基于云计算的数据挖掘系统设计与实现

王晓妮[1],段 群[2],韩建刚[3]
(1. 咸阳师范学院 信息中心,陕西 咸阳 712000;
2. 咸阳师范学院 计算机学院,陕西 咸阳 712000;
3. 西北机电工程研究所 生产部电调室,陕西 咸阳 712000)

摘 要: 为了解决数据出现指数式增长所导致的海量数据与传统数据挖掘系统计算能力有限的矛盾日益尖锐这个问题,提出了一种将云计算技术和数据挖掘有机结合的解决方案。通过采用Map/Reduce这种能够处理大量半结构化数据集合的并行编程模型方法,将云计算技术融入海量数据挖掘过程中,设计并实现了基于云计算的数据挖掘系统。通过对高校师生在图书馆的电子文献资料查阅日志数据集的挖掘分析,对该系统的性能进行了测试,表明该系统能够实现根据用户需求为其提供即时服务。实验结果表明,该系统的运行效率和挖掘速度均高于单机系统,而且随着数据量的增加,挖掘效率的优势愈发明显。故该系统能够满足用户需求,可以有效解决传统数据挖掘系统中的技术瓶颈。

关键词: 云计算; 数据挖掘; 海量数据; Map/Reduce
中图分类号: TP302 **文献标识码:** A **文章编号:** 1673-629X(2019)03-0178-05
doi: 10.3969/j.issn.1673-629X.2019.03.037

Design and Implementation of Data Mining System Based on Cloud Computing

WANG Xiao-ni[1], DUAN Qun[2], HAN Jian-gang[3]
(1. Information Center, Xianyang Normal University, Xianyang 712000, China;
2. School of Computer Science, Xianyang Normal University, Xianyang 712000, China;
3. Electric Control Room of Production Department, Northwest Electrical and Mechanical Engineering Research Institute, Xianyang 712000, China)

Abstract: In order to solve the problem of the ever-increasing contradiction between the massive data and the limited computing capacity of traditional data mining system caused by the exponential growth of data, we propose a solution combined cloud computing technology and data mining organic. By using Map/Reduce, a parallel programming model method that can handle a large number of semi-structured data collections, cloud computing technology is integrated into massive data mining process, and a cloud-based data mining system is designed and implemented. This system is tested by excavating and analyzing log datasets of university educators and students in library e-documents. The results prove that the system can provide even services for users according to their needs. The experiment shows that the running efficiency and speed of the system are higher than that of the single machine system, and with the increase of data volume, the advantage of mining efficiency is more obvious. Therefore, the system can meet users' needs and effectively solve the technical bottleneck of traditional data mining systems.

Key words: cloud computing; data mining; massive data; Map/Reduce

0 引 言

互联网、大数据和云计算等信息技术的飞速发展使人类社会进入信息时代,人们经常通过网络来访问和接受各种各样的数据信息。面对这些鱼龙混杂的海量数据和个人对数据的不同需求,使得从海量数据中提取和挖掘有用信息显得非常重要,于是就出现了数据挖掘技术。数据挖掘能够处理信息庞大、数据模糊和组成结构相对复杂的数据,应用范围较广[1]。例如淘宝商就是采用数据挖掘技术来分析消费者的个人需求、喜好、心理价位和消费层次,以此为买家推荐合适

收稿日期: 2018-04-09 **修回日期:** 2018-08-14 **网络出版时间:** 2018-12-19
基金项目: 陕西省教育科学"十三五"规划2017年课题(SGH17H196);咸阳师范学院专项科研基金资助项目(13XSYK087)
作者简介: 王晓妮(1977-),女,硕士,工程师,研究方向为计算机应用、网络安全和云计算。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181219.1511.040.html

的商品或商家,便于买家进行合理快速的选择。数据挖掘为人们解决了燃眉之急,促进了信息时代的发展,使数据更加复杂和庞大[2]。随着数据量出现几何级爆炸式的迅猛增长,导致数量级从最初的 MB 发展到 TB 或 PB,这就使传统的数据挖掘系统面临严重的挑战和威胁[3]。

