ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 442－452
DOI:10. 16337/j. 1004-9037. 2019. 03. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

面向数据集成的多真值发现算法

陈烈锋　许青林

(广东工业大学计算机学院,广州 ,510006)

摘　要: 大数据时代,大规模数据往往由多个数据源组成并服务于多个数据驱动型应用程序。由于数据源的可信度不同,不同数据源往往会产生数据冲突,使得难以判断哪些信息是真实的。近年来,真值发现方法通过从多个数据源中找到最符合现实的真值来解决冲突而成为研究热门。当前真值发现算通常假设实体某个属性只有一个真值,然而在现实中,实体具有多个真值的情况更为常见。针对多值实体提出了一个多真值发现算法,该算法将多真值发现转化为一个函数优化问题。根据对目标函数的求解选取置信度最高的多个值作为实体的真值。同时在计算描述值的置信度时,提出一种非对称的支持度计算方法,结合相似值的支持对其置信度进行修正。通过多个真实数据集上的实验表明本文算法的准确性优于现有的真值发现算法。

关键词: 数据集成;数据冲突;真值发现;多真值;数据源可信度

中图分类号: TP181　　**文献标志码:** A

Multi-Truth Finding Algorithms for Data Integration

Chen Liefeng, Xu Qinglin

(Department of Computer Science and Technology, Guangdong University of Technology, Guangzhou, 510006, China)

Abstract: In the era of big data, large-scale data are often contributed by numerous data sources and used by many data-driven applications. Because of different trustworthiness of sources, different sources often produce data conflicts, making it difficult to determine which information is true. In recent years, truth finding has become a research hotspot by finding the most credibility values from multiple sources. The current truth finding methods usually assume that the entity has only one truth, while in reality, entities may have multiple true values. In this paper, we present an approach for multi-truth finding, which transforms the multi-truth finding into an optimization problem. In so doing, we select the values with the highest credibility as truths of entities. We also propose an asymmetric approach to compute support between values and incorporate influences of similar values to measure value credibility for better truth finding. Experiments on several data sets show that the effectiveness of our algorithm outperform the existing state-of-the-art techniques.

Key words: data integration; data conflicting; truth finding; multi-truth finding; source trustworthiness

---

基金项目: 广东省重大科技厅重大专项(2016B030306003)资助项目。
收稿日期: 2018-07-11;**修订日期:** 2018-09-20

引 言

随着网络技术的飞速发展以及智能设备的广泛使用,数据以前所未有的速度生成和创建。然而,在大数据改变现代社会许多层面的同时,我们也经常可以观察到不同的数据源对同一实体提供了相互冲突的描述。这些冲突往往是由于输入错误、数据过时、记录丢失等原因造成的[1-2],如果应用于实际可能会造成巨大的损害和经济损失。例如,数据在医疗系统中被用于药物推荐或者在股票市场上数据被用于股票价格预测[3]。给定一个大规模数据,手工确定数据的真实性是不现实的,而真值发现方法能从多个数据源中找到最符合现实的真值来解决冲突,因此成为研究热门。

到目前为止,已有大量工作来处理真值发现问题。现有的算法[1-2,4-12]大多能通过迭代的方法来联合推导数据源的可信度和描述值的置信度,它们综合考虑各种方面的影响如数据源的依赖关系,先验知识和数据源的质量等来提高真理发现的准确率。当前算法通常假设每个实体只有一个真值,然而在现实世界中,实体拥有多个真值的情况可能更为常见。例如,一本书通常有多个作者,一部电影可能有几位导演。尽管先前的算法通过将某个数据源上提供的一组值简单地看作一个值集,并选择置信度最高的值集作为真值集,以此来处理多真值发现问题,但是不同的数据源提供的值集通常是有关联的:同一个实体上两个数据源提供的值集之间可能存在重叠,并不是完全冲突的。例如,数据源"Powell's Books"给图书"Rapid Contextual Design"提供值集 V_1={Karen Holtzblatt},而数据源"Barnes & Noble"提供值集 V_2={Karen Holtzblatt,Jessamyn Wendell,Shell Wood}。采用投票的思想,V_1和 V_2各得1票,但如果将值集里的值分开进行投票,值"Karen Holtzblatt"得2票,其他值得1票,因此该值成为真值的可能性更大,如果忽略这层含义将会降低真值发现的准确性。

