计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009809] [CSTR: 32024.14.csa.009809]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 融合主题模型的图卷积神经网络知识图谱实体对齐①

李腾腾, 杨　光

(南京航空航天大学 计算机科学与技术学院, 南京 211106)
通信作者: 李腾腾, E-mail: lt_teng7@nuaa.edu.cn

**摘　要**: 实体对齐技术旨在识别并匹配不同知识图谱中指代同一实体的项, 对于知识图谱的融合具有至关重要的作用, 其在知识补全、社交网络分析等多个领域已经展现出广泛的应用潜力与显著的实用价值. 随着基于知识表征学习的实体对齐方法的不断演进, 研究者们开始探索利用实体之间的多种信息维度来计算相似度, 从而评估源实体与目标实体之间的相似性. 尽管如此, 实体的部分属性信息在目前已有的方法中仍未得到充分利用, 尤其是实体属性中的主题信息, 通过主题模型能够识别出实体间更为显著的语义联系. 针对这一研究, 以实体属性的主题信息为核心, 提出了一种实体对齐框架 EAGT (knowledge graph entity alignment via graph convolutional network with biterm topic model), 通过实体主题结合图卷积神经网络进行实体对齐. 为了验证所提方法的有效性, 在开源的数据集上进行了实验, 结果表明, EAGT 在大多数情况下均实现了性能提升.

**关键词**: 知识图谱; 实体对齐; 主题模型; 图卷积神经网络

引用格式: 李腾腾,杨光.融合主题模型的图卷积神经网络知识图谱实体对齐.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9809.html

## Knowledge Graph Entity Alignment via Graph Convolutional Neural Network with Topic Model

LI Teng-Teng, YANG Guang

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

**Abstract**: Entity alignment technology aims to identify and match items that refer to the same entity across different knowledge graphs. It plays a crucial role in the integration of knowledge graphs and demonstrates broad application potential and significant practical value in multiple fields such as knowledge completion and social network analysis. With the continuous evolution of entity alignment methods based on knowledge representation learning, researchers begin to explore the use of multiple information dimensions among entities to calculate similarity, thereby evaluating the similarity between source and target entities. Nonetheless, some of the attribute information of entities is not fully exploited in existing methods, especially the thematic information within entity attributes. By using topic models, more prominent semantic connections between entities can be identified. Focusing on this research, with the thematic information of entity attributes as the core, this study proposes an entity alignment framework called EAGT (knowledge graph entity alignment via graph convolutional networks with biterm topic model), which aligns entities by combining entity topics with graph convolutional neural networks. To verify the effectiveness of the proposed method, experiments are conducted on open-source datasets. The results show that EAGT achieves performance improvements in most cases.

**Key words**: knowledge graph; entity alignment; topic model; graph convolutional neural network (GCN)

① 收稿时间: 2024-09-24; 修改时间: 2024-10-21; 采用时间: 2024-11-01; csa 在线出版时间: 2025-02-28

知识图谱 *KG* (knowledge graph) 作为自然语言处理领域的基石, 其构建过程通常涉及不同的机构, 这导致了构建过程中的不一致性, 以及数据冗余和异构性问题. 知识图谱的结构通常由数百万个实体组成, 这些实体通过关系或属性相互连接, 其三元组表示法通常为<头实体, 关系, 尾实体>或<头实体, 属性, 值>. 鉴于知识图谱的规模庞大, 人工链接实体以获取完整信息变得不切实际. 因此, 如何有效地链接不同知识图谱中相同实体, 成为一个亟待解决的关键问题[1].

实体对齐 *EA* (entity alignment)[2]的核心目标是识别并匹配出不同知识图谱中指代同一现实世界实体的项. 近年来, 随着知识图谱的快速发展, 如 DBpedia[3]、YAGO[4]和 BabelNet[5]等, 它们各自以独特的方式构建, 导致同一实体在不同图谱中可能以多种形态呈现. 借助实体对齐技术, 可以有效地整合分散在不同数据源中的信息, 还能够促进对实体全面理解的深化. 实体对齐的关键问题在于解决异构数据源之间的实体匹配问题, 通过建立跨数据源的一致性关系, 以识别、关联并消除重复实体[6].

在实体对齐的研究领域中, 传统的实体对齐方法主要侧重计算实体之间的标签和字符距离, 包括基于相似性计算[7–9]、基于关系推理[10–12]的方法, 具体而言, 它们采用直观的计算过程和有限的参数集来学习实体的低维嵌入表示. 尽管这些方法在模型训练上相对高效, 但它们在处理知识图谱的复杂结构和丰富的语义信息时, 往往难以捕捉实体间的深层次关系, 导致对齐效果不尽如人意. 相对而言, 基于知识表示学习的实体对齐方法, 已经成为该领域的研究前沿, 并取得了较为显著的成效[13]. 该方法首先通过实体表征嵌入技术为两个知识图谱中的实体生成高维向量表示, 然后将这些嵌入映射在同一向量空间中对齐, 最终通过计算实体向量间的相似度得分来识别和匹配跨图谱的实体[14]. 典型的方法是基于图卷积神经网络 GCN (graph convolutional neural network)[15]的模型[16–18]. 此外, 利用 BERT 和 Transormer 等先进的自然语言处理模型, 研究者们开发了更为复杂的实体对齐框架, 如 BERT-INT[19]和 MEAformer[20], 以及多模态[21]和时态[22]等不同研究角度. 目前的工作主要关注实体的统计特征和高维度的语义表征信息, 未对实体属性中的主题信息进行利用, 这限制了对齐效果的进一步提升. 具体而言, 在对齐过程中, 若目标实体缺乏明确的对齐邻居节点, 可能会导致在实体嵌入的学习过程中发生信息丢失. 以图 1 为例, 上图“汕尾市”邻居结点包含“惠州市”“客家语”等, 而下图“Shanwei”的邻居结点包含“Guangdong”“Suki_Chui”等, 源实体“汕尾市”与目标实体“Shanwei”其各自的邻居结点是无法对齐的. 这种情况下, 由于缺少对齐的相邻节点, 实体在利用结构特征进行卷积的过程中会有信息丢失, 造成汕尾市和 Shanwei 这两个实体嵌入的差距较大, 在对齐阶段几乎没有证据支持这两个实体对齐.

