基于云计算的海量数据挖掘研究

贺 瑶[1],王文庆[2],薛 飞[1]
(1. 西安邮电大学 管理工程学院,陕西 西安 710061;
2. 西安邮电大学 自动化学院,陕西 西安 710061)

摘 要:为了实现高效率低成本的海量数据挖掘,为企业决策提供参考,提出了基于云计算的海量数据挖掘模型。该模型中海量数据的处理和存储都是在云计算环境中进行的,首先对海量的数据进行一定的预处理,形成结构一致的数据后,应用云计算平台上的 MapReduce 模型进行高效的并行数据处理,最后得到所需的数据挖掘结果。基于云计算的海量数据挖掘的效率明显高于传统的数据挖掘,并且数据挖掘结果的准确性有了一定的提高,而且随着数据量的增多,该模型的优势会愈发明显。

关键词:云计算;数据挖掘;海量数据;MapReduce;数据预处理
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)02-0069-04
doi:10.3969/j.issn.1673-629X.2013.02.017

Study of Massive Data Mining Based on Cloud Computing

HE Yao[1], WANG Wen-qing[2], XUE Fei[1]
(1. School of Management Engineering, Xi'an University of Post & Telecommunication, Xi'an 710061, China;
2. School of Automation, Xi'an University of Post & Telecommunication, Xi'an 710061, China)

Abstract: In order to achieve high efficiency and low cost of massive data mining, and provide decision references for enterprise, the model of massive data mining based on cloud computing has been proposed. The massive data's processing and storage of the model were carried on the cloud computing environment. Firstly, take some certain preprocessing for the massive data to form data with the same structure. Then, use the MapReduce model on the cloud computing platform to parallelly process the data efficiently. Finally, get the needed result of data mining. The efficiency of massive data mining based on cloud computing is clearly higher than traditional data mining. Meanwhile, the accuracy of data mining will be improved. Along with the increase of data, the advantage of the model will increasingly obvious.

Key words: cloud computing; data mining; massive data; MapReduce; data preprocessing

0 引 言

近年来,随着信息技术的高速发展,如今每 18 个月产生的数据量大约等于过去几千年产生的总和,并且有不断增加的趋势。如此多的数据无疑能为人们带来广阔的信息量,但需要从海量数据中发现对企业或个人有用知识的难度随之增加。而云计算平台能够进行动态资源调度和分配、具有高度虚拟化和高可用性等特点,正好能满足高效数据挖掘的需求[1,2]。将云计算技术与现有的数据挖掘技术进行有效结合不失为一种可行的途径。文献[2]就是在 Google App Engine 平台提出了一个并行数据挖掘模型,并通过实验得出基于云计算平台的数据挖掘系统执行的效率要比单机系统高,随着数据量的增大,效率优势越明显。文中提出的基于云计算的数据挖掘模型,就是要为不同企业提供高效便捷的数据挖掘服务,用来提高企业生产工作效率及降低成本。

1 云计算

1.1 云计算的定义

云计算[3]从发起至今,还没有一个统一的定义。维基百科中对云计算的定义如下:云计算是一种能够通过互联网为用户提供服务的计算模式,它提供的主要是能够进行动态伸缩的虚拟化了的资源,用户不需要了解如何管理那些支持云计算的基础设施。简单来说,云计算是一种新颖的商业模式,它使用大量廉价

收稿日期:2012-06-24;**修回日期**:2012-09-26
基金项目:国家自然科学基金资助项目(61100165/F020508);陕西省自然科学基金(2007F18)
作者简介:贺 瑶(1987-),女,陕西榆林人,硕士,研究方向为信息管理;王文庆,教授,研究方向为智能信息处理、信息系统分析、复杂系统结构分析与鲁棒控制。

的、相互连接在互联网上的计算机进行任务的处理,为各种应用系统提供所需要的存储资源、计算资源和其他服务资源等[4]。

从技术层面来说,云计算技术早已存在,它是虚拟化技术的扩展、分布式计算技术的演进、SOA 架构的延伸、信息资源的集中管理和智能调配机制的体现。与传统 IT 技术有所区别的是,云计算带来了理念创新。从商业角度来看,云计算的核心理念是以服务的形式提供计算资源,用户在需要时进行使用和购买,可以更好地满足组织业务快速变更和创新升级的需求。

