XKG: A Large-scale Knowledge Graph for Academic Data Mining

Yuchen Yan, 515030910564

ABSTRACT

Most existing knowledge graphs (KGs) in academic domains suffer from problems of insufficient multi-relational information, name ambiguity and improper data format for large-scale machine processing. In this paper, we present XKG[1], a new large-scale KG in academic domain. XKG not only provides clean academic information, but also offers a large-scale benchmark dataset for researchers to conduct challenging data mining projects including link prediction, community detection and scholar classification. Specifically, XKG describes 3.13 billion triples of academic facts based on a consistent ontology, including necessary properties of papers, authors, field of studies, venues and institutes, as well as the relations among them. To enrich the proposed knowledge graph, we also perform entity alignment with existing databases and rule-based inference. Based on XKG, we conduct experiments of three typical academic data mining tasks and evaluate several state-of-the-art knowledge embedding and network representation learning approaches on the benchmark datasets built from XKG. Finally, we discuss propose several promising research directions that benefit from XKG.

KEYWORDS

Knowledge Graphs, Academic Data Mining, Benchmarking

1 INTRODUCTION

Knowledge graphs have become very crucial resources to support many AI related applications, such as graph analytics, Q&A system, web search, etc. A knowledge graph, which describes and stores facts as triplets, is a multi-relational graph consisting of entities as nodes and relations as different types of edges. Nowadays, many companies and research teams are trying to organize the knowledge in their domain into a machine-readable knowledge graph, e.g., YAGO [5], NELL [8], DBpedia [6], and DeepDive [2]. Although these large-scale knowledge graphs have collected tremendous amount of factual information about the world, many fields still remain to be covered.

With information of papers, scholars, institutes, venues, field of studies and other useful entities, data mining on academic networks aims to discover hidden relations and to find semantic-based information. Several academic databases or knowledge graphs have been built with structured academic data [11, 12, 15]. The public academic knowledge graphs can provide scholars with convincing academic information, and offer large-scale benchmark datasets for researchers to conduct data mining projects.

However, there are some limitations in existing databases or knowledge graphs. First, most of existing works provide homogeneous academic graphs, while relations among different types of entities remaining lost [11, 15]. Second, some databases only concentrate on one specific field of study, limiting the projects which aim at discovering cross-field knowledge [11]. Third, synonymy and ambiguity are also the restrictions for knowledge mining [12]. Allocating the unique IDs to the entities is the necessary solution, but some databases use the names of the entities as their IDs directly.

In this paper, we propose Academic Knowledge Graph (XKG), 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, field of studies, venues, institutes and relations among them. Apart from the knowledge graph itself, we also perform entity alignment with the existing KGs or datasets and some rule-based inference to further extend it and make it linked with other KGs in the linked open data cloud. Based on XKG, we further evaluate several state-of-the-art knowledge embedding and network representation learning approaches in Sections 3 and 4. Finally we discuss several potential research directions that benefit from XKG in Section 5 and conclude in Section 6.

Compared with other existing open academic KGs or datasets, XKG has the following advantages.

(1) XKG 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) XKG 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 XKG more convincing and of practical value.
(3) XKG 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) XKG is fully organized in structured triplets, which is machine-readable and easy to process.

2 THE KNOWLEDGE GRAPH

The XKG dataset can be freely accessed online. XKG is a large academic knowledge graph with 3.13 billion triples. It covers almost the whole academic area and offers a heterogeneous academic network.

2.1 Ontology

All objects (e.g., papers, institutes, authors) are represented as entities in the XKG. Two entities can stand in a relation. Commonly used attributes of each entities including numbers, dates, strings and other literals are represented as well. Similar entities are grouped into classes. In total, XKG defines 5 classes of academic entities: $Papers$, $Authors$, $Fieldofstudies$, $Venues$ and $Institutes$. 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. The ontology of XKG is shown in Figure 1.

To deal with synonymy and ambiguity, each entity in defined classes are allocated with a URI. For example, `XKG:7E7A3A69` and `ace:7E0D6766` are two scholars having the same name: Jiawei Han, one of whom is the influential data mining scientist. Compared with

[1] XKG name is anonymized for double-blind review.

