数据质量评估方法研究

杨青云[1] 赵培英[2] 杨冬青[1] 唐世渭[1] 童云海[1]
[1](北京大学信息科学技术学院,北京 100871)
[2](山东轻工业学院计算机系,济南 250100)
E-mail:qingyun@db.pku.edu.cn

摘 要 数据质量管理已经成为当今数据管理的关键问题,并得到了广泛的研究和应用。数据质量评估作为数据质量管理中的必要过程和基础部分,目前缺乏一种定量的系统的方法。针对数据质量评估中的这一问题,该文介绍了一些基本的数据质量评估指标,提出了一种数据质量评估模型,并阐述了该模型的构造技术和计算方法。

关键词 数据质量 数据质量评估 数据质量评估模型

文章编号 1002-8331-(2004)09-0003-02 文献标识码 A 中图分类号 TP311

Research on Data Quality Assessment Methodology

Yang Qingyun[1] Zhao Peiying[2] Yang Dongqing[1] Tang Shiwei[1] Tong Yunhai[1]
[1](School of Information Science and Technology,Peking University,Beijing 100871)
[2](Computer Department,Shandong Institute of Light Industry,Jinan 250100)

Abstract: Data quality management,which is widely researched and applied,has become the key problem of data management.However,data quality assessment,the necessary process and elementary component of data quality management,is short of a quantitative and systematic method at present.To solve this problem,this paper introduces some foundational data quality assessment indicators,proposes a data quality assessment model and describes the constructing technology and computing method of the model.

Keywords: Data quality,Data quality assessment,Data quality assessment model

1 引言

在现代社会,数据是企业走向信息化的必要基础,然而随着企业应用系统数据量的急剧扩大、新应用的不断出现以及应用之间的相互整合,数据质量问题变得日益突出,这些问题主要表现在数据不正确、数据不完整、数据不一致等方面。质量低劣的数据已经成为影响企业进行正确决策的重要因素,所以数据质量管理必将成为企业进行信息化进程中一个必不可少的环节。

针对数据质量问题的各个环节,包括数据清洗、数据整合、相似记录检测、数据质量评估、数据质量过程控制和管理等方面,业界已进行了大量的学术研究和实际应用的探索。在这些环节中,数据质量评估是提高数据质量的基础和必要前提,它能对应用系统的整体或部分数据的质量状况给出一个合理的评估,从而可以帮助数据用户了解应用系统的数据质量水平,并采取相应的处理过程来提高数据质量。

对于数据质量评估,一些研究人员也进行了许多工作。文献[2,3,4,5]从不同的方面提出了数据质量的评估方法,文献[6]介绍了数据质量的评估过程,文献[7]设计了一个数据质量的分析和浏览的工具,文献[1]给出了一种基于属性的数据质量评估模型。但总的来讲,它们都没有提供一个定量的系统方法来进行数据质量的评估。该文在此基础上,提出了一个数据质量的评估模型,并阐述了该模型的构造技术和计算方法。

该文的结构安排是:第二节介绍了数据质量评估的一些基本指标,第三节提出了一个数据质量评估模型并结合具体实例阐述了其构造技术和计算方法,第四节进行总结并介绍了今后进一步的工作。

2 数据质量评估的基本指标

文献[2,3,4,5]提出了一些数据质量的评估指标。在进行数据质量评估时,要根据具体的数据质量评估需求对数据质量评估指标进行相应的取舍。但是,数据质量评估至少应该包含以下两方面的基本评估指标:

(1)数据对用户必须是可信的(believable),其中包括精确性、完整性、一致性、有效性、唯一性等指标。这些指标的具体含义:

精确性(accurate):描述数据是否与其对应的客观实体的特征相一致。

完整性(complete):描述数据是否存在缺失记录或缺失字段。

一致性(consistent):描述同一实体的同一属性的值在不同的系统或数据集中是否一致。

有效性(valid):描述数据是否满足用户定义的条件或在一定的域值范围内。

唯一性(unique):描述数据是否存在重复记录。

基金项目:国家973重点基础研究发展规划项目(编号:G1999032705)资助
作者简介:杨青云,博士研究生,研究方向为数据仓库与数据挖掘。杨冬青,教授,博士生导师,研究方向为数据库与信息系统。唐世渭,教授,博士生导师,研究方向为数据库与信息系统。

