

基于云计算的 Web 数据挖掘

程 苗

(中国科学技术大学管理学院 合肥 230026)

摘 要 因特网是一个巨大的、分布广泛的信息服务中心,其上产生的海量数据通常是地理上分布、异构、动态的,复杂性也越来越高,若用已有的集中式数据挖掘方法则不能满足应用的要求。为了解决这些问题,提出了一种基于云计算的 Web 数据挖掘方法:将海量数据和挖掘任务分解到多台服务器上并行处理。采用 Hadoop 开源平台,建立一个基于 Apriori 算法的并行关联规则挖掘算法来验证了该系统的高效性。还提出“计算向存储迁移”的设计思想,将计算在数据存储节点就地执行,从而避免了大量数据在网络上的传递,不会占用大量带宽。

关键词 云计算,数据挖掘,Map/Reduce,关联规则

Web Data Mining Based on Cloud-computing

CHENG Miao

(School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract Internet is a huge and widely distributed information service center, the vast amounts of data generated on the Internet are usually geographically distributed, heterogeneous, dynamic and become more complex, it can not meet the requirements if we use the existing centralized data mining methods. To solve these problems, proposed a cloud computing-based Web data mining method, the massive data and mining tasks will be decomposed on multiple computers parallelly processed. We use open platform—Hadoop to establish a parallel association rules mining algorithm based on Apriori, and it tests and verifies the efficiency of system. This paper proposed a design thinking that “migrate the calculation to the store”, the calculation will be implemented on the local storage nodes, thus it can avoid the large amount of data transmission on the network, and will not take a lot of bandwidth.

Keywords Cloud-computing, Data mining, Map/Reduce, Association rules

1 概述

随着 Internet 技术的迅猛发展,互联网上的数据呈指数形式飞速增长,如何在这个全球最大的数据集中发现有用信息成为数据挖掘研究的热点。Web 数据挖掘是建立在对 Web 上海量数据分析的基础上,利用数据挖掘算法有效地收集、选择和存储所感兴趣的信息以及在日益增多的信息中发现新的概念和它们之间的关系,实现信息处理的自动化。这对企业获取有用可靠的外界信息,商业运作过程中收集、分析数据从而做出正确决策有着十分重要的意义。

Web 数据挖掘主要是以网络日志为研究对象,利用数据挖掘技术发现用户行为的潜在规律。目前,基于网络日志的用户行为模式研究已在网络安全、电子商务、远程教育等多个领域得到了广泛的应用,是当前的热点研究之一。网络日志文件中的数据主要包括 URL 请求、页面间链接的拓扑结构、注册用户特征等。采用关联规则分析,可获取用户页面访问行为间的关系;采用聚类分析,可将特征相似的用户或页面归并分组;采用分类分析,可对用户行为特征进行归类识别;采用频繁序列模式分析,可获取用户访问习惯。这些常用数据

挖掘方法获取的用户行为模式,解决了页面自动导航、页面重要性评价以及改进网站设计、提高网站运营效益等问题。由于因特网本身所具有的分布广泛、用户众多等特性,也使得其上所产生的数据是海量的、地理上分布的、异构的、动态的,这给现有的数据挖掘系统带来了难题:处理这些数据的复杂度很高,系统的计算能力很难达到要求。目前,Web 日志挖掘还有待研究的问题主要有两个:一是如何整合与处理分布式的 Web 日志;二是如何开发出高性能、可伸缩的分布并行的挖掘算法,保证挖掘的效率。

为了解决高性能计算问题,国内外学者提出了基于集群、基于 Agent 等的各种分布式并行数据挖掘平台,提高了数据挖掘系统的处理能力,但实现却相对复杂且只能针对特殊应用。之后, M Cannataro^[8]等人基于 Globus Toolkit 设计了一种分布式并行知识发现平台,该平台利用 Globus Toolkit 所提供的网格计算能力,解决了传统数据挖掘计算能力不足的问题。近几年的研究^[8,9]集中在基于 Globus Toolkit 平台并行数据挖掘算法的实现与改进方面。但网格计算缺少商业化实现,且 Globus Toolkit 是基于中间件技术,需要通过编程或安装设置来搭建底层架构,增加了系统实现的难度^[5]。

