计算机科学
COMPUTER SCIENCE

一种基于变分多跳图注意力编码器的深层协同真值发现

张国昊, 王轶, 周喜, 王保全

引用本文

张国昊, 王轶, 周喜, 王保全. 一种基于变分多跳图注意力编码器的深层协同真值发现[J]. 计算机科学, 2024, 51(3): 109-117.

ZHANG Guohao, WANG Yi, ZHOU Xi, WANG Baoquan. Deep Collaborative Truth Discovery Based on Variational Multi-hop Graph Attention Encoder [J]. Computer Science, 2024, 51(3): 109-117.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200063

一种基于变分多跳图注意力编码器的深层协同真值发现

张国昊　王　轶　周　喜　王保全
中国科学院新疆理化技术研究所　乌鲁木齐 830011
中国科学院大学　北京 100049
新疆民族语音语言信息处理实验室　乌鲁木齐 830011
(zhangguohao20@mails.ucas.ac.cn)

摘　要　大数据时代,数据价值的释放经常需要融合多源数据,数据冲突成为这一过程中无法避免的关键问题。为了从冲突数据中筛选出真实声明以及可靠数据源,研究人员提出了真值发现方法。然而,现有的真值发现大多注重数据源与声明之间的直接协同信息,忽略了更深层的间接协同与对抗信息,导致不足以表达出数据源与声明的特征。针对此问题,提出了基于变分多跳图注意力编码器的真值发现方法(TD-VMGAE),基于数据源与声明之间的包含关系构建二分图网络,采用多跳图注意力层为每个节点表征汇聚间接协同信息以及对抗信息,并设计真值发现变分自编码器,抽取节点表征中所需的分类分布,对数据源和声明进行协同分类。实验结果表明,所提方法在3个不同尺度的数据集中均有不错的表现,消融实验和可视化也验证了所提方法的有效性和泛化能力。

关键词:数据质量;冲突消解;真值发现;多跳图注意力;变分自编码器
中图分类号　TP391

Deep Collaborative Truth Discovery Based on Variational Multi-hop Graph Attention Encoder

ZHANG Guohao,WANG Yi,ZHOU Xi and WANG Baoquan
Xinjiang Technical Institute of Physics & Chemistry,Chinese Academy of Sciences,Urumqi 830011,China
University of Chinese Academy of Sciences,Beijing 100049,China
Xinjiang Laboratory of Minority Speech and Language Information Processing,Urumqi 830011,China

Abstract　In the era of big data,the release of data value often requires the fusion of multi-source data,and data conflict has become an inevitable key problem in this process. In order to filter out true claims and reliable sources from conflicting data,researchers have proposed truth discovery methods. However,the existing truth discovery methods pay more attention to the direct collaborative information between sources and claims,and ignore the deeper indirect collaborative and confrontational information,which is insufficient to express the characteristics of sources and claims. To solve this problem,this paper proposes a truth discovery method based on variational multi-hop graph attention encoder(TD-VMGAE). It constructs a bipartite graph network based on the inclusion relationship between sources and claims,uses a multi-hop graph attention layer to gather indirect cooperative information and antagonistic information for of each node,and a truth discovery variational auto-encoder is designed to extract the categorical distribution required in node characterization,and collaborative classification of data sources and claims is carried out. Experiments show that the proposed method has good performance in three datasets with different scales,and the effectiveness and generalization ability of the method are verified by ablation experiments and visualization.

Keywords　Data quality,Conflict resolution,Truth discovery,Multi-hop attention graph neural network,Variational auto-encoder

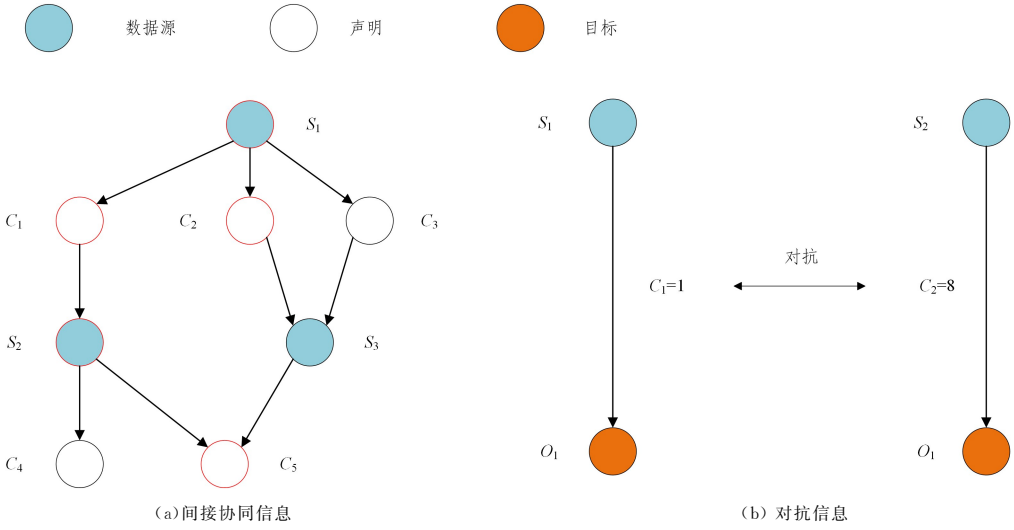
到稿日期:2022-12-12　返修日期:2023-04-04
基金项目:新疆维吾尔自治区重点实验室开放课题(2020D04050);新疆自然科学基金杰出青年基金(2022D01E04);新疆维吾尔自治区自然科学基金(2022D01B67);中科院青年创新促进会项目(2021434)
This work was supported by the Xinjiang Key Laboratory for Minority Speech and Language Information Processing(2020D04050),Natural Science Foundation for Distinguished Young Scholars of Xinjiang Uygur Autonomous Region,China(2022D01E04),Natural Science Foundation of Xinjiang Uygur Autonomous Region(2022D01B67) and Youth Innovation Promotion Association of Chinese Academy of Sciences(2021434).
通信作者:王轶(wangyi@ms.xjb.ac.cn)

1 引言

大数据时代，各类信息化服务和应用系统融入了人们生产生活的方方面面，产生了前所未有的丰富数据。但由于数据的领域来源、关注重点等各不相同，同一实体往往在信息空间中被投影为结构各异、分散自治的多源数据描述。数据融合能够对这些海量多源异构数据加以关联治理，形成真正有用的数据资源，这是进行深度挖掘分析、释放数据价值的必要手段。但在这一过程中，不同数据源存在质量差异，在数据准确性、完整性、表达方式等方面各不相同，数据冲突是无法避免的关键问题[1]。针对这一问题，学者们进行了大量研究，真值发现[2-3]逐渐成为主流的解决方案，在自然语言处理[4-7]、众包[8-13]、知识图谱构建[14-15]等众多领域任务中得到了广泛应用。

