基于图嵌入的社交账号与知识图谱实体对齐

郭　强,谭菊仙,刘家祝
(江南计算技术研究所,江苏 无锡 214085)

摘　要:社交网络与知识图谱的之间的数据融合对于知识图谱构建和社交网络分析具有重要的应用价值,而社交账号与知识图谱实体的对齐是两类数据融合的关键。针对社交账号与知识图谱实体的对齐问题,结合社交网络与知识图谱的结构特点,文中提出了一种基于图嵌入特征的社交账号实体对齐方法,旨在给定社交账号的情况下,能够在知识图谱中找到正确的对应实体。该方法在目标实体选择阶段将社交关系子图映射成知识图谱子图,利用图嵌入特征选取子图中的核心实体集,并根据核心实体集构造特征向量,选用多层感知机作为分类器,从而确定社交账号所对应的目标实体。使用基于 Twitter 与 Wikidata 的实体对齐数据集进行了实验验证,通过与基线方法的对比,实验结果表明该方法能够达到较好的对齐效果。

关键词:社交网络;知识图谱;数据融合;图嵌入特征;实体对齐
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2021)09-0019-05
doi:10.3969/j.issn.1673-629X.2021.09.004

Graph Embedding Based Alignment Between Social Media Account and Knowledge Graph Entity

GUO Qiang, TAN Ju-xian, LIU Jia-zhu
(Jiangnan Institute of Computing Technology, Wuxi 214085, China)

Abstract: The data fusion between social network and knowledge graph has important application value for knowledge graph construction and social network analysis, and the alignment of social account and knowledge graph entity is the key to the fusion of two kinds of data. In view of the alignment of social account and knowledge graph entity, combining the structure characteristics of social network and knowledge graph, we put forward a social account entity alignment method based on the embedded characteristics of the graph, so as to find the correct corresponding entity in the knowledge graph given the social account. This method maps the social relationship subgraph into a knowledge graph subgraph during the target entity selection stage, uses graph embedding features to select the core entity set in the subgraph, constructs the feature vector according to the core entity set, and choose the multi-layer perceptual machine as the classifier to determine the target entity corresponding to the social account. Experimental validation is performed by the entity alignment data set based on Twitter and Wikidata. By comparing with the baseline method, the experiment shows that the proposed method can achieve better alignment.

Key words: social network; knowledge graph; data fusion; graph embedding features; entity alignment

0　引　言

随着社交网络的日益普及,微博、Twitter、Facebook 等社交媒体成为人们传播新事件、分享新知识的主要媒介。特别是知识图谱中大量的人物、组织在社交网络中都开设有账号。社交媒体实时更新的信息可以帮助扩充知识图谱内容,而知识图谱在社交网络分析中可以起到知识引导的作用[1]。社交媒体和知识图谱互相融合对知识图谱扩充与社交网络分析等具有重要作用,而社交账号与知识图谱实体对齐是这两类数据融合的关键问题。

以实体为中心的知识图谱和以账号为中心的社交网络在数据上呈现出不同的特点。第一,知识图谱的质量一般要求较高,YAGO 具有 95% 的准确度[2]。而对于社交媒体来说,数据通常是嘈杂的,甚至存在虚假信息。第二,知识图谱一般采用标准的、易于计算机访问的数据结构。而对于社交媒体来说,数据主要呈现非结构化特征,多数受限于社交媒体平台的 API 访问限制。第三,社交媒体能够提供实时的最新信息,而知

收稿日期:2020-04-07　　**修回日期**:2020-08-08
基金项目:国家科技部重点研发计划项目(2018ZX01028101)
作者简介:郭　强(1986-),男,工程师,研究方向为人工智能、知识图谱。

识图谱的更新一般滞后数小时到数月之间不等[3]。这种知识更新的滞后,限制了知识图谱在实时性要求较高场景中的应用。这些特点给社交媒体和知识图谱之间的数据融合带来挑战。

文中充分利用社交网络和知识图谱的结构特点,研究社交账号与知识图谱实体的对齐技术,将社交账号与知识图谱中的实体链接起来。与实体链接过程类似,社交账号与知识图谱实体也为两个步骤[1]:候选实体集生成与目标实体选择。以 Twitter 账号与 Wikidata 实体对齐为例,图 1 给出了一个对齐过程。

