文章编号：1003-0077(2021)04-0035-09

基于多相似性度量和集合编码的属性对齐方法

伍家豪[1,2]，陈 波[1]，韩先培[1]，孙 乐[1]

（1. 中国科学院 软件研究所，北京 100190；
2. 中国科学院大学，北京 100049）

摘 要：属性对齐的目标是发现异构知识图谱中表示同一概念的属性之间的对应关系，是实现跨图谱知识融合的关键技术之一。现有模型通常利用基于规则和词嵌入的方法进行属性对齐，但这些方法仍存在以下两个问题：相似性度量不全面和属性实例信息未被充分利用。针对上述问题，该文提出了基于多相似性度量的属性对齐模型，通过多个角度设计相似性度量方法来获取属性间的相似性特征，并利用机器学习模型进行特征聚合。同时，为了充分利用属性的实例信息，在上述模型框架下提出了属性实例集合表示学习算法，通过将属性实例集合编码为向量来提取集合间的主题相似性，从而辅助属性对齐。在属性对齐数据集上的实验验证了模型的有效性，实验还表明，集合的表示学习算法能够有效捕捉属性实例的主题特征，并显著提升属性对齐结果。

关键词：属性对齐；表示学习；多相似性度量；集合编码

中图分类号：TP391 **文献标识码**：A

Attribute Alignment Based on Multi-Similarity Measure and Set Encoding

WU Jiahao[1,2], CHEN Bo[1], HAN Xianpei[1], SUN Le[1]

(1. Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: The goal of attribute alignment is to find the corresponding relationship which representing the same concept in heterogeneous knowledge graph. It is one of the key technologies to knowledge fusion. The existing models based on rules and word embedding are defected in incomplete similarity measurement and insufficient using of attribute instance information. To address this issue, this paper proposes an attribute alignment model based on multi similarity measures. We design similarity measures from multiple perspectives, and use machine learning model to aggregate this kind of features. At the same time, this paper proposes the attribute instance set representation learning algorithm. We extract the topic similarity between sets by encoding the attribute instance set as vectors, so as to assist attribute alignment. Experiments prove the validity of the model, and show that the set representation learning algorithm can effectively capture the subject feature of attribute instances and significantly improve the attribute alignment results.

Keywords: attribute alignment; representation learning; multi-similarity measures; set encoding

0 引言

随着计算机和通信技术的进步，许多垂直领域都形成了丰富的领域知识图谱。在知识图谱的构建过程中，设计者对相同事物描述的用词差异和形式化程度不同都会导致属性的异构。作为知识图谱中描述实体的框架，属性的异构会导致实体层面无法进行统一的描述，进而导致实体对齐、跨知识库检索等一系列任务无法很好地完成。属性对齐任务的目标是将不同知识图谱中表示同一概念的属性进行归并，从而解决知识图谱间的异构问题。在实际场景中，经常需要融合同一领域中的异构知识图谱，属性对齐作为数据融合[1]的基本操作之一就产生了丰富的应用需求。一个典型的案例是电子商务平台在发展过程中的数据融合，当一个平台的商品希望在另

收稿日期：2019-12-12 **定稿日期**：2020-01-12

一个平台进行销售时，需要融合两个平台的属性和目录来克服平台间的异构性，从而达到统一而高效的管理及检索。

目前属性对齐的主流方法通常综合利用属性名称和属性实例等信息来计算属性间的相似度，但这些方法中仍存在以下两个问题：①相似性度量不全面。属性对齐问题需要从多个角度综合考虑，但每种衡量属性相似度的方法都有一定侧重，从而导致相似性度量不全面的问题。例如，“宠物类型”和“猫品种”，当两个属性名称的描述词汇出现异构时，基于字符串的相似性度量就会失效。此时需要引入语义相似性的度量方法，同时语义相似性度量方法也会有其相应的局限性。②属性实例信息未被充分利用。在实际问题中，对属性的定义不准确会导致无法从属性的名称获取有效特征，例如，两个名称分别为“牛奶种类”和“脂肪含量”的属性，“牛奶种类”可以以脂肪含量为标准进行划分，两者的实例都在描述牛奶的脂肪含量，此时则难以从名称判断属性是否对齐，因此实例信息的利用对于属性对齐任务具有重要作用。针对属性对齐问题，如何从属性的实例集合中高效地提取特征仍然是一个挑战。现有的模型通过杰卡德相似系数(Jaccard index)、采样、词嵌入等方法从属性实例集合中提取特征，但以下两个原因导致现有模型在实际场景中的应用效果均不理想：一方面，数据集可能包含大量噪声；另一方面，两个描述相同领域的实例集合可能存在无交集的情况。例如，两个属性实例集合{“红色”，“黄色”，“橙色”，“lkg”}和{“绿色”，“蓝色”，“紫色”}，现有方法都无法很好地解决这些问题。