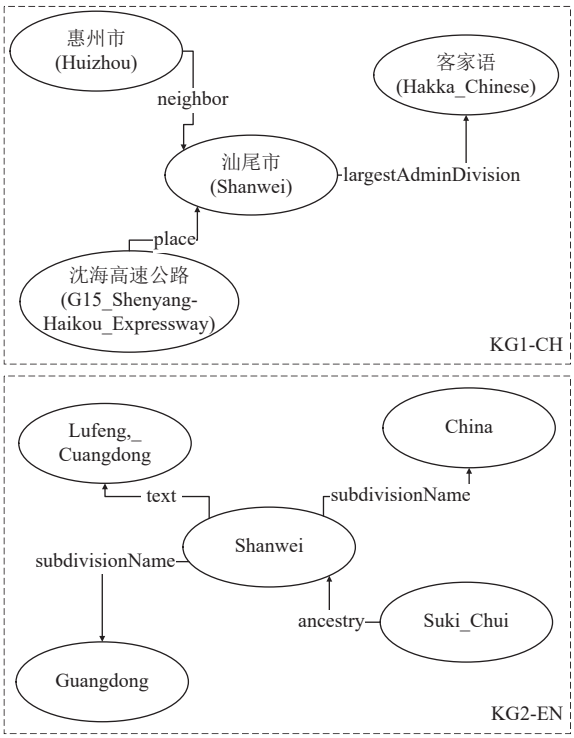


图 1 源实体“汕尾市”和目标实体“Shanwei”对齐邻居

为了解决这个问题, 提出了一种新颖的实体对齐框架 EAGT (knowledge graph entity alignment via graph convolutional networks with biterm topic model), 通过引入主题模型来学习源实体和目标实体之间的主题表征. 通过主题模型的分析, 能够识别出实体间更为显著的语义联系, 从而提高结构上不够相似的目标实体排名. 实验结果表明, EAGT 不仅提高了实体对齐的准确性, 而且可以通过实体的主题分布加强模型的可解释性.

本文主要贡献如下.

(1) 提出了一种基于主题模型和图卷积神经网络的实体对齐框架 EAGT, 该框架旨在通过获取实体属性的主题表征来提升实体对齐的准确性和可解释性.

(2) 为了有效地学习实体的主题信息, 构建了一个属性集合, 该集合利用实体属性信息作为训练数据, 通

过主题模型来训练并捕获实体的属性主题表征.

(3) 在公开的 DBP15K 和 WK3L-15K 数据集上的实验中, 结果表明加入主题模型可以提升实体对齐的性能.

## 1 相关工作

传统的实体对齐方法通常分为两大类: 基于相似度计算的实体对齐和基于关系推理的实体对齐. 其中基于相似度计算有 Cohen 等人[7]使用 TFIDF (term frequency-inverse document frequency) 计算实体名称之间的距离来对齐实体; Sarawagi 等人[8]将实体对齐看作是分类问题; Arasu 等人[9]提出了结合过滤机制和主动学习的方法来对齐实体. 基于关系推理有 Suchanek 等人[10]将实体关系推理进行概率化计算来对齐实体; Lacoste-Julien 等人[11]利用贪心算法思想进行局部搜索来对齐实体; Song 等人[12]采用启发式搜索进行过滤并结合无监督学习来对齐实体. 这些传统的实体对齐方法涉及各个领域, 缺少通过统一的标准进行衡量, 而且忽略了很多隐藏的语义信息, 使得对齐效果表现一般.

目前, 许多研究者使用基于知识表示学习的实体对齐方法, 核心思想就是首先将知识图谱进行嵌入, 然后其映射到同一个向量空间, 最后通过相似度计算衡量实体之间的距离[23]. 基于知识表示学习的实体对齐方法已经成为解决实体对齐的主流技术, 其大多数使用翻译模型[24,25]或图神经网络 GNN (graph neural network)[26]进行实体嵌入. 基于知识表示学习的方法通常利用实体的结构信息、属性信息[27]、实体名信息、实体描述信息等进行对齐. 此外, 还用利用 BERT, Transformer 模型进行的实体对齐模型, 如 BERT-INT[19]和 MEAformer[20]效果更佳.

Wang 等人[16]首次提出利用 GCN 对结构信息和属性信息联合嵌入进行实体对齐, 然而对实体之间关系缺少嵌入, 因此实体对齐效果较差. Wu 等人[18]对关系进行了嵌入表示, 并引入了高速门机制. 以上基于 GCN 的实体对齐方法都偏重关注实体结构信息, 由于缺少对实体本身信息的深度挖掘, 实体对齐效果表现一般. BTM (biterm topic model) 模型是由 Yan 等人[28]提出的一种更适合短文本主题学习的概率模型, 其提出的双向主题模型, 采用一种新颖的方式来构建短文本语料库. 这里文档中的任何两个不同的单词都会构成一个双项. 在提取每个文档中的双术语后, 整个语料库现在变成了双术语集. 双项提取过程可以通过对文档的单次扫描来完成. 通过对实体属性主题进行训练, 能够表达出实体的主题. 受到上述内容启发, 提出了 EAGT 来对实体对齐进行研究.

## 2 EAGT 方法

### 2.1 数据定义

在知识图谱的研究与应用中, 现实世界的实体及其相关知识被结构化为三元组的形式, 以便于机器的理解和处理. 知识图谱中的三元组主要分为两类: 关系三元组和属性三元组. 关系三元组用于表达实体间的相互联系, 遵循 $(entity_1, relation, entity_2)$ 的格式, 其中 $entity_1$ 和 $entity_2$ 分别是关系 $relation$ 的头实体和尾实体. 而属性三元组则用于描述实体的具体属性, 格式为 $(entity, attribute, value)$, 其中 $value$ 是 $entity$ 属性 $attribute$ 的具体描述.

形式上, 将 $KG$ 表示为 $KG=(E,R,A,T_r,T_a)$, 其中 $E$ 代表实体集合, $R$ 代表关系集合, $A$ 代表属性集合; $T_r=\{(h,r,t)\in H\times R\times T\}$ 代表关系三元组的集合, 其中 $H\subseteq E$ 是头实体集合, $T\subseteq E$ 是尾实体集合; $T_a=\{(e,a,v)\in E\times A\times V\}$ 代表属性三元组的集合, 其中 $V$ 是属性值的集合. 具体到跨语言知识图谱实体对齐的问题, 考虑到两个不同语言的知识图谱 $KG_1=(E_1,R_1,A_1,T_{r_1},T_{a_1})$ 和 $KG_2=(E_2,R_2,A_2,T_{r_2},T_{a_2})$ 以及它们之间预先对齐的实体对集合 $EA=\{(s,t)|s\in E_1,t\in E_2\}$. 实体对齐就是找到一个映射 $M: s\rightarrow t$, 它能够将源图谱 $KG_1$ 中的实体 $s$ 映射到目标图谱 $KG_2$ 中的实体 $t$, 基于此映射进一步识别和对齐更多的实体. 在实体对齐的不同类型的任务中, 映射 $M$ 的性质可能有所不同, 它可以是单射 (一对一)、满射 (多对一)、双射 (一对一且每个实体都有映射). 在本研究中, 假设对于测试中的源实体 $s$, 存在一个唯一的目标实体 $t$ 与之对应. 因此, 跨语言知识图谱对齐任务可以定义为, 基于现有的实体对齐集合 $EA$, 发现并建立新的实体对齐映射 $M$.

