컬럼 임베딩과 개체명 인식을 통합한 조인 가능한 데이터셋 검색

최윤석, 김한준*
서울시립대학교 전자전기컴퓨터공학과

choiys8819@naver.com, khj@uos.ac.kr

Joinable Dataset Discovery with Integrating Column Embedding and Named Entity Recognition

Yoonseok Choi, Han-joon Kim*
Department of Electrical and Computer Engineering, University of Seoul

요 약

우리는 둘 이상의 관계형 테이블을 하나의 테이블로 조인(join)하여 새로운 유의미한 정보를 생성할 수 있다. 하지만 매우 많은 데이터셋에서 조인 가능한 테이블을 찾는 것은 노동 집약적이고 시간 소모가 큰 작업이다. 본 논문은 컬럼명 임베딩과 개체명 인식 기법을 통합하여 자동으로 조인 가능한 테이블을 찾아내는 새로운 기법인 CNE-join 을 제안한다. 우리는 Kaggle 에서 수집한 50 개의 테이블 데이터셋을 사용한 실험을 통해 제안 기법의 우수성을 보인다.

Ⅰ. 서 론

둘 이상의 관계형 테이블을 특정 컬럼을 기준으로 조인(join) 연산을 수행하여, 조인하기 이전에는 알 수 없었던 유의미한 정보를 얻는 것이 가능하다. Table1 은 그 예시를 보여준다. 좌측 테이블은 '대학명'과 '위치' 컬럼, 우측 테이블은 '음식점'과 '위치' 컬럼으로 구성되어 있다. 우리는 이 2 개의 테이블을 '위치' 컬럼을 기준으로 조인하여 각 대학교 주변에 있는 음식점을 알아낼 수 있다. Table1 의 예시에서는 우리가 주어진 2 개의 테이블이 조인 가능한지를 쉽게 판단할 수 있다. 여기서, 테이블의 개수가 수백 또는 수천개 이상으로 많아진다면, 사람이 수작업으로 조인 가능한 테이블을 찾아내는 것은 매우 어려운 작업이 될 것이다. 본 논문은 많은 수의 테이블 데이터가 주어질 때, 조인 가능한 테이블을 자동으로 찾아내는 기법을 제안한다.

Table 1. 조인 가능한 관계형 테이블 예시

대학명	위치
서울시립대학교	동대문구
서울대학교	관악구
고려대학교	성북구

음식점	위치
스테이크&치즈	동대문구
비빔밥하우스	동대문구
파스타킹	관악구
피자몰	성북구

둘 이상의 테이블이 조인 가능한지 자동으로 판단하는 기법은 크게 2 가지로 나뉜다. 첫째는 컬럼명이 유사하면 조인 가능하다고 판단하는 기법이고, 둘째는 컬럼에 포함된 값들이 유사하면 조인 가능하다고 판단하는 기법이다. 첫째 기법은 한 컬럼명이 다른 컬럼명과 부분적으로 일치하는지 N-gram 알고리즘을 사용하여 확인하는 기법[1], 컬럼명에 대한 임베딩 벡터를 생성하여 그 벡터간의 코사인 유사도 계산을 통해 두 컬럼명이 의미적으로 유사한지 판단하는 기법[2] 등을 포함한다. 둘째 기법은 세부적으로 특정 컬럼의 값들이 다른 컬럼의 값들과 100% 일치하는지 확인하는 기법[3], 한 컬럼에 포함된 값의 집합과 다른 컬럼에 포함된 값의 집합간의 유사도를 계산하는 기법[4] 등을 포함한다. 본 논문은 조인 가능한 테이블을 탐색하기 위해 컬럼명과 컬럼값의 유사성을 동시에 고려하면서 개체명 인식 기술을 병합한 CNE-join 기법을 제안한다.

