

面向领域的数据挖掘云平台研究

曾 安¹, 陈 云¹, 潘 丹², 杨海东³, 谢光强¹

(1 广东工业大学 计算机学院, 广东 广州 510006; 2 美国 Batteries Plus 公司,
美国 哈特兰 53029; 3 中油管道物资装备总公司, 河北 廊坊 065000)

摘 要: 为促进数据挖掘开发和商业应用, 通过将云计算技术、领域工程和数据挖掘相结合, 从领域驱动设计和云计算框架出发, 提出了一种面向领域的数据挖掘云平台框架, 阐述了云计算环境下面向领域的数据挖掘服务的机理, 并给出了面向领域的数据挖掘云平台服务开发模式, 从软件开发生命周期的角度定义了面向领域的数据挖掘服务开发过程。

关键词: 云计算; 领域工程; 应用工程; 数据挖掘

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2014)08-0015-05

Domain-oriented Cloud Platform for Data Mining

ZENG An¹, CHEN Yun¹, PAN Dan², YANG Hai-dong³, XIE Guang-qiang¹

(1 Faculty of Computer, Guangdong University of Technology, Guangzhou 510006, China;
2 Batteries Plus LLC, Hartland 53029, USA; 3 Pipe Material & Equipment Corp. of CNPC, Langfang 065000, China)

Abstract: To promote the development and commercial application of data mining, with the help of the combinations of the cloud computing technology, domain engineering and data mining based on the domain-driven design and the cloud computing framework, a domain-oriented data mining cloud platform framework was proposed and the mechanism of domain-oriented data mining services under cloud computing environment was given. Also, the development mode of domain-oriented data mining cloud platform services was advocated and the development process of domain-oriented data mining services from the view of software development and life cycle was described.

Key words: cloud computing; domain engineering; application engineering; data mining

1 引言

目前, 关于云计算和大数据的研究已成为计算机科学中的研究热点之一^[1-4], 在基于云计算环境^[5-6]的数据挖掘过程中, 每个领域人员精通自身的领域知识、并有着丰富的经验, 但是对跨领域的知识的掌握相对就薄弱, 这将使得此类数据挖掘系统不能很好地适应跨领域或多领域的数据挖掘任务. 当需要完成跨领域数据挖掘任务时, 对此类系统的改动将会非常大, 并造成资源的浪费. 基于此, 本研究结合云计算、数据挖掘和领域工程应用, 提出面向领域的数据挖掘云平台架构^[7].

2 相关技术

2.1 领域工程

领域(domain)的定义有多种形式. 在软件工程中, 领域是指一组具有相似或相近软件需求的应用系统所覆盖的功能、问题、问题解决方案或知识区域^[8-10]. 文献[11-12]中提及到领域工程是指为相似和相近功能的系统提供构建基础的过程, 同时为生成可复用构件提供支持. 领域工程有助于产生具有较高可复用性的构件. 它把同一领域的知识进行整合并提炼出适合的知识规约、领域设计和领域架构. 基于现有的领域系统, 通过领域工程产生的可复

收稿日期: 2013-10-18; 修回日期: 2013-12-09

基金项目: 国家自然科学基金项目(61300107); 广东省自然科学基金项目(S2012010010212)

用构件^[13-14]体现了其本质需求.因此,这些构件具有较高的可复用性.领域工程包括3个主要阶段.

领域分析:结合领域知识,对领域中的问题域进行分析和研究,并对特定领域中的共性和特性进行描述.在领域分析过程中,首先通过领域专家获取相应的领域知识,再利用领域知识和用户需求进行分析,获取领域内可复用需求,最后形成需求规约,并对领域知识进行分析建模.

领域设计:以领域分析模型为基础,获取特定领域的软件架构(domain-specific software architecture, DSSA).DSSA结合了领域模型,对领域中一组或一系列系统需求进行高层次设计,不仅可为单个系统,还可为多个系统提供解决方案.并且,由于领域需求的可变性,DSSA也具有可变性这一特点.

领域实现:实现 DSSA 和构件,并依据领域分析模型和 DSSA 组织可复用构件库.

领域工程和应用工程是两个并行的过程.应用工程中可以利用领域工程中可复用的资源,来进一步提高设计开发效率.

