基于模糊偏序关系支持度模型的真值发现算法

李少波[1,2]，王继奎[1]，杨观赐[2]

(1.中国科学院成都计算机应用研究所 成都 610041；2. 贵州大学现代制造技术教育部重点实验室 贵阳 550003)

【摘要】为了解决主数据集成、web数据集成中的真值发现问题，提出了一种基于模糊偏序关系支持度计算模型的真值发现算法(FA-SDCM)。针对已有算法中，以描述相似度替代描述支持度进行计算，忽视了描述所含真值信息的不对称性问题，在分析描述本身特性的基础上，提出了描述蕴含概念，定义了基于模糊偏序关系的支持度计算模型，较好地解决了描述所含真值信息的不对称性问题。在考虑了数据源可信度及描述之间支持度对真值发现影响的基础上，基于迭代思想，提出了FA-SDCM算法。在Books-Authors数据集上进行实验，结果表明FA-SDCM算法比Vote算法与TruthFinder算法具有更高的准确率。

关 键 词 不对称性； 描述蕴含； 模糊偏序关系； 支持度模型； 真值发现

中图分类号 TP311 **文献标志码** A doi:10.3969/j.issn.1001-0548.2014.06.017

True Value Finding Algorithm Based on a Support Degree Calculation Model Using Fuzzy Partial Order Relation

LI Shao-bo[1,2],WANG Ji-kui[1], and YANG Guan-ci[2]

(1. Chengdu Institute of Computer Applications, Chinese Academy of Sciences Chengdu 610041;

2. Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University Guiyang 550003)

Abstract In order to find the true values in master data integration and web data integration, we propose a true value finding algorithm (FA-SDCM) based on a support degree calculation model using fuzzy partial order relations. In existing algorithms, support degrees are usually substituted by similarity, which ignores the asymmetry in the true vales. In this paper, the concept of description containing is proposed through analyzing characteristics of descriptions, and then a support degree calculating model is developed based on fuzzy partial order relations to solve the description of asymmetric problems in the true values. Considering the influence of the data source reliability and the support degrees among descriptions on true value finding, the FA-SDCM algorithm is realized iteratively. An experiment has been carried on the Books-Authors data set, and the result shows that the FA-SDCM algorithm has better accuracy than the Vote and the TruthFinder algorithms.

Key words asymmetry; description containing; fuzzy partial order relations; support degree calculation model; true value finding algorithm

针对冲突数据的真值发现问题，研究者们进行了一系列探讨。文献[1]对数据集成中的冲突处理策略进行了总结。文献[2-4]注意到了web数据的特点，考虑了web数据源之间的复制关系，给出了刻画数据源复制依赖关系的方法。文献[5]首次提出了web世界的真值发现问题，提出了TruthFinder算法。文献[6-8]等进一步考虑了数据源的准确性因素，并将其与数据源的依赖关系结合起来。文献[9-10]采用一种不同的概率投票方法，另外还考虑了投票数据源的权威性。文献[11]提出了基于Markov逻辑网的两阶段数据冲突解决方法，依据冲突程度分两个阶段解决冲突问题。上述均是以描述之间相似度代替支持度进行计算。而不同描述间的支持度计算是冲突数据真值发现的核心环节，值得进一步研究。本文的工作包括：1) 提出了描述蕴含概念，并提出了基于模糊偏序关系的支持度计算模型。2) 基于迭代思想，提出了基于模糊偏序关系支持度模型的真值发现算法(true value finding algorithm based on support degree calculation model using fuzzy partial order relation，FA-SDCM)；3) 在Books-Authors数据集上进行实验，结果表明FA-SDCM算法比Vote算法、TruthFinder算法具有更高的准确率。

收稿日期： 2013 – 07 – 04；修回日期：2014 – 01 – 16

基金项目：国家科技支撑计划项目(2012BAF12B14)；国家自然科学基金(51475097)；贵州省科技项目(黔科合JZ字[2014]2001、黔科合计Z字[2012]4009)

作者简介：李少波(1973 –)，男，教授，博士生导师，主要从事大数据、制造物联、计算智能等方面的研究.