云计算主要有三种主流的商业模式,分别是平台即服务(PaaS)、基础架构即服务(IaaS)和软件即服务(SaaS)。其基本功能逻辑图如图 1 所示。

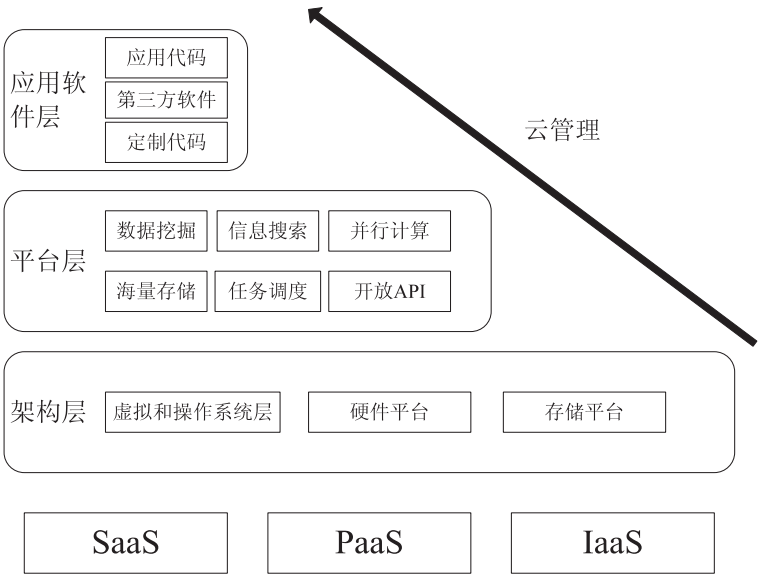

图 1 云计算主流商业模式逻辑图

1.2 云计算环境下的并行计算模型

MapReduce[5,6] 是 Google 实验室提出的一个分布式并行编程模型或框架,主要用于大规模数据的并行处理。一个 MapReduce 作业由大量 Map 和 Reduce 任务组成,它将大规模数据处理作业拆分成若干个可独立运行的 Map 任务,分配到不同的机器上去执行,生成某种格式的中间文件,再由若干个 Reduce 任务合并这些中间文件获得最后的输出文件。MapReduce 与分布式文件系统 GFS、分布式的锁机制 Chubby、大规模分布式数据库 Big Table 和集群管理 Borg 一起被人们认为是 Google 在云计算领域的五大技术精髓。有了以上的技术,云计算才可以提供高效的海量数据处理和分析平台。具体来说,将互联网中海量的数据分解成大小相同的数据块,并且分布式地存储在云计算网络中的各个服务器上,接下来的数据处理使用到了 MapReduce 并行计算模型,这种技术可以说是 Google 公司在搜索引擎应用中获得极大成功的至关重要的法宝。然而海量数据要想经过 MapReduce 计算模型进行并行化计算,必须要求结构一致,并且计算要简单。对于像数据挖掘任务这种大量的数据密集型应用,往往需要牵扯到近似求解、程序迭代、数据降维等比较复杂的算法,真正计算起来是非常困难的。因此,基于云计算的海量数据挖掘技术受到了学术界和工业界的共同关心,并且成为了现今的热点技术中的一员。

2 基于云计算的海量数据挖掘

2.1 数据挖掘

数据挖掘(Data Mining),也称为数据库中的知识发现过程,就是从海量数据中发现新颖的、有效的或者可能有潜在作用的、最终可被理解的模式的过程。对于企业来说,最终的目的是从海量数据中提取出可理解的知识,并且希望数据规模越大越好,这样挖掘出的知识才更加准确。这么高要求的数据挖掘对开发环境和应用环境有比较高的要求。在这种情况下,基于云计算的方式是比较适用的。云计算平台中数据中心可以存储海量数据,并可以根据数据挖掘应用的需求对资源进行动态分配,保证数据挖掘算法的可扩展性,并采用容错机制来保证数据挖掘应用的可靠性。

