文章编号：2095-2163(2022)10-0009-06 中图分类号：TP301 文献标志码：A

一种基于图嵌入模型的关联感知真值发现

吕 航，Xiu Susie Fang，司苏新，王 康
（东华大学 计算机科学与技术学院，上海 201620）

摘 要：为解决多源数据间广泛存在的冲突问题，真值发现成为一个热门的研究课题。现有的真值发现算法通常基于这一原则：如果一个信息源总是提供真实的信息，那么就会更加可信；如果一条信息由可信的信息源支持，那么就更有可能是真实的。现有的真值发现算法虽然在大部分场景下取得了较好的效果，但大多忽略了实体属性之间的关系。在本文中，提出了一种新的模型，该模型采用图嵌入方式在真值发现的同时捕捉了实体属性间的关系。通过构建 4 种异构网络，包括源-源、源-属性值、实体属性-实体属性、实体属性-实体属性值网络，以对数据之间的关系建模。接着将这些网络嵌入到低维空间中，使得可靠的来源和可靠的属性值彼此接近，实体属性之间的关系反映在属性值上，从而进行真值发现推理。在 2 个真实数据集上的实验，表明本文的算法优于现有的真值发现算法。

关键词：图嵌入；真值发现；实体属性关系；异构网络

A graph embedding model for correlation aware truth discovery

LÜ Hang, Xiu Susie Fang, SI Suxin, WANG Kang
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 In order to solve the widely existing conflicts among multi-source data, truth discovery has become a hot topic. Existing truth discovery algorithms are usually based on such a principle: when a source always offers true information, it would be more trustworthy. When the information is supported by trustworthy sources, it would be believed to the truth. However, even though existing truth discovery algorithms have achieved good results in most scenarios, the relationship between entity attributes could be mostly ignored. This paper proposes a new model that uses graph embedding to obtain truth and captures the relationship between entity attributes. This model builds four heterogeneous networks, including source-source, source-attribute value, entity attribute-entity attribute, and entity attribute-entity attribute value networks, to capture the relationship among data. These networks are then embedded into the low dimensional space, so that the reliable sources and reliable attribute values are close to each other, and the relationships between entity attributes can be reflected in the attribute values, therefore truth discovery inference could be carried out. Experiments on two real-world datasets show that the proposed algorithm is superior to the state-of-the-art truth discovery algorithms.

【Key words】 graph embedding; truth discovery; entity attributes relation; heterogeneous network

0 引 言

在过去的几十年里，从搜索引擎、社交媒体平台、众包平台等各种网络渠道收集的数据量急剧增加。人们往往可以从不同的数据源收集同一实体的声明信息，然而由于记录错误、机器故障、噪音、恶意攻击等原因，这些信息可能会相互冲突。如果不解决这些冲突，从网络上检索到的信息将毫无用处。为了得到可靠的信息（即真实的事实），就需要研究多源数据的聚合技术。

近些年来，许多研究提出了多源数据聚合的方法。这些方法可以分为 3 类：

（1）迭代法[1]。迭代计算来源的可靠性和声明值的可信度。

（2）基于最优化的方法[2-4]。使每个声明值与真实值之间的源加权距离最小。

（3）概率法[5-6]。对源和声明值的联合分布进行建模，使联合分布可能性最大化。

虽然现有方法已取得了不错的效果，但大部分方法都忽视了实体属性之间存在的各种关联[7]。研究可知，充分利用实体属性之间的关联能提升真值发现结果的准确性。

这里通过表 1 中的实例来阐述这一点。由表 1 看到，实体具有年龄、出生日期、居住城市和邮编等

作者简介：吕 航（1998-），男，硕士研究生，主要研究方向：真值发现；Xiu Susie Fang（1988-），女，博士，讲师，主要研究方向：大数据分析、网络挖掘、数据集成；司苏新（1998-），男，硕士研究生，主要研究方向：众包；王 康（1998-），男，硕士研究生，主要研究方向：真值发现。
通讯作者：吕 航 Email：2202516@mail.dhu.edu.cn

收稿日期：2022-03-17

属性。这些属性中存在如下关联:年龄取决于出生日期,城市和邮编具有一一对应关系。如果采用多数投票的方法,可能会在实体 1 的年龄属性上得到错误的结果为 18 岁。然而,通过考虑年龄和出生日期之间的依赖关系,就可以先获得出生日期的真值,即 2004 年 1 月 1 日,从而得到正确的年龄为 19 岁。这说明如果在本文的方法中,能够捕捉属性间的关系,可以获得更准确的结果。