针对以上问题，本文提出了一种基于多相似性度量和集合编码的属性对齐模型，该模型综合考虑多方面的特征，解决了相似性度量不全面的问题。具体地，本文针对属性对齐问题设计了多种相似性度量方法，并利用这些方法来提取候选属性对的相似性特征，再利用机器学习模型对特征进行聚合，从而完成属性对齐。同时，为了充分利用属性的实例信息，本文在上述模型框架下提出了属性实例集合表示学习算法，利用深度学习的特征抽取能力来提取属性实例的主题特征，有效解决了属性实例集合无交集的问题。同时，模型中基于内容的注意力机制可以自动学习集合中元素的重要程度，解决了数据中的噪声问题。

本文在实际场景中的数据集(淘宝、天猫、盒马鲜生和某电商平台)上验证我们的模型。实验表明，我们的模型取得了最好的效果，并且验证了集合编码算法对于特征提取的有效性及其对属性对齐模型的提升作用。本文的主要贡献包括：①提出了基于多种相似性度量的属性对齐模型，并在实际任务中验证了模型的有效性；② 针对属性对齐问题，从属性名称和属性实例等通用信息出发设计了多种相似性度量方法；③提出了属性实例集合表示学习算法，有效地从集合中学习到属性实例的主题特征，并验证了该特征对于属性对齐任务的提升作用。

1 相关工作

1.1 属性对齐

目前的属性对齐方法按照使用信息的不同可以分为三类[2]：基于属性名称(schema-level)[3]、基于属性实例集合(instance-level)[4-5]以及前两者的结合[6]。其中基于属性名称的方法通常利用词典、规则匹配或词嵌入来计算两个属性名称的相似性。基于属性实例集合方法的基本思想是：属性的相关性与公共实例的重叠程度成正比。这里的挑战在于如何定义集合的重叠程度，以往的工作[7-8]利用杰卡德相似系数、词嵌入或采样技术[9]等方法来定义重叠程度。为了更好地定义属性实例集合在语义上的重叠程度，本文借鉴集合编码[10]的思想，提出了属性实例集合表示学习算法。我们将一个属性的所有实例看作一个集合，利用集合编码的方法提取属性实例集合的主题特征，从而判断两个集合的相似性。

1.2 本体匹配

本体匹配[11]问题与属性对齐问题相似，目前的本体匹配算法大致可以分为两类：单一匹配算法和基于组合方法的匹配算法。单一匹配算法通过设计一个强匹配器来完成本体匹配，例如，基于规则的匹配[12]、基于图算法的结构级匹配[13]、基于语料库的匹配[14]等。基于组合方法的匹配算法通过设计多个弱匹配器，充分利用本体中包含的所有信息进行匹配，并一直保持领先的效果，最具代表性的模型包括AML[15-17]、CroMatcher[18]等。对于这类组合算法，相似性度量方法的选择和聚合是两个巨大的挑战。在聚合多个特征时需要设置聚合权重，但人工设置的权重难以达到最好效果，另外，不同情况对应的最优权重可能不同。针对这一问题，CroMatcher模型可以根据匹配问题自动计算特征的聚合权重。

本文提出的模型同样基于多种相似性度量组合的思想，但本文利用机器学习模型进行特征聚合，避免了聚合权重的设定。2013 年后，一些工作[19-20]将预训练词向量引入本体匹配领域并取得了很好的效果。由于匹配问题的标注数据有限，基于大量训练数据的深度学习算法没能很好地应用于本体匹配领域，但仍有一些工作利用小规模训练数据使用机器学习算法进行本体匹配，例如，GLUE[21]、YAM＋＋[22]和 Rafcom[19]等。其中，Rafcom 使用了随机森林算法对匹配问题进行建模，并引入了预训练词向量。本文借鉴了本体匹配领域中方法的思想，针对属性对齐问题设计了多种相似性度量方法，并引入中文预训练词向量，利用机器学习的方法进行特征聚合。

2　任务定义

属性对齐的目标是发现异构知识图谱中表示同一概念的属性之间的对应关系，对齐的结果以映射形式呈现，并为上层应用提供统一的属性描述。

其形式化定义为：给定两个知识图谱 G_1 和 G_2，其中 G_1 包含属性 $\{A_1, A_2, \cdots, A_m\}$，$G_2$ 包含属性 $\{B_1, B_2, \cdots, B_n\}$，属性对齐的结果是一个映射集合 $f: A \rightarrow B$，映射集合中的元素 $A_i \rightarrow B_j$ 表示知识图谱 G_1 中的属性 A_i 与知识图谱 G_2 中的属性 B_j 指代相同，如图 1 所示。

