Computer Science and Application 计算机科学与应用, 2023, 13(5), 1166-1177
Published Online May 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.135114

Hans汉斯

基于属性增强的对偶图实体对齐算法

姚　荣

成都信息工程大学软件工程学院，四川 成都

收稿日期：2023年4月27日；录用日期：2023年5月24日；发布日期：2023年5月31日

摘　要

本文提出结合属性信息的对偶图实体对齐算法针对基于关系感知的双对偶关系图算法中没有考虑到的属性信息进行优化，对属性结构嵌入向量使用图卷积神经网络算法对邻居节点抽取信息，并使用对偶关系图和注意力机制抽取实体对中的关系信息，最后通过结合实体对的关系信息和属性信息的相似度，判断是否为同一实体。针对原算法中识别效率不高的异构知识图谱实体对提升效果明显。在数据集DBP15K的三个跨语言数据集ZH-EN，JA-EN，FR-EN上实验，实验结果验证了对偶注意力以及属性信息对实体对齐方法的有效性。

关键词

属性增强，对偶关系图，实体对齐，知识图谱

Attribute Augmentation Based Alignment Algorithm for Pairs of Dyadic Graph Entity Alignment

Rong Yao

School of Software Engineering, Chengdu University of Information Engineering, Chengdu Sichuan

Received: Apr. 27th, 2023; accepted: May 24th, 2023; published: May 31st, 2023

Abstract

In this paper, we propose the dyadic graph entity alignment algorithm DAI combining attribute information to optimize the attribute information that is not considered in the dual dyadic relational graph algorithm RDGCN based on relationship awareness, use the graph convolutional neural network algorithm for attribute structure embedding vector to extract information from

文章引用: 姚荣. 基于属性增强的对偶图实体对齐算法[J]. 计算机科学与应用, 2023, 13(5): 1166-1177.
DOI: 10.12677/csa.2023.135114

neighbor nodes, and use the dyadic relational graph and attention mechanism to extract the relational information in entity pairs, and finally by combining the relational. The similarity of relationship information and attribute information of entity pairs is finally judged whether they are the same entity or not. The improvement effect is obvious for the heterogeneous knowledge mapping entity pairs which are not recognized efficiently in the original algorithm. Experiments on three cross-lingual datasets ZH-EN, JA-EN, and FR-EN of dataset DBP15K are conducted, and the experimental results verify the effectiveness of pairwise attention and attribute information on entity alignment methods.

Keywords

Attribute Augmentation, Dyadic Relationship Graphs, Entity Alignment, Knowledge Graphs

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

实体对齐是指对于异构数据源知识图谱中的多种实体，找到物理世界中的同一实体指代对象。随着信息技术的高速发展，网络资源也越来越多，知识图谱中的实体可以从网络百科页面抽取出实体，并对不同来源的实体进行对齐，构建高质量的异构知识图谱。实体对齐就是解决如何从各个来源的数据中找到同一实体，并对进行知识融合的答案。

传统的实体对齐主要是针对句法和实体结构采用相似度对齐的方法[1]，利用实体之间的标签或者字符距离映射到同一空间进行相似度对比。Bhamidipaty 等研究人员通过人工标注的方法来标记不方便对齐的实体[2]。Laccoste 等提出了 SiGMa 方法，该方法的本质是使用了贪心算法的思想，通过对实体属性和机构化信息进行局部搜索来完成实体对齐的任务[3]。Scharffe 等学者通过匹配模糊实体字符串，对单词的关系使用分类来进行实体对齐[4]。Bizer 等人通过使用规范的实体语法定义来进行对齐任务[5]。这类方法的缺点在于需要根据实体对齐任务有针对性地使用不同的相似度函数，因此会耗费大量的人力资源，同时极大地增加工作量。

随着知识图谱嵌入方法的快速发展，越来越多的研究人员开始使用知识图谱嵌入向量来完成实体对齐任务。这类方法主要思路是将知识图谱中实体和关系嵌入到同一个低维度的向量空间，并在该空间中通过向量来计算相似度系数[6] [7] [8]。翻译距离模型 TransE 首先被提出用于知识图谱的嵌入表达任务中，其表达方式是将每一个知识图谱三元组都表达成头实体向量和关系向量向尾实体向量的映射，但是该模型难以处理各种错综复杂的关系映射。于是，研究人员陆续提出一系列的翻译距离模型来进行改进[7]。IEAJKE 模型则是利用共同嵌入的方式来进行对齐后再用迭代训练优化模型，其目的在于增强模型的实体对齐效果[8]。Guan 等学者提出了 SEEA 模型，该模型使用自学习的方式来进行实体对齐[9]。Sun 和其他学者建立模型利用迭代的方式进行实体对齐，把实体对齐任务变成分类任务进行求解[10]。Sun 等人为了对多种特征进行利用，提出了一种实体对齐模型 JAPE，该模型从知识图谱的结构嵌入方式和属性嵌入方式这两个两方面来学习实体特征，进一步优化实体对齐的准确率[11]。AttrE 模型也是以知识图谱嵌入向量为基础，提出了一种融合实体结构信息与属性信息的方法[12]。He 等学者使用属性三元组来实现实体对齐与属性对齐的信息交互，从而产生大量优质对齐实体对[13]。这说明实体的属性信息包含了很多未被

挖掘的内容，但这些方法仍然不能充分地利用属性，知识图谱嵌入向量中还存在很多深层次特征，值得研究学者进一步研究。

2. 研究动机

RDGCN (Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs)基于关系感知双图卷机网络主要是通过构造对偶图来实现以关系作为顶点的对偶关系图，并通过对偶注意力机制来捕获原始的知识图谱和关系对偶图的之间的相互关系，然后通过多层的图神经网络来捕获邻居顶点的结构信息，增强感受野，然后再对通过最终表示结构来判断实体对中两个实体的相似度。

