doi:10.3772/j.issn.2095-915x.2017.01.007

# 面向学术资源集成的真值发现算法

中国科学技术信息研究所 北京 100038
董微 杨代庆

**摘 要** 在构建多渠道元数据资源建设体系时，往往存在着大量的元数据冲突的问题，即对同一对象的属性存在多种描述，造成了元数据的组织与揭示的困难。本文处理的原则是遵从原文，优先选取原文的值作为唯一的真值，将数据冲突问题视为单真值冲突问题。考虑到数据提供商均需要加工数据，将数据源之间的关系视为相互独立。根据以上，本文提出了一种面向学术资源集成的真值发现算法。该算法基于贝叶斯算法，考虑了有关联关系的属性。实验证明本文方法所构造的真值发现算法在保证准确率的同时，大大节省了人力的工作时间。

**关键词：** 资源建设，元数据集成，冲突数据，真值发现
**中图分类号：** G250.7

# Academic Resource Integration Oriented Truth Discovery Algorithm

Center of Information Technical Support, Institute of Scientific and Technical Information of China, Beijing 10038, China
DONG Wei YANG DaiQing

**Abstract** Metadata resources construction in constructing multi-channel system, there are often a lot of metadata conflict problem, namely, there are many description of the same object attribute, which lead to difficulties for organization and reveal of the metadata. In this paper, the principle of treatment was to follow the original document, preferred to select the value of the original as the only true value, and took the data conflict as a single true value conflict. Considering the data provider all need data processing, this study took the data source as independent to each other. According to above, this paper proposed an algorithm of the true discovery for integration of academic resources. The algorithm was based on the Bayesian algorithm, considering the relationship between the related properties. Experiments showed that the method can main the accuracy of truth discovery and greatly reduced manpower work time.

**Keywords:** Resources construction, metadata integration, conflicting data, truth discovery

**基金项目：** 本文受NSTL专项基金项目：开放学术资源建设（2016XM16）的资助。
**作者简介：** 董微（1987–），博士，研究方向：数据挖掘、网络爬取，email：dongw@istic.ac.cn；杨代庆（1975–），高级工程师，研究方向：大数据分析、数据挖掘、图书情报。

## 1 引言

随着信息时代与互联网的发展，Web 数据已经成为了信息的主要表现形式和传播模式。目前图书馆以印本为主的传统的数据加工模式已难以适应时代发展，而数字出版潮流正在日益繁荣。为适应数字出版环境，图书馆在印本加工的同时，构建多渠道元数据资源建设体系。通过构建多渠道元数据资源建设体系，将学术资源进行全面地采集、组织与揭示，实现开放获取学术资源的高效传播与有效利用，形成满足用户需要的新的信息资源体系。在元数据集成过程中，存在诸多问题，其中一个重要的问题是：常出现多个数据源对同一实体对象的描述存在冲突；将多数据源的元数据进行整合，解决数据冲突往往需要大量的人工辅助，十分耗费人力，而且造成数据不能及时更新。Yin 等人 [1] 首次提出将上述冲突处理问题定义为真值发现问题。给定一个数据源的集合（例如图书网站集合），以及针对某个属性（例如作者属性），各数据源为同一对象（例如一本书）提供不同数据值（例如作者列表）的集合，据此判断每个数据值的准确性以及各数据源的可信性。

在学术资源元数据集成过程中，可将元数据的组织视为真值发现过程。真值发现问题最简单直观的方法是采用投票机制，根据数据值的投票数来判断其准确性。但是这种方法对每个数据源同等对待，没有考虑不同数据源可信性的差异。在集成出版社、集成商、数据商提供的元数据时，均需要对数据进行加工，因此可以认为各个数据源之间均提供自身所包含的事实，而非从别的数据源复制。因此本文针对学术资源元数据集成进行研究，考虑了数据源投票值与数据源准确率不一致的情况，提出了一种面向学术资源集成的真值发现算法，提高数据值的准确性，同时也可大大的节省人力资源。

## 2 相关工作

在复杂的 Web 环境下，数据源的质量作为最直接的因素，被众多早期的研究考虑到真值发现算法中。文献 [1] 首次正式定义真值发现问题，提出 TruthFinder 算法来联合推导真值和数据源质量。在此基础上 Galland 等人 [2] 介绍了一个概率模型，同时考虑了数据源的覆盖率，提出了 Cosine、2-Estimates、3-Estimates 等算法。Pastenack 等人 [3] 提出 Sums、AverageLog、Investment 等算法，将用户的先验知识作为约束条件构建真值发现算法。

