http://www.jsebs.org
pISSN 2288-3908
eISSN 2765-3846

The Journal of Society for e-Business Studies
Vol.26, No.4, November 2021, pp.1-13
https://doi.org/10.7838/jsebs.2021.26.4.001

공간 정보를 가지는 데이터셋의 준자동 융합 기법

Semi-automatic Data Fusion Method for Spatial Datasets

윤종찬(Jong-chan Yoon)*, 김한준(Han-joon Kim)**

초 록

빅데이터 관련 기술이 발달함에 따라 이전에는 처리할 수 없었던 방대한 규모의 데이터를 처리할 수 있게 되었다. 이에 따라 데이터 선정 및 융합 자동화 프로세스 구축은 빅데이터 기반 서비스 구현에 있어 선택이 아닌 필수인 시대가 되었다. 본 논문은 공간 정보를 담고 있는 데이터셋을 융합하여 유의미한 새로운 정보를 생성하기 위한 준자동화 기법을 제안한다. 우선 Node2Vec 모델을 활용하여 주어진 데이터셋의 키워드를 이용해 데이터셋의 임베딩 벡터를 생성한다. 생성된 각 임베딩 벡터를 이용해 코사인 유사도를 계산하여 데이터셋 간의 시멘틱 유사도를 구한다. 이후 사람이 개입하여 그 시멘틱 유사도가 상대적으로 높은 데이터셋 쌍 중에서 공간 정보를 가진 데이터셋을 선별하고, 데이터셋 쌍을 융합하여 시각화한다. 이러한 일련의 준자동 융합 프로세스를 통해 단일 데이터셋으로부터는 얻을 수 없는 유의미한 융합 정보를 생성할 수 있음을 보인다.

ABSTRACT

With the development of big data-related technologies, it has become possible to process vast amounts of data that could not be processed before. Accordingly, the establishment of an automated data selection and fusion process for the realization of big data-based services has become a necessity, not an option. In this paper, we propose an automation technique to create meaningful new information by fusing datasets containing spatial information. Firstly, the given datasets are embedded by using the Node2Vec model and the keywords of each dataset. Then, the semantic similarities among all of datasets are obtained by calculating the cosine similarity for the embedding vector of each pair of datasets. In addition, a person intervenes to select some candidate datasets with one or more spatial identifiers from among dataset pairs with a relatively higher similarity, and fuses the dataset pairs to visualize them. Through such semi-automatic data fusion processes, we show that significant fused information that cannot be obtained with a single dataset can be generated.

키워드 : 공간 데이터, 데이터 융합, 임베딩, 시멘틱 유사도, 빅데이터, 데이터셋
Spatial Data, Data Fusion, Embedding, Semantic Similarity, Big Data, Dataset

---

본 논문은 과학기술정보통신부 및 정보통신기술진흥센터의 대학 ICT 연구센터지원사업의 연구결과로 수행되었으며(IITP-2021-2018-0-01417), 또한 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-00121, 데이터 품질 평가기반 데이터 고도화 및 데이터셋 보정 기술 개발)을 받아 수행된 연구임.

* First Author, MS., Department of Electrical and Computer Engineering, University of Seoul (pletory94@gmail.com)

** Corresponding Author, Professor, Department of Electrical and Computer Engineering, University of Seoul (khj@uos.ac.kr)

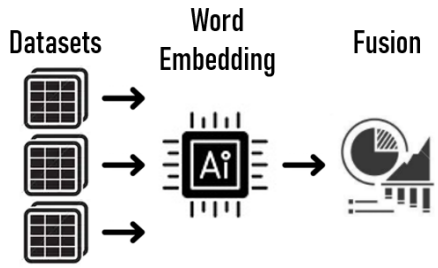
Received: 2021-07-27, Review completed: 2021-08-26, Accepted: 2021-09-17

1. 서 론

4차 산업혁명 시대가 도래함에 따라 인류가 가진 데이터의 양은 기하급수적으로 증가하고 있다. 데이터가 부족했던 과거에는 데이터를 수집하는 능력이 중요했지만, 데이터의 양이 매우 방대해진 지금은 데이터 수집 능력보다는 데이터를 선별, 가공, 융합하는 능력이 훨씬 중요해졌다. 특히 다수개의 데이터셋에 대한 융합 과정을 통해 하나의 단일 데이터셋으로부터는 추출할 수 없었던 유의미한 정보를 추출하는 것은 필수적인 단계로 포함되고 있다.

데이터 융합은 2개 이상의 데이터셋을 가지고 유의미한 정보를 생성할 수 있는 새로운 융합된 데이터셋을 생성하는 방법이다[1, 6, 8, 12]. 그런데 데이터 융합을 위해서, 방대한 규모의 데이터셋 중에서 의미있는 융합 후보 대상이 되는 데이터셋을 찾아내는 작업은 매우 시간이 많이 소요되는 작업이다. 만약 융합 후보 데이터를 찾는 과정이 자동적으로 수행된다면 해당 작업 시간을 줄이는데 크게 기여할 것이다. 게다가, 융합 데이터 선정 과정에서 사람의 경험과 직관에 의존하는 것은 많은 시행착오를 유발할 가능성이 높다. 데이터셋 간의 의미적 연관성을 정량적으로 산출하여 특정 기준에 부합하는 결과가 사용자에게 제시된다면 보다 객관적이고 정확한 융합 결과를 생성할 수 있을 것이다.

