Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 5, pp. 406-416, October 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.5.406

메타데이터 간 유사도를 이용한 그래프 기반의 공공데이터맵 구축: 서울특별시 사례를 대상으로

최준혁[1] · 권민지[2] · 김준철[2] · 정준각[3†]

[1]포항공과대학교 산업경영공학과 / [2]서울기술연구원 데이터사이언스센터 / [3]한양대학교 산업융합학부

A Framework for Developing Public Data-map Using Similarity between Meta-data and Graph: The Case of Public Data from Seoul

Junhyuk Choi[1] · Minji Kwon[2] · Junchul Kim[2] · Junegak Joung[3]

[1]Department of Industrial and Management Engineering, Pohang University of Science and Technology
[2]Center for Data Science, Seoul Institute of Technology
[3]School of Interdisciplinary Industrial Studies, Hanayng University

The South Korean government is actively working to make data available to the public. However, as data from different departments is integrated and made accessible, efficient search algorithms for big data have become a major issue. This paper proposes a framework for developing a public data-map that uses metadata similarity and graph concepts to suggest ways to visualize and search related data. Additionally, to improve the performance of measuring similarity, we develop the domain-specific data pre-processing for public data and incorporate the step into the framework. To validate the framework, an empirical study was conducted using the case of the Seoul Metropolitan Government's Big Data Division. The results show that this framework can significantly improve the usability of public data and facilitate its open access.

Keywords: Public Data, Data-map, Word Embedding, Network Analysis

1. 서 론

공공데이터란 공공기관이 생성, 취득, 관리하는 모든 종류의 데이터를 의미하며, 대한민국 정부는 2013년에 공공데이터법을 제정하여 다양한 공공데이터를 적극적으로 민간에 공개하고 있다. 이를 위해 정부는 포털을 구축하여 데이터를 공개하고, 나아가 정부 부처 간 협력 증진과 민관 협력을 통한 사회문제 해결 지원 등의 새로운 부가가치 창출을 목표로 하고 있다. 또한 정부뿐만 아니라 서울특별시 등 각 지자체에서도 부서 간 따로 저장 및 관리되고 있는 공공데이터를 한곳에 모아 개방하는 사례가 많아지고 있다.

공공데이터 개방의 대표적인 사례로는 행정안정부에서 운영하는 공공데이터 개방 플랫폼인 공공데이터포털(DATA.go.kr)가 있다. 대한민국 정부 각 부처에서 보유하고 있는 여러 공공데이터를 한 곳에 모아서 제공한다. 2021년 10월 기준 해당 포털에 개방된 공공데이터는 총 6.5만 건이며 데이터 다운로드 및 신청 건수는 3,155만 건이다. 최근 국가중점데이터 46개의 분야를 공개하였으며, 정형데이터만 아니라 비정형데이터도 함께 개방하고 있다.

그 외에도 서울시는 열린데이터 광장(DATA.seoul.go.kr)을 통해 서울시의 시정 활동 중에 수집된 환경, 인구, 교육 등과 관련된 다양한 공공데이터를 민간에 공개하고 있다. 이를 통

이 논문은 한양대학교 교내연구지원사업, 한국연구재단, 서울기술연구원의 지원을 받아 수행되었음 (HY-202200000003529, NRF-2021R1I1A1A01044552, 2022-AH-001).
† 연락저자 : 정준각 교수, 04763 서울특별시 성동구 왕십리로 222 한양대학교 산업융합학부, Tel : 02-2220-2363, E-mail : june30@hanyang.ac.kr
2023년 6월 1일 접수; 2023년 7월 31일 게재 확정.

해, 공공기관과 민간 사용자 간의 연결 고리를 구축하고 데이터 기반의 비즈니스 기회 창출을 목표로 한다. 해당 서비스는 활용 목적에 따라 공공데이터의 유형을 분류하고, 기보유한 공공데이터를 환경, 안전 등 10개의 상위 분류와 49개의 하위 분류로 나눠서 공공데이터를 개방하고 있다. 또한, 사용자가 자신의 목적에 맞게 데이터를 쉽게 탐색하게 하도록, 활용도 높은 공공데이터를 20개 분야로 구분하여 제공하는 “인기그룹데이터” 기능을 제공한다.

최근에는 단순한 데이터 개방에서 나아가, 공공데이터 수요자의 편의성 향상을 위한 다양한 서비스가 함께 제공되고 있다. 예를 들어, 공공데이터 수요자는 사전에 분류된 데이터 유형에 따라 자신이 원하는 데이터를 탐색하거나, 검색어를 입력하여 원하는 데이터를 찾을 수 있다. 하지만, 수요자 측면에서 공공데이터의 실질적 활용성은 아직 부족한 상황이다(Song and Kim, 2022). 그 이유로는 공공데이터의 낮은 데이터 품질과 수요에 적합한 데이터의 부재도 있지만, 검색 기능이 부족하여 수요에 적합한 데이터 혹은 이와 연관성 높은 데이터를 탐색하기 어려운 점도 있다(Kim, 2021).

수요자의 목적과 연관성이 높은 데이터를 제공하기 위해, 기존 서비스에는 담당자가 공공데이터의 키워드를 직접 라벨링하고 라벨링된 키워드에 기반하여 연관성 높은 데이터셋을 제공한다. 하지만, 이는 키워드 라벨링 시간이 크게 소요되고 담당자 별로 키워드 라벨링이 상이하여 부정확한 결과를 초래할 수 있다. 그 이외에도, 공공데이터 분류체계를 개발하고 이에 따라 연관 데이터를 제공하고 있지만, 해당 방법 또한 담당자의 주관에 의존하여 공공데이터를 분류체계에 따라 분류해야하고 분류체계를 주기적으로 업데이트해야 하는 한계점이 있다(Kim *et al.*, 2019). 즉, 방대한 양과 다양한 종류의 공공데이터를 고려한다면, 위 방법들에 기반한 연관 데이터 추천은 비객관적이며 비효율적인 한계점이 있다.

담당자의 개입이 최소화되어 객관적으로 연관 데이터를 제공하기 위해서, 자연어 처리 기법을 활용하여 데이터명 혹은 생성 및 관리 부서 간의 유사도를 산출하는 방법이 있다. 하지만, 동일한 부서에서 방대한 양의 공공데이터가 공개되는 경우 해당 부서에서 수집된 공공데이터만을 연관 데이터로 추천하는 문제점이 있다. 또한, 일반적으로 쓰이지 않고 공공데이터에서만 쓰이는 단어들로 인해 정확한 유사도 산출이 어려운 한계점이 있다. 이러한 한계점을 극복하기 위해, 공공데이터맵 구축 시 공공데이터의 특성에 적합한 유사도 산출 방안 및 자연어 전처리 방안을 고려하여 라벨링된 키워드 없이도 실질적 연관성 높은 데이터를 제공해야 한다.

본 논문은 메타데이터 간 유사도를 활용한 그래프 기반의 공공데이터맵 구축 프레임워크를 제안하고, 이를 서울특별시 사례를 대상으로 실증 분석한다. 본 논문의 구성은 다음과 같다. 제2장은 공공데이터 개방 서비스의 국내외 현황과 이를 시각화하는 연구를 소개한다. 제3장은 공공데이터 간 유사도를 활용한 그래프 기반의 공공데이터맵 구축 방법론을 설명한다. 제4장에서는 제3장에서 제안한 프레임워크를 서울시 공공데이터에 적용하고 결과를 분석한다. 마지막으로 제5장은 본 연구를 요약하고 향후 연구 방향성을 소개한다.

2. 이론적 배경

2.1 공공데이터맵

공공데이터맵이란 공공데이터의 소재와 데이터 간 연관관계를 수요자가 직관적으로 이해하기 용이한 맵 형태로 시각화한 데이터 관계도를 의미한다(Kim, 2019). 공공데이터맵을 기반으로 수요자가 원하는 데이터를 용이하게 검색하고, 연관 데이터를 쉽게 탐색할 수 있다. 다양한 공공데이터 개방 서비스에서 수요자들의 편의성 향상을 위해 데이터맵 서비스를 제공하고 있다.