Ⅱ. 제안 기법

Ⅱ-1. CNE-join 기법

Figure 1 은 CNE-join 기법의 개략적인 과정을 보여준다. 그림에서 보이는 바와 같이, CNE-join 은 다수의 관계형 테이블을 입력 받아 각 테이블에 대해 컬럼명 임베딩(column embedding), 개체명 인식(named entity recognition) 및 포괄성(containment) 계산을 수행하여, 결과적으로 사용자에게 조인 가능한 테이블쌍과 조인 기준이 되는 컬럼쌍을 출력한다.

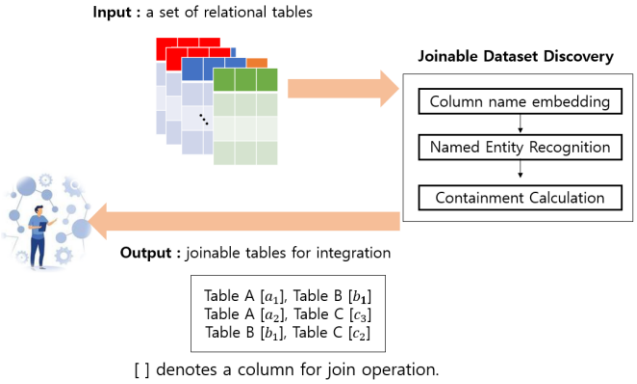

Figure 1. CNE-join 의 개략적 수행과정

Figure 2 는 CNE-join 의 구체적인 수행 과정을 보여준다. CNE-join 은 입력으로 주어진 N 개의 관계형 테이블에 대해, 우선 서로 다른 테이블에 포함된 컬럼 2 개를 묶어 컬럼쌍으로 만드는 작업을 진행한다. 예를 들어, Figure 2 에서 3 개의 테이블이 주어질 때, 각 테이블이 3 개의 컬럼을 가지고 있으므로, 총 27 개의 컬럼쌍이 생성된다.

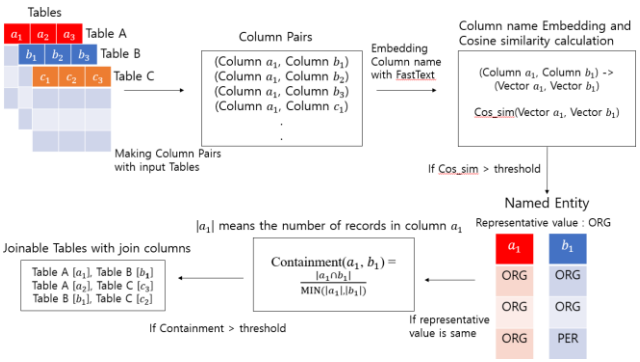

Figure 2. CNE-join 의 구체적 내부 수행과정

CNE-join 은 만들어진 모든 컬럼쌍에 대해 컬럼명 임베딩을 진행하여, 하나의 컬럼쌍마다 하나의 벡터쌍을 생성한다. 이어서 그 벡터쌍에 대해 코사인(cosine) 유사도 계산을 진행한다. 벡터쌍의 코사인 유사도 값이 임계값 이상이면, 그 벡터쌍에 해당하는 컬럼쌍에 대해 개체명 인식을 진행한다. 개체명 인식 단계에서는 컬럼쌍의 각 컬럼에 포함된 값들에 대해 개체명 인식기를 사용해 고유명사 수준의 개체명을 추출한다. 여기서 개체명은 기관(ORG), 인물(PER), 지리(GEO)등을 포함한 8 개의 카테고리로 분류된다. 그리고 컬럼값으로 가장 많이 출현한 개체명 카테고리는 그 컬럼의 대표 카테고리가 된다. 예를 들어, Figure 2 에서 a_1컬럼의 3 개 컬럼값에 대한 개체명 카테고리는 모두 ORG 이므로 a_1 컬럼의 대표 개체명 카테고리는 ORG 이다. 유사하게 b_1컬럼에 대한 대표 개체명 카테고리는 ORG 가 된다. 그 다음 단계는 컬럼쌍에서 두 컬럼의 대표 개체명 카테고리의 일치 여부를 확인하고, 해당 컬럼쌍에 대해 포괄성 계산을 진행한다. 이어서 모든 컬럼쌍들 중에서 포괄성 수치가 임계값 이상인 컬럼쌍들을 가려낸다. 포괄성(containment) 계산 수식은 (1)과 같다.

