数据质量研究综述*)

韩京宇[1] 徐立臻[2] 董逸生[2]

(中科院软件所基础软件国家工程中心 北京 100080)[1] (东南大学计算机科学与工程系 南京 210096)[2]

摘 要 数据质量管理是信息系统建设的首要问题。本文首先回顾了数据质量的定义和质量提高策略的分类,然后对数据质量研究涉及的两个主要方面,即数据质量评估和数据质量提高技术的各种方法进行了比较和分析,并对有代表性的数据质量提高工具进行了介绍。最后提出了一个评估驱动的数据质量提高框架,并对数据质量研究方向进行了展望。

关键词 数据质量,数据清洗,机器学习,数据审计

An Overview of Data Quality Research

HAN Jing-Yu[1] XU Li-Zhen[2] DONG Yi-Sheng[2]

(National Engineering Research Center for Fundamental Software of Institute of Software, Chinese Academy of Sciences, Beijing 100080)[1]

(Department of Computer Science and Engineering, Southeast University, Nanjing 210096)[2]

Abstract Data quality management is an essential problem for information systems. First, the definitions of data quality are overviewed and the strategies for improving quality are summarized. Then, the two main aspects of data quality research, that is data quality assessment and data quality improvement methods are analyzed respectively. At last, some data quality tools are briefly touched on. Based on above analysis, an assessment-driven data improvement framework is proposed and the future research directions are discussed.

Keywords Data quality, Data cleansing, Machine learning, Data auditing

随着信息处理技术的不断发展,各行各业已建立了很多计算机信息系统,积累了大量的数据。为了使数据能够有效地支持组织的日常运作和决策,要求数据可靠无误,能够准确地反映现实世界的状况。数据是信息的载体,好的数据质量使各种数据分析(如 OLAP 分析、数据挖掘等)能够得到有意义结果的基本条件。人们常常抱怨所谓的"数据丰富,信息贫乏",其中一个原因是缺乏有效的数据分析技术,而另一个重要原因则是数据质量不高,如数据残缺不全、数据不一致、数据重复等,导致数据不能有效地被利用[18]。数据质量管理如同产品质量管理一样贯穿于数据生命周期的各个阶段[65],但目前尚缺乏一个系统的思路。数据质量的研究由来已久,涉及到统计学、人工智能、数据库等各个领域[65]。

1 数据质量定义及提高策略分类

数据质量评估是数据管理面临的首要问题。目前对数据质量有不同的定义[15],文[33]认为数据质量是数据适合使用的程度(fit for use)。文[6,10]认为数据质量是数据满足特定用户期望的程度。文[14]认为,数据质量主要指一个信息系统在多大程度上实现了模式(schema)和数据实例(instance)的一致性,及模式和数据实例在多大程度上实现了正确性(correctness)、一致性(consistency)、完整性(completeness)和最小性(minimality)。

提高数据质量的策略多种多样,可以从不同的角度进行分类。数据质量提高策略可以分别从问题的发生时间或质量问题解决依赖的知识两个角度来划分。从数据的整个生命周期来看,数据质量提高主要分两个角度来考虑:一类是从预防的角度,即在数据生命周期的任何一个阶段,都有严格的数据规划和约束来防止脏数据产生。另一类是事后诊断,即由于数据的演化或集成,会有脏数据逐渐涌现,须采取特定的算法检测出现的脏数据[65]。从数据质量问题解决依赖的知识来看,数据质量提高策略分成两类:一类提高策略不依赖特定业务规则,是应用独立的,如数据拼写错误、数据分布异常、某些缺失值处理等,这类问题的解决不依赖于特定的业务规则,可以从数据本身中寻找特征来解决[1,35];另一类解决方法与特定业务规则相关,是应用依赖的,这些相关的领域知识是消除数据逻辑错误的必需条件[74]。由于数据质量问题涉及方方面面,成功的数据质量提高方案必然是综合应用上述各种策略。目前,数据质量的研究主要围绕两个方面展开:(1)数据质量的评估和监控;(2)从技术的角度如何保证和提高数据质量。下面第 2、3 节分别对此加以分析。

2 数据质量评估

数据质量评估和监控是解决数据质量问题的一个源头性问题。尽管对数据质量的涵义有不同的看法[10,78,14,69,2,5],但一般认为数据质量是一个层次分类(category)的概念,每个质量类最终分解成具体的数据质量维度[15]。数据质量评估的核心在于如何具体地评估各个维度,目前方法主要分成两类:定性的策略和定量的策略。

对各个维度从定性的角度来分析其"好"或"坏",这是目前数据质量评估方法的主流。文[14]从理论的角度对数据质

*)江苏省"十五"高科技项目(BG2001013)。**韩京宇** 博士,助理研究员,主要研究方向为数据质量、数据仓库、移动对象数据库等。**徐立臻** 教授,主要研究方向为数据网格、信息集成等。**董逸生** 教授,博士生导师,主要研究方向为数据库、信息集成等。

