基于数据源依赖关系的信息评价方法研究

张志强[1)] 刘丽霞[1),2)] 谢晓芹[1)] 潘海为[1)] 方一向[1)]

[1)](哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

[2)](闽南理工学院信息管理系 福建 石狮 362700)

摘 要 当前很多的数据管理应用都需要从多个数据源集成数据,每个数据源都会提供一组值,并且不同的数据源常常提供相互冲突的数据值.为了提供给用户高质量的数据值,关键是数据集成系统能够解决数据冲突问题,提取出正确的数据值.文中对已有的真值发现算法进行了分析与总结,通过考虑处理同一个值的不同表现形式和改进的选票算法,作者对现有方法给出了改进,改进后的方法可以更有效地在众多冲突数据中找出正确的数据值.

关键词 数据源;数据值;数据集成系统;真值;选票算法

中图法分类号 TP311 **DOI号**: 10.3724/SP.J.1016.2012.02392

Information Evaluation Based on Sources Dependence

ZHANG Zhi-Qiang[1)] LIU Li-Xia[1),2)] XIE Xiao-Qin[1)] PAN Hai-Wei[1)] FANG Yi-Xiang[1)]

[1)](*College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)

[2)](*Department of Information and Management, Minnan University of Science and Technology, Shishi, Fujian* 362700)

Abstract Many data management applications require integrating data from multiple sources. Each of these sources provides a set of values and different sources can often provide conflicting values. To present quality data to users, it is critical that data integration systems can resolve conflicts and discover true values. In this paper, we improve the existing algorithm with using a new voting algorithm and considering the diverse expressions of the same value, e.g. person's name. The experiment results shown it is very effective for discovering true values among conflicting values.

Keywords data sources; data value; system of data integrating; true value; vote algorithm

1 引 言

随着互联网的快速发展,Web上的数据越来越多,已经成为一个巨大的数据库,里面的信息随着时间的推移不断在变化,这种变化主要体现在两个维度,一个是指客观世界对象信息的增删变化,例如亚马逊网站上增加了某本新书的信息;另外一个维度是指同一个客观世界对象的信息随着时间的变化而发生变化,例如一个人的联系方式、家庭住址、工作单位等信息会随着时间的变化而变化.与此同时,我们还经常发现会有多个不同的数据源提供同一对象的信息,例如多个网站提供飞机航班/火车车次或宾馆的信息.提供这些信息的数据源也要不断地更改它们的数据,以获取对应对象的最新状态.那么当我们需要从互联网上获取一个对象的信息时,就需要

收稿日期:2012-06-05;最终修改稿收到日期:2012-09-04.本课题得到国家自然科学基金(60803037,61202090,61272184)、教育部新世纪人才支持计划(NCET-11-0829)、黑龙江省自然科学基金(F201130,F201016)、哈尔滨市科技创新人才研究专项基金(RC2010Q010024)和中央高校基本科研业务费专项资金(HEUCFZ1010,HEUCFT1202)资助.**张志强**,男,1973年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为信息检索、智能信息处理、数据库.E-mail: zqzhang@hrbeu.edu.cn. **刘丽霞**,女,1986年生,硕士,助教,主要研究方向为信息检索.**谢晓芹**,女,1973年生,博士,副教授,主要研究方向为服务计算、知识工程、社会网络、智能信息处理.**潘海为**,男,1974年生,博士,副教授,主要研究方向为数据挖掘、智能信息处理.**方一向**,男,1988年生,硕士,主要研究方向为信息检索与数据挖掘.

确定哪个数据源提供的信息更准确，质量更高.

影响数据值质量的原因有很多.首先，数据值随着时间的变化在不断地更新变化，错误的信息可能会潜入到数据中，一些数据可能会过时.这对于决定哪些值曾经是正确的，并且在哪个阶段是正确的是一个难题.其次，数据源通常有不同的质量，一个自然的想法是当我们判断真值的时候将数据源的因素考虑进去.影响数据低质量的原因有很多，一些数据源在它们初始提供数据的时候就造成了一些错误，一些数据源虽然提供了正确的数据，但是没能及时进行信息的更新.第三，一个数据源会从其它的数据源那里复制一些数据，通常不会知道所复制的这些数据是否正确或是否是过时的.此外，数据源之间的复制依赖关系也会随着时间的变化而不断地变化.本文的主要工作就是研究如何解决数据冲突问题，查找准确的数据值.

数据冲突问题早期在数据处理领域中就已经被提了出来[1].后续的一些数据集成系统也相继提出了一些解决策略[2-4]，文献[5]对数据集成系统中的冲突处理策略进行了总结.Cho 等人[6]提出了一种自动维护本地数据库拷贝与初始数据源真值一致性的方法.由于这些工作的背景，在解决数据冲突问题时，往往假设数据源都是独立的，相互之间没有关联.但是在 Web 的环境下，同一个对象的信息往往会分布在多个数据源中，而这些数据源的数据具有较强的关联，文献[7]注意到了 Web 数据的这个特点，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，并且引入链接分析的方法[8]来对数据源的可信度进行评判.Berti-Equille 等人[9]也注意到了数据源依赖这个问题，并给出了问题的详细定义和描述.文献[10-11]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，获得了较好的效果.Dong 等人[12]对集成问题中的冲突解决研究工作给出了详细的总结和分析.文献[13]采用一种不同的概率投票方法，将数据源的可靠性和描述的准确性之间的关系运用在投票的思想中，同时考虑不同描述之间的影响，另外还考虑了投票数据源的权威性.本文则在分析总结上述工作的基础上，进一步考虑了同一对象数据值具有不同表达形式这个因素以及结合了数据源投票值与数据源准确率不一致的情况，提出了一种改进的方法，实验结果表明，本文的方法提升了结果的准确率.

本文第 2 节将对信息评价问题的主要相关工作进行总结与分析；第 3 节将介绍本文提出的基于数据源依赖的信息评价方法；第 4 节介绍了本文采用的实验方案，通过实验对提出的新方法进行评估，并对实验结果进行分析；第 5 节对全文进行总结并给出今后的研究方向.

2　主要相关工作介绍

本节将主要介绍该领域中的几个经典工作.由于本文工作是在文献[7,10]的基础上进行的改进，并且选取与这两个文献相同的数据集进行测试.因此为了能够获得更客观的对比结果，本文选择这两个文献的工作作为对比对象.下面将主要介绍这两个文献的主要方法.

