데이터 활용률 제고를 위한 기술 용어의 상호 네트워크 생성과 통제

Generating and Controlling an Interlinking Network of Technical Terms to Enhance Data Utilization

정도헌 (Do-Heon Jeong)*

초 록

빅 데이터 시대에 접어들면서 저장 기술과 처리 기술이 급속도로 발전함에 따라, 과거에는 간과되었던 롱테일(long tail) 데이터가 많은 기업과 연구자들에게 관심의 대상이 되고 있다. 본 연구는 롱테일 법칙의 영역에 존재하는 데이터의 활용률을 높이기 위해 텍스트 마이닝 기반의 기술 용어 네트워크 생성 및 통제 기법을 제안한다. 특히 텍스트 마이닝의 편집 거리(edit distance) 기법을 이용해 학문 분야에서 사용되는 기술 용어의 상호 네트워크를 자동으로 생성하는 효과적인 방안을 제시하였다. 데이터의 활용률 향상 실험을 위한 데이터 수집을 위해 LOD(linked open data) 환경을 이용하였으며, 이 과정에서 효과적으로 LOD 시스템의 데이터를 활용하는 기법과 용어의 패턴 처리 알고리즘을 제안하였다. 마지막으로, 생성된 기술 용어 네트워크의 성능 측정을 통해 제안한 기법이 롱테일 데이터의 활용률 제고에 효과적이었음을 확인하였다.

ABSTRACT

As data management and processing techniques have been developed rapidly in the era of big data, nowadays a lot of business companies and researchers have been interested in long tail data which were ignored in the past. This study proposes methods for generating and controlling a network of technical terms based on text mining technique to enhance data utilization in the distribution of long tail theory. Especially, an edit distance technique of text mining has given us efficient methods to automatically create an interlinking network of technical terms in the scholarly field. We have also used linked open data system to gather experimental data to improve data utilization and proposed effective methods to use data of LOD systems and algorithm to recognize patterns of terms. Finally, the performance evaluation test of the network of technical terms has shown that the proposed methods were useful to enhance the rate of data utilization.

키워드: 롱테일 법칙, 개방형 연결 데이터, 언어 자원, DBLP, 편집 거리 알고리즘
long tail theory, linked open data, language resources, DBLP, edit distance algorithm

* 덕성여자대학교 문헌정보학과 조교수(doheonjeong@duksung.ac.kr)

■ 논문접수일자 : 2018년 2월 18일 ■ 최초심사일자 : 2018년 3월 8일 ■ 게재확정일자 : 2018년 3월 26일
■ 정보관리학회지, 35(1), 157-182, 2018. 〔http://dx.doi.org/10.3743/KOSIM.2018.35.1.157〕

1. 서 론

세계적인 온라인 유통업체인 아마존닷컴의 롱테일(long tail) 비즈니스가 성공하면서 빅 데이터 환경에서 작은 데이터들의 가치에 대해 많은 기업과 연구자들이 더욱 관심을 갖게 되었다. 롱테일 법칙은 80:20의 집중 현상을 설명하는 파레토 법칙을 그래프에 나타냈을 때 꼬리처럼 긴 부분을 형성하는 80%의 부분을 일컫는다. 발생 확률 혹은 발생량이 상대적으로 적은 데이터들의 경우, 과거에는 간과되는 경향이 있었으나 빅 데이터 시대에 접어들면서 저장 기술과 처리 기술이 비약적으로 발전함에 따라 전사적인 관점의 데이터 프로세싱이 가능하게 되었다. 아마존닷컴은 전체 매출의 절반 이상을 랭킹 13만 위 이하의 책에서 올리고 있으며, 전날 팔았던 책들 보다 팔지 않았던 책이 오늘 더 많이 판매된다고 밝히고 있다. 즉, '잘 팔리는 책' 보다 '잘 팔리지 않는 방대한 양의 책'들의 판매량이 더 높다는 사실이 롱테일 법칙을 잘 설명하고 있다(Wikipedia, 2018a).

빅 데이터와 관련한 또 다른 유형의 데이터는 다크 데이터(dark data)이다. 세계적인 비즈니스 컨설팅 그룹인 가트너(Gartner Inc.)는 '현행 비즈니스 수행을 위해 정보를 수집, 가공 및 저장 처리를 하였지만, 분석 등의 다른 목적으로 활용하고 있지 않는 정보 자산'으로 다크 데이터를 정의하고 있다(Gartner, 2018). 아직까지 대부분 기업의 입장에서는 다크 데이터가 필요성이 불분명한 데이터일 뿐이며 저장, 관리 및 보안 조치를 위해 막대한 비용을 지불해야 하는 골치 덩어리일 수 있다. 실제로 글로벌 정보보안 기업인 베리타스가 2016년 발표한 설문조사에 따르면, 많은 세계적 기업들이 자사가 보유한 절반 이상의 데이터가 다크 데이터일 것으로 진단하고 있다고 한다(Veritas, 2016). 이러한 현황들을 고려해 볼 때, 향후 많은 기업들이 다크 데이터를 재분류하고 관리하는 프로세스를 적극적으로 도입하고 활용 방안을 모색할 것으로 예상할 수 있다. 최근에는 인공지능 기반의 데이터 분석회사 래티스(Lattice)를 애플사가 인수하기도 하였는데, 이 신생 회사는 기계학습 등의 기법을 통해 비정형 데이터인 다크 데이터를 활용할 수 있는 가용 데이터로 변환시키는 기술을 개발한 기업이다(Fortune, 2017). 이러한 다크 데이터 처리기술을 통해 애플은 시리(Siri)를 비롯한 인공지능 기술을 더욱 정교하게 발전시킬 것으로 예상할 수 있다. 향후 빅 데이터 프로세싱의 관점이 이러한 롱테일 데이터와 다크 데이터와 같은 숨은 데이터를 발견하고 발굴해 내는 기술로 더욱 집중될 것으로 보인다.

본 연구는 (1) 대량의 데이터를 LOD 시스템으로부터 수집하여 패턴 기반의 정보 추출 기법을 이용해 기술 용어를 추출하여, (2) 롱테일 영역에 존재하는 기술 용어 데이터를 통계적으로 확인하고, (3) 텍스트 마이닝 기법을 이용하여 생성한 용어 네트워크를 이용해 데이터의 활용도를 높이는 것을 궁극적인 목표로 한다.

상세한 연구의 내용은 다음과 같다. 첫째, 비정형 학술 데이터로부터 기술 용어 자원을 추출하여 구축하기 위해 사용한 일련의 기법을 소개한다. 웹으로부터 수집한 대량의 LOD 데이터를 이용해 학술논문의 초록정보를 수집하고 기술 용어로 인식되는 패턴 처리를 통해 데이터를 생성한다. 둘째, 다양한 형태가 존재하여 통제

가 이루어지지 않은 용어 자원 간의 상호 네트워크를 생성함으로써 비정형 용어 데이터의 처리가 가능하도록 하였다. 이를 위해 텍스트 마이닝 기법인 편집 거리(edit distance) 기법을 이용함으로써 초록 정보에 대량으로 존재하는 롱테일 데이터를 군집화 하였다. 마지막으로, 생성된 용어 네트워크를 기반으로 재현율(recall) 성능 실험을 통해 정보 검색과 정보 분석 시 정확성(accuracy)이 향상되는 현상을 설명하고자 하였다.

본 연구의 방법 및 수행 범위와 관련하여 첫째, 실험을 위한 데이터 집합은 개방 데이터인 DBLP의 학술 데이터를 사용하였고, 초록 데이터 구축을 위해 DOI 링크 정보가 존재하는 데이터에 한해 수집을 실시하였다. 둘째, 비정형 초록 정보로부터 자동 추출한 색인어는 일정한 패턴을 보이며 두문자어(acronym)와 확장어(expansion) 형식을 갖는 학문 분야 별 주요 기술 용어를 이용하였다. 기술 용어는 해당 분야의 내용과 추세를 가장 잘 반영하는 주요 색인어이기 때문에 기술 용어를 이용하여 학문 분야의 연구 경향과 추세를 연구하는 많은 연구가 수행되고 있다(황미녕 외, 2011; Abe & Tsumoto, 2010; Kim et al., 2012; Hwang et al., 2014). 셋째, URL이 부여되어 고유 식별이 가능한 기술 용어의 네트워크는 본 연구를 통해 시스템을 통한 공개 단계까지 수행되지 않았으므로, 후속 연구를 통해 시스템 개발과 자원 공개를 통한 다양한 응용 연구가 수행되어야 할 것으로 보인다.

2장에서는 데이터의 활용성과 관련한 연구와 사례로써 LOD 관련 응용 연구 사례, 롱테일과 다크 데이터 관련 연구를 소개하고 기술 용어 등의 언어 자원 생성 및 응용 연구를 소개한다. 3장에서는 웹으로부터 데이터를 수집하기 위한 과정과 기술 용어를 추출하는 방법을 상술한다. 텍스트 마이닝 기법을 이용해 추출된 데이터를 상호 연결하는 방법을 제안하고 네트워크 구축 결과를 통계적으로 설명하고자 한다. 4장에서는 용어 네트워크를 이용한 검색 실험을 실시하고 성능을 진단함으로써 본 연구의 효과와 활용 가능성을 언급하고자 하였다. 마지막으로 결론에서는 본 연구의 의의를 정리하고, 연구의 한계와 향후 연구를 밝힌다.

2. 관련 연구

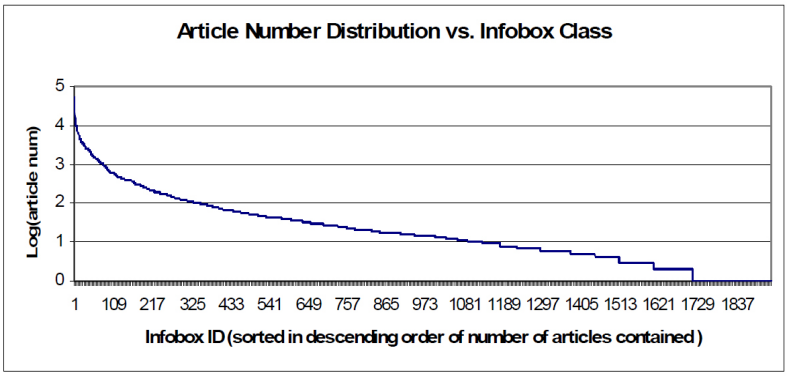
2.1 롱테일 데이터와 다크 데이터 활용 연구

이미 언급한 바와 같이 LOD는 공개된 수많은 데이터를 어떻게 엮어서 시너지를 만들 것인가에 궁극적인 목표를 두고 있다. 데이터를 공개하는 절차도 매우 중요하지만 그 목적은 역시 데이터의 활용에 있다고 할 수 있다. 더불어 빅 데이터 환경에서의 롱테일 데이터와 다크 데이터 처리 기술은 개체 수가 작아서 통계적으로 잘 활용되지 않거나 시스템에 숨겨진 데이터들을 얼마나 활용할 수 있느냐에 목적을 두고 있다. 특별할 수 있는 작은 데이터들을 발견하기 위한 연구들이 향후 빅 데이터 프로세싱에서 더욱 중요한 역할을 하게 될 것이다. Wu et al.(2008)은 대표적인 오픈 데이터인 Wikipedia의 인포박스(infobox) 내 정보 개체들이 롱테일 분포 현상을 보임을 언급하였다