网络技术能为人们提供大量的信息,但它也使人们从海量数据中提取有用信息的难度越来越大,这就必须加快数据挖据技术的发展。因此,云计算技术应运而生,它的出现和发展给数据挖掘造成了技术瓶颈[4],但也为数据挖掘和云计算有效结合的新模式创造了发展机遇。云计算的 SaaS 标准化服务模式为数据挖掘提供了良好的理论和技术支持,低成本、存储能力强,可伸缩和动态的计算能力等特点能够实现海量数据的高效挖掘[5]。将云计算技术灵活地应用于数据挖掘领域,例如随着高校数据馆数字化的推进,通过挖掘电子文献资料查阅日志信息资源来了解读者需求,为管理者提供可靠的科学依据,为用户提供及时服务变得非常重要。因此开发一个基于云计算的数据挖掘系统,并把它应用在高校图书馆电子文献资料的查阅过程中显得迫在眉睫。

1　理论知识

1.1　云计算技术

云计算(cloud computing) 是并行计算、效用计算、虚拟化、网络存储、负载均衡、网格计算、热备份冗余和分布计算等传统计算机和网络技术迅速发展融合的产物,它采用的是一种基于互联网相关服务的使用、增加、交付和按使用量付费的模式,利用网络提供易扩展可动态变化的虚拟化资源。通过把计算任务分布在由大量计算机组成的资源上,便于各个应用系统按需获得存储空间、信息服务和计算力。云计算通常为用户提供 LaaS(基础设施即服务) 、SaaS(软件即服务) 和 PaaS(平台即服务) 这三种服务形式[6],其中 LaaS 能提供以硬件设备为基础的网络、计算和存储服务,对硬件资源实现了抽象和服务化提供,能够完成分布式存储和分布式计算。云计算具有超大规模、高可靠性、虚拟化、高扩展性、通用性、极其廉价、按需服务和潜在的危险性等特点[7]。

1.2　数据挖掘

数据挖掘(data mining) 是指从模糊的、大量的、有噪声的、随机的和不完全的实际应用数据中,提取那些人们事先未知的、隐含在其中的、但又非常有用的知识和信息的过程[8]。数据挖掘就是通过在数据库中的大量信息中分析每个数据,从中找出其规律并挖掘出有用信息的技术,主要包括三个步骤: 数据准备、寻找规律和规律表示。典型数据挖掘系统的总体结构如图 1 所示。数据挖掘按照其挖掘目的可分为关联分析、分类分析、类聚分析、特异分析、演变分析和异常分析等挖掘任务[9]。

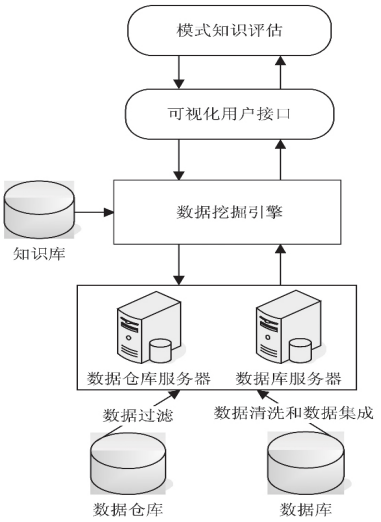

图 1　典型数据挖掘系统的总体结构

数据挖掘过程[10]是一个通过原始数据不断修改、调整和循环的挖掘知识过程,可分为四个步骤: 数据预处理(数据采集、数据清洗、数据集成、数据抽样和数据转换) ; 数据挖掘(通过各种挖掘算法,对预处理的数据进行知识发现的过程) ; 模式评估(根据用户特点、爱好等来识别如何表示知识模式) ; 知识表示(将最终的挖掘结果通过可视化的知识表示技术展示给用户) 。