공공데이터포털(Data.go.kr)에서도 “국가데이터맵”이란 데이터맵 서비스를 제공한다. 해당 서비스를 통해 포털 내 모든 부서가 보유한 데이터의 소재 정보와 해당 데이터 간의 연관관계를 쉽게 검색할 수 있도록 시각화한 데이터 관계도를 의미한다. 사용자는 해당 서비스를 통해 분류된 유형에 따라 자신의 목적에 맞는 데이터를 찾거나, 검색어를 입력하여 탐색할 수 있다. <Figure 1>의 오른쪽 데이터맵은 “버스정류장 데이터베이스”라는 검색어를 입력하여, 해당 검색어와 관련성이 높은 약 20개의 데이터셋을 시각화한 맵이다. 산림빅데이터거래소에서도 공공데이터포털과 마찬가지로 “산림빅데이터맵”이란 기능을 통해 키워드와 유관한 데이터셋과 데이터셋 간의 관계를 그래프로 표현하여 제공한다. <Figure 2>와 같이 키워드와 유관한 데이터셋과 해당 데이터셋의 정보과 데이터맵 형태로 제공된다.

데이터맵 형태를 제공하기 위해선 데이터셋 간의 유사도 산출이 필수적이다. 유사도 산출을 위해서 기존에는 공공데이터포털의 담당자 혹은 공공데이터 제공자가 공개된 공공데이터에 적합한 키워드를 선정하여 라벨링을 한다. 예를 들어, ‘농림식품기술기획평가원_농림수산식품 식품・유통 R&D 과제 정보’ 라는 공공데이터의 키워드로 ‘전통식품’, ‘식품안전’, ‘식품가공’이 라벨링되었다. 담당자가 라벨링한 해당 키워드를 기준으로 연관성 높은 공공데이터가 제공된다. 공공데이터의 방대한 양을 고려한다면, 해당 방법은 키워드 라벨링에 소요되는 시간 및 비용이 큰 문제점이 있다. 또한, 동일한 공공데이터셋에도 담당자 별로 서로 다른 키워드가 라벨링될 수 있는 등 사람의 주관이 개입되어 결과가 객관적이지 않은 한계점이 있다.

이외에도 공공데이터포털에서는 사전에 공공데이터 분류체계를 개발하여 공공데이터셋을 분류체계에 따라 공개하고 있다. 해당 방법 또한 담당자가 개입하여 공공데이터셋에 적합한 분류를 선정해야 하며, 적합한 분류가 없는 경우 분류체계를 지속적으로 업데이트해야하는 어려움이 있다(Kim *et al.*,

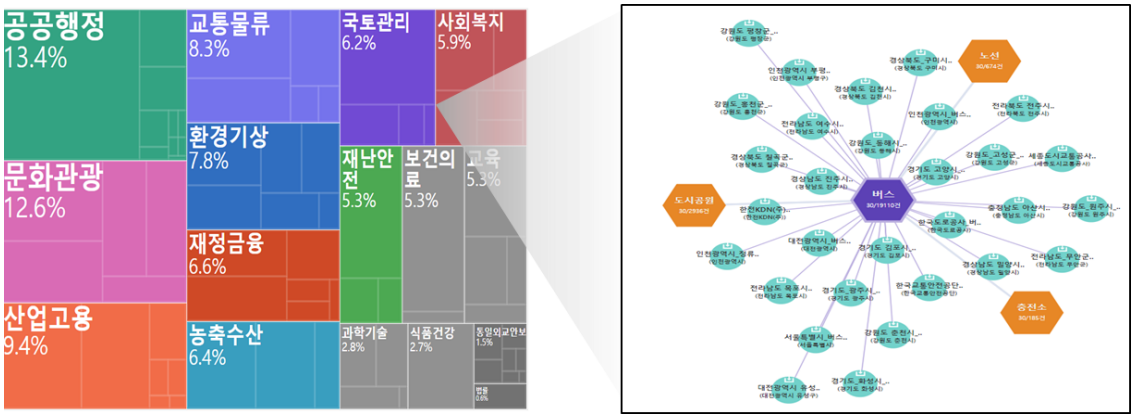


Figure 1. Data Map for "Bus Stops". Adapted from "DATA.GO.KR" by Ministry of the Interior and Safety, accessed 3 March, 2023, www.data.go.kr/tcs/opd/ndm/view.do.

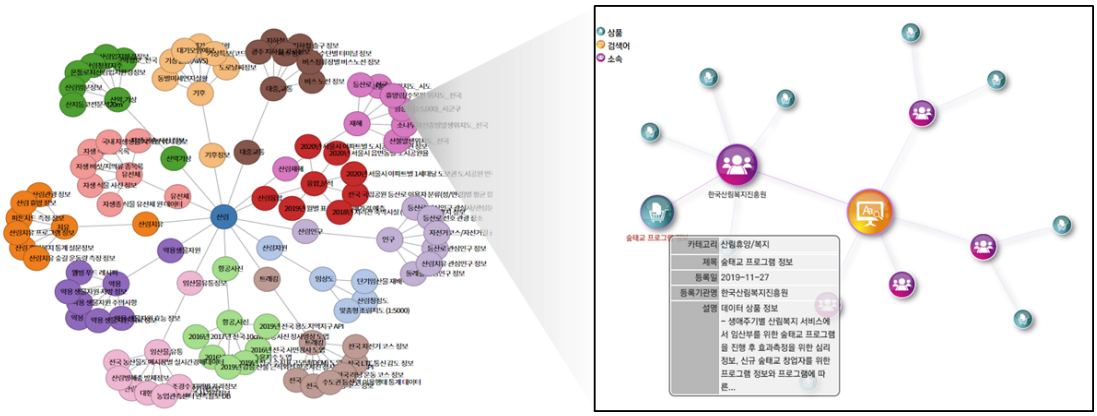


Figure 2. Data Map for Forest. Adapted from "Forest Big Data Exchange Platform" by Korea Forest Service, accessed 3 March, 2023, https://www.bigdata-forest.kr/service/alldatamap.

2019). 즉, 해당 방법 또한 사람의 개입이 필요하여 그 분석 결과가 비객관적인 한계점을 지닌다.

2.2 자연어 처리 기반의 유사도 산출 방안

사람의 개입이 최소화되고 객관적인 유사도 산출 결과를 위해 자연어 처리 기법이 주로 활용되고 있다. 이를 위해, 먼저 데이터의 태그 정보인 메타데이터(Meta-data)를 연산가능한 형태인 벡터로 표현한 후 벡터 간 유사도를 산출한다(Sakaji *et al.*, 2021). 이처럼, 텍스트 단어를 벡터로 표현하는 방법을 워드임베딩이라고 하며, 대표적으로 Word2Vec, GloVe, FastText 모델 등이 있다.

Word2vec 모델은 2013년 Google 연구팀이 제안한 모델로 대표적인 워드임베딩 모델이다(Mikolov *et al.*, 2013). Word2vec 모델은 신경망 모델을 활용하여 학습 데이터인 방대한 자연어 코퍼스에서 단어의 연관성을 학습한 모델이며, 단어 간 유사도를 산출하는데 주로 활용된다(Han *et al.*, 2018).

FastText는 2016년 Facebook 연구팀에서 제안한 워드임베딩 모델로, Word2vec의 skip-gram 과 CBOW(Continuous Bag Of Word) 아이디어를 발전시켜 개발되었다. 기존 Word2Vec에서는 학습 데이터 내 단어만 학습하는 것과 달리, FastText는 단어 내 Sub-word를 고려하여 형태학적인 특징을 분석한다(Bojanowski *et al.*, 2017). 즉, 각 단어 내 Sub-word인 N-gram으로 나눈 후, N-gram 벡터를 함께 학습하여 형태소적 특징을 포함하여 벡터화를 수행한다. 이와 같은 방법으로, 해당 모델은 학습에 포함되지 않은 단어인 OOV(Out of vocaburary)에 대해서도 벡터화를 수행할 수 있는 장점이 있다.