在多真值发现问题中,由于可能涉及大量的数据源和实体,想获得完整的真值集是很困难的。例如,文献[1]通过手工检查每本书的封面制作用来实验的图书作者数据集,花费了大量的时间和精力。因此需要一种无监督的方法来解决真相发现问题。多真值发现问题的另一个挑战是数据源的质量(即可信度)是未知的;因对数据源的了解很少,且数据源的质量通常是不同的。如果不评估和区分它们,真值发现算法很容易被低质量的数据源误导。为了解决这些挑战,文献[13-17]已经提出了一些方法来处理多值实体,然而它们并没有考虑描述值不同表现形式的影响,忽略了相似值对实体真值的支持。例如,"Jessamyn Wendell"和"Jessamyn Burns Wendell"可能是同一作者姓名的不同表现形式,因此它们是相互支持的,忽略这些值的影响可能会降低真值发现的准确率。

本文主要贡献如下:

(1) 基于启发式思想,本文将多真值发现转化为一个函数优化问题,目标函数是每个实体的真值集与数据源对该实体提供的所有值集之间的相似度加权和达到最大,权重为数据源可信度。

(2) 在计算实体真值的过程中,根据对目标函数的求解并采用贪心策略来选取实体的真值。同时,我们定义一种非对称的支持度计算方法来度量相似值之间的影响,并结合相似值的支持到描述值置信度的计算当中,提高了真值发现的准确率。

(3) 通过多个数据集上的实验表明,本文算法的准确率优于现有的真值发现算法。

1 文献综述

目前真值发现问题已有了广泛的研究。最简单的方法是采用基于投票的方法,当实体的某个描述值所获得的票数达到某个阈值时,该值则被认为是真值。然而,该方法没有考虑数据源的可信度对描述值置信度的影响。文献[2]提出了一种可以迭代计算数据源可信度和描述值置信度的算法 TruthFinder。该算法基于启发式思想:可信度越高的数据源提供的描述值置信度越高,同时提供越多高置信度描述值的数据源的可信度也越高,因此可以利用两者的关系进行迭代计算。之后在这一思想基础上,研

究人员通过考虑不同场景或不同影响因素对基本算法进行扩展。文献[3-7]考虑数据源之间的依赖关系,如复制关系[3-6]和分组关系[7]等,显著提高了真值发现的准确率。文献[6]考虑信息的时效性,作者采用隐马尔科夫模型来判断数据源之间的复制关系和复制时间,建立一个贝叶斯模型从数据源中聚合信息从而确定信息的真实性。文献[9]通过估计实体每个描述值的获取难度从而避免数据源从获取难度较低的描述值中获得较高的可信度。文献[10]通过将先验知识引入到真值发现中而得到更高的精度。文献[11-12]可以进行真值发现的在线计算。文献[8-9]对数据源可信度采取不同的度量方法来提高真值发现的准确率。

尽管真值发现问题已经进行了大量研究,然而大多数研究集中在单真值发现问题,多真值发现问题的研究相对较少[18]。文献[13]通过构建一个概率图模型LTM来联合推导实体描述值的置信度和数据源可信度,是第一个处理多真值发现的模型。然而,该模型假设数据源的准确率和召回率服从某一特定分布,如果真实数据集不满足假设的分布,该算法的效率则受到很大影响。文献[14]通过分析多真值发现问题的特性,结合数据源对描述值置信度的影响和一种更优的拷贝检测技术到贝叶斯模型中,提高了真值发现的效率。文献[15]也提出了一种考虑多值实体的概率模型,然而该模型需要初始化多个参数,如每个实体真值的个数和假值的个数等,对真值发现的准确率有一定影响。文献[16]设计3种模型(即副产品模型、联合模型和合成模型)用于增强现有的真实发现算法。最近,Fang等[17]提出了一种基于图的模型,通过对两类数据源关系的建模来估计数据源的可信度和检测数据源之间的恶意复制,并考虑实体流行度对真值发现的影响,提高了真值发现的准确率。然而,上述方法没有考虑实体描述值的不同表现形式,忽略了相似值对实体真值的影响。

