근거 기반 지식 그래프
병합 모델 성능 분석서

2022. 12. 10

수탁 기관: 연세대학교 산학협력단

제 1 장 서론

지식 그래프(Knowledge Graph)[1,2]는 질의응답, 추천 시스템, 대화 생성[3,4,5] 등과 같은 자연어 처리 응용 분야에서 여러가지 방법으로 활발하게 활용되고 있다. 방대한 양의 시멘틱 웹 기반 정보들을 엔티티(Entity)와 릴레이션(Relation)으로 구성된 그래프로 인코딩하여 효율적인 정보 처리를 할 수 있어 지식 그래프에 대한 관심이 점점 높아지고 있다. 인터넷의 성장과 함께 지식 그래프의 크기도 거대해짐에 따라 지식 그래프 병합, 추출, 분석 등과 같은 지식 그래프 처리에 대한 많은 방법들이 제안되고 있다. 특히, 지식 그래프 데이터를 통합하여 정보 검색 성능을 높이는 지식 그래프 병합 연구[6,7,8,9]에 대한 필요성이 지속적으로 요구되고 있다.

지식 그래프 병합 연구에서는 병합되는 그래프끼리의 종류가 같거나 다른지에 따라 동종 지식 그래프(Homogeneous knowledge graph) 병합과 이종 지식 그래프 (Heterogeneous knowledge graph)로 각각 구분된다. 기존 동종 지식 그래프 병합 방법에서는 그래프 구조의 계층적 정보를 통해 같은 상위 레벨에 속하는 엔티티들을 하나의 그룹으로 묶어주어 그래프를 병합한다[10]. 이종 지식 그래프 간의 병합 방법에서는 각 그래프에 존재하는 동등 엔티티들(Equivalence entities)을 하나로 통합하여 그래프를 병합한다[11]. 하지만 이러한 기존 지식 그래프 병합 연구에서는 엔티티 중심으로만 그래프 병합을 수행하여 그래프의 또 다른 구성 요소인 릴레이션(Relation) 정보를 충분히 활용하지 못하고 있다. 또한, 이러한 방법을 통해 병합된 그래프가 정보 검색에 직접적으로 얼마나 영향을 미치는지에 대한 분석이 제대로 이루어지지 않았다.

본 문서에서는 이러한 문제를 해결하고자 동종 지식 그래프의 릴레이션 중심 지식 그래프 병합 모델을 제안하고 병합된 지식 그래프가 실질적으로 정보 검색 성능 향상에 도움이 되는 것을 보인다. 릴레이션이 가지는 의미적 정보를 토대로 의미적 유사 릴레이션(Semantic similar relation)을 통합하여 그래프를 병합한다. 뿐만 아니라 지식 그래프의 릴레이션 규칙(Relation rule)을 통해 통계적 동등 릴레이션(Statistic similar relation) 병합을 수행하여 최종적으로 병합된 릴레이션 집합을 생성한다. 이를 통해 얻어진 그래프와 병합 전 그래프에 대해 정보 검색 성능을 비교하여 릴레이션 중심 동종 지식 그래프 병합의 필요성을 확인한다.

다음으로 2장에서는 지식 그래프 병합 연구에 대한 기존 연구의 특징 및 장단점에 대해 기술한다. 3장에서는 본 문서에서 제안하는 릴레이션 기반의 지식 그래프 병합 모델의 세부 동작 과정을 제시한다. 4장에서는 제안하는 방법을 적용한 지식 그래프 병합 모델을 통한 검색 성능 분석 결과를 제시한다. 5장에서는 본 문서에서 제안하는 모델의 의미와 결론을 제시한다.

제 2 장 기술 동향 분석

구글의 검색 서비스 기술로써 지식 그래프가 도입된 이후 다양한 분야에서 지식 그래프와 융합한 신흥 기술들이 자리 잡고 있다[12]. 인터넷이 급속도로 발달함에 따라 지식 그래프의 규모 또한 거대해지고 있는 상황에서 지식 그래프를 병합하여 중복된 데이터를 하나로 통합하거나 서로 관련있는 데이터들끼리 묶어 정제된 그래프를 생성하는 연구가 활발히 이루어지고 있다[13].

