软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(4):810−824 [doi: 10.3724/SP.J.1001.2013.04262]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种面向关系型数据的可视质量分析方法*

滕东兴[1], 曾志荣[1,2], 杨海燕[1], 王宏安[1], 戴国忠[1]

[1](中国科学院 软件研究所 人机交互技术与智能信息处理实验室,北京 100190)

[2](中国科学院 研究生院,北京 100049)

通讯作者: 滕东兴, E-mail: dongxing@iscas.ac.cn

摘 要: 由于信息系统所提供数据的质量不高(如数据残缺、数据不一致、数据重复等)导致管理者决策过程中经常面临“数据丰富,信息匮乏”的困惑是目前企业普遍存在的现象.为了切实提高信息系统所提供数据的可用性,研究了影响关系数据库数据质量的主要因素,提出了面向多数据源的统一元数据模型和数据库数据质量评估模型,构建了用于数据质量评估的交互式可视形态集.建立了一个面向关系数据库的数据质量可视分析系统,并结合具体企业应用实例进行验证.结果表明,该系统能够有效分析数据质量,提高企业分析决策的可靠性和准确性.

关键词: 信息可视化;可视分析;人机交互;数据质量

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 滕东兴,曾志荣,杨海燕,王宏安,戴国忠.一种面向关系型数据的可视质量分析方法.软件学报,2013,24(4):810−824. http://www.jos.org.cn/1000-9825/4262.htm

英文引用格式: Teng DX, Zeng ZR, Yang HY, Wang HA, Dai GZ. Visual quality analysis method for relational data. Ruanjian Xuebao/Journal of Software, 2013,24(4):810−824 (in Chinese). http://www.jos.org.cn/1000-9825/4262.htm

Visual Quality Analysis Method for Relational Data

TENG Dong-Xing[1], ZENG Zhi-Rong[1,2], YANG Hai-Yan[1], WANG Hong-An[1], DAI Guo-Zhong[1]

[1](Intelligence Engineering Laboratory, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

Corresponding author: TENG Dong-Xing, E-mail: dongxing@iscas.ac.cn

Abstract: Because of the low quality of data provided by information systems such as data missing, data conflicts and data duplicate, it is widespread in enterprise that decision-makers are often faced with “rich data but poor information”. To improve the data availability of information systems, the main factors affecting data quality of relational database are studied in this paper, also a unified metadata model based on multi data sources and a data quality assessment model are proposed, and a set of interactive visual analogues for data quality assessment is built. Finally, a visual analysis system for data quality in relational database is developed, which is verified with several enterprise practical cases. It is indicated that the built system can analyze data quality effectively, and then improve the reliability and accuracy of enterprise decision-making.

Key words: information visualization; visual analysis; human-computer interaction; database quality

利用信息系统从海量数据中分析、挖掘和存储具有决策意义的信息已成为引人注目的商业机遇之一.然而,面对业务信息系统中长期积累的大量数据,决策者却经常面临“数据丰富,信息匮乏”的窘况.究其原因,除了缺乏有针对性的数据挖掘手段以外,一个很重要的原因是业务数据质量不高,具体表现为数据残缺不全、数据不一致、数据重复等[1].以制造企业经营环节的信息系统为例,数据质量存在重大问题的达 30%以上.数据质量

* 基金项目: 国家自然科学基金(61173057, 61100162); 国家重点基础研究发展计划(973)(2011CB302205, 2013CB329305); 国家高技术研究发展计划(863)(2012AA02A608, 2012AA02A613); 创新基金重大项目专项计划(ISCAS-2010-01)

收稿时间: 2012-03-07; 修改时间: 2012-04-20; 定稿时间: 2012-05-25

缺陷直接影响了数据挖掘算法的结果,降低了决策者对信息系统建设的信赖程度.信息系统不能给决策者提供可信的支撑,严重影响了信息化应用的有效推广.因此,提高企业信息化系统的数据质量,构建满足业务需求的数据挖掘系统,是目前信息化建设的两个重点和研究热点.

在数据质量分析管理领域中,大多数研究人员侧重于从模式层面和集成层面研究智能化数据转化和数据清洗方法[2],并开发了一系列 ETL 工具用于数据清洗过程,以期达到提高数据质量的目的.但是由于数据质量问题往往取决于应用领域本身的特点,众多实践表明,忽视人的发现问题和分析问题的认知能力而单纯依赖机器学习和智能分析等自动化技术并不能有效发现数据质量的不足,通过可视化技术将人的强大认知能力与计算机强大的自动处理能力相结合才是提高数据质量的有效途径.目前,数据管理领域的可视化研究偏重于挖掘过程和内容的可视展示,如可视化数据挖掘、可视化 OLAP、基于信息可视化的数据简化等研究[3],对数据质量分析过程中的交互式可视化技术研究存在不足.

为此,本文侧重于研究一种基于可视化形态的数据质量可视分析方法,用于分析提高企业信息化系统中的数据质量,进而提高企业分析决策活动过程中的有效性和准确性.本文首先分析了信息系统中影响数据库数据质量的主要因素,提出了面向多数据源的统一元数据模型和数据库数据质量评估模型,进一步构建了用于数据质量评估的交互式可视形态集,以减轻用户的数据认知负担,帮助用户进行数据库数据质量的可视分析与管理.

1 相关工作

1.1 数据质量管理

数据质量问题的研究,在统计领域始于 1960 年代末期,管理领域始于 1980 年代初,而计算机领域始于 1990 年代初[4].质量的概念最初由 Bendell 总结出来,他列举了西欧、美国、日本的 9 个具有代表性的实例,将质量定义为“产品的功能符合满足客户需求”.其后,Johannsen 等人从管理角度对质量的概念作了重新的定义,认为“质量是符合需求的,关注与产品规格说明的最小偏差”[5].Aebi 将数据质量定义为“数据的一致性(consistency)、正确性(correctness)、完整性(completeness)和最小性(minimality)这 4 个指标在信息系统中得到满足的程度”,构建了数据质量的理论框架,并指出了在信息系统中一些常见的数据质量问题以及针对系统中不同的数据质量问题提出了不同的解决方案[6].在数据质量的研究过程中,数据库作为信息系统中存储数据的主流载体,成为数据质量研究的主要课题.英国成立了信息质量管理中心(Centre of Information Quality Management,简称 CIQM),用于数据库数据质量的管理,CIQM 项目是由英国图书馆成立的项目,用于调查数据库的质量问题及这些问题对用户产生的影响,收集解决数据库数据质量问题的方案,制定一个国际标准用于提高数据库的数据质量[7].Wang 等人提出了数据工程中数据质量的需求分析模型,他认为衡量数据质量的候选指标很多,用户必须根据实际需求选择合适的维度进行数据质量的管理和评价.数据质量指标可以分为两类:数据质量指示器(data quality indicator)和数据质量参数(data quality parameter).前者包含用于描述有关数据客观信息的数据维度,如数据源、创建时间等;后者包含用于展示用户对数据质量的主观性评价的数据维度,如数据来源的可信度(credibility)、数据的及时性(timeliness)等[8].随着数据库数据质量研究的深入,研究人员提出用元数据来表示数据质量以方便数据质量管理,主要从模式层和实例层来进行数据质量评估和提升.模式层主要用于研究数据库设计过程数据结构的缺陷,如缺乏完整性约束的数据库设计;实例层主要用于研究数据库中数据内容问题,如数据缺失、数据重复[9]等.并针对数据库中存在着的数据冲突、数据错误、数据缺失等一系列问题提出了相应的数据清洗、数据缺失弥补等算法,形成了一些比较成熟的 ETL 工具,比如 Merge,Cluster[10,11]就是在 SQL 基础上扩展起来的数据清洗工具.