2　需求分析

伴随云计算的发展和信息时代的到来,呈现指数式增长的海量数据导致信息超载,个性化和多样化的数据挖掘需求越来越明显,数据复杂度与传统的数据挖掘系统计算能力有限的尖锐矛盾日益突出[11]。效率太低、速度太慢、能耗太高等缺陷使传统的单机数据挖掘系统在并行计算中黔驴技穷,集中式传统数据挖掘系统早已无法适应,出现了技术瓶颈,只有并行计算才能满足海量数据的大规模计算。而目前传统的数据挖掘技术或解决方案大多数都是围绕数据挖掘系统这个中心,着重于数据挖掘算法和系统设计工程,缺乏从广大用户实际应用的角度去考虑数据挖掘技术的具体实施。

许多现有的数据挖掘系统只用于少数具有专业挖掘知识的技术人员,只有了解和熟悉相应的数据挖掘算法,才能取得理想的挖掘结果,这让大多数的普通用户望而生畏。这些因素阻碍了数据挖掘技术的推广和应用,使其具有一定的局限性,无疑增加了各企业选择或开发适合自己业务的数据挖掘系统的投资成本。面

对急速增长的海量数据,如何快速准确地从这些杂乱无章的海量数据中挖掘出有价值的信息已成为目前许多数据挖掘系统急需解决的问题。具备高可用性、并行计算、虚拟化、动态资源分配和调度特点的云计算平台能够满足信息时代数据挖掘计算性能的基本要求。

具有可弹性变化的计算能力和海量存储能力使云计算技术成为解决传统数据挖掘系统缺陷的有效策略,云计算平台通过不同的部署模式突破了传统数据挖掘技术瓶颈[12],成为优化传统数据挖掘系统较好的解决方案。例如在高校中数字图书馆如果要适应师生特定需求,要准确及时地向用户提供所需数据,那么就要对用户查阅日志数据集进行必要的数据挖掘分析,然后才能动态地组织和呈现出用户所需的相关数据。基于云计算的数据挖掘系统是实现数字图书馆信息资源整合的最优方案。

3 系统设计

3.1 系统设计目标

该系统是通过云计算和数据挖掘技术相结合而搭建起来的,不仅能为各种终端用户提供友好便捷的界面服务,还能集成基于本系统开发的其他应用程序开放接口。用户可以通过各种终端设备登录用户界面直接访问或通过相关应用程序调用系统的开放接口间接访问这两种方式来使用系统。系统应该满足这几个基本要求:

(1) 数据的分析深度。系统必须能够满足分析多种类型数据的需求,可以从多方面和多角度进行数据分析,便于快捷地加入比较复杂的计算机学习和概率统计等算法。

(2) 数据的适应性。系统能够处理各种类型的数据,尽量满足用户对数据挖掘的个性化和多样化需求。

(3) 友好的用户界面。能够满足大多数用户操作习惯,通过开放接口的直接调用实现外部服务的使用。

(4) 系统的便捷性。能够满足和快速适应数据指数增长和频繁更新的应用场景,利用虚拟技术对计算资源进行调度和自主分配。

(5) 挖掘算法的通用性。系统中的算法要大众化,能够根据挖掘数据和用户需求的不同而随时调整。

(6) 数据的安全性。根据用户需求和级别划分相应的操作权限,防止用户隐私数据和信息被非法访问。

3.2 系统架构设计

云计算海量数据挖掘系统架构体系如图 2 所示,利用面向组件的分层设计思想,从上至下共分为用户层、数据挖掘云服务层和云计算支撑平台。其中用户层主要负责用户与云平台的信息数据交互,开放接口便于用户获取数据集和算法调用,用户界面为用户提供快捷友好的访问操作和挖掘结果展示; 数据挖掘云服务层封装了大量的方法和类,算法和函数等为数据预处理模块和数据挖掘模块提供服务,方便其随时调用; 云计算平台为数据挖掘云服务层提供其所需的所有应用程序接口。该架构具有硬件投入少、软件消耗低、各模块间相互配合、后期维护简单和安全性高等特