최근에는 기존 워드임베딩 모델에 Transformer 등 딥러닝 기법을 접목하여 대규모 말뭉치를 학습한 BERT(Bidirectional Encoder Representations from Transformers)와 GPT(Generative Pre-Training) 등의 대형 언어모델이 공개되고, 다양한 자연어 처리 분야에서 우수한 성능을 보이고 있다. 대표적으로 Google에서 개발된 BERT는 문장 내 관계와 문장 간 관계를 Birecdirectional 학습하여, 문장과 전체 텍스트 내 문맥을 고려하여 워드임베딩을 수행한다. 문맥을 고려하는 장점으로 인해 BERT 등 대형 언어모델은 질의응답 생성과 텍스트 내 감성 분류 등 문맥이 중요한 텍스트 분석에서 우수한 성능을 보이고 있다.

하지만, 문장 속 혹은 문장 간 문맥이 중요하지 않은 텍스트

분석 작업에는 BERT 등 대형 언어모델과 기존 워드임베딩 모델 간의 성능 차이는 크지 않다. 이러한 경향성은 실제 연구 결과에서도 확인할 수 있다(Chawla *et al.*, 2022; d'Saet *et al.*, 2020). 예를 들어, 단어 수준의 워드임베딩이 필요한 문서의 표절 여부 감지 작업을 대상으로 Fasttext와 BERT의 성능을 비교한 결과, Fasttext 모델이 BERT에 비해 더 우수한 결과를 보였다(Chawla *et al.*, 2022). 또한, 짧은 문장 내 혐오 표현 여부 작업을 대상으로 Fasttext와 BERT의 성능을 비교한 결과, 모델 간 성능 차이가 유의미하지 않았다(d'Sa*et al.*, 2020).

또한, Fasttext 등 기존 워드임베딩 모델은 대형 언어모델과 비교하여 학습 및 추론 시간이 작은 장점을 보인다. 예를 들어, 동일한 분석 환경에서 수행된 Text 분류 작업에서 BERT 모델은 Fasttext 모델의 추론 시간을 비교해본 결과, Fasttext 모델이 유사한 성능을 보이면서도 약 7배 높은 계산 효율을 보여주었다(Aksoy *et al.*, 2023). 즉, 기계 번역 혹은 질의응답 등의 문맥 파악이 필수적인 작업이 아닌 텍스트 분석에서는 Fasttext 등 기존 임베딩 모델과 최신 대형 언어모델의 성능 차이가 유의미하지 않고, 기존 워드임베딩 모델이 계산 효율적인 장점을 보인다.

3. 공공데이터맵 구축을 위한 제안 프레임워크

본 연구에서는 <Figure 3>과 같이 데이터 서비스 설계를 위한 그래프 기반 공공데이터맵 구축 프레임워크를 제안한다. 본 연구의 프레임워크는 크게 세 부분으로 나눌 수 있다. 제3.1절에서는 공공데이터의 특성을 반영한 텍스트 데이터 전처리 방안을 설명하고, 다음 제3.2절에서는 워드임베딩 모델을 활용하여 전처리된 텍스트에 대해 벡터화를 수행한 후 유사도를 산출하는 방안을 설명한다. 마지막으로, 제3.3절에서는 산출된 데이터 간 유사도를 기반으로 데이터맵을 구성하는 방안에 관해 설명한다.

3.1 텍스트 데이터 전처리

(1) 토큰화

텍스트 데이터 전처리를 위해, 먼저 KoNLPy 내 Okt 분석기를 활용하여 문장의 형태소를 분석하여 데이터명과 부서명을 토큰화한다. 이때, 토큰화란 주어진 문장 혹은 단어를 가장 작은 의미 단위인 "토큰"으로 나누는 작업을 의미하며, 한국어 토큰화에 주로 오픈소스 패키지인 KoNLPy가 활용된다(Park and Cho, 2014). 본 연구에서도 해당 패키지를 활용하여 토큰화를 수행하며, 영어가 아닌 한국어를 대상으로 분석하며 모든 품사의 토큰을 저장 후 분석을 진행한다.

(2) 공공데이터에 특화된 불용어 및 카테고리성 사전 구축

토큰화된 단어 중 빈번하게 출현하지만, 텍스트의 의미에는 기여하지 않는 불필요한 어휘를 불용어라고 한다. 불용어는 텍스트 간 유사도 산출 시 부정확한 결과를 초래하므로, 자연어 처리 시 불용어를 사전에 제거하는 작업이 필요하다(Kil, 2018). 예를 들어, 두 데이터셋인 "서울시 인구 정보"와 "서울시 하천 정보"는 실제로 연관된 데이터가 아니지만 "서울시"와 "정보"란 단어가 공통적으로 포함되어 유사도가 높게 산출된다. 즉, 유사도 산출의 정확도를 높이기 위해선, 불용어를 제거한 후 주제어인 "인구"와 "하천" 간 유사도만 산출해야 한다.

불용어 제거를 위해선 주로 자연어 처리 라이브러리에서 제공하는 한국어의 불용어 목록을 활용한다. 하지만, 불용어 목록은 도메인에 따라 달라질 수 있다. 예를 들어, 서울시 공공데이터와 다르게, 다른 분야에서는 "서울시", "정보"는 출현 빈도가 낮고 불용어가 아닐 수 있다. 그러므로, 본 연구에서는 공

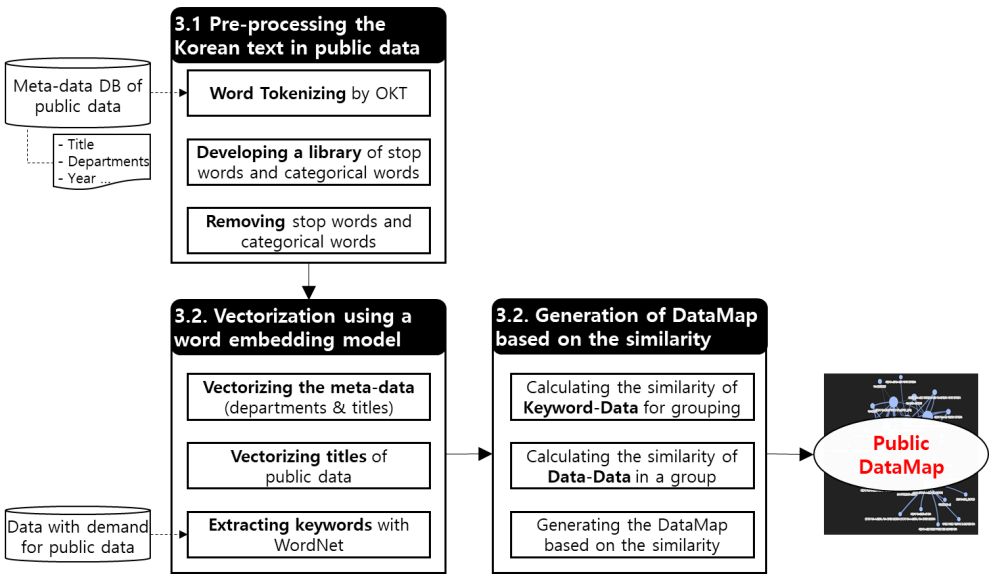


Figure 3. A Proposed Framework for generating the DataMap

공데이터에서 데이터명과 관리 및 생성 부서에서 출현 빈도수가 100회 이상인 단어 239개로 구성된 불용어 후보 목록을 생성하였다. 그 후, 해당 불용어 후보 목록을 서울기술연구소의 연구원과 본 논문의 저자들이 함께 검토하여 <Figure 4>와 같이 최종 불용어 사전을 구축하였다.

불용어와 마찬가지로 유사도 산출 시 부정적 영향을 미치는 카테고리성 단어도 선정 후 전처리한다. 공공데이터의 특성상 데이터명에 지역과 연도가 포함되는 경우가 많고, 이는 최종 데이터맵의 품질에 악영향을 미친다. 예를 들어, “서울시 광진구 모범음식점 지정 현황”, “서울시 금천구 모범음식점 지정 현황” 등의 데이터셋은 동일한 모범음식점 관련 정보이지만 지역구 별로 따로 저장 및 관리되고 있다. 이처럼 동일한 데이터가 지역과 연도 별로 따로 저장된다면, 모범음식점 키워드로 검색했을 때의 최종 데이터맵은 지역 별로 나눠진 동일한 데이터셋으로만 구성된다. 그러므로, 시각화된 데이터맵을 한눈에 파악하기 어렵고 동일한 공공데이터로만 구성되어 새로운 데이터 서비스 설계가 어려워진다. 이를 해결하기 위해서, 불용어와 마찬가지로 서울시 내 행정구, 행정동 목록과 연도 등의 카테고리성 단어를 선정하고 단어 사전을 구축한다.