2.2 云计算数据挖掘服务的优势

1)基于云计算的模式可以进行分布式并行数据挖掘,实现高效实时的挖掘。同时可以适应规模不同的组织,为中小企业带来新型低成本计算环境,大企业云计算平台对某些特定数据的计算对大型高性能机的依赖性会得到减轻。

2)基于云计算的数据挖掘开发方便,底层被屏蔽掉了。对于用户来说,无需考虑数据的划分、数据分配加载到节点以及计算任务调度等。

3)在并行化条件下利用原先的设备,可以在很大程度上提高大规模处理数据能力。在增加结点方面也比较自由和方便,同时容错性得到了提高。

4)基于云计算的数据挖掘保证了挖掘技术的共享,降低了数据挖掘应用的门槛,使海量数据挖掘需求得到了满足。

2.3 基于云计算的海量数据挖掘模型

基于云计算的海量数据挖掘服务的主要目标是利用云计算的并行处理和海量存储能力,解决数据挖掘面临的海量数据处理问题。图 2 给出基于云计算的海量数据挖掘模型的层次结构图。

基于云计算的海量数据挖掘模型[7]大体上可以分为三层。位于最底层的是云计算服务层,提供分布式

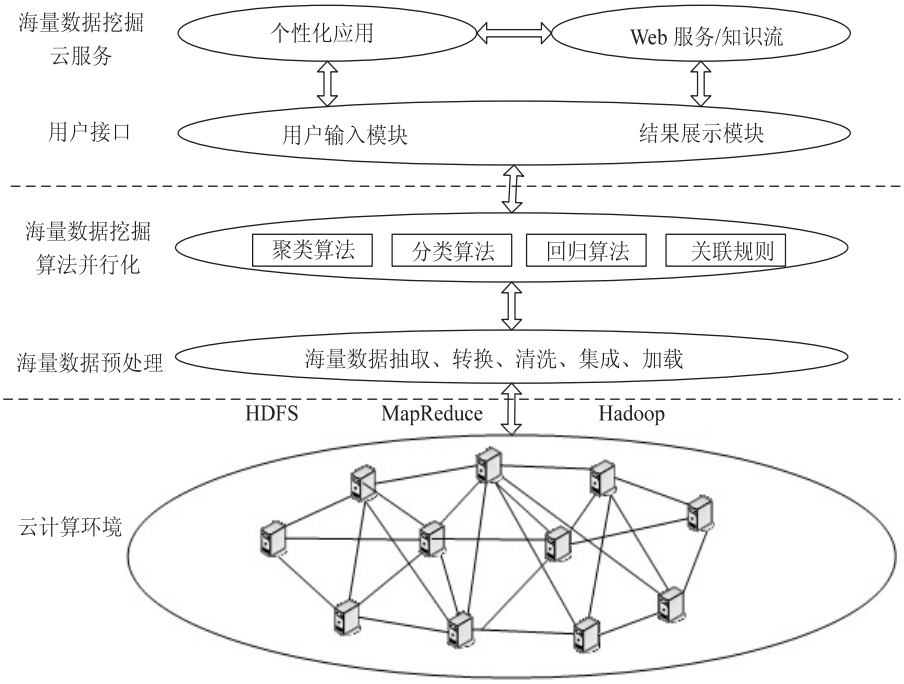

图 2　基于云计算的海量数据挖掘模型的层次结构图

并行数据处理及数据的海量存储。云计算环境中对海量数据的存储既要考虑数据的高可用性,又要保证其安全性。云计算采用分布式方式对数据进行存储,为数据保存多份副本的冗余存储方式保证了当数据发生灾难时不影响用户的正常使用。目前常见的云计算数据存储技术有非开源的 GFS(Google File System)和开源的 HDFS(Hadoop Distributed File System)[8,9],其中 GFS 是由 Google 开发的,HDFS 是由 Hadoop 团队开发的。此外,云计算使用并行工作模式,能够在大量用户同时提出请求时,迅速给予回应并提供服务。

