数据质量和数据清洗研究综述*

郭志懋, 周傲英

(复旦大学 计算机科学与工程系,上海 200433);
(复旦大学 智能信息处理开放实验室,上海 200433)
E-mail: zmguo@fudan.edu.cn
http://www.fudan.edu.cn

摘要: 对数据质量,尤其是数据清洗的研究进行了综述.首先说明数据质量的重要性和衡量指标,定义了数据清洗问题.然后对数据清洗问题进行分类,并分析了解决这些问题的途径.最后说明数据清洗研究与其他技术的结合情况,分析了几种数据清洗框架.最后对将来数据清洗领域的研究问题作了展望.

关 键 词: 数据质量;数据清洗;数据集成;相似重复记录;数据清洗框架
中图法分类号: TP311 文献标识码: A

在当今时代,企业信息化的要求越来越迫切,其中一个很重要的方面就是企业数据的管理.根据“进去的是垃圾,出来的也是垃圾(garbage in, garbage out)”这条原理,为了支持正确决策,就要求所管理的数据可靠,没有错误,准确地反映企业的实际情况.因此,企业数据质量的管理正在获得越来越多的关注.数据质量管理牵涉到的方面很多,本文主要从数据集成和数据清洗的角度加以探讨.

最初,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.在研究数据集成的过程中,很多工作的重点放在如何解决模式冲突上.其实,在数据实例层次上同样有很多数据质量问题发生.数据清洗过程的目的就是要解决这些“脏数据(dirty data)”的问题.数据质量问题的一种情况是一个现实实体可能由多个不完全相同的记录来表示,这样的记录称为相似重复记录(duplicate record).为了检测并合并这些相似重复记录,研究人员提出了很多记录匹配算法.近年来,研究人员在数据清洗系统的框架、模型和语言以及如何利用专家知识、如何结合数据清洗过程和数据挖掘方法等方面做了很多工作.本文对与数据质量相关的将来可能的研究主题进行了展望.

1 研究背景

当建立一个信息系统的时候,即使进行了良好的设计和规划,也不能保证在所有情况下,所存放数据的质量都能满足用户的要求.用户录入错误、企业合并以及企业环境随着时间的推移而改变,这些都会影响所存放数据的质量.因此,有必要用元数据来表示数据质量[1,2].文献[1]以形式化的方法定义了数据的一致性(consistency)、正确性(correctness)、完整性(completeness)和最小性(minimality),而数据质量被定义为这 4 个指标在信息系统中得到满足的程度.文献[2]提出了数据工程中数据质量的需求分析和模型,认为存在很多候选的数据质量衡量指标,用户应根据应用的需求选择其中一部分.指标分为两类:数据质量指示器和数据质量参数.前者是客观的信息,比如数据的收集时间,来源等,而后者是主观性的,比如数据来源的可信度(credibility)、数据

* 收稿日期: 2002-03-12; 修改日期: 2002-07-02
基金项目: 国家自然科学基金资助项目(60003016);霍英东教育基金青年教师基金资助项目;教育部跨世纪优秀人才培养计划资助项目
作者简介: 郭志懋(1978－),男,湖南宁乡人,博士生,主要研究领域为数据清洗,XML 数据发布;周傲英(1965－),男,安徽宣城人,博士,教授,博士生导师,主要研究领域为数据挖掘,数据清洗,XML 数据管理,P2P 对等计算.

的及时性(timeliness)等.

在单个数据源中可能存在质量问题.例如,某个字段是一个自由格式的字符串类型,比如地址信息、参考文献等;错误的字段值,由于录入错误或者其他原因,数据库中一个人的年龄为 485 等.考虑多个数据源的情形,比如数据仓库系统、联邦数据库系统,或者是基于 Web 的信息系统,问题更加复杂.来自不同数据源的数据,对同一个概念有不同的表示方法.在集成多个数据源时,需要消解模式冲突,主要就是为了解决这个问题.还有相似重复记录的问题,需要检测出并且合并这些记录.解决这些问题的过程称为数据清洗过程.数据清洗(data cleaning,data cleansing 或者 data scrubbing)的目的是检测数据中存在的错误和不一致,剔除或者改正它们,这样就提高了数据的质量[3].

