

大数据环境下 Web 数据源质量评估方法研究

赵 星¹, 李石君¹, 余 伟^{1,2}, 杨 莎¹, 丁永刚¹, 胡亚慧¹

(1. 武汉大学 计算机学院, 武汉 430072; 2. 纽约州立大学宾汉姆顿分校 计算机科学系, 纽约 13905)

摘 要: 在大数据环境下 Web 数据资源的开放性和多源性使得不同互联网平台提供的的数据质量参差不齐, 严重影响人们从互联网中有效准确地获取信息。为此, 提出一种 Web 数据源质量评估方法。建立面向多源互联网平台的统一数据模型和数据质量标准模型, 给出针对大数据全样本数据分析的质量标准度和表示方法, 并通过多维数据质量的综合评估实现 Web 数据源质量的统一度量。实验结果表明, 该方法能全面度量互联网平台的数据质量, 为用户提供准确高效的质量评价结果。

关键词: 大数据; Web 数据源; 数据质量评估; 全样本分析; MapReduce 框架

中文引用格式: 赵 星, 李石君, 余 伟, 等. 大数据环境下 Web 数据源质量评估方法研究[J]. 计算机工程, 2017, 43(2): 48-56.

英文引用格式: Zhao Xing, Li Shijun, Yu Wei, et al. Research on Web Data Source Quality Assessment Method in Big Data[J]. Computer Engineering, 2017, 43(2): 48-56.

Research on Web Data Source Quality Assessment Method in Big Data

ZHAO Xing¹, LI Shijun¹, YU Wei^{1,2}, YANG Sha¹, DING Yonggang¹, HU Yahui¹

(1. School of Computer, Wuhan University, Wuhan 430072, China;

2. Department of Computer Science, Binghamton University, State University of New York, New York 13905, USA)

[Abstract] The irregularity of data quality from different Internet platforms, which is caused by openness and multi-source, has affected negatively knowledge acquiring from Internet in big data environment. Aiming at this problem, this paper proposes a Web data source quality assessment method. It establishes a unified data model and data quality standard model for multi-source Internet platform, gives quality standards measurement and representation methods for full sample data analysis of big data, and achieves the unity of Web data source quality metrics by comprehensive assessment of multidimensional data quality. Experimental results show that this method can comprehensively measure data quality of Internet platforms, provide accurate and efficient quality evaluation results for users.

[Key words] big data; Web data source; data quality assessment; full sample analysis; MapReduce framework

DOI: 10.3969/j.issn.1000-3428.2017.02.009

0 概述

“互联网+”通过发挥互联网和计算机技术在人-机-物三元融合和应用中的资源优化和数据集成作用, 形成了大量高价值的数据资源和信息服务。这一新形态正快速地重塑互联网大数据的构造、获取和使用方式, 有效地提升了实体经济的创新力和生产力。“互联网+”作为计算机和其他学科交叉的新型技术, 成为未来网络发展的重要内容, 对于推动

新型产业创新发展具有重要作用。互联网的爆发式增长带来了前所未有的丰富数据, 使得人们能更为方便广泛地利用互联网获取信息, 但由于其开放性和多源特性, 不同互联网平台提供的的数据质量参差不齐, 严重影响了人们有效准确地从互联网中获取知识。由于这些互联网平台中部分数据存在一些不一致、缺失、冲突等错误和冗余数据, 而数据的优劣直接影响用户使用和互联网数据融合应用的可靠性, 因此需对其数据质量进行评估以便为用户提供

基金项目: 国家自然科学基金(61502350, 61672393, U1536114); 中央高校基本科研业务费专项资金项目(2042014kf0057); 湖北省自然科学基金(2014 CFB289)。

作者简介: 赵 星(1990—), 男, 硕士研究生, 主研方向为数据分析、数据库管理; 李石君, 教授、博士; 余 伟、杨 莎, 讲师、博士; 丁永刚, 副教授、博士; 胡亚慧, 讲师、博士。

收稿日期: 2016-01-05 **修回日期:** 2016-02-25 **E-mail:** 1920775356@qq.com

更丰富的数据信息。

本文所研究的数据质量评价问题是提高大数据利用价值的基础和关键,通过科学、客观的方法评价数据的真实情况,使用户能清晰、明确地了解当前数据的状态,以便对数据进行判断和抉择,并采取适当的手段提高数据的质量。

1 相关研究分析

数据是信息的载体,数据质量评估和数据获取是数据管理面临的重要问题。目前在关系数据库下的数据质量问题已取得了较好的研究成果,文献[1-2]定义数据质量是数据满足特定用户期望的程度;文献[3]以形式化的方法定义了数据的一致性、正确性、完整性和最小性,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。随着互联网的广泛应用和深刻变革,Web数据科学的高速发展给信息社会带来了巨大的机遇和挑战,大数据的数据质量和数据可用性引起了学术界和工业界的共同关注。文献[4-6]在数据质量管理的基础性研究中提出了影响数据质量的5个关键因素,即数据一致性、数据唯一性、数据精确性、信息完备性和数据时效性,并针对每个因素提出了理论框架、基础模型与算法。文献[7]提出了3Cs模型,从上下文一致性、操作一致性和时间一致性维度评估大数据质量。文献[8]从数据建模、数据存储和质量度量方面,研究了Deep Web中网页的数据质量评估问题。文献[9]介绍了基于数据管理理论进行地震数据质量控制的应用。

大数据环境下的数据质量问题普遍存在高维、多样、海量样本和不确定等问题,文献[10]研究了数据融合过程中多数据源冲突导致的质量损失问题,针对多个数据源定义了数据质量标准和损失质量标准,使用损失质量标准和统一方程舍弃高开销的数据源,并利用模糊多属性组决策来挑选合适的数据源,然后提出数据融合中的数据不一致解决方案,采用数据源质量标准扩展了近似面向对象的数据模型^[11]。文献[12]关注于数据质量控制中数据时效性的判定问题,并提出了求解算法。文献[13]从数据准确性、完备性和一致性入手,提出异构多源多模态感知数据综合数据质量的评价方法以及基于数据质量的感知数据获取算法,并对物联网中异构多源多模态数据质量进行建模和评估。

