基于 Markov 逻辑网的两阶段数据冲突解决方法

张永新 李庆忠 彭朝晖
(山东大学计算机科学与技术学院 济南 250014)

摘 要 在数据集成中,如何准确地解决数据冲突是关系集成数据质量的关键问题.现有的方法主要针对单个属性进行冲突解决,由于没有区分不同属性的冲突程度,也没有考虑不同属性间冲突解决的相互影响,导致数据冲突解决的准确率不高.针对现有方法存在的不足,文中提出一种基于 Markov 逻辑网的两阶段数据冲突解决方法.该方法可以根据冲突程度对属性进行划分,并分两阶段进行处理:(1)在第 1 阶段,对于弱冲突属性,利用投票规则及事实之间相互印证等简单规则进行冲突解决;(2)在第 2 阶段,利用了第 1 阶段冲突解决的结果,在规则中加入数据源与事实之间的相互影响规则、数据源之间相互依赖规则及弱冲突属性对强冲突属性影响规则,对强冲突属性进行冲突解决.通过在大量真实数据上的实验结果证明,该方法能够有效地解决集成数据的冲突问题,具有较高的准确率.

关键词 数据冲突解决;Markov 逻辑网;数据集成;冲突程度;推理规则
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2012.00101

2-Stage Data Conflict Resolution Based on Markov Logic Networks

ZHANG Yong-Xin LI Qing-Zhong PENG Zhao-Hui
(*School of Computer Science and Technology, Shandong University, Jinan* 250014)

Abstract In data integration, how to resolve the data conflicts accurately is a key issue that is closely related to the quality of integrated data. Current methods only consider single attribute, neither conflict degree nor mutual influence of different attributes are considered in data conflict resolution. It causes their accuracy not to be high. For the shortcomings of existing methods, a 2-stage approach for resolving data conflict based on Markov Logic Networks is proposed. This approach can divide different attributes according to their conflict degree and carry on 2-stage data conflict resolution: (1) In the first stage, the attributes which conflict degree is low can be resolved by simple rules such as voting and mutual verification of facts; (2) In the second stage, with the aid of the results from the first stage, the attributes which conflict degree is high can be resolve via adding some more complex rules such as mutual influence between sources and facts, inter-dependency of sources and low conflict degree attributes to high conflict degree attributes influence. Experimental results using a large number of real-world data show that the proposed approach can resolve the integrated data conflict effectively, which is more accurate.

Keywords data conflict resolution; Markov logic networks; data integration; conflicting degree; inference rule

收稿日期:2010-09-17;最终修改稿收到日期:2011-07-28.本课题得到国家科技支撑计划(2009BAH44B02)、国家自然科学基金(90818001,61003051)、山东省科技攻关计划(2010GGX10108)资助.**张永新**,男,1978 年生,博士研究生,主要研究方向为 Web 数据集成、Web 数据融合. E-mail: waterzyx@gmail.com. **李庆忠**(通信作者),男,1965 年生,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据集成、软件即服务. **彭朝晖**,男,1978 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为数据库关键词检索、Web 数据管理.

1 引 言

数据集成的目的是通过集成同一领域中的多个数据源,为用户提供更加全面、高质量的数据.但由于不同数据源的数据质量不同,它们对同一实体有着不同的描述,这些数据有的能正确反映现实世界,有的则不能,这些不同的描述数据之间可能存在冲突.而这些冲突数据的存在会严重影响集成数据的质量.因此,如何有效解决数据冲突,即从众多冲突数据中选择与现实世界一致的数据(可以称为真值)已成为一个亟待解决的问题.这个问题正逐渐引起研究者的关注,成为数据集成领域中最新的研究热点之一[1-2].

现有数据冲突解决方法大都通过关系扩展的方式实现,并定义了若干冲突解决策略和冲突解决函数[3].通过扩展关系操作及聚合函数,由用户或领域专家根据实际的需求或领域知识为冲突值指定不同的冲突解决函数,从而得到一致、无冲突的结果[4].但是,这类方法主要是根据具体的冲突类型人为指定冲突解决函数,在适应性和准确性方面还存在以下问题:

(1)针对某一属性上冲突值指定的冲突解决函数,当有新数据源及新数据加入时可能对现有的指定产生影响,从而需要重新指定甚至定义冲突解决函数.而数据集成是个动态的过程,随着数据源信息的变化以及新数据源的加入,不断会有新的数据集成到系统中.显然这类方法的适应性不强,无法满足数据集成动态性的需要.

(2)在现有的冲突解决策略中,应用比较广泛的是 Trust your friends 和 Cry with your wolves[3],其原理是通过信任某一数据源及少数服从多数的投票方式从多个冲突值中选择一个作为真值.但在数据集成中,如何从多个的数据源中选择"优质"数据源是一个挑战,并且选择单一的"优质"数据源作为信任数据源的做法也有些武断.另外,特别是在 Web 环境下,Web 数据易于传播的特点使得在传播真实值的同时也传播了虚假值,单纯通过投票的方式确定真值也不尽合理.因此,利用现有策略进行冲突解决很难保证准确性.

(3)现有冲突解决方法为每个属性分别指定相应的冲突解决函数,但没有区分不同属性的冲突程度,也没有考虑不同属性上冲突解决之间的相互影响,这也是现有冲突解决方法不够准确的一个原因.

为了适应数据集成的需要,针对于现有方法的不足,本文提出一种基于 Markov 逻辑网[5]的两阶段数据冲突解决方法.本文创新点主要体现在以下几个方面:

(1)提出一种基于 Markov 逻辑网的两阶段数据冲突解决方法,该方法可以根据冲突程度对属性进行划分,并分两阶段进行处理;由于充分利用了弱冲突属性对强冲突属性的影响因素,该方法有效地提高了冲突解决的准确度;

(2)通过对冲突数据及数据源特点的观察和分析,该方法综合运用了多角度的特征和规则,保证了数据冲突解决的有效性和准确率;

(3)通过在多个真实数据集上的实验表明,所提方法能够较好地完成数据冲突解决任务,具有较强的准确性和可适应性.

本文第 2 节分析数据冲突解决的相关工作;第 3 节将对数据冲突解决的相关概念及问题进行说明;第 4 节详细介绍本文提出的数据冲突解决方法;第 5 节给出实验结果与分析;第 6 节对全文做总结.

