基于 Hadoop 平台的并行数据挖掘算法工具箱与数据挖掘云

来源：南京大学计算机科学与技术系　**作者**：高 阳，杨育彬，商 琳　**时间**：2011-06-27　**浏览次数**：60

一 基于云计算的海量数据挖掘

2008 年 7 月，《Communications of the ACM》杂志发表了关于云计算的专辑，云计算因其清晰的商业模式而受到广泛关注，并得到工业和学术界的普遍认可。目前工业界推出的云计算平台有 Amazon 公司的 EC2 和 S3，Google 公司的 Google Apps Engine, IBM 公司的 Blue Cloud，Microsoft 公司的 Windows Azure, Salesforce 公司的 Sales Force, VMware 公司的 vCloud，Apache 软件开源组织的 Hadoop 等。在国内，IBM 与无锡市共建了云计算中心，中石化集团成功应用 IBM 的云计算方案建立起一个企业云计算平台。阿里巴巴集团于 2009 年初在南京建立电子商务云计算中心。

严格的讲，云计算是一种新颖的商业计算模型，它可以将计算任务分布在大量互连的计算机上，使各种应用系统能够根据需要获取计算资源、存储资源和其他服务资源。Google 公司的云平台是最具代表性的云计算技术之一，包括四个方面的主要技术：Google 文件系统 GFS、并行计算模型 MapReduce、结构化数据表 BigTable 和分布式的锁管理 Chubby。基于以上技术，云计算可以为海量数据处理和分析提供一种高效的计算平台。简单来说，将海量数据分解为相同大小、分布存储，然后采用 MapReduce 模型进行并行化编程，这种技术使 Google 公司在搜索引擎应用中得到了极大的成功。

然而 MapReduce 计算模型适合结构一致的海量数据，且要求计算简单。对于大量的数据密集型应用（如数据挖掘任务），往往涉及到数据降维、程序迭代、

近似求解等等复杂的算法，计算非常困难。因此，基于云计算的海量数据挖掘技术成为了工业界和学术界共同关心的热点技术之一。

分布式计算是解决海量数据挖掘任务，提高海量数据挖掘效率的方法之一。目前，分布式数据挖掘技术主要有基于主体（agent）的分布式数据挖掘、基于网格的分布式数据挖掘、基于云的分布式数据挖掘等。海量数据挖掘另一个核心问题是数据挖掘算法的并行化。图 1 给出基于云计算的海量数据挖掘服务的层次结构图。

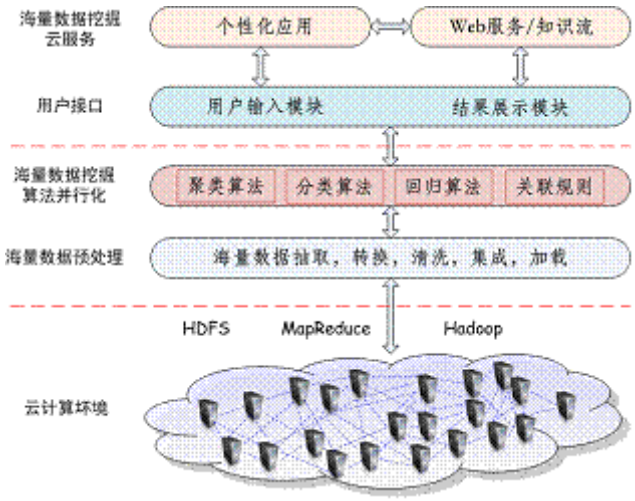

图 1　基于云计算的海量数据挖掘服务的层次结构图

中国移动研究院从 2007 年 3 月份启动“大云”的研发工作。2008 年，中国移动研究院已建设有 256 个节点、1024 个 CPU、256TB 存储的云平台。中国移动“大云”平台主要为数据挖掘、系统评估、搜索等应用提供计算服务。在开源 Hadoop 云平台上，中科院计算所研制了并行数据挖掘工具平台 PDMiner。针对海量数据，云计算分别从数据挖掘模式和方法等方面进行相关的研究。与此同时，中科院深圳先进研究院还研制了一个分布式数据挖掘系统 AlphaMiner。

