中国科技资源导刊 ISSN 1674－1544
2008年11月 第40卷第6期 9－20
CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
ISSN1674-1544 Vol.40 No.6 9-20,Nov. 2008

对大规模数据集的公正与效用导向的质量评估的理解

[美国] G. Shankaranarayanan Adir Even
(波士顿大学管理学院,信息系统系,波士顿 MA 02215)

摘 要:在复杂的数据环境中建立并维持非常高的数据质量是很昂贵并且往往不可能实现的。对质量的定量评估能够为制定改进措施的优先顺序提供重要的帮助。本文探索了一套方法,能够对数据质量的公正评估与效用导向评估两种方法进行评价。公正评估可以评价和衡量数据缺陷的程度,效用导向的评估能够衡量在一个特定的使用环境中,质量缺陷的存在对降低数据效用的程度。本文介绍的质量评估方法通过一个实际的校友数据库得到了实际验证。这个数据库是一个巨大的数据资源,它对校友关系和发起的承捐活动进行管理。质量评估的结果能够为这个数据库执行和管理改进数据质量的策略提供帮助。

关键词:数据质量;TDQM(综合数据质量管理);信息产品;数据管理;CRM(客户关系管理)
中图分类号: G203 **文献标识码**: A DOI: 10.3772/j.issn.1674－1544.2008.06.002

Understanding Impartial Versus Utility－Driven Quality Assessment in Large Datasets

G. Shankaranarayanan, Adir Even
(Information Systems Department, Boston University School of Management, Boston MA 02215)

Abstract: Establishing and sustaining very high data quality in complex data environments is expensive and often practically impossible. Quantitative assessments of quality can provide important inputs for prioritizing improvement efforts. This study explores a methodology that evaluates both impartial and utility－driven assessments of data quality. Impartial assessments evaluate and measure the extent to which data is defective. Utility－driven assessments measure the extent to which the presence of quality defects degrades utility of that data, within a specific context of usage. The quality assessment methodology is empirically assessed using real－life alumni data－a large data resource that supports managing alumni relations and initiating pledge campaigns. The results provide important inputs that can direct the implementation and management of quality improvement policies in this data repository.

Keywords: data quality, TDQM, information products, data management, CRM

第一作者简介:G. Shankaranarayanan(1963－),男,副教授,美国波士顿大学商学院信息系统系移动客户实验室主任,研究方向是数据质量、元数据质量、数据质量管理。

收稿日期:2008年9月1日。

1 引 言

从数据使用者的角度看，建立并维持非常高的数据质量是最理想的，但是，数据量的快速增加和高昂的维护费用，使得这样的目标几乎不可能实现。从经济学的角度看，以追求完美的数据质量为目标并不是最好的选择，因为提高数据质量的费用抵消了提高数据质量可获得的好处。由于难以实现完美的数据质量以及权衡考虑维护运行需要的巨大经济投入，显然有必要制定改进措施的优先顺序，并优先关注特定的数据元素或数据子集。定量评估能够为数据质量管理以及指导改进相关措施和策略提供帮助。现在，这样的评估通常是公正的，它可以衡量质量缺陷存在的程度，却忽略了使用环境。在本文的研究中，我们认为，可以通过增加数据的效用以及在特定使用环境中数据的价值贡献提升质量评估的效果，我们称之为效用导向的数据质量评估。我们开发了一套方法用于沿着不同的质量维度（在这里用完整性和时效性表示）对公正评估和效用驱动质量评估进行测度。这个测度的结果有助于对质量特性的深入了解，并指导质量改善政策的完善。我们在客户关系管理（CRM）环境中对这种方法进行了验证，数据样本来自于大学用于管理校友关系、开展募捐和承捐活动等事宜的数据库。信息资源的效用目标来源于实践[1]，并且随着使用环境而定（例如一个决策任务）。相同的数据资源在不同的环境中可以发挥不同的效用，相应地，数据缺陷的存在会对效用的降低有不同程度的影响。因此，我们建议通过衡量效用高低的程度来测度质量，进而评估不同环境中数据质量缺陷的影响。尽管在数据质量的研究中非常强调关于环境评估的重要性（例如文献[2][3][4]），但是并没有削弱公正质量评估的作用。我们现在要做的是在一个现实的数据环境中，演示说明效用导向数据质量评估方法的应用过程，并揭示它对数据质量管理的意义，进而说明这种方法在数据质量管理中的重要性。这篇论文在以下几个方面有重要的贡献。

第一，通过一个现实数据环境中的演示，说明效用导向数据质量评估方法的应用过程验证了参考文献[5]中提出的效用导向数据质量评估的正确性。第二，本文提供了一个公正质量评估和效用驱动质量评估的对比分析，并通过它们之间的相互关系揭示了一些重要的观点。第三，数据的效用分析显示，单个的数据记录之间在对效用的贡献方面存在很大的不同，也就是说，不同类型的数据缺陷可以不同程度地影响数据对效用的贡献，并且，这种不同可以在效用导向评估中得到体现。本研究阐明了在大型数据集中这些不同对管理数据质量的影响，例如在制订质量改进措施的优先顺序方面。需要注意的是，对效用的不确定性及其在质量上反映的测度是基于具体环境的。概括这些结果并应用于其他的数据集（即使是属于同一领域的）需要进行重新评价。第四，本文描绘了本项评估技术如何用于理解效用和其他相关的事物，并且还说明了本文的一些观点如何指导数据质量改善方法和策略的落实。在本文的其他部分，我们首先探讨了在管理大型数据资源质量方面面临的挑战，简短回顾了对我们的研究有影响的一些质量评价和改善方法。然后提出了一套效用导向的用于质量评价的方法，我们将这种方法应用于校友数据库，并且用获得的结果规范、改善质量。最后，我们明确了本研究的贡献，讨论了一些管理问题，并提出了深入研究的建议。