<Figure 1>에서 보는 바와 같이, 본 논문은 사전에 학습된 기존 워드임베딩 모델을 활용하여 융합 데이터셋을 임베딩 벡터(embedding vector)로 표현하고, 데이터셋 간의 시멘틱 유사도를 기반으로 데이터 융합을 수행하는 프로세스를 제안한다. 본 연구는 융합 대상을 공간 정보를 가지는 데이터셋으로 한정한다. 융합 대상이 되는 공간 데이터셋 간의 시멘틱 유사도가 높고, 공간 정보를 가지는 데이터셋의 표현 단위(granularity)를 통일시켜 각 컬럼에 대한 융합 조건을 설정하여 조인 연산(join operation)을 수행함으로써 새로운 융합 데이터셋이 만들어진다[14]. 본 논문은 이러한 일련의 과정을 포함하는 준자동 데이터 융합 프로세스를 제안하며, 이를 통해 보다 신속하게 의미적 연관도가 높은 데이터셋에 대한 융합을 수행하여, 이를 가지고 유의미한 분석 서비스를 제공할 수 있음을 보일 것이다.



〈Figure 1〉 Data Fusion Process

2. 이론적 배경 및 관련 연구

2.1 워드임베딩(Word Embedding)

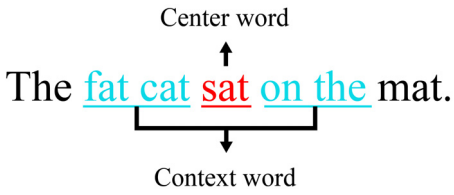
데이터셋 간의 의미적 연관성을 정량적으로 산출하기 위해서는, 데이터셋 자체가 내포하는 의미를 수치로 표현할 수 있는 체계가 필요하다. 이를 위해 본 논문은 워드임베딩(word embedding) 모델을 활용한다. 워드 임베딩은 단어가 가지고 있는 의미를 저차원의 벡터로 표현하는 기법이다[11]. 워드 임베딩의 결과로 생성된 임베딩 벡터는 단어의 존재 여부에 따라 단

순히 0과 1로 표현하는 기존의 원-핫 벡터(one-hot vector)와는 달리, 단어가 내포하는 의미를 정교하게 표현할 수 있을 뿐만 아니라, 벡터의 차원 수를 사용자 임의대로 적게 설정하여 공간 효율성을 높일 수 있다.

워드임베딩 모델은 문장 내 단어간 인접 정보를 인공신경망에서 학습하여 의미적으로 가까운 단어의 입베딩 벡터들이 임베딩 공간 내 인접한 영역에 위치하게 된다. 초기에 이를 위한 실제 학습 알고리즘으로서 NNLM(Neural Network Language Model), RNNLM(Recurrent Neural Network Language Model) 등이 제안되었는데 학습 시간이 오래 걸리는 단점이 있었다. 2013년 Google에서 발표한 연구로서 학습 시간을 대폭 단축시킨 것이 Word2Vec 모델[13]이다. 그리고 단어의 위치와 문장 내 구조적 역할 정보를 그래프(graph)에 표현하여 이를 학습한 모델이 Node2Vec[7] 모델이다. 본 연구는 데이터셋에 부여된 키워드 임베딩을 위해 Word2Vec과 Node2Vec을 연계한 이중 임베딩 구조[3, 4]를 활용하였다(<Figure 3> 참조).

2.1.1 Word2Vec

Word2Vec 모델은 텍스트 데이터를 임베딩하기 위한 모델로서, <Figure 2>와 같이 중심 단어와 주변 단어를 나눠서 문장 내 의미를 학습하는 모델이다. 예를 들면, 그림 2에서 중심 단어가 'sat'인 경우 'sat' 좌우에 인접한 단어가 주변 단어가 된다. Word2Vec 모델의 학습 방식은 CBOW(Continuous Bag of Words)와 Skip-Gram으로 구분되는데, CBOW는 주변 단어로 중심 단어를 예측하는 방식이고, Skip-Gram은 중심 단어로 주변 단어를 예측하는 방식이다.

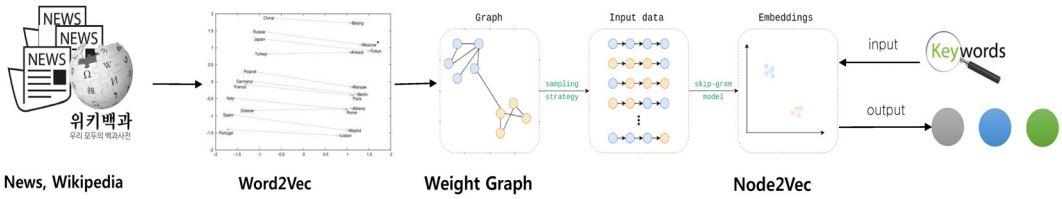


<Figure 2> Word2Vec

2.1.2 Node2Vec

Node2Vec 모델은 단어 간의 관계, 문장 내 위치와 구조적 역할을 학습하기 위해 단어들 간의 인접 정보가 그래프로 표현되고 인접 노드 간의 관계를 학습하여 임베딩 벡터를 만들어낸다.