$$\text{Containment}(a_1,\ b_1) = \frac{|a_1 \cap b_1|}{\text{MIN}\ (|a_1|,\ |b_1|)} \qquad (1)$$

여기서, $|a_1|$은 컬럼 a_1의 레코드 개수를 의미한다.

Ⅱ-2. CNE-join 에 활용된 기법

컬럼명 임베딩 단계에서 우리는 사전학습된 FastText[5] 모델을 사용한다. FastText 는 특정 단어를 입력 받으면 그 단어에 해당하는 300 차원의 벡터를 출력하는 모델이다. CNE-join 은 FastText 에 컬럼명을 입력하여 얻어낸 벡터를 코사인 유사도 계산에 사용한다. 개체명 인식 단계에서 사용된 개체명 인식기는 BERT[6]를 Kaggle 의 NER 데이터셋[7]으로 학습하여 구축되었다. 개체명 인식기는 문장 수준의 데이터를 입력 받아야 하므로, 단어 수준의 컬럼값에 대해 '[컬럼명] + [is] + [컬럼에 포함된 값]' 형식의 문장을 인위적으로 구성하여 개체명 인식기에 입력한다.

Ⅲ. 실험

제안기법의 우수성을 보이기 위해, 우리는 Kaggle 에서 50 여개의 관계형 테이블 데이터셋을 수집하였다. CNE-join 기법과 상대적으로 비교되는 최근 기법은 Equi-join[3], Jaccard-join[4], CE-join 이다. Table2 는 CNE-join 과 비교 기법의 차이점을 보여준다. Equi-join 은 한 컬럼에 포함된 값이 다른 컬럼의 포함된 값과 완전히 겹치면, 두 컬럼을 기준으로 테이블이 조인 가능하다고 판단하는 기법이다. Jaccard-join 은 두 컬럼의 Jaccard 유사도 값이 임계값 이상이면, 두 컬럼을 기준으로 테이블이 조인 가능하다고 판단하는 기법이다. CE-join 은 제안기법인 CNE-join 에서 컬럼명 임베딩 단계만을 사용한 기법이다. 성능 평가 지표는 [4]의 논문을 참고하여 Precision(수식 2)과 Recall(수식 3)의 조화평균인 F1-score 를 사용하였다.

$$\text{Precision} = \frac{\text{Number of retrieved joinable column pairs}}{\text{Number of retrieved column pairs}} \qquad (2)$$

$$\text{Recall} = \frac{\text{Number of retrieved joinable column pairs}}{\text{Number of joinable column pairs in the retrieved pool}} \qquad (3)$$

여기서 retrieved column pairs 는 각 기법이 조인 가능하다고 판단한 컬럼쌍, 이 중 실제로 조인 가능한 컬럼쌍이 retrieved joinable column pairs 이다. 데이터셋의 모든 테이블을 보고 조인가능한지를 라벨링하는 것은 노동집약적이고 시간 소모가 큰 작업이기에, 우리는 CNE-join 과 비교기법의 retrieved joinable column pairs 에 대해 합집합 연산을 수행하여 joinable column pairs in the retrieved pool 을 생성하였다. Figure 3 은 CNE-join 과 비교기법의 F1-score 값을 비교한 그래프이다. CNE-join 의 파라미터인 컬럼명 임베딩 임계값과 포괄성의 임계값을 조율하며 성능을 측정하고, 비교기법의 성능과 비교를 진행하였다. 우리는 컬럼명 임베딩의 임계값을 낮추면 Recall 값이 오를 것이라 예상하였고 포괄성의 임계값을 높이면 Precision 값이 오를 것이라 예상하여, 낮은 컬럼명 임베딩 임계값과 높은 포괄성 임계값으로 실험을 진행하였다. 컬럼명 임베딩 임계값이 0.3 이고 포괄성 임계값이 0.6 일 때 가장 좋은 성능을 보였다. Figure 3 은 CNE-join 이 비교기법보다 성능이 매우 우수함을 보여준다.

