doi:10.3969/j.issn.1001 - 893x.2013.04.020

网络舆情的云计算监测模式分析与实现*

吴建军**

(浙江财经学院 现代教育技术中心,杭州 310018)

摘 要:针对传统网络舆情监测方式的不足,选取较为成熟的云计算架构,采用统一收集各局域网日志,并对海量数据使用云计算技术进行存储和分析的方法,提出了一种网络舆情云计算监测模式及其具体实现方法,并通过实验给出在设备配置、分析效率等方面有参考意义的实验数据。

关键词:广域网;局域网;舆情监测;云计算;大数据

中图分类号:TP393 **文献标志码**:A **文章编号**:1001 - 893X(2013)04 - 0476 - 06

Analysis and Implementation of Network Public Opinion Monitoring Based on Cloud Computing

WU Jian-jun

(Modern Education Technology Center, Zhejiang University of Finance & Economics, Hangzhou 310018, China)

Abstract: To avoid the shortcoming of traditional network public opinion monitoring mode, by selecting mature cloud computing framework, gathering all sub network logs unified, and getting the mass data stored and analyzed using the cloud computing technology, this paper proposes a network public opinion monitoring mode of cloud computing and its implementation. Experimental data on equipment allocation and efficiency analysis is given, which has reference significance.

Key words: WAN; LAN; network public opinion monitoring; cloud computing; big data

1 引 言

随着国民素质的不断提高,网民的社会责任感和政治参与热情也日渐增强,他们往往对社会事件有着较高的敏感性和参与度,因而把握网络舆情对于控制社会情绪、正确引导社会中坚力量有着极其重要的意义。现有的舆情监测技术大都存在监测盲点较多、准确率不高等问题,而随着云计算技术的发展,对较大网络范围内的大数据量进行获取和分析成为可能。通过对传统舆情监测技术的问题剖析,结合目前较为成熟的云计算技术架构,本文提出了网络舆情的云计算监测模式,并分析和给出了一种具体实现。该模式的核心是近两年兴起的大数据获取、存储及分析技术,将大数据技术用于舆情监测目前仍然是一个较新的应用研究领域。

2 网络舆情监测的现状和问题

舆情监测是对网络热点舆论在一定时间内发生的频率及趋势的监测和分析。随着网络和信息技术发展,网络舆情在监测方式方法、分析数据量等方面已经发生很大变化。

2.1 网络舆情的主要监测方法

舆情监测的要点是信息的采集和分析,按信息来源和采集方式的不同,网络舆情主要有下列主要监测方法。

* **收稿日期**:2013 - 02 - 07;**修回日期**:2013 - 04 - 15 Received date: 2013 - 02 - 07; Revised date: 2013 - 04 - 15
基金项目:浙江省杭州市哲学社会科学常规性规划课题研究项目(D12JY06)
Foundation Item: Supported by Philosophy and Social Sciences General Planning Project of Hangzhou, Zhejiang Province (D12JY06)
** **通讯作者**:sandrain@sina.com **Corresponding author**: sandrain@sina.com

(1)网页抓取和分析[1]

这是目前网络舆情最主流的监测方法,该方法通常采用网络爬虫类软件对互联网信息进行抓取、清洗和归并,并给出综合分析结果。信息源通常为论坛、博客、微博、贴吧等交友、互动类网站。对于信息源范围,也就是爬虫检索和抓取对象的确定,一种方式是通过搜索引擎得出[2],另一种是人工搜集的网站,两种方式各有优劣。

(2)日志分析

在大型网络和电信运营商的出口部位截取网络设备日志并加以分析,这是另一种常见的网络舆情分析方法。由于网络日志相当庞大,并且记录了流经网络出口的所有信息,信息内容杂乱,需要采用高性能、大容量设备和系统进行层层过滤和分析,才能获得和舆情相关的价值信息,因此时间和软硬件成本都较高,目前采用并不广泛。该方式最大的优点是对某段网络内产生的舆情信息能完全截获。