本文受博士点基金项目(200803580024),创新研究群体科学基金(70821001)资助。

程 苗(1986—),女,硕士生,主要研究方向为云计算、数据挖掘。

Web 数据挖掘处理的是海量数据,而且以指数级增长,同时所设计到的挖掘算法相当复杂,有的算法需要多次扫描数据库,当数据量增加时会增加扫描的代价;有的算法需要存储各序列的相关信息,当信息量很大时,会带来存储上的问题。因此,将云计算融入 Web 数据挖掘中将具有非常重要的现实意义,可以解决 Internet 上广域分布的海量数据挖掘问题。

2 Map/Reduce 编程模式

Map/Reduce 是一个用以进行大数据量计算的编程模型,同时也是一种高效的任务调度模型,它将一个任务分成很多更细粒度的子任务,这些子任务能够在空闲的处理节点之间调度,使得处理速度越快的节点处理越多的任务,从而避免处理速度慢的节点延长整个任务的完成时间。它将大型分布式计算表达为一个对数据键/值对集合进行串行化分布式操作,包括 Map(映射)和 Reduce(化简)两个阶段。Map 是一个分的过程,用于将输入数据结合拆分为大量的数据片段,并将每一个数据片段分配给一个计算机处理,达到分布式运算的效果,而 Reduce 则把分开的数据合到了一起,最后将汇总结果输出。Map/Reduce 的执行由两种不同类型的节点负责,Master 和 Worker。Worker 负责数据处理,Master 负责任务调度及不同节点之间的数据共享。执行一个 Map/Reduce 操作需要 5 个步骤:输入文件、将文件分割并分配给多个 worker 并行执行、本地写中间文件、合并中间文件、输出最终结果。具体流程如下^[12]:

① Map/Reduce 库将输入文件分成 16 到 64MB 的 M 份,并在集群的不同机器上执行程序的备份。

② Master 节点的程序负责找出空闲的 worker 节点并为它们分配子任务(M 个 Map 子任务和 R 个 Reduce 子任务)。

③ 被分配到 Map 子任务的 Worker 节点读入已经分割好的文件作为输入,经过处理后生成 key/value 对,并调用用户编写的 Map 函数,Map 函数的中间结果缓存在内存种并周期性地写入本地磁盘。

④ 这些中间数据通过分区函数分成 R 个区,并且将它们在本地的位置信息发送给 Master,然后再由 Master 将位置信息发送给执行 Reduce 子任务的节点。

⑤ 执行 Reduce 子任务的节点从 Master 获取子任务后,根据位置信息调用 map 工作节点所在的本地磁盘上的中间数据,并利用中间数据的 key 值进行排序,将具有相同键的对合并。