真值发现的目的是从多个数据源对目标的不同描述（即数据冲突）中筛选出真实数据及可靠数据源，其中，数据源对目标的描述统一称为声明。传统工作中，研究人员基于“质量高的数据源提供的信息更可靠，真值多的数据源质量更高”这一假设做了大量工作，大致可以分为迭代方法[3,16-19]、基于优化的方法[8,20-22]和基于概率图模型的方法[9-12,23-25]3类，这些方法主要利用了数据源与声明之间的包含关系。随着深度学习技术的发展，研究人员开始尝试引入神经网络[26-31]挖掘更深层次的关联信息，并取得了不错的效果。

其中，图网络能够很好地建模数据源与声明之间的包含、共现等直接协同信息，在真值发现任务上有更大的潜力。深入分析真值发现任务的图网络建模过程可知，仍有深层次协同信息值得继续挖掘。部分数据源与声明虽然没有直接连接（如图1(a)中的 S_1 和 C_5），但是根据间接关联对象（如图1(a)中的 S_3 和 S_2）的相似性仍会传导一定的潜在支持信息（如图1(a)中 $S_1 \xrightarrow{sim(S_1,S_2)} C_5$），我们称之为间接协同信息。另外，若不同的数据源对同一目标进行了差异化的声明（如图1(b)中的 C_1 和 C_2），声明之间产生了数据冲突，则表示数据源观察到了倾向之间的差异，我们将这种差异信息称为数据源之间的对抗信息。



(a)间接协同信息 (b) 对抗信息

图1 深层协同信息

Fig. 1 Deeper collaborative information

现有少量基于图的真值发现工作[13,26,28]虽然取得了很好的效果，但部分工作[13,26]仅关注数据源与声明之间邻域的协同信息以及声明之间的冲突现象，忽略了更深层节点的协同信息以及数据源观察倾向之间的差异产生的对抗信息，导致学习出的数据源与声明的特征区分度不明显，影响了识别效果。还有工作[28]虽然尝试使用图注意力网络挖掘更多深层信息，但是单层的邻域挖掘深度不够，并且需要半监督的标注数据支持，在海量数据场景下难以发挥作用。

为了解决上述问题，本文提出了无监督的基于变分多跳图注意力自编码器真值发现模型 TD-VMGAE。该模型以数据源和声明的包含关系为边，用 CASE[26]预训练出的数据源与声明的表征作为节点属性构建了数据源和声明之间的二分图。同时，引入残差多跳图注意力层基于节点间的相互关系学习每个节点对高阶领域的注意力，通过消息传递聚集每个节点的多跳信息，为每个节点表征汇聚了深层协同信息，并使用BPR损失函数[32]来为节点加入对抗信息。最后，参考ETM[33]的基本思想设计了真值发现变分自编码器抽取残差多跳图注意力层输出的节点表征中所需的分类特征，无监督的半平摊推理出每个节点表征代表的可靠度或者可信度。相比现有的图网络工作，本文模型考虑了数据源与声明的深层协同信息以及对抗信息，无监督地对数据源和声明进行协同分类。实验结果表明，在多个有代表性的数据集上，本文算法都可以较好地解决多源数据的冲突问题，并且，本文通过消融实验解释了所提模型的泛化能力，以及用可视化解释了每个模块处理数据时起到的作用。

本文第2章介绍真值发现相关工作；第3章给出真值发现问题以及模型相关参数的定义；第4章介绍模型利用多跳图注意力层以及真值发现变分自编码器为数据源与声明融合深层协同信息和对抗信息并推断真值的方法；第5章给出了在3个不同尺度数据集上的实验结果、消融实验以及可视化展示；最后总结全文并展望未来。

2 相关工作

2.1 传统真值发现

当前传统的方法按照计算方式的不同主要可分为3类：迭代方法、优化方法和概率方法。

迭代方法通过两阶段迭代计算数据源的可靠性和声明的可信性，第一阶段固定数据源的可靠性计算声明的可信性，第二阶段固定声明的可信性计算数据源的可靠性，直至收敛。Yin等[3]最早提出了真值发现算法，并以高可靠性数据源所提出的声明更加可行为准则，设计了最初的迭代真值发现算法TruthFinder。Pasternack等[16]认为数据源给出的声明带有一定的主观性，因此设计了一个框架，将先验信息融入真值发现算法中。Dong等[18]通过贝叶斯分析来确定信息源之间的依赖关系，并设计了一种算法来迭代地检测依赖关系，同时发现真值。

基于优化的方法通过定义优化函数，最小化每个声明与数据源可靠加权的真实主张之间的距离，来使真实值与可靠权重大的数据源的声明更接近。Li等[20]考虑了数据源提出声明可能会出现长尾现象，设计优化函数降低了小信源的影响，在长尾数据集上取得了较好的效果。Li等[21]设计了一个可以引入不同的损失函数来识别各种数据类型特征的优化框架，在异构数据上取得了不错的效果。

真值发现的概率方法通过设计一些概率模型或者概率图模型，考虑了一些可能影响数据源生成声明的潜在变量，建立了数据源和声明的联合分布模型。LTM[23]通过建模数据源两类错误(假阳性和假阴性)的生成过程，对数据源可靠性的两个不同方面建模，提升了多真值发现的一个精度效果。BCC[24]通过贝叶斯网络建模数据源的观察生成过程，使用混淆矩阵来表示每个数据源的可靠性。CBCC[9]用群体可靠性代替数据源个体可靠性，解决了因数据源声明太少而无法评估数据源可靠性的问题。BCCTD[10]通过建立半监督共聚类贝叶斯图网络，突破了混淆矩阵只能判别固定数量声明的限制，在稀疏数据上也有很好的表现。

2.2 深度学习真值发现

随着数据量越来越大，深度学习逐渐成为主流方法，基于神经网络的真值发现方法尝试从各个角度深度挖掘数据源可靠性与声明可信度的非线性关系。Chang等[4]利用Bi-GRU建模文本声明的语义信息，并利用双层注意力机制进行细粒度的真值学习，突破了传统模型无法直接应用于文本的限制。ART[27]利用自编码器抽取了数据源与声明之间的非线性关系，并在编码器与解码器中采用了贝叶斯网络，在学习过程中考虑了数据源社区关系，在抽取数据源与声明高维关系的同时增加了可解释性。

其中，少部分基于图的工作取得了不错的效果。CASE[26]分别构建了数据源与数据源、数据源与声明、声明与真值3种网络，通过概率优化将数据源、声明与真值嵌入高维空间，使可靠数据源、可信声明与真值在空间中更加接近，达到了良好的分类效果。GETD[31]在CASE[26]的基础上增加了实体属性以及实体属性集网络，进一步丰富了各种嵌入的信息。由于传统方法很大程度上会受到初始参数设置的影响，TIGE[13]基于优化函数，提出了泛化性能更好、参数更简洁的图嵌入模型。BAT[28]利用数据源与目标具体内容的表征作为节点、声明作为边，建模了二分图，通过双向注意力图神经网络从整个图中半监督地挖掘真值。