同时 Web 数据源之间往往存在复制依赖关系，数据源之间的复制关系也是影响真值计算的重要因素之一。Dong 等人 [4] 提出基于贝叶斯的方法判断数据源之间的依赖关系，根据“独立数据源提供的错误应该互不相同”判断数据源之间的复制关系，但无法处理多数据源间可能存在的更为复杂的依赖关系。Dong 等人 [5] 在文献 [4] 的基础上考虑了信息的时效性问题，采用隐马尔可夫模型来判断数据源之间的拷贝关系和拷贝的时间点。文献 [6] 从单一属性方面来判断数据源之间的依赖关系。文献 [7] 提出了综合考虑对象多个属性信息来改进数据源复制关系的判断方法。

除此之外，真值发现问题还有一系列研究工作，通过考虑真值发现判断的各种因素，采用

doi:10.3772/j.issn.2095-915x.2017.01.007

基于聚类算法[8]、数据抽取器[9]、半监督学习[10]等方法来提高真值发现的准确率和计算的效率。

上述的真值发现方法基本上是以整个 Web 站点作为研究对象，鲜少有以学术资源的元数据对象作为分析的对象。在元数据集成过程中，若涉及多值属性，可能存在多个真值，如作者名称，由于涉及元数据的组织与揭示，本文处理的原则是遵从原文，优先选取原文的值作为唯一的真值，因此本文主要处理单真值下的数据冲突。考虑到数据提供商均需要加工数据，可认为数据源之间相互独立。基于此，本文以学术资源的元数据对象作为研究对象，提出了一种面向学术资源集成的真值发现算法。该算法基于贝叶斯算法，针对特定属性，将数据源分为提供此属性值的数据源与不提供此属性值的数据源，考虑其对真值概率的奖惩，采用数据源的精确率衡量数据源的质量，并对有关联的属性采用联合概率计算其真值。实验证明本文方法所构造的真值发现算法在保证准确率的同时，大大节省了人力的工作时间。

## 3 真值发现算法

### 3.1 相关定义

本文研究学术资源集成过程中的真值发现问题，其目标是在多数据来源中，自动发现一个对象在某个属性上的真值，并且本着遵从原文的原则。下面将该问题进行定义。

数据源：真值发现中数据的提供者。

对象：表示某种类型实体的具体对象，它可由多个属性描述。

属性与属性值：属性是针对对象某个方面的描述，每个属性可能有多个冲突的属性值，提供每个属性值的数据源可能有多个。

真值：真值是符合现实世界描述的事实，对于属性，有的可能有单个属性值，有的可能有多个属性值，由于本文本着遵从原文的原则，因此本文只针对单真值问题进行处理。

令冲突数据集 $CD=\{r_1,r_2,\ldots\ldots,r_n\}$ 为输入信息，其中 $r_i$ 是形如（$O,V,S$）的三元组，$o$ 是一个对象，$V$ 是对象的属性值集合，$S$ 是数据源集合。假设输入数据已经做了去重处理，每条记录都是唯一的，那么 $r_i$ 表示针对一个对象的多来源描述。其中，$O=\{o_1,o_2,\ldots\ldots,o_n\}$，表示对象的集合；$S=\{s_1,s_2,\ldots\ldots,s_m\}$，表示数据源的集合。数据源 $s\in S$ 可以为对象 $o\in O$ 的特定属性提供一个属性值 $v$。本文假设在不同数据源为对象 $o\in O$ 的某个属性 $k$ 提供的属性值 $V(k)=\{v_1,v_2,\ldots,v_K\}$ 中，只有一个属性值为真值。

### 3.2 问题描述

令 $\varphi_k$ 表示与属性 $k$ 相关的属性值 $\varphi_k=V(k)=\{v_1,v_2,\ldots\ldots,v_K\}$ 的集合。在计算属性值 $v_i$ 为真值的概率时，需要同时考虑提供 $v_i$ 的数据源对它的支持以及没有提供 $v_i$ 的相关数据源对它的惩罚。

部分学术资源的属性 $k$，如国别、页码等，其真值的选择具有独立性，其它与属性 $k$ 无关的数据源提供的属性值对 $v_i$ 为真值的概率计算并没有影响，即属性值 $v_i$ 的真值概率为 $P(v_i)=\rho(v_i\,|\,\varphi_k)$。