⑥ 执行 Reduce 子任务的节点遍历所有排序后的中间数据,并传递给用户定义的 reduce 函数。Reduce 函数的结果将被输出到一个最终的输出文件。

⑦ 当所有的 map 子任务和 reduce 子任务完成时,Master 节点将 R 份 Reduce 结果返回给用户程序,用户程序将这些数据合并得到最终结果。

3 基于云计算的 Web 数据挖掘系统设计与实现

3.1 概述

基于云计算的 Web 数据挖掘系统是在 Internet 上广域

分布的海量数据和计算资源的环境中发现数据模式和获取新的知识和规律。基于云计算的 Web 数据挖掘同传统 Web 数据挖掘的基本过程一致,分为数据预处理、数据挖掘、模式评价 3 个阶段,只是在数据的处理方式上有所不同,其区别有:借助 Hadoop 的 MapReduce 思想,1)在收集数据时,一改传统将所有数据、文件统一存储在数据仓库中的做法,将 Web 上广域分布的海量数据经过过滤、清洗、转换和合并,并转化为半结构化的 XML 文件后,保存到分布式文件系统中。同一文件都会复制副本并将其保存在不同的存储节点上,这样不仅可以解决传统 Web 数据挖掘中普遍存在的存储容量扩展和 I/O 操作问题,还可以有效地避免因机器故障而带来的数据丢失问题。2)在执行某一具体挖掘任务时,由任务主节点(Master)负责整个的控制工作,创建子节点的从属任务,然后交由 Web 上空闲的计算资源(ServiceNode)去处理,ServiceNode 将状态和完成的信息向 Master 汇报。最后再由 Master 负责将所有结果进行合并。

3.2 计算与存储整合

在 Internet 中,网络带宽是相对稀缺的资源。Map/Reduce 的 Map 在各节点进行操作,处理过程中一般没有数据的传输工作,只是在 Reduce 过程中需要向 Master 传送计算结果,对于 Web 数据挖掘这种数据密集型的计算任务,这种方法节省了大量的数据传输时间。由于网络传输速度远小于 CPU 计算速度,因此有人提出了以计算来换通信的编程策略。可以通过让输入数据保存在构成集群机器的本地磁盘上的方式来减少网络带宽的开销。我们可以将数据文件分成 64M 大小的块,在不同的机器上保存块的拷贝。由 Master 保存这些块的位置信息,并在保存相应输入数据块的设备上执行 Map 任务。这种方法使得大部分输入数据都是在本地机器读取的,并不占用网络带宽。

3.3 数据文件的备份

在设计云计算系统时,不但要考虑计算与存储的整合,还必须在节点失效时考虑计算和存储的迁移。一般的云计算系统(Hadoop)实现存储的迁移,但对计算和存储同时迁移则做得不好,实现计算迁移的基础是数据块必须采用副本策略,这样计算迁移时才能重新找到所要处理的数据。一般来看信息通过网络进行迁移是比较慢的,而计算的迁移可以由系统很快完成,在有副本策略的系统中,只需要找到副本所在地,将计算迁移过去就完成了存储和计算的迁移工作,所以效率非常高。

3.4 系统架构

在本文设计的基于云计算的 Web 数据挖掘系统(见图 1)中,节点分为 3 类。一类是主控节点(Master),在云中,Master 只有一个,负责调度与协调计算节点之间的工作进程;一类节点是算法存储节点,负责存储数据挖掘所需的算法;还有一类节点是服务节点(ServiceNode),负责存储分好块的 XML 文件以及执行由 Master 分配的任务,并把计算结果返回给 Master。相应地,基于云计算的 Web 数据挖掘系统分为 3 层:数据存储层、挖掘算法层和业务处理层。

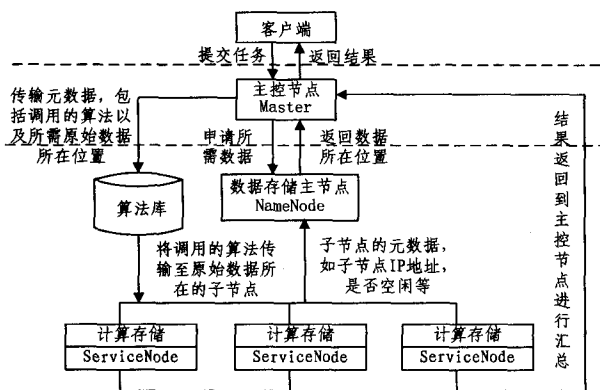


图1 基于云计算的 Web 数据挖掘系统架构

3.4.1 数据存储层

该层应具备的如下功能：

- ①能够将 Web 上收集到的文件，如 Web 日志文件等自动解析成半结构化 XML 文件，并装入分布式存储系统中；
- ②能够自动复制 XML 文件，复制的 XML 文件被随机地存储在一个 DataNode 上，防止因某个 DataNode 瘫痪而带来的数据丢失问题；
- ③能够长期存储包含用户使用信息、用户基本信息的文件；
- ④提供大量分布式数据集的访问接口；
- ⑤分布式文件系统中新的 DataNode 加入或有旧的 DataNode 删除时，能够自动更新。