以上研究成果表明,大数据环境下的数据质量问题已经引起了广泛的关注,并已在结构化数据库领域取得非常有效的成果,然而在复杂的互联网环境下进行数据质量分析和评估还有一些不足,因此,本文围绕互联网环境下的大数据质量问题展开

深入研究。

2 Web数据源质量评估模型

影响互联网数据质量的因素有很多,在不同的应用领域和使用场景对数据质量标准的需求也不一样。由于网络数据不全是以结构化的方式展示给用户,而是通过不同组织方式的网页进行半结构化的内容输出,且每次只能获取到部分的互联网数据,为互联网数据质量的评估带来了挑战。同时以“互联网+”、移动互联网为创新应用的服务方式,以及大数据的理论和环境,为互联网数据质量提供了新的内涵。本节从数据质量标准的度量、数据质量值的标准化和数据质量值的估计优化3个部分展开研究。

2.1 数据质量标准的建立和度量

从数据质量的层次来划分,可以分为数据的质与量;从数据质量的角度来划分,可以分为客观标准和主观标准;从数据质量的度量方法来划分,可以分为独立标准和相对标准,图1给出了数据质量标准在不同维度的划分和关系。

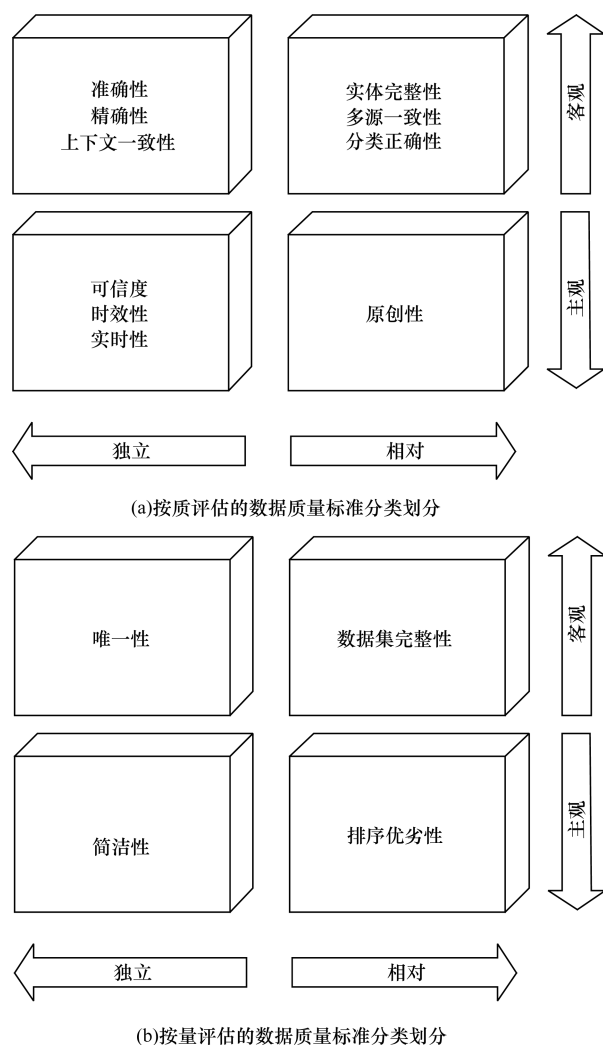


图1 数据质量标准在不同维度的分类

数据源 S 中一共有 n 个对象实例,数据源可以表示为 $S = \{D_1, D_2, \dots, D_n\}$, 其中, D_i 为每一个对象实例,该对象中一共有 m_i 个特征,可表示为 $C_i = \{ \langle K_1, V_1 \rangle, \langle K_2, V_2 \rangle, \dots, \langle K_{m_i}, V_{m_i} \rangle \}$, 其中, K 表示该特征的属性; V 表示该特征的值;不同的数据源和对象数据可能具有不同的特征数 m_i 。在已有研究基础上^[14],根据互联网应用的特征、用户的使用体验和国内外近年来的研究现状进行总结,影响互联网数据质量的质量标准主要包括以下内容:

1) 准确性。准确性表示数据源中个体数据的各个属性值描述的正确和准确程度。数据源的准确性用 q_1 表示。

$$t_{1j} = \frac{\sum_{j=1}^{m_i} F_1(V_j)}{m_i}, q_1 = \frac{\sum_{j=1}^n t_{1j}}{n} \quad (1)$$

其中, F_1 为判断该特征值正确与否的函数,若正确,则为 1;若不正确,则为 0。因此, t_{1j} 表示第 j 条数据的正确性。

2) 精确性。精确性表示描述中数据值的精度表示情况,例如出产日期“2015-02-18”和“2015 年”的精确性是不一样的。数据源的精确性用 q_2 表示。

$$t_{2j} = \frac{\sum_{j=1}^{m_i} \text{Sim}(V_j, \hat{V}_j)}{m_i}, q_2 = \frac{\sum_{j=1}^n t_{2j}}{n} \quad (2)$$

其中, Sim 表示该特征的值 V 和真实值 $\hat{V}$ 的相似度,在本文中采用编辑距离来计算 2 个值的相似度,取值范围为 $[0, 1]$ 。因此, t_{2j} 表示第 j 条数据的精确性。

3) 实体完整性。实体完整性又称为深刻性,表示对于个体的实体数据中,描述的特征属性的完备程度以及已有属性的非空值的情况。

根据多数据源的融合所映射的完整数据空间^[15],对象 D_i 所描述的实体的完备特征空间为 $\hat{C}_i = \{K_1, K_2, K_{m_i}\}$, 实体完整性可以用 q_3 表示。

$$t_{3j} = \frac{\sum_{j=1}^{m_i} F_3(V_j)}{\hat{m}_i}, q_3 = \frac{\sum_{j=1}^n t_{3j}}{n} \quad (3)$$

