计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140684
52(9):1931-1940,2015

基于数据源分类可信性的真值发现方法研究

马如霞[1,2]　　孟小峰[1]

[1](中国人民大学信息学院　北京　100872)

[2](首都师范大学教育技术系　北京　100048)

(maruxia@126.com)

Truth Discovery Based Credibility of Data Categories on Data Sources

Ma Ruxia[1,2] and Meng Xiaofeng[1]

[1](*Department of Information*, *Renmin University of China*, *Beijing* 100872)

[2](*Department of Education Technology*, *Capital Normal University*, *Beijing* 100048)

Abstract　The popularization of the network and the development of e-commerce have changed the way people access information and consume. For most of people, Web has been the important source of information. Meanwhile, information quality issue is becoming increasingly prominent. There is a lot of information which is outdated, incorrect, false and bias. Particularly, the problem of conflicting information provided by different websites is obvious. It has to be solved that how to find the truth from conflicting information. As we know, there is not a method which considers the credibility of data categories on data sources during discovering truth. So, we propose a problem which is truth discovery based credibility of data categories on data sources. In this paper, two methods are proposed to detect the credibility differences of data categories on sources, and a Bayesian method is used to iteratively compute the data sources quality and data accuracy. Additional, data coverage and the difficulty of each object is considered to improve the accuracy of truth finding. The experiments on a real data set show that our algorithms can significantly improve the accuracy of truth discovery.

Key words　truth discovery; data conflicting; credibility of data categories on data sources; quality of information; data fusion

摘　要　网络的普及和电子商务的发展改变了人们信息获取以及消费的方式. Web 已经成为大多数人获取信息的重要来源. 与此同时,互联网信息质量问题也逐渐凸显. Web 中存在大量过时、错误、虚假、片面的信息. 其中,不同网站为相同对象提供冲突信息的问题尤为突出. 如何从这些冲突信息中找到正确信息成为亟待解决的问题,这类问题又被称为真值发现问题. 通过对现有真值发现问题解决方法的调研,发现现有方法均未考虑数据源分类可信性差异对真值发现的影响. 因此,提出基于数据源分类可信性的真值发现问题. 提出 2 种方法探测数据源分类可信性差异,并采用贝叶斯的方法迭代计算数据源分类可信性和属性值准确性. 另外,通过考虑数据源覆盖率和对象难度对真值发现的影响,进一步提高真值发现算法的准确性. 一个真实数据集的实验结果表明,所提方法可以显著提高真值发现的准确性.

收稿日期:2014-07-25;修回日期:2014-11-05

基金项目:国家自然科学基金项目(61379050, 91224008);国家"八六三"高技术研究发展计划基金项目(2013AA013204);高等学校博士学科点专项科研基金项目(20130004130001);中国人民大学科学研究基金项目(11XNL010)

通信作者:孟小峰(xfmeng@ruc.edu.cn)

关键词 真值发现;数据冲突;数据源分类可信性;信息质量;数据融合

中图法分类号 TP311

信息时代和互联网的发展给人们生活带来了巨大改变. Web 已经迅速发展为一个浩瀚的信息海洋,其数据量仍以惊人的速度剧增. 网络的普及和电子商务的发展改变了人们获取信息以及消费的方式. Web 已经成为大多数人信息获取的重要来源. 例如,人们从网上购物时,常常到各个购物网站查看商品介绍、价格以及评论. 与此同时,互联网信息质量问题也日益凸显. 博客、微博、论坛、合作知识库等社会媒体的出现在很大程度上降低了信息的准入门槛. 另外,由于信息时效性、传播性、信息发布者的主观故意性和导向性等因素,使得大量过时、错误、虚假、片面信息充斥于网络. 其中,不同网站为相同对象提供冲突信息的问题尤为突出. 如表 1 所示,《数据库系统概念》(第 6 版)一书不同图书网站提供的信息相互冲突. 其中,China-pub、京东和华章提供了完整的作者列表,而亚马逊和当当提供了错误的作者信息,各网站提供的页数信息也存在差异. 因此,如何从这些冲突信息中找到正确信息成为亟待解决的问题.

Table 1 Conflicting Information of "Database System Concepts" (6th Edition)

表 1 各网站提供《数据库系统概念》(第 6 版)信息

Website	Author	Page No.
China-pub. com	Abraham Silberschatz, Henry F Korth, S Sudarshan	804
amazon. cn	Silberschatz A	805
JD. com	Abraham Silberschatz, Henry F Korth, S Sudarshan	805
dangdang. com	西尔伯沙茨等	805
hzbook. com	Abraham Silberschatz, Henry F Korth, S Sudarshan	832

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可信性.

解决这一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,而这种差异在实际中普遍存在. 因此,投票算法往往产生错误的结果.

针对上述问题,许多方法采用一种类似 Authority-Hub[2]的方法迭代地计算数据源的可信性和数据值的准确性. 由越多高可信性数据源提供的数据值其正确的可能性越大,提供越多正确数据的数据源其可信性也越高,以此来建模数据源的质量. 为了进一步提高真值发现的准确性,已有的研究工作除了考虑数据源质量外还考虑了如下一些影响因素:数据的难度、数据值之间的相似性、数据源之间的拷贝关系等. 但这些工作都假设同一数据源呈现给其所有对象相同的准确性,然而实际情况并非如此. 我们发现,来自相同数据源的不同类别对象其数据值准确性不尽相同. 例如,一个图书网站对计算机类别的图书有较高的可信性,而对经济类图书的可信性较低. 本文针对这一问题,通过对数据源信息进行分类的方法自动检测这种差异,为不同类别数据区分数据源的质量,从而提高真值发现的准确性.