分布式文件系统负责 XML 文件的存储和读取，它由一个主节点(NameNode)和多个子节点(DataNode)构成。在实际中，单个存储节点失效的情况是经常存在的，在系统设计时必须将不可信节点的失效屏蔽在系统之内，因此该文件系统使用副本复制存储策略来实现文件系统的高可靠性。本文将每个 XML 复制一个，分别存储在 2 个 DataNode 上。

如图 2 所示，NameNode 存储着每一个 XML 文件的元数据，这些元数据包括 XML 文件的 IP 地址等，通过该节点可以对存储在系统分布式文件系统的 XML 文件进行访问和处理。NameNode 还负责管理文件的存储等服务，但实际的数据并不存放在 NameNode 上。DataNode 用于实际数据的存放，对 DataNode 上数据的访问并不通过 NameNode，而是与用户直接建立数据通信。DataNode 每隔一段时间向 NameNode 发送一个信号，以证明该 DataNode 工作正常，没有出现故障。如果 NameNode 没有收到该信号，则表示 DataNode 出现故障，NameNode 则将保存在其他节点上的副本复制到另一个 DataNode 上，始终保持系统中每个 XML 文件都有 2 个，从而保证了系统的高可靠性。

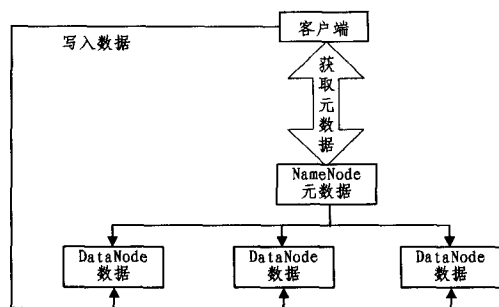


图2 文件存储系统的结构

用户保存 XML 文件的操作过程如下：首先向 NameNode 提交保存请求，NameNode 将 XML 文件分割为多个大小为 64M 的子文件，并查询元数据表找到空闲的 DataNode，然后将存储数据的 DataNode 的 IP 地址返回给用户，并通知其它接收副本的 DataNode，同时将文件的元数据(分成几个子文件、每个子文件存储在哪个 DataNode 上)写入元数据表中。用户根据结果直接与相应的 DataNode 建立连接，将子文件写入 DataNode 中。

3.4.2 挖掘算法层

该层(算法库)存储了用于数据挖掘的各种算法，这些算法都是基于传统挖掘算法改进后的适用于云计算平台的并行数据挖掘算法。在实际调用时，该节点首先从 Master 获取元数据(调用何种算法，执行该算法的节点的所在位置)，然后将相应算法传输到原始数据所在的节点上。本文实现的并行关联规则挖掘算法是基于 Apriori 算法改进的。

3.4.3 业务处理层

分布式数据挖掘子系统设计的核心是任务调度，所有挖掘器统一由 Master 负责调度，执行流程如下：ServiceNode 每隔一段时间向 Master 发送一个信号，以证明该 ServiceNode 工作正常。Master 将该 ServiceNode 放入空闲节点列表。Master 接收用户的业务申请，获得各数据块的存储信息以及所需调用的挖掘算法，然后向挖掘算法存储节点申请所需挖掘算法，算法节点直接将算法发送到原始数据所在的 ServiceNode 节点上，计算任务立即在文件存储服务器就地启动计算工作，完成后只向 Master 传送相关结果，并不向 Master 传送文件数据块，Master 汇总后生成最终的结果返回给用户。这一过程中没有了文件的传送和重组过程，计算和存储都在一个节点上面，节省了数据传输的时间。

