doi: 10.12068/j.issn.1005-3026.2018.11.004

表示学习知识图谱的实体对齐算法

朱继召, 乔建忠, 林树宽
(东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

摘 要: 与现有的根据知识图谱的结构信息或实体属性特征进行相似度匹配的实体对齐的方法不同,提出了一种基于表示学习的知识图谱实体对齐方法. 首先,在低维向量空间下,通过机器学习方法学得实体和关系的语义表示,这种表示形式蕴含了知识图谱的内在结构信息及实体属性特征;其次,将人工标注的实体对作为先验知识,学习知识图谱间实体对的映射关系. 经实验验证表明:与基于特征匹配的方法 SiGMa 相比,本文方法能够有效提高知识图谱实体对齐的精确率,同时保持较高的 $F1$ 值.

关 键 词: 机器学习;表示学习;知识图谱;知识融合;实体对齐
中图分类号: TP 182 **文献标志码**: A **文章编号**: 1005-3026(2018)11-1535-05

Entity Alignment Algorithm for Knowledge Graph of Representation Learning

ZHU Ji-zhao, *QIAO Jian-zhong*, *LIN Shu-kuan*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: QIAO Jian-zhong, professor, E-mail: qiaojianzhong@mail.neu.edu.cn)

Abstract: A novel supervised method for knowledge graph entity alignment based on representation learning was proposed, which is different from the existing methods due to the similarity of structural information or attributive characters. First, the method automatically learns the semantic representations for the entities and relations of a knowledge graph in a low-dimensional vector space was proposed, and these embeddings contain the intrinsically structural information of a knowledge graph and the attributive features of entities. Afterwards, taking the manually aligned entity pairs as prior knowledge, the cross-KG mapping relationship between entities could be learned, which will be used for predicting entity alignment. Experiments conducted on real datasets demonstrated that our method can effectively improve the precision of knowledge graph entity alignment while keeping a high $F1$ score, when compared with the feature matching based method SiGMa.

Key words: machine learning; representation learning; knowledge graph; knowledge fusion; entity alignment

随着万维网技术的迅猛发展,各类 Web 应用不断涌现,引发了网络数据的爆炸式增长. 如何有效地组织和利用大规模网络数据中蕴含的知识,构建人、机器都能理解的智能化网络,是基于知识互联“Web 3.0”时代的目标. 自 2012 年 5 月 Google 发布知识图谱产品以来,国内外一些著名的搜索引擎公司纷纷建立各自的知识图谱产品,如:微软 Bing Satori、百度知心、搜狗知立方等. Google 构建知识图谱的初衷是增强搜索引擎的能力、优化搜索结果、提升用户的搜索体验. 目前,知识图谱被用来泛指各种大规模的知识库,并被广泛应用到智能搜索、深度问答、自动驾驶及语音识别等领域. 然而,知识图谱的构建缺乏统一标准,任何组织机构或个人都可以根据自己的需求和设计理念构建知识图谱,从而导致了知识图谱之间存在严重的异构和冗余. 因此,研究多源知识

收稿日期: 2017-08-01
基金项目: 国家自然科学基金资助项目(61272177).
作者简介: 朱继召(1986-),男,山东菏泽人,东北大学博士研究生; 乔建忠(1964-),男,辽宁兴城人,东北大学教授,博士生导师; 林树宽(1966-),女,吉林长春人,东北大学教授,博士生导师.

的融合技术,整合已有知识资源,从顶层创建一个大规模的统一的知识图谱,从而帮助机器理解底层数据[1],并能够提升相关应用领域服务水平.

实体对齐作为知识融合过程中的关键技术,又被称为实体匹配,是推断来自不同数据集合中的不同实体是否映射到物理世界中同一对象的处理过程,并受到了工业界和学术界的高度关注. 文献[2-3]基于字符串相似性匹配原则,根据待对齐实体属性特征的字面量判定实体匹配与否. 文献[4]基于属性信息,将实体对齐问题转化为分类问题,建立了相应的概率模型. 由于文献[4]建立模型过程中忽略了实体不同属性的重要程度存在差别,文献[5]为属性增加了权重系数,从而提高了实体对齐的精确率. 以上方法的思想简单,并在应用中表现出了不错的效果,但面对跨语言知识图谱或实体的属性字面量描述不统一的情况,这些方法将变得低效甚至不再适用.