Node2Vec 모델은 그래프의 인접 노드와의 관계를 파악하기 위해 BFS(Breath-First Search) 또는 DFS(Depth-First Search) 그래프 탐색 알고리즘을 사용한다. BFS는 같은 레벨에 존재하는 노드를 우선적으로 탐색하고, DFS는 다음 분기로 넘어가기 전 해당 분기를 완벽하게 탐색



〈Figure 3〉 Double Embedding Architecture for Keyword Embedding

하고 넘어가는 방식이다. DFS 알고리즘을 중심으로 학습하면 네트워크 동질성(network homogeneity)을 위주로 학습하고, BFS 알고리즘을 중심으로 학습하면 그래프의 구조적 동치(structural equivalence)를 위주로 학습한다. 다시 말해서, DFS 알고리즘은 가까이 있는 노드들부터 학습하므로 인접 노드들 간에 높은 유사성을 부여한다. 비교하여 BFS 알고리즘은 같은 레벨에 존재하는 노드를 중심으로 탐색하므로 문장 내 비슷한 위치에 있고, 유사한 구조적 역할을 수행하는 단어 간에 높은 유사성을 부여한다.

2.2 공간 정보 식별자의 표현

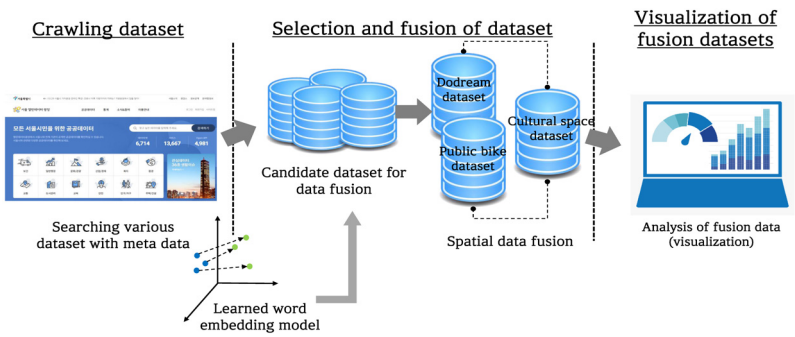
공간 정보 내 위치를 식별하는 방법은 주소, 좌표계, PNU 코드등이 있다. 국내의 경우, 주소는 크게 지번 주소와 도로명 주소로 나뉜다[2]. 지번 주소는 구획 내 소유주를 기준으로 번호를 나눠 할당하는 방식이다. 도로명 주소는 도로의 너비를 기준으로 “대로”, “로”, “길”로 나누고, 도로의 기점으로부터 20m 간격으로 건물 번호를 부여하는 방식이다. 좌표계는 지구상 임의의 지점을 좌표의 형식으로 표현한다. 좌표값의 결정을 위해서, 우선 지구를 타원 형태에 해당하는 수학적 모델로 표현해야 하고, 중심 좌표를 설정하여 3차원 좌표값을 2차원 지도상으로 투영해서 표현해야 한다. 각 세부 단계마다 다양한 방법론이 존재하기 때문에 좌표계의 종류가 매우 많은데, 다양한 좌표계를 국제적 표준으로 통일하기 위해 제정한 코드가 EPSG 코드이다. 예를 들어, 실생활에서 흔히 사용하는 GPS 서비스는 지구를 WGS84 타원체로 설정하여 EPSG:4326 코드를 사용한다.

3. 공간 정보 융합

본 절은 본 논문이 제안하는 준자동 데이터 융합 프로세스와 그것의 주요 요소를 소개한다.

3.1 개요

본 논문의 목표는 방대한 규모의 데이터셋 중에서 유의미한 정보를 생성할 수 있는 최적의 데이터셋 쌍을 선정하여 융합하는 것이다. 이 목표를 달성하기 위해 본 모델은 우선 <Figure 3>과 같이 사전 연구된 이중 임베딩 기법으로 단어 수준의 임베딩 벡터를 생성한다. 데이터셋과 메타데이터로서의 키워드를 수집



〈Figure 4〉 Data Analysis Process with Data Fusion

하여 각 데이터셋에 대한 임베딩 벡터를 생성한다. 여기서 데이터셋을 표현한 임베딩 벡터는 임베딩 공간에서 유의미한 융합이 가능할 수준의 연관성이 큰 데이터셋이 근접한 영역에 표시된다. 데이터셋 간의 시멘틱 유사도를 산출하기 위해서는 해당 임베딩 벡터간의 코사인(cosine) 유사도 함수를 활용한다.