(2)数据对用户必须是可用的(Useful),其中包括时间性、稳定性等指标。这些指标的具体含义:

时间性(Timely):描述数据是当前数据还是历史数据。

稳定性(Volatile):描述数据是否是稳定的,是否在其有效期内。

3 数据质量评估模型

3.1 数据质量评估模型的基本概念

为了对应用系统的数据质量进行评估,给出了一个数据质量的评估模型,该模型是一个六元组:

$M=<D,I,R,W,E,S>$

D:需要进行评估的数据集。对于关系数据库来讲,一个数据集相当于一个表或视图。

I:数据集 D 上需要进行评估的指标,如精确性、完整性、一致性等。

R:与评估指标相对应的规则。规则可以使用规范化的自然语言或形式化语言来书写,以便于转换成程序脚本。

W:赋予规则 R 的权值(大于 0 的整数),描述了该规则在所有规则中所占的比重。

E:对规则 R 给出的期望值(介于 0 到 100 之间的实数),是在评估之前对该规则所期望得到的结果。

S:规则 R 对应的最终结果(介于 0 到 100 之间的实数),是在检测该规则后所得的结果。

数据集随应用的不同可能有不同的质量评估需求,所以,一个数据集可以对应多个质量评估模型。在一个数据质量评估模型中,一个数据集可以对应多个评估指标,一个评估指标可以对应多个规则(如图 1 所示)。

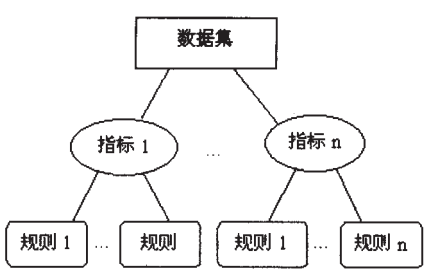


图 1 数据质量评估模型图

3.2 数据质量评估模型的构造技术

构造数据质量评估模型要经过 4 个步骤:确定数据集评估应用视图,选择评估指标,制定规则集,计算规则结果得分。下面将结合具体的实例来说明如何构造数据质量评估模型。

假设银行某应用系统有一个客户信息关系表 Customer(如表 1 所示)。

表 1 客户信息表

NO	Name	Sex	ID	City
0001	张三	M		北京
0002	李四	0	102323196602011341	北京
0003	王五	F	370102510923122	

3.2.1 确定数据集评估应用视图

在进行数据质量评估时,首先要提出数据质量评估的需求,要确定哪些数据是用户感兴趣的(包括数据库、数据库中的数据集和数据集上的字段),对它们建立对应的用户视图。如对表 Customer 要评估客户的性别和身份证的数据质量情况,则生成视图 Customer_Quality_View(NO,Sex,ID)(其关键字要保留)。

3.2.2 选择评估指标

对于每个给定的数据集,选择所需要的评估指标。对于 Customer_Quality_View,选择完整性和有效性两个指标。

3.2.3 制定规则集

根据选择的评估指标,制定数据质量评估规则,并确定它们相应的权值和期望值。对于 Customer_Quality_View,针对完整性和有效性指标制定以下规则:

(1)ID 非空(权值:5,期望值:90):完整性

(2)ID 长度为 18(权值:10,期望值:90):有效性

(3)Sex 值为'F'或'M'(权值:10,期望值:98):有效性

3.2.4 计算规则结果得分

对于规则集中的每条规则 R,检查数据集上的数据实例,计算满足 R 的数据元组的百分比,得到 R 对应的结果 S。

在这一过程中,可以为规则 R 编写相应的程序脚本或工具来执行。对于上面例子中的三条规则,可以编写其对应的 SQL 语句:

(1)select count(*)from Customer_Quality_View WHERE ID is not null

(2)select count(*)from Customer_Quality_View where length(ID)=18

(3)select count(*)from Customer_Quality_View where Sex in('F','M')

这三条 SQL 语句的结果与 Customer_Quality_View 中数据元组总数的百分比,就是最终结果。假设它们的结果分别为 95,90,90。

根据上面所定义的内容和计算的结果,形成 Customer 的数据质量评估模型(如表 2 所示)。

其中:CQV 是 Customer_Quality_View

表 2 Customer 数据质量评估模型

T	I	R	W	E	S
CQV	完整性	ID 非空	5	90	95
CQV	有效性	ID 长度为 18	10	90	90
CQV	有效性	Sex 值为'F'或'M'	10	98	90