2.1　基于数据源和信息可信度的数据评价方法

这类方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源所提供的信息，所以数据源的信任度对数据信息准确性的影响是存在的，而数据源的信任度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，利用迭代算法的思想去计算数据源的信任度和数据值的可信度.

这类方法的典型代表是 Yin 等人[7]提出的 ACCUNOD 算法，该方法的思路是每个数据源的信任度是由该数据源提供的所有数据值的可信度期望值得到的，由式(1)计算每个数据源的信任度就是求该数据源 w 所提供的所有数据值可信度的平均值：

$$t(w)=\frac{\sum_{f\in F(w)} s(f)}{|F(w)|} \tag{1}$$

式中，$F(w)$是数据源 w 提供的数据值的集合，$s(f)$是数据值的可信度.

相比之下，估计一个数据值的可信度是很困难的事，对于每一个对象都可能会有很多的数据值，这些数据值之间是相互冲突的.假设 f_1是对象 o 的数据值，数据源 w_1和 w_2都提供 f_1这个值，假定数据源 w_1和 w_2是相互独立的，这样 f_1是错误的概率是$(1-t(w_1))(1-t(w_2))$，f_1不是错误的概率是$1-(1-t(w_1))(1-t(w_2))$.如果 f 是对象 o 的唯一的数据值，那么 f 的可信度是 $s(f)=1-\prod_{w\in W(f)}(1-t(w))$，$W(f)$是提供 f 值的数据源集合，经过变换

后,上式改成 $\tau(w)=-\ln(1-t(w))$,那么数据值的可信度可以表示成 $\sigma(f)=-\ln(1-s(f))$.

对于每一个对象实体总会有很多的数据值,这些数据值之间是有一定关联的,如存在两个数据值 f_1 和 f_2,f_1 是由很多信任度很高的数据源提供的数据值,而 f_1 和 f_2 具有很强的关联,那么有理由认为 f_2 也得到了这些信任度高的数据源的支持,所以要增加 f_2 的可信度值,即

$$\sigma^*(f_2)=\sigma(f_2)+\rho\cdot\sum_{o(f_1)=o(f_2)}\sigma(f_1)\cdot imp(f_1\rightarrow f_2) \tag{2}$$

式中,ρ 是 0 到 1 的一个变量,它控制着数据值之间的相关关系;$imp(f_1\rightarrow f_2)$ 表示 f_1 与 f_2 的相关性;$o(f)$ 表示数据值 f 对应的对象.

该模型考虑到了数据源的准确性对数据值的准确性也有影响.通过计算数据源的准确性来计算数据的可信度.而数据的可信度又是由数据源的准确性决定的.不同质量的数据源提供的数据的准确性是不同的,我们当然会更相信那些准确性比较高的数据源,但是本算法没有考虑数据源之间复制关系对数据值的影响.

2.2 基于数据源依赖关系的数据评价方法

Dong 等人[10]考虑了数据源之间的复制依赖关系,提出了 BENE 算法、MAL 算法和 ACCU 算法.

2.2.1 BENE 算法

善意模型是假设提供数据信息的数据源都是好的,它们会尽量提供正确的信息,检测到错误的数据并将其改正,正确的数据要比错误的数据更加稳定.具体的思想是在某一时刻对某一对象的数据信息进行监测,根据概率公式计算数据源之间的依赖关系.依据提供该数据值的数据源依赖关系图对数据值进行选票,比如对于一个数据值 v,总共有 5 个数据源提供这个值,那么 v 获得的选票数就是 5,然而这样却忽略了数据源之间可能存在的复制依赖关系.假设有 3 个数据源是相互复制的,那么它们相当于重复地进行了投票,所以调整后的选票数就是 $x(x<5)$,当然这里面考虑了数据源的复制概率影响.选票数越多说明该数据值越准确,因为提供该值的独立数据源比较多.具体的思路是先计算数据源之间的依赖关系,这里认为所有数据源的初衷都是善意的,所以只要求出数据源之间的依赖关系就可以了,然后根据这个依赖关系计算数据值的选票,利用贪心迭代算法得到稳定的结果.

数据值的选票算法是这样的,首先考虑一个特定对象的某一数据值 v,令 $\overline{S}_0(v)$ 表示为对象 o 提供 v 值的数据源集合,假设在集合 $\overline{S}_0(v)$ 中的一对数据源,如果我们知道其中的一个数据源复制了另一个数据源,并且还知道哪个是复制者,那么我们就可以画出一个依赖关系图,每一个数据源就是一个结点,对于每对数据源,如果有 S_1 复制 S_2 的数据值,那么就存在一条边,从 S_1 指向 S_2,如($S_1\rightarrow S_2$).

对于每一个属于 $\overline{S}_0(v)$ 的数据源 S,我们定义 $d(S,G)$ 为图 G 中数据源 S 的出度,表示数据源从多少个数据源复制了数据.如果 $d(S,G)=0$,说明数据源 S 是独立的,它的选票值为 1,否则,数据源 S 复制数据源 S' 的数据值.数据源 S 独立于数据源 S' 提供一个数据值的概率是 $1-c$,c 是数据源 S 复制数据源 S' 的概率.那么有数据源 S 独立于其它数据源提供 v 值的概率是 $(1-c)^{d(S,G)}$,式(3)表示了关于图 G,数据值 v 的总选票数:

$$V(v,G)=\sum_{S\in\overline{S}_0(O)}(1-c)^{d(S,G)} \tag{3}$$

该方法虽然可以知道依赖关系,但是不能确定依赖关系的方向,所以必须要计算所有图的选票数,这样计算起来很麻烦,所以引入了估计选票数的算法,并将其应用在 ACCU 算法中.

2.2.2 MAL 算法

恶意模型考虑到了提供信息的数据源不全都是好的事实,因为有一些数据源会刻意隐藏自己的复制行为,为了不让其它数据源发现它是复制源,它会将复制过来的数据进行一些修改,我们称这类数据源为恶意数据源.为了查找出恶意的数据源,我们必须要知道数据源之间依赖关系的方向,所以该算法解决了找出数据源复制方向的问题.数据源的依赖关系方向不同,它们共同提供正确的值和错误的值的概率也不同.该算法是在 BENE 算法的基础上考虑了数据源之间的复制方向.

