Journal of KIIT. Vol. 22, No. 4, pp. 17-22, Apr. 30, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.4.17

오픈 데이터셋 분류 및 검색을 위한 메타 데이터셋 생성 연구

윤서희*[1], 하고은*[2], 유석종**

Meta-Dataset Creation Method for Classification and Search of Open Datasets

Seo-Hee Yoon*[1], Go-Eun Ha*[2], and Seok-Jong Yu**

요 약

인공지능 학습에 필요한 오픈 데이터셋 제공 사이트들이 증가하면서 효과적인 데이터셋 검색에 대한 중요성이 커지고 있다. 현재 메타 데이터셋 연구는 특정 사이트에서 제공하는 데이터셋들에 초점을 맞추고 있으며, 다중 데이터셋 포털에 대한 통합 플랫폼 연구는 미미한 실정이다. 본 논문에서는 통합 데이터셋 검색 지원을 위해 메타 데이터셋 분류 및 생성 기법을 제안하고자 한다. 시스템 구현을 위해 국내·외 오픈 데이터셋 포털에서 제공하는 데이터셋을 수집하여 메타데이터를 생성하고, 딥러닝 모델로 기존 카테고리를 표준 카테고리로 분류 생성하였다. 다중 언어 검색을 지원하기 위해 파파고 API를 통해 한영 검색 도구를 구현하였고, 메타데이터에 대한 추가적인 분석 결과를 그래프로 시각화하였다. 본 연구를 통하여 오픈 데이터셋의 검색 효율성을 높일 수 있으며, 인공지능 학습 및 데이터셋 분석연구에 기여할 수 있다.

Abstract

With the increasing number of sites providing open datasets for AI learning, effective dataset search has become crucial. Current metadata research is focusing on datasets offered by specific sites, while research on integrated platforms for multiple datasets remains limited. This paper aims to propose techniques for metadata classification and generation to support integrated dataset search. To implement the system, datasets provided by domestic and international open dataset portals were collected to generate metadata, and deep learning models were used to classify existing categories into standard categories. A search tool for Korean-English was implemented through the Papago API to support multilingual search, and additional analysis results for metadata were visualized using graphs. Through this study, it contributes to artificial intelligence learning and dataset analysis research by increasing the search efficiency of open datasets.

Keywords

Meta-Dataset, LSTM, deep learning, classification, web crawling, morphological analyzer

* 숙명여자대학교 소프트웨어학부 학사과정
- ORCID[1]: https://orcid.org/0009-0006-4910-7185
- ORCID[2]: https://orcid.org/0009-0000-8006-5768
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Feb. 19, 2024, Revised: Mar. 12, 2024, Accepted: Mar. 15, 2024
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Womens's University,
Cheongpa-ro 100, 47-gil, Cheongpa-ro, Yongsan-gu, Seoul, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Ⅰ. 서 론

교육 및 연구 분야에서 인공지능에 대한 관심이 높아지면서 오픈 데이터셋을 활용한 데이터 분석에 대한 수요가 크게 증가하고 있다. 데이터셋은 특정 목적을 위해 조직되고 형식화된 관찰 결과의 모음[1]을 말하며, 대표적인 오픈 데이터셋 포털로 공공데이터포털[2], 국가통계포털, Data.gov[3], Kaggle[4] 등이 있다. 그러나 사용자는 원하는 데이터셋을 찾기 위하여 개별 오픈 데이터셋 포털을 방문해야 하며 분류 방법이 상이하여 데이터셋 획득에 많은 시간이 소요되는 문제점이 발생한다. 공공데이터포털은 공공기관이 보유하고 있는 데이터를 제공하며, Kaggle의 경우는 기업 또는 개인이 수집한 데이터가 대부분을 차지한다. 예를 들어, ‘문화관광’ 데이터셋을 탐색하는 경우, 사용자는 공공데이터포털에서 키워드 검색을 통해 11,797개를 검색한 후, Kaggle에 또다시 4,508개를 찾은 후 검토해야 한다. 이처럼 각 데이터셋 포털에서 개별적인 탐색 도구를 사용하여 탐색 과정을 반복 수행해야 한다.