CIKM'18, 22-26 October 2018, Lingotto, Turin, Italy
2018. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

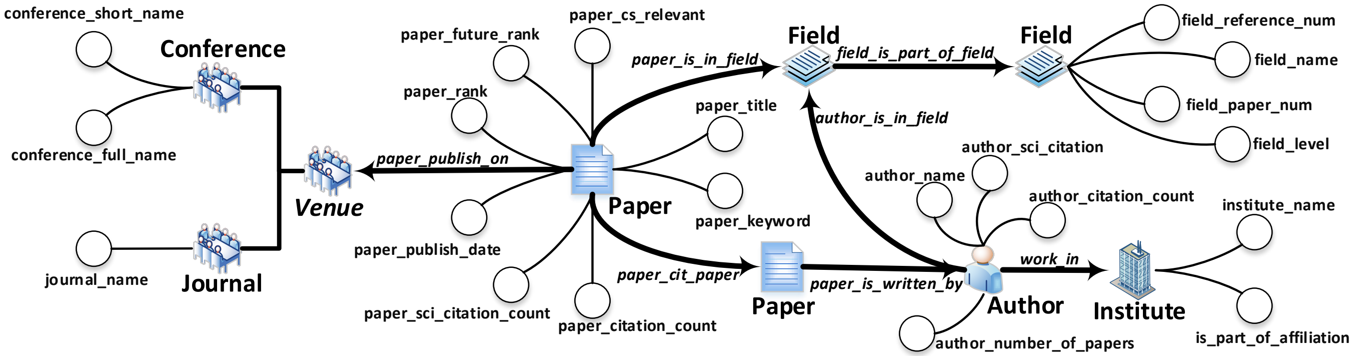

Figure 1: An overview of XKG Ontology

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

Table 2: Statistics of node mapping

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

the datasets which uses entity names to represent entities directly, XKG can avoid mistakes caused by synonymy and ambiguity,

The statistics of XKG are shown in Table 1. All the facts are represented as *subject-predicate-object* triplets (SPO triplets). And we release the Turtle format XKG online. It can be queried by Apache Jena framework[2] with SPARQL easily.

2.2 Entity alignment

In order to make XKG more connected and comprehensive, we map a large part of papers in computer science of XKG to the papers stored in IEEE, ACM and DBLP databases. All the latest papers in those three databases have been aligned with XKG . Some mapping statistics are shown in Table 2. The knowledge graph is updated with the latest academic information periodically.

2.3 Inference

Rule-based inference on knowledge graph is a typical but critical way to enrich the knowledge graph. The selected inference rules that we design are shown in Figure 2. With those inference rules, we can define the new relations on XKG, 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 our knowledge base XKG.

3.1 Task Definition

Given a set S of triplets (h, r, t) composed of two entities $h, t \in E$ (the set of entities) and a relation $r \in R$ (the set of relationships), knowledge graph embedding maps each entity to a k-dimensional vector in the embedding space, and defines a scoring function to evaluate the plausibility of the triplet (h, r, t) in the knowledge base. 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. The commonly used benchmark datasets are FB15K and WN18, which are extracted from Freebase[4] and Wordnet[7]. We construct a new benchmark dataset (denoted as XK18K in the rest of this section) extracted from XKG for knowledge embedding. We will show how it differs from FB15K and WN18 in section 3.2. We compare the following algorithms in our experiments: **TransE** [1], **TransH** [17], **DistMult** [18], **ComplEx** [16], **HolE** [9].

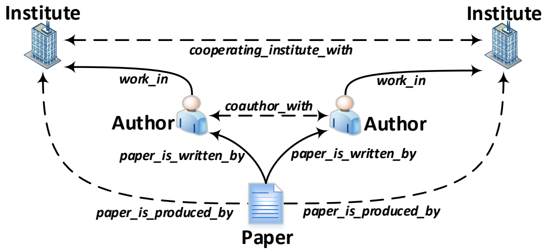

Figure 2: Example of rule-based inference. The dotted arrows are inferred predicates.

