DOI：10.13334/j.0258-8013.pcsee.2015.03.004 文章编号：0258-8013 (2015) 03-0527-11 中图分类号：TM 76

电力用户侧大数据分析与并行负荷预测

王德文，孙志伟

(华北电力大学控制与计算机工程学院，河北省 保定市 071003)

Big Data Analysis and Parallel Load Forecasting of Electric Power User Side

WANG Dewen, SUN Zhiwei

(School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, Hebei Province, China)

ABSTRACT: With the development of smart grids, communication network and sensor technology, the electric power user side data is growing exponentially, more complexi, and gradually forms the big data of electric power user side. Now the traditional data analysis model can't meet the demand of big data, so a new data analysis model aiming at analyzing and processing big data of power user side is urgently necessary. The source of the big data of electric power user side is analyzed in this paper. Those challenges facing data storage, availability, processing of the power user side are pointed out based on volume, variety and speed and other characteristics of the big data. Combining cloud computing technology, an analysis and processing platform of big data of electric power user side is given, which integrates smart meter data, SCADA systems data and various sensors data to be processed by MapReduce or Spark. A load forecasting method based on parallel random forests algorithm is proposed. Parallelization random forest algorithm is used to analyze data, such as load data, temperature, wind speed. The method shortens the time of load forecasting and improves random forests algorithm on data processing capability. Parallel load forecasting prototype system of electric power users side big data based on Hadoop is designed and implemented, including cluster management, data management, predictive classification algorithms library functions and so on. By using data sets of different sizes to do load forecasting experiment with parallelization random forest algorithm, the experiment results show that the prediction accuracy of the parallel random forest algorithm is significant higher than that of the decision tree. The prediction accuracy of different data sets is generally higher than the forecast accuracy of the decision tree, and applying the parallel random forest algorithm to analyze and processing big data is a better choice.

KEY WORDS: big data; electric power user side; load forecasting; parallel processing; cloud computing

摘要：随着智能电网、通信网络技术和传感器技术的发展，电力用户侧数据呈指数级增长、复杂程度增大，逐步构成了用户侧大数据。传统的数据分析模式已无法满足需求，迫切需要解决电力用户侧的大数据在分析与处理方面的难题。该文分析电力用户大数据的来源，针对电力用户侧大数据的数据量大、种类繁多与速度快等特点，指出电力用户侧的大数据在数据存储、可用性、处理等方面面临的挑战。结合云计算技术提出一种电力用户侧大数据分析处理平台，将智能电表、SCADA 系统和各种传感器中采集的数据整合，并利用并行化计算模型 MapReduce 与内存并行化计算框架 Spark 对电力用户侧的大数据进行分析。提出基于随机森林算法的并行负荷预测方法，将随机森林算法进行并行化，对历史负荷、温度、风速等数据进行并行化分析，缩短负荷预测时间和提高随机森林算法对大数据的处理能力。设计并实现基于 Hadoop 的电力用户侧大数据并行负荷预测原型系统，包括数据集群的管理、数据管理、预测分类算法库等功能。采用不同大小的数据集对并行化随机森林算法进行负荷预测实验，实验结果表明，并行化随机森林算法的预测精度明显高于决策树的预测精度，且在不同数据集上预测精度普遍高于决策树的预测精度，能够较好的对大数据进行分析处理。

关键词：大数据；电力用户侧；负荷预测；并行处理；云计算

0 引言

智能电网是当前全球电力工业关注的热点，而用户作为智能化用电的行为主体，在智能电网需求响应中起着至关重要的作用[1]。对电网用户侧实时数据的采集、传输和存储，并结合累积的海量多源历史数据进行快速分析能够有效的改善需求侧管理，对用户侧数据进行管理与处理支撑着智能电网

基金项目：国家自然科学基金项目(61074078)；中央高校基本科研业务费专项资金资助项目(12MS113)。

Project Supported by National Natural Science Foundation of China (61074078); Fundamental Research Funds for the Central Universities (12MS113).

安全、坚强及可靠运行。

随着各类传感器和智能设备数量的不断增加，设备中进行获取与传输的各类数据也在发生着指数级的增长，这些数据不仅包括智能电表收集的用电量，还包括各类传感器按照固定频率采集的温度、天气、湿度、地理信息和风速信息等。用户侧数据复杂程度增大，数据存储规模将从目前的 GB 级增长到 TB 级，甚至 PB 级[2]，逐步构成了用户侧大数据。

大数据目前已成为学术界和产业界共同关注的研究主题[3]。2013 年中国电机工程学会信息化专委会发布了《中国电力大数据发展白皮书》[4]，文中阐述了电力大数据的特征，将会给社会带来的价值和在电力行业中的发展前景以及在发展过程中面临的技术挑战。

如何对电力用户侧大数据进行可靠存储、高效管理和快速分析，是当前重要的研究课题。电力用户侧大数据主要来源于智能电表的广泛使用、各类传感器的普及、智能家电的使用和用户消费模式的改变，其中智能电表覆盖率在 2013 年 1 月底已达到为 40.5%，其中直供直管范围的智能电能表覆盖率为 55%[5]，而智能家电随着物联网和大数据的发展将使更多可控的智能家电进入居民生活中。根据其来源总结出电力用户侧大数据的特点如下：

1）数据量巨大。美国太平洋天然气电力公司每个月从 900 万个智能电表中收集超过 3 TB 的数据，每年将存储超过 39 TB 的数据[6]。一个地区如果有 10 000 套传感器终端，按每套终端每 5 min 采集一次数据计算，每月产生数据总量约 9.3 TB，每年产生数据接近 1 PB。随着电网智能化程度的加深，以及为了保证精细化、准确化控制，数据维度也从几十向上百过渡，同时影响电力负荷因素采集频率的提高和采集种类的增多，使上述数据量更加快速增长，而且在多数情况下还需要存储所有的历史数据值以满足溯源处理和复杂数据分析的需求[7]。

2）数据结构类型繁多。随着各类传感器的广泛使用，收集的数据包括各种结构化数据、半结构化数据和非结构化数据，这些数据在采集、传输、存储和处理的过程中形成了多源异构数据。

3）速度快。一次采集频度的提升就会带来数据体量的“指数级”变化，如对 100 万智能电表的数据采集中，采集频率 15 min 将产生 3.18 TB 的数据，频率为 1 min 将产生 47.7 TB 的数据，频率为 1 s 将产生 11.2 PB 的数据[4]。电力系统中的高级应用需要对海量的历史数据进行离线分析处理，这要求数据平台能够提供并行化的海量历史数据批处理的能力，以及能够快速传输与存储采集到的新数据。

