doi:10.11887/j.cn.202204019 http://journal.nudt.edu.cn

多源文本数据真值发现方法*

曹建军[1],常 宸[2],陶嘉庆[1,3],翁年凤[1],蒋国权[1]
(1. 国防科技大学 第六十三研究所,江苏 南京 210007;2. 陆军工程大学 指挥控制工程学院,江苏 南京 210007;
3. 南京工业大学 工业工程系,江苏 南京 211800)

摘 要:针对传统真值发现算法无法直接应用于文本数据的问题,提出基于深度神经网络面向多源文本数据的真值发现算法(NN_Truth)。根据文本答案多因素性、词语使用多样性以及文本数据稀疏性等特点,将"数据源-答案"向量作为网络输入,识别答案真值向量作为网络输出,依据真值发现的一般假设,无监督学习各数据源答案向量间关联关系,并最终获得答案真值。实验结果表明,该算法适用于文本数据真值发现场景,较基于检索的方法及传统真值发现算法效果更优。

关键词:数据质量;真值发现;神经网络;文本挖掘
中图分类号:TP311 **文献标志码**:A **文章编号**:1001-2486(2022)04-172-08

Truth discovery method for multi-source text data

CAO Jianjun[1], *CHANG Chen*[2], *TAO Jiaqing*[1,3], *WENG Nianfeng*[1], *JIANG Guoquan*[1]
(1. The Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China;
2. Command and Control Engineering College, Army Engineering University, Nanjing 210007, China;
3. Department of Industrial Engineering, Nanjing Tech University, Nanjing 211800, China)

Abstract: In order to solve the problem that the traditional truth discovery algorithm cannot be applied to text data directly, a truth discovery algorithm(NN_Truth) for text data based on deep neural network was proposed. For the features of multifactorial property of text answers, the diversity of word usages, and the sparsity of the text data, the "source-answer" vector was used as the network input, and the truth vector was recognized as the network output. The relationship between answers from each source could be unsupervised learned according to general hypothesis of truth discovery, and finally obtained the truth. The experiment results show that the proposed algorithm is suitable for text data truth discovery, and it is better than the retrieval methods and traditional truth discovery algorithm.

Keywords: data quality; truth discovery; neural network; text mining

大数据时代,随着信息技术的发展,互联网信息量呈爆炸式增长,但开放多源的互联网使得不同数据源所提供的信息有所差别,不仅存在大量虚假和错误的信息,而且还存在许多恶意的数据源。谣言和低质量信息通过这些恶意数据源向外界传播,严重影响对正确信息的判断[1]。如今,数据质量问题日益严重,如何将正确信息从这些低质量数据中筛选出来是一个意义重大且具有挑战性的研究。

真值发现是解决数据质量问题的重要方法,研究从不同数据源提供的关于多个真实对象的大量冲突描述信息中,为每一个真实对象找出最准确的描述。传统的真值发现算法主要基于两个假设:若数据源提供越多的可信事实,则该数据源越可靠;若数据源的可靠性越高,则该数据源提供的事实越可信。根据这两个基本假设,传统真值发现算法可分为三类。一是基于迭代的方法:基于迭代的方法用简单函数来表达数据源可靠度与观测值可信度之间的关系,迭代计算真值和数据源可靠度直至损失收敛[2]。二是基于优化的方法:基于优化的方法与基于迭代的方法类似,首先来通过假设条件设置目标函数,再优化目标函数来求解真值[3-4]。三是基于概率图模型的方法:该方法假设观测值的实际分布情况服从概率分布,通过参数估计和数据采样对真值进行估算[5]。

随着信息时代的推进,数据的形式在不断丰富,多样化的数据也给真值发现带来新的挑战。如今,用户可以通过访问互联网平台的公开信息

* **收稿日期**:2020-11-24
基金项目:国家自然科学基金资助项目(61371196);中国博士后科学基金资助项目(20090461425);中国博士后科学基金特别资助项目(201003797)
作者简介:曹建军(1975—),男,山东郓城人,副研究员,博士,硕士生导师,E-mail:caojj@nudt.edu.cn

寻找某个特定问题的答案,但这些答案大部分由互联网用户所提供而并非由专家提供,因此存在错误和冲突的答案。这些答案大多以文本的形式发布在互联网平台,如何克服文本数据所特有的自然语言特性对真值发现的影响,使得真值发现在文本数据领域有了新的挑战。

首先,文本数据具有词语使用多样性的特性,用户提供的答案可能会表达与正确答案关键词非常相似的语义,例如 exhausted 和 fatigue 都可表达疲惫的含义,但由于传统真值发现算法缺少对文本语义信息的充分挖掘,会将它们视为两个完全不同的答案。其次,传统真值发现算法在对结构化数据进行真值发现时,根据数据源的众多观测值对数据源可靠度进行评估,而在文本数据真值发现场景中,大量网络用户可以对同一问题进行回答,但是同一用户回答的问题却很少,数据稀疏性增大了对用户可靠度进行评估的难度,所以传统真值发现算法并不适用于文本数据真值发现场景。