3 定义

3.1 真值发现任务

真值发现的主要目的是从多个数据源对多个目标给出的冲突声明中找出可靠数据源与每个目标的真实声明。形式化定义如下：

需要收集 T 个目标的真实数据，给定目标集合 $O=\{o_1,o_2,\cdots,o_T\}$。从 M 个数据渠道收集这些目标的数据，并将其定义为数据源集合 $S=\{s_1,s_2,\cdots,s_M\}$。同时，从 M 个数据源所收集到的关于 T 个目标的全部 N 个声明的集合定义为 $C=\{c_1,c_2,\cdots,c_N\}$。全部声明集合 C 中分为真实声明集合 $C^T=\{c_1,c_2,\cdots,c_{n_t}\}$ 以及虚假声明集合 $C^F=\{c_1,\cdots,c_{n_f}\}$。其中 $C^T\cap C^F=\varnothing$，$C^T\cup C^F=C$。

3.2 数据源-声明二分图

为了挖掘每个数据源以及声明的深层协同信息，我们以数据源与声明为节点构建了二分图，并将其作为TD-VM-GAE模型的输入，以此来传递节点之间的深层协同信息。

数据源-声明二分图定义为 $G=\{N,V,E\}$，它由一个节点集合 $N=S\cup C$、一个节点属性集 $V=VS\cup VC$ 和一个边集 E 组成。其中，节点集合 N 由数据源集合 S 以及声明集合 C 组成。VS 和 VC 分别是我们使用CASE[19]以数据源集合 S 与声明集合 C 作为基础进行预训练得到的数据源表征集合和声明表征集合，$h_{s_i}^{(m)}\in R^d$ 代表数据源 s_i 在 m 层输入的表征，$h_{c_k}^{(m)}\in R^d$ 代表声明 c_k 在 m 层输入的表征，边集 E 表示数据源与声明之间的边集合。当数据源 s_i 的观测集合 c_{s_i} 包含 c_k，表示边 (s_i,c_k) 在边集 E 中。由于数据源对于声明的重要性与声明对于数据源的重要性是不对等的，因此我们将边 (c_k,s_i) 也加入到边集 E 中。同时，根据边集 E，我们构建二阶边集合 $E^{(2)}$，当数据源 s_i 与声明 c_k 相连，c_k 又与 s_{i2} 相连，则将边 $\langle(s_i,c_k),(c_k,s_{i2})\rangle$ 加入边集 $E^{(2)}$ 中。同理，根据需要构建高阶边集 $E^{(3)}$，$E^{(4)}$ 等边集。

3.3 对抗信息

大多数目标会被多个数据源观察，产生冲突数据。当一个目标被多个数据源观察，并产生了差异化声明，我们将声明层面的冲突称为数据冲突，将数据源观察倾向层面的差异称为对抗。由于数据源之间会频繁地产生对抗与合作，两个数据源之间没有绝对的对抗，因此我们将与数据源声明冲突的声明作为该数据源的对抗信息，所有冲突声明的集合作为对抗声明集合，并在下一节将其作为对抗信息融入节点。相关定义如下：

目标数据源 s_i 只观测到了 t_j 个目标，该目标集合定义为 $O_{s_i}=\{o_1,o_2,\cdots,o_{t_j}\}$，这 t_j 个目标被所有 M 个数据源所观测到的声明集合为 $C_{o_{t_j}}=\{c_1,c_2,\cdots,c_{n_{t_j}}\}$；数据源 s_i 对 t_j 个目标所得出的 n_{s_i} 个观察的观察声明集合定义为 $C_{s_i}=\{c_1,c_2,\cdots,c_{n_{s_i}}\}$，则数据源 s_i 对目标集合 O_{s_i} 的对抗声明集合定义为

$C_{-s_i}=\{c_1,c_2,\cdots,c_{n_{-s_i}}\}$，其中，$C_{s_i}\cap C_{-s_i}=\varnothing$，$C_{s_i}\cup C_{-s_i}=C_{o_{s_j}}$。

4 基于变分多跳图注意力编码器的真值发现

为了获取深层次信息来为数据源与声明进行更好的协同分类，本文参考文献[34-35]的方法提出了 TD-VMGAE 方法。整体流程如图 2 所示，首先，以节点集合 N 为例，根据边集 E 和节点属性 V 构成数据源-声明二分图 G；其次，将构成的二分图 G 与每个数据源的对抗声明集合 C_{-s_i} 输入多跳残差注意力网络，得到汇聚了深层协同信息的节点表征 $h^{(m)}$；最后，将得到的节点表征输入真值发现变分自编码器中推测出每个目标的真值声明。