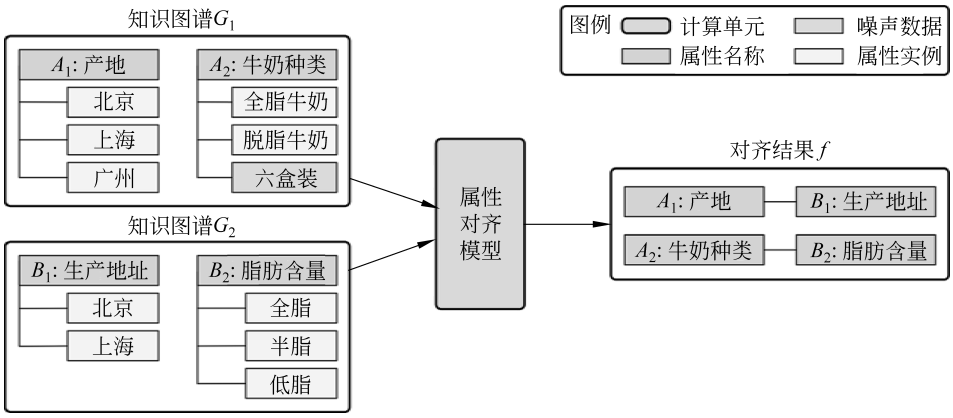

图 1　属性对齐任务示例

属性对齐问题可利用的元素主要有两类：第一类是属性的名称（也称属性项），对属性内容进行概括性定义，例如，食品类实体的属性中会包含“食品口味”“包装方式”等。第二类是属性的实例（也称属性值），同一个属性在不同实体中会对应不同的实例，属性的实例信息通常以短语、句子或数字符号的形式呈现，例如，属性“产地”可能对应“北京”“上海”“广州”等实例。

3　基于多相似性度量和集合编码的属性对齐模型

针对现有方法中相似性度量不全面的问题，本文提出了基于多种相似性度量的属性对齐模型，并针对属性对齐问题设计了多种相似性度量方法。本文将属性对齐问题建模为二分类问题，利用有监督的分类器模型判断候选属性对是否对齐。模型的架构如图 2 所示。

首先，在预处理阶段抽取得到两个异构知识图谱中所有属性的名称信息和实例信息，并根据数据特点利用分块技术（blocking）得到合适数量的候选属性对。然后，基于本文提出的相似性度量方法抽取每个候选属性对的相似性特征。对这些特征进行组合并做归一化等处理，使得所有特征具有相同的衡量尺度。最后，利用标注数据训练二分类模型来判断每个候选对是否对齐。模型每次输入一个候选对的所有相似性特征，输出 1 或 0 分别表示候选对对齐或不对齐。

由于本文设计的相似性特征从不同信息和角度出发，所以特征之间具有一定的互补关系。故本文使用了集成学习中的梯度提升决策树算法的变体 XGBoost[23]作为分类模型，可以充分利用特征之间的互补关系。同时 XGBoost 模型中对于缺失值的处理可以避免一些缺失数据或脏数据带来的问题。

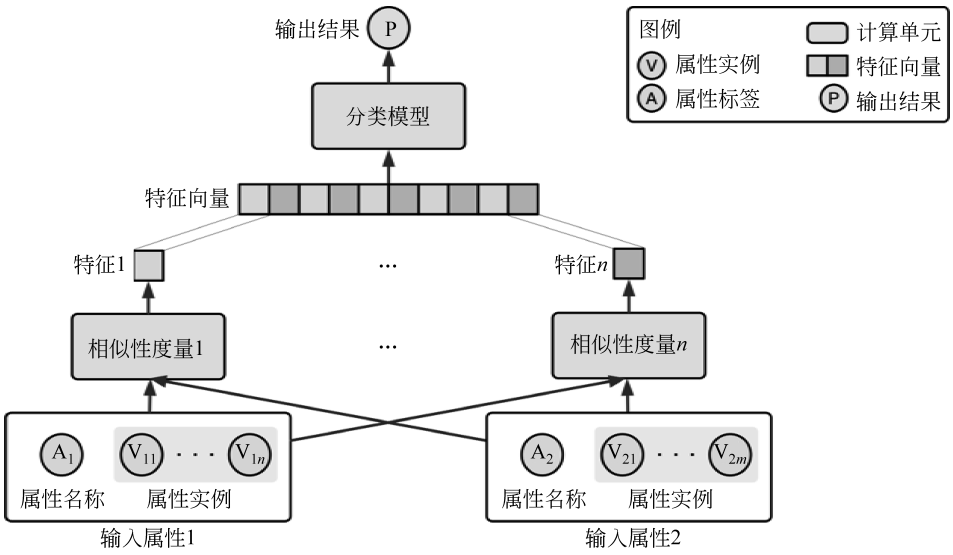

图 2 基于多相似性度量的属性对齐模型架构图

3.1 多种相似性度量的设计

本节主要介绍本文针对属性对齐问题设计的多种相似性度量方法。这些相似性度量方法利用属性名称和属性实例信息进行度量。所有相似性度量方法的概览如表 1 所示。

相似性度量方法 1 至方法 6 利用属性的名称信息进行度量，属性的名称是人为定义的对属性进行概括性描述的自然语言文本，可以非常直观地辅助属性对齐任务。针对每组候选属性对，可以获取这两个属性的名称对应的字符串。对于方法 4(杰卡德相似度 1)，将两个字符串 S_1 和 S_2 分别进行 2-gram 切分成为两个切分集合 Set_1 和 Set_2，例如，“生产地址”经过切分后成为{“生产”，“产地”，“地址”}。计算两个切分集合的杰卡德相似系数作为相似性度量。计算方法如式(1)所示。

$$J(\mathrm{Set}_1,\mathrm{Set}_2)=\frac{|\mathrm{Set}_1 \cap \mathrm{Set}_2|}{|\mathrm{Set}_1 \cup \mathrm{Set}_2|} \tag{1}$$