数据清洗过程必须满足如下几个条件:不论是单数据源还是多数据源,都要检测并且除去数据中所有明显的错误和不一致;尽可能地减小人工干预和用户的编程工作量,而且要容易扩展到其他数据源;应该和数据转化相结合;要有相应的描述语言来指定数据转化和数据清洗操作,所有这些操作应该在一个统一的框架下完成.

在模式转化和集成方面,人们已经做了很多的研究工作.而对于数据清洗,尽管工业界已经开发了很多数据抽取、转化和装载工具(ETL tool),但是它并没有得到足够多的研究人员的关注.一些研究人员研究相似重复记录的识别和剔除[4~8],还有一些与数据清洗相关的工作[9~12].在通过模式转化和集成获得了一致模式以后,在实例层次上仍然需要消除不一致性.同一个现实实体在两个数据源的记录中可能用不同的主键来标识,它们的信息可能存在冗余,有些互为补充,甚至有些互相矛盾.为了识别并且合并这些相似重复记录,研究人员提出了很多算法.这些算法有两个重要的评价标准:记忆率(recall)和准确率(precision).记忆率是识别出的相似重复记录占所有相似重复记录的百分比;准确率是指在算法识别出的相似重复记录里,那些真正的相似重复记录所占的百分比.一般来说,在这两个指标之间需要权衡,在提高记忆率的同时会损害准确率,反之亦然.文献[13]提出的办法较好地解决了这个问题,这种办法可以同时获得较高的记忆率与准确率.

绝大多数相关领域的研究人员认为,要很好地完成数据清洗过程,一定要结合特定应用领域的知识.因此,人们通常将领域知识用规则的形式表示出来.文献[13]利用专家系统的外壳,以方便规则的表示和利用.在清洗过程中,需要专家的干预.当系统碰到不能处理的情况时,报告异常,要求用户辅助作出决定.同时,系统可以通过机器学习的方法修改知识库,以后碰到类似情况时,它就知道怎样作出相应的处理了[14].

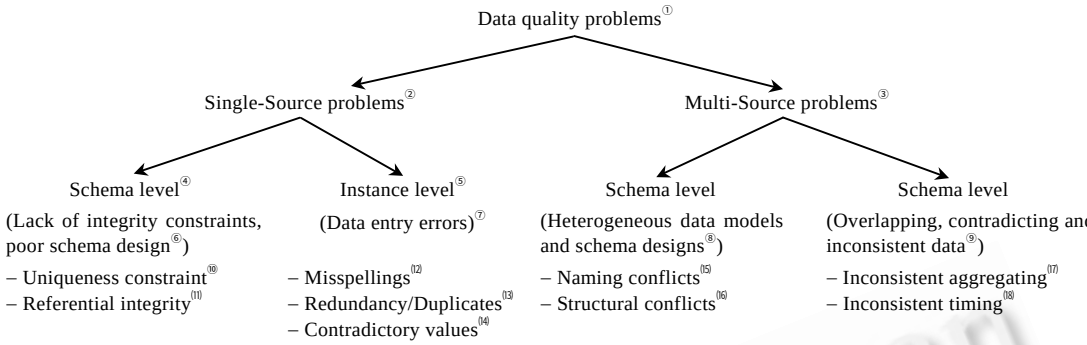
市场上的各种数据抽取、转化和装载工具或多或少提供了一些数据清洗功能,但是都缺乏扩展性[15].鉴于此,一些研究人员提出了数据清洗系统的框架.他们围绕这样的框架,提出了数据清洗的模型和语言.这些语言在 SQL 基础上扩展了新的数据清洗操作,比如 Merge,Cluster 等[16,17].有的框架模型采用了分层抽象的方法,最上面是逻辑层,可以用来定义数据清洗的流程,不需要关心具体采用的算法;最下面是物理实现层,这一层根据用户定义的逻辑流程,采用合适的算法和参数,对算法的优化过程也是在这一层完成的[16].

最初,数据清洗过程应用在数据仓库装载数据之前,其目的是为了提高数据的质量.这样,后继的 OLAP 和数据挖掘应用才可能得到正确的结果,决策支持系统才能够辅助管理者作出正确的决策.但是,一些研究人员提出,可以反过来将数据挖掘的技术应用到数据清洗过程之中[18].他们结合数据挖掘和数据清洗,利用数据挖掘技术可以发现数据中的模式和匹配规则.

