Embedding and Network Representation Learning in AceKG

Project Report For EE447: MOBILE NETWORKS, 2017-2018 Spring

Luhua Jin(with teammate Yuchen Yan)
515030910585(515030910564)

Supervisor: Prof. Xinbing Wang and Prof. Luoyi Fu
School of Electronic, Information and Electrical Engineering
Shanghai Jiao Tong University

1 Introduction

This project is based on Academic Knowledge Graph (AceKG), an academic semantic network, which describes 3.13 billion triples of academic facts based on a consistent ontology, including commonly used properties of papers, authors, fields of studies, venues, institutes and relations among them.
Compared with other existing open academic KGs or datasets, AceKG has the following advantages:
(1) AceKG offers a heterogeneous academic information network, i.e., with multiple entity categories and relationship types, which supports researchers or engineers to conduct various academic data mining experiments.
(2) AceKG is sufficiently large (3.13 billion triples with nearly 100G disk size) to cover most instances in the academic ontology, which makes the experiments based on AceKG more convincing and of practical value.
(3) AceKG provides the entity mapping to computer science databases including ACM, IEEE and DBLP, which helps researchers integrate data from multiple databases together to mine knowledge.
(4) AceKG is fully organized in structured triplets, which is machine readable and easy to process.

The main purpose of this project to further evaluate different state-of-the-art knowledge embedding and network representation learning approaches to test the reliability of AceKG. In the end, some promising methods to improve the performance of AceKG.

2 Basic Concepts of AceKG

2.1 Ontology

AceKG defines 5 classes of academic entities: *Papers*, *Authors*, *Field of studies*, *Venues* and *Institutes*. Between two entities there is a relation. And the facts including the frequently used properties of each entities and the relations between the entities are described as triplets in the knowledge graph. Figure 1 briefly demonstrates the ontology of AceKG.

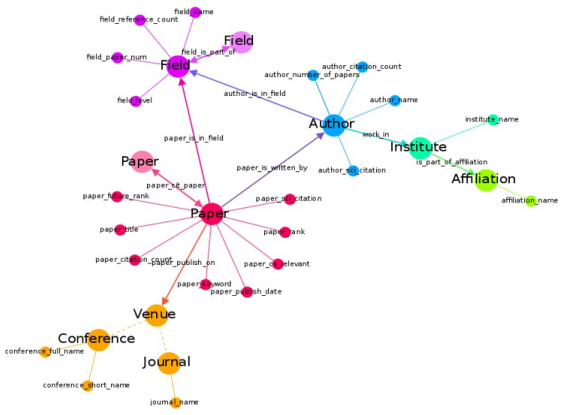

Figure 1: AceKG Ontology

The statistics of AceKG are shown in Table 1. All the facts are represented as subject-predicate-object triplets (SPO triplets).

Table 1: Statistics of XKG

Class	Number	Class	Number
Paper	61,704,089	Institute	19,843
Author	52,498,428	Field	50,233
Journal	21,744	Conference	1,278
Total Entities	114,295,615	Total Relations	3,127,145,831

2.2 Entity Alignment

A large part of papers is mapped in computer science of AceKG to the papers stored in IEEE, ACM and DBLP databases. All the latest papers in these three databases have been aligned with AceKG. Some mapping statistics are shown in Table 2.

Table 2: Statistics of node mapping

Database	IEEE	ACM	DBLP
Mapping number	2,332,358	1,912,535	2,274,773

2.3 Inference

In AceKG, there are some inference rules which have been designed. With these inference rules, we can define the new relations on AceKG, which provides more comprehensive ground truth.

3 Knowledge Embedding

In this section, we will evaluate several state-of-the-art approaches for knowledge embedding using AceKG.

