不确定数据的重复记录检测

邓慧挺,毛宇光

(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

摘 要:随着不确定数据成为研究的热点,不确定数据管理吸引了研究者的极大兴趣。目前业界已经使用概率数据库来存储和管理不确定数据。为合并多个自治概率数据库中的数据,需要对不确定数据进行集成。现有对数据集成的研究主要集中于对确定数据(关系型数据和半结构化数据)的研究,对不确定性数据的集成没有相关工作。重复记录检测是集成过程中必要和具有代表性的组成部分,文中讨论了重复检测的基础,研究了有依赖和无依赖的不确定数据重复检测,最后提出了两个不确定数据重复记录检测的模型。

关键词:不确定数据;重复记录;数据整合;比较向量;决策模型

中图分类号:TP311 **文献标识码**: A **文章编号**:1673-629X(2012)08-0060-03

Duplicate Record Detection of Uncertain Data

DENG Hui-ting, MAO Yu-guang

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: As uncertain becomes a hot research, the management of uncertain data has attracted tremendous interest from research. Probabilistic databases have been proposed to manage uncertain data. In order to combine data from multiple autonomous probabilistic databases, an integration of probabilistic data has to be performed. Existing approaches have focused on the integration of certain source data (relational and semi-structure). There is no related work on the uncertain data integration. Duplicate detection is an essential and representative component. In this paper, discuss the foundation of duplicate detection. Then study duplicate detection of uncertain data with (without) dependency. At last, present two models of duplicate record detection of uncertain data.

Key words: uncertain data; duplicate record; data integration; comparison vector; decision model

0 引 言

在大量的应用领域中(如传感器网络[1]、地理信息系统[2]和数据挖掘[3]等),存储和管理不确定数据的需求与日俱增。传统确定的数据模型已不能满足概率性数据的表示和呈现要求。为此,研究者提出了许多概率数据模型[4,5]和概率数据库原型。

为了联合管理多个自治的概率数据库,对不确定数据库的集成非常必要。数据集成过程主要包括四个步骤[6]:

(1)模式匹配;

(2)模式映射;

(3)重复检测;

(4)数据融合。

文中研究如何在概率数据的重复检测中利用已有技术,提出了不确定数据的重复检测模型。

1 相关工作

重复记录检测是识别同一现实世界实体的过程。文献[7]中提出了一种构造聚类树和增设阈值的相似检测方法。文献[8,9]提出了大数据量的相似记录检测。文献[10,11]分别将内码序值和遗传神经网络用于重复记录检测。文献[12]中提出一种基于信息熵的半监督分类方法。文献[13,14]使用一种半结构化的概率模型以处理 XML 数据检测中的不确定性。

2 重复检测的基础

识别同一现实世界实体是重复记录检测的目标。由于数据采集、建模和管理中的缺陷,收集的数据经常是不正确或不完整的。因此,需要设计有效地处理各种数据的重复检测技术。通常,重复记录检测需要五

收稿日期:2011-12-10;**修回日期**:2012-03-12

基金项目:国家自然科学基金(60873025)

作者简介:邓慧挺(1986-),男,福建永安人,硕士研究生,CCF 会员,主要研究领域为不确定数据库与数据仓库;毛宇光,副教授,主要研究领域为数据库系统及理论、数据挖掘与数据仓库、特种数据库、多值逻辑及其应用。

个步骤[15]:

2.1 数据准备

数据是标准化的(如有统一的规范和约定)和干净的(清除明显的错误数据)以获得所有源数据的相同表现。

2.2 缩减搜索空间

将元组间的所有组合进行比较需要指数的时间复杂度,往往用一些启发式算法(贪心算法,最近邻居算法等)缩小搜索空间。

2.3 属性值匹配

元组相似度可以从属性的相似度得出。由于数据准备、语法和语义的不规范,可以用语法的方法(如编辑距离、n-gram 算法和 Jaro 距离等)和语义的方法(如本体论和专业词典等)得到属性值相似度。论文采用比较向量 $c=[C_1,\cdots,C_n]$,其中 C_i 代表第 i 个属性的值的相似度。

2.4 决策模型

决策模型以比较向量为输入,决定一个元组对 (T_1,T_2) 的归类:匹配元组(M)、不匹配元组(U)或可能匹配元组(P)。决策结果存储在结果 $\mathfrak{L}(T_1,T_2)\in\{m,p,u\}$ 中,其中 m 代表 (T_1,T_2) 归为 M 类的情况,p 和 u 类似。决策模型通常基于领域知识技术和概率性原理的方法。