2 数据质量问题的分类

根据处理的是单数据源还是多数据源以及问题出在模式层还是实例层,文献[3]将数据质量问题分为 4 类(如图 1 所示):单数据源模式层问题、单数据源实例层问题、多数据源模式层问题和多数据源实例层问题.图 1 表示了这种分类,并且分别列出了每一类中典型的数据质量问题.

单数据源情形中出现的问题在多数据源的情况下会变得更加严重.图 1 对多数据源没有列出在单数据源情形中就已经出现的问题.模式层次上的问题也会体现在实例层次上.糟糕的模式设计、缺少完整性约束的定义以及多个数据源之间异质的数据模型、命名和结构冲突等,都属于该类问题.可以通过改进模式设计、模式转化和模式集成来解决模式层次上的问题.实例层次上的问题在模式层次上不可见,一些可能的情况有数据拼写错误、无效的数据值、重复记录等.下面我们主要考虑实例层次上的问题.



①数据质量问题,②单数据源问题,③多数据源问题,④模式层,⑤实例层,⑥缺少完整性约束,糟糕的模式设计,⑦数据记录的错误,⑧异质的数据模型和模式设计,⑨冗余、互相矛盾或者不一致的数据,⑩唯一性约束,⑪引用约束,⑫拼写错误,⑬相似重复记录,⑭互相矛盾的字段,⑮命名冲突,⑯结构冲突,⑰不一致的汇总,⑱不一致的时间选择.

Fig.1 Classification of data quality problems
图 1 数据质量问题的分类

3 清洗实例数据

我们首先来看一些数据存在问题的情形[3],见表 1.

Table 1 Examples of problems at instance level
表 **1** 实例层上的数据质量问题的典型例子

Problem①	Dirty data②	Causes③
Missing values④	phone=9999–9999999	Unavailable values during data entry⑤
Misspellings⑥	city="Londo"	Error introduced during data entry⑦
Strange abbreviations⑧	Experience="B", occupation="DB Pro."	
Text in free-form⑨	name="J. Smith 12.02.70 New York"	Multiple values entered in one attribute⑩
Misfielded values⑪	city="Germany"	
Violated attribute dependencies⑫	city="Redmond", zip=77777	city and zipcode should correspond⑬
Word transpositions⑭	name1="J. Smith", name2="Miller P."	In a free-form field⑮
Duplicated records⑯	emp1=(name="John Smith"); emp2=(name="J. Smith")	Two records represent the same real-world entity⑰
Contradicting records⑱	emp1=(name="John Smith", bdate=12.02.70); emp2=(name="John Smith", bdate=12.12.70)	Some attribute of the same real-world entity is described by different values⑲
Wrong references⑳	emp=(name="John Smith", depno=17)	John Smith is not in the department with depno 17㉑

①问题,②脏数据,③原因,④缺少值,⑤录入数据时,不知道,⑥拼写错误,⑦录入时引入的错误,⑧不同的缩写,⑨自由格式的文本串,⑩单一字段中,存放了多种信息,⑪值与字段名不匹配,⑫字段之间不对应,⑬城市和邮政编码不对应,⑭词移位,⑮该字段无固定格式,⑯相似重复记录,⑰两条记录对应于同一个现实实体,⑱互相矛盾的记录,⑲同一个现实实体的某个属性有多个不同的值,⑳错误的引用,㉑John Smith 并不在 17 所对应的部门.

对于第 1 种情形,由于在数据输入时不知道电话字段的值,因此在数据库中以存放一个无效值来表示.如果针对电话字段定义一个规则存放在数据清洗库中,清洗工具就能够根据这条规则判断出哪些是无效值.对于第 2 种拼写错误的情形,需要在数据清洗库中建立一个存放所有城市名的查找表,通过与该查找表中的城市名相比较,就可以判断出数据库中存放的本来应该是哪个城市.对于第 3 种情况,一般也需要利用外部的查找表才能检测出来并加以改正.在数据清洗工具中,一些典型的查找表应该是内建的,此外也应该具备可扩展性,允许用户加入新的查找表.对于第 4 种情形,在一个自由格式的文本类型的字段里包括了很多部分,每个部分都可以单独作为一个字段.如果每个部分的先后顺序一定,且互相之间有分隔符或者保留字,比如 Street,Road 等等,就比较容易处理.但是,实际中的情况往往不是这样,因此要通过机器学习或者其他办法来解决.由领域专家选定学习样本(相对于所要处理的数据集,样本数量少得多)来训练系统,等训练好了以后,再由系统自动处理大规模的数据集.由于采用机器学习的办法,因此一般来说,需要折衷考虑记忆率和准确率.文献[19]提出了一种利用隐马尔科夫模型(HMM)的解决办法.