对于文本数据真值发现,现有方法都是通过简化问题,从粗粒度角度对文本数据进行分析,判断互联网上的文本数据是否为真,也就是将问题转化为二分类的问题进行求解。Broelemann 等将受限玻尔兹曼机隐含层应用在真值发现场景之中,通过学习真值概率分布判断真值,但是受限玻尔兹曼机本身特性的限制,也只能将问题转换为二值属性的问题[6]。Sun 等提出了一种基于合同的个性化隐私保护激励机制,用于众包问答系统中的真理发现,该机制为具有不同隐私需求的工人提供个性化付款,以补偿隐私成本,同时确保准确的真值发现[7]。Li 等提出一种适用于少量观测值的移动众包真值发现算法,通过重复使用各数据源的观测值,挖掘数据源间的相关性[8]。Marshall 等首次运用神经网络解决真值发现问题,利用全连接神经网络学习数据源可靠度与观测值可信度间的关系,并将数据源和观测值信息输入网络进行真值发现[9]。Li 等将长短时记忆网络用于真值发现,将数据源可靠度矩阵和对象属性值作为输入,输出观测值为真的概率,最小化真值与观测值间的距离加权,使得网络参数达到最优[10]。

与传统真值发现算法不同,本文提出基于深度神经网络面向多源文本数据的真值发现算法,将“数据源 - 答案”向量输入神经网络,通过训练神经网络自主学习答案语义关系,输出答案的可信度矩阵。所提算法降低了数据稀疏性对评估数据源可靠度的影响,并且解决了传统真值发现算法强假设数据分布而导致真值发现效果不佳的问题。

1　问题定义

考虑文本数据真值发现的一般模型,给定问题集合 $Q=\{q_i|i=1,2,\cdots,M\}$,其中 q_i 表示第 i 个问题;用户集合 $U=\{u_j|j=1,2,\cdots,N\}$,其中 u_j 表示第 j 名用户;每个问题 q_i 由 N_{q_i} 名用户回答,并构成该问题的候选答案集合 $A_{q_i}=\{a_k^{q_i}|k=1,2,\cdots,N_{q_i}\}$,解决从 N_{q_i} 名用户提供的众多文本答案中找到每个问题 q_i 的最佳答案。表 1 为不同用户对“What are the symptoms of flu?”进行回答的实例。

表 1　不同用户的回答实例

Tab. 1　Example of answers from different users

问题	What are the symptoms of flu?
用户 1	People will feel cold and cough, sometimes they will feel exhausted.
用户 2	The symptoms of flu are fever and freezing.
用户 3	Maybe chills, cough, and fatigue.

由表 1 可知,对于同一问题,3 名用户分别给出了不同的答案。首先从文本数据细粒度的角度分析,每名用户所提供的答案包含了正确答案的不同关键因素,具有部分正确率。然后通过神经网络对答案语义进行提取,更准确地对答案间的关系进行度量。本文旨在通过充分运用神经网络的强表达能力来挖掘文本自然语言特性,学习用户答案间的关系,为问题寻求最优答案。

2　算法模型与分析

NN_Truth 文本数据真值发现算法共分为3 个步骤:第一步对文本进行语义表征,挖掘文本的自然语言特性,将文本表征为多维向量;第二步利用神经网络进行文本向量的真值发现,通过网络优化,最终依据众多答案向量计算识别真值向量;第三步根据各答案向量与识别真值向量的相似度计算各个答案的分数并排序。

2.1　文本的语义表征

图 1 给出了对各用户提供的文本答案进行语义表征的示意图。

如图 1 所示,对于问题 q_i,将用户及答案文本用实值向量表示,通过文本的语义表征,得到“数

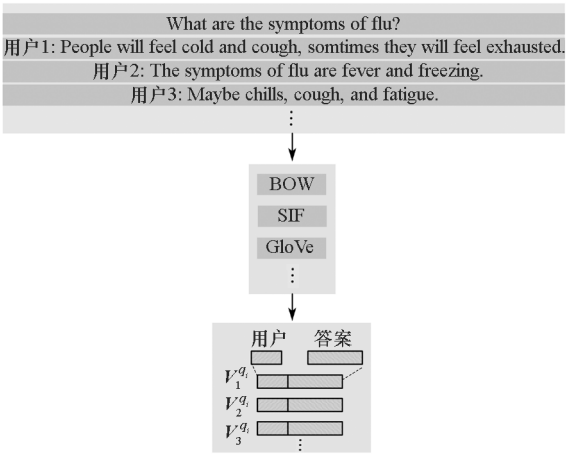


图 1 文本答案的语义表征

Fig. 1 Semantic representation for text answers

据源 - 答案"向量集合 $\{\boldsymbol{V}_k^{q_i} | k=1, 2, \cdots, N_{q_i}\}$,任一向量 $\boldsymbol{V}_k^{q_i}$ 包含了该答案的来源及内容信息,即 $\boldsymbol{V}_k^{q_i} = \boldsymbol{Y}_k^{q_i} + \boldsymbol{\Gamma}_k^{q_i}$。$\boldsymbol{Y}_k^{q_i} = \langle \gamma_{k1}^{q_i}, \gamma_{k2}^{q_i}, \cdots, \gamma_{kL}^{q_i} \rangle$ 为数据源向量,L 为数据源向量的维度,$\boldsymbol{\Gamma}_k^{q_i} = \langle \tau_{k1}^{q_i}, \tau_{k2}^{q_i}, \cdots, \tau_{kL}^{q_i} \rangle$ 为答案向量,P 为答案向量的维度。采用以下 4 种模型对用户及其答案进行语义表征:

词袋(bag-of-word,BoW)模型:以词为最小单元,将用户答案看作是词的集合,忽略词序及句法信息。将答案表示为一个多维向量(维度为答案中的词表大小)。向量中某个维度的值为 1 代表当前词存在于词表中,不存在则为 0。

词频 - 逆文档频次(term frequency-inverse document frequency,TF-IDF)算法:基于分布假说"上下文相似的词,其语义也相似",能够反映一个词的重要程度,模型假设词与词之间相互独立。

全局向量的词嵌入(global vectors for word representation,GloVe):将答案所包含的关键词使用 GloVe 工具进行向量化,之后使用答案内关键词向量的平均值作为答案向量。此方法获得的句向量与单词的顺序无关,且所有单词具备相同的权重。

平滑逆频率[11](smooth inverse frequency,SIF):对答案中的每个词向量,乘以一个权重,出现频率越高的词,其权重越小,计算句向量矩阵的第一个主成分,并让每个句向量减去它在第一主成分上的投影,对句向量进行修正。

在以上 4 种语义表征方法中,BoW 与 TF-IDF 方法不对答案中关键词的语义相似性进行细粒度度量,对答案的准确性要求较高,要求用户提供准确一致的答案关键因素。GloVe 与 SIF 向量由于包含了答案中关键词的语义信息,能够对答案进行更加细粒度的度量,有效克服了词语使用多样性带来的影响,适用于比较开放和主观的问题。

2.2 真值发现

图 2 为利用答案向量通过神经网络进行真值发现的示意图。

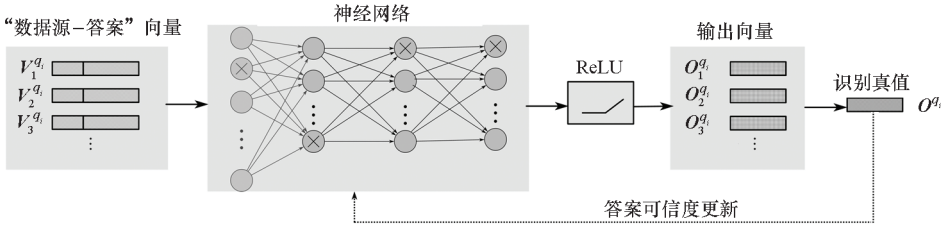


图 2 真值发现过程

Fig. 2 Process of truth discovery

如图 2 所示,将进行文本语义表征后问题 q_i 的"数据源 - 答案"向量集合 $\{\boldsymbol{V}_k^{q_i} | k=1, 2, \cdots, N_{q_i}\}$ 输入网络,通过神经网络 3 层隐藏层,计算得到输出向量集合 $\{\boldsymbol{O}_k^{q_i} | k=1, 2, \cdots, N_{q_i}\}$。输出向量经由修正线性单元(rectified linear unit,ReLU)激活函数处理后,再平均所有节点的输出得到向量 $\boldsymbol{O}^{q_i}$ 作为答案 q_i 的识别真值,网络输入层节点个数为"数据源 - 答案"向量的维度 $L+P$,输出层节点个数为答案向量维度 P。通过反向传播不断更新识别真值可信度,神经网络参数迭代更新至最优值时,得到最优化的识别真值向量。

文本数据真值发现基于两个假设:①数据源的可靠度越高,则其提供的答案相似度越高[12];②问题答案的真值情况应该与各数据源提供的观测值尽可能地接近。根据这两个假设,定义模型损失函数如式(1)所示。

$$f_{\text{loss}} = \sum_{k=1}^{N_{q_i}} d(\theta; \boldsymbol{O}^{q_i}, \boldsymbol{\Gamma}_k^{q_i}) + \frac{1}{2}\|\boldsymbol{W}\|^2 \quad (1)$$

式中:θ 表示模型中的全部参数;$\boldsymbol{O}^{q_i}$ 为模型输出结果,即完成本次迭代训练输出的识别真值;$\boldsymbol{\Gamma}_k^{q_i}$ 为 q_i 的第 k 个答案的句向量表示;$d(\theta; \boldsymbol{O}^{q_i}, \boldsymbol{\Gamma}_k^{q_i})$ 为数据源提供的答案向量与网络输出的识别真值向

量之间的距离;$\frac{1}{2}\|\boldsymbol{W}\|^2$ 为正则项,使权重更加接近原点。该损失函数的优化目标为各数据源提供的答案向量与识别真值向量距离之和最小。

对于 GloVe 与 SIF 文本特征,根据表征向量有正有负的特性,$d(\theta;\ \boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i})$ 定义为向量 $\boldsymbol{O}^{q_i}$ 与 $\boldsymbol{\Gamma}_k^{q_i}$ 的规范化后的余弦距离,用式(2) 计算。

$$d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i}) = \frac{1}{\pi}\arccos\left(\frac{\boldsymbol{O}^{q_i}\cdot\boldsymbol{\Gamma}_k^{qi}}{\sqrt{\boldsymbol{O}^{q_i}}\ \sqrt{\boldsymbol{\Gamma}_k^{qi}}}\right) \quad (2)$$