Table 2. CNE-join 과 비교 기법의 차이점

기법	컬럼명 비교	컬럼값 비교	임베딩	개체명 인식
CNE-join	O	O	O	O
Equi-join [3]	X	O	X	X
Jaccard-join [4]	X	O	X	X
CE-join	O	X	O	X

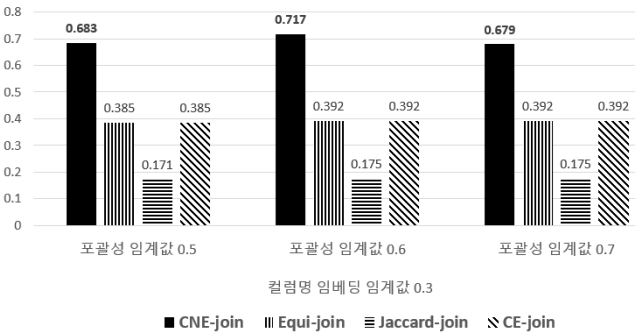

Figure 3. CNE-join 과 비교기법의 F1 score 비교

Ⅳ. 결론

본 논문은 입력으로 N 개의 테이블이 주어지면 그 중에서 조인 가능한 테이블을 자동 탐색하는 기법인 CNE-join 을 소개한다. 이는 N 개의 관계형 테이블에 대해 컬럼명 임베딩, 개체명 인식, 포괄성 계산을 진행하여 조인 가능한 테이블을 찾아낸다. 우리는 Kaggle 에서 수집한 관계형 테이블 데이터셋을 가지고 조인 가능한 테이블을 탐색하는 실험을 수행하였으며, 그 결과, CNE-join 의 성능이 기존 최신 기법보다 우수함을 확인하였다. 우리는 향후 테이블 데이터셋의 개수를 대폭 확장하여, 수평융합 조인 연산과 수직융합 유니온 연산을 모두 고려한 자동 융합 기법을 개발할 예정이다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(교육부)의 재원으로 한국연구재단의 지원(No. NRF-2022R1A2C1011937, 정형 테이블 데이터셋에 대한 딥러닝 기반 데이터 융합 기술 개발)과 과학기술정보통신부 및 정보통신기술진흥센터의 대학 ICT 연구센터지원사업(IITP-2023-2018-0-01417)의 연구결과로 수행된 연구임.

참 고 문 헌

[1] A. Bogatu, A. A. A. Fernandes, N. W. Paton, and N. Konstantinou, “Dataset Discovery in Data Lakes,” in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 2020, pp. 709–720.

[2] J. Pilaluisa, D. Tomás, B. Navarro-Colorado, and J.-N. Mazon, “Contextual word embeddings for tabular data search and integration,” *Neural Computing and Applications*, vol. 35, no. 13, pp. 9319–9333, 2023.

[3] E. Zhu, D. Deng, F. Nargesian, and R. J. Miller, “JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes,” in *2019 International Conference on Management of Data*, 2019, pp. 847–864.

[4] Y. Dong, K. Takeoka, C. Xiao, and M. Oyamada, “Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach,” in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 456–467.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Trans. Assoc. Comput. Linguist*, vol. 5, pp. 135–146, 2017.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *ArXiv Prepr. ArXiv181004805*, 2018.

[7] “NER Data.” https://www.kaggle.com/datasets/rajnathpatel/ner-data.