数据源集合 S 里面有善意的独立源和恶意的复制源.假设每个对象共有 n 个错误的数据值($n>1$),当一个恶意的复制源独立地提供一个数据值时,选择其中一个错误数据值的概率是 $1/n$.考虑两个数据源 S_1 和 S_2,直观来讲,恶意的复制源比善意的复制源更有可能将正确的值改成错误的值,如果 S_2 提供的正确值是 S_1 提供的值的子集,那么很可能数据源 S_2 复制了数据源 S_1,但是如果数据源 S_2 提供的值和数据源 S_1 提供的值很少或没有相同的,那么 S_1 和 S_2 就可能有一个是恶意的复制源.

$\overline{O}_t$ 为数据源 S_1 和数据源 S_2 提供相同的正确值

的对象实体集合. $\bar{O}_f$ 为数据源 S_1 和数据源 S_2 提供不相同的错误值的对象实体集合. $\bar{O}_d$ 为数据源 S_1 和数据源 S_2 提供不同值的对象实体集合，可以将其划分成 3 个集合：$\bar{O}_{d1}$ 代表的是数据源 S_1 提供正确数据值的对象实体集，数据源 S_2 提供错误的数据值；$\bar{O}_{d2}$ 代表的是数据源 S_2 提供正确数据值的对象实体集，数据源 S_1 提供错误的数据值；$\bar{O}_{d0}$ 代表的是 S_1 和 S_2 提供了不同错误数据值的对象实体集. 此外，定义 S_1 复制 S_2 表示成 $S_1 \rightarrow S_2$，相反 S_2 复制 S_1 表示成 $S_2 \rightarrow S_1$.

若 S_1 与 S_2 具有依赖关系，计算 Φ 的概率（某一时刻对数据集进行观测可能会出现的所有情况的空间[9]，用 Φ 来表示，例如数据源 S_1 和 S_2 同时为对象 o 提供数据值，可能会是同一正确的值，也可能会是同一错误的值，还有可能是不同的错误值. 这 3 种可能构成了一个观测空间）. 如果 S_1 与 S_2 是相互独立的，由条件概率公式得出

$$P(o\in\bar{O}_t \mid S_1 \perp S_2)=(1-\varepsilon)^2 \tag{4}$$

$$P(o\in\bar{O}_f \mid S_1 \perp S_2)=\frac{\varepsilon^2}{n} \tag{5}$$

$$P(o\in\bar{O}_{d1} \mid S_1 \perp S_2)=\varepsilon\cdot(1-\varepsilon) \tag{6}$$

$$P(o\in\bar{O}_{d2} \mid S_1 \perp S_2)=\varepsilon\cdot(1-\varepsilon) \tag{7}$$

$$P(o\in\bar{O}_{d0} \mid S_1 \perp S_2)=\varepsilon^2\cdot\frac{n-1}{n} \tag{8}$$

其中，ε 是 S_1 或 S_2 提供错误数据值的概率，ε^2 是 S_1 和 S_2 都提供错误数据值的概率，$(n-1)/n$ 是 S_2 提供一个不同于 S_1 的错误数据值的概率.

如果 S_2 复制 S_1 的数据值，由条件概率公式得出

$$P(o\in\bar{O}_t \mid S_2 \rightarrow S_1)=(1-\varepsilon)\cdot c \tag{9}$$

$$P(o\in\bar{O}_f \mid S_2 \rightarrow S_1)=\varepsilon\cdot c \tag{10}$$

$$P(o\in\bar{O}_{d1} \mid S_2 \rightarrow S_1)=(1-\varepsilon)\cdot(1-c) \tag{11}$$

$$P(o\in\bar{O}_{d2} \mid S_2 \rightarrow S_1)=\varepsilon\cdot(1-c)\cdot\frac{1}{n} \tag{12}$$

$$P(o\in\bar{O}_{d0} \mid S_2 \rightarrow S_1)=\varepsilon\cdot(1-c)\cdot\frac{n-1}{n} \tag{13}$$

式(9)和(10)考虑的情况是 S_2 复制 S_1 的数据值，这个数据值是正确值的概率是 $1-\varepsilon$，错误值的概率是 ε. 式(11)～(13)考虑的情况是 S_2 独立的提供 v 值（其概率为 $1-c$，c 为复制其它数据源的概率），如果 S_1 提供一个正确的数据值，那么 S_2 提供一个正确的值的概率是 $1/n$. 提供一个不同的错误的值的概率是 $(n-1)/n$.

当 S_1 复制 S_2 的数据值时，由条件概率公式得出

$$P(o\in\bar{O}_t \mid S_1 \rightarrow S_2)=(1-\varepsilon)\cdot c \tag{14}$$

$$P(o\in\bar{O}_f \mid S_1 \rightarrow S_2)=\varepsilon\cdot c \tag{15}$$

$$P(o\in\bar{O}_{d1} \mid S_1 \rightarrow S_2)=\varepsilon\cdot(1-c)\cdot\frac{1}{n} \tag{16}$$

$$P(o\in\bar{O}_{d2} \mid S_1 \rightarrow S_2)=(1-\varepsilon)\cdot(1-c) \tag{17}$$

$$P(o\in\bar{O}_{d0} \mid S_2 \rightarrow S_1)=\varepsilon\cdot(1-c)\cdot\frac{n-1}{n} \tag{18}$$

这样具有不同的依赖关系方向，就会得出不同的概率值. 下面是利用数据源之间的依赖关系来计算数据值的选票数.

每一个数据值对应一些依赖关系图，如图 1 所示 3 个数据源 S_1、S_2 和 S_3 具有 4 个依赖关系图，所有依赖关系图的选票数之和就是这个数据值的选票数（每一个依赖关系图边的权重为数据源之间的依赖关系概率值）. 存在的一个问题就是当数据源的个数增加时，依赖关系图的数量也会成指数的增加. 在这里 Dong 等人设计了一个简单的计算数据值选票数的算法，该算法大大降低了时间复杂度和空间复杂度. 其基本思路是先计算提供 v 值的数据源的选票数，再将这些数据源的选票数求和，就是数据值 v 的选票数，具体过程在文献[2]中有详细描述.