总的来说,本文主要有以下 3 个贡献:

1) 提出了数据源分类可信性的概念,并且提出 2 种方法来探测数据源分类可信性差异.

2) 在真值发现过程中,设计了一个基于贝叶斯的方法考虑数据源分类可信性. 在计算数据源分类可信性的过程中,同时还考虑了数据源的数据覆盖率和对象的难度因素.

3) 通过在真实数据集上的实验说明我们的算法明显地提高了真值发现的准确性.

1 相关工作

解决信息冲突的一般问题通常称为数据融合,其在数据集成领域已经有了大量研究工作. 文献[3]是一篇数据融合的综述,其定义了数据融合的目标,分析了面临的挑战,并对现有冲突处理策略进行分类,详细介绍了 9 个基于关系代数和 SQL 操作的融合技术. 但这些数据融合技术并未考虑数据源质量对冲突处理结果的影响.

文献[1]首次正式定义真值发现问题,并提出采用类似 Authority-Hub 方法的迭代机制(TruthFinder)来联合推导真值和数据源质量. 许多链接分析方法[2,4]均采用这种迭代机制分析数据源的权威性.

但与 Authority-Hub 不同的是,TruthFinder 对数据源可信性的判断并不依赖于其提供信息的数量,而是取决于信息的准确性.更为重要的是,TruthFinder 还考虑了不同信息之间相互关系对信息准确性判断的影响,从而进一步提高冲突处理结果的准确性.

随后,真值发现问题先后出现了一系列研究工作,通过考虑影响真值发现判断的各种因素采用基于信息检索[5]、Web 链接分析[6]、贝叶斯[1,7-11]和半监督学习[12]等方法来提高真值发现的准确率和计算效率.

文献[5]介绍了一个概率数据模型,其不仅考虑数据源提供值的不确定性,而且考虑了数据源的覆盖率.在此模型基础上提出 3 个算法:COSINE,2-ESTIMATES,3-ESTIMATES,用以估计真值和数据源可信性.其中,COSINE 基于信息检索中常用的 cosine 相似性度量方法;2-ESTIMATES 分别估计事实为真的概率和数据源出错的概率,其适用于一些统计的场景;3-ESTIMATES 改进了 2-ESTIMATES,通过估计每个事实的难易程度,避免数据源从相对容易的事实那里获得过高的可信性分值.

文献[6]在 Authority-Hub 的基础上进行了改进,提出 3 种新的方法来发现真值.作者通过取平均值、对数等方式减少信息数量对信息准确性判断的影响;同时还提出了一个框架将先验知识整合到任何真值发现算法,并用整数规划的方法对数据添加约束条件.该算法可以应对更大规模的真值发现问题,既减少了错误,也可以实现个性化的真值判断.

文献[1,5-6]根据"越多高质量数据源提供的事实越可能正确"这一假设,为每个事实分配准确性分值,该假设在数据源相互独立的情况下成立.然而在 Web 中数据源之间的拷贝现象非常普遍,因此这一假设就不能成立了.Dong 等人[7]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"这一启发式判断数据源之间的拷贝关系.文献[8]考虑信息的时效性问题,即信息的真值随时间可能发生变化,对数据源的覆盖面、准确性和新鲜度进行建模,从而提高真值发现的准确率.作者首先采用隐马尔可夫模型来判断数据源之间的拷贝关系以及拷贝的时间点;然后,采用贝叶斯模型聚集数据源的信息来判断真值.文献[9]对文献[7]中的方法进行了改进,加入了数据值之间的相似度问题处理.文献[7-9]仅从单一属性判断数据源之间的依赖关系,文献[10]提出考虑综合对象的多个属性信息改进数据源拷贝关系的判断方法.

为了简化问题,文献[1,5-10]均假设每个对象有唯一真值.然而实际情况并非如此,很多对象可能有多个真值.例如,每个人可能有多个电话号码.文献[11]提出了一个概率图模型用于在没有任何监督的情况下自动地推导真值记录和数据源质量.与其他工作不同的是,该方法通过 2 类错误——错误的肯定(false positive)和错误的否定(false negative)——来建模数据源的准确性和完整性.这样一来,可以解决多真值的问题.另外,作者还采用一个基于采样的推导算法使得算法时间复杂度达到线性量级.

文献[12]针对无监督学习方法存在的问题,通过使用数据记录之间的相似性提出一个半监督学习方法,通过标定好的真实数据来帮助寻找真值.与已有只提供迭代算法的研究不同,作者推导出该问题的最优解决方案,并且提供了一个迭代算法收敛到该最优解.

现有研究工作在建模数据源质量时均基于一个假设:数据源对其提供的对象具有相同的质量.但我们发现,来自相同数据源的不同类别对象其数据值准确性不尽相同.因此,本文提出基于数据源分类可信性的真值发现问题.

2 问题定义

本文研究基于数据源分类可信性的真值发现问题,其目标是自动发现数据源在不同类型对象上的质量差异,再根据这种差异来判断一个对象在某个属性上的真值.下面我们介绍该问题的形式化定义.

令冲突数据集 $DB=\{r_1,r_2,\cdots,r_n\}$ 为输入信息,其中 r_i 是形如 (o,v,s,c) 的四元组,o 是一个对象,v 是对象的属性值,s 是数据源,c 是该对象的所属类别.假设输入的记录已经作了去重处理,每条记录都是唯一的.从 DB 我们可以得到如下信息:$O=\{o_1,\cdots,o_n\}$ 是对象的集合;$S=\{s_1,\cdots,s_n\}$ 是数据源集合.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,因此不同数据源可能为对象 o 提供了冲突的属性值.这里假设:在不同数据源提供给对象 $o\in O$ 的所有属性值 $V(o)=\{v_1,\cdots,v_L\}$ 中,只有一个属性值为真.本文要解决的问题是:给定 DB,为每个对象 $o\in O$,从各数据源提供的属性值集合 $V(o)$ 中找出真值,并求出各数据源在不同类别对象上的可信性.