量从若干个维度进行分析：准确性(accuracy)、完整性(completeness)、一致性(consistency)和最小性(minimality)等，但没有探讨具体的维度评估方法。文[58]认为数据工程中有许多数据质量衡量指标，分成客观的数据质量指示器(data quality indicator)和主观的数据质量参数(data quality parameter)。它提出用户应根据需要选择重要的指标进行衡量。文[18]提出应从一个更宽广的角度来理解数据质量，它将数据质量评估指标分成四类：内在质量(intrinsic DQ)、可访问性质量(Accessibility DQ)、上下文质量(Contextual DQ)、表达质量(Representational DQ)，每个类又细分成具体的维度来评估，从而拓宽了仅仅对内在质量关注的认识。文[78]提出应在本体的概念上来理解数据质量各个维度的含义，避免数据质量维度的定义缺乏统一的语义基础。数据新鲜度(freshness)在分布式系统的数据利用中尤为重要，文[44]从数据抽取延误(currency)和原始数据更新频度(timeliness)两个因素进行了分析。上述这些方法在评估每一个具体的维度时，采取定性分析的方法获得某一个具体维度的评估值，缺乏客观、量化的分析。

由于定性的分析缺乏客观性和可重现性，定量评估技术成为一个值得关注的方向，目前这个方面的研究主要集中在关系数据库数据的质量评估技术。文[4,5]采取取样计算的方法，对关系数据库数据质量的两个重要维度即精确度(accuracy)和完整度(completeness)进行量化，并具体分析了数据质量对四种常见的关系代数操作：选择、投影、笛卡尔积、连接的影响。文[2]等提出对关系数据库划分成不同的数据质量区域，分别计算各区域的数据质量好坏，从而获得相对准确的质量评估值。但这种量化评估方法要靠人工逐一验证，在实际执行时是很困难的。文[49]提出对于特定的领域可以通过元数据来定义质量视图(quality view)，通过这个质量视图来指导具体的数据处理过程。

可以看到数据质量的评估尽管已引起大家的重视和研究，但如何客观有效地进行质量评估，仍亟待进一步探索。

3 数据质量提高技术

数据质量提高技术主要涉及实例和模式两个层面[56]。数据清洗(data cleansing, data scrubbing)是数据质量提高技术研究的主要内容，它主要关注数据实例层面的问题，主要集中在几个方面：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。在数据库界，关于模式层的研究很多，而站在数据质量提高技术的角度主要讨论如何根据已有的数据实例重新设计和改进模式。下面分别论述。

3.1 数据清洗研究

数据清洗主要研究如何检测并消除数据中的错误和不一致，以提高数据质量[18]。数据清洗的研究主要是从数据实例(instance)层的角度考虑来提高数据质量。

3.1.1 重复对象检测

现实世界中的实体在不同的数据源中常常有多个表达，在数据集成的时候经常要判断不同数据源中的表达是否代表现实世界中的同一实体[44,45,47,62]，它是数据清洗研究的一个重要方面。这类问题的解决主要是采用数据库和人工智能的方法。目前的研究主要集中在两个方面：一方面是关系数据库数据的重复记录检测，另一方面是 XML 重复元素检测。

(1)重复记录检测

语法上相同或相似的不同记录可能代表现实世界中的同一实体，因此结构化数据的相似重复记录检测是数据清洗研究中的重要问题。相似重复记录识别主要有以下几类方法：排序 & 合并的方法[1,41]、建索引的方法[34,61]、机器学习的方法[60,40,22]、根据上下文信息识别[44,27]、基于特定领域知识的方法[74,59]、根据数据特征的方法[62,59]。

传统的排序 & 合并方法是一种通用的、不依赖于特定应用领域的识别算法。文[41]采用排序 & 合并的方法，根据用户选定的若干个属性字符串作为键进行排序后，采用固定大小的滑动窗口进行聚类来识别相似重复记录；文[1]根据不同的属性进行多次排序合并，并采用优先队列取代固定大小的滑动窗口来进行聚类。这两种方法字符串外排序引起的高频度的 I/O，是很耗资源的。另外由于字符串排序对字符和单词位置太敏感，并不一定能保证将相似的重复记录排在邻近的位置，导致随后的聚类操作不一定能将相似重复记录识别出来。文[24,52]提出采用 n-gram 聚类的方法来解决相似重复记录的识别。

采用 R 树建索引的方法[34]首先依据 fastmap 方法[11]选取若干个字符串作为轴(pivot)，各个记录依据这若干个轴计算它在多维空间中的坐标，然后采用 R 树进行多维相似性连接来实现相似重复记录的识别。由于“维度灾难”(curse of dimensionality)决定了维度不能过高，使得这种方法不具有通用性。文[61]提出一种具有代表性的在线数据清洗方法，它是在有干净(clean)参照表的条件下进行数据清洗的方法，其基本思路是首先对干净的参照表数据建立一个 ETI(error tolerance index)索引，每一个在线输入的数据根据这个索引迅速找到与之最匹配的干净记录，然后用它来取代，从而完成对输入数据的在线清洗。