동종 지식 그래프 병합에서는 서로 공통된 속성을 갖는 엔티티들끼리 같은 그룹으로 묶어주어 그래프 데이터 처리 시 같은 그룹끼리는 동일한 계산이 이루어지도록 한다. 지식 그래프에 대한 질의들을 정규 표현식으로 변환한 후 같은 정규 표현식을 가지는 엔티티들끼리 하나의 집합으로 간주하여 그래프를 표현하는 방법[14]이 고안되었고, 엔티티의 동등 클래스 (Equivalence class) 속성을 고려한 쿼션트 그래프(Quotient graph) 개념을 도입하여 엔티티의 동일 관계(Equivalence relationship)을 찾아 하나로 합쳐주는 방법[15]이 시도되었다. 또한, 엔티티들의 계층적 정보를 활용하여 동일한 상위 개념으로 묶일 수 있는 엔티티들끼리 하나의 노드에 구성하여 그래프 데이터 처리를 빠르게 할 수 있는 방법[16]이 연구되었다.

이종 지식 그래프 병합에서는 서로 다른 그래프 구조 프레임워크를 사용하지만 같은 엔티티를 뜻하는 노드들을 찾아 이종 그래프 간의 병합을 수행한다. 엔티티가 가지는 이름, 속성값 등의 심볼릭 특징(Symbolic features)을 통한 엔티티 유사도를 측정하여 동일 엔티티를 찾아내는 방법[17]이 고안되었다. 최근엔 인공신경망을 이용하여 엔티티를 벡터로 표현하여 암시적 의미 매칭으로 동일 엔티티를 찾는 연구가 활발히 이루어지고 있다[18].

이러한 지식 그래프 병합 연구를 통해 방대한 지식 그래프가 사용자에게 필요한 데이터를 중점으로 정제되어 더욱 정확한 사용자 요구에 대한 결과를 추출해주거나 그래프 데이터 이해에 도움이 되는 것을 보여준다[19].

제 3 장 모델 설계

본 문서에서 제안하는 동종 지식 그래프의 릴레이션 중심 병합 모델은 엔티티 중심 병합이었던 기존 연구와 달리 그래프 릴레이션의 의미적, 통계적 정보를 통해 그래프를 병합한다. 또한, 릴레이션 중심으로 병합된 그래프가 검색 성능을 향상시켜 도움이 된다는 점을 보여준다.

첫 번째 단계에서는 동종 지식 그래프 내에서 의미적으로 유사한 뜻을 가지는 릴레이션끼리 결합한 의미적 유사 릴레이션 집합(Semantic similar relation set)을 구성하여 그래프 데이터를 병합한다. 두 번째 단계에서는 그래프가 가지는 릴레이션 규칙을 통해 서로 다른 의미적 유사 릴레이션 집합들을 통합하여 통계적 동등 릴레이션 집합(Statistic equivalence relation set)을 생성한다. 이렇게 생성된 동등 릴레이션 집합은 유사한 의미를 내포하는 릴레이션 정보를 공유할 수 있어 그래프 데이터 처리 시 기존 전체 그래프의 정보를 병합된 정보를 통해 처리할 수 있도록 만든다.