2 相关工作

现有数据冲突解决的研究主要集中在关系扩展上,其中比较具有代表性的是德国波茨坦大学 Naumann 等人的工作.Naumann 等人总结了现有的冲突解决策略和冲突解决函数[3],提出了相应的研究原型,并扩展、实现了诸如 minimum union[6] 等关系操作.对于此类方法的不足在引言中已经提及,这里不赘述了.

除了通过关系扩展的方式外,还有一些研究从多个冲突值中选择真值的方法.Wu 等人[7] 提出一种对搜索引擎返回结果集进行聚合的方式以选取真值的方法,该方法利用了 Web 数据源的重要度及各数据源提供结果值之间的相似性.但这种方法仅能对数值型结果值识别真伪,具有一定的局限性,其中 Web 数据源的重要度通过网站的排名及流行度来衡量,而事实上这种通过 Page Rank 的方式获得的排名与网站的可信度仍有一定的差距.Yin 等人[8] 将从描述同一对象的多个冲突值中识别真值的问题定义为 Veracity,利用数据源的可信度和数据值的准确性之间的相互影响以及正确数据值之间的相互印证关系,提出一个名为 TruthFinder 的迭代算法,以实现真值的识别.但这种方法并没有考虑数据源之间的依赖关系,由于 Web 数据易于传播的特点,

使得在传播正确值的同时也传播了错误值，仅考虑正确值之间相互印证的关系显然是不足的，这在一定程度上会降低真值识别的准确率. Dong 等人[9]主要考虑数据源之间的依赖关系，提出一种真值识别的贝叶斯推理模型. 但对于贝叶斯模型而言，当有新的推理规则加入时需要重新建模，因此该方法在适应性上稍显不足.

另外，现有真值识别方法主要是针对单个属性进行数据冲突解决，而对于不同属性的数据冲突采取同等对待的方式. 这些方法并没有区分不同属性的冲突程度，也没有考虑不同属性间冲突解决的相互影响因素，这在一定程度上也影响了冲突解决的准确率.

Markov 逻辑网[5]是一种将 Markov 网络与一阶谓词逻辑相结合的统计关系学习模型，它将领域知识引入 Markov 网，为大型 Markov 网提供了一种简洁的描述语言，为一阶逻辑增加了不确定性处理能力，可以作为很多统计关系学习任务的统一框架. Markov 逻辑网已经在诸如实体统一[10]、信息抽取[11]等方面的研究中得到应用.

综上所述，现有方法在准确性、适应性方面分别存在着一定的不足，难以胜任集成数据的冲突解决任务，迫切需要一种准确率更高、适应性更强的数据冲突解决方法.

3　数据冲突解决问题

在集成数据的冲突解决问题中，数据冲突主要表现在对同一实体的相同属性的不同数据源提供了不同的值，这些值有的能正确反映现实世界，有的不能. 为了便于描述，对文中相关概念及问题解释如下：

数据源(Source). 提供冲突数据的不同来源，这里的数据源可以是数据库、Web 网站等. 数据源集合表示为 $S=\{s_1,s_2,\cdots,s_n\}$，其中 $s_i(1\leq i\leq n)$ 表示第 i 个数据源.

实体(Entity). 实体是现实世界存在的，唯一的、可识别的对象，比如一本书、一部电影. 实体集合可表示为 $E=\{e_1,e_2,\cdots,e_m\}$，其中 $e_i(1\leq i\leq m)$ 表示第 i 个实体.

实体类型(Entity Type). 一个实体类型即为拥有相同属性的实体集合. 即实体类型是实体的一个类别，而实体是给定实体类型的一个实例. 比如书、电影等.

实体属性(Entity Attribute). 某实体的一个属性，比如某一本书的作者、某部电影的导演. 实体属性集合可表示为 $EA=\{ea_{11},ea_{12},\cdots,ea_{mk}\}$，其中 $ea_{ij}(1\leq i\leq m,1\leq j\leq k)$ 表示第 i 个实体的第 j 个属性.

事实(Fact). 对于一个实体属性，由某一数据源提供的值. 如对于实体属性 ea(《Flash CS3：The Missing Manual》这本书的作者)，数据源 s(图书在线网站 ABC Books)提供的事实为 f(“Chris Grover，E. A. Vander Veer”).

数据冲突(Data Conflict). 多个数据源对同一实体属性提供的事实不一致时产生数据冲突.

真值(True Value). 与现实世界一致的事实.

不同的数据源对于不同实体属性提供大量的事实，而对于同一实体属性不同数据源提供的不同事实之间可能存在着数据冲突，图 1 展示了数据源、实体、实体属性及事实之间的关系. 数据冲突解决就是从实体属性的不同冲突事实中识别真值的过程. 对于数据冲突解决，给出如下定义.

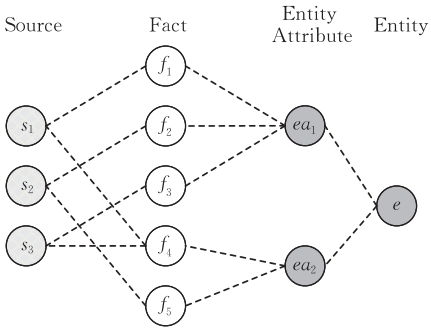


图 1　数据源、实体、实体属性及事实之间的关系

定义 1(数据冲突解决).　输入数据源集合 S、实体集合 E、实体属性集合 EA、事实集合 F 及其之间的关系，对于一个实体属性 $ea\in EA$，从各数据源对于实体属性 ea 所提供的事实集合 $F_{ea}=\{f_1,f_2,\cdots,f_L\}$ 中选择对应的真值 f_i，其中 $f_i\in F_{ea}$.

4　数据冲突解决方法

本文提出一种基于 Markov 逻辑网的两阶段数据冲突解决方法，其流程如图 2 所示. (1) 首先，对不同属性计算各属性上的数据冲突程度，根据冲突程度将属性分为两个集合：弱冲突属性和强冲突属性. (2) 在第 1 阶段冲突解决中，对于弱冲突属性，根据观察的特征，设定投票规则及事实之间相互印证规则，通过训练集训练 Markov 逻辑网模型并进行真值推理；由于该部分属性冲突程度较小，通过这

些简单的规则即可得到很高的数据冲突解决准确度.(3)在第2阶段冲突解决中,将第1阶段数据冲突解决的结果加入到训练集中,重新训练模型,对强冲突属性进行数据冲突解决;由于该部分属性冲突程度较大,为了有效利用第1阶段数据冲突解决的结果,提高对强冲突属性的冲突解决准确率,该阶段规则中加入数据源与事实之间的相互影响规则、数据源之间相互依赖规则及弱冲突属性对强冲突属性影响规则.(4)最后,根据两阶段冲突解决的结果,进行数据合并,得到无冲突的数据集.