2 研究背景

数据的高质量对于机构信息系统的整合非常重要，数据会存在各种各样的质量缺陷，例如数据缺失、数据损坏、数据不准确、数据无效和数据过时[5]。数据质量的缺陷降低了数据质量，危害数据的可用性，并且损害机构的收入和信誉[6]。最近的发展趋势（例如数据仓储、企业资源规划、RFID技术、供应链和Clickstream技术等）形成了对复杂数据分析的巨大需求，所以，机构需要管理好大型的复杂数据资源。在复杂的数据管理环境中，如果以建立无缺损的数据集为目标是非常昂贵并且几乎是不可能实现的。因此，如果定位于根据多个维度（例如精确性与及时性、完整性与

一致性)确定数据质量，就能换来内部均衡[7,8]的实现。高效率的质量管理需要评估这些均衡，优化(不是最大化)质量水平，同时允许一些瑕疵存在[9]，并相应地制订改进措施的优先顺序。用于数据质量改善的方法有三大类[6]：

(1) 错误的检测与修正。错误可以通过与正确的基线相比较而检测出来(例如实际存在的物体、预先设定的规则/计算结果、数值范围或者一个经过验证的数据集)。自动检测和修正的运算规则已经被提出来(例如[10][11])，有几个商业化软包也支持自动错误检测和数据清理[12]。当自动修正功能无法得到满意结果时，公司可以考虑手工修正，或者雇用外部的专业数据清理机构。虽然错误检测/修正有助于提高质量水平，但是它并不能改变产生错误的根本原因并阻止错误再次发生。

(2)过程控制和改进。整体数据质量管理理论(tdqm)[13]建议将质量需求定义成一个连续的过程，并沿着这个过程进行测度，分析结果并相应地改进数据处理过程。与错误检测和修正方法不同，这套方法能够帮助检测并改正产生错误的根本原因，并且有成功的例子显示其也能帮助避免错误的再次发生。不同的方法都支持整体数据管理周期，例如文件信息处理路线图(IPMAP)[3]，质量权衡最优化[14]和一些质量趋势可视化工具等。

(3) 过程设计。数据过程最好从起点开始创建(或者已有构成的重新设计)，这样可以使得质量更加容易管理，并且错误出现的可能性会降低。过程设计的原理已经有大量的研究进行了探讨(例如文献[6][13][14])，例如管理与参与、嵌入式控制、数据建模、处理流程、运作效率等。

我们提出的方法对以上的分类都进行了研究，本研究更加适合于满足制定改进措施的优先顺序的需要，尤其是针对大型表格型数据集，它们的特点是多条记录具有同样的属性结构。尽管提出的方法可以被用于任何表格型数据库，但是我们关注的是数据仓库中的大型数据表。数据仓库由 Fact 表和 Dimension 表组成(图 1)。Fact 表存储用户实际感兴趣的事件数据，根据数据库的设计，一条记录可以代表一个单独的事件或者是一个事件集合。一个 Fact 表的记录包含量测结果(例如数量和总量)数值或者其他的可计算的数据及其描述(例如时间标记、付款/装载指南)。Fact 表还包含 Dimension 的标识符，用于连接事件和与之相关的企业单位。Dimension 表存储 Dimension 实例的列表及相关的描述信息(例如时间标记、顾客姓名、人口特征、地理位置、产品和分类等)。

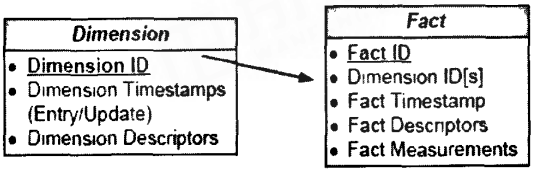

图 1　维度与事务表

本研究涉及 Dimensional 数据的质量改善，而 Fact 数据主要是用于评估 Dimensional 数据的质量及开发相应的改善策略，在本研究中并不涉及。在决策支持环境中，Dimensional 数据的质量至关重要。例如，数据库销售专家使用销售数据分析消费行为并且在特定的地区面向特定的消费者推出新的促销手段和产品[15]，因此，维护相关的 Dimensional 数据(例如消费者、产品和地区)的高质量就变得非常重要，否则，促销手段就有可能达不到目的。在数据仓库中有一个普遍的问题是“慢慢改变的 Dimensions”[16]：Dimension 的特点(例如收入、婚姻状态、职业)会随着时间而改变，并且没有固定的规律，事件数据及其相关联的 Dimensional 数据也会变得不同步，因此会影响决策的准确性。

提高数据集的质量必须考虑以下因素。①目标：确定质量水平时，可以将质量水平看成是一个连续体，它的一端是最高的质量水平，即数据没有任何缺陷，它的另一端可以不管，不用采取任何措施提高它。在这两端之间，我们可以考虑制定相关的政策等来提高质量水平满足我们的需要，但是要允许缺陷的存在。②范围：我们可以考虑对所有的记录和属性用相同的方式对待，也可以区别对待，即可以对特定的记录赋予高优先级以提高它的质量，而不用采取很多措施提高其他记录的质量。这样就可以对不同类型的政策进行评估。

(1)预防:可以采取某些措施用于防止、减少质量缺陷,或者减少在数据获取及处理过程中出现缺陷的几率。例如,改善数据获取界面,驳回遗漏值,执行完整性约束或者使用另外的(但可能会比较贵的)干净的数据。

(2)审核:质量缺陷不仅会出现在数据获取阶段,也会出现在数据处理阶段(例如由于对字段的错误计算或者由于错误整合了多个数据源的数据而导致与编码不匹配),或者甚至出现在数据已经被存储以后(例如由于Dimension数据描述的现实世界的事物发生了改变)。因此,需要审核数据,进行过程监测,检测缺陷的存在。

(3)修正:有时候甚至是当数据缺陷被发现以后,如何修正它们也是个问题。在某些情况下,修正缺陷很费时间,而且所需代价很高(例如联系客户,或者重新购买遗失的数据)。当价格不合理时,可以不进行修正。