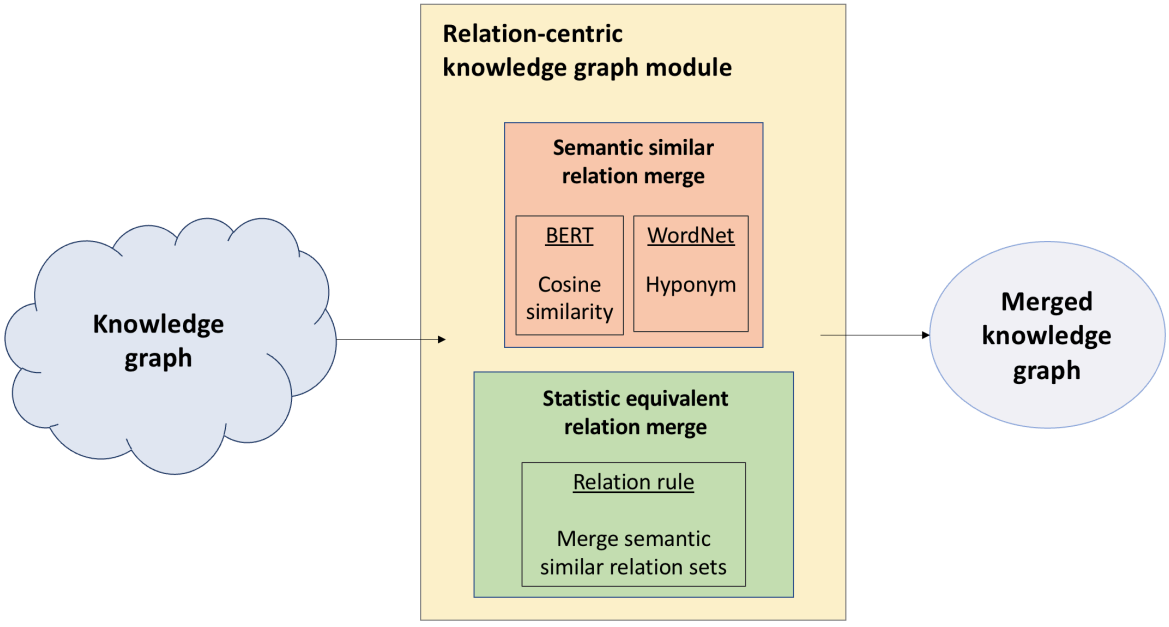

그림 1. <모델 전체 구조도>

첫 번째 릴레이션 병합 단계인 의미적 유사 릴레이션 집합 생성을 위해서는 릴레이션이 가지는 단어들끼리의 의미적 유사도(Semantic similarity)를 구하여 특정 값(Threshold) 이상인 유사도 값을 갖는 릴레이션들끼리 하나의 집합으로 구성한다. 방대한 말뭉치 데이터를 통해 단어들의 문맥적인 의미 정보(Context semantic information)를 사전 학습한 언어 모델인 BERT에 릴레이션 단어들을 입력하여 나온 임베딩 값들 간의 코사인 유사도(Cosine similarity)를 계산하여 의미적 유사도 점수를 산출한다.

그래프 구성요소인 릴레이션 $R = \{r_1, r_2, r_3, \dots r_n\}$ 을 BERT 언어 모델에 입력하면 릴레이션 단어의 임베딩 값 $R' = e_1, e_2, \dots e_n$ 을 출력한다. 출력된 릴레이션 임베딩 값끼리 코사인 거리를 계산을 통해 두 릴레이션 벡터 간의 유사한 정도를 측정하여 의미적 유사도 점수 sim_score를 산출한다. 특정 값 t 이상의 sim_score 를 갖는 릴레이션끼리 문맥적 유사 릴레이션 집합 C 를 구성한다.

$$e_1 = BERT(r_1) \tag{1}$$

$$e_2 = BERT(r_2) \tag{2}$$

$$sim\_score = cosine(e_1, e_2) \tag{3}$$

$$C_{r_1} = \{r_n | sim\_score(e_1, e_n) > t, \quad n \neq 1\} \tag{4}$$

문맥적 의미 정보를 통해 유사도를 구하여 문맥적 유사 릴레이션 집합을 생성한 후 어휘에 대한 유의어 집단(Synonym set)끼리 분류한 유의어 사전인 WordNet[21]을 활용하여 유의어 관계의 릴레이션을 같은 집합으로 포함한다.

먼저 릴레이션을 토크나이즈(Tokenize)하여 각 문맥적 유사 릴레이션 집합을 이루고 있는 릴레이션들을 단어 토큰 단위로 분해한다 ($r_1 = \{w_1, w_2, \dots, w_m\}$). 분해된 각 단어들에 대응되는 유의어 집합 S_w 을 통해 릴레이션을 이루고 있는 모든 단어에 대한 유의어 집합 S_{r_1}을 생성한다. 이 유의어 집합에 속한 단어를 가지고 있는 다른 릴레이션이 존재한다면 그 릴레이션을 같은 문맥적 유사 릴레이션 집합으로 포함하여 의미적 유사 릴레이션 집합 $S_{\{r_1, r_2, \dots\}}$을 생성한다.

$$S_{r_1} = \{s | s \in S_w, \quad w = w_1, w_2, \dots, w_m\} \tag{5}$$

$$S_{\{r_1, r_2, \dots\}} = \{s | s \in S_{r_1} or\ s \in S_{r_2} \dots\} \tag{6}$$

