云计算环境下的数据挖掘服务模式

丁 静 杨善林 罗 贺 丁 帅

(过程优化与智能决策教育部重点实验室 合肥 230009) (合肥工业大学管理学院 合肥 230009)

摘 要 为了求解网络环境下分布式海量数据的分析处理、促进数据挖掘的开发集成和商业应用,提出了云计算环境下的数据挖掘解决方案,通过云环境计算能力和云计算服务模式,阐述了对数据挖掘服务问题的解决机理。云计算环境下的数据挖掘是一种网络环境下的信息资源服务模式。基于此,构建了数据挖掘服务的架构,设计了数据挖掘服务的创建流程,给出了数据挖掘服务模型的体系结构,并从生命周期的角度定义了数据挖掘的服务过程,从而形成了云计算环境下的数据挖掘服务模式。

关键词 数据挖掘,服务,云计算

中图法分类号 TP274,TP391 **文献标识码** A

Data Mining Service Model in Cloud Computing Environment

DING Jing YANG Shan-lin LUO He DING Shuai

(Key Laboratory of Process Optimization and Intelligent Decision of Ministry of Education, Hefei 230009, China)

(Management School of Hefei University of Technology, Hefei 230009, China)

Abstract Data mining in cloud computing environment was proposed as a solution, in order to solve the task of distributed massive data analyzing in network and promote development integration and business application of data mining. The solving mechanism of data mining service was explained by computing capability and service model of cloud computing. Data mining in cloud computing environment is a service model of information resources in network. Based on these, the data mining service architecture was constructed, and the data mining service creating procedure was designed. The system architecture of data mining service model was depicted. The service process of data mining was defined. The data mining service model in cloud computing was formed consequently.

Keywords Data mining, Service, Cloud computing

1 引言

云计算已成为当前的一个研究热点,Google、Amazon、IBM等主流信息技术公司先后提出了各自的云计算体系架构,多家研究机构也提出了各种云计算实践平台,如芝加哥大学和佛罗里达大学开发的用于科研教育的弹性云计算平台Nimbus Cloud和Florida Cloud[1]。云计算是借助高速带宽和虚拟化技术,在分布式计算、并行计算、网格计算和效用计算基础上的进一步发展。目前,云计算还处于研究与应用的初级阶段,尚未形成统一的标准和定义。分析和综合众多云计算定义,可以得出其基本特点:云环境具有超大规模的存储和计算能力,资源和结构具有动态伸缩性,并且通过虚拟化技术和庞大的资源池按需提供服务。云计算的这些特点使数据存储、分析和应用的商业化成为可能,也使云计算环境下的数据挖掘成为一个具有理论和应用价值的研究领域。

随着业务量的增长和业务范围的扩展,企业数据库中积累了海量的商业数据,传统的数据挖掘模式无法满足海量数据挖掘对计算能力的需求,因而需要建立具有高性能计算能力的新型数据挖掘模式。同时,网络环境下,为了适应数据量的增长和跨地区的业务操作,企业的数据多存储在分布式的数据仓库或数据中心上,现有的大量数据管理软件和商业决策软件不支持网络环境下的分布式挖掘技术,因而需要构建一个能够处理分布式数据存储、分布式执行数据分析任务的数据挖掘模式。云计算为网络环境下的数据挖掘提供了良好的解决方案,解决了传统数据挖掘方法在网络数据分析中存在的问题。云计算环境下的资源以分布式的形式存储,数据挖掘任务的执行模式有别于传统的本地单机挖掘模式,符合网络环境下数据挖掘的要求。云计算超大规模的服务器集群具备超强的计算能力,云存储具备强大的存储能力、数据分析能力和数据管理能力,其共同构成了海量数据挖掘开发和应用的有利基础。

