Journal of KIIT. Vol. 20, No. 1, pp. 31-39, Jan. 31, 2022. pISSN 1598-8619, eISSN 2093-7571 31
http://dx.doi.org/10.14801/jkiit.2022.20.1.31

지식 그래프 엔티티 매칭을 위한 확장된 GCN 모델 설계 및 구현

김재웅*, 이용주**

Design and Implementation of Extended GCN Model for Knowledge Graph Entity Matching

Jaewoong Kim*, Yongju Lee**

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2016R1D1A1B02008553).

요 약

컴퓨터가 정보와 정보 사이의 관계를 이해하고 추론할 수 있는 시맨틱 웹은 최근 대량의 링크드 데이터에서 온톨로지를 활용하여 지식 통합을 효율적으로 수행할 수 있다. 하지만 이러한 온톨로지는 아직까지 이질성 문제 때문에 실제로 사용하기에는 정확성이 낮으며 간단한 엔티티 매칭도 실패하는 경우가 많이 발생한다. 이러한 이질성 문제를 해결하기 위해 본 논문에서는 보다 정교한 매칭을 위한 이중 그래프 컨볼루션 네트워크를 활용함과 동시에 누락되는 엔티티 매칭까지 할 수 있는 유사도 기법을 함께 결합한 확장된 GCN(Graph Convolutional Network) 모델을 제안한다. 실험 결과 본 논문에서 제안한 기법은 기존의 모델에 비해 약 10%의 성능 향상을 보인다.

Abstract

The Semantic Web, which can be understood by computers and inferred between information and information by utilizing a large amount of linked data, can efficiently perform knowledge integration by utilizing ontology recently. However, this ontology is still inaccurate for practical use due to the heterogeneity problem, and even simple entity matching often fails. To solve this heterogeneity problem, in this paper, we propose an extended Graph Convolutional Network (GCN) model that utilizes a double graph convolutional network for more sophisticated matching and combines similarity techniques that can match missing entities at the same time. As a result of the experiment, the method proposed in this paper shows a performance improvement of about 10% compared to the existing model.

Keywords
graph convolutional network, semantic similarity, ontology heterogeneity, knowledge graph

* 경북대학교 IT대학 컴퓨터학부
- ORCID: https://orcid.org/0000-0002-9589-2667
** 경북대학교 IT대학 컴퓨터학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1705-4967

· Received: Oct. 12, 2021, Revised: Nov. 11, 2021, Accepted: Nov. 14, 2021
· Corresponding Author: Yongju Lee
School of Computer Science and Engineering, Kyungpook National University, 80, Daehak-ro, Buk-gu, Daegu 41566, Korea,
Tel.: +82-53-950-7285, Email: yongju@knu.ac.kr

Ⅰ. 서 론

최근 “한국판 뉴딜 정책”의 대표 과제인 “데이터 댐” 사업을 통해 공공 데이터 개방 및 AI 학습용 데이터 구축을 촉진하고 있으며, 해외의 경우에도 Open API 프로젝트 등을 통해 엄청난 규모의 초대용량 빅데이터(Big data) 클라우드를 구축하고 있다. 그러나 많은 예산으로 구축된 오픈 데이터(Open data)들에 대한 효율적인 검색 방법이나 활용 방안을 위한 깊이 있는 연구는 상대적으로 매우 부족한 상황이다. 최근 신경망 분야의 비약적인 발전으로 빅데이터에 대한 기계 학습이나 데이터 분석 등으로 스스로 학습하고 추론할 수 있는 방식으로 진화되고 있지만, 오픈 데이터 구축은 연결된 서비스가 아니라 고립된 개별적 서비스였고 데이터 자체가 연결되어 있는 통합된 모습을 가질 수가 없었다. 이에 구글, 아마존, LOD(Linked Open Data)[1] 등을 중심으로 지식 그래프(Knowledge graph)와 같은 검색 고도화 연구가 활발히 수행되고 있다[2].

그러나 현재까지 시맨틱(Semantic) 시스템에 지식 그래프 기술을 적용한 예는 거의 찾아보기 힘들다. 대규모 RDF(Resource Description Framework) 데이터 셋에서 시맨틱이 어떻게 임베딩(Embedding)되고, 딥러닝(Deep learning)되며, 어떻게 온톨로지 매칭(Ontology matching)이 적용되는지는 거의 연구가 되지 않고 있다. 이에 대한 가장 근본적인 문제는 서로 다른 온톨로지 어휘를 사용하는 애플리케이션들의 상호 작용이다. 현재까지 모든 데이터 셋에 적용 가능한 표준 어휘 사전이 없기 때문에 많은 종류의 어휘 이질성(Heterogeneity) 문제에 당면하고 있다. 예를 들면, ‘음료수’와 ‘과일 주스’는 같은 의미이지만 동의어 집합에 속하지 않기 때문에 매칭되지 않는다.