第 6 种情形的问题是字段之间不对应.为了改正,需要知道哪个字段更可信,这必须利用其他信息才能决定.

第 8 种和第 9 种情形表示的是相似重复记录的情况.在第 8 种情形里,一个记录的 name 没有简写,而另一个记录的 name 被简写了,通过定义合适的编辑距离函数,或者内建常用的缩写规则,清洗工具可以检测出这类重复记录.在第 9 种情形中,同一个现实实体(两个记录的 name 值相同),但是两个记录的 bdate 值不一样,在合并这两条记录时,如何选择一个合适的 bdate 值,是一个棘手的问题.相似重复记录的匹配和合并,是数据清洗过程中一个很重要的问题.首先,选择一个好的距离函数很重要.另外,记录的匹配过程非常耗时.如果采用最简单的方法,所有记录之间两两进行比较,以此来决定是否匹配,其计算复杂度为 $O(n^2)$,这里 n 为数据库中的记录数.对很大的数据库来说,这样的时间开销是无法忍受的.

在检测相似重复记录之前,需要先对数据进行一些处理.典型的处理操作包括:

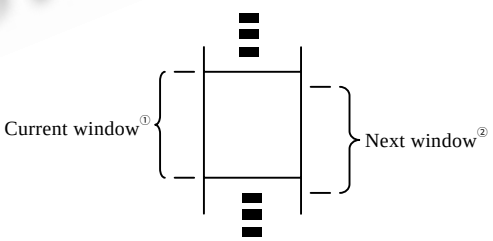
- 字段分裂.从自由格式的文本字段中抽取结构,分离各个部分.
- 验证和改正.根据查找表来验证字段值的正确性,若发现错误,则加以改正.如果提供合适的领域知识,该过程也可以验证字段之间的依赖关系.
- 数据标准化.将同一类型的数据用统一的格式来表示,比如日期、电话号码、性别等.

在完成大部分的数据转化和其他清洗步骤以后,就可以执行相似重复记录的匹配和合并了.文献[5,7,8]在清除重复记录上作了很多研究,提出了较好的算法.重复记录的匹配和合并也被称为对象标识问题和重复记录清除问题.通常情况下,指向同一个现实实体的两条记录的信息是部分冗余的,它们的数据互为补充[3].因此,通过将其合并,能够更准确地反映该实体.

相似重复记录清除可以针对两个数据集或者一个合并后的数据集.首先,需要识别出标识同一个现实实体的相似重复记录,即记录匹配过程.随后,将相似重复记录合并成一个包含该实体的更多属性,而且无冗余信息的记录,同时从数据集中删除多余的记录.

最简单的情况是,数据记录具有这样的属性集(或者属性),它总能够惟一标识一个实体.这时,只要对两个记录集在该属性集上作等值连接,就完成了记录匹配过程[3].对单个记录集的情形,先根据该属性集进行排序,然后通过检查相邻的记录,就可以判断出它们是否为相似重复记录.如果不存在这样的键属性集,而且数据中可能还存在错误,例如拼写错误等,上面的简单办法就不合适了.这时可以通过引入匹配规则来完成模糊匹配,规则是描述性的,而且可以利用用户自定义的函数[20].例如,可以有这样的规则:如果 name 字段相同,而且 address 字段相似度也很大,那么这两条记录是重复记录.字段之间的相似度,一般用 0~1 之间的数值来表示,而且不同的字段对记录之间总的相似度的贡献,具有不同的权值.相似度的定义和权值的分配,要由领域专家来确定.对字符串类型的数据,精确匹配或者基于通配符、词频、编辑距离、键盘距离和发音相似度的模糊匹配是很有用的[5,6],文献[8]还考虑了字符串的缩写形式.有些研究人员结合信息检索的向量空间模型来定义文本元素之间的相似度[21].