现有的数据挖掘解决方案大多以系统为中心,特别重视算法和系统工程,没有从用户的角度探讨数据挖掘技术的应用,使系统难于操作和使用。一些数据挖掘工具只适合专业

本文受国家自然科学基金(71131002,71071045)资助。

丁 静(1987-),女,博士生,主要研究方向为云计算、数据挖掘、服务计算等,E-mail:ahdingjing@163.com;杨善林(1948-),男,硕士,教授,博士生导师,主要研究方向为智能决策、云计算等;罗 贺(1982-),男,博士,讲师,主要研究方向为信息资源管理、智能决策、云计算;丁 帅(1984-),男,博士,讲师,主要研究方向为智能决策、云计算、可信软件。

技术人员，如果对算法不了解，则难以得出好的模型，这也增加了企业纵向开发数据挖掘平台的技术成本，阻碍了数据挖掘的企业应用。云计算环境从面向服务的角度为数据挖掘提供了良好的解决方案。在云平台中，存储、平台、应用都是可共享的资源，这些资源被封装成具有统一接口的组件，以服务的形式提供给用户和开发者。此外，作为一种商业计算模式，云计算的软件即服务（Software as a Service，软件即服务）模式将数据挖掘程序作为服务按需出售，降低了中小企业的数据挖掘成本，为数据挖掘商业应用的推广提供了良好的平台。

在传统的分布式数据挖掘和网格数据挖掘的基础上，结合现有的云计算相关研究，国内外的专家学者们对云计算环境下的数据挖掘进行了开拓性的探索，提出了初步的设计构想。现有的研究成果主要集中在3个方面：一是云计算环境下的数据挖掘算法研究，即通过算法在云计算环境下的移植或改进，来提高算法的性能；二是云计算环境下数据挖掘的体系架构研究，即分析设计数据挖掘平台的体系结构；三是云计算在数据挖掘应用中的研究，亦即将云计算平台作为数据挖掘商业应用的解决方案。文献[2]认为可将云计算用于数据挖掘和机器学习领域，并在云计算平台上运行了基于 K 邻近和约束玻尔兹曼机算法的客户兴趣预测模型，有效提高了模型的精度。文献[3,4]将云架构分成资源和服务两部分，并设计了用于高性能广域网中海量分布式数据抽取、管理、分析和分配的云环境数据挖掘架构，该架构主要由两部分组成：负责大规模数据集存储的存储云 Sector 和支持海量分布式数据集并行分析的计算云 Sphere。文献[5]将云计算作为大型社会性网络分析的解决方案，并在 Amazon 云架构下用 PageRank 算法对社交网络 Twitter 的用户进行排序。

纵观上述文献，用于云计算环境下的数据挖掘的技术、工具和平台已经取得了一定的研究成果，然而对数据挖掘解决方案的研究尚且不足。Gregory Piatetsky-Shapiro 在 KDD2000 上从应用角度将数据挖掘系统的发展归纳为3个阶段：独立的数据挖掘软件、横向的数据挖掘工具集、纵向的数据挖掘解决方案[6]。数据挖掘算法和技术可视为前两个阶段发展的显著成果，目前面临的难题是，以公共设施服务的形式为用户和开发者提供数据挖掘的解决方案，也即第三阶段的核心内容，面向服务的云计算环境正是解决这一问题的有效途径。

云计算环境通过基础设施即服务（IaaS）、平台即服务（PaaS）和软件即服务（SaaS）3种服务模式，将数据存储、计算设备、开发平台、应用软件等软硬件资源以服务的形式提供给用户，形成一种按需获得的计算服务。在这种计算服务的模式下，用户使用云计算环境下的数据挖掘，关心的不是各种数据挖掘应用在云平台中的实现，而是根据数据挖掘任务的需求，最大限度地使用云平台中服务于数据挖掘的各种资源，包括计算资源、存储资源、应用程序资源等。云计算环境下的数据挖掘是一种网络计算资源的应用，其实质是一种服务模式。用户向云端提出的数据挖掘任务就是一种云服务；执行数据挖掘运算的处理器和存储空间均视为服务的资源，与任务相关的数据库、数据仓库、数据挖掘算法等可看作是支持数据挖掘服务的专用资源。基于此，本文提出了一种面向服务的数据挖掘模式——基于云计算的数据挖掘服务。