文[40]提出通过学习不同类型的字符串记录的相似性度量函数和相似度阈值来提高重复检测的精度，文[60]提出采用主动学习的方法，通过少量的样本学习即可以获得好的检测精度。在文[68]中将重复记录检测当作分类问题来解决，任何一对数据库记录可以根据其相似度划分到三个决策空间：重复、不重复和可能重复，据此通过学习贝叶斯分类模型来解决重复记录检测问题。Telcordia[20]是一个具有代表性的采用机器学习的方法进行模糊重复记录识别的系统。机器学习方法的最大优点在于尽可能自动化地发现重复识别规则，减少人工干预的工作量。

重复记录识别有一个重要的现象是仅仅根据数据内容判断重复常常会失效。文[44]提出用上下文信息消除可疑链接(spurious link)，即对于无法确认的相似重复记录根据与它相关的隐藏在其它记录中的上下信息计算上下文相似度来确认。文[27]首次提出用迭代的方法来解决作者识别的问题，它根据作者名字的选择性和同时发生(co-occurence)的频率来作为识别的依据。这类利用上下文相关信息的思想为识别疑难重复记录对象提供了一条好的途径。

现实的信息系统都是面向某一个具体应用的，都与特定的业务规则相联系，为此文[74,22]提出了根据特定领域知识建立规则来识别相似重复记录的框架。这类方法能够根据业务规则，取得好的识别精度。这种方法的主要问题在于：为了识别相似重复记录，必须建立相应领域的规则库，对领域知识的要求比较高，人工定义的工作量比较大。

文[62]提出了一种鲁棒性强的模糊重复记录识别方法，它根据模糊重复所具有的紧缩集(compact set, CS)特征，即重复记录间比不重复的记录间更临近和重复记录的本地邻居

是稀疏的特征(sparse neighbor, SN)来识别相似重复记录。文[59]提出在数据仓库环境下,根据维表中的维度层次逐步将多维数据表中的重复记录识别出来。这两种方法都是根据数据特征来解决问题。

(2)重复 XML 元素检测

XML 数据作为具有代表性的半结构化数据,已成为网上数据传输和交换的标准,例如现在涌现出的大量 RSS 数据都是 XML 格式的。当多个不同数据源以 XML 格式来描述现实世界的实体时,格式和内容不尽相同的元素可能代表现实世界中的同一实体。相比于传统的关系数据库数据,识别层次状的 XML 数据中的重复元素时,要面临着两个挑战性的问题:一是结构的多样性[39],不同于具有严格数据模式的关系数据库数据,同一类型的实体可能采用不尽相同结构的元素来描述;另一方面,不同于平面关系数据库数据,层次状的 XML 数据具有复杂的元素和子元素间的依赖关系。这些使得识别重复的 XML 元素成为一个更具有挑战性的问题。

文[39]中给出了一种在 XML 文档中识别重复 XML 元素的方法。为了解决 XML 元素多样性的问题,它采用 XQuery 语言将具有不同子树结构的 XML 元素变换成统一的结构,并将同层次的 XML 元素的内容合并为一个元素来处理。但这种方法由于混淆了具有不同标签(元素名)的数据进行相似性计算,会损失精度。文[12]提出将具有相似结构的 XML 元素进行合并的方法,它提出了三类启发式聚类算法来实现相似重复元素的合并:全部比较聚类、选择比较聚类、M 树聚类方法来实现重复元素的有效合并。其主要不足在于它没有解决 XML 数据结构多样性的问题。归结起来,XML 数据清洗还有三个方面的问题有待于进一步的探索:(1)如何有效地处理结构多样性的问题;(2)XML 数据是一种层次状的数据,如何发现复杂的父元素和子元素间的依赖关系;(3)如何有效地降低计算的复杂度。

3.1.2 缺失数据(missing value, missing data)处理

实际的数据集中经常会存在缺失数据,常常会对分析结果有很大的影响[75]。如何尽可能合理地将缺失数据填补是一个重要的问题。缺失数据处理的研究主要分布在统计领域和数据库领域。

在统计领域关于缺失值处理的研究由来已久,主要分成单一填补法(single imputation)和多重填补法(multiple imputation)。其中单一填补法是指对缺失值构造单一替代值来填补,常见的方法有取平均值或中间数填补法、回归填补法、最大期望(expectation maximization)填补法、hot deck 填补等方法[30,57]。其中 hot deck 填补法也叫就近补齐,是指采用与有缺失的观测最"相似"的那条观测的相应变量值作为填充值。但单值填充方法常常不能反映原有数据集的不确定性,会造成较大的偏差。多重填补法是指用多个值来填充,然后用针对完整数据集的方法对它们进行分析得出综合的结果。比较常见的有趋势得分法(propensity score, ps)和(predictive mean matching, PMM)法等。这类方法的优点在于通过模拟缺失数据的分布,可以较好地保持变量间的关系,其缺点在于计算复杂。