近年来,以深度学习为代表的表示学习技术得到快速发展,促使了知识表示学习的提出[6-8]. 知识表示学习是将知识图谱中的实体和关系映射到低维空间,学习得到实体和关系的向量表示[9]. 这种低维稠密的向量表示不仅蕴含了知识图谱内在的结构信息及实体和关系的属性特征,还具有丰富的语义信息. 与基于特征相似性匹配的方法不同,本文基于表示学习,提出了一种用于知识图谱实体对齐的方法. 首先,将待对齐的两个知识图谱分别转化为向量表示形式(称为知识表示);然后,基于得到的知识表示,根据先验对齐数据学习知识图谱间实体对的映射关系. 经实验验证表明:与基于特征匹配的方法 SiGMa[2] 相比,本文方法能够有效提高知识图谱实体对齐的精确率,同时保持较高的 $F1$ 值.

1 问题描述

知识图谱被看作是对客观世界中事物及其相互关系的一种形式化描述[1]. 目前,知识图谱一般采用资源描述框架模式(resource description framework schema, RDFS)或万维网本体语言(Web ontology language, OWL)等形式化方式构建. RDFS 采用定义事实(fact)三元组的形式,通常由主语、谓语、宾语组成,即 SPO(subject - property - object). OWL 通过本体描述客观世界中的知识,其中定义了实例、属性、类别等基本元素. 一般情况下,知识图谱中的实体包含知识图谱中的实例、属性等元素. 本文知识图谱实体对齐是指实例的对齐. 下面给出知识图谱和知识图谱实体对齐的形式化定义.

定义 1 知识图谱. 知识图谱是由以下方式构成的三元组:$\mathrm{KG}=(E,R,F)$,其中 $E=\{e_1,e_2,\cdots,e_{N_e}\}$ 代表实体集合,包括实例及其属性的取值;$R=\{r_1,r_2,\cdots,r_{N_r}\}$ 代表二元关系集合,用来描述实体与实体间的关系;$F\subseteq E\times R\times E$ 代表事实三元组集合.

定义 2 知识图谱实体对齐. 给定两个知识图谱 $\mathrm{KG}_1,\mathrm{KG}_2$,分别找出知识图谱 KG_1(或 KG_2)中的能对齐到 KG_2(或 KG_1)中的所有实体. 即:$\mathrm{Align}_{\mathrm{entity}}(\mathrm{KG}_1,\mathrm{KG}_2)=\{(e,e')\mid e\in E_1,e'\in E_2\}$.

2 基于表示学习的知识图谱实体对齐算法

2.1 算法概述

基于表示学习的知识图谱实体对齐算法由两部分组成:知识表示的学习和实体间映射关系的学习. 首先,将待对齐知识图谱 KG_s 和 KG_t 分别映射到低维空间得到对应的知识表示,分别记作:$\mathbf{KG}_s$ 和 $\mathbf{KG}_t$;其次,基于知识表示 $\mathbf{KG}_s$ 和 $\mathbf{KG}_t$,根据人工标注的实体对齐数据集 N,学得实体对间的对应关系,即:$\varphi:\mathbf{KG}_s\leftrightarrow\mathbf{KG}_t$. 整个算法学习过程中的目标是最小化全局损失 L:

$$L=L_{\mathrm{emb}}(\mathrm{KG}_s,\mathrm{KG}_t,\mathbf{KG}_s,\mathbf{KG}_t)+L_{\mathrm{mat}}(\varphi,N)\ . \tag{1}$$

其中:L_{emb} 代表根据知识图谱 KG_s 和 KG_t 学习得到知识表示 $\mathbf{KG}_s$ 和 $\mathbf{KG}_t$ 过程中的损失;L_{mat} 代表基于知识表示 $\mathbf{KG}_s$ 和 $\mathbf{KG}_t$,在给定的标注数据集 N 上,映射函数 φ 的错误率.