단어 임베딩은 유사한 단어일수록 가까운 거리에 위치하도록 각 단어에 해당하는 벡터값을 찾는 것으로, 대표적인 연구로는 NNLM[3], Word2Vec[4], Glove[5] 등이 있다. 하지만 단어 임베딩의 입력 자료는 사전에 전처리(Preprocessing)된 텍스트 기반 말뭉치(Corpus)를 사용하는데 반해, 지식 그래프 자료는 RDF 형태의 트리플(Triple) 구조로 저장되므로 단어 임베딩보다는 그래프 임베딩 기법이 사용되어야만 한다. RDF2Vec은 그래프를 임베딩하는 방법이지만 RDF2Vec으로부터 임베딩된 벡터는 특정 릴레이션과 관련된 엔티티 매칭(Entity matching)에는 최적화되어 있지 않다.

그동안 시맨틱 웹 분야에서는 LOD 프로젝트를 통해 여러 가지 괄목할 만한 성과를 거두었으나 연구 기간에 비해 현실 세계에 그다지 큰 영향을 미치지 못했다는 비판이 있었다. 이러한 측면에서 구글이 제시한 지식 그래프는 시맨틱 웹의 개념과 실용성을 결합한 사례로 높이 평가되고 있다. 구글이 제안한 TransE[6] 모델은 head와 relation 벡터의 합은 tail과 같아지도록 학습한다. 그러나 TransE는 1:1 관계일 때는 우수한 성능을 나타내지만 1:N, N:1, N:N과 같은 경우에는 문제가 발생한다. TransH[7] 모델은 관계를 hyperplane에 투영(Projection)시킴으로써 TransE 문제를 해소하였다. 그러나 엔티티(Head & Tail)와 관계(Relation)는 명확히 다른 객체이기 때문에, TransR[8]은 엔티티와 관계를 구분된 공간에 임베딩하는 아이디어를 제안했다. 최근에 제안된 ConvE[9]는 임베딩 벡터에 대하여 2D convolution을 수행하고 벡터의 표현력을 높이기 위하여 비선형성을 적용하였다.

최근까지 링크 자동 완성 분야 등에 지식 임베딩 기법들을 적용한 연구는 상당한 진전이 이루어졌지만, 이러한 기법들을 활용하여 엔티티 매칭을 수행하기에는 아직까지 정확도가 부족하다. 그 이유는 임베딩으로부터 만들어지는 그래프 구조만을 고려하여 엔티티 매칭이 이루어지기 때문에 단말 엔티티에 대한 세밀한 매칭이 어렵다. 본 논문에서는 이러한 엔티티 매칭의 문제점을 해결하기 위해 그래프 기반의 GCN(Graph Convolutional Network)[10]과 유사도 기법을 활용한 확장된 GCN 모델을 제안한다. 기존의 신경망(Convolution network)을 사용하여 복잡한 그래프 구조를 학습하여 찾아내는 엔티티 매칭을 유지하고, 매칭 가중치를 기반으로 하여 유사도를 활용해 기존의 신경망 모델이 찾지 못하는 세밀한 엔티티들을 매칭시켜 정확성을 높인다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구를 통해 온톨로지와 지식 그래프, 그리고 최신 온톨로지 매칭 연구들을 소개한다. 3장은 본 논문에서 제안하는 확장된 GCN 모델의 전체 시스템을 자세

히 설명한다. 4장에서는 데이터 셋을 활용해 제안한 기법의 성능을 평가하고 분석한다. 마지막으로 5장에서는 결론을 내리고 향후 연구 내용을 기술한다.

II. 관련 연구

2.1 온톨로지 이질성 문제

온톨로지는 지식(정보)의 개념, 지식과 지식의 관계, 속성 등을 컴퓨터가 이해하고 처리할 수 있도록 표현하는 모델을 말한다[11]. 온톨로지는 의료, 기계, 금융 등 다양한 분야에서 응용이 가능하도록 정보들을 개념과 개념의 관계를 정의하고 구조화시킴으로써 컴퓨터와 사람이 지식의 공유가 가능하도록 도움을 주는 역할을 한다. 그러나 다른 사람이나 조직 또는 다른 데이터 셋이나 언어에서 일반적으로 다른 개념화를 요구하기 때문에 단일 온톨로지를 사용하는 것이 항상 가능한 것이 아니다. 따라서 다른 온톨로지를 사용할 경우, 서로 다른 온톨로지 간 엔티티가 중복되거나 엔티티 불일치가 포함될 수 있다.

그림 1은 간단한 자동차 관련 온톨로지 예를 통해 온톨로지의 이질성 문제를 보여준다. 여기서 온톨로지 매칭 분석 결과 다음과 같이 두 쌍의 매칭된 결과를 추출할 수 있다.