2.4.1 基于领域知识技术

基于知识的重复检测技术[15]中,领域专家定义识别规则,规则指定元组以某个确定因数作为判断元组重复的条件。根据确定因数与临界值的比较结果,决定元组对是否以某个概率为重复记录。

2.4.2 概率性的技术

元组对 (T_1,T_2) 的两个条件概率定义如下[15]:

$$\begin{cases} m(c)=p(c(T_1,T_2)\in M \\ m(c)=p(c(T_1,T_2)\in U \end{cases} \tag{1}$$

基于匹配权重 $T=m(c)/u(c)$ 和临界值 μ 和 λ。当 $T>\mu$ 时,(T_1,T_2) 为匹配元组对;当 $T<\lambda$ 时,(T_1,T_2) 为不匹配元组;否则,为可能匹配元组对,需进一步对其检验。元组对 (T_1,T_2) 匹配过程的两个步骤为:

第一步:用联合函数 $\mathrm{sim}(T_1,T_2)=w(c)$,$w:[0,1]^n\to\mathbb{R}$ 决定元组的相似度。而且,当使用基于领域知识的技术(确定向量)时,产生的相似度是标准化的;当使用基于概率的技术(匹配权重)时,产生的相似度是非标准化的。

第二步:在 $\mathrm{sim}(T_1,T_2)$ 的基础上使用两个临界值将元组对归类到 M、P 或 U。

2.5 确认和验证

通常用准确率、召回率、伪阴性和伪阳性等验证重复记录检测方法的效果[15]。若对识别效果不满意,可用其它临界值和方法再次进行检测。

3 概率性数据中的重复检测

概率数据库的形式化定义为 $\mathrm{PDB}=(W,P)$,其中 $W=\{I_1,\cdots,I_n\}$ 是可能世界集合,P 是所有可能世界的概率分布且其概率和为 1。可能世界的组合非常多,通常无法存储所有可能世界,需要对可能世界模型做精简的描述。在概率关系模型中,有两个层次上的不确定性:元组层次上的存在性概率和属性值层次上表示候选属性值的概率。

一个关系中的元组成员往往来自应用的上下文。例如,一个人可以被存储在两个不同的关系中:一个存储未成年人,另一个存储有职业的人员。考虑一个 13 岁并且有 10% 的概率有工作的人,该元组属于第一个关系的概率为 $P(T_1)=1.0$,属于第二个关系的概率为 $P(T_2)=0.1$。这表明应根据属性值层次的不确定性进行重复检测。

3.1 无依赖的重复检测

考虑表 1 的概率关系 R,R 包含了元组层次的不确定性和属性值层次上的不确定性。注意到元组 T_{11} 表示的人有 10% 概率没有职业,论文用“Λ”来指示这种不存在性。属性间不存在依赖性,仍可以检测出基于属性到属性的相似性。

显然,一个不存在的值和任何存在的值均不相似。则有:

$$\mathrm{sim}(\Lambda,\Lambda)=1,\mathrm{sim}(a,\Lambda)=\mathrm{sim}(\Lambda,a)=0,(a\neq\Lambda) \tag{2}$$

文中对无错误数据和有错误数据相似度的形式化定义如式(3)和式(4)所示:

$$\mathrm{sim}(a_1,a_2)=\sum_{d\in D}P(a_1=d,a_2=d) \tag{3}$$

$$\mathrm{sim}(a_1,a_2)=\sum_{d_1\in D}\sum_{d_2\in D}P(a_1=d_1,a_2=d_2)\cdot\mathrm{sim}(d_1,d_2) \tag{4}$$

由于处理属性值层次上的不确定性,比较向量 c 上的匹配结果不变。因此,可以直接使用已有的决策模型。

表 1　概率关系 R

元组	姓名	职业	概率
T_{11}	张三	{教师:0.8, 商人:0.1}	1.0
T_{12}	{李四:0.5, 王五:0.5}	{医生:0.8, 作家 0.2}	1.0
T_{13}	{张三:0.7, 张思:0.3}	教师:0.2	0.8

3.2 x-tuples 模型的重复检测

为建立属性值间的依赖性模型,Trio 的 ULDB 模型引入了 x-tuples(x 元组)概念。一个 x 元组 T 由一个或多个互斥的候选元组 $(t^1,\cdots,t^n)$ 组成。ULDB 模型不支持无限的候选元组(比如在连续域上的不确定

性)。为了避免海量的候选值,往往把属性值与概率分布相关联。比如"南京"代表所有以字符串"南京"开头的所有名称(比如"南京火车站")。包含一个或多个 x 元组的关系称为 x 关系。