4）数据的交互性。智能电网的一个重要特性之一是交互性，包括与用户的交互实现智能用电和与相关行业的数据交互融合进行全方位的挖掘分析，如将负荷数据与收集到的民生数据、气象数据进行融合进行电力负荷预测。

目前，云计算是解决大数据管理的一种基础平台和高效支撑技术。开源 Hadoop 技术已经成为大数据管理与并行处理的主流技术，主要包括分布式文件系统(Hadoop distributed file system，HDFS)和并行编程框架 MapReduce 两部分，该技术具有高性能、高可靠性和强大的可扩展能力等适合管理大数据的优点，已被淘宝、百度、京东等众多互联网公司使用。电力行业也已开始对其进行研究与应用，目前的研究成果主要集中在系统架构设计、系统模型和存储等方面。例如，文献[8]针对智能电网数据的特点，结合 Hadoop 云计算技术，提出智能电网云计算平台的解决方案，分析了基础设施层、平台层、业务应用层与服务访问层，但尚未讨论实现细节。文献[9]利用 Hadoop 技术对海量电网设备状态监测数据进行存储，设计并实现了一个数据存储原型系统，包括存储客户端和查询客户端，能够对数据进行高效的存储和快速的查询。

国家电网公司在发输电系统的技术与欧美差别不大，但在配用电侧特别是用户侧存在较大差异，不仅技术领域的名称不同，技术内涵和解决方案也有很大差别。由于相适应的市场机制尚未形成，中国实施智能用电技术的条件不够成熟，难以支持智能配电系统和用户侧管理系统的有效集成[6]。电力用户侧的大数据管理存在如下挑战：

1）大数据整合。

传感器网络在智能电网中的广泛使用，智能电表和物联网技术的快速发展，使其产生的大数据模态千差万别，各单位数据口径不一，加工整合困难。针对海量异构数据，如何构建一个模型来对其进行规范表达，如何基于该模型来实现数据融合是亟需解决的问题。

2）大数据可用性。

大数据可用性问题是大数据的重要挑战之一。由于数据的采集方式多种多样，各个通信信道质量不一，不仅接收的数据质量低劣，而且对数据的管控能力也不足，从而导致利用这些低劣的数据进行

挖掘分析发现的知识也是不科学的，不能做出精准的决策。这已经在全球范围内造成了恶劣后果，严重困扰着信息社会。

3）大数据存储。

我国目前已累计实现 1.55 亿户用电信息采集，构建了大规模的高级量测体系(advanced metering infrastructure，AMI)系统，并在 26 个省区建成投运了电动汽车充换电站 360 座、充电桩 15333 个，这些设备汇集到后台将会产生庞大的数据量[8]，而电力数据对储存时间的要求以及海量电力数据的爆发式增长对 IT 基础设施提出了更高的要求。大数据的数据类型复杂，传统的关系型数据库和文件存储格式已不能满足大数据快速增长的需求。

4）大数据分析技术。

智能电网的交互性决定了用户侧大数据处理具有实时性与精准性。而大数据作为一种技术变革的标志，传统的对数据分析技术不能对大数据进行快速数据挖掘分析，已不再适合大数据。从大量数据中挖掘发现可用知识越发困难，迫切需要新的分析技术对大数据进行分析以支持智能电网的推进。

针对电力用户侧大数据的特点和其面对的挑战提出一种基于 Hadoop 的电力用户侧大数据管理方案。利用 Hadoop 集群搭建大数据的基础存储平台，将各电网子系统采集到的数据整合成大数据存储，并利用并行化计算框架对电力用户侧的大数据进行快速挖掘分析。本文以电力负荷预测应用为例，将传统的负荷预测迁移到云计算平台，利用随机森林算法实现并行负荷预测，并与决策树算法进行对比。利用不同大小的数据集对并行化随机森林算法进行实验，分析其算法处理大数据的性能。

1　电力用户侧大数据分析平台

1.1　大数据分析平台

本文参照云计算技术体系结构[10]与处理工具，并结合电力用户侧大数据分析的实际需要，搭建以分析计算为主的电力用户侧大数据管理平台，其基本架构如图 1 所示，分为应用层、私有云计算层、数据管理层。

此框架主要是结合云计算技术，利用 Hadoop 搭建电力用户侧大数据管理平台，在平台上采用 HDFS、HBase 与 Hive 建立大数据存储系统，在平台上搭建 MapReduce 并行化计算框架和 Spark 内存并行化计算框架作为大数据计算分析系统，对电力用户侧的大数据进行分析。

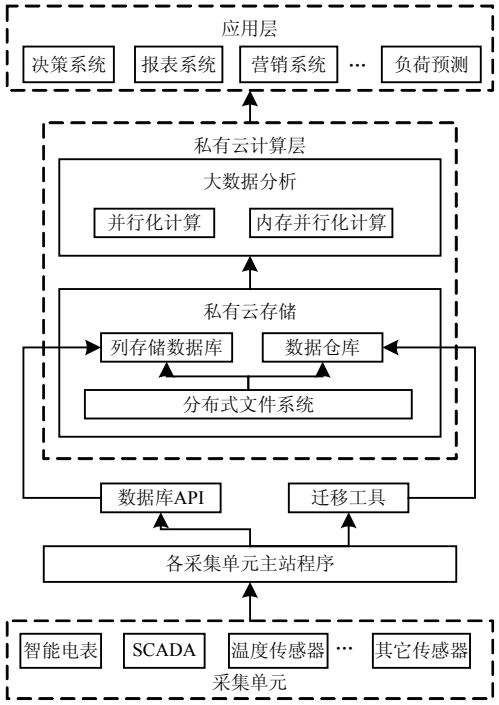

图 1　电力用户侧大数据管理平台架构图

Fig. 1　Architecture diagram of big data manage platform of electric power user side

1.2　数据管理层

数据管理层主要是对数据进行采集和集成整合。数据采集主要包括从智能电表、SCADA 系统和各种传感器中采集的数据，这些数据不仅包括电网内部的数据，还包括大量相关的数据，这些数据由不同产商的设备产生，模态千差万别，各单位数据口径不一，形成了海量异构数据流，加工整合困难。这些数据的集成整合主要是指对传统系统的产生的数据迁移至私有云平台，进行高效的管理。

虽然各产商都提供了相应的应用程序编程接口(application programming interface，API)，但其自动化程度并不高。简单的使用 API 对大数据进行操作效率不高，需要使用第三方工具进行操作，例如 Sqoop 和 Datanucleus 等。Sqoop 是一款在 Hadoop 和关系数据库之间进行相互转移数据的工具。利用 Sqoop 可以使各个子系统的数据在大数据平台上进行整合[11]。Datanucleus 是一款开源的 java 持久化工具，可以对 HBase、Cassandra 多种非关系型数据库进行操作。