3.1 Backgound

The target of knowledge embedding is to project triplets(h, r, t) in a given knowledge base to d-dimensional vectors, where $h, t \in E$(set of entities) and $r \in R$(set of relations). We also defines a scoring function to evaluate the plausibility of the triplet (h, r, t) in the knowledge base. There are quite a few algorithms in knowledge embedding. Figure 2 gives a simple illustration of two commonly used algorithms, transE[1] and transH[2]: TransE aims at connecting h, t by translation vector and transH modifies TransE by embedding triplets on hyperplanes.

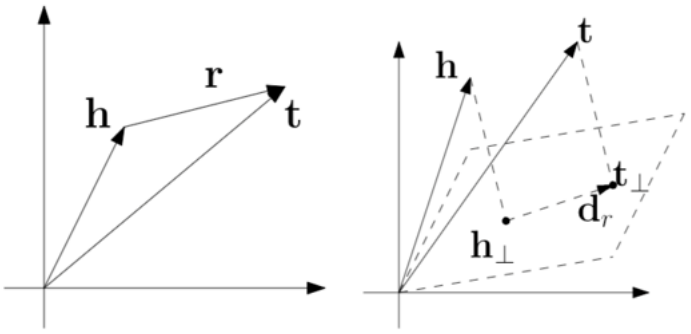

Figure 2: transE and transH

3.2 Experiment Design

We study and evaluate related methods on link prediction proposed by Bordes et al. [1]: given one of the entities and the relation in a latent triplet, it aims to predict the other missed entity. In stead of using benchmark datasets FB15K[3] and WN18[4], we construct XK18K, a new benchmark dataset extracted from AceKG for knowledge embedding. Table 3 shows the statistics of the WN18, FB15K and XK18K. XK18K is sparser than FB15K but denser than WN18, and it provides only 7 types of relations.We compare the following algorithms in our experiments: TransE[1], TransH[2],DistMult[5], ComplEx[6], HolE [7].

Table 3: Datasets used in knowledge embedding.

Dataset	#R	#E	#Trip. (Train/ Valid/ Test)		
WN18	18	40,943	141,442	5,000	5,000
FB15K	1345	14,951	483,142	50,000	59,071
XK18K	7	18,464	130,265	7,429	7,336

3.3 Experiment Results

The results of the experiment is shown in table 4. We can divide the algorithms we use into two classes:(1) traditional algorithms(transE, transH) and (2)compositional algorithms(DistMult, ComlEx, HolE). TransE outperforms all counterparts on hit@10 as 89.2%. Although 94.4% of relations in the knowledge base are many-to-many, TransE shows its advantages on modeling sparse and simple knowledge base. On the other hand, ComlEx performs quite well when it comes to hit@1(83.8%) and hit@3(87.1%). We hypothesize that it confirms their advantages on modeling antisymmetric relations because all of our relations are antisymmetric.

Table 4: Results of link prediction task on XK18K

	MRR		Hits at		
Model	Raw	Filter	1	3	10
TransE	0.358	0.719	62.7	82.5	**89.2**
TransH	0.315	0.701	61.0	77.2	84.6
DistMult	0.432	0.749	68.7	79.5	86.1
HolE	**0.482**	**0.864**	**83.8**	**87.1**	88.2
ComplEx	0.440	0.817	75.4	85.8	89.0

4 Network Representation Learning

In this section, we will evaluate several state-of-the-art approaches for network representation learning (NRL) on AceKG.

4.1 Background

Network embedding assigns nodes in a network to low-dimensional representations and effectively preserves the network structure and the content of the nodes. Figure 3 is a brief illustration of the 2-dimensional representations of conferences after Network embedding. We evaluate related algorithms including DeepWalk[8], PTE[9], LINE[10] and metapath2vec[11] on two tasks: scholar classification and scholar clustering.