文中用 x 元组 $t_1=\{t_1^1,\cdots,t_1^k\}$,$t_2=\{t_2^1,\cdots,t_2^l\}$ 候选元组的相似性计算元组的相似性。在属性值匹配阶段,t_1 的所有可能元组的属性值要和 t_2 的所有可能元组的属性值进行两两对比。用3.1节中的公式处理单个属性值的确定性。因此,需要一个 $k\times l$ 维的比较向量和将元组对 (T_1,T_2) 分类到 M、P 或 U 集合中的决策模型。

3.3 两个 x-元组决策模型

论文提出了两种决策模型,输入为 x 元组对 (T_1,T_2)、一个由候选元组对的比较向量组成的比较矩阵、属性优先权重向量 AW 和匹配临界值 UT。

模型一:

输入:x - 元组 $(T_1=\{T_1^1,\cdots,T_1^k\},T_2=\{T_2^1,\cdots,T_2^l\})$

比较矩阵 $(\boldsymbol{c}(T_1,T_2)=[c_{11},\cdots,c_{kl}])$

AW 向量 $\boldsymbol{a}=\{at_1,\cdots,at_n\}$,临界值 UT

1. 对每个候选元组对 (T_1^i,T_2^j) 的比较向量 $\boldsymbol{c}_{ij}$ 执行联合函数 $\Gamma(at_1,\boldsymbol{c}_{ij})$

=> 结果:$\mathrm{sim}(T_1^i,T_2^j)$

2. $mVal=\varphi((mVal(T_1^{i-1},T_2^{j-1}),(T_1^i,T_2^j)))$

若 $mVal>UT$,则将 (T_1,T_2) 分类到 $\{M\}$ 并退出,否则转步骤1

3. 基于 $\mathrm{sim}(T_1,T_2)$ 分类 (T_1,T_2) 到 $\{M,P,U\}$

输出:(T_1,T_2) 是否为重复记录。

模型二:

输入:x - 元组 $(T_1=\{T_1^1,\cdots,T_1^k\},T_2=\{T_2^1,\cdots,T_2^l\})$

比较矩阵 $(\boldsymbol{c}(T_1,T_2)=[c_{11},\cdots,c_{kl}])$

AW 向量 $\boldsymbol{a}=\{at_1,\cdots,at_n\}$,临界值 AT

1. 对每个候选元组对 (T_1^i,T_2^j) 的比较向量 $\boldsymbol{c}_{ij}$ 执行联合函数 $\Gamma(at_1,\boldsymbol{c}_{ij})$

=> 结果:$\mathrm{sim}(T_1^i,T_2^j)$

2. 基于 $\mathrm{sim}(t_1^i,t_2^j)$ 将 (t_1^i,t_2^j) 分类到 $\{M,P,U\}$

$\Rightarrow$ 结果:匹配度 $\eta(t_1^i,t_2^j)\in\{m,p,u\}$

$\Rightarrow mVal=\varphi(mVal(\eta(t_1^{i-1},t_2^{j-1}),\eta(t_1^i,t_2^j)))$

若 $mVal>UT$,则为重复记录,将 (t_1,t_2) 分类到 $\{M\}$ 并退出,否则转步骤1

3. 基于 $mVal$ 将 (t_1,t_2) 分类到 $\{P,U\}$

输出:(t_1,t_2) 是否为重复记录。

模型一中的 $\mathrm{matchVal}(\mathrm{sim}(t_1^{i-1},t_2^{j-1}))$ 为已经得出相似性的所有候选元组的相似性值,其中 $\varphi(mVal(\mathrm{sim}(t_1^{i-1},t_2^{j-1}),\mathrm{sim}(t_1^i,t_2^j)))$ 为进行 (t_1^i,t_2^j) 之后的相似性函数,若 $mVal$ 大于定义的匹配临界值,则为重复记录,不需进一步计算,否则,执行完所有候选元组相似性计算后,根据最终的 $mVal$ 分类到 $\{M,U\}$ 中。第三个步骤是将 (t_1,t_2) 对归类到 $\{P,U\}$,因为如果属于重复记录,(t_1^i,t_2^j) 已经在第二个步骤分类到 $\{M\}$ 中。注意到第三个步骤给出了归类情况和匹配值,以方便进一步分析。

