数据是组织机构的资产。数据质量的高低直接影响组织机构的业务开展，甚至直接影响组织领导者的科学决策。一个组织所拥有的数据一般有两类，一是事务型数据，二是非事务型数据。事务型数据主要是由组织开展业务工作所产生的，而非事务型数据主要是通过社会化媒体获取的。加强非事务型数据的管理和分析研究，将能够增强组织的市场竞争力。美国巴布森学院的数据质量管理研究专家 Shankaranarayanan 教授等从精确度、完整性、时效性和相关性等质量维度对社会化媒体的数据质量进行了深入的分析研究。现将他们关于社会化媒体数据质量评价的研究论文编译后刊发出来，以飨读者。

——编者

社会化媒体数据质量评价初探

G. Shankaranarayanan Bala Iyer Donna Stoddard
（美国巴布森学院，美国马萨诸塞州 02457）
刘润达 〔编译〕
（中国科学技术信息研究所，北京 100038）

摘 要：首先阐述事务型数据与社会化媒体数据的概念与特性以及评价数据质量的内容和方法，然后对传统的数据质量维度如准确性、完整性、一致性、可信性、时效性、可获取性和相关性等进行分析，认为在事务型数据和社会化媒体数据中，各个质量维度的重要性有明显的不同，并确定可用于评价社会数据质量的质量维度。最后论述利用确定的质量维度对社会化媒体数据质量进行管理与评估，通过实例分析提出相应的启示与建议。

关键词：事务型数据；社会化媒体；社会化媒体数据；数据质量；质量维度；机构数据；数据质量评估
中图分类号：C931.3 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2012.02.012

Preliminary Study on Data Quality Assessment for Socialized Media

G. Shankaranarayanan, Bala Iyer, Donna Stoddard
(Technology, Operations and Information Management Babson College Babson Park, MA 02457, USA)
Translate and Edit: Liu Runda
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Firstly, this paper expatiates the concepts of transactional data and Socialized media, and the content and the method of data quality assessment, and then analyzes the dimensions of traditional data quality such as accuracy, integrality, uniformity, credibility, limitation, acquisition, relativity and so on. From it, the fundamentality of each quality dimension is different in transactional data and Socialized media, and the quality dimensions are confirmed

第一作者简介：G. Shankaranarayanan（1964- ），男，美国巴布森学院教授，博士，研究方向：技术、运营和信息管理，数据质量、元数据质量、数据质量管理。
基金项目：巴布森学院研究基金项目（BFRF）；国家软件科学项目（2011GXQ4K029）；国家自然科学基金项目（70831003）。
收稿日期：2011 年 9 月 7 日。

to evaluate Social data quality. Ultimately discusses the methods of the management and assessment for Socialized media data quality by decided quality dimension, and put forward relevant apocalypse and advice by analysis of example.

Keywords: transactional data ,Socialized media, social media data, data quality, quality dimensions, organizational data, assessment of data quality

1 引言

信息技术的迅猛发展，使组织更加容易地获取情报，获取、收集、存储、处理、分析和分发数据。以往一个组织的数据主要是围绕业务活动的开展而产生的，而近来机构数据池中增加了社会化媒体及社交网络产生的数据。这些新增加的数据可以帮助组织开展各项活动，比如营销、跟踪、客户支持、新产品创意及市场状况了解等。为了便于区分，本文将通过业务活动获得的数据称为传统数据，亦称事务型数据，而通过社会化媒体获取的非事务型数据称为社会化媒体数据。数据是一个组织机构的资产，在 ERP[1]、直销 [2] 及数据产品 [3−4] 等生产活动中，数据发挥了重要的作用。保存并维护高质量的数据资源对组织来讲至关重要，高质量的数据可以获得较高的商业回报，而低质量的数据将对组织产生负面的影响。加强数据资源的分析与研究，可以使组织避免误导，减少失败，增进信任。但是，组织机构的数据质量可能存在一定缺陷 [5−6]。实践证明，加强数据质量的管理可以事半功倍。在过去的 20 年中，关于数据质量管理研究提出的技术方法可以粗略地划分为 3 类：评测数据质量的方法和技术（如文献 [2,7−9]）、提升数据质量的技术（如文献 [10−14]）、探究数据质量对机构决策的影响（如文献 [15−19]）。这 3 类方法是在数据质量多维度结构的基础上提出的 [20−22]。目前，利用此方法研究传统的事务型数据质量比较普遍，而对于非传统的数据质量的探究却很少。本文将重点利用精确度、完整性、时效性和相关性等数据质量维度对社会化媒体数据的质量进行探讨。