2 基于云计算的数据挖掘服务架构

数据挖掘服务是数据挖掘所涉及的功能、行为的集合，包括数据选择、数据预处理、数据集成、挖掘、分析、结果表示和评价等，通过混合并搭配这些功能，形成新的复合应用。云计算构建了一个实现计算机设备、存储设备、服务器集群、集成开发环境、应用软件等共享的网络环境[7]。在此基础上，通过虚拟化、组件、接口和集成技术，将软硬件封装打包成相应的服务模块，响应基础设施、平台开发和应用3个不同层次上用户的服务请求，即 IaaS（Infrastructure as a Service，基础设施即服务）、PaaS（Platform as a Service，平台即服务）和 SaaS（Software as a Service，软件即服务），从而实现了一套完整的服务模式。基于此，云计算环境下的数据挖掘可以为用户提供一整套数据挖掘开发和应用所必须的能力，为数据挖掘服务提供良好的解决方案。根据数据挖掘的行为和需求，结合云计算的体系结构以及 SOA（Service Oriented Architecture，面向服务的体系结构）架构[11]，设计了基于云计算的数据挖掘服务架构，见图1。

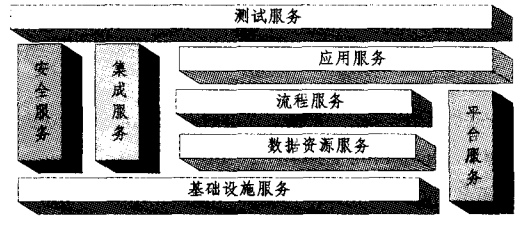

图1 基于云计算的数据挖掘服务架构

基础设施服务是基于数据中心的服务，以服务形式提供数据挖掘所需计算资源，并提供远程访问这些资源的能力。数据资源服务提供远程托管数据库的服务，使用户可以像使用本地数据库一样使用远程数据资源，并提供需求驱动的数据库、数据仓库技术。流程服务提供数据挖掘业务流程服务，可跨多个系统，将数据挖掘的关键模块与数据信息绑定起来，形成挖掘流程的元应用，创建挖掘流程的远程资源。应用服务也称为软件即服务，将数据挖掘应用程序作为整体，通过网络平台交付给终端用户。测试服务通过远程托管的测试工具对本地的数据挖掘系统或云平台中交付的数据挖掘系统进行测试。平台服务提供了数据挖掘应用的远程开发服务，包含应用程序开发、接口开发、数据库开发、存储、集成、部署、测试和运行维护等功能，使用户可以创建数据挖掘的企业级应用。集成服务基于应用抽象接口、语义仲裁、流控制、整合设计等技术，提供数据挖掘应用中异质系统和异构数据资源的集成功能，并以服务的形式交付给用户。安全是云计算的一个弱点，安全服务通过提供数据挖掘中的加密服务、身份管理服务和安全等级服务，为数据挖掘构建安全的云计算环境。

基于云计算的数据挖掘服务架构使用户能够更加灵活地使用服务资源，同时使开发者按照业务需求进行动态的服务组合成为可能。

3 基于云计算的数据挖掘服务建模流程

实现数据挖掘服务，最重要的是根据上述服务架构的分析，为每一个数据挖掘服务建立起服务组件模型。通过服务的识别和描述、数据的关联，可构造出一个完成的服务组件，

其创建流程如图 2 所示。

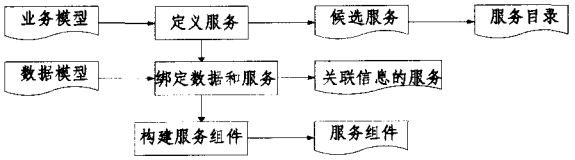

图 2　数据挖掘服务的创建流程

(1)定义服务。在问题域中理解并收集数据挖掘服务相关的信息,定义该服务的核心功能、所需的数据支撑以及服务的产出。通过服务的定义,获得与数据挖掘业务相关的服务描述,形成数据挖掘服务目录中的候选服务。