下面介绍3个基本概念：

定义1. 对象属性值的准确性. 一个对象属性值 v 的准确性 $A(v)$ 是指 v 正确的概率.

定义2. 数据源分类可信性. 数据源分类可信性指数据源在各类对象上的可信性. 数据源 s 在 c 类对象上的可信性 $T(s,c)$ 是 s 为 c 类对象提供的属性值为真的概率.

定义3. 对象难度. 对象 o 的难度是指为对象 o 提供正确属性值的难易程度.

基于观察我们发现，不同数据源为同一对象提供的真值相同或者相似，而提供的错误值则有很大不同. 另外，在某类对象上，提供越多真值的数据源为其他对象提供真值的可能性也越高. 因此，对象真值的判断和数据源分类可信性的计算互相关联. 数据源在各类信息上的可信性可以帮助判断其所提供信息的准确性，信息准确性又用来计算数据源在各类信息上的可信性. 通过多次迭代计算的方法，最后达到收敛.

为了简化计算，我们作了如下2个假设：

假设1. 相同领域中，不同数据源的对象分类大致相同.

假设2. 数据源之间相互独立地提供对象的属性值.

3 基于数据源分类可信性的真值发现

本文提出基于数据源分类可信性的真值发现方法(CTruthFinder)，其流程如图1所示：

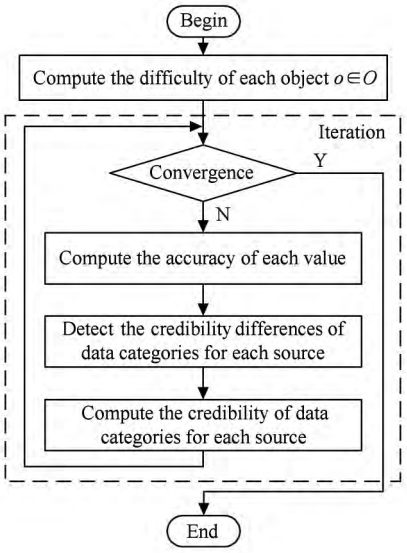

Fig. 1 The flow chart of CTruthFinder.

图1 基于数据源分类可信性的真值发现方法流程图

首先，进行初始化工作，计算每个对象的难度. 如果各数据源为某个对象提供的属性值一致，则该对象的难度较低. 与难度较高的对象相比，其对数据源可信性的贡献应该较低. 然后，进行数据源分类可信性和对象属性值准确性的迭代推导.

下面介绍数据源分类可信性和属性值准确性的基本推导.

3.1 基本推导

3.1.1 数据源分类可信性计算

根据定义，数据源分类可信性是数据源为某类对象提供真值的概率. 最直观的计算方法是：数据源 s 在类别 c 上的可信性 $T(s,c)$ 为 s 提供的所有 c 类对象属性值准确性 $A(v)$ 的均值，如式(1)所示：

$$T(s,c)=\frac{\sum_{o\in Obj(s,c)\wedge v=Value(s,o)}A(v)}{N(s,c)},\tag{1}$$

其中，$Obj(s,c)$ 表示数据源 s 提供的所有 c 类对象的集合；$Value(s,o)$ 表示数据源 s 为对象 o 提供的属性值；$N(s,c)$ 表示数据源 s 提供的 c 类对象的数目，即 $|\{o|o\in Obj(s,c)\}|$.

在实际中，对象难易程度也影响了数据源分类可信性的判断. 数据源不应该从相对容易的对象那里获得太多的可信性. 因此，我们在 $T(s,c)$ 中加入了对象难易度的考虑，如下所示：

$$T'(s,c)=\frac{\sum_{o\in Obj(s,c)\wedge v=Value(s,o)}Difi(o)\times A(v)}{N(s,c)},\tag{2}$$

其中，$Difi(o)$ 表示对象 o 的难度，$0\leqslant Difi(o)\leqslant 1$，具体的计算方法将在3.2节中详细讨论.

另外，数据源的覆盖面也一定程度上对数据源可信性的判断产生影响. 例如，数据源 s_1 提供了100个属性值准确性为0.9的数据，而另一个数据源 s_2 提供了10 000个属性值准确性也为0.9的数据，直观来看数据源 s_2 的可信度应该高于数据源 s_1. 因此，我们对式(2)进一步改进，得到式(3)：

$$T''(s,c)=\frac{N(s,c)}{N(-,c)}\times\frac{\sum_{o\in Obj(s,c)\wedge v=Value(s,o)}Difi(o)\times A(v)}{N(s,c)},\tag{3}$$

其中 $N(-,c)$ 表示冲突数据集中类别为 c 的所有对象的数目.

3.1.2 对象属性值的准确性计算

根据第2节的假设，数据源集合 S 为任意对象 o 提供的属性值集合中，有且仅有一个属性值为真，

并且各数据源独立地提供属性值.我们根据提供属性值的数据源数目以及数据源在该类对象上的可信性,采用贝叶斯的方法计算属性值为真的概率.令 $o\in O$,$\Omega(o)$为冲突数据集 DB 中各数据源为对象 o 提供的属性值情况.这里要求 v 为真的概率 $P(v)$实际上是:在 $\Omega(o)$条件下,v 为真的条件概率 $p_r(v|\Omega(o))$.我们首先可以求出,在 v 为真的条件下$\Omega(o)$的概率为:为对象 o 提供属性值 v 的所有数据源 $Sour(o,v)$提供正确值、并且为对象 o 提供除 v 以外属性值的所有数据源 $Sour(o,\neg v)$提供错误值的概率,如式(4)所示:

$$p_r(\Omega(o)\mid v)=\prod_{s\in Sour(o,v)\wedge c=Cg(s,o)}T''(s,c)\times\prod_{s\in Sour(o,\neg v)\wedge c=Cg(s,o)}\frac{1-T''(s,c)}{|V(o)|},\tag{4}$$

其中,$Cg(s,o)$表示数据源 s 中对象 o 所属的类别.