平台针对数据集成整合这一难点采用 sqoop 工具对数据进行抽取整合工作，将各个独立的系统产生的数据及历史数据利用 sqoop 抽取整合到 Hive 与 HBase 中。使用 Datanucleus 对列存储数据库进行操作，将基于云计算的应用产生的在线数据写入到 HBase 中。大数据的抽取整合流程如图 2 所示。

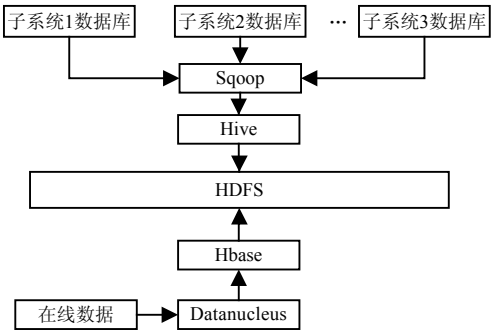

图 2 电力用户侧大数据的抽取与整合流程

Fig. 2 Extraction and integration of electric power user side big data

1.3 私有云计算层与应用层

私有云计算层主要负责大数据的存储和计算分析功能。

云计算层利用 Hadoop 搭建而成，大数据存储在分布式文件系统 HDFS 中，利用 Hive、Pig 和 HBase 对数据进行管理，电力大数据在存储方面已进行了一些研究，例如文献[12-13]提出利用云计算存储、运算技术进行电力数据中心的搭建。文献[14]在云计算平台上将数据映射成数据空间的点集，充分利用计算存储资源，实现数据集到数据中心的布局方案。文献[15]在对数据进行存储时考虑到数据的安全性，利用 HBase 高性能优势和现代密码技术，将密钥与密文的管理分离，开发了基于 Hadoop 的智能电网数据安全存储原型系统。

该平台利用 HBase 存储电力负荷数据和相关数据，HBase 数据库是列为存储单元的，方便对整列数据进行查询，而随后使用的随机森林算法在学习过程中需要多次对整列数据进行读取计算，对数据的操作需求符合 HBase 数据存储的特点。

利用并行化计算模型 MapReduce 对大数据进行并行化批量计算分析，而对数据密集型的迭代计算采用基于内存的并行化计算模型 Spark。Spark 是一个开源的分布式集群系统，用于大数据的快速处理分析。Spark 克服了 Hadoop 在迭代计算上的不足，现已成为 Apache 的顶级项目。Spark 提供了一种内存并行化计算框架，框架将作业所需数据读入内存，所需数据时直接从内存中查询，这样比基于磁盘的 MapReduce 访问数据的速度快，减少了作业的运行时间，也减少了 IO 操作。

并行计算模型主要是对大量的数据进行挖掘，其计算模型主要有 MapReduce[16]、Dremel[27]、Dryad[18]和 Cascading 等，该平台主要利用 Map Reduce 模型对电力用户侧大数据进行挖掘分析。

应用层主要是利用私有云计算集群强大的存储和计算分析能力为企业各部门提供决策和指导功能接口。

2 基于随机森林算法的并行负荷预测

2.1 电力负荷预测

负荷预测是电网规划中的关键环节，是变电站、网架规划重要计算依据[19]，高精度的短期负荷预测能够有效降低发电成本，有关键作用[20]。目前，短期负荷预测常用的方法主要包括以下几种：决策树、极限学习、遗传算法等。其中，决策树在传统预测算法中得到广泛研究，文献[21]分析了决策树 ID3 在扩展时易偏向属性值多的属性及属性间相关性考虑较少的缺点，对其进行改进，提出了属性–值对的两次信息增益优化算法，并用此算法进行日特征负荷决策树预测，预测结果能够满足并超过负荷预测实用化标准的要求，并具有较高的预测精度。文献[22]根据各时段负荷和平均负荷受相关因素影响的不同，结合决策树和解耦法提出解耦决策树方法进行预测，并将决策树前两层由实际经验指定，其余节点自动形成，该方法已在北方某市进行实际应用。随着大数据的产生，云计算技术也越来越多的应用在电力系统中，王保义等[23]针对智能电网中负荷数据的特性，结合云计算技术，利用极限学习进行负荷预测，使其具有分布式能力和多 Agent 思想，提升了负荷预测算法预测准确率和速度。

以上方法均已取得了相应的研究成果，其中决策树是解决短期电力负荷的主流算法之一，但其自身原因和外界因素也存在很多不足，总结如下：

1）在建树初始要把所有属性读入内存，这限制了可以处理的数据量，无法对大数据进行分析；

2）容易出现过生长现象，使决策树过于复杂，导致对训练数据集可以进行很好的分类，但对测试数据集分类效果不佳；

3）随着智能电网的不断发展，用电信息的采集频率不断提高，以及对预测的精度要求越来越高，采集到的影响负荷变化的随机因素也越来越庞大，不确定性也越来越大；因此，传统的数据挖掘算法已经不能满足大数据环境下短期负荷预测的要求。

随机森林是一种集成学习方法，以决策树为基本学习单元，包含多个由 Bagging 集成学习理论和随机子空间方法训练得到的决策树，输入待分类的样本，由各个决策树产生各分类结果，最终的分类

结果由各个决策树的结果进行投票决定。随机森林是多个决策树的集成学习方法，不仅可以克服决策树的一些不足，而且具有良好的可扩展性和并行性，能够有效解决大数据的快速处理问题，针对大数据环境下的电力负荷预测有较好的应用前景。

2.2　随机森林算法原理

随机森林[24]是由一系列分类回归树组成的，在2001 年由 Leo Breiman 根据他的 Bagging 集成学习理论和 Ho 提出的随机子空间理论相结合提出的。在随机森林中，每个分类回归树都有各自独立的样本训练集 TS，TS 是由 Bagging 算法从总样本 S 中有放回的抽取与 S 等数量的样本组成。算法在利用各个 TS 进行分类回归树的训练学习，形成各个分类器过程中，每个内部节点的分支是根据随机子空间理论随机选取若干个属性值进行的，最后形成一个具有分类规则或者回归功能的决策树群。随机森林的最终结果为各个分类回归树进行投票选择或者各分类回归树结果的平均值。随机森林中单个决策树训练过程如图 3 所示。