(3)特殊客户端及人工监测

将具备监测甚至控制功能的客户端安装在特定人群或场合内的上网计算机上,以达到对该类人群进行舆情监测甚至控制的目的。该方式监测面较窄,并且客户端的安装本身已经在心理上对上网者产生约束,不能体现上网者的真实心理情绪,因此管理和控制的色彩更浓,只在特殊情况下使用,类似的如 2008 年国家教育部面向青少年推广的“绿坝 – 花季护航”软件。传统的人工监测具有灵活、快速等优点,但面对浩如烟海的互联网,目前只作为舆情监测手段的补充在特殊情况下采用。

2.2　现有网络舆情监测模式的问题

通过对网络舆情主要监测方法的分析可以看到,相关网站日志分析和网页抓取等互联网手段的监测方法实施较为简便,但普遍存在信息来源不精确问题,无论是通过人工还是搜索引擎,都无法确定舆情的准确来源,在这种情况下,舆情的漏报和误报就几率较高,得出的监测结果事实上并不能完全表现舆情发展趋势,有时舆情可能会在监测系统所不熟知的网站中传播;在现有技术条件下,只能在大型网络和电信运营商的出口部位截取网络设备日志并加以监测才能较为准确地反应舆情信息,但是软硬件投资代价又太高,而且监测数据量的增长速率远远超出现有硬件处理能力的增长。

3　网络舆情云计算监测模式的提出

针对现有网络舆情监测模式的不足,业界迫切需要一种既能较准确监测舆情,又具有大数据处理能力、较大样本集合,具备一定普遍性,同时又有一定可操作性的舆情监测方案,在此思路指导下,本文提出一种新的网络舆情的云计算监测模式。舆情云计算并非是一个新名词,但以往提出的这个概念通常是指在舆情的分析阶段基于大数据技术,采集和存储阶段使用传统方式,并且深入进行理论和实践研究的学者也很少,而采集反而是舆情监测是否准确的重要环节。本文提出的模式将在舆情数据的采集、存储和分析各个环节采用成熟的云计算技术,是一套较为完整和具有新思路的舆情监测解决方案。

3.1　云计算监测模式的导出

网络舆情的监测对象是全体网民,对应的网络概念是广域网(Wide Aera Network,WAN)。广域网由众多局域网(Local Aera Network,LAN)组成,横向来看有多种主要的局域网,例如各大型企业局域网、各科研机构局域网、各级政府政务网、各学校校园网及各城区电信城域网等;而从纵向来看,很多局域网在自身体系内拥有相近的技术架构及行政管理机构,例如各级政府政务网、各学校校园网及各城区电信城域网。各局域网横向纵向结合,构成了广域网,云计算监测模式因此将重点放在各局域网的舆情监测和监测结果的整合,只要解决了这个关键问题,推广到全部局域网只是系统堆叠和行政管理机制的问题,这里将抛开行政管理许可问题而重点讨论其技术实现。

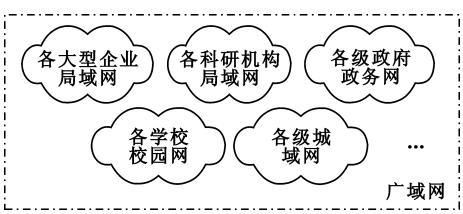

图 1　广域网中包含的主要局域网类型

Fig.1 The main types of LAN included in wide area network

3.2　云计算监测模式的基本架构

网络舆情云计算监测模式是在出口日志监测方式基础上的架构扩展。本文在架构上设计了对多个局域网网络出口数据的监测,这个设计较好地解决了监测网络单一、样本集合较小的问题,可以对某省甚至更大区域内的局域网进行数据的集中监测分析。为实现良好的扩展性、可用性,对整个架构模式提出了更高的要求,即要求日志数据分布式获取、海量存储及分布式计算分析,因此在监测中心引入了

