计算机科学与技术在大数据挖掘中的应用与挑战研究

何江川
加州大学圣地亚哥分校，加利福尼亚 圣地亚哥 92093
DOI: 10.61369/TACS.2025060024

摘　　要： 在信息技术飞速发展的当今时代，数据呈现出爆炸式增长的态势，大数据时代已然来临。大数据挖掘作为从海量、复杂的数据中提取有价值信息和知识的关键技术，正受到广泛关注。计算机科学与技术作为大数据挖掘的核心支撑，在数据预处理、挖掘算法设计以及数据可视化等环节发挥着至关重要的作用。本文旨在深入探讨计算机科学与技术在大数据挖掘中的具体应用，分析当前面临的主要挑战，并针对这些挑战提出相应的应对策略，以期为大数据挖掘技术的进一步发展和应用提供理论参考，推动该领域在理论研究和实际应用方面的双重进步。

关 键 词： 计算机科学与技术；大数据挖掘；应用；挑战

Research on the Application and Challenges of Computer Science and Technology in Big Data Mining

He Jiangchuan
University of California, San Diego, San Diego, California 92093

Abstract: In the contemporary era of rapid advancements in information technology, there has been a considerable proliferation of data, which has led to the emergence of the era of big data. Big data mining has become a pivotal technology for the extraction of valuable information and knowledge from substantial and intricate data sets, thus garnering considerable attention from the research community. It is evident that computer science and technology, as the fundamental underpinnings of big data mining, assume a pivotal role in data preprocessing, the design of mining algorithms, and the realm of data visualisation. The purpose of this paper is to thoroughly examine the specific applications of computer science and technology in big data mining. In addition, it will analyse the main challenges currently being faced and propose corresponding countermeasures for these challenges. The final aim is to provide a theoretical reference for the further development and application of big data mining technology. Furthermore, it will promote the dual progress of this field in theoretical research and practical application.

Keywords: computer science and technology; data mining; application; challenge

一、大数据挖掘的基本概念

大数据挖掘是从海量、不完全、有噪声的数据中提取隐含价值信息的过程，目标是发现数据模式、关联及趋势，涉及计算机科学、统计学等多学科。其数据具备“5V”特征：规模达 PB 级以上的 Volume（量大）、实时处理需求的 Velocity（高速）、文本图像等多类型的 Variety（多样）、有用信息占比低的 Value（低价值密度）及准确性要求的 Veracity（真实性）[1]。

二、计算机科学与技术在大数据挖掘中的应用路径

（一）数据预处理

预先处理过的数据，可以提升数据质量。主要步骤包括：清洗、集成、变换以及削减等四部分。第一步，利用统计学的方法来清除数据中的异常值，并用均值对缺失值进行填补；第二步，通过数据集成来处理异构数据；第三步，进行数据变换。它主要包括了对数据进行标准处理（Z-scorestandardization）、规范化（将其置入区间 [0，1] 内）、离散化（如将年龄分成几个类别）和特征提取（如通过图像获取边缘特征）等；第四步，可通过采样、特征选择（如采用滤波器）以及主成分分析（PCA）等来降低数据维度，这样可使数据分析的过程变得更简单，进一步提升探索效率。

（二）挖掘算法

在各种算法中，决策树采用树状图进行表现规律；而朴素贝叶斯采用的是独立假设；支持向量机是找一个最佳的超平对象；神经网络则是生物体结构方式的一种拟真。大规模数据，比如分布式决策树和并行的 SVM，能够通过集群运算进行提升工作效率。聚类中 K- 均值依据最小误差平方和的方式对簇进行分割；层次聚类构建树状的结构方式；DBSCAN 则针对密度的方式去判定簇的形貌；如果将 PCA 降维和分布式 K- 均值法联系在一起，就能够对高维度的进行有效的处理。最后，关于关联规则挖掘这一板块，Apriori 则是采用频繁项目集合的方式不断去探索；FP-

growth 则是用创建 FP 树的方式进一步提升；另外，这种模式也早已应用在包括文本、图等多种形式的数据中。

（三）数据可视化

可视化的主要形式是采用各类视觉通道（如折线图、柱状图的图表形式，地图、网络图形式，放大 / 筛选功能等）呈现信息的结果，通过如 Tableau、PowerBI 等软件的使用可使其更易上手，基于虚拟现实（VR）/ 增强现实（AR）技术的应用则有希望实现更好的用户体验，如对数据分布情况的分析研究开展于三维环境之中[2]。