但由于属性名称文本较短以及式(1)在实际计算时可能出现分母为零和未归一化等情况，为了更好地衡量相似性特征，实际使用的变形公式如式(2)所示。

$$J(\mathrm{Set}_1,\mathrm{Set}_2)=\frac{|\mathrm{Set}_1 \cap \mathrm{Set}_2|}{\max(\min(|S_1|,|S_2|)-n+1,1)} \tag{2}$$

其中，n 为 n-gram 的切分长度。对于方法 5(词向量相似度 1)和方法 6(词向量相似度 2)，对两个字符串 S_1 和 S_2 进行分词，得到两个词语列表 $[W_{11},W_{12},\cdots,W_{1m}]$ 和 $[W_{21},W_{22},\cdots,W_{2n}]$，利用预训练词向量对词语进行词嵌入得到两个向量列表 $[\boldsymbol{V}_{11},\boldsymbol{V}_{12},\cdots,\boldsymbol{V}_{1m}]$ 和 $[\boldsymbol{V}_{21},\boldsymbol{V}_{22},\cdots,\boldsymbol{V}_{2n}]$，再对列表中

表 1 多种相似性度量方法及其描述

信息源	序号	方法名称	方法描述
属性名称	1	完全匹配	判断两个字符串是否完全相同，若相同则置 1，否则置 0
	2	编辑距离	计算两个字符串的编辑距离作为相似性度量
	3	最大连续匹配	计算两个字符串的最大连续匹配序列长度作为相似性度量
	4	杰卡德相似度 1	利用 2-gram 切分字符串成为集合，计算集合的杰卡德相似系数作为相似性度量
	5	词向量相似度 1	利用预训练词向量嵌入，计算向量的余弦距离
	6	词向量相似度 2	利用预训练词向量嵌入，计算向量的欧氏距离
属性实例	7	杰卡德相似度 2	利用 2-gram 切分处理属性实例集合，计算集合的杰卡德相似系数作为相似性度量
	8	TF-IDF 相似度	利用 TF-IDF 信息将两个实例集合分别嵌入稀疏向量，计算向量之间的余弦距离
	9	词向量相似度 3	利用预训练词向量将两个实例集合分别嵌入稠密向量，计算向量之间的余弦距离
	10	集合编码相似度	通过训练表示学习模块计算得到集合相似度

的向量求平均得到两个向量 $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$，分别表示字符串 S_1 和 S_2。方法 5 和方法 6 分别利用两个向量 $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$ 的余弦距离和欧氏距离作为相似性度量，为了保证相似度 1 代表完全相同的特性，将其结果乘“－1”后再归一化至[0，1]。

相似性度量方法 7 至方法 10 利用了属性的实例信息，属性实例作为属性的具体形式，可以为属性对齐提供准确的判别信息。针对候选属性对，我们可以抽取这两个属性对应的所有实例并以集合形式呈现，集合中的一个元素即为一个实例。对于方法 7(杰卡德相似度 2)，将两个实例集合中的每个实例进行 2-gram 切分，成为切分集合；再分别将两个属性实例集合中的所有切分集合取并集，形成两个切分并集，通过计算两个切分并集的杰卡德相似系数的对数作为相似性度量。但由于两个属性实例集合的相似性随着集合元素交集数量的增加边际递减，所以实际应用中利用指数函数对杰卡德相似系数进行了非线性变换。

对于方法 8(TF-IDF 相似度)，将每个属性实例集合看作一个文档，计算词语的 TF-IDF 值，并利用 TF-IDF 值将每个词语表示成词表长度的独热向量(one-hot)。将两个属性实例集合中的 TF-IDF 词向量求和作为实例集合的表示，计算这两个向量的余弦距离作为相似性度量。对于方法 9(词向量相似度 3)，将两个属性实例集合中的所有词语利用预训练词向量进行词嵌入，再将两个属性实例集合中的词向量求平均得到稠密向量作为实例集合的表示，计算这两个向量的余弦距离作为相似性度量，为了保证相似度 1 代表完全相同的特性，将其结果乘“－1”后再归一化至[0，1]。

属性实例的信息分布在集合的每个元素中，以上方法通过将实例集合向量化或对实例集合进行处理，从而计算两个实例集合的相似度。但属性实例具有信息密度低的特点，并且包含噪声问题和相关属性实例无交集问题，这导致以上方法对实例信息的利用不够充分。针对这些问题我们提出了属性实例集合表示学习算法作为方法 10(详见 3.2)，利用深度学习的特征抽取能力来学习集合的主题特征，并利用基于内容的注意力机制解决噪声问题。

3.2　属性实例集合表示学习模块

本主要介绍属性实例集合表示学习模块，该模块仅利用属性实例信息进行属性对齐，训练结束后将其作为一个相似性度量方法加入上述属性对齐模型中。属性的实例以集合形式呈现，实例与实例之间并没有顺序关系，但它们共同反映了属性的主题信息。对于输入或输出是集合形式的问题，需要算法具有以下两个特点：①输入或输出对顺序不敏感。②输入或输出长度可变。针对这一问题，本文借鉴集合编码的思想，提出了属性实例集合表示学习模块。模块的架构如图 3 所示，首先利用外部预训练词向量将每个属性实例嵌入稠密向量，再通过集合编码单元将属性实例集合转化为向量，对两个实例集合的向量进行拼接后通过浅层神经网络得到两个集合的相似性特征。