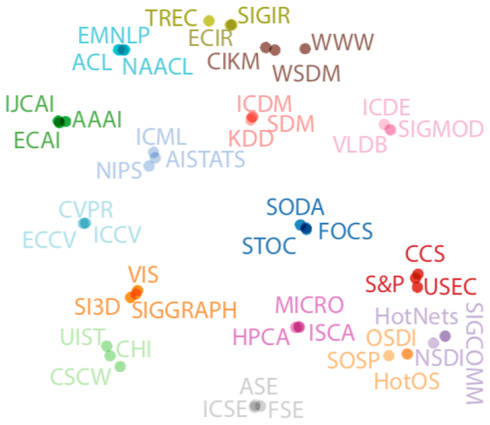

Figure 3: Network Embedding

4.2 Experiment Design

We select 5 fields of studies (FOS) from AceKG(Biology, CS, Economics, Medicine and, Physics) and 5 main subfields of each field above. Then we extract all scholars, papers and venues in those fields of studies respectively to construct 5 heterogeneous collaboration networks.
We also construct 2 larger academic knowledge bases: 1) We integrate 5 networks above into one graph which contains all the information of 5 fields of studies; 2)Select art of related papers and scholars in Google Scholar to construct one large heterogeneous collaboration networks.
The statistics of these networks are shown in Table 5.

Table 5: Datasets used in network representation learning.

Dataset	#Paper	#Author	#Venue	#Edge
FOS_Biology	1211664	2169820	13511	5544376
FOS_CS	452970	738253	10726	1658917
FOS_Economics	412621	597121	8269	1163700
FOS_Medicine	182002	491447	7251	819312
FOS_Physics	449844	596117	5465	1602723
FOS_5Fields	2578185	3868419	18533	10160137
Google	600391	635585	151	2373109

4.3 Experiment Results

(1) **Classification**

Logistic regression is applied for classification after network embedding. Table 6 shows the classification results evaluated by Micro-f1 and Macro-f1. Metapath2vec performs relatively better than other methods. The modified heterogeneous sampling and skip-gram algorithm may be the reasons. We also notice that DeepWalk and LINE also perform well, showing their scalability on heterogeneous networks. The reason may be that the kinds of relations are limited so that homogeneous algorithm can also learn a comprehensive network representation.

Another interesting result is that there is significant performance gap between FOS-labeled datasets and Google-labeled datasets. Since there are more cross-field papers and scholars in FOS-labeled datasets than in Google-labeled ones, it adds up to difficulty of classification. For instance, a professor is in CS but he publishes more papers in biology with AI algoritms than in CS, then our classification may be wrong because we may mistake him as a biology scholar.

What's more, cross-field papers and scholars also lead to differences of performance among different fields with the same algorithm. For example, the highest Micro-F1 shows that the sub-fields of Biology are the most independent, while the lowest Micro-F1 means that the sub-fields of CS cross mostly.

Table 6: Results of scholars classification

Metric	Method	FOS_BI	FOS_CS	FOS_EC	FOS_ME	FOS_PH	FOS_5F	Google
Micro-F1	DeepWalk	0.792	0.545	0.692	0.663	0.774	0.731	0.948
	LINE(1st+2nd)	0.722	0.633	0.717	0.701	0.779	0.755	0.955
	PTE	0.759	0.574	0.654	0.694	0.723	0.664	0.966
	metapath2vec	0.828	0.678	0.753	0.770	0.794	0.831	0.971
Macro-F1	DeepWalk	0.547	0.454	0.277	0.496	0.592	0.589	0.942
	LINE(1st+2nd)	0.445	0.542	0.385	0.577	0.640	0.655	0.949
	PTE	0.495	0.454	0.276	0.555	0.571	0.528	0.961
	metapath2vec	0.637	0.570	0.485	0.659	0.635	0.682	0.968

(2) **Clustering**

K-means algorithm is applied for clustering after network embedding. Table 7 shows the clustering results evaluated by normalized mutual information (NMI). Interestingly, we can draw similar conclusions with the case of classification. Metapath2vec performs much better than the other algorithms and the significant performance gap between FOS-labeled datasets and Google-labeled datasets still exists. The reason may also be similar: the modified heterogeneous sampling and skip-gram algorithm gives advantage to metapath2vec.