改进的模型二中基于 $\mathrm{sim}(t_1^i,t_2^j)$ 将 (t_1^i,t_2^j) 分类到 $\{M,P,U\}$ 中,进而对已计算的候选元组对计算 $mVal=\varphi(mVal(\eta(t_1^{i-1},t_2^{j-1}),\eta(t_1^i,t_2^j)))$,此时的 $mVal(\eta(t_1^{i-1},t_2^{j-1}))$ 函数可设计为 (t_1^i,t_2^j) 的归类为 $\{M\}$ 的概率密度函数,第三个步骤和模型一类似,只需将 (t_1,t_2) 分类到 $\{P,U\}$ 中并给出匹配值。步骤3中的导出函数还可以基于概率性原理,比如通过将元组的相似性定义为匹配权重。两个 x 元组的相似性基于定义在离散域 $\{m,p,u\}$ 中的值,因此 x 元组的相似性比第一种模型更精确。尽管步骤2中的结果未标准化,但可避免产生无法描述元组相似度的情况。

4 结束语

论文研究了不确定性数据的重复检测。给出了属性值和元组相似度的计算方法,提出了无依赖和有依赖的不确定数据重复记录检测模型。其中第一种模型适合基于知识的技术,第二种模型适合概率技术。在未来工作中,将进一步研究不确定数据重复记录检测的缩小搜索空间、模式匹配、模式映射和数据融合的实现等技术。

参考文献:

[1] Deshpande A, Guestrin C, Madden S, et al. Model-driven Data Acquisition in Sensor Networks[C]//Proceedings of the 30th VLDB Conference. Toronto: [s. n.], 2004: 588-599.

[2] 周迪民,段国云. 地理信息系统属性数据不确定性的研究[J]. 计算机技术与发展,2009,19(12):174-177.

[3] 邓玮舛,余永权. 数据挖掘中粗糙决策规则及其不确定性研究[J]. 计算机技术与发展,2008,18(8):50-57.

[4] Barbara D, Garcia-Molina H, Porter D. The Management of Probabilistic Data[J]. IEEE Transactions on Knowledge and Data Engineering, 1992, 4(5): 487-502.

[5] Keulen M, Keijzer A, Alink W. A Probabilistic XML Approach to Data Integration[C]//Proceedings of the 21st International Conference on Data Engineering. [s. l.]: [s. n.], 2005: 459-470.

[6] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate Record Detection: A Survey[J]. IEEE Transactions on Knowl-

(下转第66页)

```
($_POST['pwd']); //密码
    }
    }
    ?>
```

在loginController类里定义一个Action方法loginAction(),那么控制器会指定这个Action方法,选择同名的View文件(login.phtml)输出。在loginAction()里采用$_POST[]方法获取来自method="post"的表单中提交的用户名和密码,并且调用Model里的方法loginManage($params),若数据正确,则通过$this->_redirect('/Index/main')这句代码转向main.phtml页面,否则提示登录失败。

3.2.3 模型的实现

在MVC设计模式中,Model负责对数据库的操作,并处理控制器Controller传来的数据请求以及当数据发生改变时将这些变化告知视图,然后视图会做出相应的调整。在基于Zend Framework框架的Web OA系统中,模型类文件存放在\application\models目录下。Model获取从控制器Controller传来的用户名和密码等数据,在模型文件中定义一个方法,此方法调用存储过程P_System_login,获取数据库中的数据经控制器的调用将结果返回给相关的视图(View)输出给客户端。主要实现代码如下:

```
$db=get_db();
$params = array(
array(&$username, SQLSRV_PARAM_IN),
array(&$password, SQLSRV_PARAM_IN),
)
$results=$db->queryporc('P_System_login',
$params); //调用存储过程
return $results;
```

4 结束语

文中对于MVC设计模式和Zend Framework以及系统的设计与实现进行了详细的阐述。MVC设计模式不仅能够适宜的把整体分成局部,同时利用三层架构的思想以及低耦合性的结构使得开发人员各司其职,而且简化了应用程序的复杂性,提高了系统的可扩展性。Zend Framework的模块化的结构设计和丰富的组件使得程序更易扩展和灵活,系统也更加稳定。实践表明,采用MVC模式开发出来的系统结构更加清晰,性能更加稳定,并且易于管理。

参考文献:

[1] 王映辉,王英杰,王彦君,等.基于MVC的软件界面体系结构研究与实现[J].计算机应用研究,2004(9):188-199.

[2] 孙卫琴.精通Struts:基于MVC的JavaWeb设计与开发[M].北京:电子工业出版社,2004.

[3] 刘春花,王忠民.基于MVC模式的远程评议系统的设计与实现[J].计算机工程与设计,2008,29(13):3648.

[4] 王晓楠. MVC的设计和实现[J].计算机系统应用,2004(3):56-58.

[5] 陈名箴.基于Zend框架的项目管理系统设计与实现[D].杭州:浙江大学,2010.