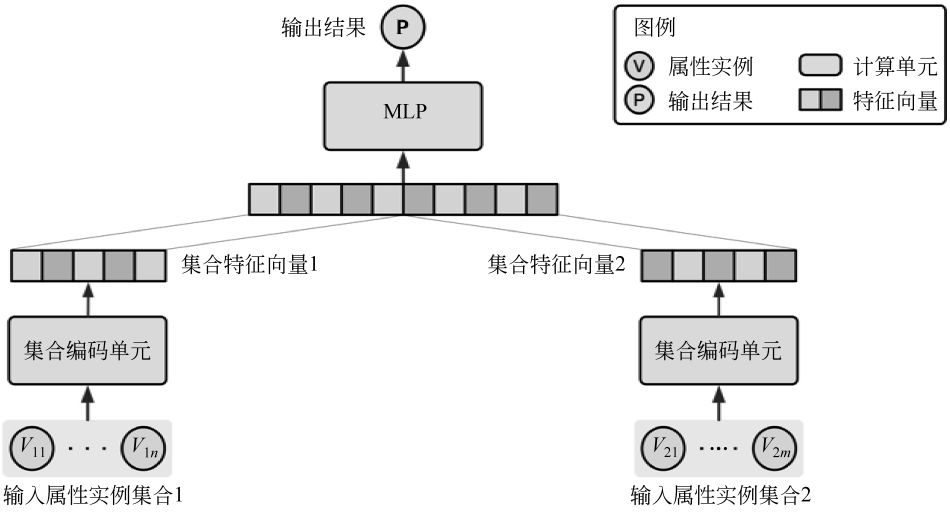

图 3　属性实例集合表示学习模块架构图

本文对循环神经网络(RNN)的输入进行修改从而实现输入不定长和顺序不敏感的特点。使用的集合编码单元与 RNN 的结构对比如图 4 所示。

RNN 在每个时间步输入一个单位的内容，所有

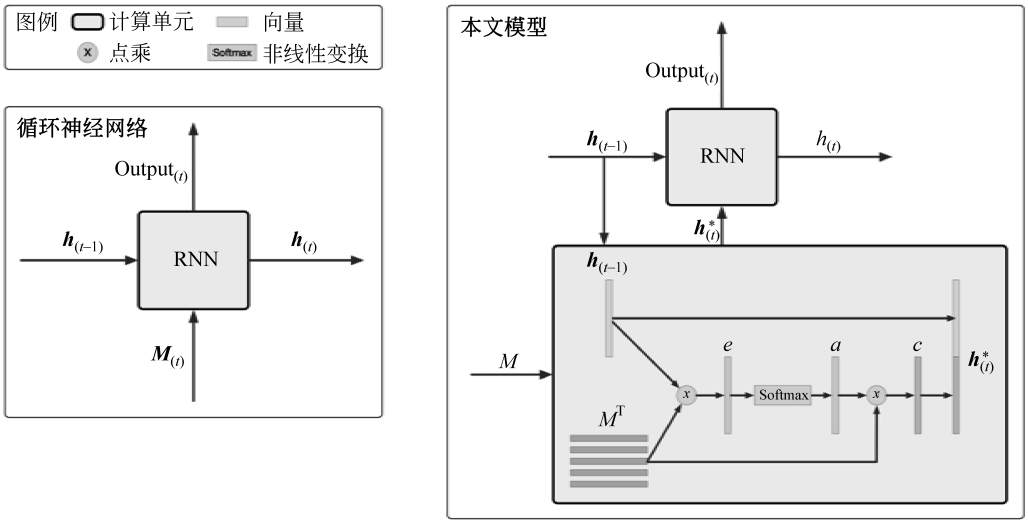

图 4 集合编码单元与 RNN 的架构对比

输入内容的长度决定了网络的时间步数量，从而导致输入具有顺序性。本文在每个时间步输入所有内容而不是部分内容，时间步的长度则由超参数设定。这样做使得输入元素的顺序不影响模块的输出结果，并且输入长度可变，同时保留了 RNN 的特征提取能力。