算法的整体流程如图 1 所示.

2.2 知识表示的学习

为方便起见,本小节在描述知识表示的学习过程中,统一使用符号 KG 表示知识图谱,对应的知识表示记作 $\mathbf{KG}$. 按照翻译模型的思想[6],对于事实三元组 (h,r,t),尾实体 t 被看作头实体 h 通过关系 r 的翻译过程,三元组打分函数定义为

$$s(h,r,t)=-\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_{l_2}. \tag{2}$$

基于公式(2),事实三元组 (h,r,t) 出现的概率被定义为

$$p(h,r,t)=\delta(s(h,r,t))\ . \tag{3}$$

其中,$\delta(x)=1/(1+\mathrm{e}^{-x})$ 是 sigmoid 函数.

为了学得知识图谱的向量表示 $\mathbf{KG}$,本文把最大化知识图谱中所有三元组出现概率的对数作为学习的目标函数,即:

$$\sum_{(h,r,t)\in\Delta}\lg p(h,r,t)=\sum_{(h,r,t)\in\Delta}\lg\delta(s(h,r,t)). \quad (4)$$

其中，Δ 表示知识图谱中的事实三元组集合.

按照公式(4)，对知识图谱中出现的事实三元组进行概率最大化等同于最小化 L_{emb}. 求解过程中采用 Word2Vec[9] 中提出的负采样技术，则目标函数可改写为

$$L_{\text{emb}}=\max_{(h,r,t)\in\Delta}\Big[\lg\delta(s(h,r,t))+\sum_{k=1}^{N}E_{(h',r,t')\sim P(\Delta')}\times\lg(1-\delta(s(h',r,t')))\Big]. \quad (5)$$

其中：Δ'表示负三元组集合；n 是负采样的次数；$E_{(h',r,t')\sim P(\Delta')}$表示从 Δ'中随机取出的 n 个负三元组的期望. 负三元组是通过替换正三元组(h,r,t)中的头或尾实体生成的，并保证负三元组$(h',r,t')\notin\Delta$. 需强调的是：在生成负三元组过程中，头实体和尾实体不能同时被替换，即：$\Delta'=\{(h',r,t)\mid h'\in E\}\cup\{(h,r,t')\mid t'\in E\}$.

在整个知识表示学习过程中，采用随机梯度下降算法进行优化求解. 此外，学习实体和关系的表示时分别满足约束限制：

$$\|\boldsymbol{h}\|_{l_2}\leqslant 1,\ \|\boldsymbol{r}\|_{l_2}\leqslant 1,\ \|\boldsymbol{t}\|_{l_2}\leqslant 1.$$