O1.Vehicle ↔ O2.Means of Transportation
O1.Car ↔ O2.Automobile

기존의 신경망을 활용한 온톨로지 매칭 기법을 이용하여 매칭을 수행한 결과 온톨로지 O1의 Vehicle과 O2의 Means of Transportation 개념이 매칭되었고, O1의 Car와 O2의 Automobile 개념이 매칭되었다. 하지만 그림에서 O1의 'Benz E 250'과 O2의 'Mercedes Benz E class 250'은 매칭이 되지 않았다. 이는 기존의 온톨로지 매칭 기법들이 동의어 집합을 사용하여 매칭을 수행하는데, 'Benz E 250'과 'Mercedes Benz E class 250'은 같은 동의어 집합에 속하게 하지 않았기 때문에 매칭이 되지 않은 것이다.

2.2 구조적 기반 온톨로지 매칭 기법

온톨로지 이질성 문제를 해결하기 위해서 온톨로지 매칭(정렬)을 해야 한다. 온톨로지 매칭 분야는 지난 수년간 광범위하게 연구가 되어 왔으며 온톨로지의 스키마를 매칭시키는 온톨로지 매칭 시스템과 인공지능을 활용하여 지식 그래프의 엔티티를 매칭시키는 두 방향으로 크게 나뉘었다. 온톨로지 스키마를 매칭시키는 대표적인 모델들로는 LogMap [12], AML[13], ALIN[14] 등이 있다.

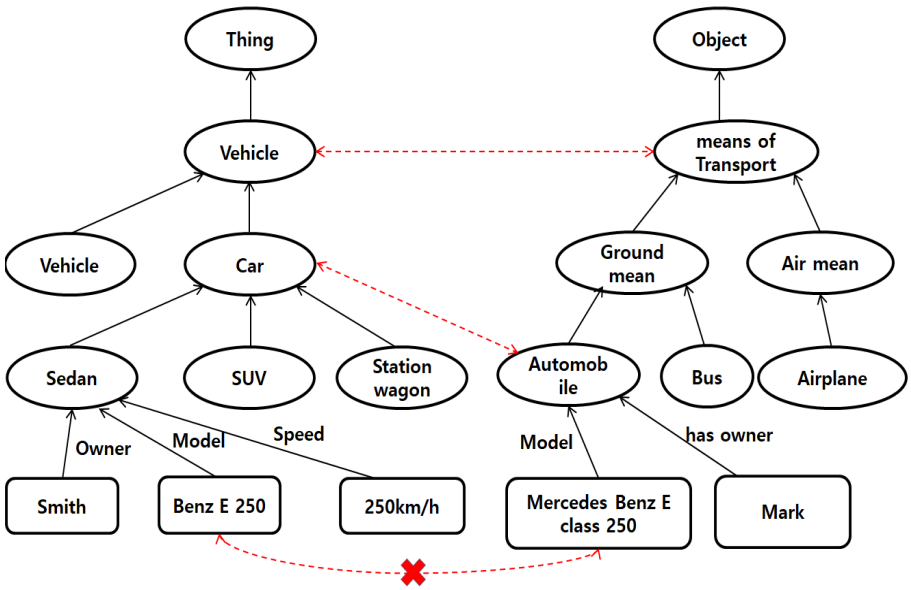

그림 1. 온톨로지 이질성 문제의 예
Fig. 1. Example of ontology heterogeneity problem

이 기법들은 데이터의 유형과 값, 그리고 관련 엔티티와 같은 속성 및 정의를 비교하는데, 여기에 관계 구조를 활용하여 매칭을 한다. 대부분의 방법들은 데이터 셋의 그래프 구조를 통해 유사한 하위 그래프 또는 상위 그래프를 찾는다. 하지만 구조 기반의 매칭 방법은 단독으로는 적용이 쉽지 않고 정확도가 낮기 때문에 대부분 문자열 유사성 측정 방법들과 함께 사용한다.

2.3 신경망을 활용한 온톨로지 매칭 기법

머신러닝 분야의 발전에 따라 신경망을 활용하여 그래프 구조를 학습시켜 두 지식 그래프의 대응 관계를 발견하는 것과 같은 작업이 성공적으로 이루어졌다. 또한 기술의 발전으로 인해 방대한 양의 데이터를 학습시키는 것이 가능해져 큰 데이터 셋에서 더 복잡한 모델을 훈련시키는 것이 가능해졌다. MtransE[15] 모델과 IPTransE[16] 모델은 지식 그래프의 엔티티 매칭을 위한 좋은 방법임을 보였다. 그 중에서도 GCN을 활용한 지식 그래프 엔티티 매칭이 최근 활발히 연구되고 있다.