(3) 불용어 제거 및 카테고리성 단어 전처리

토큰화된 텍스트 데이터 중 구축된 불용어 사전에 포함되는 토큰을 제거한다. 불용어를 단순히 제거하는 것과 달리, 카테고리성 단어는 제거 후 상위 데이터셋을 생성하는 작업을 추가한다. 예를 들어, “서울시 광진구 모범음식점 지정 현황”, “서울시 금천구 모범음식점 지정 현황” 와 같이 행정구별로 나눠진 데이터를 모두 제거하고, 해당 데이터를 포함하는 “서울시 모범음식점 지정 현황” 데이터를 생성하여 목록에 추가한다. 그다음, 기존에 제거된 데이터셋은 새로 생성된 데이터에 종속되도록 하위노드로 데이터맵에 시각화한다(제 3.3.3절 참조). 즉, 카테고리성 단어 사전에 포함된 제거한 뒤 상위 노드를 인위적으로 생성하고, 기존 카테고리성 단어가 포함된 데이터는 따로 저장하여 추후에 하위노드로 네트워크를 구성한다.

3.2 워드임베딩 모델 기반 벡터화

(1) 실국본부 및 데이터명 벡터화

본 절에서는 FastText 기반의 워드임베딩 모델을 활용하여 전처리된 토큰의 벡터화를 수행한다. Word2Vec 등 여러 워드임베딩 모델이 제안되었지만, 공공데이터의 특성상 데이터명 내에 공공행정에서만 쓰이는 용어가 포함되므로 OOV에 강건한 FastText 모델로 워드임베딩 모델을 수행한다. 예를 들어, 서울의 둘레길을 포함하는 산책로를 의미하는 “두드림길”은 일반적인 한국어 단어 사전에 포함되지 않기 때문에 기존 워드임베딩 모델에서 처리할 수 없지만, FastText 모델은 단어 내 형태소적 특징을 학습하여 벡터화를 수행한다.

또한, 공공데이터셋의 데이터명과 생성 및 관리 부서명은 문장이 아닌 단어 수준의 텍스트이므로, 해당 텍스트 데이터는 문장 혹은 전체 문맥이 포함되어 있지 않은 특성이 있다. 그러므로, 문맥이 고려되지 않은 텍스트 분석에서 Fasttext 모델이 보인 우수한 실험 결과를 고려했을 때, 본 연구에 BERT 등 대형언어 모델 보다 Fasttext 모델이 보다 더 적합하다고 판단하였다, 또한, 계산 효율성 측면에서도 Fasttext 모델로 워드임베딩을 수행하는 것이 적합하다고 판단하였다. 본 연구에서는 한국어 Wiki를 사전 학습한 FastText 모델을 활용하여 벡터화를 수행한다(Grave *et al.*, 2018).

공공데이터의 태그 정보, 즉 메타데이터에는 생성 부서를 의미하는 실국본부와 데이터명이 포함되어 있다. 실국본부와 데이터명을 구분하지 않고 유사도를 산출한다면, 하나의 실국본부에서 많은 데이터가 생성 및 관리되기 때문에 유사도 산출 시 실국본부에 의존도가 높아진다. 하지만, 부서 간 협력을 통한 새로운 데이터 서비스 설계를 위해서는 동일 부서가 아닌 서로 다른 부서 간 데이터 활용이 필요하다. 이를 위해, 본 연구에서는 부서명과 데이터명을 각각 따로 벡터화한 후에 추후 가중치를 두어 최종 유사도를 산출한다.

(2) 키워드 추출 및 벡터화

약 2만 건이 넘는 공공데이터를 한눈에 보이는 데이터맵으로 표현하는 것은 현실적으로 어려우므로, 키워드를 선정한 후 키워드별로 데이터맵을 제공해야 한다. 기존 공공데이터 포털에서도 공급자 관점에서 분류체계를 개발하고 이에 따라 데이터맵을 제공하고 있다. 예를 들어, 공공데이터포털(Data.go.kr)은 정부기능분류모델을 기반으로 하는 업무 중심의 분류체계에 따라 16개의 상위 키워드(공공행정, 과학기술, 교통물류 등)를 선정하고 이에 따라 데이터맵을 제공하고 있다. 산림청 빅데이터거래소(bigdata-forest.kr)에서는 산림재해, 산림휴양, 산림자원 등의 산림 데이터의 특성에 적합한 키워드를 선정하고 이에 따라 데이터맵을 제공한다.

기존 공공데이터 분류체계를 참조하여 키워드를 선정할 수 있지만, 본 연구에서는 실증 대상 공공데이터의 활용성을 제고하기 위하여 공공데이터 활용 측면을 고려하여 키워드를 선정하였다. 공공데이터 활용 관점에서 서울시 산하기관의 공공데이터는 일반적인 정보 공개 외에도 서울시의 사회 이슈 해결 및 정책 추진 과정에서 활용 가능성이 높다(Kim *et al.*, 2014). 즉, 공공데이터의 수요자 관점에서 활용 범주는 서울시의 해결과제 및 추구하는 미래 상과 연결될 수 있다(Kim *et al.*, 2014). 예를 들어, 서울시의 핵심 정책 중 하나인 “주택 공급 확대”와 관련된 공공데이터가 데이터맵으로 제공한다면, 공공데이터의 수요자들을 해당 공공데이터를 활용하여 새로운 아이디어를 발굴할 수 있다. 이렇게 발굴된 아이디어는 서울시의 핵심과제와 연결되어있으므로, 제품·서비스로 현실화될 가능성이 크다.

공공데이터 활용 측면을 고려하여 키워드를 선정하기 위해, 본 연구에서는 서울시의 정책 방향 및 향후 전략이 포함된 “Seoul Vision 2030 핵심과제” 문서를 활용한다. “Seoul Vision

선정된 불용어 및 카테고리성 단어

	단어	출현 빈도		단어	출현 빈도
1	서울시	6888	20	서울	281
2	정보	5808	21	역	270
3	인허가	3117	22	도시	268
4	현황	2330	23	계획	266
5	통계	1470	24	서비스	247
6	별	1313	25	이용	244
7	업	1158	26	운영	242
8	시설	862	27	대한	227
9	관리	796	28	인구	223
10	대장	401	29	평균	222
11	판매업	396	30	노선	218
12	사업	338	31	주택	218
13	구별	327	32	이력	210
14	유치원	320	33	강북구	207
15	식품	320	34	일반	197
16	및	319	35	도봉구	193
17	물	314	36	강동구	193
18	위치	298	37	측정	190
19	코드	284	38	강남구	184
			39	성동구	183

Figure 4. Example of Stop-words or Categoric Words

2030 핵심과제"는 서울시의 향후 10년 시정 운영 방향을 20개의 핵심과제로 정리하여 공개된 문서이다. 즉, 해당 문서에 포함된 핵심과제에 기반하여 키워드를 추출한다. 먼저, '서울시'의 공공데이터에 적합한 키워드 선정을 위해 "Seoul Vision 2030 핵심과제" 내 20개의 핵심과제 내 키워드를 <Figure 5>와 같이 추출한다. 그 후, 유사도 산출의 정확도를 높이기 위해서 각 키워드별로 유의어를 탐색하여 키워드를 확장한다. 유의어 탐색을 위해선 NLTK 내 Wordnet을 활용한다. 예를 들어, 핵심과제 내 "주택" 이란 키워드가 있다면, "주택"의 유의어인 "주거", "건설" 등을 탐색하여 이를 키워드 List에 추가한다. 이러한 과정을 거쳐, 20개의 그룹별로 확장된 키워드 List가 도출된다. 마지막으로, 제 3.2.1절과 동일하게 FastText 모델을 활용하여 각 그룹 별 키워드를 벡터화한다.

본 연구에서는 서울시 산하기관에서 공개된 공공데이터를 대상으로 실증 연구를 수행하므로 "Seoul Vision 2030 핵심과제" 문서에 기반하여 키워드를 선정하였다. 서울시뿐만 아니라 대부분의 공공기관은 핵심과제와 향후 마스터플랜을 정리하여 문서로 공개하고 있다. 추후, 다른 분야에 본 연구의 프레임워크를 적용 시, 공공데이터 공급자의 핵심과제 및 향후 전략 등이 포함된 문서를 확보하여 동일한 작업을 수행할 수 있다.

