

异构数据联合式的真值发现算法

陈超^{1,2}, 申德荣¹, 寇月¹, 于戈¹

(1. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110169; 2. 渤海大学 信息科学与技术学院, 辽宁 锦州 121007)

摘 要: 互联网上提供的同一事实的信息通常会存在冲突,影响数据集成和知识发现. 为了甄别真值,提出了一种基于距离的异构数据联合真值发现算法. 首先,关于同一数据项,基于数据源声明值与真值的距离,计算数据项向量;采用 KMeans 聚类算法,获得数据项初始聚类. 然后,迭代进行信任分析和聚类,即在每个类簇内,采用最优化思想,联合异构类型数据,更新事实的可信度和数据源的类簇内可靠性,重新计算每个数据项向量,再次聚类,迭代直至类簇达到稳定. 实验结果表明:由于细粒度的数据源质量划分,联合考虑异构数据类型,可以获得更高的真值发现准确度.

关 键 词: 真值;真值发现;KMeans 聚类;最优化;异构数据

中图分类号: TP 301 **文献标志码:** A **文章编号:** 1005-3026(2017)10-1373-05

Joint Truth Finding on Heterogeneous Data

CHEN Chao^{1,2}, SHEN De-rong¹, KOU Yue¹, YU Ge¹

(1. School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China; 2. College of Information Science & Technology, Bohai University, Jinzhou 121007, China. Corresponding author: SHEN De-rong, professor, E-mail: shenderong@ise.neu.edu.cn)

Abstract: The value of an entity attribute on the web is usually provided by multiple data sources, but the values provided by them are not always the same, which affects the effective integration of data, so it is necessary to find out the true value among these given values. The existing truth finder algorithms mainly focus on the single type data kind, so a distance-based truth finding algorithm was proposed by considering heterogeneous data jointly. Firstly, for a specific data item, the data item vectors were calculated on the basis of the distance between the claimed value from every source and the truth value. The KMeans algorithm was used to get initial clustering. Then, alternate clustering and trust analysis were iteratively performed, i. e., within each cluster, confidence of facts and trustworthiness of sources were updated with the idea of optimization and joint heterogeneous data. Each data item vector was recalculated and reclustered, and when each cluster was stable, the iteration would be terminated. The experiment results showed that the proposed algorithm has a higher accuracy for truth finding because of the fine grained partition of source quality and the joint model of heterogeneous data.

Key words: truth; truth finding; KMeans clustering; optimization; heterogeneous data

信息时代,为了更有效地为用户返回相关知识,互联网上相继出现了一些大规模的语义知识库,如 Freebase, YAGO 和 Linked Data 等,它们已成为了互联网知识获取的主要途径. 为了构建和维护这些语义知识库,需要从 Web 中抽取数量庞大的实体和各种类型的海量数据作为语义数据库

的支撑数据,并且要求这些数据具有很高的准确度. 然而,由于 Web 数据源本身存在数据时效性和准确度问题,再加上数据抽取器带来的误差^[1-2],使得同一个实体的某些属性往往存在冲突值的情况^[3-4],影响知识库的构建和维护. 本文提出一种有效的真值发现算法,从多个抽取值中

甄别关于该实体属性的真值,即真值发现.

目前提出的真值发现算法主要用于对实体的某一个属性进行真值发现,即关于单属性的同构数据真值发现算法^[5-9].但实际上,许多应用,尤其在知识库领域,需要发现同一个实体多个属性多种类型数据的真值,这就需要研究关于多属性的异构数据真值发现问题.文献[10]提出了一个关于多属性的异构数据的通用真值发现框架,利用适当的损失函数可以获取数据类型特有的属性特征.然而,在该框架下,真值计算简单地遵循着可靠性最高的数据源提供的所有观察值即是真值的原则,这是不合适的.因为,任何一个可靠的数据源都有可能在某些数据项上提供错误的观察值.

本文提出一种基于聚类的多属性异构数据联合式的真值发现算法.该算法不仅考虑了各种数据类型独有的特征,还利用了多个属性的联合特征,能够更加全面地推理出数据源的可靠程度,从而有效提高了真值发现算法的准确度.另外,突破传统方法为每个数据源只赋予一个整体质量水平.考虑了不同数据对象类簇下,数据源具有不同可靠程度的情况.实验结果表明,本文提出的算法可有效提高多属性异构数据真值发现的准确度.

1 问题描述

本文关注的是异构数据类型上的真值发现问题,基于目标优化的思想,采用迭代计算方法,发现事实的真值,并采用聚类思想,细粒度评估数据源在不同数据项上的可靠性.