本文以下内容的组织是：第二节对相关文献综述，重点阐述本研究的范围；第三节分析研究传统的数据质量评价维度，讨论其评价社会化媒体数据质量的可行性；第四节论述社会化媒体数据对传统数据质量的影响及其在实践中的应用；第五节归纳总结本研究成果，强调指出要加强对社会化媒体数据质量影响的研究。

2 文献综述

2.1 事务型数据与社会化媒体数据

在组织机构中，事务是指在商业活动中与业务相关的部分或整体活动。由事务而产生的数据是独立的，价值是自包含的，它传达了一个明确的含义，不需要任何额外的上下文来解释其重要性和意义。事务型数据有一个明确的结构，有一个预先定义的数据模型，并可以被关系数据库存取。其语义是明确的，其含义可以准确无误地从数据和结构中推断出来。该数据有一个定义良好的域，每个数据元素值都在这个域内。通常情况下，事务型数据存储在数据库中，可以方便地获取并可以明确地对这些数据进行解释。

不同于事务型数据，社会化媒体数据不是独立的，需要从上下文和语气中推断其含义。它没有一个明确的结构，即使有一个结构也由于不符合任何已知或预先定义的数据模型而不易被识别。社会化媒体数据的数据元素值是不受预先定义域限制的，其产生的目的只有创作者才知道，如博客、微博或 Facebook 等。这些目的可能不被使用的组织机构所真正发现和理解。因此，对于社会化媒体产生的数据，必须重新处理，赋予一个特定的结构。然而，这样的数据结构化可能是非常昂贵的，甚至是非常困难的。

近年来，人们越来越关注社会化媒体数据的质量问题，已在一些关于数据质量管理的文献对这个问题进行探讨。数据质量管理的传统方法依赖于对数据的结构和语义的理解。由于社会化媒体数据的结构和语义不同于事务型数据的结构和语义，因此传统管理数据质量的技术可能不适用。作为一个新的并不断发展的主题，本文的研究目的在于了解如何利用社会化媒体工具对社会化媒体上用户生成的

内容进行数据质量的管理。

2.2　数据质量主观评估

数据质量具有多维度结构。国内外许多学者对数据质量的评估进行了探究，提出了一些很好的方法及评估维度。Wang 和 Strong 建议沿准确性、完整性、有效性和传播性等维度对数据质量进行定义和评价，以更好地反映数据质量的概念 [22]。如文献 [7]、文献 [21] 等已经提出了利用不同维度、采用 0（差）和 1（完美）之间的数值定量地评价事务型数据的质量。也有人提出在复杂决策环境中计算准确性、完整性和时效性的结构比率和加权平均数 [23]。

Wang 和 Strong 指出，用户对准确性、完整性、有效性和传播性等评价维度上数据质量的认识仅仅停留在数据本身。Pipino 等认为，同一个维度既可以直接、公平、公正地利用，也可以根据上下文的内容加以利用，这主要取决于数据质量评价的目的 [21]。因此在对数据质量进行整体评估、提供数据质量管理解决方案时，既要考虑公平公正的评估方法，也要兼顾利用上下文内容进行评估的方法 [16,21]。由于公平公正的评估是基于数据的客观特征，而上下文内容的评估是主观的。因此，在评估数据质量时可能会受到以下方面因素的影响：（1）在适用范围方面：个人、机构部门作为一个整体分别对数据质量评估是不同的。个人用户更多关注的是使用特定的数据，而机构部门则更注重数据的质量。（2）在工作任务方面：使用数据的目的不同也可能影响数据质量的评估。如，用于战略决策的数据主要是汇总数据，涵盖了广泛的商业范围，而开展业务工作则主要需要详细的、琐碎的事务性数据。这两方面工作对数据质量的要求是不同的。（3）在使用角色方面：利用数据的相关者如数据获取者、数据保管者、数据消费者等的职责和工作阶段的不同，对数据质量的要求也不同 [22]。（4）在使用时机方面：不同的数据使用时机将对评估数据质量产生不同的影响。数据使用频率可以影响用户对数据质量的评估 [15,17]。（5）在个体方面：语境的评估可能会受到诸如动机、经验和参与程度等情况的影响 [17,19]。

