ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 49, No. 9, pp. 773-779, 2022. 9
https://doi.org/10.5626/JOK.2022.49.9.773

개체 유형 정보를 활용한 지식 그래프 임베딩

(Knowledge Graph Embedding with Entity Type Constraints)

공 승 환 † (Seunghwan Kong)　정 찬 영 † (Chanyoung Chung)　주 수 헌 †† (Suheon Ju)　황 지 영 ††† (Joyce Jiyoung Whang)

요 약 지식 그래프 임베딩은 그래프의 구조적 특성을 반영하여 개체와 관계를 특성 공간에 나타내는 기술이다. 대부분의 지식 그래프 임베딩 모델은 그래프 구조 이외의 정보를 가정하지 않고 특징 벡터를 생성한다. 하지만 실생활과 밀접한 지식 그래프는 개체의 유형 정보 등 추가적인 정보를 얻을 수 있다. 본 논문에서는 개체의 유형이 클러스터의 역할을 수행할 수 있다는 점에 착안하여, 유형 정보를 반영할 수 있는 손실 함수를 통한 지식 그래프 임베딩 모델을 제시한다. 또한, 지식 그래프 내 관계의 주어/술어에 해당하는 유형이 제한적이라는 관찰을 토대로 개체 유형 제한에 특화된 네거티브 샘플링 기법을 제시한다. 본 논문에서 제시한 모델에 대한 링크 예측을 평가하기 위해 개체 유형 제한을 가진 지식 그래프인 SMC 데이터 셋을 생성하여 실험을 진행하였다. 링크 예측 결과는 본 모델이 네 개의 베이스라인 모델과 비교해서 뛰어난 성능을 보이는 것을 확인하였다.

키워드: 지식 그래프, 임베딩, 개체 유형, 네거티브 샘플링, 링크 예측

Abstract Knowledge graph embedding represents entities and relationships in the feature space by utilizing the structural properties of the graph. Most knowledge graph embedding models rely only on the structural information to generate embeddings. However, some real-world knowledge graphs include additional information such as entity types. In this paper, we propose a knowledge graph embedding model by designing a loss function that reflects not only the structure of a knowledge graph but also the entity-type information. In addition, from the observation that certain type constraints exist on triplets based on their relations, we present a negative sampling technique considering the type constraints. We create the SMC data set, a knowledge graph with entity-type restrictions to evaluate our model. Experimental results show that our model outperforms the other baseline models.

Keywords: knowledge graph, embedding, entity type, negative sampling, link prediction

· 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-00153, 기계학습 모델 보안 역기능 취약점 자동 탐지 및 방어 기술 개발)과 한국연구재단의 지원(2022R1A2C4001594) 및 삼성전자의 지원(I0201209-07880-01)을 받아 수행된 결과임

† 비 회 원 : 한국과학기술원 전산학부 학생
shkong@kaist.ac.kr
chanyoung.chung@kaist.ac.kr
†† 비 회 원 : 삼성전자 메모리사업부 FAB QA팀
suheon92.ju@samsung.com
††† 종신회원 : 한국과학기술원 전산학부 교수(KAIST)
jjwhang@kaist.ac.kr
(Corresponding author임)

논문접수 : 2022년 3월 7일
(Received 7 March 2022)
논문수정 : 2022년 4월 25일
(Revised 25 April 2022)
심사완료 : 2022년 4월 26일
(Accepted 26 April 2022)

1. 서 론

지식 그래프(knowledge graph)는 사람의 지식을 단방향 그래프로 표현한 것으로 여러 개의 삼중항(triplet)으로 구성되어 있다. 각 삼중항은 주어(head), 관계(relation), 술어(tail)로 구성되어 각 개체(entity) 간의 관계에 관해 서술한다. 지식 그래프는 대부분 불완전하기 때문에 표현되지 않은 지식을 추가하는 것은 하위 태스크(task)의 성능 향상에 필수적이다. 이를 위해 지식 그래프 임베딩(knowledge graph embedding)을 통해 지식 그래프 내의 개체와 관계를 특징 벡터(feature vector)로 표현하여 학습에 사용될 수 있게 한다[1]. 지식 그래프 임베딩 모델은 특징 벡터를 학습하여 링크 예측(link prediction), 삼중항 예측(triplet prediction) 등을 통하여 누락된 지식을 추가할 수 있다.

많은 실생활 데이터의 경우 개체의 유형(type) 등 그래프 구조 이상의 정보를 얻을 수 있다. 하지만 대부분의 지식 그래프 임베딩 방식은 주어진 지식 그래프의 구조만을 사용하여 특징 벡터를 학습한다. 최근 연구에 의하면, 개체의 유형 정보 등 지식 그래프에 존재하는 추가 정보를 이용하면 특징 벡터의 품질을 높일 수 있다[2,3].