与上述多真值发现方法相比,本文算法有两个创新点:(1)本文算法将多真值发现转化为一个函数优化问题,通过对目标函数的求解直接返回实体的真值列表;(2)本文算法考虑描述值不同表现形式的影响,提出一种非对称的相似值支持度计算方法,结合相似值的支持到描述值可信度的计算当中。

2 问题描述

数据源通常会提供实体多个属性的描述值信息,然而对于每个属性来说,数据源的可信度可能不同,因此每个属性类型需要进行单独处理。本文假设实体只有一个属性来简化讨论。

本章首先给出一些相关定义,然后在此基础上对本文所提问题进行形式化定义。

2.1 相关定义

定义1 数据源为真值发现问题提供相互冲突的数据,可以来自网站、数据库等等。

定义2 一个实体表示一个能在真实世界中被识别的、唯一的对象。例如:一本书或一部电影。

定义3 一个数据源可以为一个实体提供多个描述值,这些值可以组成一个值集。

定义4 实体的可能值集表示所有数据源对该实体提供的值集的并集。

定义5 在实体的可能值集中,所有与真实世界一致的值构成了一个真理集。

定义6 实体的属性上只有一个真值的称为单真值发现问题,不止一个真值的称为多真发现问题。本文研究的是多真值发现问题。例如,电影可能有多个导演,一本书可能有多个作者。

定义7 不同的数据源为同一个实体提供不同的值集,从而产生数据冲突。

2.2 问题定义

假设有数据源集合$S=\{s_1,s_2,s_3,\cdots,s_m\}$,这里$m$表示数据源的数量。联合提供实体集合$E=\{e_1,e_2,e_3,\cdots,$

$e_n\}$,n表示实体的数量。在多真值发现问题中,数据源s可以给实体e提供一个值集,用V表示。实体e的可能值集则是S对e提供的所有值集的并集,用V^*($V^*=\bigcup_{V\in V(e)}V$)表示,同时用L表示V^*的长度。实体e可以有多个真值,用Truth表示该实体的真值集,Truth是V^*的子集。由此,本文问题可以定义为:给定一个冲突数据源集合S和一个实体集合E,本文的任务是为每个实体在该实体的可能值集V^*中找到真值集Truth。

3 多真值发现算法

本节译述了方法细节,包括值集之间相似度的定义,多真值发现的框架以及相应的算法。本节中使用的所有变量如表1所示。

表1 变量描述

Tab.1 Variable description

Name	Description
M	Number of data sources
N	Number of entities
E	Set of entities
S	Set of data sources
V	Value set provided by source s for entity e
V^*	Value set provided by all sources for entity e
L	Length of V^*
Truth	Truth set for entity e
$t(s)$	Credibility of source s
$V(s)$	Set of value sets provided by source s
$S(V)$	Set of sources providing value set V
$S(v)$	Set of sources providing value v
$c(v)$	Credibility of value v of entity e
$c^*(v)$	Adjust credibility of the value v of entity e
$\boldsymbol{W}$	Credibility vector of values to be true for entity e

3.1 值集之间的相似度计算

余弦相似度常用来计算文档向量之间的相似性,将文本中的词语映射到向量空间,形成文档中词频与向量数据的映射关系,通过计算两个向量之间的余弦相似度得出文档之间的相似度。例如计算下面两个句子的相似度:

A:“我爱母亲,也爱父亲”;

B:“我爱父亲,更爱母亲”。

首先对句子进行分词,得到分词集合{我 爱 母亲 也 父亲 更},计算词频:

A:我1,爱2,母亲1,也1,父亲1,更0;

B:我1,爱2,母亲1,也0,父亲1,更1。

得出词频向量$\boldsymbol{A}$(1,2,1,1,1,0)和$\boldsymbol{B}$(1,2,1,0,1,1)。通过计算两个向量的余弦相似度即可得两个句子的相似度。

因此,本文将余弦相似度引入到值集之间的相似度计算当中来。令向量$\boldsymbol{A}$表示值集V的二值向量,$\boldsymbol{A}$的长度为实体可能值集V^*的长度,则向量$\boldsymbol{A}$的第i个元素值为

$$\boldsymbol{A}[i]=\begin{cases}1 & V^*[i]\in V\\ 0 & V^*[i]\in V\end{cases} \tag{1}$$

