计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(5):80−85 [doi: 10.15888/j.cnki.csa.006341]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Hadoop 的生物质能源工程数据资源管理平台①

李海涛[1], 刘云生[1], 兰长杰[2]

[1](青岛科技大学 信息科学技术学院, 青岛 266061)
[2](青岛励图高科信息技术有限公司, 青岛 266000)

摘　要: 针对我国生物质能源工程信息化程度比较低的现状, 结合当今流行的大数据处理技术, 基于 Hadoop 开源框架设计并且实现了生物质能源工程数据资源管理平台. 介绍了平台的主要功能模块, 包括生物质能源工程的管理、监测指标的管理、实时监测、反欺诈模型的管理、统计分析等. 对平台建设中的关键技术包括数据获取、大数据存储、大数据处理、负载均衡等做了深入研究. 旨在将物联网技术, 互联网技术和大数据处理技术与生物质能源工程有机的结合起来, 提高生物质能源工程的信息化水平, 保障生产安全, 优化工艺流程, 实现效益最大化, 为同类工程的建设提供理论和实践依据.

关键词: 生物质能源; 大数据; Hadoop; 负载均衡

引用格式: 李海涛,刘云生,兰长杰.基于 Hadoop 的生物质能源工程数据资源管理平台.计算机系统应用,2018,27(5):80–85. http://www.c-s-a.org.cn/1003-3254/6341.html

Data Resource Management Platform of Biomass Energy Engineering Based on Hadoop

LI Hai-Tao[1], LIU Yun-Sheng[1], LAN Chang-Jie[2]

[1](School of Information Science &Technology, Qingdao University of Science &Technology, Qingdao 266061, China)
[2](Qingdao LiMap Information Technology Co. Ltd., Qingdao 266000, China)

Abstract: In view of the current situation of China's relatively low informatization level of biomass energy engineering, we combine today's popular big data processing technology, design and implement a biomass energy engineering data resource management platform based on Hadoop open source framework. This paper introduces the main function modules of the platform, including the management of biomass energy engineering, the management of monitoring indicators, real-time monitoring, the management of anti-fraud model, statistical analysis, etc. In-depth research on key technologies of platform construction has been carried out, including data acquisition, big data storage, big data processing, load balancing, and so on. Combining Internet of Things technology, Internet technology, and big data processing technology with biomass energy engineering organically, we can improve the informatization level of biomass energy engineering, ensure the production safety, optimize the manufacturing process, and maximize the benefits, thus provide theoretical and practical basis for the construction of similar projects.

Key words: biomass energy; big data; Hadoop; load balancing

随着我国的经济建设的快速发展, 能源消耗量持续攀升; 作为经济发展的重要因素, 传统能源存量有限, 由于使用效率不高, 浪费大, 传统能源日渐紧缺; 生物质能源作为清洁的可再生能源, 是理想的替代能源[1]. 我国的生物质能源工程也越来越多, 但是由于信息化水平较低, 管理起来费时费力.

① 收稿时间: 2017-08-31; 修改时间: 2017-09-20; 采用时间: 2017-09-25; csa 在线出版时间: 2018-04-23

针对我国生物质能源信息化水平较低的现状, 利用先进的物联网技术, 互联网技术, 大数据处理技术, 数据挖掘技术等, 构建了生物质能源工程数据资源管理平台. 提高了生物质能源工程的信息化水平, 便于了生物质能源工程的统一管理和运作, 节省了大量的人力物力, 提高生产效益.

1 系统概述

生物质能源工程数据资源管理平台借助物联网技术将生物质工程的生产数据实时上传到平台, 然后利用大数据处理技术帮助监管人员实时掌握生物质能源工程的运行情况. 提供了指标的工程监测, 实时指标监测, 视频监控, 统计分析及异常报警等服务.

生物质能源工程数据资源管理平台主要由数据采集层, 数据存储层, 数据处理层, 数据应用层四大部分组成. 系统总体架构示意图如图 1.