其中, F_3 为判断该特征值非空的函数,若非空,则为 1;若为空,则为 0。因此, t_{3j} 表示第 j 条数据的完整性。

4) 上下文一致性。上下文一致性表示在数据源中具有相同指代的不同实体之间的数据一致情况,上下文一致性指数据源内的一致性。数据源的上下文一致性可以用 q_4 表示。

$$t_{4j} = \frac{\sqrt{\sum_{j=1}^{m_i} \sum_{k=1, k \neq j}^{m_i} \text{Re}(K_j, K_k) \cdot \text{Dis}(V_j, V_k)}}{m_i}$$

$$q_4 = \frac{\sum_{j=1}^n t_{4j}}{n} \quad (4)$$

其中, Re 为 2 个特征之间的相关度; Dis 为 2 个值之间的逻辑距离。因此, t_{4j} 表示第 j 条数据的上下文一致性。

5) 多源一致性。多源一致性表示在多个数据源之间具有相同指代的不同实体之间的数据一致情况,多源一致性指其与其他数据源的一致性。数据源的多源一致性可以用 q_5 表示。

$$t_{5j} = \frac{\sum_{j=1}^{m_i} \sum_{k=1}^h \text{CS}(D_j, \hat{D}_{jk})}{h \cdot m_i}$$

$$q_5 = \frac{\sum_{j=1}^n t_{5j}}{n} \quad (5)$$

其中, CS 为 2 个数据之间的一致性^[16]; $\hat{D}_{jk}$ 为数据源 k 中表示和数据 D_j 同一实体的数据对象^[15]; h 为其他相关领域数据源的数量。

6) 可信度。可信度表示数据内容的可信程度,即被用户认可其信息正确的期望。

$$q_6 = \frac{\sum_{j=1}^n \sum_{k=1}^{P_j} R_k(D_j)}{n} \quad (6)$$

其中, R_k 表示根据用户评论得到的第 k 个用户对数据 D_j 的描述准确的评价,通过标准化使值域缩放在区间 $[0, 1]$; P_j 为评论数量。

7) 数据集完整性。数据集完整性表示满足某一个条件所筛选出的数据集的丰富和完整程度^[17]。数据集完整性通过对多个相关数据源进行抽样计算,对于 h 个数据源 $S_1, S_2, \dots, S_h$, 进行 p 次查询抽样,第 i 次、第 j 个数据源返回的有效查询结果集合为 $QR_{ij} = \{D_1, D_2, \dots\}$, 因此,第 j 个数据源的数据完备性为:

$$q_7 = \frac{1}{p} \sum_{i=1}^p \frac{QR_{ij}}{\bigcup_{j=1}^h QR_{ij}} \quad (7)$$

8) 排序优劣性。排序优劣性表示该数据源在进行数据集输出时,数据结果的排序满足用户需要的程度^[14]。数据源 S 的列表页或查询结果页所返回的结果数量为 m , 结果为 $s = \{D_1, D_2, \dots, D_m\}$, 根据用户的选择和认可程度,对数据记录倒序排列,得到新的向量 $s' = \{D'_1, D'_2, \dots, D'_m\}$ 。通过 p 次查询抽样,可以得到排序优劣性为:

$$q_8 = \frac{1}{p} \sum_{i=1}^p \left(1 - \frac{\text{Dis}(s, s')}{m_i(m_i - 1)/2} \right) \quad (8)$$

其中, Dis 表示 2 个向量的编辑距离; $m_i(m_i - 1)/2$ 表示第 i 次抽样中最坏情况下的编辑距离。

9) 时效性。时效性表示数据源中数据满足时效的程度,若数据源中过时无效数据越多,则时效性越差。网页的过时程度的度量函数如式(9)所示。

$$t_9 = \begin{cases} \frac{t - T_{\text{publish}}}{T_{\text{expire}} - T_{\text{publish}}}, & T_{\text{publish}} < t \leq T_{\text{expire}} \\ 1 + \ln\left(1 + \int_{T_{\text{expire}}}^t (T_{\text{expire}}/T_{\text{occur}} wte^t) dt\right), & \\ t > T_{\text{expire}} \end{cases} \quad (9)$$

其中, T_{occur} , T_{publish} , T_{forward} , T_{expire} 分别是关于网页与其描述信息的4维时间向量^[18]。

数据源的时效性可以定义为:

$$q_9 = \frac{\sum_{j=1}^n (1 - t_{9j})}{n} \quad (10)$$

10) 实时性。实时性表示数据源中数据更新的程度,若数据源总能最快地提供最新的数据,则实时性较好。

对于 h 个数据源 $S_1, S_2, \dots, S_h$, 在 p 个时刻 $T_1, T_2, \dots, T_p$ 进行 q 次同领域的查询抽样,每次得到的结果集合为 $QR_{ijk} = \{D_{111}, D_{222}, \dots, D_{hpq}\}$, 则第 i 个数据源的实时性可表示为:

$$q_{10} = \frac{1}{pq} \sum_{j=1}^p \sum_{k=1}^q \|NT(QR_{ijk}) - MT(\bigcup_{i=1}^h QR_{ijk})\|$$

其中, NT 为获取该结果的更新时间函数; MT 为获取该结果在各数据源中最早的更新时间函数; $\| \|$ 表示2个向量的距离范式。

11) 唯一性。唯一性表示实体在数据源中唯一存在的程度,若描述同一对象在数据源中有多个数据副本,则影响到唯一性。例如数据源 S 的列表页或查询结果页所返回的结果数量为 m , 结果为 $s = \{D_1, D_2, \dots, D_m\}$, 通过去重得到的结果向量为 $s' = \{D'_1, D'_2, \dots, D'_m\}$, 则通过 p 次查询抽样,可以得到排序优劣性为:

$$q_{11} = \frac{1}{q} \sum_{i=1}^q \left(\frac{m'}{m} \right) \quad (11)$$

12) 原创性。原创性表示数据源中数据的原创程度。对于 h 个数据源 $S_1, S_2, \dots, S_h$, 在 p 个时刻 $T_1, T_2, \dots, T_p$ 进行 q 次同领域的查询抽样,每次得到的结果集合为 $QR_{ijk} = \{D_{111}, D_{222}, \dots, D_{hpq}\}$, 则第 i 个数据源的原创性可表示为:

$$q_{12} = \frac{1}{pqh} \sum_{j=1}^p \sum_{k=1}^q \sum_{i=1}^h \|MT(QR_{ijk} \cap QR_{ijk}) - NT(QR_{ijk})\|$$

13) 简洁性。简洁性表示数据服务时内容组织的简洁性和非冗余程度,可表示为:

$$q_{13} = \frac{\sum_{j=1}^n \text{Path}(D_j) + (1 - RP(D_j)) + RL(D_j)}{n} \quad (12)$$