RDGCN 使用关系对偶图的目标是为了捕捉实体之间的关系表示，通过以关系为顶点的对偶图通过对偶注意力机制来加强实体对齐的可靠性，并通过对偶关系图的关系注意力机制和原始图的图注意力机制来共同加强关系结构的特征表达，最后将融入关系注意力机制的实体表达送入到多层的图卷积神经网络中，通过图卷积对齐的思想实现实体对齐，因此 RDGCN 模型是属于实体对齐分类中的基于图神经网络模型的实体对齐算法，其本质还是使用了图神经网络算法模型来进行实体对齐，并在此基础上通过对偶关系图和注意力机制捕获关系结构的表达。

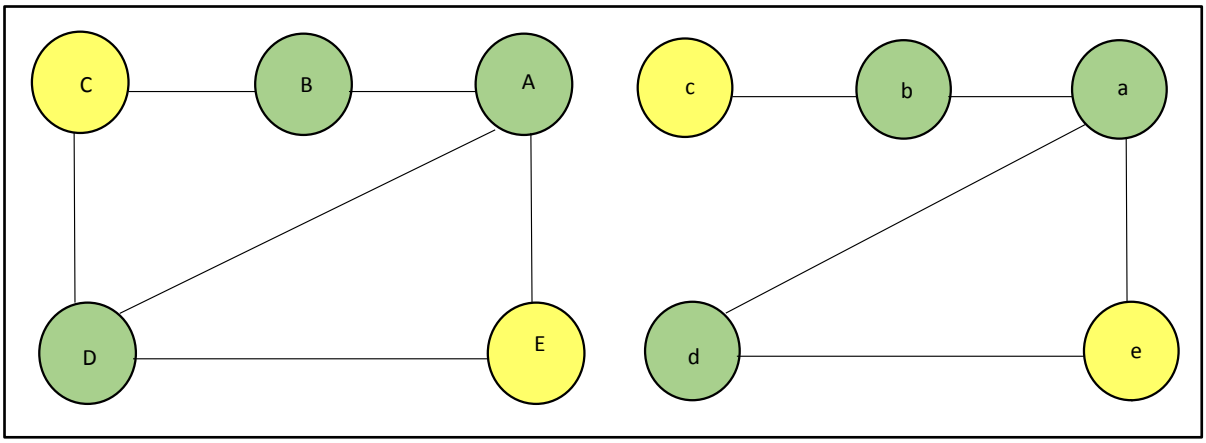

Figure 1. Isomorphic entity alignment schematic
图 1. 同构实体对齐示意图

因此本文在分析 RDGCN 模型时，实体对齐主要考虑在图卷积神经网络中实体嵌入向量的表达，如图 1 所示是多个需要待对齐的实体对，假设根据初始化种子条件知道，顶点 A 的实体与顶点 a 的实体已经对齐，顶点 B 的实体与顶点 b 的实体已经对齐和顶点 C 的实体与顶点 c 的实体已经对齐，用向量来表示就是 $V_A^0 = V_a^0$，$V_B^0 = V_b^0$，$V_D^0 = V_d^0$，其中 V_A^0 代表顶点 A 在经过图卷积聚合前的向量表示，而 V_A^1 代表顶点 A 在经过一次图卷积聚合操作之后的向量表示，由于通过训练种子得到已经对齐的实体对，因此也可以知道 $V_A^1 = V_a^1$，$V_B^1 = V_b^1$，$V_D^1 = V_d^1$，根据图神经网络的传播公式，如公式(1)可知

$$V^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}V^{(l)}W^{(l)}\right) \tag{1}$$

其中 $\tilde{A}$ 是邻接矩阵和单位矩阵的和矩阵，$\tilde{D}$ 是度矩阵，W 是权重矩阵，因此我们可以通过图神经网络的传播公式可以推断出各个层次间的实体向量表示关系，其中可知 $V_A^1 = \sigma\left(W^{(0)}\left(V_A^0 + V_B^0 + V_D^0 + V_E^0\right)/4\right)$，$V_B^1 = \sigma\left(W^{(0)}\left(V_A^0 + V_B^0 + V_C^0\right)/3\right)$，同理对齐实体中也存在 $V_a^1 = \sigma\left(W^{(0)}\left(V_a^0 + V_b^0 + V_d^0 + V_e^0\right)/4\right)$，$V_b^1 = \sigma\left(W^{(0)}\left(V_a^0 + V_b^0 + V_c^0\right)/3\right)$，因为实体对(A, a)，(B, b)和(D, d)已经完成了对齐，因此可知实体 E 和实

体 e 应该对齐，同时可知实体 C 和实体 c 应该完成对齐，因此可以通过这个实例看出，在部分实体已经对齐的情况下，如果待对齐的实体对邻居顶点的关系结构是同构的，那么可以使用图神经网络的方法来完成知识图谱的实体对齐。