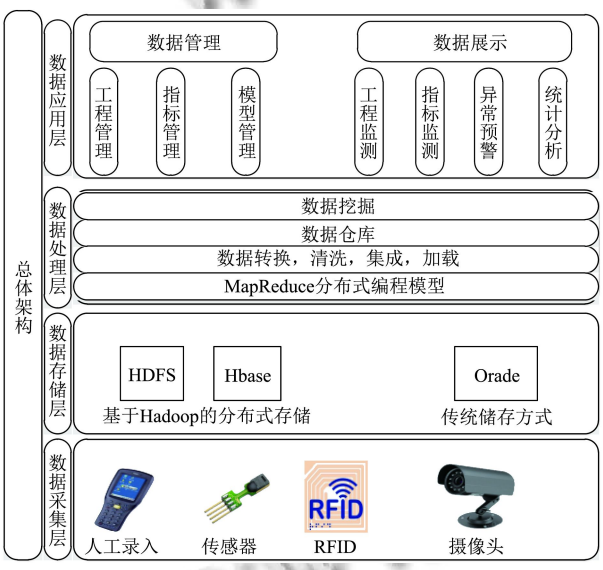

图 1 系统总体架构示意图

1.1 数据采集层

数据采集层主要主要包括 3 部分: 1) 传感器模块, 传感器模块是通过集成电路将温度传感器, 湿度传感器, 浓度传感器等组合起来, 通过 PLC 编程, 将监测数据通过互联网或者电信网络传送到接口服务器. 2) 人工数据的录入, 生物质能源工程的工人可以通过手机端和电脑端将生产过程中的一些数据进行填写, 并且提交到服务器[2]. 3) 高清网络摄像机, 将生物质能源工程的监控视频实时的传输到云平台服务器, 实现对工程的实时监控[3].

1.2 数据存储层

根据生物质能源工程数据的非结构化和结构化数据采用了关系型数据库和非关系型数据库两种数据库分别存储不同的数据. 对于工程的基本信息, 用户信息, 指标信息等采用传统的关系型数据库 Oracle 进行存储. 对于生物质能源工程产生的生产数据, 结合其数据量大的特点采用基于 Hadoop 平台的 Hbase 分布式数据库进行存储.

1.3 数据处理层

由于生物质能源工程数据比较复杂, 接口服务器将上传上来的指标信息进行解析, 异常分析, 然后进行存储. 由于生物质能源工程数据是海量增长的, 传统的关系型数据库无法满足数据的存储. 采用 Hbase 进行指标数据的存储. 对于上传上来的数据, 通过定时任务, 通过 MapReduce 进行反欺诈模型分析, 判断工程是否异常. 利用常用的数据挖掘算法对数据进行价值挖掘.

1.4 数据应用层

数据应用层主要包括数据管理和数据展示.

生物质能源工程的监管人员可以通过电脑或者手机等终端设备访问系统平台, 就可以随时随地进行数据的管理, 包括工程管理, 指标管理, 模型管理等. 也可以对工程的相关情况进行查看, 包括工程监测, 指标监测, 统计分析等.

2 系统设计与实现

2.1 软件系统总体设计

系统本着在达到预定目标, 具备所需功能的前提下. 遵循简单性原则, 灵活性和适应性原则, 一致性和完整性原则, 可靠性原则, 进行了系统的设计.

使用可以跨平台的 Java 语言进行开发, 采用开源框架 Spring+Spring MVC +Mybatis 进行[4]. 采用关系型数据库 Oracle 和基于列存储的 Hbase 数据库. Oracle 用来存储一些工程的基本信息, 监测指标的基本信息以及用户的基本信息包括登录账号和密码等. Hbase 用来存储生物质能源工程上传的生产数据包括温度, 压强, 以及各种气体的浓度等. 使用 Spring Tool Suite 作为开发工具.

系统功能架构示意图如图 2.

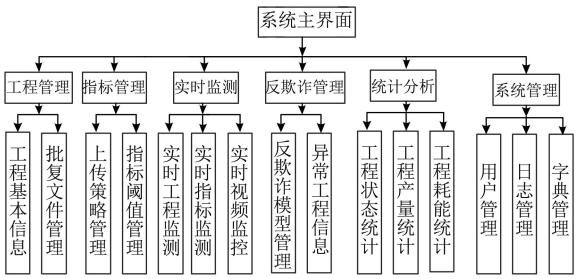

图 2 功能架构图

2.2 功能详细设计与实现

2.2.1 工程管理

工程管理是对监管区域的生物质能源工程的基本信息进行管理, 包括对生物质能源工程的增加, 编辑, 设置为监管, 地图定位等.