总的来看,数据质量相关领域的研究人员侧重于结合应用领域知识[12,13]实现数据管理的智能化,但往往忽视了人类自身强大的认知能力,导致数据质量分析水平不高.

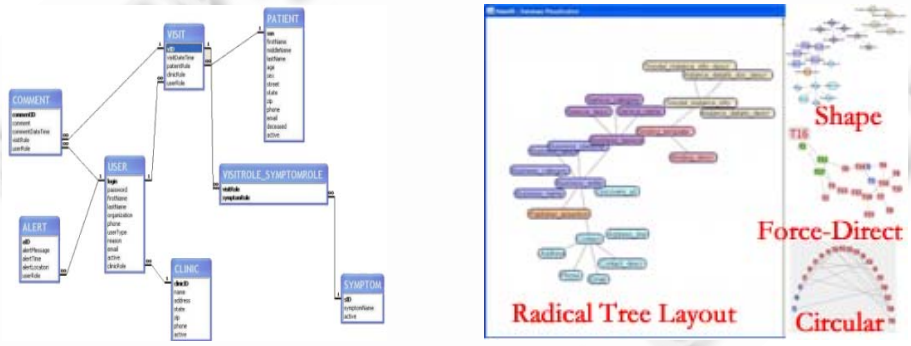
1.2 可视分析与数据库可视化

可视分析(visual analytics)最早在 2005 年美国安全局建立的国家可视化及分析中心组织的研讨会上首次

提出,它重点研究如何通过交互式可视化界面辅助用户进行分析推理,提供辅助用户分析决策的工具和技术,使用户能够从海量、动态、模糊并且可能存在冲突的数据中综合信息并洞察隐藏的规律和模式、检测预期事件、发现意外事件,或为指导行动进行有效的评估和交流[14].可视分析是将可视化、人的因素和数据分析进行组合用于决策的整合方法[15],它在分析推理过程研究中更加强调用户行为及对可视化的有效使用方式等,旨在提供更多的智能化数据分析支持.

目前已有的可视分析系统主要利用 OverView+Detail、动态过滤等交互技术在数据空间中进行探索,能够很好地借助可视隐喻来增强人们对特定数据的认知.典型的可视分析系统有:Sandbox[16]是一个集中人类信息交互的灵活的、表达力强的推理环境,它为支持视觉思维提供了文本编辑来完成形成假设、充实证据、分组、评注等活动,并提供了分析过程模板的功能,用户可以直接填充已有模板来完成分析推理;Aruvi[17]是为支持分析推理过程而设计的可视分析系统,其中的知识视图使用户能够记录分析思路等知识,导航视图提供了可视分析过程的总览视图,允许用户在各个历史分析状态进行迭代分析.

可视分析研究人员进行了相应的数据库可视化的研究工作,如 Delaunay[18]数据库可视化工具是一个可视搜索系统.与传统的支持一套预定义好的可视表示的可视搜索系统不同,它能够支持用户使用基于约束的查询语言表述的面向对象数据库可视化.NakeDB[19]是数据库结构可视化(database schema visualization)工具,如图 1 所示,图 1(a)为传统的数据库软件 E-R 图,如 PowerDesigner,SqlServer 等;图 1(b)为 NakeDB 的 E-R 关系图.与传统的数据库软件 E-R 图相比,NakeDB 的可视形态能够更直观地展示数据库中的信息.该可视形态中的节点代表数据库的实体,颜色、形状代表数据实体簇集,节点大小代表数据实体的关联程度.与此同时,NakeDB 构建了 4 种动态可交互的数据库结构布局视图,能够实时地构建数据库架构的视图模型,有利于用户通过自然交互方式查找和过滤数据库表模型信息,从而帮助软件工程师更好地理解和使用数据库.这些数据库可视化工具侧重于数据库本身的可视化,并没有针对数据库的数据质量构建完整的可视化分析工具,在数据库的数据质量评估和管理方面存在不足.



(a) 传统的数据库表结构 E-R 图 (b) NakeDB 数据库架构视图可视化

Fig.1 Database visualization

图 1 数据库可视化

2 面向数据库数据质量管理的评估模型

影响数据库数据质量的因素有很多,本节分析产生数据质量问题的主要原因以及数据质量问题的分类.同时,为了解决数据质量的评价与管理过程中多数据源数据接口不统一的问题,构建了数据源元数据模型.进一步地,基于数据质量问题分析和数据源数据模型,给出了数据质量评价指标的评价方法以及相应的数据质量评价指标模型.

2.1 数据质量问题分析

信息系统的质量控制是指在对数据进行采集、录入、传输交换、存储管理、加工处理过程中有效地进行质量控制和管理,确保数据的真实准确、安全可靠.从另一方面说,数据质量是指数据满足用户需求、适合用户使用的程度.因此,数据质量问题不仅仅是数据错误,更进一步的是数据的一致性(consistency)、正确性(correctness)、完整性(completeness)和最小性(minimality)这 4 个指标在信息系统中得到满足的程度.在信息系统中,产生数据质量问题的原因主要包括以下几个方面:

(1) 多数据源:多源异构数据源的数据存储方式、存储类别不同,增加了数据的多重性,导致了数据共享和转换过程中出现很多数据质量问题,如命名冲突、结构冲突;

(2) 数据架构设计缺陷:在信息系统的数据架构的设计过程中,缺乏对数据架构的完整性约束、安全性约束的考虑;

(3) 数据生产中的主观判断:由于录入错误,数据源中的数据未及时更新,或不正确的计算等,导致数据源中数据过时,或者一些数据与现实实体中字段的值不相符;

(4) 有限的数据资源:为了降低数据存储空间,常常对数据作一些简化,数据简化过程中会导致一些无法理解的数据值,如伪值、多用途域、异常的格式、密码数据等;

(5) 安全性和可获取性之间的平衡:信息建设过程中,人们经常以损失系统的安全性来提高系统的效率,然而系统安全性的降低,必然产生了许多系统数据质量问题.如,为了提高数据的添加效率,经常删除数据的索引约束,从而导致不同数据表之间的数据冲突;

(6) 信息系统升级:信息系统的升级经常产生数据质量的缺陷,由于新的信息系统的更高的功能要求,导致原有的信息系统的数据架构无法满足要求,在系统数据架构的升级过程中,出现数据的版本兼容问题.

如图 2 所示,数据质量问题可以分为模式层数据质量问题和实例层数据质量问题.

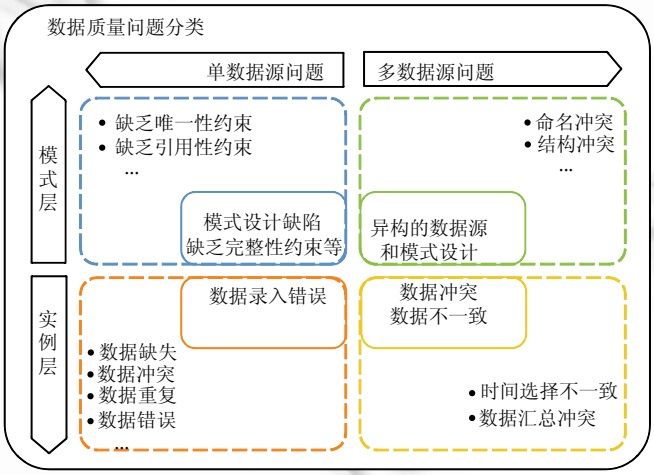


Fig.2 Classification of data quality problems

图 2 数据质量问题分类图

模式层数据质量问题主要反映由于模式设计缺陷导致的数据质量问题,如缺乏完整性约束、唯一性约束、结构冲突等;而实例层数据质量问题主要是用于描述数据记录方面的数据质量问题,如数据缺失、数据重复等.实际应用中,模式层上的数据质量问题也影响到数据的实例层面问题[1].此外,由于数据质量问题在多数据源和单数据源上存在着明显的区别,单数据源下的数据质量问题在多数据源情况下显得更加严重,如数据冲突,由于多数据源情况下数据格式、标准等方面的不一致,数据冲突变得更为普遍.

