具有关系敏感嵌入的知识库错误检测

缪 琦,杨昕悦

(辽宁工程技术大学 电子与信息工程学院,辽宁 葫芦岛 125105)

摘 要:准确性与质量对于知识库而言尤为重要,尽管已经有很多关于知识库不完整性的研究,但是很少有工作者考虑到对于知识库存在的错误进行检测,按照传统方法通常无法有效捕捉知识库中错误事实内在相关性。本文提出了一种知识库具有关系敏感嵌入式方法 NSIL,以获取知识库各关系之间的相关性,从而检查出知识库中的错误,以此提高知识库的准确性与质量。该方法分为相关性处理和错误检测两阶段。在相关性处理阶段,使用 NSIL 的相关函数以分值形式获取各关系之间的相关度;在错误检测阶段,基于相关度分值进行错误检测,对于缺失主体或客体的三元组进行缺失成分预测。最后在知识库之一 Freebase 生成的基准数据集"FB15K"上进行了广泛验证,证明了该方法在知识库错误知识检测方面有着很高的性能。

关键词:知识库;嵌入模型;错误检测

中图分类号:TP183 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2020.10.005

引用格式:缪琦,杨昕悦. 具有关系敏感嵌入的知识库错误检测[J].信息技术与网络安全,2020,39(10):23-27,37.

Knowledge base error detection with relation sensitive embedding

Miao Qi, Yang Xinyue

(School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: Accuracy and quality are very important for the knowledge base. Although there have been many researches on the incompleteness of knowledge base, few workers consider the detection of errors in the knowledge base. According to the traditional methods, it is usually unable to effectively capture the internal correlation of errors in the knowledge base, so as to check the errors. In this paper, a relational sensitive embedded method NSIL for knowledge base is proposed to obtain the correlation among the relationships between them, so as to check out the errors in the knowledge base, so as to improve the accuracy and quality of the knowledge base. This method is divided into two stages: correlation processing and error detection. In the correlation processing stage, correlation function of NSIL is used to obtain the correlation degree of each relationship in the form of score; in the error detection stage, error detection is based on the score of correlation degree, and missing component prediction is carried out for the triplet of missing subject or object. At last, the method is verified on the benchmark data set "FB15K" which is generated by Freebase, one of the largest knowledge bases. It is proved that the method has high performance in knowledge base error detection.

Key words: knowledge base; embedding model; error detection

0 引言

如今,知识库已经成为各种研究和应用越来越重要的和常用的数据源,如语义搜索、实体链接、问答系统和自然语言处理等。为了使庞大数据库更易于操作,研究者提出了一种新的研究方向——知识库嵌入。关键思想是嵌入 KB(Knowledge Base)组件,包括将实体和关系转化为连续的向量空间,从而简化操作,同时保留 KB 原有的结构。实体和关系嵌入能进一步应用于各种任务中,如 KB 补全、关系提取、实体分类和实体解析。虽然庞大的知识库中有数以亿计的事实,但是在信息爆炸的时代远远不够。大部分的研究工作聚焦知识库对缺失边的扩充,很少有人考虑到其中过时的、不正确的信息[1-3]。许多扩充知识库研究将事实投射到 k 维向量空间,通过聚类来找到关系的相关性,很难实现高效有效处理。

1 关系敏感知识库错误检测

知识库错误的检测仍然是一个艰巨的挑战:(1)知识库的知识具有离散性,因此通过传统嵌入方法[4-6]难以在知识库中进行广泛推理和检测;(2)知识库中的关系几乎没有上下文可以捕获其语义的相关性,所以大部分著作都是对实体和实体进行研究,忽略了关系和关系之间的相关性;(3)对于纠错大部分是建立在实体-实体或者建立在字符成本上的。

为了解决上述挑战,提出嵌入一个新颖简单的关系敏感方法(NSIL),该函数由 RSEA[7]方法的思想启发产生,但是性能更高。该函数计算了主体与客体之间的相关性,能在大规模的知识库中准确地对三元组进行识别和错误检查,并且对纠正三元组的错误具有较高精准性。