工程地理分布如图 3 所示.

图 3 工程地理分布图

2.2.2 指标管理

上传策略管理是针对不同的生物质能源工程所监测的指标不同, 因为生物质能源工程的规模不同, 信息化程度不同, 能够获取到指标数据也不同, 所以针对每一个沼气工程指定一套与之符合上传的策略. 上传策略包括上传的频率, 上传的单位, 自动上传还是手动填报进行上传等.

指标阈值的设定基于生物质能源工程的实践基础上制定的标准规范, 因原料工艺的不同而不同. 所以阈值的设定也要根据生物质能源工程的工艺进行针对性的设定.

2.2.3 实时监测

实时工程监测是从工程的维度进行数据的监测, 根据工程类别可以查看到该类别下的工程, 然后有每个工程的每个工段的指标正常和异常的数量. 通过详情可以查看每个指标的详细情况包括指标的上下阈值, 当前值等信息. 也可以通过图表的方式查看各个时间段的指标变化曲线. 实时工程监测效果如图 4 所示.

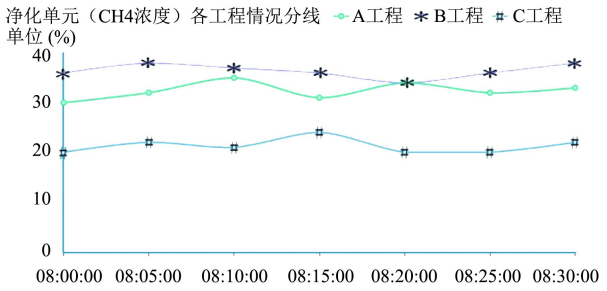

图 4 实时工程监测

实时指标监测是从指标的维度进行数据的监测, 根据监测指标, 可以查看各个工程下该监测指标的当前值, 并进行对比. 实时指标监测效果如图 5 所示.

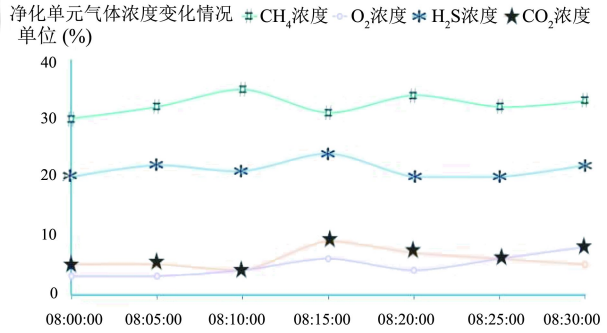

图 5 实时指标监测

实时视频监控可满足随时随地的对生物质能源工程的视频监控区域的状况进行了解, 对突发的异常情况作出快速的处理解决. 实时指标监测效果如图 6 所示.

图 6 实时视频监控

2.2.4 反欺诈管理

通过对生物质能源工程的长期观察分析, 可以发现各指标的变换范围都是有内在联系的, 不仅表现在严谨的的能量守恒上, 还表现在事物发展的一般规律上. 通过建立算法模型, 对生物质能源工程的数据进行

分析, 可以判断数据是否符合真实有效, 异常与否. 反欺诈模型管理主要包括对反欺诈模型的启动, 运算频率的设定等. 用到的反欺诈模型如表 1 所示.

表 1 反欺诈模型算法

模型名称	样本指标个数	样本时间选择	运行周期 (min)
产气压力相关性分析模型	≥150	近2天	15
产气差异性分析模型	≥150	近3天	15
产气离散型分析模型	≥150	近2天	15

异常工程信息管理可以查看被反欺诈模型处理过的异常工程, 可以查看反欺诈模型处理后的详细结果, 方面监管人员进行决策. 实现界面如图 7.

工程名称	模型名称	分析时间	分析结果	操作
A工程	产气-压力相关性分析模型	2017-05-06 08:25:00	异常	分析结果详情查看
B工程	产气-压力相关性分析模型	2017-05-06 08:25:00	异常	分析结果详情查看
C工程	产气-压力相关性分析模型	2017-05-06 08:25:00	异常	分析结果详情查看
A工程	产气差异性分析模型	2017-05-06 08:20:00	异常	分析结果详情查看
B工程	产气差异性分析模型	2017-05-06 08:20:00	异常	分析结果详情查看
C工程	产气差异性分析模型	2017-05-06 08:20:00	异常	分析结果详情查看
A工程	产气离散性分析模型	2017-05-06 08:15:00	异常	分析结果详情查看

图 7 异常工程信息

2.2.5 统计分析

通过对上传的工程生产数据进行处理和分析, 将分析结果通过图表的形式更加直观的展示给监管人员, 方便对工程运行状态的及时了解.