### 2.2 模型框架

在现有的基于嵌入的实体对齐模型中, 当源实体和目标实体的邻居节点数量有限时, 模型往往难以充分捕捉实体间的共性. 这一局限性表明, 仅依赖神经网络进行实体对齐不足以全面描述实体间的相似性. 鉴于此, 实体的属性信息, 作为其固有特征的直接体现, 不仅提供了丰富的语义内容, 而且在邻居节点匮乏的

情况下, 能够有效增加实体之间的相似度, 从而增强实体对齐的准确性.

本研究提出的 EAGT 模型, 即结合图神经网络和主题模型的实体对齐框架, 旨在通过整合图卷积神经网络嵌入和语义主题信息来提升实体对齐的性能. EAGT 模型的工作流程如图 2 所示: (1) 主题学习模块, 构建实体属性集合, 并应用主题模型挖掘实体属性背后的主题特征; (2) 实体嵌入模块, 利用图卷积神经网络对两个待对齐的知识图谱中的实体进行嵌入表示, 以捕获实体的结构信息; (3) 实体对齐模块, 结合 GCN 得到的嵌入和 BTM 得到的主题嵌入, 对源实体和目标实体执行对齐, 以实现对实体间相似性的全面评估.

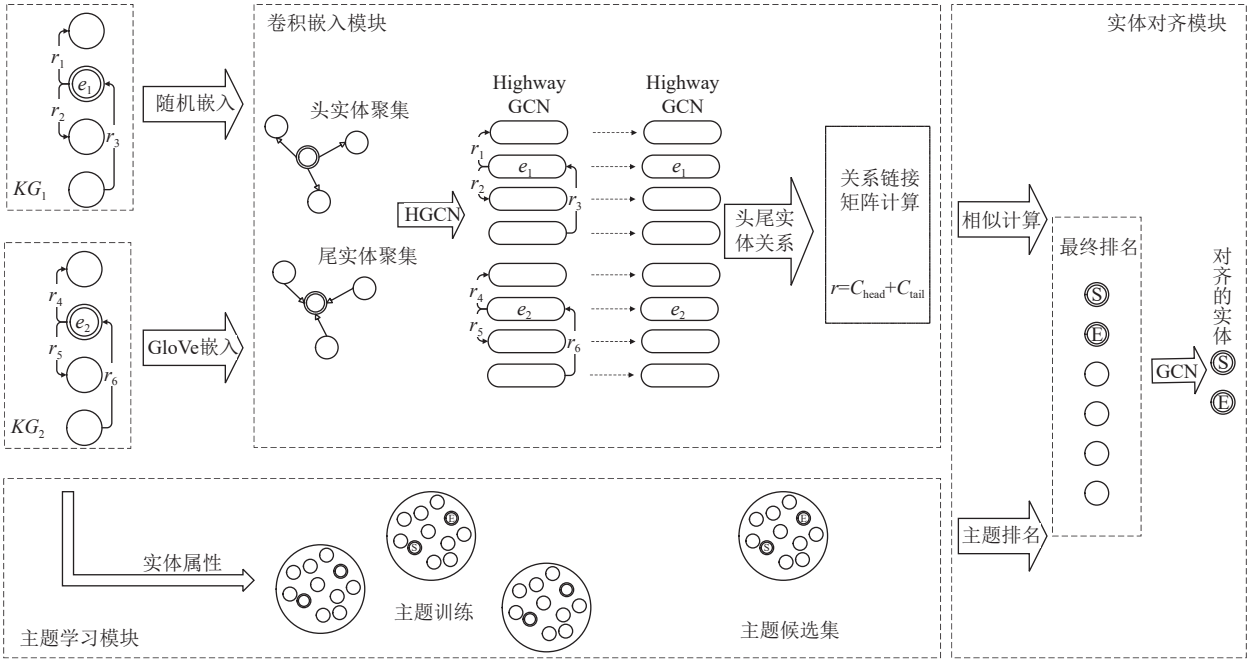


图 2 EAGT 数据流图

2.2.1 主题学习模块

BTM 基本思想是通过直接建模词对 (biterm) 的共现关系来捕捉潜在的主题结构, 短文本中的所有词对都被提取出来, 一个词对是指文本中两个不同的词的组合, 顺序无关. 例如对于短文本“$abc$” (其中$a,b,c$分别表示不同的词语) 可以生成以下词对$(a,b)$、$(a,c)$、$(b,c)$. 通过统计所有词对在文档集合中的共现次数来构建一个全局的词对共现矩阵. 主题模型能够识别出实体间更为显著的语义联系, 提取更高级的语义信息, 进而提高结构上不够相似的目标实体排名.

为了对实体属性集合进行主题建模, 将实体的属性视为主题建模的基本单位, 并假设属性相似度高的实体更有可能归属于同一主题. 例如, 实体“北京话 (Beijing_dialect)”的属性“语系 (familycolor)”和“地区 (region)”通常共同出现用于描述实体信息, 因此它们可以被归纳为同一主题.

在主题模型的理论框架下, 一个主题被表征为一组相关实体属性的集合. 实体属性数据集中每个实体平均拥有约 16 个属性, 传统是主题建模方法通常使用 LDA (latent Dirichlet allocation) 模型[28], 然而通过训练发现该方法并不适合短文本数据, 因此选择在短文本主题建模表现更好的 BTM 模型. 为了对齐不同知识图谱中的实体, 将待对齐的源实体属性和目标实体属性作为训练的语料库, 即整个文档$D$如下所示:

$$D=\{t \mid t \in (T_{a1},T_{a2})\} \tag{1}$$

其中, $t$是$KG_1$和$KG_2$一个实体属性.

令$z \in [1,K]$为主题指示变量, 整个属性集合为$K$个主题. 使主题$K$符合多项式分布$\theta \in \{\theta_k\}_{k=1}^{K}$来表示语料库的主题分布概率, 即$P(z)$, 其中$\theta_k=P(z=k)$且$\sum_{k=1}^{K}\theta_k=1$. 每个主题的属性可以用$K\times t$矩阵$\varphi$表示, 其中第$\varphi_k$是所有属性在第$k$个主题上的多项式分布, 即$P(t \mid z)$, 其中$\varphi_{k,t}=P(t \mid z=k)$且$\sum_{t=1}^{T_a}\varphi_{k,t}=1$.

同样, 遵循 BTM 主题模型的约定, 对$\theta$和$\varphi_k$使用对

称狄利克雷先验, 分别具有超参数$\alpha$和$\beta$. $\alpha$越小, 主题之间相关性越小; $\beta$越大, 主题复杂度可能更高. 主题模型学习的过程定义如算法 1 所示.

算法 1. BTM 的吉布斯采样算法

输入: 实体属性集合数量上限$K$; 超参数$\alpha$和$\beta$; 双向词数据集合$D$.
输出: 主题概率$\theta$和$\varphi$.