对社会化媒体数据质量的评估只能在一定的上下文语境中或采用一定的主观方式进行，主要原因有 3 个：一是因为对数据质量的评估必须依赖于数据的结构，在可解释的数据结构基础上进行评估。而社会化媒体数据缺乏一个正式的数据结构，用户只能依靠数据的用途来解释数据的结构。二是因为社会化媒体数据的含义是模糊的，用户只能依靠上下文内容对数据进行评估。三是因为只有数据的生产者才知道数据产生的目的，数据用户只能根据数据的需要使用相关的数据。因此，评估具有半结构性的社会化媒体数据只能采取相对正确性的质量维度，而不采用精确的质量维度。

事实证明，社会化媒体数据质量的主要研究方法是链接分析和基于链接的方法 [24]。PageRank[25]、HITS[26] 等采用基于链接技术的排名算法评价问答门户网站的数据质量，Yahoo! Answers、Google Answers、Yedda 等采取问题及答案评分定级的方式评估数据质量。利用 PageRank，ExpertiseRank[27] 可以判断专家的水平。目前，有学者对在 Epinions（http://epinions.com）用户间传播问题的信任度进行了研究 [28]。Su 等 [29] 和 Jeon 等 [30] 也对问答门户网站的问题进行了探究。但是，这些研究评价数据质量的标准是每个问题的回复长度和每个问题的用户点击数，有的也使用最佳答案片段和答案的数量等特征作为评价标准。本研究是对 Agichtein 等人研究成果的扩展，不仅有对问题回复质量的评价，而且有对提出问题质量的评价。这里主要探究了社会化媒体数据的质量评估，尽管没有提出评价数据质量的方法，但利用了社会化媒体技术提供的工具对社会化媒体数据质量进行了主观评价。

2.3　研究内容与方法

事务型数据与社会化媒体数据的特性明显不同，因此事务型数据采用的数据质量管理技术不能照搬用于社会化媒体数据的管理。在事务型数据管理中，数据质量被看作是多维度结构数据，可以从几个不同的维度进行数据质量的评估 [18]。本文首先通过考察每一个维度对社会化媒体数据质量的评估作用探究其可用性。然后再利用社会化媒体技术提供的工具对社会化媒体数据质量进行度量与评估，以此探究用户对社会化媒体数据质量评价的方法。

目前，关于组织如何使用社会化媒体以及如何管理社会化媒体数据的质量的相关研究和信息很少。本文的研究目的是了解评价社会化媒体数据质量维度的含义以及组织是如何应用这几个特殊维度来管理社会化媒体数据质量的。访问调查是评价研究数据质量的一个很好的工具。据 Kvale[31] 称，访问调查可以直接与被访问者进行交流，发现其中隐含的信息。通过访谈，并对访谈数据进行分析，可

以获得新的发现。在对数据理解的基础上进行访谈，首先需要确定收集数据的计划。基于文献 [31] 所描述的技巧，可以确定进一步获取额外详细信息的备用问题。然后再确定访谈小组对象。无论是电话访问还是登门拜访，在获得访谈对象同意的情况下，对于所有的谈话内容都要进行记录，并且由研究者独立地分析谈话内容并做出总结。最后由 3 方人员进行讨论，提炼结论。

3 质量纬度与社会化媒体数据评估

质量维度与数据的管理评价有着密切的关系。分析研究质量维度可以科学地、公正地管理评价社会化媒体数据。下面本文将通过对传统的常用的数据质量评价维度，如准确性、完整性、一致性、可信性、时效性、可获取性和相关性等进行分析 [6]，确定可用于评价社会数据质量的质量维度。