(2)绑定数据和服务。数据挖掘服务是由功能和数据共同构成的,根据服务的定义,为候选服务及其关联的数据建立联系。通过这一过程,为服务找到其行为所需的信息,规定服务对信息的调用,得到关联信息的服务。

(3)构建服务组件。整合服务描述中的方法和服务绑定的数据,实现服务定义的功能,并将服务封装打包成独立的组件;定义调用服务的统一接口,形成独立完整的服务组件。

4　基于云计算的数据挖掘服务模型体系结构

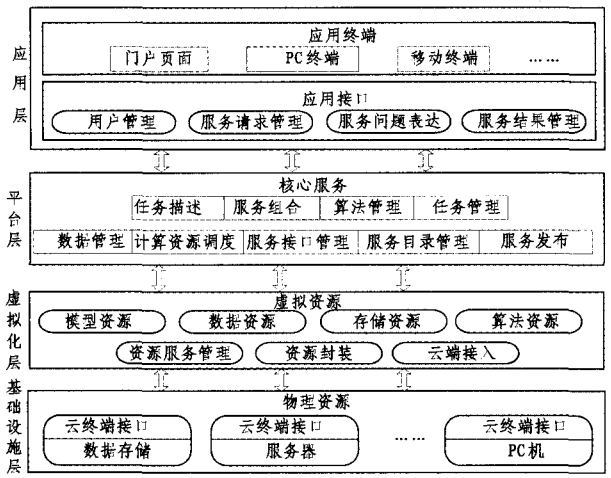

图 3　云数据挖掘服务模型的体系结构

为了实现上述的云计算环境下的数据挖掘服务架构,提出图 3 所示的云数据挖掘服务模型的体系结构,该结构包括以下 4 个层次:

(1)基础设施层　基础设施层提供数据挖掘服务所需的计算、存储资源。该层通过终端接口,将各种物理资源接入到网络中,实现物理资源的互联和共享,并为虚拟化过程提供接口。

(2)虚拟化层　虚拟化层利用虚拟化工具,将云计算环境下的各种分布式资源汇聚,并封装成逻辑集中的、统一透明的服务资源;通过对资源的管理,实现数据挖掘过程中资源的合理分配和调度,并将资源封装,提供给平台层的应用和开发。

(3)平台层　平台层是数据挖掘服务的核心服务层,为数据挖掘服务的实施和综合管理提供各种核心服务和功能,包括面向服务建模的服务目录管理和服务组合,用于服务实施的任务描述、数据管理和计算资源调度等。

(4)应用层　应用层由应用接口层和终端层两部分构成。应用接口层提供面向用户的认证、用户管理、请求处理和请求表达,以及服务结果的管理。终端层为不同的访问介质提供了不同的访问接入,通过门户页面、PC 机、移动终端和各种专用终端等,用户均可以访问和使用云数据挖掘服务。

5　基于云计算的数据挖掘服务过程

云计算环境下的数据挖掘从管理角度看,是一个服务过程;从技术角度看,是一种软件产品。结合软件的生命周期,在云计算的分布式开发过程中,探讨云数据挖掘的服务过程。将基于云计算的数据挖掘服务过程定义为分析、设计、开发、维护和衰亡 4 个阶段,如图 4 所示。

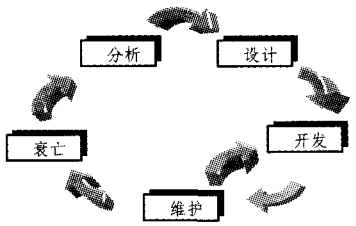

图 4　云数据挖掘服务的过程

(1)分析阶段:根据需求发现或识别服务,定义相应的服务描述,标志着服务生命周期的开始。

(2)设计阶段:根据服务定义和描述,生成服务的说明书,设计服务的接口和契约,包括服务的语义和非功能性特性,以及服务提供商、服务消费者和服务代理之间的契约。

(3)开发阶段:实现服务的功能性特性,在服务定义的范围内进行低耦合、高内聚的功能集成,最终形成独立的服务组件,并通过服务接口进行功能的沟通与调用。

(4)维护阶段:服务处于运行状态时,在不影响服务设计的前提下修复开发的缺陷,或根据服务的需求更新已有的服务版本。服务的开发与维护是同时进行的,使服务的生命周期处于不断成长、成熟的循环发展状态。