3.2 Experimental setup

To extract XK18K from XKG, we firstly select 68 critical international venues and influential papers published on them. Then we add the triplets of authors, fields and institutes. Finally, the Train/ Valid/ Test datasets are divided randomly. Table 3 shows the statistics of the WN18, FB15K and XK18K. XK18K is sparser than FB15K but denser than WN18 (indicated by the value of $\#Trip/\#E$), and it provides only 7 types of relations. We will evaluate the models' scalability on the knowledge base which has simple relation structure but tremendous amount of entities. The code we used is based on the OpenKE, an open-source framework for knowledge embedding.

3.3 Evaluation Results

We show the link prediction results based on knowledge embedding in Table 4. The reported results are produced with the best set of

[2] https://jena.apache.org

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

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

Table note: Filtered and Raw Mean Reciprocal Rank (MRR) and Hits@{1,3,10} for the models tested on the AK18K dataset. Hits@{1,3,10} metrics are filtered. Filtered metrics means removing from the test list the other triplets that appear in the dataset while evaluation.

hyper-parameters after the grid searches reported in the papers. The compared state-of-the-art models can be divided into two categories: i: translational models (TransE, TransH); ii: compositional models (DistMult, HolE, ComplEx). TransE outperforms all counterparts on hit@10 as 89.2%. Although 94.4% of relations in our knowledge base are many-to-many, TransE shows its advantages on modeling sparse and simple knowledge base, while TransH not achieving better results. However, HolE and ComplEx achieve the most significant performance on the other metrics, especially on hit@1(83.8%/75.4%) and on filtered MRR (0.482/0.440). We hypothesize that it confirms their advantages on modeling antisymmetric relations because all of our relations are antisymmetric, such as $field\_is\_part\_of$ and $paper\_is\_written\_by$.

Compared with the experiment results on FB15K and WN18 reported in [9], performances evaluated using XK18K is noticeably different. First, results on XK18K are lower than those on WN18 but higher than those on FB15K. It is caused by the limited relation types and large amount of potential entities per relation. Some relation such as $paper\_is\_in\_field$ can have thousands of possible objects per triplet, limiting the prediction performance. Second, the performance gap between two model categories grows more pronounced as the knowledge graph become more complicated, which indicates the translational models with simple assumptions can not model the complicated graph well.

4 NETWORK REPRESENTATION LEARNING

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

4.1 Task Definition

Given a network $G = (V, E, A)$, where V denotes the vertex set, E denotes the network topology structure and A preserves node attributions, the task of NRL is to learn a mapping function $f : v \mapsto r_v \in R_d$, where r_v is the learned representation of vertex v and d is the dimension of v_r. We study and evaluate related methods including **DeepWalk** [10], **PTE** [13], **LINE** [14] and **metapath2vec** [3] on two tasks: scholar classification and scholar clustering.

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

4.2 Experimental setup

Based on XKG, we firstly select 5 field of studies (FOS) [3] and 5 main subfields of each. Then we extract all scholars, papers and venues in those field of studies respectively to construct 5 heterogeneous collaboration networks. We also construct 2 larger academic knowledge base: 1) We integrate 5 networks above into one graph which contains all the information of 5 field of studies; 2) We match the eight categories of venues in Google Scholar[4] to those in XKG . 151 of 160 venues (8 categories × 20 per category) are successfully matched. Then we select all the related papers and scholars to construct one large heterogeneous collaboration networks. The statistics of these networks are shown in Table 5. Moreover, the category of scholars are labeled with the following approach:

(1) To label the papers, we adopt the field of study information and Google scholar category directly as the label of papers in 6 field of study networks and 1 Google scholar network respectively.
(2) As for the label of the scholars, it is determined by the majority of his/her publications' labels. When some labels have equal quantity of papers, they are chosen randomly.