2.2 领域驱动设计和经典分层架构

图1所示为为领域驱动设计的经典分层架构,为有效构建基于领域的数据挖掘系统,需要不同领域专家、行业人员、开发人员等进行有效的沟通合

作.所以,结合领域工程和云计算技术提出面向领域的数据挖掘云平台体系结构和开发模式是非常关键的.

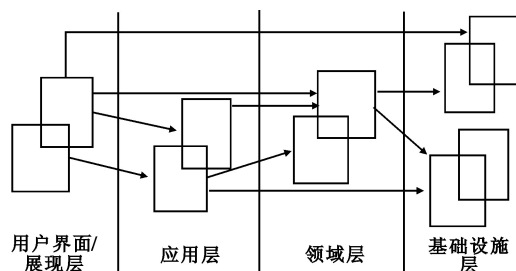


图1 领域驱动设计的经典分层架构

3 面向领域的数据挖掘云平台体系结构

针对云计算和大数据时代的发展,结合领域设计和领域驱动设计经典分层架构,从多层次角度出发,提出一种面向领域的数据挖掘云平台框架如图2所示.主要呈现的是积木式的框架层次以及构件和接口.从系统结构中可以看出,在基于云平台的环境下,我们更加容易把数据挖掘和领域知识有效地结合和管理,使面向领域的数据挖掘变更加开源化.在云计算环境中根据不同的抽象层提供面向领域数据挖掘应用、特殊服务资源和平台.

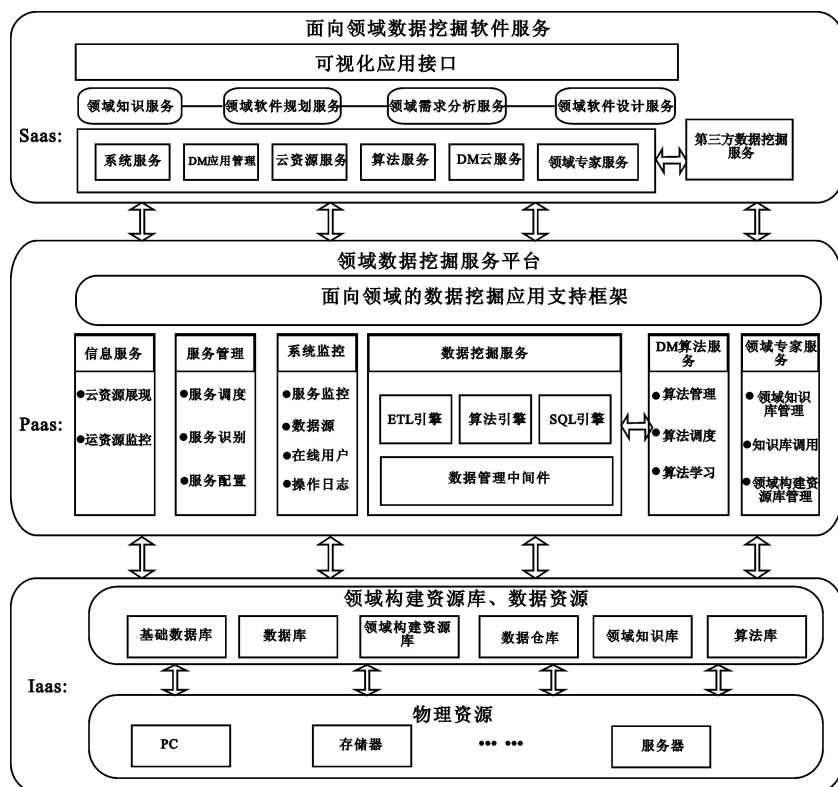


图2 面向领域数据挖掘云平台应用架构

考虑到面向领域数据挖掘云平台的全面建设和有效运行,面向领域数据挖掘云平台应用架构模型由4部分组成:物理资源、领域构件资源库和数据资源、领域数据挖掘服务平台、面向领域的数据挖掘软件服务。在这些层次中,自下而上依次满足物理资源整合、数据和领域构件整合、数据挖掘服务整合和面向领域数据挖掘的整合需要,向用户提供领域数据挖掘服务。