사이트별로 고유한 분류 체계도 데이터셋 탐색의 어려움을 가중시키는 요인이다. 공공데이터포털은 ‘교육’, ‘국토관리’, ‘공공행정’, ‘재정금융’, ‘산업고용’, ‘사회복지’, ‘식품건강’, ‘문화관광’, ‘보건의료’, ‘재난안전’, 등으로 카테고리를 사용하고 있으나, Kaggle에서는 명확한 카테고리가 존재하지 않으며 대신 ‘Popular Culture’, ‘Culture and Humanities’ 등 키워드 태그를 제공하고 있다. 기존의 데이터셋 분류 및 검색 연구에는 데이터셋 검색 전용 시스템 연구[6]와 메타 데이터셋 자동 추출 연구[7], 공공데이터의 메타데이터를 분석한 연구[8]가 있으나, 대부분 단일 사이트 내의 데이터셋을 중심으로 수행되었다는 한계가 있다. 이외에도 데이터셋 검색 기법[9]에 대한 연구가 존재하나 질의 검색을 중심으로 데이터 통합을 기술하고 있다.

본 연구에서는 기존 연구를 확장하여 국·내외 다수의 오픈 데이터셋 포털에 적용가능한 통합 검색 시스템을 제안하고자 한다. 이를 위하여 웹크롤링으로 수집된 데이터셋들을 분류하여 메타 데이터셋을 생성하고 이에 기반한 통합 검색 방법을 구현한다. 본 연구의 대상 데이터셋 사이트로 국외의 Kaggle와 Data.gov을 포함하고, 국내의 공공데이터포털, 서울열린데이터광장[10], AI-Hub[11]를 포함하여 메타 데이터셋을 구축한다. 기존의 개별적인 카테고리 분류 체계를 딥러닝 모델인 LSTM을 사용하여 메타 데이터셋을 위한 표준 카테고리로 재생성하였다. 또한 Streamlit을 활용하여 메타 데이터셋 검색을 제공하는 웹사이트를 구현하였으며, Papago API를 사용하여 다중 언어 통합 검색을 제공하였다.

본 논문은 2장에서 기존의 오픈 데이터셋 활용 연구를 소개하고, 3장에서는 제안하는 메타 데이터셋 검색 플랫폼과 처리 과정을 기술한다. 4장에서는 구현 시스템의 수행 결과와 분류 성능 평가 지표를 제시하고, 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

2.1 메타 데이터셋 활용 연구

공개 데이터셋 제공 사이트가 증가하면서 데이터셋 검색 지원을 위한 메타데이터 분석 연구도 활발해지고 있다. W. Y. Choi 연구[6]에서는 데이터셋 항목의 특성을 고려하여 차별화된 가중치를 부여하는 방식의 데이터셋 검색 연구를 수행하였다. J. J. Jung[7]은 데이터셋 검색을 위해 텍스트, 이미지, 비디오 형태 별로 데이터들의 주요 특징들을 추출하고 군집화하여 메타데이터를 자동 추출하는 시스템을 제안하였다. H. L. Kim의 연구[8]에서는 메타데이터 분석을 통해 데이터 목록의 특성을 파악하고 서로 다른 데이터 사이의 연결 가능성을 판별하여 데이터의 품질을 개선하였으며, R. Miller[9]는 유사도 측정 방식을 활용하여 대용량 데이터 저장소에서 높은 정확도를 제공하는 데이터셋 검색 기법을 제안하였다. 기존 연구는 공통적으로 개별 사이트에서 초점을 맞추고 있으며 다중 사이트에서의 데이터셋 통합 검색에 대한 연구는 미흡한 실정이다.