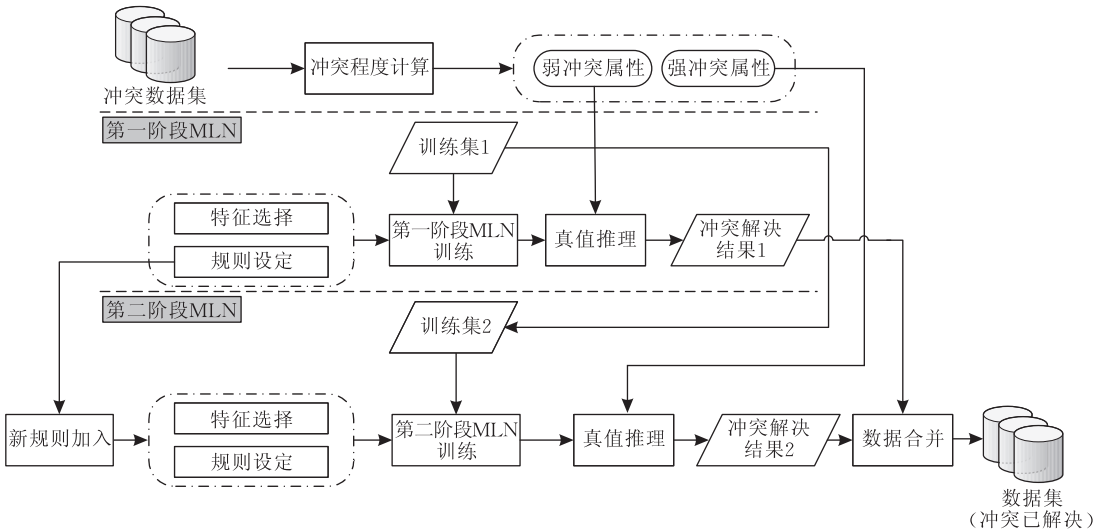


图2 基于Markov逻辑网的两阶段数据冲突解决整体流程

数据冲突解决流程的输入是来自不同数据源的带有冲突的数据集,其中重复检测已经完成;输出的是数据冲突已解决的数据集.数据冲突解决整体算法如下.

算法1.

输入:冲突数据集 D_C,其中包括属性集 A、实体集 E,训练集为 D_{Train}、测试集为 D_{Test};数据冲突阈值 T

输出:数据集 D_R,其中数据冲突已解决

主要步骤:

1. 初始化:属性集合 $A_L=\varnothing$,$A_H=\varnothing$,其中 A_L、A_H 分别表示弱冲突属性集合和强冲突属性集合;$D_R=\varnothing$
2. for $a_i\in A$
3. 计算 a_i 属性上的数据冲突程度 $Conflict(a_i)$;
4. if $Conflict(a_i)<T$
5. $A_L=A_L\cup\{a_i\}$;
6. else
7. $A_H=A_H\cup\{a_i\}$;
8. 定义谓词及公式,利用训练集 D_{Train} 训练 Markov 逻辑网模型并对属性集 A_L 上的 D_{Test} 进行冲突解决,得到结果集 D_1;
9. $D_{\text{Train}}=D_{\text{Train}}\cup D_1$;
10. 公式中加入新规则,利用新的训练集 D_{Train} 训练 Markov 逻辑网模型并对属性集 A_H 上的 D_{Test} 进行冲突解决,得到结果集 D_2;
11. for $e_i\in E$
12. for $a_j\in A$
13. 根据结果集 D_1、D_2 选择相应的真值;
14. 将对应属性的真值组成记录 r_i,$D_R=D_R\cup\{r_i\}$;
15. 返回数据集 D_R.

4.1 冲突程度的度量

在集成数据集中,不同属性的冲突程度各不相同.表1是从多个图书网站收集的关于图书《Flash CS3: The Missing Manual》(ISBN:0596510446)的信息,内容包括数据源(Source)、书名(Title)和作者(Authors).对于该图书实体的书名属性,不同数据源提供的信息趋于一致,其冲突程度较小;而相对于作者属性,各数据源提供的信息差别较大,

表1 图书实体属性上的冲突信息

Source	Title	Authors
ABC Books	Flash CS3: The Missing Manual	Chris Grover, E. A. Vander Veer
A1 Books	Flash CS3: The Missing Manual	Veer, E. A. Vander, Grover, Chris
Auriga Ltd	Flash CS3: The Missing Manual	E A Vander Veer, Chris Grover, Vander Veer E., Grover Chris
textbooksNow	Flash CS3: Missing Manual	Vander Veer
Powell's Books	Flash Cs3: The Missing Manual	Vander Veer, E A
Book Lovers USA	Flash CS3: the Missing Manual, by Moore	Moore, Emily
Stratford Books	FLASH CS3	Glover

其冲突程度也较大. 显然，对于一个实体属性而言，数据源提供的不同事实越多，其不确定性越大，则冲突程度也越大. 因此，利用信息论中熵的概念定义实体属性的冲突程度.

定义 2(实体属性的冲突程度). 对于实体属性 $ea\in EA$，各数据源提供的不同事实集合为 $F=\{f_1,f_2,\cdots,f_L\}$，$|f_i|$ 表示事实 $f_i(1\leq i\leq L)$ 出现的次数，则实体属性 ea 的冲突程度可以定义为

$$EAConflict(ea)=-\frac{1}{\log L}\sum_{i=1}^{L}p(f_i)\cdot\log p(f_i) \tag{1}$$

其中 $p(f_i)$ 为事实 f_i 出现的先验概率，$p(f_i)=\frac{|f_i|}{\sum_{j=1}^{L}|f_j|}$；$\log L$ 为归一化因子.

定义 3(属性冲突程度). 对于属性集合 A 中的一个属性 $a\in A$，其对应的实体属性集合为 $EA=\{ea_1,ea_2,\cdots,ea_m\}$，则属性 a 的冲突程度可以定义为

$$Conflict(ea)=\frac{\sum_{i=1}^{m}EAConflict(ea_i)}{m} \tag{2}$$

4.2 利用 Markov 逻辑网进行数据冲突解决

在对属性进行冲突程度划分后，本方法将利用 Markov 逻辑网分别对弱冲突属性及强冲突属性进行两阶段的数据冲突解决. 本小节将首先对 Markov 逻辑网进行介绍，然后分别介绍利用 Markov 逻辑网进行数据冲突解决的几个重要步骤：特征选择、规则设定、真值推理.