云计算平台架构设计。

舆情云计算监测模式的基本架构如图 2 所示。

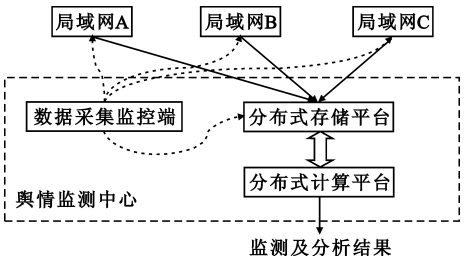

图 2 舆情云计算监测模式架构示意图

Fig.2 The architecture of network public opinion monitoring based on cloud computing

3.3 监测所采用的信息来源

监测所采用的信息来源是各局域网出口网络日志。舆情监测是政府主导的稳定企事业单位、稳定社会的行为,纵向看很多相同管理体系内的局域网所属相同行政管理部门,因此通过行政管理途径集中、统一获得区域内多局域网日志信息来监测网路舆情在信息来源上是可行的。

按照中华人民共和国公安部 2005 年颁布的《互联网安全保护技术措施规定(公安部令第 82 号)》,规模局域网必须提供网络日志记录功能。经过近几年的发展建设,具备一定规模的局域网网络出口都已经配备了网络日志记录和上网行为审计设备。网络日志为文本流格式并遵循一定的国际标准,是舆情监测相较为可靠的信息源[3]。上网行为审计系统近年来也发展较快,该设备能提供更多、更灵活的日志及内容审计信息,包括记录 web 访问、邮件、聊天等多种协议和行为,并可以根据需要调节审计粒度,但由于其审计内容较丰富,目前各厂家大多采用自定义格式保存日志,而另一方面国家公安等有关部门正在对行为审计设备制定相关标准,相信更丰富的审计日志在将来也会形成相对统一的数据格式,成为舆情监控更丰富的信息源。

3.4 海量分布式日志数据的获取和传输

局域网出口日志,在1 Gb/s出口链路,记录常规日志情况下,按经验值每天产生日志量约为5 GB,对于数万人中等规模局域网每日日志量约为10 GB,该数值在出口带宽充裕的大型网络中可能会达到上百 GB。为稳定、可靠地采集、传输海量日志,我们引入分布式、高可用的海量日志收集系统 Flume。Flume 支持在日志系统中定制各类数据发送方,用于收集数据,并对数据进行清洗、加密等处理,写入到定制的数据接收端。在局域网出口日志记录设备上进行配置,让日志数据流转存到网内服务器上,同时在服务器上安装 Flume 的 Agent 代理客户端,即可以将数据流分别传送到 Flume 日志收集器,实现分布式的数据收集。

3.5 模式所依据的云计算模型

处理几十 GB 数据,对于单台大中型服务器来说,效率已显不足,而当我们需要同时处理一个地区几十个甚至更多局域网的日志数据时,面对每天上百 GB 的数据规模,单台设备在存储和计算能力上已经完全失去扩展能力。针对海量数据存储和处理,我们引入 Apache Hadoop 即 HDFS(Hadoop Distributed Filesystem)分布式存储及 MapReduce 分布式计算模型[4]。

完整的舆情云计算监测架构如图 3 所示。

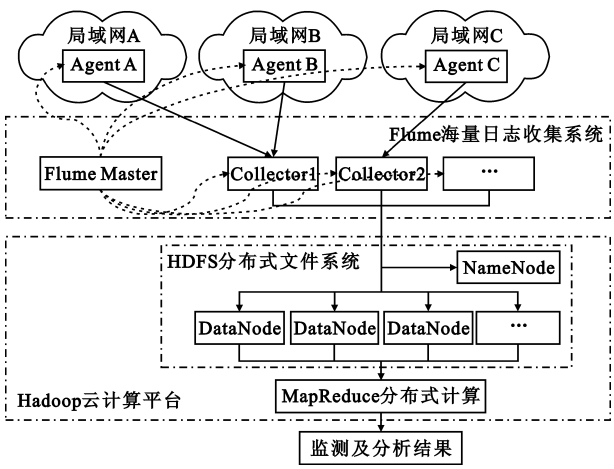

图 3 基于 Hadoop、Flume 的舆情云计算监测架构

Fig.3 The architecture of network public opinion monitoring based on Hadoop and Flume