2.2 数据源元数据模型

企业各部门业务及信息化程度的差异性导致各信息化系统构建过程相对独立,分散的信息系统难以统一数据标准或信息共享问题,容易导致数据源之间的数据冲突(如命名冲突、结构冲突等).多数据源的数据质量评估需要考虑不同数据源的存储模式和数据获取方式,一定程度上增加了数据质量在实例层和模式层的评估难度.为了简化数据质量分析过程中异构数据的获取过程,使用户能够专注于数据质量的评估和分析,本文根据数据源关注的数据质量指标构建了元数据模型,并为企业中经常使用的异构数据源构建了相应的数据获取接口,用于数据源的数据转换.每个数据源都可以转化为元数据模型的实例,元数据模型是数据源关于数据、数据间关系、数据语义和数据约束的概念描述.本文采用 BNF 范式对数据源的元数据模型作如下定义:

定义 1(元数据模型). 用于描述数据源内部实体、实体关系、实体模式、实体约束的概念集合,可以由数据源类型、实体、实体关系、实体模式表示等组成:

⟨*MetaData*⟩::=⟨*Entity*⟩⟨*Relation*⟩⟨*Schema*⟩⟨*Source*⟩.

定义 2(实体). 用于描述数据源中独立的、区别于其他数据载体的描述,可以是信息系统的数据表、视图,或者是数据源的系统表.每个实体可以用名称、属性、键值、模式和备注进行描述:

⟨*Entity*⟩::=[*Table*]|[*View*]|[*SysTable*];

⟨*Table*⟩::=⟨*Name*⟩⟨*Columns*⟩⟨*Key*⟩⟨*Category*⟩⟨*Schema*⟩⟨*Remark*⟩;

⟨*View*⟩::=⟨*Name*⟩⟨*Columns*⟩⟨*Key*⟩⟨*Category*⟩⟨*Schema*⟩⟨*Remark*⟩;

⟨*SysTable*⟩::=⟨*Name*⟩⟨*Columns*⟩⟨*Key*⟩⟨*Category*⟩⟨*Schema*⟩⟨*Remark*⟩.

定义 3(属性). 用于描述数据源中实体内部各个字段的概念描述,由含义、类型、长度和备注组成:

⟨*Columns*⟩::=⟨*Column*⟩⟨*Type*⟩⟨*Size*⟩⟨*Remark*⟩.

定义 4(约束). 用于描述数据源中实体之间的数据约束,由实体及约束属性组成形成数据源的实体关系网:

⟨*Relation*⟩::=⟨*Entity*1⟩⟨*Column*1⟩⟨*Entity*2⟩⟨*Column*2⟩.

定义 5(模式). 用于描述数据源数据设计的逻辑设计概念:

⟨*Schema*⟩::=[*dbo*]|[*sys*]|[*other*].

定义 6(数据源). 定义数据来源的类型.本文主要针对企业信息化系统中常用的数据源(如 Oracle,MSSQL,MySql 等):

⟨*Source*⟩::=[*Microsoft Sql Server*]|[*Oracle*]|[*MySql*]|[*Access*]|[*Excel*]|[*XML*].

2.3 数据质量评估模型

数据质量的评估应根据数据库中存在着的数据质量问题的特点,结合数据领域知识制定相应的数据质量评估指标.如第 2.1 节中图 2 所示,将数据质量的问题分为两类:模式层次的数据质量问题和实例层次的数据质量问题.为了便于数据质量指标模型的研究,本文将数据库的数据质量评价指标也划分为两个层次:模式层和实例层.模式层是数据库关于实体数据的含义的抽象描述集合,用集合 $T=\{t_0,t_1,\ldots,t_n\}$表示;实例层是数据库关于模式层实体元素的一个实例数据集合,用集合 $I=\{i_0,i_1,\ldots,i_n\}$表示.那么,数据库的评估 M 相当于对模式层 T 和实例层 I 评估的并,即 $M=\{\{t_0,t_1,\ldots,t_n\},\{i_0,i_1,\ldots,i_n\}\}$.

为了进行数据质量指标评估,我们假设 $M=\{\{rt_0,rt_1,\ldots,rt_n\},\{ri_0,ri_1,\ldots,ri_n\}\}$为完美的数据库指标特征集,则现有的数据库数据质量指标特征集为 $S=\{\{st_0,st_1,\ldots,st_n\},\{si_0,si_1,\ldots,si_n\}\}$,那么,如果数据库具有良好的数据质量,$S$ 和 R 之间存在完美的映射.为此,本文构建了一个 S 到 R 的映射函数 $f_t,f_i:S\to R$,表示 S 到 R 的一个映射过程;那么,$f_t^{-1}, f_i^{-1}: R\to S$ 表示 R 到 S 的一个逆映射过程.$g_t,g_i:S\times R\to\{\text{true,false}\}$用于反映 S 特征集上的数据是否满足 R 特征集上的描述,那么,S 数据质量的特征是良好的数据质量应满足如下标准:

- SI 集合中的任意一个元素实例都是 ST 的一个实体元素的具体实例,即$\forall si_i\in SI\to\exists si_k\in ST$;
- S 集合中任意一个模式元素都是 R 集合中某个模式元素的正确反映:$\forall st_i\in ST\wedge f_t(st_i)\in R\to g_t(st_i,f_t(st_i))=$true;$S$ 集合中任意一个实例元素在 R 集合中都存在对应的实例元素,即

$$\forall si_i \in SI \wedge f_i(si_i) \in R \rightarrow g_i(si_i, f_i(si_i)) = \text{true};$$

- S 集合中的模式集包含 R 集合中每一个模式集,即 S 集合是完备的描述实体特性的模式集:

$$\forall rt_i \in RT \rightarrow f_t^{-1}(rt_i) \in ST;$$

- S 集合中的模式集应该不包含 R 集合以外的任意元素,$\forall st_i \in ST \rightarrow \exists f_t(st_i) \wedge (f_t(st_i) \in RT)$;$S$ 集合中的实例集不包含 R 集合中不存在的实例元素,$\forall si_i \in SI \rightarrow \exists f_i(si_i) \wedge (f_i(si_i) \in RI)$.

针对 R 和 S 集合中模式层和实例层元素的评价角度和方法,Wang 等人从数据工程中数据质量的需求分析和模型的角度出发,将指标分为两类:数据质量指示器(data quality indicator,简称 DQI)和数据质量参数(data quality parameter,简称 DQP).DQI 用于一些客观数据维度的评估,DQP 用于一些主观信息的评估.本文针对数据质量常见的问题,对数据质量的评估指标进行了深化,提出数据质量评估至少应该包含以下两方面的基本评估指标[20,21],如图 3 所示.