(4)使用:在某些情况下,可能会建议使用者不要使用某些记录或者属性的子集,或者全部阻止使用。例如,当数据质量过低并且不能得到很大提高的时候,或者当某些子集被发现不适合在某些环境下使用的时候。当设定某些策略的目标和范围时,必须要考虑能够真正获得的改善效果,它对数据可用性的影响,还有效用/成本的权衡。对预期的效用/成本权衡和全面的经济方面的影响进行定量评估,有助于评价不同的政策并对其进行选择[9]。本研究用到的测度方法能够为以上的评估提供参数输入。

3 公正与效用导向的评估

数据质量可以沿多个维度进行测度(例如精确性、完整性和时效性)。质量一般用0-1之间的数字表示,0表示差,1表示完美[2,6]。一些评价方法是基于物理特征参数(例如项目数量、时间标签和失误率)的,预设一个绝对的客观的质量标准,但是不考虑数据应用的环境。一个可以替代的方法是从数据内容推导出测量指标,并且在特定的使用环境中对这些指标进行评定。前者称为基于结构的方法,后者称为基于内容的方法[8]。所谓的公正质量评估,是一种基于数据本身评估的思想,而不用考虑数据的用途和使用环境。在某些情况下,相同的维可以使用公正的或者基于环境的方法进行评估,这取决于评估的目的[2]。公正和基于环境的这两种评估方法都可以为反映情况的全面性做出贡献,因此同时满足它们的要求。本研究探索的这套方法是评估质量缺陷的存在(这是公正的观点)及其对效用损失的影响(这是基于环境的观点)。对公正的观点和基于环境的观点进行研究可以为我们提高质量及制定相关策略提供帮助。

本研究采用参考文献[5]中建议的评估框架。简要地说,这个框架允许沿着不同维进行基于环境的评估,也同样允许进行公正地评估。这个框架中的质量评估是基于数据集效用驱动的。这个框架用比率表示质量,它与效用评估的单位没有不同。在这个研究中,我们只考虑效用这一个用途,然而文献[5]介绍的框架中也可以代表多个用途。

评估的数据集有N条记录(用[n]表示)和M种属性(用[m]表示),记录[n]的属性[m]的数据内容表示为 $f_{n,m}$。对 $f_{n,m}$ 的质量评估反映了记录[n]的属性[m]受质量缺陷影响的程度(用0-1之间的数字表示,0表示有一些缺陷,1表示没有缺陷)。总体效用 U^D 来源于效用集合{U^R n},是基于相对重要性的,因此 $U^D = \sum_{n=1\cdots n} U_n^R$。本框架中使用了效用规划函数 U,它把记录内容和效用质量连接起来。

$$U_n^R = u(\{f_{n,m}\}_{m=1\cdots M}, \{q_{n,m}\}_{m=1\cdots M}) \quad (1)$$

对于一个给定的属性内容集合 $\{f_{n,m}\}$,当所有属性都有完美的质量时,该记录的效用达到上限 U_n^{RMAX},但是如果某些属性有缺陷时,效用也会在一定程度上被减小。记录质量 Q_n^R 被定义为0-1之间的比率[0,1],它表示的是在实际效用 Q_n^R 和上限效用之间的一个值:

$$\begin{aligned} Q^R{}_n &= U^R{}_n / U_n^{RAMX} \\ &= (u(\{f_{n,m}\}_{m=1\cdots M}, \{q_{n,m}\}_{m=1\cdots M})) \\ &\quad / (u(\{f_{n,m}\}_{m=1\cdots M}, \{q_{n,m}=1\}_{m=1\cdots M})) \end{aligned} \quad (2)$$

相似地,数据集质量 Q^D 也是在实际和最大可能效用之间的一个比率数值:

$$Q^D = \left(\sum_{n=1\cdots N} U_n^R\right) / \left(\sum_{n=1\cdots N} U_n^{RMAX}\right) = \left(\sum_{n=1\cdots N} U_n^{RMAX} Q_n^R\right) / \left(\sum_{n=1\cdots N} U_n^{RMAX}\right) \quad (3)$$

当效用独立于属性内容（也就是常数 $U_n^{RMAX} = U^D/N$）时，结果就是一个公正评估的结果，它表示的是完美记录的数目与总记录数的比率，这与一般的结构式的定义是一样的[2,6]：

$$Q_n^R = (1/M) \sum_{m=1\cdots M} q_{n,m}$$
$$Q^D = (1/MN) \sum_{n=1\cdots N} \sum_{m=1\cdots M} q_{n,m} \quad (4)$$

这个定义允许沿不同维进行评估，每一个评估可以反映特定的质量缺陷。例如，完整性反映遗漏值，合法性反映不属于一个 value - domain，精确性反映不正确的内容，时效性反映数据不是最新的。

在某些数据集中，效用不均衡的程度比其他数据集相对较大。尽管一条记录中质量出现缺陷的可能性与它的效用无关，但如果承认数据记录有更高的效用价值会有利于鼓励人们为了减少它的数据缺陷而采取改进措施，那么效用导向的评估可以反映出缺陷的存在对效用的减小程度。将公正质量评估和效用导向质量评估的结果进行对比，对于数据集的质量管理非常重要。在一个高水平上，我们可以就 3 种可能的情况进行对比讨论：①如果效用驱动评估的得分比公正评估高很多，这说明具有高效用的数据缺陷很少。此外，还可以有两个补充解释：一是有缺陷数据的用途本来就很小。二是一些错误改进策略已经被采用，也就是说，已经采取了一些措施，例如删除数据中的缺陷，使得一些具有高效用的数据具有高质量。②如果效用驱动评估的得分与公正评估的得分差不多，这说明了无关联性——有质量缺陷的数据的比例并不取决于某些记录的效用，这个比例可能高也可能低。这也可能说明该数据集的质量普遍是高的，所有记录的效用分布很平均。③如果效用驱动评估的得分比公正评估低很多，这意味着高效用的数据含有很高比例的缺陷。造成这种不正常现象的一个可能原因是有缺陷的数据反而具有高效用，也可能因为数据集中存在很高的不均衡性（例如，很少一部分的记录却承担了很大比例的效用），还有可能是高效用数据发生了大量的损坏。就像我们在前面利用校友实际数据进行评估演示的那样，对公正质量评估和效用导向质量评估进行分析能够帮助我们开发 DQM 策略。