(〈그림 1〉 참조). 많은 기계 학습 실험이 대체로 충분한 자질이 확보되는 경우에는 훌륭한 성능을 보이지만, 불균등하고 빈도가 낮은 데이터를 이용하여 학습을 하는 경우에는 상대적으로 저조한 성능을 보이게 됨을 지적하였다. 특히 롱테일 영역의 데이터를 이용함으로써 정보 검색 시 재현율(recall)을 향상시키기 위한 기법을 소개하였다. Li et al.(2014)은 실제 정보 처리 환경에서는 동일한 데이터 항목들이 서로 다른 자원에서 수집되기도 하고, 이로 인해 데이터의 중복과 같은 충돌이 일어나기도 한다는 점을 상기하였다. 이러한 데이터 신뢰의 문제를 해결하기 위해 롱테일 데이터를 활용하고자 하였다. 다양한 자원을 신뢰도에 따라 가중치를 부여하는 최적화 기반의 수집 프레임워크를 구축하고, 실제 롱테일 데이터 셋에 적용하는 실험을 통해 제안된 신뢰성 탐지 알고리즘의 우수한 성능을 보여주었다. Heidorn(2008)은 도서관을 비롯한 많은 기관들이 롱테일 법칙에 따라 간과되기 쉬운 많은 데이터를 다크 데이터의 개념으로 이해하는 것이 중요하다고 지적하였다. 숨겨진 가치 있는 데이터를 활용함으로써 많은 사회적 이슈와 기술적 이슈를 이해하는 데 기여할 수 있다는 점을 강조하였다. Zhang et al.(2016)은 대량의 텍스트, 표, 이미지 등으로부터 관계 데이터베이스를 생성하는 소프트웨어인 DeepDive를 소개하며, 과학기술 문헌, 웹 광고, 고객 서비스 데이터 등의 방대한 다크 데이터로부터 가치있는 빅 데이터를 창출할 수 있음을 강조하였다. 기존의 정보 추출(information extraction) 시스템에 비해 높은 정확률과 재현율을 동시에 보장하는 시스템을 이용함으로써 기업, 정부, 과학자, 정보 분석가에게 보다 많은 가치 정보를 제공할 수 있을 것으로 기대하였다.

2.2 기술 용어 생성 및 응용 연구

시소러스(thesaurus), 텍사노미(taxonomy), 기술용어 등 전문용어 사전(terminology dictionary), 동의어 사전(synonym dictionary), 시맨틱 웹을 위한 온톨로지(ontology) 등 다양한 유형의 언어 자원은 정보 검색 시스템의 성능 향상, 인공지능과 기계학습 기술 향상, 정보 분석 기술의 향상을 위해 다양한 응용 연구에 활용된다.

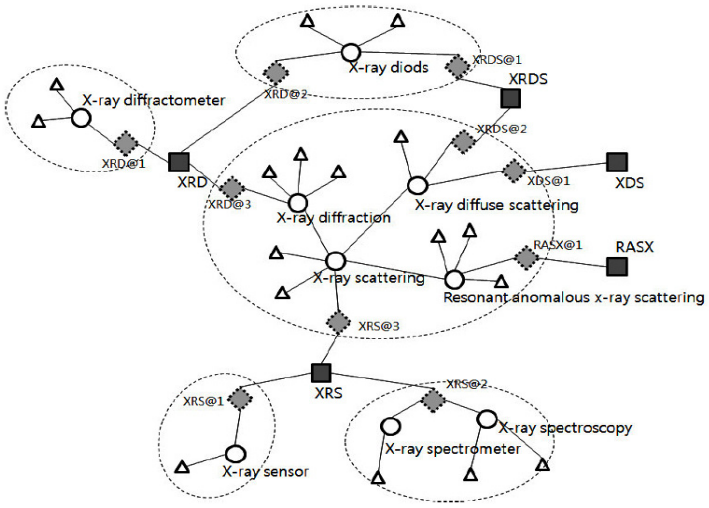


〈그림 1〉 롱테일 현상을 보이는 위키피디아 인포박스 클래스 당 논문 수(Wu et al., 2008)

Jeong et al.(2011)은 기술용어의 지식 맵 처리 방안을 제시하여 언어 자원의 시맨틱 네트워크를 생성하는 연구를 수행하였다. 두문자어 표현형에 대한 여러 종류의 기술 용어가 연결되었을 때 효과적으로 처리하기 위해 의미그룹을 생성하는 방법을 제안하였으며, 인스턴스의 매핑 규칙을 통해 의미 식별을 명시적으로 처리하였다(〈그림 2〉 참조). 또한 Jeong et al.(2013)은 기계 학습을 통한 의미 식별 연구를 통해 두문자어와 확장어 간의 복잡한 의미적 관계를 해석하기 위한 방법론을 제시하였다. 학술 논문 초록의 자연어 처리를 통해 추출한 다양한 유형의 자질(features)을 이용하였으며, 마이크로 평균 정확률 측정 결과, 단일 명사인 NN 유형이 명사 상당어구인 NP 타입보다 우수한 식별 성능을 보여 줌을 확인하였다. Abe et al.(2010)과 Hwang et al.(2014)은 패턴 인식 방식을 기반으로 기술 용어의 추세를 분석하여 기술 용어의 생명 주기를 측정하고 유형화하는 용어 기반의 분석적 연구를 수행하였다. Kim et al.(2012)은 결정 나무(decision tree) 기반의 기술 용어의 트렌트 분석 및 예측 방법을 제안하였다. 특히 용어의 발전 단계를 진입(Irruption), 급성장(Frenzy), 추세전환(Turning point), 융합상승(Synergy), 성숙(Maturity)의 5단계 과정으로 자동 분류하여 기술 용어의 현재 상태를 진단하고 향후 기술의 발전 정도를 예측하였다.

2.3 편집 거리(edit distance) 알고리즘

본 연구는 유사 어형을 가진 기술용어의 통제를 위해 편집 거리 알고리즘을 사용한다. 편집 거리 알고리즘은 특정 문자열 S를 다른 유사 문자열인 S'로 변환시키는 과정에서 필요한 최소 편집 연산 횟수를 측정하는 기법이다. 문자열 또는 문헌 벡터의 유사도를 측정하는 방법으로 사용되는 유사 계수(similarity coefficient)는 유사 정도를 벡터의 가중치 값으로 나타내는



〈그림 2〉 다양한 확장어를 갖는 두문자어의 시맨틱 네트워크(Jeong et al., 2011)

반면, 편집 거리는 몇 개의 글자가 차이나는 가를 측정하므로 문자열 간의 비유사성을 계산한다. Jeong et al.(2011)은 기술 용어의 유사성을 측정하여 시맨틱 네트워크를 생성하기 위해 다이스(Dice) 계수를 이용하였으며, 임계치의 변화를 관찰하여 유사도 0.9 이상을 용어 군집화의 기준으로 설정하였다. 이 연구에서는 용어의 길이가 군집 결과에 영향을 끼치게 되므로 임의의 기준치를 설정하는 데 어려움이 존재하였으나, 이에 반해 편집 거리 알고리즘을 사용하면 용어의 길이와 상관없이 두 문자열을 비교하고 변환해야 하는 문자의 수를 반환하는 장점이 있다. 문자 편집 시 사용하는 세 가지 방법은 삽입, 삭제, 대체 연산이며, 최적 알고리즘을 통해 가장 적은 횟수로 변환할 수 있다(Wikipedia, 2018d). Cormode & Muthukrishnan(2007)은 부분 문자열의 이동을 가능하게 하는 추가 기능의 알고리즘을 제안하고 기존의 편집 거리 알고리즘을 개선하는 연구를 제안하였다. 또한 Reis et al.(2004)은 대용량 자원인 월드 와이드 웹으로부터 tree edit distance 기법을 이용해 웹 뉴스 데이터를 추출하는 방법을 제안하였다. 안광모 등(2013)은 편집 거리 알고리즘인 레벤슈타인(Levenshtein) 거리를 이용해 영화평 내의 감성을 분류하고자 하였다. 분류 실험을 위한 감성 자질을 추출하기 위해 편집 거리 알고리즘을 활용하였으며, 레벤슈타인 거리 값과 학습 기법을 변화시키면서 최적의 분류 성능을 찾고자 하였다. 편집 거리 알고리즘은 문자열 편집, 철자 오류 검색, 자연어 번역 등의 자연어 처리 분야에서 많이 응용되고 있으며, 유전자 유사도를 판별하는 등 다양한 패턴처리를 위해 광범위하게 활용되고 있다.

2.4 LOD(linked open data) 활용 연구

Paulheim & Fümkranz(2012)은 기계 학습의 성능은 좋은 데이터로부터 충분히 많은 자질을 구축하는 데 달려있다고 언급하면서, 공개된 대량의 데이터인 LOD로부터 확보한 자질을 비지도 학습(unsupervised learning) 기법을 이용해 확장하는 연구를 수행하였다. Jain et al.(2010)은 LOD가 데이터의 웹(Web of Data)으로 발전하기 위해서는 상호 연계성이 보장되어야 한다고 주장하면서, 데이터 셋 간의 효과적인 온톨로지 매칭(ontology alignment) 기법을 제안하였다. 오픈 데이터의 활용성을 높이기 위한 또 다른 연구로, Jeong et al.(2011)은 대표적인 오픈 데이터 중 하나인 DBPedia의 학술 정보 메타데이터를 수집한 후 이로부터 추출한 온톨로지 인스턴스(instance)들의 상호 매칭 방안을 제시하고 해석을 위한 추론 알고리즘을 시각적인 지식 맵과 함께 설명하였다. Noia et al.(2012)은 대표적인 LOD 사이트인 DBPedia, Freebase, LinkedMDB로부터 데이터를 수집하여 시맨틱 기술 기반의 새로운 영화 추천 시스템을 개발하였다. LOD로부터 수집한 대량의 RDF 데이터 간의 유사도를 측정하기 위해 벡터 공간(vector space) 모델을 사용하였으며, 성능 실험을 통해 LOD 기반의 추천 시스템이 정확률과 재현율 면에서 기존의 추천 시스템보다 나은 성능을 보여주었다.

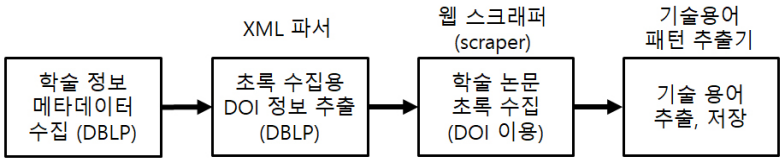
다음 장에서는 LOD 자원을 활용하여 기술 용어 네트워크를 생성하기 위한 일련의 과정인 데이터 수집과 패턴 추출 및 네트워크 생성 방안 등에 대해 상세히 소개한다.