其中, Path 表示从索引页到达该数据页的路径长度函数; RP 表示数据中各特征属性之间的重复度; RL 表示页面中相关信息所占的信息量比重。

14) 分类正确性。分类正确性表示数据源中数据的分类正确与合理程度,可表示为:

$$q_{14} = \frac{\sum_{j=1}^n \sum_{i=1}^{m_j} R(D_j, C_i)}{n} \quad (13)$$

其中, C_i 表示数据 D_j 中第 i 个分类特征; m_j 为数据 D_j 中分类特征的维度大小; R 表示数据和分类特征的相关程度函数。

2.2 数据质量值的标准化

各质量标准存在不同的值域和方向,会对数据质量的融合产生影响。例如 q_1 表示的准确度越高,数据质量越好; q_{10} 表示的过时程度越大,则数据质量越差; q_{13} 的取值范围为 $[0, 3]$; q_{12} 的取值范围为 $(-\infty, 0]$ 。因此,需要对各质量值进行标准化,本文采用极差标准化,根据质量值对数据源数据质量的正相关和负相关影响分别采用正向极差标准化和负向极差标准化。

对于数据集中的所有样本数据,质量标准 i 在第 j 个数据源中的值为 q_{ij} , 其中,最大值为 $\max(q_{ij})$; 最小值为 $\min(q_{ij})$, 因此,第 j 个数据源的质量标准 i 的正向极差标准化值为:

$$q'_{ij} = \frac{q_{ij} - \min(q_{ij})}{\max(q_{ij}) - \min(q_{ij})} \quad (14)$$

同理,负向极差标准化值为:

$$q''_{ij} = \frac{\max(q_{ij}) - q_{ij}}{\max(q_{ij}) - \min(q_{ij})} \quad (15)$$

其中,特殊情况当最大值与最小值相等时,则 $q'_{ji}, q''_{ji} = 1$ 。经过此标准化后各数据源的质量标准的值域都为 $[0, 1]$ 。

2.3 数据质量值的估计优化

互联网平台的数据质量来自于对各质量标准的评估计算,质量标准的依赖于各数据的质量标准,因此,传统方法一般采用在数据集中进行抽样,然后计算各抽样数据的质量标准后进行平均求值。然而随着互联网的爆发式增长,使得互联网数据服务呈现出规模性和多样性特点,使得采用抽样方法获取数据样本来度量互联网平台的总体数据质量的效果越来越差。实验发现式(1)~式(13)中对各样本数据的质量值求平均值不能有效地反映数据的内在规律和特征,为大数据环境下的数据质量评估带来了巨大的挑战。同时由于大数据技术的发展,使得对数据源中的海量数据进行全样本分析成为可能,也对高效评估数据质量提出了更高的要求。

数据源 S_j 中有 p 个数据,对于质量标准 v_i , 第 k 个数据可度量为 b_{ik} , 因此数据源 S_j 在质量标准 v_i 上的质量向量为 $(b_{i1}, b_{i2}, \dots, b_{ip})$, 例如某数据源的各数据的准确度分布如图2所示。由于 p 代表了数据源中的所有数据的数量,因此该向量为全抽样数据,反映了数据源的完整数据质量情况。由于在不同的数据中具有不同的质量分布,因此均值法不能有效表示数据源在该质量标准上的质量情况。建立

数据集 $G_1, G_2, \dots, G_t$, 对于任意 $h = 1, 2, \dots, t$, 满足 $G_h \subset S_i$, G_h 为满足该标准的质量在 $\left(\frac{h-1}{t}, \frac{h}{t}\right]$ 区间的个体数据的集合, 集合的元素个数为 C_h 。 t 表示将该指标等分的个数, 取值为保持等分区间内数据量连续的最大个数。例如对于准确度, 实验统计得到的最佳等分个数为 10。如图 3 所示, 为某数据源的准确度划分成 t 等分后的各区间元素个数分布。

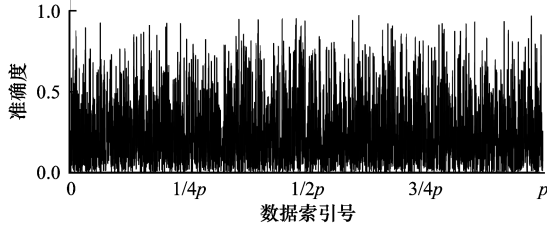


图 2 某数据源的各数据准确度分布

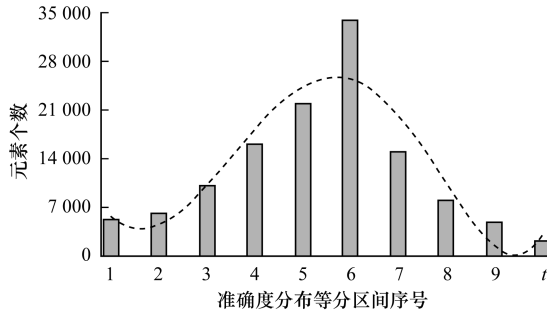


图 3 某数据源准确度在各区间的元素个数分布

设每个数据第 i 个质量标准的质量分布函数为 $f_i(x)$, 则 $f_i(x)$ 的核密度估计为:

$$\hat{f}_i(x) = \frac{1}{pt} \sum_{j=1}^p K\left(\frac{x - b_{ij}}{t}\right) = \frac{1}{p} \sum_{j=1}^p K_i(x - b_{ij}) \quad (16)$$

其中, K_i 为针对该质量标准所设定的高斯核函数:

$$K_i(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \quad (17)$$

该函数反映了在该数据源中质量标准 i 在全样本数据中的分布情况, 求解其在数据质量上的总体期望可表示为:

$$E(\hat{f}_i) = -\frac{1}{nt} \int K^2(u) du f(x) + o\left(\frac{1}{nt}\right) \quad (18)$$