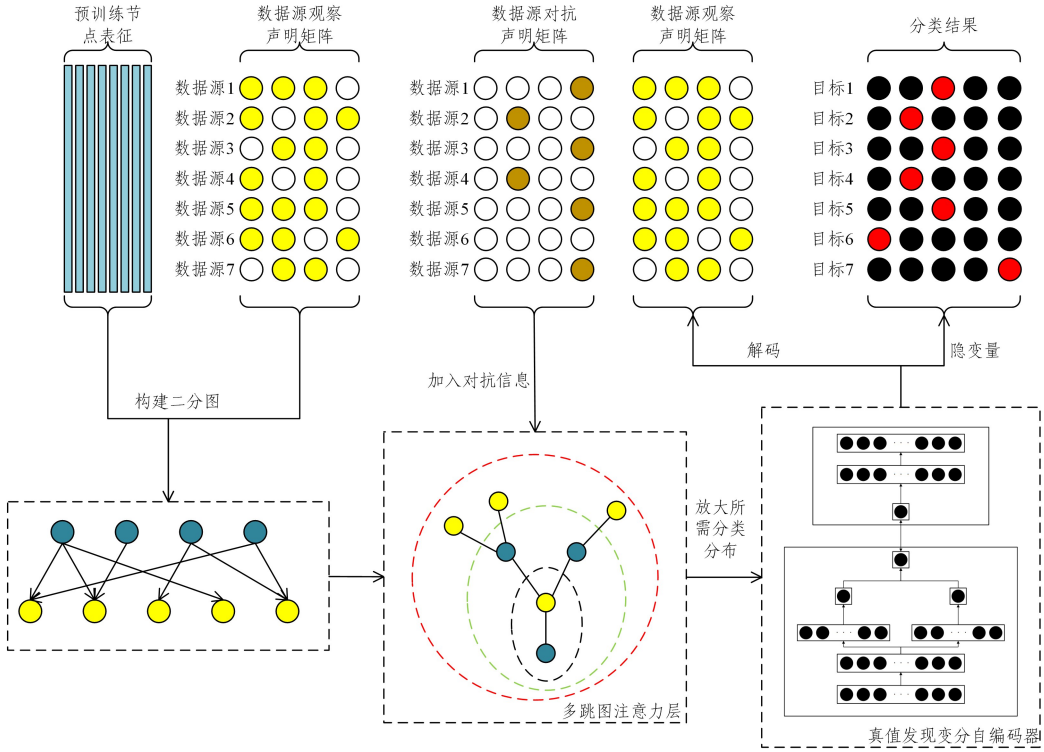


图 2 模型流程图

Fig. 2 Model flowchart

4.1 多跳残差注意力网络

4.1.1 浅层双向注意力计算

由于每个声明对数据源可靠性的影响并不相同，反之亦然，因此引入注意力机制，双向汇聚不同节点间差异化的协同信息。

在 l 层，输入是边集 E 中的图节点对，以及边所连接的节点表征 $h_{s_i}^{(l)}$ 和 $h_{c_k}^{(l)}$，这里我们使用 n_j^l 表示两种节点，$h_{n_j}^{(l)}$ 表示节点表征。为了计算 l 层中节点 n_j^l 输入到第 $l+1$ 层的表征，将所有 l 层中到 n_j^l 的消息聚合为单个消息，然后使用该消息将 n_j^l 的表征 $h_{n_j}^{(l)}$ 更新为 $h_{n_j}^{(l+1)}$。

其中，二分图中有数据源节点和声明节点两种节点。由于它们所代表的信息不同，数据源对声明的影响与声明对数据源的影响不同，因此，数据源对声明的注意力计算与声明对数据源的注意力计算函数相同、参数不同，两种注意力由以下公式计算得出：

$$a_{s\to c}^{l}=\sigma(b_{s\to c}^{(m)}(\mathrm{LeakyReLU}(W_h^{(m)}h_{s_i}^{(m)}\,\|\,W_t^{(m)}h_{c_k}^{(m)}))) \tag{1}$$

$$a_{c\to s}^{l}=\sigma(b_{c\to s}^{(m)}(\mathrm{LeakyReLU}(W_t^{(m)}h_{c_k}^{(m)}\,\|\,W_h^{(m)}h_{s_i}^{(m)}))) \tag{2}$$

两个公式结构一致，但是参数不同，$a_{s\to c}$ 和 $a_{c\to s}$ 分别代表声明对数据源的注意力和数据源对声明的注意力，σ 表示 softmax 层，$b_{s\to c}$ 和 $b_{c\to s}$ 分别表示数据源对声明的影响参数和声明对数据源的影响参数，$\|$ 表示连接算子，W_h 和 W_t 是可学习的全连接层。

以声明对数据源的注意力 $a_{s\to c}^{l}$ 为例，根据计算出来的注意力，我们构建注意力矩阵 $\boldsymbol{A}_{s\to c}^{l}$：

$$\boldsymbol{A}_{i,j}^{l}=\begin{cases}a_{i,j}^{l}, & \text{if } (n_i,n_j)\in E\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

同理可得数据源对声明的注意力矩阵 $\boldsymbol{A}_{c\to s}^{l}$。

4.1.2 深层协同信息汇聚

我们进一步计算没有直接连接的深层节点之间的注意力，通过下面的注意力扩散机制来实现这一点，如图 3 所示。

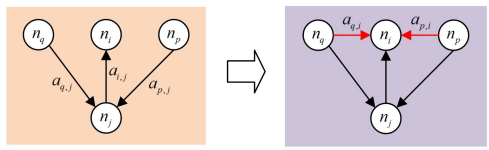


图 3 注意力扩散

Fig. 3 Attention diffusion

基于直连注意力矩阵的幂，通过图扩散计算深层节点的注意分数 $\mathcal{A}$：

$$\mathcal{A}=\sum_{i=0}^{\infty}\theta_i\boldsymbol{A}^i,\ \text{where}\ \sum_{i=0}^{\infty}\theta_i=1 \tag{4}$$

由于两种注意力的图扩散方式一致，因此我们采用 $\boldsymbol{A}$ 来统一表示注意力矩阵。其中，$\boldsymbol{A}^i$ 为注意力矩阵 $\boldsymbol{A}$ 的 i 次幂，

我们给出了路径长度为 i 的从节点 n_p 到节点 n_q 的注意力，增加了每个节点接收协同信息的深度。θ_i 为可学习注意力衰减因子，同时每个节点并不一样并且越远的节点对本节点的影响越小，即 $\theta_i>\theta_{i+1}$。在得到多跳邻居的注意分数矩阵后，节点 n_i^l 在 $l+1$ 层输出的特征 h_i^{l+1} 的计算式如下：

$$h_i^{l+1}=LN\left(h_i^l+W^lLN\left(h_i^l+\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N(i)}\mathcal{A}_{i,j}h_j^l\right)\right) \tag{5}$$

4.1.3 对抗信息融入

上述计算过程中，多跳图神经网络将节点的多层信息汇聚，可以使相似节点表征在空间中更近。为了使真假声明的表征区别更加明显，我们加入了对抗信息，采用 BPR 损失函数[32]，最大化数据源 s_i 与对抗声明 c_{-s_i} 的距离，进一步最小化数据源 s_i 与观察声明 c_{s_i} 的距离，损失函数定义如下：

$$L=-\sum_{i,j\in S}\log(\sigma(s_ic_{s_i}-s_ic_{-s_i})) \tag{6}$$

其中，s_i 表示图网络 m 层输出的数据源节点表征 $h_{s_i}^{(m)}$，同样地，c_{s_i} 表示数据源 s_i 的观察声明节点表征 $h_{c_{-s_i}}^{(m)}$，c_{-s_i} 为对抗声明节点表征 $h_{c_{-s_i}}^{(m)}$。

4.2 真值发现变分自编码器

在得到处理过的节点表征后，我们设计了真值发现变分自编码器（Truth Discovery Variational Auto-Encoder，TDVAE），以图网络输出的数据源和声明的表征作为输入，通过抽取含有所需分类分布信息的隐变量对数据源与声明进行协同分类。TDVAE 补充学习了两个潜在维度的概念：首先，它抽取了声明表征表示在高维空间中隐含的可信度分布特征，即声明为真实数据的概率；其次，它将每个数据源表征转化为潜在可靠性，即数据源提出的声明为真的概率。基于越可靠的数据源提出的声明越可信这种关系，我们将可靠性与可信度视为同一种概念，以下统一用可信度表示两种概念，并将可信度量化到[0,1]之间。TDVAE 结构如图 4 所示。