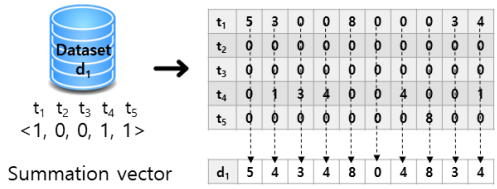
유의미한 정보를 도출할 수 있는 융합 후보 데이터셋들은 상호간 의미적 연관성이 높아야 할 것이다. 이에 따라, 데이터셋 쌍에 대한 코사인 유사도가 임계값(예를 들어, 0.5) 이상이면서, 공간 정보를 가지는 데이터셋 쌍을 탐색한다. 결과적으로 하나의 융합 데이터셋을 생성하기 위해서는 주어진 융합 후보 데이터셋에 내재된 위치 정보 표현 단위를 통일하는 과정이 필요하다. 표현 단위가 통일된 데이터셋 내 공간 영역(spatial zone)의 개념을 차용하여 두 지점 사이의 거리가 특정 조건을 만족하는 레코드가 융합 연계되어 최종 융합 데이터셋을 생성하게 된다. <Figure 4>는 지금까지 소개한 데이터 융합 프로세스를 보여준다.

3.2 임베딩 기반 데이터셋 간 시멘틱 유사도 계산

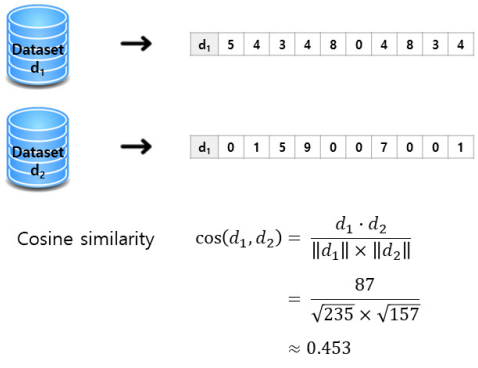
데이터셋이 내포하는 의미를 임베딩 벡터로 표현하기 위해서는 데이터셋에 부여된 유의미한 메타데이터를 충분히 수집하는 과정이 필요하다. 메타데이터는 데이터에 대한 데이터를 일컫는다. 데이터셋에 대한 메타데이터는 데이터셋에 대한 구조 및 전반적인 개요를 포함하는 데이터이다. 예를 들어, 카테고리, 주요 키워드, 데이터의 크기, 저장 방식, 데이터의 공개일자 등이 그것이다. 결국, 주어진 데이터셋의 메타데이터를 활용하여 임베딩 벡터로 표현할 때, 메타데이터의 품질이 데이터셋 임베딩 벡터의 신뢰도를 결정하게 된다.

수집한 데이터셋의 메타데이터로부터 데이터셋의 키워드를 추출하고, 이미 학습된 이중 임베딩 모델로부터 해당 키워드에 상응하는 임베딩 벡터를 가져온다. 그리고 <Figure 5>와 같이 각 키워드의 임베딩 벡터를 차원별 합산을 수행한 합(合)벡터(summation vector)를 생성하여 이를 주어진 데이터셋에 대한 임베딩 벡터로 정의한다.

이어서 융합 대상이 되는 후보 데이터셋들을 가려내기 위하여 데이터셋의 임베딩 벡터 간의 코사인 유사도를 계산한다. 주어진 2개의 데이



〈Figure 5〉 Generation of an Embedding Vector for a Dataset

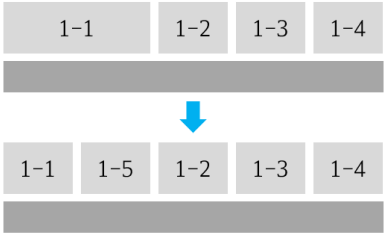


〈Figure 6〉 Example of Computing a Cosine Similarity between Datasets d_1, d_2

터셋의 연관성이 높을수록 해당하는 각 임베딩 벡터간의 각도가 0°에 가까워져 코사인 1에 가까워진다. 코사인 유사도는 차원 수와 상관없이 두 벡터 간 유사도를 신속하게 측정할 수 있기 때문에, 데이터셋 임베딩 벡터를 활용하여 데이터셋들에 대한 클러스터링 연산 작업도 효율적으로 수행할 수 있다[16].

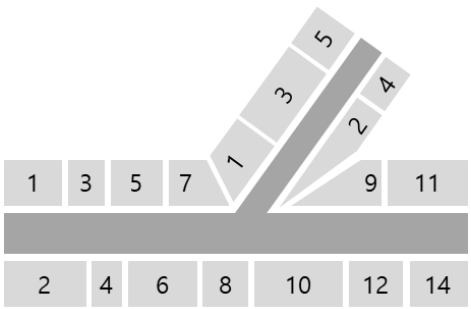
3.3 위치 표현 체계 통일

기본적으로 공간 데이터에 대한 융합을 수행하기 위해서는 특정 위치에 대한 표현 단위를 통일한 후, 두 위치 사이의 거리가 임계 조건 내 들어오는지를 평가하여 해당 레코드를 조인한다. 이러한 융합 과정을 실현하기 위해서는 크게 두 가지 사항이 해결되어야 한다. 첫째, 두 지점을 나타내는 표현 체계가 동일해야 한다는 것이다. 두 지점 사이의 거리를 계산할 때 위치 표현 체계가 다르면 계산 과정이 원활히 이뤄지기 어렵기 때문이다. 둘째, 거리 계산 알고리즘을 적용하기 용이한 표현 체계를 채택해야 한다. 예를 들어, '서울시청'과 '서울역' 사이의 거리를 계산할 때, '서울시청'과 '서울역'이라는 문자열 데이터만으로는 두 지점 사이의 거리를 계산할 수 없다.