据此可根据全样本分布数据来求解各质量标准的评估值。

3 多维数据质量综合评估方法

上文介绍了对互联网平台的数据质量具有不同方面影响的 14 个质量标准, 而各平台在不同的质量标准上可能效果不同, 因此, 需要一个统一的综合评估方法来度量互联网平台的数据质量, 本文所提出的基于质量标准度量的全数据质量评估方法 (Data Quality Assessment Method Based on Quality Standard Measurement, QABQSM), 可以将每个互联网平台的数据源质量函数表示为:

$$Q(S) = \sum_{i=1}^n \omega_i \hat{q}_i \quad (19)$$

其中, $\hat{q}_i$ 表示标准化和估计优化后的质量标准值; 权重 $\sum_{i=1}^n \omega_i = 1, \omega_i \geq 0$ 为第 i 个质量标准的质量权重, $i = 1, 2, \dots, n, n$ 为质量标准的数量, 在本文研究中 $n = 14$, 则 Web 数据源质量评估的问题就转换成多属性的决策分析问题, 其最基本问题为确定各数据质量标准的权重。由于互联网服务的多样性和广泛性, 不同质量标准的权重有不同的确定方法, 根据质量标准的特性来进行确定, 称为客观权重, 根据用户对质量标准的敏感和偏好来确定, 称为主观权重, 综合主客观的确权方法, 称为主客观权重。本文首先分别分析客观确权 and 主观确权方式, 再综合给出主客观综合权重的计算方法。

在客观确权中, 根据客观的数据关系和数据规律来确定, 不考虑用户的特性和偏好。在 m 个互联网数据源平台中, 有决策矩阵 $A = [a_{ij}]_{mn}$, 其中 $a_{ij} = q_i(s_j)$ 表示第 j 个服务 s_j 在质量标准 v_i 上的取值, 按式 (14)、式 (15) 将其标准化后得到标准矩阵 $B = (b_{ij})_{mn}$, 构造以下决策函数:

$$\begin{cases} \min f_1 = \sum_{i=1}^n \sum_{j=1}^m (b_i^* - b_{ij})^2 \omega_i^2 \\ \text{subject to } \sum_{i=1}^n \omega_i = 1 \\ 0 \leq \omega_i \leq 1; i = 1, 2, \dots, n \end{cases} \quad (20)$$

其中, $b_i^* = \max\{b_{i1}, b_{i2}, \dots, b_{im}\}$ 表示属性 v_i 在决策矩阵中的理想值; 目标函数 f_1 表示满足理想值与其他候选服务对应 v_i 属性值的方差达到最小, 则得到的各数据源的质量 $Q^*(S_j)$ 所形成的有序队列 S_{sort}^* 满足客观评价标准, 所得到的权重向量解为:

$$\omega^* = (\omega_1^*, \omega_2^*, \dots, \omega_n^*)^T \quad (21)$$

其中, ω_i^* 为式 (20) 均方差法所确定的权值。

在主观确权中, 根据用户给出的属性比较矩阵 $D = [d_{kj}]_{nn}$, 其中 d_{kj} 表示属性 v_k 较 v_j 的相对重要程度, 则求解权重向量使得到的各数据源的质量 $Q^{**}(S_j)$ 所形成的有序队列 S_{sort}^{**} 满足主观评价顺序, 因此, 构造决策函数如下:

$$\begin{cases} \min f_2 = \sum_{i=1}^n \sum_{k=1}^n (d_{ik} \omega_k - \omega_i)^2 \\ \text{subject to } \sum_{i=1}^n \omega_i = 1 \\ \sum_{i=1}^n d_{ik} = \frac{\sum_{i=1}^n \omega_i}{\omega_k} \\ 0 \leq \omega_i \leq 1; i = 1, 2, \dots, n \\ d_{ik} > 0; d_{ik} = 1/d_{kj}; d_{kk} = 1 \end{cases} \quad (22)$$

所得到的权重向量解为:

$$\omega^{**} = (\omega_1^{**}, \omega_2^{**}, \dots, \omega_n^{**})^T \quad (23)$$

主观确权 and 客观确权各有优缺点, 主观确权反映

了用户的喜好和要求,所得到的结果更能满足用户的偏好,但由于人为因素的不确定性会造成评价结果的随意性;客观确权根据逻辑关系和理论规律进行评价,但没有体现用户的观点。因此,本文结合主观确权和客观确权的优点,采用主客观综合权重来进行度量,根据式(20)和式(22)构造主客观综合决策函数如下:

$$\begin{cases} \min f_1 = \sum_{i=1}^n \sum_{j=1}^m (b_i^* - b_{ij})^2 \omega_i^2 \\ \min f_2 = \sum_{i=1}^n \sum_{k=1}^n (d_{ik} \omega_k - \omega_i)^2 \\ \text{subject to } \sum_{i=1}^n \omega_i = 1 \\ 0 \leq \omega_i \leq 1; i=1, 2, \dots, n \end{cases} \quad (24)$$

式(24)可转换为如下多目标规划问题:

$$\begin{cases} \min f_1 = \alpha \sum_{i=1}^n \sum_{j=1}^m (b_i^* - b_{ij})^2 \omega_i^2 \\ \quad + \beta \sum_{i=1}^n \sum_{k=1}^n (d_{ik} \omega_k - \omega_i)^2 \\ \text{subject to } \sum_{i=1}^n \omega_i = 1 \\ 0 \leq \omega_i \leq 1; i=1, 2, \dots, n \end{cases} \quad (25)$$

其中, α, β 表示主、客观确权模式的相对重要程度,并且满足 $\alpha + \beta = 1, 0 < \alpha, \beta < 1$ 。

在不考虑权重的非负约束的情况下,由式(25)可构造 Lagrange 函数:

$$L = \alpha \sum_{i=1}^n \sum_{j=1}^m (b_i^* - b_{ij})^2 \omega_i^2 + \beta \sum_{i=1}^n \sum_{k=1}^n (d_{ik} \omega_k - \omega_i)^2 + 2\lambda (\sum_{i=1}^n \omega_i - 1) \quad (26)$$

设 $\frac{\partial L}{\partial \omega_h}, h=1, 2, \dots, n$, 则有:

$$\alpha \sum_{j=1}^m (b_h^* - b_{hj})^2 \omega_h + \beta [\sum_{i=1}^n (d_{ih} \omega_h - \omega_i) d_{ih} - \sum_{k=1}^n (d_{ih} \omega_k - \omega_h)] + \lambda = 0 \quad (27)$$

结合约束 $\sum_{i=1}^n \omega_i = 1$, 将这 $n+1$ 个等式构造成矩阵:

$$\begin{bmatrix} \mathbf{Z} & \mathbf{e} \\ \mathbf{e}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega} \\ \kappa \end{bmatrix} = \begin{bmatrix} \mathbf{O} \\ 1 \end{bmatrix} \quad (28)$$