3.1 面向领域的数据挖掘软件服务

面向领域的数据挖掘应用服务是一个SAAS服务平台,在这个服务平台中向用户提供云服务、数据挖掘应用管理、数据挖掘云服务以及针对专家的领域专家服务。当然也可结合第三方数据挖掘服务(如Weka数据挖掘算法)给用户提供服务。该应用服务平台主要包括面向用户的可视化应用接口层、面向领域的数据挖掘过程服务和功能服务。

(1)面向用户的可视化应用接口:提供用户面向领域的数据挖掘可视化应用接口,包括ETL(数据的抽取-转换-加载)、数据挖掘模型、算法演示以及云计算平台各服务展示。

(2)面向领域的数据挖掘过程服务:在该服务过程中,把领域工程和基于数据挖掘的应用工程相结合。从领域工程角度出发,构造能够在应用工程中广泛使用的可复用的领域构件资源。因此,从领域构件的有效利用角度出发,将面向领域的数据挖掘过程服务分解成四个主要过程:领域知识服务、领域软件规划、领域需求分析服务和领域软件设计服务。基于这四个过程,可以把领域工程产生的领域软件构件应用到数据挖掘应用系统中,同时将这些领域构件应用结果通过云平台服务机制反馈到领域构件资源库,使领域构件资源库不断地更新,更好地服务于数据挖掘用户。

(3)功能服务:功能服务主要从云计算平台、数据挖掘云服务以及领域构件资源的角度出发,给系统提供一系列的服务。如:系统服务、DM应用管理服务、云资源服务、算法服务、DM云服务、领域专家服务等。

3.2 面向领域数据挖掘服务平台

该层在云服务的基础上为用户提供数据挖掘核心服务。不仅具备云计算的各种优势,而且为面向领域的数据挖掘提供通用的、共性的和基础性的服务,从而保证面向领域数据挖掘的灵活运用性、可扩展性和可持续性,这样才能更好地为用户提供服务。大致包括信息服务、服务管理、系统监控、数据挖掘服

务、DM算法服务、领域服务等。在PAAS服务中,不仅实现了给用户提供领域数据挖掘相关服务、云资源管理服务以及服务调度和系统监控,还为可复用领域构件资源库的开发提供技术支撑。

(1)数据挖掘服务:该服务功能中包括面向领域数据挖掘的核心功能。在该服务中提供ETL引擎、算法引擎和SQL引擎以及数据管理中间件。该服务功能实现了数据挖掘中的强大的数据处理功能,它的基础性、通用性必须毋庸置疑。

(2)DM算法服务:众所周知数据挖掘算法是整个数据挖掘的核心思想,没有算法服务的数据挖掘系统也就不能更好为用户提供决策和支持。正因为如此在DM算法服务中提供算法学习、算法调度、算法管理等服务功能,同时在数据层提供算法库,这样才能针对不同领域,提供适合该领域的ETL数据处理算法、数据挖掘算法等。

(3)领域服务:在该服务功能中包括对领域知识库的管理、调度以及领域构件资源库的相关操作功能。

(4)面向领域的数据挖掘应用支持框架:可作为数据挖掘应用系统的参考原型。基于领域模型、领域驱动设计和领域架构,为数据挖掘系统所需的基础应用提供开发环境,为领域构件组装和数据交换提供标准接口,并根据用户的个性化需求,实现快速的构件开发与交付。

3.3 领域构件资源、数据资源

领域构件资源和数据资源为面向领域数据挖掘云平台不同层次上的服务和功能提供数据资源交换和共享、领域知识、算法知识、数据挖掘模型以及领域构件资源的相关整合等。通过虚拟化技术,把云环境下的各种分布式资源集中汇集,并封装成服务资源,提供给上层的服务和应用。

(1)数据库:集中存放面向自适应领域的服务构件所需的领域数据信息,为面向领域数据挖掘的运行提供强大的数据支撑;基础数据库用来存放领域数据挖掘服务平台所需的基础数据信息,如服务平台日志数据、信息服务数据、用户管理数据等。

(2)领域构件资源库:存放各种可复用的领域构件资源。由于针对不同领域的数据挖掘,且不同领域具有不同的规则,所以在具体的数据处理过程、处理算法等服务中需要根据不同的领域规则来处理。根据领域工程和应用工程把这些过程开发或封装成可复用的领域构件。这些领域构件资源不仅为面向领域数据挖掘服务提供有利的支撑,而其增加了系统

的灵活有效性.