1 问题描述

在主数据集成、web数据集成等过程中，经常碰到针对同一客观实体描述不一致的情况；数据集成必须解决这些冲突。为了方便讨论，本文只对客观实体的某一属性进行研究。比如，针对《Computing Essentials》书籍实体的作者属性，不同的Web数据源提供了相互冲突的信息，如表1所示。

表1 《Computing Essentials》书籍作者的冲突信息

数据源	作者
Collegebooksdirect.com	O'Leary, Timothy J.; O'Leary, Linda I.
LGTextbooks.com	O'LEARY
G.T.S.	Timothy J O'Leary; Linda I O'Leary
A1Books	Timothy J Oandapos;Leary, Linda I Oandapos;Leary
Limelight Bookshop	O'Leary, Linda I. J.

从表1可以看出，不同的网上书城对于同一本书作者的描述有很大差别。

1.1 假设

1) 真值是唯一的，并且存在于冲突数据之中。此假设旨在简化讨论；

2) 独立数据源提供的描述是相同或者是很相似的，每个数据源提供的描述都是真值的一个视图，都部分或全面地反映了真值信息；

3) 独立数据源提供的相同错误值的可能性很小。这符合人们的直觉，判断两个学生作弊的有力证据就是两个学生的答卷错误相同；

4) 提供了更多真值的数据源更可信。这也符合人的直觉，如果一个人总是说真话，那么他人也更容易相信他所说的话。

通过以上假设认识到作为真值的描述往往与较多的描述存在支持关系，也获得较多的支持值；而错误的描述往往差别较大，一致性差。所以通过描述之间支持度修正描述的可信度，会使正确的描述可信度增加，从而增加被算法识别为真值的概率。

1.2 定义

定义 1 描述 f

客观实体的一个取值，定义为该客观实体的一个描述。$o(f)$ 代表描述 f 对应的实体。$o(f_i)=o(f_j)$ 表示描述 f_i 与描述 f_j 描述了同一实体。

定义 2 提供描述 f 的集合 $w(f)$

$w(f)=\{w \mid f \in F(w)\}$，$F(w)$ 表示数据源 w 提供的描述集合。

定义 3 蕴含关系

设set_1是字符串s_1的义原词集合，set_2是字符串s_2的义原词集合，义原词集合调用中科院ICTCLAS分词软件的Java接口，经过自定义函数格式化后获取。

若set_1为set_2的子集，称set_2蕴含set_1，用$s_2 \neg s_1$表示。假设：$\text{set}_1=\text{set}_2$，可推导出$s_2=s_1$，即忽略义原词位置对字符串造成的影响。