本文首先讨论了海量数据挖掘的研究热点；其次基于开放的 Hadoop 平台，讨论并行数据挖掘算法工具箱和数据挖掘云的设计。

二 技术热点

云计算是一种资源利用模式，它能以简便的途径和以按需的方式通过网络访问可配置的计算资源，快速部署资源。在这种模式中，应用、数据和资源以服务的方式通过网络提供给用户使用。大量的计算资源组成资源池，用于动态创建高度虚拟化的资源以供用户使用。但对于海量数据分析任务，云平台缺乏针对海量数据挖掘和分析算法的并行化实现。因此面向海量数据挖掘的新型云计算模式，主要包括海量数据预处理、适合于云计算的海量数据挖掘并行算法、新型海量数据挖掘方法和云计算数据挖掘工具箱等技术。

（1）海量数据预处理。为了适合并行处理，云平台应可以提供海量数据的概念分层组织以及海量数据的并行加载；并实现高维度约减和数据稀疏化技术，提高数据管理和挖掘的效率。

（2）适合于云计算的海量数据挖掘并行算法。海量数据挖掘的关键问题是数据挖掘算法的并行化。而云计算采用 MapReduce 等新型计算模型，这意味着现有的数据挖掘算法和并行化策略不能直接应用于云计算平台下进行海量数据挖掘，需要进行一定的改造。因此需要深入研究数据挖掘算法的并行化策略，继而实现高效的云计算并行海量数据挖掘算法。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。在此基础上，针对海量数据挖掘算法的特点对已有的云计算模型进行优化和扩充，使其更适用于海量数据挖掘。

（3）新型海量数据挖掘方法。新型海量数据挖掘方法包含面向同构数据、异构数据和跨域数据的不同的数据挖掘新方法。在同构海量数据挖掘系统中，各个节点存储的数据都具有相同的属性空间。云平台采用集成学习的方式来生成最终的全局预测模型。并在同构节点的元学习基础上，实现数据挖掘增量学习方法，已满足实时要求；在异构海量数据挖掘系统中，云平台根据数据模态，将数据节点分类，并提供异构数据相关性度量和集成机制。除此之外，由于数据挖掘应用的特殊性，云平台能提供对海量数据迁移挖掘方法的支撑，以便扩充云计算环境下数据挖掘应用的适用范围，更好地满足数据挖掘终端用户的需求。

（4）并行数据挖掘工具箱。海量数据挖掘应用系统开发前，都会对采用的算法进行性能的评估。目前已有的 Weka 工具箱采用的是单机算法，不能应用在基于云计算的海量数据挖掘应用中。Apache 组织近年来组织了 Mahout 开源项目，设计用于云平台的数据挖掘算法。但 Mahout 项目目前还缺少数据准备、数据展示和用户交互，还不完全适合海量数据挖掘并行算法的性能评估。因此，云平台应可以提供一个基于 MapReduce 计算模型的并行数据挖掘工具箱，用于海量数据挖掘并行算法的性能评估。

在网格计算研究中，国际研究者研发了多个基于网格的复杂数据分析任务的服务系统，如 Data Mining Grid、Grid Miner 等等。在这些系统中，实现了复杂数据分析任务的工作流定义、资源调度和管理的透明化、具体算法的注册和服务化等。以上部分技术可以直接迁移到云计算平台上，但由于云计算模式和数据挖掘服务的特殊性，仍需在按需服务、多任务调度和分配等技术上进行进一步的突破。具体技术内容包括：

（1）按需服务的自治计算模式。 将海量数据挖掘任务的服务化，设计并实现并行数据挖掘软件自配置、自优化、自修复和自保护的方法，以及自适应用户需求的数据挖掘服务的自动发现和组合算法。

（2）多任务的动态分配机制。海量数据挖掘应用往往是数据密集，且具有突发性的特点；除此之外，不同的数据挖掘应用对算法精度、性能要求也不一致。因此，基于云计算的海量数据挖掘必须优化负载调节的策略与任务迁移策略等。

