Link Prediction between Group Entities in Knowledge Graphs (Student Abstract)

Jialin Su,[1] **Yuanzhuo Wang,**[1] **Xiaolong Jin,**[1] **Yantao Jia,**[2] **Xueqi Cheng**[1]
[1]CAS Key Lab of Network Data Science & Technology, Institute of Computing Technology, CAS, Beijing, 100190, China
[2]Huawei Technologies Co., Ltd, Beijing, China
{sujialin17g, wangyuanzhuo, jinxiaolong, cxq}@ict.ac.cn, jamaths.h@163.com

Abstract

Link prediction in knowledge graphs (KGs) aims at predicting potential links between entities in KGs. Existing knowledge graph embedding (KGE) based methods represent individual entities and links in KGs as vectors in low-dimension space. However, these methods focus mainly on the link prediction of individual entities, yet neglect that between group entities, which exist widely in real-world KGs. In this paper, we propose a KGE based method, called GTransA, for link prediction between group entities in a heterogeneous network by integrating individual entity links into group entity links during prediction. Experiments show that GTransA decreases mean rank by 5.4%, compared to TransA.

Introduction

Link prediction over a knowledge graph (KG) aims at predicting missing or possible future links in the KG. Namely, link prediction finds the missing head entity h or tail entity t of a triple (h,r,t) in the KG. Existing KGE based link prediction methods, such as TransE(Bordes et al. 2013) and TransA(Jia et al. 2016), represent entities and links in a heterogeneous network into a low-dimension vector space. These methods learn a score function according to the distance between different entities, and generate a list of candidate entities in order to find missing entities during prediction. In practical tasks, entities in the network can be classified into two categories, individual and group ones. Individual entities, such as researchers, are single nodes in the network which are not composed of any other nodes. Group entities, such as institutions, are nodes that consist of individual nodes in the network. Classical KGE methods take only the links between individual entities (i.e., individual links) into consideration. For instance, in scientific networks, “Apple” and “Google” are group entities, while their members, such as “Timothy Cook” and “Sundar Pichai”, are individual entities. Since group entities can be represented by a set of individual entities, links between group entities (i.e., group links) can be seen as a collection of links between their respective components. On the other hand, group links cannot be simply represented by the link between two specific individual entities. For example, in order to represent the group link between “Apple” and “Google”, we need to take into account the complete set of individual links between their members, instead of using the link between “Timothy Cook” and “Sundar Pichai” alone to represent that between “Apple” and “Google”. Namely, the prediction of group links are much more complicated than that of individual links, which existing KGE based methods failed to take into account.

In this paper, we propose a KGE based method for link prediction between group entities in heterogeneous networks, called GTransA, to address the above issues. GTransA considers hierarchies between group entities and individual entities, as well as within group entities, and represents group links with individual links by adding a weight function into the score function of existing KGE based methods. Experiments on a real-world dataset show that the mean rank of GTransA decreased by 5.4% as compared to TransA.

The Link Prediction Method GTransA

The idea of GTransA is to predict group links by representing group links using individual links. Therefore, a weight function $w(r)$, which assigns greater weights to group entities with more group links, limits the impact of the number of components in group entities and maps individual links to group links, needs to be defined. In GTransA, we have an entity set E and a relation set R and represent the KG by a set of triples $S = (h, r, t)$, where entities $h, t \in E$ and their relation $r \in R$, which are represented by embedding vectors during prediction. GTransA divides E into two subsets, individual entity set E_I and group entity set E_G. For a pair of given group entities U, V ($\forall U, V \in E_G$), we define the weight function $w(r)$ as:

$$w(r) = \begin{cases} (\delta\alpha_{U,V} + (1-\delta)\beta_{U,V})r & r = r^*, |T_{U,V}| > 0 \\ 0 & r = r^*, |T_{U,V}| = 0 \\ r & r \neq r^* \end{cases} \tag{1}$$

$$\alpha_{U,V} = \frac{log|T_{U,V}|}{log|E_U| + log|E_V|} \tag{2}$$

$$\beta_{U,V} = \frac{log|T_U| + log|T_V|}{log|E_U| + log|E_V|} \tag{3}$$

where $|\cdot|$ represents the number of elements in the set. r^* is the embedding vector of the relation to be predicted. We define Δ as the positive triple set (triples that actually exist in the KG). $E_U = \{h|(h,r,U) \in \Delta, h \in E_I\}$ and $E_V = \{t|(t,r,V) \in \Delta, t \in E_I\}$ represent the set of individual entities constituting group entities U and V, respectively. $T_{U,V} = \{(h,r^*,t)|(h,r^*,t) \in \Delta, h \in E_U, t \in E_V\}$ represents triples which take individual entities constituting group entity U as head entities and those constituting V as tail entities. $T_U = \{(e,r^*,h)|(e,r^*,h) \in \Delta, \exists(h,r^*,t) \in T_{U,V}, h,e \in E_U\}$ and $T_V = \{(e,r^*,t)|(e,r^*,t) \in \Delta, \exists(h,r^*,t) \in T_{U,V}, t,e \in E_V\}$ represent triples with head and tail entities both from group entity U and V, respectively. $\delta \in [0,1]$ is a manually set coefficient, determined by the importance of different types of relations. Specially, if $|T_{U,V}| = 0$, i.e. there is no link between individual entities from group entities U and V, we set $w(r)$=0. For each triple (h,r,t) in the KG, the score function of GTransA is defined as:

$$f_r(h,t) = ||h + w(r) - t||_{L1/L2-norm} \tag{4}$$

Similar to TransA, GTransA constructs the following loss function, allowing positive triples to get higher scores while negative triples get lower ones:

$$L = \sum_{(h',r,t')\in\Delta'} \sum_{(h,r,t)\in\Delta} \max(0, (f_r(h,t) - f_r(h',t') + M_{opt})) \tag{5}$$

where Δ' is the negative triple set (triples that are generated by randomly replacing the head or tail entity of positive triples). M_{opt} is the optimal margin, which is defined as:

$$M_{opt} = \mu M_{ent} + (1-\mu)M_{rel}, 0 \le \mu \le 1 \tag{6}$$

$$M_{ent} = \frac{\sum\limits_{r\in R_h} \min\limits_{t,t'} \sigma(||h-t'|| - ||h-t||)}{nr_h} \tag{7}$$

$$M_{rel} = \min_{r_i \in\ R_{h,r}} ||r_i - r|| \tag{8}$$

where M_{ent} is the optimal entity-specific margin and M_{rel} is the optimal relation-specific margin. For a given head entity h and related relation r in the KG, denote P_r and N_r as the positive and negative entity set of relation r, respectively. $t \in P_r = \{t|(h,r,t) \in \Delta\}$, $t' \in N_r = \{t|(h,r,t) \notin \Delta, (h,r',t) \in \Delta, \exists r' \in R\}$, where Δ is the set of positive entities and R is the set of relations in the KG. To be specific, P_r contain entities that are connected to head entity h by relation r, while N_r contain those which are connected to h with a relation other than r. Moreover, nr_h is the number of different types of relations with h as one end, and R_h is the set of relations related to h. $\sigma(x)$ returns the absolute value of x. $R_{h,r} = \{r_1, r_2, \ldots, r_{nh_r-1}\}$ is the set of relations that entity h has except r, when $N_r \neq \emptyset$. In particular, when $N_r = \emptyset$, we set $M_{ent} = 0$.

Experiments

The experiments are carried out on the network we generate from DBLP[1] and Aminer[2], two academic datasets. The network contains 7 types of relations and 4 types of entities, i.e., researcher, paper, institution and venue. The experiments aim at predicting future collaboration links ($r^* = collaboration$) between institutions, group entities composed of researchers. In the experiments, we set the embedding dimensionality as 100, learning rate as 0.01, μ=0.5, δ=0.9. We train the models with a batch size of 100 and a maximum of 1000 epochs. The baseline methods include classical KGE methods and traditional methods shown in Table 1. Here, we employ mean rank and hits@k(k = 10) (the proportion of correctly predicted entities ranked in the top k predictions) of correctly predicted entities. Higher hits@k and lower mean rank indicate better performance. The experimental results are shown in Table 1.

Table 1: Experimental results of link prediction.

Method	Mean Rank		Hits@10	
	Raw	Filtered	Raw	Filtered
Common Neighbors	6230.57	6012.76	0.104	0.193
Resource Allocation	6368.07	6012.76	0.104	0.193
TransE	4467.69	4417.43	0.153	0.203
TransH	4332.97	4283.95	0.175	0.244
TransA	4400.63	4340.75	0.213	0.234
GTransA	4162.65	4103.78	0.222	0.245

The results in Table 1 show that KGE methods outperform traditional methods, which indicates the effectiveness of KGE methods, since KGE methods can address the problem of feature sparsity caused by traditional methods. Besides, GTransA decreases the mean rank of TransA by 5.4%, since GTransA deals with individual links and group links differently by measuring group links based on individual links using a weight function. Thus, the group link contains the global information about its components, which improves the performance of prediction.

Conclusion

In this paper, we proposed GTransA for predicting links between group entities in KGs, which employs a weight function to integrate individual entity links into group entity links during prediction. The experiments demonstrate the effectiveness of GTransA.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1402600, 2017YFC0820404), the National Natural Science Foundation of China (No. U1836206, 61572469, 61772501, 91646120) and the GF Innovative Research Program.

References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, 2787–2795.

Jia, Y.; Wang, Y.; Lin, H.; Jin, X.; and Cheng, X. 2016. Locally adaptive translation for knowledge graph embedding. In *Thirtieth AAAI conference on artificial intelligence*.

[1] https://dblp.uni-trier.de/

[2] https://www.aminer.cn/