三、计算机科学与技术应用下大数据挖掘面临的挑战

（一）数据隐私保护

大规模数据挖掘时对数据私密性的保障是大数据挖掘的一大难点，随着数据收集和分享的范围不断扩大，含有个人生活信息、企业内部商业信息、国家保密信息等信息的数据风险也水涨船高，而泄露个人信息，则可能引发隐私权或者隐私权利的侵犯的问题。传统保密方法诸如访问控制以及数据加密等手段能够在一定程度上保障数据隐私性，但在大数据环境下遭遇了大量困难。一方面由于大规模分散式存储和运算使得数据在各个节点之间的间接传送、共享增多，因而数据的泄露概率增加；另一方面是大数据挖掘技术从大量的噪声数据中提取敏感数据，如果原始数据采用的是匿名化处理，依然可能被使用关联分析等方式再次辨认个人身份。因此，如何既能最大限度地发挥大数据挖掘的作用又能保证数据隐私性的挑战，成为大数据挖掘的一大痛点。

（二）数据质量

数据质量是大数据挖掘成功的非常重要的因素。高质量的数据为获得有用的资料提供了基础，而如果质量低，会造成分析的不准确或者可信度不足，甚至会得出偏颇结论[3]。在一个大数据背景下，这个问题更为突出，主要有重要属性和记录缺失、数据与事实不符、同一数据在不同来源中呈现的形式或值不同，以及同一个数据被多次录入等情况。出现问题的原因主要有采集阶段的设备损坏或人工操作失误、传递过程中的网速延时或信息丢失和储存过程的数据库设计缺陷或资料陈旧等问题。此外，大数据由于其多样性及复杂性使数据品质管理难度加大，各种类型资料的质素性质及评价指标各不相同，要求有不同的数据品质管理策略。

（三）计算资源

由于数据具有海量和多样性等特点，对计算机硬件资源提出了非常苛刻的要求。传统单机式的计算模式不能满足海量数据的计算需求，如运算速度缓慢、存储容量低、网络流量小等[4]。具体表现在：一是运算能力不足，大数据分析过程中会进行大量的计算，如矩阵计算、统计分析、机器学习训练等，这些都需要 CPU 和 GPU 强大的硬件能力支持；二是存储空间不足，大数据中有很多规模的数据，现有的机械硬盘式存储系统容量小且读取数据的速度有限，均不能满足大数据空间的需求；三是网络传输速度低，在分布式计算中，数据需要在各个节点之间进行传输、共享，如果网络传输速度低，则会造成数据传输时间长，降低了整体的计算效率。除此之外，计算成本也是其问题之一。为了适应大数据分析对于计算资源的要求，我们不得不花巨额成本，购买并支持高效服务器、存储系统及网络设备等，对一些中小企业来说成本大。因此，如何在有限的计算机计算资源条件下高效、快速实现大数据挖掘，成为大数据挖掘亟待解决的关键问题。

四、应对挑战的策略

（一）加强隐私保护技术

隐私安全问题需要不断加强隐私保护技术和应用的研究力度。新出现的隐私保护措施，例如差分隐私（differentialprivacy）、同态加密（homomorphicencryption）、联邦学习（federtatedlearning）等都为我们在大数据背景下保护个人隐私提供了新的思路。差分隐私采用在数据中加入少量噪声使得除了已经掌握某一组信息的情况下其他所有的信息都不能明确得知这一组信息是否出现的方式来保护个人的隐私信息，这种方法是在严密的数学原理指导下保护隐私的一种方法，在保护隐私的同时不丢数据[5]。同态加密是一种特殊的加密算法，它能允许我们在经过加密的数据上进行计算，计算后所得的解在解密之后与解密之前在未加密状态下的原数据上进行操作所得出的结果相同，这样我们可以通过加密的数据来研究和分析数据而无需公布原始数据，从而保证数据安全。联邦学习是一种分布式机器学习模式，多个合作伙伴不需要提供原始数据，在彼此不交换原始信息情况下仅通过交换模型参数或其他中间结果的方式共同构建出一个全球分布的模型。这种模式可以在隐私保护下充分利用各参与方的数据优势，提升模型效果。还需健全数据隐私保护法律法规和行业制度规范，加强对采集数据和存储、加工、共享的数据流程管理，厘清数据提供者和使用者的权利和义务，从法律制度层面保障数据隐私安全[6]。