同理，我们可以推断出，假如待对齐的实体对之间的邻居顶点关系图不是同构的，是否可以和上述方式一样的推断出实体对齐的实体对，如图 2 所示：

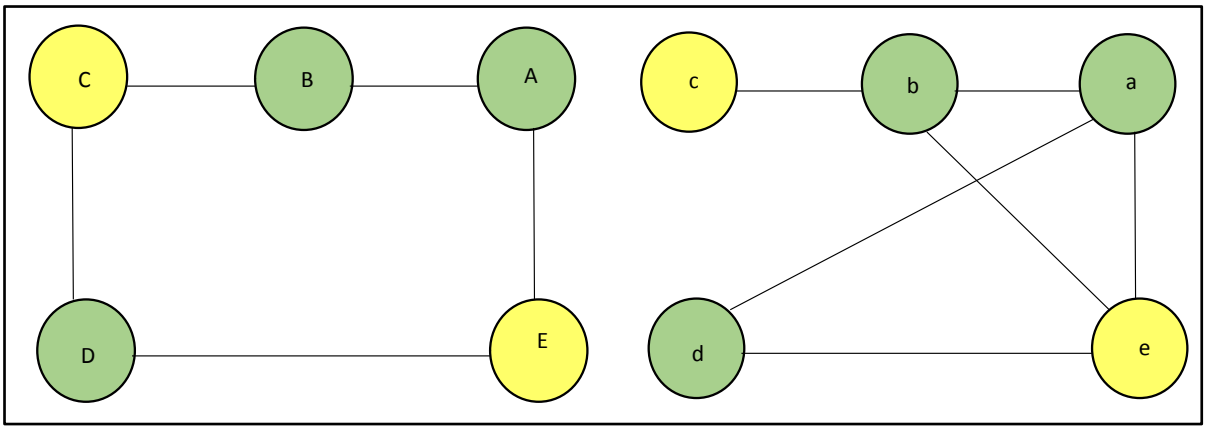

Figure 2. Heterogeneous entity alignment schematic
图 2. 异构实体对齐示意图

当前我们已经部分实体对已经对齐，实体 A，实体 B 和实体 D 与异构知识图实体 a，实体 b 和实体 d 已经对齐，但是由于此时知识图是异构的，因此我们可以通过公式发现 $V_A^1=\sigma\left(W^{(0)}\left(V_A^0+V_B^0+V_D^0+V_E^0\right)/4\right)$，$V_a^1=\sigma\left(W^{(0)}\left(V_a^0+V_b^0+V_d^0+V_e^0\right)/4\right)$，同理可以通过已对齐的实体知道 $V_A^0=V_a^0$，$V_B^0=V_b^0$，$V_D^0=V_d^0$ 和经过一次图卷积后的向量关系 $V_A^1=V_a^1$，$V_B^1=V_b^1$，$V_D^1=V_d^1$，不难得出异构知识图谱中实体 E 和实体 e 的实体对齐关系是相等的，但是对于待对齐实体对(C, c)就很难判断实体是否对齐，已知 $V_B^1=\sigma\left(W^{(0)}\left(V_A^0+V_B^0+V_C^0\right)/3\right)$，而此时 $V_b^1=\sigma\left(W^{(0)}\left(V_a^0+V_b^0+V_c^0+V_e^0\right)/4\right)$，通过实体对齐已知关系，我们并不能得出 $V_C^0=V_c^0$，因此直观上来看并不能直接得到实体对齐的信息，虽然因为实体对的邻居结构相似，向量表示的相似度较高，可能最后通过结果计算可以判断出实体对(C, c)存在一定的相似度，但依然可以证明，只通过关系结构相似度来判断实体对齐存在片面性不足之处。

因此本文考虑使用知识图谱中的属性结构来辅助关系结构对实体对齐进行判断。但是使用属性信息又会存在一定的矛盾和冲突，常见的属性三元组一般表示为(实体名称，属性，属性值)，但是知识图谱的属性值描述可能存在一定的差异性，因为属性值一般是使用字符类型进行存储的，一般是没有固定的格式规范的，例如(墨石公园景区，门票价格，60 元起)和(墨石公园，门票价格，成人票 60 元)，从中可以看出来虽然这两个属性三元组描述的事实约定是相同的，但是属性值不经过一定的约束和格式化，是无法进行使用的，因此本文考虑只使用属性信息来进行编码判断，不使用属性值进行约束。即构造属性图，两个实体之间不考虑关系之间的结构联系，只考虑属性之间的联系，两个实体之间如果有相同的属性，则在两个实体之间存在一条属性边，这条属性边的名称就是该属性的名称。如图 2 所示，使用到前面的关系结构图神经网络传播异构表示中，我们可以知道如果只使用关系结构很难知道实体 C 和实体 c 的实体对齐关系，但是我们引入了属性结构提取特征，可以提取到实体 C 和实体 c 中存在大量相同的属性，因此实体 C 和实体 c 的属性表征向量的相似度很高，并结合关系结构向量表征的相似度计算，就可以大大提高异构实体对的实体对齐相似度。

3. 本文算法

3.1. 算法整体框架

知识图谱 $KG=(E,R,T,A,V)$ 是一个有向图，其中包括所有实体集合 E、所有关系集合 R、所有三元组集合 T、所有实体属性 A 以及对应的属性值 V。给定源知识图谱 $KG_1=(E_1,R_1,T_1,A_1,V_1)$，目标知识图谱 $KG_2=(E_2,R_2,T_2,A_2,V_2)$ 以及已对齐实体对种子 $S=\{(u,v)|u\in E_1,v\in E_2,u\equiv v\}$，其中 $u\equiv v$ 代表实体 u 和实体 v 等价，也就是两个实体对指向的是现实世界中同一个物体，实体对齐任务的主要目的就是找到两个知识图谱中等价的实体对。如图 3 所示：