图 1 中,对于关系“家人近期病史”,可以对应的不同主体是离散的,主体可以是“市民 A”、“市民 B”等,客体也是离散的,可以是“SARS-COV 病毒”、“COVID-19 病毒”等,彼此之间没有必然的关系。但是对于关系对而言,它们会产生局部相交,如同对于三元组(“市民 A”,“患病”,“COVID-19 病毒(Corona Virus Disease 2019)”)和(“市民 A”,“家人近期病史”,“COVID-19 病毒”)之间,它们不仅有相同的主体集合(“市民 A”)和客体集合(“COVID-19 病毒”),而且有与“家人近期病史”和“患病”的主体和客体都有内在直接关系的共同实体“市民 C”,即“市民 A”的“客体”,“COVID-19 病毒”的“主体”。于是认为关系“家人近期病史”和“患病”相关。从关系的主客体集合和与主客体关联的实体集出发,使得发现关系之间的相关性具有可能。就像上文中说的那样,如果在一定程度上认定“家人近期病史”≈“患病”,那么对于三元组(“市民 B”,“家人近期病史”,“SARS-COV 病毒(SARS-associated coronavirus)”),可以得出市民 B 最可能患病的事实是 SARS-COV 病毒。如果认定“市民 B 有家人近期病史是 SARS-COV 病毒”,如果给了需要判定的三元组是(“市民 B”,“患病”,“COVID-19 病毒”),那么会判断它是错的。如果它的客体需要纠正,它最有可能被纠正为“SARS-COV 病毒”。

本文提出了一种新的关系敏感函数 NSIL,该方法从关系的角度出发,能有效识别关系库中关系的相关性,该关系在多对多、一对一、一对多、多对一的实体中具有不错的识别效率。

2 知识库嵌入技术的发展

近来,提出了许多知识库嵌入技术来将离散知识图编码为连续向量空间。首先介绍一些常用的符号。知识库中的事实三元组(h,r,t),即(主体,关系,客体)。其对应关系的矢量表示表示为(h,r,t),

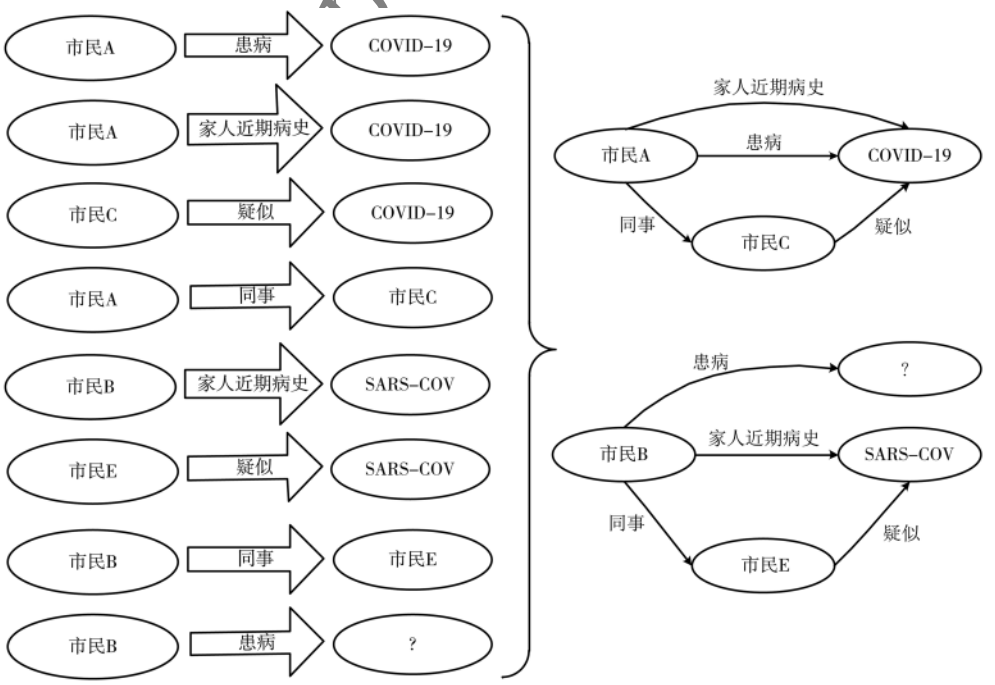


图 1

分数函数 $f(h,t)$ 对于属于知识库的正三元组将自动获得较高的分数,而对于负三元组[8]将自动生成较低的分数。

表 1 为在相同的 k 维嵌入空间 $\boldsymbol{R}^{k\times k}$ 中,得分函数、参数数量和时间复杂度的模型比较。

表 1 模型比较

模型	得分函数	参数数量	时间复杂度
TransE	$\|\|h+r-t\|\|_2^2$	kN_e+kN_r	N_t
TransH	$\|\|h_\perp+r-t_\perp\|\|_2^2$	kN_e+2kN_r	$2kN_t$
TransR	$\|\|h\mathrm{Mr}+r-t\mathrm{Mr}\|\|_2^2$	$kN_e-k(k+1)N_r$	$2k^2N_t$