根据贝叶斯公式,我们还需要求出 $\Omega(o)$的概率 $p_r(\Omega(o))$,假设每个属性值 $v\in V(o)$为真的先验概率 $p_r(v)$相同,设为 γ,则:

$$p_r(\Omega(o))=\sum_{v'\in V(o)}(\gamma\times\prod_{s\in Sour(o,v')\wedge c=Cg(s,o)}T''(s,c)\times\prod_{s\in Sour(o,\neg v')\wedge c=Cg(s,o)}\frac{1-T''(s,c)}{|V(o)|}).\tag{5}$$

由 $p_r(\Omega(o)|v)$和 $p_r(\Omega(o))$可以求出 $P(v)$:

$$P(v)=p_r(v\mid\Omega(o))=\frac{p_r(\Omega(o)\mid v)\times p_r(v)}{p_r(\Omega(o))}=\left\{\prod_{s\in Sour(o,v)\wedge c=Cg(s,o)}T''(s,c)\times\prod_{s\in Sour(o,\neg v)\wedge c=Cg(s,o)}\frac{1-T''(s,c)}{|V(o)|}\right\}\Big/\left\{\sum_{v'\in V(o)}\left(\prod_{s\in Sour(o,v')\wedge c=Cg(s,o)}T''(s,c)\times\prod_{s\in Sour(o,\neg v')\wedge c=Cg(s,o)}\frac{1-T''(s,c)}{|V(o)|}\right)\right\}.\tag{6}$$

由于对所有的 $v\in V(o)$来说 $p_r(\Omega(o))$都相同,所以在计算 v 的准确性分值时为了简化可以将其忽略.另外,为了防止下溢我们采用了取对数的方法,得到 v 的准确性分值计算式(7):

$$A'(v)=\sum_{s\in Sour(o,v)\wedge c=Cg(s,o)}\ln(T''(s,c))+\sum_{s\in Sour(o,\neg v)\wedge c=Cg(s,o)}\ln\left(\frac{1-T''(s,c)}{|V(o)|}\right).\tag{7}$$

为了使得准确性分值在[0,1]区间内,进一步对 $A'(v)$进行标准化得到式(8):

$$A(v)=\frac{A'(v)}{\sum_{v'\in V(o)}A'(v')}.\tag{8}$$

3.2 对象难度计算

对象难易度是指为该对象提供真值的难易程度.直观地看,如果各数据源为给定对象提供的属性值基本一致,则很容易能判断出该对象的真值;但如果提供的属性值差异较大,则判断其真值的难度就比较大.因此,可以用属性值的一致性衡量对象的难易程度.

属性值一致性判断最简单的方法是直接用属性值的数目来度量.例如,如果数据源为对象 o_1 提供了 1 个相同的属性值,而为另一个对象 o_2 提供 5 个不同的属性值,则很容易看出 o_1 的真值比 o_2 容易判断.也就是说,不同属性值的数量越少则该对象的真值判断难度越小.另外,为了保证对象难度值在区间[0,1]之间,我们进行了标准化处理,得到对象难度计算式(9):

$$Difi(o)=\frac{|V(o)|}{\max(\{|V(o')|\mid o'\in O\})}.\tag{9}$$

但是上述方法仅考虑不同属性值的个数,并未考虑属性值的分布情况.例如,数据源为对象 o_1 和对象 o_2 均提供了 2 个不同的属性值,对象 o_1 的属性值分布为$\{(v_1,9),(v_2,1)\}$,对象 o_2 的分布为$\{(v_1',4),(v_2',6)\}$,则可以看出 o_1 的真值判断相对比较容易.可以用信息熵来刻画属性值分布的一致性.对象$o\in O$的信息熵计算公式为

$$Entropy(o)=-\sum_{v\in V(o)}\left(\frac{|Sour(o,v)|}{|Sour(o,-)|}\times\mathrm{lb}\frac{|Sour(o,v)|}{|Sour(o,-)|}\right),\tag{10}$$

其中,$|Sour(o,v)|$表示为对象 o 提供属性值 v 的数据源数目,$|Sour(o,-)|$表示为对象 o 提供属性值的数据源数目.信息熵越高,则判断真值的难度越大.与式(9)类似,式(10)也进行了标准化处理得到对象难度计算式(11):

$$Difi(o)=\frac{Entropy(o)}{\max(\{Entropy(o')\mid o'\in O\})}.\tag{11}$$

3.3 自动探测数据源分类可信性差异

在基本推导中,在已知分类结果的基础上,计算了数据源分类可信性.该分类结果区分了数据源分类可信性的差异,那么如何得到感知数据源分类可信性差异的分类结果?例如,数据源 s_1 在计算机类图书上的可信性一致,则不需要进一步区分其可信性;而在经济类图上的可信性差异很大,需要对它的子类的可信性进行进一步区分.

该问题可以定义为:给定数据源 $s\in S$,其各个对象 $o\in O$ 的属性信息以及数据源 s 提供的每个属性值 v 的准确性 $A(v)$,发现数据源 s 在各类对象上的可信性差异.针对该问题,我们提供了 2 种方法:1)基于固定分类的可信性差异探测;2)基于贪心聚类的分类可信性差异探测.