模式采用业界主流开源 Hadoop 云计算架构,Flume 也是 Hadoop 生态图谱中非结构化数据收集的典型系统(Flume 的最新分支版本在架构上有所改变,但尚未稳定推广)[5]。整个日志数据处理过程可以描述如下。

受监测局域网出口处需配备行为审计等日志记录设备,并将日志数据引出到网内服务器上,该服务器预装 Flume 的代理(Agent),这些代理由舆情监测中心的 Flume 主控制器(Master)进行管理和配置,代理每 5 s 与 Master 进行通信一次交换管理信息。Flume 代理将日志数据进行格式转化、压缩、加密等预处理后,通过教科网、因特网等线路传输到同样受 Master 管理的日志收集器(Collector)集群内,收集器

根据接收监测目标数据量情况可以设置多个,以保证数据处理速度。最终所有日志由收集器集群并行写入 HDFS 分布式文件系统,写入时可以设置按照数据大小、行数或者间隔时间自动分割为多个文件。

HDFS 分布式文件系统对于日志数据这样一次写入不必更改的大文件是理想的存储架构,HDFS 主要由 NameNode 和 DataNode 组成。NameNode 是 HDFS 的管理者,提供数据存取的查询、写入和删除等管理操作,DataNode 是数据服务器集群,所有数据默认被切分成 64 MB,并复制 3 份分布存放在 DataNode 中。系统数据因此具有网络冗余功能,集群数据节点越多,节点同时损坏的几率越低,数据安全性则越高。分布式存储容量可以通过简单增加 DataNode 数据节点几乎无限制扩展。

MapReduce 是基于 HDFS 的分布式计算架构。他根据数据存放地就近进行计算作业,是典型的把计算带给数据的云计算架构。日志数据分布在 DataNode 中,Hadoop 会在包含指定日志文件数据块的多个数据节点中启动 MapReduce 计算,因为每个文件有 3 个副本,文件块分散度又大,因此能最大限度减少数据复制传输量。每个 DateNode 节点在 MapReduce 阶段可以变成计算节点,在自行编制的舆情分析算法导引下,经过 Map、Shuffle and Sort 及 Reduce 3 个步骤后形成演算结果保存在 HDFS 中。

监测中心编写的舆情监测分析程序可根据监测周期要求由 Hadoop 定期加载和运算,例如在每天夜间 HDFS 文件系统相对空闲时开始对过去一天收集到的日志数据进行统计分析,运算时间视分析的数据量和分析所包含的程序及代码数量而定,通常在数十分钟到数小时之间。

4　云计算监测模式的关键问题及实践

4.1　信息源的法律问题

舆情监测的信息源是局域网日志数据,与之相关的法律法规主要有《互联网安全保护技术措施规定(公安部令第 82 号)》、《计算机信息网络国际联网安全保护管理办法(公安部令第 33 号)》等,这几个法规主要从网络安全角度规范了网络建设、维护方记录上网信息的责任和义务,但并未在网络日志等信息的使用管理上做更细致的规定;从网络隐私权角度来看,我国的立法也相对欠缺,在实际运用当中则把网络隐私权部分作为隐私权并划归为名誉权进行保护,部分则归入一般财产侵权案件进行保护[6]。综合来看,在日志数据上进行商业和非商业统计分析尚没有相关法律的约束,前提是不利用数据泄露和追溯个人敏感信息,否则会陷于民事纠纷当中。但商业性质的统计分析从一定角度上来说无法保障数据的安全,因此网络舆情分析应该由非商业团体即政府相关部门或研究机构开展,其数据源的获取和分析才能得到政策和数据上的安全保障。