式中,$V^*[i]$表示V^*的第i个元素。例如:可能值集合$V^*=\{a,b,c,d,e\}$,值集$V=\{a,c,d\}$,则V的二值向量$\boldsymbol{A}=(1,0,1,1,0)$。

通过余弦相似度来度量两个值向量之间的相似性为

$$\mathrm{sim}(\boldsymbol{A}_1,\boldsymbol{A}_2)=\frac{\boldsymbol{A}_1\bullet\boldsymbol{A}_2}{\|\boldsymbol{A}_1\|\times\|\boldsymbol{A}_2\|} \tag{2}$$

3.2 数据源可信度计算与实体多真值发现

3.2.1 基本推导

数据源可信度越高,则其提供的值集与实体的真值集相似度越高,反之,两者的相似度越低。因此,本文通过计算数据源提供的所有值集与实体真值集的平均相似度来度量数据源的可信度,用$\boldsymbol{A}^*$表

示实体的真值集,可得

$$t(s)=\frac{\sum_{V\in V(s)}\mathrm{sim}(\boldsymbol{A}^*,\boldsymbol{A})}{|V(s)|} \tag{3}$$

实体的真实集应该最大程度地接近冲突数据源提供的所有值集。为了找到最可能正确的真值集,结果应该在所有数据源提供的值集中相似度达得最大。因此,提出本文多真值发现的目标函数

$$\max\left(\sum_{e\in E}\sum_{s\in S}t(s)\mathrm{sim}(\boldsymbol{A}^*,\boldsymbol{A})\right) \tag{4}$$

到目前为止,已经将多真理发现转化为一个优化问题。根据目标函数可以在实体的可能值集中选取置信度最高的几个值成为真值。在多真值发现问题中,实体可能值的置信度通常是不同的,置信度高的描述值会更大可能成为实体的真值。因此,本文采用一种贪心选择策略:根据置信度的大小对实体的可能值进行排序,然后优先选择高可信度的描述值作为实体的真值。

对于描述值v,通过各数据源的加权投票来计算其置信度

$$c(v)=\frac{\sum_{s\in S(V)\wedge v\in V}t(s)}{\sum_{s\in S(V)}t(s)} \tag{5}$$

由式(5)得到实体每个可能值的置信度大小,从而生成置信度向量$\boldsymbol{W}$。根据$\boldsymbol{W}$按从大到小的顺序将可能值放入候选真值集,然后计算该真值集与实体的所有值集之间的相似度,保留相似度之和较大的真值集,最后相似度之和最大的真值集就是所求解。具体算法如算法1所示,算法的时间复杂度为$\boldsymbol{O}(\mathrm{ML})$。

算法1 实体真值发现

输入:V^*,$\{t(s)|s\in S\}$

输出:Truth

```
construct an empty set C;
while(v ∈ V*)
   compute c(v) according to Eq. (5);
end while
k=1;
while(k≤L)
   v=SelectTop(W, k);
   put v into C;
   while(s ∈ S)
      temp+=t(s)×sim(V, C);
   end while
   if temp>temp_max then
      Truth=C, temp_max=temp;
   else
      take v out of C;
   end if
   k++;
while(change)
```

3.2.2 结合相似值的影响

在现实中,同一个值有不同表现形式的情况是很常见的,例如,“Shell Wood”和“Wood”很可能是同个真值的不同表现形式。现有的多真值发现算法忽略了它们对真值的支持。同时,“Shell Wood”包含“Wood”,所以“Shell Wood”有更高的概率成为一个真值。在实际中,许多错误的值可能是由于数据不完整或缺少某系部分造成的,然而它们可以用来提高真值的置信度,从而提升真值发现的准确率。因此,本文提出了一种非对称的支持度计算方法:

令Z_1,Z_2分别表示描述值v_1,v_2包含的单词集合,m,n分别表示Z_1,Z_2中的单词个数,则v_1对v_2的支持度为

$$\sup(v_1,v_2)=\frac{\sum_{i\in[1,m],j\in[1,n]}\text{isSame}(Z_1[i],Z_2[j])}{n} \tag{6}$$