2.2 선행 연구와의 차별성

표 1은 기존 연구와 제안 연구의 특징을 비교한 것이다. 기존 연구는 단일 데이터셋 포털을 대상으로 분석하였으나, 본 연구에서는 국내·외 다수의 데

이터 포털에서 수집한 데이터셋을 대상으로 진행하였다. 또한 개별 카테고리를 사용하는 기존 연구와 달리, 제안 연구는 딥러닝 모델을 사용하여 데이터셋을 통합 카테고리로 자동 분류하였으며, 데이터셋 통합 검색을 위해 다중 언어 검색을 지원한다.

표 1. 선행 연구와의 특징 비교
Table 1. Comparison with previous methods

Feature	Previous methods	Proposed study
Target site	Single site	Multiple sites
Multilingual search	not support	Korean, English
Integrated category	not support	deep learning classification

III. 메타 데이터셋 생성 및 검색

3.1 제안 시스템의 구성

그림 1은 본 연구에서 제안하는 메타 데이터셋 검색 시스템의 전체 구조 및 처리 과정이다.

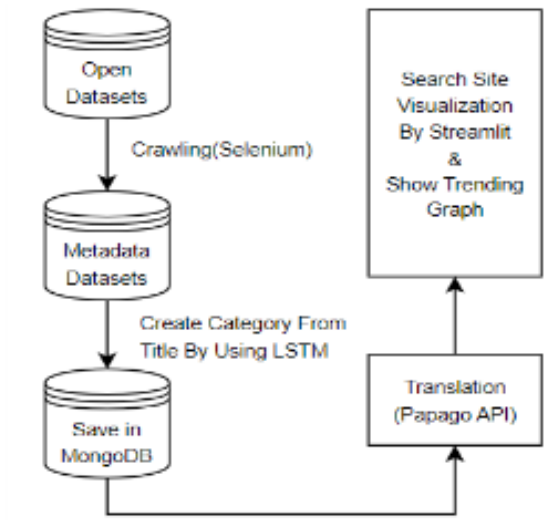


그림 1. 메타 데이터셋 검색 시스템 구조
Fig. 1. Architecture of metadata search systems

1) 오픈 데이터셋 사이트에서 웹크롤링을 통해 데이터셋과 메타 정보를 수집한다.

2) 수집한 메타 데이터를 바탕으로 LSTM을 사용하여 통합 카테고리로 자동 분류를 진행한다.

3) 분류한 데이터셋을 MongoDB에 저장한다.

4) Papago API를 사용한 검색어 번역을 통해 한영 변환 검색 기능을 제공한다.

5) Streamlit으로 사용자 인터페이스를 구현하여 데이터셋 검색과 데이터셋 시각화 분석 기능을 제공한다.

3.2 오픈 데이터셋 수집

데이터셋 수집을 위해 대표적인 오픈 데이터셋 사이트 중 국내(공공데이터포털, 서울열린데이터광장, AI-Hub) 3곳과 국외(Kaggle, Data.gov) 2곳을 선정하여 크롤링을 진행하였다. 이들 사이트는 데이터셋 다운로드가 가능하고 메타 데이터 정보가 공개되어 메타 데이터셋 구축에 적합하다. 웹크롤링은 웹 애플리케이션 자동화 테스트 프레임워크인 셀레니움(Selenium)을 활용하였다. 표 2는 각 사이트별로 크롤링한 최근 3년간 데이터셋 수치 자료이다 (2023년 11월 기준).

표 2. 크롤링한 데이터셋의 개수
Table 2. Number of crawling datasets

Public dataset portal	Number of datasets
Public Data Portal (PDP)	1720
Seoul Open Data Square (SDS)	343
AI-Hub (AIH)	510
Kaggle (KAG)	1545
Data.gov (DAG)	1717
Total	5835

크롤링한 총 데이터셋의 수는 5,835개로, 메타 데이터로 데이터셋 제목, url, 생성날짜, 수정날짜, 태그, 제공자, 조회수, 다운로드수, 포맷, 카테고리를 수집하였다.

