DOI:10.12361/2661-3263-06-05-138621

大数据时代信息挖掘的价值风险及其规避方法

靳高强

对外经济贸易大学，中国·北京　100029

【摘　要】随着大数据时代的到来，信息挖掘成为了重要的技术手段，帮助我们从海量数据中提取知识和信息，以支持决策和创新。然而，信息挖掘也伴随着一些价值风险，包括隐私泄露、偏见和误解等。本文将探讨大数据时代信息挖掘的价值风险，并提供一些规避方法，以确保信息挖掘的合理和可靠性。

【关键词】大数据时代；信息挖掘;价值风险；规避方法；讨论与展望

The value risk of information mining and its avoidance method in the era of big data

Gaoqiang Jin

University of International Business and Economics, Beijing 100029

[Abstract] With the advent of the era of big data, information mining has become an important technical means, which helps us to extract knowledge and information from massive data to support decision-making and innovation. However, information mining is also accompanied by some value risks, including privacy leakage, prejudice and misunderstanding. This paper will explore the value risk of information mining in the era of big data, and provide some avoidance methods to ensure the rationality and reliability of information mining.

[Keywords] Big data era; information mining; value risk; avoidance method; discussion and outlook

1　引言

1.1　大数据时代与信息挖掘的背景介绍

随着互联网的普及和技术的进步，现如今我们生活在一个大数据时代。大数据指的是以极其庞大的数据规模、快速的数据流转和多样化的数据类型为特征的信息体系。在这样的时代下，信息挖掘成为了一种重要的处理和分析大数据的手段。信息挖掘可以帮助我们从海量的数据中提取有价值的知识和信息，发现隐含的模式和关联规律，用于支持决策和创新。如何让信息挖掘过程更加稳健可靠，成为了一个重要的问题。

1.2　研究目的和重要性

在大数据时代，信息挖掘具有广泛的应用前景，可以帮助企业、政府和个人从海量数据中发现商业机会、提高决策效率和推动社会进步[1]　。然而，信息挖掘也伴随着一些价值风险，例如隐私泄露、数据偏见和误解等。因此，本文将针对这些风险展开讨论，并提供相应的规避方法，旨在确保信息挖掘的合理和可靠性。

2　大数据时代信息挖掘的价值风险

2.1　隐私泄露风险

隐私泄露风险是当今社会面临的一个重要问题。在大数据时代，个人的敏感信息收集、存储和使用已经变得更加普遍。如果不加以适当的保护，个人隐私有可能被滥用，导致一系列问题。隐私泄露可能对个人造成直接的损失。个人身份信息的泄露可能导致身份盗窃、账号被盗等风险，给个人带来财务损失甚至信用危机。此外，偏好和消费行为的泄露也可能被商业机构利用，产生个人无法掌控的后果，如个性化广告骚扰、个人秘密泄漏等。另外，隐私泄露可能破坏个人和机构之间的信任。关键是在个人提供敏感数据的同时，需要相信机构能够安全地存储和使用这些数据。一旦发生数据泄露事件，用户的信任度将大幅下降，可能导致用户流失、声誉受损等问题，对企业可持续发展造成不良影响。更甚隐私泄露还可能引发法律纠

纷。许多国家都制定了相关的隐私保护法规，如欧洲的《通用数据保护条例》（GDPR）等。如果企业未能妥善处理个人数据，出现泄露事件，可能会受到法律制裁，面临巨额罚款和声誉损失。

2.2 数据偏见和误解风险

信息挖掘算法的设计和参数设置可能会导致数据偏见和误解的产生。算法本身可能存在对某些数据类型的偏好或忽视，进而影响了模型和结果的准确性。此外，由于大数据时代的数据多样化和复杂性，数据的质量和完整性也会对信息挖掘的可靠性产生影响。具体存在的风险为以下四个方面。

2.2.1 算法设计：算法设计中的某些偏好或忽视可能导致数据偏见和误解。例如，在分类算法中，如果算法对某个特定类别的样本进行识别的能力较弱，则会造成该类别被错误地分类或被忽视。

2.2.2 数据质量：数据质量对信息挖掘的结果产生重大影响。如果数据存在缺失、重复、错误或不一致等问题，那么在进行信息挖掘分析时就很容易产生误导性的结果。因此，在进行数据采集和预处理时，需要严格遵循数据质量管理的标准，尽量减少数据质量问题对结果的影响。