본 논문에서는 개체의 유형 정보가 클러스터(cluster) 구조를 생성할 수 있다는 점에 착안하여, 클러스터링(clustering) 아이디어[4,5]를 접목한 손실 함수를 제안하였다. 해당 손실 함수를 사용하여 같은 유형 내 개체 간의 특징 벡터를 비슷하게 학습하도록 하였다. 또한, 본 논문에서는 지식 그래프 내의 관계에 대해 주어와 술어에 등장할 수 있는 개체의 유형이 제한된다는 점에 주목했다. 예시로, '졸업했다'라는 관계가 주어졌을 때, 주어에 등장할 수 있는 개체의 유형은 '사람'으로 한정되어 있고, 술어에 등장할 수 있는 개체의 유형은 '교육기관'으로 한정되어 있다. 본 논문에서는 확실한 유형 제한을 가진 지식 그래프에 특화된 네거티브 샘플링(negative sampling) 기법을 제안하였다.

확실한 유형 제한을 가진 지식 그래프에 대해 제안한 모델의 성능 평가를 진행하기 위해, 본 논문에서는 반도체 공정 내에 등장하는 다양한 사실을 바탕으로 SMC(SeMiConductor) 데이터 셋(data set)을 생성하였다. SMC 데이터 셋에서의 개체 유형 제한을 사용한 링크 예측 결과를 통해 본 논문에서 제시하는 모델이 타 모델에 비해 높은 성능을 보이며, 지식 그래프의 구조적 특성과 더불어 개체의 유형 정보를 잘 반영하는 특징 벡터를 생성하는 것을 확인할 수 있다. 또한, 주성분분석(principal component analysis)를 통한 특징 벡터 시각화로 본 논문에서 제안한 모델이 같은 유형에 속한 개체들에 대해 비슷한 특징 벡터를 생성하는 것을 확인할 수 있다.

2. 관련 연구

최근 많은 지식 그래프 임베딩 모델이 연구되고 있다. 그중 관계를 포함하지 않는 일반적인 그래프의 임베딩을 생성하는 모델인 Graph Neural Networks (GNN)로부터 발전된 지식 그래프 임베딩 모델이 존재한다. 해당 지식 그래프 임베딩 모델은 레이어(layer)를 쌓아 가중치(weight)를 학습한다. 또한, 각 개체의 특징 벡터를 이웃으로부터 직접 얻는(aggregate) 특성이 있다. R-GCN[6]은 GNN을 이용한 모델 중 가장 유명한 모델 중 하나다. R-GCN은 GNN의 특성에 추가로 이웃 정보를 받을 때 다른 관계마다 다른 가중치를 사용하여 지식 그래프의 특성을 활용한다.

변환 거리 모델(translational distance model)은 점수 함수(scoring function)를 거리 기반으로 정의한 모델이다. 변환 거리 모델은 삼중항의 관계를 바탕으로 주어와 술어의 특징 벡터를 변환한 후 둘 사이의 거리를 측정하여 점수를 계산한다. TransE[7]는 대표적인 변환 거리 모델로, 개체와 관계의 특징 벡터를 같은 공간에 표시하여 점수를 계산한다.

시맨틱 매칭 모델(semantic matching model)은 변환 거리 모델과는 다르게 유사도(similarity)에 기반하여 점수를 측정한다. 대표적인 시맨틱 매칭 모델로는 DistMult[8]가 있으며, 해당 모델은 각 임베딩 차원에서의 주어와 술어 사이의 상관관계를 통해 삼중항의 신뢰도를 계산한다. 하지만 변환 거리 모델과 시맨틱 매칭 모델의 대부분은 개체의 유형 정보를 직접 학습할 수 없다는 단점이 존재한다.

임베딩 모델과 다르게 규칙 기반 모델은 그래프 내의 패턴을 찾아 컨피던스(confidence) 점수를 이용하여 삼중항의 점수를 계산한다. AnyBURL[9]은 각 패턴을 개체 혹은 개체 변수를 사용한 경로로 나타내며, 해당 경로의 빈도를 통해 컨피던스 점수를 계산한다. 개체 변수가 유형에 가까운 역할을 수행하지만, 대부분의 규칙 기반 모델 역시 직접적인 유형 정보를 반영하지 않는다.