〈Figure 7〉 Assignment of Land-Lot Address

제2절에서 서술한 바와 같이, 우리는 공간 정보의 식별을 위해서 크게 주소 체계와 좌표 체계를 사용한다. 국내 주소 체계에서 지번 주소는 땅 구획에 번호를 매겨 주소를 정하는 방식이다. 예를 들어, <Figure 7>과 같이 1번지의 땅을 소유하고 있는 사람이 4명이라고 가정할 때, 각 소유주에게 1-1번지부터 1-4번지까지 할당하는 방식이다. 그런데 만약 1-1번지 땅을 소유하고 있는 사람이 땅을 나눠서 주소를 새로 할당해야 하는 상황이 발생하면 1-1-1, 1-1-2가 아닌 1-1, 1-5로 주소를 할당한다. 즉, <Figure 7>과 같이 1-1번지와 1-5번지는 인접해있지만, 번호는 연속적이지 않다. 따라서 지번 주소는 각 지점 간의 거리를 계산할 수 있는 근거가 되기에는 충분하지 않다[10].



〈Figure 8〉 Assignment of Street Address

비교하여, 도로명 주소는 크게 도로명과 건물번호로 구성된다. 도로는 너비에 따라 "대로", "로", "길"의 세 가지 등급으로 구분된다. 그리고 도로의 기점을 기준으로 건물번호가 부여된다. 건물번호를 부여하는 방식은 <Figure 8>과 같이 도로의 기점을 기준으로 왼쪽 건물은 홀수, 오른쪽 건물은 짝수 번호를 부여하되, 기점으로부터 20m씩 떨어질 때마다 건물번호

가 2씩 증가하도록 부여한다[9]. 하지만 각 건물의 너비가 다를 수 있음을 고려하지 않기 때문에, 도로명과 건물번호만으로 두 지점 사이의 거리를 정확하게 계산할 수 없다.

좌표계를 활용하는 경우, 지구상 두 좌표 간의 거리를 계산하는 알고리즘만 구현되어 있다면, 좌표계가 두 지점 사이의 거리를 계산하기에 가장 용이하다. 지구상 임의의 지점을 좌표체계로 표현하기 위해서는 크게 세 가지 과정이 필요하다. 첫째, 지구는 불완전한 타원의 형태를 가지고 있는데, 이 지구 타원체를 수학적 모델로 표현해야 한다. 둘째, 좌표의 중심을 어디로 설정할 것인지에 대한 합의가 필요하다. 셋째, 3차원 좌표계로 표현된 좌표를 2차원의 지도로 투영하는 투영법에 대한 정의가 필요하다[5].

통일된 위치 정보 체계를 선정하는 데 있어 가장 중요한 요소는 얼마나 두 지점 사이의 거리를 쉽게 계산할 수 있느냐이다. 앞서 서술한 바와 같이, 두 지점 사이의 정확한 거리 계산을 위해서 지번 및 도로명 주소 체계 보다는 좌표계를 사용하는 것이 바람직하다. 우리는 다양한 좌표계 중에서 가장 많은 데이터셋에서 사용하고 있는 좌표계로서 WGS84 타원체를 이용한 경위도 표현 시스템 EPSG : 4326을 사용한다.

3.4 공간 데이터 융합

구체적으로, 본 논문은 구면상 두 좌표의 거리를 계산하기 위해 하버사인 공식(Haversine formula)[15]을 사용하였다. 2차원상의 두 좌표의 거리를 계산할 때 사용하는 일반 공식은 지구의 곡률을 고려하지 않으므로 오차가 생길 수밖에 없는데, 이를 고려한 하버사인 공식은 구면상에서 정확한 거리를 구하는데 적합하다. 데이터셋의 융합시 내부 레코드를 연계할 때, 하버사인 공식을 이용하여 주어진 레코드에 대한 좌표 사이의 거리가 임계값 이하인 경우 이를 융합 연계하는 것이다.

4. 실 험

4.1 실험 데이터

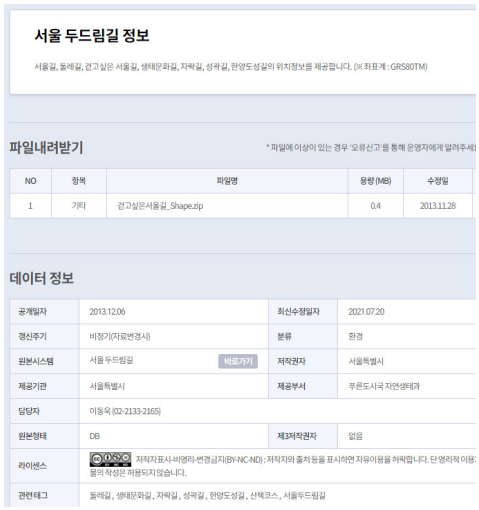
본 연구의 핵심인 데이터셋 임베딩 벡터의 생성을 위해서는 데이터셋의 의미적 연관성을 반영한 임베딩 공간이 필요하다. 이를 위해서 우리는 2018년 1월 1일부터 2020년 1월 1일까지의 뉴스 데이터를 가지고, <Figure 3>의 이중 임베딩 아키텍처에서 사용한 Word2Vec 및 Node2Vec 모델을 학습하였다. 결과적으로 175,000개가량의 단어에 대한 임베딩 벡터를 생성하였다.