式中,$\boldsymbol{O}^{q_i}\cdot\boldsymbol{\Gamma}_k^{qi}$ 表示向量 $\boldsymbol{O}^{q_i}$ 与 $\boldsymbol{\Gamma}_k^{qi}$ 的内积,$\sqrt{\boldsymbol{O}^{q_i}}$ 与 $\sqrt{\boldsymbol{\Gamma}_k^{qi}}$ 分别表示向量 $\boldsymbol{O}^{q_i}$ 与 $\boldsymbol{\Gamma}_k^{qi}$ 的模。该距离适用于向量系数同时有正值与负值的情况,$d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i}) \in [0,\ 1]$。当 $d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i}) = 0$ 时,两向量的夹角为 0°,它们的指向完全相同,此时距离最小;当 $d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i}) = 1$ 时,两向量指向完全相反,此时距离最大。

对于 BoW 与 SIF 文本特征,网络增加 σ 层,将最后一层的输出 $\boldsymbol{O}^{q_i} = \langle O^{q_{i1}},\ O^{q_{i2}},\ \cdots,\ O^{q_{ip}}\rangle$ 映射到[0,1] 范围。$d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i})$ 定义为 $\boldsymbol{O}^{q_i}$ 与 $\boldsymbol{\Gamma}_k^{q_i}$ 的交叉熵损失,度量识别真值与数据源观测值分布的接近程度,用式(3) 计算,式(4) 为 σ 公式。

$$d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i}) = \frac{1}{P}\sum_{p=1}^{P}\left[-\boldsymbol{\Gamma}_k^{q_i}\lg(\sigma(\boldsymbol{O}^{q_i}))\right] - (1-\boldsymbol{\Gamma}_k^{q_i})\lg(\sigma(1-\boldsymbol{O}^{q_i})) \quad (3)$$

$$\sigma(\boldsymbol{O}^{q_i}) = \frac{1}{1+\mathrm{e}^{-\boldsymbol{O}^{q_i}}} \quad (4)$$

式(3) 中,P 表示文本答案向量的维度,交叉熵损失越小,该用户提供的答案与识别真值越接近。化简式(3),得到式(5)。

$$d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i}) = \frac{1}{P}\sum_{p=1}^{P}\boldsymbol{O}^{q_i} - \boldsymbol{O}^{q_i}\boldsymbol{\Gamma}_k^{q_i} + \lg(1+\mathrm{e}^{-\boldsymbol{O}^{q_i}}) \quad (5)$$

式中,当 $\boldsymbol{O}^{q_i} < 0$,则 $\mathrm{e}^{-\boldsymbol{O}^{q_i}} \to \infty$。为避免溢出并确保计算稳定,对式(5) 进行修正,使用 $\max\{\boldsymbol{O}^{q_i},\ 0\}$ 代替 $\boldsymbol{O}^{q_i}$,则最终交叉熵损失用式(6) 计算。

$$d(\boldsymbol{O}^{q_i},\ \boldsymbol{\Gamma}_k^{q_i}) = \frac{1}{P}\sum_{p=1}^{P}\max(\boldsymbol{O}^{q_i},\ 0) - \boldsymbol{O}^{q_i}\boldsymbol{\Gamma}_k^{q_i} + \lg(1+\mathrm{e}^{-\boldsymbol{O}^{q_i}}) \quad (6)$$

在网络优化过程中使用随机梯度下降方法优化网络模型参数,并使用 ReLU 作为激活函数,将非线性特性引入网络中。

2.3　用户答案评分

通过深度神经网络多次训练优化迭代,直至网络参数收敛,网络输出最终的识别真值向量记为 $\boldsymbol{O}^{q*}$,依据识别真值向量与各数据源提供的答案向量的相似度定义各个答案的分数,对于 GloVe、SIF 向量,用式(7) 计算。

$$Score_k^{q_i} = 1 - \frac{1}{\pi}\arccos\left(\frac{\boldsymbol{O}^{q_i}\cdot\boldsymbol{\Gamma}_k^{qi}}{\sqrt{\boldsymbol{O}^{q_i}}\ \sqrt{\boldsymbol{\Gamma}_k^{qi}}}\right) \quad (7)$$

对于 BoW、TF-IDF 向量,用式(8) 计算。

$$Score_k^{q_i} = \left[\frac{1}{P}\sum_{p=1}^{P}\max(\boldsymbol{O}^{q*},\ 0) - \boldsymbol{O}^{q*}\boldsymbol{\Gamma}_k^{q_i} + \lg(1+\mathrm{e}^{-\boldsymbol{O}^{q*}})\right]^{-1} \quad (8)$$

分数越高,则答案越可靠,根据分数对问题提供的答案进行排名,找到众多回答中的可靠回答。

3　实验与分析

3.1　实验环境与数据集

在真实数据集上进行对比实验验证基于深度神经网络的多源文本数据真值发现算法的实验效果。实验框架为 tensorflow,CPU 为 Inter Xeon E5 - 2630,内存为 192 GB,GPU 为 Nvidia Tesla P40 ×2,操作系统为 CentOS 7 64 位。