4 评估校友录数据

本部分我们将演示效用驱动质量评估及其对制定改进措施优先顺序的意义，我们将利用校友录中大量的数据进行评估，这个大学大部分的收入来自于这个数据集记录的捐赠。不同的院系利用校友录的数据与捐赠者进行联系，跟踪他们的捐赠历史并管理各种承捐活动。这些数据及其管理系统都可以被看成是一个客户关系管理（CRM）表。在通常情况下，客户关系管理表用来管理客户关系，跟踪他们过去的贡献，分析促销模式并进行分类，以利于将来更好地进行促销活动。

下面讨论数据收集和评估方法。本研究使用的数据来自于两个数据集，即 Profiles 和 Gifts（表 1）。

Profiles（Dimensional 数据）数据集有 358372 条记录，包括同学录中的记录和其他潜在捐赠者的联系信息和人口数据。Profiles 数据集的源数据库记录有 100 多个属性，许多属性用于管理、索引和审计，与大多数数据使用者关系不大。在本研究中，我们主要关注在决策过程中应用很广的 6 个属性：毕业学校、性别、婚姻状况、收入、种族和宗教信仰。它们都是分类属性，也就是说，它们每一个都和一个值域相关联，而这个值域含有有限的数值可供选择，并且被存放在一个相关联的查找表中。另外，我们还要关注两个时间标签：毕业日期（Graduation Year）和更新日期（Update Year）。更新日期（Update Year）是数据最后更新的时间。

Gifts 数据集（Fact 数据）有 1415432 条记录，主要是关于捐赠历史的记录，一些记录是关于已经捐赠的礼物，还有一些是关于将要捐赠的礼物（通过 Record Type 属性加以区分）。每一条记录

表1 同学录数据

Dataset	Records	Growth	Attributes	Description
Profiles – data on alumni, parents, and friends One record per name listed	358,372	Annual average: 7.044 STD 475	Profile ID	A unique identifier of the profile record
			Graduation Year	The year in which a profile record was added
			Update Year	The year in which a profile record was last updated
			School	The school from which the person graduated (28 categories)
			Gender	Male Female
			Marital	Marital status (7 categories)
			Income	Income category (3 categories)
			Ethnicity	Ethnic group (7 categories)
			Religion	Religion (31 categories)
			Other Attributes	Contact information (e.g., address, phone), demographics, administrative fields
Gifts – detailed historical	1,415,432	Annual average: 45.884	Gift ID	A unique identifier of the gift record
			Record Type	Some records represent pledges that have been paid later, or multiple payments on behalf of a gift
archive of gift transactions		STD 6147	Profile ID	A foreign key to the *Profiles* dataset Each record is associated with one profile, but some profiles are not associated with any gifts.
			Gift Amount	The gift value (in USD)
			Gift Year	The year in which the gift record was added to the dataset
			Other Attributes	Additional details – e.g. pledge efforts, gift allocation, payment methods

也包括礼物数量和捐赠时间。

出于保密的原因，我们使用的数据集只有实际数据量的40%左右，某些属性被用编码掩盖起来，礼物数量是其与一个常量因子相乘的结果。源数据是在1983－2006年收集的。1983－1984年，一个很大的关于以前活动的数据量被增加进来（203359条Profile数据和504969条Gift数据），从1985年开始，两个数据集都是被定期更新，数据量也在稳定地增加。

我们的评估按照以下步骤实施：

（1）初步评估：我们为所有质量评估需要的参数收集了一些摘要统计数据，并且研究可能存在的相互关系和依存关系。

（2）公正质量评估：我们使用了比率测度方法[它基于项目数量(方程4)]评估公正质量。对于我们评估的Profile属性，我们考虑4种类型的质量缺陷。

①遗漏数值：当记录一条新的Profile(或者更新已经存在的)，如果系统允许该字段记录为空，则根据初步评估显示，该类记录有很大部分可能会是缺失的。

②无效数据：我们最初的评估显示没有无效数据，所有不允许为空的属性值都与其所属的值域相一致。

③是否时新：有很多的Profile数据已经很长时间没有更新，有的数据甚至自从创建后就再也没更新过，这是这个数据集存在的一个严重的问题。我们使用一个二进制的变量用于表示数据是否为最新的，1表示数据最近被更新过，0表示最近没有更新。我们用这个变量对1年期(2006年)和5年期(2002－2006年)的数据进行了评估。此外，还有更好的方法就是指数变换[5]，它将日期转换为[0,1]之间的数字表示。

$$t = \exp\{-\alpha(Y^C - Y^U)\} \qquad (5)$$

Y^C，Y^U分别表示当前的年代和上次更新的年代。

α表示敏感度因子，反映Profiles中过期数据的比例。这里$\alpha = 0.25$，表示每年有大约20%～25%的Profiles数据过期。

t表示时新性的级别。0表示数据很长时间没有被更新过（也就是$Y^C \geq Y^U$），1表示数据很新(也就是$Y^C = Y^U$)。

④不准确：相当部分的Profile数据记录不够

表 2　校友录 profile 演示样本

ID	Gender	Marital Status	Income Level	Record Complete (Absolute)	Record Complete (Grade)	Last Update	Recent Updated (1Y)	Up-to-date Rank	Inclination	Amount
A	Male	Married	Medium	1	1	2006	1	1	1	200
B	Female	Married	*NULL*	0	0.667	2003	0	0.47	1	800
C	*NULL*	Single	*NULL*	0	0.333	2005	0	0.78	0	0
D	*NULL*	*NULL*	*NULL*	0	0	1996	0	0.08	0	0
									2	*1000*

