

基于云计算的数据挖掘技术分析

王荣平

广州华立学院 511325

摘要: 随着互联网、大数据技术的飞速发展,人们如何从海量的信息中挖掘获取有用的信息已成为当前计算机人员需要解决的难题。传统的数据挖掘技术虽能为人们解决数据处理、开发利用等,但往往效率低下。而云计算平台很好地解决了这一难题,云计算具有处理大规模数据计算和处理,将传统的数据挖掘算法与云计算平台紧密联系起来,将最大化发挥云计算技术功能,有利于各行业量化分析有序进行。

关键词: 大数据;云计算;数据挖掘

随着现代技术的发展,数据量不断增加,数据种类和结构也日益多样化,传统的数据处理方法已经无法满足对数据的挖掘和分析需求。应用大数据挖掘技术可以帮助人们从海量数据中发现隐藏的模式和规律,预测客户需求、市场趋势和企业内部运营的方向等,可以为企业的战略决策和业务创新提供高效的决策支持,有助于提高企业竞争力。由于云计算技术具有高可用性、高扩展性、灵活性和经济性等特点,有助于企业和组织降低成本、提高效率、加速创新,从而获得竞争优势。而数据挖掘技术则可以帮助企业和组织发现数据中的有价值的信息,进行预测、分类、聚类、关联、异常检测等分析,提供决策支持,探索商业机会。因此,现在越来越多的企业和组织开始关注并利用云计算和大数据技术进行数据挖掘和分析。在未来的发展中,基于云计算技术的数据挖掘分析将具有广阔的运用前景。

1. 云计算和数据挖掘技术概述

云计算是一种基于互联网的计算模式,通过计算机技术及相关程序,实现资源虚拟化,将用户收集的信息进行隐蔽计算,及时把结果反馈给用户。云计算的数据存储方式采用分布式存储为主,对服务器规模和安全性和可靠性有较高要求具有服务器扩充自由、高可靠性、供需相适应,实现资源共享等特点。云计算服务主要有三种模式即基础设施即服务(IaaS)、平台即服务(PaaS)、软件即服务(SaaS)。

数据挖掘是指通过建立复杂计算模型和算法,从大规模数据中发现隐藏的、有价值的信息和知识。数据挖掘技术与计算机技术、统计技术、在线分析处理技术等紧密联系,为数据存储、共享、运用、预测等提供了保障。它的特点是:数据的多样性和复杂性、挖掘高效性、知识发现能力强、决策支持性较好^[1]。数据挖掘的主要任务包括分类、聚类、关联规则挖掘、异常检测、预测分析等。主要步骤包括:1、明确问题,理解业务需求;2、确保数据的质量和可用性,完成数据清洗;3、数据集成和选择,构建适合挖掘的数据集;4、数据变换和表示,便于数据挖掘;5、选择合适的模型及算法,建立预测和分析模型;6、模型评估和验证,确定它们的准

确性和预测能力;7、模型部署和应用。

2. 云计算在数据挖掘技术的优势

基于云计算的数据挖掘技术具有以下优势:一、便利性,用户可以随时随地无需事先购买硬件设备、进行软件安装等直接使用云服务。不仅提高了工作效率,而且操作十分简单。二、灵活性,云计算可以根据用户需求快速扩展和收缩计算和存储资源。三、成本效益,用户只需按实际使用量支付费用,避免了传统IT模式中需要购买及维护硬件设备的高成本。四、高可用性,云计算提供了高度的可用性及冗余备份,保证了数据的可靠性^[2]。

3. 云计算在数据挖掘技术的应用

3.1 算法的应用

在云计算环境下,数据种类各种各样,形式千变万化,数据规模呈几何倍数增长。如果希望按照用户的需求,将这些数据分析处理,挖掘他们之间的关联,算法是最为重要的工具和技术。在云计算时代,传统的数据挖掘如聚类分析、分类分析、回归分析等,难以精确估计数据量的大小,通过小规模分布式集群来实现大批量数据处理是达不到预期目标,随着运算部署难度不断扩大,成本投放也将不断增长。因此,基于云计算的数据挖掘技术,将充分发挥效率。简单来说,就是在Hadoop集群上构建开源并行的数据挖掘平台,通过Map Reduce计算框架实现对数据的处理工作。

数据挖掘技术涉及各种算法,不同的算法应用于不同的环境和范围,效果也将不尽相同。用户从实际需求及数据本身考虑选择合适的算法或算法转化,利用云计算技术优化算法提高数据挖掘水平。

基于云计算的数据挖掘技术,一般都和其他类型算法结合使用。如深度学习算法既能处理大批量数据,又可以通过样本集训练实现自动数据处理和挖掘。对于超多数据的处理和分析,可以将人工神经网络系统与数据挖掘技术结合起来。遗传算法通过可视化技术,给用户动画、影像等更形象化的体验,有利于生产更好地推广,它的应用也是十分广泛的。

3.2 用户数据处理

根据自己的需求,用户合理安排数据的来源,有时还需要根据实际购买新的需求数据。用户在获得所需数据之后,对数据进行有效地管理往往采用云计算的 DaaS 服务模式。用户通过数据挖掘系统能够迅速、准确地找到所需处理的数据,为接下来数据处理打下基础。另外用户可以将个人的数据共享,方便他人使用和处理。不同的用户只得到自己的数据结论而不相互干扰。此外,为了实现对用户数据的动态管理,可借助数据挖掘的分类聚类功能,科学划分用户的数据。对数据处理不改变数据的属性,这样数据可以实现重复利用。