表 1 实体的信息表

Tab. 1 Entity information table

源	实体 1			
	年龄	出生日期	城市	邮政编码
源 1	18	2004.1.1	上海	200000
源 2	18	2004.1.1	上海	200000
源 3	19	2004.1.1	北京	200000
真值	19	2004.1.1	上海	200000

考虑实体属性相关性的真值发现研究仍处于起步阶段,仅有的一些方法对实体属性关系的捕捉还不全面,比如现有的研究集中于属性的几种特定关系,如时间关系[8]、空间关系[9]或常识[6,10],或将属性之间的关系采用数据间约束[4]来表示。本文提出的异构网络模型不仅捕捉了数据源间的相似关系、数据源对声明值的偏好选择关系,还考虑了实体属性的一般化关系来推断实体属性的真实值。接下来将基于 2 个真实世界数据集的实验结果证明了本文的算法优于现有方法。

1 图嵌入模型

1.1 问题定义

假设有 N 个实体,每个实体具有 M 个属性。这些实体的信息由 K 个来源提供。第 n 个实体的第 m 个属性的第 i 个声明值表示为 V_{mi}^{n}。第 n 个实体的第 m 个属性的声明值集合为 V_{m}^{n},第 i 个实体的第 m 个属性为 T_{m}^{i},第 i 个实体的属性集合为 T_{i},源的集合用 S 表示,第 i 个源为 S_{i}。考虑一个具体的例子:李华作为第一个实体,具有 2 个属性:年龄和出生日期,这 2 个属性间存在关系。源的集合为 $S=\{S_1,S_2,S_3\}$,源关于实体属性做出的声明值集合为:S_1:“22,2000 年 2 月 1 号”、S_2:“21,2000 年 2 月 1 号”、S_3:“22,2000 年 2 月 1 号”。在这个声明值集合中 V_{11}^{1} 为 22,V_{1}^{1} 为 $\{22,21\}$,属性 T_{1}^{1} 为年龄。本文的研究目标就是基于对年龄和出生日期属性间关系的捕捉,找出实体属性的真实值。

1.2 异构网络

本节将创建 4 个网络,这 4 个网络一起构建了一个大型的异构网络,如图 1 所示,用于处理存在实体属性关联的真值发现问题。

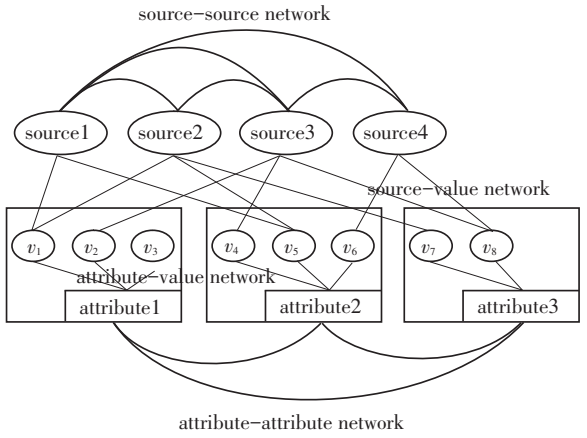

图 1 异构网络

Fig. 1 Heterogeneous network

定义 1 源-属性值网络 源与属性值之间的网络定义为 $G_{sv}=(S\cup V,E_{sv})$,S 是源的集合,V 是属性值的集合,E_{sv} 是源与属性值之间的边集合,边上定义了权重 W_{mij}^{n},当源 S_j 声明了属性值 V_{mi}^{n} 时,源 S_j 与属性值 V_{mi}^{n} 之间的边上的权重为 1,反之则为 0。

通过定义源与属性值之间的网络,能够对源声明一个属性值的过程建模,这种建模可以将源的可靠性体现在其对声明值的偏好选择上。

定义 2 源-源网络 源与源之间的网络定义为 $G_{ss}=(S\cup S,E_{ss})$,这里 S 是源的集合,E_{ss} 是源与源之间的边,边上定义了 2 种不同权重。第一种 W_{ij} 为源 S_i 和源 S_j 对给定的同一实体的同一属性做出相同的声明值的数量,第二种 D_{ij} 为这 2 个源对给定的同一实体的同一属性做出不同的声明值数量。