图 1 Twitter 账号与 Wikidata 对齐

在候选实体生成步骤中,综合使用多种搜索策略,对搜索结果的实体类型进行过滤,只保留人物实体和组织实体。在目标实体选择阶段中,提出了一种新的方法:基于图嵌入特征的算法,利用从社交媒体账户中提取的社交关系,通过知识搜索服务映射成知识图谱的子图,利用知识图谱的图嵌入特征来生成候选实体特征向量,然后通过感知机分类器来选择目标实体。

1 相关工作

实体链接一般是指将文本中的实体提及(entity mention)链接到知识图谱实体的过程[4]。文中研究的问题是将社交账号链接到知识图谱的实体,与通常的实体链接过程类似。Usbeck R 等人[5]发布的 AGDISTIS 系统试图挖掘知识图谱中的子图的节点主题一致性规律,完成批量的实体链接工作。在目标实体选择阶段他们采用 HITS[6]或 PageRank[7]算法,选取重要程度最高的实体为目标实体。AGDISTIS 系统用于社交实体对齐存在一定局限性,主要原因在于 AGDISTIS 系统使用启发式算法,没有考虑图节点的潜在语义特征。

社交账号与知识图谱实体的对齐问题近年来受到学者的关注。2017 年 Trendo 大学的 Nechaev Y 等人[8]首次提出该问题,他们研究了 Twitter 账号与 DBpedia 之间的链接问题,基于监督学习给出了初步解决方案并提出了 SocialLink 问题,指出跨社交网站的账号链接是其中的难点和重点。文献[9]提出了对 SocialLink 问题的改进,引入了 Social Embedding 的概念,与知识图谱中的知识表示学习方法配合使用,以提高对齐的效果。

文献[1]提出一个基于子图相交的启发式算法用于对齐社交账号与知识图谱实体,并利用 Twitter 数据与 Wikidata 数据构建了一个社交账号与知识图谱实体对齐数据集,在该数据集上实现了 0.637 的准确率。这个研究揭示了基于社交关系映射的知识图谱子图,在目标实体“附近”存在聚集特性,利用这一特性预测目标实体能够取得了一定的准确率。然而这种启发式算法没有考虑实体的语义特征,特别是近年来知识图谱表示学习取得较好进展[10],实体的图嵌入特征能够表达实体的语义信息,对实体对齐具有十分重要的作用。文中采取文献[1]的研究框架,探讨实体的图嵌入特征在实体对齐上的应用。

2 问题定义与方法

目的是针对给定的 Twitter 账号 t,在知识图谱 KG 中找出对应的实体 e_t。令集合 C 为账号 t 在 KG 中生成的候选实体集,$C=\{c_1,c_2,\cdots,c_n\}$,函数 φ 表示根据账号 t 在知识图谱 KG 中生成候选实体集,函数 ψ 表示计算候选实体 c_i 为正确实体的概率。链接过程可以形式化地描述为如下两个部分:

(a)候选实体集生成:$C=\varphi(t,\mathrm{KG})$。

(b)目标实体选择:$\tilde{e}=c_q$,其中 $c_i\in C$ 且 $q=\arg\max_i(\psi(c_i))$ 成立。

2.1 候选实体集生成

在候选实体生成阶段,主要对实体类型进行过滤。由于社交账号对应的实体只能是人物或组织,对于知识图谱搜索服务的返回结果,进行实体类型过滤,只保留人物和组织实体。为了使社交账号对应的实体尽可能在返回结果中,使用文献[1]中的用户名策略、用户名去符号策略、用户名分割策略等三个搜索策略,对搜索结果取并集。算法描述过程如下:

算法 1:候选实体生成算法 getCandidates。

输入:社交账号 t;

输出:候选实体列表 C。

步骤:

1. $C \leftarrow$ NULL

2. $\bar{C} \leftarrow$ KGSearch(t.name) $\cup$ KGSearch(remove_tag(t.name)) $\cup$ KGSearch(split(t.name))

3. for c in $\bar{C}$:

4. if fitDomain(c) then:

5. $C = C \cup c$;

2.2 基于图嵌入特征的实体对齐算法(A_{rep})

2.2.1 社交子图生成

从获取的数据中提取与目标账号相关的社交账号以组成社交子图 SG_{sub}。具体来说,从目标账号数据中提取关注(following)、提及(mention)、转发(retweet)和引用(quote)中出现的账号,定义目标账号社交子图实体集合为 SG_{sub},那么提取过程可以形式化表述如下:

$$SG_{sub} = SG_{following} \cup SG_{mention} \cup SG_{retweet} \cup SG_{quote} \tag{1}$$