工程状态统计: 通过选择某个省或者市, 然后对该区域的生物质能源工程的状态进行统计, 便于决策人员因地制宜的进行生物质能源工程项目的建设和指导. 实现界面如图 8 所示.

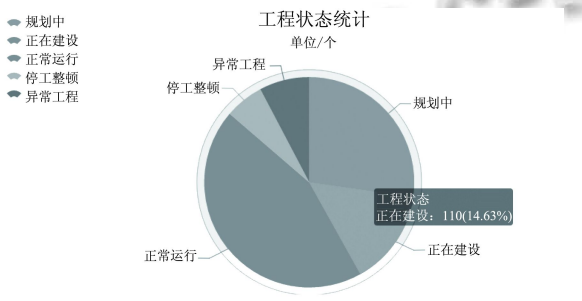

图 8 工程状态统计

工程产气量统计: 对于每个生物质能源工程的计划产气量和实际量产气量进行对比分析. 便于决策人员对于生物质能源工程项目的规模进行控制. 实现界面如图 9 所示.

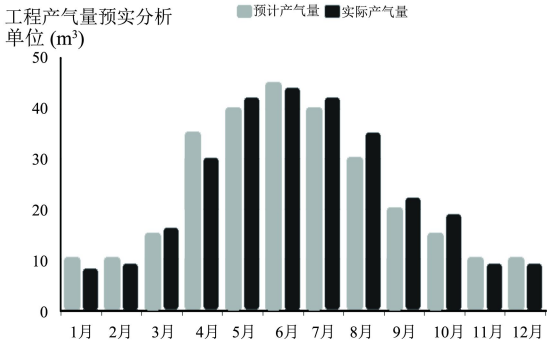

图 9 工程产气量预实分析

工程耗能情况统计: 通过对于生物质能源工程的数据指标进行大数据处理, 计算出生物质能源工程的耗电量, 耗水量, 催化剂量等. 将耗能量的统计结果进行显示, 方面了决策人员对于对物料消耗情况的了解, 便于开展下一步工作. 实现界面如图 10 所示.

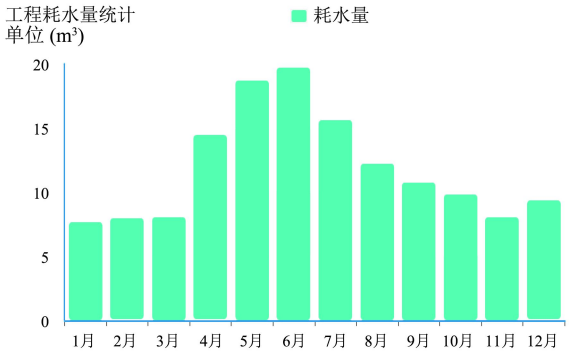

图 10 工程耗能情况统计

3 系统应用与测试

生物质能源工程数据资源管理平台设计开发完成后, 对系统进行的完善的系统性能测试, 业务测试. 下一阶段将逐步推广到各大生物质能源工程, 提高生物质能源工程的信息化水平, 统一监管.

4 系统关键技术

4.1 数据获取技术

数据采集采用智能数据盒子进行采集, 数据盒子是集传感器, 处理模块和通信模块为一体的, 采用 PLC 编程, 根据指定的指标上传策略对原始数据进行压缩加密处理, 通过 GSM 模块或者 WIFI 模块连接互联网, 将数据传递给接口服务器. 考虑到生物质能源工程的数量大, 数据上传频率高, 上传数据量大的特点, 对接口服务器采用负载均衡的技术.

4.2 大数据存储技术

生物质能源工程数量多, 而且每个生物质能源工程每天产生的数据也多达上万的级别, 面对如此海量的工程数据, 采用了面向列的分布式数据库 Hbase, Hbase 是依托 Hadoop 的 HDFS 作为最基本的存储单元, 因此可以解决随时读写和访问大数据集的难点, 这是普通的关系型数据库难以做到的[5].