(1)TransE[9]:TransE 是基于实体和关系的分布式向量表示,由 Bordes 等人于 2013 年提出,受 word2vec 启发,利用了词向量的平移不变现象。例如:$C(\text{king})\_C(\text{queen})\approx C(\text{man})\_C(\text{woman})$。其中,$C(w)$就是 word2vec 学习到的词向量表示。TransE 定义了一个距离函数 $d(h+r,t)$,它用来衡量 $h+r$ 和 t 之间的距离。

(2)TransH[10]:为了解决 TransE 在面对自反关系,以及多对一、一对多、多对多关系的不足,2014 年 WANG Z 等[10]提出了 TransH 模型,其核心思想是对每一个关系定义一个超平面 W_r 和一个关系向量 d_r。$h\perp$、$t\perp h$ 是 h、t 在 W_r 上的投影,这里要求正确的三元组需要满足 $h_r+d_r=t_r$。这样能够使得同一个实体在不同关系中的意义不同,同时不同实体在同一关系中的意义也可以相同。

(3)TransR[11]:TransR 是在 TranE 的基础上的改进,在数学上的描述看起来会更加直观:对于每一类关系,不光有一个向量 r 来描述它自身,还有一个映射矩阵 $\boldsymbol{M}_r$ 来描述这个关系所处的关系空间,即对于一个三元组 (h,r,t),需要满足 $d(h,r,t)=\|\|h\boldsymbol{M}_r+\boldsymbol{r}-t\boldsymbol{M}_r\|\|_2^2\approx 0$。TransE、TransH、TransR 等方法无法很好地解决非一对一关系,而且受限于知识图谱的数据稀疏问题。

(4)其他方法。2019 年 9 月由 KIM S 提出一种基于概率的知识库的新型检测方法,也是通过研究关系与关系之间的相关性来检错,其主要算法是通过计算两个关系的共同前后节点关系和各个节点的前后关系来得到相关性的。本文受该方法启发,但是准确率更高。

3 基于 NSIL 函数的知识库错误检测方法总览

本节主要介绍知识库的基本模型,给出问题定义和工作流程。

3.1 知识库

知识库的常用表现形式是 RDF,以(subject,predicate,object)的三元组形式表示实体之间的许多复杂联系。

3.2 问题描述

正如前文介绍的那样,知识库中有很多过时的不正确的事实,但是大量的研究都在不断发现知识库中缺失的边来填充缺失的成分而忽略了对错误事实的检测。因此,旨在利用关系(谓词)之间的关联性来对不断扩充的知识库进行准确的错误检测。

定义 1 关系关联:在给定知识库 KB 中将各关系(谓词)通过关系函数关联到同一个组别。

定义 2 检测错误:在给定的知识库 KB 中将找到最不可能的事实三元组 F,F 可能是放错的边或者客体或者主体。

定义 3 缺失三元组:在给定的知识库中对缺失的三元组(客体或者主体缺失)进行预测缺失的实体。

定义 4 可信三元组:忽略落单的①谓词②主体,客体。

3.3 工作流程

如图 2 所示,工作流程主要有三个部分:给定 KB,然后通过相关函数来测定各个三元组中关系的关联程度,得到关系间相关性分值并进行划分,最后给定一个知识库,找到其中最不可能的事实三元组,即通过 NSIL 判定的错误事实,对于判定的错误事实,进行预测三元组客体与主体。