3.3 데이터셋 전처리

카테고리 분류를 위해 크롤링한 메타데이터를 이용하여 전처리를 진행하였다. 국내 사이트, 국외 사이트로 분리하여 형태소를 분석하였고, 국내 사이트인 공공데이터포털, 서울열린데이터광장, AI-Hub은 KoNLPy 패키지를 사용하였다.

형태소 분석기의 성능 비교 연구의 결과[7]에 따라 공백 삽입 기능과, 신조어 처리가 가능한 Okt를 분석에 사용하였다. 국외 사이트인 Kaggle와 Data.gov는 NLTK를 사용하여 단어를 토큰화하고 불용어를 제거하였다. NLTK는 다양한 클래스와 함수를 응용하여 언어의 분류, 토큰화 등이 가능한 라이브러리로 영문 형태소 분석에 광범위하게 사용된다.

3.4 LSTM을 이용한 표준 카테고리 분류

본 연구에서는 데이터셋 카테고리 자동 분류 모델로 LSTM을 채택하였다. LSTM은 RNN 모델의 그래디언트 소실문제(Gradient Vanishing Problem)를 개선하기 위해 Hochreiter와 Schmidhuber가 제안한 머신러닝 모델이며, RNN 모델보다 형태소 기반 텍스트 데이터에서 높은 성능을 보인다[12]. 셀 값의 유지 기간을 결정하는 게이트를 통해 입력 길이에 상관없이 필요한 정보만 가질 수 있다는 장점이 있으며, 분류 및 예측 문제에 뛰어난 정확도를 제공한다[8].

모델 적용을 위해 형태소 분석과 원핫인코딩을 이용하여 입력 값에 대한 전처리를 진행하였다. 이후 각 포털별로 구축한 데이터를 8:2의 비율로 나누어 훈련, 검증 데이터셋으로 사용하였다. 이진 분류를 위한 출력층이 시그모이드 함수일 때 성능이 가장 높은 것을 고려하여 활성화 함수로 선택하였다. 모델을 최적화하기 위해 GridSearchCV를 이용하여 하이퍼 파라미터 튜닝을 진행하였다. 에포크는 50으로, 임베딩층은 256개로, 은닉층은 64개로 설정하였을 때 시험 데이터셋의 성능이 제일 높아 해당 파라미터를 최적으로 설정하였다.

IV. 실험 및 성능평가

4.1 데이터셋 분류 실험 결과

대상 데이터셋 사이트에 따라 데이터를 분류하는 카테고리명에 차이가 존재하였다. 각 포털 별로 Public Data Portal(PDP)는 17개, Seould Open Data Square(SDS)는 10개, AI-hub(AIH)는 14개의 고유 카테고리를 사용하고 있으며, Kaggle(KAG)과 Data.gov(DAG)는 별도의 카테고리가 사용하지 않는다. 데이터셋 분류를 위하여 기존 카테고리를 분석하여 12개의 통합된 표준 카테고리를 선정하여, LSTM으로 데이터셋을 자동 분류하였다. 표 4는 각 사이트별로 12개의 표준 카테고리에 매핑된 결과를 나타낸 것이다.

표 3. 표준 카테고리 분류 결과
Table 3. Mapping result by standard categories

Standard category	Number of datasets					Total
	PDP	SDS	AIH	KAG	DAG	
Agriculture	122	15	56	141	144	478
Disaster safety	110	35	27	144	152	468
Education	108	38	14	125	60	345
Environment	182	20	27	140	147	516
Finance	219	29	36	130	111	525
Food health	202	21	24	120	105	472
Healthcare	107	27	68	107	80	389
Law	127	36	34	126	456	779
Science technology	109	24	110	148	43	434
Social welfare	214	25	29	112	237	617
Transportation	118	52	65	131	94	460
Travel	102	21	20	121	88	352

표 4는 LSTM을 사용하여 데이터셋을 표준 카테고리로 분류한 정확도를 측정한 결과이다. PDP의 경우 82%로 가장 높은 정확도를 보여준다.