GCN은 그래프의 노드와 노드 사이의 상관관계를 모델링하는 GNN(Graph Neural Networks)[17]의 convolutional 연산을 사용하는 기법으로 새로운 방향을 제시하였다. 하지만 이 접근 방법은 관계 정보를 적절히 모델링 하는데 한계점이 존재하는데, 일반적인 GCN은 방향성이 존재하지 않으며, 레이블이 지정되지 않은 그래프에서 작동되기 때문에 지식 그래프의 유용한 관계 정보가 적절하게 활용하지 못하고 누락되는 경우가 존재한다. R-GCN (Relational GCN)[18]을 사용하여 다중 관계 그래프를 모델링할 수 있지만, 이는 단순히 각 관계에 대해 각각의 가중치 행렬을 사용하여 너무 과도한 매개변수 세트가 필요하다. DPGCNN(Dual-Primal Graph CNN)[19]은 정점(Vertex) 중심의 원본 그래프와 엣지(Edge)에 해당하는 이중 그래프에서 컨볼루션 연산을 교대로 수행하여 정점과 엣지 특징(Feature)을 모두 학습하고, 그래프 주의(Attention) 메커니즘을 반복적으로 적용하여 엣지 표현을 향상시켰다.

한편, DPGCNN의 영감을 받아 관계를 더 잘 표현하고 다른 지식 그래프 간의 관계를 특징화할 수 있는 RDGCN(Relation-aware Dual-Graph Convolutional Network)[20]이 제안되었다. 하지만 DPGCNN과 RDGCN은 좋은 출발점 역할은 할 수 있지만 딥러닝 기능만 활용할 경우 매칭 정확도는 아직 떨어지기 때문에 단어 및 의미론적 유사성을 함께 고려하여 정확도를 보다 더 높일 필요가 있다.

III. 엔티티 매칭을 위한 확장된 GCN 모델

3.1 이중 그래프 컨볼루션 네트워크

본 논문에서는 기존의 딥러닝 방법을 통한 지식 그래프 엔티티 매칭의 단점을 보완하여 의미론적 유사성을 함께 고려하는 확장된 GCN 모델 기법을 제안한다. 이는 방향성을 가지고 있지 않는 일반적인 GCN 모델보다 복잡한 엣지 구조와 관계를 더욱 더 잘 표현하게 하는 이중 그래프 컨볼루션 네트워크를 사용한다. 여기서 이중 그래프란 이중 관계 그래프를 의미하는데 대부분의 지식 그래프의 임베딩은 기본 구조로 head + relation = tail의 Trans 시리즈를 사용하게 된다. 하지만 이 구조는 복잡한 다중 관계형 그래프에서는 비효율적이다.

그림 2는 삼각형 데이터 구조의 한 예를 보여준다. V는 Vertex로 정점을 의미하고 R은 Relation으로 정점과 정점 사이의 관계를 의미한다. 이와 같은 데이터는 $v_1+r_a=v_2$, $v_2+r_a=v_3$,$v_1+r_a=v_3$를 만족해야 한다. 하지만 $v_1+2r_a=v_3$, $v_1+r_a=v_3$가 모순이 되기 때문에 예시와 같은 복잡한 구조가 많으면 많을수록 딥러닝 모델의 정확성은 떨어지게 된다.

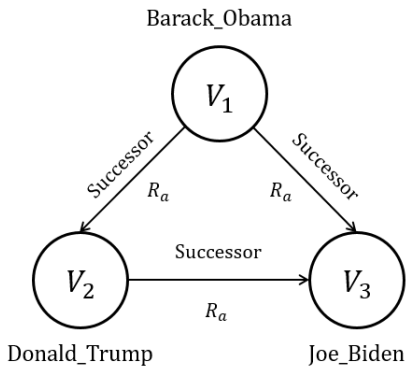

그림 2. 삼각형 데이터 구조의 예
Fig. 2. Example of triangular data structure

이러한 문제를 해결하기 위해 본래의 그래프로부터 이중 관계에 있는 데이터의 그래프를 구성한 다음, 이들 그래프를 함께 처리하는 이중 그래프 처리를 활용하여 기존의 본래 그래프와 함께 상호 작용을 통한 메커니즘으로 학습을 한다.

이중 관계 그래프의 역할은 기존 그래프의 표현에 더 잘 통합이 되도록 도움을 준다. 그림 3은 이중 관계 그래프의 학습 모델의 개요를 보여준다. Dual Awareness Layer는 Primal Layer에서 그래프의 표현에 더욱 정확한 관계 통합을 위한 상호 작용을 통해 영향을 준다.