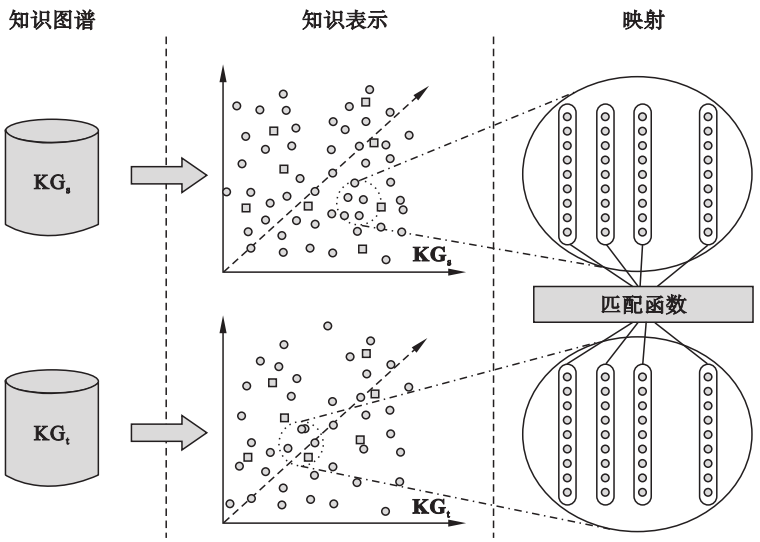

图 1　知识图谱实体对齐算法整体流程

Fig. 1　Oerall flow of our method for knowledge graph entity alignment

2.3　实体间对齐关系的学习

基于 2.2 小节描述的算法学得相应知识表示，本小节将在人工标注的对齐数据集 N 上，学习实体对间的对应关系 φ. 为了方便起见，引入(e_s,e_t)表示数据集 N 中的任一实体对，即：$(e_s,e_t)\in N$，其中 $e_s\in E_s$，$e_t\in E_t$. 公式(1)中的匹配损失 L_{mat}具体定义为

$$L_{\text{mat}}=\sum_{(e_s,e_t)\in N}\sum_{(e_s',e_t')\in N^-}[\varphi(e_s,e_t)+\gamma-\varphi(e_s',e_t')]_+. \quad (6)$$

其中：$[\cdot]_+$表示合页函数；φ 是用来度量两实体间匹配程度的函数；$\gamma>0$ 用来分离正与负实体对的间隔. N^-表示负实体对集合：

$$N^-=\{(e_s',e_t)\mid e_s'\in E_s\}\cup\{(e_s,e_t')\mid e_t'\in E_t\}.$$

本文通过语义匹配来度量实体间的对齐程度. 目前，存在多种语义匹配方式，如：欧氏距离、余弦相似度、双线性函数等. 本文分别使用双线性和张量两种类型的匹配算法.

1) 双线性(bilinear). 双线性函数可以捕获两个实体向量表示间的相互关系，本文引入该类型函数来度量实体表示间的匹配程度，定义为

$$\varphi(e_s,e_t)=\boldsymbol{e}_s^{\mathrm{T}}\boldsymbol{M}\boldsymbol{e}_t+b. \quad (7)$$

其中：$\boldsymbol{M}\in\mathbf{R}^{d\times d}$是双线性变换矩阵；$d$ 表示空间的维度；$b\in\mathbf{R}$ 是偏置量.

2) 张量(tensor). 与双线性函数相比，张量函数更加强大. 线性、双线性函数都可被看作张量函数的简化形式. 张量函数在建模两向量间相互关系方面表现出了不错的效果[10-11]，定义为

$$\varphi(e_s,e_t)=\boldsymbol{u}^{\mathrm{T}}f\left(\boldsymbol{e}_s^{\mathrm{T}}\boldsymbol{M}^{[1:c]}\boldsymbol{e}_t+\boldsymbol{W}\begin{bmatrix}e_s\\e_t\end{bmatrix}+\boldsymbol{b}\right). \quad (8)$$

其中：$f=\tanh$ 是元素级非线性函数；$\boldsymbol{M}^{[1:c]}\in\mathbf{R}^{d\times d\times c}$是三维张量；$d$ 是表达空间的维度；c 是张量分片的数目；$\boldsymbol{W}\in\mathbf{R}^{c\times 2d}$和 $\boldsymbol{b}\in\mathbf{R}^c$是公式(8)中线性部分的参数，$\boldsymbol{u}\in\mathbf{R}^c$.

2.4　实体对齐预测

在预测阶段，对于给定的任意实体 e_s(不妨假设 e_s 是来自知识图谱KG_s 的实体集 E_s，即：$e_s\in E_s$)，预测的目的是找出待对齐知识图谱KG_t

中,与 e_s 指向物理世界中的同一实体 e_t($e_t \in E_t$). 具体做法是:首先,遍历知识图谱 KG_t 中的每个实体 e_t,分别与 e_s 构造为实体对(e_s,e_t),按照匹配算法 φ 进行打分;然后,将打分结果升序排列,分值越低的实体对意味着两实体对齐程度越高. 本文选取 Top1 作为预测结果对算法进行评估.