(1) 数据对用户必须是可信的,包括精确性、完整性、一致性、有效性、唯一性等评估指标.精确性评估指标是指描述数据是否与其对应的客观实体的特征相一致;完整性评估指标是指描述数据是否存在缺失记录或缺失字段,不能缺项,尤其是重要的项目;一致性评估指标是指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;有效性评估指标是指描述数据是否满足用户定义的条件或在一定的值域内;唯一性是指描述数据是否存在重复记录;

(2) 数据对用户必须是可用的,包括时间性、稳定性等指标.时间性评估指标是指描述数据是当前数据还是历史数据,如数据在多源数据库中的时间等;稳定性评估指标是指描述数据是否是稳定的,是否在其有效期内.

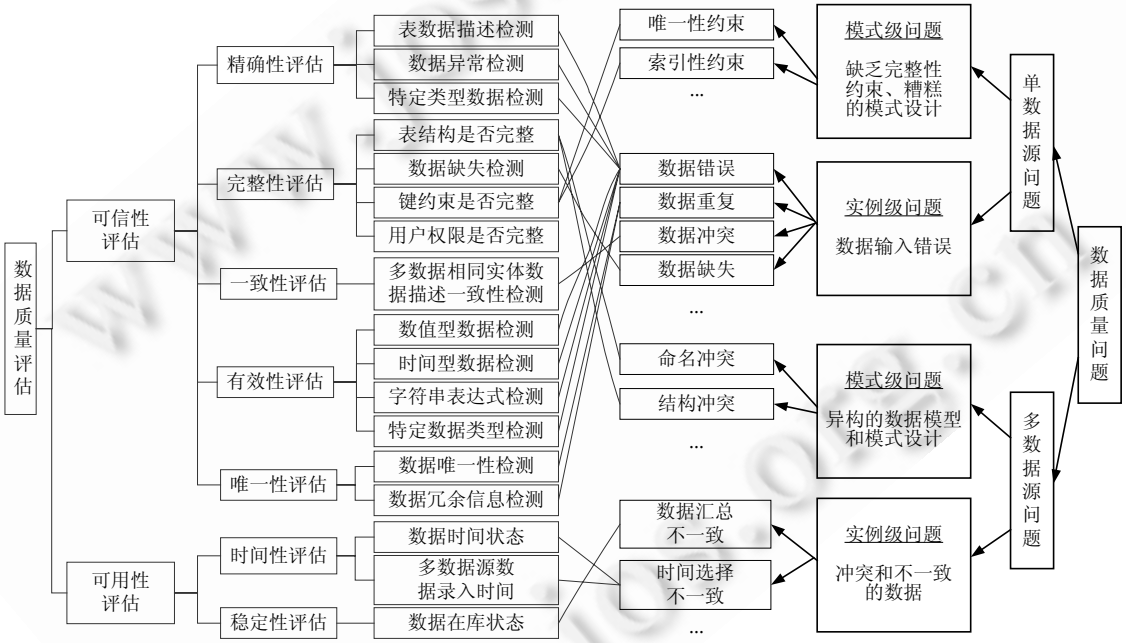


Fig.3 Mapping between data quality indicator and database problems

图 3 数据质量评估指标与数据质量问题关系映射图

3 面向数据质量评估指标的可视形态

数据质量的分析管理过程是一个循环往复的渐进式的过程.管理人员需要从数据源、数据库设计模式、数据库数据内容等多个数据质量指标角度渐进地分析数据库的数据质量问题.因此,为了辅助管理人员的分析决策过程,本文针对第 2.2 节和第 2.3 节构建的数据库元数据模型和数据库数据质量评估模型构建了自然的、对用户思维活动干扰少的交互方式、降低分析决策者的交互负担的可视化形态,用于辅助用户的数据质量管理与

分析过程.

3.1 数据质量评估指标的可视形态构建方法

Card认为,信息可视化是从数据到可视化形式再到人的感知系统的可调节的映射过程[22].在该模型中,从原始数据到人的感知系统,数据变换把原始数据映射为数据表;可视化映射把数据表转换为可视化结构(空间基、标记和图形属性的结构);视图变换通过定义位置、缩放比例、裁剪等图形参数创建可视化结构的视图;用户的交互动作则用来控制这些变换的参数,例如把视图约束到特定的数据范围,或者改变变换的属性等.可视化及其控制过程最终服务于分析任务.本文结合 Card 的可视模型特点,针对数据库元数据模型和数据库数据质量评估模型构建了数据质量的可视形态模型.如图 4 所示,可视形态从统一的元数据模型中获取数据,简化了从多数据源中获取数据的过程;然后,根据相应的数据质量指标评估选取相应的数据质量评估算法进行分析和评估,将评估结果转化为可视形态可以映射和绘制的可视结构,映射到用户的可视视图上;用户再根据领域知识和数据质量管理的先验知识对可视视图进行交互,如拖拽、圈选、钻取等;可视视图再根据用户交互的视图映射到相应的可视结构上,可以是可视视图的视图变换(如放缩、平移)、可视结构的数据变换(如数据表格式变化)、分析结果的数据变换(如数据过滤、数据检索)等任务集,甚至可以是重新选择算法和数据进行循环往复的可视分析.

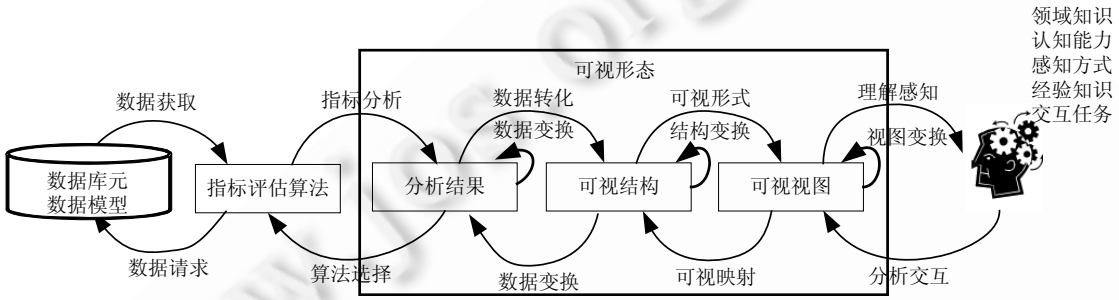


Fig.4 Data quality visual component and data transfer

图 4 数据质量可视形态交互与数据迁移图

因此,可视形态定义了数据信息的可视化输出形态、交互任务集、底层数据模型.将可视形态定义为一个三元组:

$$Morphology=\langle Visual,Actions,Model\rangle.$$

其中,

- *Visual* 表示可视表示,代表数据质量数据模型通过映射到可视视图上的具体可视形式,包含数据模型的具体的可视隐喻;
- *Actions* 是用户与可视形态交互的任务集,包括图元操作、数值操作、视图操作、内容操作等[23].采用四元组表示如下:

  $$Actions=\langle 图元操作,数值操作,视图操作,内容操作\rangle;$$

  其中,图元操作表示针对可视形态中的图形元素的操作,包括平移、缩放、选定、旋转;数值操作指针对可视形态中的数据运用求和、平均、比例、搜索、过滤等方式进行操作;视图操作指针对可视形态中的数据采用排序、上钻、重置等方式重新布局可视形态中的数据情况;内容操作是指用户可以对可视形态进行标记、注释等操作;
- *Model* 表示指标评估算法分析的数据经过数据变换后可以映射到可视形态的数据集.

3.2 面向数据质量评估的可视分析形态集

可视分析环境通过交互式形态呈现数据,并提供用户交互任务途径,是用户与分析环境的数据质量管理与分析的界面接口.本文结合数据质量的元数据模型特点和数据质量指标评价标准,针对数据质量问题不同层次

的问题构建了一系列可视形态集合,如图 5 所示.