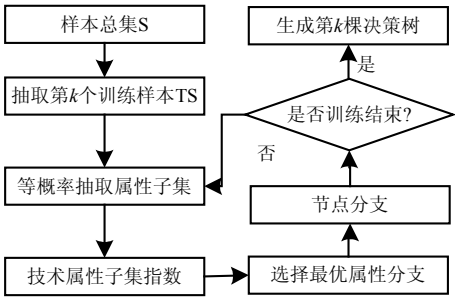

图 3　随机森林中单个决策树训练过程

Fig. 3　Training process of a single decision tree in random forests

单个分类回归树的构造过程主要包括从属性集中选择合适的属性值进行分支，然后在其产生的子树上分别重复划分搜索过程，直到满足一个停止生长规则为止。

进行分支属性值的选择依据为 Gini 指数和最小二乘偏差，其中 Gini 指数适合于分类树，而最小二乘偏差适合于回归树，具体计算如下：

1）Gini 指数。

Gini 指数可以度量节点的不纯性，其公式为

$$\mathrm{GINI}(t)=1-\sum_{j}p^2(j/t) \tag{1}$$

式中：t 为当前节点分支属性；$p(j/t)$表示目标类别 j 在节点 t 中出现的比例。

节点 t 按属性值 s 划分的 Gini 标准定义为

$$\mathrm{GINI}(s,t)=p_L\mathrm{GINI}(t_L)+p_R\mathrm{GINI}(t_R) \tag{2}$$

划分标准为使 GINI(s,t)最小。

2）最小二乘偏差。

最小二乘偏差多用来度量回归树，节点 t 的拟合误差公式为

$$\mathrm{Err}(t)=\frac{1}{n_t}\sum_{D_t}(y_i-k_t)^2 \tag{3}$$

式中：n_t 为节点 t 中实例的个数；k_t 为每个节点中实例的目标值的平均值，$k_t=\frac{1}{n_l}\sum_{D_l}y_l$。

节点 t 按属性值 s 划分的最小二乘偏差标准定义为

$$\mathrm{Err}(s,t)=\frac{n_{tL}}{n_t}\mathrm{Err}(t_L)+\frac{n_{tR}}{n_t}\mathrm{Err}(t_R) \tag{4}$$

为简化在计算机中的计算过程，避免多次遍历属性值，对式(4)进行简化，可得

$$\mathrm{Err}(s,t)=\frac{S_L^2}{n_{tL}}+\frac{S_R^2}{n_{tR}} \tag{5}$$

式中 $S_L=\sum_{D_{tL}}y_i$，$S_R=\sum_{D_{tR}}y_i$，划分标准是使式(5)最大。

2.3　随机森林算法的并行化依据

随机森林算法是一个集成学习算法，这成为并行化的基础，但随机森林并不是 K 个决策树模型的简单组合，这使随机森林并行化并不只是量的变化。随机森林的 Bagging 与随机子空间为算法并行化提供了理论依据，分别叙述如下。

1）Bagging 思想。

从总样本 S 中有放回的抽取 K 个训练样本，其样本数为$|S|$，其中约有 37%的数据没有被抽到，使得每个学习单元的训练样本不一样，构建的过程是相互独立的。这确保了随机森林的训练过程不仅可以进行数据的并行化，也可以进行学习单元的并行化，提高模型生成的速率，能够更有效的进行大数据的处理分析。

2）随机子空间思想。

学习单元在每个节点进行属性测试时，随机的从样本属性中抽取若干个属性进行测试。这就避免了一次把所有测试属性读入内存，也避免了决策树在形成过程中容易产生过度拟合的问题。

这两个思想确保了随机森林算法是一个可以并行，能够对大数据进行分类预测的算法，具有较高的分类预测精度，而且对噪声和异常值有较好的稳健性，具有较强的泛化能力。

目前，随机森林算法的应用仍然是单机版的串

行应用。随机森林算法是多个决策树的集成，每个决策树的训练是很耗时的，使得随机森林预测模型的建立需要较长的时间，无法直接应用于大数据环境下的电力负荷预测中。本文针对随机森林在大数据负荷预测中的不足提出了基于 MapReduce 的并行化随机森林算法(MapReduce-random forests，MR-RF)。

2.4 基于并行随机森林的负荷预测过程

整个过程利用 3 个 MapReduce 作业类执行算法的训练过程，每一个 MapReduce 的输出作为其后一个的输入。训练结束后得到的随机森林模型保存在 Hadoop 的分布式集群中，其分为三部分：生成数据字典；生成决策树；形成随机森林。

生成数据字典就是对进行训练的样本数据进行描述，产生一个文件来描述样本中条件属性和决策属性，记录条件属性值的类型和决策属性的位置，以及要创建的模型是进行分类还是回归运算。这个过程由第一个 MapReduce 完成，每个 Map 过程读取实验数据的一部分，记录数据的属性类型和负荷值或者类型标识。产生的描述文件以 key/value 的形式存储在 Hadoop 的文件系统 HDFS 中，以备随后的 MapReduce 使用。

生成决策树过程为整个并行化算法的核心，其并行化过程主要其中在以下几方面：

1）对原数据集 Dataset 利用 Bagging 算法进行随机有放回的抽取 K 个与原样本数据集大小一样的样本数据 $TS_{1,2,\ldots,k}$。因为是有放回的抽取，所以可以并行对原数据集进行抽取，而不会对 TS 产生影响。一个 TS 对应一个决策树的训练集，每个 TS 都有所不同，并且与原数据集大小一样，这样既保证了各个决策树的不同，又不会失去原数据集的知识规模。

2）根据样本数据中属性的个数 M 确定每个节点随机选择的属性个数 $m(m<<M)$，分类模型中 m 为 M 的开方根，回归模型中 m 为 M 的 1/3。计算 m 个属性中每个属性的信息量，选择最佳的属性进行分支；

3）递归的进行节点的建立，生成决策树。

生成第 k 个决策树的 map 伪代码如下：

输入：n_t 个实例，实例中 y 值的总和 S_t，属性值 X_v；

输出：第 k 个决策树；

方法：按 X 属性值升序排序

$S_R = S_t$；$S_L = 0$；

$n_R = n_t$；$n_L = 0$；

BestValue = 0；

For 实例 i in 所有的实例{

$S_L=S_L+y_i$；$S_R=S_R-y_i$；

$n_L=n_L+1$；$n_R=n_R-1$；

//$X_{i,v}$ 为 X 属性值排序后第 i 个属性值

if($X_{i+1,v}>X_{i,v}$){

SplitValue =(S_L^2/n_L)+(S_R^2/n_R)；

If(SplitValue > BestValue){

BestValue = SplitValue;

BestSplitPoint=($X_{i+1,v}+X_{i,v}$)/2;