3.3.1 基于固定分类的数据源可信性差异探测

根据数据源各对象的分类情况,可以构造出其分类层次树 T.在已知数据源分类层次树的情况下,可以通过其各类别对象属性值准确性的离散程度来判断是否需要对该类别的子类进行进一步可信性差异探测.例如,如果发现当当网在小说类别图书的作者信息准确性差异很大,则需要对小说的子类(如恐怖小说、言情小说等)进行进一步差异探测.这里我们用标准差来度量数据源 s 在 c 类对象上属性值准确性的离散程度 $SDeviation(s,c)$.

如算法 1 所示,从分类层次树的根结点开始,自顶向下采用广度遍历的方法遍历各个类结点,如果当前结点的 $SDeviation(s,c)$ 大于阈值 δ,则需要对其子分类结点进行进一步探测;否则删除当前结点的所有子树.通过上述处理对原有的分类层次树进行了修整,根据修整后的分类层次树生成感知分类可信性差异的对象分类结果 $C(R)$.

算法 1. 基于固定分类的数据源可信性探测算法(fixTrustProbe).

输入:数据源 s 的分类层次树 T,以及数据源 s 提供的对象详细信息表 $R=\{(o,c,v,A(v))\mid$ 对象 o 属于 c 类,值为 v,属性值 v 的准确性分值为 $A(v)\}$;

输出:分类结果 $C(R)=\{c_1,\cdots,c_n\}$.

初始化栈 $stack$;
$root=$ 分类层次树 T 的根;
PUSH($stack$,$root$);
WHILE(not $IsEmpty(stack)$)
  $node=$POP($stack$);
  c 为 $node$ 结点对应的类别信息;
  IF $SDeviation(s,c)<\delta$ THEN
    删除结点 $node$ 的所有子树;
  ELSE
    FOR EACH $p\in ChildNode(node)$
      PUSH($stack$, p);
    ENDFOR
  ENDIF
ENDWHILE
FOR 修剪后的分类树 T 的每个叶子结点 $leaf$
  $path=$从 T 的根结点到 $leaf$ 分类信息组成的字符串;
  $c_i=$属于该 $path$ 类别的所有对象组成的集合;
ENDFOR
RETURN $C(R)$.

算法 1 受到固定分类的限制,只能针对现有分类探测数据源分类可信性差异.但是,在实际情况中可能会出现属性值准确性分布和已有分类不一致的情况.例如,属性值准确性分值在分类树的叶子分类结点上的离散程度仍旧很高,则需要对叶子结点进行进一步的划分,而基于固定分类的方法不能满足这一需求.因此,我们提出了另一种基于贪心聚类的方法来探测数据源在不同对象类别上的可信性差异.

3.3.2 基于贪心聚类的数据源可信性差异探测

根据数据源 s 提供的各对象类别信息以及属性值准确性分值信息对对象进行聚类.下面首先介绍目标函数,然后给出聚类算法.

1) 目标函数

一个理想的聚类应该满足高内聚性和低耦合的特点.目前已经有一些考虑高内聚低耦合的目标函数,本文从稳定性和计算代价来考虑选择 Davies-Bouldin 指数[13].

形式化地定义为,给定一个聚类 $C=\{c_1,c_2,\cdots,c_n\}$,它的 Davies-Bouldin 指数定义为

$$\Phi(C)=\mathop{Avg}_{i=1}^{n}\left(\max_{j\in[1,n]\wedge j\neq i}\frac{d(c_i,c_i)+d(c_j,c_j)}{d(c_i,c_j)}\right),\tag{12}$$

其中 $d(c_i,c_j)$ 表示 c_i 和 c_j 之间的距离.当 $i=j$ 时,衡量 c_i 的内聚性,$i\neq j$ 时衡量 2 个类之间的耦合性.当该目标函数达到最小值时表示达到高内聚低耦合的标准.

接下来计算 2 个类之间的距离.如果 2 个类的距离越大,则表示相同属性的值差别越大.这里考虑对象的 2 个属性:分类信息和属性值准确性分值 $A(v)$.取这 2 个属性距离的线性组合作为 2 个类的距离,即:

$$d(c_i,c_j)=\frac{d_c(c_i,c_j)+d_v(c_i,c_j)}{2},\tag{13}$$

其中,$d_c(c_i,c_j)$ 表示在分类信息上 2 个类之间的距离,$d_v(c_i,c_j)$ 表示在属性值准确性分值上 2 个类之间的距离.

下面分别计算分类信息和属性值准确性分值的

相似性.对象的分类信息是文本类型,如"图书→教材→研究生→本科→专科教材→工学",计算其相似性时考虑文本的顺序,因此采用编辑距离的方法来计算2个分类信息之间的相似性.令在类 c_i 中分类信息的集合为 R_i,类 c_j 中分类信息的集合为 R_j,则分类信息在2个类 c_i 和 c_j 上的距离计算公式为

$$d_c(c_i,c_j)=1-\underset{r\in R_i\wedge r'\in R_j}{Avg}\ sim(r,r'). \quad (14)$$

属性值准确性分值是数值类型,令类 $c_i(c_j)$ 中属性值准确性分值的集合为 $Q_i(Q_j)$,则其在2个类上的距离计算公式为

$$d_v(c_i,c_j)=\underset{q\in Q_i\wedge q'\in Q_j}{Avg}\frac{|q-q'|}{\max(q,q')}. \quad (15)$$

2)贪心聚类算法

要找到最优的分类结果是几乎不能实现的,这里采用贪心算法找到一个近似最优解.该聚类算法首先初始化一个聚类,然后迭代地选择2个最合适的类 c_i 和 c_j 进行合并.