3.3 유사도 기반 네트워크 구성

(1) 키워드별 공공데이터 그룹화

공공데이터맵 구축을 위해 먼저 키워드와 공공데이터 간 유사도를 산출하여 공공데이터를 그룹화를 수행한다. 모든 공공데이터간 유사도를 산출하여 그룹화를 수행할 수 있지만, 계산 부하가 크고 비효율적이므로 사전에 선정한 키워드를 기준으로 그룹화를 수행한다.

공공데이터와 키워드 간 유사도 산출 및 그룹화 알고리즘은 아래 Pseudo Code 1과 같다. 먼저, 공공데이터의 데이터명과 키워드 간 유사도를 산출한다. 제3.2절에서 벡터화된 데이터명과 키워드 각각에 대해 Cosine 유사도를 산출한 후 평균을 구한다. 이때, 공공데이터별로 데이터명 내 토큰의 개수가 다르므로, 키워드와 모든 토큰과의 유사도를 산출한다면 평균에 왜곡이 발생할 수 있다. 이러한 왜곡을 줄이기 위해, 모든 토큰과의 유사도를 산출한 후 상위 5개의 평균을 구해서 이를 키워드와 데이터명 간 유사도로 정의한다. 실국본부와 키워드도 동일한 방식으로 상위 5개의 평균을 산출한다. 그 후, 데이터명-키워드 유사도와 실국본부-키워드 간 가중치 평균을 산출한다. 최종적으로 공공데이터와 각 그룹 키워드 간의 유사도가 산출되면, 가장 유사도가 높은 그룹으로 해당 공공데이터를 분류한다. 모든 공공데이터에 대해 동일한 작업을 수행하여, 각 공공데이터를 그룹별로 분류한다. 단, 모든 그룹과의 유사도가 Threshold를 넘지 않는 공공데이터는 분석에서 제외하며, 해당 Threshold는 실험을 통해 선정한다.

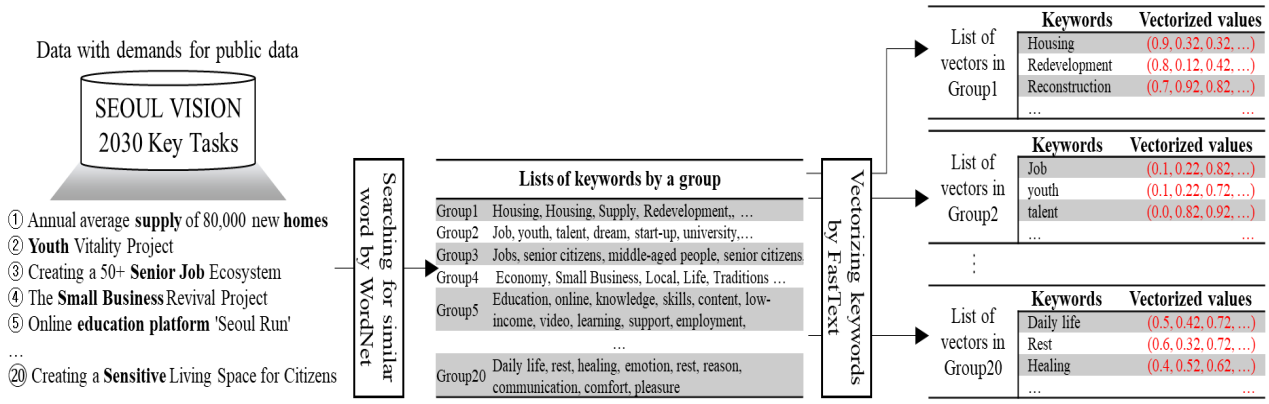


Figure 5. Keyword Extraction Process According to Seoul Vision 2030

Pseudo Code 1 : Selecting the group id for a public data

input :
Vector representations for the title and departments of a public data, $D = (DT, DP)$
Set of vector representations of keywords, $K = \{K_1, K_2, ..., K_{20}\}$,
A weight of the title against departments, $w \in (0, 1)$
output : Selected group id g^*, $g^* \in \{1, ..., 20\}$

// Calculate a similarity between each keywords
for each K_g, $g \in \{1, ..., 20\}$ **do**
  // Calculate data title similarity
  for each vector in DT, $t \in \{1, ..., T\}$ **do**
    for each vector in K_g, $e \in \{1, ..., E\}$ **do**
      Calculate a cosine similarity $cos(DT_t, K_{g,e})$;
      Add $cos(DT_t, K_{g,e})$ to L_{title}, a list of similarity ;
    end
  end
  Get L^*_{title}, the top five items with the highest value in L_{title} ;
  $SimTitle_g \leftarrow$ average(L^*_{title}) ;

  // Calculate data department similarity
  for each vector in DP, $p \in \{1, ..., P\}$ **do**
    for each vector in K_g, $e \in \{1, ..., E\}$ **do**
      Calculate a cosine similarity $cos(DP_p, K_{g,e})$;
      Add $cos(DP_p, K_{g,e})$ to L_{Dpt} a list of similarity ;
    end
  end
  Get L^*_{Dpt}, the top five items with the highest value in L_{Dpt} ;
  $SimDpt_g \leftarrow$ average(L^*_{Dpt}) ;

  $Sim_g = w$ x $SimTitle_g$ + (1-w) x $SimDpt_g$
  Add Sim_g to S, a list of similarity;
end
$g^* \leftarrow$ the id of maximum value in S ;
return (g^*)

(2) 그룹 내 공공데이터 간 유사도 산출

그룹별 공공데이터맵 구성을 위해서, 각 그룹 내에 속한 공공데이터 간 유사도를 산출한다. 공공데이터 간 유사도 산출 방안은 제 3.3.1절의 공공데이터-키워드 유사도 산출 방식과 동일하다. 벡터 간 Cosine 유사도를 활용하여, 데이터명 간 유사도와 실국본부 간 유사도를 산출한 후 이를 가중치 합하여 최종 유사도를 산출한다.

(3) 유사도 기반 공공데이터맵 구축

본 절에서는 그래프 분석 및 시각화에 주로 활용되는 Python 내 NetworkX와 PhyVis 라이브러리를 활용하여, 공공데이터 간 유사도를 기반으로 그룹별 공공데이터맵을 구축 및 시각화한다. 그래프 형태인 공공데이터맵 구축을 위해 각 공공데이터를 노드로 설정하며, 해당 노드에 공공데이터의 데이터명과 실국본부 정보를 추가한다. 제 3.3.2 절에서 산출한 공공데이터 간 유사도를 기반으로 간선으로 그래프를 모델링하여 노드 간 연결 그래프를 구축한다. 이때, Threshold보다 작은 유사도는 연결 그래프에서 제외하고 유사도 값이 클수록 두꺼운 간선으로 연결한다.

4. 실증 분석: 서울특별시 공공데이터의 데이터맵 구축

4.1 데이터 수집

본 연구에서는 서울기술연구원의 지원을 받아 서울특별시에서 생성 및 관리하는 공공데이터 총 20,026건을 수집하여 제안 프레임워크를 실증 분석하였다. 2022년 6월 말 기준으로 공공데이터의 메타데이터를 수집하였으며, 운영시스템 별로는 통합저장소(빅데이터서비스플랫폼)에서 11,134건, 디지털시장실에서 307건, 서비스팀(열린데이터광장, 빅데이터캠퍼스)에서 7,373건과 통계조사팀(통계정보시스템)에서 1,212건을 수집하였다. 공공데이터의 메타데이터에는 “데이터명”과 “실국본부”(데이터 생성 및 관리 부서) 이외에도 “관리자”, “갱신일”, “데이터 형식” 등의 정보가 포함되어 있지만, 데이터 유사도에 기여도가 낮으므로 분석에서 제외하였다. 최종적으로 수집 데이터의 컬럼 중 “데이터명”과 “실국본부”만 분석에 활용하였다. 자세한 수집 데이터의 예시는 <Figure 6>과 같다.