}

}

}

K 个决策树的生成是并行产生的，一个 Map 生成一个决策树，实现了算法的并行。这个过程由第二个 MapReduce 过程完成。此 MapReduce 只有 Map 过程没有 Reduce 过程。

形成随机森林也就是把每个决策树分类器组合起来。每个决策树都会产生一个结果，如果随机森林用来分类其最终结果为投票选取，当它用来回归预测时，K 个树会给出 K 个值，最终值为各树的平均值。此过程由第三个 MapReduce 完成。

利用随机森林算法的并行化进行短期负荷预测的具体预测流程如图 4 所示。整个模型是建立在 Hadoop 的分布式集群上，对大数据进行分布式存储，利用 MapReduce 将随机森林算法并行化，使算

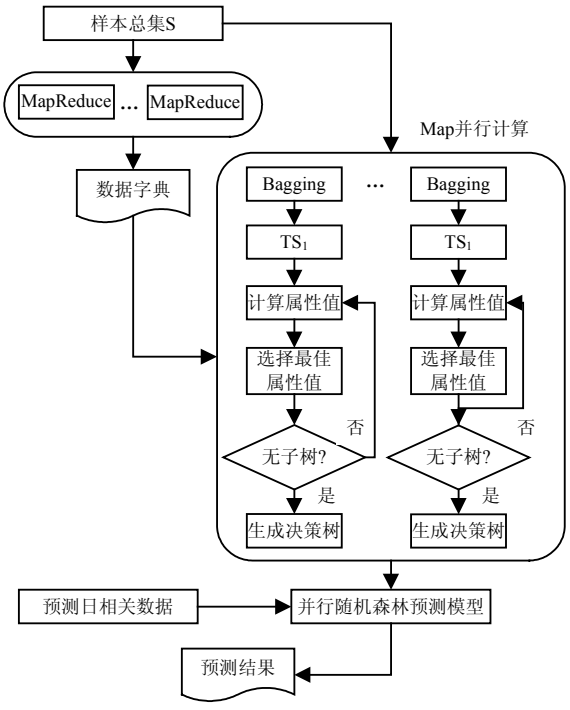

图 4 并行化随机森林短期负荷预测流程图

Fig. 4 Flow chart of parallel random forests for short-term load forecasting

法能够依托 Hadoop 集群的存储能力和计算能力对数据的挖掘和计算预测，整个过程都是并行执行的，能够有效提高预测的精度和改善负荷预测系统处理大数据的能力。

3　电力用户侧大数据并行负荷预测原型系统与实验分析

3.1　电力用户侧大数据实验平台

课题组在实验室构建了一个电力用户侧大数据实验平台。实验环境由 35 台配置相同的 PC 机组成，每台 PC 机 CPU 为双核 Inter i5-2400，主频 3.10 GHz，4.00 GB 内存，500 GB 硬盘，其中一台 PC 机作为主节点，其它的 PC 均作为数据节点。主节点作为一个中心服务器，负责整个集群的资源分配和作业的调度，也是整个文件系统的管理节点，负责文件系统名字空间的管理与维护。数据节点主要是存储和运行任务。主节点将文件进行分块并存储与文件分块信息相关的名字空间和元数据，各个分块数据被冗余的存储在各个数据节点，每块数据默认存储在 3 个数据节点上。一个 MapReduce 作业提交到主节点之后，由主节点将此作业分解成多个小任务，并根据整个集群的资源和任务所需资源将小任务分配给各个数据节点进行运行，并对其运行过程进行监控。实验集群的拓扑结构如图 5 所示。

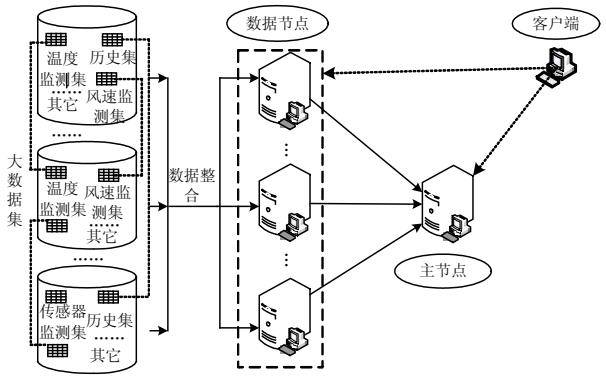

图 5　实验集群拓扑结构图

Fig. 5　Topology map of experimental cluster

图 5 中大数据集为各个系统在关系型数据库中存储的数据，包括传感器的监测数据、电力负荷的历史数据以及相关数据等。大数据集通过相关的数据整合工具迁移到 Hadoop 集群中，目前仍没有一个标准的高效的数据整合方法将各数据集整合到 Hadoop 集群中。本文采用 3 种整合方法，包括编写 MapReduce 作业、开源的 java 持久化插件 DataNucleus、开源的 Sqoop 工具，分别叙述如下。

1）编写 MapReduce 作业。

根据需求编写 MapReduce 作业对操作者有较高的要求，需要操作者根据需求编写高效快速的程序，其运行效率与程序的编写有着直接的关系。在测试中将 20 万条测试数据插入到 HBase 中，单线程的逐条插入所需时间为 9 min，多线程的并发插入所需时间为 1 min，而利用 HBase 中表在 HDFS 中存储特点，利用 BulkLoad 插入 HBase 所需时间为 30 s。

2）开源的 java 持久化插件 DataNucleus。

DataNucleus 是一个开源的持久化插件，支持当前众多主流存储系统。DataNucleus 屏蔽了个存储系统的差异，提供统一的对外接口，操作者容易掌握，但其灵活度不高，操作者只能使用其提供的 API 对程序进行优化。在测试中将 20 万条测试数据插入到 HBase 中，所需时间为 3 min。

3）开源的 Sqoop 工具。

Sqoop 可以将本地文件或者数据库表与 HDFS 文件进行相互迁移。Sqoop 是基于 MapReduce 实现的，使操作者不用过多的去关注 MapReduce 的实现和优化。在测试中将 20 万条测试数据插入到 HBase 中，所需时间约为 1 min。

上述 3 种方法各有利弊，Sqoop 虽然效果较好，但使用不够灵活；DataNucleus 虽然效率较低，但利于集成开发；直接编写 MapReduce 作业效率最快，但对操作者有较高的要求。由于研究过程中需求的多样性，采用了 3 种整合方法。