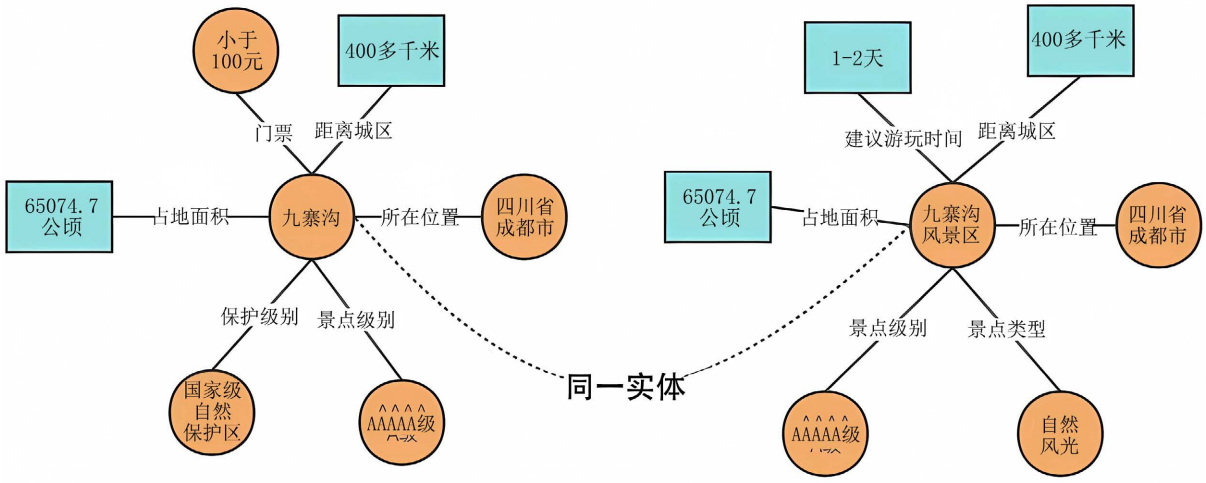

Figure 3. Entity alignment illustration
图 3. 实体对齐图示

本文根据 RDGCN 模型中没有对实体属性信息进行利用这一不足之处，提出了基于结合属性信息的对偶图实体对齐算法(Alignment Algorithm of Dual Graph Entities Combining Attribute Information, DAI)。DAI 算法使用知识图谱的对偶图构造出关系图，关系图的顶点表示关系，通过多个注意力机制使得关系对偶图和原知识图谱进行信息的交换，然后通过带有高速路神经网络门控的双层 GCN 来合并邻居结点的结构信息，扩大感受野。DAI 算法不仅使用对偶图来对关系信息进行提取，还使用了图卷积神经网络对实体的属性进行分析，将属性作为节点来及进行分析，得到属性结构信息的嵌入向量，最后再将关系嵌入向量和属性嵌入向量计算距离得分函数，得到实体是否一致的结果。该算法模型可以分为三个部分：1) 对偶图关系结构提取模块，主要是通过对偶图和注意力机制来计算实体的关系嵌入向量。2) 图卷积属性结构提取模块，主要是通过图卷积神经网络来计算实体的属性嵌入向量。3) 联合对齐模块，主要是通过关系提取模块和属性提取模块得到的结果来计算实体是否对齐。算法的模型示意图如图 4 所示。

3.2. 对偶图关系结构提取模块

对偶图关系结构提取模块主要是通过构造对偶关系图，然后通过注意力机制将关系图和原知识图谱进行信息交换，最后通过 RGCN 网络获取相邻节点的信息，并对干扰噪声信息过滤，将邻居信息聚合给实体向量。构造对偶关系的过程就是将源知识图谱 KG_1 和目标知识图谱 KG_2 放在一起作为输入知识图谱 KG_{merge}，其中 KG_{merge} 的实体和关系就是 KG_1 和 KG_2 的实体和关系的总和。构造对偶关系图 KG_{dual} 时，应该遵守以下规则：