4.2.1 Markov 逻辑网

Markov逻辑网(Markov Logic Networks MLN)[5]是一种将 Markov 网络与一阶谓词逻辑相结合的统计关系学习模型，它是对关系数据进行建模的一阶逻辑的概率扩展. 其基本思想是将一阶逻辑的约束软化：一个可能世界违反的公式越多，其发生的概率越小，但其概率未必为零. 对于数据冲突解决问题而言，需要组合不确定及不完美的知识，因此 Markov 逻辑网是个比较适合的模型.

定义 4(Markov 逻辑网). Markov 逻辑网[5] L 是一个二元组集合 $\{(F_i,w_i)\}_{i=1}^{m}$，其中 F_i 为一阶逻辑公式，实数 w_i 为公式 F_i 的权值. 已知 Markov 逻辑网中 L 和有限个体常项集合 $C=\{C_1,C_2,\cdots,C_{|C|}\}$，则可以产生一个以闭谓词为节点、闭谓词关系为边的 Markov 网 $M_{L,C}$，其中，

(1) 每个闭谓词对应 $M_{L,C}$ 的一个二元节点，如果闭谓词为真则对应二元节点状态值为 1，否则为 0.

(2) 每个闭公式对应 $M_{L,C}$ 的一个特征，如果闭公式为真则对应的特征为 1，否则为 0.

Markov 逻辑网可以看作是一个构建 Markov 随机场的模板，其对应 Markov 网络中的一个状态 x 的概率定义为

$$P(X=x)=\frac{1}{Z}\exp\Big(\sum_i w_i n_i(x)\Big)=\frac{1}{Z}\prod_i \phi_i(x_{\{i\}})^{n_i(x)} \tag{3}$$

其中 Z 为归一化常数，$n_i(x)$ 为 F_i 在 x 中所有取真值的基本规则数量，$x_{\{i\}}$ 是 F_i 中为真的原子，$\phi_i(x_{\{i\}})=e^{w_i}$，$w_i$ 为第 i 个公式的权重.

式(3)所定义的是一个产生式 Markov 逻辑网模型，即它定义了所有谓词的联合概率分布. 在本文的数据冲突解决应用中，预先已知了证据谓词和查询谓词，需要对谓词的条件概率分布建模，因此，需要一个判别式 Markov 逻辑网模型[12]. 相比于产生式模型，判别式模型可以充分利用问题中的一切有用特征作为证据，也更有应用前景. 首先，将谓词划分为两个集合——证据谓词 X 和查询谓词 Q. 给定一个常量 x，判别式 Markov 逻辑网将条件概率定义如下[12]：

$$P(q|x)=\frac{1}{Z_x(w)}\exp\Big(\sum_{i\in F_Q}\sum_{j\in G_i}w_i g_j(q,x)\Big) \tag{4}$$

其中 $Z_x(w)$ 为归一化因子，F_Q 是至少包括一个查询谓词的公式集合，G_i 为第 i 个一阶谓词公式的基本公式集合. $g_j(q,x)$ 是一个二元函数，当第 j 个基本公式为真时其值为 1，否则为 0.

本文中的数据冲突解决是判别多个冲突事实的真伪性，从中选择与现实世界一致的真值. 因此，查询谓词只有一个，用来判定一个事实是否准确，即 $IsAccurate(fact)$. 证据谓词可以是冲突事实所呈现出来的特征，在式(4)所定义的判别式 Markov 逻辑网模型中，任意有用的特征都可以作为证据谓词. 根据这些预定义的特征，可以定义一些 Markov 逻辑网中的谓词公式.

4.2.2 特征选择

根据对数据源、数据及数据冲突特点的观察和分析，本文主要从 4 种不同的角度搜集特征：基本特征、数据源的可信性与事实的准确性、事实之间的相互印证、数据源之间的相互依赖. 对于不同的特征，在 Markov 逻辑网中，将为其定义相关谓词，表 2 展示了本方法中用到的不同谓词.

表 2 数据冲突解决中的谓词

搜集角度	谓词	描述
基本特征	$Provide(s,f)$	数据源 s 提供事实 f
	$About(f,ea)$	f 是关于实体属性 ea 的一个事实
	$Belong(ea,e)$	实体属性 ea 是实体 e 的一个属性
	$MaxFrequence(ea,f)$	f 是关于实体属性 ea 的所有事实中出现次数最多的事实
数据源的可信性与事实的准确性	$IsAccurate(f)$	事实 f 是准确的
	$IsTrustworthy(s)$	数据源 s 是可信的
事实之间的相互印证	$Equal(f_1,f_2)$	事实 f_1 与事实 f_2 的内容等价
	$Contain(f_1,f_2)$	事实 f_1 的内容包含事实 f_2 的内容
数据源之间的相互依赖	$InterDepend(s_1,s_2)$	数据源 s_1 与 s_2 之间存在依赖关系

(1) 基本特征

基本特征展示了数据源、实体、实体属性及事实之间的基本关系. 其中,数据源 s 提供事实 f,其谓词可表示为 $Provide(s,f)$;实体属性 ea 是实体 e 的一个属性,其谓词可表示为 $Belong(ea,e)$;事实 f 是关于实体属性 ea 的一个事实,其谓词可表示为 $About(f,ea)$. 另外,为了引入"投票规则",这里使用谓词 $MaxFrequence(ea,f)$ 表示 f 是关于实体属性 ea 的所有事实中出现次数最多的事实.

(2) 数据源的可信性与事实的准确性

在数据冲突解决中,应该存在这样的"可信"数据源:相对于其它数据源,它提供的事实准确度比较高,如表 1 中的数据源 ABC Books 和 A1 Books,事实上这种情况是存在的. 于是,如果多个这样的"可信"数据源对同一实体属性提供相同的事实,那么认为该事实也很可能是真值;同样,如果一个数据源对很多实体属性都能提供真值,那么认为该数据源也很可能是"可信的". 为了表示数据源的可信性和事实的准确性,这里引入谓词 $IsTrustworthy(s)$ 和 $IsAccurate(f)$.