<Table 1>은 실제로 융합 연산을 수행할 데이터를 보여주며, 이는 '서울 열린 데이터 광장'(https://data.seoul.go.kr/) 사이트에서 활용도가 높은 32개 분야 내 총 138개 데이터셋을 포함한다. 또한 선정한 데이터셋의 메타데이터로서 관련 키워드, 제공 부서, 분류, 데이터 설명 등 데이터셋의 의미를 유추할 수 있는 다양한 정보를 수집하여(<Figure 9> 참조), 이를 정교한 임베딩 공간을 생성하는데 활용하였다. 메타 정보의 하나인 '데이터 설명' 항목은 대개 문장 형식으로 작성되었기 때문에 별도의 단어 추출 전처리 과정을 수행하였다. 이를 위해 Konlpy 라이브러리의 Twitter 패키지를 사용

하였으며, 주어진 문장을 파싱하여 핵심 단어를 임베딩 학습모델에 포함시켰다.

〈Table 1〉 List of Datasets

No	datasets
1	individual public land price
2	construction reminder
3	market analysis service (estimated floating population)
…	…
15	English public service reservation
16	cultural event public service reservation
17	education public service reservation
18	facility rental public service reservation
19	reservation of public service for sports facilities
20	public bicycle use (by time)
21	public bicycle rental
22	rental of public bicycles for foreigners (by date)
…	…
29	public bicycle usage
30	public parking lot information
31	number of passengers by subway line
…	…
62	street vendor location
63	dodream street location
64	mosquito repellent alert
…	…
73	cultural event information
74	cultural space information
75	price for agricultural and livestock products
…	…
107	fire department status
108	fire department location
109	fire water location
…	…
137	lifelong learning portal education institution
138	lifelong learning portal offline courses



〈Figure 9〉 Example of Metadata Given to a Dataset

준비된 138개 데이터셋 간의 모든 코사인 유사도를 계산하여(<Table 2> 참조), 이 중 유사도가 0.5 이상인 데이터셋 쌍 2,402개 중에서 공간성 컬럼명(예: 주소, 위도, 경도 등)을 가지는 쌍에 한정하여 융합을 수행하였다. 최종적으로 생성된 융합 데이터셋의 목록은 <Table 3>과 같다.

〈Table 2〉 Dataset Pairs with Higher Similarities

dataset pair		cosine similarity
number of passengers on each subway station	subway station information	0.962
number of passengers on each subway station	subway real-time train location	0.959
public transportation route	subway last train timetable	0.951
number of passengers at each bus stop by time	public transportation route	0.949

dataset pair		cosine similarity
…	…	…
fire department location	traffic history	0.656
public bicycle rental	cultural space information	0.656
traffic bearkout road information	subway last train timetable	0.656
…	…	…
dodream street lcoation	cultural space information	0.572
information on using puhblic bi-cycles (by date)	location of fire-fighting objects	0.572
…	…	…
population living in seoul (long-term for-eigners)	kids cafe in-formation	0.500
education public service reser-vation	ultra-fine dust issued by year	0.500