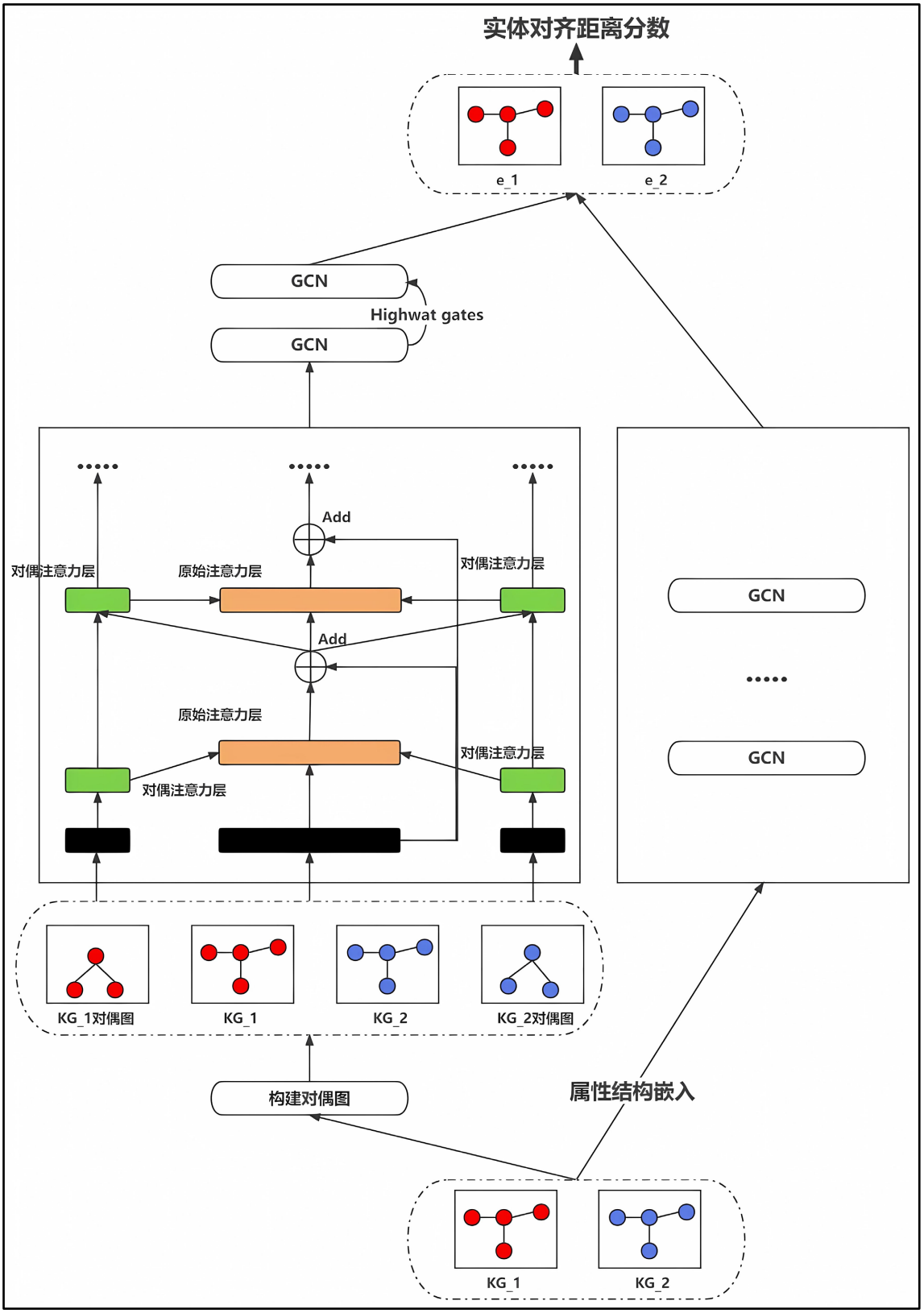

Figure 4. DAI model structure diagram
图 4. DAI 模型结构图

KG_{merge} 中的每一个关系在对偶图 KG_{dual} 中都有一个节点进行表示，因此 KG_{dual} 的节点集合就是 KG_{merge} 的关系集合。

如果在 KG_{merge} 中存在两种关系 r_i 和 r_j 共享同一个实体，那么在对偶图 KG_{dual} 中应该构建一条边 u_{ij}^r 来连接这两个关系节点，其中这条边 u_{ij}^r 的权重值 w_{ij}^r 计算公式如(2)~(4)，其中 H_i 和 T_i 分别对应的是关系 r_i 的头实体集合和尾实体集合。

$$w_{ij}^r = H\left(r_i, r_j\right) + T\left(r_i, r_j\right) \tag{2}$$

$$H\left(r_i, r_j\right) = \frac{H_i \cap H_j}{H_i \cup H_j} \tag{3}$$

$$T\left(r_i, r_j\right) = \frac{T_i \cap T_j}{T_i \cup T_j} \tag{4}$$

构造得到对偶图 KG_{dual} 后，与原知识图谱 KG_{merge} 一起作为输入数据，进行注意力机制的信息交换。在 DAI 模型算法中，主要使用了两种注意力机制算法，分别是针对对偶图的对偶注意力机制和针对知识图谱的原始注意力机制。

1) 对偶注意力机制

将关系信息引入对偶关系图中，把关系信息和原始图相结合进行关系特征描述，并采用统一的向量来进行表示。将图注意力机制用于选代地得到对偶关系图与原始图中节点间的向量表示，在该向量表示中注意力机制有利于促进关系图与对偶图之间的信息交互，其中包括两层注意力机制，分别是对偶注意力层与原始注意力层。

其中对偶注意力机制首先构造 $X^r \in R^{m\times 2d}$ 为输入知识图谱的对偶关系顶点表示矩阵，其中矩阵中的每一行都对应着对偶关系图中的一个关系顶点。利用原始注意力机制生成的节点向量特征来计算对偶注意力分数

$$\tilde{x}_i^r = \sigma^r\left(\sum\nolimits_{j\in N_i^r} \alpha_{ij}^r x_j^r\right) \tag{5}$$

$$\alpha_{ij}^r = \frac{\exp\left(\eta\left(w_{ij}^r a^r\left[c_i \| c_j\right]\right)\right)}{\sum_{k\in N_i^r}\exp\left(\eta\left(w_{ik}^r a^r\left[c_i \| c_k\right]\right)\right)} \tag{6}$$

在公式可以看到，$\tilde{x}_i^r$ 表示对偶关系顶点 v_i^r 处结合注意力分数的输出向量表示，x_j^r 表示在顶点 v_j^r 的对偶关系表示向量；N_i^r 表示顶点 v_i^r 的所有领域顶点集合；α_{ij}^r 表示对偶注意力机制的权重值，a^r 表示为全连接层，σ^r 是指非线性激活函数 ReLU，η 指的是非线性激活函数 Leaky ReLU；c_i 代表的是从原始注意力层获取的关系 r_i 的嵌入向量表现。通过对知识图谱中的平均头实体和尾实体近似表示 r_i 的关系为 c_i，c_i 由原始注意力层求得，其公式如(7)所示：

$$c_i = \left[\frac{\sum_{k\in H_i}\hat{x}_k^e}{\left|H_i\right|} \| \frac{\sum_{l\in T_i}\hat{x}_l^e}{\left|T_i\right|}\right] \tag{7}$$

$\hat{x}_k^e$ 和 $\hat{x}_l^e$ 指的是从原始注意力层中关系 r_i 第 k 个头节点和第 l 个尾节点的输出向量表示。

2) 原始注意力机制

本层是对原始知识图中应用图注意力机制，根据对偶关系图中的关系顶点表示来计算原始注意力得分 α_{qt}^e，其目的是通过对偶注意力层的关系向量表示来计算原始实体图顶点嵌入向量表示，如公式(8)所示。输入原始知识图中顶点的向量表示矩阵 $X^e \in R^{n\times d}$，原始知识图中实体 e_q 的嵌入向量表示 $\tilde{x}_q^e$ 的计算方式如

公式(9)所示：

$$\alpha_{qt}^{e}=\frac{\exp\left(\eta\left(a^{e}\left(\tilde{x}_{qt}^{r}\right)\right)\right)}{\sum_{k\in N_{q}^{e}}\exp\left(\eta\left(a^{e}\left(\tilde{x}_{qk}^{r}\right)\right)\right)} \tag{8}$$

$$\tilde{x}_{q}^{e}=\sigma^{e}\left(\sum_{t\in N_{q}^{e}}\alpha_{qt}^{e}x_{t}^{e}\right) \tag{9}$$

其中，$\tilde{x}_{qt}^{r}$ 代表的是对偶关系图中获得的实体 e_q 与实体 e_t 之间的关系 r_{qt} 的对偶表示，N_q^e 代表原始知识图中实体 e_q 的所有领域节点的集合，a^e 则是表示全连接层，σ^e 是原始注意力层的非线性激活函数。在该模型中，原始节点的初始表示向量矩阵是通过实体名来进行初始化的，这为实体对齐的有效性提供了重要依据。因此，实体嵌入向量的表示是通过叠加原始注意力层的输出向量和初始向量表示来进行修改的。