式中,isSame($Z_1[i]$,$Z_2[j]$)$\in\{0,1\}$,两个单词相等时取值1,否则取值0。例如,当v_1=“Wood”,v_2=“Shell Wood”时,sup(v_1,v_2)=1但是sup(v_2,v_1)=1/2。v_2有更高的概率成为真值。因此,根据式(6)可以对描述值的置信度进行修正,定义调和置信度

$$c^*(v)=c(v)+\beta\times\sum_{v'\in\overline{\text{Sim}}(v)}c(v')\sup(v',v) \tag{7}$$

式中,ρ是一个0和1之间的参数,控制相似值的影响。为了获得值v的调和置信度c^*,需要得到它的相似值列表。基于启发式思想:一个描述值的不同表现形式与该描述值不可能出现在同一值集中。因此,采用一种简单的方法,值v相似值列表中的值需满足两个条件:(1)对值v的支持度大于零。(2)不会出现在包含v的值集中。例如,“o'leary timothy j”根据条件1有两个相似值“o'leary linda i”和“timothy j”。然而,“o'leary timothy j”和“o'leary linda i”同时出现在一个值集中,因此很有可能是不同的值,根据条件(2),排除了“o'leary linda i”。具体算法如算法2所示。

算法2 相似值列表计算.

输入:V^*,$B=\{c(v)|v\in V^*\}$

输出:$\{\overline{\text{Sim}}(v)|v\in V^*\}$

```
if V* ∉ ∅ then
while(v ∈ V*)
  while(v' ∈ V*)
    temp= sup(v, v');
    if c(v') ∈ B∧temp ≠ 0 then
      put v' into Sim(v);
    end if
    if v' and v appear in a value set together
      take v' out of Sim(v);
    end if
  end while
end while
return Sim(v);
```

3.3 迭代计算

如上所述,若知道数据源的可信度,那么可以推导实体的真值集,反之亦然。与TruthFinder算法类

似,采用迭代的方法来联合推导数据源的可信度和实体的真值集。算法一开始并不知道关于数据源和真值集的信息,但每次迭代都进一步了解数据源的质量信息和实体的真值集,直到满足收敛条件时算法则会停止。下面给出了算法的总体流程。

首先,为所有数据源的可信度设置初始值 T_0(T_0为估计的平均可信度,本文设 $T_0=0.9$),然后开始迭代计算。每次迭代分两步:(1)使用从上一次迭代获得的数据源可信度来计算实体的真值集;(2)使用上一次迭代获得的真值集计算数据的源可信度。如此迭代直到算法达到稳定状态。稳定状态通过数据源可信度的变化来度量,用向量 $\boldsymbol{T}$ 来表示。使用余弦相似度来度量两次迭代之间 $\boldsymbol{T}$ 的变化。如果只在迭代之后 $\boldsymbol{T}$ 只改变了一点点,则算法停止。

算法3 算法框架.

输入:S,E

输出:$\{\text{Truth}|e\in E\}$

initialize the credibility of data sources $\{t(s)|s\in S\}$;

$n=0$;

do

$n++$;

while($e\in E$)

compute Truth according algorithm 2;

end while

while($s\in S$)

compute $t(s)$ according Eq.(4);

end while

until the algorithm satisfies the convergence condition;

return $\{\text{Truth}|e\in E\}$;

如果算法迭代 K 次,则本文算法的时间复杂度为 $O(\text{KMNL})$。

4 实 验

本节通过3个真实数据集比较了本文算法和现有的真值发现算法,并给出了实验结果。

4.1 实验设计

4.1.1 对比算法

Voting:该方法基于投票,如果一个描述值获得的票数占总票数的比例超过0.5,则认为该值为真。

Truthfinder[1]:该方法能联合推导数据源的可信度和值集的置信度,并考虑了不同值集之间的影响。

LTM[18]:该方法通过构建一个概率图模型来联合推导实体描述值的置信度和数据源的可信度。

MBM[13]:该方法定义一种新的描述值之间的互相排斥,同时结合更优的复制检测到一个贝叶斯模型中来进行真值发现。

MTD-hrd[14]:该方法通过结合两种影响,非平衡的肯定和否定断言的分布和描述值在同一值集共同出现的频数来增强它的概率模型。

SmartMTD[16]:该方法通过对两种类型的数据源关系进行建模来计算数据源的可信度和探测数据源的恶意复制。

OptMTF:本文提出的多真值发现算法。

4.1.2 数据集

(1) 图书-作者数据集

从O'Reilly官网中提取了一部分已出版的图书数据,包括书名、作者、出版年份、ISBN,并将这些数据当作真值。随机抽取100本图书,以ISBN为关键字,在abebooks.com网站上爬取相关图书数据作为图书数据集。为使问题更具挑战性,删除了只有轻微冲突的记录。处理后的数据集共包含来自685个数据源的10 583个冲突记录,平均每本书有8个可能的作者。作者可能值的数量分布如图1所示。