具体方法如算法2所示.首先,完成初始化工作,将分类信息和 $A(v)$ 值相同的对象放到同一个类中,生成初始分类 $C^0(R)$.然后,对上一次生成的聚类结果 $C^{t-1}(R)$ 中任意2个类 c_i 和 c_j,计算将它们合并且不改变其他类的情况下的Davies-Bouldin指数.将Davies-Bouldin指数最小的2个类进行合并,直到本次聚类结果和上次聚类结果相同,聚类算法结束;否则继续执行类的合并操作.

算法2. 基于贪心聚类的数据源可信性探测(ClusterTrustProbe).

输入:数据源 s 提供的对象详细信息表 $R=\{(o,c,v,A(v))|$对象 o 属于 c 类,值为 v,属性值 v 的准确性分值为 $A(v)\}$;

输出:R 的聚类结果 $C^t(R)$.

初始化 $C^0(R)$;

$t=0$;

REPEAT

  $t=t+1$;

  FOR EACH $c_i,c_j\in C^{t-1}(R)$ 且 $c_i\neq c_j$

    将 c_i 和 c_j 合并其他类不变生成 C_{temp};

    $Score(i,j)=BouldinIndex(C_{\text{temp}})$;

  ENDFOR

  选 $Score(i,j)$ 值最小的2个类 c_i,c_j 进行合并,生成 $C^t(R)$;

UNTIL $C^{t-1}(R)==C^t(R)$;

RETURN $C^t(R)$.

这里我们假设数据源提供了对象的分类信息,因此对分类信息和属性值准确性分值进行聚类.如果数据源未提供对象分类信息,也可以采用同样的聚类方法对其他属性(如图书的书名等)与属性值准确性分值进行聚类.

3.4 迭代计算

数据源分类可信性需要利用其提供对象的属性值准确性计算,而属性值的准确性反过来又需要通过提供它的数据源分类可信性计算.因此,数据源分类可信性和属性值的准确性计算是一个迭代的过程,如算法3所示.首先,根据每个数据源的层次分类树,将数据源分类可信性初始化为统一的估计值 t_0;在第 n 次迭代中,CTruthFinder可以从第 $n-1$ 次迭代得到的数据源分类可信性 $T''_{n-1}(s,c)$ 计算属性值准确性 $A_n(v)$;接下来,根据得到的属性值准确性信息对数据源分类可信性进行差异探测,生成分类 C_n;再进一步计算数据源分类可信性.如此迭代,直到2次迭代生成的属性值准确性向量 $\bar{\boldsymbol{A}}_n$ 与 $\bar{\boldsymbol{A}}_{n-1}$ 的余弦相似性大于 $1-\mu$ 才停止迭代.这里阈值 μ 表示2次迭代得到的属性值准确性向量之间的差异.

算法3. 面向数据源信息分类的真值发现算法(CTruthFinder).

输入:冲突数据集 DB;

输出:数据源分类可信性、对象属性值的准确性分值.

FOR EACH $o\in O$

  计算对象 o 的难易程度 $Diff(o)$;

ENDFOR

FOR EACH $s\in S$ /*初始化 $T''_0(s,c)$*/

  FOR 数据源 s 的每个分类 c

    $T''_0(s,c)=t_0$;

  ENDFOR

ENDFOR

$n=0$;

REPEAT

  $n=n+1$;

  根据式(6)~(8),利用 $T''_{n-1}(s,c)$ 计算 $A_n(v)$;

  对每个数据源进行数据源可信性差异探测生成对象分类 C_n;

  根据式(3),利用 $A_n(v)$ 计算 $T''_n(s,c)$;

UNTIL 属性值准确性向量 $\bar{\boldsymbol{A}}_n$ 与 $\bar{\boldsymbol{A}}_{n-1}$ 的余弦相似性大于 $1-\mu$;

RETURN 每个数据源分类可信性和每个对象准确性分值.

4 实 验

本节通过在一个真实数据集上的实验检验本文算法的效果. 我们实现了如下 6 个算法:

1) Voting. 采用投票机制计算真值,即选取由最多数据源提供的属性值作为真值.

2) TruthFinder. 文献[1]提出的方法,在计算属性值准确性时考虑了数据源的可信性.

3) CTF_Fix. 本文提出的 CTruthFinder 方法,其中数据源分类可信性差异探测采用固定分类的探测方法,使用信息熵的方法计算对象难度.

4) CTF_Cluster. 本文提出的 CTruthFinder 算法,其中数据源分类可信性差异探测采用贪心聚类的探测方法,使用信息熵的方法计算对象难度.

5) CTF_Fix_ND. 为了说明数据源覆盖率和对象难度的作用,我们实现了不考虑数据源覆盖率和对象难度的 CTF_Fix 算法.

6) CTF_Cluster_ND: 为了说明数据源覆盖率和对象难度的作用,实现了不考虑数据源覆盖率和对象难度的 CTF_Cluster 算法.

4.1 实验设置

4.1.1 数据集

本文采用的数据集是一个真实的图书数据集,该数据集也被用于文献[1,7]中. 该数据集是从 Abebooks.com 的计算机科学类图书爬取. 该网站为每一本书提供了各图书数据源的报价、图书描述等信息. 为了分析数据源的分类可信性,我们对该数据集进行了扩充,添加了工程和艺术类图书,并到 amazon.com 网站搜索每本图书,在查询结果页面获得该图书的分类信息,并且将这些分类信息抓取下来. 然后,对新的数据集进行去重处理,并且对作者的姓名进行了预清洗生成统一的形式,从而避免因为姓名格式不一致而产生影响. 预处理后的数据集包含 3 213 本图书、950 个图书网站以及 74 364 条记录.