通过定义源与源之间的网络,能够挖掘源与源之间的相似性。如果相同声明值权重 W_{ij} 越大于不同声明值权重 D_{ij},则说明 2 个源越相似。同时结合源-属性值网络中捕捉的关系,源之间的关系表明了源提供可信声明的偏好。

定义 3 实体属性-实体属性网络 实体属性-实体属性之间的网络定义为 $G_{TT}=(T\cup T,E_{TT})$,这里 T 是同一实体属性的集合,E_{TT} 是同一实体属性之间的边的集合,边上定义了 2 种不同权重。第一种 W_{nm}^{i} 为属性 T_{n}^{i} 和属性 T_{m}^{i} 的值成对出现大于一次的数量,第二种为 D_{nm}^{i} 为属性 T_{n}^{i} 和属性 T_{m}^{i} 的值成对出现等于一次的数量。

通过定义实体属性之间的网络，能建模实体属性的相关性，如果权重 W_{nm}^{i} 越大于 D_{nm}^{i}，则说明实体属性之间的相关性越强。

定义 4　实体属性-实体属性值网络　实体属性-实体属性值之间的网络定义为 $G_{TV}=(T\cup V,E_{TV})$，这里 T 是实体属性的集合，V 是实体属性值的集合，E_{TV} 表示实体属性和其属性值之间的边。边上定义了权重 W_{mi}^{n}，若实体属性 T_m^n 具有属性值 V_{mi}^{n}，则为 1，反之为 0。

通过定义实体属性与实体属性值之间的网络，能将建模的实体属性之间的关系体现在属性值层面上。

4 种异构网络中捕捉的各种连接关系为规范化真值发现建模提供了更多的证据。

1.3　网络的嵌入

在本节中，提出将 4 种异构网络嵌入到低维空间的处理方法。由于这个异构的网络由 4 个子网络组成，这里采用的是对每个子网络进行嵌入，再嵌入整个异构网络的方法。

定义 5　(源-属性值)网络嵌入　使用 o_j 代表源 S_j 在嵌入空间的表示，使用 a_{mi}^{n} 代表属性值 V_{mi}^{n} 在嵌入空间的表示。由于源-属性值的网络能够捕捉源对于某一属性值的偏好，图 G_{sv} 中的每一条边上属性值被源所声明的概率为：

$$p(V_{mi}^{n}\mid S_j)=\frac{\exp\left(a_{mi}^{n\mathrm{T}}o_j\right)}{\sum_{V_{mr}^{n}\in V_m^n}\exp\left(a_{mr}^{n\mathrm{T}}o_j\right)}\tag{1}$$

通过对源与属性值的建模，较高的条件概率 $p(V_{mi}^{n}\mid S_j)$ 表示源与属性值的嵌入表示相似度较高，也就表明在这个属性上，源会趋向于选择这个声明值，因此如果有了声明值的可信度，就可以进行源可靠性建模。即通过源与声明值的异构网络将源的可靠性与属性值的可信度之间的关系建模。

通过最大化条件概率 $p(V_{mi}^{n}\mid S_j)$，能使源及在某一实体属性上做出的声明值在嵌入空间中相近，即最小化损失函数 O_{sv}：

$$O_{sv}=-\sum_{(S_j,V_{mi}^{n})\in G_{sv}}W_{mij}^{n}\log p(V_{mi}^{n}\mid S_j)\tag{2}$$

定义 6　(源-源)网络嵌入　对于图 G_{ss} 的每一条边，定义了源 S_i 和源 S_j 的联合概率：

$$p_{ij}=p(S_i,S_j)=\frac{1}{1+\exp\left(o_i^{\mathrm{T}}*o_j\right)}\tag{3}$$

越高的条件概率 p_{ij} 表明源 S_i 和源 S_j 具有越高的相似性。即源 S_i 和源 S_j 在同一实体属性上做出相同声明值的概率为 p_{ij}。

图 G_{ss} 对于每一条边定义了 2 种不同的权重，在已知 p_{ij} 的条件下，可以得到 2 种权重的条件概率：

$$p(W_{ij},D_{ij}\mid p_{ij})=p_{ij}^{W_{ij}}(1-p_{ij})^{D_{ij}}\tag{4}$$

研究用 *beta* 分布去解决数据稀疏问题。此处需用到的数学公式为：

$$p(p_{ij}\mid\alpha,\beta)=\frac{p_{ij}^{\alpha-1}(1-p_{ij})^{\beta-1}}{B(\alpha,\beta)}\tag{5}$$