두 번째 릴레이션 병합 단계인 통계적 동등 릴레이션 병합을 위해서는 지식 그래프로부터 생성된 릴레이션 규칙을 활용하여 규칙 관계를 가지는 릴레이션끼리 병합한다. 각 릴레이션이 가지는 규칙 $Rule = Rule_{r_1}, Rule_{r_2}, \dots, Rule_{r_n}$을 통해 만약 다른 릴레이션과 규칙 관계를 갖는 경우 두 릴레이션을 병합한다. 만약 릴레이션이 의미적 유사 릴레이션 집합 $S_{\{r_1,r_2,\dots\}}$에 속해있는 경우라면 그 집합으로 포함하여 최종적인 유사 릴레이션 집합 을 생성한다.

$$\text{Merge } S_{\{r_1,r_2,\dots\}} \text{ and } S_{\{r_3,r_4,\dots\}} \text{ if } \{s|s \in S_{\{r_3,r_4,\dots\}}\} \in Rule_{r_1} \quad (7)$$

제 4 장 구현 및 검증

이번 장에서는 본 문서에서 제안하는 지식 그래프 병합 모델이 생성한 그래프의 정보와 병합된 그래프를 통한 검색 성능 결과에 대해 설명한다. 검색 성능을 평가하기 위한 데이터셋은 정보 검색 표준 데이터셋인 DBpedia-entity v2[20]를 사용하고 병합 모델을 적용할 지식 그래프는 1차년도 검색 모델을 통해 추출된 근거 지식 그래프를 사용한다. 근거 지식 그래프는 1차년도 검색 모델에서 추출된 질의와 연관된 엔티티로부터 1-hop 그래프를 생성하였다. 생성된 근거 지식 그래프에 대한 정보는 표 1의 첫번째 열과 같다. 근거 지식 그래프를 릴레이션 중심 그래프 병합 모델에 입력하여 병합하였을 경우 변형된 그래프에 대한 정보는 표 1의 두번째 열과 같다. 병합 후 3,869개의 릴레이션이 80개의 릴레이션 집합으로 구성되어 병합되었다.

	병합 전	병합 후
Triple 개수	5,817,108	2,353,591
Entity 개수	2,289,940	2,289,940
Relation 개수	3,869	80

표 1. <근거 지식 그래프>

표 2에서는 릴레이션이 두 가지 병합 단계를 통해 병합되는 과정의 예시를 보여준다. 첫 번째 의미적 릴레이션 병합 수행 시 문맥적 의미 정보를 통해 <artist> 릴레이션과 유사도가 0.7 이상인 문맥적 유사 릴레이션 집합을 구성하였을 경우 <artist>와 같은 유사한 의미의 릴레이션끼리 병합된다. WordNet을 통해 <artists> 릴레이션의 유의어가 포함된 <creator> 릴레이션까지 결합하면 의미적 유사 릴레이션 집합이 생성된다.

	병합 전	**병합 후**
문맥적 유사 릴레이션 병합	{artist}, {artists}	**{artist, artists}**
의미적 유사 릴레이션 병합	{artist, artists}, {creator}	**{artist, artists, creator}**
통계적 동등 릴레이션 병합	{artist, artists, creator}, {musicalArtist, musicalBand}	**{artist, artists, creator, musicalArtist, musicalBand}**

표 2. <릴레이션 병합 과정 예시>

<artist> 릴레이션에 대한 통계적 동등 릴레이션 병합을 수행하기 위해 <artist> 릴레이션과 관련된 규칙 {?a <artist> ?b → ?a <musicalBand> ?b}을 기반으로 병합을 수행하면 각 릴레이션 <musicalArtist>, <musicalBand>가 포함된 의미적 유사 릴레이션 집합이 결합되어 최종적인 유사 릴레이션 집합 {artist, artists, creator, musicalArtist, musicalBand}이 생성된다.

병합된 그래프를 질의에 대한 검색 모델에 평가하기 위해 질의 문장과 지식 그래프를 벡터 공간에 임베딩시킨 후 질의 문장의 임베딩 값과 엔티티의 임베딩 값을 비교하는 기존 모델[22]를 활용하였다. 표 3에서 보이는 것과 같이 해당 모델을 통해 검색된 엔티티 순위를 nDCG 지표로 성능 비교를 진행하였다. 병합 전 지식 그래프에서 검색을 했을 경우보다 병합 후 지식 그래프를 통한 검색 성능이 향상된 것을 확인할 수 있다.