2.2.3 数据选择偏见：信息挖掘算法所使用的数据可能存在选择偏见，即有些数据被优先选择，而其他数据被忽略。这可能导致对某些数据类型或特定场景的误解[2]。

2.2.4 结果解释的困难：信息挖掘得到的结果可能很难解释，容易被误解。这是因为算法对于模型中的决策过程通常是隐式的，并且有时很难理解其中的逻辑。

2.3 信息安全风险

信息安全风险是指在信息挖掘过程中可能发生的威胁和问题，对个人、组织和社会造成潜在的大量安全风险。具体而言，以下是一些可能遇到的信息安全风险及其影响：

2.3.1 数据泄露：数据泄露是指未经授权的个人或组织获取、使用或公开敏感数据的情况。在信息挖掘中，存储、传输和处理大量的数据是必不可少的，而这些数据可能包含个人身份信息、财务数据、商业秘密等敏感信息。如果数据存储或传输过程中存在漏洞或不安全环节，黑客或内部人员可能会获取这些数据，并将其用于非法活动，例如身份盗窃、欺诈等。此外，数据泄露还可能导致个人隐私被侵犯和声誉受损，对个人和组织带来严重损失。

2.3.2 黑客入侵：黑客入侵是指未经授权的个人或组织通过非法手段进入系统或网络，并进行非法活动的行为。在信息挖掘中，数据库、服务器、云平台等关键设备可能成为黑客攻击的目标。黑客入侵可能导致数据被篡改、删除或窃取，严重时甚至可能对业务运营造成瘫痪。此外，黑客入侵还可能导致个人隐私被泄露，客户信誉受损，对个人和组织的声誉和财务状况产生不可逆转的影响。

2.3.3 恶意攻击：恶意攻击包括病毒、木马、钓鱼等行为，旨在破坏、盗取或篡改数据。在信息挖掘中，恶意软件可能通过电子邮件、下载链接、网络广告等方式传播，在用户不知情的情况下植入恶意代码，并对系统和数据进行破坏或盗取。这些攻击行为可能导致数据的完整性和可用性受损，影响信息挖掘的准确性和可靠性[3]。同时，恶意攻击还可能导致个人隐私被泄露，个人和组织面临财务损失、法律责任以及声誉风险。

3 大数据时代信息挖掘的规避方法

3.1 隐私保护与数据脱敏

隐私保护是在信息挖掘过程中至关重要的一环。为了保护个人隐私，数据脱敏技术被广泛应用。数据脱敏是通过合适的措施对原始数据进行处理，降低敏感性和可识别性，以确保数据在分析中仍能发挥作用，同时减少个人身份暴露的风险。我们可以采用以下数据脱敏技术：

3.1.1 去标志化（De-identification）：去标志化是指从原始数据中删除或替换个人身份信息，使得数据无法直接与特定个人相关联。这其中包括删除姓名、身份证号码、电话号码等直接识别个人的信息。去标志化后的数据仍然可以用于某些数据分析和研究，但无法将其追溯回个人身份。

3.1.2 数据聚合（Data Aggregation）：数据聚合是将原始数据进行合并和汇总，以减少个体数据的详细度。例如，将个人的年龄按照一定范围进行分组，将精确的出生日期替换为年代等。通过数据聚合，可以减少个人敏感信息的暴露程度，降低个人被识别的风险。

3.1.3 噪声添加（Noise Addition）：噪声添加是在原始数据中引入一定程度的随机性和干扰，以混淆数据中的个人特征。这样可以使得数据集中的个体更难以被识别出来。例如，在位置数据中添加一些随机的偏移量，或对数值数据进行微小的加减操作。噪声添加技术可以平衡数据的分析效果和个人隐私的保护程度[4]。