3.1 准确性

准确性是指数据值与已知的基准数值相比的正确程度 [6−7,22]。它在一些数据质量文献中被广泛研究探讨，是事务型数据质量管理评价的一个重要维度。尽管有文献认为，准确性是一个固有 [18] 的维度，并在一定程度上显示出了与上下文的相关性，因此数据的准确性应该由数据应用的场景决定 [21]。而在度量时，人们并不确定基准数值是多少，因此准确性很难度量。同样地，社会化媒体数据的准确性也是难以度量的。为了验证社会化媒体数据的准确性，就要借助社会化媒体社区的附加信息对原始数据的准确性进行推断，其他数据源作为补充或佐证。与事务型数据相同，社会化媒体数据的准确性是与上下文相关联的，主要取决于数据的用途，取决于用户希望所使用的数据达到的准确程度。因此，对于社会化媒体数据质量管理与评价来讲，准确性是一个重要的维度。度量准确性的方法将在下一节介绍。

3.2 完整性

完整性是指被测验数据所表现或包含的数据元素的程度 [7−8,21, 32]。对于事务型数据，完整性主要通过以下 3 个方面进行衡量：数据模式的完整性、数据行列的完整性和数据元素的完整性 [6]。数据模式是指数据库的结构，数据模式的完整性可以衡量所有实体及属性在总体数据框架中的表示程度。数据行列的完整性可以衡量某一特定属性值的表现程度（如果某列遗漏一个值，那么就被认为数据元素不完整）。数据元素的完整性可以衡量对数据内容的揭示程度。从以上对完整性的描述可以看出，这些度量标准是基于数据结构的。有文献认为，完整性是基于上下文的，用户是根据数据的用途来感知和度量数据完整性的。对于社会化媒体数据的上下文相关与非结构化的特点，很难定义其完整性并进行度量。所以，不适合用完整性来度量社会化媒体数据质量。

3.3 一致性

一致性主要通过数值一致性和格式一致性对数据质量进行度量[21−22]。如果对于同一个商务实体（如顾客）的同一个属性（如顾客姓名）在两个不同的数据源或者同一个数据源的两个不同部分中有不同的表现值，那么就出现了数值上的不一致。如果相同数据其表现形式不同（如顾客名字在一种情况下是一个字符串，而在另一种情况下则分为名和姓两部分），那么就出现了格式上的不一致。一致性是与上下文无关并存在于数据内部的。对于事务型数据中，由于数值是在一个事先定义好的域中抽取的，并且具有数据结构，因此是可以度量的。而对于非结构化或者没有正式定义值域的社会化媒体数据，要度量其一致性是非常困难的甚至是不可能的。特别是，社会化媒体数据允许用户使用非正式的缩写，由于缺乏规范性，即使数值和格式存在一致性，也要借助软件解析“社会化媒体语言”。因此，一致性也不适用于对社会化媒体数据质量的评价。

3.4 可信度

可信度是评价数据质量中与上下文相关非常密切的维度。研究发现，数据的可信度由以下 3 个因素决定：数据来源的可靠性、数据的规范性、数据产生的时间 [22]。如果数据来源可靠并且很知名，那么数据的可信度较高。如果数据是在已知或者被接受值的范围内，那么数据的可信度更高。由于在数据质量评价的所有维度中，可信度并不是一个十分重要的维度，因此数据质量在可信度维度方面的研究并没有像其他维度（如准确性、完整性、时效性）那样有深入的探究。但是，由于任何人都可以产生数据，因此度量数据源的可靠程度是十分必要的。对于社会化媒体数据，可信度是一个非常重要而关键的维度。

3.5 时效性

时效性是数据质量另一个与上下文极其密切的维度。有文献将时效性定义为数据在完成任务或者

利用中的更新程度 [6,8,22]。由于数据从产生到获取再到利用，可能会有一个很显著的时间差。特别是，数据被手工获取并被数字化存储再到最终被理解、获取和访问，这个过程的时间差更加明显。因此，时效性对于事务型数据是一个非常重要的质量维度。而对于社会化媒体数据，时效性更显重要。社会化媒体数据主要是描述实时事件或者活动的，数据内容变化快。因为每天获取的社会化媒体数据都是有时间标识的，数据是实时被获取和传播的，因此时效性也是衡量社会化媒体数据质量的一个重要维度。

3.6　可获取性

可获取性是度量数据获得的容易程度 [33]。对于事务型数据，由于受隐私 / 安全的限制、数据敏感性的限制，一些数据可能很难获取。而对于社会化媒体数据，由于它的开放性，数据本身并非私有。随着移动终端和无线技术的发展，社会化媒体数据的访问与获取变得更加容易和迅捷，可获取性已不再是评价社会化媒体数据的重要维度。我们认为，隐私是受到保护的，具有隐私的数据是不容易获得的。在本研究中，假设社会化媒体数据是很容易访问的，不具有隐私性。