位于云计算服务层之上的是数据挖掘处理层,这一层又包括海量数据预处理和海量数据挖掘算法并行化。海量数据预处理主要是对海量不规则数据事先进行处理。没有好的数据就没有好的数据挖掘结果。由于云计算环境下的 MapReduce 计算模型适用于结构一致的海量数据,因此,面对形态各异的海量数据,首先就要对它们进行预处理。数据预处理方法包括数据抽取、数据转换、数据清洗和集成、数据规约、属性概念分层的自动生成等。经过预处理的数据能提高数据挖掘结果的质量,使挖掘过程更有效、更容易。

海量数据挖掘的关键是数据挖掘算法的并行化[10,11]。由于云计算采用的是 MapReduce 等新型计算模型,需要对现有的数据挖掘算法和并行化策略进行一定程度的改造,才有可能直接应用在云计算平台上进行海量数据挖掘任务。因此需要在数据挖掘算法的并行化策略上进行更为深入的研究,从而使云计算并行海量数据挖掘算法的高效性得以实现。并行海量数据挖掘算法包括并行关联规则算法、并行分类算法和并行聚类算法,用于分类或预测模型、数据总结、数据聚类、关联规则、序列模式、依赖关系或依赖模型、异常和趋势发现等。基于此,针对海量数据挖掘算法的固有的特点对已经存在的云计算模型进行优化升级以及适当扩充,使其对海量数据挖掘的适用型得到最大程度的提升。

最顶层是面向用户的用户层,该层主要接收用户的请求,并将传递给下面两层,并将最终的数据挖掘结果展示给用户。用户通过友好的可视化界面管理和监视任务的执行,并且可以很方便地查看任务执行结果。

用户的数据挖掘请求通过用户输入模块传递到系统内部,系统根据用户提交的一些数据挖掘参数和基本数据,在算法库中选择合适的数据挖掘算法,然后调用经过预处理阶段的数据,分配到 MapReduce 平台上进行并行数据挖掘,挖掘出的结果通过结果展示模块传递给用户。海量数据的存储和并行化处理都依赖于云计算环境。

2.4　基于云计算的数据挖掘模型的不足及后续工作开展的方向

由于云计算还处于高速发展时期,必然会面临很多不足与挑战,基于云计算的数据挖掘中也同样存在着一些问题。

1)云计算带来的需求问题。基于云计算的数据挖掘,最终会发展成为一种云服务模式,必然会面临着多样化和个性化的需求。

2)海量数据的问题。从数量上来说,可能需要处理数量级达到 TB 级乃至 PB 级的数据,另外还有高维数据、各种噪声数据以及动态数据等,这都为数据处理带来了极大的困难。

3)算法的选择问题。选择合适的算法及并行策略来完成任务是最关键的问题。还有算法的设计、参数的调节都会直接影响到最终的结果。

4)不明确性问题[12]。数据挖掘过程中可能会存在许多不明确性,进行数据挖掘的目的就是要将这些不明确性带来的影响降低到最低。这些不明确性包括对数据挖掘任务描述的不明确性、进行数据采集和预处理时会出现的不明确性、数据挖掘方法选择和最终

结果的不明确以性及对如何评价数据挖掘结果的不明确性等。

针对以上提出的问题,后续工作可以从以下几个方面着手:

1)基础设施建设方面,根据多样化和个性化需求,并综合考虑到各领域各行业的特点,构建专属的数据挖掘云服务平台。

2)虚拟化技术为数据挖掘云服务提供了重要的技术支持,后续应加大对虚拟化技术的研究开发,并促进其成果的广泛应用,高效地对计算资源实现自主分配和调度。

3)在云服务应用产品的研发环节中,应多考虑社会实际需求,并大力引导公众积极参与其中,这样就可以更好地满足数据挖掘个性化、多样化的需求。

4)在可信性方面,使用的算法最好具有通用性,并且可以随时进行检查、调整以及查看。

5)数据安全[13]问题不能像一般的信息安全那样直接加密,应该是由客户根据自己的需求,在自己的平台终端上自主通过适当加密措施对数据进行保护。

3 结束语

由于云计算高效率、低成本、高可用性等优势,文中将云计算技术引入到数据挖掘中,提出了基于云计算的海量数据挖掘服务,分析了基于云计算的数据挖掘服务的具体层次结构和优势,总结了云数据挖掘中存在的一些缺陷及问题,并针对这些问题提出了一些简单建议。