3.5 基于云计算的 Web 数据挖掘算法

用于数据挖掘的算法种类繁多，例如关联规则、聚类、分类等，其中关联规则挖掘在 Web 日志分析、个性化信息推荐等诸多方面发挥着重要的作用，普遍应用于 Web 数据挖掘领域。关联规则挖掘分两步进行，第一步是找出所有的频繁项集；第二步是在频繁项集的基础上产生关联规则。为了找出所有的频繁项集，目前普遍采用迭代的方法，即：首先找出频繁 1-项集 L_1 ，接着找出频繁 2 项集 L_2 ，一直到某个 k 使得 L_k 为空，最终算法结束。当求 L_k 时，首先通过 L_{k-1} 的自连接生成候选项集 C_k ，然后检查 C_k 的每一个元素，满足用户自定义的最小支持度阈值的元素就是 L_k 的元素。显然，在 Web 这个广域数据源上验证 C_k 的元素是算法的一个瓶颈，会产生大量的候选项集合和重复扫描数据库。本文提出的基于云计算平台的 Apriori 算法将以上两项工作分配给“云”中多个计算节点 ServiceNode 并行处理，即各个计算节点 ServiceNode 分别求出各自局部频繁项集，再由 Master 统计出各频繁项集的全局支持合计数，并最终确定全局频繁项集，这可以大大提高 Apriori 算法的挖掘效率。

本文实现的 Web 并行数据挖掘算法是在传统 Apriori 算法上改进的，挖掘过程如下：

- ①用户通过 Web 浏览器提出数据挖掘服务请求，指定关联规则的最小支持度和最小置信度。

② Master 接收到挖掘请求后,向 NameNode 申请所需的 XML 数据文件,同时访问空闲节点列表,将 ServiceNode 的元数据(机器名、IP 地址是否空闲)返回到 Master。Master 将元数据发送给算法存储节点,算法存储节点将 Apriori 算法发送到原始数据所在节点。

③各 ServiceNode 首先扫描本地数据库,统计库中事务的个数、每个项的出现次数,然后根据挖掘流程和 Apriori 算法,得到局部的候选 1-项集,再把统计结果和局部候选 1-项集发送到 Master 计算得出全局 1-项集,然后再把全局频繁 1-项集发送到各个 ServiceNode 生成更精确的局部频繁 1-项集,再由局部 1-项集得出局部候选 2-项集,扫描本地数据库中的事务,统计每个项的出现次数,把新的局部候选 2-项集和统计结果发往 Master……如此重复,直到生成符合用户定义的满足最小支持度的频繁项集,最后根据置信度阈值生成规则。

④Master 将得到的关联规则返回给用户。

3.6 算法结果

该系统由 7 台服务器(均安装 Linux 以及 Hadoop 云计算系统)组成,其中 1 台作为客户端和主控节点,1 台作为算法存储节点,5 台作为服务节点 ServiceNode。在并行执行过程中,时间消耗主要在各节点之间建立连接以及数据的传输。首先,将所有数据放在主节点上直接调用 Aprior 算法,计算出执行时间;然后将数据集分割成 5 个子文件分别保存在 5 个 ServiceNode 上,将 Aprior 算法从算法存储节点上并行传到 1、3、5 个 ServiceNode 上执行,计算出时间;最后将 Aprior 算法分别拷贝到 5 个 ServiceNode 上,将数据文件传输到 1、3、5 个 ServiceNode 上执行,计算出时间。通过 3 个实验对比,可以发现执行效率随着数据量的增明显得到提高。同时,随着数据量的增加,向存储节点传输算法的时间也明显少于向算法节点传输数据。