3.7　相关性

相关性是度量数据与上下文的相关程度 [22−33]。它是一个上下文相关的维度，用户只有根据数据及数据使用情况才能确定。对于研究事务型数据质量的文献并没有就这个维度进行深入的研究，除了用户赋予权重或分值外，并没有提出正式的其他的度量方法。而在社会化媒体数据质量度量中，相关性是一个重要的质量维度。社会化媒体数据量在不断扩大，用户不得不对其进行过滤以获得相关的有效数据。用户经常被问及某条评论或博文是否“有用”。由此可见，这时数据的相关性被归入“可用性”，但并不能明确地对“可用性”进行评价。

综上所述，在事务型数据中成功使用的数据质量维度并不是都适合社会化媒体数据。在事务型数据和社会化媒体数据中，各个质量维度的重要性有明显的不同。由于社会化媒体数据的非结构化和上下文相关的特质，在事务型数据质量度量中采用的方法和工具，在度量社会化媒体数据质量中不一定适合。准确性依然是社会化媒体数据的一个重要的质量维度。时效性、完整性、可获取性等质量维度对事务型数据来说非常重要，但是对社会化媒体数据却并不重要。而可信性在度量社会化媒体数据质量时较在度量事务型数据质量时的重要性明显增强。相关性和一致性对于事务型数据相对比较容易衡量，但在社会化媒体数据质量评价中还有待深入研究，并且社会化媒体数据的相关性评价常被可用性评价所代替。

4　社会化媒体数据质量管理与评估

由用户和多种已有社会化媒体技术（如 Facebook，Twitter 和 Myspace）产生了大量的社会化媒体数据。下面，将采用前面提到的不同的适用于社会化媒体数据评价的质量维度，即准确性、可信性、时效性以及可用性等对社会化媒体数据质量进行管理与评估。

4.1　数据质量管理与评估的方法

（1）提高数据准确性的方法。为确定数据的准确性，需要将数据值与基准数值或者已知的正确值进行对比。由于基准数值经常是未知的或者在度量时尚未确定，因此数据的准确性很难度量。数据的准确性可以根据历史数据利用统计的方法（见文献 [34]）进行估算。当历史数据不存在或者是不可用 [23] 时，可以利用社会化媒体数据和技术估算基准数据，还可以采用众包（crowd−sourcing）变形的方法，即将一个任务外包给一大群未知人员或者一个群体 [35]，获得难以获取的数据估值 [36]。一些大型组织机构利用众包的方式获取更加准确的数据。他们通过内部预测市场的方式，即采取激励的方式，鼓励雇员对所参与的市场进行预测，以此获得更加真实的数据。这种方式比访问调查更容易发现观点的变化。采用众包的方式，可以通过社会化媒体获取变化的数据，组织业内专家对基准数据进行估值。尽管这是一个很廉价的方案，但为了获取真实的数据，需要制定明确可行的激励措施。

（2）提高数据可信性的方法。社会化媒体数据和技术能够提升数据的可信性。数据可信性可以通过数据来源的可靠性进行判断。来自更加可靠或者更加知名的数据源的数据其可信性较高 [18,33]。文献往往采用“数据起源”或“数据世系”等词汇描述这个问题（如文献 [37]）。为方便用户判断数据源的可信性，可以随数据一起提供描述来源的元数据。但因此产生大量的信息而往往被很多用户所忽视。研究表明，美国 80% 的人对产品以及品牌的信任来自 Facebook。Facebook 上的一个公司页面被认为

是获取信任的最大数据来源[38]。可见，社会化媒体的影响力可以用来确定数据来源的可信度。此外，数据可信度还可以通过公共感知的数值范围来判断。当被接受的值域不确定时，可以通过众包的方式来确定。社会化媒体可以为数据及其产生数据的数据源创造一个代理分值。如可以加入 Amazon 或者 Salesforce.com 的一个讨论组。讨论组成员就会对社区的用户进行评价，对问题进行回复，而每一个问题或评价又会被读者根据对他们的有用程度按 5 或 7 分阶进行评分。这种方式将不断地对数据进行验证，并使其保持最近的更新状态。如果有需要更正的数据，则由数据来源媒体或者社区成员对其进行更改或添加，从而提高了数据的准确性和时效性。社区会根据回答问题的数量及质量为这些专家赋予一定的分值。数据来源的可靠度可以通过考察社区对专家的回复内容及专业知识水平的评价进行判断。