其中, $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_n)^T$; $\mathbf{Z} = [z_{ij}]_{n \times n}$; $\mathbf{e} = (1, 2, \dots, 1)^T$, $\mathbf{O} = (0, 0, \dots, 0)^T$, 矩阵 $\mathbf{Z}$ 的每一个元素为:

$$\begin{cases} z_{ij} = -\beta(d_{ij} + d_{ji}), i \neq j \\ z_{ii} = \beta(\sum_{k=1}^n d_{ki}^2 + n - 2) + \alpha \sum_{k=1}^m (b_i^* - b_{ki})^2 \end{cases} \quad (29)$$

求解方程可得:

$$\boldsymbol{\omega} = \mathbf{Z}^{-1} \mathbf{e} / \mathbf{e}^T \mathbf{Z}^{-1} \mathbf{e} \quad (30)$$

其中, $\boldsymbol{\omega}$ 是主客观综合确定的权值。由此,可以根据式(19)求解各互联网平台的综合数据质量。

4 实验与结果分析

为检验本文方法针对大数据环境下的互联网平台数据质量评估的有效性,在真实数据环境下进行实验。电子商务是目前“互联网+”应用最为广泛和深入的领域,本文实验的数据集来源于对京东、苏宁易购、国美在线、易讯网和一号店共5个中国主流的综合型B2C电子商务平台的实时数据。实验采用MapReduce框架的开源实现。

4.1 实验环境及数据集

本文实验利用虚拟化技术,通过开源的虚拟化软件 Oracle Virtualbox 将高性能 NF8560M2 服务器虚拟出30个主机节点,并以此为底层的分布式硬件环境。每个节点虚拟出一个 XENO E7-4807 的 CPU 和 4 GB 的内存,主机采用 Windows Server 2008 R2 的操作系统,节点采用 Ubuntu 12 操作系统的 Hadoop 0.20.2 平台。实验环境是在已有的 Web 大数据分析处理框架基础上建立,框架结构如图4所示。

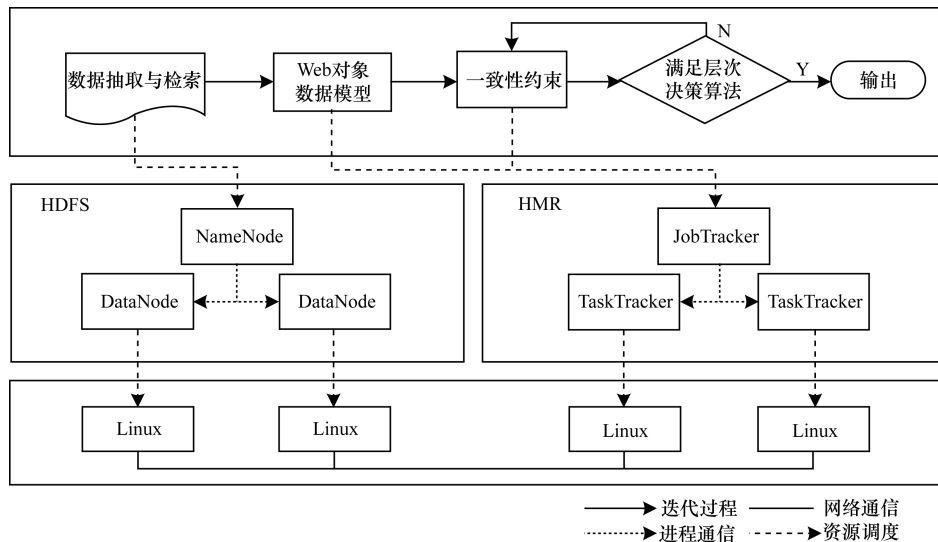


图4 实验原型系统框架

为更充分地体现算法效果,实验采用真实的数据集,通过自定义爬虫采集 5 个中国主流的综合 B2C 电子商务平台的商品信息,并对部分抽样数据加入人工标注。数据源平台分别定义为 S_1, S_2, S_3, S_4, S_5 , 分别对应京东、苏宁易购、国美在线、易讯网和一号店。整个采集的商品信息包括 10 大类 38 个二级分类 938 781 件商品(部分分类进行了整合,不含图书类和虚拟类产品),不同平台的不同大类商品数量如表 1 所示。

表 1 数据集各分类的分布数量

类别名	S_1	S_2	S_3	S_4	S_5
手机	10 192	7 353	22 527	11 856	10 509
数码	575 640	12 580	36 738	66 521	45 393
电脑	411 752	31 163	83 173	18 741	9 473
办公用品	386 072	31 065	40 509	98 379	64 852
家居用品	101 740	21 213	24 309	33 452	38 280
服饰首饰	314 144	270 138	121 134	62 143	74 997
个护化妆	94 944	168 250	53 928	20 541	13 988
运动户外	61 040	6 954	32 948	5 214	2 448
食品饮料	155 692	26 647	10 294	5 784	6 211
生活用品	16 843	91 514	22 125	24 649	11 041
合计	2 128 059	666 877	447 685	347 280	277 192

所有数据按数据源的键值型存储在 Redis 数据库并通过 HFDS 持久化,每个键值对 $\langle K, V \rangle$ 表示一个特征,例如 $\langle \text{Material}, \text{Plastic} \rangle$, 每一组键值对构成一个数据对象的实例并通过主键建立关联,其结构为 $D_i = \{ \langle K_1, V_1 \rangle, \langle K_2, V_2 \rangle, \dots, \langle K_{m_i}, V_{m_i} \rangle \}$, 每个数据源都由相关的数据 D_i 组成。

4.2 实验结果

实验首先通过对各互联网平台的数据进行抽取,再对每条个体数据进行各数据质量标准的计算,求解每条数据的各标准质量值。表 2 描述了对各数据源的所有数据求解对应质量标准所耗费的计算时间。

然后根据所有个体数据的质量标准值,在大数据全样本分析的基础上通过式(3)求解各数据源的数据质量标准(值区间 $[0, 1]$),结果如表 3 所示。分布图如图 5 所示,结果表明,各平台在不同质量标准下有不同的优劣性。因此,需要对各质量标准建立质量权重,通过主客观综合权重进行度量,根据式(30)对主观分析和客观分析进行综合,求解质量权重向量为: $w = (0.117\ 5, 0.035\ 9, 0.062\ 9, 0.045\ 3, 0.017\ 0, 0.066\ 2, 0.041\ 3, 0.100\ 4, 0.061\ 9, 0.100\ 8, 0.017\ 2, 0.155\ 3, 0.087\ 2, 0.091\ 2)$ 。