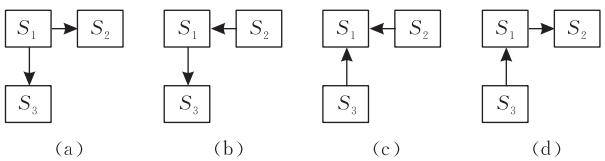

图 1　数据源的依赖关系图

两种模型相互比较，它们都会从原始的数据源那里共享错误的信息，BENE 算法模型能够判断出这些错误的信息，但是 BENE 算法模型计算得到的恶意复制依赖关系概率要比 MAL 算法模型的低，并且也不能够判断出复制依赖关系的方向，甚至当它发现该复制是恶意复制时，它还要假定为善意复制，并进行投票计算. 换句话说，恶意模型适合于恶意数据源的查找，但是却忽略了一些将错误数据值改成正确数据值的善意数据源. 即使它发现了善意的复制源，也会将它所提供的所有值忽略掉，尽管有一些值确实是独立的.

2.2.3　ACCU 算法

ACCU 算法是在 BENE 算法和 MAL 算法的基础上提出的，该方法既考虑了数据源的依赖关系，也考虑了数据源的可信度，主要的思想是通过利用式(19)计算数据源的可信度：

$$A(S)=\frac{\sum_{v\in\bar{V}(S)}P(v)}{m} \tag{19}$$

式中，$P(v)$ 表示数据值 v 是正确的概率；$\bar{V}(S)$ 表示数据源 S 提供的所有数据值的集合；m 表示数据源

提供数据值的数量，即集合 $\overline{V}(S)$ 中数据值的数量.

计算 $P(v)$ 是通过式(20)和(21)的概率公式和贝叶斯公式完成的：

$$P(\psi(o)|v\ \text{true})=\prod_{S\in\overline{S}_o(v)}A(S)\cdot\prod_{S\in\overline{S}_o-\overline{S}_o(v)}\frac{1-A(S)}{n} \tag{20}$$

$$P(v)=P(v\ \text{true}|\psi(o)) \tag{21}$$

式中，$\psi(o)$ 表示对象 o 的数据值空间；$\overline{S}_o$ 表示给对象 o 提供数据值的所有数据源集合，$\overline{S}_o(v)$ 表示给对象 o 提供数据值 v 的数据源集合；v 值是对象 o 的某一特定的数据值. 利用式(22)计算每个数据值的可信度 $C(v)$.

$$C(v)=\sum_{S\in\overline{S}_O(v)}A(S)I(S) \tag{22}$$

式中，$I(S)$ 表示的是数据源的选票数，该算法虽然考虑了数据源的可信度和数据源的复制依赖关系，但是没有对数据值进行很好的处理，例如实际中一个对象的同一数据值在不同的数据源中往往具有不同的表现形式，导致不同表示形式的数据值会得到不同的可信度值，同时也降低了真实数据值在结果集中的影响程度，有可能正确数据值的可信度分值低于错误数据值的可信度分值.

3 本文方法

根据对现有几种方法的总结与分析，我们发现现有方法虽然考虑了数据源可信度、数据值可信度、数据源的依赖关系以及三者之间的关联关系等因素，而且可以较好地刻画数据源之间的复制依赖关系，但是仍然存在以下两点不足：

(1) 没有考虑对象的同一数据值具有不同表现形式的实际情况，而这种现象在实际中十分普遍，且对真值的确定有直接影响.

(2) 实际中一个数据值的选票值与其正确性概率两者并非总是一致的，有时可能会相差较大，而一个数据值的可信度与两者都有关联. 目前的方法只考虑了数据源的可信度和选票值，没有考虑数据值正确性概率的因素.

为此，本文对 ACCU 算法进行了改进，增加了对同一对象数据值不同表示形式的判断，并增强了数据值可信度的计算，算法的基本流程如图 2 所示.

图 2 中(1)～(4)步骤，数据源依赖关系和数据值选票数的计算与 ACCU 算法完全一致，由于篇幅原因这里不再详述. 下面将分别介绍数据源可信度和数据值的正确概率以及数据值可信度的计算方法，并且对数据值之间的相似性判断给予详细介绍.

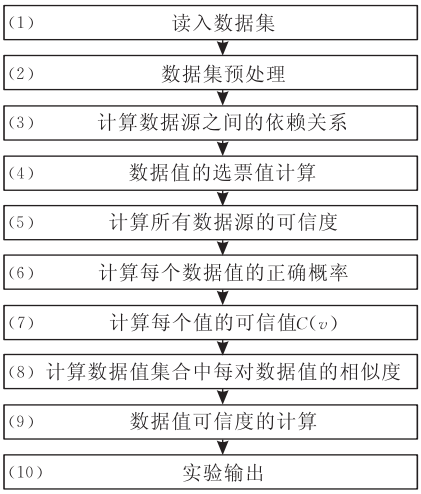

图 2 真值查询算法流程图

3.1 数据源的可信度计算

数据源的可信度影响着数据值的准确程度，通常我们会更加相信那些可信度比较高的数据源，就好像我们向别人打听消息一样，有的人爱说实话，他的话可信度就会很高，而有些人，却很喜欢毫无根据地乱说，他的话可信度就会很低，所以我们不会去听第二类人的话，而比较相信第一类人的话. 对于数据源也一样，可信度越高的数据源它所提供的数据值的可信度也就越高. 依据这一理论，考虑数据源可信度对数据值的正确性影响是必不可少的. 我们采用以下计算数据源的可信度公式：

$$T(S)=\frac{\sum_{v\in V(S)}Vote(v)}{m} \tag{23}$$

式中，m 是数据源 S 提供的值的个数；$V(S)$ 是数据源 S 提供的数据值的集合；$T(S)$ 为数据源 S 的准确性；$Vote(v)$ 是数据值 v 的选票数. 该公式不同于 ACCU 算法中求数据源准确性的方法，该公式的优点是在于数据值的选票分值是表示每个数据值准确程度的. 式(19)是 ACCU 算法中计算数据源准确度的公式，它只是利用数据值的正确概率，然后计算数据源的可信度，而式(23)更能够体现出数据源的可信度，因为每个数据源所提供的数据值的可信程度是不同的，为了更好地体现出数据源的可信度，利用数据值的选票数来计算每个数据源的可信度要比式(19)的方法好，而且通过实验验证了该方法的准确度要高于 ACCU 算法的准确度.