准确，这主要是因为多年来缺乏跟踪，使得有些人口数据没有及时更新而造成的，也有很少是因为数据输入错误造成的。然而，由于缺乏适当的基准，在这个研究中我们没法评估不准确性的影响。

初步评估以后，我们重点研究两种数据缺陷类型：一种是遗漏数值和数据过期及其相关的质量测度指标：完整性和时效性。完整性在数据条目水平和记录水平进行评估。在数据条目水平，完整性是完整条目数量和总数据数量的比率。为了能在记录水平评估完整性，我们考虑两种方法：(1)绝对排名方法：如果一条记录有至少一个属性值丢失，则将整条记录标记为有缺陷(0 表示有缺陷，1 表示无缺陷)。(2)分级排名方法：用没有缺陷的属性数目与属性总数目相除得出的数字表示 (0 表示所有的属性值丢失，0.5 表示一般的属性值丢失，1 表示没有属性值丢失)。在后面的计算实例中，可以看到，当所有属性合并计算时，在记录水平和条目水平计算完整性是等同的。因此，我们计算时效性时只需要记录水平进行就可以。

(3) 效用导向质量评估：我们使用效用测度作为比例因子(方程 1－3)，使用 Gifts 数据集，重复进行质量评估，对每个 Profile 我们评估两个效用测度。

①趋势 (Inclination)：是一个二进制变量，它反映一个人捐献的趋势。这个效用测度在两个时间段被测量：最后一年 (2006 年) 和以前 4 年 (2002－2005 年)，21485 个 Profiles(约等于总量的 6%) 与 2006 年的捐赠有关系，43157 个 Profiles 发生在 2003－2005 年。

②总数量 (Amount)：捐赠的物品总数量。评估最后一年和前 4 年的。

这两个效用测度会有不同的潜在应用。例如，趋势 (Inclination) 一般是用于承捐活动，是建立在大的捐赠者的基础上。总数量(Amount)则针对大量特定的有潜力的捐赠者。

(4) 分析：对比公正评估和效用导向评估的结果能够提供很有用的知识，并且能够为开发 DQM 策略做出贡献。

为了演示这个计算方法，我们使用表 2 中的校友录 Profile 数据作为演示样本，在这个数据库中，一些属性丢失了 (高亮显示)，一些记录最近没有更新。

我们可以看到，性别 gender 字段 4 个记录中有 2 个丢失了值，因此，这个性别字段的公正完整性是 0.5，相应地，婚姻状况的公正完整性是 0.75(4 个中有 1 个丢失)，收入水平的公正完整性是 0.25，所有属性合并在一起，公正完整性是 0.5(12 个中有 6 个丢失)。对于记录水平的完整性，只需计算绝对排名。4 个中有 3 个缺少记录，因此完整性是 0.25。如果使用分级排名方法，记录水平的完整性是 0.5，对于效用导向的完整性测度，我们看到 4 个里面有 2 个 Profile 记录是与效用有关系，并且我们使用 Inclination 和 Amount 作为比例因数。对于性别和婚姻状况，没有与效用有关的属性值丢失，因此，效用导向的完整性是 1。对于收入水平，1 个与效用有关系的属性值丢失，与 Inclination 因子相运算，完整性是 (1＊1＋1＊0)／2＝0.5，与 Amount 相运算，完整性是 (1＊200＋0＊800)／1000＝0.2。在记录水平，用 Inclination 因子与绝对排名相运算，又产生一个完整性水平 (1＊1＋0＊1)／2＝0.5，与 Amount 运算则产生 (1＊200＋0＊800)／1000＝0.2。用 In-

clination 分级排名方法的结果运算则产生 (1 * 1 + 0 * 0.667) / 1.667 = 0.6，用 Amount 运算则是 (1 * 200 + 0.667 * 800) / 1000 = 0.733。公正的时效性是 0.25，使用 up - to - date 方法是 0.58。对于效用导向时效性计算，用 Inclination 计算分别是 (1 * 1 + 1 * 0) / 2 = 0.5 和 (1 * 1 + 1 * 0.47) / 2 = 0.74，Amount 参与运算则分别是 (1 * 200 + 0 * 800) / 1000 = 0.2 和 (1 * 200 + 0.47 * 800) / 1000 = 0.58。

5 结 果

首先，我们已经为每一个 Profile 记录计算了以下变量。

(1) 遗漏数据指标：对于每一个属性（总共 6 个），相应的变量反映出其数值是否消失（0 表示消失，1 表示没有消失）。我们还为每一个记录计算了绝对排名（0 表示至少 1 个属性值遗漏，1 则表示没有遗漏），还计算了分级排名。

(2) 是否时新 (Up - to - date)：我们计算了一个二进制指标反映一条记录是否已经在上一年更新或者一个 5 年时段。我们也基于方程 (5) 使用了时新性 (Up - to - date) 排名。

(3) 效用测度：我们计算了 Inclination 和 Amount 效用测度。

以上变量和统计数字及其相互关系参见表 3。

从上面的结果得到见解如下：

(1) 遗漏数据的比例相对较高，达到 64% 的记录至少有一个属性值遗漏（绝对排名），平均 22% 的属性值遗漏 (分级排名)。

(2) 很大比例的 Profile 记录更新不及时，只有 17% 的 Profiles 在上一年被更新或增加，49% 在最近 5 年没有更新。

(3) 不同属性遗漏数据的数量不同。学校和性别几乎没有遗漏，然而收入、种族和地区遗漏的比例很高。

(4) 遗漏数据指标之间的关系意义很大。这意味着都是一个记录的某个属性值遗漏，则它有可能也会遗漏其他属性值。

(5) 大多数情况下，遗漏数据指标与时新性 (up - to - date) 关系比较小，但是却意义很大。这意味着老的数据更有可能遗漏数据，但不是大规模的。

(6) 近年来，在捐赠趋势 (Inclination) 和捐赠数量 (Amounts) 两个测度之间有很密切的关系。捐赠数量比较多的记录一般时新性很高，但数据完整性比较低。