4.2 데이터맵 시각화를 위한 하이퍼파라미터 설정

데이터맵 시각화를 위해 먼저 총 3가지의 하이퍼파라미터인 그래프 내 최대 공공데이터 개수, 데이터-실국본부 간 가중치, 유사도 Threshold 를 설정해야 한다. 그래프 내 최대 공공데이터 개수가 적으면 연관 데이터셋을 탐색하기 어렵고, 반대로 많으면 그래프가 복잡해져서 사용성이 감소한다. 데이터명-실국본부 간 가중치가 낮으면 실국본부가 유사한 데이터셋으로 공공데이터맵이 구성되며, 반대로 높으면 실국본부가 다르더라도 데이터명이 유사한 데이터셋으로 공공데이터맵이 구성된다. 마지막으로, 유사도 Threshold가 낮으면 유사도가 낮은 데이터셋도 공공데이터맵에 시각화되므로, 사용자의 수요와 관련된 데이터셋의 수가 적은 경우에는 유사도 Threshold를 낮게 설정하여 최대한 많은 수의 공공데이터셋을 시각화하는 것이 유리하다.

실제 공공데이터맵 서비스 제공 시, 해당 하이퍼파라미터는 사용자가 직접 조정해가면서 선호도에 맞게 최적화할 수 있다. 하지만, 공공데이터 사용자의 편의성을 증진하기 위해, 본 연구에서는 실험과 정성적인 평가를 통해 하이퍼파라미터의 기본값을 선정하였다. 구체적으로는, 다양한 설정값 조건의 하이퍼파라미터 하에서 여러 버전의 공공데이터맵을 생성한 후, 서울기술연구소의 연구원과 저자들이 이를 검토하고 자체 만족도를 평가하여 최종 기본값을 선정하였다. 먼저, 데이터맵 내의 공공데이터 수를 최대 100개로 선정하였다. 약 2만 건인 공공데이터를 20개의 키워드로 그룹화 시, 그룹당 약 1,000개의 공공

구분	실국본부	부서명	대분류	중분류	분류	운영시스템정보	서비스	원천	컨텐츠	서비스데이터명	광역시	시군구	행정동	최초등록일	갱신주기	최종갱신일	제공데이터형식(/	담당팀	담당자
1	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_CNTRWK_MTRIL_STNDRD		공사 자재 규격				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
2	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_CNTRWK_PBCWRK_STNDRD		공사 토공 규격				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
3	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_CNTRWK_SNPS_EQP_STNDRD		공사 부설공 규격				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
4	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_CNTRWK_SNPS_STNDRD		공사 부대공 규격				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
5	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_SM_BDONG		요금시스템 법정동 코드				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
6	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_SM_HDONG		요금시스템 행정동 코드				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
7	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_SM_MIDDLE_BLCK		요금시스템 중블록 코드				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
8	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TC_SM_SMALL_BLCK		요금시스템 소블록 코드				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
9	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TH_CLS_SCTAP_HIST		급수전폐전이력				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
10	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TH_EDC_MTC_HIST		교육회의이력				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
11	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TH_LGZ_FCLTS_CHK_MEND_HIST		대형시설물 점검보수이력				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
12	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TH_LGZ_FCLTS_MNTNCE_HIST		대형시설물 유지보수이력				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
13	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TH_LGZ_FCLTS_SAFECHK_HIST		대형시설물 안전점검이력				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
14	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TH_REDEV_WTLREC_HIST		재개발재건축누수복구이력				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
15	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TH_SVRPL_RECLS_HIST		잔존관로철거폐쇄이력				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
16	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_BSNS_DETAIL_PLAN		사업 상세계획				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
17	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CHCK_MEND_STRCTU_INFO		점검보수구조물정보				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
18	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_ATCHMNFL		공사 첨부파일				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
19	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_CNTRCT_INFO		공사계약정보				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
20	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_COMP_BAS		공사 산출근거				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
21	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_FLAW_MENDNG		공사 하자보수				2021-10-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱		
22	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_MANAGE		공사 관리				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
23	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_MTRIL		공사 자재				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
24	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_PBCWRK		공사 토공				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
25	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_SNPS		공사 부대공				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
26	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_SNPS_EQP		공사 부설공				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
27	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CNTRWK_STRCTU		공사 구조물공				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
28	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CTK_SPP_CRS_CP_VAL		공사 직관횡단면도산출값				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
29	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_CURRENT_CAP		급수공사화면저장				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
30	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_DEPT_INFO		부서 정보				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
31	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_DIUSER_PLINE_INFO		불용관정보				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
32	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_ENLF_PIPE_SETUP		노후관설정				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
33	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_ENTRPS_CNTRWK_MNGR		공사업체관리				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
34	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_ENTRPS_COST_MANAGE		업체단가관리				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
35	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_ENTRPS_MNGR		업체관리				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
36	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_FCLTS_MNTNCE		시설물유지보수				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
37	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_FLTBED_INFO		여과지정보				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
38	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_FLTPLT_FCTRY		정수장공장				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
39	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_FLTS_RELATE_DTA		시설물 관련자료				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
40	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_LYR_MANAGE		레이어 관리				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
41	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_MAIN_DSCSS_CS_FLFL		주요협의 내용 이행				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
42	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_MIDDL_BLCK_FCLTY		중블록 시설				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱
43	상수도사업본부	공간정보과				빅데이터서비스플랫폼	원천테이블	TN_MIDDL_BLCK_MNGR		중블록 관리자				2021-04-23 0:00	월	2022-03-18 0:00	csv,xml,xlsx,json	빅데이터인프라팀	오재욱

Figure 6. An Example of Collected Data

데이터가 포함되어 수요자가 한눈에 파악하기 어려운 문제점이 있었다. 또한, 제 3.3.2절의 데이터명-실국본부 간 가중치를 0.7로 설정했을 때, 이종 실국본부 간 유사한 공공데이터 탐색이 용이하다고 판단되었다. 마지막으로 유사도 Threshold의 기본값 설정을 위해, 유사도가 0,5~0.8인 공공데이터 List를 탐색한 결과 0.8로 설정했을 때 가장 만족도가 높았다.

4.3 키워드별 공공데이터맵 구축 결과

본 연구에서는 제안 프레임워크를 활용하여 키워드별 공공데이터를 그룹화한 후 각각의 공공데이터맵을 구축하고 시각화하였다. <Figure 7>은 "안전"과 관련된 공공데이터를 데이터 간 유사도에 기반하여 그래프로 시각화하여 나타낸 것이다. 각 노드는 하나의 공공데이터를 의미하며, 서로 연결된 공공데이터는 관련성이 높은 공공데이터임을 의미한다. 예를 들어, "화재 발생 정보"와 "원인별 화재 발생"과 같이 관련성이 높은 공공데이터는 공공데이터맵에서 가까이 위치하게 된다. 또한, 다른 노드와 연결된 간선이 많은 공공데이터는 다른 데이터와 관련성이 높고 활용도가 높은 데이터라고 해석할 수 있다.

주목할만한 점은, 유사도 기반으로 그래프가 모델링 되어 있으므로 공공데이터맵 내의 데이터는 모두 "안전"과 관련된 데이터이지만 더 유사한 관계에 있는 데이터는 밀접하게 모여

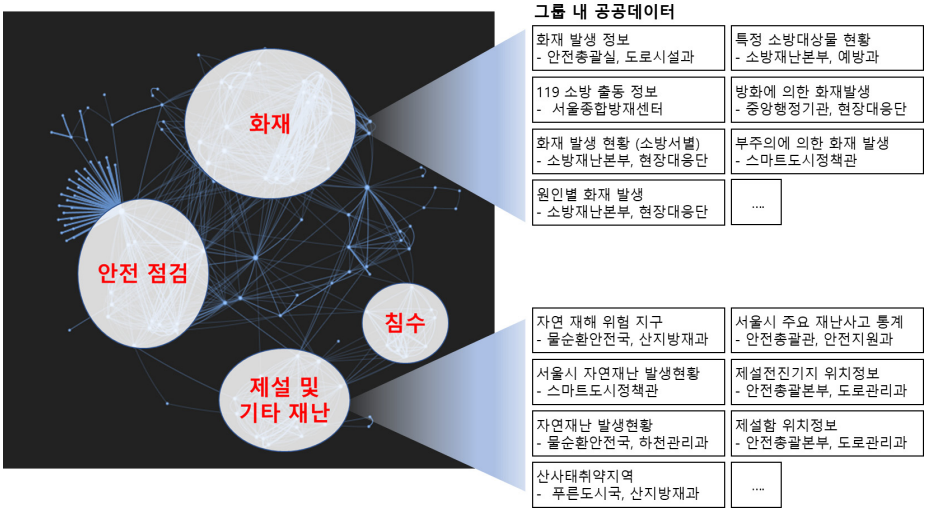


Figure 7. A Result of Datamap for the "Safety"

있다는 것이다. <Figure 7>에서도 크게 “화재”, “안전 점검”, “제설 및 기타 재난”, “침수”라는 4개의 밀접한 그룹이 있다는 것을 확인할 수 있다. 예를 들어, “화재” 그룹에는 “화재 발생 정보”, “119 소방 출동 정보” 데이터가 포함되어 있고, “제설 및 기타 재난” 그룹에는 “자연 재해 위험 지구”, “제설전진기지 위치정보” 등이 포함되어 있다. 이처럼 서로 관련성 높은 데이터가 밀접하게 모여있는 것을 보아, 데이터 간 유사도 산출의 성능이 높다고 해석할 수 있다.