3. 기술 용어 네트워크 생성과 통제

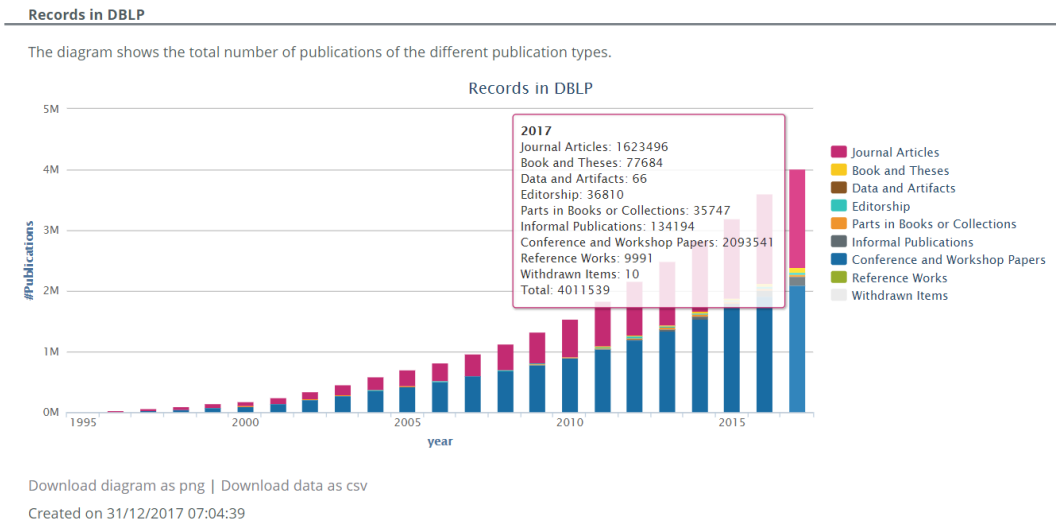
3.1 데이터 수집

본 연구에서는 기술 용어의 동의어 네트워크를 자동 생성하여 데이터의 활용성을 증대시키는 방안을 제시한다. 〈그림 3〉은 학술 논문의 초록 데이터를 이용한 기술 용어의 효과적인 활용을 위한 기술 용어 생성을 위한 전체 프로세스를 설명한다.

기술 용어 데이터를 생성하기 위한 첫 번째 단계는 학술 정보 메타데이터 수집 단계이다. 언어 자원 수집을 위한 자원(resource)으로 DBLP (http://dblp.org/) 사이트를 활용하였다. DBLP는 주요 컴퓨터 과학 저널과 학회로부터 생산되는 서지정보를 제공하는 대표적인 개방형 데이터(open data) 사이트들 중 하나이다. 2017년 12월 말 현재를 기준으로 구축된 DBLP의 총 보유 자원의 양은 〈그림 4〉와 같다. 저널 논문 약 162만 건, 학술 대회 논문 약 209만 건 등 총 400만 건 이상의 학술 자원이 구축되어 있으며 매년 지속적으로 증가하고 있다. 구축된 메타데이터 자원은 공개되어 있으며, DBLP 사이트를 통해 2GB 이상의 dblp.xml 논문 파일을 다운로드 받을 수 있다.



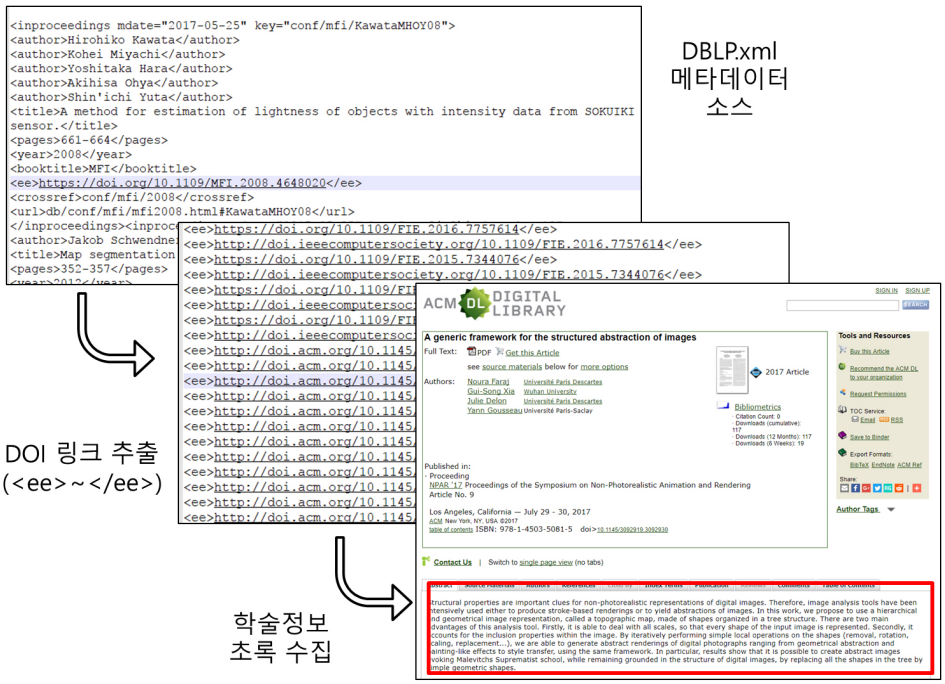
〈그림 3〉 기술 용어 생성 프로세스



〈그림 4〉 DBLP 보유 레코드 증가 추세와 구축 데이터 통계(2017년 12월 31일자)

데이터 프로세싱의 두 번째 단계는 DBLP 메타데이터를 이용해 초록 정보를 수집하기 위한 과정으로, 자원의 영구한 링크를 제공하는 디지털 객체 식별자인 DOI(digital object identifier) 정보를 활용한다. 이를 위해 데이터 소스인 DBLP 자원의 내용과 구조 분석이 선행되어야 한다. 〈그림 5〉는 DBLP 메타데이터의 주요 항목과 구조 정보 및 초록 정보 수집의 주요 과정을 예시 화면으로 설명하고 있다. XML 파서를 직접 개발하여 필요 데이터 항목인 해당 자원의 원본 링크 정보를 담은 '〈ee〉~〈/ee〉' 영역을 추출한 후 DOI 주소를 이용해 특정 웹 사이트 페이지 내의 학술 정보를 수집할 수 있게 된다.

세 번째는 다양한 웹 사이트에서 초록 정보를 수집하는 단계이다. 우선 본 연구의 데이터 수집 방법에 따라 DBLP 학술 정보 중 DOI가 존재하는 데이터로 필터링 하였으므로 DOI 정보가 존재하지 않는 박사 학위 논문 등의 자료들은 자동으로 배제되었다. 본 연구를 위해 수집한 DBLP XML 데이터는 2017년 9월 3일자(파일명: dblp-2017-09-03.xml.gz)이며, DOI가 존재하는 레코드는 총 3,505,031 건으로 전체 3,878,600 건의 약 90.37% 수준이다. 대량의 데이터를 빠르게 처리하기 위해 우선 독일어 등 영어가 아닌 학술 정보를 사전 제거한 후, 무작위로 후보 리스트를 생성하고 일정 접속 시간 내에 수집되는 사이트를 중심으로 초록 수집을 위한 웹 스크래퍼(web scraper)를 직접 개발하여 실행하였다. 일반적으로 웹 사이트 단위의 구조 정보까지 처리하는 웹 크롤러



〈그림 5〉 DBLP 기반 초록 정보 수집 과정

(web crawler)에 비해 웹 스크래퍼는 해당 페이지의 구조를 파싱하여 원하는 초록 부분만을 인식하여 수집하는 간단한 기능을 갖는다. 최대한 빠르게 데이터를 수집하기 위해 랩탑 PC 1대와 데스크탑 PC 2대를 동시에 활용하여 약 50만 건의 초록 데이터 수집을 목표로 2017년 9월 중순부터 12월 중순까지 약 3개월간 수집하였다. 또한 통계 분석 및 성능 평가 등 일련의 실험은 데이터 셋 구축 완료 후 2018년 1월까지 수행되었다.

데이터 수집 및 구축의 마지막 단계는 기술 용어 추출 단계이다. 특정한 규칙으로 이루어진 기술 용어 패턴을 감지하고 초록 정보로부터 이들을 추출하는 과정이다. 본 연구에서 제안하는 기술 용어 추출기의 패턴 처리 방안은 다음 장에서 상세히 설명한다. 앞서 소개한 일련의 정보 추출 과정을 통해 최종적으로 기술 용어가 존재하는 초록 데이터 약 50만 건(494,992건)을 구축하였고, 이로부터 약 9만 건(90,292종)의 고유한 기술 용어를 생성하였다.

3.2 기술 용어의 상호 네트워크 생성

3.2.1 패턴 기반 기술 용어 추출

본 장에서는 비정형 텍스트인 초록 정보로부터 기술 용어를 추출하기 위한 방안을 상술한다. 본 연구는 〈그림 6〉과 같이 기술 용어가 대부분 두문자어를 가지며 학술 논문에서 기술을 설명할 때 흔히 사용된다는 점에 착안하였다. 예를 들면 성능이 우수한 자동 분류 알고리즘의 하나인 SVM이 초록 상에서는 흔히 'support vector machines(SVMs)' 또는 'SVMs(support vector machines)'의 형태로 표현되는 패턴을 처리하여 기술 용어를 추출하였다. 본 연구 수행을 위해 오류 패턴 처리 능력과 처리 속도를 개선한 기술 용어 추출기(technical term extractor)를 파이썬(python) 프로그래밍 언어를 사용하여 직접 개발 하였다.

[ABSTRACT]
Identifying bacterial promoters is the key to understanding gene expression. Promoters lie in tightly constrained positions relative to the **transcription start site (TSS)**. Knowing the TSS position, one can predict promoter positions to within a few base pairs, and vice versa. As a route to promoter identification, we formally address the problem of TSS prediction, drawing on the database of known (mapped) Escherichia coli TSS locations. The accepted method of finding promoters (and therefore TSSs) is to use **position weight matrices (PWMs)**. We use an alternative approach based on **support vector machines (SVMs)**. In particular, we quantify performance of several SVM models versus a PWM approach, using area under the **detection-error tradeoff (DET)** curve as a performance metric. SVM models are shown to outperform the PWM at TSS prediction, and to substantially reduce numbers of false positives, which are the bane of this problem.