采用源自 Kaggle 竞赛的 Short Answer Scoring (https://www.kaggle.com/datasets/harshdevgoyal/short-answer-scoring) 数据集,该数据集包含英语与艺术、生物、科学、英语 4 个学科数据集,每个学科数据集由问题和答案组成,所有答案经由竞赛学生撰写,答案长度为 40 ~ 60 个单词,并经过相关人员对答案进行手动打分(0 ~ 3 分)。

3.2　评价指标及实验超参数

实验使用 Top k 值作为评价指标,即取答案分数从高到低排名的前 k 个分数求平均分。表 2 为实验超参数设置。

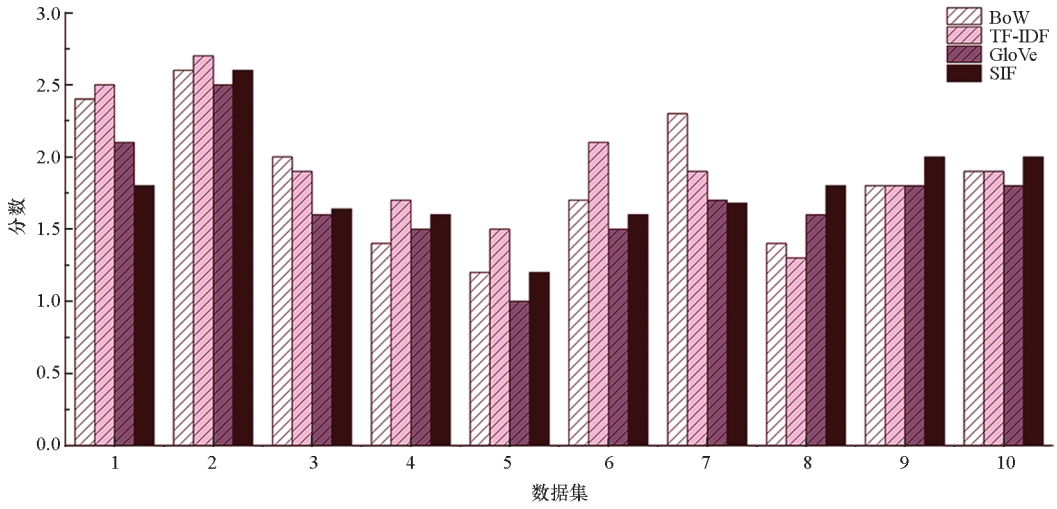
表 2　实验超参数设置

Tab. 2　Experimental super parameters setting

参数名称	参数值
batch_size	200
学习率	0.001
隐藏层层数	3
隐藏层节点数	10

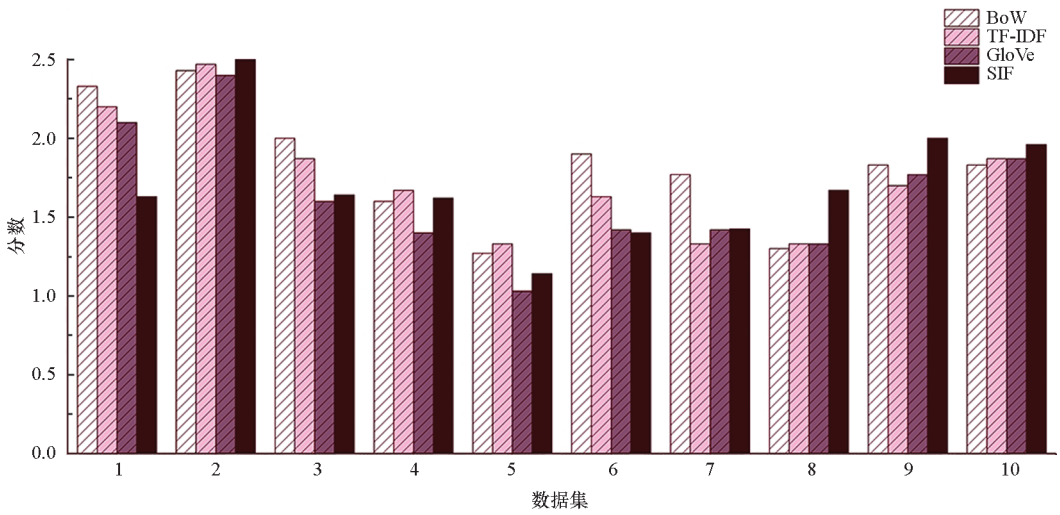
3.3　语义表征方法对比

为验证不同语义表征方法对真值发现结果的影响,分别使用 BoW,TF-IDF,GloVe,SIF 进行文本的语义表征,数据集 1 ~ 3 为学科科学,4 ~ 5 为英语与艺术,6 ~ 7 为生物,8 ~ 10 为英语,对比结果如图 3 所示。