Table 7: Results of scholar clustering

Model	FOS-labeled	Google-labeled
DeepWalk	0.277	0.394
PTE	0.153	0.602
LINE(1st+2nd)	0.305	0.459
metapath2vec	0.427	0.836

5 Promising Methods

5.1 GAN Classification

In the paper of Salimans T. et al[12], an algorithm of node classification with ”GAN” based on network embedding is proposed. Figure 4 demonstrates the framework of this algorithm. The core theory of the algorithm is to use GAN to generate a new class of data and transform K-classification to K+1-classification. This algorithm combines supervised learning and unsupervised learning. In this case, the system can still learn the knowledge presentation even if some labels of data are missed.

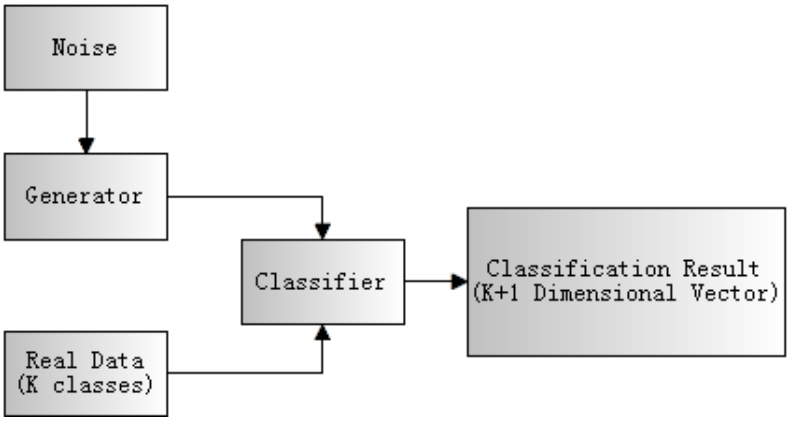

Figure 4: ”GAN” Classification

5.2 Struc2vec[13]

Structural identity is a concept of symmetry in which network nodes are identified according to the network structure and their relationship to other nodes. struc2vec can be leveraged for this task when labels for nodes are more related to their structural identity than to the labels of their neighbors. In every field of study, there are scholars and papers with different ”levels”. Top scholars are likely to have similar structural identity in the knowledge graph even if their distance is quite large. With struc2vec, we can classify and cluster entities with different ”levels” in AceKG and then we can construct a recommendation system with differen ”levels” by using the result.

6 References

[1]Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In NIPS, 2013.

[2]Bishan Yang, Wen tau Yih, Xiaodong He, and Li Deng Jianfeng Gao. Embedding entities and relations for learning and inference in knowledge bases. In ICLR, 2015.

[3]Google. Freebase data dumps. https://developers.google.com/freebase/data.

[4]George A. Miller. Wordnet: A lexical database for english. Commun. ACM, 38(11):39-41, November 1995. ISSN 0001-0782.

[5]Bishan Yang, Wen tau Yih, Xiaodong He, and Li Deng Jianfeng Gao. Embedding entities and relations for learning and inference in knowledge bases. In ICLR, 2015.

[6]Théo Trouillon, Christopher R. Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. Knowledge graph completion via complex tensor factorization. J. Mach. Learn. Res., 18(1), January 2017.

[7]Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In AAAI, 2016.

[8]Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In KDD. ACM, 2014.

[9]Jian Tang, Meng Qu, and Qiaozhu Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In KDD. ACM, 2015.

[10]Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In WWW, 2015.

[11]Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In KDD, 2017.

[12] Salimans T, Goodfellow I, Zaremba W, et al. Improved Techniques for Training GANs[C]. In NIPS, 2016 .

[13]Ribeiro, Leonardo F. R, P. H. P. Saverese, and D. R. Figueiredo. struc2vec : Learning Node Representations from Structural Identity. 2017:385-394.