1) 随机初始化所有双向主题词分配
2) 对于每个在$N_{\text{iter}}$里面的$n_{\text{iter}}$
3) 对于每个在集合$D$里面的$d$=$(t_i,t_j)$
4) 从$P(z_i|z_{-i},D)$中抽取主题$k$
5) 更新 $n_k$, $n_{w_i|k}$和$n_{w_j|k}$
6) 结束此次循环
7) 结束此次循环
8) 计算并返回$\theta$和$\varphi$

通过设定合适的超参数$\alpha$和$\beta$, 以训练出每个属性的主题概率分布, 进而推导出整个实体属性的主题概率分布.

2.2.2 卷积嵌入模块

(1) 结构和属性嵌入

首先将实体名称序列嵌入为初始实体表示. 通过预训练的 GloVe (global vectors for word representation)[29]词向量实现, 将实体名称中的每个单词映射到低维向量空间中. GloVe 词向量捕捉了单词之间的语义关系, 让语义相近的单词在向量空间中更接近. 然后将嵌入后的单词序列输入到 LSTM (long short-term memory) 网络中, 以生成实体的初始表示向量. LSTM 能够捕捉序列中的长距离依赖关系, 从而更好地表示实体名称的语义信息.

对于每个头实体, 将其所有尾实体视为邻域; 同理, 对于每个尾实体, 将其所有头实体视为邻域. 通过 HGCN (highway-graph convolutional network)[18]模型将每个头尾实体分别赋予结构特征向量$H_s$和属性特征向量$H_a$. 为了提升模型的表征能力, 对属性进行了筛选, 选取了出现频率在特定范围内的属性, 即$30<Counter(h)<2000$. 将卷积计算重新定义为:

$$[H_s^{l+1};H_a^{l+1}]=ReLU\left(\hat{D}-\frac{1}{2}\hat{A}\hat{D}-\frac{1}{2}HW\right) \tag{2}$$

$$HW=[H_s^lW_s^l;H_a^lW_a^l] \tag{3}$$

其中, $W_s^l$和$W_a^l$分别是第$l$层结构特征和属性特征的权重矩阵; $\hat{A}=A+I$为带自环权重矩阵, $I$为单位矩阵; $\hat{D}$为$\hat{A}$的对角节点度矩阵; [;]表示两个矩阵的串联.

为了控制 GCN 网络中的信息传播, 采用了逐层高速门 (highway gate) 机制, 写成以下函数$T$的形式:

$$T(H^l)=\sigma(H^lW^l+b^l) \tag{4}$$

$$H^{l+1}=T\left(H^l\right)\odot H^{l+1}+\left(1-T\left(H^l\right)\right)\odot H^l \tag{5}$$

其中, $H^l$是第$l$层的输出, $\odot$是点乘元素.

(2) 连接矩阵计算

实体关系的指向对两个实体等效性评估也有较大影响, 因此为了连接两个实体需要统计两种关系$R_i\rightarrow E$和$R_j\leftarrow E$, 为了连接两个实体, 让$r_{ij}$表示比对信息从第$i$个实体传播到第$j$个实体的程度:

$$C_{\text{head}}=\frac{\#count\_head\_of\_r}{\#count\_all\_of\_r} \tag{6}$$

$$C_{\text{tail}}=\frac{\#count\_tail\_of\_r}{\#count\_all\_of\_r} \tag{7}$$

$$r_{ij}=\sum_{(e_i,r,e_j)\in KG}C_{\text{head}}+\sum_{(e_j,r,e_i)\in KG}C_{\text{tail}} \tag{8}$$

其中, $\#count_all_of_r$是关系$r$的所有三元组的数量; $\#count_head_of_r$是连接关系$r$头实体的数量; $\#count_tail_of_r$是连接关系$r$尾实体的数量; $(e_i,r,e_j)$代表关系$r$的三元组, $e_i$是头实体, $e_j$是尾实体.

(3) 卷积训练

使用预对齐的实体$S$来训练 GCN 模型. 模型训练是通过最小化以下损失函数$L$:

$$F(s,t)=f(h(s),h(t)) \tag{9}$$

$$L=\sum_{(s,t)\in S}\sum_{(s',t')\in S'}\max[F(s,t)+\gamma-F(s',t'),0] \tag{10}$$

其中, $f(a,b)=\|a-b\|_1$, $h(\cdot)$表示实体嵌入, $(s',t')\in S'$表示通过破坏$(s,t)$构造的负实体对齐集合, 即将$s$和$t$替换为$KG_1$或$KG_2$中随机选择的实体; $\gamma$是分隔正负实体对齐的边际超参数. 采用随机梯度下降法来最小化上述损失函数, 获得实体的结构嵌入$h_a(\cdot)$和属性嵌入$h_s(\cdot)$.

2.2.3 实体对齐模块

通过主题模型训练后, 可以得到实体属性主题的概率分布, 然而这些概率分布并不能直接转化为实体的主题概率分布. 鉴于属性对于实体具有代表性, 通过计算属性的主题概率分布的数学期望来推断实体的主题分布. 具体而言, 属性$t^a$的期望主题分布$E(t^a)$可以通过以下公式计算:

$$E(t^a)=\sum P(t^a\mid z)P(z) \tag{11}$$

其中, $P(t^a \mid z)$表示属性$t^a$属于主题$z$的条件概率, 而$P(z)$是主题$z$的后验概率.

为了进一步揭示实体之间的语义关联, 并进行实体相似度的比较, 结合 Word2Vec 词向量模型来训练实体属性的向量表示, 并引入属性主题概率的期望值作为权重. 实体的向量表征$V$定义为:

$$V = Vec(t^a) = E(t^a) \cdot Word2Vec(t^a) \tag{12}$$

通过余弦相似度, 可以筛选出与源实体主题向量相似度高的候选目标实体, 其计算如下:

$$D_{s,t} = Similarity(V_1, V_2) = \frac{V_1 \times V_2}{\|V_1\| \times \|V_2\|} \tag{13}$$

使用实体的主题筛选出对齐的实体, 为每个源实体选择一个候选集合. 通过这种方式, 可以在对齐阶段缩小源实体和目标实体之间的距离, 从而提升对齐效率.

通过 GCN 卷积嵌入模块分别得到实体的结构嵌入$h_a(\cdot)$和属性嵌入$h_s(\cdot)$, 结合实体主题相似度, 因此, 对于知识图谱$KG_1$中的源实体$s$和知识图谱$KG_2$中的目标实体$t$, 计算它们之间的以下距离度量, 即实体之间的相似度表示:

$$H_s = \frac{f(h_s(s), h_s(t))}{d_s} \tag{14}$$

$$H_a = \frac{f(h_a(s), h_a(t))}{d_a} \tag{15}$$

$$D(s,t) = (1-x)(yH_s + (1-y)H_a) + xD_{s,t} \tag{16}$$

其中, $x$和$y$作为平衡因子, 分别代表实体 GCN 嵌入和主题向量的重要性指标. $d_s$和$d_a$是结构嵌入和属性嵌入的维数. 对齐过程如算法 2 所示.