（a）前10名平均分

（a）Average scores of Top 10



（b）前30名平均分

（b）Average scores of Top 30

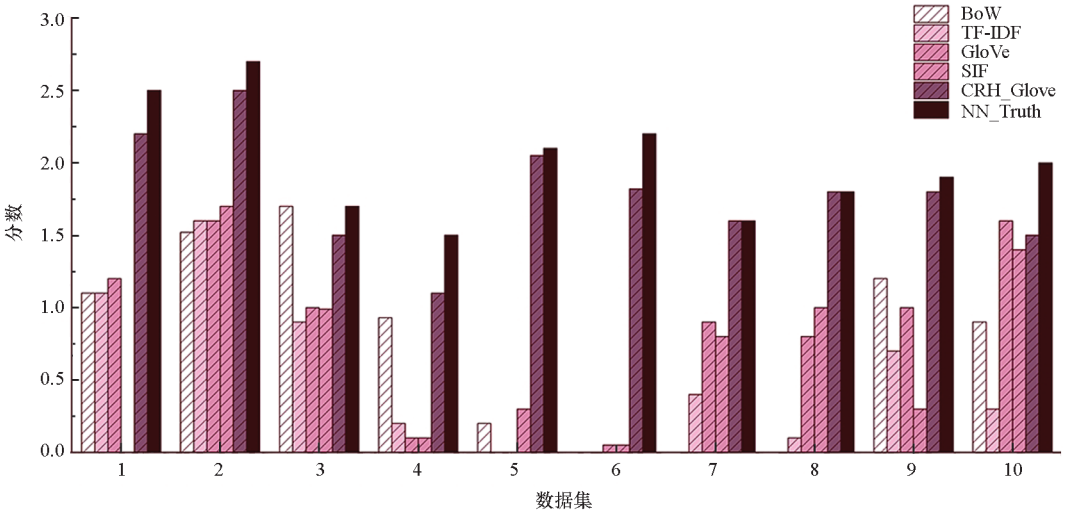
图3 不同文本表征方法结果比较

Fig. 3 Comparison of different text representation methods

由图3可知，针对不同的数据集，不同的语义表征方法结果有微小差别，BoW与TF-IDF表示方法粒度较粗，不对答案中关键词语义相似性进行度量，要求用户必须给出准确的关键词，适用于客观严谨的问题，在科学与生物两门学科的答案评估中，结果较优。而GloVe与SIF表示方法粒度较细，考虑文本的语义信息，答案中语义相近的词被赋予相同的可信度，用户给出类似或含义相近的词，获得近似的可靠度，适用于主观开放性问题，在英语学科具备优势。TF-IDF方法在不同数据集中结果相对稳定，在之后的对比实验中，以TF-IDF方法作为答案语义表征方法，展示所提文本数据真值发现方法的优越性与稳定性。

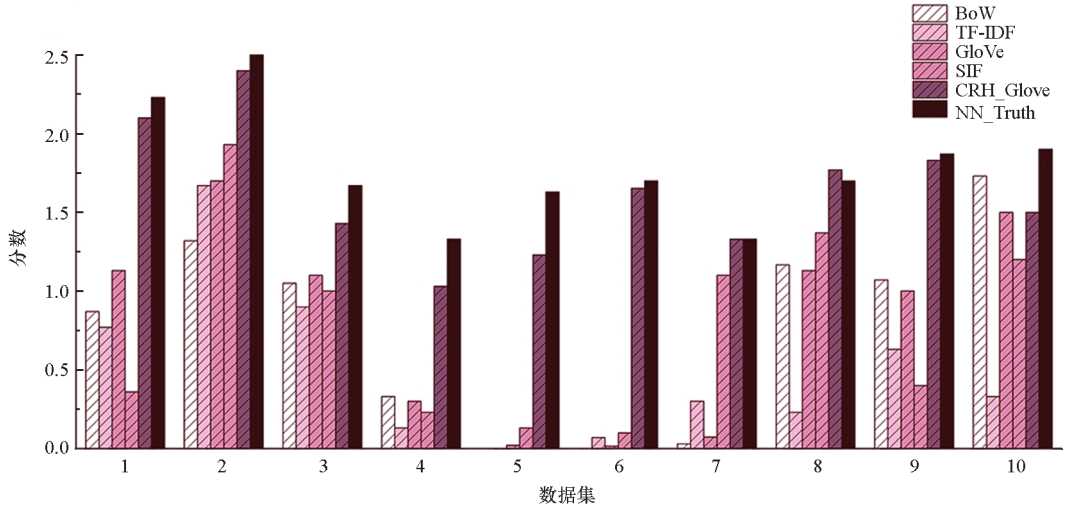
3.4 对比算法

将所提方法分别与基于检索的方法BoW Similarity，TF-IDF Similarity，GloVe Similarity，SIF Similarity及表现优异的异质数据冲突消解（conflict resolution on heterogeneous data，CRH）算法[13]进行比较。由于传统真值方法均不适用于文本数据的真值发现，本文对CRH真值发现算法进行改进，同样对文本数据进行了语义表征，并使用本文提出的距离函数度量文本答案间的相似性，当使用BoW与TF-IDF时，由于文本向量稀疏，真值发现的结果为全0向量，此时CRH方法失效，本文使用GloVe向量进行语义表征的真值发现结果作为CRH方法的结果，对比结果如图4所示。