4.3 Evaluation Results

4.3.1 Classification. We adopt logistic regression to conduct scholar classification tasks. Note that in this task 5-fold cross validation are adopted. Table 6 shows the classification results evaluated by micro-f1 and macro-f1. metapath2vec learns heterogeneous node embeddings significantly better than other methods. We attribute it to the modified heterogeneous sampling and skip-gram algorithm. However, DeepWalk and LINE also achieve comparable performance, showing their scalability on heterogeneous networks. Another reason for the comparable performance is that our edge types and node types are limited, homogeneous algorithm can also learn a comprehensive network representation.

It should be noted that there is significant performance gap between FOS-labeled datasets and Google-labeled dataset. We hypothesize that it is because of the different distribution of papers and scholars. Papers collected in the Google-labeled dataset are published on Top-venues and consequently few scholar could be active in multiple categories, while there are more cross-field papers and scholars in FOS-labeled datasets.

Moreover, The performance indicates the level of interdiscipline in these fields. 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. Finally, the dramatical decline from micro-f1 to macro-f1, especially in Economy, indicates the imbalance of sub-fields in some field of studies.

[3] biology, computer science, economics, medicine and physics

[4] https://scholar.google.com/citations?view op=top venues&hl=en&vq=eng

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

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

4.3.2 Clustering. Based on the same node representation in scholar classification task, we further conduct scholar clustering experiment with k-means algorithm to evaluate the models' performance. All clustering experiments are conducted 10 times and the average performance is reported.

Table 7 shows the clustering results evaluated by normalized mutual information (NMI). Overall, metapath2vec outperform all the other models, illustrating the modified heterogeneous sampling and skip-gram algorithm can preserve the information of the knowledge graph better. Another interesting result is the performance gap between FOS-labeled dataset and Google-labeled dataset, which indicates the hypothesis we proposed in section 4.3.1.

5 FUTURE DIRECTIONS

There are other research topics which can leverage XKG. In this section, we propose three potential directions in this section.

Cooperation prediction. To predict a researcher's future cooperation behavior is an interesting topic in academic mining, and many current works have contributed to it by considering previous cooperation, neighborhood, citation relations and other side information. However, all these factors can be thought as obvious feature in an academic knowledge graph, which is incomplete and may always ignore some other features like the same institution or the same field. Given this situation, one may perform cooperation prediction based on the NRL results, which can represent the feature of a researcher better and may provide some help to cooperation prediction task.

Author disambiguation. Author disambiguation is a traditional problem in social network, which means distinguishing two people with the same name in a network. With the help of XKG, author disambiguation can be conducted conveniently. The structure and node information in XKG can enhance the author disambiguation performance. Then, some author disambiguation algorithms with good performance can be applied to XKG. The author disambiguation problem can be solved and the quality of XKG will be improved in such an iterative way.

Finding rising star. Finding academic rising star is important in academic mining in that it can provide helpful reference for universities and companies to hire young faculty or new scientist. Researchers have raised various algorithms for this based on publication increasing rate, mentoring relations and some other factors. In order to make the classification better, we will first embed the XKG to uncover the hidden structure features of rising star and then apply some clustering algorithms on the embedding results.

6 CONCLUSION

In this paper we propose XKG, a large-scale knowledge graph in academic domain, which consists of 3.13 billion triples of academic facts based on a consistent ontology, including commonly used properties of papers, authors, field of studies, venues, institutes and relations among them. Based on XKG, we design three experimental evaluations and further compare several state-of-the-art approaches using XKG. Besides, we propose several potential research topics that can also benefit from the dataset. We will keep maintaining and updating the coverage of XKG for wider usage in this direction.

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] Christopher De Sa, Alex Ratner, Christopher Rē, Jaeho Shin, et al. Deepdive: Declarative knowledge base construction. In *ACM SIGMOD Record*, 2016.

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

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

[5] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, et al. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

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

[8] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, et al. Never-ending learning. In *AAAI*, 2015.

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

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

[11] Amit Singhal. Computer science bibliograph, 1996. URL https://dblp.uni-trier.de.

[12] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, et al. An overview of microsoft academic service (mas) and applications. In *WWW '15 Companion*, 2015.

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

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

[15] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *SIGKDD*, 2008.

[16] 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.

[17] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.

[18] 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.