(5)衰亡阶段:当服务的功能无法通过维护满足服务需求时,必须撤销该服务,防止使用中服务的数量出现急剧膨胀。

结束语　鉴于现有数据挖掘服务研究和应用的不足,本文根据云计算的信息资源服务模式,它以及云计算面向服务的架构和云计算平台的体系结构,提出了云计算环境下的数据挖掘服务,构建了数据挖掘服务的架构,设计了数据挖掘服务的创建流程,给出了数据挖掘服务模型的体系结构,并定义了数据挖掘服务的过程,从而形成了基于云计算的数据挖掘服务模式,它对云计算环境下数据挖掘服务模式的研究以及数据挖掘服务的开发和应用具有一定的参考价值。在以后的研究中,需要在云计算平台下,结合数据挖掘的应用实例,构建数据挖掘的服务模型,实现相应的数据挖掘服务,从而实施并进一步验证本文提出的模型。

参 考 文 献

[1] Keahey K, Figueiredo R, Fortes J, et al. Science Clouds: Early Experiences in Cloud Computing for Scientific Applications [C]// Proceedings of High Performance Computing and Communications. 2008:825-830

[2] Wang Jian-zong, Wan Ji-guang, Liu Zhuo, et al. Data mining of mass storage based on cloud computing [C]// Proceedings of 2010 Ninth International Conference on Grid and Cloud Computing. 2010:426-431

(下转第 237 页)

3.2 仿真实验及结果分析

仿真实验参数设定如下：$\alpha=0.1$，$\gamma=0.95$，所有 $Q(s_t,a_t)$ 表的初值均为 0。猎人可移动的最大步数为 300，任一猎人移动超过 300 步则此次围捕失败。实验分两组，第一组为单一奖惩标准的 Q 学习算法，第二组为多奖惩标准的 Q 学习算法。每组实验分为 100 轮，每一轮实验均是在上一轮学习的基础上继续学习，每轮实验开始时猎人的位置都会被随机分配。

图 3 所示的是基于单奖惩标准的 Q 学习算法的实验结果，图 4 是基于多奖惩标准的 Q 学习算法的实验结果。将 100 轮实验分成 20 组，每组 5 轮，统计这 5 轮围捕中成功的次数，如图 3 和图 4 的左图所示，横坐标是组次，纵坐标是每组成功次数。同时将 100 轮实验中所有围捕成功的实验统计出来，求出每次成功围捕每个猎人行走的平均步数，如图 3 和图 4 的右图所示，横坐标是成功围捕的轮次，纵坐标是每个猎人行走的平均步数。