其中，B 是 *beta* 函数，α,β 是 2 个超参数。

通过最大化由源-源网络得到的概率，能使具有相似可靠性的源在嵌入空间内的距离相近。即最小化损失函数 O_{ss}：

$$O_{ss}=-\sum_{(S_i,S_j)\in G_{ss}}[\log p(W_{ij},D_{ij}\mid p_{ij})+\log p(p_{ij}\mid\alpha,\beta)]\tag{6}$$

定义 7　(实体属性-实体属性)网络嵌入　使用 t_n^i 代表属性 T_n^i 在嵌入空间的表示，对于 G_{TT} 的每一条边，定义了属性 T_n^i 和 T_m^i 的联合概率：

$$p_{nm}=p(T_n^i,T_m^i)=\frac{1}{1+\exp\left(t_n^{i\mathrm{T}}*t_m^i\right)}\tag{7}$$

越高的条件概率 p_{nm} 表示属性 T_n^i 和 T_m^i 之间存在关联的可能性越高，即 2 个属性之间存在关联的概率为 p_{nm}。

图 G_{TT} 对于每一条边定义了 2 种不同的权重，在已知 p_{nm} 的条件下，可以得到 2 种权重的条件概率为：

$$p(W_{nm}^{i},D_{nm}^{i}\mid p_{nm})=p_{nm}^{Wi_{nm}}(1-p_{nm})^{Di_{nm}}\tag{8}$$

接着提出用 *beta* 分布去解决数据稀疏问题。推理得出的数学公式为：

$$p(p_{nm}\mid\alpha,\beta)=\frac{p_{nm}^{wi_{nm}}(1-p_{nm})^{Di_{nm}}}{B(\alpha,\beta)}\tag{9}$$

其中，B 是 *beta* 函数，α,β 是 2 个超参数。

通过最大化属性-属性网络得到的概率，能使具有关系的属性在嵌入空间内的距离相近。即最小化损失函数 O_{tt}：

$$O_{tt}=-\sum_{(T_n^i,T_m^i)\in G_{TT}}[p(W_{nm}^{i},D_{nm}^{i}\mid p_{nm})+p(p_{nm}\mid\alpha,\beta)]\tag{10}$$

定义 8　(实体属性-实体属性值)网络嵌入　通过定义实体属性-实体属性值之间的网络，能将建模的实体属性之间的关系体现在属性值层面上。定义图 G_{TV} 上的 2 条边上的属性值成对出现的概率为：

$$p(V_{mi}^{l},V_{nj}^{l}\mid T_m^l,T_n^l)=\frac{\exp\left(a_{mi}^{l\mathrm{T}}t_m^l v_{njn}^{l\mathrm{T}}t_n^l\right)}{\sum_{a_{mr}^{l}\in V_m^l,a_{ns}^{l}\in V_n^l}\exp\left(a_{mr}^{l\mathrm{T}}t_m^l a_{ns}^{l\mathrm{T}}t_n^l\right)}\tag{11}$$

概率 $p(V_{mi}^l, V_{nj}^l \mid T_m^l, T_n^l)$ 值越高,表明这对声明值满足属性之间的约束,这也说明在值的嵌入表达式中反映了属性之间的关联。通过最大化条件概率 $p(V_{mi}^l, V_{nj}^l \mid T_m^l, T_n^l)$,能使满足实体属性关联的属性值与实体属性在嵌入空间中相近。即最小化损失函数 O_{tv}:

$$O_{tv} = -\sum_{(T_n^l, T_m^l, V_{mi}^l, V_{nj}^l) \in G_{TV}} W_{mi}^l W_{nj}^l \log p(V_{mi}^l, V_{nj}^l \mid T_m^l, T_n^l) \tag{12}$$