又因为准确率应该是一个 0 到 1 之间的数，而上面等式的结果可能是一个大于 1 的数，所以需要进行指数变换，将其值限制在 0 到 1 之间，上式改为

$$T^*(S)=1-\mathrm{e}^{-T(S)} \tag{24}$$

3.2　数据值的正确概率和可信度计算

我们利用得到的数据源可信度计算得到数据值 v 的正确概率 $P(v)$，由条件概率贝叶斯公式得

$$P(\psi(o)\mid v\ \text{true})=\prod_{S\in\bar{S}_0(v)}T^*(S)\cdot\prod_{S\in\bar{S}_0-\bar{S}_0(v)}\frac{1-T^*(S)}{n} \tag{25}$$

$$P(\psi(o))=\sum_{v\in\nu(O)}(P(\psi(o)\mid v\ \text{true})\cdot P(v\ \text{true})) \tag{26}$$

$$P(v)=P(v\ \text{true}\mid\psi(o))=\frac{\prod_{S\in\bar{S}_0(v)}\frac{nT^*(S)}{1-T^*(S)}}{\sum_{v_0\in\nu(O)}\prod_{S\in\bar{S}_0(v_0)}\frac{nT^*(S)}{1-T^*(S)}} \tag{27}$$

式中，n 是对象 o 具有的数据值总数；$\psi(o)$ 表示所有提供给对象 o 的数据值空间.

通过以上公式重新调整数据值 v 的可信度，如式(28)所示：

$$C(v)=\frac{P(v)+Vote(v)}{2} \tag{28}$$

经过改进确定数据值的可信度利用式(28)来计算，该公式涉及到求数据值的选票数和数据值的正确概率，然后求它们的平均值. 因为每个数据值的选票数和正确性概率的值差距可能很大，可以综合两者的值，通过不同的角度来考虑数据值的准确度，能够更接近数据值 v 的准确率，$P(v)$ 是通过条件概率的理论基础和数据源的准确度来计算的，$Vote(v)$ 是通过计算数据源之间的依赖关系概率，利用贝叶斯网络概率的原理计算的，通过两种方法的结合让数据值的准确性判断更加接近于真实世界的描述，提高查找正确数据值的概率.

上述方法在数据源可信度和数据值正确率以及数据值可信度等方面对原 ACCU 算法进行了改进. 下面我们将介绍本文的另一个主要改进：数据值之间的相似性计算.

3.3　数据值之间的相似性

同一个数据值的表示形式也会有很多种，例如，“Jia-wei Han”、“Han Jiawei”，或“J. W. Han”等均指同一个人的名字；又如日期 2012 年 8 月 20 日，可以表示成“08/20/2012”，或者是“2012-08-20”等形式. 这种情况在实际中十分常见，幸运的是实际当中很多的值往往是有规律的，我们可以利用这些规律进行相似性的判断. 本文采用的方法很简单，首先将所有的值都转换为字符串，然后再根据不同的特点将这个串分解成单元的集合，最后根据分解得到的单元集合进行相似性处理. 本文考虑了 3 种主要的数据值表示方式，第 1 种用英文表示，第 2 种用中文表示，第 3 种用数据值型数据表示. 下面分别介绍一下这 3 种相似度处理方法.

3.3.1　英文字符串型的数据值

例如两个英文字符串“O’Leary, Timothy J, and O’Leary, Lina I”和“O’Leary, Timothy J./O’Leary, Linda I.”，在不考虑字符串的相似度时，之前的方法将它们看作是两个不同的串，因为这两个串不能完全匹配. 但我们知道这两个串表示的是同一个值，为此本文采用了如下的分词方法对字符串进行处理.

字符串是由一组不同含义的单词所组成，将单词集作为二元变量表中的属性集合，设定为集合 Z，假设字符串 $String_1$ 和字符串 $String_2$ 的单词包含于集合 Z 中，设 q 是字符串 1 和字符串 2 中共有的单词总数，s 是字符串 1 中存在，字符串 2 中不存在的单词总数，r 是字符串 2 中存在，字符串 1 中不存在的单词总数，那么采用传统的相似度计算方法中的非恒定相似度评价系数方法(Jaccard 系数)来处理，即两个字符串间的相似度公式如下：

$$Sim(String_1,String_2)=q/(q+r+s) \tag{29}$$

3.3.2　汉字形式的数据值

所谓义原词[14]是指用来描述对象某一特征最小意义的词. 数据值是一个词组或一段文字描述的，那么对于一个数据值 A 来说，可以将其分解成一组义原词集合，这样计算一对数据值的相似度就是对这两个数据值所对应的义原词词组进行相似度计算，其计算公式如下：

$$Sim(A,B)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}Sim(A(a_i),B(b_j))}{\left(m+n-\sum_{i=1}^{m}\sum_{j=1}^{n}Sim(A(a_i),B(b_j))\right)} \tag{30}$$

式中，A、B 表示两个数据值，数据值 A 分解成 $\{a_1,a_2,\cdots,a_m\}$，其中 a_i 代表数据值 A 分解的第 i 个义原词，数据值 B 分解成 $\{b_1,b_2,\cdots,b_n\}$，其中 b_j 代表数据值 B 分解的第 j 个义原词，例如：A =“我爱中国”和 B =“我爱母亲”，那么 A 的义原词集

合为{我,爱,中国},B 的义原词集合为{我,爱,母亲},利用式(30)计算数据值 A 和 B 的相似度就应该是 0.5.

3.3.3 数值型的数据值

由于数值型的数据以数字的形式表示,如果是简单的数据值,我们可以将其看作是义原词,但是对于有些数据,如日期的表现形式,我们还需要进一步划分. 如 20110609、06/09/2011 和 2011-06-09,它们具有不同的表现形式但是却表示相同的意义,利用数值型数据的特点我们采用树形结构模型将数据值拆分,分解成最小的义原词然后进行对比.