3) 合并结构信息

从 GCN 层的输出中收集最终的实体表示向量 $\bar{X}$，通过实体对之间的对齐距离作为得分函数，其公式如(10)所示，

$$D\left(e_1,e_2\right)=\left\|\bar{x}_{e_1}-\bar{x}_{e_2}\right\|_{L_1} \tag{10}$$

其中，$\bar{x}_{e_1}$ 表示知识图谱中实体 e_1 的嵌入向量表示，$\bar{x}_{e_2}$ 表示知识图谱中实体 e_2 的嵌入向量表示，$D\left(e_1,e_2\right)$ 则是代表两个实体 e_1 和 e_2 在 L_1 范数规则空间的距离。

3.3. 图卷积属性结构提取模块

属性结构提取模块是根据图卷积神经网络进行实体对齐的思想将实体属性进行单独提取分析。属性结构嵌入主要使用图卷积神经网络模型，构造实体属性矩阵，实体之间不考虑关系边的连接，将实体当作顶点，从而形成以顶点为实体，使用属性边连接起来的属性图，如果两个实体拥有同一个属性，那么这两个实体之间就有边进行连接，连接边称为属性边。因此在属性结构嵌入模块中只选用属性，不考虑属性值的影响，同时排除实体和实体之间的关系连接，因为实体之间的关系结构已经在关系结构模块进行了提取，因此在属性提取模块里，连接的是实体和属性的关系，实体相互间是没有关系边进行连接的，实体所有属性边的总数作为属性矩阵的维度。属性提取结构采用和关系结构相同的卷积过程，属性模块第一层的输入，用随机初始化的节点向量，其卷积过程如公式(11)所示，

$$H_{a}^{(l+1)}=\sigma\left(\hat{D}_{a}^{-\frac{1}{2}}\hat{A}_{a}\hat{D}_{a}H_{a}^{(l)}W_{a}^{(l)}\right) \tag{11}$$

其中，$\hat{A}_a\in R^{N_a\times N_a}$ 代表属性图的结构信息，$H_a^{(l)}\in R^{N_a\times N_a}$ 代表 l 层的特征向量；N_a 表示所有属性的数量；下标 a 表示属性结构的嵌入模块，$W_a^{(l)}$ 是 l 层的权重矩阵。

3.4. 联合对齐模块

通过上述两个提取模块计算出知识图谱的关系相似度，也计算完属性结构相似度后，还要合并整体计算出两个实体之间的相似度信息，首先计算出实体关系结构中实体对齐的结果，再计算实体属性结构中实体对齐的结果，分别得到两个提取模块的实体相似性后，然后通过对两个提取模块中实体相似度加权求和的方式，调节权重占比参数，以此设置关系结构与属性结构的比重大小，最后得到待对齐实体对之间的相似度大小，往往在实际数据训练中，因为属性边相似的实体会更多一些，因此属性提取相似度权重占比一般会比关系提取相似度权重占比小一些。公式(12)定义了实体相似度的得分函数，如下所示

$$D\left(e_i,v_j\right)=\alpha_{sa}\frac{f\left(h_s\left(e_i\right),h_s\left(v_j\right)\right)}{d_s}+\beta_{sa}\frac{f\left(h_a\left(e_i\right),h_a\left(v_j\right)\right)}{d_a} \tag{12}$$

其中，$f\left(x,y\right)=\left\|x-y\right\|_1$，代表计算实体之间的距离大小，$h_s$ 和 h_a 分别表示实体对的关系嵌入向量和属性嵌入向量，d_s 和 d_a 分别表示模型中关系嵌入向量和属性嵌入向量的维度大小。α_{sa} 和 β_{sa} 是平衡两种嵌入方式相似度的重要超参数。

4. 实验与分析

4.1. 数据集介绍

为了本文在实体对齐对比实验中使用的是 DBP15K 数据集(见表 1)，该数据集是由南京大学提出的一个用于实体对齐的数据集，包含多种不同翻译语言的知识图谱对齐实例，该数据集是从 DBpedia 的知识图谱中抽取出来的一个子集，DBpedia 是一个大规模的维基百科的语义网络，它包含着不同语言中丰富的语义信息，不同语言之间的对应都有链接。DBpedia 的中文、英文、日文和法文版本的子集是通过一定选择规则获取的。

Table 1. Data set DBP15K information
表 1. 数据集 DBP15K 信息

		实体	属性	关系三元组	属性三元组
ZH-EN	Chinese	66,469	8113	153,929	379,684
	English	98,125	7173	237,674	567,755
JA-EN	Japanese	65,744	5882	164,373	354,619
	English	95,680	6066	233,319	497,230
FR-EN	French	66,858	4547	192,191	528,665
	English	105,889	6422	278,590	576,543

4.2. 实验结果与分析

4.2.1. 预测结果对比

实验结果如表 2 所示，可以明显看出来经过改进后的结合属性信息的对偶图实体对齐算法 DAI 提升效果较为明显。

Table 2. Comparison results analysis of different entity alignment methods
表 2. 不同实体对齐方法对比结果分析

	ZH-EN			JA-EN			FR-EN		
	Hits@10	Hits@50	MRR	Hits@10	Hits@50	MRR	Hits@10	Hits@50	MRR
JAPE	74.46	88.90	0.490	68.50	85.35	0.476	66.68	83.19	0.430
GCN-Align	74.38	86.23	0.549	74.46	86.10	0.546	74.49	96.73	0.532
RDGCN	84.55	93.40	0.749	89.54	96.10	0.812	95.72	97.24	0.908
DAI	86.43	94.05	0.767	91.76	96.19	0.827	96.44	97.66	0.923

可以从表中的结果看到使用了属性结构提取特征的对偶图实体对齐算法 DAI 相比于原来的 RDGCN 算法有了不错的提升效果，由于属性提取特征采用的是属性名称的方式来对比的，因此实体之间的类型