在处理大的数据集时,匹配重复记录是一个非常耗时的过程.因为是模糊匹配,所以整个过程相当于要对两个记录集做笛卡尔积.然后,根据相似度进行排序,那些相似度超过某一阈值的记录被认为是重复记录,低于某一阈值的记录被认为不是重复记录,而相似度介于这两个阈值之间的记录是候选的相似重复记录,需要用户作出决定.因为这类记录的数量不多,所以由用户来决定是可行的.



①当前窗口,②下一个窗口.

Fig.2 Window sliding during data cleaning

图 2 数据清洗中的窗口滑动

文献[5]提出在不同的键上多次排序,并分别计算邻近记录的相似度,最后综合多次计算的结果完成记录匹配过程.其具体方法是这样的,对单一数据源,每次排序在不同的属性集上进行.对排过序的记录集,如图 2 所示,只检查固定大小的窗口内部邻近的记录,看它们是否满足匹配规则.这样就大大减少了记录之间匹配规则的检查次数.综合多次排序的匹配结果,并计算传递闭包,最终就得到了所有的匹配记录对.

4 数据清洗与其他领域的结合

最初,数据清洗是数据仓库应用中数据准备阶段的一项重要工作.但是,它并不是一个全新的领域,相反,它可以借用其他很多领域的研究成果.文献[14]是先用抽样的方法从大的数据集中取出样本,在此基础上通过专家的参与产生预处理规则和匹配规则.在得到初步的规则之后,把它们应用到样本数据上,通过观察中间结果,用户可以修改已有规则,或者添加新的领域知识.如此反复,直到用户对所得结果满意为止.这时,就可以将这些规则应用到整个数据集中.系统利用了机器学习和统计方法来帮助建立匹配规则,以减少手工分析的工作量.

文献[13]提出了数据清洗的两个基准指标:记忆率和准确率,并分析了已有的相似重复记录的匹配方法,认为这些方法在提高记忆率的同时会损害准确率;反之亦然.这被称为记忆率-准确率两难选择(recall-precision dilemma).文章指出,多次排序的算法虽然提高了记忆率,但是很可能降低了准确率.如果记录A和记录B很相似,记录B和记录C也很相似,通过求闭包的过程,就得出结论A和C也是相似记录.实际上,A和C可能已经相差很大了.由于数据清洗过程是领域相关的,领域知识是成功的数据清洗中一个必不可少的部分,所以文献[13]提出了一个基于知识管理的智能型数据清洗系统的框架,他们采用专家系统,用规则来表示领域知识,实现了知识的高效表示和灵活管理.文章通过指定有效的规则,并且在传递闭包的计算过程中引入不确定因子,在一定程度上解决了记忆率-准确率两难问题.

文献[18]提出使用数据挖掘的办法来检测数据质量问题,并加以改正.例如,通过发现字段之间的关联规则,如果该关联规则的置信度非常接近 100%,那么违反该规则的记录很可能有问题.文献[22]提出了一种高效的基于N-Gram的聚类算法,复杂度仅为$O(n)$,而且能够较好地聚类相似重复记录.

5 数据清洗框架

文献[16]分析了已有的数据清洗方法和工具.一部分数据清洗工具只提供了有限的清洗功能,另一部分则专门针对数据清洗.对一些常规的应用领域,比如在美国对客户数据库的清洗,人们已经积累了足够多的经验,知道该如何来设计和实现这样的一个数据清洗程序.他们知道需要对数据进行哪些转化,需要哪些操作符和哪些参数.但是,对那些非常规的应用领域,比如将大量的历史遗留数据迁移到关系数据库中,已有的数据清洗平台不能有效地支持开发新的数据清洗程序.人们很难设计一个能够高效执行的数据处理流程,其中的原因是:(1) 数据转化缺少逻辑规范和物理实现的分离;(2) 这些工具不记录中间数据的来龙去脉,而且缺乏允许用户交互的设施,因此用户很难分析和逐步调整数据清洗过程.因此,文献[16]提出了一个数据清洗框架,试图清晰地分离逻辑规范层和物理实现层.用户在逻辑层设计数据处理流程,确定清洗过程需要执行的数据转化步骤;物理层实现这些数据转化操作,并对它们进行优化.例如,用户为了计算记录之间的相似度,对输入数据集指定执行匹配操作,而在物理层,数据清洗系统会根据实际数据的特点选择一个特定的实现方法.该算法可以进行优化,它无须计算所有记录对的相似度,而同时又不会损害所要求的结果.除了分离逻辑层和物理层以外,文献[16]提出了一种描述性语言.该描述性语言可以在逻辑层上指定数据清洗过程所需采取的数据转化操作,并指定何时可以弹出例外,要求用户的交互.该描述性语言还可以指定一些数据转化操作的参数,比如记录匹配操作所使用的距离函数等.文献[16]实现了一个可扩展的数据清洗工具AJAX[4],其实验结果证明了该框架的价值.