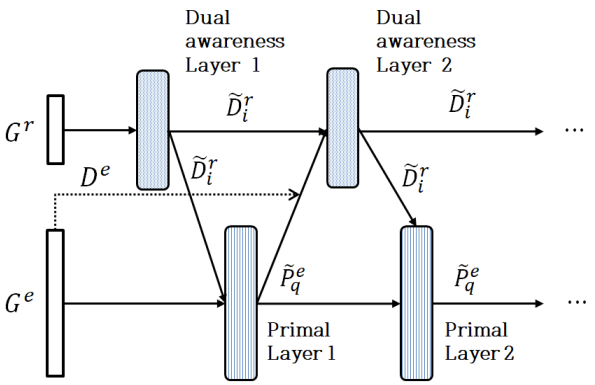

그림 3. 이중 관계 그래프의 학습 모델
Fig. 3. Learning model of relation-aware dual graph

식 (1)의 $\widetilde{D_i^r}$ 는 이중 정점 v_i^r 에서 d 차원의 출력 표현을 나타내고 N_i^r 은 이웃하는 인덱스의 집합을 나타낸다. α_{ij}^r는 이중 집중 점수이며, α_r 는 $2d'$ 차원의 입력을 스칼라로 매핑하는 완전 연결 계층이다. λ^r는 활성화 함수 ReLU를 나타내며, c_i는 이전의 Primal Layer로부터 생성된 r_i의 관계 표현을 나타낸다.

$$\widetilde{D_i^r} = \lambda^r\left(\sum_{j \in N_i^r} \alpha_{ij}^r D_j^r\right) \tag{1}$$

$$\alpha_{ij}^r = \frac{\exp\left(\lambda\left(\omega_{ij}^r M_{ij}^r\right)\right)}{\sum_{k \in N_i^r} \exp\left(\lambda\left(\omega_{ik}^r M_{ik}^r\right)\right)} \tag{2}$$

$$M_{ij}^r = a^r[c_i][c_k] \tag{3}$$

Primal Layer에서는 Dual Layer에서 생성되는 관계 표현을 사용하여 정점 임베딩에 영향을 준다. 식 (4)와 식 (5)는 Primal Layer에서의 Primal 주의 점수를 나타낸다. 여기에서 $\widetilde{x_{qt}^r}$ 는 관계 그래프로부터 얻은 이중 표현을 나타내며, σ^e 는 Primal Layer의 활성화 함수이다.

$$\widetilde{D_q^e} = \sigma^e\left(\sum_{t \in N_q^e} \alpha_{qt}^e D_t^e\right) \tag{4}$$

$$\alpha_{qt}^e = \frac{L_{qt}^e}{\sum_{k \in N_q^e} L_{qk}^e} \tag{5}$$

$$L_{qt}^e = \text{Exp_func}\left(\lambda\left(a^e\left(\widetilde{x_{qt}^r}\right)\right)\right) \tag{6}$$

이중 관계 그래프와 Primal 그래프 사이의 여러 라운의 상호 작용 후 관계 인식 엔티티 표현을 얻을 수 있다. 마지막으로 인접 구조 정보 통합을 위해 highway gate가 존재하는 GCN을 적용한다. 최종적으로 GCN 계층 l에서 생성하는 출력 표현 $R^{(l)}$은 다음과 같다.

$$R^{(l+1)} = \xi(G^{(l)} R^{(l)} \widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}}) \tag{7}$$

여기서 $\widetilde{A} = A + I$는 본래의 그래프 G의 인접 행렬이며 I는 단위행렬이고 ξ 는 활성화 함수 ReLU이다.

$$\widetilde{D}_{ij} = \sum^{k} \widetilde{A}_{jk}, \quad G^{(l)} \in R^{d^{(l)} \times d^{(l+1)}} \tag{8}$$

식 (8)은 계층별 학습이 가능한 가중치의 행렬이다. 그리고 마지막으로 GCN 계층의 출력에서 생성되는 최종 엔티티 표현을 두 개체 간의 거리를 통해 정렬을 진행한다.

3.2 엔티티 의미 유사도 계산

딥러닝의 방법으로 지식 그래프 엔티티 매칭을 할 경우, 그래프 구조를 기반으로 매칭을 하기 때문에 세밀하고 정확한 끝 노드에 대한 매칭에서 정확성이 떨어진다. 매칭의 정확도를 높이기 위해 단어

의 유사도 점수를 함께 측정하여 정확도를 높인다. 지식 그래프 엔티티를 구성하는 URI(Uniform Resource Identification)는 어휘적인 정보를 담고 있는 경우가 대다수이기 때문에 유용하게 사용할 수 있다. 또한, 단순히 레이블만 비교하는 것은 예외와 오류가 발생할 수 있기 때문에 본 논문에서는 단어 유사도 점수와 WordNet 유사도 점수를 함께 고려하여 비교한다.

우선, 단어 유사도 점수는 가장 보편적으로 사용되면서 높은 정확도를 보이는 코사인 거리(Cosine similarity)(즉, 식 (9))를 사용한다. 코사인 유사도는 기존의 유클리드 거리 계산 방식의 단점을 보완한다. 코사인 유사도는 벡터의 크기가 아닌 벡터의 방향에 중점을 두기 때문에 정확한 매칭이 가능하다.