실생활에서는 개체에 유형 정보가 주어져 있는 경우가 존재한다. 일부 지식 그래프 임베딩 방법은 유형 정보를 반영한 특징 벡터를 학습하는 데 초점을 두고 있다. SSE[2]는 같은 유형에 속한 개체 간의 거리가 가까워야 한다는 가정에 기인하여 특징 벡터를 학습하며, TKRL[3]은 각 유형에 해당하는 벡터 공간으로의 투영을 통해 특징 벡터를 학습한다. SSE와 TKRL이 제시하는 임베딩 방법은 상기한 방법들과 다른 유사도 가정을 통해 특징 벡터를 학습하며, 확실한 유형 정보가 주어진 경우의 학습 전략을 제시한다.

3. 실험 데이터

유형이 주어진 개체에 대한 임베딩 방법의 성능을 평가하기 위해, 현업 엔지니어들의 반도체 공정 지식을 기반으로 하여 지식 그래프인 SMC 데이터 셋을 제작하여 사용하였다. 데이터셋의 개체와 관계는 반도체 공정 내에 등장하는 다양한 요소 및 요소 간의 관계를 나타낸다.

SMC 데이터 셋 내의 개체에는 해당하는 유형에 대한 정보가 암호화 되어 주어져 있다. 예를 들어, '불량:FMaa'는 불량 유형에 해당하는 개체이며, '베이:Bkg'는 베이 유형에 해당하는 개체다. 또한, SMC 데이터 셋 내의 각 관계는 주어 및 술어에 등장할 수 있는 유형이 정해져 있다. 예를 들어 '위치해 있다' 관계의 주어로는 설비 유형의 개체만 등장할 수 있으며, 술어로는 베이 유형의 개체만 등장할 수 있다. 이와 같은 유형 정보를 지키며(설비: Efg, 위치해 있다, 베이:Bcr), (불량: FMah, 모니터링된다, 지수: Yak)와 같은 삼중항이 만들어진다. 그림 1은 SMC 데이터 셋이 포함하고 있는 유형과 유형 간의 관계에 대해 정리한 그림이다.

본 논문에서는 SMC 데이터 셋을 바탕으로 만들어진 총 세 가지의 데이터 셋(SMC_v1, SMC_v2, SMC_v3)을 사용하여 임베딩 모델의 성능 평가를 진행한다. SMC_v2는 SMC_v1에서 관계 '포함한다'가 추가되었고, SMC_v3는 SMC_v2에서 관계 '먼저 이루어진다'가 추가된 데이터 셋이다. 각 데이터 셋은 다양한 크기를 가져 임베딩 모델의 성능을 다각적으로 평가할 수 있다. 각 데이터 셋의 수치 정보는 표 1에 정리되어 있다.

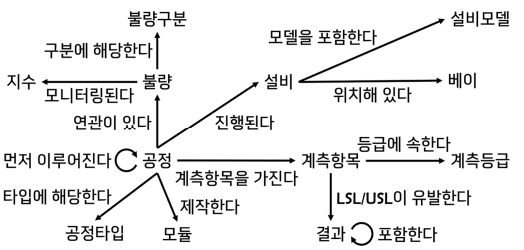


그림 1 SMC 데이터 셋의 유형과 유형 간의 관계
Fig. 1 Types and relations in SMC datasets

표 1 SMC 데이터 셋의 요약
Table 1 Summary of SMC datasets

	SMC_v1	SMC_v2	SMC_v3
No. Entities	2,370	2,411	2,432
No. Relations	12	13	14
No. Types	12	12	12
No. Triplets	5,412	5,509	14,534

4. 모 델

본 논문에서는 개체 유형 제한을 가진 지식 그래프 모델에 대해, 유형 정보를 효과적으로 반영할 수 있는 지식 그래프 임베딩 모델 및 학습 방법을 제시한다.

4.1 지식 그래프 임베딩

지식 그래프 G는 $G=(V,R,E)$로 정의되며, V와 R은 개체 및 관계의 집합, $E=\{(h,r,t): h,t\in V, r\in R\}$는 삼중항의 집합을 의미한다. 유형의 집합은 $T=\{t_1,...,t_k\}$로 정의한다. 개체 $v\in V$에 대한 유형은 $\tau(v)\in T$로 표현하며, 유형 t에 속하는 개체의 집합은 V_t로 표기한다.

지식 그래프 임베딩은 지식 그래프에 속해 있는 각 개체 $h\in V$와 관계 $r\in R$에 대한 벡터 표현 $\mathbf{h}$와 $\mathbf{r}$을 생성하는 것을 목표로 한다. 일반적으로, 지식 그래프 임베딩 모델은 삼중항 (h,r,t)에 대해 모델의 가정을 반영하는 점수 함수 $f(h,r,t)$를 사용하여 삼중항의 신뢰도를 측정한다. 예를 들어, TransE[7]는 $f(h,r,t)=-\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|$를 점수 함수로 사용하여 $\mathbf{h}+\mathbf{r}$과 $\mathbf{t}$가 가까울수록 더 높은 점수를 부여한다. 임베딩 모델은 손실 함수를 사용하여 올바른 삼중항에 대해서는 높은 점수, 틀린 삼중항에 대해서는 낮은 점수를 부여하도록 벡터 표현을 학습한다.