（3）提高数据时效性的方法。社会化媒体技术可以提高数据的时效性。随着移动和无线技术的进步，在实时访问社会化媒体数据时，新的信息技术提升了社会化媒体数据的时效性，同时也提升了事务型数据的时效[39]。这些技术的应用可以消除在数据获取和数据发布之间的时间差，并且显著地降低数据获取时的差错率。但是，社会化媒体本身对组织机构中事务型数据的时效性管理没有显著的作用。

（4）提高数据可用性的方法。数据的“可用性”是一个重要的数据质量属性。在收集用户评论的实际工作中，典型方法是向用户提出类似“这个数据对你多有用”的问题。然而，回答“是的，有用”或者“不，没用”并不能明显提升所评价的数据质量。如果允许用户能够根据数据质量维度对所评价数据的可用性进行评级打分，那么数据收集者将能从中获得提升数据质量的切实有效的办法。

4.2 社区用户参与数据质量评价的实例分析

社区用户的参与也可以提升数据的可靠性和准确性。数据开放范围直接影响数据质量评价的结果。组织机构数据开放的程度不同，如是对其雇员开放，还是对合作伙伴开放，还是对外全面开放，所获得的数据质量评价结果是不同的。下面介绍几家公司利用社区用户对数据质量进行评价的成功经验。

（1）Jigsaw。知识员工流动性较高。为了确保其销售数据库的完整性，Jigsaw 提出向社区开放数据库，并对更新其数据的用户奖励点数。用户可以将获得的点数兑现为对 Jigsaw 的免费访问权。结果，通过对社区用户开放数据库并采取激励机制，提升了系统中数据的完整性；通过随社区用户访问的限制，吸引了一群有知识、有效的用户，确保他们对数据库的贡献是有效的积极的；通过允许社区对数据库的更新及点数奖励，增强了对数据更新、治理的能力。

（2）Wikipedia。Wikipedia 对数据的治理也采用了同样的办法。读者可以对每一篇文章进行编辑修改。由于在系统中长期保存和维护所有的编辑记录，因此很容易找到谁在何时对文章进行了怎样的编辑和修改。文章的修改、数据的维护不再是几位编辑的工作，而是扩大到整个社区中的用户，从而提升了文章的可靠性。

（3）DARPA。美国 DARPA 主办了气球准确定位竞赛。DARPA 承诺为能够准确找到随机分布在美国的 10 个气球的人颁发 4 万美元奖金。其中有一个团队在比赛前承诺对每一个帮助找到定位气球的人给予奖励。最终，这个团队在 48 小时内找到了 10 个定位气球，获得此次比赛的胜利。

（4）NetFlix。当 NetFlix 要提升所推荐数据的质量时，他们决定加入到社区中，对获得最佳质量的团队颁奖。他们首先制定了一个完成任务的目标，然后发布了需要评价和测度的数据集。为更好地完成数据的评测，社区组建了若干小团队。几年来，Netflix 的数据质量显著提升。尽管还不清楚这样的做法主要针对的是哪一个质量维度，但可以表明，有着明显的相互关联的用户社区是能够提升数据质量的，甚至能够提升机构内部事务型数据的质量。

（5）Google。Google 在数据质量管理中成功应用了被制造企业称作源头质量管理的完整数据管理（TQM）。他们将要提升的项目数据以及雇员提供的个人共享数据发布到局域网上，并根据雇员提出的空闲时间和过去的表现分配项目任务。雇员根据项目任务输入自己的数据。因为所有的雇员可以浏览这些数据，数据的差错很快被修复，并及时对数据质量进行校验和平衡。Google 就是这样通过项目任务和数据提供者收集准确而完整的数据。

4.3 启发与建议

（1）关于社区构成：用户社区对提高数据质量

至关重要。用户社区能够提高数据的准确性，确保数据的及时更新，增强数据的实时性。用户社区还可以度量数据源的有效性。当决定让用户社区介入时，必须考虑数据的性质和个人的隐私以及组织机构的知识产权等问题，如社区的用户是否应该来自机构内部，机构数据是否应该对公众开放，对于输入信息的社区用户是否应该有要求规定等。