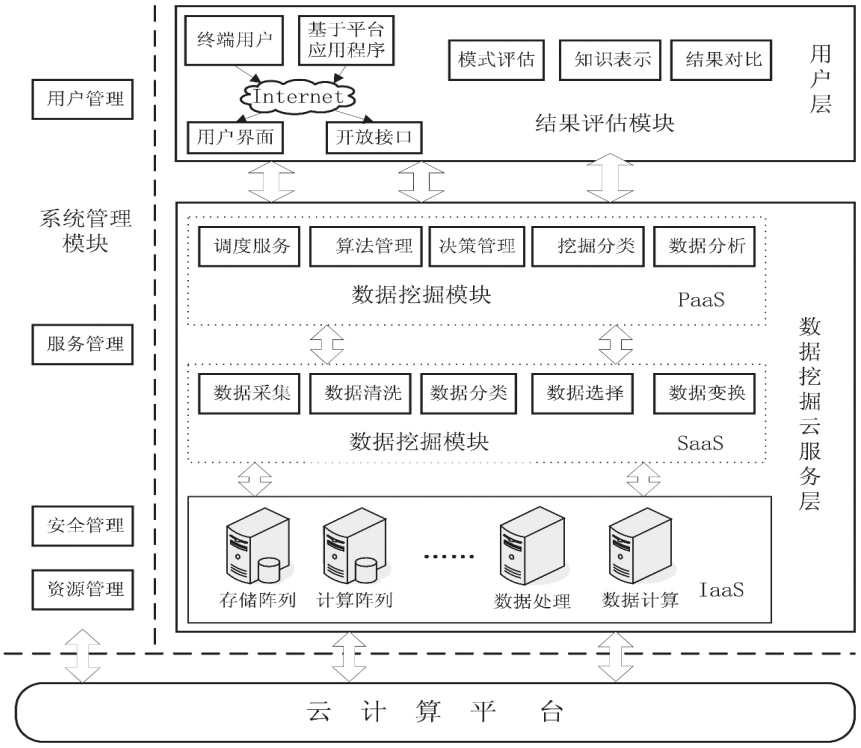

图 2 云计算挖掘系统的架构

点。系统通过修改部分接口和添加类来方便地增加新的功能模块,分层而治的架构设计模式增强了系统的安全性。

3.3　系统功能模块的设计

基于云计算的数据挖掘系统主要有以下四个功能模块:

(1) 系统管理模块。主要负责统一管理系统的用户、安全、服务和资源。系统能够通过友好的界面为用户直接提供其所需的相关资源、软件和服务。

(2) 数据预处理模块。主要对数据进行采集、清洗、集成、选择和变换等操作。系统首先要从海量数据源中采集相关的原始数据,然后调用接口服务和清洗算法对采集来的相关数据进行清洗,对清洗结果按类别进行集成,选择出有用数据进行类型变换后通过数据层存储在云平台上。

(3) 数据挖掘模块。利用云计算技术首先在接口处对预处理后的数据进行合理分类,然后对其按类别进行调用服务和详细的分析,找出隐藏的有用信息,最后根据此分析结果通过决策管理选择合适的挖掘类型和最优的挖掘算法,实现数据挖掘。