在属性实例集合中，部分关键实例更能反映出属性的主题信息而离群点则引入了噪声。所以本文在输入部分加入基于内容的注意力机制，通过有监督学习算法自动判别每个元素的重要程度，从而更加有效地提取集合的主题特征并忽略噪声实例。本文使用的注意力机制如式(3)～式(7)所示。

$$\boldsymbol{h}_{t-1}=\mathrm{RNN}(\boldsymbol{h}^{*}_{t-1}) \tag{3}$$

$$\boldsymbol{e}_{i,t}=f(M_i,\boldsymbol{h}_{t-1}) \tag{4}$$

$$\boldsymbol{a}_{i,t}=\frac{\exp(\boldsymbol{e}_{i,t})}{\sum_j \exp(\boldsymbol{e}_{j,t})} \tag{5}$$

$$\boldsymbol{c}_t=\sum_i \boldsymbol{a}_{i,t}\cdot M_i \tag{6}$$

$$\boldsymbol{h}^{*}_t=[\boldsymbol{h}_{t-1}\ \boldsymbol{c}_t] \tag{7}$$

其中，$\boldsymbol{h}^{*}_{t-1}$表示 $t-1$ 时间步的输入，$\boldsymbol{h}_{t-1}$表示 $t-1$ 时间步输出的隐藏状态。$\boldsymbol{M}$ 表示所有输入组成的 m 行矩阵，矩阵的第 i 行 $\boldsymbol{M}_i$ 表示集合的第 i 个元素所对应的向量表示。f 表示一个计算函数(本文使用向量的点积)。$\boldsymbol{e}_{i,t}$是 $\boldsymbol{M}_i$ 与 $\boldsymbol{h}_t$ 的点积，表示第 i 个输入元素在时间步 t 的权重。$\boldsymbol{e}_t$是一个 m 维向量 $\{\boldsymbol{e}_{1,t},\boldsymbol{e}_{2,t},\cdots,\boldsymbol{e}_{m,t}\}$。对 $\boldsymbol{e}_t$进行归一化处理得到 $\boldsymbol{a}_t$，即表示归一化后的权重向量。利用权重向量对每个输入元素 $\boldsymbol{M}_i$进行加权求和得到 $\boldsymbol{c}_t$。再将加权求和的结果 $\boldsymbol{c}_t$与 $\boldsymbol{h}_t$进行拼接得到的 $\boldsymbol{h}^{*}_t$ 作为 t 时间步的输入。在最后一个时间步 T，将 $\boldsymbol{h}_T$ 当做集合的向量表示。

4 实验

4.1 实验数据

本文使用实际场景中的数据集进行实验，数据来自淘宝、天猫、盒马鲜生和某电商品台。平台中的实体为商品，按商品类目进行划分，其中的实体以属性为描述框架。我们从中选取 142 对商品类目，抽取属性信息并标注对齐属性作为实验数据集。其中包含 41 980 个候选属性对，标注的正例(对齐属性对)数量为 1 082 对，标注的负例(不对齐属性对)40 898 对。例如，正例包括对齐属性“产地”和“生产地址”等，负例包括指代不相同的属性“产地”和“脂肪含量”。

4.2 模型设置

基于多相似性度量的属性对齐模型使用 scikit-learn 实现，选择 XGBoost 模型进行分类。

属性实例集合表示学习模块基于 Pytorch 实现。模型采用 RMSprop[24] 算法进行参数更新；batch size 设置为 300；epoch 设置为 50；学习率初始化为 0.000 5；正负样本的损失更新权重设置为 $\{1,\boldsymbol{r}/3\}$，其中 $\boldsymbol{r}$ 表示负样本与正样本数量的比值；集合编码单元中设置层数为 1，时间步总长为 10，隐藏状态向量维度为 200；浅层神经网络的维度设置为{1 200，100，10，2}。模型中词向量使用了外部的

中文预训练词向量[25]，词向量维度为 200，由于训练样本规模较小，所以本实验使用了固定的预训练词向量。

4.3　实验结果

实验将本文提出的模型与本体匹配领域最好的方法之一 CroMatcher[18] 进行对比，CroMatcher 模型可以根据匹配问题动态计算多种相似性度量的聚合权重。同时与 Rafcom[19] 模型进行对比，Rafcom 使用随机森林分类模型进行本体匹配，并且引入预训练词向量辅助对齐。由于任务的差异，本文保留了以上方法中所有可用的相似性度量方法。Baseline 仅利用字符串相似性度量进行属性对齐。MSSE（multi-similarity with set encoding）和 MS（multi-similarity）为本文模型，MS 在 MSSE 的基础上去除了属性实例集合表示学习模块。评价指标采用正例的准确率（P）、召回率（R）和综合指标（F_1）。实验结果如表 2 所示，从表中可以得到以下结论。

表 2　基于多相似性度量的属性对齐实验结果

序号	模型	P	R	F_1
1	Baseline(stringba)	0.854 0	0.713 5	0.777 4
2	CroMatcher	0.922 1	0.851 9	0.885 6
3	Rafcom	0.877 8	0.945 7	0.910 5
4	MSSE (ours)	0.976 8	0.925 2	**0.952 6**
5	MS (-SE)	0.980 6	0.889 9	0.933 0

- 本文提出模型在属性对齐任务中 F_1 值达到 0.952 6，相比于其他模型取得了最好的结果。
- 对比模型 Rafcom 和 MSSE 可以看出，本文设计的相似性度量方法能够有效地提取属性的相似性特征，并将 F_1 值提升约 4 个百分点。
- 对比模型 MS 和 MSSE 可以看出，本文提出的属性实例集合表示学习算法充分利用了属性的实例信息，并将属性对齐结果提升了接近两个百分点。