表 2 MapReduce 框架下各数据源计算时间 s

指标	S_1	S_2	S_3	S_4	S_5
准确性	2 069	611	428	352	255
精确性	2 044	655	471	327	283
实体完整性	3 911	1 298	957	759	600
上下文一致性	6 982	2 136	1 356	966	770
多源一致性	7 530	2 274	1 620	1 229	1 036
可信度	2 175	697	463	371	261
数据集完整性	959	316	226	182	127
排序优劣性	1 839	544	338	252	218
时效性	3 052	986	645	470	457
实时性	977	331	203	171	136
唯一性	4 710	1 403	954	666	578
原创性	5 914	1 728	1 239	918	730
简洁性	2 230	745	460	391	334
分类正确性	3 848	1 427	848	737	508

表 3 各数据源的质量标准值

指标	S_1	S_2	S_3	S_4	S_5
正确性	0.92	0.86	0.97	0.92	0.89
精确性	0.87	0.78	0.79	0.73	0.71
实体完整性	0.78	0.7	0.93	0.64	0.79
上下文一致性	0.80	0.62	0.86	0.70	0.72
多源一致性	0.76	0.68	0.50	0.76	0.60
可信度	0.93	0.89	0.99	0.92	0.93
数据集完整性	0.96	0.97	0.70	0.91	0.63
排序优劣性	0.87	0.86	0.70	0.88	0.73
时效性	0.89	0.78	0.61	0.85	0.70
实时性	0.93	0.75	0.77	0.81	0.78
唯一性	0.78	0.81	0.82	0.92	0.78
原创性	0.83	0.79	0.76	0.86	0.94
简洁性	0.69	0.74	0.88	0.79	0.85
分类正确性	0.96	0.98	0.84	0.92	0.84

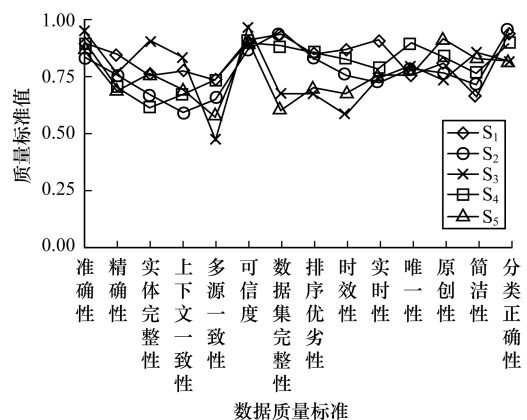


图 5 各数据源的质量标准值分布

根据式(19)得到各数据源的数据质量得分如表4所示。

表4 各数据源的总体数据质量得分

数据源	数据质量得分
S_1	0.865
S_2	0.812
S_3	0.814
S_4	0.842
S_5	0.816

4.3 结果分析

目前尚缺少基于多维质量标准综合评估的Web数据源质量评估方法,为验证本文方法的有效性,一方面通过专家训练和评价分析^[19]得到真实结果集,另一方面与Web文档中基于词法或语法特征分析的质量评估方法^[20-21]进行比较。在专家训练方面,通过20名对实验数据集知识领域熟悉的专家从多个角度进行综合评估打分,最后进行平均汇总,并将百分制转换到 $[0,1]$ 区间。在评价分析方面,通过分析各数据源中具有评价反馈页面的评价内容,计算出用户对数据源的质量认可程度^[19]。在实验比较上,通过本文的QABQSM方法与FQA方法^[20]和SVR方法^[21]进行比较。FQA通过识别主题相关的上下文档,实现基于事实提取的Web文档内容数据质量评估。SVR方法综合文本、评论和网络特征来训练模型,实现了基于特征训练向量机模型的质量评估方法。

实验结果如图6所示,针对有效数据,专家评估结果和评价分析结果保持了相同的走势,这说明专家评估的结论能作为真实结果集进行结论验证。由于页面评论的稀疏性,评价分析方法不能作为通用的质量评估方法,仅适合对有效数据进行验证。实验结果表明,QABQSM方法所得到的数据质量排序(S_1, S_4, S_5, S_3, S_2)保持了和真实结果集相同的排序,平均误差小于FQA和SVR方法。这是因为QABQSM方法从数据源各数据的本质进行分析,完全符合数据质量的评估标准,考虑了不同维度和不同程度的数据质量偏移,比其他方法更具全面性。但实验中发现,QABQSM方法的计算性能要逊色于其他方法,这是因为其所要计算的数据样本大小和分析复杂度都比较高,这将在今后研究中进行优化。

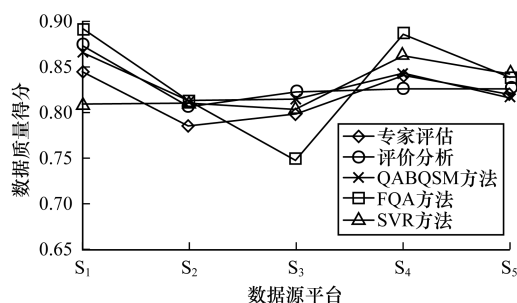


图6 不同评估方法的数据质量度量结果

5 结束语

针对Web大数据环境下异构数据源的数据量庞大、数据来源复杂等现象,以及其所导致的数据源质量参差不齐、用户体验不佳等问题,本文提出了有效评估大数据环境下互联网数据质量的方法。通过建立面向多源互联网平台的统一数据模型和数据质量标准模型,构建面向大数据全样本数据分析的质量标准度量 and 表示方法,研究多维数据质量的综合评估方法,实现互联网数据质量的统一度量。通过对真实的互联网数据进行抽取和分析,并在MapReduce并行框架下进行计算,实验结果表明该方法的数据质量度量耗时与数据量成线性关系,能较好地适应大数据环境,并且所得到的数据质量标准权重符合质量的评估要求,结果也满足实际情况,体现了本文方法的有效性。然而大数据环境下的互联网还有更多的复杂性,如各种不同的数据模态、各种不确定的数据误差以及数据质量标准之间的非线性关系,如何在复杂的应用环境下更高效地解决问题将是下一步的研究重点。