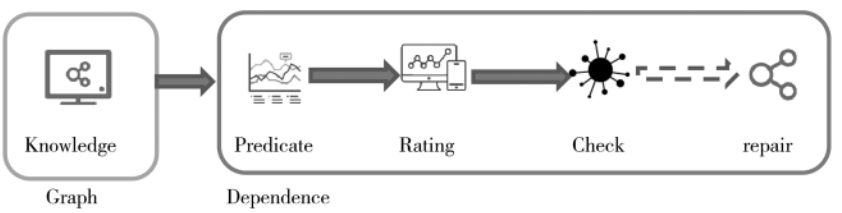


图 2 NSIL 方法工作流程

4 NSIL 关系相关函数

在直觉上,如果两个关系在知识库中拥有更多关联的共同节点(即与主体的关联节点,与主体和客体关联的节点),那么两个关系相关性也越高。在

图 1 的例子中,对于关系“市民 C”和“患病”,它们都有共同的关联性实例“市民 C”,即它们的主体都能够通过另一个相同的实例“市民 C”和它们的客体相连接,并且都拥有两个相同的实例“市民 A”和“COVID-19 病毒”。如果两个关系之间共性越多,就越相信这两个关系相关。

4.1 函数算法

首先介绍符号:$H(r_i)$表示 r_i 主体作为客体时其前置主体的集合,$R(r_i)$表示对于事实三元组(h_i, r_i, t_i)的事实三元组(h_j, r_j, t_j)的集合,其中存在关系 r_x、r_y 构成事实三元组(h_i, r_x, h_j)、(t_j, r_x, t_i),将该三元称为伴生三元组。

两个关系的关联前置主体集合:

$$S(r_i, r_j)=H(r_i)\cap H(r_j) \tag{1}$$

两个关系共同的关联的伴生三元组:

$$I(r_i, r_j)=R(r_i)\cap R(r_j) \tag{2}$$

直接关联集合是两个联系相同的直接关联的主体-客体集合:

$$L(r_i, r_j)=(h_x, r_x)\in U1\cap(h_y, r_y)\in U2,\quad U1.t=t_i, U2.t=t_j \tag{3}$$

定义 SIL 分数函数:

$$\mathrm{SIL}(r_i, r_j)=S(r_i, r_j)+(I(r_i, r_j)+1)\cdot L(r_i, r_j) \tag{4}$$

归一化[12]:

$$\mathrm{NSIL}(r_i, r_j)=\frac{2a\times\tan(\mathrm{NSIL})}{\pi} \tag{5}$$

4.2 NSIL 应用示例

对于图 1:如果加上三元组(“野味市场”,“逗留过”,“市民 A”),那么对于“家人近期病史”和“患病”就有共同的前置节点 S(“家人近期病史”,“患病”)={“野味市场”},共同伴生三元组为 I(“家人近期病史”,“患病”)={“市民 C”},直接关系集合为 L(“家人近期病史”,“患病”)={(“市民 A”,“COVID-19 病毒”)},所以:

$$\mathrm{NSIL}(r_i, r_j)=\frac{2a\times\tan(\text{“野味市场”}+(\text{“市民 C”}+1)\times(\text{“市民 A, COVID-19”}))}{\pi}$$

$$=\frac{2a\times\tan(1+2\times1)}{\pi}=79\%$$;而对于关系“同事”和“患病”,

$$\mathrm{NSIL}(r_i, r_j)=\frac{2a\times\tan(0+0\times0)}{\pi}=0\%$$。可见“家人近期病史”比“有同事”对于“患病”的相关性更高,对“患病”的影响力更大。如果一个人家人近期病史为 MERS 病毒(Middle East Respiratory Syndrome Coronavirus),那么从他的家人近期病史出发推测出他很有可能患上了 MERS。

5 实验

5.1 实验环境

(1)数据集:在实验中,采用了一个基准数据集,即从 Freebase[13]生成的“FB15K”。Freebase 是最大的知识库之一,对应用户构建的现实真实情况如表2所示。

表 2 基准数据集的详细信息

Dataset	#Relation	#Entity	#Training	#Valid	#Test	Source
FB15K	1 345	14 951	483 142	50 000	59 071	Freebase

(2)评估:对于两个数据集进行广泛的错误检测和预测纠错。错误检测将用最新的算法与 NSIR 判断三元组(h, r, t)对错,比较其性能。预测纠错则会用缺失的三元组,即主体或者客体残缺的三元组来进行预测缺失部分。