3.2　基于 Hadoop 的电力用户侧大数据并行负荷预测原型系统

在此实验平台上搭建电力用户侧大数据管理平台，并在该平台上实现了基于 Hadoop 的电力用户侧大数据并行负荷预测原型系统(见附录 A 中图 A1)。

系统功能包括数据集群的管理、文件管理、数据管理、序列化文件管理、预测分类算法库和预测结果展示。集群管理主要是向 Hadoop 集群注册用户，得到操作集群的权限，并对 MapReduce 作业进行实时监控。文件管理在 HDFS 文件系统上实现了文件的上传、下载、在线查看和编辑功能。数据管理是对各系统中的数据进行抽取整合到 Hadoop 集群中，并进行数据的展示。序列化文件管理主要是对较小的文件和 MapReduce 作业中的中间结果进行管理，将较小的文件进行序列化为 SequenceFile 文件进行存储，有利于提高小文件的存储效率，对 MapReduce 作业的中间结果进行查看容易更好的理解 MapReduce 的执行过程。预测分类算法库包括

了多种并行化的预测分类算法，包括 K 均值算法、决策树算法、随机森林算法、贝叶斯算法。结果展示则是对预测分类的结果利用图表工具进行展示，提高数据的可视化程度。本文中实验均在此原型系统中进行。

3.3 实验数据和特征值的选取

实验数据来自某地区 2011 年 3、4 月的负荷信息和天气信息，负荷信息采集频率是 0.5 h(共 2 880 行数据)，天气信息为最高气温、最低气温、降雨量。实验中数据量虽然没有达到大数据的规模，但可以用此实验数据进行算法正确性实验，随后对实验数据进行人为的扩充达到大数据规模进行算法预测速率实验。进行反复多次的测试，取平均值为最终实验结果。

对大数据环境下负荷预测数据的研究发现这些数据呈现一种延续性、周期性、相关性特点，根据这些特点和大量文献[25-26]的研究成果确定样本属性为星期、是否周末、最高气温、最高气温变化率、最低气温、最低气温变化率、降雨量、上月同时期负荷、上周同时期负荷、昨天负荷、预测负荷，其样本数据如表 1 所示。此外，负荷数据又是一个时间序列数据，有着近大远小的特点，则对上述属性设置加权值为上月同时期负荷的权值为 0.2，上周同时期负荷的权值为 0.5，昨天负荷的权值为 1。

表 1 部分负荷训练数据集

Tab. 1 Part of load training data set

属性	值
星期	星期日
是否周末	1
最高气温	9.0
最高气温变化率	–2.5
最低气温	8.3
最低气温变化率	0.5
降雨量	25.9
上月同时期负荷	1 529.05
上周同时期负荷	1 453.64
昨天负荷	1 517.89
负荷	1 470.08

3.4 实验评价指标

负荷预测结果的评价指标采用平均绝对百分比误差(mean absolute percentage error，MAPE)，表达式为

$$\mathrm{MAPE}=[\sum_{t=1}^{n}(|Y_t-y_t|/y_t)]/n\times 100\% \qquad (6)$$

式中：Y_t 为预测值；y_t 为真实值；n 为预测点的个数。电力负荷预测中，MAPE 值越小，负荷预测值越准确。

算法并行性评价指标采用通用的加速比，表达式为

$$S_{\mathrm{peedup}}=t/T \qquad (7)$$

式中：t 为单机运行的时间；T 为集群运行的时间。

3.5 基于随机森林算法的并行负荷预测实验分析

实验一：本次实验将 MR-RF 算法与传统决策树算法进行比较，以某地区 2011-3-7 至 2011-4-27 的历史数据为训练样本数据集，训练 MR-RF 算法和决策树算法来预测 2011-4-28 当天的负荷值，实验进行多次求取平均值为最终实验结果，采用公式(6)作为评价函数，实验结果如表 2 所示。由表 2 可见，MR-RF 的 MAPE 为 1.43%，而决策树的 MAPE 为 2.12%，这表明 MR-RF 的预测精度高于决策树，这是因为 MR-RF 是由若干个随机抽取的决策树集成在一起的，具备决策树优点的同时又克服了决策树的一些缺陷，表现出比决策树更好的特性。

表 2 MR-RF 与决策树的 MAPE

Tab. 2 MAPE of MR-RF and decision tree

预测方法	MAPE/%
MR-RF	1.43
决策树	2.12

图 6 是 2011-4-28 当天真实负荷值和用 MR-RF 算法、决策树算法的进行电力负荷预测的预测值的对比图。由图 6 可知，当真实负荷值变化比较平缓时，决策树算法和 MR-RF 都表现出较高的准确度，但随着真实负荷谷峰的出现，决策树算法的预测精度有所下降，MR-RF 算法依然表现出较高的预测精度，可见 MR-RF 算法更适合在生产实际中使用。

实验二：本次实验主要是对 MR-RF 算法中决策树的个数 K 进行实验确定。不同树大小的 MAPE 值如表 3 所示。由于数据集大小和数据集属性个数

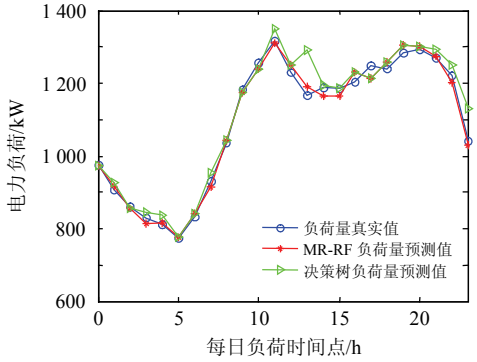

图 6 电力负荷真实值与预测值的对比

Fig. 6 Comparison of actual and forecast load values

表 3　不同大小森林的 MAPE

Tab. 3　MAPE of different sizes forest

树的大小 K	MAPE/%
100	1.84
150	1.61
200	1.43
250	1.48
300	1.61

的影响，K 取值过小会使预测模型倾向决策树模型，预测精度不高，取值过大会随机森林模型分类过细，使集群计算量剧增，因此 K 是决定 MR-RF 预测精度的重要变量。由表 3 可以看出 K=200 时有较好的 MAPE。

实验三：本实验主要是比较数据量的增长对算法预测精度的影响。实验数据文件大小和包含的数据元组数如表 4 所示。

表 4　数据文件大小和元组数

Tab. 4　Size of data files and number of tuples

文件大小/MB	410	820	1 640	3 280	6 560
元组数	4×10^6	8×10^6	1.6×10^7	3.2×10^7	6.4×10^7

将各不同的数据集分别用来进行并行负荷预测实验，然后分别计算各个 MAPE 值，其实验结果如表 5 所示。

表 5　不同数据集的 MAPE

Tab. 5　MAPE of different data sizes

元组数	MAPE/%
4×10^6	1.63
8×10^6	1.61
1.6×10^7	1.82
3.2×10^7	1.76
6.4×10^7	1.91

由表 5 可以看出，不同大小数据集的预测精度不一样，没有明显的变化规律，但其精度均小于实验一中决策树的预测精度，证明并行化的随机森林算法适合用户侧大数据的负荷预测。