在数据库研究领域,文[37]提出对于含有空缺值的数据的相似记录的度量方法。文[76]提出一种对多维数据集的度量值的缺失数据进行填补的方法,它采用对数模型和对数线性模型来填补数据立方的缺失值。文[76]还提出一种利用约束来实现数据立方缺失值填充的方法。文[TW84]提出利用条件表(conditional table)来实现对关系数据库的数据表缺失数据的填补。文[27]提出采用 K 近邻方法来实现对缺失数据的填补。文[54]用马尔可夫链蒙特卡罗法(Markov chain Monte Carlo MCMC)实现对缺失数据的填充。需要指出,缺失值填补主要是为了防止数据分析时,由于相当部分的值空缺导致的分析偏差。但这种填补方法,对于填补的单个数据,只具有统计意义,不具有个体意义。

3.1.3 异常数据检测

数据中异常一般是由两种原因造成的:一是数据固有变异性造成的,另外一种则是由于度量或执行错误导致的,在数据清洗时这两者都应予以关注。在数据清洗领域对异常数据的自动化发现主要采用数据审计的方法来解决[15],也称为数据质量挖掘(Data Quality Mining, DQM)[31]。其基本步骤主要分两个环节来解决,第一步是数据概化(data profiling),即采用数理统计的方法对数据分布进行概化(profiling)描述,以自动化地获得数据的总体分布特征[19,28],以此作为进一步分析的基础。然后针对某一特定的数据质量问题进行挖掘以发现异常。DataGliches[66]首先将数据根据距离划分成层(layer),然后对每个层统计数据特征,根据定义的距离计算层中的各个数据点与中心距离的远近来判断可能的异常。文[31]提出采用关联规则挖掘的方法来实现标称型(Nominal)异常数据的发现,但这种方法不能直接应用于数值型属性的处理。文[15]指出采用数据审计的方法发现异常数据的效果,很大程度上依赖于数据挖掘算法能否准确分辨异常和非异常数据,为此它采用 C4.5 的决策树算法来产生模拟数据,用来检测和验证特定的数据挖掘算法发现偏差数据的能力。传统的数据挖掘方法在异常数据检测方面积累了大量方法如基于统计模型的、基于距离的、基于偏离的等。但在数据清洗领域,由于数据常常是不干净的,在特定的挖掘算法执行前的数据概化(data profiling)常常是非常重要的,这有助于探测型挖掘(exploratory mining)的进行[65],更有针对性地发现异常数据。

3.1.4 逻辑错误检测

实际的信息系统都是面向某一个应用领域的,对于一个具体的应用如何采用自动化的方法来解决数据中不符合业务逻辑的错误,是一个有实际应用价值的问题[25,71]。这类问题是数据编辑修正(editing and imputation)所研究的主要内容,其思路是根据应用依赖的领域知识建立规则体系来自动处理。1976 年,Fellegi[25]等提出了一个严格的形式化的数学模型(Fellegi-Holt 模型),后来几乎所有的成功的数据审计系统都在不同程度上采用了这一模型。其基本思想是希望对记录的各个变量做最少的改动,以满足所有的编辑规则的要求。主要思路是:对于具体的应用根据领域知识定义显式约束规则(explicit rules),然后根据一定的数学方法求出规则的整个闭集,对于每条记录,自动判断其是否违反规则约束。这种方法的优点在于:数学基础严密,规则自动生成,发展相当成熟。它在审计、统计领域已广泛地成功应用。根据这一思想,一些统计系统 SPEER[71]和 DISCRETE[71]等相继出现。鉴于实现 FH 的关键在于(error localization)问题的解决,文[64]提出了一个快速解决算法。

3.1.5 不一致数据处理

多个数据源的数据集成清洗的时候,几个独立维护的数据源经常提供相互重叠的数据内容,会出现不一致的数据,如何从若干个不一致的数据结果中获得理想的数据答案是数据

清洗中经常面临的问题[73]。目前常用的方法有排序、融合和根据规则的方式[3]。DIRECT[73]系统中将数据间的不一致划分成上下文独立的冲突(context independent conflicts)和上下文依赖的冲突(context dependent conflicts)两种。上下文依赖的冲突是指不同数据源的数据由于不同系统或应用间固有的数据设计和表达因素造成的,这种数据冲突须用数据转换规则来解决。上下文独立的冲突是指由于一些外部的随机造成的不一致,这类问题的解决一般需要人工干预和特定的方法。DIRECT提出对于上下文依赖的冲突可以采用机器学习的方法学习转换规则来解决不一致的问题。文[3]对每个数据值从不同性能参数,即特征(feature)来评估,整体的评估值是各个特征评估值的线性组合,然后根据这个评估值确定哪个值为唯一正确的。Standford大学的Trio系统提出了一种完备的处理遗留(lineage)和不一致数据的扩展关系模型[46],对不一致数据的处理从数据模型的角度给出了一条新的思路。