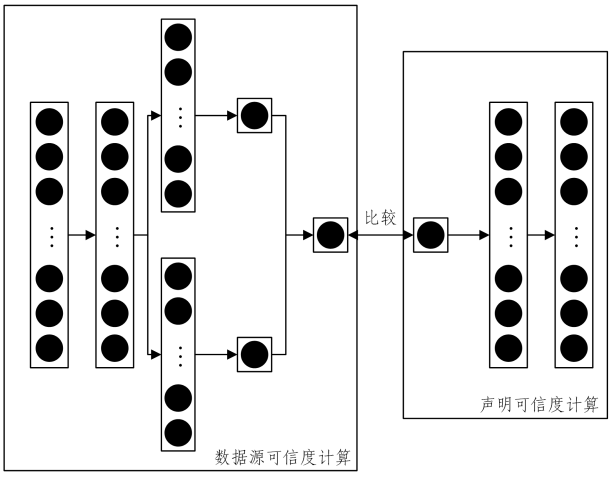


图 4 真值发现变分自编码器

Fig. 4 Truth discovery variational auto-encoder

我们将 $L\times N$ 声明表征矩阵表示为 $\boldsymbol{\rho}$，其中 L 为图网络输出的表征高维空间，N 为全部声明的数量，列 ρ^n 是声明 c_n 的嵌入。同时在空间中设置了向量 $\boldsymbol{\alpha}\in R^L$ 代表可信度 1，即完全可信向量，它是声明表征空间中的一个点。

TDVAE 的输入参数为声明表征矩阵 $\boldsymbol{\rho}$ 和完全可信向量 α。给定数据源集合 $S=\{s_1,s_2,\cdots,s_M\}$ 以及每个数据源 s_m 所提出的声明集合 $c_{s_m}=\{c_1,c_2,\cdots,c_{N_{s_m}}\}$，最大化所有数据源声明集合的对数边际似然。

$$\mathcal{L}(\boldsymbol{\rho},\boldsymbol{\alpha})=\sum_{m=1}^{M}\log p(c_{s_m}|\boldsymbol{\rho},\boldsymbol{\alpha}) \tag{7}$$

为了简化边际似然 $p(c_{s_m}|\boldsymbol{\rho},\boldsymbol{\alpha})$ 的计算，用 δ_m 来表示数据源 s_m 的可信度分布，公式转换如下：

$$p(c_{s_m}|\boldsymbol{\rho},\boldsymbol{\alpha})=\int p(\delta_m)\prod_{n=1}^{N_{s_m}}p(c_{n_{s_m}}|\delta_m,\boldsymbol{\rho},\boldsymbol{\alpha})d\delta_m \tag{8}$$

使每个声明的条件分布 $p(c_{n_{s_m}}|\delta_m,\boldsymbol{\rho},\boldsymbol{\alpha})$ 边缘化可信度分布，公式如下：

$$p(c_{n_{s_m}}|\delta_m\boldsymbol{\rho},\boldsymbol{\alpha})=\theta_m^{\mathrm{T}}\varphi_{c_{n_{s_m}}}+(1-\theta_m)^{\mathrm{T}}(1-\varphi_{c_{n_{s_m}}}) \tag{9}$$

这里，$c_{n_{s_m}}$ 只有真假两种可能的类别，θ_m 表示数据源表征转换后的可信度，即数据源提出声明为真的概率，$(1-\theta_m)$ 表示数据源提出声明为假的概率，θ_m 通过下面的变分推断，由 δ_m 可信度分布得到。$\varphi_{c_{n_{s_m}}}$ 表示声明的可信度，由计算声明表征 $\boldsymbol{\rho}$ 和完全可信向量 $\boldsymbol{\rho}$ 的相似度得出：

$$\varphi_{c_{n_{s_m}}}=\mathrm{sigmoid}(\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\alpha})|_{c_{n_{s_m}}} \tag{10}$$

为了保证模型按照预期方向收敛，我们参考 BAT[22] 使用极少量的低成本数据，将投票法选取置信度高的少量数据以及基于每轮高于一定阈值的 $\varphi_{c_{n_{s_m}}}$ 作为伪标签 $Q^*=\{c_1^*,\cdots,c_{n_Q}^*\}$，以此保证 θ_m 为真实类，将涉及引导目标的数据源用以下损失函数进行计算：

$$L=-\sum_{i,j\in S}\log(\sigma(\theta_m^{\mathrm{T}}\varphi_{c_{n_{s_m}}})) \tag{11}$$

为了计算出 $\varphi_{c_{n_{s_m}}}$，我们需要得到等式(8)中二次计算出的 $\boldsymbol{\rho}$ 和 $\boldsymbol{\alpha}$，但是直接计算积分是十分困难的。参考文献[33]的变分推断，使用平摊推理，又由 υ 参数化的变分自编码器，将数据源 s_m 的观察声明嵌入表示 h_{s_m} 输入，得到可信度分布 δ_m 的均值和方差（其中 δ_m 为高斯分布），从而可以得到 δ_m 服从于 $q(\delta_m;h_{s_m},\upsilon)$。

用 $q(\delta_m;h_{s_m},\upsilon)$ 来限定等式(8)中边际似然的对数，得到 ELBO，将其作为神经网络的参数和变分参数的优化函数，具体如下：

$$\begin{aligned}\mathcal{L}(\rho,\alpha,\nu)=&\sum_{m=1}^{M}\sum_{n=1}^{N_{s_m}}\mathbb{E}q[\log p(c_{n_{s_m}}|\delta_m,\boldsymbol{\rho},\boldsymbol{\alpha})]-\\&\sum_{m=1}^{M}KL(q(\delta_m;h_{s_m},\upsilon)\parallel p(\delta_m))\end{aligned} \tag{12}$$

式(11)的第一项表达了 δ_m 通过 $\boldsymbol{\rho}$ 和 $\boldsymbol{\alpha}$ 构成的解码器还原原始数据源观察声明集合的能力；第二项的目标是使变分自编码器输出的可信度分布 $q(\delta_m;h_{s_m},\upsilon)$ 尽量接近真实的可信度先验分布 $p(\delta_m)$。同时，最大化 ELBO 对模型参数 $(\boldsymbol{\rho},\boldsymbol{\alpha})$ 的影响相当于最大化期望的完全对数似然 $\sum_{m=1}^{M}q(\delta_m,h_{s_m}|\boldsymbol{\rho},\boldsymbol{\alpha})$。式(12)中由于期望难以计算，因此参考文献[36]的处理方法，对期望进行蒙特卡洛估计，将其转换为如下近似：

$$\begin{aligned}\widetilde{\mathcal{L}}(\boldsymbol{\alpha},\boldsymbol{\rho},\upsilon)=&\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N_s}\sum_{m=1}^{M}\log p(c_{n_{s_m}}|\delta_m^{(k)},\boldsymbol{\rho},\boldsymbol{\alpha})-\\&\sum_{m=1}^{M}KL(q(\delta_m;h_{s_m},\upsilon)\parallel p(\delta_m))\end{aligned} \tag{13}$$