3.3 Map Reduce 模型

Map Reduce 是一种常用的云计算数据挖掘并行计算模型,常用于商业、电商、科学领域。MapReduce 模型包含两个主要的计算阶段: map 和 reduce。Map 阶段将数据映射到一个或多个 key-value 对中,然后将其传递到 reduce 阶段,reduce 阶段对这些 key-value 对进行归约计算,生成一个或多个输出结果。整个过程包括分配任务、数据划分、节点间通信等多个环节。MapReduce 模型的优势在于可以方便地处理大规模数据,并充分发挥分布式计算的并行处理优势。此外,它还可以通过跨多台服务器、数据与计算分离等技术来提高可用性和容错性。

4. 云计算的数据挖掘分析平台设计

4.1 平台的架构和功能

基于云计算的数据挖掘分析平台的架构可以包括以下组件^[1]:

数据采集层:负责从不同的数据源(例如数据库、网络、文件等)中采集数据,并进行预处理和转换。

数据存储层:将采集到的数据进行存储,并支持快速检索和查询。可以使用云存储服务(例如 AWS S3、Google Cloud Storage)或分布式数据库(例如 Spark SQL、Hadoop HDFS)实现。

数据处理层:执行各种数据挖掘分析任务,包括预测分析、聚类分析、关联规则挖掘等。可以使用云计算平台(例如 AWS EC2、Azure VM)或者分布式计算框架(例如 Hadoop、Spark)进行计算和分析。

数据可视化层:将分析结果以图表、报告等形式呈现给用户,可以使用各种可视化工具(例如 Tableau、Power BI)或自定义可视化组件实现。

4.2 平台的实现技术和工具

平台的实现技术和工具需要考虑云基础设施方案的选择、数据仓库的设计、数据挖掘算法的实现、分析工具和可视化组件的细节等方面。一些常见的技术和工具包括:PaaS(平台即服务)提供商如 Azure、AWS、Google Cloud 等,Hadoop 生态圈、Spark 计算引擎、Python 编程语言、TensorFlow、Apache Zeppelin 等。或者使用 Python 语言编写爬虫程序进行数据采集,并使用 MySQL 等数据库进行数据存储和管理;使用 python numpy、pandas 等库进行数据清洗和预

处理;使用 scikit-learn、Keras 等库进行机器学习算法的调用和训练;使用 matplotlib、seaborn 等库进行可视化展示^[2]。

4.3 平台的实验结果和性能评估

平台的实验结果和性能评估,需要进行具体的实验和性能评估,以证明平台的有效性和实用性。这可能包括基准测试、数据集的选择和处理、关注如何提升平台的性能和处理速度等。同时,考虑到云计算常常需要清晰的成本评估和引入外部资源等方面,还需要花费更多的精力来实施测试和评估。

5. 云计算应用在数据挖掘技术中存在的问题

当前云计算处于发展起步阶段,各方面发展还不是非常完善,安全性、可靠性等还有待进一步提高。因此,当前云计算仍然存在以下挑战:第一、安全问题,云计算服务商需要保证用户数据的安全,避免数据泄露、滥用等问题。第二、性能问题,由于云计算需要通过网络提供服务,因此网络带宽、延迟等因素会对服务的性能产生影响。标准问题:云计算领域内缺乏统一的标准和规范,可能会导致不同服务商之间的互操作性问题。第三,算法问题,面对不同的问题,需要选择合适的算法和策略,建立相应的模型,完成数据处理工作。第四,不稳定性,云计算服务商可能因为系统故障、维护等原因导致服务不稳定。因此,在使用云计算服务时需要全面考虑这些优势和挑战,确保数据安全、性能满足要求,并避免不必要的风险。

6 结语

综上所述,高效处理海量数据已成为当今各个行业亟待解决的难题,基于云计算技术的数据挖掘分析正是应对这一需求的有效解决方案。不断发展的数据挖掘技术积极推动了经济、社会、生活等方面发展,而通过基于云计算的数据挖掘技术优势明显,正呈现蓬勃发展趋势。将云计算与数据挖掘技术完美结合起来,解决云计算在数据挖掘实践中存在的问题,将对数据挖掘技术不断发展起到推动作用,给人们和社会发展带来现实意义。

参考文献:

- [1]白萍.云计算与物联网技术结合的数据挖掘分析[J].互联网周刊, 2023 (03): 84-86.
- [2]文新懿.对基于云计算技术的数据挖掘分析[J].中国新通信, 2021, 23 (21): 33-34.
- [3]胡素娟.基于云计算技术的分布式网络海量数据处理系统设计 [J]. 科技与创新, 2021 (17): 84-85.DOI : 10.15913/j.cnki.kjyx.2021.17.032.
- [4]黄敏.基于云计算技术视角的大数据挖掘技术分析[J].数字技术与应用, 2019, 37 (11): 96+98.DOI : 10.19695/j.cnki.cn12-1369.2019.11.54.
- [5]李慧.关于云计算环境下数据挖掘技术分析[J].电脑迷, 2018 (08): 38.