3.1 质量相关的模式问题研究

数据模式的研究有很多,但站在数据质量的角度,其研究关注的问题侧重如何理解数据[67]和如何根据已有的数据实例重新设计数据模式[50]。Evoke提出采用数据挖掘的方法对数据的分布特征进行概化,以发现各个属性的数据值分布状况、数据冗余和函数依赖关系,以帮助数据质量的检测和数据模式的改造。Bellman[67]系统采用数据挖掘的方法对数据进行分析,力图自动化地理清数据结构,揭示数据涵义。不同于Evoke系统,它不仅自动化地对表的记录、字段内容信息进行概括,而且能有效地发现不同表间字段的对应关系和主外键的依赖关系等,为多数据源的自动化的数据清洗提供了有力的帮助。文[50]采用信息论的理论对数据进行聚类,这种方法可以克服对不同类型的数据、距离函数的定义困难的问题,有效地发现不同属性的函数依赖关系,以优化模式设计。应该看到,自动化地进行数据模式再设计是密切依赖于数据实例的,是数据清洗时必须面对的问题。

4 若干个有代表性的数据质量框架和工具

数据质量和数据清洗的工具有很多,它们有各自不同的特色和侧重点。

数据审计(data auditing)作为自动化数据清洗的方法,其思路是采用机器学习的方法来发现数据中蕴涵的语义结构,与此不符的数据作为数据清洗时关注的对象。根据这一思路,文[15]提出了一个具有代表性的数据质量提高框架。首先,根据应用由领域专家分析确认数据的分布特征参数。然后,采用不同机器学习算法来推导数据隐含的结构并进一步发现数据偏差。这个过程不断调整,直到正确识别偏差的比率用户满意为止。为了比较不同算法检测脏数据的效果,它提出了一套系统的模拟数据生成方法,以解决算法学习中要人工判断数据对错的问题。最后这种定制的数据审计算法被数据质量引擎调用,完成对实际脏数据的检测和错误纠正。这个工具在QUIS(QUailty information system)应用,证明是有效的。

DaQuinCIS[38]提出了一个在协作信息系统中通过各个不同数据源数据的比较、纠正来提高整个协作信息系统的数据质量的框架。这个框架主要分成两个部分:数据质量代理和数据质量通知服务,分别负责数据质量的交换和数据质量的改进。系统采用一个D^2Q模型来描述数据和数据的质量,D^2Q的数据模式用XML数据来表达,作为各不同子系统间数据质量元数据交换的标准。GuardianIQ[13]是一个具有代表性的基于业务规则的数据质量提高工具。它提出数据是面向应用的,数据质量的好坏要在具体的上下文环境中衡量,产生数据质量问题的原因也应根据具体的应用环境来寻找:同一种现象可能是由不同的原因造成的,因此具体的数据质量问题解决方法必须根据应用灵活地定制。在GuardianIO系统中,根据不同的业务规则可以自动生成相应的可执行的数据清洗代码,实现对特定应用的数据清洗。另外,Telcordia[20]、Intelliclean[74]等都是具有代表性的依赖业务规则进行数据清洗工具。这类工具由于充分考虑了具体应用的特殊性,具有很好的实用性。其它还有许多类似的数据清洗工具。AJAX[23]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。它在SQL语言的基础上提出了一种从逻辑上描述数据清洗转换过程的语言,并在物理实现层给出了具体的优化算法。Potter wheel是一个交互式的数据清洗框架[26],辅助用户渐进式地定义数据的清洗转换规则,提供即时的反馈。

5 数据质量驱动的研究框架和相关展望

由于数据质量问题涉及应用依赖和应用独立两个方面的知识。在整个数据生命周期中,原有的数据质量问题解决了,往往还会发现有新的质量问题,这主要表现为两点:其一是质量问题的某些"症状"会随着另外一些"症状"的解决而显现;其二是随着时间的推移和数据的演化,会有新的数据质量问题产生。因此,不能指望任何一种方法能够毕其功于一役。数据质量的保证和提高遵循这样一个过程:数据质量分析→发现问题→应用独立清洗→应用依赖清洗→数据质量分析。在这个不断反复的过程中,数据中的问题逐步被发现解决,从而使数据质量得到保证和提高。这个过程周而复始,伴随数据的整个生命周期。鉴于以上的原因,我们提出三个同心环的数据质量研究框架:居于核心的是质量维度监控评估;中间一层是不依赖于知识的数据清洗,如重复消除、缺失数据处理、异常数据处理等;最外层是依赖于应用逻辑的数据清洗。在这个框架下系统地逐步解决数据质量问题。在这个框架下,我们强调数据质量的评估居于核心位置,它是数据质量问题捕捉和清洗的基础,是系统地解决问题的关键。