实验四：本次实验主要是测试不同数据集导入不同大小 Hadoop 集群的影响。将实验三中的数据分别导入 5 台和 8 台 Hadoop 集群所用的时间如图 7 所示。

由图 7 可以看出，不同大小数据文件导入集群的时间随着数据集的变化而变化，但所花费的时间较少。在 5 个节点和 8 个节点的集群上导入相同大小数据文件的时间曲线基本吻合，说明数据文件导入 Hadoop 集群受集群规模的影响较小。

实验五：算法并行性的好坏用加速比来衡量，

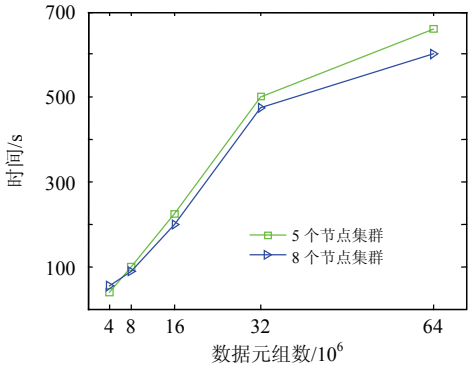

图 7　数据文件导入两个集群的时间对比图

Fig. 7　Time comparison of two clusters on data files importing

即式(7)。由于实验数据有限，人为的把原数据集扩大 2.4 G、12.4 G、124 G，分别运行在 1、5、15、25、35 台大小的分布式集群上，运行结果如图 8 所示。

由图 8 可以看出，并行的随机森林算法在不同数据量不同大小的分布式集群中显示了接近线性增长的趋势，并且在相同集群大小的情况下数据量越大加速比也越大，但是随着集群的增多加速比会减少，但总的来说随着集群数量的增多加速比会变大。

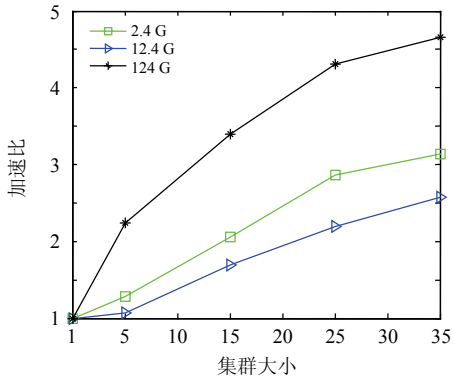

图 8　并行随机森林算法的加速比

Fig. 8　Speedup of parallel random forest algorithm

4　结论

本文结合国内外电力大数据的研究现状，针对电力用户侧大数据展开了研究。分析了用户侧大数据的特点，提出了一个大数据分析平台，并在此平台上开发了基于 Hadoop 的电力用户侧大数据并行负荷预测原型系统，在此原型系统上利用并行化后的随机森林算法进行负荷并行预测实验，经试验表明该方法提高了负荷预测的精度。

受实验环境的影响，实验中使用的数据集最大只达到 GB 级，但是所进行的实验已从不同的角度模拟数据量的增加，其实验结果仍具有可参考性，下一步的工作准备对更大数据集进行分析和并行处理，对内存并行化计算框架 Spark 进行深入研究。

参考文献

[1] Rusitschka S，Eger K，Gerdes C．Smart grid data cloud：A model for utilizing cloud computing in the smart grid domain[C]//First IEEE International Conference on Smart Grid Communications．Gaithersburg，MD，USA：IEEE，2010：483-488.

[2] 丁杰，奚后玮，韩海韵，等．面向智能电网的数据密集型云存储策略[J].电力系统自动化，2012，36(12)：66-70.
Ding Jie，Xi Houwei，Han Haiyun，et al．A smart grid-oriented data placement strategy for data-intensive cloud environment[J]．Automation of Electric Power Systems，2012，36(12)：66-70(in Chinese).

[3] 孟小峰，慈祥．大数据管理：概念、技术与挑战[J]．计算机研究与发展，2013，50(1)：146-169.
Meng Xiaofeng，Ci Xiang．Big data management：concepts，techniques，and challenges[J]．Computer Research and Development，2013，50(1)：146-169(in Chinese).

[4] 中国电机工程学会信息化专委会.中国电力大数据发展白皮书[S]．北京：中国电力出版社，2013.
Chinese society for electrical engineering．The white paper on the development of big data on china electric [S]．Beijing： China Electric Power Press，2013(in Chinese).

[5] 宋亚奇，刘树仁，朱永利，等．电力设备状态高速采样数据的云存储技术研究[J]．电力自动化设备，2013，33(10)：150-156.
Song Yaqi，Liu shuren，Zhu Yongli，et al．Cloud storage of power equipment state data sampled with high speed[J]．Electric Power Automation Equipment，2013，33(10)：150-156(in Chinese).

[6] 张东霞，姚良忠，马文媛.中外智能电网发展战略[J].中国电机工程学报，2013，33(31)：1-14.
Zhang Dongxia，Yao Liangzhong，Ma Wenyuan．Development strategies of smart grid in China and abroad[J]．Proceedings of the CSEE，2013，33(31)：1-14(in Chinese).

[7] 宋亚奇，周国亮，朱永利．智能电网大数据处理技术现状与挑战[J]．电网技术，2013，37(4)：927-935.
Song Yaqi，Zhou Guoliang，Zhu Yongli．Present status and challenges of big data processing in smart grid[J]．Power System Technology，2013，37(4)：927-935(in Chinese).

[8] 王德文，宋亚奇，朱永利．基于云计算的智能电网信息平台[J]．电力系统自动化，2010，34(22)：7-12.
Wang Dewen，Song Yaqi，Zhu Yongli．Information platform of smart grid based on cloud computing [J]．Automation of Electric Power Systems，2010，34(22)：7-12(in Chinese).

[9] 刘树仁，宋亚奇，朱永利，等．基于 Hadoop 的智能电网状态监测数据存储研究[J].计算机科学，2013，40(1)：81-84.
Liu Shuren，Song Yaqi，Zhu Yongli，et al．Research on data storage for smart grid condition monitoring using Hadoop[J]．Computer Science，2013，40(1)：81-84(in Chinese).

[10] 曹子健，林今，宋永华．主动配电网中云计算资源的优化配置模型[J]．中国电机工程学报，2014，34(19)：3043-3049.
Cao Zijian，Lin Jin，Song Yonghua．Optimization model for resources allocation of cloud computations in active distribution networks[J]．Proceedings of the CSEE，2014，34(19)： 3043-3049(in Chinese).