算法 2. 实体对齐过程

输入: 结构嵌入向量$H_s$; 属性嵌入向量$H_a$; 主题相似度向量$T$.
输出: 实体对齐结果.

1) 对于每个实体
2) 计算实体的结构嵌入和属性嵌入
3) 组合结构嵌入和属性嵌入
4) 计算并添加主题相似度
5) 获得相似度
6) 结束此次循环

## 3 实验结果与分析

### 3.1 数据集合

在实验中, 使用 DBP15K[16]和 WK3L-15K[17]作为数据集. DBP15K 包含 3 个基于 DBpedia 多语言版本构建的跨语言数据集: DBP15K (zh_en) (汉译英)、DBP15K (ja_en) (日译英)、DBP15K (fr_en) (法译英). WK3L-15K 包含两个基于 Wikidata 多语言版本构建的跨语言数据集: WK3L-15K (fr_en) (法译英)、WK3L-15K (de_en) (德译英). 与大多数使用这些数据集的算法一样, 使这些数据对中的 30% 作为训练数据, 剩下的作为测试数据. 表 1 和表 2 分别列出了 DBP15K 及 WK3L-15K 数据集的详细统计数据.

表 1 DBP15K 数据集分布

| DBP15K | Entities | Relations | Triples |
|---|---|---|---|
| zh_en | Chinese | 66 469 | 2 830 |
| | English | 98 125 | 2 317 |
| ja_en | Japanese | 65 744 | 2 043 |
| | English | 95 680 | 2 096 |
| fr_en | French | 66 858 | 1 379 |
| | English | 105 889 | 2 209 |

表 2 WK3L-15K 数据集分布

| DBP15K | Entities | Relations | Triples |
|---|---|---|---|
| en_fr | English | 15 170 | 2 228 |
| | French | 15 393 | 2 422 |
| en_de | English | 15 127 | 1 841 |
| | German | 14 603 | 596 |

### 3.2 实验参数

在表 3 中列出 EAGT 方法的超参数及对应的取值. 此外, 实验在基于 Ubuntu 18.04 LTS 操作系统的个人工作站上执行. 该工作站配备了 Intel Xeon(R) CPU E5-2640 v3 处理器, 系统内存为 128 GB, 显卡型号为 NVIDIA GeForce RTX 2080Ti. 为了确保实验的公平性, 所有编码工作均采用 Python 3.6.5 版本完成, 而深度学习模型的训练则依赖于 TensorFlow 1.19.1 库.

表 3 超参数选取

| 名称 | 释义 | 取值 |
|---|---|---|
| $\alpha$ | 文档主题超参数 | 0.1 |
| $\beta$ | 属性主题超参数 | 0.8 |
| $x$ | 主题特征占比重 | 0.4 |
| $y$ | 结构特征占比重 | 0.9 |

### 3.3 评估指标

使用 *Hit*@1 和 *Hit*@10 作为评估指标来评估 EAGT 的性能. *Hit*@*k* 排名前$k$的正确比对比例, 计算公式为:

$$Hit@k = \frac{card(\{e_s \in Sour_{\text{test}} \mid rank_{e_s} \leqslant k\})}{card(Sour_{\text{test}})} \tag{17}$$

其中, $rank_{e_s}$是实体$e_s$的真实目标的排名, $Sour_{\text{test}}$是测试集合. *Hit*@*k*值越大, 证明模型的效果越好.

### 3.4 基准模型

(1) MTransE[30]模型: MTransE 是 TransE[24]模型的一个改进版本, 旨在通过将实体和关系映射到低维连续向量空间中, 来实现实体间的对齐.

(2) JAPE[31]模型: JAPE (joint attribute preserving embedding) 通过结合关系三元组和属性信息来进行实体嵌入.

(3) GCN[16]模型: GCN-Align 利用实体的结构特征和属性特征作为输入, 通过图神经网络框架来处理和学习实体间的关系.

(4) HGCN[18]模型: HGCN 在传统 GCN 的基础上引入了高速门机制, 该机制能够更有效地捕捉实体间的复杂交互.

(5) DATTI[32]模型: DATTI 通过对知识图谱中的邻接信息和实体本身信息来增强实体对齐的解码过程, 该实体对齐解码算法基于三阶张量同构的方式.

(6) NAMN[33]模型: NAMN 采用分层的思想分别处理每跳邻居信息, 通过门控机制进行聚会以学习图结构的表征.

### 3.5 结果与分析

#### 3.5.1 问题设置

为了证明 EAGT 方法的有效性, 提出以下 4 个研究问题.

(1) 相比于现有的基准方法, EAGT 是否能取得更好的对齐效果?

(2) 不同数量的候选集合和主题数目对 EAGT 的影响?

(3) 不同主题模型对属性集合训练的效果影响?

(4) 不同的超参数取值对 EAGT 的影响?

#### 3.5.2 有效性分析

表 4 展示了 EAGT 模型与其他基线模型在 DBP15K 数据集上的对比结果, 表 5 展示了 EAGT 模型与其他基线模型在 WK3L-15K 数据集上的对比结果. 其中, “*”表示模型按照源文献实验复现的结果.

表 4 DBP15K 数据集上对齐效果 (%)

| 模型 | | zh_en | | ja_en | | fr_en | |
|---|---|---|---|---|---|---|---|
| | | *Hit*@1 | *Hit*@10 | *Hit*@1 | *Hit*@10 | *Hit*@1 | *Hit*@10 |
| Trans | MTransE | 30.83 | 61.41 | 27.86 | 57.45 | 24.41 | 55.55 |
| | JAPE | 41.18 | 74.46 | 36.25 | 68.50 | 32.39 | 66.68 |
| GCN | GCN | 41.25 | 74.38 | 39.91 | 74.46 | 37.29 | 74.49 |
| | HGCN | 72.03 | 85.70 | 76.62 | 89.73 | 89.16 | 96.11 |
| | DATTI | 83.50 | 95.30 | 80.60 | 96.90 | 87.30 | 97.90 |
| | NAMN | 76.80 | 89.40 | 79.20 | 93.60 | 92.90 | 97.40 |
| Ours | EAGT | **78.49** | **90.32** | **82.43** | **92.01** | **83.52** | **92.53** |
| 模型 | | en_zh | | en_ja | | en_fr | |
| | | *Hit*@1 | *Hit*@10 | *Hit*@1 | *Hit*@10 | *Hit*@1 | *Hit*@10 |
| Trans | MTransE | 24.78 | 52.42 | 23.72 | 49.92 | 21.26 | 50.60 |
| | JAPE | 40.15 | 71.05 | 38.37 | 67.27 | 32.97 | 65.91 |
| GCN | GCN | 36.49 | 69.94 | 38.42 | 71.81 | 36.77 | 73.06 |
| | HGCN | 71.45* | 83.73* | 72.52* | 86.81* | 82.69* | 91.22* |
| | DATTI | 80.83* | 92.92* | 78.34* | 93.34* | 83.28* | 96.77* |
| | NAMN | 70.23 | 86.26 | 77.46 | 91.52 | 89.26 | 97.42 |
| Ours | EAGT | **77.94** | **89.27** | **81.36** | **89.61** | **82.59** | **92.15** |