3.2 数据质量与样本选择的优化

数据质量管理和样本选择优化是确保数据分析的准确性和可信度的重要环节。以下是关于数据质量与样本选择的优化的详细信息：

3.2.1 数据质量管理：

数据质量是指数据的准确性、完整性、一致性和可靠性等方面的度量和保障。为了确保数据质量，需要采取以下措施：

- 数据收集：在数据收集阶段，应确定清晰的数据收集目标并明确数据需求。合理选择数据来源，优先选择可信度高、可靠性强的数据源。

- 数据清洗：在数据分析前，对原始数据进行清洗。这包括检查和修复数据中的错误、缺失值和异常值，并确保

数据的一致性和完整性。

- 数据验证：验证数据的准确性和可信度。可以通过与其他可靠数据源进行比较、进行逻辑验证、重复测量等方式来验证收集的数据。

- 数据文档化：建立数据文档，记录数据的来源、收集时间、采集方法、处理过程等信息，便于后续数据的追溯和审计。

3.2.2 样本选择优化：

样本选择是指从总体数据中选择一部分具有代表性的样本进行分析和推断。样本选择的优化需要考虑以下因素：

- 总体定义：明确研究的总体范围和目标。清楚定义总体有助于确定合适的样本选择策略。

- 样本容量：根据数据需求和分析目标确定所需的样本容量。通常，样本容量越大，可信度越高，但也要兼顾成本和时间因素。

- 随机抽样：采用随机抽样的方式选择样本，使得每个个体被选入样本的概率相同。这样可以减小样本的偏倚，并增加样本的代表性。

- 分层抽样：如果总体具有明显的分层特征，则可以采用分层抽样的方法。将总体划分为若干子群，然后从各子群中进行随机抽样，确保样本能够充分反映总体的分布特征。

- 重要性抽样：对于关注特定子群的问题，可以采用重要性抽样的方法。根据某些重要属性进行抽样，以提高对指定子群的统计推断能力。

- 样本比率调整：在某些情况下，样本选择后可能发现与总体的某些属性比例不一致。在此情况下，可以进行样本比率调整以修正这种偏差。

4 讨论与展望

4.1 价值风险规避的问题与挑战

4.1.1 个体隐私权利保护的平衡：在进行数据分析和挖掘的过程中，涉及到大量的个人信息，并可能会对个人隐私造成潜在风险。如何在保护隐私的同时充分利用数据的价值，是一个重要的问题和挑战。

4.1.2 数据质量管理的标准化：数据的质量对于分析结果的准确性和可靠性至关重要。然而，现实中数据质量参差不齐，包括数据缺失、噪声、不一致等问题。如何制定和执行数据质量管理的标准化方法，以提高数据的准确性和可信度，是一个需要解决的问题。

4.1.3 信息安全技术的创新：随着技术的不断发展，网络攻击和安全威胁也日益增加。现有的信息安全技术需要不断创新和升级，以适应日益复杂的安全威胁。同时，也需要解决安全技术与业务需求之间的平衡问题，确保安全防护与业务运行的有效结合。

4.2 未来发展方向与趋势

4.2.1 加强隐私保护技术研究：未来需要研究和开发更加隐私友好的数据分析和挖掘技术，例如差分隐私、同态加密等技术，以实现在不泄露具体个人信息的前提下，实现对数据的有效分析和利用。

4.2.2 推动数据质量管理的标准化：建立和推广数据质量管理的标准和方法，包括数据清洗、去重、增强和集成等方面的技术，以提高数据的准确性、完整性和一致性，保证分析结果的可信度。

4.2.3 强化信息安全管理和技术创新：加强信息安全管理的规范化和标准化，制定和执行安全策略和流程，同时推动信息安全技术的创新，研究和应用先进的防御和监测技术，如人工智能在安全领域的应用[5]。

4.2.4 关注伦理和社会责任：在进行数据分析和挖掘时，应注重个体权益的保护，尊重用户的隐私和选择权。同时，积极推动数据使用的透明化和公正性，避免数据滥用和不当行为的发生。

综上所述，未来的发展方向是在保护个体隐私权利的前提下，加强数据质量管理和信息安全技术的研究和创新，推动信息挖掘技术的发展与社会应用。

结论：大数据时代信息挖掘的价值风险不可忽视，但通过适当的规避方法，可以有效地降低这些风险。在保护隐私、减少偏见和误解以及保障信息安全等方面都有相应的规避方法可供选择。未来的研究需要更加注重个体隐私权保护、数据质量管理和信息安全等方面，以进一步推动信息挖掘技术的发展和应用。大数据时代带来了前所未有的机遇和挑战，我们应积极应对，并寻求解决方案，以推动社会的可持续发展。

参考文献：

[1]燕道成,高紫叶.大数据时代信息挖掘的价值风险及其规避[J].新媒体研究, 2020, 6(22):4.

[2]禹露.大数据时代科技风险的特征及其规避[J].环球人文地理,2014(1X):1.DOI:10.3969/j.issn.2095-0446.2014.02.230.

[3]程龙.大数据时代数据处理过程中的风险控制[J].科技传播,2019,11(10):2.DOI:CNKI:SUN:KJCB.0.2019-10-064.

[4] 刘磊.大数据分析的经济价值评价与过度挖掘风险研究[D].天津财经大学,2017.DOI:CNKI:CDMD:1.1018.033801.

[5]惠志斌.大数据时代国家信息安全风险及其对策研究[J].复旦国际关系评论,2015(2):9.DOI:CNKI:SUN:FDGJ.0.2015-02-006.

作者简介：

靳高强（1994.3—）男，汉族，籍贯山西省运城市，硕士研究生在读，对外经济贸易大学统计学院在职人员高级课程研修班学员，研究方向：大数据科学与应用方向。