[11] 王德文，肖凯，肖磊．基于 Hive 的电力设备状态信息数据仓库[J]．电力系统保护与控制，2013，41(9)：152-130.
Wang Dewen，Xiao Kai，Xiao Lei．Data warehouse of electric power equipment condition information based on hive[J]．Power System Protection and Control，2013，41(9)：152-130(in Chinese).

[12] 王德文.基于云计算的电力数据中心基础架构及其关键技术[J]．电力系统自动化，2012，36(11)：67-71.
Wang Dewen．Basic framework and key technology for a new generation of data center in electric power corporation based on cloud computation[J]. Automation of Electric Power Systems，2012，36(11)：67-71(in Chinese).

[13] 赵俊华，文福拴，薛禹胜，等．云计算：构建未来电力系统的核心[J]．电力系统自动化，2010，34(15)：1-8.
Zhao Junhua，Wen Fushuan，Xue Yusheng，et al．Cloud computing：Implementing an essential computing platform for future power systems[J]．Automation of Electric Power Systems，2010，34(15)：l-8(in Chinese).

[14] 丁杰，奚后玮，韩海韵，等．面向智能电网的数据密集型云存储策略[J].电力系统自动化，2012，36(12)：66-70.
Ding Jie，Xi Houwei，Han Haiyun，et al．A smart grid-oriented data placement strategy for data-intensive cloud environment[J]．Automation of Electric Power Systems，2012，36(12)：66-70(in Chinese).

[15] 张少敏，李晓强，王保义．基于 Hadoop 的智能电网数据安全存储设计[J]．电力系统保护与控制，2013，41(14)：136-140.
Zhang Shaomin，Li Xiaoqiang，Wang Baoyi．Design of data security storage in smart grid based on hadoop [J]．Power System Protection and Control，2013,41(14)：136-140(in Chinese).

[16] Tom White．Hadoop 权威指南：中文版[M]．曾大聃，周傲英，译．清华大学出版社，2010.
Melnik S，Gubarev A，Long Jingjing，et al．Dremel：Interactive analysis of web-scale datasets[J]．PVLDB，2010，3(1)：330-339.

[17] Isard M，Budiu M，Yu Yuan，et al．Dryad：distributed data-parallel programs from sequential building blocks [C]//Proc of EuroSys 2007．New York：ACM，2007：59-72．

[18] 钟清，孙闻，余南华，等．主动配电网规划中的负荷预测与发电预测[J]．中国电机工程学报，2014，34(19)：3050-3056．
Zhong Qing，Sun Wen，Yu Nanhua，et al．Load and power forecasting in active distribution network planning [J]．Proceedings of the CSEE，2014，34(19)：3050-3056(in Chinese)．

[19] 毛李帆，姚建刚，金永顺，等．中长期电力组合预测模型的理论研究[J]．中国电机工程学报，2010，30(16)：53-59．
Mao Lifan，Yao Jiangang，Jin Yongshun，et al．Theoretical study of combination model for medium and long term load forecasting[J]．Proceedings of the CSEE，2010，30(16)：53-59(in Chinese)．

[20] 栗然，刘宇，黎静华，等．基于改进决策树算法的日特征负荷预测研究[J]．中国电机工程学报，2005，25(24)：36-41．
Li Ran，Liu Yu，Li Jinghua，et al．Study on the daily characteristic load forecasting based on the optimizied algorithm of decision tree[J]．Proceedings of the CSEE，2005，25(24)：36-41(in Chinese)．

[21] 李响，黎灿兵，曹一家，等．短期负荷预测的解耦决策树新算法[J]．电力系统及其自动化学报，2013，25(3)：13-19．
Li Xiang，Li Canbing，Cao Yijia，et al．New algorithm of short-term load forecasting according to decision tree and decoupling[J]．Proceedings of the CSU-EPSA，2013，25(3)：13-19(in Chinese)．

[22] 王保义，赵硕，张少敏．基于云计算和极限学习机的分布式电力负荷预测算法[J]．电网技术，2014，38(2)：526-531．
Wang Baoyi，Zhao Shuo，Zhang Shaomin．Distributed power load forecasting algorithm based on cloud computing and extreme learning machine[J]．Power System Technology，2014，38(2)：526-531(in Chinese)．

[23] Breiman L．Random forests[J]．Machine Learning，2001，45(1)：5-32．

[24] 焦润海，苏辰隽，林碧英，等．基于气象信息因素修正的灰色短期负荷预测模型[J]．电网技术，2013，37(3)：720-725．
Jiao Runhai，Su Chenjun，Lin Biying，et al．Short-term load forecasting by grey model with weather factor-based correction[J]．Power System Technology，2013，37(3)：720-725(in Chinese)．

[25] 张素香，刘建明，赵丙镇，等．基于云计算的居民用电行为分析模型研究[J]．电网技术，2013，37(6)：1542-1546．
Zhang Suxiang，Liu Jianming，Zhao Bingzhen，et al．Cloud computing-based analysis on residential electricity consumption behavior[J]．Power System Technology，2013，37(6)：1542-1546(in Chinese)．

[26] Matei Zaharia，Mosharaf Chowdhury，Tathagata Das，et al．Resilient distributed datasets：A fault-tolerant abstraction for in-memory cluster computing[C]//Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation．USA：USENIX Association Berkeley，2012：2-2．

[27] Matei Zaharia，Mosharaf Chowdhury，Michael J，et al．Spark：Cluster computing with working sets [C]//Proceedings of the 2nd USENIX conference on Hot topics in cloud computing．USA：USENIX Association Berkeley，2010：10-10．

[28] 刘萌，褚晓东，张文，等．负荷分布式控制的云计算平台构架设计[J]．电网技术，2012，36(8)：140-144．
Liu Meng，Chu Xiaodong，Zhang Wen，et al．Design of cloud computing architecture for distributed load control [J]．Power System Technology，2012，36(8)：140-144(in Chinese)．

附录　A

笔者所在实验室搭建了电力用户侧大数据管理平台，并在该平台上实现了基于 Hadoop 的电力用户侧大数据并行负荷预测原型系统，界面如图 A1 所示。

图 A1　基于 Hadoop 的并行负荷预测原型系统

Fig. A1　Parallel load forecasting prototype system based on Hadoop


王德文

收稿日期：2014-09-29。

作者简介：

王德文(1973)，男，博士，副教授，研究方向为电力系统自动化与智能信息处理，wdewen@gmail.com；

孙志伟(1987)，男，硕士研究生，研究方向为大数据与电力用户用电行为分析，sunzw20120901@126.com．

(责任编辑　李婧妍)