（3）数据挖掘服务的动态按需迁移。云平台提供支持海量数据挖掘任务的服务重定位方法，即当一个服务器上运行中的服务按需迁移到另一个服务器上去时，能同时有效地为后继工作流任务提供可用的资源空间，并满足整合服务器资源的需要。在资源管理和配置中，针对海量数据的大规模和异构等特点，运用虚拟化技术进行存储管理，并设计一种新型的动态迁移架构。

（4）复杂数据挖掘任务服务平台。 在 Hadoop 等云平台上，设计支持复杂数据挖掘任务服务化的中间件系统。支持复杂数据分析任务的流定义、复杂数据分析任务的动态配置、并行算法的注册、云平台资源的调度和管理的透明化，最终实现复杂数据分析任务的按需服务。

三 基于 Hadoop 的并行数据挖掘算法工具箱——Dodo

Weka 是由新西兰 Waikato 大学研发的数据处理和知识发现软件包。其可以实现数据预处理、聚类、分类、回归、特征选择、可视化等各种数据挖掘的任务。Weka 被广泛用于各种数据挖掘任务中算法的评估。但其中数据挖掘算法的实现是基于单机实现的。与 Weka 不同的是，Apache 组织基于 Hadoop 平台的，采用 MapReduce 计算模型，实现大量机器学习算法的并行化，并将其封装在 Mahout 项目。但由于 Mahout 并不提供一种图形界面交互，用户需要大量手工配置数据和参数，同时目前实现的并行数据挖掘算法也不完全。因此有必须借鉴 Weka 和 Mahout 的优点，研发一个基于 Hadoop 的并行数据挖掘算法工具箱——Dodo。表 1 给出三个工具箱目前的主要异同点。

表 1　Weka, Mahout 和 Dodo 主要异同

	数据源	数据格式	数据存储	算法	用户界面
Weka	支持文本文件：包括本地的数据文件以及网络数据文件； 支持数据库文件：通过 JDBC 连接。	标准格式是 Arff,行表示实例，列表示各个属性。另外还支持 CSV，C45 以及 BSI。	数据文件加载存储于内存之中	在单机上实现分类、聚类、关联规则等数据挖掘算法	包括发现模式的表示，数据挖掘原语的操作，界面功能主要包括 4 个部分：Simple CLI、Explorer、Experimenter Knowledge Flow
Mahout	仅支持文本文件	每个算法自己根据算法的情况自己设定的文件格式	存储于 Hdfs 上	基于 MapReduce 计算模型，实现….	命令行交互
Dodo	支持文本文件、网络文件和数据库文件	支持 Arff 等通用标准格式，也支持顺序文件，文本文件等格式，并提供预处理	存储于 Hdfs 上	迭代和非迭代类数据挖掘算法的 MapReduce 化	数据管理：上传、删除、修改。 Hadoop 平台管理：启动、关闭。 算法管理：选择算法、修改算法参数。 任务提交。 任务进度显示。