文献[17]提出了数据清洗的一个交互式系统框架,它紧密地集成数据转化和差异检测(discrepancy detection).用户面对表单风格的界面,能够以直观的图形化方式逐步建立起整个数据转化过程.在此过程中,用户可以执行或者撤消转化操作,而且操作的结果马上就可以在屏幕上看到.后台进程以增量方式检测转化后的数据中存在的问题,如果检测到,则及时报告给用户.用户利用系统提供的基本的数据转化操作,无须书写复杂的程序就能够完成数据清洗任务,而且用户能够随时看到每一步转化操作后的结果,没有很长的延迟.因此,这个系统框架的交互性很好.

6 结论和展望

数据清洗问题的重要性是不言而喻的.从市场上如此多的相关产品,可以明白这一点.然而,目前在学术界,

它并没有得到足够的关注.有些人认为数据清洗是一个需要大量劳动力的过程,而且往往过于依赖特定应用领域.其实不然,在数据清洗系统的灵活框架上仍然有很多东西值得研究.在多语言环境中,如何准确地识别多语言文本的相似重复记录,也有很多工作可以做.文献[23]在这方面做了一些工作.当前 Web 数据量迅速增长,对 Web 搜索引擎返回的结果进行清洗也是一个有价值的问题.随着 XML 数据处理标准的日见成熟,如何定义 XML 文档数据的质量标准以及针对它们的清洗过程和针对关系数据的清洗过程的区别,都是值得研究的.文献[24]提出了 XML 键的概念,完全有理由相信它们可以促进针对 XML 数据的清洗,正如关系表的键在数据集成中扮演了特殊的角色一样.

References:

[1] Aebi, D., Perrochon, L. Towards improving data quality. In: Sarda, N.L., ed. Proceedings of the International Conference on Information Systems and Management of Data. Delhi, 1993. 273~281.

[2] Wang, R.Y., Kon, H.B., Madnick, S.E. Data quality requirements analysis and modeling. In: Proceedings of the 9th International Conference on Data Engineering. Vienna: IEEE Computer Society, 1993. 670~677.

[3] Rahm, E., Do, H.H. Data cleaning: problems and current approaches. IEEE Data Engineering Bulletin, 2000,23(4):3~13.

[4] Galhardas, H., Florescu, D., Shasha, D., *et al*. AJAX: an extensible data cleaning tool. In: Chen, W.D., Naughton, J.F., Bernstein, P.A., eds. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Texas: ACM, 2000. 590.

[5] Hernandez, M.A., Stolfo, S.J. Real-World data is dirty: data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 1998,2(1):9~37.

[6] Lee, M.L., Ling, T.W., Lu, H.J., *et al*. Cleansing data for mining and warehousing. In: Bench-Capon, T., Soda, G., Tjoa, A.M., eds. Database and Expert Systems Applications. Florence: Springer, 1999. 751~760.

[7] Monge, A.E. Matching algorithm within a duplicate detection system. IEEE Data Engineering Bulletin, 2000,23(4):14~20.

[8] Monge, A.E., Elkan, C. The field matching problem: algorithms and applications. In: Simoudis, E., Han, J.W., Fayyad, U., eds. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Oregon: AAAI Press, 1996. 267~270.

[9] Savasere, A., Omiecinski, E., Navathe, S.B. An efficient algorithm for mining association rules in large databases. In: Dayal, U., Gray, P., Nishio, S., eds. Proceedings of the 21st International Conference on Very Large Data Bases. Zurich: Morgan Kaufmann, 1995. 432~444.

[10] Srikant, R., Agrawal, R. Mining Generalized Association Rules. In: Dayal, U., Gray, P., Nishio, S., eds. Proceedings of the 21st International Conference on Very Large Data Bases. Zurich: Morgan Kaufmann, 1995. 407~419.

[11] Abiteboul, S., Cluet, S., Milo, T., *et al*. Tools for data translation and integration. IEEE Data Engineering Bulletin, 1999,22(1):3~8.