在数据质量评估模型构造完成之后,如果需要再次评估数据质量,只需要执行第 4 步,计算每条规则的结果,然后再计算数据集的数据质量结果得分。

3.3 数据质量评估模型的计算方法

当数据质量评估模型构造完成并计算了每条规则的结果之后,便可以利用该模型来计算每个数据集的数据质量结果。

假设数据集 T 对应的规则集为 $R^T=(R_1,R_2,...R_n)$,赋予 R^T 中规则 R_i 的权值为 Wi,期望值为 E_i,R_i 计算的结果得分为 S_i,$i=1,2,...n$,由此计算数据集 T 的数据质量:

数据质量绝对量化值

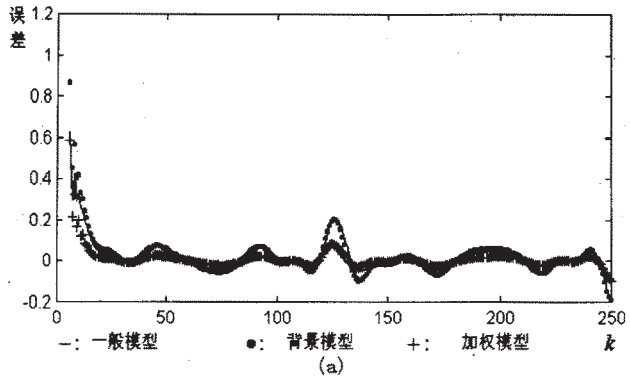
$$SA=\frac{\sum_{i=1}^{n}W_i\times S_i}{\sum_{i=1}^{n}W_i} \quad (1)$$

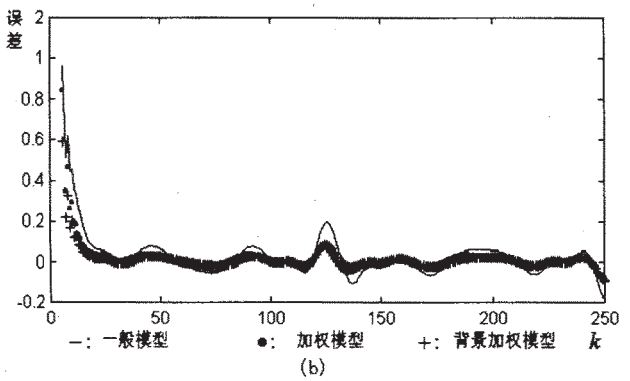
数据质量相对量化值

$$SR=SA=\frac{\sum_{i=1}^{n}W_i\times E_i}{\sum_{i=1}^{n}W_i} \quad (2)$$

SA 是规则集 R^T 所得结果得分的加权平均值,它反映了数

(下转 15 页)





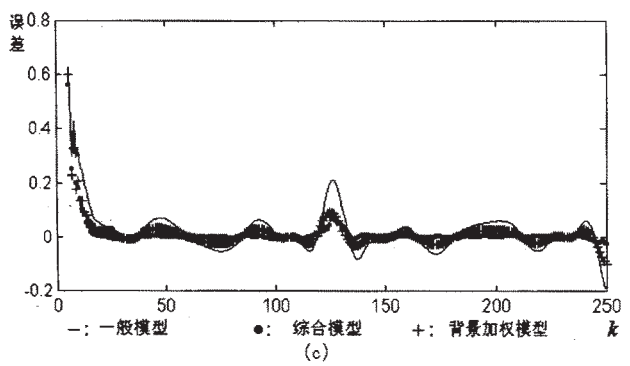


图 2　仿真结果

型)，背景加权 GM (,1)模型 即背景值构造和加权 GM (,1)模型的集成)以及该文所提出的综合 GM (,1)模型对下述差分方程进行仿真验证：

$$y(k+1)=0.3y(k)+0.6y(k-1)+g(u(k)) \quad (2)$$

其中

$$g(u(k))=0.6\sin(6\pi u(k))+0.3\sin(3\pi u(k))+0.1\sin(5\pi u(k))+0.8 \quad (3)$$