4.2　数据的传输和处理

一方面基于法律问题,另一方面也为降低数据传输量,对于日志数据不论在传输环节和处理环节都需要进行一定的技术过滤。我们建议在海量日志收集系统 Flume 的 Agent 在传输前就应该对数据进行初步的清洗和过滤,例如过滤明文登录、网上银行以及支付系统等访问信息,甚至邮件信息,过滤程度取决于各局域网管理者与舆情监测中心的合作和信任程度,毕竟很多个人信息对于舆情监测的统计分析还是具有一定价值的。

不同局域网所采用的日志记录设备不尽相同,并且不同审计深度其数据格式也不尽相同,虽然遵循一定的标准,但在日志格式细节上仍然会有差异,Flume Agent 在传输前也可以进行一些格式的清洗和转换,以尽量消除格式差异的困扰,并在传输时对数据进行压缩和加密。

4.3　舆情监测中心的建设

舆情监测中心应在政府相关部门或所属研究机构主导下进行建设。政府应与各局域网所属人达成较深层次的合作并听取和参考局域网信息化相关部门意见和建议,形成严格的数据获取、传输及处理等环节的规章制度和流程。数据的收集和传输应以尽量减少对局域网影响为原则,并充分考虑各网络信息化建设的实际情况,例如在东部较发达地区,大多数多出口链路企事业单位租用电信运营商出口链路作为网络主出口,而其他链路相对较为空闲,这时可以选择空闲链路传输数据。

监测中心的 Hadoop 集群应根据舆情监测的要求建设。典型地,当接入 20 个局域网时,我们预计日数据量约为 200 GB,年数据量约为 75 TB,按冗余 3 个副本计算,共需磁盘空间 225 TB,按两年建设容量配置为 450 TB。单台数据节点服务器按照 Hadoop 推荐配置比值:1 磁盘 + 2CPU 内核 + 6 ~ 8 GB 内存来配比,则可以测算出每服务器建议配置为:8 × 2 TB 硬盘 + 2 颗 8 核 CPU + 64 GB 内存,根据目标容量该 2U 机架式服务器共需 28 台。具体配置可根据

服务器参数、性价比和需要的总容量进行调整,数量则需根据计算复杂度和分析时间要求进行调整,如果一段时间后如系统容量不足,或希望提高计算效率,只需向集群添加服务器即可。舆情监测中心服务器及配置可按表 1 进行初步测算。

表 1 监测中心服务器配置测算表

Table 1 Server of the monitoring center configuration schedule

接入局域网数/个	预计日数据量/GB	预计年数据量/TB	需磁盘容量/TB	预计服务器台数/台
20	200	75	450	28
50	500	183	1 098	69
100	1 000	365	2 190	137

备注:服务器配置按 8×2 TB 硬盘 + 2 颗 8 核 CPU + 64 GB 内存计,所需磁盘容量按 3 个数据副本及 2 年存储需求计,实际配置和台数还要参考实际算法的计算量及内存消耗。

4.4 舆情监测报告

舆情监测固然是滞后于已经发生的舆论的,但仍具有一定的实时性,这取决于监测分析的间隔和效率。不同间隔的舆情报告其着重点是不同的,例如人民网舆情监测室按年发布《中国互联网舆情分析报告》,报告以年为单位分析中长期舆情的产生、发展和处理及平息的趋势,意在总结整个舆情周期的发展规律和处理经验,对于以月、周甚至日为周期的分析,重点在于观测短期舆情的爆发情况和趋势,以应对和预防为主。在我们的云计算监测模式中,利用云计算和云存储平台将日志数据作为宝贵的资源不断积累,既可以做按天为单位的短期应对和预测研究,也可以做长周期的经验总结研究。

由于数据量较大,同时为保持一定的实时性,舆情监测通常会按日计算。考虑在每天流量较小的午夜 12 点至早晨 6 点之间对前一日的累积数据做演算,根据不同监测指标和要求,基于同一批数据可能需要进行多次演算,典型的算法有每日网站按访问量排序,涉及词汇(话题)排序,搜索引擎关键字排序,用户活动频繁度按时间变化曲线等,每个算法耗时因程序效率、Hadoop 集群大小等因素而有显著不同[7]。