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A^{2_i}} \sqrt{\sum_{i=1}^{n} B^{2_i}}} \tag{9}$$

이에 더해 단어의 유사성과 함께 의미의 유사성으로 더욱 정확한 매칭을 위해 WordNet 유사도 기법을 활용한다. WordNet은 영어의 대규모 어휘 데이터베이스로서 명사, 동사, 형용사 및 부사의 각각 별개의 개념을 표현하는 동의어(Synsets)들의 집합으로 이루어져 있다. 동의어는 개념의 의미와 어휘들의 관계를 통해 서로 연결이 되어 있으며 의미 있는 관련된 단어와 개념들을 네트워크를 통해 탐색할 수 있다. 또한 단어와 단어들의 관계를 상/하 위계 구조로 구성한다. 이 구조를 활용하여 단어와 단어 간의 의미 유사도를 측정할 수 있다. 측정 방법에는 여러 가지가 존재하지만 여러 실험 끝에 가장 성능이 좋다고 판단되는 Wu-Palmer 유사도(즉, 식 (10))를 사용한다.

$$wup(a,b) = \frac{2 * depth(LCS(a,b))}{depth(a) + depth(b)} \tag{10}$$

여기서 depth는 단어의 위계 구조에서 단어의 깊이를 나타낸다. LCS는 Least Common Subsumer로서 단어 간의 최소 공통 포함 수를 의미한다. Wu-Palmer 유사도를 활용하여 두 엔티티 사이의 의미적 유사도를 최종적으로 나타낼 수 있도록 프로그래밍한다. WordNet 의미 유사도를 구하는 과정은 다음과 같다. 먼저 토큰화를 통해 문장을 단어들로 나눈 뒤 품사에 따라 분류를 한다. 그리고 동의어 집합 중 가장 대표되는 단어를 반환하고 Wu-Palmer 유사도를 계산해서 최종값을 생성한다. 그림 4는 위에서 언급한 딥러닝 기법과 유사도 기법을 결합한 확장된 GCN 모델의 전체 개요를 보여준다.

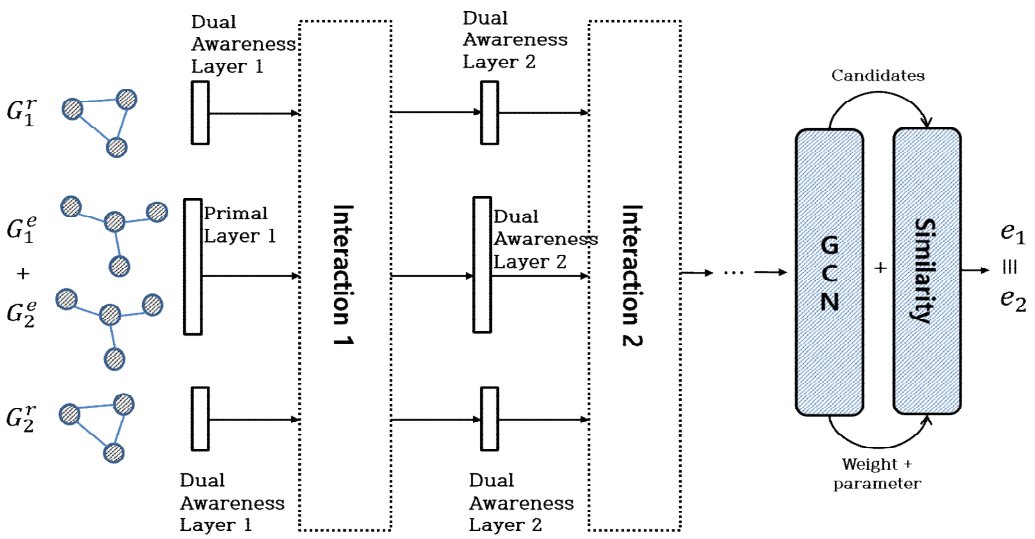

그림 4. 확장된 GCN 모델의 개요
Fig. 4. Overview of extended GCN model

이 모델에서 이중 그래프 컨볼루션 네트워크의 딥러닝 기법으로부터 만들어지는 두 지식 그래프 엔티티의 거리를 기준으로 정렬을 한다. 이때 거리가 가장 가까운 엔티티가 바로 딥러닝이 찾은 매칭 값이 된다. 우리는 딥러닝이 찾은 이 매칭 값을 기준으로 아직 정밀하게 찾지 못한 나머지의 엔티티들을 유사도 기법과 함께 측정하여 가중치들을 통해 최종적으로 모델이 찾은 엔티티를 매칭시킨다.