4.2 유형 정보를 반영한 손실 함수

지식 그래프의 구조적 특성을 유지하면서 동시에 유형 정보를 반영한 특징 벡터를 생성하기 위해, 본 논문에서는 서로 다른 가정을 가진 두 가지의 손실 함수를 통해 최종적인 손실 함수를 제시한다.

첫 번째 손실 함수는 지식 그래프의 구조적인 특성을 반영하도록 설계되었으며, 이는 기존의 지식 그래프 모델에서 사용하는 가정과 동일하다. 해당 손실 함수를 식으로 나타내면 다음과 같다.

$$L_s=\sum_{(h,r,t)\in E}\sum_{(h^*,r^*,t^*)\in E^*}[\gamma-f(h,r,t)+f(h^*,r^*,t^*)]_+$$

위 식에서 γ는 마진, (h^*,r^*,t^*)은 (h,r,t)를 오염시켜(corrupt) 생성한 삼중항, $[x]_+=\max(0,x)$를 의미한다. 본 모델에서는 TransE의 점수 함수를 사용하여 모델을 학습한다.

두 번째 손실 함수는 개체의 유형 정보를 반영하도록 설계하였다. 본 논문에서는 각 유형에 속한 개체들을 하나의 클러스터로 취급할 수 있다는 가정에 주목하여, k-Means 클러스터링의 목적 함수를 바탕으로 손실 함수를 정의하였다. 이를 식으로 표현하면 다음과 같다.

$$L_t=\sum_{t\in T}\sum_{e\in V_t}\|\mathbf{e}-\mathbf{c}_t\|,\ \mathbf{c}_t=\left(\sum_{e\in V_t}\mathbf{e}\right)/|V_t|$$

위 식에서 $\mathbf{c}_t$는 클러스터 V_t의 중앙을 의미하며, 이는 V_t에 속한 모든 개체의 특징 벡터의 평균으로 정의한다.

모델의 최종적인 손실 함수는 다음과 같이 표현된다.

$$L = L_s + \lambda \cdot L_t$$

두 손실 함수 간의 중요도를 조절하기 위해 가중치 λ를 사용한 가중합 형식의 손실 함수를 사용한다. 모델은 위 손실 함수를 줄이는 방향으로 최적화를 진행하며, 이를 통해 지식 그래프의 구조와 유형 정보를 동시에 반영하는 지식 그래프 임베딩을 생성한다.

4.3 유형 정보를 반영한 네거티브 샘플링

구조적 특성을 반영하는 손실 함수에서 옳은 삼중항과 쌍을 이루기 위해 네거티브 샘플링이 필요하다. 네거티브 샘플링은 지식 그래프 임베딩 모델의 학습 시 임의로 삼중항을 오염시키는 것이며, 임베딩 모델은 오염된 삼중항에 대해 낮은 점수를 부여하도록 학습한다. 지식 그래프의 불완전성으로 인해, 네거티브 샘플링은 지식 그래프 임베딩 모델의 중요한 부분을 차지한다. 효과적인 네거티브 샘플링은 정교한 특징 벡터를 생성하는 데 도움을 준다. 본 논문에서는 유형 제한이 확실한 데이터 셋에 특화된 네거티브 샘플링을 제시한다.

많은 수의 기존 모델에서는 네거티브 샘플링이 주어진 삼중항에 대해 임의로 주어 혹은 술어를 다른 요소로 대체하는 방식으로 진행된다. 하지만 각 관계 별로 유형 제한이 확실한 경우, 위의 방식은 모델 학습에 그다지 효과적이지 않다. 예를 들어, '위치해 있다' 관계는 주어로 '설비' 유형의 개체만, 술어로 '베이' 유형의 개체만 등장가능하다. 하지만 주어를 '계측등급' 유형의 개체로 오염한 경우, 이미 관계의 유형 제한으로 인해 해당 삼중항은 자동으로 틀린 삼중항으로 판단할 수 있다. 따라서 모델 학습에 도움을 주기 위해서는, 네거티브 샘플링을 진행할 때 주어 혹은 술어를 같은 유형 내의 다른 개체로 대체하여야 한다.