(7) 高的捐赠趋势 Inclination 与大多数质量指标有密切的关系，例如，高的捐赠数量 Amount 与所有的 up - to - date 和一些遗漏数据指标关系密切。

(8) 高的公正质量 (很小的缺陷，更多的近期更新) 与高的效用相关联。我们下一步测度这个关联的程度。对于二进制指标 (遗漏数据，近期更新) 我们使用一个 2 路 ANOVA 测量缺陷和非缺陷数据的效用的差异相关性。对于可用一个系列数值表示的参数，我们使用线性回归的方法。P 值见表 4。

结果显示，捐赠的趋势 (对于两个时段) 与几乎所有公正指标都有很强的关联度。另一方面，数量 (Amount) 与时新性 (up - to - date) 有很强的关联性，但是与完整性只有一部分关系。很重要的是，调整后的 R - SQR 结果很低 (低于 0.1)，这意味着效用的可变性与高的公正数据质量相关联。下一步我们测量了公正和效用导向的完整性和时效性。结果见表 5。

大多数效用导向质量评估得分要比相应的公正评估得分高。这是因为对大多数指标来讲，强的效用与高的公正质量有很强的关联性：

在属性水平和记录水平上，用 4 个效用指标进行效用驱动的完整性评估基本差不多。这就意味着，在计算效用驱动完整性时，用 4 个指标评估并不比用一个指标有多大的好处。

(1) 对于天生具有高的公正完整性的属性 (如学校和性别)，效用导向评估与公正评估没有很大的区别。婚姻状态中会存在一些差别，但是由于公正完整性相对较高，这些差别就显得很小。

(2) 对于天生具有低的公正完整性的属性，我们看到在效用驱动评估和公正评估得分上具有实质性的区别。对于种族属性，差别相对较小，它稍稍比宗教高一点，但是比收入高很多。这意

表 3　Profile 变量一统计结果及其关系

				Correlation*														
		Avg.	STD.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Attribute	1. School	0 99	0 01	-	L*	L*	L*	L*	L*	L*	L*	L	L	L	L	L	L	L
	2. Gender	0.99	0 10	L*	-	M*	M*	M*	L*	L*	M*	L*	L*	L*	L*	L*	L	L
	3. Marital	0.89	0.30	L*	M*	-	M*	M*	M*	M*	H*	L*	L*	L*	L*	L*	L	L*
	4. Income	0.63	0.48	L*	M*	M*	-	M*	M*	H*	H*	L*	L*	L*	M*	M*	L*	L*
	5. Ethnicity	0.59	0.49	L*	M*	M*	M*	-	M*	H*	H*	M*	M*	M*	L*	L*	L	L
	6. Religion	0.60	0.49	L*	L*	M*	M*	M*	-	H*	H*	L*	L*	L	L*	L*	L	L*
Record	7. Absolute	0 36	0 48	L*	L*	M*	H*	H*	H*	-	H*	L*	L*	L*	L*	M*	L	L
	8. Grade	0.78	0 20	L*	M*	H*	H*	H*	H*	H*	-	L*	L*	L*	L*	M*	L*	L*
	9. Recent-1	0.1*	0.37	L	L*	L*	L*	M*	L*	L*	L*	-	M*	H*	L*	L*	L*	L*
	10 Recent-5	0.51	0.50	L	L*	L*	L*	M*	L*	L*	L*	M*	-	H*	M*	L*	L*	L*
	11 Up-to-date	0.42	0 35	L	L*	L*	L*	M*	L*	L*	L*	H*	H*	-	L*	L*	L*	L*
Utility	12 Inclin 1	0.06	0.24	L	L*	L*	M*	L*	L*	L*	L*	L*	M*	L*	-	H*	L*	L*
	13 Inclin 2/5	0.12	0.33	L	L*	L*	M*	L*	L*	M*	M*	L*	L*	L*	H*	-	L*	L*
	14 Amt -1	50	7 1K	L	L	L	L*	L	L	L	L*	L*	L*	L*	L*	L*	L*	-	L*
	15 Amt.-2/5	190	11 7 K	L	L	L*	L*	L	L*	L	L*	L*	L*	L*	L*	L*	L*	L*	-

H. 0.5, M. 0 1 to 0.5, L 0 1, (*) Significant (P-value 0.02)

表 4　效用变量的差异(P 值)

	Variable	Inclination (1 Year)	Inclination (2-5 Years)	Amount (1 Year)	Amount (2-5 Years)
Attribute	School	0 620	0.945	0 972	0.939
	Gender	~0**	~0**	0 759	0.393
	Marital	~0**	~0**	0 207	0.008**
	Income	~0**	~0**	0 021*	~0**
	Ethnicity	~0**	~0**	0 598	0 048*
	Religion	~0**	~0**	0 060*	0.004**
Record	Absolute	~0**	~0**	0 067*	0.486
	Grade	~0**	~0**	0 029*	0.007**
	Recent-1	~0**	~0**	~0**	~0**
	Recent-5	~0**	~0**	0 001**	~0**
	Up-to-date	~0**	~0**	~0**	~0**

(**) Highly Significant (P-value 0.01), (*) Marginally Significant (P-Value 0.1)

表 5　质量评估

		Impartial Completeness	Utility-Driven Completeness			
			Inclination (1Y)	Inclination (2-5 Y)	Amount (1 Y)	Amount (2-5 Y)
Attributes	School	0.999	0.999	0.999	0 999	0 999
Completeness	Gender	0.990	0 997	0 998	0 997	0 999
	Marital	0.894	0.950	0.958	0 984	0 977
	Income	0.631	0.872	0.896	0 891	0 836
	Ethnicity	0 596	0.646	0 654	0 656	0 496
	Religion	0.605	0.717	0.715	0 819	0 751
	All	0.786	0.863	0.870	0 891	0 843
Record Completeness	Absolute	0.356	0.497	0.511	0 561	0 608
	Grade	0.786	0.863	0.870	0 891	0 843
Record Currency	Recent-1	0.171	0.282	0.219	0 635	0 552
	Recent-5	0.510	0.635	0.635	0 899	0 860
	Up-to-date	0.425	0.540	0.518	0 807	0 775