（二）提高数据质量

我们需要从多个维度——数据的采集、管理、存储、利用等方面，建立完整的数据质量控制体系。第一，从数据采集环节入手，应做好数据采集的规则和流程设置工作，选择合适的工具和技术手段获取准确完整有效的信息[7]。例如在形成数据表的时候，通过适当的设置所需要填写的项目以及项目的数据格式校验规则等，避免人为误差；第二，如果是通过传感器获取的信息，应定期校准、保养传感器，确保数据采集的准确性。第三，在数据处理上要做好数据整理、合并、转换等工作，通过高科技手段的数据质检软件和算法，快速发现消除数据中的噪声、缺失字段、矛盾数据、重复字段等。如能通过数据质量评价工具给出数据评分，并根据评分判断数据质量的好坏，之后通过相应手段对数据进行整改；第四，可通过使用数据融合工具解决不同数据源间数据差异的问题，确保数据的一致性与完整性。最后，在数据存储方面要做好数据库结构设计、数据索引设计、数据备份设计，以确保数据的安全可访问[8]。例如，采用关系数据库存储结构化数据，采用 NoSQL 数据库存储非结构化或部分结构化数据，并按照

信息访问频率和重要性进行数据分区和缓冲；定时执行数据库的备份和恢复测试以防数据丢失或损坏。在数据使用过程中，建立数据质量反馈制度，及时接受用户反馈的数据质量建议，不断地提高数据质量。此外，还可用数据质量培训和教育，提高数据管理人员和使用者对数据质量的意识，形成整个企业齐心协力关注数据质量管理的良好氛围。

（三）优化计算资源

解决计算资源的问题，须利用分布式计算、并行计算与云计算等高科技手段优化和合理利用计算资源[9]。比如，Hadoop 和 Spark 等分布式计算平台会将海量数据分发至不同计算节点，以便并行计算处理，尽最大可能利用全集群的运算能力和存储资源，进而强化大数据分析的质量。Hadoop 在采用分布式的文件系统（HDFS）和 MapReduce 的计算模式的基础上，实现了数据的分布式存储以及分布式运算；而 Spark 是在 Hadoop 的基础上进行二次优化改进，形成了更为有效的内存计算和更丰富的分布式计算框架，同时支持的计算模型也较为丰富，包括大规模批处理、实时处理以及机器学习等。除此之外，云计算技术的应用也使得大数据的分析具备了更加弹性灵活的计算和存储资源，客户可以按照自身的真实业务需要对计算资源进行弹性需求控制，进而避免过多的硬件开支与运营成本的投入。通过阿里云、腾讯云、AWS 等云计算云服务平台，客户可以快速建立大数据分析环境并且可根据自身的需求进行伸缩计算资源。其次，借助边缘计算，我们将一部分计算任务置于与原始数据更靠近的地方处理，降低了数据在网络之间的传输量，减少了对中心服务器的计算压力和提高了数据处理的时效性和效率。例如，将物联网中的数据处理的一些预处理和基础分析等工作下放至感知终端或者是边缘路由上，那么将大大降低向云端上传的信息量、减少网络延时和计算成本[10]。

五、结论

综上所述，计算机科学与技术在大数据挖掘中具有至关重要的作用，涵盖了数据预处理、挖掘算法设计和数据可视化等多个关键环节。随着信息技术的不断发展，大数据挖掘技术将面临更多新的机遇和挑战。未来，需要进一步结合人工智能、区块链、物联网等新兴技术，不断创新大数据挖掘方法和技术，提高大数据挖掘的智能化水平和应用范围。同时，还需要加强跨学科研究和合作，整合不同领域的理论和方法，共同推动大数据挖掘技术的发展，为解决实际问题提供更有效的解决方案。

参考文献

[1] 高瑞，杨洋 . 数据挖掘技术在计算机网络病毒防御中的实践应用 [J]. 中国信息界 ,2024,(09):176-178.

[2] 杨军 . 试论大数据信息时代计算机科学技术的应用 [J]. 软件 ,2024,45(10):166-168.

[3] 刘翠芳，徐舒 . 计算机数据挖掘技术在互联网中的应用 [J]. 软件 ,2024,45(08):69-71.

[4] 何钊 . 数据挖掘技术在计算机软件工程中的应用研究 [J]. 中国信息界 ,2024,(03):133-135.

[5] 马延龙 . 智能技术与物联网大数据的挖掘算法分析 [J]. 集成电路应用 ,2024,41(06):160-161.

[6] 杨博宁 . 计算机科学与技术助力金融大数据分析与风险预测 [A] 全国绿色数智电力设备技术创新成果展示会论文集（七）[C]. 中国电力设备管理协会，中国电力设备管理协会 ,2024:3.

[7] 陈敬予 . 大数据背景下计算机科学与技术运用研究 [J]. 数字通信世界 ,2024,(05):15-17.

[8] 李斌 . 基于数据库技术的计算机数据挖掘平台设计 [J]. 信息与电脑（理论版）,2024,36(09):96-99.

[9] 张晓，李军丹 . 计算机数据挖掘技术的开发及应用探究 [J]. 软件 ,2024,45(02):146-148.

[10] 徐琴；刘智珺 .Python 数据分析与挖掘 [M]. 华中科技大学出版社 :202401.377.