尽管数据质量的研究取得了一定的成果,但在许多方面尚待进一步探索:(1)现在的许多系统都假定数据质量的元数据是可得的,如何自动化地获取是一个有理论和现实意义的问题。尽管有许多质量评估的论述,定性的或是定量的,但其有效性尚待检验。无论是传统的关系数据库数据还是Web数据的质量评估,现有的方法从可行性和效率上都值得进一步研究和改进的[55]。(2)自动化地进行数据清洗是值得进一步探索的问题。自动化的数据分析和自动化的转换清洗是不可分割的,因此如何更好地利用统计、人工智能等技术实现自动化的问题发现和解决是目前大家关注的一个方向[15]。(3)数据质量问题产生的根本原因在于人们对现实世界的同一实体有多侧面的认识和多形式的表达,人们之所以能识别出数据所代表的含义,是因为数据在人们的知识范畴内是可识别的[13]。因此,如何从知识范畴的角度和语义的层面对数据质量问题进行认识是值得考虑的。

参考文献

1 Monge A, Elkan C. An efficient domain-independent algorithm

for detecting approximately duplicate database records [C]. In: Proceedings of the ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining, Tucson, AZ, 1997
2 Motro A, Rakov I. Estimating the quality of data in relational databases [C]. In: Proceedings of the 1996 Conference on Information Quality, Cambridge, Massachusetts, October 1996
3 Motro A, Anokhin P, Acar A C. Utility-based resolution of data inconsistencies [C]. IQIS 2004. 35～43
4 Parssian A, Sarkar S, Jacob V S. Assessing data quality for information products [C], 1999
5 Parssian A, Sarkar S, Jacob V S. Assessing information quality for the composite relational operation joins [C]. In: Proc. of Seventh International Conference on Information Quality, 2002
6 Kahn B K, Strong D M. Product and Service Performance Model for Information Quality: An Update. IQ 1998. 102～115
7 Barnett V , Lewis T. Outliers in statistical data. New York: John Wiley and Sons Inc ,1994
8 Liu B, Hsu W, Ma Y. Integrating classification and association rule mining [C]. In: Proc. of 4th International Conference on Knowledge Discovery and Data Mining (KDD98), ACM press, 1998. 80～86
9 Pluempitiwiriyawej C. A new hierarchical clustering model for speeding up the reconciliation of XML based, semistructured data in mediation systems [D]: [Doctoral Thesis]. 2001
10 Cappiello C, Francalanci C, Pernici B. Data quality assessment from user's pespective [C]. IQIS, 2004
11 Faloutsos C, Lin K-I. FastMap: A Fast Algorithm for Indexing Data-mining and Visualization of Traditional and Multimedia Datasets [C]. ACM SIGMOD 1995. 163～174
12 Pluempitiwiriyawej C, Hammer J. Element matching across data-oriented XML sources using a multi-strategy clustering model [J]. Data & Knowledge Engineering, 2004
13 Loshin D. Rule-based data quality. CIKM, 2002
14 Aebi D, Perrochon L. Towards improving data quality [C]. In: Proc. of the International Conference on Information Systems and Management of Data, 1993. 273～281
15 Lueebber D, Grimmer U. Systematic development of data mining based data quality tools [C]. In: 29th vldb, 2003
16 Strong D M, Lee Y W, Wang R Y. Data quality in context. communications of the ACM, 1997
17 Bertino E, Guerrini G, Mesiti M. A matching algorithm for measuring the structural similarity between an XML document and a DTD and its applications [J]. Information Systems, 2004
18 Rahm E, Do Hong Hai. Data cleaning: problems and current approaches [J]. IEEE Data Engineering Bulletin, 2000, 23(4): 3～13
19 Evoke software. Data profiling and mapping, the essential first step in data migration and integration projects. available at: http://www.evokesoftware.com/pdf/wtpprDPM.pdf, 2000
20 Caruso F, Cochinwala M, Ganapathy U, et al. Telcordia's database reconciliation and data quality analysis tool. In: the Proceedings of the 26th International Conference on Very Large Databases. Egypt, 2000
21 Batisa G E A P A, Monard M C. A study of k-nearest neighbour as an imputation method. 2003
22 Shahri H H, Barforush A A Z. Data mining for removing fuzzy duplicates using fuzzy inference [J], 2004 IEEE
23 Galhardas H, Florescu D, Shasha D. Declarative Data Cleaning: Language, Model, and Algorithms [C] . In: 27th vldb, 2001
24 韩京宇,徐立臻,董逸生. 一种大数据量的相似记录检测方法. 计算机研究与发展, 2005
25 Fellegi I P, Holt D. A systematic approach to automatic edit and imputation [J]. Americ Statistics of Association, 1976. 17～35
26 Raman I, Hellerstein J M. Potter's Wheel: An Interactive Data Cleaning System [C]. In: Proceedings of the 27th VLDB Conference, Roma, Italy, 2001
27 Bhattacharya I, Getoor L. Iterative Record Linkage for Cleaning and Integration [C]. In: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004. 11～18
28 Olson J. Data profiling: the key to success in integration projects. eAI ournal, 2002
29 Kubica J, Moore A. Probabilistic noise identification and data cleaning [C]. In: Proc. of the Third IEEE Conference on Data Mining, 2003
30 Grzymala-Busse J, Hu M. A comparison of several approaches to missing attribute values in data mining. In: Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing, 2000
31 Hipp J, Guntzer U, Grimmer U. Data quality mining: making a virtue of necessity [C]. In: Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara, 2001
32 Kukich K. Techniques for automatically correcting words in text [J]. ACM Computing Surveys 24, 1992. 377～439
33 Huang K-T, Lee Y W, Wang R Y. Quality information and knowledge management. New Jersey: Prentice Hall, 1998
34 Jin Liang, Li Chen, Mehrotra S. Efficient Record Linkage in Large Data Sets [C]. In: Eighth International Conference on Database Systems for Advanced Applications , March 2003. 137
35 Last M, Kandel A. Automated detection of outliers in real-world data [C]. In: Proc. of Second Conference on Intelligent Tech ,2001
36 Lee M L, Ling, et al. Cleansing data for mining and warehousing [J]. In: Springer, 1999. 751～760
37 Zhao Li, Yuan Sung Sam, Yang Qi Xiao, Sun Peng. Dynamic similarity for fields with null values [C]. In: Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, 2002. 161～169
38 Scannapieco M, Virgillito A, Marchetetti C, et al. The DaQuinCIS architecture: a platform for exchange and improving data quality in cooperative information systems [J]. Information systems, 2004, 29: 551～582
39 Weis M, Naumann F. Detecting Duplicate Objects in XML Documents [C]. In: IQIS, 2004. 10～19
40 Bilenko M, Mooney R J. Adaptive duplicate detection using learnable string similarity measures [C]. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Ddata Mining, 2003. 39～48
41 Hernandez M, Stolfo S. The merge/purge problem for large databases [J]. ACM SIGMOD Record, 1995. 127～138
42 Lee M, Ling T W, Low W L. Intelliclean: a knowledge-based intelligent data cleaner. In: Proceedings of ACM Knowledge Discovery and Data Mining, Aug 2000
43 Bouzeghoub M, Peralta V. A framework for analysis of data freshness [C]. In: Proc. of 2004 International Information Quality Conference on Information System, 2004
44 Lee Mong Li, Hsu Wynne, Kothari V. Cleaning the spurious links in data [J]. IEEE Intelligent Systems, 2004
45 Newcombe H B, Kennedy J M, Axford S J. et al. Automatic linkage of vital records [J]. Science, 1959
46 Benjielloun O, Sarma A D, Halevy A. ULDBs: Databases with uncertainty and lingeage. VLDB, 2006
47 Christan P. A probliilistic deduplication record linkage and geocoding systems. http://datamining.anu.edu.au/linkage.html
48 Bernstein P, Brodie M, Ceri S, et al. The Asilomar Report on Database Research [R]. SIGMOD Record ,1998, 27(4)
49 Missier P, Embury S M, Greenwood M, et al. Quality Views: Capturing and Exploiting the User Perspective on Data Quality, vldb, 2006
50 Andritsos P, Miller R J, Tsaparas P. Information-theoretic tools for mining database structure from large data sets [C]. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, 2004