味着这些属性与其效用有很大的区别。收入的完整性数据与高效用和低效用 Profile 记录有很大区别(对于 Inclination 和 Amount 都是这样)。宗教属性的完整性数据也是这样，与它们有很多差异。而种族属性的完整性与 Profile 记录区别不大。

(3) 测度所有属性和在一起的完整性，或者用记录水平测度它，会有一个平均的效果。在公正评估和效用导向评估之间会存在一些差异，但是并没有相对于其一种属性的差异大。

(4)不像完整性。关于时效性,对于所有的指标，Amount - driven 的得分比 Inclination - driven 的得分高很多。这意味着一条记录及时更新的程度与捐赠的数量有很密切的关系。显而易见，与平均值相比，捐赠者之间的捐赠数量差别极大(表 3)，这可能会在捐赠数量和 Profile 记录之间引起很大的不平衡，即一小部分 Profiles 与很大数量的捐赠相关联，而大量的 Profiles 与很小部分的捐赠相联系。

(5) 关于效用导向的时效性测度，在使用 Inclination 和 Amount 两个指标作为效用因子之间有很大的差别。然而，在分别用 Inclination 和 Amount 作为因子评估 1 年和前 4 年的效用时差别不大。

6 讨 论

关于 Profile 数据，结果说明了在效用和质量之间的关联性。不管是基于捐赠趋势的测度还是基于总的捐赠数量，更新及时、遗漏数据很少的 Profiles 与高效用有着很强的关联性，因此，效用导向的测度得分比公正测量高。基于我们的讨论，数据管理者对质量和效用之间的关联解释如下：

(1) 新的 Profiles 主要从学生注册系统中输入，只能获得所需要属性的一部分数据 (如收入水平没有提供，种族和宗教只有一部分)。结果是，大多数 Profile 记录进入系统时缺少属性值，这对评估他们的潜在贡献有负面影响。

(2) 一些 Profile 属性可能会随着时间推移而发生改变(如地址、电话、收入和婚姻状况)，如果更新不及时会限制联系校友、收集其他信息和评估它们潜在贡献的能力。

(3) 数据管理者和系统用户倾向于只有当一个人捐赠时才更新其 Profile 信息。结果是，如果一个人最近捐赠了，其 Profile 数据可能会被更新并且很少会丢失数据。而如果一个人很多年没有捐赠,关于其 Profile 质量就会变坏。

(4)在一些情况中,数据管理者和/或用户通过付钱给专业增强数据的机构更新某些捐赠者的数据。但是这样往往只是更新了有限的那些有潜力捐赠人的信息。因此,与捐赠有关系的 Profile 记录可能会被经常更新。

效用和质量之间存在关联是被数据管理者和决策者公认的，并在一定程度被反映在制定相关的管理策略中，我们的研究也揭示出了一些问题还需要继续深入研究。

(1)区别对待:在一般情况下,数据管理者在审计数据记录和属性、纠正质量缺陷甚至组织缺陷再次发生的过程中都需要采取不同的策略，有时他们会建议使用者在某些情况下不要使用某些数据或者属性。我们的研究也显示,Profile 记录的效用贡献存在一定的不确定性，不同的数据属性可能会指向不同的效用。效用与时效性存在着很敏感的关系。最后，沿着不同质量维度进行质量测度结果有很大的区别。数据质量管理的措施和策略应该有区别地应用在不同的记录子集中，这有利于发挥数据的最大效用。

(2) 归于实用：我们的结果突出显示了测度和实用的好处。我们的测度指标 Inclination 和 Amount 都反映出质量缺陷对效用的作用。而这两个测度指标在计算两个时间段的得分却没有很大的区别，这可能是因为在捐赠模式之间的关联度造成的。一个捐赠者在某一年捐赠了，其有可能也会在后面的年份里捐赠。基于这样的观察，对效用评估可以有一个很重要的改进，就是不但考虑过去的捐赠行为，也可以使用 Customer Lifetime Value (CLV) 测度技术预测将来捐赠的可能性[15]。

(3)提高完整性:结果显示，只是在记录水平分析遗漏数据的影响是不够的，还需要对属性水平的影响进行分析。一些属性的公正的完整性天

生就很高，例如学校和性别，因此，改进这些数据质量可得到的潜在效用并不大。但是对于那些完整性很低的属性，改进其质量会有很大的效用提高，如收入，我们能看到在数据质量和效用之间会有很强的关联性，因此可以为此付出更多的努力。另外一些属性，如婚姻状况、宗教等，收效会相对较低。因此，在以后的案例中，就可以考虑是否值得对于每一种数据缺陷都需要提高投资，甚至可以考虑不再存储或者管理这些属性。显而易见，我们这个研究中使用的数据源中还含有很多其他的 Profile 数据，他们同样可以使用我们提到的这种思路进行管理。

(4) 提高时效性：效用与时效性之间存在着密切的联系，过期的 Profiles 是与低的 Inclination 和 Amount 联系在一起的。这需要对 Profiles 经常进行审查。目前，数据库中近一半的 Profiles 在近 5 年内没有被更新，前面提到的属性对效用的价值关系可以帮助制定改进措施。如前面的研究显示，在过去的捐赠行为和过去 4 年的捐赠行为之间存在着联系，因此，与近期捐赠 Inclination 相关的 Profiles 应该在改进计划中有较高的优先级，例如收入。这样的属性可以在改进优先级中作为分类的标志，对那些与高收入相关联的属性的审查和更新应该更频繁。一旦一个属性被选定为分类标志，它的质量就必须维护在一个高水平，例如收入被认为是效用的一个很好的预测指标，就需要对其经常更新并消除遗漏。我们也需要改进它的分类力度(目前只有 3 个收入类别)，加入一个时间标签跟踪其变化(目前的跟踪只是停留在记录水平)。