（2）关于回复质量：评价回复的质量是数据源可信性评价的一个重要维度。对回复质量的评价可以通过给回复数据评分进行。数据用户可以在评价回复内容的过程中获益。他们还可以通过回复者提供的简介对回复的质量进行评判。同时，数据用户可以根据公司对每个回复提供的全面的度量标准，提出关于提升回复数据可信性的办法。

（3）关于激励机制：与对数据源的评价一样，对数据的打分评价也可以提高数据的质量。为了激励数据用户对数据评价的积极性，首先应制定一个切实可行的激励方案。激励方案必须在评价开始之前确定并公布。让每一位评价者清楚地知道其中的条款。激励的办法有多种，可以是晋升职位和颁发奖章等内部奖励的办法，也可以是向参与评价者颁发奖金的办法。

（4）关于其他细节：仅仅让数据用户社区对数据质量的可用性进行评级是远远不够的。因为可用性的评价对提升数据质量并没有明显的作用。最好的办法是对“可用性”分成几个具体的质量维度，如数据是否准确、数据是否完整、数据是否可信等。这样，数据管理员可以发现在当前的形式下可能出现的问题，并能利用一个或更多特殊的评估维度提出提升数据质量的办法。特别是，可以允许数据用户对重要的质量维度赋予较高的权重。数据用户利用这些质量维度对数据的当前状态打分评价。

5　结语

本文展示了社会化媒体对数据质量影响的初步研究成果。将事务型数据质量评价的质量维度应用到社会化媒体数据质量评价中，并检验各质量维度的可用性。研究比较后认为，这些维度在社会化媒体数据质量管理和评价中其重要性是会发生变化的。于是提出，在管理评价社会化媒体数据时应进一步确定新的质量维度。最后研究利用社会化媒体工具管理社会化媒体数据质量的可行性，探讨利用社会化媒体工具管理数据质量的方法。这为更好地研究社会化媒体对数据质量的影响迈出了重要的一步。

参考文献

[1] Gattiker TF, Goodhue DL. Understanding the Local−Level Costs and Benefits of ERP through Organizational Information Processing Theory[J].Information and Management, 2004, 41: 431−443.

[2] Roberts M L, Berger P D. Direct Marketing Management [M]. 2nd ed. Prentice−Hall, Englewood Cliffs, NJ, 1999.

[3] Jain S, Kannan P K. Pricing of Information Products on Online Servers: Issues, Models, and Analysis[J]. Management Science, 2002,48(9): 1123−1142.

[4] West L A Jr.Private Markets for Public Goods: Pricing Strategies of Online Database Vendors[J]. JMIS, 2000, 17(1):59−84.

[5] Eckerson W W. Data Quality and the Bottom Line[M]. Seattle, WA: The Data Warehousing Institute, 2002.

[6] Lee Y W, Pipino L L, Funk J D, et al Journey to Data Quality[M]. Cambridge, MA: MIT Press, 2006.

[7] Redman T C. Data Quality for the Information Age[M]. Boston, MA: Artech House, 2006.

[8] Ballou D, Wang R Y, Pazer H, et al. Modeling Information Manufacturing Systems to Determine Information Product Quality[J].Management Science,1998, 44(4): 462−484.

[9] Shankaranarayanan G, Cai Y. Supporting Data Quality Management in Decision−making[J]. Decision Support Systems ,2006,42 (1): 302–317.

[10] English L. Improving Data Warehouse and Business Information Quality[M]. John Wiley & Sons, NY, 1999.

[11] 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.

[12] Hernandez M A,Stolfo S J. Real World Data Is Dirty: Data Cleansing and the Merge/Purge Problem[J]. Journal of Data Mining and Knowledge Discovery, 1998, 2(1):9−37.

[13] Kahn B K, Strong D M, Wang R Y. Information Quality Benchmarks: Product and Service Performance[J]. Communications of the ACM, 2002, 45(4):184−193.

[14] Parssian A, Sarkar S, Jacob VS. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian product[J]. Management Sci−

ence,2004, 50(7): 967−982.