〈그림 6〉 초록으로부터 추출 가능한 기술 용어 패턴 예시

〈그림 7〉은 학술 논문의 초록 데이터로부터 기술 용어를 추출하는 알고리즘을 설명한 의사 코드(pseudo code)이다. 우선 프로그램을 실행하면 초록 정보를 한 문장씩 읽어 들인다(line 1-3). 이 때 문장에 나타난 "support vector machines(SVMs)" 또는 "SVMs(support vector machines)"의 형태들을 모두 찾아내 기술용어의 후보 리스트에 임시 저장한다. 이 때 두문자어는 항상 대문자로 표기되며, 기술 용어는 각 어절의 첫 글자가 대문자인 경우도 흔히 존재한다. 따라서 글자의 길이가 짧으면서 모두 대문자로 표기된 경우에만 두문자어로 인식하도록 하고 다어절의 문자열은 기술 용어 후보가 된다(line 5-6, line 8-9). 인식에 성공하면 다음 패턴 매칭으로 갈 수 있도록 플래그(process_flag)를 설정하고 조건에 맞지 않는 경우에는 플래그를 'No'로 선언하여 패턴 처리를 건너뛰도록 한다

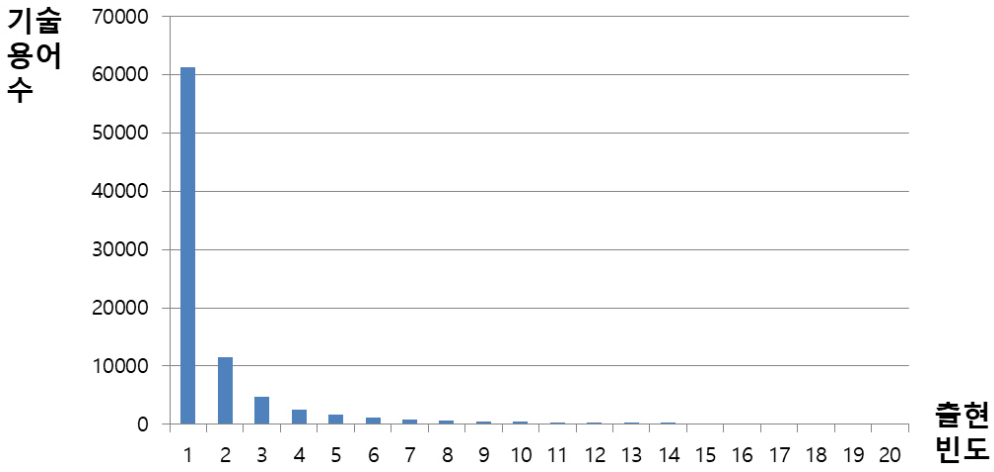
(line 7, line 10-13). 앞서 후보 기술 용어 추출에 성공한 경우에만 패턴 매칭을 실시하므로(line 14), 기술 용어인 확장어(expansion)의 각 어절의 두문자를 합친 것이 두문자어(acronym)와 일치하는가를 비교한다(line 15-16). 이 때 패턴으로 처리해야 하는 경우의 수가 증가하게 된다. 대표적인 경우가 하이픈 '-'의 처리 방식으로, 〈그림 6〉의 detection-error tradeoff(DET)에서는 하이픈(-) 있는 단어를 두 개의 어절 'DE'로 처리해야 하지만, anti-lock braking system(ABS)의 경우에는 하이픈으로 이어진 하나의 어절로 처리하기도 한다. 이와 같이 데이터 패턴의 다양한 경우를 처리할 수 있도록 프로그램의 로직을 지속적으로 추가하였다. 입력된 규칙(rule)에 따라 처리되는 패턴에서 예외인 경우에는, 별도로 작성된 특별한 기술 용어 형태를 처리하도록 한다. 예를 들면, 온톨로지 표준 언어인 OWL은 "Web Ontology Language"의 약어로 두문자 표기의 순서가 바뀌는 경우이며, XML은 "Extensible Markup Language"의 약어로 두문자 E가 아닌 X를 사용하여 두문자어를 만든다. 이러한 예외 처리는 간단하게 예외 처리용 사전을 이용하여 처리한다(line 18). 두 가지 경우를 벗어난 패턴에 대해서는 후보 두문자어와 기술 용어를 모두 버리게 된다(line 19-20). 한 줄 처리가 종료되면 리스트에 대기 중인 다음 문자열을 처음부터 반복 처리하도록 하고(line 4), 하나의 초록 데이터에 대해 모든 처리가 완료되면 프로그램을 종료한다(line 21).

```
 1) Start technical term extractor program
 2) Input String (abstract)
 3) Read sentences of abstract
 4) While every "string1 (string2)" form exists in the sentences ;
 5)         if string1 is shorter than string2 and string1 is upper case ;
 6)               acronym = string1, expansion = string2
 7)               process_flag = 'Yes'
 8)         else if string2 is shorter than string1 and string2 is upper case ;
 9)               acronym = string2, expansion = string1
10)               process_flag = 'Yes'
11)         else ;
12)               acronym, expansion are abandoned
13)               process_flag = 'No'
14)         if process_flag equals 'Yes' ;
15)               if all initial composition of expansion equals acronym ;
16)                     final_acronym = acronym, final_expansion = expansion
17)               else if acronym is in the exception_dictionary ;
18)                     final_acronym, final_expansion assigned by exception_dictionary
19)               else ;
20)                     acronym, expansion are abandoned
21) End technical term extractor program
```

〈그림 7〉 기술 용어 추출 알고리즘

3.2.2 롱테일 기술 용어 데이터의 통제

특정 주제 분야의 기술 용어는 대체적으로 통일된 형태를 지니게 되지만, 저자들은 통제된 어휘집을 사용하지 않고 자연어 상태로 용어를 기입하기 때문에 다양한 형태가 나타나게 된다. 대량의 초록 정보로부터 추출된 기술 용어들은 상호 연결되지 않은 상태로 개별 존재하고 있어, 서로 다른 형태로 표기된 경우에는 동일어 임에도 개별 정보로 인식된다. 구축된 기술 용어의 전체적인 빈도 통계를 살펴보면 롱테일 법칙에 따라 분포되어 있음을 알 수 있다. 〈그림 8〉은 저빈도 기술 용어의 데이터 분포 일부를 나타내고 있으며 〈표 1〉은 상위 20위와 하위 20위의 빈도 데이터 일부를 보여준다. 데이터 구축 결과, 두문자어 당 평균 기술 용어는 약 2.9개 정도가 추출되었지만, 자연어 말뭉치 표현에 나타나는 단어들을 그 사용 빈도가 높은 순서대로 나열하였을 때 모든 단어의 사용 빈도는 해당 단어의 순위에 반비례 한다는 지프의 법칙(Zipf's Law)을 보이며 롱테일 데이터의 형태를 나타내고 있음을 확인할 수 있다. 1회 발생한 기술용어는 무려 6만 종이 넘는데, 대부분의 기술용어가 약 50만 건의 초록 문서 집합에서 1회만 출현하고 있다는 의미이다. 기술 용어의 종 수는 출현 빈도 증가에 따라 급격하게 줄어들어 2회 발생한 용어의 수는 약 1만 1천여 종 수준으로 급감하여 멱함수(power-law)의 분포 경향이 지속된다. 〈표 1〉의 우측 컬럼인 고빈도 상위 20위까지의 발생 빈도를 살펴보면, 가장 많이 출현한 용어는 15,919회이며, 이 용어는 엑스선결정학 'X-ray diffraction(XRD)'이었다. 고빈도 2위의 기술 용어는 high-performance liquid chromatography (HPLC)로 6,585건의 장서 빈도(CF; collection frequency)를 보여준다. XRD는 결정 내로 입사된 엑스선의 회절을 이용하여 결정의 원자, 분자 구조를 밝혀내는데 이용하는 도구이며(Wikipedia, 2018b), HPLC는 합성물의 요소들을 분리하고 구별짓는 분석적 화학 기법 중의 하나이다(Wikipedia, 2018c).



〈그림 8〉 출현 빈도에 따른 기술 용어 종 수의 롱테일 현상(저빈도 20위까지)

〈표 1〉 출현 빈도에 따른 기술 용어 종 수 통계(상하위 각 20위까지 표기)

출현 빈도(하위 20위)	기술용어 종 수	출현 빈도(상위 20위)	기술용어 종 수
1	61,344	2,250	1
2	11,553	2,268	1
3	4,650	2,349	1
4	2,591	2,389	1
5	1,708	2,615	1
6	1,196	2,634	1
7	872	2,693	1
8	703	2,852	1
9	550	2,924	1
10	446	2,980	1
11	353	3,190	1
12	328	3,250	1
13	274	3,645	1
14	256	3,780	1
15	203	4,156	1
16	186	4,680	1
17	143	4,988	1
18	144	5,554	1
19	131	6,585	1
20	121	15,919	1

3.2.3 기술 용어의 네트워크 생성 방법

본 연구에서 어형 통제를 위한 기계적인 처리는 크게 두 번의 단계를 거치는데, 우선 전처리를 거쳐 간단한 통제를 한 후 편집거리 알고리즘을 이용하여 용어의 어형들을 통제하게 된다. 첫 번째 단계는 하이픈 기호('-')의 처리 및 특수문자의 처리이다. 하이픈은 앞서 추출 알고리즘에서 설명한 바와 같이 기술 용어에서 빈번하게 사용된다. 모든 하이픈은 스페이스로 대체하는 전처리를 거치게 되며, 상당 수의 용어 데이터가 하이픈 처리만으로도 상호 연결되는 것을 알 수 있다(〈표 2〉 참조). 웹 상의 초록 정보를 수집하는 경우에, 원본에 포함된 각종 html 특수문자를 처리해야 한다. 자주 등장하는 태그나 특수문자 코드의 몇 가지 예를 들면, " "는 한 칸 띄어쓰기, "〈sup〉~〈/sup〉"와 "〈sub〉~〈/sub〉"는 각각 위첨자와 아래첨자를, "®"는 트레이드마크 심볼, "'"는 "'"문자를 의미한다. 그밖에 "&bgr;, β, ç, ú, è, é, /, ’, ', ..." 등 특수 문자 목록을 만들어 전처리에 사용하였다. 〈그림 9〉는 1단계 전처리 과정을 통해 처리된 용어 데이터의 예시이며, 첫 행의 'ML [STR1] Müller-Lyer 3 [STR2] Müller-Lyer 6 [ED] 0'를 살펴보면, ML은 두문자어, 연결할 대상 용어는 'Müller-Lyer'와 'Müller-Lyer'이며 각각 발생 빈도가 3과 6임을 뜻한다. 맨 마지막의 '[ED] 0'은, 독일어 ü 문자를 서로 다른 형식으로 표기한 부분을 간단하게 처리한 후 비교한 결과 두 문자열의 편집 거리는 0임을 나타낸다. 편집 거리 알고리즘은 2단계 과정에

ML [STR1] Müller-Lyer 3 [STR2] Müller-Lyer 6 [ED] 0
SLS [STR1] Senior-Løken syndrome 3 [STR2] Senior-Løken syndrome 1 [ED] 0
SLS [STR1] Sjögren-Larsson syndrome 9 [STR2]Sjögren-Larsson syndrome 6 [ED] 0

〈그림 9〉 어형 통제 처리 1단계: 문자열 치환을 통한 전처리(특수문자 처리 예)

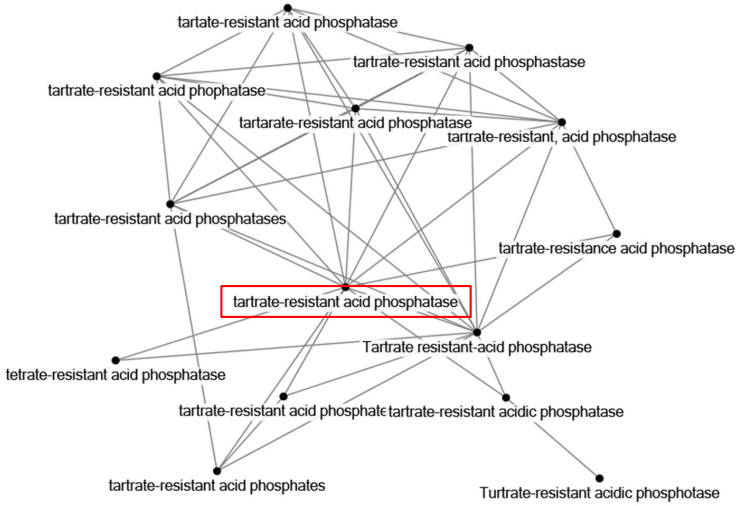
서 이어 설명하도록 한다.