本次研究的目标是使源、实体属性、实体属性值的联合概率最大化,等价于最小化损失函数 O_{sv}、O_{ss}、O_{tt}、O_{tv},即最小化 O_{sum}:

$$O_{sum} = O_{sv} + O_{ss} + O_{tv} + O_{tt} \tag{13}$$

1.4 模型的学习

一种直观的解法:可以同时使用所有子网络来学习并更新各种嵌入的表示。即公式(13)的优化,也就是通过合并所有子网络的边,并对边抽样,抽样的概率与其在网络中的权重成正比,再根据抽样的边对参数的嵌入表示进行更新。但由于不同子网络的权重是不可比的,因此迭代采样每个子网络的边,基于偏导对每个子网络的嵌入表示进行更新。

对于源-属性值网络,计算 O_{sv} 关于 E_{sv} 的每一条 (S_j, V_{mi}^n) 上 a_{mi}^n 的偏导:

$$\frac{\partial O_{sv}}{\partial a_{mi}^n} = o_j \left(\frac{\exp(a_{mi}^{n\mathrm{T}} o_j)}{\sum_{v_{mr}^n \in V_m^n} \exp(a_{mr}^{n\mathrm{T}} o_j)} - 1 \right) \tag{14}$$

对于同一属性上未被源 S_j 提供的声明值 V_{ml}^n,同样求 O_{sv} 关于其的偏导:

$$\frac{\partial O_{sv}}{\partial a_{ml}^n} = o_j \frac{\exp(a_{ml}^{n\mathrm{T}} o_j)}{\sum_{v_{mr}^n \in V_m^n} \exp(a_{mr}^{n\mathrm{T}} o_j)} \tag{15}$$

紧接着,计算 O_{sv} 关于源 S_j 的偏导:

$$\frac{\partial O_{sv}}{\partial o_j} = -a_{mi}^n + \frac{\sum_{v_{mr}^n \in V_m^n} \exp(a_{mr}^{n\mathrm{T}} o_j) a_{mr}^n}{\sum_{v_{mr}^n \in V_m^n} \exp(a_{mr}^{n\mathrm{T}} o_j)} \tag{16}$$

通过式(14)~(16)对源,属性值的嵌入表示的更新,能使可靠的源和可信度较高的属性值在嵌入空间距离相近。

对于源-源网络,同样计算源 S_i 关于 O_{ss} 的偏导:

$$\frac{\partial O_{ss}}{\partial o_i} = (2 - W_{ij} - D_{ij} - \alpha - \beta) \frac{o_j \exp(-o_i^{\mathrm{T}} o_j)}{1 + \exp(-o_i^{\mathrm{T}} o_j)} + o_j (D_{ij} + \beta - 1) \tag{17}$$

通过对源的嵌入表示的更新,能使具有相似可靠性的源在嵌入空间相近。

对于实体属性-实体属性网络,同样计算实体属性 T_n^i 关于 O_{tt} 的偏导:

$$\frac{\partial O_{tt}}{\partial t_n^i} = (2 - W_{nm}^i - D_{nm}^i - \alpha - \beta) \frac{t_m^i \exp(t_m^{i\mathrm{T}} t_n^i)}{1 + \exp(t_m^{i\mathrm{T}} t_n^i)} + t_m^i (D_{nm}^i + \beta - 1) \tag{18}$$

通过对实体属性的嵌入表示的更新,能使具有关联的实体属性在嵌入空间相近。

对于实体属性-实体属性值网络,计算 O_{tv} 关于 t_m^l 的偏导:

$$\frac{\partial O_{tt}}{\partial t_m^l} = \frac{\sum_{a_{mr}^l \in V_m^l, a_{ns}^l \in V_n^l} \exp(a_{mr}^{l\mathrm{T}} t_m^l a_{ns}^{l\mathrm{T}} t_n^l) a_{mr}^{l\mathrm{T}} a_{ns}^{l\mathrm{T}} t_n^l}{\sum_{a_{mr}^l \in V_m^l, a_{ns}^l \in V_n^l} \exp(a_{mr}^{l\mathrm{T}} t_m^l a_{ns}^{l\mathrm{T}} t_n^l)} - a_{mi}^{l\mathrm{T}} a_{nj}^{l\mathrm{T}} t_n^l \tag{19}$$