3.4 数据值的可信度修正

基于以上理论,依据数据值之间的相似性求解 $sim(v,v')$,我们有式(31):

$$C^*(v)=C(v)+\rho\cdot\sum_{v\neq v'}C(v')\cdot sim(v,v') \quad (31)$$

由上面公式可知计算数据值的可信度 $C^*(v)$,需要先求解数据值的可信度 $C(v)$和数据值之间的相似度 $sim(v,v')$,具体算法如下.

算法 1. 数据值相似性算法.

输入:E_o,F_o

输出:$\overline{S}im_o(v_j)$

//E_o集合:对象 o 的数据值集合

//F_o集合:对象 o 的每一个数据值的可信度$C(v)$集合

//$\overline{S}im_o(v_j)$:数据值 v_j与其它数据值的相似度列表

1. 当集合 E_o不为空时
2. $j=k=1$;
3. 当变量 j 小于等于 E_o中数据值的最大数量时
4. 变量 k 小于等于 E_o中数据值的最大数量时
5. 通过式(29)或(30)计算 $sim(v_j,v_k)$;
6. 如果 $C(v_k)$属于集合 F_o
7. 计算 $C(v_k)*sim(v_j,v_k)$,并将该值加入到 $\overline{S}im_o(v_j)$中;
8. 变量 k 自加返回到 4 处,直到 k 大于 E_o中数据值的最大数量时;
9. 变量 j 自加返回到 3 处,直到 j 大于 E_o中数据值的最大数量时;
10. 程序结束,返回集合 $\overline{S}im_o(v_j)$.

通过对比每个数据值的可信度值来最终断定哪些值才是查询对象的正确数据值.

4 实验与结果分析

4.1 实验数据

本实验所用的数据集来自于 Dong Xin Luna 和 Yin Xiaoxin 两位博士. 下面介绍一下本文的数据集. 第 1 个数据集是 Books_Authors,该数据集是网上书店提供的有关书的信息,需要查询书的作者信息. 第 2 个数据集是 MovieRunTime,该数据集提供了电影的播放时长信息,需要查询电影的准确播放时间长度. 运用本文提出的改进算法对两个数据集进行处理,实验结果表明本文提出的算法无论在数据源数量的变化,还是数据对象数量的变化等方面,得到的准确率都是最好的. 我们定义本文的算法为 NEWACCU 算法.

4.2 Books_Authors 数据集上的实验

本实验是针对在网上书店上查找书的作者信息的应用背景,对于同一本书不同网站提供的该书作者信息可能是不同的,我们将考察不同算法在冲突的数据值中查找正确作者信息的能力.

我们随机选取了 10 个书的对象,分别执行需要进行对比的各种算法. 表 1 给出了 ACCU 算法和 NEWACCU 算法的部分对比结果.

表 1 ACCU 和 NEWACCU 的部分实验结果对照表

对象编号	ACCU 算法实验结果	NEWACCU 算法实验结果	真实结果
9780073-516677	"O'Leary, Linda I. J."	"o'leary, timothy j.; o'leary, linda i."	"o'leary, timothy j.; o'leary, linda i."
9780072-999389	"Yacht,Carol/Crosson, Susan"	"Yacht, Carol/Crosson, Susan"	"yacht, carol; crosson, susan"
9780072-843996	"Haag,Stephen/Perry, James T./Sosinsky, Barrie/Estevez, Efren"	"Haag, Stephen/Perry, James T./Sosinsky, Barrie/Estevez, Efren"	"haag, stephen; perry, james t.; sosinsky, barrie; estevez, efren"
9780072-232172	"Dennis Suhanovs,"	"Dennis Suhanovs"	"suhanovs, dennis"
9780072-230611	"Scambray, Joel"	"scambray, joel; mcclure, stuart"	"scambray, joel; mcclure, stuart"
155860-9350	"Courage, Catherine"	"Courage, Catherine; Baxter, Kathy"	"courage, catherine; baxter, kathy"
155860-8893	"Goldman,Ron"	"Goldman,Ron; Gabriel, Richard P."	"goldman, ron; gabriel, richard p."
155860-846X	"Almeroth, Kevin C."	"Almeroth, Kevin C.; Makofske, David"	"makofske, david; almeroth, kevin"
131872-893	"Dann, Wanda P."	"Dann, Wanda P."	"dann, wanda p.; cooper, stephen; pausch, randy"
131463-055	"Geary, David"	"Geary, David; Horstmann, Cay"	"geary, david; horstmann, cay"

从表 1 中很容易就能看出 NEWACCU 算法的结果准确率要高于 ACCU 算法的准确率. 当数据集数据增加时,我们发现每种算法的准确率会随着数据值的数量增加而改变,并呈现递增的趋势,也就是

说，数据值的数量越多，真值查找的准确率就越高. 当数据值的数量增加到 110 个的时候，NEWACCU 算法、ACCU 算法、BENE 算法和 MAL 算法的准确率变得很接近，都达到了最大值，并接近于 1，如图 3 所示.

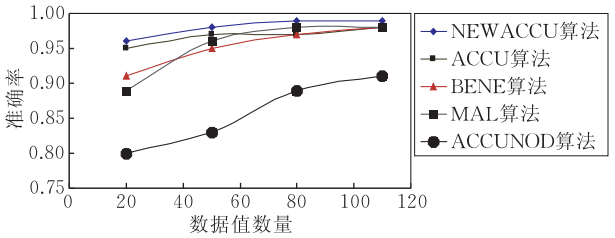

图 3　数据值的数量对准确度的影响

图 4 给出了数据集中数据源数量的变化对数据值准确率的影响情况. 数据量增加时各算法模型的准确率都呈现递增的变化趋势. BENE 算法是将所有数据源都认为是善意的，即不承认恶意数据源的存在. 这样，当数据集中恶意的数据源如果很多的话，就会影响到 BENE 算法判断正确数据值的准确率. 由于 NEWACCU 算法中加入并增强了计算数据源的可信度环节，大大降低了 BENE 算法中出现的问题. 同样在 MAL 算法中由于过多地考虑恶意数据源对数据值的影响，而错把那些善意的数据源认为是恶意的数据源，即使这些数据源是将错误的数据值更正为正确的数据值. 因此这也影响到了 MAL 算法的运行结果，使得在计算数据值准确率时不能有很高的准确率. ACCUNOD 算法中只考虑了数据源的可靠性，假设数据源是独立存在的，没有考虑数据源之间还存在复制的关系，但是实际中这种理想的独立关系是不存在的，所以大量的数据源进行复制时就严重影响了真值判断的准确率. ACCU 算法虽然也考虑了数据源的依赖关系和可靠性，但是在性能上不如 NEWACCU 算法，因为 NEWACCU 算法不但利用数据值的选票值来计算数据源的可信度，而且还在数据值的可信度计算上将数据值的选票值和正确概率进行均值计算，目的就是使得结果更加符合真实情况，实验证明我们的假设是成立的.