또한, 동일한 그룹에 포함되어 있더라도 서로 다른 실국본부에서 생성 및 관리하고 있는 데이터가 포함되어 있음을 알 수 있다. 즉, 제안된 공공데이터맵을 활용하면 다른 부서에서 관리되는 데이터라 하더라도 자신의 부서에서 생성 및 관리하는 데이터와 관련성이 높은 데이터를 쉽게 탐색할 수 있다. 나아가, 자신이 보유한 데이터와 관련성 높은 데이터를 시각화하여 보여주므로, 다양한 공공데이터를 기반으로 새로운 데이터 서비스를 설계할 수 있는 기회를 제공할 수 있다.

4.4 검색어별 데이터맵 구축 결과

본 절에서는 사전에 정의된 키워드가 아닌 수요자가 직접 입력한 검색어와 유관한 공공데이터의 공공데이터맵을 구축한 결과를 설명한다. <Figure 8>은 “버스 정류장 데이터베이스”와 관련된 공공데이터를 데이터 간 유사도에 기반하여 그래프로 시각화하여 나타낸 것이다. 가장 관련도가 높은 데이터셋은 노드의 크기가 큰, “버스노선의 정류장별 이용통계”, “버스정류장 위치정보”과 “버스노선별 승하차 인원 정보”이다. 해당 공공데이터를 중심으로 다른 공공데이터가 연결되어 있다. 예를 들어, 왼쪽 중앙에 위치한 “버스정류장 위치정보”를 중심으로 “버스 특정 차량 위치정보”, “교통약자전용 노선버스 위치정보”, “정류장 현황통계”가 있고 “버스노선별 승하차 인원 정보”를 중심으로 “버스몰림정류장”과 “정류장별 총 버스 운행횟수 정보” 등이 위치하고 있다. 이처럼 수요자는 자신의 관심사인 검색어를 기준으로 가장 관련성이 높은 공공데이터가 무엇인지 확인할 수 있고, 해당 공공데이터와 연관성이 높은 공공데이터는 무엇인지 쉽게 탐색할 수 있다.

<Figure 9>는 “자동차 운전 사고”를 검색어로 국가데이터맵(Data.go.kr)과 제안 프레임워크의 데이터맵 시각화 결과를 비교한 그림이다. 동일한 검색어로 시각화했을 때, 국가데이터맵의 결과는 <Figure 9(a)>이고 제안 프레임워크의 결과는 <Figure 9(b)>이다. 두 데이터맵 모두 교통사고 현황, 음주운전 등의 데이터가 포함되어 있다. 하지만 국가데이터맵의 결과에는 운전면허학원 정보, 대리운전 업체 정보 등의 연관성 낮은 데이터가 다수 포함되어 있다. 또한 해당 데이터맵에는 연관 데이터 간의 연관성은 나타내고 있지 않다. 반면, <Figure 8(b)>의 오른쪽 아래에는 교통사고의 시간대별, 월별 등의 통계 데이터가 모여 있고, 오른쪽 위에는 운전 사고의 원인 혹은 운전자와 관련된 데이터가 모여 있다. 이처럼, 제안 프레임워크를 활용하여 서로 관련성이 높은 공공데이터셋을 효과적으로 그래프 기반으로 시각화할 수 있는 것을 확인하였다.

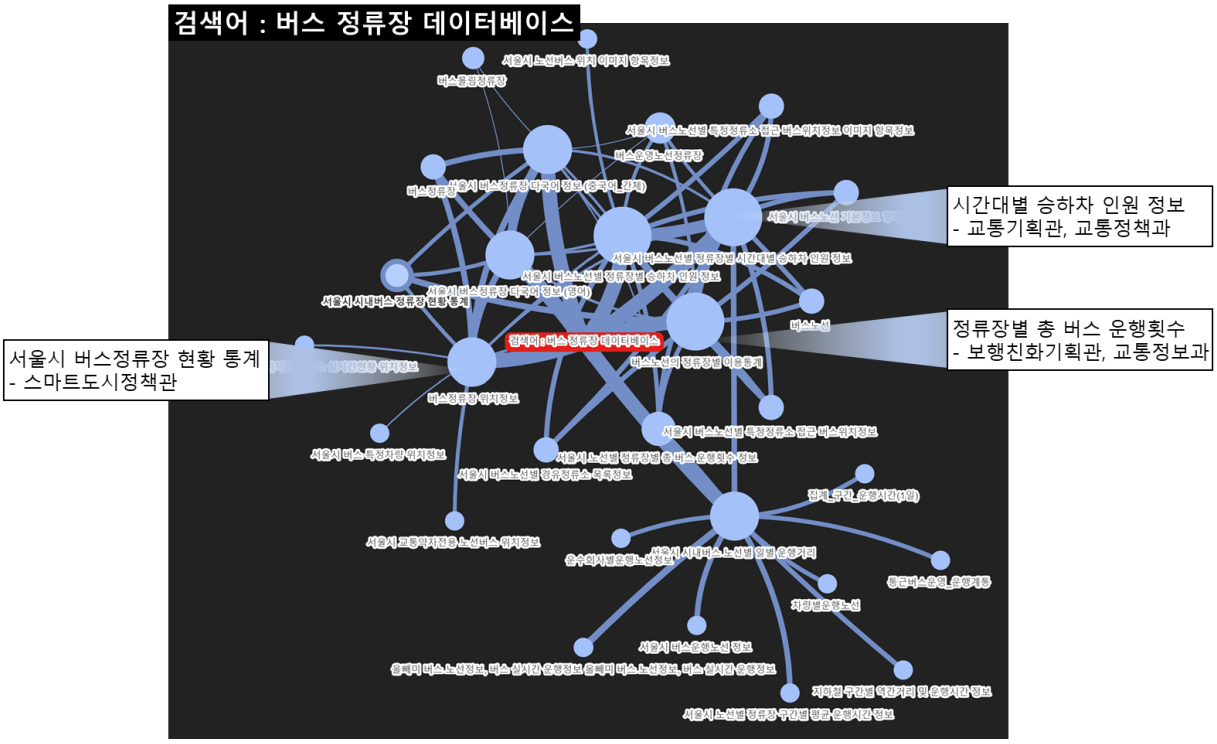
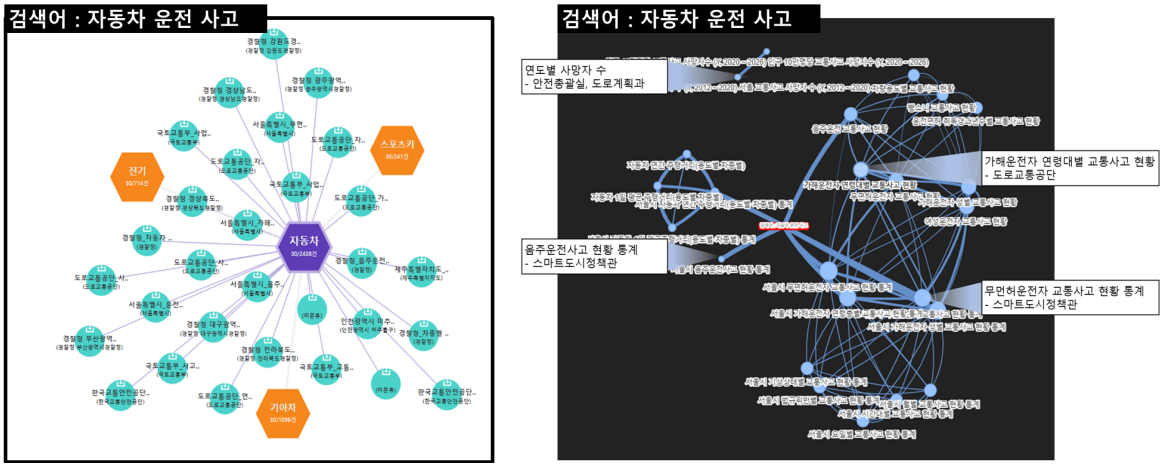