计算 O_{tv} 关于 t_n^l 的偏导:

$$\frac{\partial O_{tt}}{\partial t_n^l} = \frac{\sum_{a_{mr}^l \in V_m^l, a_{ns}^l \in V_n^l} \exp(a_{mr}^{l\mathrm{T}} t_m^l a_{ns}^{l\mathrm{T}} t_n^l) a_{mr}^{l\mathrm{T}} t_m^l a_{ns}^{l\mathrm{T}}}{\sum_{a_{mr}^l \in V_m^l, a_{ns}^l \in V_n^l} \exp(a_{mr}^{l\mathrm{T}} t_m^l a_{ns}^{l\mathrm{T}} t_n^l)} - a_{nj}^{l\mathrm{T}} a_{mi}^{l\mathrm{T}} t_m^l \tag{20}$$

计算 O_{tv} 关于 a_{mi}^l 的偏导:

$$\frac{\partial O_{tv}}{\partial a_{mi}^l} = \frac{t_m^l a_{nj}^{l\mathrm{T}} t_n^l \exp(a_{mi}^{l\mathrm{T}} t_m^l a_{in}^{l\mathrm{T}} t_n^l)}{\sum_{a_{mr}^l \in V_m^l, a_{ns}^l \in V_n^l} \exp(a_{mr}^{l\mathrm{T}} t_m^l a_{ns}^{l\mathrm{T}} t_n^l)} - t_m^l a_{nj}^{l\mathrm{T}} t_n^l \tag{21}$$

计算 O_{tv} 关于 a_{nj}^l 的偏导:

$$\frac{\partial O_{tv}}{\partial a_{nj}^l} = \frac{t_n^l a_{mi}^{l\mathrm{T}} t_m^l \exp(a_{mi}^{l\mathrm{T}} t_m^l a_{in}^{l\mathrm{T}} t_n^l)}{\sum_{a_{mr}^l \in V_m^l, a_{ns}^l \in V_n^l} \exp(a_{mr}^{l\mathrm{T}} t_m^l a_{ns}^{l\mathrm{T}} t_n^l)} - t_n^l a_{mi}^{l\mathrm{T}} t_m^l \tag{22}$$

通过式(19)~(22)对实体属性、属性值的嵌入表示更新让满足实体属性之间关联的属性值在嵌入空间靠近实体属性。

综上所述,就可使用随机梯度下降(SGD)方法去更新实体属性、源、属性的嵌入表示。

1.5 真值的推断

在模型学习中,得到了属性值嵌入、源嵌入和属性嵌入,同时对于各种嵌入的优化让真值嵌入和真实属性值嵌入在嵌入空间中接近。因此通过计算集合 V_m^n 中每个属性值和真值之间的相似性,相似性最高的属性值即为属性的真实值。

但是由于本文算法是无监督的,并没有真值的嵌入。为了构造真值嵌入,文中采用对所有属性多数投票来找到真实值,对得到的真实值的集合按照

真实值的可信度进行排序，从排序的真实值集合选取前 L 个真实值并取其平均值作为真值的嵌入。

2　实验

研究采用 Python(3.6)实现了所有的基线方法和本文提出的模型(GETD)，所有的实验都是在 Intel Core i5-7200U CPU@2.50 GHz 的电脑上运行的。

2.1　数据集

(1) Restaurant[11]：该数据集包括来自 5 个源提供的信息。每个餐厅是一个实体，每个实体有 5 个分类属性：餐厅名称、建筑编号、街道名称、邮政编码和电话号码。

(2) Weather[3]：该数据集包含 9 个来源提供的信息，每个城市是一个实体，每个实体具有 30 个分类属性，即一个月内的天气情况。

真实数据集的统计结果见表 2。

表 2　真实数据集的统计

Tab. 2　The statistics of real-world datasets

数据集	大小/KB	实体数	源的数量	属性数量
Restaurant	563.0	10 763	5	5
Weather	59.9	20	9	30

2.2　评价指标

错误率 (*error rate*)：推断的实体属性真实值与 *ground truth* 中不同的数量占 *ground truth* 的百分比，越小的错误率表明实验结果越好。