각 유형에 속한 개체 수의 극명한 차이로 인해 기존과 같은 방법으로 주어 및 술어를 동일한 확률로 오염시키는 것은 불균형한 네거티브 샘플링을 유발할 위험이 있다. '등급에 속한다' 관계에 대한 네거티브 샘플링을 진행한다고 가정했을 때, 주어에 위치하는 '계측항목' 유형의 개체는 총 318개 존재하는데 반해, 술어에 위치하는 '계측등급' 유형의 개체는 5개밖에 존재하지 않는다. 위의 비율을 반영하지 않은 네거티브 샘플링을 진행한다면 '계측항목' 유형에 속한 개체 간의 구분을 학습하기 힘들 것이다. 따라서 본 논문에서는 유형의 크기를 반영한 네거티브 샘플링 기법을 제시한다.

삼중항 $(h,r,t) \in E$가 주어질 때, 주어와 술어의 유형은 각각 $\tau(h)$와 $\tau(t)$로 주어진다. $|V_{\tau(h)}|/(|V_{\tau(h)}|+|V_{\tau(t)}|)$의 확률로 주어를 $V_{\tau(h)}$ 내의 다른 개체로 오염시키며, 반대로 $|V_{\tau(t)}|/(|V_{\tau(h)}|+|V_{\tau(t)}|)$의 확률로 술어를 $V_{\tau(t)}$ 내의 다른 개체로 오염시킨다. 위의 예시에서, '등급에 속한다' 관계에 대해 주어가 오염될 확률은 318/323이 되며 술어가 오염될 확률은 5/323이 된다. 이와 같은 방법을 통해 크기가 동일하지 않은 유형 간의 불균형을 해소할 수 있는 네거티브 샘플링을 진행할 수 있다.

모든 네거티브 샘플링을 진행할 때에는 오염된 삼중항이 지식 그래프 내에 포함되어 있을 가능성이 존재한다. 이를 방치한다면 올바른 삼중항을 틀린 삼중항으로 학습하여 모델의 학습 방향에 문제가 생기기 때문에 이와 같은 경우가 존재하는지 검사하여 올바른 삼중항을 틀린 삼중항으로 처리하지 않도록 하였다.

5. 실 험

5.1 실험 세팅 및 평가 방법

본 논문에서는 3장에서 언급한 SMC 데이터 셋을 이용하여 우리 모델과 4개의 서로 다른 베이스라인(baseline)의 성능을 비교하였다. 4개의 베이스라인은 AnyBURL, DistMult, R-GCN, TransE로 각각 규칙 기반 접근 모델, 시맨틱 매칭 모델, GNN 기반 모델, 변환 거리 모델에서 선택되었다.

모델의 평가를 위해 보편적으로 사용되는 링크 예측을 사용하였다. 링크 예측은 불완전한 삼중항 $(h,r,?)$ 혹은 $(?,r,t)$가 주어졌을 때 ?에 올 수 있는 개체를 점수 함수를 기반으로 예측하는 것이다. 링크 예측을 평가하기 위해 보편적으로 사용되는 평가 방식(metric)인 Mean Rank (MR), Mean Reciprocal Rank (MRR), Hit@10을 사용하였다. 추가로, 더욱 정확한 성능 평가를 위해, 세 가지 평가 방식 모두 [7]에서 제안한 'filtered' 설정으로 측정되었다. MR의 경우 낮은 값을 가질수록 좋은 성능을 기록하는 것이고, MRR과 Hit@10의 경우 높은 값을 가질수록 좋은 성능을 의미한다. 각 평가 방식은 서로 다른 의미를 가지고 있기 때문에 한 가지 평가 방식을 선택하여 모델간의 우위를 비교하기는 힘들다. 따라서 [1]에서의 평가 방식에 착안하여 3가지 평가 방식을 종합하여 모델의 성능을 평가할 수 있는 방식을 사용하였다.

평가 방식에 대한 손실률을 다음과 같이 정의한다.

$$\text{손실률}_{\text{평가 방식}} = \left| \frac{\text{최고점수} - \text{모델점수}}{\text{최고점수}} \right|$$

최고점수는 해당 평가 방식에서 가장 좋은 값을 의미하고 모델점수는 손실률을 구하고자 하는 모델의 값을 나타낸다. 낮은 손실률을 가질수록 모델의 종합적인 성능이 뛰어나다는 것을 말한다. 3가지 평가 방식을 종합하기 위해 각 평가 방식의 손실률을 모두 다 더하여 모델의 성능을 비교한다.

$$\text{손실률} = \text{손실률}_{\text{MR}} + \text{손실률}_{\text{MRR}} + \text{손실률}_{\text{Hit@10}}$$

표 2 하이퍼파라미터 요약
Table 2 Summary of hyperparameters

Hyperparameters	Values
lr_{tra}	$\{0.5, 1.0, 2.0, 3.0, 4.0, 5.0\}$
lr_{sem}	$\{0.1, 0.05, 0.01, 0.005, 0.001\}$
margin	$\{0.5, 1.0, 1.5, 2.0\}$
number of bases	$\{5, 10, 15, 20\}$
number of layers	$\{1, 2, 3\}$
λ	$\{0.001, 0.0001, 0.00001\}$
regularization rate	$\{0.5, 0.1, 0.05, 0.01\}$

각 모델의 하이퍼파라미터는 그리드(grid) 탐색을 통해 정하였다. 하이퍼파라미터에 대한 자세한 정보는 표 2에서 확인할 수 있다.