두 번째 단계는 어형 통제 과정으로, 본 연구에서는 유사 문자열을 갖는 동일 용어를 상호 연결하기 위해 편집 거리 알고리즘을 사용한다. 〈그림 10〉에서 '[ED] 1'은 편집 거리가 1인 두 개의 기술용어를 나타내며, '[ED] 2'는 편집 거리가 2인 기술용어 쌍을 의미한다. ED1은 주로 "X-ray diffraction"을 "X-ray diffaction"이나 "X-ray difraction"으로 작성한 사소한 오류 표기나 복수형 어미 's'의 탈자를 처리하는 데 유용하였다. ED2는 ED1과 같은 오류 표기를 포함하여 미국식과 영국식 표기의 차이, 즉 "high-fiber diet"와 "high-fibre diet"와 같이 "_er"과 "_re"를 추가 처리하기 위한 것이다. 지나친 링크 생성으로 데이터 품질을 저하시키지 않도록 편집거리 3 이상은 사용하지 않는다.

〈그림 11〉은 2단계의 처리과정을 통해 자동 생성된 기술 용어 네트워크의 그래프 시각화

STA [STR1] short-term average 2 [STR2] short-term averages 1 [ED] 1
FRO [STR1] free-radical oxidation 7 [STR2] free-radicals oxidation 1 [ED] 1
HFD [STR1] high-fiber diet 3 [STR2] high-fibre diet 1 [ED] 2

〈그림 10〉 어형 통제 처리 2단계: 편집거리 알고리즘 활용(ED1-2 적용 예)



〈그림 11〉 자동 생성된 기술 용어 네트워크 구조 및 대표어 시각화 예시

〈표 2〉 네트워크 에지 생성 프로세싱 과정 예시

노드 1	노드 2	노드1 빈도	노드2 빈도	편집 거리	빈도 총합
Tartrate resistant-acid phosphatase	tartrate-resistant acid phosphatase	1	543	0	544
tartrate-resistant acid phosphates	tartrate-resistant acid phosphate	1	3	1	4
tartate-resistant acid phosphatase	Tartrate resistant-acid phosphatase	1	1	1	2
tartate-resistant acid phosphatase	tartrate-resistant acid phosphatase	1	543	1	544
Tartrate resistant-acid phosphatase	tartrate-resistant acid phosphastase	1	1	1	2
Tartrate resistant-acid phosphatase	tartrate-resistant, acid phosphatase	1	1	1	2
Tartrate resistant-acid phosphatase	tartrate-resistant acid phophatase	1	1	1	2
Tartrate resistant-acid phosphatase	tartrate-resistant acid phosphatases	1	1	1	2
Tartrate resistant-acid phosphatase	tartarate-resistant acid phosphatase	1	7	1	8
tartrate-resistant acid phosphastase	tartrate-resistant acid phosphatase	1	543	1	544
tartrate-resistant acid phosphatase	tartrate-resistant, acid phosphatase	543	1	1	544
tartrate-resistant acid phosphatase	tartrate-resistant acid phophatase	543	1	1	544
tartrate-resistant acid phosphatase	tartrate-resistant acid phosphatases	543	1	1	544
tartrate-resistant acid phosphatase	tartarate-resistant acid phosphatase	543	7	1	550
tartrate-resistant acid phosphates	Tartrate resistant-acid phosphatase	1	1	2	2
tartrate-resistant acid phosphates	tartrate-resistant acid phosphatase	1	543	2	544
tartrate-resistant acid phosphates	tartrate-resistant acid phosphatases	1	1	2	2
tartrate-resistant acid phosphate	Tartrate resistant-acid phosphatase	3	1	2	4
tartrate-resistant acid phosphate	tartrate-resistant acid phosphatase	3	543	2	546
tartrate-resistance acid phosphatase	Tartrate resistant-acid phosphatase	4	1	2	5
tartrate-resistance acid phosphatase	tartrate-resistant acid phosphatase	4	543	2	547
tartrate-resistance acid phosphatase	tartrate-resistant, acid phosphatase	4	1	2	5
tartate-resistant acid phosphatase	tartrate-resistant acid phosphastase	1	1	2	2
tartate-resistant acid phosphatase	tartrate-resistant, acid phosphatase	1	1	2	2
tartate-resistant acid phosphatase	tartrate-resistant acid phophatase	1	1	2	2
tartate-resistant acid phosphatase	tartrate-resistant acid phosphatases	1	1	2	2
tartate-resistant acid phosphatase	tartarate-resistant acid phosphatase	1	7	2	8
Tartrate resistant-acid phosphatase	tartrate-resistant acidic phosphatase	1	2	2	3
Tartrate resistant-acid phosphatase	tetrate-resistant acid phosphatase	1	1	2	2
tartrate-resistant acid phosphastase	tartrate-resistant, acid phosphatase	1	1	2	2
tartrate-resistant acid phosphastase	tartrate-resistant acid phophatase	1	1	2	2
tartrate-resistant acid phosphastase	tartrate-resistant acid phosphatases	1	1	2	2
tartrate-resistant acid phosphastase	tartarate-resistant acid phosphatase	1	7	2	8
tartrate-resistant acidic phosphatase	tartrate-resistant acid phosphatase	2	543	2	545
tartrate-resistant acidic phosphatase	Turtrate-resistant acidic phosphotase	2	1	2	3
tartrate-resistant acid phosphatase	tetrate-resistant acid phosphatase	543	1	2	544
tartrate-resistant, acid phosphatase	tartrate-resistant acid phophatase	1	1	2	2
tartrate-resistant, acid phosphatase	tartrate-resistant acid phosphatases	1	1	2	2
tartrate-resistant, acid phosphatase	tartarate-resistant acid phosphatase	1	7	2	8
tartrate-resistant acid phophatase	tartrate-resistant acid phosphatases	1	1	2	2
tartrate-resistant acid phophatase	tartarate-resistant acid phosphatase	1	7	2	8
tartrate-resistant acid phosphatases	tartarate-resistant acid phosphatase	1	7	2	8

결과 예시이며, 이 때 이형 용어들의 처리 과정을 '전처리 → ED 1 → ED 2' 순으로 보여주는 데이터 프로세싱 과정은 〈표 2〉와 같다. 고유한 개체 수(node)는 14개이며, 네트워크를 형성하는 라인인 에지(edge)는 42개이다. 〈표 2〉의 '편집 거리' 컬럼을 참조하면, '전처리 → ED 1 → ED 2'처리 순서에 따라 '1 → 13 → 28'개의 에지가 생성됨을 알 수 있다. 또한 각 용어의 빈도가 상호 연결되는 순간 총 빈도로 업데이트 되며, 개체 빈도의 최종 합이 전체 네트워크의 총 빈도 수가 된다. 모든 기술 용어의 발생빈도는 568회로 모든 개별 용어는 '568'이라는 값을 장서 빈도로 상속받을 수 있다. 또한 〈그림 11〉에서 붉은 색으로 표시된 'tartrate-resistant acid phosphatase'의 빈도 수는 543회로 네트워크를 대표하는 대표어로 자동 지정된다.

본 연구는 저자가 사용한 자연어 상태의 기술 용어를 원형 그대로 유지하기 위해 용어 자체에는 변경을 가하지 않으며, 오자 및 탈자를 포함한 오류 정보도 그대로 유지하는 것을 원칙으로 한다. 즉, 모든 기술 용어의 원본 표기를 그대로 유지하여 URI를 부여하였으며, 용어 데이터의 상호 연결을 위한 용도로만 2단계의 어형 통제 방식을 수행하였다. 이러한 처리 방식은 데이터를 추출한 원문의 고유 번호(foreign key)을 통한 문헌-용어 간의 연결 뿐 아니라 원문의 용어 표기 형태의 동일성을 유지할 수 있다는 장점이 있어, 빅 데이터 환경에서 지속적으로 추가되는, 동일한 표기의 다른 문헌들도 모두 손쉽게 상호 연결 지을 수 있다는 장점이 있다.

3.2.4 URI 할당 방법

다양한 형태로 표현된 동일 기술 용어들 간의 네트워크를 생성한 후에는 논문 정보에 발생한 모든 형태의 용어를 처리할 수 있게 된다. 이에 용어 네트워크를 생성하고 데이터 관점에서 발생 빈도에 따라 대표 표현형을 자동으로 지정하는 방법을 제시한다. 앞서 설명한 예시를 다시 살펴보면, 〈그림 11〉과 〈표 2〉 상에서 빈도가 가장 높은 용어인 'tartrate-resistant acid phosphatase'를 자동으로 대표어로 선정하였다. 이 때 만약 같은 빈도의 용어가 2개 이상 동일 네트워크 상에 공존할 경우에는 알파벳 순, 혹은 임의로 후보 용어 중 하나를 지정한다. 이 후 문헌 데이터가 지속적으로 추가되어 용어의 출현 빈도 차가 발생할 경우 자동으로 대표어를 갱신하도록 한다.

〈그림 12〉는 기술 용어의 상호 네트워크 생성 및 대표어 선정과 URI 부여 등의 통제 과정을 의사 코드로 설명하고 있다. 우선 입력 데이터는 추출 용어의 목록이다(line 2). 다음 단계에서는 두 용어 간의 편집거리를 측정하기 위해 용어의 쌍을 생성한다(line 3). 입력된 N개의 용어에서 K개를 선택하여 만들 수 있는 모든 순서쌍의 개수는 조합(combination) ${}_nC_k$으로 계산할 수 있다. ${}_nC_k$ = N!/{(N-K)!*K!}에서 2개의 용어를 선택하므로 K=2를 대입하여 계산한 최종 식은 N(N-1)/2이 된다. 본 연구에서는 추출된 모든 기술 용어의 개수가 90,292개이므로 편집 거리 측정을 위해 약 4억(90,292*90,291/2 = 4,076,277,486)개의 후보 순서쌍이 프로세싱 과정에서 자동 생성된다. 생성된 모든 후보 용어의 쌍에 대해(line 4), 우선 고유한 식별자인 URI를 모든 용어에 부여한다. 이를 위해 간단하고 효과적인 URI 생성 규칙(convention rule)을 작성하였다. URI는 총

```
1) Start Network_Creation program
2) Input String (list_of_terms)
3) Generate pairs from list_of_terms
4) While a pair of terms exists (as an interlinking graph candidate) ;
5)        generate and assign 1st URI to all terms
6)        string1 = pair[0], string2 = pair[1]
7)        erase special_characters in string1, string2 from stopword_list
8)        if string1 equals sting2 ;
9)              create a new graph of interlinked network with "ED_0" value
10)       else if Edit_Distance value between string1 and string2 is "1";
11)             create a new graph of interlinked network with "ED_1" value
12)       else if Edit_Distance value between string1 and string2 is "2";
13)             create a new graph of interlinked network with "ED_2" value
14)       else ;
15)             abandon string1 and string2
16) #End of While statement: Creating semantic network
17) count Collection_Frequency(CF) of all terms in the global_network
18) For sub_network in global_network ;
19)             find a term with max(CF) and consider its URI as authorized 2nd URI
20)             assign 2nd URI to all terms in the same network
21) #End of For statement: Assigning URI to all terms with authorization
22) End the program
```