(2) 电影-导演数据集

从IMDB网站中提取了最流行的100部电影数据,包括电影和导演的名字。基于IMDB站点的权威性,将该站点的电影数据作为真值。然后根据选择的电影名称,采用了一种类似于文献[1]的做法在谷歌上进行搜索,提取了由不同站点提供的电影导演信息作为电影数据集。该数据集包括来自743个来源的403个导演的数据,平均每个电影有7个可能的导演。导演可能值的数量分布如图1所示。

(3) 父母-孩子数据集

采用文献[10]的做法在维基百科上提取与父母孩子相关的数据,同时使用最后一次编辑结果作为真值。与处理图书-作者数据集的方法类似,我们移除了较小冲突的数据。最终数据集有1 202个人的孩子的数据,平均每个人有6个可能的孩子信息。孩子可能值的数量分布如图1所示。

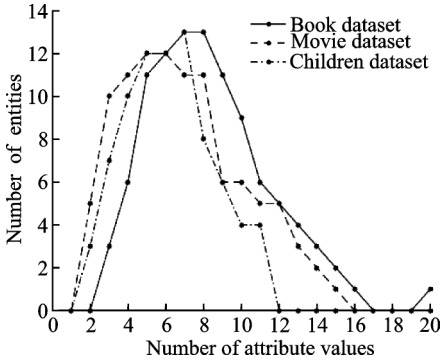

图1 实体可能值集大小的分布

Fig.1 Distribution of size of entity possible value sets

4.1.3 度量指标

使用3个指标来评估算法的性能。对于所有这些指标,较大的值表示更好的结果。

(1) 精确率Precision,表示在所有实体预测的真值集合中,预测实际真值的平均百分比;

(2) 召回率Recall,表示在所有实体实际的真值集合中,预测实际真值的平均百分比;

(3) 调和平均值F-score,精确率和召回率的调和平均值,范围从0到1。其计算公式为

$$F_1=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}} \tag{8}$$

4.1.4 实验环境

本节实验硬件环境为4 GB内存, 2.5 GHz Intel Core i5处理器和Windows 10操作系统。本文用JAVA语言实现了所有比较算法。

4.2 实验结果

4.2.1 对比现有的真值发现算法

表2展示了不同算法在3个真实数据集上准确度、召回率和的表现,已被加粗的是最优的结果。可以看到,对比现有的真值发现算法,本文算法的F-score始终能达到最好的结果。由于在图书数据集和父母数据集上消除了少量冲突的记录,所有算法在这两个数据集上准确度较低。同时,由于电影数据集的记录比其他两个数据集多,所有算法在电影数据集上运行的时间较长。

Voting算法在3个数据集上准确度较高,它的召回率是最低的,同时该算法的运行时间是最低的。这是因为大多数数据源只提供了小部分的完整真值集,同时Voting算法没有考虑数据源的可信度,可信度高的数据源提供的值没有得到更多的权重,因此降低了召回率。

表2 不同方法的比较

Tab.2 Comparison of different methods

Method	Book author dataset				Movie director dataset				Parent children dataset			
	Precision	Recall	F_1-score	t/s	Precision	Recall	F_1-score	t/s	Precision	Recall	F_1-score	t/s
Voting	**0.85**	0.65	0.73	**1.3**	**0.90**	0.65	0.75	**1.71**	0.87	0.60	0.71	**1.65**
TruthFinder	0.82	0.63	0.71	14.8	0.85	0.80	0.82	21.3	**0.82**	0.79	0.80	19.7
LTM	0.80	0.73	0.76	19.2	0.80	0.83	0.81	20.0	0.79	0.81	0.80	18.6
MBM	0.81	**0.75**	**0.78**	12.6	0.82	0.84	0.83	28.5	**0.82**	0.80	0.81	23.2
MTD-hrd	0.82	0.60	0.69	14.4	0.83	0.80	0.81	17.5	0.81	0.79	0.80	13.4
SmartMTD	0.81	0.74	0.77	10.2	0.84	**0.85**	0.84	16.4	0.81	0.82	0.82	14.5
OptMTF	0.82	0.74	**0.78**	9.8	0.86	**0.85**	**0.85**	20.3	**0.82**	**0.83**	**0.83**	15.3

Note: The best results are bolded.