(a) 前 10 名平均分

(a) Average scores of Top 10



(b) 前 30 名平均分

(b) Average scores of Top 30

图 4　不同方法结果对比

Fig. 4　Results comparison of different methods

由图 4 可知,NN_Truth 算法优于对比算法。首先,基于检索的方法对答案进行排序的依据是问题与答案的相似度,但答案所需的关键词并不一定包含在实际问题中,所以基于检索的方法所找到的答案,其关键词与问题中的关键词有很大的相似性,并不是真正意义上的真值。CRH 算法对数据源与观测值之间的关系做出假设,即可用简单线性函数来表示它们之间的关系,但简单函数难以对这种复杂关系进行准确描述,因此这种强假设使得 CRH 的实验效果不佳。文本数据真值发现场景中用户数量大,但每个用户只提供的少量的答案,将会加快 CRH 迭代训练时的收敛速度,进而对评估数据源产生影响。NN_Truth 算法通过深度神经网络寻找指定问题的真值,不需要强假设答案与各真值的关系,这种复杂关系将存储在神经网络的矩阵参数中,同时将“数据源 - 答案”向量输入深度神经网络,解决了传统真值发现算法在文本数据真值发现场景中失效的问题,并且在面对稀疏数据时有很大的优势。

3.5　参数设置实验

为验证深度神经网络应用于文本数据真值发现时的有效性及稳定性,对 NN_Truth 算法进行参数设置实验。通过设置 5 种不同的隐藏层层数,验证深度神经网络层数对算法效果的影响,选取网络最终输出结果排名前 10、30、50、80、100、200、300 名学生分数计算平均分来评价实验效果,实验结果如表 3 所示。

表3 参数设置实验 Top k 表

Tab. 3 Experimental Top k parameters setting

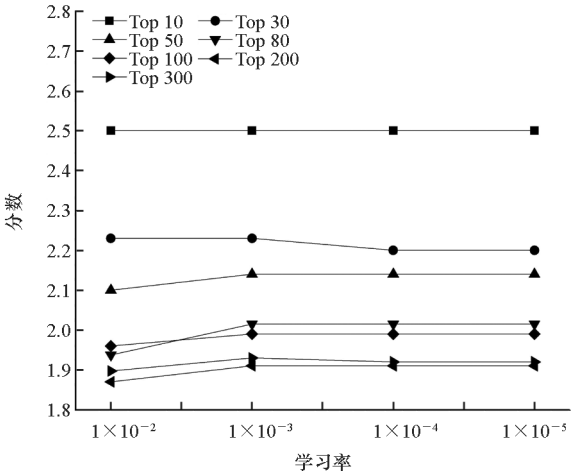
数据集	隐藏层层数	Top 10	Top 30	Top 50	Top 80	Top 100	Top 200	Top 300
科学1	1	2.413	2.031	2.057	1.966	1.992	1.903	1.874
	2	2.493	2.216	2.129	2.009	1.982	1.910	1.869
	3	2.511	2.236	2.149	2.013	2.006	1.915	1.902
	4	2.396	2.224	2.078	1.975	1.971	1.896	1.891
	5	2.455	2.083	2.116	1.989	1.920	1.890	1.860
英语与艺术1	1	1.452	1.317	1.288	1.295	1.290	1.062	1.038
	2	1.514	1.330	1.321	1.310	1.305	1.046	1.120
	3	1.508	1.335	1.340	1.357	1.311	1.115	1.124
	4	1.481	1.309	1.276	1.344	1.308	1.083	0.982
	5	1.436	1.298	1.251	1.282	1.279	1.109	0.962
生物1	1	1.593	1.103	1.138	1.068	1.136	0.966	0.950
	2	1.586	1.132	1.127	1.129	1.121	1.024	0.983
	3	1.602	1.279	1.322	1.204	1.146	1.048	1.036
	4	1.582	1.264	1.256	1.057	1.080	1.068	1.007
	5	1.560	1.252	1.140	1.133	1.061	1.049	0.970
英语1	1	1.906	1.791	1.808	1.652	1.672	1.564	1.530
	2	1.939	1.860	1.815	1.716	1.751	1.647	1.611
	3	2.003	1.898	1.841	1.782	1.805	1.656	1.709
	4	1.865	1.876	1.783	1.759	1.742	1.607	1.623
	5	1.890	1.843	1.746	1.762	1.708	1.575	1.556

由表3可知,深度神经网络隐藏层数量会影响实验结果,当隐藏层层数为3时,实验结果最好。

3.6 学习率对实验结果的影响

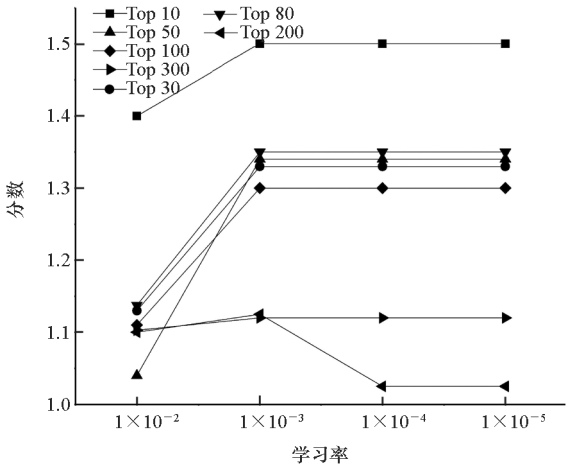
学习率控制着神经网络参数的更新速度,学习率过大或者过小,都会影响神经网络收敛速度和最优解的获取。通过设置5种不同的学习率,验证学习率对网络训练效果及收敛速度的影响,实验结果如图5所示。