〈그림 12〉 기술 용어의 네트워크 생성 및 통제 알고리즘

10자리의 알파벳과 숫자(alphanumeric)의 조합 데이터로 구성되어 있으며, 데이터베이스 저장을 고려하여 모두 대문자(upper case)로 생성한다(〈그림 13〉 참조). 이 때 각 자리 수는 36가지로 표현되므로 10자리로 표현가능 한 용어의 종 수는 36의 10승(3,656,158,440,062,976)개 이다. URI 부여를 위해 URI와 해당 용어는 "키(key):값(value)"구조로 해시(hash) 데이터 구조에 입력되어 관리한다. 해시 구조를 이용함으로써 빠르게 중복을 배제하고 해당 URI에 대한 용어를 고속으로 찾아낼 수 있도록 구성하였다. 다음 단계는 전처리 과정으로 앞서 설명한 바와 같이 몇 가지 규칙과 불용어(stop words) 사전을 이용해 어형을 정제한다(line 7). 편집거리 알고리즘은 0, 1, 2의 3단계로 진행하며, "ED_0"은 전처리 후 문자열이 완전 일치하는 용어 쌍으로 생성한 네트워크를 의미하며 "ED_1과 ED_2"를 통해 편집거리 1과 2의 용어 쌍들을 모두 그래프 데이터로 생성한다(line 8-16). 다음 단계는 용어 네트워크 생성이 끝난 후 생성된 군집별로 대표어를 자동 선정하는 과정이다(line 17-21). 초록 정보로부터 기술 용어를 추출하는 과정에서 사전 측정한 각 용어별 CF를 이용한다(line 17-18). 이 때 같은 네트워크 그룹 내의 대표어의 1차 URI를 이용하여 모든 그룹 내의 용어에 2차 URI를 할당한다. 앞서 설명한 바와 같이 동일 빈도의 용어가 2개 이상 같은 네트워크 상에 존재할 경우에는 특정 기준에 따르거나 임의로 하나를 지정한다. 이 후 문헌이 지속적으로 추가되어 용어의 빈도 차이

Id	acr	exp	exp_f...	uri	uri_freq	aut
89,588	XRD	X-ray diffraction	15,919	BOZYE2PWTR	15,967	Y
34,059	HPLC	high-performance liquid chromatography	6,585	63GH4ASZ20	8,101	Y
20,716	EBV	Epstein-Barr virus	5,554	SLN6W7KX4F	5,604	Y
14,680	CRP	C-reactive protein	4,988	X28SYPH1DZ	5,034	Y
41,449	LDL	low-density lipoprotein	4,680	ZJB15CVO4K	5,354	Y
1,199	ACE	angiotensin-converting enzyme	4,156	AZBTN4HOEP	4,220	Y
46,025	MAPK	mitogen-activated protein kinase	3,780	ZCEUW1QTIM	4,161	Y
65,103	PSA	prostate-specific antigen	3,645	0KBIG4OR9Q	3,696	Y
11,979	CGRP	Calcitonin gene-related peptide	3,250	1CE8I24ZFA	3,329	Y
31,814	HDL	high-density lipoprotein	3,190	N3XRWUBY1J	3,620	Y
56,392	NSCLC	non-small cell lung cancer	2,980	P84A0K765M	3,117	Y
89,336	WT	wild-type	2,924	JRMIHAPSF8	2,925	Y
60,901	PDGF	platelet-derived growth factor	2,852	O0FWT26GQL	2,874	Y
27,448	FSH	follicle-stimulating hormone	2,693	X3TQHDBF0G	2,723	Y
23,780	ESRD	end-stage renal disease	2,634	62NSLVGYAC	2,649	Y

〈그림 13〉 자동화된 어형 통제 시스템의 데이터베이스 구조

가 발생할 경우 자동으로 대표어를 갱신하도록 한다. 이러한 2단계 URI 관리 방식으로 각 그룹 내의 개별 용어의 원형들은 1차 URI를 통해 액세스하고, 어형 통제 결과는 네트워크 그룹 단위의 2차 URI를 이용할 수 있다. 〈그림 13〉은 자동화된 어형 통제 시스템의 데이터베이스 구조와 실제 저장된 데이터의 예시이다. 첫 번째 행의 XRD의 경우, 자신의 네트워크 그룹 내에서 가장 높은 빈도(15,919회)를 갖는 "X-ray diffraction"이 대표어로 자동 지정되어 해당 필드에 'Y'로 선언 되었으며, 네트워크의 내 모든 용어들의 총 빈도가 15,967회로 합계되었음을 확인할 수 있다.

4. 데이터의 활용률 실험 및 평가

정보 시스템이 기술 용어의 통제를 통해 다양한 형태로 나타나는 데이터를 표준화하여 처리하는 경우 얻을 수 있는 장점으로 첫째, 시스템이 자연어 상태의 이형 표기를 유지하면서 의미적으로는 하나의 대표적인 형태로 인식할 수 있기 때문에 통제된 환경에서 검색 효율을 향상시킬 수 있으며, 특히 검색 시스템의 재현율의 저하를 막을 수 있다. 대량의 정보를 검색하는 빅 데이터 환경에서 검색된 문헌의 정오를 판별하는 정확률이 시스템 성능의 기준이 될 수 있으나, 관련 문헌을 모두 찾아야 하는 법률 또는 특허 기술 문서와 같은 특정한 목적의 시스템에서는 빠짐없이 찾아내는 재현율을 보장하는 시스템이 더욱 가치를 갖는다. 둘째, 자동 분류 또는 추천 시스템과 같은, 레이블링(labeling) 기반의 응용 시스템 환경에서도 용어의 통제는 매우 중요하다. 통제가 이루어지지 않은 레이블 환경에서는 유사 데이터가 분산되기 때문에 데이터의 희소현상(sparseness)이 더욱 커지는 이유로 기계 학습 기반의 다양한 서비스와 분석적 작업을 수행하기 매우 어렵다. 본 연구에서 제안하는 기계적인 방식은 수작업 기반의 데이터 처리 방식에 비해 대량의 문헌이 지속적으로 증가하는 빅 데이터 환경에 보다 효율적일 수 있다. 대표어 선정 방식 역시 기술 용어의 빈도를 중심으로 자동으로

부여하는 방식을 사용하면 간단하게 이형 표기를 갖는 동일 용어들을 한 데 모아 처리할 수 있다. 본 장에서는 저빈도 기술 용어를 통제함으로써 일관된 레이블을 부여하는 효과를 살펴보고, 재현율 관점의 데이터 활용률 측정 실험을 통해 본 연구의 의의를 설명하고자 한다.

4.1 롱테일 데이터의 활용

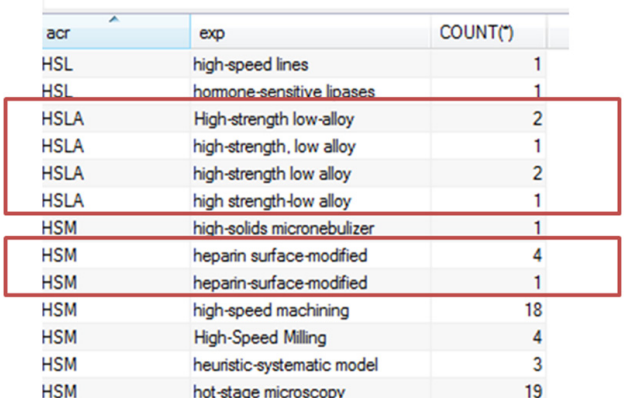
앞서 설명한 바와 같이 자연어 상태의 기술 용어 분포는 롱테일 현상을 보이며 지프의 법칙을 따르기 때문에 빈도 1, 2의 저빈도어가 전체 용어 종 수의 대부분을 차지하고 있다. 일반적으로 일정한 기준 이하의 저빈도 용어는 통계적인 의미가 적기 때문에 처리 속도 향상을 위해 전처리 과정에서 제거되기도 한다. 본 연구는 롱테일 데이터 활용을 위해서는 실제 데이터의 많은 비율을 차지하는 저빈도어를 용어 네트워크 기반으로 통제하여 색인어로서 의미가 큰 중간 빈도어로 전환하는 방법을 제안하고자 한다.

〈표 3〉은 장서 빈도를 기준으로 저빈도어와 중간 빈도어 및 고빈도어의 종 수에 대한 통계를 보여준다. 데이터 분포 통계를 통해 중간 빈도어 설정의 기준을 설정하였는데, 우선 누적 분포 비율을 통해 적절한 저빈도어의 경계를 살펴보았다. 장서 빈도 1인 용어가 67.94%로 대부분을 차지하고 있으며, 빈도 1~4의 낮은 빈도어가 전체의 88% 이상을 차지하는 것을 알 수 있다. 본 연구에서는 (1) 데이터의 희소 문제를 고려해 전체 용어 종 수 대비 최소 10% 이상의 비율을 차지하며, (2) 중간 빈도어로서 가능한 최고 빈도 수를 갖는 지점인 CF〉=5를 기준으로 설정하였다. 또한, 본 연구를 위해 구축한 데이터의 분포 통계를 감안할 때, 빈도가 증가함에 따라 급격하게 개체수가 감소하므로 중간빈도의 구간은 대략 CF 5에서 50 정도로 고려할 수 있을 것으로 보인다. 단, CF〉=5 기준은 네트워크 기반의 용어 통제를 통해 효과를 비교하기 위한 임의의 값이며, 기준에 따라 효과가 달리 표현될 수 있으므로 일례로 제시한 값이 절대적인 네트워크 효과를 보여주는 것은 아니다.

〈그림 14〉를 예로 들면, 저빈도 기술 용어인 HSLA(high-strength low alloy)는 총 4가지 형태로 각각 발생 빈도 1 또는 2를 나타내고 있

〈표 3〉 빈도별 용어 분포 통계(저빈도, 중간빈도, 고빈도 구간 일부)

빈도 (CF)	용어 종 수	비율(%)	누적비율(%)
1	61,344	67.94	67.94
2	11,553	12.80	80.73
3	4,650	5.15	85.88
4	2,591	2.87	88.75
5	1,708	1.89	90.65
10	446	0.49	94.82
50	21	0.02	98.81
100	8	0.01	99.41
500	1	0.00	99.89



acr	exp	COUNT(*)
HSL	high-speed lines	1
HSL	hormone-sensitive lipases	1
HSLA	High-strength low-alloy	2
HSLA	high-strength, low alloy	1
HSLA	high-strength low alloy	2
HSLA	high strength-low alloy	1
HSM	high-solids micronebulizer	1
HSM	heparin surface-modified	4
HSM	heparin-surface-modified	1
HSM	high-speed machining	18
HSM	High-Speed Milling	4
HSM	heuristic-systematic model	3
HSM	hot-stage microscopy	19

〈그림 14〉 중간 빈도어 전환이 가능한 저빈도 용어들의 예

다. 네트워크 생성 방법을 이용해서 하나의 기술용어로 통제하게 되면 각각의 기술 용어는 전체 빈도의 합인 6을 상속받아 나누어 갖게 된다. HSM(heparin surface-modified)의 경우에도 두 가지 이형을 상호 연결함으로써 본 연구에서 설정한 중간 빈도어 기준인 CF〉=5의 용어로 갱신될 수 있었다. 네트워크 기반의 방법을 사용함으로써 총 빈도의 합을 검색 또는 분석 결과로 보여줄 수 있으며, 이러한 용어 통제로 인해 통합 레이블이 부여되어 검색 또는 분석용 정보 시스템에서 효과적으로 활용될 수 있다. 용어 통제의 효과는 기존의 용어 네트워크에 새로운 데이터가 계속 추가 확장되는 개념이므로 대용량 데이터가 지속적으로 추가되는 빅 데이터 환경에서 더욱 잘 나타날 수 있을 것이다.