$$u(k)=\sin\left(\frac{2\pi k}{250}\right),k \text{ 为 整数} \quad (4)$$

仿真结果如图 2 (a)~ (c)所示。其中图 (a)为理论值与利用一般模型、加权模型 、背景模型的预测误差 ;图 (b)为理论值与利用一般模型、加权模型、背景加权模型的预测误差 ;图 (c)为理论值与利用一般模型、综合模型、背景加权模型的预测误差。从仿真结果可以看出，利用一般 GM (,1)模型预测算法虽然简单，但精度不高 ;利用背景模型时预测精度有所改善 ;加权模型的精度有很大提高 ;同加权模型相比较，背景加权模型略有改善 ;基于加权模型，背景模型以及神经网络补偿器的综合模型有很高精度，并且神经网络的训练过程很快。

6　结论

该文提出一种灰色-神经网络综合预测模型由背景值构造、加权 GM (,1)模型和神经网络补偿器三部分组成。仿真结果验证了所提方法的有效性，但如何更好地提高计算速度值得进一步研究。 收稿日期 :2004 年 2 月)

参考文献

1.Chyun-Shin Cheng,Yen-Tseng Hsu,Chwan-Chia.Grey Neural Network[J]. IEICE Trans Fundamentals,1998 ;E81-A (1) :2433~2442

2.Chan-Ben Lin,Shun-Feng Su,Yen-Tseng Hsu.High Precision Prediction Using Grey Models[J].Int J of Systems Science,2001 ;32 (5) :609~619

3.谭冠军.GM (,1)模型的背景值构造方法和应用 ()[J].系统工程理论与实践,2000 ;20 (4) :98~103

4.王军平.一种构建灰色预测模型的新算法[J].西北大学学报,2002 ;32 (6) :613~616

(上接 4 页)

据集 T 的真实的数据质量状况。

SR 是 SA 与期望值的差值，它反映了数据集 T 相对于其期望值的数据质量状况，若 SR 符号为正，则其数值越大，说明数据质量比预期的更好 ;若 SR 符号为负，则其数值越大，说明数据质量比预期的更差。

在上面的例子中，可以根据 Customer 的数据质量评估模型计算出 Customer 的两个数据质量量化值 (SA=91，SR=-2.2)。在进行数据质量评估的过程中，可以灵活选择这两个计算方法，从两个不同的方面来评估数据质量。

整个应用系统的数据质量同样可以使用该计算方法来计算，在此不予赘述。

4　结论和进一步工作

针对数据质量评估问题，该文介绍了一些基本的评估指标，提出了一个数据质量评估模型，并阐述了该模型的构造技术和计算方法。该模型可以通过量化的指标来对应用系统的整个或部分数据质量状况进行评估。

这一模型已在数据仓库源数据的质量评估中得到了成功的应用。笔者今后进一步的工作是将对数据质量评估模型的物化、评估规则自动化执行和数据质量分析等方面继续进行研究。 收稿日期 :2004 年 2 月)

参考文献

1.Richard Y Wang,M P Reddy,Henry B Kon.Toward Quality Data :An Attribute-based Approach[J].Decision Support System,1995 ;13 :349~372

2.Yair Wand,Rihard Y Wang.Anchoring Data Quality Dimensions in Ontological Foundations[J].COMMUNICATIONS OF THE ACM,1996 ;39 (1) :86~95

3.Richard Y Wang,Veda C Storey,Christopher P Firth.A Framework for Analysis of Data Quality Research[J].IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,1995 ;7 (4) :623~640

4.Diane M Strong,Yang W Lee,Richard Y Wang.Data Quality In Context[J].COMMUNICATIONS OF THE ACM,1997 ;40 (5) :103~110

5.Leo L Pipino,Yang W Lee,Richard Y Wang.Data Assemssment[J]. COMMUNICATIONS OF THE ACM,2002 ;45 (4) :211~218

6.Data Quality Assessment :A Methodology for Success.FirstLogic,White-paper,2003

7.Tamraparni Dasu,Theodore Johnson,S Muthukrishnan et al.Mining Database Structure ;Or,How to Build a Data Quality Browser[C].In :ACM SIGMOD 2002,2000-06 :4~6