[15] Chengalur−Smith I, Ballou D P, Pazer H L. The Impact of Data Quality Information on Decision Making: An−Exploratory Study[J]. IEEE Transactions on Knowledge and Data Engineering, 1999, 11 (6).

[16] Even A, Shankaranarayanan G. Utility−driven Assessment of DQ[J]. The DATA BASE for Advances in Information Systems,2007, 38 (2): 76–93.

[17] Fisher C W, Chengalur−Smith I, Ballou D P. The Impact of Experience and Time on the Use of Data Quality Information in Decision−making[J]. Information Systems Research, 2003, 14(2):170−188.

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

[19] Watts S, Shankaranarayanan G, Even A. Assessing Data Quality in Context: A Cognitive Perspective[J]. Decision Support Systems,2009, 48: 202−211.

[20] DeLone W H, McLean E R. Information Systems Success: The Quest for the Dependent Variable[J]. Information Systems Research, 1992,3(1): 60−95.

[21] Pipino L L, Lee Y W, Wang R Y. Data Quality Assessment[J]. Communications of theACM ,2002,45 (4).

[22] 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.

[23] Shankaranarayanan G, Ziad M, Wang R Y. Managing Data Quality in Dynamic Decision Environment: An Information Product Approach[J]. Journal of Database Management, 2003,14:14−32.

[24] Scott J P. Social Network Analysis: A Handbook[M]. Thousand Oaks, CA: Sage Publications, 2000.

[25] Brin S,Page L. The Anatomy of a Large−Scale[EB/OL]. [2011−08−01]. http:infolab.stanford.edu/~backrub/google.html.

[26] Kleinberg J M. Authoritative Sources in Hyperlinked Environment[J]. Journal of the ACM,1999, 46(5): 604−632.

[27] Zhang J, Ackerman M S, Adamic L. Expertise Networks in Online Communities: Structure and Algorithms [C]//16th International Conference on the World Wide Web (WWW ’07). New York, NY: ACM Press, 2007.

[28] Guha R, Kumar R, Raghavan P, et al. Propagation of Trust and Distrust[C]// Proceedings of the 13th International Conference on the World Wide Web (WWW’04). New York, NY: ACM Press, 2004.

[29] Su Q, Pavlov D, Chow J−H, Baker W C. Internet−scale Collection of Human−Reviewed Data[C]//16th International Conference on the World Wide Web (WWW’07). New York, NY: ACM Press, 2007.

[30] Jeon J, Croft B W, Lee J H, et al. A Framework to Predict the Quality of Answers with Non−textual Features [C]// Proceedings of the 29th Annual ACM SIGIR Conference (SIGIR ’06).New York, NY: ACM Press, 2006.

[31] Kvale S. Interviews: An Introduction to Qualitative Research Interviewing[M]. Thousand Oaks, CA: Sage Publications, 1996.

[32] Strong D M, Lee Y W, Wang RY. Data Quality in Context [J]. Communications of the ACM, 1997,40(5):103−110.

[33] Fisher C W, Lauria E, Chengalur−Smith I, et al Introduction to Information Quality, Advances in Information Quality Book Series[M]. Boston, MA: MIT IQ Publications, 2006.

[34] Morey R C. Estimating and Improving the Quality of Information in the MIS[J].Communications of the ACM, 1982,25 (5): 337–342.

[35] Howe J. The Rise of Crowd−Sourcing, Wired Magazine[EB/OL].[2011−08−01].http://www.wired.com/wired/archive/14.06/crowds.html

[36] Cowgill B, Wolfers J, Zitzewitz E. Using Prediction Markets to Track Information Flows: Evidence From Google[EB/OL].[2011−08−01]. http:// www.bocowgill.com/GooglePredictionMarketPaper.pdf.

[37] Madnick S, Wang R Y, Lee W. Overview and Framework for Data and Information Quality Research[J]. Journal of Information and Data Quality, 2009,1: 1−22.

[38] Webster T.The Uneasy Relationship between Twitter and Social Media Measurement[EB/OL].[2011−06−01]. http://brandsavant.com/the−uneasy−relationship−between−twitter−and−social−media−measurement.

[39] Gaynor M, Shankaranarayanan G. Implications of Sensors and Sensor−Networks for Data Quality Management[J]. International Journal of Information Quality,2008, 2(1): 75−93.