性质 1 $s_1 \neg s_1$永真，蕴含关系具有自反性。

性质 2 若$s_2 \neg s_1$，且$s_1 \neg s_2$，则$s_2=s_1$，蕴含关系具有反对称性。

性质 3 若$s_1 \neg s_2$，且$s_2 \neg s_3$，则$s_1 \neg s_3$，蕴含关系具有传递性。

定义 4 模糊偏序关系模型

设：$U=\{x_1, x_2, \cdots, x_n\}$，则$(U, \neg)$是个偏序集，如果二元函数$R$满足以下性质：

设：$i,j,k \in \{1,2,\cdots,n\}$

1) $R(x_i,x_j) \in [0,1]$

2) if $x_i \neg x_j$ then $R(x_i, x_j)=1$

3) if $x_i \neg x_j$ then $R(x_i, x_k) \geqslant R(x_j, x_k)$

4) if $x_i \neg x_j \neg x_k$ then $R(x_i, x_k) \geqslant R(x_j, x_k)$

则称$(U, \neg, R)$为模糊偏序关系模型[12]。

定义 5 描述 f 的可信度 $s(f)$

描述 f 的可信度与提供 f 的数据源的可信度有关，有：

$$s(f)=1-\prod_{(w\in W(f))}(1-t(w)) \tag{1}$$

式中，$t(w)$ 表示数据源w的可信度。

定义 6 数据源w的可信度$t(w)$

数据源的可信度由其提供的描述的可信度决定，即其所提供所有描述可信度的算术平均值，有：

$$t(w)=\sum_{(f\in F(W))} s(f)/|F(W)| \tag{2}$$

式中，$t(w)$ 表示数据源w的可信度；$F(W)$ 表示数据源 w 提供的描述集合。

定义 7 数据源w的可信值 $\tau(w)$

为了方便计算，避免经过多次乘法操作后溢出，定义：

$$\tau(w)=-\ln(1-t(w)) \tag{3}$$

定义 8 描述 f 的可信值 $\sigma(f)$

$$\sigma(f)=-\ln(1-s(f)) \tag{4}$$

结合式(1)与式(3)：

$$\sigma(f)=-\ln\left(1-1+\prod_{(w\in W(f))}(1-t(w))\right)= -\ln\left(\prod_{(w\in W(f))}(1-t(w))\right)=\sum_{w\in W(f)}\tau(w) \tag{5}$$

描述的可信度值为所有提供相同描述的数据源可信值之和。

2　支持度计算模型

对于字符串型描述来说：比如作者列表描述A=“Scambray, Joel”，描述B=“Scambray,Joel; McClure, Stuart”，如果采用相似度进行计算，则Support(A,B)=Support(B,A)=0.5；但是从主观上来看，描述B比描述A更可信，而且出现这种情况描述A也完全失去了成为真值的可能，因为描述B蕴含描述A，且信息量更大。由此可见描述本身的特性比数据源的可信度更重要。

2.1　支持度计算

文献[5]在计算描述间支持度时，采用了$\mathrm{imp}(s_1,s_2)=\mathrm{sim}(s_1,s_2)-\mathrm{base_sim}$计算公式，其中$\mathrm{sim}(s_1,s_2)$表示采用编辑距离法计算的相似度；base_sim为相似度阈值，设置为0.5，其含义是如果两个字符串的相似度大于阈值表示支持度为正，否则支持度为负。这个算法看起来很有道理，但是其基本思路依然是以相似度代替支持度进行计算。比如前面列举的例子：作者列表描述A=“Scambray; Joel”，描述B=“Scambray, Joel, McClure, Stuart”。利用文献[5]所使用的公式计算：$\mathrm{imp}(A,B)=\mathrm{imp}(B,A)=2/4-0.5=0$。从某种意义上来说，这也是很有道理的，毕竟两种描述存在着巨大的差距。但是通过观察，发现描述B蕴含描述A，描述B更有可能成为真值，即使其余的数据源都支持描述A，描述A也不能成为真值，这种情况往往是由于描述A是部分信息的缘故。如果存在描述C=“Scambray”，则$\mathrm{imp}(C,B)=\mathrm{imp}(s_1,s_2)=1/4-0.5=-0.25$。描述$C$是描述$B$的一部分，理应成为支持描述$B$理由，结果却反对了描述$B$。这些分析表明，采用文献[5]所采用的算法存在着改进的空间。

设$s_1(a_i)$，$s_2(a_j)$为字符串s_1，s_2的义原词；m，n代表字符串s_1，s_2包含的义原词个数，n不等于0。$\mathrm{support}(s_1,s_2)\in[0,1]$计算公式为：

$$\begin{cases}\mathrm{sup\,port}(s_1,s_2)=1 & s_1\neg s_2\\ \dfrac{\sum\limits_{i\in[1,m];\,j\in[1,n]}\mathrm{isSame}(s_1(a_i),s_2(a_j))}{n} & \text{otherwise}\end{cases}\tag{6}$$