4.4　实验分析

为了验证本文提出的属性实例集合表示学习算法的有效性，我们设计了只利用属性实例信息进行属性对齐的对比实验。Jaccard of String 对属性实例进行 2-gram 串切分后，以两个集合的杰卡德相似系数作为度量进行属性对齐。Average of Vector 利用预训练词向量对所有实例进行词嵌入，利用一个集合中所有向量的均值表示这个集合，计算两个向量的余弦相似度作为度量进行属性对齐。Set-Encoding为本文提出的属性实例集合表示学习算法。评价指标采用正例的准确率（P）、召回率（R）和综合指标（F_1），结果如表 3 所示。从表中可以得到以下结论：属性实例集合表示学习算法在本实验中 F_1 值超过 0.72，并显著提升了属性对齐的准确率和召回率。模型有效学习到了集合的主题特征，从而使得属性实例信息的利用效率大幅的提升。

表 3　属性实例集合表示学习实验结果

序号	模型	P	R	F_1
1	Jaccard of String	0.609 4	0.343 6	0.439 4
2	Average of Vector	0.525 9	0.312 7	0.392 3
3	SetEncoding(ours)	0.791 2	0.676 1	**0.729 1**

在基于多相似性度量的属性对齐模型中，特征的重要性可以反映出相似性度量方法的有效性。XGBoost 模型在训练中输出的特征权重如图 5 所示，可以看出属性实例集合表示学习模块提取的特征（集合编码相似度）权重排名第二，充分说明了该模型有效地利用了属性的实例信息，以及该特征在属性对齐模型中的重要作用。

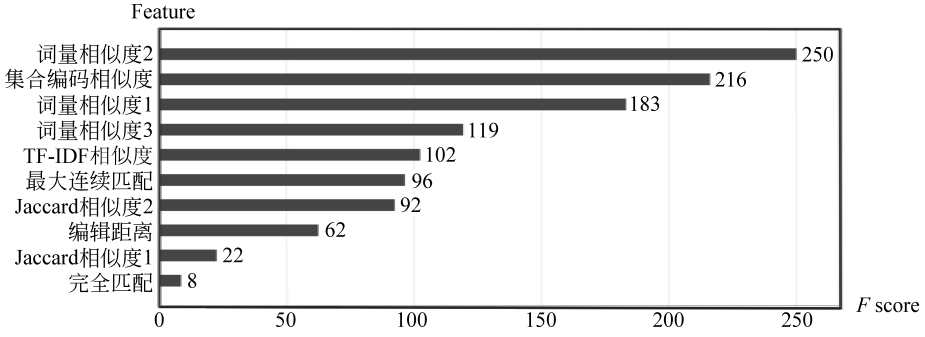

图 5　XGBoost 模型特征权重

本文提出的属性实例集合表示学习算法有效解决了属性实例信息的利用问题。对于例子{“红色”，“黄色”，“橙色”，“lkg”}和{“绿色”，“蓝色”，“紫色”}中包含的噪声问题和相关实例无交集问题，模型中的注意力机制降低了噪声实例“lkg”的权重从而减小了噪声的影响，同时通过有监督的学习算法，模型从实例集合中有效学习到了实例集合的主题特征(颜色)。

4.5 错误分析

通过对以上两个实验的错误样例分析，发现仍存在以下两个问题：①预训练词向量在对齐问题中对于相关词和同义词不敏感。Word2Vec[26]训练词向量的依据是上下文共现信息，所以高相似性词语不仅包含同义词也包含相关词，例如，“方式”和“方法”为同义词，“苹果”和“梨”为相关词。而在对齐问题中需要严格判断两个词是否指代相同，预训练词向量则会引入小部分相关词问题。设计模型训练适合对齐类问题的词向量，是解决此问题的方法之一。②模型没有针对数据特征选择合适的相似性度量方法。在电商平台的知识图谱中，属性信息一般会以文本、数字、符号等形式出现。每种形式的特征需要利用更合适的方法来表示和衡量。例如，对于属性“货号”的实例{“PCH-60”，“PCH-100”}，则无法通过词嵌入的方法有效地表示集合的主题特征。通过设计模型自动化地选择数据的特征提取方法会为模型效果带来进一步的提升。

5 结束语

本文提出了基于多种相似性度量的属性对齐模型，针对属性对齐任务从不同角度设计了多种相似性度量方法。为了充分利用属性的实例信息，本文提出了属性实例集合的表示学习算法。实验表明，我们提出的属性对齐模型在实际任务中取得了领先的结果，针对属性对齐问题设计的相似性度量方法可以有效提取属性的相似性特征，从而解决相似性度量不全面的问题。同时，属性实例集合表示学习算法充分利用属性实例信息，通过提取属性实例的主题特征解决了实例集合无交集的问题，其中基于内容的注意力机制缓解了数据的噪声问题。通过实验分析，我们注意到预训练词向量在对齐类任务中存在相关词和同义词不敏感的问题，以及模型没有针对数据特点选择合适的相似性度量方法。在未来的工作中，我们将通过修改预训练模型来获得适合属性对齐问题的词向量表示，同时在属性对齐模型中加入相似性度量方法的自动化选择机制，从而进一步提升属性对齐效果。

参考文献

[1] Dong X L, Rekatsinas T. Data integration and machine learning: A natural synergy[C]//Proceedings of the 2018 International Conference on Management of Data. Houston, Texas, USA: ACM, 2018: 1645-1650.