表 5 WK3L-15K 数据集上对齐效果 (%)*

| 模型 | | en_fr | | en_de | | en_fr | | en_de | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Hit*@1 | *Hit*@10 | *Hit*@1 | *Hit*@1 | *Hit*@1 | *Hit*@10 | *Hit*@1 | *Hit*@1 |
| Trans | MTransE | 19.45 | 38.37 | 5.65 | 8.56 | 16.54 | 26.56 | 7.21 | 11.58 |
| | JAPE | 17.31 | 36.53 | 13.80 | 29.29 | 16.59 | 28.39 | 16.72 | 35.27 |
| GCN | GCN | 18.44 | 37.45 | 15.23 | 34.23 | 17.19 | 33.42 | 19.31 | 37.84 |
| | HGCN | 50.54 | 66.28 | 51.51 | 64.76 | 46.36 | 60.32 | 48.63 | 65.32 |
| | DATTI | 53.32 | 70.73 | 55.85 | 71.39 | 52.74 | 69.06 | 52.64 | 69.38 |
| Ours | EAGT | **56.48** | **74.38** | **57.81** | **73.08** | **48.54** | **73.66** | **55.70** | **74.51** |

(1) 性能对比: 表 4 与表 5 的结果显示, 与所有基准指标相比, EAGT 在所选的下游任务数据集上的对齐效果更好. 在 DBP15K (zh_en)、DBP15K (ja_en)、WK3L-15K (fr_en)、WK3L-15K (de_en) 等任务中, 与传统基于

GCN 的实体对齐方法, EAGT 的对齐效果在对齐指标 *Hit*@1 上提升了 8.97%、7.58%、11.75%、12.23%.

(2) 结果分析: 根据表 4 与表 5 的结果显示, 表明主题模块在寻找特定源实体的真实目标时确实过滤了具有不同全局语义的实体. 其中, DATTI 对实体对齐的解码过程缺乏有效的改进, 而 NAMN 未能考虑关系信息和属性信息的作用. 主题模型能够从目标实体集中找到与原实体主题分布相似的候选实体. 然而在 DBP15K 中的 FR_EN 数据集合上的表现不如基线模型明显不如基准模型, 实体命中率反而下降了. 出现该情况的原因是这个数据集合实体之间结构上关联度较高, 其相邻节点提供了较多的对齐信息, 而加入主题过滤后, 使得实体间的结构关联程度降低, 造成真正对齐目标实体排名后移, 目标实体的命中率降低.

(3) 案例分析: 以 DBP15K 数据集合为例, 仅通过 GCN 嵌入难以实现一些实体的对齐, 结合 BTM 有效提升实体对齐的效果.

如图 3 (S1: 圣巴特里克竞技足球俱乐部; T1、MT1: St_Patrick's_Athletic_F.C.; M1: Kevin_ Doyle; S2: 安达鲁西亚; T2、MT2: Andalusia; M2: Fernando_ Llorente) 所示, 使用 PCA 选择两个实体映射到二维平面空间, 在仅通过 GCN 嵌入情况下, 源实体嵌入 S (source entity) 和错误的目标实体嵌入 M (misaligned entity) 实现了对齐, 而真正的目标实体嵌入经相似度比较则是 MT (target entity). 在加入主题模型后, 源实体嵌入 S 和真正的目标实体嵌入 T (target entity) 实现对齐.

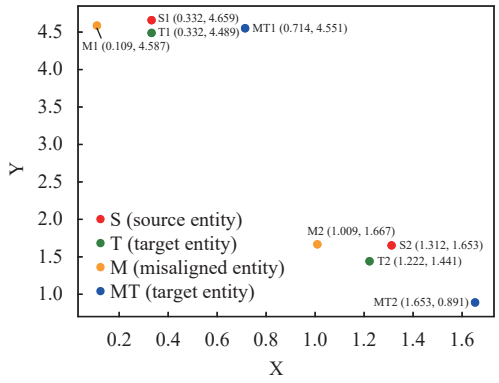


图 3　BTM 影响下的实体对齐情况

图 4 展示了实体圣巴特里克竞技足球俱乐部 (St_Patrick's_Athletic_F.C.) 实体描述语言不同, 但其属性描述高度相似. 源实体圣巴特里克竞技足球俱乐部通过 GCN 进行嵌入获得排名第 1 的实体为 Kevin_ Doyle, 而真正目标实体 St_Patrick's_Athletic_F.C.则排名为 33. 从属性角度分析几乎没有证据支持圣巴特里克竞技足球俱乐部和 Kevin_Doyle 对齐 (仅有图 4 中 fullname 属性一致), 而圣巴特里克竞技足球俱乐部和 St_Patrick's_Athletic_F.C.的属性相似度很高 (如图 4 season, …, position 等属性一致). 因此结合主题模型可以使得圣巴特里克竞技足球俱乐部和真正的目标实体 St_Patrick's_Athletic_F.C.对齐, 这解释了主题模型在实体对齐效果上的有效性.

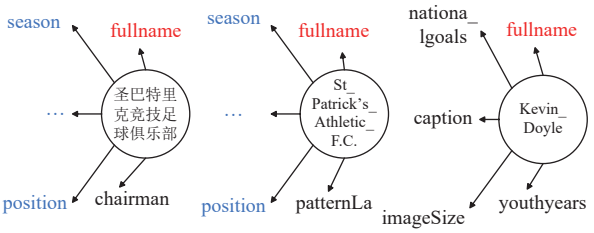


图 4　圣巴特里克竞技足球俱乐部、St_Patrick's_ Athletic_F.C.和 Kevin_Doyle 的属性

### 3.5.3　候选集和主题数目分析

由于主题数目的设定和候选集集合大小的都会对结果产生一定的影响, 因此如何确定主题数目以及候选集和的大小是其中关键的问题. 如果候选集合太大, 可能无法达到预期的筛选效果; 如果主题数目太小, 可能无法将实体类别区分开. 因此在 DBP15K 数据集上测试了不同主题和候选集大小对实体对齐效果的影响, 如图 5 和图 6 所示.

为测试候选集大小和主题数目对实体对齐效果上的影响, 设置 5 个分档 $k$=10、20、30、40、50 来测试主题数目和候选集数目在 1 000、2 000、3 000、4 000、5 000 种选择候选集数目. 通过图 5 和图 6 结果, 可以观察到主题数 30 和候选集合 3 000 的时候效果最好. 当主题数目太小, 主题模型无法充分学习到属性之间的关系; 主题数目太大, 会导致属性之间的关联度过低. 随着候选集合数目的增加, 目标实体被纳入后候选集合, 但是候选集合数目过大会导致筛选的效果降低, 无法做到区分效果.