Hbase 的服务体系结构遵循简单的主从服务器结构, 它由 HRegion 服务器群和 HMaster 服务器群组成[6]. HMaster 管理所有的 HRegion 服务器群, 它们由 Zookeeper 来进行协调, 并且处理 Hbase 服务器运行期间出现的各种问题, 保障的生物质能源工程数据的准确性, 一致性和完备性, 保证的数据安全[7].

Hbase 的优势在于接近线性水平的高度可扩展性, 因此随着生物质能源工程数据量的增加, 通过增加子节点就可以扩充其存储空间, 节省了购买高性能计算机的费用.

4.3 大数据处理技术

生物质能源工程数据满足大数据的 4V 特征定义: Volume(大量)、Velocity(高速)、Variety(多样)、Value(价值)[8].

规模大: 每个工程每天产生的数据就是 10 万条, 加上监控的视频每天也是 10 G 左右, 全国的生物质能源工程每年产生的数据量至少也是 PB 级, 并且还在指数型增长.

数据多样性: 生物质能源工程数据种类繁多, 包括生物质能源的基础信息, 上传的文件, 图片, 摄像头拍摄的视频, 传感器上传的非结构化数据等.

数据价值密度: 生物质能源工程数据规模大, 种类多, 但是有用的数据却很少. 如上传的很多指标数据, 可能也就哪个时间短的异常指标数据对监管人员来说是有价值的. 摄像头全天拍摄的视频, 可能也就发生突发安全情况的几分钟视频是有价值的.

高速性: 生物质能源工程的主要目的之一在于监管异常情况, 因此对异常情况的及时性要求比较高, 随时发现问题, 随时通知警报, 方便及时处理降低不必要的损失.

为了解决大数据处理的难点, 采用了分布式编程模型 MapReduce. MapReduce 是由 Google 公司研究提出的面向大规模数据处理的并行计算模型和方法. MapReduce 借鉴分而治之的思想, 将数据处理过程拆分为两步: Map(映射) 与 Reduce(化简). 第一步就是需要将数据抽象为键值对的形式, 接着 Map 函数的输入条件为抽象的键值对, 经过 Map 函数的运算处理后, 输出新键值对作为中间结果. MapReduce 计算框架自动将这些中间结果数据作聚合处理 (将键相同的进行归并处理), 并且会将键相同的数据分发给 Reduce 函数进行处理. 第二步就是 Reduce 函数以键和对应的的值的集合作为输入条件, 经过 Reduce 函数处理产生另外一系列键值对作为最后的输出结果[9].

用表达式表示如下:

{K1,V1}→{K2,List<V2>}→{K3,V3}

处理流程如图 11 所示.

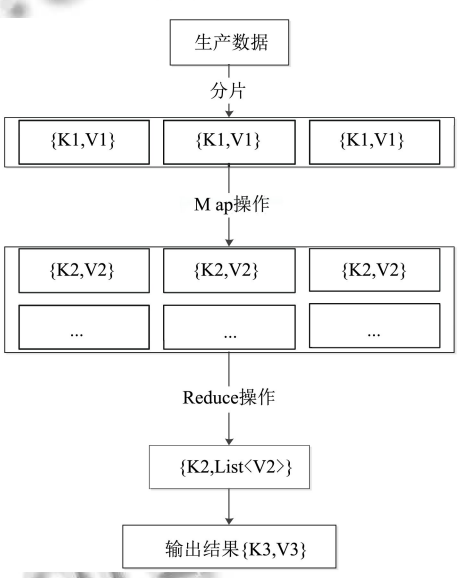

图 11 MapReduce 处理流程

4.4 负载均衡技术

针对生物质工程上传数据频繁的难题, 对接口服务器采用负载均衡技术保证系统的稳定. 负载均衡的算法有很多包括加权轮询, 源地址哈希法, 最小连接数法, 随机法, 加权随机法等[10].