种类越多，不同实体之间的包含属性差别越大，属性提取的比较效果越好。JAPE 模型同样也是使用关系结构特征和属性结构特征的模型，在关系结构特征提取中使用的是 TransE 模型来进行判别，而属性结构特征是使用 Skip-gram 模型，来获取不同属性之间的相关性。而 GCN-Align 模型是使用图神经网络来使用属性特征和关系特征进行实体对齐的模型，RDGCN 是使用对偶图来构造关系特征图的，并使用注意力价值加强关系特征提取的模型。图 5 是上述实体对齐方法的 Hits@K 可视化对比结果，主要是取了 Hits@10 和 Hits@50 的结果进行对比。

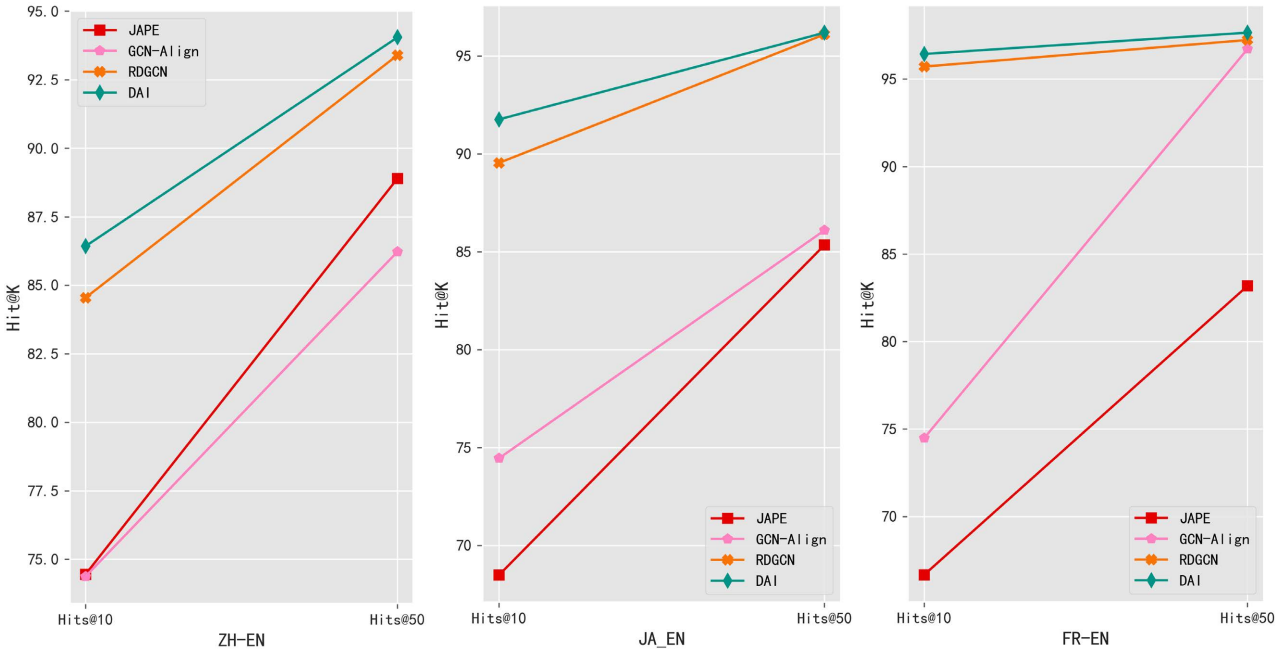

Figure 5. Experimental results of Hits@K comparison with different modeling approaches
图 5. 不同模型方法的 Hits@K 对比实验结果

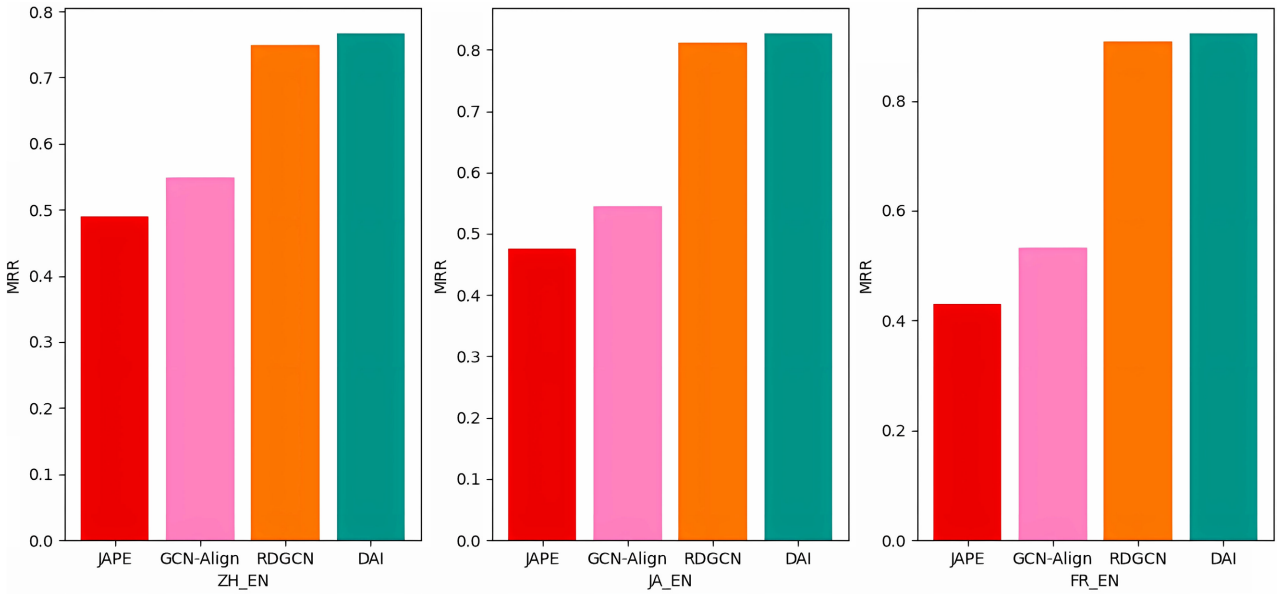

Figure 6. Experimental results of MRR comparison of different modeling approaches
图 6. 不同模型方法的 MRR 对比实验结果