其中,$SG_{following}$ 等子图表示从目标账号相关数据中提取出来的社交账号集。由于某些账号的粉丝数量巨大,且粉丝对实体对齐算法效果影响不明显,在社交子图中不考虑粉丝账号。

2.2.2 结构投影子图生成

社交子图生成之后,根据每个社交账号的候选实体,构建候选实体之间的知识图谱子图。特定账号的社交子图投影到知识图图谱子图的过程见算法 2。

算法过程描述如下:

算法 2:结构投影算法。

输入:目标账号 t;有关目标账号 t 的爬取数据 data;知识图谱 KG;

输出:结构投影子图 KG_{sub}。

步骤:

1. $KG_{sub} \leftarrow \varphi$

2. $SG_{sub} \leftarrow$ GetSubSocialGraphFromSavedData(t,data)

3. RA $\leftarrow$ getRelateAccount(t,m)

4. FOR $ra_i \in$ RA

5. $C_E \leftarrow C_E \cup$ KGSearchService(ra_i,Tr,k,KG)

6. END FOR

7. KG_{sub} = subgrpah(C_E)

8. RETURN KG_{sub}

2.2.3 图嵌入特征构建

在知识表示学习领域,以 TransE[11] 为代表的翻译模型在知识图谱补全问题上取得较好的效果,能一定程度捕获实体的语义信息,文中使用 TransE 模型的实体嵌入特征。结构投影子图 KG_{sub} 在结构上存在聚集特征[1],文中充分利用这个特点来构造候选实体的特征向量。使用一种迭代删除 KG_{sub} 中离散实体,保留最“密集”处核心实体的算法。该算法每次迭代计算 KG_{sub} 的质心,删除一定数量离质心最远的实体,最终保留特定个数核心实体。算法描述如下:

算法 3:核心实体集生成算法。

输入:投影子图 KG_{sub};实体嵌入特征列表 W;离散实体删除率 p;核心实体保留数 m。

输出:核心实体集 S_{core}。

步骤:

1. $S_{core} \leftarrow$ GetVectors(KG_{sub},W)

2. $L_{distance} \leftarrow \varphi$

3. LOOP

4. $I_{centroid} \leftarrow \dfrac{\sum_{i=0}^{|S_{core}|} I_i}{|S_{core}|}$

5. FOR I_i IN S_{core}

6. $L_{distance} \leftarrow L_{distance} \cup$ EuclideanDistance($I_{centroid}$,I_i)

7. END FOR

8. $L_{distance} \leftarrow$ SortListByDistanceDesc($L_{distance}$)

9. IF $|S_{core}| \times p > 1$ THEN

10. $k \leftarrow |S_{core}| \times p$

11. ELSE

12. $k \leftarrow 1$

13. END IF

14. $S_{core} \leftarrow$ RemoveTopKElement(S_{core},$L_{distance}$,k)

15. IF $|S_{core}| \leqslant m$ THEN

16. BREAK

17. END IF

18. $L_{distance} \leftarrow \varphi$

19. END LOOP

20. RETURN S_{core}

利用该算法得到的核心实体集 S_{core},构造每个候选实体特征向量 $I_{feature}$ 如下:

$$I_{feature} = \sum_{i=0}^{|S_{core}|} I_i - c_j \tag{2}$$

其中,$I_i \in S_{core}$,c_j 为候选实体的特征向量。知识表示学习工具 OpenKE[12] 使用 TransE 模型对 wikidata 全量数据进行了训练,文中直接使用其训练结果。

2.2.4 目标实体选择

目标实体选择以特征向量 $I_{feature}$ 为输入,计算候选实体为目标实体的匹配值,最后根据这一组候选实体匹配值,选择最终对齐实体。

在目标实体匹配值计算的设计中,为了能够更好处理 $I_{feature}$ 这一类特征向量,引入多层感知机(MLP)模型[13]的神经网络来计算匹配值。MLP 模型在结构上是一个多层的全连接网络,除了输入层(input layer)和输出层(output layer)外,中间还有若干隐层(hidden layer),层与层之间全连接,隐层和输出层存在激活函数。MLP 模型采用梯度反向传播算法训练参数。

