo-Jung (2018). Timeline-Based Topic Trend Analysis of Archives Management in Korea. Korean Society of Archives and Records Management, 18(1), 29-47. https://doi.org/10.14404/JKSARM.2018.18.1.029

Ree, Sangbok (2019). Analysis of Research Trends in Journal of Korean Society for Quality Management by Text Mining Processing. Journal of Korean Society for Quality Management, 47(3), 597-613. https://doi.org/10.7469/JKSQM.2019.47.3.597

Sohn, Hye In & Nam, Young Joon (2016). A Study on the Research Trends of Archives Management in Korea: Focused on the Journal of Records Management & Archives Society of Korea and The Korean Journal of Archival Studies. Journal of the Korean Society for Information Management, 33(1), 85-110. https://doi.org/10.3743/KOSIM.2016.33.1.085

Yoon, Hee-Young & Kwak, Il-Youp (2020). The Association Modeling on Keywords and Documents of Korea International Trade Research using Paper Abstract data. Korea International Commerce Review, 35(2), 45-64. https://doi.org/10.18104/kaic.2020.35.2.45