 DOI: https://doi.org/10.12349/iser.v5i7.3111

Review of Statistical Analysis of Information Quality in Contemporary Big Data Environment

Yajing Wei

Health Comprehensive Law Enforcement Brigade of Zhen'an County, Shaanxi Province, Shangluo, Shaanxi, 711500, China

Abstract

With the advent of the era of big data, the problem of information quality has been paid attention to by more researchers. This paper summarizes and summarizes the progress of statistical analysis of information quality in the current big data environment. We evaluate information quality from four dimensions: accuracy, consistency, comprehensiveness, and timeliness. We also introduce mainstream statistical analysis methods for information quality, including descriptive statistics, regression analysis, cluster analysis, etc. Research has shown that new statistical methods adapted to the big data environment have the potential to address information quality issues. For structured and unstructured data, data quality issues exhibit different characteristics, and there should be differences in the methods used to solve the problems. For example, for structured data, traditional methods such as descriptive statistics and regression analysis can to some extent discover and address quality issues in the data; Unstructured data requires complex methods such as text analysis and semantic analysis.

Keywords

big data environment; information quality; statistical analysis; data security; artificial intelligence technology

当代大数据环境下信息质量统计分析研究综述

卫亚静

陕西省镇安县卫生健康综合执法大队，中国 · 陕西 商洛 711500

摘　要

随着大数据时代的到来，信息质量的问题也被更多研究者关注。论文详细梳理和总结了当前大数据环境下信息质量的统计分析研究进展。我们从信息的精确性、一致性、全面性和时效性四个维度来评估信息质量，同时介绍了主流的信息质量统计分析方法，包括描述统计、回归分析、聚类分析等。研究表明，适应大数据环境的新型统计方法具有解决信息质量问题的潜在性能。对于结构化和非结构化数据，数据质量问题表现出不同的特点，以及解决问题的方法也应有所差异。例如，对于结构化数据，用描述统计、回归分析等传统方法可以在一定程度上发现和牵引数据的质量问题；而非结构化数据则需要用到文本分析、语义分析等复杂的方法。

关键词

大数据环境；信息质量；统计分析；数据安全；人工智能技术

1 引言

随着信息技术的飞速发展，大数据时代已然来临。每天，我们都会产生大量的数据信息，这些巨大的数据量为我们提供了丰富的信息，并使我们能够更准确地分析和设计策略，优化决策。然而，数据的增长速度并没有带来信息质量的相应提高，反而使数据的管理和利用面临更大的挑战。进入大数据时代后，对数据质量的问题提出了更高的要求，也带来了更多的挑战，如何更准确、全面、及时地获取和利用信息，成为当前亟待研究的重要课题。据统计，每年因为低质量信息造成的经济损失高达数千亿，这显示了提升信息质量，特别是在大数据环境下提升信息质量的紧迫性。因此，论文将着重梳理和总结当前大数据环境下的信息质量统计分析研究情况，以期为解决当前面临的信息质量问题，提供理论依据和技术手段。我们将从信息的精确性、一致性、全面性、时效性等多个维度出发，探索适应大数据环境的新型统计分析方法，以期望在大数据和人工智能技术的持续发展过程中，能够更好地提高信息质量，并有效地解决当前属于大数据特有的问题。

【作者简介】卫亚静（1981–），女，中国陕西商洛人，本科，副高级统计师，从事统计研究。

2 大数据环境下的信息质量问题

2.1 大数据环境简介

近年来，伴随着科技的飞速发展和信息化进程的加快，人类社会迅速进入了大数据时代 [1]。大数据环境指的是在现代科技手段支持下，通过对海量数据的采集、存储、处理、

分析和应用，来发现新的商业模式、提升社会治理能力和推动科学研究进步的一种新型信息技术环境。大数据不仅在数据量上达到前所未有的规模，而且在数据的多样性、速度和高价值方面也表现出显著特征。

大数据环境的核心特性之一是数据量的巨大规模。每天，全球各类传感器、网络平台和移动设备等都会产生海量的数据，根据国际数据公司的估测，全球数据总量将在未来几年内呈现指数级增长。数据类型的多样性也是大数据环境的另一个重要特征，不仅包括传统的结构化数据，如数据库记录，还涵盖了非结构化和半结构化数据，如文本、图片、音视频等。

大数据环境下，数据的产生和传输速度显著加快，实时数据的处理需求不断增加，为信息处理技术带来不小的挑战。大数据中的信息往往蕴含着极高的潜在价值，通过深入的数据挖掘和分析，能够揭示出隐藏在数据背后的规律和趋势，为决策提供有力支持。在这一环境中，数据的真实性、完整性和安全性等问题也变得尤为突出，需要先进的信息质量管理和统计分析方法来加以解决。