文中设计了一个单隐层的 MLP 模型,具体结构如图 2 所示。

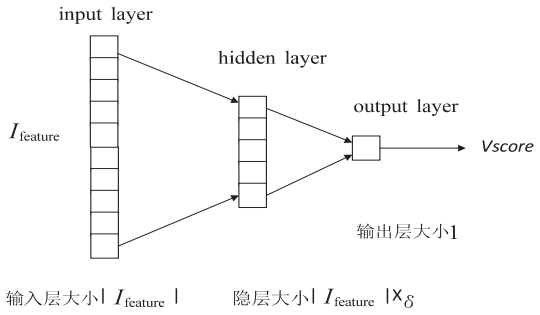

图 2 目标实体匹配值计算模型结构示意图

其中,隐层的大小等于 I_{feature} 的维数乘以一个给定的隐层大小系数 σ ,隐层激活函数为 ReLU 函数,输出层大小等于 1,输出的结果即为目标实体匹配值, $v_{\text{score}} \in [0,1]$ 。损失函数采用均方差损失函数(MSELoss),由于模型输出大小为 1,故损失函数公式如下:

$$L_{\text{MSELoss}} = (V_{\text{score}} - V_{\text{label}})^2 \tag{3}$$

其中, V_{label} 为训练数据标签值。训练过程的反向传播调整参数的过程使用了 Adam[14] 算法作为优化器。为了防止过拟合,模型采用 L2 正则化方法。

针对一个候选实体集 C ,计算每个候选实体的匹配值 $V_{\text{score}} = \{v_1, v_2, \cdots, v_n\}$, n 为候选实体个数,选择分值最高的候选对象为目标实体。

3 实验与分析

实体对齐旨在从候选实体集中选择最有可能的实体作为目标实体,故最终的结果只有“成功”或“失败”两种结果。参考文献[1]的评价方法,文中衡量方法性能的指标为准确率(Accuracy)。

3.1 对比算法

为了验证基于图嵌入特征的实体对齐算法(A_{rep})的有效性,引入了三种对比算法,分别是:标题匹配法、AGDISTIS[5] 算法、子图相交算法[1]。

3.1.1 标题匹配算法(A_{title})

标题匹配法以 Twitter 账号用户名与候选实体标题字符串的相似度为选择标准,选择第一个与 Twitter 账号用户名完全相同的候选实体为目标实体。

3.1.2 AGDISTIS 算法(A_{HITS})

AGDISTIS 算法对知识图谱子图进行深度为 2 的广度优先搜索,从而生成新子图,然后使用 HITS 算法计算新子图的节点权威值,选取权威值最高的节点作为链接结果。

3.1.3 子图相交算法(A_{sub})

文献[1]在 AGDISTIS 算法的基础上提出了子图相交算法,它将候选实体进行深度为 3 的广度优先搜索,为每个候选实体生成一个子图,然后将社交账号相关联的账号投影到知识图谱生成目标子图,计算目标子图和候选实体子图的交集,选择交集元素最多的候选实体作为最终对齐结果。

3.2 实验数据

文献[1]通过 Wikidata Query Service[15],利用 SPARQL[16] 语言获取了 3 024 条具有 Twitter 账号的 Wikidata 实体,其中包含 1 379 个人物账号,1 645 个组织账号。然后根据 Twitter 账号名,利用网络爬虫技术,爬取相关账号的基本信息、推文及关注账号列表。为了保证能够获取较为可靠的社交关系,去除了推文总数在 300 条以下且关注总数在 100 以下的账号,最终保留账号 2 281 个,其中人物账号 1 086 个,组织账号 1 195 个。

为了进行实验对比,根据文献[1]的方法对数据集进行扩充,重新获取 15 962 个 Twitter 账号作为训练集和验证集,其中人物账号 10 256 个,组织账号 5 706 个,将文献[1]中的 2 281 个账号作为测试集。

3.3 目标实体选择

基于图嵌入特征的实体对齐算法涉及的主要超参数如下:

(a)核心实体保留数 m 。该参数表示核心实体集生成算法返回的核心实体集最终包含的实体个数,取值范围[20,40,60,80,100]。

(b)隐层大小系数 σ 。该参数用于 MLP 模型根据输入层确定隐层神经元个数的系数, $\sigma \in (0,1]$,取值范围[0.2,0.4,0.6,0.8,1]。

为了选择最优超参数,按照 7∶1 的比例将训练数据集划分为训练集、验证集,使用验证集进行网格搜索,确定最优超参数组合,见表 1。

表 1 最优实验参数组合

参数	值
核心实体保留数 m	20
隐层大小系数 σ	0.6

在测试数据集上,应用最优超参数组合进行性能评估,我们得到基于图嵌入特征的实体对齐算法的最终实验结果 A_{rep} 。实验数据集将按照数据类型分为人员、组织、综合(人员+组织)分别进行实验。实验结果 A_{rep} 与标题匹配算法、AGDISTIS 算法、子图相交算法进行对比,得到实验结果见表 2。