另外,我们在原有的 100 本图书标准集的基础上,对另外 2 个分类分别人工确认了 100 本图书的作者信息. 将每个方法计算结果与标准集作比较,分析算法的准确率. 计算准确率时除了考虑作者是否相同,还要考虑作者出现的顺序是否一致.

4.1.2 实验环境

本文所有实验运行在一个双核、中央处理器为 Intel Core™ i5-2430M 2.40 GHz、内存为 4 GB、操作系统为 Windows8 的环境下. 所有算法均使用 Python 语言编写.

4.2 实验结果

4.2.1 真值发现的准确性评估

我们比较了 Voting, TruthFinder, CTF_Fix_ND, CTF_Fix, CTF_Cluster_ND, CTF_Cluster 六个算法的真值发现准确性. 其中, CTF_Fix_ND 和 CTF_Fix 算法的参数 $\delta=0.15$, 阈值 $\mu=10^{-4}$.

为了测试对象数量对真值发现算法准确率的影响,分别抽取 20%, 40%, 80%, 100%的图书进行实验. 为了判断算法的准确性,每个子数据集都包含标准集的 300 本图书,其余图书则随机抽取. 实验结果如图 2 所示,随着数据集的增大,各算法的准确率平稳增高. 其中, Voting 算法的准确率最低, CTF_Cluster 算法收获的准确率最高. 按照本文思想实现的 4 个算法中,考虑数据源覆盖率和对象难度的算法准确性比不考虑这些因素的算法准确率高. 考虑数据源分类可信性的算法普遍比 Voting 和 TruthFinder 算法的准确性高.

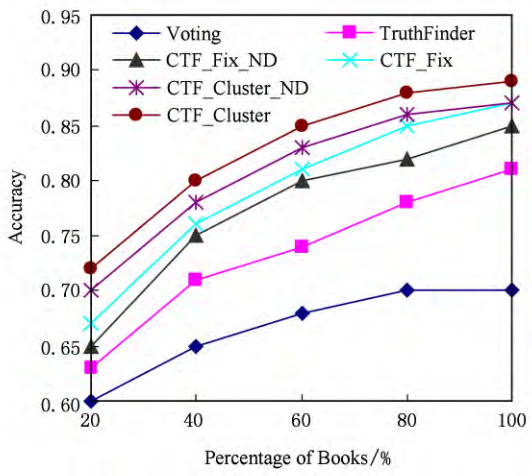

Fig. 2 Accuracy of each algorithm with different number of objects.

图 2 对象数量对真值发现算法准确率的影响

为了测试数据源数目对真值发现准确率的影响,我们分别随机抽取 20%, 40%, 80%, 100%的数据源进行实验. 实验结果如图 3 所示,各算法的准确性随着数据源数量的增加呈递增的趋势,但随着数据源的增加递增趋势逐渐趋缓.

4.2.2 真值发现的效率评估

本节观察算法的收敛情况,讨论各算法的迭代次数以及运行时间.

表 2 列出了各算法在整个图书数据集上的准确性、迭代次数和运行时间. CTF_Cluster 算法获得了

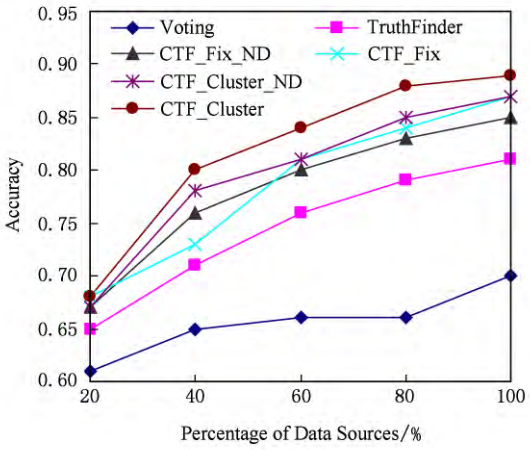

Fig. 3 Accuracy of each algorithm with different number of data sources.

图 3 数据源数量对真值发现准确性的影响

最好的准确性;与 CTF_Fix_ND 和 CTF_Cluster_ND 算法相比,CTF_Fix 和 CTF_Cluster 算法考虑了数据源覆盖率和数据对象的难度获得了更高的准确率.采用基于贪心聚类的数据源分类差异探测方法比基于固定分类的数据源分类差异探测方法获得更高的准确率.我们算法由于需要考虑数据源分类可信因素,CTF_Fix 和 CTF_Cluster 算法还考虑了数据源覆盖率和对象难度因素,因此算法的执行时间较之 Voting 和 TruthFinder 算法较高,时长约 60~180 s,但由于真值发现算法通常只需运行一遍,因此这样的时长是可以接受的.

Table 2 Accuracy, Number of Iterations and Running Time of All Algorithms

表 2 各算法的准确性、迭代次数以及总运行时间

Algorithm	Accuracy	Number of Iterations	Running Time/s
Voting	0.70	1	0.5
TruthFinder	0.82	10	30.1
CTF_Fix_ND	0.85	15	78.6
CTF_Fix	0.87	20	116.8
CTF_Cluster_ND	0.87	19	143.9
CTF_Cluster	0.89	22	183.2

4.2.3 参数敏感性

在基于固定分类的数据源分类可信性差异探测算法中,参数阈值 δ 用于判断从某个分类结点上是否需要进一步探测其子分类结点.我们对参数 δ 分别取值为:0.05,0.10,0.15,0.20,0.25,…,观察算法 CTF_Fix_ND 和 CTF_Fix 准确率的变化情况.