其中，$\delta_m^{(k)}\sim q(\delta_m;h_{s_m},\upsilon)$，$k=1,\cdots,K$。为了无偏估计 ELBO 及其梯度，在对 $\delta_m^{(1)},\cdots,\delta_m^{(k)}$ 采样时使用了重参数化技巧，从 $q(\delta_m;c_{s_m},\upsilon)$ 中采样 $\delta_m^{(k)}$，公式如下：

$$\epsilon_m^{(k)}\sim\mathcal{N}(0,I)\text{ and }\theta_m^{(k)}=\mu_m+\sum_{m}^{\frac{1}{2}}\epsilon_m^{(k)} \tag{14}$$

其中，μ_m 和 $\sum_m^{\frac{1}{2}}$ 分别为 $q(\delta_m; c_{s_m}, \upsilon)$ 的均值和协方差，它们通过变分自编码器隐式地依赖于 υ 和 c_{s_m}。

使用 $\mathfrak{B}$ 表示数据源的批量大小，则利用数据子抽样的 ELBO 公式为：

$$\tilde{\mathcal{L}}(\alpha,\rho,\upsilon)=\frac{D}{|\mathfrak{B}|}\sum_{d\in\mathfrak{B}}\sum_{n=1}^{N_d}\sum_{s=1}^{S}\log p(c_{n_{s_m}}|\delta_m,\boldsymbol{\rho},\boldsymbol{\alpha})-\frac{D}{|\mathfrak{B}|}\sum_{d\in\mathfrak{B}}KL(q(\delta_m;h_{s_m},\upsilon)\parallel p(\delta_m)) \tag{15}$$

假设 $q(\delta_m; c_{s_m}, \upsilon)$ 和 $p(\delta_m)$ 都是高斯变量，则公式如下：

$$KL(q(\delta_m;h_{s_m},\upsilon)\parallel p(\delta_m))=\frac{1}{2}\{\mathrm{tr}(\sum_d)+\mu_d^{\mathrm{T}}\mu-\log\det(\sum_d)-K\} \tag{16}$$

计算过程中，用 h 统一表示多跳图神经网络输出的数据源表征与声明表征，$NN(h;\upsilon)$ 表示输入为 h，参数为 υ 的神经网络。具体算法如算法 1 所示。

算法 1 真值发现变分自编码器

输入：$(h, \boldsymbol{\rho}, \boldsymbol{\alpha}, C)$

输出：$(\theta, \varphi, \boldsymbol{\rho}, \boldsymbol{\alpha})$

初始化变分参数与神经网络参数

1. for iteration i=1,2,…,do
2. 计算所有声明的可信度先验 $\beta=\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\alpha}$
3. 计算所有声明的可信度 $\varphi=\mathrm{sigmoid}(\beta)$
4. 从数据从抽取出 $\mathfrak{B}$ 个数据源
5. for each source s_m in $\mathfrak{B}$ do
6. 从输入中找到 s_m 的表征 h_{s_m}
7. 计算均值 $\mu_s=NN(h_{s_m};\upsilon_\mu)$
8. 计算方差 $\sum_s=NN(h_{s_m};\upsilon_\sum)$
9. 使用式(14)采样数据源可靠度 θ_m
10. for each claim in the $C_{s_i}=\{c_1,\cdots,c_{n_{s_i}}\}$ do
11. 使用式(10)计算 $p(c_{n_{s_m}}|\delta_m,\boldsymbol{\rho},\boldsymbol{\alpha})$
12. end for
13. for each claim in the $Q^*=\{c_1^*,\cdots,c_{n_Q}^*\}$ do
14. 使用式(11)计算损失
15. end for
16. end for
17. 使用式(15)和式(16)计算 ELBO
18. 反向传播 ELBO 产生的梯度
19. 更新模型参数 $\boldsymbol{\rho}, \boldsymbol{\alpha}$
20. 更新变分参数 $(\upsilon_\mu, \upsilon_\sum)$
21. end for

5 实验

5.1 实验设计

为了验证本文工作的有效性，我们选取了 3 个具有代表性的公开数据集进行实验，并与 6 个现有的 state-of-the-art 方法进行对比。

实验数据为 3 个不同尺度的数据集，它们来自于 Zheng 等在 VLDB 上发表的真值发现系统性评测工作[37]及其在 github 上公开的数据集[1)]。其中，Duck 数据集是一组决策数据集，数据源要检查每个目标是否包含动物。脸部情感识别(FSI)是一个人脸识别任务集，数据源识别特定目标面部的情绪，并进行判断。Product 数据集中每个目标都是两段描述，数据源需要观察每组描述，并判断这两个描述是否指向同一实体。

这些数据集的对比情况如表 1 所列。从规模角度看，Duck 和 FSI 数据集分别有 4212 和 5242 个声明，两者大致相当，但 Product 数据集的规模是前两者的数倍；从数据稀疏度来看，Duck 中每个数据源都做出了 39 个声明，而 FSI 和 Product 则较为稀疏，FSI 中每个数据源都做出了 8.9 个声明，Product 中每个数据源做出了 3 个声明。

表 1 使用的 3 个数据集

Table 1 Three used datasets

Dataset	Duck	FSI	Product
目标数量	108	584	8315
数据源数量	39	27	176
声明数量	4212	5242	24945
声明/目标	39.0	8.9	3.0

对比的 6 种算法为：PM，ZC，CATD，GLAD，GTIC，TIGE。PM 基于优化模型[8]，该算法的主要思想是通过最小化数据源权值与目标真值权值之间的总加权距离，使推断的真值在分布上更接近于真实真值。CATD[20]是另一种基于优化模型的真值推理方法，它额外考虑了数据源出现的次数，认为数据源提出的声明越多，占的比重就应该越大。ZenCrowd[11]利用概率推理和众包技术构建了大规模实体链接系统，引入了可靠性和真实度的似然函数，并通过 EM 算法求解。GLAD 算法是模型[25] ZenCrowd 算法的扩展，与 ZenCrowd 不同的是，它认为观察每个目标的真值的难度都是不相同的，在推断真值的过程中额外考虑了观察目标的难度。GTIC 使用聚类算法[12]的基础真值推断，通过聚类而不是传统的概率方法来获取真值，在特征计算和聚类方法方面与其他对比算法有很大的不同。TIGE[13]以目标和数据源为节点、数据源提出的声明为边构建了图结构，并使用优化函数基于图结构进行了真值推断，在多个数据集上都有很好的表现。

本实验中的算法均为无监督真值发现算法。本实验是在一台 16GB 内存的一体机上进行的，显卡为 K80。实验代码主要使用 python 语言完成，实验环境为 Ubuntu 系统下的 python 3.8.7。我们基于 Pytorch 1.12 构建了 TD-VMGAE。