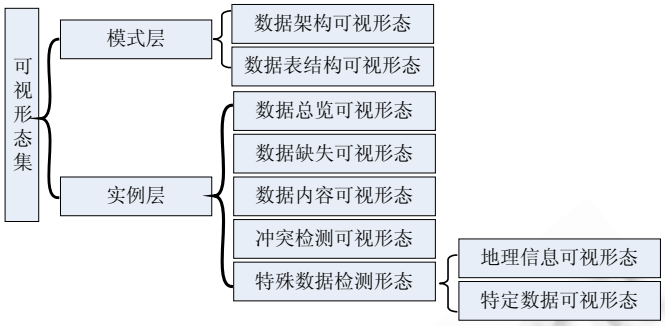


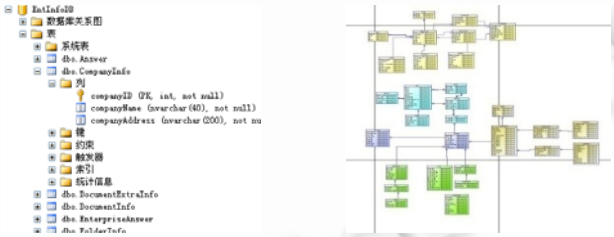
Fig.5 Classfication of visual component

图 5 可视形态分类图

针对模式层的问题,本文构建了数据架构可视形态和数据表结构可视形态,其中,数据架构可视形态用于分析数据库模式层架构问题,数据表结构可视形态用于分析数据库模式架构独立表的结构问题.针对实例层的问题,本文构建了数据总览可视形态、数据缺失可视形态、数据内容可视形态、冲突检测可视形态、特殊数据检测形态等,数据总览可视形态用于反映整个数据库内部数据的概括,可以进行数据内部数据冲突检测;数据缺失可视形态可以用于分析数据库内部表的数据缺失情况;数据内容可视形态辅助用户进行数据检索和数据过滤;数据冲突检测形态用于分析数据表内部数据的冲突情况;特殊数据检测形态用于分析一些具有特殊格式的数据或者特殊数据,如城市地理信息、电话、邮编等.

(1) 数据架构可视形态

数据架构可视形态用于对构建的数据库元数据模型的可视化,主要让分析人员能够了解数据模型的整体结构,是用户进行数据质量分析的出发点,对于用户的数据库模型的理解和数据质量分析具有重要的作用.图 6 是传统的数据库架构图,这种数据库架构可视形态存在明显不足:第一,采用传统的普通树形表示的表结构按照表名进行组织,展示信息不足,不利于获取表之间的信息;第二,E-R 图表虽然可以获取比较详细的信息,但是不利于复杂的企业级数据模型的展示,并不能够获取表内数据的重要性和状态.



(a) SQL server 表关系展示 (b) Powerdesigner 实体关系图

Fig.6 Tranditional database schema visualization

图 6 传统的数据库架构图

为了弥补可视信息的不足,本文基于 prefuse[24]有向图布局的设计开发了元数据可视形态,它能够反映出数据模型之间的关系,并且能够根据节点大小、颜色迅速找到数据模型中的关键表和核心实体模型.如图 7 所示,图形中的节点圆表示实体数据模型表,圆的大小表示模型蕴含数据量大小,圆越大,数据量越大.圆中的中心圆代表不同的实体类型(如图 7 右侧说明所示),圆中的彩色 Ring 环代表数据模型中列,不同的颜色代表不同的数据类型(如图 7 右侧说明所示),环占有的比例代表不同列占有的数据比例.数据连线表示实体模型之间的实体关系.

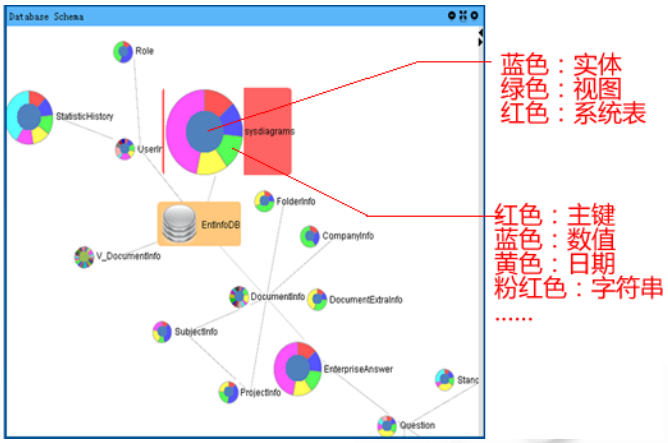


Fig.7 Database schema visual component base on undigraph

图 7 基于无向图的数据架构可视形态

(2) 数据表结构可视形态

数据表结构可视形态用于分析数据库模式架构独立表的结构问题.如图 8(a)所示,该可视形态主要获取表的基本结构信息,包括字段名、键值、字段宽度、基本描述等信息.

(3) 数据总览可视形态

数据总览可视形态用于反映整个数据库内部数据的概括,可以进行数据内部数据冲突检测.如图 8(b)所示,图中圆柱代表不同的实体表,表的结构越复杂,圆柱越大;数据量越大,圆柱越高.表中连线代表不同表中数据之间的联系,对于正常数据和冲突数据,系统可自定义颜色,分别加以表征.

(4) 数据缺失可视形态

数据缺失可视形态:为了检验数据在数据模型中是否完整,如图 8(c)所示,用户可以查看单个实体模型表各类数据占用比重.对于已有数据和缺失数据,系统可自定义颜色,分别加以表征.

(5) 数据内容可视形态

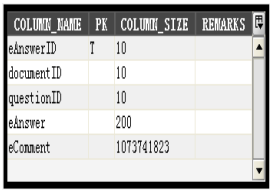
数据内容可视形态辅助用户进行数据检索和数据过滤.如图 8(d)所示,用户可以使用数据项进行检索过滤,以发现异常的数据信息.

(6) 数据冲突检测形态

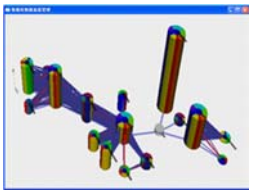
数据冲突检测形态用于辅助用户进行内部数据的冲突检测.如图 8(e)所示,本文采用一组平行轴对具有约束关系的实体进行数据展示,维度代表数据属性,每一条线代表数据实例在实体中实例数据的数据值(字符串类型的,用记录 ID 表示).因此,如果数据存在数据冲突,在可视形态上数据线存在断裂,帮助用户开始发现数据的冲突情况.

(7) 特殊数据检测形态

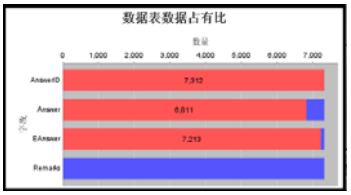
针对数据库中具有特殊含义的数据进行错误的检测,如邮编信息、邮箱信息、地址信息等具有特殊格式的数据.如图 8(f)所示,采用点图对数值型数据进行数据聚类,有利于发现异常离群点数据.如图 8(g)所示,针对具有地域信息的数据,我们采用数据 GIS 形态,图中每个标签代表相应地域的数据内容信息.