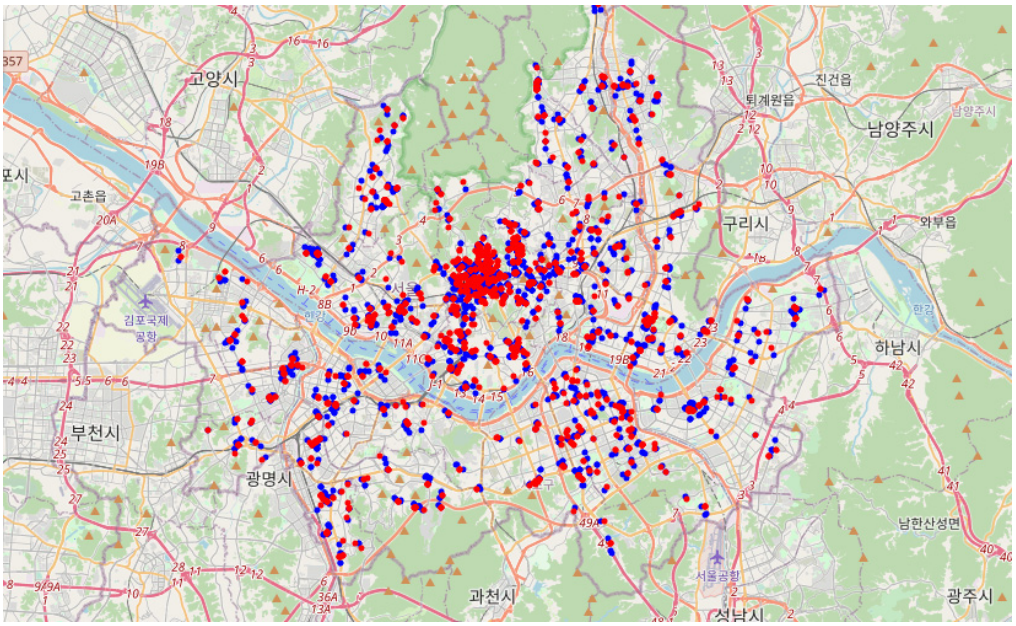
〈Table 3〉 List of Data Fusion

ID	Dataset
16	cultural event public service reservation
21	public bicycle rental
30	public parking lot information
63	dodream street location
74	cultural space information
108	fire department location

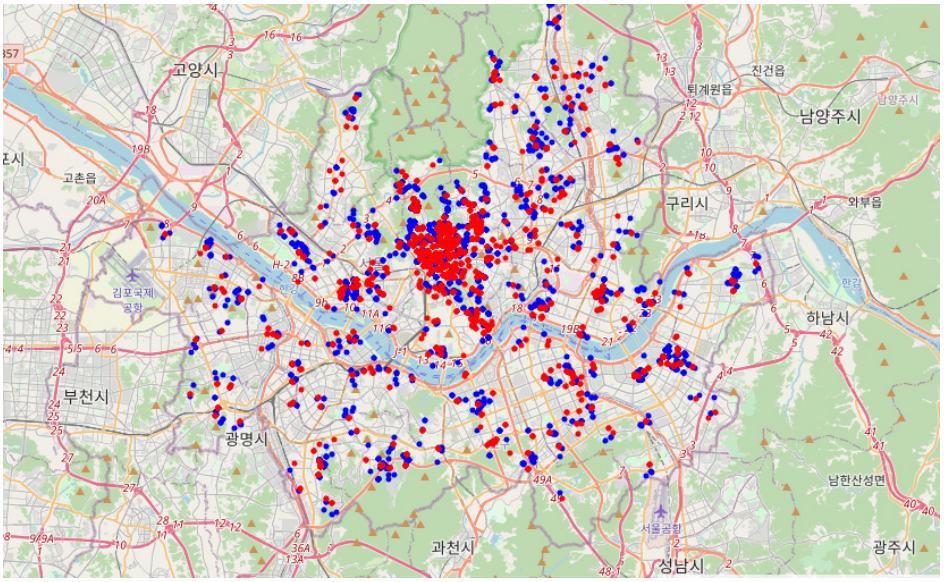
4.2 실험 결과

자동 융합 수행 결과 중에서 유의미한 정보를 생성하고, 사회적 현상에 대한 통찰에 도움을 줄 수 있는 유스케이스(usecase) 몇 가지를 소개한다.

<Figure 10>은 '서울시 문화 공간 정보' 데이터셋과 '서울시 공공자전거(따릉이) 대여소 위



〈Figure 10〉 Visualization of the Fusion Data with Cultural Place and Public Bike Station Location Datasets



〈Figure 11〉 Visualization of the Fusion Data with Cultural Place Information and Dodream Street Datasets

치 정보' 데이터셋 쌍을 융합한 결과를 시각화한 것이다. 여기서 빨간 점은 문화 공간 위치를 표시한 것이고, 파란 점은 공공자전거 대여소 위치를 나타낸 것이다. 최근 지속 가능한 도시 조성을 위해 서울시는 '자전거 타기 좋은 도시' 만들기 사업을 진행하고 있는데, 여전히 서울시의 몇몇 문화 공간의 자전거에 대한 접근성은 좋지 못하다는 사실을 알 수 있다.

<Figure 11>은 '서울시 두드림길 정보' 데이터셋과 '서울시 문화 공간 정보' 데이터셋 쌍을 융합한 결과를 보여준다. 도심 속의 문화 공간은 도심과는 또 다른 공간으로 사람들에게 문화 여행을 떠나는 기분을 만끽할 수 있게 하는 공간이다. 그 여행을 문화 공간에서 끝내는 것이 아닌 주변의 두드림길과 같은 생태적, 역사적으로 의미 있는 공간과 어우러질 수 있도록 도시 설계에 대한 아이디어를 제공할 수 있다. 이러한 융합 데이터셋은 단일 데이터셋만으로는 알 수 없었던 사회적 현안에 대한 통찰을 제공할 수 있다. 더 나아가 그 통찰을 통해 사회적 문제를 해결할 수 있는 단초를 제공할 수 있다.

5. 결 론

본 논문은 방대한 규모의 데이터셋 간 유사도를 계산하여 융합 가능성이 높은 데이터셋의 쌍을 선정하고, 그 중에서 공간 정보를 가지는 데이터셋에 대한 융합 연산을 포함한 융합 프로세스의 준자동화를 시도하였다. 사용자는 제안 기법을 활용함으로써 방대한 규모의 데이터셋으로부터 융합 후보가 되는 데이터셋의 범위를 대폭 축소할 수 있다. 또한, 공간 데이터의 특수성을 감안하여 데이터셋에 주어진 위치 표현 체계를 하나로 통일하고, 공간 조인 연산을 통하여 융합 작업을 수행하였다.