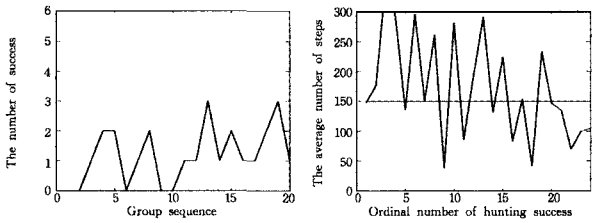

图 3 基于单奖惩标准的 Q 学习算法实验结果

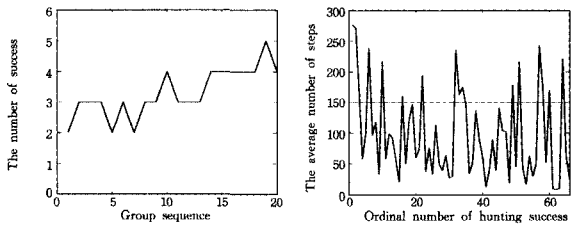

图 4 基于多奖惩标准的 Q 学习算法实验结果

从图 3 和图 4 的左图可以明显地看到，随着实验轮数的增加，基于多奖惩标准的 Q 学习算法的成功次数在明显增加，上升趋势也很明显。从图 3 和图 4 的右图可以更明显地看到，在 100 轮试验中基于单奖惩标准的 Q 学习算法只成功了 20 多次，而基于多奖惩标准的 Q 学习算法则成功了 60 多次。图 3 和图 4 的右图中有一条以 150 为基准的水平线，可以看到，基于单奖惩标准的 Q 学习算法在实验中每次成功围捕所需要行走的平均步数大部分在 150 步以上，而基于多奖惩标准的则大部分在 150 步以下。

由此可以看出，基于多奖惩标准的 Q 学习算法较基于单奖惩标准的 Q 学习算法，无论是在成功的次数上还是行走的平均步数上都有很明显的优势。

结束语 本文提出了基于多奖惩标准的 Q 学习算法，该算法是针对环境复杂、状态较多的学习场景而提出的。传统的单奖惩标准 Q 学习算法过于单一，无法灵活地适应环境或状态的变化，而多奖惩标准的 Q 学习算法减少了单一标准的束缚，避免了许多重复的工作，可以较灵活地适应不同的环境和状态。同时在学习过程中制定阶段目标，分段完成任务，真正做到“因地制宜、因时而异”。从结果上我们也可以明显看到，基于多奖惩标准的 Q 学习算法的成功次数是单奖惩标准 Q 学习算法的 2～3 倍，所需步数减少了近一半，整体性能也有很大的提高。因此，基于多奖惩标准的 Q 学习算法能够灵活适应动态环境，高效地完成学习任务。

参 考 文 献

[1] 徐昕. 增强学习与近似动态规划[M]. 北京：科学出版社，2010

[2] 范波，潘泉，等. 多智能体学习中基于知识的强化函数设计方法[J]. 计算机工程与应用，2005，3：77-79

[3] 陈宗海，段家庆，等. 针对机器人觅食任务的强化学习算法及其仿真研究[C]// 系统仿真技术及其应用. 2008：252-256

[4] 宋清昆，胡子婴. 基于经验知识的 Q-学习算法[J]. 自动化技术与应用，2006，25(11)：10-12

[5] Notsu A, Ichihashi H. State and action space segmentation algorithm in Q-learning[C]// IEEE International joint conference on neural networks. 2008：2384-2389

[6] 黄炳强. 强化学习方法及其应用研究[D]. 上海：上海交通大学，2007

[7] 李铁. 基于多 Agent 交互的团队学习仿真研究[D]. 山西：山西大学，2009

[8] 叶超群. 多 Agent 复杂系统分布仿真平台中的关键技术研究[D]. 长沙：国防科学技术大学，2006

[9] 刘杰. 基于强化学习的多机器人围捕策略的研究[D]. 长春：东北师范大学，2009

[10] 胡子婴. 基于智能体系统的 Q-学习算法的研究与改进[D]. 哈尔滨：哈尔滨理工大学，2007

[11] Stone P, Veloso M. Multiagent Systems: A Survey from a Machine Learning Perspective[J]. Autonomous Robots 8, 2000：345-383

(上接第 219 页)

[3] Robert G, Gu Yun-hong. Data mining using high performance data clouds: experimental studies using sector and sphere [C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008

[4] Grossmam R L, Gu Yum-homg, Michael S, et al. Compute and storage clouds using wide area high performance network [J]. Future Generation Computer Systems, 2009, 25：179-183

[5] Noordhuis P, Heijkoop M, Lazovik A. Mining twitter in the cloud: a case study [C]// Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing. 2010：107-114

[6] Piatetsky-Shapiro G. Knowledge discovery in databases: 10 years after[J]. SIGKDD Explorations, 2000, 1(2)：59-61

[7] 王鹏. 走进云计算[M]. 北京：人民邮电出版社，2009