4.5 舆情云计算监测数据分析效率实验

为了对监测中心集群建设规模、监测报告出具的时间等方面的初步测算及规律提供参考,在实验室中进行了初步的模拟计算。实验环境如下:单机配置为 1 个 Intel 双核 CPU,2 GB 内存,1 TB SATA 硬盘,系统环境为 CentOS 6,Hadoop 0.20.2,Java 1.6.0。实验 1 以 500 MB 日志数据为分析对象,计算前 100 个访问量最大的网站并排序,考察集群在不同节点数量下的演算效率变化情况,实验结果如图 4 所示。

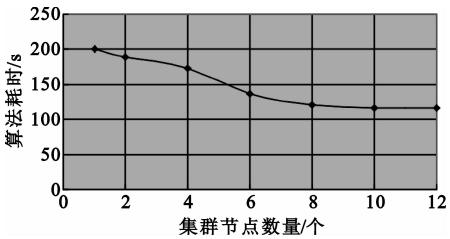

图 4 Hadoop 集群日志分析效率随节点数量变化情况

Fig.4 Log analysis efficiency varies with the number of nodes in Hadoop cluster

由该实验可以观察到 Hadoop 集群的日志分析效率并非简单随节点增加而线性增加,当集群计算量足够大时,继续增加节点数量已基本不能对集群效率产生较大影响,这时数据从磁盘存取的时间成为集群分析时间的重要组成部分,无论如何增加节点数量也无法超越和降低该基本时间。

实验 2 以不同大小的日志数据为分析对象,计算前 100 个访问量最大的网站并排序,考察集群在不同分析数据量情况下的演算效率变化情况,实验结果如图 5 所示。

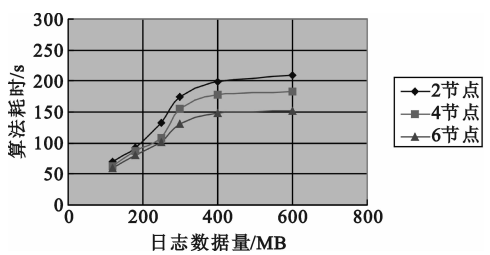

图 5 Hadoop 集群日志分析效率随数据量变化情况

Fig.5 Log analysis efficiency varies with the amount of data in Hadoop cluster

由实验 2 可以观察到,集群效率随着处理数据量增长其效率增长可能会降低,但节点数量越大,其效率下降相对平缓。这个实验给我们的启示是 Hadoop 集群处理效率与节点数量、处理数据量及单机配置都有关联,而且随着节点数量增大其增加的处理效能并不一定能被充分利用,但是 Hadoop 集群对未来可能无限增大的数据提供了处理的可行性[8]。

4.6 舆情云计算监测在校园网中的应用

选取本校校园网络日志数据进行了应用实践。

学校教学区校园网拥有电信、移动、联通及教科网 4 个出口,总出口带宽 1.5 GB,校园网全体师生用户约 1.5 万个。在实验室条件下,获取了 1 个月的校园网日志文件约 500 GB,编制了关键字(话题)每日排名、用户访问最多网站每日排名、用户每日活跃趋势等若干与网络舆情相关的统计分析程序。通过分析,可以清晰地观察到关键字(话题)每天的活跃度发展趋势,如果有较大量的历史数据积累,应该可以判断话题活跃到何种程度是为舆情发展的何种阶段,当然为避免片面性,需要多个局域网在较长的历史时间内的数据积累,样本数据越丰富,舆情监测越全面和准确。除此之外,还能观察到一些有意思的现象,例如教师用户在近中午时段较为活跃,而学生用户通常在下午 2~3 点到达活跃高峰,这些数据对于分析引导用户行为有很好的参考价值。