4차 산업혁명과 인공지능 산업이 블루오션으로 각광받고 있는 지금, '데이터 융합'의 중요성은 나날이 커지고 있다. 이는 한 분야에 국한된 데이터셋보다는 서로 다른 분야의 데이터셋에 대한 융복합을 통하여 얻어낸 분석 결과가 더 큰 가치를 발휘하기 때문이다. 그래서 본 연구는 데이터셋의 가치를 평가하는 척도로서 '데이터 융합 가능성'을 포함시켜야 함을 시사한다.

향후 연구로서, 우리는 공간 데이터의 융합에 초점을 맞춘 본 연구를 일반화하여 시계열 데이터의 융합 자동화 기법을 개발할 예정이다. 최근 IoT 및 클라우드 기술의 발전과 더불어 시계열 데이터의 유형이 다양해지고, 그 규모도 매우 커지고 있다. 이에 시공간 데이터셋에 대한 준자동화 융합 모델은 보다 활용 범위가 클 것으로 기대한다.

References

[1] Bleiholder, Jens, and Felix, N., "Data fusion," ACM computing surveys (CSUR), Vol. 41, No. 1, pp. 1-41, 2009.

[2] Chang, T. W., "A Study on Integration and Application Plans of Address and Location Information," The Journal of Society for e-Business Studies, Vol. 15, No. 2, pp. 93-105, 2010.

[3] Cho, S. R. and Kim, H. J., "A Preliminary Study on Improving Korean Text Embedding Model," Proceedings of KICS Winter Conference, 2020.

[4] Cho, S. R. and Kim, H. J., "Topic Re-modeling System using Node2Vec," Proceedings of Fall Conference of 2020 Korea Associations of Information Systems, 2020.

[5] Choi, Y. S., Park, H. G., and Kim, G. S., "Establishment of th Plane Coordinate System for Framework Data(UTM-K) in Korea," Korean Journal of Geomatics, Vol. 22, No. 4, 2004.

[6] Gao, J., Li, P., Chen, Z., and Zhang, J., "A Survey on Deep Learning for Multimodal Data Fusion," Neural Computation, Vol. 32, No. 5, pp. 829-864, 2020.

[7] Grover, A. and Leskovec, "Node2Vec: Scalable feature learning for networks," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[8] Khan, S., Nazir, S., García-Magariño, I., and Hussain, A., "Deep learning-based urban big data fusion in smart cities: Towards traffic monitoring and flow-preserving fusion," Computers & Electrical Engineering, Vol. 89, 106906, 2021.

[9] Korea Ministry of the Interior and Safety, Road Name Address System, http://www.juso.go.kr/.

[10] Lee, S. H., Yang, C. M., and Baek, S. C., "Improvement on Location Based Parcel Numbering System," Journal of Cadastre & Land Informatix, Vol. 42, No. 1, pp. 148-149, 2012.

[11] Li, Y. and Yang, T., "Word embedding for understanding natural language: A survey," Guide to big data applications,

pp. 83-104, Springer, 2018.

[12] Liu, J., Li, T., Xie, P., Du, S., Teng, F., and Yang, X., “Urban big data fusion based on deep learning: An overview,” Information Fusion, Vol. 53, pp. 123-133, 2020.

[13] Ma, L. and Zhang, Y., “Using Word2Vec to process big text data,” Proceedings of IEEE International Conference on Big Data, pp. 2895-2897, 2015.

[14] Wiemann, S., and Lars, B., “Spatial data fusion in spatial data infrastructures using linked data,” International Journal of Geographical Information Science, Vol. 30, No. 4, pp. 613-636, 2016.

[15] Winarno, E., Hadikurniawati, W., and Rosso, R. N., “Location based Service for Presence System using Haversine Method,” Proceedings of 2017 International Conference on Innovative and Creative Information Technology (ICITech), pp. 1-4, 2017.

[16] Xia, P., Zhang, L., and Li, F., “Learning Similarity with Cosine Similarity Ensemble,” Information Sciences, Vol. 307, pp. 39-52, 2015.

저 자 소 개

윤종찬 (E-mail: pletory94@gmail.com)
2020년 서울시립대학교 전자전기컴퓨터공학부 (공학사)
2020년～현재 서울시립대학교 전자전기컴퓨터공학과 석사과정
관심분야 딥러닝, 빅데이터, 데이터융합

김한준 (E-mail: khj@uos.ac.kr)
1994년 서울대학교 계산통계학과 (이학사)
1996년 서울대학교 전산과학과 (이학석사)
2002년 서울대학교 컴퓨터공학부 (공학박사)
2002년～현재 서울시립대학교 전자전기컴퓨터공학부 교수
관심분야 데이터마이닝, 머신러닝, 빅데이터, 데이터베이스, 지능형 정보검색