表 2 基于图嵌入特征的实体对齐算法实验结果

方法	人员	组织	综合
A_{title}	0.524	0.337	0.426
A_{HITS}	0.532	0.541	0.537
A_{sub}	0.787	0.500	0.637
A_{rep}	0.943	0.760	0.842

从表 2 中可以看出,基于图嵌入特征的实体对齐

方法 A_{rep} 在整个数据集上达到了最好的性能,相比于基于子图相交的启发式算法综合准确率提升了 32%。A_{rep} 算法的核心是以实体图嵌入特征为基础的目标实体匹配值计算模块,它既能利用图的聚集特点,又能够利用实体的语义特征,在获得更多标记数据的情况下,对齐方法的准确率可以进一步提升。

4　结束语

文中提出了一种将社交账号与知识图谱实体进行对齐的算法——基于图嵌入特征的实体对齐算法(A_{rep})。通过将目标账号的社交关系图映射到知识图谱中形成子图,充分利用子图存在聚集特征的特点,以核心实体集的表示学习向量为基础构造特征向量,最终通过多层感知机来选择目标实体。该研究表明了基于图嵌入特征的实体对齐方法,能够利用实体的语义特征,从而达到更好的实体对齐效果。该方法在测试数据集上实现了 0.842 的准确率。A_{rep} 算法所利用的社交媒体的社交关系图以及知识图谱的图结构等信息,是普遍存在于社交媒体和知识图谱中的,所以该对齐方法可以应用于其他的社交媒体和知识图谱。

下一步的工作可以从两个方面开展。首先是应用更为高效和准确的投影方法来生成投影子图,将会有助于提高投影子图的聚集特征。其次是扩充数据集,加入在知识图谱中不存在对应实体的社交账号用于扩展算法和评估算法的性能。

参考文献:

[1] 刘家祝,郭　强,吴碧伟,等. 基于子图相交的社交账号与知识图谱实体对齐[J]. 计算机技术与发展,2020,30(5):10-15.

[2] HOFFART J,SUCHANEK F M,BERBERICH K,et al. YAGO2:a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence,2013,194:28-61.

[3] FETAHU B,ANAND A,ANAND A. How much is Wikipedia lagging behind news? [C]//Proceedings of the ACM web science conference. Oxford,England:ACM,2015:28.

[4] 陆　伟,武　川. 实体链接研究综述[J]. 情报学报,2015(1):105-112.

[5] USBECK R,NGOMO A C N,RÖDER M,et al. AGDISTIS-graph-based disambiguation of named entities using linked data[C]//The semantic web - ISWC 2014. Riva del Garda,Italy:Springer,2014:457-471.

[6] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM,1999,46(5):604-632.

[7] PAGE L,BRIN S,MOTWANI R,et al. The PageRank citation ranking:bringing order to the web[R]. Stanford:Stanford InfoLab,1999.

[8] NECHAEV Y,CORCOGLIONITI F,GIULIANO C. Linking knowledge bases to social media profiles[C]//Proceedings of the symposium on applied computing. Marrakech Morocco:ACM,2017:145-150.

[9] NECHAEV Y,CORCOGLIONITI F,GIULIANO C. SocialLink:exploiting graph embeddings to link DBpedia entities to Twitter profiles[J]. Progress in Artificial Intelligence,2018,7(4):251-272.

[10] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展,2016,53(2):247-261.

[11] BORDES A,USUNIER N,GARCIA-DURAN A,et al. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems. Lake Tahoe,Nevada,United States:NIPS,2013:2787-2795.

[12] HAN X,CAO S,LV X,et al. OpenKE:an open toolkit for knowledge embedding[C]//Empirical methods in natural language processing. Brussels,Belgium:EMNLP,2018:139-144.

[13] NIELSEN M A. Neural networks and deep learning[M]. San Francisco,CA,USA:Determination Press,2015.

[14] KINGMA D P,BA J. Adam:a method for stochastic optimization[C]//International conference on learning representations. VenueSan Diego,CA,USA:ICLR,2015.

[15] ERXLEBEN F,GUNTHER M,KROTZSCH M. Introducing Wikidata to the linked data web[C]//International semantic web conference. Riva del Garda,Italy:Springer International Publishing,2014.

[16] BIZER C,SCHULTZ A. The berlin sparql benchmark[J]. International Journal on Semantic Web and Information Systems,2009,5(2):1-24.