(3) 事实之间的相互印证

对于同一实体属性不同事实之间存在着相互印证的关系. 这里主要考虑两种关系:等价和包含. 如在表 1 中,一个数据源提供的一本图书的作者为"Chris Grover, E. A. Vander Veer",另一个数据源提供的则是"Veer, E. A. Vander, Grover, Chris",这两个事实内容是一致的,只是表现形式不同,这里将这种关系定义为等价关系,使用谓词 $Equal(f_1,f_2)$ 表示. 又如两个数据源分别提供事实"E. A. Vander Veer"和"Vander Veer",则前一个事实的内容包含后一个事实的内容,这里将这种关系定义为包含关系,使用谓词 $Contain(f_1,f_2)$ 表示.

(4) 数据源之间的相互依赖

如果两个数据源对很多实体的实体属性都提供一致的事实,那么认为这两个数据源之间存在着依赖关系,从而它们对于其它实体属性提供的事实也极有可能具有相同的准确性. 本文引入谓词 $InterDepend(s_1,s_2)$ 来表示数据源之间的依赖关系. 为了更严谨地定义和计算数据源之间的依赖,这里给出依赖的定义.

定义 5(数据源之间的依赖). 对于两个数据源 s_1、s_2,如果存在 $\frac{|F_1 \cap F_2|}{|EA_1 \cap EA_2|} \geq \alpha$,则称 s_1、s_2 存在依赖关系. 其中,F_1、F_2 分别表示 s_1、s_2 提供的事实集合,EA_1、EA_2 分别表示 s_1、s_2 提供了事实的实体属性集合,参数 $\alpha \in [0,1]$ 为阈值. 需要说明的是,两个事实相等当且仅当它们是关于同一实体属性的事实且值相等.

4.2.3 规则设定

通过对数据源及其数据特点的观察和分析,针对上一小节定义的特征,本节引入一些用于推理的规则,这些规则体现出一些启发式的特点,以谓词公式的形式展现. 由于 Markov 逻辑网具有强大的表述能力,当需要新的规则加入时只要定义相应的谓词公式,再重新学习公式权重以用于推理即可,这使得本方法具有较强的可适应性. 另外,本文所提规则大都为不确定性规则,而 Markov 逻辑网恰恰具有这种组合不确定的甚至是相互矛盾的知识的能力,也使得所有对数据冲突解决有用的规则都可以加入到本方法中.

4.2.3.1 第 1 阶段规则

在第 1 阶段数据冲突解决中,对于弱冲突属性,由于冲突程度较小,通过一些简单的规则即可得到很高的数据冲突解决准确度,这里介绍投票规则和事实之间相互印证规则.

(1) 规则 1. 投票规则

针对于从多个冲突事实中选择真值的问题,投票规则是最为朴素的规则,通常对同一实体属性出现次数最多的事实往往是准确的.

$$MaxFrequence(ea,f) \Rightarrow IsAccurate(f) \quad (5)$$

(2) 规则 2. 事实之间相互印证规则

如果同一实体属性 ea 的两个事实的内容是等价的,则它们具有相同的准确性;通常,详尽的信息会比简略的信息好,因此,对于同一实体属性的两个

事实 f_1 和 f_2，如果 f_1 的内容包含 f_2 的内容并且已知 f_2 是准确的，那么 f_1 也是准确的.

$$Equal(f_1,f_2)\Rightarrow(IsAccurate(f_1)\Leftrightarrow IsAccurate(f_2)) \tag{6}$$

$$About(f_1,ea)\wedge About(f_2,ea)\wedge Contain(f_1,f_2)\wedge IsAccurate(f_2)\Leftrightarrow IsAccurate(f_1) \tag{7}$$

4.2.3.2　第 2 阶段规则

在第 2 阶段数据冲突解决中，对于强冲突属性，由于冲突程度较大，这里除了第 1 阶段利用的简单规则外，还加入一些比较复杂的规则：数据源与事实之间的相互影响规则、数据源之间相互依赖规则及弱冲突属性对强冲突属性影响规则.

(1) 规则 3. 数据源与事实之间的相互影响规则

基于 4.2.3 小节的分析，通常准确事实的提供者是可信的，而可信数据源提供的事实是准确的. 这里定义以下规则

$$IsAccurate(f)\wedge Provide(s,f)\Rightarrow IsTrustworthy(s) \tag{8}$$

$$IsTrustworthy(s)\wedge Provide(s,f)\Rightarrow IsAccurate(f) \tag{9}$$

(2) 规则 4. 数据源之间相互依赖规则

如果两个数据源 s_1、s_2 对很多实体属性都提供一致的事实，那么认为这两个数据源之间存在着依赖关系，从而它们对于其它实体属性提供的事实也极有可能具有相同的准确性.

$$InterDepend(s_1,s_2)\wedge About(f_1,ea)\wedge About(f_2,ea)\wedge Provide(s_1,f_1)\wedge Provide(s_2,f_2)\Rightarrow (IsAccurate(f_1)\Leftrightarrow IsAccurate(f_2)) \tag{10}$$

(3) 规则 5. 弱冲突属性对强冲突属性影响规则

对于一个实体而言，如果一个数据源对同一实体的多个实体属性都提供正确的事实，那么它对其它实体属性提供的事实很可能也是正确的.

$$IsAccurate(f_1)\wedge Provide(s,f_1)\wedge About(f_1,ea_1)\wedge Belong(ea_1,e)\wedge Provide(s,f_2)\wedge About(f_2,ea_2)\wedge Belong(ea_2,e)\Rightarrow IsAccurate(f_2) \tag{11}$$

4.2.4　真值推理

除了定义好的特征及规则以外，Markov 逻辑网还必须包括每个规则的相对权重. 在数据冲突解决问题中，由于事先并不知道各规则之间的依赖关系，因此需要通过自动训练模型来学习每个公式的权重.

目前，对于判别式 Markov 逻辑网中权重学习的比较好的方法是表决感知算法[13]. 表决感知算法是一种梯度下降算法，它首先将所有权重设为 0，然后根据训练集中预定值与真值是否匹配，来迭代训练数据并更新权重. 最后，为了防止过拟合，权重设置为整个学习阶段的平均值而不是最终的权重. 另外，为了使用表决感知算法，还需要知道每个子句为真的期望值. 不过，通常这种期望值的获取是很难解决的，本文采用 MC-SAT[14] 算法来近似解决这一问题.

通过训练模型得到每个公式的权重后，即可通过 Markov 逻辑网对测试集进行推理，从而从冲突数据中识别出真值. 传统的概率模型推理通常使用 MCMC(Markov Chain Monte Carlo)[15] 算法，而纯逻辑系统通常采用可满足性算法. Markov 逻辑网中的推理既具有概率性又具有确定性，因此，在本文的方法中采用了 MC-SAT 算法来确定查询谓词的值. MC-SAT 算法综合了 MCMC 和可满足技术，因此在 Markov 逻辑网的推理中得到广泛应用.