2.3　对比算法

(1) Majority Voting：该方法认为在所有源中出现次数最多的声明值为真值。

(2) TruthFinder[1]：通过给定源的可靠性，去推断真值，再根据真值去推断源的可靠性，迭代更新源的可靠性和真值至收敛。

(3) CRH[3]：将真值发现视为一个最优化问题，采取两步迭代更新，一步更新源权重，一步更新值的可信度。

(4) CATD[2]：采用最优化的方法解决真值发现问题，将源的权重采用置信区间的方式建模，以解决数据稀疏问题。

(5) CASE[12]：通过使用一种嵌入方法，解决真值发现问题，但不考虑属性之间的关系。

(6) CTD[4]：将真值发现视为一个最优化问题，同时使用数据库约束来捕捉属性关联的方法。

2.4　实验设置

为了确保公平的对比，研究运行了一系列的实验来为每个基线方法设定最优的参数。对于本文的方法，设置嵌入维度为 12。对于 SGD 方法，设置学习率为 0.1。设置 *beta* 函数中 α 和 β 为 1.1。对于真值的推断步骤中 L 值设为 3%。

2.5　实验结果

在表 3 和表 4 中列出了不同真值发现算法在 Restaurant 数据集和 Weather 数据集上实验 3 次的运行结果及平均的错误率。从实验结果中，可以看到本文提出的 GETD 方法在 2 个真实世界数据集上都优于其他方法，这是因为这些基线方法大多都没有考虑属性之间的关系，只是单纯地考虑数据源的可靠性或相似性，并不能捕捉属性之间的关系，导致实验精度不够。CTD 算法虽然考虑了属性之间的关系，但是算法的 reduction 部分并未考虑迭代，导致了精度的丢失。本文提出的 GETD 模型在考虑源的可靠性与源的相似性的基础上，全面捕捉了一般化的属性关系，能够更加精准地挖掘底层数据之间的关联。

表 3　基于 Restaurant 数据集的对比结果

Tab. 3　Comparison results based on Restaurant dataset

方法	Test1	Test2	Test3	*Mean*
MV	0.063 0	0.063 0	0.063 0	0.063 0
CRH	0.060 1	0.060 5	0.060 9	0.060 5
CATD	0.061 9	0.061 9	0.061 9	0.061 9
CASE	0.060 1	0.060 5	0.059 6	0.060 1
CTD	0.058 7	**0.058 7**	0.059 0	0.058 8
TF	0.064 9	0.064 9	0.064 9	0.064 9
GETD	**0.057 8**	0.058 9	**0.058 2**	**0.058 3**

表 4　基于 Weather 数据集的对比结果

Tab. 4　Comparison results based on Weather dataset

方法	Test1	Test2	Test3	*Mean*
MV	0.424 1	0.424 1	0.424 1	0.424 1
CRH	0.415 5	0.415 5	0.415 5	0.415 5
CATD	0.386 2	0.386 2	0.386 2	0.386 2
CASE	0.409 3	0.407 5	0.405 8	0.407 5
CTD	0.384 5	0.384 5	0.384 5	0.384 5
TF	0.417 2	0.417 2	0.417 2	0.417 2
GETD	**0.378 2**	**0.376 5**	**0.381 7**	**0.378 8**

2.6　不同 L 值对实验结果的影响

通过在 Restaurant 数据集上采用不同 L 值，研究该参数对实验结果准确性的影响。在这个实验中采用实验效果最好的 CTD 作为对比算法。实验结果如图 2 所示，当值低于 2%时，GETD 的错误率较高，效果不如 CTD，但当值较大时，本文的方法始终优于 CTD 算法。根据实验结果，文中将 L 设置 3%。

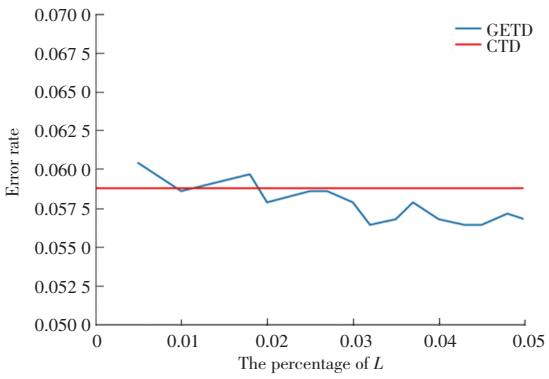

图 2 Restaurant 数据集上不同 L 值对错误率的影响

Fig. 2 Influence of different L on error rate in Restaurant dataset

3 结束语

本文采用基于异构网络的图嵌入方法解决了存在属性关联的真值发现问题。提出的模型构建了 4 个异构网络,包括源-属性值、源-源、实体属性-实体属性和实体属性-实体属性值网络。同时,通过源-属性值网络捕捉源可靠性与属性值可信度的关系、源-源网络捕捉源之间的相似性关系、实体属性-实体属性网络捕捉属性之间的关系,实体属性-实体属性值网络将建模的实体属性之间的关系体现在属性值层面上,对每个子网络采取随机梯度下降的方法来更新嵌入表示,最后根据嵌入表示来推断真值。在 2 个真实世界数据集上的实验证明了该模型的有效性。