本文基于云计算平台改进的 Aprior 算法,由于其对各个节点频繁项集的筛选都是在全局端进行的,因此既不会流失有效的关联规则,也不会产生无效的关联规则^[7]。

结束语 传统数据挖掘系统运行于 UNIX 小型机的集中平台上,这在海量数据以及应用愈加复杂的 Web 挖掘中受到

很多限制。与传统 Web 数据挖掘相比,基于云计算的 Web 数据挖掘系统通过“云”中多个资源完成原先由一个节点承担的挖掘工作,使资源得到了充分利用,提高了数据挖掘过程的效率。基于云计算的数据挖掘工作意义重大,它不仅能够提高挖掘效率,还克服了网格环境的弊端,能够面向商业应用,更具有价值。

参考文献

- [1] 李健,徐超,谭守标.一种 Web 数据挖掘系统的设计和研究[J]. 计算机技术与发展,2009,19(2)
- [2] 张涛. Web 数据挖掘现状分析[J]. 科学之友,2009,6(17)
- [3] 潘正高. Web 数据挖掘技术综述[J]. 电脑知识与技术,2009,5(15)
- [4] 席景科,闫大顺. Web 数据挖掘中数据集成问题的研究[J]. 计算机工程与设计,2006,8(27)
- [5] 纪俊. 一种基于云计算的数据挖掘平台架构设计与实现[D]. 青岛:青岛大学,2009
- [6] 郑晶. 基于网格的并行数据挖掘算法的实现[J]. 福建工程学院学报,2010,2(8)
- [7] 齐玉成,郑丽英,高三营. 基于网格的数据挖掘算法[J]. 电脑知识与技术
- [8] Cannataro M, Talia D, Trunfio P. KNOWLEDGE GRID: High Performance Knowledge Discovery on the Grid[C]// Lecture Notes In Computer Science, Vol. 2242, Proceedings of the Second International Workshop on Grid Computing. 2001:38-50
- [9] Ye Yan-bin, Chiang C-C. A Parallel Apriori Algorithm for Frequent Item sets Mining[C]// Proceedings of the Fourth International Conference on Software Engineering Research Management and Applications(SERA'06). 2006:87-94
- [10] Armbrust M, Fox A, Griffith R, et al. Above the Clouds: A Berkeley View of Cloud Computing
- [11] 万至臻. 基于 MapReduce 模型的并行计算平台的设计与实现[D]. 杭州:浙江大学,2008
- [12] 王鹏. 云计算的关键技术与应用实例
- [13] 郑庆华,刘均,田锋,等. Web 知识挖掘:理论、方法与应用[M]. 北京:科学出版社,2010

(上接第 135 页)

结束语 本文针对有野外数据采集需求的 e-Science 应用,提出了一种数据采集传输系统的设计思想。在此基础上构建了原型系统,并完成了相关软件的开发,实现了系统设计的基本功能。本文提出的系统,对于提高野外 e-Science 应用的工作效率、系统管理能力有重要的意义。下一步,将选择在青海湖国家级自然保护区、黑河流域等实际的野外科研环境中部署该系统,并根据野外应用的特点进行系统的优化和改进。

参考文献

- [1] 宋琳琳. E-Science 发展情况简介[J]. 图书馆学研究,2005(07): 21-23
- [2] Taylor J. The Definition of e-Science [OL]. <http://www.lesc.>

ic. ac. uk/admin/escience. html,2005-10-13

- [3] Hey T, Trefethen A E. Cyberinfrastructure for e-Science[J]. Science,2005,308(5723):817-821
- [4] ETSI GSM 02. 60: Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS)[S]. Service Description Stage 1. 1998
- [5] ETSI GSM 03. 60: Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS)[S]. Service Description Stage 2. 1998
- [6] IEEE 802. 15. 4. Standard-2003, Standard for Part 15. 4. Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANS)[S]. 2003
- [7] 孙利民,李建中,陈渝,等. 无线传感器网络[M]. 北京:清华大学出版社,2005