在方法的最后，需要根据两阶段数据冲突解决的结果将数据集进行合并. 这里，对于每个实体根据对应实体属性上的真值，将表示同一实体的各条记录合并至一条记录，从而组成一致、准确的数据集，主要步骤参考算法 1.

5　实验评价

本文所提方法将在两个真实数据集上进行验证. 首先随机选择一定量的数据作为训练集进行标注，在 Markov 逻辑网的学习和推理中，使用了开源工具 Alchemy. 为了验证算法的有效性，本文将从 4 个方面对实验结果进行分析评价：(1) 数据冲突解决的准确率；(2) 训练样本数量对准确率的影响；(3) 两阶段数据冲突解决的有效性；(4) 不同规则及组合对准确率的影响.

5.1　测试数据集

(1) 图书数据集

首先从 O′Reilly 官网抽取其出版的图书信息(共 1258 条)，抽取信息包括书名、作者、出版年份、ISBN(10 位)，并以此真值，作为最后冲突解决结果的比较标准. 然后利用自制的信息收集程序，将 O′Reilly 出版图书的 ISBN 信息作为关键字，在 www.abebooks.com 中检索并收集相应的图书信息，作为图书数据集. 本数据集共包括来自 881 个数据源的 26 891 条图书信息. 在图书信息中，由于 ISBN 信息不会出现冲突，因此本文通过对书名、作者、出版年份信息进行数据冲突解决来验证本方法

的有效性.由于不同网站提供的作者姓名含有一些杂质信息,本文对数据集进行了一些预处理以去除这些噪音信息.

(2)电影数据集

在图书数据集中,由于出版年份信息出现冲突的可能性较小,本方法主要对书名、作者这样的字符型数据的冲突进行处理;为了验证本方法对不同类型数据冲突的处理能力,本文搜集了一些电影信息,并测试本方法对电影时长这种数字型数据的冲突解决的有效性.首先从 IMDB 网站抽取最热门的 250 部电影的信息,其中包括电影名称、导演、电影时长,基于 IMDB 网站的权威性,将来自 IMDB 的电影信息作为真值,并以此作为比较标准.然后根据抽取的电影名称,利用类似文献[8]的方法从 Google 上搜索并抽取不同站点提供的电影时长及导演信息.本数据集共包括来自 952 个数据源的 7119 条电影信息.

5.2 实验结果与分析

为了全面验证算法的有效性,本文主要从 5 个方面对实验结果进行分析评价.

(1)数据冲突解决的准确率

针对于数据冲突解决问题,准确率(Accuracy)可定义为正确识别真值的实体属性数占实体属性总数的比例.本文在两个数据集上比较了基于 Markov 逻辑网的两阶段数据冲突解决方法(用 2-Stage MLN 表示)、投票方法(Voting)、TruthFinder[8] 及文献[9]中方法(用 Bayes 表示)的准确率.需要说明的是,在 TruthFinder 和 Bayes 中,对于不完整的信息仍有得分,而在本文的实验中,不完整的信息被认为是不准确的.另外,对于内容等价的信息,本文将不考虑其展现形式,如一本图书有多个作者,只要作者数量及每个作者信息都正确则认为该图书作者信息是准确的,而不考虑作者的顺序.

实验中,对于图书数据集,随机选择 600 个实体(不同 ISBN)对应的记录作为训练集,利用定义 3 计算各属性的冲突程度,根据冲突程度的不同,首先选择书名、出版年份作为弱冲突属性,利用 Markov 逻辑网进行第 1 阶段的数据冲突解决,而对于冲突程度较强的作者属性,在第 2 阶段中进行处理.对于电影数据集,选择 120 个实体对应的记录作为训练集,其它作为测试集,同样通过属性冲突程度计算,将导演、电影时长信息分别归为弱冲突属性及强冲突属性,然后进行两阶段的数据冲突解决.其中,判断数据源相互依赖的阈值设置为 $\alpha=0.8$,而数据冲突阈值设置为 $T=0.5$.

图 3 显示了本方法与其它 3 个方法在准确率这个指标上的比较,从图中可以看出,在两个数据集上,本方法都能得到较高的准确率.相对而言,Bayes 的准确率接近于本方法,但在对新特征和新规则的适应能力上较本方法有一定欠缺.其中在图书数据集上,本方法优势明显(92.9%),这主要是因为对于作者信息而言,其不完整、不准确的情况比较多,这在一定程度上也说明本方法处理数据冲突的能力;而对于电影数据集,本方法与 TruthFinder 及 Bayes 在准确率上差别不是特别明显,这是因为相比于图书数据集,同一电影的时长及导演信息变化不大,投票方法也可以获得较高的准确率(87.0%).实验证明,通过两阶段的数据冲突解决,综合运用多角度的特征和规则,本方法能够有效的提高数据冲突解决的准确率.

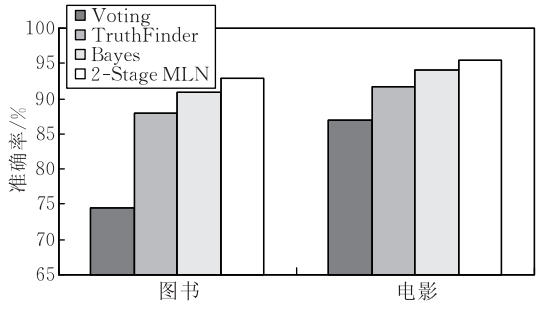


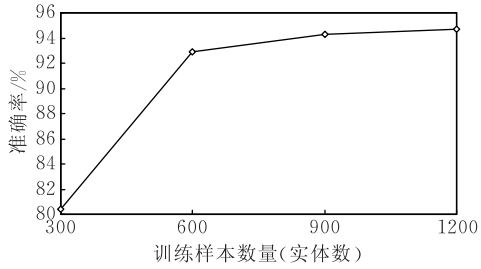
图 3 数据冲突解决的准确率

(2)训练样本数量对准确率的影响

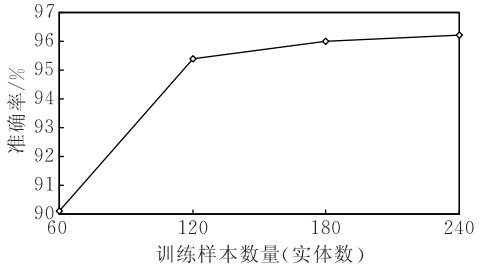
为了进一步验证本方法的影响因素,通过实验测试了训练样本数量的变化对准确率的影响.从图书数据集中随机选择 300、600、900、1200 个实体对应的记录作为训练集,通过本方法进行数据冲突解决;而对于电影数据集,随机选择 60、120、180、240 个实体对应的记录作为训练集.图 4(a)和(b)分别展示了在两个数据集上本方法准确率随着训练样本数量的增加而变化的曲线.通过观察发现,本方法准确率随着训练样本数量的增加而逐步提高.但是曲线会随着训练样本数量的增加而逐渐变得扁平.这个实验表明,训练样本的数量对本方法的性能有着明显的影响;但是随着训练样本数量的继续增加,这种影响将逐步降低.