[12] Milo, T., Zohar, S. Using schema matching to simplify heterogeneous data translation. In: Gupta, A., Shmueli, O., Widom, J., eds. Proceedings of the 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann, 1998. 122~133.

[13] Lee, M.L., Ling, T.W., Low, W.L. IntelliClean: a knowledge-based intelligent data cleaner. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000. 290~294.

[14] Caruso, F., Cochinwala, M., Ganapathy, U., *et al*. Telcordia's database reconciliation and data quality analysis tool. In: Abbadi, A.E., Brodie, M.L., Chakravarthy, S., *et al*., eds. Proceedings of the 26th International Conference on Very Large Data Bases. Cairo: Morgan Kaufmann, 2000. 615~618.

[15] Galhardas, H. Data cleaning and integration. 2001. http://caravel.inria.fr/~galharda/cleaning.html.

[16] Galhardas, H., Florescu, D., Shasha, D., *et al*. Declarative data cleaning: language, model and algorithms. In: Apers, P., Atzeni, P., Ceri, S., *et al*, eds. Proceedings of the 27th International Conference on Very Large Data Bases. Roma: Morgan Kaufmann, 2001. 371~380.

[17] Raman, V., Hellerstein, J. Potter's wheel: an interactive data cleaning system. In: Apers, P., Atzeni, P., Ceri, S., *et al*, eds. Proceedings of the 27th International Conference on Very Large Data Bases. Roma: Morgan Kaufmann, 2001. 381~390.

[18] Hipp, J., Guntzer, U., Grimmer, U. Data quality mining: making a virtue of necessity. In: Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara, 2001.

[19] Borkar, V., Deshmukh, K., Sarawagi, S. Automatically extracting structure from free text addresses. IEEE Data Engineering Bulletin, 2000,23(4):27~32.

[20] Hellerstein, J., Stonebraker, M., Caccia, R. Independent, open enterprise data integration. IEEE Data Engineering Bulletin, 1999,22(1):43~49.

[21] Cohen, W. Integration of heterogeneous databases without common domains using queries based on textual similarity. In: Haas, L., Tiwary, A., eds. Proceedings of International Conference on Management of Data. Seattle: ACM Press, 1998. 201~212.

[22] Qiu, Yue-feng, Tian, Zeng-ping, Ji, Wen-yun, *et al*. An efficient approach for detecting approximately duplicate database records. Chinese Journal of Computers, 2001,24(1):69~77 (in Chinese).

[23] Yu, Rong-hua, Tian, Zeng-ping, Zhou, Ao-ying. A synthetical approach for detecting approximately duplicate database records of multi-language data. Computer Science, 2002,29(1):118~121 (in Chinese).

[24] Buneman, P., Davidson, S., Fan, W., *et al*. Keys for XML. In: Proceedings of the 10th International World Wide Web Conference. Hong Kong: ACM Press, 2001. 201~210.

附中文参考文献:

[22] 邱越峰,田增平,季文赟,等.一种高效的检测相似重复记录的方法.计算机学报,2001,24(1):69~77.

[23] 俞荣华,田增平,周傲英.一种检测多语言文本相似重复记录的综合方法.计算机科学,2002,29(1):118~121.

Research on Data Quality and Data Cleaning: a Survey*

GUO Zhi-mao, ZHOU Ao-ying

(*Department of Computer Science and Engineering, Fudan University, Shanghai* 200433, *China*);

(*Laboratory for Intelligent Information Processing, Fudan University, Shanghai* 200433, *China*)

E-mail: zmguo@fudan.edu.cn

http://www.fudan.edu.cn

Abstract: Data quality, especially data cleaning, is surveyed in this paper. The importance of data quality, and its measurement metrics are described. The data cleaning problems are defined and classified. The approaches to solving data quality problems are detailed. How to combine the techniques in other research areas with data cleaning is overviewed, and several data cleaning frameworks proposed previously by others are introduced. The future research topics related to data cleaning problems are also discussed.

Key words: data quality; data cleaning; data integration; duplicate record; data cleaning framework

* Received March 12, 2002; accepted July 2, 2002

Supported by the National Natural Science Foundation of China under Grant No.60003016; the Young Teacher Foundation of Huo Yingdong Education Foundation of China; the Cross-Century Excellent Talent Raising Program of Education Ministry of China