(3)领域知识库:是系统最基础的服务标准.结合领域专家知识,并针对不同领域的标准、结构标准、信息标准等,为领域工程和面向领域的数据挖掘相关服务提供统一的标准规范.

(4)算法库:存放不同领域的算法,为数据挖掘过程中的算法服务提供强大的算法资源.如决策树算法、关联规则算法、分类算法、聚类算法等.

3.4 物理资源

物理资源即基础设施层,为面向领域的数据挖掘云平台服务提供所需的计算存储资源.利用相应接口,把分布式的软硬件资源通过云计算和网络技术,实现资源的互联和共享,同时为虚拟化数据资源过程提供服务.

4 面向领域的数据挖掘云平台服务开发模式和相关实现

基于云计算环境下的面向领域数据挖掘应用就是利用云计算服务和可复用的领域构件为用户提供更多的数据挖掘服务.数据挖掘应用包括运行数据挖掘算法、领域构件、数据挖掘工具、领域知识库、数据仓库等,这些数据挖掘应用都在云计算环境中被实现成灵活的、可扩展的以及分布式的服务资源.结合传统的数据挖掘模式和服务流程,研究了多个数据挖掘应用的设计和开发过程,提出了如图3所示的面向领域的数据挖掘云平台开发模式.

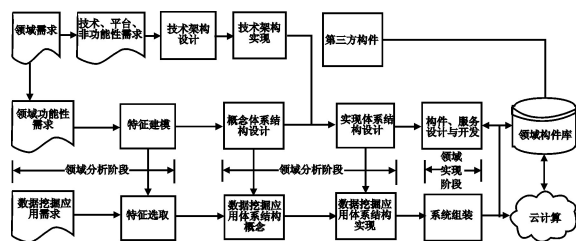


图3 面向领域的数据挖掘云平台开发模式

图3所示的开发模式中按领域工程的分析、设计、实现的流程逐步结合应用工程,最后与云计算技术融合构造数据挖掘应用系统.在具体的实现中,面向领域的数据挖掘云平台开发模式可以归纳为由图4所示的三个主要服务开发过程构成.

4.1 面向领域的数据挖掘分析与建模

领域分析是指基于领域中若干典型成员系统的需求分析,结合预期的需求变化、技术演化和约束条件等因素,识别领域的共性和可变特征,获取一组可

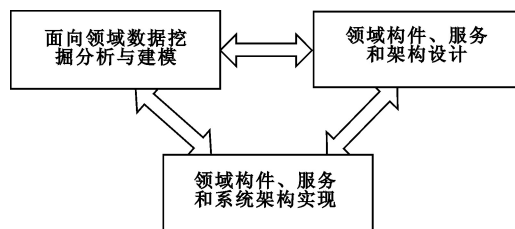


图4 服务开发过程

复用性高的领域需求,并将其抽象形成领域模型.而在面向领域的数据挖掘云平台中,我们首先从领域专家和领域知识库中获取相应的领域知识,对特定领域的数据挖掘进行分析,提取领域内的可复用需求,并形成领域服务规约.在分析基础上为面向领域的数据挖掘建立静态、动态等领域需求模型.

4.2 领域构件、服务和架构设计

根据需求模型和服务规约,在云计算环境中为面向领域的数据挖掘系统设计可复用领域构件和面向特定领域的服务,并设计出基于特定领域的数据挖掘系统架构.这些抽象的、灵活的、可复用的领域构件和服务进一步提高开发的效率,降低系统的复杂性,同时也增强了系统的可扩展性、可重构性和可维护性.

4.3 领域构件、服务和系统架构实现

在这个阶段中,结合云计算环境特性和现有的开发技术,实现面向领域数据挖掘的可复用构件和支撑DM系统以及云计算平台服务.这些构件和服务,在领域规约和应用工程的基础上进行高内聚、低耦合的功能集成.当然在云计算中我们要充分考虑到并行计算、分布式调用等问题,如MapReduce并行计算编程模型,它可屏蔽分布式处理的诸多问题.