〈표 4〉는 네트워크 생성 실험을 통한 기술용어 그룹 수의 변화와 빈도의 통계를 나타낸다. 전체 문헌집단에 발생한 총 빈도인 CF를 5로 기준 설정하고, 5 미만의 기술용어 수와 5 이상의 기술용어 수의 변화를 비교하였다. 상호연결 방식을 통한 의미그룹 수의 변화를 살펴보면, 네트워크 생성을 통해 통제된 그룹의 개수(2차 URI 부여)는 78,688개로 약 12.85% 정도 감소하였다. 그룹의 수가 감소하였다는 것은 별개로 검색될 많은 용어들이 하나로 통제됨으로써 적합 문헌을 재현해 내는 효율이 높아짐을 의미한다. 이에 따라 빈도 5이상 용어 수와 빈도 5미만의 용어 수의 비율 역시 각각 -2.77%와 -14.13%로 감소하였다. 롱테일 영

〈표 4〉 네트워크 생성을 통한 기술 용어와 군집 수의 비율 변화

	네트워크 생성 전 (개별 용어)	네트워크 생성 후 (그룹화)	네트워크 생성 전후의 비율
기술 용어 수/군집수	90,292	78,688	-12.85%
빈도 5이상 용어/군집 수	10,154	9,873	-2.77%
빈도 5미만 용어/군집 수	80,138	68,815	-14.13%

역에 존재하는 작은 데이터들의 감소 비율이 상대적으로 높다는 점을 통해, 본 연구에서 제안하는 네트워크 기반의 용어 통제 효과는 롱테일 데이터 영역에서 더욱 잘 나타나고 있음을 알 수 있다.

본 연구를 통해 추출된 기술 용어는 49만 여 건의 구축된 초록 정보에 대한 양질의 색인어이다. 결국 롱테일 색인어의 활용률 제고는 초록 정보의 활용률을 높이는 방안이라 할 수 있다. 용어 네트워크 생성 전후의 기술 용어의 초록 분포 비율을 살펴보면, 장서빈도 5회 미만의 데이터는 전체 용어의 약 88.75%(80,138/90,292)로 대부분을 차지하며, 전체 초록 데이터의 약 21.97%(108,764/494,992)에 색인어로 존재하고 있다. 반대로 빈도 5 이상인 기술용어는 전체 용어의 약 11.25%(10,154/90,292)만을 차지하나 초록 데이터의 약 78.03%(386,228/494,992)를 커버하고 있다. 네트워크 생성 후 변화를 살펴보면, 5회 미만의 저빈도 데이터는 전체 용어의 약 76.12%(68,815/90,292) 수준으로 약 12.63% 포인트 감소하였으며, 전체 문헌(초록)을 커버하는 비율은 약 19.15%(94,775/494,992)수준으로 낮아졌다. 이 때, 5회 이상의 빈도를 갖는 용어들은 전체 용어의 약 10.93%(9,873/90,292)를 차지하며, 이 때 초록을 커버하는 비율은 약 80.85%(400,217/494,992)로 약 2.82% 포인트 증가하였음을 알 수 있다(〈표 5〉 참조). 네트워크 생성의 효과로 저빈도를 갖는 롱테일 기술 용어의 초록 정보에 대한 검색 효율이 높아졌음을 확인할 수 있다. 다만, 앞서 언급한 바와 같이 CF〉=5의 중간 빈도어 기준은 현상의 일부를 설명하기 위한 설정 값으로 수치별 성능 측정을 위한 임계치가 아님을 다시 밝힌다. 다양한 구간에 대한 재현율 기반 데이터 활용률 평가 결과는 다음 장에서 소개한다.

〈표 5〉 기술 용어의 빈도 기준에 따른 초록 정보 매칭 비율

	네트워크 생성 전 초록 매칭 건수	네트워크 생성 후 초록 매칭 건수
CF〈5	108,764	94,775
CF〉=5	386,228	400,217

4.2 롱테일 기술 용어의 데이터 활용률 실험 및 평가

본 연구의 통제 방법은 검색과 분석을 위한 정보 시스템에 활용할 수 있다. 학술 정보 검색을 비롯한 특허 검색 등의 정보 검색이나 특정 분야의 추세 파악이 필요한 정보 분석 시스템의 경우, 누락되는 데이터를 줄이고 재현율을 높이는 것이 매우 중요하다. 이와 유사하게 재현율을 높이기 위한 연구로 Wu et al.(2008)은 충분한 자질 확보를 통해 기계 학습의 성능을 높이고자 Wikipedia의 인포박스(infobox) 내 존재하는 롱테일 영역의 데이터를 이용함으로써 정보 검색 시 재현율(recall)을 향상시키기 위한 기법을 소개하였다. 본 장에서는 재현율의 관점에서 기술 용어 네트워크의 생성 효과를 측정하고자 한다.

〈표 6〉과 〈표 7〉은 용어 빈도를 변화시키면서 재현율을 비교한 검색 실험 결과를 보여준다. 앞서 제시한 빈도별 용어 분포 통계인 〈표 3〉을 참고하여 장서 빈도를 기준으로 CF〉=500, 100, 50, 10, 5, 4, 3, 2, 1의 9개 대표 구간으로 나누어 이 용어로 초록을 검색했을 때의 마이크로 평균 재현율(micro-averaged recall), 매

크로 평균 재현율(macro-averaged recall)을 각각 비교 평가하였다. 각 구간 별로 'CF〉=500'의 96개(약0.1%)에서 'CF〉=1'의 90,292개(100%)까지 해당 기술 용어를 실험에 사용하였다. 추세를 잘 관찰하기 위해 저빈도어 구간은 CF=1, 2, 3, 4까지 모두 측정하였고, 용어의 수가 급감하여 개체 수가 적은 희소구간에서는 용어의 비율을 고려하여 몇 개의 대표 지점을 선정하고 측정하였다. CF〉=1은 발생 빈도 1 이상을 갖는 10,154개의 기술 용어를 모두를 사용하였다는 의미이다. 정확률(precision)은 '검색된 문헌 전체 중에서 몇 개가 적합한 문헌인가'를 의미하고, 재현율(recall)은 '적합 문헌 전체 중에서 몇 개가 실제 검색 되었는가'를 의미한다. 데이터베이스 내 구축된 초록에 대한 색인어로써 기술 용어가 부여되어 있는 상태이므로, 이 실험에서의 검색 정확률은 네트워크 생성 전과 후 모두 이론 상 100%라 할 수 있다. 따라서 본 연구에서는 정확률은 따로 언급하지 않는다. 재현율의 경우, 네트워크 생성 후에는 모든 용어가 통제되어 초록에 색인된 것으로 간주하여 용어 네트워크 기반의 재현율을 100%로 기준 설정하고, 이에 대한 상대적 성능을 측정하여 성능의 변화를 비교하였다. 기술 용어별 재현율을 측정한 결과의 일부 데이터 예시는 〈표 6〉과 같다. 마이크로 평균 재현율은 기술 용어 빈도의 총합을 적합문헌 수의 총합으로 나눈 결과이며, 매크로 평균 재현율은 각 용어별 재현율의 평균이므로 최우측 컬럼인 '기술 용어의 재현율(%)' 전체 값의 평균을 구하여 측정한다.

〈표 7〉은 롱테일 데이터의 활용률을 살펴보기 위한 재현율의 성능 측정 최종 결과이다. 기술용어의 빈도 구간별로 구분하여 마이크로, 매크로 평균 재현율 성능을 표시하였다. 실험 결과, 모든 구간에서 매크로 평균 재현율은 80% 중후반의 일정한 성능을 보여준 데 반해, 마이크로 평균 재현율은 고빈도어 영역인 CF〉=500에서는 87.94%를 기록한 후 저빈도 구간으로 갈수록 성능이 급격하게 저하되어 롱테일 데이

〈표 6〉 기술 용어의 재현율 측정(빈도 180-190대 일부 예시)

기술 용어	URI (용어 네트워크)	기술 용어 빈도	적합문헌 (URI 빈도)	기술 용어의 재현율(%)
CREB-binding protein	R1AU3IYNC7	188	189	99.47
large-eddy simulation	CVKOAMPSFX	188	248	75.81
ankle-brachial index	NDJK6E0MQO	189	194	97.42
corticosteroid-binding globulin	KENZ3SP7W1	189	202	93.56
methylation-specific PCR	MSB324J8OI	189	192	98.44
corticotrophin-releasing hormone	NUL3VHGMOF	190	1327	14.32
monocyte-derived macrophages	0RY8LQBT94	190	215	88.37
sister-chromatid exchanges	J9PRHKW485	190	474	40.08
fast spin-echo	Y0FWTHL5IR	191	195	97.95
anterior-posterior	H0DIGXB7LW	194	283	68.55

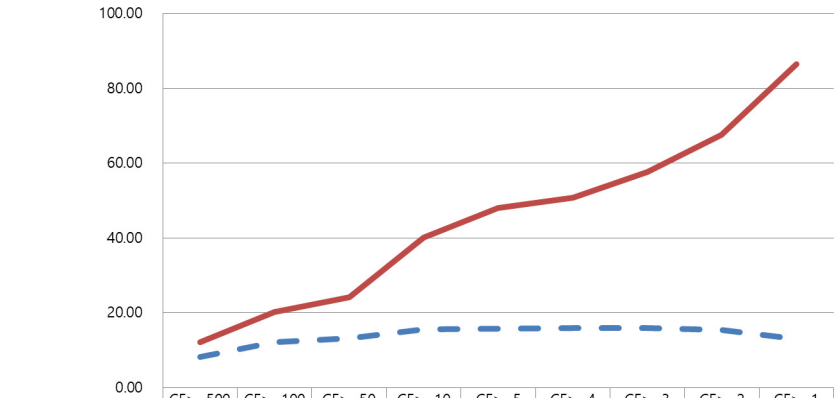
터 구간인 CF〉=1에서는 매우 저조한 13.51%의 재현율을 보여주었다. 구간 CF〉=1은 네트워크 생성 전의 원본 데이터를 의미하는 것으로, 네트워크 기반의 데이터 활용률 100%에 대비한 편차가 큰 만큼 네트워크 기반의 용어 통제 효과가 크다고 해석할 수 있다. 따라서 재현율을 기반으로 데이터 활용률을 용이하게 해석할 수 있도록, 롱테일 영역의 개선된 데이터 활용률(enhanced data utilization rate)은 다음과 같은 간단한 식으로 표현할 수 있다.

```
Enhaced Data Utilization Rate(EDUR)
 = Maximum Utilization Rate - Evaluated
   Performance Value
```

〈그림 15〉는 개선된 데이터 활용률의 추이를 나타낸다. 앞서 언급한 바와 같이 저빈도 기술 용어로 갈수록 매크로 관점의 데이터 활용률은 크게 개선이 없는 반면, 마이크로 관점의 데이터 활용률은 급격한 성능 향상을 보이고 있다. 매크로 평균 기반의 롱테일 용어 데이터의 활용률은 최고 15.96%까지 향상되었으며, 마이크로 평균 기반의 활용률은 최저 10% 대에서 최대 86.49%까지 비약적으로 개선되었다. 마이크로 평균 기반의 활용률이 상대적으로 큰 폭으로 향상되었다는 점은 실제 초록 데이터의 검색 시스템에서, 저빈도의 이형이 매우 많이 나타나는 특정 용어가 존재하며, 이 용어들이 마이크로 평균 재현율을 저해하는 요소로 파악된다. 또한 이러한 특징을 갖는 용어들이 검색어로 사용되었을 경우, 적합 문헌의 재현 성능에 영향을 끼칠 것으로 예상할 수 있으므로 특허 검색과 같은 재현율 보장이 필요한 시스템에서는 이러한 용어를 사전에 분석하고 파악하는 과정이 매우 중요할 것으로 보인다. 이와 관련한 상세한 논의는 추후 연구를 통해 진행하고자 한다. 이상의 실험을 통해 본 연구에서 제안한 네트워크 기반의 기술 용어 생성 및 통제 방법을 통해 롱테일 영역에 존재하는 데이터의 활용률이 기존 대비 향상됨을 확인하였다.