当分类中所有属性值准确性的标准差大于 δ 时,继续进行探测.直观来看,δ 值越大则算法对属性值离散程度的敏感性越低,则算法的准确率越低.从图 4 我们发现,2 个算法的准确率随着参数 δ 值的降低而增加.当参数 δ 降低到一定程度($\delta \leqslant 0.2$),算法的准确率达到基本稳定.因此,为了减少计算量,δ 的取值并不是越小越好,本文前面的实验中 $\delta=0.15$.当 $\delta \geqslant 0.5$ 时,算法的准确率也逐渐趋于稳定达到最低值,此时 CTF_Fix_ND 算法的准确性和 TruthFinder 算法接近,而 CTF_Fix 算法由于考虑了数据源覆盖率和对象难度因此准确率高于 TruthFinder 算法.

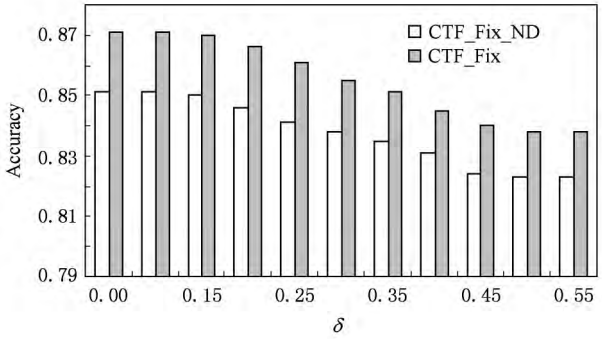

Fig. 4 Accuracy of each algorithm with different δ.

图 4 随参数 δ 的变化算法准确率的变化图

在数据源分类可信性和属性值准确性的迭代计算中,阈值用于控制迭代过程的结束.当 2 次迭代得到的属性值准确性向量的余弦相似性大于 $1-\mu$,则停止迭代.阈值 μ 的选择将影响算法的准确性.如图 5 所示,对阈值 μ 分别取值:10^{-1},10^{-2},10^{-3},10^{-4},10^{-5},10^{-6},10^{-7},观察算法准确性随 μ 的变化情况.随着阈值的减小,算法的准确性也逐渐增高;

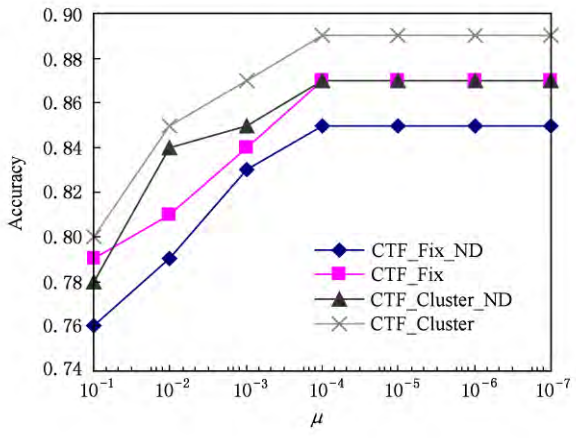

Fig. 5 Accuracy of each algorithm with different μ.

图 5 随阈值 μ 的变化算法准确率的变化图

但当阈值 $\mu<10^{-4}$ 时,算法准确性也趋于稳定达到最高值.

本节实验可以看出,我们提出的基于数据源分类可信性的真值发现算法准确性显著高于 Voting 和 TruthFinder 算法.

5 结 论

本文研究了如何通过检测数据源分类可信性差异提高真值发现的准确性. 提出 2 种自动探测数据源分类可信性差异的方法:基于固定分类的数据源分类可信性差异探测和基于贪心聚类的数据源分类可信性差异探测方法,并在真值发现的过程中,通过数据源分类可信性计算属性值的准确性. 为了进一步提高算法的准确性,在计算数据源分类可信性时还考虑了数据覆盖率和对象难易因素. 在真实数据集上的实验表明,我们的算法明显提高了真值发现的准确率. 接下来,我们将对该工作进一步扩展,用于处理多真值属性的真值发现问题.

参 考 文 献

[1] Yin X, Han J, Yu P S. Truth discovery with multiple conflicting information providers on the Web [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(6): 796-808

[2] Kleinberg J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604-632

[3] Bleiholder J, Naumann F. Data fusion [J]. ACM Computing Surveys, 2008, 41(1): 1-41

[4] Lawrence P, Sergey B, Rajeev M, et al. The PageRank citation ranking: Bringing order to the Web, SIDL-WP-1999-0120 [R]. Palo Alto, CA: Stanford University, 1999

[5] 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

[6] Pasternack J, Roth D. Knowing what to believe (when you already know something) [C] //Proc of the 23rd Int Conf on Computational Linguistics. Beijing: Tsinghua University Press, 2010: 877-885

[7] 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

[8] 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

[9] Zhang Zhiqiang, Liu Lixia, Xie Xiaoqin, et al. Information evaluation based on sources dependence [J]. Chinese Journal of Computers, 2012, 35(11): 2392-2402 (in Chinese)
(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)

[10] Blanco L, Crescenzi V, Merialdo P, et al. Probabilistic models to reconcile complex data from inaccurate data sources [C] //Proc of the 22nd Int Conf on Advanced Information Systems Engineering. Berlin: Springer, 2010: 83-97

[11] Zhao B, Rubinstein B, 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

[12] Yin X, Tan W. Semi-supervised truth discovery [C] //Proc of the 20th Int Conf on World Wide Web. New York: ACM, 2011: 217-226

[13] Davies D, Bouldin D. A cluster separation measure [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227

Ma Ruxia, born in 1977. PhD candidate at Renmin University of China. Student member of China Computer Federation. Lecturer in Capital Normal University. Her main research interests include Web data management, the credibility of Web information etc.

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Executive member of China Computer Federation. His main research interests include cloud data management, Web data management, flash-based databases, privacy protection etc.