(3)两阶段数据冲突解决的有效性

本文所提方法的一个重要特点是可以根据冲突程度将属性分为两个集合,然后利用 Markov 逻辑网进行两阶段的分别处理,为了验证两阶段数据冲突解决的有效性,本文进行以下实验.首先将所有属性同等对待,利用文中所提特征及规则(由于无法考



(a) 图书数据集



(b) 电影数据集

图 4　训练样本数量对准确率的影响

虑弱冲突属性对强冲突属性影响规则，仅使用前 4 个规则)，通过训练集训练 Markov 逻辑网模型并进行真值推理，这里将这种方法称为一阶段 Markov 逻辑网方法(简称为 MLN). 然后利用本文所提的 2-Stage MLN 方法对数据集进行同样的实验，比较两种方法的准确率.

图 5 展示了两种方法在两个数据集上的准确率比较，相比于 MLN 方法，2-Stage MLN 可以比较明显地提高数据冲突解决的准确率. 首先对于弱冲突属性，利用第 1 阶段 Markov 逻辑网即可获得较高的数据冲突解决准确率；其次，第 1 阶段冲突解决的结果有效地扩充了第 2 阶段冲突解决的训练集，再利用弱冲突属性对强冲突属性影响等规则，在第 2 阶段可以训练出更为精确的 Markov 逻辑网模型，从而有效地提高了数据冲突解决的准确性.

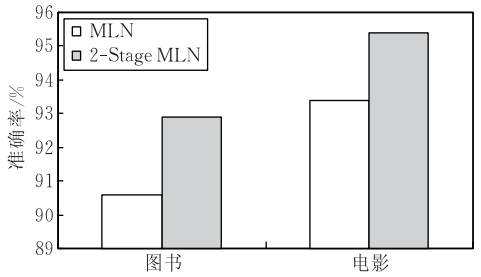


图 5　两阶段数据冲突解决的有效性

(4) 不同规则对准确率的影响

为了验证本文所提规则的有效性，通过实验测试了不同规则对准确率的影响. 由于弱冲突属性对强冲突属性影响规则需要先对弱冲突属性进行冲突解决，该规则仅能在第 2 阶段 Markov 逻辑网中应用，并且其对准确率的影响在图 5 所示实验中已有验证，因此本实验主要针对本文所提的前 4 类规则：投票规则(简称为 V)、事实之间相互印证规则(简称为 I)、数据源与事实之间的相互影响规则(简称为 SF)、数据源之间相互依赖规则(D). 将投票规则作为基本规则，在此基础上分别加入其它 3 种规则形成 3 种规则组合以及所有规则的组合，通过一阶段 Markov 逻辑网实验分别测试使用这 5 种规则及组合所得到的算法准确率.

实验在图书数据集上进行，其它设置与图 3 所示实验相同. 图 6 展示了不同规则对应的准确率直方图. 通过观察发现，所提规则都在一定程度上提高了本方法的准确率，这证实了所提规则的有效性. 在所有规则中，以数据源与事实之间的相互影响规则、事实之间相互印证规则对准确率的提升尤为明显，而数据源之间相互依赖规则对准确率的影响相对较小. 这一方面证实了“可信”数据源的存在，不同数据源的可信性对冲突解决有着很大影响；另一方面，也说明了冲突信息会经常出现不完整或显示形式不一致等现象，这是导致难以识别真值的一个重要原因；而对于数据源之间的相互依赖规则，本文并没有考虑依赖的方向，这是导致依赖规则作用不显著的原因之一.

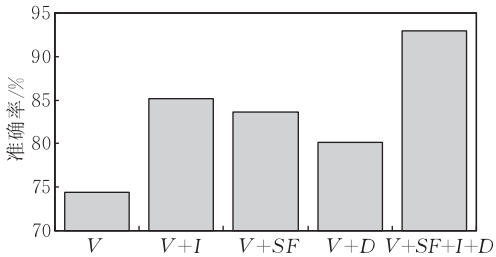


图 6　不同规则对准确率的影响

另外，实验也验证了本方法能够方便地组合不同规则，只需要添加或删除相应的谓词公式即可. 对于数据集成而言，由于集成的动态性特点，会不断地出现新的冲突类型，本方法可以根据新冲突类型的特征及展示出的规则，定义相应的谓词及公式，重新学习即可用于真值推理，这也体现了本方法具有较强的适应能力.

(5) 参数敏感性

在本文的方法中存在两个重要的阈值参数：α 和 T，其中 α 为数据源相互依赖阈值(见 4.2.2 节)，用以判断两数据源是否存在相互依赖关系，而 T 为数据冲突阈值，用以区分弱冲突属性和强冲突属性.

根据经验，将阈值分别设置为 $\alpha=0.8, T=0.5$. 下面的实验表明，在一定的范围内，阈值的选择对方法的准确度影响很小.

图 7(a)显示了在两个数据集上不同 α 值对方法准确率的影响，其中 $T=0.5$. 在区间[0.4，0.9]上，参数 α 对方法准确率影响很小. 这是因为如果 α 取值过小，此时大部分数据源之间都存在着相互依赖关系，这使得通过 MLN 学习得到的规则 4(数据源之间相互依赖规则)的权重很低，在推理过程中该规则起到的作用会很小，特别地，当 $\alpha=0$ 或 $\alpha=1$ 时，规则 4 的作用将降到最低；同理，当 α 取值过大时效果类似. 图 7(b)显示了在两个数据集上不同 T 值对方法准确率的影响，其中 $\alpha=0.8$. 由于数据集中同一实体类型的属性个数较少，数据冲突阈值对准确率的影响有限，一般而言，参数 T 取值在区间[0.2，0.8]上时方法准确率趋于稳定.