IV. 실험 및 분석

4.1 데이터 셋

본 논문에서 제안한 확장된 GCN 모델의 성능을 평가하기 위하여, 본 실험에서는 실제 데이터 셋을 활용하여 확장된 GCN 모델을 통해 실험을 진행하고 성능을 비교, 분석, 평가한다. 실험을 위해서 한 가지 데이터 셋보다는 다양한 유형의 지식 그래프 엔티티 매칭을 위해 표 1과 같이 DBP15K의 데이터 셋을 기반으로 지식 그래프의 엔티티 매칭에 맞게 변형했다. 데이터 셋에는 정치, 사회, 산업, 지명, 문화 등의 다양한 지식 정보들과 관계 정보들이 포함되어 있다. 엔티티들은 URI 형식으로 구성되어 있다.

표 1. 실험 데이터 세트
Table 1. Experimental data sets

Data set	No. of entities	No. of relations	No. of attributes	No. of rel. triples
KG 1	67,511	3,083	8,221	160,421
KG 2	102,298	2,421	7,612	240,864

4.2 성능 평가

실험은 다음과 같이 구성한다. 우선 데이터 셋을 활용하여 확장된 GCN 모델을 통해 엔티티 매칭을 하여 2개의 지식 그래프의 엔티티들이 정확히 일치하는지 평가한다. 예를 들어, KG 1의 'Benz E 250'과 KG 2의 'Mercedes Benz E class 250'이 매칭되는지 확인한다. 실험의 결과 값들은 엔티티 매칭 연구에서 보편적으로 사용되는 Hit@K rate를 사용하였다. Hit@K는 매칭 후보군들 중 K번째 만에 정답을 찾았는지를 나타낸다.

표 2는 현재까지 가장 성능이 좋은 것으로 알려져 있는 RDGCN 모델과 우리의 확장된 GCN 모델에 대한 실험 결과를 표로 나타내고 있다. 엔티티 매칭을 한 번에 찾은 Hit@1에서는 기존의 69.99%에서 79.04%로 약 10%의 성능 향상을 보여주었다. Hit@10에서는 기존의 84.5% 정확도에서 87.77%의 소폭 향상이 나타났다. 또한 Hit@50과 Hit@100에서는 큰 향상 폭은 나타나지 않았지만 약간의 향상을 보여주었다. 그림 5는 표 2의 성능 분석 결과를 그래프로 표현한 그림이다.

표 2. 실험 결과표
Table 2. Experimental result table

Model	Hit@1	Hit@10	Hit@50	Hit@100
RDGCN	69.99	84.5	90.33	92.52
Our approach	79.04	87.77	91.27	92.54

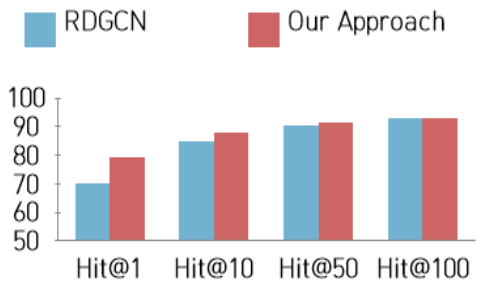

그림 5. 확장된 GCN의 모델의 실험 결과 그래프
Fig. 5. Experimental result graph of extended GCN model

V. 결론 및 향후 과제

대규모 시맨틱 빅데이터에서 온톨로지를 활용하기 위한 지식 통합(Knowledge integration)은 엔티티 이질성 매칭 문제에 당면하게 된다. 이에 단어 임베딩과 그래프 임베딩을 활용한 연구들이 많이 선행되었지만 아직까지 실생활에 적용될 만큼 정확도가 그리 높지 않으며, 그래프 구조를 기반으로 학습되기 때문에 엔티티들의 미세한 매칭은 아직까지 완벽하게 이루어지지 않는다는 문제점이 존재한다.

본 논문에서는 정확한 엔티티 매칭을 위해 그래프 구조를 학습하여 엔티티를 찾아내는 GCN 모델을 기반으로 유사도 기법들을 결합한 확장된 GCN 모델을 제안하였다. RDF의 그래프 데이터 특성상 발생하는 삼각형 데이터 구조의 문제점을 고려하여

이중 그래프를 활용한 GCN 모델을 통해 엔티티들을 찾는다. 또한 유사도 기법의 엔티티 매칭은 1:1로 확인하는 것은 큰 시간을 소요하기 때문에 전체 엔티티를 확인하는 대신 매칭 가중치를 통하여 GCN 모델의 후보군들을 활용하여 소요 시간을 줄였다.

본 논문에서 제안한 기법은 기존의 모델에 비해 정확도를 약 10% 향상시키는 효과를 보였다. 본 논문에서 제안하는 연구를 통한 기여는 기존의 GCN 모델보다 복잡한 엣지 구조와 관계를 더 잘 표현하는 이중 그래프 컨볼루션 네트워크를 사용하고, 단어 및 WordNet 유사도 기법을 함께 적용함으로써 엔티티 매칭 정확도를 향상시킬 수 있다는 가능성을 보인 것이다. 하지만 본 논문에서 제안한 모델은 유사도 기법의 확장으로 인해 소요 시간이 증가하였다. 이러한 문제점을 좀 더 보완하기 위해 향후 연구로는 기존의 다른 딥러닝 기반 엔티티 매칭 기법들과의 성능을 비교하고, 인공지능 모델과 유사도 기법의 상호 작용을 고려한 최적의 확장된 GCN 모델로 발전시킬 필요가 있다.