Figure 8. A Result of Datamap for the “Bus Stop Information”



(a) From the existing platform (b) From the proposed framework

Figure 9. A Result of Datamap for the "Car Accident"

5. 결 론

본 연구에서는 공공기관에서 생성 및 관리하는 공공데이터를 수요자가 효과적으로 탐색하기 위해 메타데이터 간 유사도를 활용한 그래프 기반 공공데이터맵 구축 프레임워크를 제안하였다. 제안 방법론은 기존 방법론과 달리 공공데이터셋 간 유사도 산출 시 사람의 개입을 최소화하여 보다 객관적인 결과를 얻을 수 있었다. 또한, 공공데이터셋의 데이터명과 관리 및 생성 부서명 등의 메타데이터의 워드임베딩에 적합한 모델을 선정하고, 공공데이터의 수요를 반영한 키워드 추출 등의 도메인 특화된 전처리 과정을 추가하였다.

본 연구에서 제안한 프레임워크를 통해 서울특별시 총 20,026건의 공공데이터는 키워드별 및 검색어별 공공데이터맵으로 시각화되었다. 본 연구에서 제안한 공공데이터맵 구축 프레임워크는 수요자가 필요로 하는 이종 부서 간 유사한 데이터를 용이하게 탐색할 수 있다는 점에서 기존 공공데이터맵과 차별화된다. 제시된 공공데이터맵을 통해 서울특별시 데이터 수요자들은 다양한 공공데이터를 효율적으로 탐색하여 새로운 데이터 서비스 아이디어를 도출할 수 있을 것으로 기대된다.

기존 방법론 대비하여 공공데이터맵이 개선된 것을 확인하였으나, 구축된 공공데이터맵의 정량적인 평가 방법에 관한 추가적인 연구가 필요하다. 저자들은 본 연구의 결과물에 기반하여 실제 서비스를 기획하여 일반 시민에게 제공할 계획에 있으며, 해당 서비스를 통해 공공데이터맵 만족도 조사 등의 방법을 통해 공공데이터맵을 정량적으로 평가하고 개선하고자 한다. 또한, 데이터명과 관리 및 생성 부서명 뿐만 아니라 데이터셋의 설명(Description)도 공공데이터셋 간의 유사도 산출에 고려된다면 연관 데이터 추천 성능이 개선될 것으로 기대한다. 즉, 향후에 공공데이터셋의 설명을 추가로 확보하고 해당 데이터에 BERT 등의 최신 대형 언어모델을 접목하여 본 제안 방법론을 고도화하는 것이 필요하다.

참고문헌

Aksoy, M., Yanık, S., and Amasyali, M. F. (2023), A comparative analysis of text representation, classification and clustering methods over real project proposals, *International Journal of Intelligent Computing and Cybernetics*, **16**(3), 595-628.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017), Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, **5**, 135-146.

Chawla, S., Aggarwal, P., and Kaur, R. (2022), Comparative analysis of semantic similarity word embedding techniques for paraphrase detection. In Emerging Technologies for Computing, *Communication and Smart Cities: Proceedings of ETCCS 2021*, Singapore: Springer Nature Singapore, 15-29.

d'Sa, A. G., Illina, I., and Fohr, D. (2020), Bert and fasttext embeddings for automatic detection of toxic speech, In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, IEEE, 1-5.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018), Learning word vectors for 157 languages, arXiv preprint arXiv:1802.06893.

Hagberg, A., Swart, P., and S Chult, D. (2008), Exploring network structure, dynamics, and function using NetworkX, *Proceedings of the 7th Python in Science Conference*, 11-15.

Han, E. J., Chae, H., Woo, H., and Sohn, S. Y. (2018), Word2vec Algorithm Applied to Identify Gender-Related Vocabulary Appeared in News Articles, *Journal of the Korean Institute of Industrial Engineers*, **44**(4), 272-282.

Kil, H. (2018), The Study of Korean Stopwords list for Text mining, *URIMALGEUL: The Korean Language and Literature*, **78**, 1-25.

Kim, D. and Joo, W., Kim, E., and Lee, Y. (2014), A Case Study on Classification System Design for Public Sector Information Typology, *Journal of Digital Convergence*, **12**, 51-68.

Kim, D., Kim, H., Song, C., Yang, J., and Kim, H. (2021), Methods for Utilising Local Government's Public Data Released to The Public Data Portal, *Journal of Digital Contents Society*, **22**(3), 445-452.

Kim, E. J., Kim, M., and Kim, H. (2019), Data Standardization for the Enhanced Utilization of Public Government Data, *Knowledge Management Review*, **20**(4), 23-38.

Kim, H. (2019), A Concept and Model of Public Datamap, *TTA Journal*, **182**, 28-33.
Kim, H. L. (2021), A Knowledge Model of Data Map for Semantically Representing National Data, *Journal of Digital Contents Society*, **22**(3), 491-499.
Kwon, S. B. and Yoo, J. E. (2022), Online career counseling text classification using BERT and FastText, *Journal of the Korean Data And Information Science Society*, **33**(6), 991-1006.
Lim, J. and Choi, G. (2017), The Influence of Open Data Policies on Public Innovation, *Journal of the Korean Institute of Industrial Engineers*, **43**(1), 19-29.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... and Stoyanov, V. (2019), Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013), Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
Park, E. L. and Cho, S. (2014), KoNLPy: Korean natural language processing in Python, *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, 133-136.
Sakaji, H., Hayashi, T., Fukami, Y., Shimizu, T., Matsushima, H., and Izumi, K. (2021), Retrieving of Data Similarity using Metadata on a Data Analysis Competition Platform, In *2021 IEEE International Conference on Big Data*, 3480-3485.
Song, C. and Kim, H. (2022), Improvements of public data policy through data portal analysis of local governments, *Journal of Digital Contents Society*, **23**(4), 697-705.

저자소개

최준혁 : 포항공과대학교 산업경영공학과에서 2017년 학사를 취득하고 동 대학원에서 산업경영공학과 박사과정에 재학 중이다. 관심 연구분야는 머신러닝 기반 예측 시스템 개발 및 AI를 위한 데이터 관리다.

권민지 : 권민지 연구원은 서울과학기술대학교 산업정보시스템학과에서 2016년 학사 학위를 취득하고 포항공과대학교 산업경영공학과에서 2018년 석사 학위를 취득하였다. LG GNS, 한국과학기술연구원을 거쳐, 현재 서울기술연구원에서 전임연구원으로 재직 중이다. 관심 연구분야는 데이터 시각화 및 분석, 머신러닝 기반 자연어 처리이다.

김준철 : 서울기술연구원에서 수석연구원으로 인공지능, 빅데이터, 데이터사이언스 전문성을 기반으로, 서울특별시의 복잡한 도시문제에 대해 데이터 기반의 솔루션 제공 및 첨단 과학기술을 서울시정에 접목하여 해결책을 제시하기 위한 실증연구를 담당한다.

정준각 : 포항공과대학교 산업경영공학과에서 2013년 학사, 2019년 석박사 통합 학위를 취득하였다. 일리노이 대학교 어바나-샴페인과 울산과학기술원 산업공학과에서 박사후연구원으로 근무했으며, 현재 한양대학교 산업융합학부에서 조교수로 재직 중이다. 관심 연구분야는 텍스트 마이닝, 품질 데이터 분석, 설명가능 인공지능 응용이다.