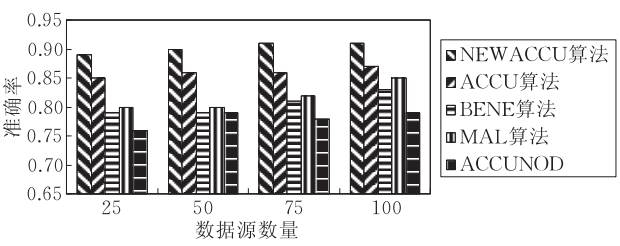

图 4　数据源数量对准确率的影响

我们知道影响某一数据值 v 准确率的因素有两个，一个是数据源的可信度，另一个是独立数据源的个数. 在选票计算中如果一个数据值被多个独立数据源所提供的话，那么它的选票值就会相对较高，在计算数据值的可信度时，如果提供该值的数据源的可信度很高，那么这个数据值的可信度也会相对较高.

通过实验我们观察到当数据源为 100 个的时候各种方法的准确率如表 2 所示.

表 2　准确率对照表

算法模型	准确率
ACCUNOD	0.79
BENE	0.83
MAL	0.85
ACCU	0.87
NEWACCU	0.91

本文提出的 NEWACCU 算法的实验结果准确率比 ACCU 算法提高了 4%多，当数据源数量增大时，准确率还会更高一些.

4.3　MovieRunTime 数据集上的实验

这个数据集记录了 500 部电影，由于本数据集的查询内容是数值型数据，所以在数据处理中要比第 1 个数据集容易一些.

利用 BENE 算法、MAL 算法、ACCUNOD 算法、ACCU 算法和 NEWACCU 算法分别进行实验，查找电影的播放时长. 表 3 描述了实验结果.

表 3　MovieRunTime 数据集实验结果对照表

(a)

电影	BENE 查找时长/min	MAL 查找时长/min	ACCUNOD 查找时长/min
12 angry men	52	52	52
2001：a space odyssey	105	148	148
24 hour party people	115	115	95
42nd street	90	90	176
a beautiful mind	136	134	134
a christmas story	97	97	97
a clockwork orange	137	137	137
a fistful of dollars	101	101	101
a fistful of dynamite	138	138	138

(b)

电影	ACCU 查找时长/min	NEWACCU 查找时长/min	真实结果/min
12 angry men	52	52	52
2001：a space odyssey	105	105	105
24 hour party people	115	117	117
42nd street	90	90	90
a beautiful mind	134	134	134
a christmas story	97	97	97
a clockwork orange	137	137	137
a fistful of dollars	101	101	101
a fistful of dynamite	138	138	138

通过上面的实验结果可以看出，本文提出的NEWACCU算法的准确率要高于其它算法的准确率.

随着数据源数量的增加，错误值的变化趋势如图5所示.从图中可以看出当数据源的数量增加时，ACCU算法和NEWACCU算法的错误值数量都在减少，说明当数据源的数量增加时，数据源之间的依赖关系体现得更加明显，准确刻画数据源依赖关系是提高数据值可信度计算的必要前提.当数据源的数量增加时，ACCUNOD算法的错误率反而增多，因为ACCUNOD算法没有考虑数据源的依赖关系，重复投票的比例变大，所以导致当数据源数量增加时，使得ACCUNOD算法计算出的结果错误率增加.

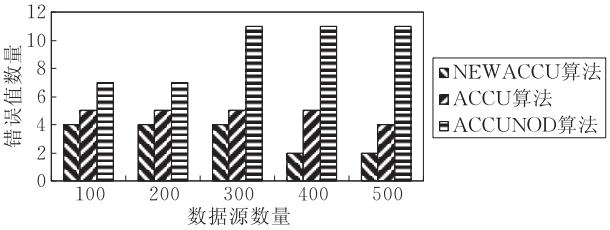

图5 错误值数量与数据源数量的关系

从图5中还可以很容易的看出来，NEWACCU算法的错误值数量总是要小于ACCU算法和ACCUNOD算法，因为NEWACCU算法中计算数据源的可信度时是利用数据值的选票值进行计算的，这样在数据源的可信度计算中考虑了数据源之间的复制依赖关系，更能够体现出每个数据源的可信度，从而提高了结果集的准确率.由于ACCU算法的性能要比BENE算法、MAL算法的好，所以这里只利用了ACCU算法、ACCUNOD算法和NEWACCU算法的实验结果进行了比较.

随着数据对象数量的增加，错误值的变化趋势如图6所示.从图中可以看到ACCUNOD算法的错误值的数量一直都很高，原因是当数据对象的数量增加时，数据源之间的复制依赖关系也会越明显，由于数据源复制行为并不是一直都存在的，复制源在某一时刻也可以作为独立数据源存在，提供自己的数据信息，或者更改错误的数据值.在这种情况下理想的独立数据源空间是不存在的，ACCUNOD算法是没有办法准确地判断出正确信息的.所以还要考虑到数据源之间的复制关系，ACCU算法就能够较好地解决这一问题.ACCU算法是BENE算法和MAL算法的改进，它是基于数据源依赖关系求解数据值的可信度，通过计算数据源的可信度来重新调整数据值的可信度.NEWACCU算法在处理数据值可信度时要优于ACCU算法，不仅在计算数据源可信度时提高了性能，而且还考虑到了数据值的相似度问题，这也是NEWACCU算法在运行结果中准确率高于其它算法的原因.

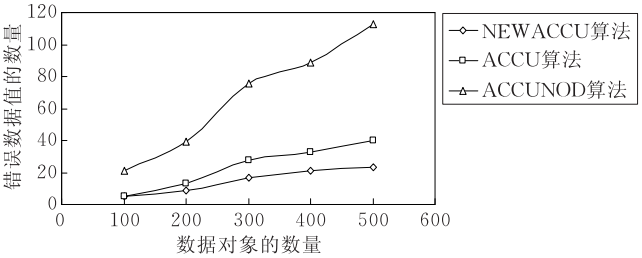

图6 错误值数量与数据对象数量的关系

表4是当数据对象为100个的时候各种算法的准确率对照表.