通过这三个过程,最终可以实现面向领域的数据挖掘云平台应用原型系统的开发.理论上,这3个主要阶段应该逐一进行,但在实际应用开发中,它是一个反复迭代、逐步求精的过程,也就是后一阶段很可能会返回到前一阶段,进行修改和完善前一阶段的结果.

5 结束语

本文在云平台服务架构基础上,把领域工程技术、应用工程技术和数据挖掘技术结合在一起,提出了面向领域的数据挖掘云平台体系结构,定义了面向领域的数据挖掘云平台应用开发模式^[15-16].它对面向领域的数据挖掘、云计算环境下的数据挖掘服务模式以及数据挖掘服务的开发和应用都有一定的参考价值.在下一步研究中,进一步完善系统架构,

使其能服务更多的领域,同时在云计算平台下,针对特定领域,实现相关的数据挖掘服务,从而进一步验证本文提出的面向领域的数据挖掘云平台应用。

参考文献:

- [1] Wu Xindong, Zhu Xingquan, Wu Gongqing, et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(1):97-107.
- [2] 丁静,杨善林,罗贺,等. 云计算环境下的数据挖掘服务模式[J]. 计算机科学, 2012, 39(6):217-219.
- [3] Zhang Xuyun, Liu Chang, Nepal Surya, et al. SaC-FRAPP: a scalable and cost-effective framework for privacy preservation over big data on cloud[J]. Concurrency and Computation-Practice & Experience, 2013, 25(18):2561-2576.
- [4] Dinh Hoang T, Lee Chonho, Niyato Dusit, et al. A survey of mobile cloud computing: architecture, applications, and approaches[J]. Wireless Communications & Mobile Computing, 2013, 13(18), 1587-1611.
- [5] 于戈,谷峪,鲍玉斌,等. 云计算环境下的大规模图数据处理技术[J]. 计算机学报, 2011, 34(10):1753-1767.
- [6] 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):47-56.
- [7] Chunye Gong, Jie Liu, Qiang Zhang, et al. ,The characteristics of cloud computing[C]// 2010 39th International Conference on Parallel Processing Workshops (ICPPW). Italy, Bari, 2010:275-279.
- [8] 李克勤,陈兆良,梅宏,等. 领域工程概述[J]. 计算机科学, 1999, 26(5):21-25.
- [9] 钱琳,秦亮曦. 基于云计算的数据流系综分类算法研究[J]. 微电子学与计算, 2012, 29(2):99-102.
- [10] Prieto-Diaz R. Domain analysis for reusability[C] // Proceedings of the COMPSAX'87. Tokyo: IEEE Computer Society, 1987:23-29.
- [11] 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):99-102.
- [12] 张伟,梅宏. 一种面向特征的领域模型及其建模过程[J]. 软件学报, 2003, 14(8):1345-1356.
- [13] 杨芙清,梅宏,李克勤. 软件复用与软件构件技术[J]. 电子学报, 1999, 27(2):68-75.
- [14] 张强. 云计算时代的软件复用[J]. 电脑知识与技术, 2008, 4(8): 2199-2209.
- [15] Chang Mingfeng, Hsu Minhsiung, Chen Kueihui. An introduction to cloud computing service platform-google app engine[J]. Compute and communication, 2008(12):24-33.
- [16] Tang Shaojie, Yuan Jing, Wang Cheng, et al. A framework for amazon EC2 bidding strategy under SLA constraints[J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 25(1):2-11.

作者简介:

曾安女,(1978-),博士,副教授.研究方向为智能信息处理、数据挖掘和生物信息学。

陈云男,(1990-),硕士研究生.研究方向为数据挖掘。

潘丹男,(1975-),博士,高级工程师.研究方向为机器学习。

杨海东男,(1977-),工程师.研究方向为数据挖掘。

谢光强男,(1979-),博士,副教授.研究方向为智能控制。

(上接第14页)

李海启男,(1987-),硕士.研究方向为混合信号 CMOS 集成电路设计。

宋焱男,(1987-),博士.研究方向为混合信号 CMOS 集成电路设计。

范世全男,(1987-),博士.研究方向为模拟/混合信号电路设计、电源管理电路及系统设计。

耿莉男,(1987-),教授,博士生导师.研究方向为数模/射频混合集成电路设计及新型半导体器件。