本文涉及的真值发现问题的基本概念如下.

定义 1 数据源:将每个 Web 源上提供的数据经过抽取器抽取处理后的结构化表示称为一个数据源. S 表示数据源的集合,即 $S = \{s_1, s_2, \dots, s_k\}$.表 1 中的数据是抽取自不同天气预报网站的信息,每个天气预报网站用 s_k 表示.

表 1 天气预报数据 Table 1 Weather forecast data				
地名	变量	s_1	s_2	s_3
圣迭戈	湿度/%	94	89	87
圣迭戈	天气	雾	雾	霾
圣何塞	湿度/%	91	96	84
圣何塞	天气	多云	大部多云	晴
圣何塞	气压/kPa	104.0	103.6	101.9

定义 2 数据项:实体是现实世界中客观存

在的具体对象,数据源通过属性加以描述,实体的一个属性表示一个数据项,每个数据项用 o_i 表示,如用 o_1 表示数据项{圣迭戈,湿度},表 1 中的数据项可以用 o_1 到 o_5 依次标号.

定义 3 声明值:每个数据源提供某些实体属性的值,称为声明值, $v^k(o_i)$ 表示数据源 k 关于数据项 o_i 的声明值.

不同数据源提供的同一个数据项的声明值之间会存在冲突,如关于数据项{圣迭戈,天气},数据源 s_1 提供的值为“雾”, s_2 提供的值为“雾”,而 s_3 提供的值为“霾”.表 1 中 3 个数据源提供了关于 5 个数据项的 15 条事实.本文模型设定每个数据项只有一个真值(或值集),即单真值发现问题.

在本文模型中,输入数据为 K 个数据源上提供的关于 N 个数据项的事实集合.

模型输出为实体属性的真实值,即每个数据项的真值,以及描述数据源可靠程度的数据源整体信任度和类簇内的信任度.

定义 4 真值:数据源所描述的实体的相关属性在真实世界中的值,称为真值, $v^*(o_i)$ 表示数据项 o_i 的真值.

定义 5 数据源可靠性:数据源可靠性是数据源质量的评估.数据源全局可靠性 $w(k)$ 描述数据源 k 的整体可信赖程度,数据源类内可靠性 $w_c(k)$ 描述数据源在类簇 c 内的可信赖程度.

2 异构数据联合式的真值发现算法

本文算法的基本思想:同一个数据源的不同数据项具有不同的可靠性,而一个类簇内的数据源的可靠性是一致的.在进行数据源的可靠性分析时,联合不同数据类型的属性值进行推理,可以获得更高的真值准确度.

2.1 数据项聚类

传统的真值发现算法为每个数据源可靠性赋予一个数值,表示数据源的整体质量水平,而实际应用中,同一数据源在不同数据项的事实上可能具有不同的可靠程度.数据源整体可靠性和类簇可靠性差异如图 1 所示.可知 3 个数据源对于所有 5 个数据项,数据源的可靠程度是 s_1 最大, s_2 最小.但就数据项 o_2 和 o_3 而言,数据源 s_1 和 s_2 提供了具有较低可信度的值, s_3 提供了较高可信度的声明值;相反,对于数据项 o_1, o_4 和 o_5 ,数据源 s_1 和 s_2 提供了较高可信度的值,数据源 s_3 提供了较低可信度的值.根据数据源可靠性分析,对数据

项进行聚类,将 o_2 和 o_3 聚为一类,用 c_1 表示;将 o_1, o_4 和 o_5 聚为一类,用 c_2 表示,进而根据各个类簇内数据项的可信度计算数据源的类簇可靠性。

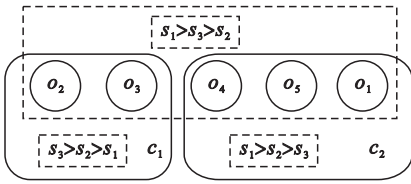


图 1 数据源整体可靠性和类簇可靠性差异

Fig. 1 Difference between global trustworthiness and cluster trustworthiness of sources