2.2 大数据环境下的信息质量概述

大数据环境下的信息质量问题日益凸显，这主要是由于数据规模的急剧扩增和数据来源的多样性所致。信息质量的定义涉及多个层面，包括精确性、一致性、全面性和时效性，每个层面对信息的有效性和可用性都有深远影响。精确性要求数据真实且无误，一致性则确保不同来源的数据在逻辑上没有矛盾。全面性反映数据是否能够覆盖所研究问题的所有方面，而时效性则关注信息更新的及时性 [2]。

在大数据环境中，信息质量面临着前所未有的挑战。大数据不仅在数量上呈现几何级增长，数据结构也趋于复杂化，包括结构化和非结构化数据。传统数据处理方法已无法完全适应这种变化，特别是在处理不完善或缺乏标准化的数据时，质量问题尤为显著。大数据的实时性要求也提高了对信息时效性的要求，实时数据处理成为必然趋势。

2.3 大数据环境对信息质量的影响

大数据环境极大地增加了数据的体量和复杂性，使得信息质量面临前所未有的挑战。数据的多样性要求不同类型的数据处理方法，不一致的数据源可能导致数据不准确和时效性差。数据生成速度加快，信息存储和管理难度显著增加，对数据安全和隐私保护也提出了更高的要求。有效评估和优化信息质量成为关键。

3 信息质量的统计分析研究

3.1 信息质量评估的四个维度

在大数据环境下，信息质量评估依赖于四个主要维度：精确性、一致性、全面性和时效性。精确性指的是信息与真实值或标准值的接近程度，高精确性的信息能够准确反映真实情况。在大数据环境中，数据源多样且数据量庞大，如何保障数据的精确性是一大挑战。

一致性涉及数据在不同数据集或不见点之间的相互匹配和协调。如果信息在不同系统或时间段内不一致，将会影响决策的可靠性。大数据环境中的各种数据源和格式增加了一致性管理的复杂性。

全面性衡量的是数据是否覆盖了所需信息的所有方面，确保每个重要因素都能被捕捉和记录。在大数据背景下，各种数据形式和数据源使得信息的全面性评估工作变得困难 [3]。不同的数据源可能缺乏某些重要的信息，这对于全面性提出了更高要求。

时效性指信息被更新或收集的时间与实际事件的发生时间之间的间隔。在实时决策和分析中，时效性显得尤为重要。在大数据环境下，数据的生成和处理速度迅猛，及时更新和获取最新信息变得至关重要。

3.2 描述统计回归分析聚类分析等信息质量统计分析方法介绍

在信息质量的统计分析中，描述统计、回归分析和聚类分析是常用的方法。描述统计通过集中趋势和离散趋势指标，如均值、中位数、标准差等，直观展示信息的基本特征，有助于发现数据异常和趋势。回归分析则用于探讨变量之间的关系，通过建立数学模型，可以预测信息质量的变化，揭示数据内在关联。聚类分析通过将数据划分为若干组别，使得同组内数据具有较高的相似性，而不同组之间差异显著，用以发现信息质量问题的分布模式，帮助识别数据集中的不同质量水平。结合大数据环境的特点，这些传统的统计分析方法依然发挥着重要作用，能够在信息质量评估和改进中提供有力支持。

3.3 对大数据环境下的新型统计方法进行潜在性能评估

大数据环境下，新型统计方法在解决信息质量问题方面展现出显著的潜在性能。这些方法不仅能够处理大量数据，还能应对数据的复杂性和多样性。机器学习算法在大数据环境中得到广泛应用，如决策树、随机森林、支持向量机等，这些算法可以通过训练模型来有效地分类和预测数据。这些方法能够显著提高数据的精确性和一致性，有助于提升信息质量。基于深度学习的方法，如卷积神经网络（CNN）和循环神经网络（RNN），在处理图像和文本数据方面具有独特优势，能够更有效地解决非结构化数据的信息质量问题。

另一类新型统计方法是大数据分析中的流数据处理。由于数据产生和传输速度的快速增长，实时信息处理变得非常重要。流数据处理技术如 Apache Kafka 和 Apache Flink 提供了强大的实时数据处理能力，确保信息的时效性和准确性。分布式计算技术如 Hadoop 和 Spark 在大数据处理中的应用也大幅提升了数据的处理速度和效率。

总的来看，新型统计方法与传统方法相比，在大数据

环境中展现出了更强的适应性和性能。通过这些方法，能够更好地解决大数据环境下的信息质量问题，提高数据的精确性、一致性、全面性和时效性。这为未来改善信息质量提供了广阔的前景。