式中，$\mathrm{isSame}(s_1(a_i),s_2(a_j))\in\{0,1\}$，两个义原词相等时候取值1，否则取值0。利用式(6)进行计算$\mathrm{support}(C,B)=0.25$，$\mathrm{support}(B,C)=1$。这种方法避免了被蕴含的描述由于出现次数较多成为真值的可能，通过观察发现，描述A很大可能为不完整数据。

由计算公式可以看出通常情况下$\mathrm{support}(s_1,s_2)$与$\mathrm{support}(s_2,s_1)$不相等，它表明字符串之间的支持度是非对称的。从现实中来看，这比较符合实际，因为如果在一次对比中的整体不能成为真值，那么部分成为真值的概率很小，即使在部分为真值的情况下，选择含有更多信息的描述作为真值也是个不错的选择，至少没有遗失信息。

针对不同的数据类型，需要设计不同的支持度计算模型。

2.2　模糊偏序关系支持度计算模型

对冲突数据进行分析发现，很多的描述为不完整的真值信息，或者是少量的错写、简写。文献[12]表明模糊偏序关系在信息融合计算中起着重要的作用，冲突数据的真值发现也是一种信息融合。

设：$F=\{f_1,f_2,\cdots,f_n\}$，set_i为f_i的义原词的集合，¬ 为定义3的蕴含关系。支持度计算模型(F，¬，support)为模糊偏序关系模型。

证明：由式(6)可知$\mathrm{sup\,port}(f_i,f_j)\in[0,1]$；如果$f_i\neg f_j$，由式(6)可知$\mathrm{sup\,port}(f_i,f_j)=1$；如果$f_i\neg f_j$，$\forall f_k$，则$\mathrm{set}_i\cap\mathrm{set}_k\supseteq\mathrm{set}_j\cap\mathrm{set}_k$，由式(6)可知$\mathrm{sup\,port}(f_i,f_k)\geqslant\mathrm{sup\,port}(f_j,f_k)$；如果$f_i\neg f_j\neg f_k$，由式(6)可知$\mathrm{sup\,port}(f_i,f_k)=1$，且$\mathrm{sup\,port}(f_j,f_k)=1$。

所以，二元关系$\mathrm{sup\,port}$在偏序集(F，¬)上满足定义4的四条性质。

3　FA-SDCM算法

3.1　描述 f 的可信值$\sigma(f)$的修正

由于不同描述之间相互关系会对描述的可信值产生影响，所以利用以下公式进行调整：

$$\sigma^*(f)=\sigma(f)+\rho*\sum_{(o(f')=o(f))}\sigma^*(f')*\mathrm{support}(f',f)\tag{7}$$

$\rho\in[0,1]$代表支持度调节因子。显然，$\sigma^*(f)\geqslant\sigma(f)$，此调整函数增大了相互间有支持关系的描述的可信度，从而使获得较多支持的描述增大了成为真值的可能。

3.2　$s(f)$的修正

考虑到描述间的支持度对描述可信值的影响，为了使可信度与可信值保持一致，从而对描述的可信度做相应的修正，参考文献[5]的方法，γ为调节因子，有：

$$s(f)=1/(1+\mathrm{e}^{(-\gamma*\sigma^*(f))})\tag{8}$$

3.3　FA-SDCM算法实现

参数：数据源可信度向量$\boldsymbol{t}$，数据源可信值向量$\boldsymbol{\tau}$，描述可信度向量$\boldsymbol{s}$，描述可信值向量$\boldsymbol{\sigma}^*$。

矩阵$\boldsymbol{A}$为：

$$A_{ij}=\begin{cases}\dfrac{1}{|F(w)|} & f_j\in F(w_i)\\ 0 & \text{otherwise}\end{cases}$$

矩阵$\boldsymbol{B}$为：

$$B_{ji}=\begin{cases}1 & f_j\in F(w_i)\\ \rho*\text{support}(f_i,f_j);o(f_i,)=o(f_j) & \\ 0 & \text{otherwise}\end{cases}$$