因为只有很少部分的 Profiles 与捐赠联系在一起，很大一部分数据中存在质量缺陷，因此，同学录数据库的质量和效用还有很大的改进空间。我们的分析并不能提供一个全面的解决方案，但是却演示了相关的方法。更全面的解决方案需要对所有的相关属性进行分析、评估，还可能需要对其他效用的评估、统计工具等。

7 结 论

量化的质量评估对于不断提高数据质量非常重要。通常的评估方法主要是从公正评估的角度出发的，忽略了数据应用的环境。本研究从基于应用环境的角度探讨了一套方法，不但研究了数据缺陷的存在，还研究了它们对可获得的效用的影响。同时，应用公正和效用导向的评估方法能够使我们更深入地观察现在的数据质量管理实践的优点和缺点，并且能够指导这些实践和制订新的改进措施。本研究通过管理校友录这样一个具体的环境演示了其应用，说明了如何对比目前的集中质量管理方法，最后还提出了评估和提高数据质量的建议方法。

研究结果也显示了理解并评估数据资源效用的重要性，数据库中不同的要素对效用的贡献会有很大的区别。在一些情况下，主要的效用可能是由很小一部分的数据自己贡献的，然而在一些其他的情况中，它们的贡献可能比较平均。建模并量化效用分布，研究可能存在的不均衡可以指导改进措施并帮助排定优先级别。效用评估对于经济权衡也很重要，有些改进措施费用需求很大，所获得的效益可能会被投入抵消。对效用和费用同时考虑有助于评估经济回报并发现经济优化的策略。

本研究并不是没有局限的，它评估的是单个表格型数据库，数据管理环境中包括多个数据集，其中一些可能还会使用非表格数据结构。本研究使用的 Inclination 和 Amount 属于客户关系管理领域，其他的应用领域和商业环境(如金融、健康、保险)将需要其他不同的效用测度。本研究评估的是相对近期的用途，但是在大多数商务应用中，还需要考虑未来潜在的效用并开发定量评估工具。这对于评估以前没有使用过的新的数据质量或者通过其他的记录和属性增强已经存在的数据非常重要。

本研究考察了两种类型的数据质量缺陷：数据遗漏和是否及时更新，它们分别反应完整性和时效性。数据是否合法有效反映数据项目与所处的值域不符，这种缺陷的检查相对容易一些，能够用建议的方法测度。用不准确性表示不正确的数值，这对许多数据环境是很大的危害，包括校友录数据。我们的评估方法也适用于评估准确性，但是找出并修正这些不准确数据是很难的。

确认所有记录和属性的准确性代价很昂贵并且在实践中难以开展。因此，要想满足要求的准确性，就需要开发创新的统计采样方法。最后，我们的评估强调效用和质量之间关系的因果性，通常的观念认为，质量优先于效用。一般是减小数据缺陷，提高数据质量从而增加其有用性。我们的研究结果建议，在一些情况下，相反的因果关系可能存在，即频繁地使用和高效用能够促进某些数据元素质量的提高，而那些不经常使用的数据的质量级别会降低。这种相互的依存关系有积极的作用(例如，成本－效益质量改善，就是主要关注具有高效用的数据)，同时也有负面的作用。这种关系以及用途将被更深地研究和理解。

参考文献

[1] Shapiro C., Varian H. R.. Information Rules[M]. Cambridge, MA: Harvard Business School Press, 1999.

[2] Pipino L. L., Yang, W. L., Wang, R. Y.. Data Quality Assessment[J]. Communications of the ACM, 2002, 45 (4): 211－218.

[3] Shankaranarayanan, G., Ziad, M., Wang, R. Y.. Managing Data Quality in Dynamic Decision Making Environments: An Information Product Approach[J]. J. of Database Management, 2003, 14 (4): 14－32.

[4] Wang, R. Y., Strong, D. M.. Beyond Accuracy: What Data Quality Means to Data Consumers[J]. Journal of Management Information Systems, 1996, 12 (4): 5－34.

[5] Even, A., Shankaranarayanan, G.. Assessing Data Quality: a Value－Driven Approach[J]. The DATA BASE for Advances in Information Systems, 2007, 38 (2): 76－93.

[6] Redman, T. C.. Data Quality for the Information Age, Artech House[M]. Boston, MA, 1996.

[7] Ballou, D. P., Pazer, H. L.. Designing Information Systems to Optimize the Accuracy－timeliness Tradeoff[J]. Information Systems Research, 1995, 6 (1): 51－72.

[8] Ballou, D. P., Pazer, H. L.. Modeling Completeness Versus Consistency Tradeoffs in Information Decision Systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15 (1): 240－243.

[9] Even, A., Shankaranarayanan, G., Berger, P. D.. Economics－Driven Data Management: An Application to the Design of Tabular Datasets[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19 (6): 818－831.

[10] Lee, Y. W., Pipino, L. Strong, D. M., Wang, R. Y.. Process－Embedded Data Integrity[J]. Journal of Database Management, 2004, 15 (1): 87－103.

[11] Tayi G. K., Ballou D. P.. An Integrated Production－Inventory Model with Reprocessing and Inspection[J]. International Journal of Production Research, 1988, 26 (8): 1299－1315.

[12] Shankaranarayanan, G., Even, A.. Managing Metadata in Data Warehouses: Pitfalls and Possibilities[J]. Communications of the AIS, 2004, 14(13): 247－274.

[13] Wang R. Y.. A Product Perspective on Total Quality Management[J]. Communications of the ACM, 1998, 41(2): 58－65.

[14] Ballou D. P., Wang R., Pazer H., Tayi G. K.. Modeling Information Manufacturing Systems to Determine Information Product Quality[J]. Management Science, 1998, 44 (4): 462－484.

[15] Roberts, M. L., Berger, P. D.. Direct Marketing Management[M]. Englewood, NJ: Prentice－Hall, 1999.

[16] Kimball R. Reeves L. Ross M., Thornthwaite W.. The Data Warehouse Lifecycle Toolkit[M]. New York, NY: Wiley Computer Publishing, 2000.