表4 准确率对照表

算法模型	准确率
ACCUNOD	0.86
BENE	0.90
MAL	0.89
ACCU	0.93
NEWACCU	0.95

NEWACCU算法的实验结果准确率比ACCU算法提高了2%多，当数据量增大的时候，准确率还会更高一些.

4.4 小 结

本节介绍了实验环境并深入地分析了实验结果，通过几种算法结果的对比，我们分析了问题出现的原因，进而提出更有效的解决办法来提高算法的性能.其中，用两个数据集分别进行实验，在第1个数据集实验中，我们通过调节数据源和某一个数据对象所对应的数据值的数量，观察了各个算法模型的实验结果变化，从中我们发现所有算法的实验结果准确率都会随着数据值或者数据源的数量增加而增加；在第2个数据集实验中，我们也得到了相类似的结论，即随着数据源或数据对象数量的增加，所有算法的准确率都增加了.通过实验我们了解到当数据集的数据量增大时，数据源之间的依赖关系体现得越明显，BENE算法、MAL算法、ACCU算法和NEWACCU算法的计算结果就越准确.由于NEWACCU算法中考虑了数据值之间的相似性判定问题，而且在计算数据源的可信度时与ACCU算法有所不同，在计算数据源的可信度时利用了每个数据值的选票值进行计算，通过这两点的改进使得

我们的算法 NEWACCU 的准确率高于 ACCU 的准确率.

5 结　论

本文的主要工作是改进了 ACCU 算法，加入了数据值之间的相似度问题处理，并且改进了数据源的可信度计算. 在计算数据源的可信度时利用数据值的选票值来计算，数据值的选票值是基于数据源的依赖关系计算的，所以利用数据值的选票值计算数据源的可信度更能体现数据源的特征. 实验结果表明改进后的算法确实提高了结果的精度，在众多冲突数据中可以更好地找到正确的数据值. 后续工作将继续扩展数据集的类型以及建立一个更全面的真值发现方法评价框架，进一步深化相关的研究工作.

致　谢　本实验所用的数据集由 Dong Xin Luna 和 Yin Xiaoxin 两位博士提供，在此感谢她们的指导和答疑解惑！

参　考　文　献

[1] Dayal U. Processing queries over generalization hierarchies in a multidatabase system//Proceedings of the VLDB. Florence, Italy, 1983: 342-353

[2] Papakonstantinou Y, Abiteboul S, Garcia- Molina H. Object fusion in mediator systems//Proceedings of the VLDB. Bombay, India, 1996: 413-424

[3] Motro A, Anokhin P, Acar A C. Utility-based resolution of data inconsistencies//Proceedings of the IQIS Workshop. Paris, France, 2004: 35-43

[4] Schallehn E, Sattler K-U, Saake G. Efficient similarity-based operations for data integration. Data and Knowledge Engineering, 2004, 48(3): 361-387

[5] Bleiholder J, Naumann F. Conflict handling strategies in an integrated information system//Proceedings of the International Workshop on Information Integration on the Web. Edinburgh, UK, 2006: 36-41

[6] Cho J, Garcia-Molina H. Synchronizing a database to improve freshness//Proceedings of the SIGMOD. Dallas, Texas, USA, 2000: 1-30

[7] Yin X, Han J, Yu P S. Truth discovery with multiple conflicting information providers on the web//Proceedings of the SIGKDD. San Jose, California, USA, 2007: 1-5

[8] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117

[9] Berti-Equille L, Sarma A D, Dong X L, Marian A, Srivastava D. Sailling the information ocean with awareness of currents: Discovery and application of source dependence//Proceedings of the CIDR. Aoilomar, CA, USA, 2009: 1-6

[10] Dong X L, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence. PVLDB, 2009, 2(1): 550-561

[11] Dong X L, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world. PVLDB, 2009, 2(1): 562-573

[12] Dong X L, Naumann F. Data fusion-resolving data conflicts for integration. PVLDB, 2009, 2(2): 1654-1655

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

[14] Liu Qun, Li Sujian. Word similarity computing based on how-net. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76

ZHANG Zhi-Qiang, born in 1973, Ph. D., professor. His main research interests include information retrieval, database and intelligent information processing.

LIU Li-Xia, born in 1986, M. S. assistant professor. Her current research interest is information retrieval.

XIE Xiao-Qin, born in 1973, Ph. D., associate professor. Her current research interests include service-oriented computing, knowledge engineering, social network, intelligent information processing.

PAN Hai-Wei, born in 1974, Ph. D., associate professor. His current research interests include data mining, intelligent information processing.

FANG Yi-Xiang, born in 1988, M. S.. His current research interests include information retrieval and data mining.

Background

With advanced network technology, more and more sources are available either over the Internet or in enterprise intranets. Modern information management applications often require integrating data from a variety of data sources, some of which may copy or buy data from other sources. Sources often provide out-of-date data. Errors can also creep into data when sources are updated often. Given out-of-date and erroneous data provided by different, possibly dependent, sources, it is challenging for data integration systems to provide the true values.

Now, the popular approach considers dependence between data sources in truth discovery, accuracy of data sources and similarity between values. Intuitively, if two data sources provide a large number of common values and many of these values are rarely provided by other sources (e.g., particular false values), it is very likely that one copies from the other.

In this paper, we study the problem of finding true values and determining the copying relationship between sources, when the update history of the sources is known. And we improved the existing algorithm with using a new voting algorithm and considering the diverse expressions of the same value (e.g., person's name), and found it is the best effective for discovering true values among conflicting values.

This paper is supported by the National Natural Science Foundation of China under Grant Nos. 60803037, 61202090, 61272184, the Program for New Century Excellent Talents in University No. NCET-11-0829, the Natural Science Foundation of Heilongjiang Province under Grant No. F201130, F201016, the Science and Technology Innovation Talents Special Fund of Harbin under Grant No. RC2010QN010024, and the Fundamental Research Funds for the Central Universities under Grant Nos. HEUCFZ1010, HEUCFT1202.