由式(2)可得:

$$\boldsymbol{t}=\boldsymbol{A}\,\boldsymbol{s}$$

由式(5)和式(7)可得：

$$\boldsymbol{\sigma}^{*}=\boldsymbol{B}\,\boldsymbol{\tau}$$

3.3.1 FA-SDCM算法实现

输入：冲突的数据源集合W，描述的集合F，数据源与描述的联系对象集$O(W,F)$

输出：每个实体对象的真值

1) 计算蕴含关系与并初始化矩阵$\boldsymbol{A}$，$\boldsymbol{B}$

2) 给向量$\boldsymbol{t}$赋初值$\boldsymbol{t}$

3) 根据式(3)计算$\boldsymbol{\tau}$

repeat /*迭代计算*/

4) $\boldsymbol{\sigma}^{*}=\boldsymbol{B}\,\boldsymbol{\tau}$

5) 利用式(8)计算向量$\boldsymbol{s}$

6) 保存$\boldsymbol{t}$到临时向量$\boldsymbol{t}'$

7) $\boldsymbol{t}=\boldsymbol{A}\,\boldsymbol{s}$

8) 根据式(3)计算$\boldsymbol{\tau}$

9) until $\boldsymbol{t}$与$\boldsymbol{t}'$余弦相似度大于阈值$1-\beta$

10) 输出可信度最大的描述，即实体对象真值。

3.3.2 算法分析

设总的描述数为m，不同的描述个数为n，由于很多描述相同，所以通常m远大于n。设每个实体平均有k个描述，则每个描述与其他$k-1$个描述冲突。矩阵$\boldsymbol{A}$计算的时间复杂度为$O(m)$，矩阵$\boldsymbol{B}$的时间复杂度为$O(km)$;考虑到算法的迭代实现，设迭代次数为i，则算法的时间复杂度为$O(im+ikm)$。算法的空间复杂度为$O(m+kn)$。从以上分析可以看到，算法有着线性的时间与空间复杂度。这种算法采用为数据源赋初始可信度的方法，不需要通过学习获得初始值。

4 实验验证

4.1 实验条件设置

实验所使用的数据集为Books-Authors数据集(哈尔滨工程大学张志强教授提供)；对比的算法为Vote算法、TruthFinder算法及FA-SDCM算法。使用Eclipse3.4、jdk1.6作为IDE编程环境，操作系统为64位Win7，内存8 G。FA-SDCM采用中科院ICTCLAS分词系统获取义原词集合。实验表明在ρ>0.1的情况下对实验结果影响很小，不同的γ，t同样对实验结果影响同样很小。为了使结果具有可比性，选取了与文献[5]相同的运行参数。运行参数设置为t=0.8，ρ=0.5，γ=0.3，$\beta=1-10^{-5}$。

4.2 冲突数据集中有蕴含关系描述分析

实验选取Books-Authors部分数据，共11本书，ISBN ∈{9780596006877，9780072999389，9780072843996，9780072232172，1558608893，155860846X，131872893，131463055，130284467，9780072230611}，从Annex Books Inc等84个数据源取得数据。实验表明没有蕴含关系的描述数由总的描述数272个减少至74个，没有蕴含关系的描述仅为总描述数的27.2%。实际观察也发现，很多描述正确但不完整。

4.3 算法精确率对比

查阅书籍的封面获取真值信息，计算准确率。

如表2所示，FA-SDCM算法的准确率比Vote算法提高了36.37%，比TruthFinder算法提高了9.09%。多数的网站只列出了部分作者，是部分完全正确的信息，所以使用Vote算法选出的很多不是真值，这体现了不能忽视描述本身的特性；而TruthFinder使用了基于编辑距离的字符串相似度算法替代支持度算法，计算的结果经常遗漏作者，这是忽视了描述本身的特性；而FA-SDCM算法采用了非对称的支持度计算模型，有意的降低了被蕴含描述成为真值的可能，对于存在大量部分正确描述的情况准确率较高。