〈표 7〉 기술 용어의 빈도 구간별 재현율 및 데이터 활용률 비교

빈도 구간	기술 용어 종수	기술 용어 원형(%)		개선된 데이터 활용률(%)	
		Macro_Recall	Micro_Recall	Macro_EDUR	Micro_EDUR
CF〉=500	96	91.90	87.94	8.10	12.06
CF〉=100	544	87.95	79.76	12.05	20.24
CF〉=50	1,098	86.88	75.93	13.12	24.07
CF〉=10	5,125	84.49	59.84	15.51	40.16
CF〉=5	10,154	84.20	51.95	15.80	48.05
CF〉=4	12,745	84.04	49.21	**15.96**	50.79
CF〉=3	17,395	84.17	42.43	15.83	57.57
CF〉=2	28,948	84.65	32.42	15.35	67.58
CF〉=1	90,292	87.02	13.51	12.98	**86.49**



	CF>=500	CF>=100	CF>=50	CF>=10	CF>=5	CF>=4	CF>=3	CF>=2	CF>=1
Macro_EDUR (%)	8.10	12.05	13.12	15.51	15.80	15.96	15.83	15.35	12.98
Micro_EDUR (%)	12.06	20.24	24.07	40.16	48.05	50.79	57.57	67.58	86.49

〈그림 15〉 롱테일 영역에서 나타나는 개선된 데이터 활용률 추이

5. 결 론

본 연구는 빅 데이터 환경에서 간과되기 쉬운 롱테일 데이터의 활용률을 높이기 위해, 텍스트 마이닝 기법을 이용하여 자원을 구축하고 해석하는 방법론을 제안하는 것을 목표로 한다. 이를 위해 LOD 환경으로부터 수집한 대량의 데이터를 기반으로 정보 추출 기법을 이용해 기술 용어를 추출하고, 구축된 대량의 기술 용어들이 롱테일 현상을 보임을 통계적으로 확인하였다. 롱테일 영역의 저빈도 데이터의 활용률을 높이기 위해 텍스트 마이닝 기법을 이용하여 용어 네트워크를 생성하고 활용하는 방안을 제안하였다.

연구의 구체적인 내용으로 첫째, 대량의 데이터를 수집하기 위해 우선 DBLP 사이트로부터 학술정보 메타데이터를 1차 수집한 후 DOI 필드 정보를 활용해 웹으로부터 대량의 초록 정보를 2차 수집하였다. 둘째, 학술 논문에서 흔히 두문자 형태로 표기되는 기술 용어의 패턴을 찾아내는 용어 추출기를 개발하여 대량의 기술 용어를 자동 추출하였다. 셋째, 텍스트 마이닝의 편집 거리 기법을 이용해 학술 논문에 나타나는 기술 용어 간의 의미 네트워크를 자동으로 생성하는 효과적인 방안을 제시하고 생성 결과를 통계적으로 해석하였다. 마지막으로, 성능 평가 실험을 통해 기술 용어 네트워크를 활용하는 방법이 재현율을 향상시킬 수 있었으며, 롱테일 영역에 존재하는 데이터의 활용률을 높이는 데 효과적임을 확인하였다.

본 연구의 기여점은 다음과 같다. 첫째, LOD 정보를 이용해 대용량 학술 정보의 초록을 수집하는 과정을 설명하였다. 대량의 메타 데이터로부터 초록 데이터를 수집하는 과정에서 다양한 정보원을 활용하는 방안을 제시하였다. 둘째, 비정형 데이터인 초록 정보로부터 기술 용어를 추출하는 패턴 기반의 정보 추출 기법을 제안하였다. 데이터 패턴에 대한 분석을 통해 2차 가공된 정보를 생산하는 방안과 구체적인 알고리즘을 제시하였다. 마지막으로 텍스트 마이닝 기법

을 이용한 용어 네트워크의 생성 방법과 실험을 통한 활용 효과를 제시하였다. 네트워크 구조를 활용하여 기술 용어의 어형 통제를 자동화하는 방안을 제시하였고, 성능 평가를 통해 롱테일 데이터의 활용률을 비약적으로 향상시킬 수 있음을 확인하였다.

본 연구의 한계점은 첫째, 수집 데이터를 수백만 건 수준으로 충분히 확장하지 못한 점이다. 개별 웹 사이트를 탐색하여 정보를 수집하는 시간의 제약으로 초기 수집된 약 370만 건의 학술정보 중 무작위 접속을 시도하여 약 50만 건의 수준까지 데이터 셋을 구축하여 실험을 실시하였다. 향후, 활용 가능한 모든 데이터를 수집하여 빅 데이터 기반의 연구로 발전시킬 예정이다. 둘째, 본 연구를 통해 구축된 기술 용어의 발행(publishing) 시스템이 아직 공개되지 않은 점이다. URI 기반의 기술 용어 집합은 향후 공개를 통해 정보 검색, 정보 분석 연구 및 자동 분류, 의미 식별(word sense disambiguation; WSD) 등 관련 응용 연구에 활용될 수 있을 것으로 기대한다.

참 고 문 헌

안광모, 김윤석, 김영훈, 서영훈 (2013). Levenshtein 거리를 이용한 영화평 감성 분류. 디지털콘텐츠학회 논문지, 14(4), 581-587. http://dx.doi.org/10.9728/dcs.2013.14.4.581

황미녕, 조민희, 황명권, 정도헌, 성원경 (2011). 기술 용어의 용어지배값을 이용한 활용주기 모델링 방법. 한국정보과학회 학술발표논문집, 38(1C), 139-141.

Abe, A., & Tsumoto, S. (2010). Analysis of research keys as temporal patterns of technical term usage in bibliographical data. Lecture Notes in Computer Science book series (LNCS, volume 6496), International Conference on Active Media Technology AMT 2010, 150-157. https://doi.org/10.1007/978-3-642-15470-6_16

Cormode, G., & Muthukrishnan, S. (2007). The string edit distance matching problem with moves. ACM Transactions on Algorithms, NY, USA, 3(1), No.2. https://doi.org/10.1145/1186810.1186812

Fortune (2017). Apple just acquired this little-known artificial intelligence startup. Retrieved from http://fortune.com/2017/05/13/apple-lattice/

Gartner (2018). Dark data (Gartner IT Glossary). Retrieved from https://www.gartner.com/it-glossary/dark-data

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. Library Trends, 57(2), 280-299. https://doi.org/10.1353/lib.0.0036

Hwang, M. N., Cho, M. H., Hwang. M., Lee, M., & Jeong, D. H. (2014). Technical terms trends analysis method for technology opportunity discovery. Information, An International Interdisciplinary Journal, 17(3), 877-883.

Jain, P., Hitzler, P., Sheth, A. P., Verma, K., & Yeh, P. Z. (2010). Ontology alignment for linked open data. Lecture Notes in Computer Science book series (LNCS, volume 6496) ISWC 2010: The Semantic Web, 402-417. https://doi.org/10.1007/978-3-642-17746-0_26

Jeong, D. H., Hwang, M., & Sung, W. K. (2011). Generating knowledge map for acronym-expansion recognition. In the Proceedings on U- and E-Service Science and Technology (UNESST 2011), 287-293. https://doi.org/10.1007/978-3-642-27210-3_38

Jeong, D. H., Hwang, M., Kim, J., Jung, H., & Sung, W. K. (2013). Acronym-expansion recognition based on knowledge map system. Information, An International Interdisciplinary Journal, 12(A), 8403-8408.

Kim, J., Hwang, M., Jeong, D. H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree. Expert Systems with Applications and Statistical Feature Analysis, 39(2012), 12618-12625. https://doi.org/10.1016/j.eswa.2012.05.021

Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., & Han, J. (2014). A confidence-aware approach for truth discovery on long-tail data. Journal Proceedings of the VLDB Endowment, 8(4), 425-436. https://doi.org/10.14778/2735496.2735505

Noia, T. D., Mirizzi, R., Ostuni, V. C., Romito, D., & Zanker, M. (2012). Linked open data to support content-based recommender systems. Proceedings of the 8th International Conference on Semantic Systems, 1-8. https://doi.org/10.1145/2362499.2362501

Paulheim, H., & Fümkranz, J. (2012). Unsupervised generation of data mining features from linked open data. Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, No. 31. https://doi.org/10.1145/2254129.2254168

Reis, D. C., Golgher, P. B., Silva, A. S., & Laender, A. F. (2004). Automatic web news extraction using tree edit distance. Proceedings of the 13th International Conference on World Wide Web, 502-511. https://doi.org/10.1145/988672.988740

Veritas (2016). Veritas global databerg report finds 85% of stored data is either dark or Redundant, Obsolete, or Trivial (ROT). Retrieved from https://www.veritas.com/news-releases/2016-03-15-veritas-global-databerg-report-finds-85-percent-of-stored-data

Wikipedia (2018a). Long tail. Retrieved from https://en.wikipedia.org/wiki/Long_tail

Wikipedia (2018b). X-ray diffraction (redirection). Retrieved from

https://en.wikipedia.org/wiki/X-ray_crystallography

Wikipedia (2018c). High-performance liquid chromatography. Retrieved from https://en.wikipedia.org/wiki/High-performance_liquid_chromatography

Wikipedia (2018d). Edit distance. Retrieved from https://en.wikipedia.org/wiki/Edit_distance

Wu, F., Hoffmann, R., & Weld, D. S. (2008). Information extraction from Wikipedia: moving down the long tail. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 731-739. https://doi.org/10.1145/1401890.1401978

Zhang, C., Shin, J., Ré, C., Cafarella, M., & Niu, F. (2016). Extracting databases from dark data with deepdive. Proceedings of the 2016 International Conference on Management of Data, 847-859. https://doi.org/10.1145/2882903.2904442

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Ahn, K. M., Kim, Y. S., Kim, Y. H., & Seo, Y. H. (2013). Sentiment classification of movie reviews using levenshtein distance. Journal of Digital Contents Society, 14(4), 581-587. http://dx.doi.org/10.9728/dcs.2013.14.4.581

Hwang, M. N., Cho, M., Hwang, M., Jeong, D. H., & Sung, W. K. (2011). A utility cycle modeling method for technological terms based on term dominance value. Proceedings of the KIISE Conference, 38(1C), 139-141.