然而另一部分学术资源的属性，如作者名称与作者单位，其真值的选择具有相关性。例如，有的数据源将作者名称与作者单位进行拆分，将一个作者的姓与名拆成多个作者，

相应的作者单位也拆分成多个单位；同时也存在，由于作者姓名的表达方式与原文不同被判定为非真值，但其数据源所提供的作者单位正确的情况。对于此类的属性，其属性值的真值概率采用联合概率。设两个属性分别为 $k_1$、$k_2$，属性值集合为 $\varphi_{k1}=\{v_1,v_2,\ldots,v_K\}$、$\varphi_{k2}=\{u_1,u_2,\ldots,u_T\}$，那么 $k_1$ 的真值概率如公式（1），$k_2$ 的真值概率亦然。其中，若存在 $\tau$ 个数据源均提供属性值 ，那么 $\xi$ 为提供 $k_1$ 数据源的可信度的最大值。

$$P(v_i)=\rho(v_i\mid\phi_{k1})\cdot\rho(u_i\mid\phi_{k2})\cdot\xi \quad （公式 1）$$

综上所述，本文要解决的问题是：给定数据冲突集合 $CD$，为每个对象 $o\in O$，从各数据源提供的属性值集合 $V(k)$ 中找出真值。

## 3.3 算法描述

该算法的基本思想是每一个数据源都有一个信任度，直观上来说，在给出的信息中我们更相信那些信任度比较高的数据所提供的信息，所以数据源的信任度对数据准确性的影响是存在的，而数据源的信任又是根据它所提供的数据值的可信度决定的，所以数据源的信任度与数据值的可信度是相互影响的。

根据以上分析，计算属性值 $v_i$ 为真的概率时需要考虑提供 $v_i$ 的数据源集合 $S(k)$ 对它的支持，以及没有提供它的相关数据源集合 $S(\bar{k})$ 对它的惩罚，数据源中还存在对该属性没有描述的集合。根据贝叶斯公式推导，可以得出真值概率的计算公式为：

$$P(v_i)=\rho(v_i\mid\varphi_k)=\frac{\rho(\varphi_k\mid v_i)\rho(v_i)}{\rho(\varphi_k\mid v_i)\rho(v_i)+\rho(\varphi_k\mid\bar{v}_i)(1-\rho(v_i))} \quad （公式 2）$$

其中，$\rho(\varphi_k\mid v_i)$ 表示属性值为 $v_i$ 的条件下，$S(k)$ 与 $S(\bar{k})$ 提供属性集合 $\varphi_k$ 的概率。$\rho(\varphi_k\mid\bar{v}_i)$ 表示在 $S(\bar{k})$ 集合中，属性值不为 $v_i$ 的条件下，$S(k)$ 与 $S(\bar{k})$ 提供属性集合 $\varphi_k$ 的概率。$\rho(v_i)$ 表示在所有数据源中（包含不存在描述的数据源），属性值为 $v_i$ 的概率。

算法描述：

输入：（1）$n$ 个数据源 $S$；（2）数据源提供的 $m$ 个对象，每个对象包含 $t$ 个属性；训练集与测试集；

输出：针对每个对象中属性的真值集合以及数据源的质量指标；

1. 初始化每个数据源的数据质量，认为出版社数据源质量为 1，其余数据源质量为 $0<q<1$；

2. 计算数据源为每个对象的每个属性提供的属性值集合；

3. for each $v_i\in V(k)$

4. 根据公式（1）与公式（2）计算出每个 $v_i$ 的真值概率 $\rho(v_i\mid\varphi_k)$；

5. 设每个属性 $k$ 的真值为真值概率 $\rho(v_i\mid\varphi_k)$ 最大的属性值；

6. 根据计算每个属性的真值，计算每个数据源的准确率作为数据源的可信度。。

# 4 实验结果

本文选取的数据集以期刊为例，数据来源主要有 EMERALD、SCRIP、TAYLOR、WILEY、中图公司、OA 资源、自加工数据。选取的元数据字段包括：正题名、作者名称、作者单位、语种、总页数、起始页码、终止页码。数据集为 10386 条数据。