(4) 结果评估模块。对得到的挖掘结果进行模式评估和结果对比,选择出最优化的数据挖掘结果并进行可视化处理后将其以知识表示的方式展示给用户。

系统的具体功能模块如图 3 所示。

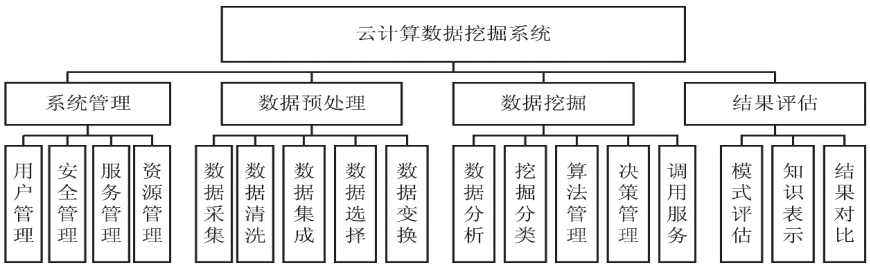

图 3　系统的功能模块

4　系统实现

4.1　实现思想和开发环境

该系统的实现思想就是利用计算技术来实现海量数据挖掘,要将云计算技术融合在数据挖掘中,通过研究挖掘算法并把算法部署在云计算平台中,将计算扩展到无限规模的服务器集群上实现数据挖掘。系统采用 Map/Reduce 分布式计算方式,把海量数据集和大量的挖掘任务划分成若干个子任务,分配到多台计算机上进行并行处理,通过不断调用系统资源来实现具体的数据挖掘任务。

该系统采用集成的跨平台开发环境,该集成开发环境是利用 Java 语言开发,具有灵活性强和系统资源免费的优点。系统的开发平台是基于 Google 的 App Engine SDK [32] 云计算开发平台,开发工具为 Eclipse3.4,开发语言为 Python。该语言具有可扩展、可移植、面向对象和丰富的资源库等优点,便于系统的协同开发。

4.2　算法实现

采用 Map/Reduce 这种并行的编程模型能够实现把云计算技术融入海量数据挖掘中,该模型是主从结构,把用户的所有请求任务看作为一个个作业。在处理杂乱无章的大规模数据时,先要将其分割成无数个块后再将大规模的计算任务扩展到由大量普通单机服务器组成的无限规模机器群集上并行完成。采用 Map/Reduce 进行海量数据挖掘算法流程如图 4 所示。

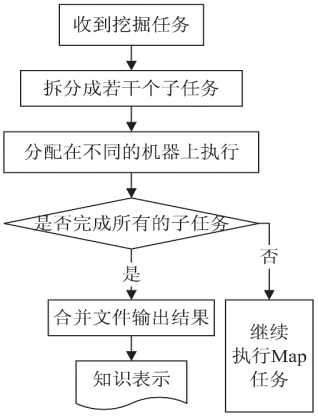

图 4　数据挖掘算法流程

算法具体实现步骤如下:

(1) 定义一个 Map(映射) 函数,从结构各异的大数据集中解析每个数据,从中提出表示数据特征的键值 key 和 value。

(2) 把用户的挖掘请求当成相应的作业,利用 Map 函数将其拆分成若干个不同的 Map 任务,并将这些子任务分配到数据挖掘平台中不同的机器上去处理,再由给定的键值对<key1,value1>生成新的对应键值对<key2,value2>。

(3) 合并 Map 输出的相同 key2 键值后,经过 Shuffle 阶段映射成一组新的键值对<key2,lsit(v2)>。

(4) 判定所有子任务是否完成映射,完成后进入下一步,否则继续映射。

(5) 定义一个 Reduce(归约) 函数,并把新的键值对<key2,lsit(v2)>指定给它,形成新的键值对<key3,

value3>,并将其写入文件。

(6) 将这些文件结合起来的目标文件就是挖掘结果,对其进行可视化处理后输出展示给用户。

4.3 系统测试及应用结果

系统验证环境为 Google App Engine SDK,CPU 为 3.3 GHz,内存 4 G,选用该校师生在电子数据资源库(知网和万方数据)中查阅学术论文的数据集对系统的可行性进行测试和验证。把所选数据集进行分组后,采用 K-means 聚类和 ID3 决策树这两种挖掘算法分别在云计算平台和本地上运行来验证系统,该系统数据挖掘结果分析对比如图 5 和表 1 所示。