5　结束语

整合行政区域各局域网开展舆情研究,可以建立各省市舆情监测中心,如果将各中心数据进行贯通,则完全可以形成全国舆情监测系统,这种监测模式对象清晰,监测较为全面,且利用最新的云计算平台处理海量数据,较好地解决了现有网络舆情分析模式的诸多缺陷,是目前相对完整和彻底的网络舆情监测解决方案,值得深入研究和探讨。

另一方面,当数据积累到一定程度时,数据价值已远远不局限于舆情监测研究了。可以深入开展不同行业网络活动的分析研究,这对于了解、掌握当代网民从生活、学习习惯到思维、心理及世界观,以及这些情况与所在企事业单位及行业的规模、信息化程度等的关系,都具有很好的参考价值,同时对把握各行业的发展和趋势也具有较大的现实意义。

参考文献:

[1] 郝文江, 武捷. 互联网舆情监管与应对技术探究[J]. 信息网络安全,2012(3):1-4.
HAO Wen-jiang, WU Jie. Internet Public Opinion Supervision and Relevant Technical Research[J]. Netinfo Security, 2012(3):1-4. (in Chinese)

[2] 叶昭晖, 曾琼, 李强. 基于搜索引擎的网络舆情监控系统设计与实现[J]. 广西大学学报(自然科学版), 2011, 36(10):303-307.
YE Zhao-hui, ZENG Qiong, LI Qiang. Design and implementation of network monitoring and analyzing system of public opinion based on search engine[J]. Journal of Guangxi University (Natural Science Edition), 2011, 36(10): 303-307.(in Chinese)

[3] 张兵. 一种网络日志挖掘的高效算法[J]. 广西师范大学学报(自然科学版), 2006, 24(1):26-29.
ZHANG Bing. An Efficient Algorithm with Incremental Data Mining for Web Usage Mining[J]. Journal of Guangxi Normal University (Natural Science Edition), 2006, 24(1): 26-29. (in Chinese)

[4] 李建江, 崔健, 王聃, 等. MapReduce 并行编程模型研究综述[J]. 电子学报, 2011, 39 (11):2635-2641.
LI Jian-jiang, CUI Jian, WANG Dan, et al. Survey of MapReduce Parallel Programming Model[J]. Acta Electronica Sinica, 2011, 39 (11):2635-2641. (in Chinese)

[5] Cloudera, Inc. Flume User Guide [EB/OL]. 2012-08. http://archive.cloudera.com/cdh/3/flume/UserGuide/.

[6] 刘琳. 论网络隐私权保护及其完善[J] . 四川教育学院学报, 2012,28(7):48-49.
LIU Lin. On Internet Privacy Protection and Its Perfection [J]. Journal of Sichuan College of Education, 2012,28(7): 48-49. (in Chinese)

[7] 朱蔷蔷, 张桂芸, 刘文龙. 基于 MapReduce 框架一种文本挖掘算法的设计与实现[J]. 郑州大学学报(工学版), 2012, 33(5):110-113.
ZHU Qiang-qiang,ZHANG Gui-yun,LIU Wen-long. The Design and Implemention of a Text Mining Algorithm Based on MapReduce Framework[J]. Journal of Zhengzhou University (Engineering Science),2012,33(5):110-113.(in Chinese)

[8] 李彬,刘莉莉. 基于 MapReduce 的 Web 日志挖掘[J]. 计算机工程与应用, 2012, 48(22):95-98.
LI Bin, LIU Li-li. Weblog mining based on MapReduce[J]. Computer Engineering and Applications, 2012, 48(22):95-98. (in Chinese)

作者简介:

吴建军(1972—),男,浙江杭州人,1998 年获浙江大学理学学士学位,现为工程师,主要研究方向为大数据处理和信息安全。

WU Jian-jun was born in Hangzhou, Zhejiang Province, in 1972. He received the B. S. degree from Zhejiang University in 1998. He is now an engineer. His research concerns big data processing and information security.

Email:sandrain@sina.com