표 4. 분류 정확도 비교
Table 4. Comparison of classification accuracy

Dataset site	Train accuracy	Test accuracy
PDP	95%	82%
SDS	92%	78%
AIH	98%	74%
KAG	83%	60%

4.2 데이터셋 통합 검색 시스템 구현

데이터셋 통합 검색 시스템 구현을 위해 MongoDB에 카테고리 분류 결과를 메타 데이터셋으로 저장하였다. MongoDB는 NoSQL(Not Only SQL) 데이터베이스로 테이블 대신 도큐먼트를 사용

하여 유연한 확장이 가능하다는 장점이 있다. 사용자 인터페이스는 데이터 분석에서 시각화를 간편하게 출력할 수 있는 Streamlit을 통해 구현하였다. 데이터셋 검색 기능은 기간별 검색, 조회순, 다운로드순, 최신순 검색, 오픈 데이터셋 사이트별 검색, 제목 키워드 검색, 카테고리 검색을 제공한다(그림 2). 국내 및 국외 사이트 통합 검색을 위해 Papago API를 사용하여 한국어, 영어 데이터셋이 모두 결과로 출력된다. 또한 Kaggle 사이트에서는 데이터셋 처리 알고리즘 기반 검색이 지원되며, 표준 카테고리 기반 검색이 제공된다. 또한 검색 기준을 설정하여 조회순, 최신순, 다운로드순으로 검색할 수 있다.

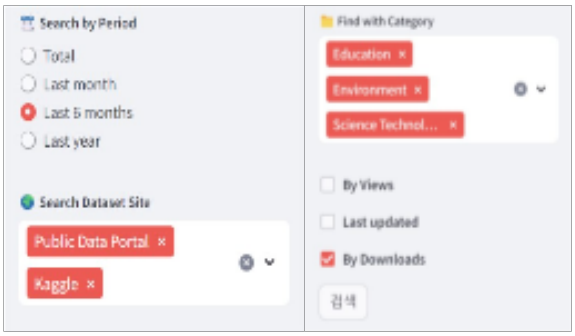


그림 2. 검색 사이드바
Fig. 2. Search sidebar

그림 3. 데이터셋 검색 결과 화면
Fig. 3. Screen of dataset search results

그림 3은 최근 6개월간의 데이터셋 중 교육, 환경, 과학기술 카테고리를 가진 캐글, 공공데이터포털 데이터셋을 다운로드순을 기준으로 검색한 결과 화면이다. 사용자는 출력된 테이블에서 데이터셋의 메타데이터 정보를 확인할 수 있다.

4.3 데이터셋 분석 결과

사용자는 메타데이터를 분석하여 시각화한 그래프를 생성할 수 있다. 그림 4는 전체 사이트에서 카테고리 상위 3개의 월별 데이터셋 개수를 그래프로 나타낸 것이다. '보건의료' 분류는 증가하다가 최근 급격히 감소하는 경향을 보이며, '사회복지', '교통물류'는 꾸준히 증가하는 패턴을 확인할 수 있다.

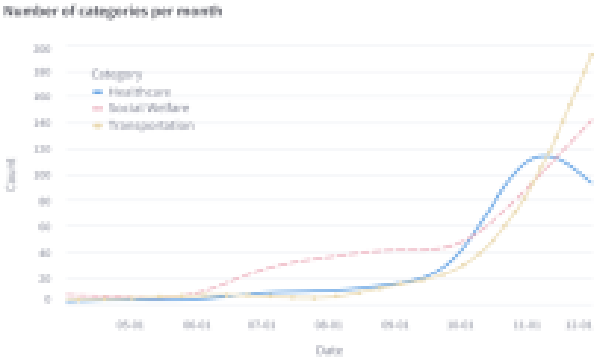


그림 4. 월별 상위 카테고리의 데이터셋 개수
Fig. 4. Number of datasets for top categories per month