2.5 复杂性分析

知识表示的学习阶段,在一轮迭代过程中本文算法的时间复杂度为 $O((N_e + N_r)d)$. 其中,d 代表空间的维度. 实体间对齐关系的匹配阶段,不同的匹配算法具有不同的时间复杂度. ①对于双线性算法,匹配阶段的复杂度为 $O(|N|d^2)$,其中 $|N|$ 代表标注的实体对数目. ②对于张量算法,匹配阶段的复杂度为 $O(|N|(d^2+2d)c)$. 在实体对齐预测阶段,知识图谱间实体匹配的时间复杂度为 $O(|N_e|^2)$.

3 实 验

3.1 数据集

为了评估本文提出的算法,在两个不同领域的真实数据集上进行验证. 实验数据的来源分别是:①对科技论文领域公开的数据集 Cora 整理而来;②通过网络爬虫工具,分别抓取百度视频和豆瓣电影的官方网站中的电影/视频信息,并对抓取的信息进行处理而来.

Cora 数据集是英语描述的科技论文书目信息集合,由 CORA 和 cora_gold 两个文件组成. 其中,CORA 描述论文书目信息,共包含 1 879 条实例信息,cora_gold 为人工标注的对齐论文实体. 由于论文书目信息的引用格式存在区别,导致数据集中存在许多重复的论文实例. 根据 cora_gold 文件中人工标注的对齐实体,随机选取对齐到相同论文的一对书目信息,将对应的实体 ID 与书目信息项分别以属性为关系类型构造三元组,然后分配到数据集 CKG1 和 CKG2 中.

百度/豆瓣数据集是中文电影/视频信息集合,分别由手工对齐的 800 部电影组成,包含以下类型信息:电影名称、导演、演员、上映时间和电影类型. 分别将电影 ID 按照属性类型,与所相应的属性值构造成为三元组,添加到相应数据集合中.

以上数据集的详细统计信息见表 1.

表 1 数据集统计信息

Table 1 Statistics of datasets

数据集	三元组数	关系数	实体数	对齐实体数
CKG1/CKG2	2 141/1 938	15	1 562/1 407	117
百度/豆瓣	12 286/14 007	6	7 481/6 954	800

3.2 实验设置

为了说明本文方法在实体对齐任务上的有效性,选择了与基于特征匹配的方法 SiGMa[2] 进行对比.

对本文算法实验验证. 首先学习 CKG1/CKG2,百度/豆瓣在语义空间的向量表示. 在该过程中,空间维度 d 选自集合{20,50,80,100,200},学习率 λ 选自集合{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}},负采样次数 n 选自集合{1,3,5,8,10,15,20,30}. 由于实验选用的两个数据集在语言类型、领域类别、数据疏密度等方面存在区别. 因此,表示学习过程中最优参数的配置往往会不同. 通过采用全网搜索的方式,分别对两组数据集进行训练,最终选定的最优参数配置分别为:①CKG1/CKG2 数据集,$d=50$,$\lambda=0.01$,$n=10$;②百度/豆瓣数据集,$d=80$,$\lambda=0.005$,$n=8$. 其次,匹配函数的学习,各数据集中的对齐实体数据按照 5:1 比例分割,分别用于训练和预测. 在该过程中,间隔 γ 选自集合{1.0,2.0,4.0},张量类型匹配函数中的参数 c 选自集合{2,3,4,5}. 最终最优参数配置为:①CKG1/CKG2 数据集,$\gamma=1.0$,$c=2$(张量);②百度/豆瓣数据集,$\gamma=2.0$,$c=3$(张量).

3.3 结果与分析

根据以上实验设置,分别在两组数据集上进行实验,实体对齐结果如表 2 所示. 通过实验结果可知,在两组数据集上,与 SiGMa 方法相比,本文方法在双线性和张量两种类型匹配算法下均表现出较好的效果. 具体地说,在 CKG1/CKG2 数据集上,SiGMa 的精确率为 82.1%,$F1$ 值为 75.3%. 本文算法达到 91.0%(双线性)和 92.6%(张量)的精确率,$F1$ 值达到 84.2%(双线性)和 86.0%(张量). 与 SiGMa 相比,精确率分别提高了 8.9%(双线性)和 10.5%(张量),$F1$ 值提高了 8.9%(双线性)和 10.7%(张量). 在百度/豆瓣数据集上,本文方法相比于 SiGMa 同样有大幅度的提升. 此外,本文算法在两种不同类型匹配方式下,张量函数表现出的效果更好. 主要原因是,

张量函数与双线性函数相比具有更强的向量间相互关系的建模能力.