根据接口服务器的业务逻辑, 选择 Nigix 的加权轮询算法进行负载均衡. 加权轮询算法分为深度优先搜索和广度优先搜索. Nigix 采用的是深度优先搜索算法, 首先是将请求都分给权重高的机器, 当该服务器的权重值降到比其他服务器低时, 才将请求分给下一个权重高的服务器; 第二, 当所有后端服务器都 down 掉时, Nigix 立即将所有服务器的标志位恢复初始状态, 避免全部的服务器因超时导致前端被夯住[11]. Nigix 轮询算法如图 12 所示.

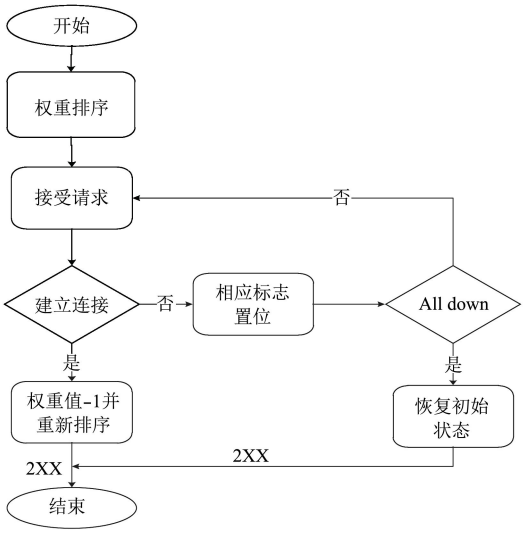

图 12 轮询算法处理流程

5 结语

生物质能源工程信息化是未来的发展趋势, 国内外很多研究机构和企业在生物质能源工程信息化方面也作出很多的探索, 借助于先进的物联网技术和大数据技术, 使生物质能源工程的信息化水平更上一个台阶. 有了大量的生物质能源工程数据资源后, 借助数据挖掘技术, 提炼出有价值的数据, 便于工程分析和决策的制定.

本文针对生物质能源工程数据资源管理平台的需求进行了设计和实现, 对设计与实现中遇到的难点, 提出了自己的解决方案, 包括数据的采集, 数据存储, 数据处理以及系统的稳定性保证.

通过对生物质能源工程数据资源管理平台的开发与研究, 实现了对生物质能源工程生产数据的采集, 监控, 预警, 保证了生物质能源工程的安全稳定运行; 通过生物质能源工程的生产数据的分析和挖掘, 为决策者提供工艺改良的依据, 减少消耗和环境污染, 提高生产效益; 通过对生物质能源工程的产气量统计分析, 进一步为政府部门的宏观调控提供理论依据. 深深的体会到生物质能源信息化是提高能源开发效率, 实现产业的可持续发展和提高市场竞争力的重要保障.

参考文献

1 杨艳华, 汤庆飞, 张立, 等. 生物质能作为新能源的应用现状分析. 重庆科技学院学报 (自然科学版), 2015, 17(1): 102–105.
2 李海涛, 王新安, 丰艳, 等. 智慧生态水产养殖系统. 计算机系统应用, 2017, 26(10): 73–76. [doi: 10.15888/j.cnki.csa.006036]
3 赵志军, 沈强, 唐晖, 等. 物联网架构和智能信息处理理论与关键技术. 计算机科学, 2011, 38(8): 1–8.
4 吕学婷. 基于 Springmvc 和 Mybatis 框架的门户网站及其内容管理系统的设计与实现[硕士学位论文]. 南昌: 东华理工大学, 2016.
5 陆婷. 基于 HBase 的交通流数据实时存储系统的设计与实现[硕士学位论文]. 北京: 北方工业大学, 2016.
6 White T. Hadoop 权威指南. 曾大聃, 周傲英, 译. 北京: 清华大学出版社, 2010.
7 瞿龙俊. 基于 HBase 的交通流数据实时存储与查询优化方案的设计与实现[硕士学位论文]. 镇江: 江苏大学, 2017.
8 郭雷风. 面向农业领域的大数据关键技术研究[博士学位论文]. 北京: 中国农业科学院, 2016.
9 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Proceedings of the 6th Symposium on Opearting Systems Design & Implementation. San Francisco, CA, USA. 2004. 137–150.
10 覃川. 基于 Nginx 的 Web 服务器负载均衡策略改进与实现[硕士学位论文]. 成都: 西南交通大学, 2017.
11 王春娟, 董丽丽, 贾丽. Web 集群系统的负载均衡算法. 计算机工程, 2010, 36(2): 102–104.