References

[1] O. Kingsley, "Linked Open Data: State-of-the-Art Mechanisms and Conceptual Framework", Linked Open Data: Applications, Trends and Future Developments, IntechOpen, Chapter 3, Oct. 2020. https://doi.org/10.5772/intechopen.94504.

[2] L. Ehrlinger and W. Wob, "Towards a Definition of Knowledge Graphs", Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems, Leipzig, Germany, Vol. 1695, Sep. 2016.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model", Neural Information Processing Systems (NIPS), Vol. 3, pp. 932-938, Jan. 2000.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", ICLR, Arizona, USA, Sep. 2013.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation", Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532-1543, Oct. 2014. http://dx.doi.org/10.3115/v1/D14-1162.

[6] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data", Neural Information Processing Systems (NIPS), NY, USA, Vol. 2, pp. 2787-2795, Dec. 2013.

[7] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge Graph Embedding by Translating on Hyperplanes", AAAI Conference on Artificial Intelligence, Québec City Québec Canada, pp. 1112-1119, Jul. 2014.

[8] T. Trouillon, C. R. Dance, E. Gaussier, J. Welbl, S. Riedel, and G. Bouchard, "Knowledge Graph Completion via Complex Tensor Factorization", Journal of Machine Learning Research, Vol. 18, No. 1, pp. 4735-4772, Jan. 2017.

[9] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D Knowledge Graph Embeddings", ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, NY, USA, pp. 113-123, Sep. 2019.

[10] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks", ICLR, Toulon, France, Feb. 2017.

[11] B. Smith, "Ontology", The Blackwell Guide to the Philosophy of Computing and Information, Wiley Online Library, pp. 155-166, Jan. 2004.

[12] E. Jimenez-Ruiz and B. Grau, "LogMap: Logic-Based and Scalable Ontology Matching", International Semantic Web Conference (ISWC), Bonn, Germany, pp. 273-288, Oct. 2011.

[13] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. Cruz, and F. Couto, "The AgreementMakerLight Ontology Matching System", OTM 2013 Conferences, Graz, Austria, pp. 527-541, Oct. 2013. http://dx.doi.org/10.1007/978-3-642-41030-7_38.

[14] J. Silva, K. Revoredo, F. Baiao, and J. Euzenat,

"ALIN: Improving interactive ontology matching by interactively revising mapping suggestions", The Knowledge Engineering Review, Vol. 35, pp. e1, Nov. 2020. https://doi.org/10.1017/S0269888919000249.

[15] M. Chen, Y. Tian, M. Yang, and C. Zaniolo, "Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment", International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne Australia, pp. 1511-1517, Aug. 2017. https://doi.org/10.24963/ijcai.2017/209.

[16] H. Zhu, R. Xie, Z. Liu, and M. Sun, "Iterative Entity Alignment via Joint Knowledge Embeddings", International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne Australia, pp. 4258-4264, Aug. 2017. https://doi.org/10.24963/ijcai.2017/595.

[17] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model", IEEE Transactions on Neural Networks, Vol. 20, No. 1, pp. 61-80, Jan. 2009. https://doi.org/10.1109/TNN.2008.2005605.

[18] M. Schlichtkrull, T. Kipf, P. Bloem, R. Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks", European Semantic Web Conference (ESWC), Heraklion, Crete, Greece, pp. 593-607, Jun. 2018.

[19] F. Monti, O. Shchur, A. Bojchevski, O. Litany, S. Gunnemann, and M. Bronstein, "Dual-Primal Graph Convolutional Networks", arXiv preprint arXiv:1806.00770, Jun. 2018.

[20] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, and D. Zhao, "Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs", International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, pp. 5278-5284, Aug. 2019. https://doi.org/10.24963/ ijcai.2019/733.

저자소개

김 재 웅 (Jaewoong Kim)

2019년 8월 : 경북대학교 컴퓨터학부(공학사)
2020년 3월 ~ 현재 : 경북대학교 컴퓨터학부(공학석사)
관심분야 : 시맨틱 웹, 지식 그래프 임베딩, 딥 러닝, 빅 데이터

이 용 주 (Yongju Lee)

1985년 : 한국과학기술원 정보검색전공(공학석사)
1997년 : 한국과학기술원 컴퓨터공학전공(공학박사)
1998년 8월 ~ 현재 : 경북대학교 IT대학 컴퓨터학부 교수
관심분야 : 링크드 데이터, 시맨틱 웹, 빅데이터, 지식 그래프