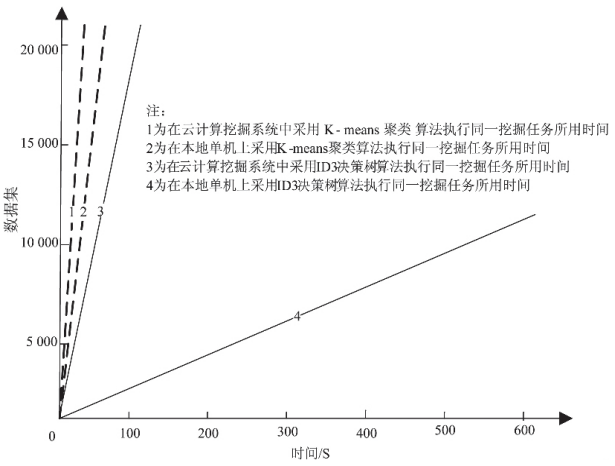

图 5 系统数据挖掘结果分析对比

表 1 数据挖掘执行时间对比

运行时间/s	K-means 聚类	ID3 决策树	数据集
本地执行时间	2.19	15.16	1 000
	4.92	57.89	2 000
	14.15	239.11	5 000
	33.25	1 390.89	10 000
	69.74	2 358.23	20 000
云计算挖掘系统执行时间	0.49	3.12	1 000
	1.13	6.89	2 000
	6.89	19.54	5 000
	14.21	68.23	10 000
	22.9	139.82	20 000

通过比较两种算法在不同环境下执行数据挖掘的时间可以得出,K-means 聚类算法和 ID3 决策树算法在云计算挖掘系统中的运行效率和挖掘速度均高于单机系统,而且随着数据量的增加,挖掘效率的优势呈现出正比关系。但同样的数据集在不同的平台中,显然 K-means 聚类算法比 ID3 决策树算法所用时间更短,效率更高。

5 结束语

把云计算技术灵活地融入数据挖掘中,设计并实现了云计算数据挖掘系统。通过该校师生在电子数据资源库(知网和万方数据)中查阅学术论文的数据集对系统的可行性进行了测试和验证,结果表明该系统能够满足大部分用户的基本需求。该系统的各模块处于云计算平台的不同服务模式中,能够为用户提供灵活多样的服务,提高了数据挖掘效率和平台的稳定性。但是系统的个别功能还不是太完善,还有待进一步深化研究和推广。

参考文献:

[1] 贺 瑶,王文庆,薛 飞. 基于云计算的海量数据挖掘研究[J]. 计算机技术与发展,2013,23(2):69-72.

[2] 王小燕. 基于云计算的大数据挖掘平台设计[J]. 电子设计工程,2017,25(13):25-27.

[3] PUTHAL D,SAHOO B P S,MISHRA S,et al. Cloud computing features,issues,and challenges: a big picture[C]//International conference on computational intelligence and networks. Bhubaneshwar,India: IEEE,2015: 116-123.

[4] 包永红. 云计算技术下数据挖掘平台设计及技术[J]. 现代电子技术,2016(16):61-63.

[5] TSAI C F,LIN W C,KE S W. Big data mining with parallel computing: a comparison of distributed and MapReduce methodolog[J]. Journal of Systems and Software,2016,122: 83-92.

[6] JIN Ran,KOU Chunhai,LIU Ruijuan,et al. Efficient parallel spectral clustering algorithm design for large data sets under cloud computing environment[J]. Journal of Cloud Computing: Advances,Systems and Applications,2013,2(1):47.

[7] 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.

[8] 应 毅,任 凯,刘正涛. 基于云计算技术的数据挖掘[J]. 微电子学与计算机,2013,30(2):161-164.

[9] 余 琦,凌 捷. 基于 HDFS 的云存储安全技术研究[J]. 计算机工程与设计,2013,34(8):2700-2705.

[10] 张治斌,李燕歌. 云计算下 MapReduce 多组容错机制架构的分析与研究[J]. 微电子学与计算机,2014,31(1):52-55.

[11] PÉREZ J,ITURBIDE E,OLIVARES V,et al. A data preparation methodology in dam mining applied to mortality population databases[J]. Journal of Medical Systems,2015,39(11):152.

[12] 何健伟. 基于 Hadoop 的数据挖掘算法研究与实现[D]. 北京: 北京邮电大学,2015.