4 大数据环境的信息质量挑战和解决策略

4.1 结构化和非结构化数据的特点及其对应的处理方法

在大数据环境下，信息质量面临着结构化数据和非结构化数据的双重挑战。结构化数据通常指的是存储在关系数据库中的数据，这类数据具有预定义的格式和固定的字段，便于进行整理和分析。对于结构化数据，常用的质量评估方法包括描述统计和回归分析，这些方法能有效检测数据中的异常值、空值和一致性问题。例如，通过描述统计可以分析数据的集中趋势和离散程度，而回归分析则能帮助发现潜在的模式和关系，从而改进数据精确性和一致性。

相比之下，非结构化数据如文本、图像和视频则缺乏固定的格式和结构，使得信息质量的评估更加复杂。针对非结构化数据，传统的统计方法显得力不从心，需要采用更复杂的技术，如文本分析和语义分析。文本分析可以通过自然语言处理技术，提取关键词、主题和情感倾向，从而评估数据的全面性和准确性。语义分析则能够识别与理解数据的语义关系，有助于提高信息一致性和时效性。

总体而言，利用不同的处理方法能够有效应对结构化和非结构化数据中的信息质量问题。在大数据环境下，整合多种分析工具和技术，不仅能提高数据的精确性和一致性，还能为解决全面性和时效性问题提供坚实的基础。这种综合性的处理方法在未来应得到进一步推广与应用，以应对不断变化和复杂的数据环境。

4.2 大数据环境下的数据安全和隐私问题

在大数据环境下，数据安全和隐私问题对信息质量提出了新的挑战。大数据包含大量的个人信息和敏感数据，这些数据可能会因不当处理或恶意攻击而泄露或被滥用，从而严重影响信息质量。隐私泄露问题尤为突出，数据持有者和使用者必须遵循严格的隐私保护原则，例如数据匿名化和加密等技术，确保个人隐私不被侵犯。

数据安全问题不仅仅涉及数据存储和传输过程中的加密和访问控制，还包括数据清洗和整合过程中的安全性。大数据处理过程中的安全漏洞可能导致数据篡改、丢失等情况，从而影响数据的精确性、一致性和时效性。为确保信息质量，数据处理环节的安全审计与监控措施至关重要。

解决这些安全与隐私问题需要综合采用技术手段和管理措施。技术方面，可以采用差分隐私、同态加密等前沿技术提升数据隐私保护水平。在管理方面，制定并严格执行数据安全与隐私保护政策，明确责任主体和法律责任，增强数据使用者的安全意识。只有在安全和隐私问题得到充分处理的前提下，高质量的信息才能在大数据环境中得以实现。

4.3 借助人工智能技术改善信息质量的前沿研究和可能性分析

借助人工智能技术可显著提升信息质量。深度学习和自然语言处理技术在非结构化数据的精确性和可靠性提升方面发挥关键作用。机器学习算法能在数据处理过程中自动纠正和填补数据缺失，提高数据一致性。图卷积网络等新技术为处理复杂数据关系提供了新的手段。通过智能化的数据预处理和修正，整体数据的全面性和时效性亦显著增强。利用人工智能技术可以更高效地监测和应对数据安全和隐私问题。

5 结语

论文全面梳理并详细总结了当前大数据环境下信息质量的统计分析研究进展。本研究从信息的精确性、一致性、全面性和时效性四个维度对信息质量进行了评估，并对主流的信息质量统计分析方法进行了综述，如描述统计、回归分析、聚类分析等。其中，对于结构化和非结构化数据的处理方法也进行了具体探讨，并指出大数据环境下也存在如数据安全和隐私问题等新的挑战。然而，值得一提的是，尽管我们已经取得了一些令人瞩目的成果，但是在大数据环境下，信息质量统计分析方法和技术的研究仍然面临着许多挑战。借由这份研究，我们期望能够激发更多的研究员对此领域产生兴趣，投入更多的时间和精力来研究和探讨在大数据环境下提高信息质量、解决大数据特有的新问题的相关方法。今后，我们应该进一步关注大数据环境下信息质量问题的研究，探索更加创新和实用的统计分析方法和技术，对特定问题提供更有针对性、更有效的解决方案，为大数据环境下的信息质量提供更加坚实和全面的保障。

参考文献

[1] 杨子佳.大数据审计下统计分析方法[J].信息周刊,2019(52):161.

[2] 张释月.大数据时代下劳动统计分析质量的研究[J].河北农机,2021(9):142-143.

[3] 张英辉,代海平.大数据环境下审计数据统计分析研究[J].会计师,2021(2):99-100.