[2] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching[J]. The VLDB Journal, 2001, 10(4): 334-350.

[3] Comito C, Patarin S, Talia D. A semantic overlay network for p2p schema-based data integration[C]//Proceedings of the 11th IEEE Symposium on Computers and Communications. Cagliari, Sardinia, Italy: IEEE, 2006: 88-94.

[4] Bernstein P A, Madhavan J, Rahm E. Generic schema matching, ten years later[C]//Proceedings of the VLDB Endowment. Seattle, WA, USA: VLDB Endowment, 2011, 4(11): 695-701.

[5] Kirsten T, Thor A, Rahm E. Instance-based matching of large life science ontologies[C]//Proceedings of the International Conference on Data Integration in the Life Sciences. Berlin, Heidelberg: Springer, 2007: 172-187.

[6] Dhamankar R, Lee Y, Doan A, et al. iMAP:Discovering complex semantic matches between database schemas[C]//Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. Paris, France: ACM, 2004: 383-394.

[7] Aumueller D, Do H H, Massmann S, et al. Schema and ontology matching with COMA++[C]//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland, USA: ACM, 2005: 906-908.

[8] Do H H, Rahm E. COMA: A system for flexible combination of schema matching approaches[C]//Proceedings of the 28th International Conference on Very Large Data Bases. Hong Kong, China: VLDB Endowment, 2002: 610-621.

[9] Köhler H, Zhou X, Sadiq S, et al. Sampling dirty data for matching attributes[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indiana, USA: ACM, 2010: 63-74.

[10] Vinyals O, Bengio S, Kudlur M. Order matters: Sequence to sequence for sets[J]. arXiv preprint arXiv: 1511.06391. 2015.

[11] Shvaiko P, Euzenat J. Ontology matching: State of

the art and future challenges[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 158-176.

[12] Jiménez-Ruiz E, Grau B C. Logmap: Logic-based and scalable ontology matching[C]//Proceedings of the 10th International Semantic Web Conference. Berlin, Heidelberg: Springer, 2011: 273-288.

[13] Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching[C]//Proceedings of the 18th International Conference on Data Engineering. Washington, DC, USA: IEEE, 2002: 117-128.

[14] Madhavan J, Bernstein P A, Doan A, et al. Corpus-based schema matching[C]//Proceedings of the 21st International Conference on Data Engineering. Tokoyo, Japan: IEEE, 2005: 57-68.

[15] Faria D, Pesquita C, Santos E, et al. AgreementMakerLight 2.0: Towards efficient large-scale ontology matching[C]//Proceedings of the International Semantic Web Conference. Riva del Garda, Trento, Italy: Springer, 2014: 457-460.

[16] Faria D, Pesquita C, Santos E, et al. The agreementmakerlight ontology matching system[C]//Proceedings of OTM Confederated International Conferences. Graz Austria: Springer, 2013: 527-541.

[17] Cruz I F, Antonelli F P, Stroe C. AgreementMaker: Efficient matching for large real-world schemas and ontologies[C]//Proceedings of the VLDB Endowment. Lyon, France: VLDB Endowment, 2009: 1586-1589.

[18] Gulić M, Vrdoljak B, Banek M. Cromatcher: An ontology matching system based on automated weighted aggregation and iterative final alignment[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2016, 41: 50-71.

[19] Nkisi-Orji I, Wiratunga N, Massie S, et al. Ontology alignment based on word embedding and random forest classification[C]//Proceedings of Joint European-Conference on Machine Learning and Knowledge Discovery in Databases. Cham, Dublin, Ireland: Springer, 2018: 557-572.

[20] Fernandez R C, Mansour E, Qahtan A A, et al. Seeping semantics: Linking datasets using word embeddings for data discovery[C]//Proceedings of the 34th International Conference on Data Engineering. Paris, France: IEEE, 2018: 989-1000.

[21] Staab S, Studer R. Handbook on ontologies[M]. Berlin, Heidelberg: Springer, 2004: 385-403.

[22] Ngo D, Bellahsene Z. YAM++: A multi-strategy based approach for ontology matching task[C]//Proceedings of the International Conference on Knowledge Engineering and Knowledge Management. Galway, Ireland: Springer, 2012: 421-425.

[23] Chen T, Guestrin C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016: 785-794.

[24] Hinton G, Srivastava N, Swersky K. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude[R]. COURSERA: Neural Networks for Machine Learning 4.2, 2012: 26-31.

[25] Song Y, Shi S, Li J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA: ACL, 2018: 175-180.

[26] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA, 2013: 3111-3119.

伍家豪(1995—)，通信作者，硕士研究生，主要研究领域为知识融合。
E-mail：diwujiahao@163.com

韩先培(1984—)，博士，研究员，主要研究领域为信息抽取、知识库构建和自然语言处理。
E-mail：xianpei@iscas.ac.cn

陈波(1988—)，博士，副研究员，主要研究领域为机器学习和自然语言处理。
E-mail：chenbo@iscas.ac.cn