(a) 数据表结构可视形态

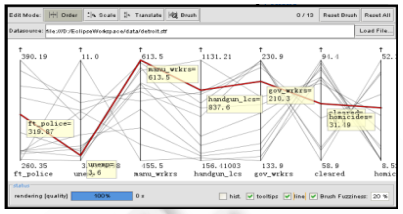


(b) 数据总览可视形态

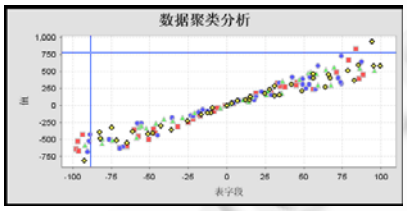


(c) 数据缺失可视形态

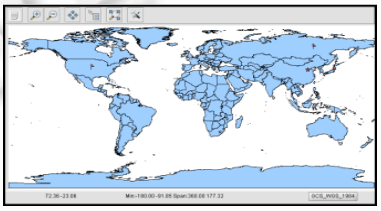
(d) 数据内容可视形态



(e) 数据冲突可视形态



(f) 特殊数据检测形态——数值聚类



(g) 特殊数据检测可视形态

Fig.8 Typical data quality analysis visual morphologies

图 8 典型数据质量分析可视化形态

4 面向数据质量管理的可视分析架构

基于前面两节的研究,本文构建了面向数据质量管理的可视分析系统,系统体系结构如图 9 所示.系统架构分为 5 层:

- 多源异构数据源层:主要用于描述在企业信息系统中多源异构的数据源.企业各个部门独立建立自己的信息系统,由于缺乏统一的标准和规划,各个部门之间的数据源存储方式有所不同,如 Oracle,MSSQL 等,因此要为各种类型的数据源编写独立的数据获取接口;
- 统一元数据模型层:针对多源异构的数据源构建相应的数据抽取、转换接口,将各个数据源转化为相同的元数据模型,用于简化指标评估过程中数据的抽取过程,从而使系统能够更加专注于数据质量的可视分析过程;
- 指标分析算法:面向数据库质量评价指标,针对元数据模型层的数据完成相应的数据质量分析.本文针对数据质量的常见的数据质量问题,设计了完整的数据质量分析指标模型,并且为数据质量的分析指标编写了相应的算法库,用于支持数据质量的可视分析过程;
- 数据质量可视形态层:主要包括两个模块:可视形态集和交互任务集.本文从易于用户理解和认知的角度构建了一套数据质量分析可视形态集,可视形态集主要用于针对指标评估算法分析的结果的可视表示,并且提供了易于用户交互使用的交互任务辅助用户的数据质量管理与分析;
- 可视分析支撑接口层:用于将用户分析和整理的可信度较高的数据用于可视分析的决策,最终提高企业分析决策的准确性和有效性.

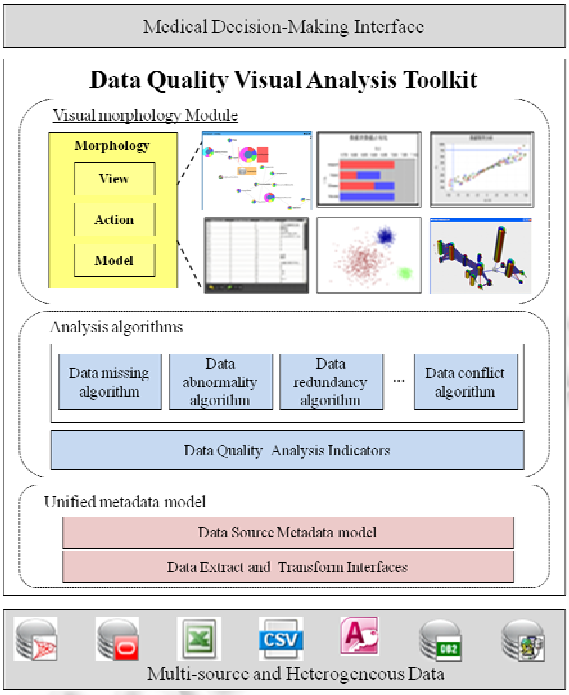


Fig.9 Database quality visual analysis system architecture

图 9 数据质量可视分析系统结构

5 关键技术

5.1 数据元数据模型图布局算法

在数据质量评估过程中,以利于用户理解和分析的方式展示数据库的 *Schema*,能使用户对数据库的信息有整体的把握,对于数据质量的评估和分析具有重要的意义.本文构建了基于 prefuse 的 *RadialGraph* 布局,构建了数据架构可视形态,形态将数据库的 *Schema* 作为无向图,*Entity* 作为无向图节点,*Entity* 之间的主外键约束作为边.由于数据库的存储架构是模块化设计的,导致所构建的 *Schema* 无向图并不是连通图,为了进行整体的布局,构建了基于 Focus+Context 等交互技术,引进了元节点 *MEntity* 连接所有的连通子图形成整个连通图.我们用 *Entity*[*n*]表示 *Schema* 里有 *n* 个节点,*relation*[*n*][*n*]存储 *Entity* 之间的约束关系.具体算法如下所示:

Step 1. 初始化 *Entity*[*n*]为 0,表示节点均未被划入任何连通子图;初始化 *relation*[*i*][*j*];初始化连通子图 *graph*=1,表示获取第 1 个连通图;初始化 BFS 路径树 *Tree*[*i*]:

$$relation[i][j]=\begin{cases}1, & \text{if } Entity\ i,\ j \text{ exist } constraint\\ 0, & \text{if } Entity\ i,\ j \text{ not exist } constraint\end{cases};$$

Step 2. 遍历 *Entity*[*n*],如果存在 *Entity*[*i*]=0,令 *Tree*[0]=*i*,跳转到 Step 3;否则,跳转到 Step 5;

Step 3. 根据 *Tree*[*i*],BFS 整个图,标记连通子图.依次取 *Tree*[*i*],如果 *Entity*[*Tree*[*i*]]=0,令 *Entity*[*Tree*[*i*]]=*graph*.如果遍历到 *Tree* 的末端,*graph*=*graph*+1,跳转到 Step 2;

Step 4. 遍历 *relation*[*Tree*[*i*]][*j*],如果 *relation*[*Tree*[*i*]][*j*]=1,将 *j* 加入到 *Tree* 中,跳转到 Step 3;

Step 5. 如果 *graph*>=2,说明存在两个以上的连通子图,则依次获取每个子图中的 *focus* 点.将连通图中入度最多的点称为这个图的 *focus* 点.因此,第 *i* 个子图的 *focus* 点为

$$Focus[i]=k;\ \text{if}\ \sum_{j=0}^{n} relation[k][j] > \max\left(\sum_{j=0}^{n} relation[l][j]\right) \& Entity[k]=Entity[l]=i;$$

Step 6. 连接元节点与 *focus* 节点,形成连通图.

进而,采用 RadialGraph 布局对构建的连通图进行布局.同时,为了帮助用户更快地发现数据库架构中的关键实体,如图 7 所示,可以用不同的实体大小,代表不同实体表的数据量大小.为了防止数据量差异过大造成实体大小差异过大,导致有些实体占据了整个形态视图,而有些实体无法正常显示,必须对实体数据进行归一化处理.定义 *Size* 为图节点尺寸,max*Size* 为根据可视形态大小变化的节点最大尺寸,min*Size* 为根据可视形态大小预算的最小尺寸,*volume* 为实体的数据量大小,max*Volume* 为数据存储架构中最大的数据量,min*Volume* 为最小的数据量.因此,节点尺寸有如下公式:

$$Size=\min Size+\frac{volume-\min Volume}{\max Volume-\min Volume}\times(\max Size-\min Size).$$

5.2 数据质量指标评估算法

在面向多源异质的数据源中,数据质量的提高主要集中在以下几个方面:重复对象检测、缺失数据处理、异常数据检测、逻辑错误检测、不一致数据处理等.针对上述问题的处理可以分为两类:一类可以通过选取恰当的可视隐喻,对原始表记录数据个体之间的关联关系进行交互式可视化,即可以有效发现不良数据,如逻辑错误检测、不一致数据处理等,这方面的数据质量问题不需要数据指标评估算法;另外一类需要对表记录集数据的某一指标的整体计算,并对计算结果进行可视化,才能实现该方面的评估,如数据重复、数据缺失、数据异常等,本文针对这类数据质量问题设计了相应的算法,用于数据质量的评估与检测.