5.2 准确率

在实验中，我们堆叠了 24 层多跳图注意力层，每层中间加上两层 ELU 层，再将输出的嵌入表示作为真值发现变分自编码器的输入。学习率设置为 0.001。

准确性测试结果如表 2 所列。观察对比的 6 种方法在 3 个数据集上的表现发现，虽然它们在不同场景下各有优势，但泛化能力较差，难以在不同尺度的任务中发挥稳定效果。其原因主要是多种方法针对数据源的长尾分布、观察难度等任务特征做出了针对性改进，在变换场景后难以发挥作用。值得注意的是，GLAD 和 TIGE 虽然在 3 个数据集上都有较稳定的效果，但模型对信息的抽取不够全面导致平均效果

1) https://github.com/TsinghuaDatabaseGroup/CrowdTI/tree/master/truth_inference_crowd/datasets

没有优势。本文方法在两个数据集上取得了最好的效果，并在另一个数据集上接近最优，平均准确率相较于现有方法有了较大提升。可以看出，本文方法具有一定的泛化能力，可以在一定程度上适应不同数据规模以及数据稀疏度的场景。

表 2 所有方法在 3 个数据集上的准确率

Table 2 Accuracy of all methods on three datasets

Dataset	PM	CATD	GLAD	GTIC	ZC	TIGE	TD-VMGAE
Duck	0.7870	0.7778	0.7593	0.8055	0.7222	0.7870	**0.8333**
FSI	0.5976	0.6164	0.6284	0.6524	0.6284	0.6061	**0.6644**
Product	0.8981	0.8266	0.9224	0.6469	**0.9280**	0.8784	0.9123
Average	0.7609	0.7403	0.7700	0.7016	0.7595	0.7572	**0.8033**

5.3 消融实验

为了探究两类深层协同信息的作用，我们设计了消融实验。以图注意力网络为基线，分别加入对抗信息(BPR Loss)以及多跳图注意力网络(Multi-Hop Attention)，实验结果如表 3 所列。

表 3 消融实验

Table 3 Ablation experiment

Dataset	Duck	FSI	Product
Attention	0.7407	0.6472	0.8536
+BPR	0.8333	0.6455	0.8888
+Multi-Hop	0.7593	0.6541	0.9020
TD-VMGAE	0.8333	0.6644	0.9123

从实验结果可以看出，对抗信息对于 2 个数据集都有提升效果，但是在数据稀疏的 FSI 中效果开始下降。深层协同信息在数据量更大、更加稀疏的 Product 数据集上效果更明显，但是在数据源提出较多声明的 Duck 中挖掘到的信息并不多，效果并不明显。分析原因如下：

当数据集中每个数据源提出了较多的声明，如 Duck 数据集，这会使数据源之间以及声明之间频繁产生对抗，对抗信息可以很好地丰富节点的表征。但此时节点邻域信息丰富，浅层的协同信息就足以表达节点表征信息，深层协同信息在丰富表征上的作用并不明显。随着数据源对目标平均提出的声明量减少，对抗以及邻域信息开始减少，例如 FSI 数据集，此时深度协同信息对丰富节点表征的帮助开始加强，对抗中蕴含的信息开始不足。但随着数据量增大，例如 Product 数据集，虽然数据源对目标平均提出声明量进一步减少，但由于数据源数量增多，对抗信息也开始丰富起来。从实验中可以看出，两种信息可以在不同场景下起到作用，且相互补充，一定程度上解释了模型的泛化能力。

5.4 数据源与声明表征可视化

本节对多 TD-VMGAE 各个阶段学习到的表征进行分析，并通过可视化探索每个模块起到的作用。以 Duck 数据集为例，我们将几个阶段所学到的数据源向量以及声明向量使用 T-SNE 进行降维，并将其嵌入空间可视化。

使用 CASE[19] 预训练的表征可视化如图 5 所示，红色点为真实声明，黑色点为虚假声明，蓝色点为数据源。从图中可以看出，该阶段模型所输出的所有表征在空间中比较聚集，不能很好地观察到虚假声明与真实声明之间的区别以及不同数据源之间的区别。

在加入多跳图注意力层后，节点表征融入了深层协同信息以及对抗信息。图 6 展示了两个数据源与其观察声明的向量空间表示。其中，红框标出的是数据源，与数据源同种颜色的点为该数据源观察到的声明。可以看出，融入更多信息后，空间中数据源表征与其观察声明表征开始靠近，模型学习到了数据源的观察倾向。图 7 展示了全部数据源视角下的表征可视化，可以看出真值与虚假声明、可信数据源与不可信数据源都有了明显区分度。相比图 5，模型在加入信息后真假声明及质量相近的数据源之间出现了聚类趋势。

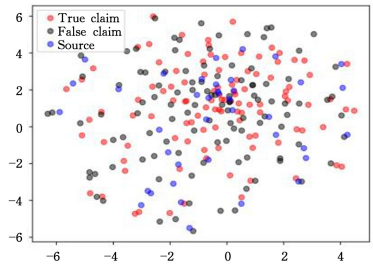


图 5 预训练表征的可视化(电子版为彩图)

Fig. 5 Visualization of pre-training representations

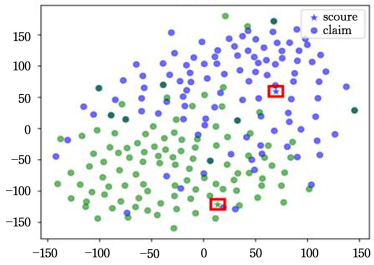


图 6 数据源与其观察声明可视化(电子版为彩图)

Fig. 6 Sources-Claims visualization

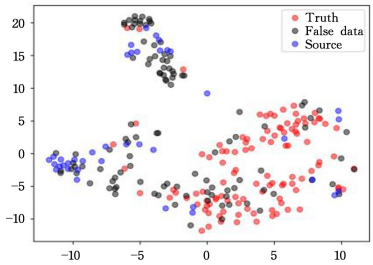


图 7 汇聚深层协同信息表征的可视化

Fig. 7 Visualization of deep collaborative information representations convergence

图 8 为加入真值发现变分自编码器后，全部声明的分类可视化结果，其中 **topic** 为完全可信向量。从图中可以看出，自编码器放大了真实声明和虚假声明的区别，真实声明更加靠近完全可信向量，而虚假声明在远离。真值发现变分自编码器抽取出了我们所需要的分类特征，相较于

图 7，声明表征在向量空间中开始分为真实与虚假两类。

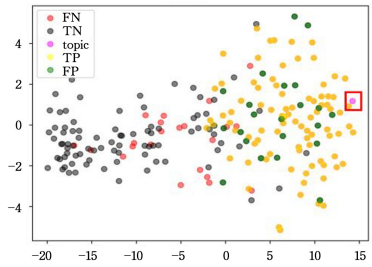


图 8 分类效果可视化

Fig. 8 Visualization of classification effects

结束语 本文提出了基于变分多跳图注意力自编码器的真值发现方法 TD-VMGAE，学习了数据源以及声明更深层次的关系，融入了其表征向量的特征，并在大多数据集上都有较好的表现，具备一定的泛化能力。在未来的工作中，我们将针对动态场景下的真值发现问题继续优化模型，使模型可以对新加入的数据源与声明进行增量学习。

参 考 文 献

[1] MENG X F, CI X. Big Data Management: Concepts, Techniques and Challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.