### 3.5.4　超参数分析

为了探究不同超参数的取值对 EAGT 的性能影响, 在保持其他参数不变的情况下, 通过多次实验调整模型的可调参数, 以进行结果分析.

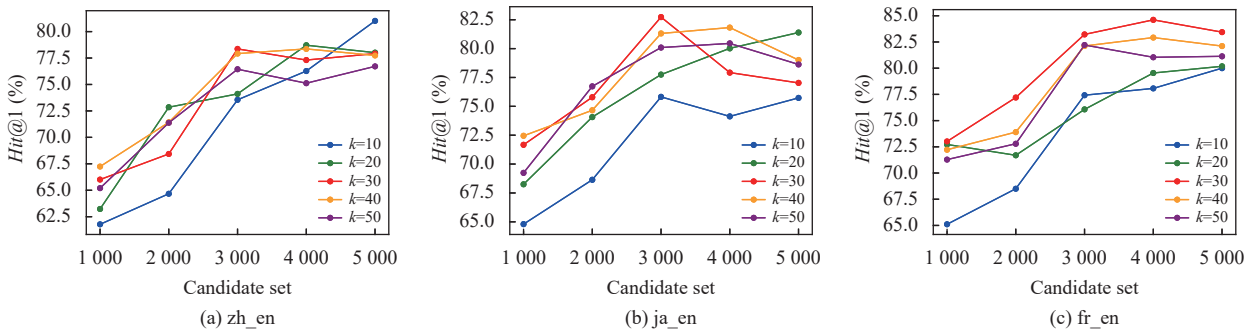


图 5 不同主题数和等待集数对 *Hit*@1 的影响

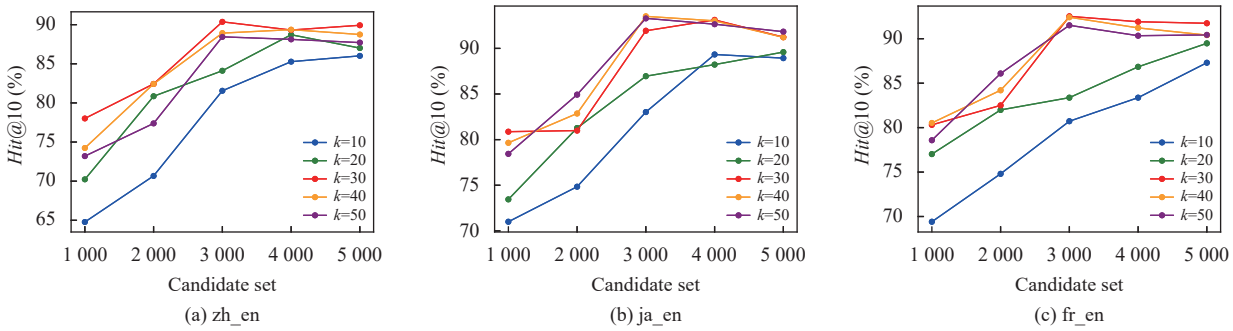


图 6 不同主题数和等待集数对 *Hit*@10 的影响

结果如表 6 所示, 实验组除当前参数取值为实验组数值, 其余参数均取对照组数值的对齐结果. 结果表明, 在 EAGT 中, 可以通过增大$\alpha$减小$\beta$来优化 BTM 对实体主题词的提取, 以降低主题之间的相关度并提高每个主题内部的聚合度. 通过对参数$x$进行调整, 获取最佳主题特征占比重, 对参数$y$进行调整, 获取最佳结构特征占比重, 使得对齐效果达到最佳状态.

### 3.5.5 主题模型分析

文本主题提取方法种类繁多, 其中概率主题模型被广泛应用. 然而, 不同的概率主题模型在不同语料库上的效果各异. 为了分析不同主题模型对 EAGT 的性能影响, 额外考虑了 LDA 主题模型作对比分析.

实验结果如表 7 所示, 以 DBP15K 数据集的 zh_en 为例, 在完成 LDA 与 BTM 模型的训练后, 对提取得到的主题词进行分析. 表 7 中, 目标实体以加粗字体表示, 相较于 LDA, BTM 模型命中的目标实体词汇数量更为可观, 并且对于非目标实体而言, BTM 所提取的词汇与之相似度更高. 表明 BTM 在目标实体搜索名字排名上优于 LDA.

表 6 DBP15K (zh_en) 超参数及其在实证研究中的价值 (%)

| 名称 | 释义 | 分组 | 取值 | *Hit*@1 | *Hit*@10 |
|---|---|---|---|---|---|
| $\alpha$ | 文档主题超参数 | 实验组 | 0.2 | 75.32 | 85.44 |
| | | 对照组 | 0.1 | 78.49 | 90.32 |
| $\beta$ | 属性主题超参数 | 实验组 | 0.7 | 74.12 | 82.92 |
| | | 对照组 | 0.8 | 78.49 | 90.32 |
| $x$ | 主题特征占比重 | 实验组 | 0.5 | 73.85 | 80.38 |
| | | 对照组 | 0.4 | 78.49 | 90.32 |
| $y$ | 结构特征占比重 | 实验组 | 0.8 | 74.47 | 83.71 |
| | | 对照组 | 0.9 | 78.49 | 90.32 |

表 7 LDA 和 BTM 提取实体主题词

| Model | Topic No. | zh_Entity | en_Entity 1 | en_Entity 2 | en_Entity 3 | en_Entity 4 |
|---|---|---|---|---|---|---|
| LDA | Topic 1 | **Jakarta** | Kent | Nagoya | Yangon | Mandalay |
| | Topic 2 | **LabVIEW** | JavaScript | BlackBerry_OS | GNU_Hurd | VP8 |
| | Topic 3 | **Sanming** | Tangshan | **Sanming** | Huai'an | Meizhou |
| | Topic 4 | **Gigi_Lai** | Candy_Chang | Florence_Kwok | Ruby_Lin | Kate_Tsui |
| | Topic 5 | **Shandong_U** | Nankai_U | Hunan_U | Duke_U | **Shandong_U** |
| BTM | Topic 1 | **Jakarta** | Mandalay | Malacca | **Jakarta** | Tokyo |
| | Topic 2 | **LabVIEW** | IW_engine | JetBrains | Android_Wear | **LabVIEW** |
| | Topic 3 | **Sanming** | **Sanming** | Meizhou | Tangshan | Lu'an |
| | Topic 4 | **Gigi_Lai** | Sunny_Chan | Florence_Kwok | **Gigi_Lai** | Blackie_Chen |
| | Topic 5 | **Shandong_U** | Fudan_U | **Shandong_U** | Sichuan_U | Hunan_U |

## 4 结论与展望

EAGT 是一种新颖的实体主题融合图卷积神经网络的实体对齐框架. 该框架利用图卷积神经网络对两个待对齐的知识图谱中的实体进行嵌入表示, 以捕获实体的结构信息, 并引入主题模型来学习源实体和目标实体之间的主题表征. 通过结合 GCN 得到的嵌入和 BTM 得到的主题嵌入, 实现对源实体和目标实体的对齐. 在 DBP15K 和 WK3L-15K 等真实公开数据集上的实验结果表明, 引入主题模型显著提升了实体对齐的性能. 设置主题数 30 和候选集 3 000 效果对齐效果最佳, 较小的主题数目, 主题模型无法得到充分的学习, 较大的候选集数量, 无法体现主题模型的区分效果. 未来研究将基于图卷积神经网络的方法考虑更多的特征信息, 如关系和属性注意力机制, 以进一步提升 EAGT 模型的性能和应用范围.