5.2 实验指标

(1)指标:通过相关得分排名列表,汇总总体测试采用两个评估指标:①Hits@10:排名前十位的实体中判断正确实体的比例,如图 3 所示;②平均等级:正确实体的平均正确率。

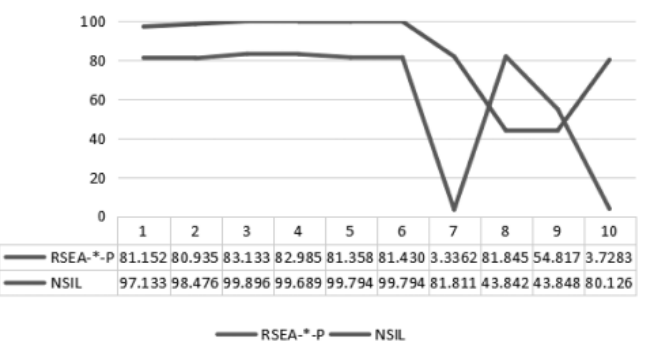


图 3 Hits@10 各排名对应训练集中准确率

(2)准确率认定:

①公式

$$\eta=\frac{\delta}{\mathscr{R}} \tag{6}$$

表示实体排名前 N 个的平均准确率,$\mathscr{R}$ 是给定的与前 N 个实体相关的三元组的总个数,破坏给定三元组的主体或客体,然后对各个三元组的真假性进行判断,δ 是对三元组真假性判断正确的个数。

②对于错误检测的准确率,遵循 Hits@10 原则,

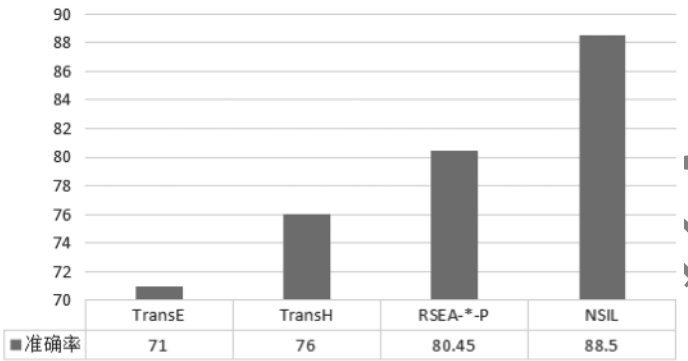
即 $N=10$。

③对于预测纠错的准确率,N 单独地作为横坐标,$N=1\sim10$。

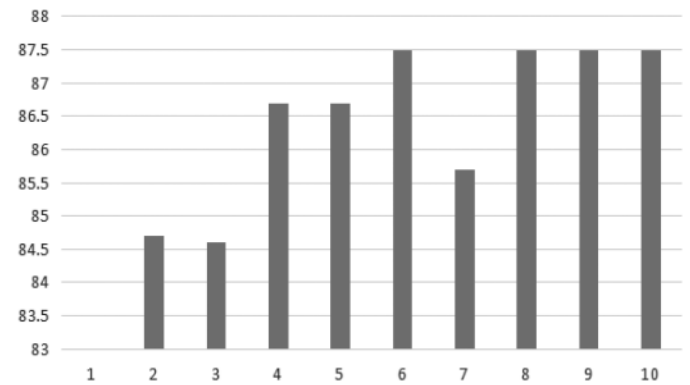
5.3 实验结论

5.3.1 错误检测

对于每一个和其他三元组有纠集联系的关系,都会有一个和其他关系的相关得分,得分低的会被划分为独立关系,得分高的为高相关性关系。对于FB15K 的错误检测的评估在图 4 中示出。在FB15K 上错误检测性能最好的是 NSIL,准确率高达 88.5%。在实验中发现随着实体相关性排名的降低,预测精度出现不稳定与下降,另外对于 NSIL 得分有着高分段密集的缺点(见表 3),这些问题会将在未来的研究内进行克服与探讨。



(1)各方法的错误检测准确率(%)



(2)基于 NSL 对于各实体排名的平均预测纠错准确率(%)

图 4 在数据集 FB15K 上的实验结果

表 4 NSIL 得分 Hits@10

排名 1	0.999 999 94	排名 6	0.999 999 881
排名 2	0.999 999 94	排名 7	0.999 999 821
排名 3	0.999 999 94	排名 8	0.999 999 821
排名 4	0.999 999 94	排名 9	0.999 999 762
排名 5	0.999 999 881	排名 10	0.999 999 762