(下转第 12 页)

[DB/OL]. http://www.ee.ucla.edu? ~paganini, 2001-09-20/2001-10-25

34 Kelly F P, Maulloo A , Tan D. Rate control for communication networks: Shadow prices, proportional fairness and stability [J]. J of Operations Research Society, 1998, 49 (3) : 237~252

35 Waldvogel M, Rinaldi R. Efficient topology-aware overlay network. ACM Communications Review, 2003,33(1):101~106

36 Hollot C, Misra V, Towsley D, et al. On designing improved controllers for AQM routers supporting TCP flows. In: Proceeding of INFOCOM 2001, Anchorage, Alaska,2001. 1726~1734

37 Christiansen M, Jeffay K, Ott D, et al. Tuning RED for Web traffic. In: Proceedings of ACM SIGCOMM 2000, Stockholm, Sweden ,2000. 139~150

38 Newman P. Backward Explicit Congest Notification for ATM Local Area Networks. In: Proc. of IEEE GLOBECOM '93, Houston, TX, 1993. 719~ 723

39 Rohrs C E, Berry R A. A Linear Control Approach to Explicit Rate Feedback in ATM Networks. In: Proc. of IEEE IN FOCOM '97, Kobe, Japan, 1997. 277~ 282

40 Zhao B Y, Huang L, Stribling J, et al. A resilient global-scale overlay for service deployment. IEEE Journal on Selected Areas in Communications, 2004,22(1):41~53

41 Habib I, Tarraf A , Saadawi T. A Neural Network Controller for Congest Control in ATM Multiplexers. Computer Networks and ISDN Systems, 1997, 29 (3) : 325~334