## 参考文献


1 Hogan A, Blomqvist E, Cochez M, *et al*. Knowledge graphs. ACM Computing Surveys (CSUR), 2022, 54(4): 71.

2 Zhao XJ, Jia Y, Li AP, *et al*. A survey of the research on multi-source knowledge fusion technology. Journal of Yunnan University: Natural Sciences Edition, 2020, 42(3): 459–473.

3 Bizer C, Lehmann J, Kobilarov G, *et al*. DBpedia—A crystallization point for the Web of data. Journal of Web Semantics, 2009, 7(3): 154–165. [doi: 10.1016/j.websem.2009.07.002]

4 Rebele T, Suchanek F, Hoffart J, *et al*. YAGO: A multilingual knowledge base from Wikipedia, WordNet, and GeoNames. Proceedings of the 15th International Semantic Web Conference on the Semantic Web (ISWC 2016). Kobe: Springer, 2016. 177–185.

5 Navigli R, Ponzetto SP. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 2012, 193: 217–250. [doi: 10.1016/j.artint.2012.07.001]

6 Sun ZQ, Zhang QH, Hu W, *et al*. A benchmarking study of embedding-based entity alignment for knowledge graphs. Proceedings of the VLDB Endowment, 2020, 13(12): 2326–2340. [doi: 10.14778/3407790.3407828]

7 Cohen WW, Richman J. Learning to match and cluster large high-dimensional data sets for data integration. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton: ACM, 2002. 475–480.

8 Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton: ACM, 2002. 269–278.

9 Arasu A, Götz M, Kaushik R. On active learning of record matching packages. Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis: ACM, 2010. 783–794.

10 Suchanek FM, Abiteboul S, Senellart P. Paris: Probabilistic alignment of relations, instances, and schema. Proceedings of the VLDB Endowment, 2011, 5(3): 157–168. [doi: 10.14778/2078331.2078332]

11 Lacoste-Julien S, Palla K, Davies A, *et al*. SIGMa: Simple greedy matching for aligning large knowledge bases. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago: ACM, 2013. 572–580.

12 Song DZ, Luo Y, Heflin J. Linking heterogeneous data in the semantic Web using scalable and domain-independent candidate selection. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 143–156. [doi: 10.1109/TKDE.2016.2606399]

13 张富, 杨琳艳, 李健伟, 等. 实体对齐研究综述. 计算机学报, 2022, 45(6): 1195–1225. [doi: 10.11897/SP.J.1016.2022.01195]

14 Zhu BB, Wang RL, Wang JY, *et al*. A survey: Knowledge graph entity alignment research based on graph embedding. Artificial Intelligence Review, 2024, 57(9): 229. [doi: 10.1007/s10462-024-10866-4]

15 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2016.

16 Wang ZC, Lv QS, Lan XH, *et al*. Cross-lingual knowledge graph alignment via graph convolutional networks. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 349–357.

17 Wu Y, Liu X, Feng YS, *et al*. Relation-aware entity alignment for heterogeneous knowledge graphs. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019. 5278–5284.

18 Wu YT, Liu X, Feng YS, *et al*. Jointly learning entity and relation representations for entity alignment. Proceedings of the 2019 Conference on Empirical Methods in Natural

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 240–249.

19 Tang XB, Zhang J, Chen B, *et al*. BERT-INT: A BERT-based interaction model for knowledge graph alignment. Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, 2020. 439.

20 Chen Z, Chen JY, Zhang W, *et al*. MEAformer: Multi-modal entity alignment Transformer for meta modality hybrid. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023. 3317–3327.

21 Zhu J, Huang CQ, De Meo P. DFMKE: A dual fusion multi-modal knowledge graph embedding framework for entity alignment. Information Fusion, 2023, 90: 111–119. [doi: 10.1016/j.inffus.2022.09.012]

22 Bai LY, Li N, Li GS, *et al*. Embedding-based entity alignment of cross-lingual temporal knowledge graphs. Neural Networks, 2024, 172: 106143. [doi: 10.1016/j.neunet.2024.106143]

23 Zhao Y, Wu YK, Cai XR, *et al*. From alignment to entailment: A unified textual entailment framework for entity alignment. Proceedings of the 2023 Findings of the ACL. Toronto: ACL, 2023. 8795–8806.

24 Bordes A, Usunier N, Garcia-Durán A, *et al*. Translating embeddings for modeling multi-relational data. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2787–2795.

25 Dettmers T, Minervini P, Stenetorp P, *et al*. Convolutional 2D knowledge graph embeddings. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: Association for the Advancement of Artificial Intelligence, 2018. 1811–1818.

26 Scarselli F, Gori M, Tsoi AC, *et al*. The graph neural network model. IEEE Transactions on Neural Networks, 2009, 20(1): 61–80. [doi: 10.1109/TNN.2008.2005605]

27 Munne RF, Ichise R. Entity alignment via summary and attribute embeddings. Logic Journal of the IGPL, 2023, 31(2): 314–324. [doi: 10.1093/jigpal/jzac021]

28 Yan XH, Guo JF, Lan YY, *et al*. A biterm topic model for short texts. Proceedings of the 22nd international conference on World Wide Web. Rio de Janeiro: ACM, 2013. 1445–1456.

29 Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 1532–1543.

30 Chen MH, Tian YT, Yang MH, *et al*. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: AAAI Press, 2017. 1511–1517.

31 Sun ZQ, Hu W, Li CK. Cross-lingual entity alignment via joint attribute-preserving embedding. Proceedings of the 16th International Semantic Web Conference. Vienna: Springer, 2017. 628–644.

32 Mao X, Ma MR, Yuan H, *et al*. An effective and efficient entity alignment decoding algorithm via third-order tensor isomorphism. Proceedings of the 60th Annual Meeting of the ACL (Vol. 1: Long Papers). Dublin: ACL, 2022. 5888–5898.

33 谭元珍, 李晓楠, 李冠宇. 基于邻域聚合的实体对齐方法. 计算机工程, 2022, 48(6): 65–72.

(校对责编: 张重毅)