表 2　实体对齐实验结果

Table2　Experimental results on entity alignment　　%

方法	CKG1/CKG2			百度/豆瓣		
	精确度	召回率	*F*1	精确率	召回率	*F*1
SiGMa	82.1	69.5	75.3	90.8	79.2	84.6
本文方法(双线性)	91.0	78.4	84.2	97.6	89.2	93.2
本文方法(张量)	92.6	80.2	86.0	98.8	91.7	95.1

在执行效率上,本文算法采用 C++实现,运行于四核服务器. 在百度/豆瓣数据集上,本文算法总耗时大约 4 min,与基于 Python 实现的 SiGMa 算法相比节省近 3 min.

从总体上来看,本文算法和 SiGMa 在数据集百度/豆瓣上的精确率比在 CKG1/CKG2 上增加了6% 以上,而 *F*1 值增加了 9% 左右. 产生这种现象主要有以下两个原因:①CKG1/CKG2 数据集中的属性类别信息较多,另外,许多论文实体的属性信息不全面,而数据集百度/豆瓣中属性类别较少,属性信息也比较完整,从而导致前者数据相比于后者较为稀疏;②由于语言类型的不同,CKG1/CKG2 数据集中论文作者通常存在多种引用格式,如:*Fahlén L E* 和 *Fahlén, Lennart E.* 被看作不同的值,而数据集百度/豆瓣中不存在这种问题.

4　结　　论

本文提出了一种基于表示学习的知识图谱实体对齐方法. 与现有的基于特征匹配的实体对齐方法不同,该方法首先学习实体和关系的语义表示,这种表示蕴含了知识图谱的内在结构信息和实体属性特征. 随后将人工标注的实体对齐数据作为先验知识,学习得到知识图谱间实体的映射关系. 在两组数据集上的实验表明:与 SiGMa 方法相比,本文方法在 CKG1/CKG2 和百度/豆瓣数据集上的实体对齐精确率平均提升 9% 左右,*F*1 值提高 10% 以上.

参考文献:

[1] 庄严,李国良,冯建华. 知识库实体对齐综述[J]. 计算机研究与发展,2016,53(1):165 - 192.
(Zhuang Yan, Li Guo-liang, Feng Jian-hua. A survey on entity alignment of knowledge base [J]. *Journal of Computer Research and Development*, 2016, 53(1): 165 - 192.)

[2] Lacoste-Julien S, Palla K, Davies A, et al. Sigma: simple greedy matching for aligning large knowledge bases [C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, 2013: 572 - 580.

[3] Sun Y, Ma L, Wang S. A comparative evaluation of string similarity metrics for ontology alignment [J]. *Journal of Information & Computational Science*, 2015, 12(3): 957 - 964.

[4] Newcombe H B, Kennedy J M, Axford S J, et al. Automatic linkage of vital records [J]. *Science*, 1959, 130(3381): 954 - 959.

[5] Herzog T N, Scheuren F J, Winkler W E. Data quality and record linkage techniques [M]. Berlin: Springer Science & Business Media, 2007.

[6] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13). Lake Tahoe, 2013: 2787 - 2795.

[7] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes [C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14). Quebec, 2014: 1112 - 1119.

[8] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion [C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15). Austin, 2015: 2181 - 2187.

[9] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning [C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, 2010: 384 - 394.

[10] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]//Advances in Neural Information Processing Systems. Lake Tahoe, 2013: 926 - 934.

[11] Qiu X, Huang X. Convolutional neural tensor network architecture for community-based question answering [C]// Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, 2015: 1305 - 1311.