42 Hespanha J P, Bohacek S, Obraczka K, et al. Hybrid modeling of TCP congestion control [A]. Hybrid Systems: Computation and Control [C],Berlin: Springer Verlag, 2001. 291~304

43 Shenker S. Making greed work in networks: A game-theretic analysis of switch service disciplines [J]. IEEE/ACM Trans on Networking , 1995, 3 (6): 819~831

44 Kelly F P, Maulloo A , Tan D. Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of Operations Research Society, 1998,49(3):237~252

45 Zhang X Y, Liu J C, Li B,et al. CoolStreaming/DONet: A data-driven overlay network for live media streaming. In: Znati T. ed. Proc of the IEEE INFOCOM, Miami: IEEE Press, 2005. 2102~2111

46 Hollot C V, Misra V, Towsley D, et al. On designing improved controllers for AQM routers supporting TCP flows. In: Sengupta B. ed. Proceedings of the IEEE INFOCOM, Anchorage, Alaska: IEEE Communications Society, 2001. 1726~1734

(上接第 5 页)

51 Christen P,Churches T, Zhu J X. Probabilistic name and address cleaning and standardization. The Australian Data Mining Workshop, November 2002. available at:http://datamining.anu.edu.au/projects/linkage.html

52 邱越峰,田增平.一种高效的识别相似重复记录的方法[J].计算机学报,2001

53 Wang R Y. A product perspective on total data quality management [J]. Communica-tions of the ACM, 1998

54 Neal R. Probabilistic inference using Markov chain Monte Carlo methods. CRGTR -93-1. Department of Computer Science, University of Toronto,1993

55 Pon R K, Cardenas A F. Ddata quality inference. In: IQIS Workshop,2005

56 Rahm E,Do H H. Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 2000

57 Little R,Rubin D B. Statistical analysis with missing data. New York: John Wiley and Sons,1987

58 Wang R Y, Kon H B,Madnick S E. Data quality requirements analysis and modeling [C]. In:Proc. of Ninth ICDE, 1993

59 Ananthakrishna R,Chaudhuri S,Ganti V. Eliminating Fuzzy Duplicates in Data Warehouses [C]. In:Proceedings of VLDB,2002. 586~597

60 Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning [C]. In: The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2002. 269~278

61 Chaudhuri S,Ganjam K,Ganti V,et al. Robust and efficient fuzzy match for online data cleaning [C]. In: Proceedings of the 2003 ACM SIGMOD International Conference on on Management of Data,2003. 313~324

62 Chaudhuri S, Ganti V, Motwani R. Robust identification of fuzzy duplicates [C]. ICDE, 2005

63 Churches T, Christen P, Lim K,et al. Preparation of name and address data for record linkage using hidden Markov models [C]. 2002

64 de Waal T, Quere R. A fast and simple algorithm for automatic editing of mixed data [J]. Journal of Official Statistics, 2003,19 (4): 383~402

65 Dasu T, Johnson T. Exploratory data mining and data cleaning [M]. John wiley, 2003

66 Dasu T, Johnson T. Hunting of the snark: finding data glitches using data mining methods [R]. AT&T lab, 1999

67 Dasu T, Johnson T, Muthukrishnan S, et al. Mining database structure; or, how to build a data quality Browser [C]. ACM SIGMOD, 2002

68 Verykios V S, Moustakides G V. A Bayesian decision model for cost optimal record matching [J]. VLDB Journal ,2003, 12(1): 28~40

69 Winkler W E. Methods for evaluating and creating data quality [J]. Information System,2004,29: 531~550

70 Winkler W E. Set-covering and Editing discrete data [J]. In: Proc. of the Section on Survery Research Methods. American statistical association, 1997

71 Winkler W E, Draper L A. The SPEER edit system. Statistical data editing (volume 2); methods and techniques, united nations. 1997

72 Winkler W E,Petkunas T F. The DISCRETE edit system. Statistical data editing (volume 2); methods and techniques, United Nations. 1997

73 Fan W, Lu Hongjun, Madnick S E, et al. Discovering and reconciling value conflicts for numerical data integration. Information Systems,2001,(26) :635~656b

74 Low Wai Lup, Lee Mong Li, Ling Tok Wang. A knowledge-based approach for duplicate elimination in data cleaning [J]. Information Systems,2001,26(8):585~606

75 Wu Xintao, Barbara D. Learning missing values from summary constraints [R]. ACM SIGKDD Explorations Newsletter, 2002

76 Wu Xintao, Barbara D. Modeling and Imputation of Large Incomplete Multidimensional Datasets. In: Proceedings of the International Conference on Data Warehousing and Knowledge Discovery. Aix-en-Provence, France, Septemeber 2002

77 Zhu Xingquan, Wu Xindong,Chen Qijun. Eliminating class noise in large data sets [C]. In:Proc. of 20th International Conference on Machine Learning ,2003

78 Yair Wand, Richard Y W. Anchoring data quality dimensions in ontological foundations [C]. Communication of ACM, 1996