5.3.2 预测纠错

在知识库中,也会出现缺失主体或者客体的情况。于是对识别出来的残缺相关实体进行主客体的预测,将预测值和真实值比较,从而判断预测纠错的性能。从实体前两名的平均情况、前三名的平均情况,到前十名的平均准确率情况如图 4 所示。Hits@10 的准确率达到了 87.5%。从实验看出,虽然预测纠错有很高的性能,但是其在非一对一关系问题上有着其局限性。

6 结束语

在本文中,提出了一种关系敏感函数 NSIL,用于知识库的错误检测,并且对知识库残缺三元组进行纠错修复。实验表明,该模型不仅可以有效地对知识库中的残缺三元组进行预测纠错,而且在大型知识库 Freebase 数据集上的错误检测均优于现模型。

参考文献

[1] ACOSTA M, ZAVERI A, SIMPERL E, et al. Crowd sourcing linked data quality assessment[M]. The Semantic Web–ISWC 2013. Springer Berlin Heidelberg, 2013.

[2] LIN P, SONG Q, WU Y. Fact checking in knowledge graphs with ontological subgraph patterns[J]. Data Science and Engineering, 2018, 3(4): 341–358.

[3] TÖPPER G, KNUTH M, SACK H. DBpedia ontology enrichment for inconsistency detection[C]. 8th International Conference on Semantic Systems, 2012.

[4] HAO S, TANG N, LI G, et al. A novel cost-based model for data repairing[C]. ICDE, 2017.

[5] HAO S, TANG N, LI G, et al. Cleaning relations using knowledge bases[C]. ICDE, 2017.

[6] LI K, LI G. Approximate query processing: what is new and where to go? A survey on approximate query processing[J]. Data Science and Engineering, 2018, 3(4): 379 – 397.

[7] KIM S, LI X, LI K, et al. Knowledge base error detection with relation sensitive embedding[C]. Inter national Conference on Database Systems for Advanced Applications. Springer, Cham, 2019.

[8] SOCHER R, CHEN D, MANNING C D, et al.

(下转第 37 页)

[2] 上海社会科学院信息所.信息安全辞典[M].上海:上海辞书出版社,2013.
[3] 林国恩,李建彬.信息系统安全[M].北京:电子工业出版社,2010.
[4] 冯登国,孙锐,张阳.信息安全体系结构[M].北京:清华大学出版社,2008.
[5] 项目管理协会.项目管理知识体系指南[Z].北京:电子工业出版社,2005.
[6] 国家质量技术监督局.质量管理体系标准[S].北京:中国计量出版社,2001.
[7] 谢宗晓,郭立生.信息安全管理体系应用手册:ISO/IEC 27001 标准解读及应用模板[M].北京:中国质检出版社,2008.
[8] 范红.信息安全风险评估规范国家标准理解与实施[M].北京:中国标准出版社,2008.

(收稿日期:2020-07-02)

作者简介:

翟亚红(1974-),女,本科,高级工程师,主要研究方向:网络安全、安全服务审核等。

吴治(1978-),男,博士,高级工程师,主要研究方向:网络安全、安全服务审核等。

段静辉(1976-),男,博士,高级工程师,主要研究方向:网络安全、安全服务审核等。

(上接第 27 页)

Reasoning with neural tensor networks for knowledge base completion[C]. NIPS, 2013.
[9] BORDES A, USUNIER N, GARC′IA-DUR′AN A, et al. Translating embeddings for modeling multi-relational data[C]. NIPS, 2013.
[10] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]. Proceedings of the Twenty-Eighth AAAI Conference on Artifificial Intelligence, 2014.
[11] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]. AAAI, 2015.
[12] BOUMA G. Normalized (pointwise) mutual information in collocation extraction[C]. Proceedings of the Biennial GSCL Conference, 2009.
[13] BOLLACKER K D, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. Sigmod Conference. ACM, 2008.

(收稿日期:2020-07-08)

作者简介:

缪琦(1998-),男,本科,主要研究方向:大数据挖掘。

杨昕悦(1996-),女,硕士,主要研究方向:知识图谱。

版权声明