표 2는 전체 베이스라인의 하이퍼파라미터를 나타내며 각 베이스라인마다 필요한 하이퍼파라미터를 선택하여 그리드 탐색을 진행하였다. 예를 들어, R-GCN은 lr_{sem}, number of bases, number of layers, regularization rate 4개의 하이퍼파라미터를 사용하여 실험이 진행되었다. 변환 거리 모델과 시맨틱 매칭 모델은 서로 다른 최적화 기법(optimizer)을 사용하기 때문에, 학습률(learning rate)의 범위를 다르게 설정하였다.

R-GCN을 제외한 모든 모델은 1,000번의 실행 횟수(epochs)를 사용하였고, 10번의 실행마다 검증을 실행하여 MR, MRR, Hit@10의 결과를 저장하였다. R-GCN의 경우는 10,000번의 실행 횟수를 사용하였고, 500번의 실행마다 검증을 실행하였다. 모든 하이퍼파라미터에 대해 실험이 종료되고 검증 성능 중에 가장 낮은 손실률을 가지는 모델의 파라미터를 불러와 실험을 진행하였다.

5.2 개체 유형 제한 조건을 이용한 링크 예측

본 실험에서 개체 유형 제한 조건을 이용한 네거티브 샘플링을 본 모델에 적용하여 성능을 확인해보았다. 실험의 테스트 과정에서는 네거티브 샘플링과 유사한 방법을 사용하여 후보 개체를 선별하여 진행하였다. 예를 들어, 삼중항 $(h, r, ?)$가 테스트 과정에서 주어졌을 때 ?에 올 수 있는 유형이 t_k라고 한다면 ? 개체의 후보는 $t \in V_{v_k}$를 만족하고 학습 때 (h, r, t)의 삼중항이 나타나지 않은 개체로 제한된다.

표 3은 세 가지 SMC 데이터 셋 상에서 베이스라인 모델과 본 모델의 실험 결과를 보여준다. 가장 좋은 실험 결과는 굵게 표시하였고 두 번째로 좋은 실험 결과는 밑줄로 표시하였다. 본 모델은 모든 데이터 셋에서 가장 좋거나 두 번째로 좋은 성능을 발휘하는 것을 확인할 수 있다. 특히 Hit@10 부분에서는 모든 모델의 성능을 뛰어넘는 것을 확인할 수 있다.

표 4는 5.1장에서 언급한 손실률과 비슷한 방법을 통해 각 데이터 셋에서 모델이 얼마나 좋은 성능을 냈는지 종합적인 평가를 제공하였다. 하이퍼파라미터의 그리드 탐색 때와는 다르게 최고점수에는 평가 방식에서 가장 좋은 결과를 삽입하고, 모델점수에는 해당 모델이 평가 방식에서 갖는 결과를 삽입하였다. 가장 좋은 결과는 굵게 표시하였고 두 번째로 좋은 결과는 밑줄로 표시하였다.

본 모델은 손실률 부분에서 모든 베이스라인 모델을 상회한다. 특히 TransE 모델과 비교하면 본 모델에 사용된 클러스터링 아이디어를 적용한 손실 함수가 제대로 작동하는 것을 확인할 수 있다.

표 3과 4를 통해 각 베이스라인과 본 모델이 링크 예측에서 어느 정도의 성능을 발휘하는지 확인할 수 있다. 본 모델이 전체적으로 상당한 성능을 보여주었고, 그 뒤로 TransE가 모든 데이터 셋에서 양호한 결과를 보였다. 다른 베이스라인의 경우 SMC_v1과 SMC_v2에서 나쁘지 않은 결과를 기록했지만, SMC_v3에서 TransE