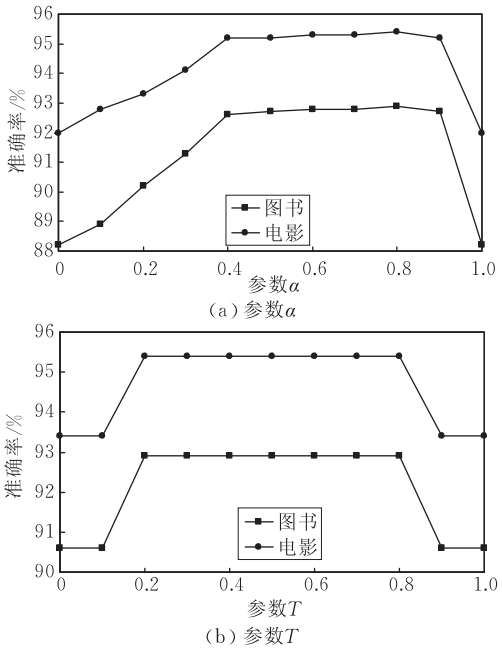


图 7 阈值参数 α 和 T 对准确率的影响

6 结 论

本文提出一种基于 Markov 逻辑网的两阶段数据冲突解决方法，该方法可以根据冲突程度对属性进行划分，并分两阶段进行处理；通过对数据源、数据及其冲突特点的观察和分析，该方法综合运用了多角度的特征和规则，有效地提高了冲突解决的准确度；基于 Markov 逻辑网强大的逻辑表述能力和处理不确定性的能力，该方法可以组合一些不完美甚至相互矛盾的知识来进行学习和推理，具有较强的可适应性. 另外，对于本文所提方法而言，当训练集规模很大时，在模型训练过程中公式权重学习所需时间相对较长，因此，如何进一步提高该方法的效率将是未来工作中面临的主要问题.

参 考 文 献

[1] Dong X, Naumann F. Data fusion—Resolving data conflicts for integration//Proceedings of the 35th International Conference on Very Large Databases (VLDB). Lyon, France, 2009: 1654-1655

[2] Galland A, Abiteboul S, Marian A, Senellart P. Corroborating information from disagreeing views//Proceedings of the 3rd International Conference on Web Search and Web Data Mining(WSDM). New York, USA, 2010: 131-140

[3] Bleiholder J, Naumann F. Conflict handling strategies in an integrated information system//Proceedings of the International Workshop on Information Integration on the Web (IIWeb). Edinburgh, UK, 2006

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

[5] Richardson M, Domingos P. Markov logic networks. Machine Learning, 2006, 62(1-2): 107-136

[6] Bleiholder J, Szott S, Herschel M, Kaufer F, Naumann F. Subsumption and complementation as data fusion operators//Proceedings of the 13th International Conference on Extending Database Technology (EDBT). Lausanne, Switzerland, 2010: 513-524

[7] Wu M, Marian A. Corroborating answers from multiple web sources//Proceedings of the 10th International Workshop on the Web and Databases (WebDB). Beijing, China, 2007

[8] Yin X X, Han J, Yu P S. Truth discovery with multiple conflicting information providers on the Web//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA, 2007: 1048-1052

[9] Dong X X, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence//Proceedings of the 35th International Conference on Very Large Databases (VLDB). Lyon, France, 2009: 550-561

[10] Singla P, Domingos P. Entity resolution with Markov logic//Proceedings of the 6th Industrial Conference on Data Mining (ICDM). Hong Kong, China, 2006: 572-582

[11] Yang J M, Cai Y, Wang Y, Zhu J, Zhang L, Ma W Y. Incorporating site-level knowledge to extract structured data from web forums//Proceedings of the 18th International Conference on World Wide Web (WWW). Madrid, Spain, 2009: 181-190

[12] Singla P, Domingos P. Discriminative training of Markov logic networks//Proceedings of the 20th National Conference on Artificial Intelligence (AAAI). Pittsburgh, Pennsylvania, USA, 2005: 868-873

[13] Lowd D, Domingos P. Efficient weight learning for Markov logic networks//Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD). Warsaw, Poland, 2007: 200-211

[14] Poon H, Domingos P. Sound and efficient Inference with probabilistic and deterministic dependencies//Proceedings of the 21st National Conference on Artificial Intelligence (AAAI). Boston, Massachusetts, USA, 2006: 458-463

[15] Gilks W R, Richardson S, Spiegelhalter D J. Markov Chain Monte Carlo in Practice. London, UK: Chapman and Hall, 1996

ZHANG Yong-Xin, born in 1978, Ph. D. candidate. His research interests include Web data integration and Web data fusion.

LI Qing-Zhong, born in 1965, Ph. D., professor, Ph. D. supervisor. His research interests include data integration and Software as a Service (SaaS).

PENG Zhao-Hui, born in 1978, Ph. D., associate professor. His research interests include keyword search in database and Web data management.

Background

Data conflict resolution is one of the key issues of data integration which is closely related to the quality of integrated data. Data conflict resolution needs to accurately identify the true fact from numerous conflict data. A lot of researches on data conflict resolution have been conducted. However, most existing methods only consider single attribute, neither conflict degree nor mutual influence of different attributes are considered in data conflict resolution. It causes their accuracy not to be high.

A majority of current approaches focus on relation expansion, which resolves data conflict by expanding relation operation or aggregation function. There are also some approaches on truth discovery, which aim at resolving data conflict by identifying true fact from conflict data. But these approaches only consider a few factors and not consider mutual influence of different conflict degree attributes, which results in low accuracy.

In this paper, a 2-stage approach based on Markov Logic Networks is proposed. This approach can divide different attributes according to their conflict degree and carry on 2-stage data conflict resolution: (1) In the first stage, the attributes which conflict degree is low can be resolved by simple rules such as voting and mutual verification of facts; (2) In the second stage, with the aid of the results from the first stage, the attributes which conflict degree is high can be resolve via adding some more complex rules such as mutual influence between sources and facts, inter-dependency of sources and low conflict degree attributes to high conflict degree attributes influence. Experimental results using a large number of real-world data show that the proposed approach can resolve the integrated data conflict effectively, which is more accurate.

This work was supported in part by the National Key Technologies R&D Program (2009BAH44B02), the National Natural Science Foundation of China (90818001, 61003051) and the Key Technology R&D Program of Shandong Province (2010GGX10108).