V. 결론 및 향후 과제

본 연구에서는 다수의 오픈 데이터셋 포털에서 원하는 데이터셋을 검색하기 위한 메타 데이터셋을 제안하고 검색 시스템을 구현하였다. LSTM을 사용하여 데이터셋의 카테고리 자동 분류를 수행하였으며 시각화 분석 시스템을 개발하였다. 본 연구의 한계로 아직 데이터셋의 실시간 갱신 기능은 제공하지 않으며, 이에 대한 후속 연구가 필요한 실정이다. 또한, 5개의 지원 포털 외에 더 많은 데이터셋 사이트가 추가될 필요가 있다.

References

[1] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. Ibáñez, E. Kacprzak, and P. Groth, "Dataset search: a survey", The VLDB

Journal, Vol. 29, No. 1, pp. 251-272, Aug. 2019. https://doi.org/10.1007/s00778-019-00564-x.

[2] Korean Public Data Portal (data.go.kr), https://www.data.go.kr/en/index.do [accessed : Jun. 13, 2022]

[3] Data.gov, https://data.gov/ [accessed : Jul. 5, 2022]

[4] Kaggle, https://www.kaggle.com/ [accessed : May 14, 2022]

[5] D. J. Kim, H. J. Kim, C. E. Song, J. W. Yang, and H. L. Kim, "Methods for Utilising Local Government's Public Data Released to The Public Data Portal", Journal of Digital Contents Society, Vol. 22, No. 3, pp. 445-452, Mar. 2021. http://dx.doi.org/10.9728/dcs.2021.22.3.445.

[6] W. Y. Choi and J. H. Chun, "Dataset Search System Using Metadata-Based Ranking Algorithm", Journal of Broadcast Engineering, Vol. 27, No. 4, pp. 581-592, Jul. 2022. https://doi.org/10.5909/JBE.2022.27.4.581.

[7] J. J. Jung, K. W. Kim, and G. H. Kim, "A Study on Automatic Metadata Extraction to Support Dataset Search", The Journal of Korean Institute of Communications and Information Sciences, Vol. 2020, No. 8, pp. 867-868, Aug. 2020.

[8] H. L. Kim, "Metadata Analysis of Open Government Data by Formal Concept Analysis", The Journal of the Korea Contents Association, Vol. 18, No. 1, pp. 305-313, Jan. 2018. https://doi.org/10.5392/JKCA.2018.18.01.305.

[9] R. Miller, "Open Data Integration", Proceedings of the VLDB Endowment, Vol. 11, No. 12, pp. 2130-2139, Aug. 2018. https://doi.org/10.14778/3229863.3240491.

[10] Seoul Open Data Square, https://data.seoul.go.kr/ [accessed : Apr. 01, 2024]

[11] AI-Hub, https://www.aihub.or.kr/ [accessed : Apr. 01, 2024]

[12] G. Y. Kim, H. J. Kong, and S. J. Yu, "Legal Case-based Insult Sentence Analysis System using Natural Language Processing Techniques", Journal of KIIT, Vol. 21, No. 7, pp. 7-11, Jul. 2023. https://doi.org/10.14801/jkiit.2023.21.7.7.

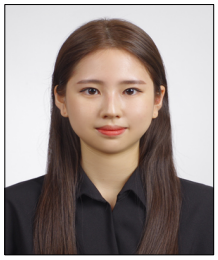
저자소개

윤 서 희 (Seo-Hee Yoon)

2021년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심분야 : 데이터 분석, 자연어처리, 인공지능

하 고 은 (Go-Eun Ha)



2021년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심분야 : 프로그래밍, 데이터 분석, 자연어 처리, 인공지능, 머신러닝

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 전산과학과(이학사)

1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)

2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)

2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수

관심분야 : 데이터마이닝, 추천시스템, 정보시각화