表2 Books_Authors数据集上准确率对比

算法	准确率/%
Vote	36.36
TruthFinder	63.64
FA-SDCM	72.73

5 结 论

目前数据集成中的冲突数据真值发现算法侧重于研究数据源间的依赖关系及数据源的可信度对于描述可信度的影响，对于描述之间的支持度对于描述可信度的影响研究较少，在计算时以相似度代替支持度，忽视了描述所含真值信息的不对称性问题。FA-SDCM算法优先考虑了描述本身的特点，其次才是数据源可信度的影响。在描述蕴含概念的基础上提出了基于模糊偏序关系的支持度计算模型，并利用探索性计算常用的迭代思想实现了FA-SDCM算法。考虑到web世界中实体真值及主数据的代表记录

是不断动态变化的，下一步拟从数据世系的角度进行研究。

参 考 文 献

[1] BLEIHOLDER J, NAUMANN F. Conflict handling strategies in an integrated information system[C]// Proceedings of the International Workshop on Information Integration on the Web. Edinburgh, UK: ACM , 2006: 36-41.

[2] 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, 20(6): 796-808.

[3] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer networks and ISDN systems, 1998, 30(1): 107-117.

[4] BERTI-EQUILLE L, SARMA A D, MARIAN A, et al. Sailling the information ocean with wareness of currents: Discovery and Application of Source Dependence[C]// Proceedings of CIDR Aoilomar. CA, USA: ACM, 2009: 1-6.

[5] YIN X, HAN J, YU P S. Truth discovery with multiple conicting information providers on the Web[J]. IEEE Transactions on Knowledge and Engineering, 2008, 20(6): 796-808.

[6] DONG X L, BERTI EQUILLE L, SRIVASTAVA D. Integrating conflicting data: The role of source dependence [C]//Proceedings of the VLDB Endowment. Lyon, France: ACM, 2009, 2(1): 550-561.

[7] DONG X L, NAUMANN F. Data fusion resolving data conflicts for integration[C]//Proceedings of the VLDB Endowment. Lyon, France: ACM, 2009, 2(2): 1654-1655.

[8] DONG X L, BERTI EQUILLE L, SRIVASTAVA D. Truth discovery and copying detection in a dynamic world[C]// Proceedings of the VLDB Endowment. Lyon, France: ACM , 2009, 2(1): 562-573.

[9] 考明军, 张炜, 高宏. 冲突数据中的真值发现算法[J]. 计算机研究与发展, 2010, 47(增刊): 188-192.
KAO Ming-Jun, ZHANG Wei, GAO Hong. Truth discovery methods in conflict data integration[J]. Journal of Computer Research and Development, 2010, 47(Supplement): 188-192.

[10] 张志强, 刘丽霞, 谢晓琴, 等. 基于数据源依赖关系的数据源信息评价方法研究[J]. 计算机学报, 2012, 11(35): 2392-2402.
ZHANG Zhi-qiang, LIU Li-xia, XIE Xiao-qin, et al. Information evaluation based on source dependence[J]. Chinese Journal of computers, 2012, 11(35): 2392-2402.

[11] 张永新, 李庆忠, 彭朝晖. 基于Markov逻辑网的两阶段数据冲突解决方法[J]. 计算机学报, 2012, 35(1): 101-111.
ZHANG Yong-xin， LI Qing-zhang， PENG Zhao-hui. 2-Stage data conflict resolution based on Markov logic networks[J]. Chinese Journal of computers, 2012, 35(1): 101-111.

[12] 仇国芳, 李怀祖. 模糊偏序关系模型的信息融合方法[J]. 工程数学学报, 2003, 20(2): 72-76.
QIU Guo-fang, LI Huai-zu. Information aggregation based on fuzzy preference relations[J]. Chinese Journal of Engineering Mathematics, 2003, 20(2): 72-76.

编 辑 蒋 晓