由图5可知,学习率对实验结果的影响较小,部分数据集在学习率为 1×10^{-2} 时,效果有所下降,对于大部分数据集,学习率为 1×10^{-3} 时效果最好。



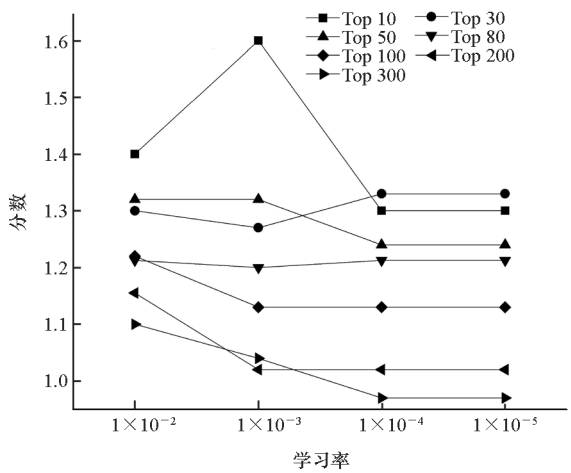
(a) 科学1

(a) Science 1



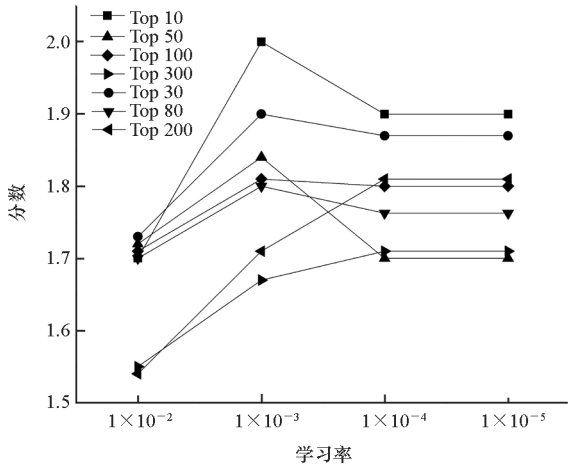
(b) 英语与艺术1

(b) English and art 1



(c) 生物1

(c) Biology 1



(d) 英语 1

(d) English 1

图 5　学习率对实验结果的影响

Fig. 5　Effect of learning rate on experiment result

4　结论

本文提出基于深度神经网络面向多源文本数据的真值发现算法 NN_Truth,对文本答案进行向量化表示,并利用神经网络寻找答案真值,区别于传统真值发现算法对数据源可靠度的依赖,NN_Truth 更加注重对答案本身语义信息的挖掘,使用深度神经网络对答案间的复杂关系进行无监督学习,为问题寻找正确可靠的答案。通过实验验证,本算法在数据源众多而观测值较少的场景中效果较好,优于 CRH 等传统真值发现算法。

参考文献(References)

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

[2] GALLAND A, ABITEBOUL S, MARIAN A, et al. Corroborating information from disagreeing views [C]// Proceedings of the 3rd ACM international conference on Web search and data mining, 2010.

[3] LI Y L, 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.

[4] LI Q, LI Y L, 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.

[5] YANG Y, BAI Q, LIU Q. A probabilistic model for truth discovery with object correlations [J]. Knowledge-Based Systems, 2019, 165: 360 – 373.

[6] BROELEMANN K, GOTTRON T, KASNECI G. LTD-RBM: robust and fast latent truth discovery using restricted boltzmann machines [C]//Proceedings of IEEE the 33rd International Conference on Data Engineering, 2017.

[7] SUN P, WANG Z B, FENG Y H, et al. Towards personalized privacy-preserving incentive for truth discovery in crowdsourced binary-choice question answering [C]// Proceedings of IEEE Conference on Computer Communications, 2020.

[8] LI Z Q, YANG S, WU F, et al. Holmes: tackling data sparsity for truth discovery in location-aware mobile crowdsensing [C]//Proceedings of IEEE the 15th International Conference on Mobile Ad Hoc and Sensor Systems, 2018.

[9] MARSHALL J, ARGUETA A, WANG D. A neural network approach for truth discovery in social sensing [C]// Proceedings of IEEE the 14th International Conference on Mobile Ad Hoc and Sensor Systems, 2017.

[10] LI L Y, QIN B, REN W J, et al. Truth discovery with memory network [J]. Tsinghua Science and Technology, 2017, 22(6): 609 – 618.

[11] ARORA S, LIANG Y Y, MA T Y. A simple but tough-to-beat baseline for sentence embeddings[C]//Proceedings of International Conference on Learning Representations, 2017.

[12] ZHANG H T, LI Y L, MA F L, et al. Text truth: an unsupervised approach to discover trustworthy information from multi-sourced text data[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.

[13] LI Y L, LI Q, GAO J, et al. Conflicts to harmony: a framework for resolving conflicts in heterogeneous data by truth discovery[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(8): 1986 – 1999.