표 4 실험 결과의 손실률
Table 4 Total loss rate of the link prediction results

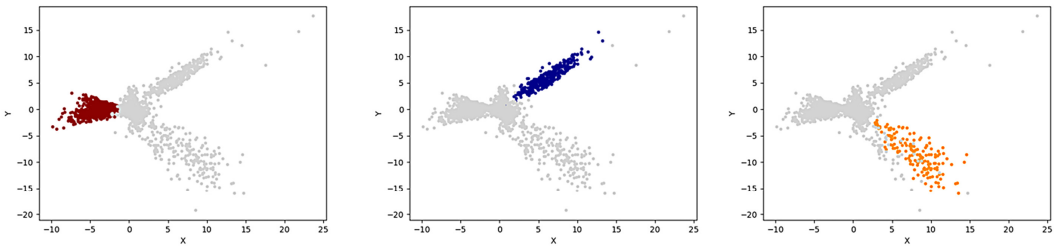
Model	SMC_v1	SMC_v2	SMC_v3
AnyBURL (10s)	2.077	1.874	2.459
AnyBURL (1000s)	0.854	0.430	1.669
DistMult	0.657	0.672	5.622
R-GCN	0.123	0.079	5.386
TransE	0.208	0.080	0.700
Our Model	**0.036**	**0.012**	**0.184**

표 3 개체 유형 제한 조건을 이용한 링크 예측 결과
Table 3 Link prediction results with entity type constraints

Model	SMC_v1			SMC_v2			SMC_v3		
	MR	MRR	Hit@10	MR	MRR	Hit@10	MR	MRR	Hit@10
AnyBURL (10s)	82.4	0.390	0.629	81.9	0.430	0.599	13.7	0.663	0.780
AnyBURL (1000s)	54.5	0.499	0.703	45.1	**0.500**	0.714	10.6	0.670	0.861
DistMult	46.9	0.475	0.699	47.6	0.465	0.645	24.4	0.226	0.470
R-GCN	**30.6**	0.461	0.721	32.5	0.474	0.721	23.6	0.265	0.485
TransE	36.3	0.503	0.736	33.9	0.490	0.728	7.7	**0.879**	0.949
Our Model	31.7	**0.504**	**0.749**	**32.1**	0.494	**0.731**	**4.5**	0.717	**0.965**

표 5 개체 유형 제한 조건을 사용하지 않은 링크 예측 결과
Table 5 Link prediction results without entity type constraints

Model	SMC_v1			SMC_v2			SMC_v3		
	MR	MRR	Hit@10	MR	MRR	Hit@10	MR	MRR	Hit@10
AnyBURL(10s)	167.4	0.358	0.568	174.4	0.429	0.601	18.7	0.660	0.780
AnyBURL(1000s)	93.8	0.477	0.684	88.4	0.499	0.714	105.9	0.570	0.691
DistMult	230.9	0.385	0.544	258.6	0.364	0.518	33.7	0.203	0.436
R-GCN	63.0	0.415	0.682	83.2	0.447	0.699	25.1	0.247	0.483
TransE	42.1	0.496	0.720	40.3	0.483	0.716	8.5	0.873	0.944
Our Model	37.2	0.498	0.737	36.4	0.491	0.720	5.1	0.716	0.964



(a) equipment (b) measure (c) process

그림 2 SMC_v1 데이터 셋에서 본 모델을 학습시켰을 때 유형별 개체의 특징 벡터
Fig. 2 Feature vectors of each type of entity in the SMC_v1 data set

와 본 모델과 비교해서 떨어지는 결과를 보여주었다. 이는 표 1의 SMC 데이터 셋의 정보에서 확인할 수 있는 추가된 하나의 관계와 이에 해당하는 약 9,000개의 삼중항에 대해 비교적 제대로 된 예측을 하지 못한다는 것을 말한다.

5.3 개체 유형 제한 조건을 사용하지 않은 링크 예측

개체 유형 제한 조건을 사용할 수 있는 데이터 셋에 특화된 모델을 사용하는 것이 어떤 이점이 존재하는지 알아보기 위해 개체 유형 제한 조건을 이용하지 않았을 때 링크 예측의 성능 변화를 알아보았다. 표 5는 개체 유형 제한 조건을 사용하지 않았을 때의 실험 결과를 나타낸다.

표 5의 결과를 확인하면 개체 유형 제한 조건을 이용한 실험과 비교해서 모든 모델에서 전체적인 성능의 하락이 있는 것을 확인할 수 있다. 특히, AnyBURL, R-GCN, DistMult에서 TransE 및 본 모델과 비교해서 큰 하락이 존재한다. 이를 통해 AnyBURL, R-GCN, DistMult은 개체 유형 제한 조건이 존재하지 않으면 유형 정보를 제대로 반영하지 못하는 것을 확인할 수 있다. TransE와 본 모델의 경우 MR을 제외하고 성능이 크게 감소하지 않아 개체 유형 제한 조건을 사용하지 않아도 어느 정도 유형 정보를 반영하고 있다는 것을 알 수 있다.