[6] Evans C. PHP Architects Guide to Programming with Zend Framework[M].[s.L.]:Marco Tabini & Associates,2008.

[7] 陈营辉,赵 伟,赵海波,等. Zend Framework技术大全[M].北京:化学工业出版社,2010.

[8] Rob A,Nick L,Steven B. Zend Framework in Action[M]. United States of America: Manning Publications,2008.

[9] 赵新燕.山东省青干院学工信息管理系统的设计与实现[D].济南:山东大学,2010.

[10] 张朝阳,熊淑华,衡 丽.基于Zend Framework的网站设计与实现[J].计算机技术与发展,2011,21(11):199-200.

[11] Flanagan D. jQuery Pocket Reference[M].[s.L.]:O'Reilly,2010.

[12] 王 锋,魏晓丽,江开耀.基于XML的C#多语言界面实现[J].计算机工程与设计,2008,29(15):4073-4074.

(上接第62页)

edge and Data Engineering,2007,19(1):1-16.

[7] 戴 颖,李兴国,赵启飞.一种相似重复记录检测算法的改进研究[J].计算机技术与发展,2010,20(7):13-16.

[8] 韩京宇,徐立臻,董逸生.一种大数据量的相似记录检测方法[J].计算机研究与发展,2005,42(12):206-212.

[9] 庞雄文,姚占林,李拥军.大数据量的高效重复记录检测方法[J].华中科技大学学报,2010,38(2):8-11.

[10] 鲁均云,李星毅,施化吉,等.基于内码序值聚类的相似重复记录检测方法[J].计算机应用研究,2010,27(3):874-878.

[11] 孟祥逢,鲁汉榕,郭 玲.基于遗传神经网络的相似重复记录检测方法[J].计算机工程与设计,2010,31(7):1550-1553.

[12] 陈锦禾,沈 洁.基于信息熵的主动学习半监督分类研究[J].计算机技术与发展,2010,20(2):110-113.

[13] 陈 伟,丁秋林.一种XML相似重复数据的清理方法研究[J].北京航空航天大学学报,2004,30(9):835-838.

[14] Keulen M, Keijzer A. Qualitative effects of knowledge rules and user feedback in probabilistic data integration[J]. VLDB journal,2009,18(5):1191-1217.

[15] Data Quality:Concepts,Methodologies and Techniques(Data-centric Systems and Applications)[M].[s.l.]:[s.n.],2006.

PCI-E 接口驱动软件设计与实现

作者： 刘娟， 田泽， 黎小玉
作者单位： 中国航空计算技术研究所,陕西西安710119
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(8)

参考文献(15条)

1. Deshpande A;Guestrin C;Madden S Model-driven Data Acquisition in Sensor Networks 2004
2. 周迪民;段国云 地理信息系统属性数据不确定性的研究[期刊论文]-计算机技术与发展 2009(12)
3. 邓玮舛;余永权 数据挖掘中粗糙决策规则及其不确定性研究[期刊论文]-计算机技术与发展 2008(08)
4. Barbara D;Garcia-Molina H;Porter D The Management of Probabilistic Data[外文期刊] 1992(05)
5. Keulen M;Keijzer A;Alink W A Probabilistic XML Approach to Data Integration 2005
6. Elmagarmid A K;Ipeirotis P G;Verykios V S Duplicate Record Detection:A Survey[外文期刊] 2007(01)
7. 戴颖;李兴国;赵启飞 一种相似重复记录检测算法的改进研究[期刊论文]-计算机技术与发展 2010(07)
8. 韩京宇;徐立臻;董逸生 一种大数据量的相似记录检测方法[期刊论文]-计算机研究与发展 2005(12)
9. 庞雄文;姚占林;李拥军 大数据量的高效重复记录检测方法[期刊论文]-华中科技大学学报 2010(02)
10. 鲁均云;李星毅;施化吉 基于内码序值聚类的相似重复记录检测方法[期刊论文]-计算机应用研究 2010(03)
11. 孟祥逵;鲁汉榕;郭玲 基于遗传神经网络的相似重复记录检测方法[期刊论文]-计算机工程与设计 2010(07)
12. 陈锦禾;沈洁 基于信息熵的主动学习半监督分类研究[期刊论文]-计算机技术与发展 2010(02)
13. 陈伟;丁秋林 一种XML相似重复数据的清理方法研究[期刊论文]-北京航空航天大学学报 2004(09)
14. Keulen M;Keijzer A Qualitative effects of knowledge rules and user feedback in probabilistic data integration 2009(05)
15. Data Quality:Concepts,Methodologies and Techniques(Datacentric Systems and Applications) 2006