未来数据挖掘云服务将会有很好的势头,更多的专业人士会成为服务的供应商,公众和各种企业组织机构会从这项服务中受益良多,数据挖掘研究受计算环境的影响将降低,其应用范围也将大大拓宽。但是由于云计算的安全还没有得到完全的证实,所以接下来的工作在云计算安全方面应得到加强。

参考文献:

[1] 张建勋,古志民,郑 超. 云计算研究进展综述[J]. 计算机应用研究,2010,27(2):429-433.

[2] 李 凯,常 征. 基于云计算的并行数据挖掘系统设计与实现[J]. 微计算机信息,2011,27(6):121-123.

[3] 刘 鹏. 云计算[M]. 北京:电子工业出版社,2010.

[4] Armbrust M, Fox A, Griffith R, et al. Above the Clouds: A Berkeley View of Cloud Computing[EB/OL]. [2011-01-10]. http://www. EECS. berkeley. edu/Pubs/TechRpts/2009/EECS-2009-28. pdf.

[5] 李成华,张新访,金 海,等. MapReduce:新型的分布式并行计算编程模型[J]. 计算机工程与科学,2011,33(3):129-135.

[6] 李军华. 云计算及若干数据挖掘算法的 MapReduce 化研究[D]. 成都:电子科技大学,2010.

[7] Wang Jianzong, Wan Jiguang, Liu Zhuo, et al. Data Mining of Mass Storage Based on Cloud Computing[C]//International Conference on Grid and Cooperative Computing. [s. l.]:[s. n.],2010:426-431.

[8] 罗军舟,金嘉晖,宋爱波,等. 云计算:体系架构与关键技术[J]. 通信学报,2011,32(7):3-21.

[9] 拓守恒. 云计算与云数据存储技术研究[J]. 电脑开发与应用,2010,23(9):1-3.

[10] 李玲娟,张 敏. 云计算环境下关联规则挖掘算法的研究[J]. 计算机技术与发展,2011,21(2):43-46.

[11] Li Lingjuan, Zhang Min. The Strategy of Mining Association Rule Based on Cloud Computing[C]//2011 International Conference on Business Computing and Global Informatization (BCGIN). [s. l.]:[s. n.],2011:475-478.

[12] 冯朝一. 云理论在数据挖掘中的应用研究[D]. 南宁:广西大学,2007.

[13] 陈丹伟,黄秀丽,任勋益. 云计算及安全分析[J]. 计算机技术与发展,2010,20(2):99-102.

(上接第 68 页)

径问题中的应用[J]. 东北大学学报(自然科学版),2011,31(7):1378-1386.

[7] 赵鹏军,刘三阳. 一种新的智能优化及其改进研究[J]. 小型微型计算机系统,2010,31(5):955-958.

[8] Al-Betar M A, Doush I A, Khader A T, et al. Novel selection schemes for harmony search[J]. Applied Mathematics and Computation,2012,218(10):6095-6117.

[9] 陈有青,徐蔡星,钟文亮,等. 一种改进选择算子的遗传算法[J]. 计算机工程与应用,2008,44(2):44-49.

[10] Mahdavi M, Fesanghary M, Damangir E. An Improved Harmony Search Algorithm for Solving Optimization Problems[J]. Applied Mathematics and Computation,2007,188(2):1567-1579.

[11] 卫 田,范文慧. 基于 NSGA Ⅱ的物流配送中车辆路径问题研究[J]. 计算机集成制造系统,2008,14(4):776-784.

[12] 罗 彪,郑金华. 多目标进化算法中基于动态聚集距离的分布性保持策略[J]. 计算机应用研究,2008,25(10):2934-2938.

基于云计算的海量数据挖掘研究

作者： 贺瑶， 王文庆， 薛飞
作者单位： 贺瑶,薛飞(西安邮电大学 管理工程学院,陕西 西安710061)， 王文庆(西安邮电大学 自动化学院,陕西 西安710061)

刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(2)