定义 6 数据项向量 T_o : 每个数据项以向量表示,每个分量是提供该数据项的各个数据源的声明值与真值的归一化距离。

采用 KMeans 聚类算法,将各个数据源为其提供了相似可信度的数据项聚为一类,计算数据源的类内可信度,从而获得更加准确的数据项真值。

2.2 信任度分析

2.2.1 异构数据的信任度分析模型

真值发现算法遵循的基本原理是可信度高的数据源提供高可信度的声明值,采用文献[10]中的模型思想,最优化公式为

$$\begin{aligned} \min_{V^{(*)}, W} f(V^{(*)}, W) &= \sum_{k=1}^K w_k \sum_{i=1}^{|O_{s_k}|} d_i(v^*(o_i) - v^k(o_i)), \\ \text{s.t. } \delta(W) &= 1, W \in D. \end{aligned} \quad (1)$$

式中: f 是损失函数,表示所有数据源的每个数据项的观察值与真值之间的加权距离之和,目标是使之最小; K 是数据源的总数量; $v^*(o_i)$ 是数据项 o_i 的真值; $v^k(o_i)$ 是第 k 个数据源提供的关于 o_i 的声明值; d_i 表示数据项 o_i 的声明值与其真值的距离; w_k 为数据源的权值,代表数据源的整体可靠性; O_{s_k} 是第 k 个数据源提供的所有数据项的集合; $V^{(*)}$ 是数据项的真值列表; W 是数据源权值列表; $V^{(*)}$ 和 W 是两个未知变量集。为保证最小化问题有最优解,令归一化函数 $\delta(W)$ 的值为 1, W 满足指定的值域 D 。

2.2.2 数据源可靠性分析

在分析数据源的可靠性时,假设数据项真值已知,根据真值和观察值的距离为每个数据源赋权值,使总损失函数最小,且满足约束条件。采用拉格朗日乘子法可得

$$w_k = -\lg \left(\frac{\sum_{i=1}^{|O_{s_k}|} d_i(v^*(o_i) - v^k(o_i))}{\sum_{k=1}^K \sum_{i=1}^{|O_{s_k}|} d_i(v^*(o_i) - v^k(o_i))} \right). \quad (2)$$

根据数据源 s_k 声明的所有数据项计算得到的 w_k , 表示数据源 s_k 的整体可靠性; 根据聚类簇内的数据项计算所得的 $w_k(c)$, 表示数据源 s_k 在类簇 c 内的可靠性。

2.2.3 事实可信度分析

在分析事实可信度时,假设数据源权值已知,最小化损失函数变成最小化每个数据项与真值的加权距离,而距离的计算又依赖声明值的数据类型。本文把数据分为连续型数据和分类型数据。

1) 连续型数据。对于连续型数据,损失函数的计算式为

$$d_i(v^*(o_i), v^k(o_i)) = \frac{|v^*(o_i) - v^k(o_i)|}{\text{std}(v^1(o_i), \dots, v^K(o_i))}. \quad (3)$$

式中, $\text{std}(v^1(o_i), \dots, v^K(o_i))$ 表示所有数据源提供的关于数据项 o_i 的声明值的标准偏差,用于归一化声明值与真值的距离,降低异常点的影响。

当数据源权值已知,将式(3)代入式(1),可得

$$v^*(o_i) = \frac{\sum_{k=1}^K w_k \cdot v^k(o_i)}{\sum_{k=1}^K w_k}. \quad (4)$$

2) 分类型数据: 对于分类型数据,损失函数的计算式为

$$d_i(v^*(o_i), v^k(o_i)) = \begin{cases} 1, & \text{if } v^k(o_i) \neq v^*(o_i); \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

将式(5)代入式(1),可得计算分类型数据项 o_i 的真值:

$$v^*(o_i) \leftarrow \arg\max_v \sum_{k=1}^K w_k \cdot 1(v, v^k(o_i)). \quad (6)$$

式中, $1(x, y)$ 函数表示如果 $x = y$, 则 $1(x, y)$ 的值为 1, 否则为 0。

2.3 算法实现步骤

首先,采用 vote/median 方法给每个数据项赋初值,计算数据项向量,采用 KMeans 聚类算法获得初始类簇。然后,遍历每一个类簇,迭代计算可信度和更新类簇,直至每个类簇都达到稳定。

1) 在每个类簇内,采用式(2)计算每个数据源类簇内可靠度 $w_c(k)$ 。

2) 在每个类簇内及每一个数据项,根据数据

类型采用式(3)或式(5),计算每个数据项向量 t_{oi} .

3) 在每个类簇内,采用式(4)或式(6),分别计算连续型和分类型数据项的真值 $v^*(o_i)$.

4) 采用 KMeans 聚类算法,输入 t_{oi} ,更新数据项 o_i 的类簇.