本文将出版社的数据源质量设为 1，其它来源的数据质量 $q$ 分别设为 0.6、0.7、0.8、

doi:10.3772/j.issn.2095-915x.2017.01.007

0.9，分别计算机器自动识别的数据冲突精准率。通过人工对比，得到本文的实验结果如表1所示。

表1 算法参数调整实验结果

| $q$ | 精准率 | 时间开销 |
|---|---|---|
| 0.6 | 74.5% | 3.5s |
| 0.7 | 78.6% | 3.7s |
| 0.8 | 84.3% | 3.6s |
| 0.9 | 79.2% | 4.1s |

根据实验结果可以看，当数据来源的数据质量 $q$ 值逐渐增大时，识别的数据冲突精准率逐渐增高；当 $q$ 值设置为0.8时，识别的数据冲突精准率最高；然而，当 $q$ 值设置为0.9时，其精准率骤降。平时时间开销大约为3.7s，大大降低了人工的工作时间，但与此同时，也在一定程度内降低了精准率。

本文选取将 $q$ 为0.8时，得到的各个数据源的可信度代入算法，将本文提出的算法，记为NewAccu，与几个单真值冲突数据的真值发现算法进行比较，它们分别是Vote、2-Estimates、3-Estimates、ACCUVOTE。通过实验，得到的实验结果如表2所示。

表2 真实数据集上的实验结果

| 算法 | 精准率 | 时间开销 |
|---|---|---|
| Vote | 81.7% | 1.2s |
| 2-Estimates | 82.6% | 3.4s |
| 3-Estimates | 84.3% | 5.6s |
| ACCUVOTE | 74.8% | 44.8s |
| NewAccu | 88. 7% | 3.8s |

表2所示的实验结果中，本文提出的算法具有较高的精确率。ACCUVOTE算法结果的精确率较差，考虑到了数据源之间的复制情况，若有一个数据源提供的数据错误，并且被复制，则会严重的影响到数据的结果。在时间开销上，Vote算法的时间开销最小，本文提出的算法时间开销较小，ACCUVOTE算法的计算需要计算两两数据源的依赖程度，因此时间开销比其它算法要大。

## 5 结束语

本文讨论了在构建多渠道元数据资源建设体系时，往往存在着大量的元数据冲突的问题，即对同一对象的属性存在多种描述，造成了元数据的组织与揭示的困难。本文处理的原则是遵从原文，优先选取原文的值作为唯一的真值，将数据冲突问题视为单真值冲突问题。考虑到数据提供商均需要加工数据，将数据源之间的关系视为相互独立。根据以上，本文提出了一种面向学术资源集成的真值发现算法。该算法基于贝叶斯算法，将数据源分为两类，即提供属性值与提供其它属性值的数据源，分别对其真值概率进行奖惩，采用数据源的召回率衡量数据源的质量，并对有关联的属性采用联合概率计算其真值。实验证明本文方法所构造的真值发现算法在保证准确率的同时，大大节省了人力的工作时间。

## 参考文献

[1] Yin X, Han J, Yu P S. Truth Discovery with Multiple Conflicting Information Providers on the Web[J]. IEEE Transactions on Knowledge & Data Engineering, 2008, 20(6): 796-808.

[2] Galland A, Abiteboul S, Marian A, et al.

Corroborating Information from Disagreeing Views[C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining. New York: ACM, 2010: 131-140.

[3] Pasternack J, Dan R. Knowing What to Believe(When You Already Know Something)[C]//Proc of the 23rd Int Conf on Computeeateional Linguisteics. Beijing University Press, 2010: 877-885.

[4] 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.

[5] 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.

[6] 张志强，刘丽霞，谢晓芹，等．基于数据源依赖关系的信息评价方法研究 [J]. 计算机学报，2012, 35(11): 2392-2402.

[7] Blanco L, Crescenzi V, Merialdo P, et al. Probabilistic Models to Reconcile Complex Data from Inaccurate Data Sources[M]// Advanced Information Systems Engineering. Springer Berlin Heidelberg, 2010: 83-97.

[8] Qi G J, Aggarwal C C, Han J, et al. Mining Collective Intelligence in Diverse Groups[C]// International Conference on World Wide Web. 2013: 1041-1052.

[9] Pochampally R, Das Sarma A, Dong X L, et al. Fusing Data with Correlations[J]. Computer Science, 2015: 433-444.

[10] Yin X, Tan W. Semi-supervised Truth Discovery. [C]// International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April. 2011: 217-226.