(1) 数据重复检测算法

数据重复(duplicate record)是指在一个数据库中,存在超过一条不完全相同的数据记录用于表示现实世界中的同一个实体,如“Daniel Keim”.“D. Keim”是一个人名的两种不同的表示形式,只是属性的存储顺序不同.

对于数据库中实体的数据,假设实体 E 存在 N 个字段,记为 $E=\{col_1,col_2,col_3,\ldots,col_n\}$.为了计算数据实体记录之间的相似性,我们针对每一个字段 col_i 进行排序,那么针对每一条数据,我们可以获得一个排序的序号 s_i,则每一条记录 R 可以用一个 n 元组表示,记为 $R=\{s_1,s_2,s_3,\ldots,s_n\}$.将每个 n 元组当做一个 n 元的向量,则两条记录之间的相似性可以用向量之间的夹角余弦值来表示:

$$Similar(R_i,R_j)=\cos(R_i,R_j)=\frac{\sum_{k=0}^{n}(S_{ik}\times S_{jk})}{\sqrt{(\sum_{k=0}^{n}S_{ik}^2)\times(\sum_{k=0}^{n}S_{jk}^2)}}.$$

为了减少记录之间的比较次数,提高检测效率,可以采用仅比较相互距离在一定范围内的记录.即先对数据表中的主记录排序,然后对邻近记录进行比较.通过以上过程,可以检测出重复记录.

(2) 数据缺失检测算法

数据缺失(data missing)是指在数据集中关于真实世界的事物描述方面的信息缺失.导致数据缺失的原因包括信息录入时数据属性无法获取、信息录入遗漏,还有可能是由于数据保存过程中设备网络故障等非人为因素造成的.

检测不完整数据的算法比较简单,本文采用的逐条记录遍历标记缺失记录的方法,假设一条记录可表示成 $R=\{col_1,col_2,col_3,\ldots,col_n\}$,表示记录 R 的 n 个属性,$R_i(col_j)$表示记录 R_i 第 j 个属性 col_j 的值,$col_j(default)$表示记录第 j 个属性 col_j 的缺省值,T 为数据表中记录的总数,则不完整数据检测算法的伪码可描述如下:

```
1) For i=1 to T;
    2) For j=1 to n;
        3) If R_i(col_j) is NULL or R_i(col_j)=R_i(col_j(default)) then
            Mark R_i out as a piece of incomplete data;
        4) End if;
    5) End;
6) End;
```

(3) 数据错误检测算法

数据异常(dirty record)指的是数据集中的数据对真实世界的事物描述存在差异性,如年龄为“249”等.产生数据异常的原因大部分是由于使用人员的疏忽或者系统缺乏有效性验证造成的.

数据异常的检测应根据数据集中不同的属性类型采用不同的检测策略,如,针对“整型数据”采用值域范围、数据聚类等方法进行分析,针对“日期”采用格式验证、有效性验证等方式.本文对数据集中的常见数据类型及其数据异常的检测方法进行总结,见表 1.其中,对于字符串错误数据的检测最为复杂,主要采用基于业务规则的检测方法,即在检测错误数据时,根据对具体业务的分析,在规则库中定义相应的业务规则,生成相应的正则表达式,如邮箱正则表达式{/^([a-zA-Z0-9_-])+@([a-zA-Z0-9_-])+((\.[a-zA-Z0-9_-]{2,3}){1,2})$/}.然后执行错误数据检测,判定每条记录是否符合所定义的业务规则.如果记录不符合所定义的业务规则,则该记录含有错误数据.对于错误数据的检测,也可以通过采用相关算法查找被审计数据中的异常数据进而发现错误数据这一过程来完成.

Table 1 Frequently-Used data type and check solutions methods

表 1 常用数据类型及检测方法

数据类型	处理类型	检测方案
Int, Smallint, Tinyint, Number, Decimal, Real	数值类型	1. [*s*,*e*],(*s*,*e*],[*s*,*e*),(*s*,*e*)值域检测 2. 数据聚类,发现异常、孤立数据点 3. 数据分布,在值域等期间内分布情况 4. 极值
Char, Varchar, Text, Nchar, Nvarchar, Ntext	字符类型	1. 正则表达式检验,如手机数据格式“13[0-9]{9}” 2. 数据长度检验
Datetime, Date	日期	1. 日期有效性验证,比如, “2012 年 2 月 30 日” 2. 数据采集是日期分布校验
Binary, Image, Varbinary	特殊数据	1. 检验字段长度在所有数据中出现的情况

6 应用实例

基于以上研究,我们开发了面向数据质量管理的交互式可视分析系统,并将该系统应用于某企业的“企业信息库管理系统”的数据质量管理过程中,图 10 为面向数据质量管理的交互式可视分析系统在企业信息库管理系统数据质量分析过程中的界面展示.“企业信息库管理系统”是一个面向某地区企业信息统计和维护的信息系统,主要用于获取企业的经营信息,然后根据企业的经营指标评审企业的经营状况,制定经营策略.因此,信息库管理系统中的数据的数据质量对于企业的经营决策的分析具有重要的意义.

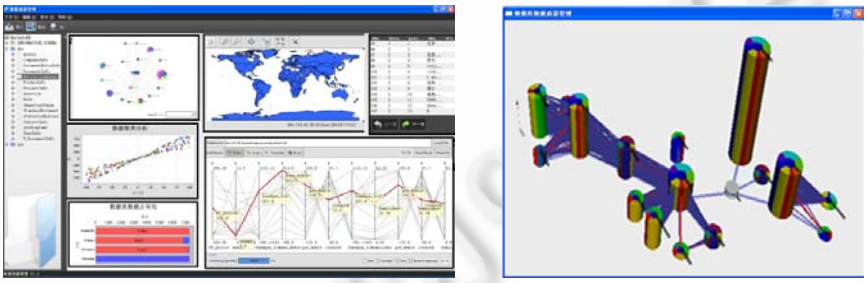


Fig.10 Example of database quality assessment system

图 10 数据质量管理系统系统实例

数据质量管理系统通过设定数据源以后,能够智能选择相应的数据源接口读取数据库的架构.采用第 5.1 节所述的数据元数据模型图布局算法对数据库的架构进行布局,如图 7 所示,使得分析人员对于“企业信息库管理系统”中的数据架构有总体认识.从数据架构可视形态中,用户可以获取数据架构的模块信息,并且可以通过平移、拖动、旋转、搜索等操作了解数据库的架构,从中发现需要关注的数据实体表.由于数据库中表的数据

量越大,说明表的数据的重要性越强,因此用户可以重点关注核心实体表,通过数据缺失可视形态观察数据表中的数据缺失情况.由图8(c)所示的可视形态中可以看到,数据表中Remark属性的数据全部缺失,则表明很多企业在填写数据信息时很少填写一些主观意见.采用数据重复检查算法查出数据表中的重复和冗余数据,针对一些数值型数据可以采用点图形态,分布出整体的数据分布情况,查找异常数据,如图 8(f)所示.针对具有地域分布信息的数据,可以采用特殊数据监测形态——GIS 形态,如图8(g)所示,进行地域型数据的可视分析.同时,用户还可以选中一些表查看这些表内部的数据分布情况,采用数据总览可视形态发现内部的异常数据,然后采用数据内容可视形态,如图 8(d)所示,展示了数据冲突的实际记录并进行修改,从而达到提高数据质量的目的.总之,数据质量管理系统对于辅助用户进行数据质量管理和分析具有重要的作用,通过可视形态之间的协同交互,有利于用户快速发现数据质量问题,保证企业分析决策的准确性和有效性.