参考文献

- [1] Charikar M S. Similarity Estimation Techniques from Rounding Algorithms [C]//Proceedings of the 34th Annual ACM Symposium on Theory of Computing. New York, USA: ACM Press, 2002: 380-388.
- [2] Cappiello C, Francalanci C, Pernici B. Data Quality Assessment from the User's Perspective [C]//Proceedings of 2004 International Workshop on Information Quality in Information Systems. New York, USA: ACM Press, 2004: 68-73.
- [3] Aebi D, Perrochon L. Towards Improving Data Quality [C]//Proceedings of International Conference on Information Systems & Management of Data. Washington D. C., USA: IEEE Press, 1993: 273-281.
- [4] Fan Wenfei, Geerts F. Foundations of Data Quality Management [J]. Synthesis Lectures on Data Management, 2012, 4(5): 211-217.
- [5] Fan Wenfei, Geerts F, Ma Shuai, et al. Data Quality

- Problems Beyond Consistency and Deduplication[M]//Tannen V, Wong L. In Search of Elegance in the Theory and Practice of Computation. Berlin, Germany: Springer, 2013:237-249.
- [6] Fan Wenfei. Data Quality: Theory and Practice [C]//Proceedings of 2012 International Conference on Web-age Information Management. Washington D. C., USA: IEEE Press, 2012:1-16.
- [7] Caballero I, Serrano M, Piattini M. A Data Quality in Use Model for Big Data [J]. Future Generation Computer Systems, 2015, 63(C):123-130.
- [8] Nie Tiezheng, Yu Ge, Shen Derong, et al. An Approach to Assess the Quality of Web Pages in the Deep Web[C]//Xu Jianliang, Yu Ge, Zhou Shuigeng. Database Systems for Adanced Applications. Berlin, Germany: Springer, 2011:514-525.
- [9] Ringler A T, Hagerty M T, Holland J, et al. Baker: The Data Quality Analyzer: A Quality Control Program for Seismic Data [J]. Computers & Geosciences, 2015, 76:96-111.
- [10] Wang Xin, Huang Linpeng, Zhang Yi, et al. A Solution of Data Inconsistencies in Data Integration-designed for Pervasive Computing Environment[J]. Journal of Computer Science and Technology, 2010, 25(3):499-508.
- [11] Wang Xin, Huang Linpeng, Xu Xiaohui, et al. A Solution for Data Inconsistency in Data Integration[J]. Journal of Information Science & Engineering, 2011, 27(2):681-695.
- [12] 李默涵, 李建中, 高 宏. 数据时效性判定问题的求解算法[J]. 计算机学报, 2012, 35(11):2348-2360.
- [13] 马 茜, 谷 峪, 张天成, 等. 一种基于数据质量的异构多源多模态感知数据获取方法[J]. 计算机学报, 2013, 36(10):2120-2131.
- [14] 余 伟, 李石君, 文利娟, 等. 基于数据质量的 Deep Web 数据源排序[J]. 小型微型计算机系统, 2010, 31(4):641-646.
- [15] 胡亚慧, 李石君, 余 伟, 等. 大数据环境下的电子商务商品实体同一性识别[J]. 计算机研究与发展, 2015, 52(8):1794-1805.
- [16] 余 伟, 李石君, 杨 莎, 等. Web 大数据环境下的不一致跨源数据发现[J]. 计算机研究与发展, 2015, 52(2):295-308.
- [17] 余 伟, 李石君, 洪 辉, 等. 基于覆盖关系的 Deep Web 数据源排名[J]. 计算机研究与发展, 2007, 44(z3):29-34.
- [18] Gan Lin, Liu Yongning, Yang Sha, et al. Web Temporal Inconsistency Modeling Based on Web Time Axis[J]. Journal of Computational Information Systems, 2012, 8(3):1063-1070.
- [19] Zhang Xiuzhen, Cui Lishan, Wang Yan. Commtrust: Computing Multi-dimensional Trust by Mining E-commerce Feedback Comments [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7):1631-1643.
- [20] 韩京宇, 陈可佳. 基于事实抽取的 Web 文档内容数据质量评估[J]. 计算机科学, 2014, 41(11):247-251.
- [21] Dalip D H, Goncalves M A, Cristo M, et al. Automatic Assessment of Document Quality in Web Collaborative Digital Libraries [J]. Journal of Data and Information Quality, 2011, 2(3):1-30.

编辑 陆燕菲

(上接第 47 页)

- [8] Kang S N, Lee S, Hur J, et al. Multi-lane Detection Based on Accurate Geometric Lane Estimation in Highway Scenarios [C]//Proceedings of IEEE Intelligent Vehicles Symposium Proceedings. Washington D. C., USA: IEEE Press, 2014:221-226.
- [9] Tran T T, Bae C S, Kim Y N, et al. An Adaptive Method for Lane Marking Detection Based on HSI Color Model[J]. Communications in Computer & Information Science, 2010, 93:304-311.
- [10] Dorj B, Lee D J. A Precise Lane Detection Algorithm Based on Top View Image Transformation and Least-Square Approaches[J]. Journal of Sensors, 2016, 2016:1-13.
- [11] Liu Weirong, Li Shutao, Huang Xu. Extraction of Lane Markings Using Orientation and Vanishing Point Constraints in Structured Road Scenes [J]. International Journal of Computer Mathematics, 2014, 91(11):2359-2373.
- [12] Cho J H, Tsogtbaatar E, Kim S H, et al. Improved Lane Detection System Using Hough Transform with Super-resolution Reconstruction Algorithm and Multi-ROI[C]//Proceedings of 2014 International Conference on Electronics, Information and Communications. Washington D. C., USA: IEEE Press, 2014:1-4.
- [13] Zhang Weiwei, Song Xiaolin. Apartitioned Approach to Real Time Lane Detection at Different Weather Conditions[J]. Advanced Materials Research, 2013, 671-674:2870-2874.
- [14] Jung S, Youn J, Sull S. Efficient Lane Detection Based on Spatiotemporal Images [J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 17(1):1-7.
- [15] Ballard D H. Generalizing the Hough Transform to Detect Arbitrary Shapes [J]. Pattern Recognition, 1987, 13(2):111-122.

编辑 陆燕菲