[2] LI Y, GAO J, MENG C, et al. A survey on truth discovery[J]. ACM Sigkdd Explorations Newsletter, 2016, 17(2): 1-16.

[3] YIN X, HAN J, YU P S. Truth Discovery with Multiple Conflicting Information Providers on the Web[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 6(20): 796-808.

[4] CHANG C, CAO J J, ZHENG Q B, et al. Truth Discovery from Text Data by Bi-GRU with Attention Mechnism[J]. Journal of Chinese Information Processing, 2020, 34(2): 46-55.

[5] SABETPOUR N, KULKARNI A, XIE S, et al. Truth discovery in sequence labels from crowds[C]// 2021 IEEE International Conference on Data Mining(ICDM). IEEE, 2021: 539-548.

[6] CHANG C, CAO J, ZHENG Q, et al. Anunsupervised approach of truth discovery from multi-sourced text data[J]. IEEE Access, 2019, 7: 143479-143489.

[7] YE C, WANG H, LU W, et al. Deep truth discovery for pattern-based fact extraction[J]. Information Sciences, 2021, 580: 478-494.

[8] AYDIN B, YILMAZ Y, LI Y, et al. Crowdsourcing for multiple-choice question an-swering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2014: 2946-2953.

[9] VENANZI M, GUIVER J, KAZAI G, et al. Community-based bayesian aggregation models for crowdsourcing[C]// Proceedings of the 23rd International Conference on World Wide Web. 2014: 155-164.

[10] DU Y, SUN Y E, HUANG H, et al. Bayesian co-clustering truth discovery for mobile crowd sensing systems[J]. IEEE Transactions on Industrial Informatics, 2019, 16(2): 1045-1057.

[11] DEMARTINI G, DIFALLAH D E, CUDRÉ-MAUROUX P. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking[C]// Proceedings of the 21st International Conference on World Wide Web. 2012: 469-478.

[12] ZHANG J, SHENG V S, WU J, et al. Multi-class ground truth inference in crowdsourcing with clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 28(4): 1080-1085.

[13] ZHOU L, ZHUO X, WU G, et al. Research on Crowdsourcing Truth Inference Method Based on Graph Embedding[C]// 2021 IEEE International Conference on Big Knowledge(ICBK). IEEE, 2021: 206-213.

[14] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: A web-scale approach to proba-bilistic knowledge fusion [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 601-610.

[15] DONG X L, GABRILOVICH E, HEITZ G, et al. From data fusion to knowledge fusion[J]. Proceedings of the VLDB Endowment, 2014, 7(10): 881-892.

[16] PASTERNACK J, ROTH D. Knowing what to b-elieve(when you already know something)[C]// Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). 2010: 877-885.

[17] LI X, DONG X L, LYONS K, et al. Truth Finding on the Deep Web: Is the Problem Solved? [J]. Proceedings of the VLDB Endowment, 2012, 6(2): 97-108.

[18] DONG X L, BERTI-EQUILLE L, SRIVASTAVA D. Integrating conflicting data: the role of source dependence[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 550-561.

[19] GALLAND A, ABITEBOUL S, MARIAN A, et al. Corroborating information from disagreeing views[C]// Proceedings of the third ACM International Conference on Web Search and Data Mining. 2010: 131-140.

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

[21] LI Q, LI Y, GAO J, 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. 2014: 1187-1198.

[22] LI Y, LI Q, GAO J, et al. On the discovery of evolving truth [C]// Proceedings of the 21th ACM Sigkdd International Conference on knowledge Discovery and Data Mining. 2015: 675-684.

[23] ZHAO B, RUBINSTEIN B I P, GEMMELL J, et al. A Bayesian approach to discovering truth from conflicting sources for data integration[J]. Proceedings of the VLDB Endowment, 2012, 5(6): 550-561.

[24] KIM H C, GHAHRAMANI Z. Bayesian classi-fier combination [C]// Artificial Intelligence and Statistics. PMLR, 2012: 619-627.

[25] WHITEHILL J,RUVOLO P,WU T,et al. Who-se vote should count more:optimal integration of labels from labelers of unknown expertise[C]// Proceedings of the 22nd International Conference on Neural Information Processing Systems. 2009:2035-2043.

[26] LYU S,OUYANG W,SHEN H,et al. Truth discovery by claim and source embedding[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017:2183-2186.

[27] YANG J,TAY W P. An unsupervised Bayesian neural network for truth discovery in social- networks[J]. IEEE Transactions on Knowledge and Data Engineering,2021,34(11):5182-5195.

[28] LIU J,TANG F,HUANG J. Truth Inference with Bipartite Attention Graph Neural Network from a Comprehensive View [C]// 2021 IEEE International Conference on Multimedia and Expo(ICME). IEEE,2021:1-6.

[29] CAO J J,CHANG C,ZHENG Q B,et al. Truth discovery method for multi-source text data[J]. Journal of National University of Defense Technology,2022,44(4):172-179.

[30] CHANG C,CAO J J,ZHENG Q B,et al. Unsupervised Multi-Attributes Truth Discovery with Deep Neural Network[J]. Computer Integrated Manufacturing Systems,2020,37(11):270-274.

[31] LU H,FANG X S,SI S X,et al. a graph embedding model for correlation aware truth discovery[J]. Intelligent Computer and Applications,2022,12(10):9-14.

[32] RENDLE S,FREUDENTHALER C,GANTNER Z,et al. BPR:Bayesian personalized ranking from implicit feedback[C]// Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. 2009:452-461.

[33] DIENG A B,RUIZ F J R,BLEI D M. Topic modeling in embedding spaces[J]. Transactions of the Association for Computational Linguistics,2020,8:439-453.

[34] VELICKOVIC P,CUCURULL G,CASANOVA A,et al. Graph attention networks[J]. arXiv:1701,10903,2017.

[35] WANG G,YING R,HUANG J,et al. Multi-hop attention graph neural network[J]. arXiv:2009. 14332,2020.

[36] KINGMA D P,WELLING M. Auto-Encoding Variational Bayes [J]. arXiv:1312. 6114,2014.

[37] ZHENG Y,LI G,LI Y,et al. Truth inference in crowdsourcing:Is the problem solved? [J]. Proceedings of the VLDB Endowment,2017,10(5):541-552.

ZHANG Guohao, born in 1997, postgraduate. His main research interests include big data governance and so on.

WANG Yi, born in 1986, Ph.D professor, is a senior member of CCF(No. 98372S). His main research interests include big data governance and block chain applications.

(责任编辑:何杨)