7 结论与展望

本文研究了影响数据质量的主要因素,提出了影响数据质量的几个主要原因,并针对所产生的数据质量问题构建了面向多源数据源的统一元数据模型和面向数据质量评估指标的数据库数据质量评估模型,提出了一系列数据质量管理和检测算法.同时,为了辅助数据质量的管理过程,减轻人们的认知负担,构建了用于数据质量评估的交互式可视形态集和交互式任务集,帮助用户进行数据库数据质量的可视分析与管理.最后,开发了一个面向数据库数据质量管理的可视分析系统,并将该系统应用于某管理部门的“企业信息管理系统”当中,结合具体企业进行验证.结果表明,该系统能够有效地进行数据质量评估与数据质量管理,提高企业分析决策的可靠性与准确性.

由于时间所限,本系统还存在以下不足:构建的可视化工具集可视形态偏少,针对关联性数据类型的数据质量可视分析比较欠缺.我们将在接下来的研究工作中进一步丰富改进可视化工具集,以提高数据库的数据质量分析水平.

致谢 在此,我们向对本文的工作提出宝贵意见的评审专家表示衷心的感谢.

References:

[1] Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 2000,23(4):3−13.

[2] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. Journal of ACM Computing Surveys (CSUR), 2009,41(3):16:1−16:52. [doi: 10.1145/1541880.1541883]

[3] Keim D, Kohlhammer J, Ellis G, Mansmann F. Mastering the Information Age—Solving Problems with Visual Analytics. Geneve: Eurographics Association, 2010.

[4] Scannapieo M, Catarci T. Data quality under the computer science perspective. Arehivi&Computer, 2002,2:1−12.

[5] Johannsen CG. Danish experiences of TQM in the library world. New Library World, 1992,93(1104):4−9.

[6] Aebi D, Perrochon L. Towards improving data quality. In: Sarda NL, ed. Proc. of the Int’l Conf. on Information Systems and Management of Data. Delhi, 1993. 273−281.

[7] Medawar K. Database quality: A literature review of the past and a plan for the future. Program: Electronic Library and Information Systems, 1995,29(3):257−272. [doi: 10.1108/eb047199]

[8] Wang RY, Kon HB, Madnick SE. Data quality requirements analysis and modeling. In: Proc. of the 9th Int’l Conf. on Data Engineering. Washington: IEEE Computer Society, 1993. 670−677. [doi: 10.1109/ICDE.1993.344012]

[9] Draisbach U, Naumann F. DuDe: The duplicate detection toolkit. In: Proc. of the VLDB 2010. Singapore: User-Centered Data Management, 2010. http://www.vldb2010.org/proceedings/files/vldb_2010_workshop/QDB_2010/Paper5_Draisbach_Naumann.pdf

[10] Galhardas H, Florescu D, Shasha D, Simon E, Saita C. Declarative data cleaning: Language, model and algorithms. In: Apers P, Atzeni P, Ceri S, *et al*., eds. Proc. of the 27th Int’l Conf. on Very Large Data Bases. Roma: Morgan Kaufmann Publishers, 2001. 371−380.

[11] Raman V, Hellerstein J. Potter’s wheel: An interactive data cleaning system. In: Apers P, Atzeni P, Ceri S, *et al*., eds. Proc. of the 27th Int’l Conf. on Very Large Data Bases. Roma: Morgan Kaufmann Publishers, 2001. 381−390.

[12] Lee ML, Ling TW, Low WL. Intelli clean: A knowledge-based intelligent data cleaner. In: Proc. of the 6th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 290−294. [doi: 10.1145/347090.347154]

[13] Guo ZM, Zhou AY. Research on data quality and data cleaning: A survey. Ruanjian Xuebao/Journal of Software, 2002,13(11): 2076−2082 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/20021103.htm

[14] Thomas J, Cook K. Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, 2005.

[15] Keim D, Andrienko G, Fekete JD, Görg C, Kohlhammer J, Melancon G. Visual analytics: Definition, process, and challenges, information visualization. Lecture Notes in Computer Science, 2008,4950:154−175. [doi: 10.1007/978-3-540-70956-5_7]

[16] Wright WD, Schroh D, Proulx P, Cort B, Jonker D. The Sandbox for analysis—Concepts and methods. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. New York: ACM, 2006. 801−810.

[17] Shrinivasan YB, van Wijk JJ. Supporting the analytical reasoning process in information visualization. In: Proc. of the 26th Annual SIGCHI Conf. on Human Factors in Computing Systems. 2008. [doi: 10.1145/1357054.1357247]

[18] Cruz IF, Averbuch M, Lucas WT, Radzyminski M, Zhang K. Delaunay: A database visualization system. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'97). ACM Press, 1997. 510−513. [doi: 10.1145/253260.253376]

[19] Cortés-Peña LM, Han Y, Pradhan N, Rigaux R. NakeDB: Database schema visualization. In: Proc. of the APRIL 2008. 2008. http://users.ece.gatech.edu/cortes/files/NakeDB.pdf

[20] Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 1996,39(11): 86−95. [doi: 10.1145/240455.240479]

[21] Strong DM, Lee YW, Wang RY. Data quality in context. Communications of the ACM, 1997,40(5):103−110. [doi: 10.1145/253769.253804]

[22] Card SK, Mackinlay JD, Shneiderman B. Readings in Information Visualization: Using Vision to Think. San Francisco: Morgan Kaufmann Publishers, 1999.

[23] Teng DX, Wang GZ, Xiong JQ, Wang HA, Dai GZ. Visual analytic system based on interaction history on-line tracking mechanism. Ruanjian Xuebao/Journal of Software, 2010,21:51−59 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/10006.htm

[24] Heer J, Card SK, Landay JA. Prefuse: A toolkit for interactive information visualization. In: Proc. of the CHI Conf. on Human Factors in Computing. New York: ACM Press, 2005. 421−430. [doi: 10.1145/1054972.1055031]

附中文参考文献:

[13] 郭志懋,周傲英.数据质量和数据清洗研究综述.软件学报,2002,13(11):2076−2082. http://www.jos.org.cn/1000-9825/20021103.htm

[23] 滕东兴,汪恭正,熊金泉,王宏安,戴国忠.基于交互历史在线跟踪机制的可视分析系统.软件学报,2010,21:51−59. http://www.jos.org.cn/1000-9825/10006.htm

滕东兴(1973－),男,山东青岛人,博士,副研究员,CCF 高级会员,主要研究领域为可视分析技术,用户界面技术.
E-mail: dongxing@iscas.ac.cn

王宏安(1963－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为实时智能,用户界面.
E-mail: hongan@iscas.ac.cn

曾志荣(1987－),男,硕士生,主要研究领域为人机交互技术,可视分析技术.
E-mail: zengzhirong88@gmail.com

戴国忠(1944－),男,研究员,博士生导师,CCF 高级会员,主要研究领域为人机交互,计算机图形学.
E-mail: guozhong@iscas.ac.cn

杨海燕(1980－),女,博士,助理研究员,主要研究领域为人机交互技术,草图用户界面.
E-mail: haiyan@iscas.ac.cn