可以从可视化的对比实验中看到改进后的 DAI 算法比 RDGCN 算法在模型中有不错的提升，尤其是在 Hits@10 的时候，提升程度较大；Hits@50 达到了一个瓶颈，提升程度并不是特别明显，但相较于原算法和其他的属性提取算法的结果仍有着不错的提升效果。

通过不同方法的 MRR 指标可视化展示，见图 6 可知，DAI 相比较于 ORGCN 在平均倒数秩上的差别不大，有略微提升，可能主要还是因为该数据集中的实体属性区分不大，多个不同实体中包含多种同样的属性关系，综合总体来分析可以看出结合属性信息的对偶图实体对齐算法具有不错的改进效果。

4.2.2. 超参数选择实验

本文将结合属性信息的对偶图实体对齐算法与其他同类型的算法进行对比实验，设置关系对偶注意力层和原始注意力迭代交互两轮，即进行两轮对偶关系信息的提取。第一轮交互主要是提取关系注意力的信息结构，减少原始注意力机制的叠加，因此混合向量权重参数为 0.1，第二轮交互增加原始注意力机制的占比，混合向量权重参数为 0.3。嵌入向量的维度为 300，学习率为 0.003，训练迭代次数为 600 次，关系结构提取参数为 0.9，属性结构提取参数为 0.1。

5. 发展与展望

本文使用的是基于属性增强的对偶图实体对齐算法，该算法通过属性增强信息能够辅助关系知识图谱的实体对齐，但是本算法只使用了实体的属性类型，没有使用属性值的具体信息，因为实体的属性值基本都是字符信息的输入，没有一定的规范化格式，在算法中不太好直接使用，以后的研究中希望能够将属性值的信息进行规范化，并与属性类型结合起来，共同辅助关系知识图谱的实体对齐。

参考文献

[1] Elfeky, M.G., Verykios, V.S. and Elmagarmid, A.K. (2002) TAILOR: A Record Linkage Toolbox. *Proceedings* 18*th International Conference on Data Engineering*, San Jose, CA, USA, 26 February 2002-1 March 2002, 17-28. https://doi.org/10.1109/ICDE.2002.994694

[2] Sarawagi, S. and Bhamidipaty, A. (2002) Interactive Deduplication Using Active Learning. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 23-26 July 2002, 269-278. https://doi.org/10.1145/775047.775087

[3] Lacoste-Julien, S., Palla, K., Davies, A., *et al.* (2013) SiGMa: Simple Greedy Matching for Aligning Large Knowledge Bases. *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL, 11-14 August 2013, 572-580. https://doi.org/10.1145/2487575.2487592

[4] Scharffe, F., Liu, Y. and Zhou, C. (2009) Rdf-Ai: An Architecture for Rdf Datasets Matching, Fusion and Interlink. https://www.researchgate.net/publication/228972747_Rdf-ai_an_architecture_for_rdf_datasets_matching_fusion_and_interlink

[5] Volz, J., Bizer, C., Gaedke, M., *et al.* (2009) Discovering and Maintaining Links on the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E. and Thirunarayan, K., eds., *The Semantic Web—ISWC* 2009, Springer, Berlin, 650-665. https://doi.org/10.1007/978-3-642-04930-9_41

[6] Hao, Y., Zhang, Y., He, S., *et al.* (2016) A Joint Embedding Method for Entity Alignment of Knowledge Bases. In: Chen, H., Ji, H., Sun, L., Wang, H., Qian, T. and Ruan, T., eds., *CCKS* 2016: *Knowledge Graph and Semantic Computing*: *Semantic*, *Knowledge*, *and Linked Big Data*, Springer, Singapore, 3-14. https://doi.org/10.1007/978-981-10-3168-7_1

[7] Bordes, A., Usunier, N., Garcia-Duran, A., *et al*. (2013) Translating Embeddings for Modeling Multi-Relational Data. https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf

[8] Lin, Y., Liu, Z., Sun, M., *et al.* (2015) Learning Entity and Relation Embeddings for Knowledge Graph Completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, **29**, 254-266. https://doi.org/10.1609/aaai.v29i1.9491

[9] Guan, S., Jin, X., Wang, Y., *et al.* (2019) Self-Learning and Embedding Based Entity Alignment. *Knowledge and Information Systems*, **59**, 361-386. https://doi.org/10.1007/s10115-018-1191-0

[10] Sun, Z., Wei, H., Zhang, Q., *et al.* (2018) Bootstrapping Entity Alignment with Knowledge Graph Embedding. *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, 13-19 July 2018,

4396-4402. https://doi.org/10.24963/ijcai.2018/611

[11] Sun, Z., Hu, W. and Li, C. (2017) Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding. In: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C. and Heflin, J., eds., *The Semantic Web—ISWC* 2017, Springer, Cham, 628-644. https://doi.org/10.1007/978-3-319-68288-4_37

[12] Trisedya, B.D., Qi, J. and Zhang, R. (2019) Entity Alignment between Knowledge Graphs Using Attribute Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 297-304. https://doi.org/10.1609/aaai.v33i01.3301297

[13] He, F., Li, Z., Qiang, Y., *et al.* (2019) Unsupervised Entity Alignment Using Attribute Triples and Relation Triples. In: Li, G., Yang, J., Gama, J., Natwichai, J. and Tong, Y., eds., *DASFAA* 2019: *Database Systems for Advanced Applications*, Springer, Cham, 367-382. https://doi.org/10.1007/978-3-030-18576-3_22