TtruthFinder算法考虑了数据源的可信度和值集之间的相互影响,但其在图书数据集上的表现比Voting算法更差,这可能归因于该算法的单真值假设。需注意的是,在本文实验中,Voting算法是基于单个描述值而不是整个值集来计算票数的。例如,如果值集(A,B)得到2票,而值集(A,C)得到3票,那么A理应得到5票。

除本文算法之外,MBM算法和SmartMTD算法跟其他算法相比也有较好的表现,这是因为考虑了数据源的否定断言,从而提高真值发现的准确率。虽然MTD-hrd算法和LTM算法也考虑了这层含义,但它们对潜在变量的先验分布做出了很强的假设。如果数据集不符合假设的分布,那么算法的表现将会很差。然而,现有的多真值发现算法没有考虑值的不同表现形式,本文算法结合相似值对真值的影响来提高描述值置信度的计算精度,同时根据所提的目标函数选取可信度较高的描述值作为实体的真值,无需对数据源做出先验假设,因此OptMTF算法实现了更高的准确度。

4.2.2 相似值的影响

为了评估相似值的影响和结合相似值计算的重要性,实现了本文算法(即OptMTF)的另一个版本用于比较。

OptMTF-s:OptMTF的另一个版本,它没有考虑相似值对模型的影响。

图2展示了两种实现方法在电影数据集上的比较。可以看到,OptMTF的准确度和召回率明显高于OptMTF-s,尽管其执行时间稍长点。图3显示了两种方法在电影数据集上的迭代,这两种方法都可以在几次迭代之后达到收敛。这些数据证明了结合相似值支持的正确性。在现实中,同一个值具有不同表现形式的情况是很常见的。表3中列出了图书“Rapid Contextual Design” (ISBN:0123540518)的作者的相似值。现有的多真值发现算法认为它们是错误的值,但它们并不是完全错误的。它们通常是因为信息不完整或缺少某些部分造成的,结合它们的支持能够提高真值发现的准确性。特别是采用非对称的方法来计算值之间的支持度,使得完整值(即包含其他值)将获得更高的支持度,它们会比其他值更优先被选为真值。例如,当“Jessamyn Burns Wendell”被加入真值集时,根据对目标函数的计算,它的相似值“Jessamyn Wendell”几乎不可能被加入真值集,即使该值的调和置信度和真值很接近,通过这种方法可以得到更准确的真值结果。

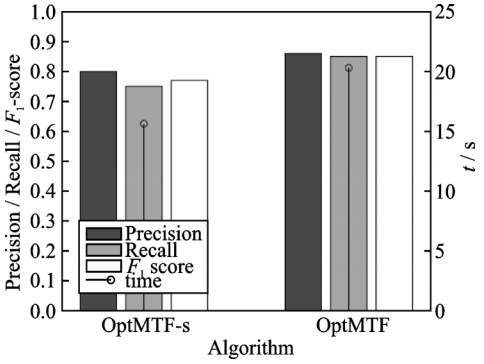

图2 两种方法在电影数据集上的对比

Fig.2 Comparison of two methods on movie dataset

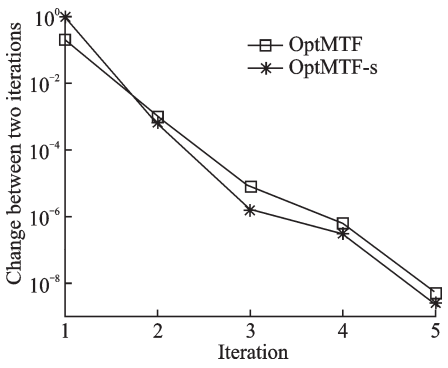

图3 两种方法在电影数据集上的迭代

Fig.3 Iteration of two methods on movie database

5 结束语

在数据集成系统中,从冲突数据中找到正确的信息是至关重要的。本文提出了一个多真值发现算法OptMTF。该算法将多真值发现转化为一个函数优化问题,其目标是实体的真值集应该与数据源对该实体提供的所有值集之间相似度最高。根据目标函数对真值的选择,设计了一个迭代算法来联合推到数据源的可信度和实体的真值集,同时,考虑值不同表现形式的影响,结合相似值的支持来计算描述值的置信度,达到更好的真值发现效果。最后通过3个真实数据集上的实验表明本文算法的有效性。