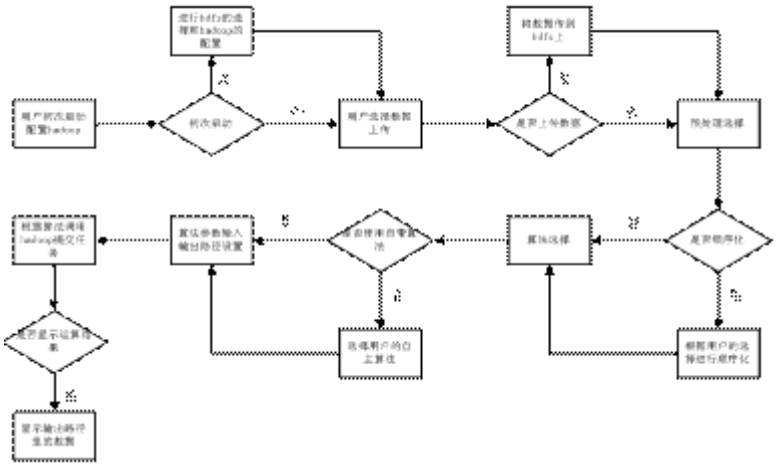

图 2　Dodo 工具箱运行流程

在图 2 中，如果用户是首次启动工具箱，需要选择连接的 Hadoop 环境并对环境进行配置；当用户需要上传数据，工具箱以树形图的形式，将用户的数据上传到指定的 Hadoop 路径上；如果不是顺序数据，工具箱则将其顺序化然后存储；在算法选择阶段，用户可以选择工具箱自带的并行化数据挖掘算法，也可以选择用户指定的、本地的 jar 文件；通过工具箱，用户能对选择的算法进行设置，其中包括输入输出路径，算法特定的参数等等；最后在 Hadoop 环境上对指定输入路径上的数据运行指定的算法，输出结果以可视化的方式展示给用户。

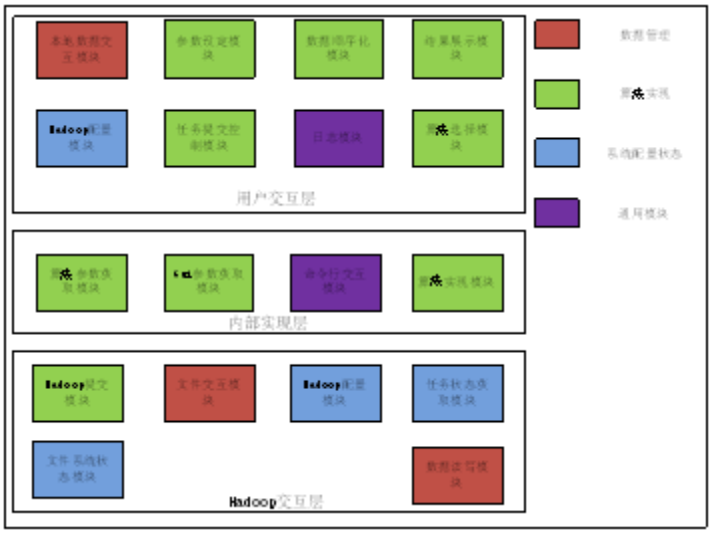

图 3　Dodo 工具箱模块结构图

图 3 中，将 Dodo 工具箱分为用户交互层、内部实现层以及 Hadoop 交互层等三个层次。用户交互层主要负责结果展示、算法选择等需要和用户进行交互的操作；内部实现层是 Dodo 的核心部分，负责与上层和下层进行交互，将一些操作进行抽象供两层进行调用；而 Hadoop 交互层主要是负责和 Hadoop 平台进行相应的操作，进行相关的平台配置或者数据上传或读写。

四 数据挖掘云

不同于其他的企业应用，将数据挖掘应用服务化，具备以下 4 个非常特殊的特点：

（1）简单化的工作流。数据挖掘应用从工作流角度来看，相对非常简单。应用中没有复杂的流程，也没有很多不同的角色。但数据挖掘应用仍然是一个工作流。因此将其服务化时，需要提供一个可视化的工作流编辑、管理界面，云平台也要提供对工作流引擎的监控。

（2）丰富的算法选择。不同于企业应用，在数据挖掘应用实现一个具体的挖掘任务有很多种算法。在很多情况下，每种算法的性能和效率都有可能不一样。

（3）结果的不确定性。数据挖掘任务中，选择不同的数据和算法，将有可能导致不同的计算结果。

（4）应用的突发性。很多的数据挖掘应用的请求会随着时间、空间呈现出突发性，这对资源提出了很高的“伸缩性”需求。

从以上特点可以看出，数据挖掘服务是一种真正的按需服务。用户可以根据自己的需求以及付费能力选择适合自己的服务模式。因此，所谓数据挖掘云是指在 hadoop 平台上提供支持复杂数据挖掘任务的服务系统，此系统能够提供复杂数据挖掘任务的工作流定义、资源调度、算法和工具以 web service 的方式向外提供服务。