3 实验结果与分析

本文实验采用的数据集和标准集来自文献[7]. Weather 数据集是从 18 个 Web 数据源抽取的关于美国 30 个主要城市的天气状况的描述.原始数据集包含 28 个不同的属性,从中选择气压、湿度和天气情况三个属性进行实验,前两个视为连续型数据,后一个视为分类型数据.

为了评估本文算法,选择准确率(A)作为评价指标,对于分类型数据,采用文献[7]的计算方法.对于连续型数据,采用文献[6]的计算方法,同时比较精准率(P)、召回率(R)和 F-Measure(F).

第一组实验验证了本文的基本假设:同一个数据源在关于它所提供的所有事实上,可靠性并不一致,即存在整体可靠性和局部可靠性(类簇内可靠性)之分.从图 2 可以看出,数据源 s_1 的整体可靠性较高,而 s_2 的整体可靠性较低,它们都围绕着某一个特定的值波动,在每个类簇内的可靠性存在差异.

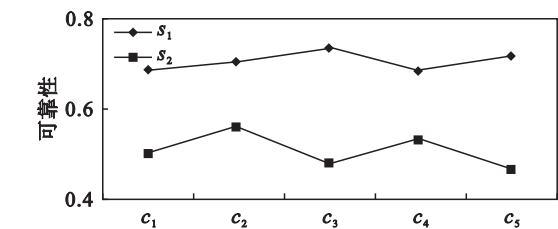


图 2 数据源 s_1 和 s_2 在各个类簇内的可靠性
Fig. 2 Trustworthiness of s_1 and s_2 in clusters

第二组实验验证了联合异构数据类型的数据源可靠性评估,可以获得更加准确的事实可信度.

基本的 Vote 方法选择分类型数据中最频繁出现的值作为真值,Median 方法将每个连续型数据项的所有声明值的中位数作为真值;CRH 方法联合考虑分类型数据和连续型数据,基于距离分两步更新数据源的权值和数据项的真值;本文方法(CBH)细粒度划分数据源的可信度,计算基于聚类的数据源权值.从图 3a 可以看出,对于分类型数据“天气条件”(Condition),不区分数据类型

的 Vote 方法的准确度明显低于考虑异构数据特征的 CRH 和 CBH.在图 3b 中,对于连续型数据“气压”(Pressure),不联合考虑类别数据的 Median 方法的准确度也明显低于 CRH 和 CBH 算法.同时可以看出,本文进行细粒度区分数据源可靠性的 CBH 算法虽然召回率略低于 CRH,但准确度获得了明显提高.

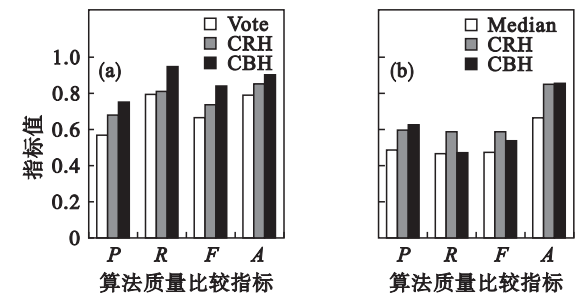


图 3 在不同类型数据上的算法性能度量
Fig. 3 Algorithm metrics on heterogeneous data
(a)一分类型数据;(b)一连续型数据.

4 结 论

本文针对 Web 数据集的真值发现问题,提出了一种异构数据联合式的数据真值发现算法,弥补了数据源在所有数据项上都采用同一可靠性这一假设的缺陷.采用最优化思想,联合考虑分类型数据和连续型数据的特点,迭代更新事实可信度和数据源类簇内可靠性,从而进一步提升了真值发现的准确性.

今后的工作,希望在不牺牲真值发现算法的准确率的前提下,改进算法的效率,尤其在动态数据流场景下,考虑时间关系,高效识别信息的可信度和数据源的可靠性.

参考文献:

[1] Dong X L, Gabrilovich E, Murphy K, et al. Knowledge-based trust: estimating the trustworthiness of web sources [J]. *Proceedings of the VLDB Endowment*, 2015, 8 (9): 938-949.

[2] Dong X L, Gabrilovich E, Heitz G, et al. From data fusion to knowledge fusion[J]. *Proceedings of the VLDB Endowment*, 2015, 8(9): 881-892.

[3] Yu D, Shen D, Zhu M, et al. A method to discover truth with two source quality metrics [C]//Web Information System and Application Conference. Jinan, 2015: 161-164.

[4] Li Y, Gao J, Meng C, et al. A survey on truth discovery[J]. *ACM SIGKDD Explorations Newsletter*, 2015, 17(2): 1-16.