表3 几个真值的相似值表

Tab.3 Table of similar values of several true values

真值	相似值
Karen Holtzblatt	Holtzblatt-Karen
	Holtzblatt
	Holtzblatt Karen
Jessamyn Burns Wendell	Wendell-Jessamyn Burns
	Jessamyn Wendell
	Wendell Jessamyn
Shelley Wood	Shelley
	Wood Shelley

参 考 文 献:

[1] Yin X, Han J, Yu P S. Truth discovery with multiple conflicting information providers on the web[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2007: 1048-1052.

[2] Dong X L, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence[J]. Proceedings of the Vldb Endowment, 2010, 2(1): 550-561.

[3] Benslimane D, Sheng Q Z, Barhamgi M, et al. The uncertain web: Concepts, challenges, and current solutions[J]. ACM Transactions on Internet Technology, 2016, 16(1): 1-6.

[4] Li X, Dong X L, Lyons K B, et al. Scaling up copy detection[C]//IEEE, International Conference on Data Engineering. [S.l.]: IEEE, 2015: 89-100.

[5] Blanco L, Crescenzi V, Merialdo P, et al. Probabilistic models to reconcile complex data from inaccurate data sources[C]// International Conference on Advanced Information Systems Engineering. [S.l.]: Springer-Verlag, 2010: 83-97.

[6] Dong X L, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 562-573.

[7] Qi G J, Aggarwal C C, Han J, et al. Mining collective intelligence in diverse groups[C]//Proceedings of the 22nd International Conference on World Wide Web. [S.l.]: ACM, 2013: 1041-1052.

[8] Pochampally R, Das Sarma A, Dong X L, et al. Fusing data with correlations[C]//Proceedings of the 2014 ACM SIGMOD

International Conference on Management of Data. [S.l.]: ACM, 2014: 433-444.

[9] Galland A, Abiteboul S, Senellart P. Corroborating information from disagreeing views[C]//ACM International Conference on Web Search and Data Mining. [S.l.]: ACM, 2010: 131-140.

[10] Pasternack J, Roth D. Knowing what to believe (when you already know something)[C]//Proceedings of the 23rd International Conference on Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2010: 877-885.

[11] Liu X, Dong X L, Ooi B C, et al. Online data fusion[J]. Proceedings of the Vldb Endowment, 2011, 4(11): 932-943.

[12] Zhao Z, Cheng J, Ng W. Truth discovery in data streams: A single-pass probabilistic approach[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. [S.l.]: ACM, 2014: 1589-1598.

[13] 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.

[14] Wang X, Sheng Q Z, Fang X S, et al. An integrated Bayesian approach for effective multi-truth discovery[C]//ACM International on Conference on Information and Knowledge Management. [S.l.]: ACM, 2015: 493-502.

[15] Wang X, Sheng Q Z, Yao L, et al. Truth discovery via exploiting implications from multi-source data[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. [S.l.]: ACM, 2016: 861-870.

[16] Wang X, Sheng Q Z, Yao L, et al. Empowering truth discovery with multi-truth prediction[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. [S.l.]: ACM, 2016: 881-890.

[17] Fang X S, Sheng Q Z, Wang X, et al. SmartMTD: A graph-based approach for effective multi-truth discovery[EB/OL]. [2017-08-07](2018-09-20).https://arxiv.org/abs/1708.02018.

[18] Li Y, Gao J, Meng C, et al. A survey on truth discovery[J]. ACM Sigkdd Explorations Newsletter, 2016, 17(2): 1-16.

作者简介:

陈烈锋(1993-),男,硕士研究生,研究方向:数据融合、真值发现,E-mail: 1151853866@qq.com。

许青林(1963-),男,副教授,研究方向:云计算、软件工程和企业信息化,E-mail: gzj2ee@126.com。

(编辑:张彤)