5.4 시각화를 통한 손실 함수 성능 확인

본 모델의 손실 함수가 유형 정보를 반영하는 특징 벡터를 생성하는지 알아보기 위해 몇몇 유형에 속하는 개체의 특징 벡터를 시각화하였다. 주성분 분석을 통해 전체 개체의 특징 벡터를 시각화한 이후 목표 유형에 속하는 개체의 특징 벡터 색을 칠했다. 그림 2는 각각 설비, 계측항목, 공정에 속하는 개체의 특징 벡터가 클러스터를 잘 이루고 있는 모습을 보여준다. 이를 통해 본 모델이 구조적 특징뿐 아니라 유형 정보 또한 반영하여 임베딩을 학습하는 것을 확인할 수 있다.

6. 결 론

본 논문에서는 데이터 셋에 존재하는 개체 유형 정보를 반영하는 지식 그래프 임베딩 모델을 제시하였다. 동일한 유형에 속해 있는 개체 간 거리를 가깝게 유지하는 손실 함수와 주어와 술어에 올 수 있는 유형의 비율에 따라 불균형을 맞춰 주는 네거티브 샘플링을 통해 모델이 데이터 셋에 특화될 수 있게 하였다. 개체 유형 정보를 가지고 있는 실제 반도체 데이터 셋에서의 링크 예측 결과는 본 모델이 상당한 성능을 보이는 것을 알려준다.

References

[1] C. Chung and J. J. Whang, "Knowledge Graph Embedding via Metagraph Learning," *Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2212-2216, Jul. 2021.

[2] S. Guo, Q. Wang, B. Wang, L. Wang, and L. Guo,

"Semantically Smooth Knowledge Graph Embedding," *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 84-94, Jul. 2015.

[3] R. Xie, Z. Liu, and M. Sun, "Representation Learning of Knowledge Graphs with Hierarchical Types," *Proc. of the 25th International Joint Conference on Artificial Intelligence*, pp. 2965-2971, Jul. 2016.

[4] J. J. Whang, Y. Hou, D. F. Gleich, and I. S. Dhillon, "Non-exhaustive, Overlapping Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 11, pp. 2644-2659, Nov. 2019.

[5] J. J. Whang, R. Du, S. Jung, G. Lee, B. Drake, Q. Liu, S. Kang, and H. Park, "MEGA: Multi-View Semi-Supervised Clustering of Hypergraphs," *Proc. of the VLDB Endowment*, Vol. 13, No. 5, pp. 698-711, Jan. 2020.

[6] M. Schlichtrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," *Proc. of the European Semantic Web Conference*, pp. 593-607, Jun. 2018.

[7] A. Bordes, N. Usunier, A. Garcid-Duran, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data," *Proc. of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, pp. 2787-2795, Dec. 2013.

[8] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding Entities and Relations for Learning and Inference in Knowledge Bases," *Proc. of the 3rd International Conference on Learning Representations*, May 2015.

[9] C. Meilicke, M. W. Chekol, D. Ruffinelli, and H. Stuckenschmidt, "Anytime Bottom-Up Rule Learning for Knowledge Graph Completion," *Proc. of the 28th International Joint Conference on Artificial Intelligence*, pp. 3137-3143, Aug. 2019.

공 승 환

2021년 KAIST 전산학부, 수리과학과 졸업(학사). 2021년~현재 KAIST 전산학부 석사과정(빅데이터 지능 연구실 소속). 관심분야는 기계학습, 지식 그래프, 빅데이터

정 찬 영

2021년 KAIST 전산학부, 수리과학과 졸업(학사). 2021년~현재 KAIST 전산학부 박사과정(빅데이터 지능 연구실 소속). 관심분야는 빅데이터, 기계학습, 지식 그래프, 그래프 마이닝

주 수 헌

2014년 서울대학교 조선해양공학과 졸업(학사). 2020년 서울대학교 조선해양공학과 졸업(박사), 2020년~현재 삼성전자 메모리 사업부 FABQA팀. 관심분야는 통계적 데이터 분석, 데이터 마이닝, 기계학습

황 지 영

2010년 이화여자대학교 컴퓨터공학과 졸업(학사). 2015년 텍사스 오스틴 대학교(The University of Texas at Austin) 컴퓨터 과학과 졸업(박사). 2016년 3월~2020년 6월 성균관대학교 소프트웨어학과 조교수(빅데이터 연구실 운영), 2020년 7월~현재 KAIST 전산학부 조교수(빅데이터 지능 연구실 운영). 관심분야는 빅데이터, 데이터 마이닝, 그래프 마이닝, 기계학습