参考文献

[1] YIN Xiaoxin, HAN Jiawei, PHILIP S Y. Truth discovery with multiple conflicting information providers on the web[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(6): 796-808.

[2] LI Qi, LI Yaliang, GAO Jing, et al. A confidence-aware approach for truth discovery on long-tail data[J]. Proceedings of the VLDB Endowment, 2014, 8(4): 425-436.

[3] LI Qi, LI Yaliang, GAO Jing, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation [C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. Snowbird, UT, USA: ACM, 2014: 1187-1198.

[4] YE Chen, WANG Hongzhi, ZHENG Kangjie, et al. Constrained truth discovery[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 205-218.

[5] ZHAO Bo, HAN Jiawei. A probabilistic model for estimating real-valued truth from conflicting sources[C]//Proc. of 10th Int. Workshop on Quality in Databases, in conjunction with VLDB 2012 (QDB'12). Istanbul, Turkey: [s.n.], 2012, 1817: 1-7.

[6] YANG Yi, BAI Quan, LIU Qing. A probabilistic model for truth discovery with object correlations [J]. Knowledge-Based Systems, 2019, 165: 360-373.

[7] FAN Weifei, GEERTS F. Foundations of data quality management [J]. Synthesis Lectures on Data Management, 2012, 4(5): 1-217.

[8] LI Yaliang, LI Qi, GAO Jing, et al. On the discovery of evolving truth[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seoul, South Korea: ACM, 2015: 675-684.

[9] MENG Chuishi, JIANG Wenjun, LI Yaliang, et al. Truth discovery on crowd sensing of correlated entities[C]//Proceedings of the 13th ACM conference on embedded networked sensor systems. 2015: 169-182.

[10] PASTERNACK J, ROTH D. Knowing what to believe (when you already know something) [C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2010: 877-885.

[11] YE Chen, LI Qi, ZHANG Hengtong, et al. AutoRepair: An automatic repairing approach over multi-source data [J]. Knowledge and Information Systems, 2019, 61(1): 227-257.

[12] LYU S, OUYANG Wentao, WANG Yongqing, et al. Truth discovery by claim and source embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(3): 1264-1275.

(上接第 8 页)

[19] EBRAHIMI M, DANESHTALAB M, FARAHNAKIAN F, et al. Congestion-aware learning model for highly adaptive routing algorithm in on-chip networks[C]//2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip. Lyngby, Denmark: IEEE, 2012: 19-26.

[20] FARAHNAKIAN F, EBRAHIMI M, DANESHALAB M, et al. Optimized q-learning model for distributing traffic in on-chip networks [C]//2012 IEEE 3rd International Conference on Networked Embedded Systems for Every Application (NESEA). Liverpool, UK: IEEE, 2012: 1-8.

[21] SAMALA J, TAKAWALE H, CHOKANI Y, et al. Fault-tolerant routing algorithm for mesh based NoC using reinforcement learning[C]//2020 24th International Symposium on VLSI Design and Test (VDAT). Bhubaneswar, India: dblp, 2020: 1-6.

[22] REZA M F. Deep reinforcement learning for self-configurable NoC[C]//2020 IEEE 33rd International System-on-Chip Conference (SOCC). Las Vegas, Nevada, USA: IEEE, 2020: 185-190.

[23] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction[M]. Cambridge: MIT Press, 1998: 182-187.

[24] GLASSC J, NI L M. Maximally fully adaptive routing in 2d meshes[C]// International Conference on Parallel Processing. St. Charles, IL, USA: dblp, 1992: 101-104.

[25] FAN Renshi, DU Gaoming, XU Pengfei, et al. An adaptive routing scheme based on Q-learning and real-time traffic monitoring for network-on-chip [C]//2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID). Xiamen, China: IEEE, 2019: 244-248.