	병합 전	**병합 후**
NDCG@10	0.477	**0.568**
NDCG@100	0.678	**0.728**

표 3. <검색 성능 결과>

제 5 장 주요 결론

지식 그래프 병합은 거대한 지식 그래프를 중복되거나 통합될 수 있는 그래프 데이터를 정제할 수 있는 작업으로 정보 검색 분야에서 점점 필요성이 요구되는 기술이다. 본 문서에서는 기존의 엔티티 중심 지식 그래프 병합 방법에서 벗어나 릴레이션을 중심으로 지식 그래프를 병합하는 방법을 제안한다. 또한, 병합된 지식 그래프가 실질적으로 검색 성능에 얼마나 영향을 미치는지를 보여준다. 이를 통해 릴레이션 중심 지식 그래프 병합을 수행하였을 경우 ~~%의 성능 향상을 보였다. 추후에는 엔티티와 릴레이션의 통합된 지식 그래프 데이터 병합에 대한 연구를 진행할 예정이다.

참고문헌

[1] Bollacker, Kurt, et al. "Freebase: a collaboratively created graph database for structuring human knowledge." Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008.

[2] Auer, Sören, et al. "Dbpedia: A nucleus for a web of open data." The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings. Springer Berlin Heidelberg, 2007.

[3] Zirui, Chen, et al. "Survey of Open-Domain Knowledge Graph Question Answering." Journal of Frontiers of Computer Science & Technology 15.10 (2021): 1843.

[4] Guo, Qingyu, et al. "A survey on knowledge graph-based recommender systems." IEEE Transactions on Knowledge and Data Engineering 34.8 (2020): 3549-3568.

[5] Zhao, Xiangyu, et al. "Multiple knowledge syncretic transformer for natural dialogue generation." Proceedings of The Web Conference 2020. 2020.

[6] Hogan, Aidan, et al. "Knowledge graphs." ACM Computing Surveys (CSUR) 54.4 (2021): 1-37.

[7] Cao, Jiahang, et al. "Knowledge Graph Embedding: A Survey from the Perspective of Representation Spaces." arXiv preprint arXiv:2211.03536 (2022).

[8] Š. Čebirić, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika. 2019. Summarizing semantic graphs: A survey. VLDB J. 28, 3 (2019)

[9] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. Labra Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A. C. Ngonga Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, and A. Zimmermann. 2020. Knowledge graphs. CoRR arXiv:2003.02320 (2020)

[10] Consens, M.P., Miller, R.J., Rizzolo, F., Vaisman, A.A.: Exploring XML web collections with DescribeX. TWEB 4(3), 11:1-11:46 (2010)

[11] Tiago Macedo and Fred Oliveira. 2011. Redis Cookbook: Practical Techniques for Fast Data Manipulation. " O'Reilly Media, Inc."

[12] Zou, Xiaohan. "A survey on application of knowledge graph." Journal of Physics: Conference Series. Vol. 1487. No. 1. IOP Publishing, 2020.

[13] Ilievski, Filip, et al. "KGTK: a toolkit for large knowledge graph manipulation and analysis." The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19. Springer International Publishing, 2020.

[14] Udrea, O., Pugliese, A., Subrahmanian, V.S.: GRIN: a graph based RDF index. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada, pp. 1465-1470 (2007)

[15] Shao, C., Hu, L., Li, J., Wang, Z., Chung, T., Xia, J., Rimom-im, 2016. A novel iterative framework for instance matching [J]. J. Comput. Sci. Technol. 31 (1), 185–197.

[16] S. Kang, K. Lee and K. Shin, "Personalized Graph Summarization: Formulation, Scalable Algorithms, and Applications," in 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 2022 pp. 2319-2332.

[17] Zhu, Q., Wei, H., Sisman, B., Zheng, D., Faloutsos, C., Dong, X.L., Han, J., 2020. Collective multi-type entity alignment between knowledge graphs [C]. In: The World Wide Web Conference. WWW), Association for Computing Machinery, New York, NY, USA, pp. 2241–2252.

[18] Sun, Z., Wang, C., Hu, W., Chen, M., Dai, J., Zhang, W., Qu, Y., 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation [C]. In:Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI). AAAI Press.

[19] Koutra, D., Kang, U., Vreeken, J., Faloutsos, C.: Summarizing and understanding large graphs. Stat. Anal. Data Min. 8(3), 183–202 (2015)

[20] Hasibi, Faegheh, et al. "DBpedia-entity v2: a test collection for entity search." Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017.

[21] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

[22] Saxena, Apoorv, Aditay Tripathi, and Partha Talukdar. "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings." Proceedings of the 58th annual meeting of the association for computational linguistics. 2020.