数据挖掘云的结构如图 4 所示：

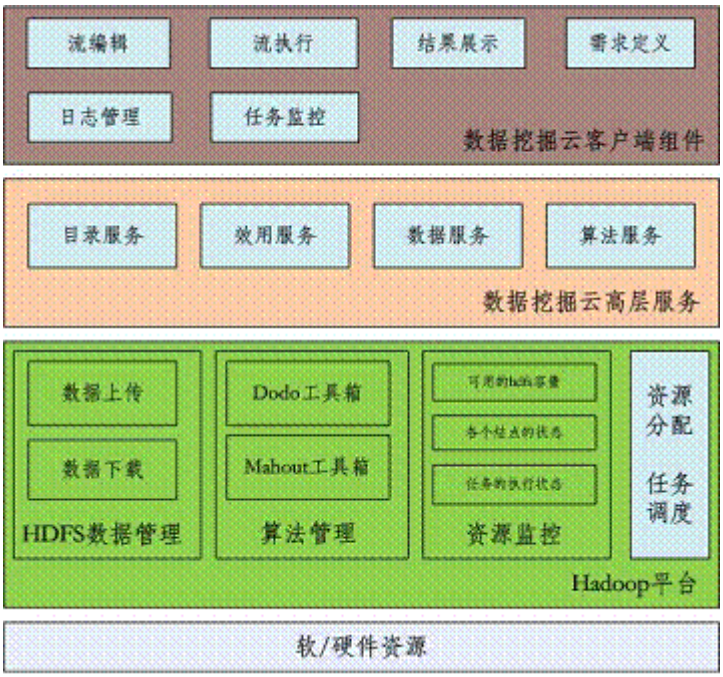

图 4 数据挖掘云

数据挖掘云的最底层是扩展云计算平台 Hadoop 的功能，实现 HDFS 数据管理、算法管理和资源监控，其中算法管理模块集成了各种基于 MapReduce 的工具箱，以向上提供算法服务。数据挖掘云的底层组件中，需要根据云服务的自适应需求，实现优化的资源分配和任务调度。数据挖掘云的中间层是数据挖掘云高层服务，包括目录服务、效用服务、数据服务和算法服务等核心组件。而最上层是客户端组件，主要用于与用户的直接交互。用户通过友好的可视化界面管理和监视任务的执行，并且很方便地查看任务执行结果。

在数据挖掘云的设计中，核心的组件有以下 6 个：

（1）目录服务：各种资源都能以目录的方式展示给用户，用户可以方便地展开目录查看所有可用的资源。

（2）资源分配和任务调度服务：把上层生成的执行计划映射到具体的计算资源和节点上，然后进行任务的调度和执行。

（3）数据访问服务：用户根据自己的任务，需要查找、上传或下载所需要的数据，数据访问服务为用户提供了良好的接口让用户方便进行这些操作。

（4）算法和应用访问服务：用户在编辑工作流的时候，需要查找满足需求的算法和应用，算法和应用服务提供了良好的接口让用户方便数据和应用的访问。

（5）流管理服务：流管理服务包括工作流的编辑和执行，以及用户对流的执行过程的监控和控制，并且在执行过程中会生成相应的日志。

（6）结果展示服务：任务执行完毕以后，用户需要查看任务的执行结果，结果展示可能包含多种方式，图状的、表格式的、文本式的等方式。

五 总结

综上所述，本文讨论了基于云计算的海量数据挖掘的进展和主要技术热点，并分析了基于 Hadoop 平台的数据挖掘算法工具箱和数据挖掘云的结构。Dodo 工具箱主要实现海量数据挖掘算法 MapReduce 化，以提高对海量数据的处理能力。在工具箱实现中，强调与 Hadoop 平台的交互式配置，迭代/非迭代类数据挖掘算法的并行化实现。在数据挖掘云服务中，为使海量数据挖掘应用服务化，提供从 Hadoop 资源分配到目录服务，再到流管理等一系列的组件服务，继而提高海量数据挖掘软件的服务能力。作为能为企业效益增值的数据挖掘应用，本质上具备了请求突发、需求多变，结果依赖于数据和算法的特点，因此必须进一步优化云计算平台，提高云平台对按需服务的支撑能力。