基于 Hadoop 的公共建筑能耗数据挖掘方法①

王 磊[1], 张永坚[1], 贾继鹏[1], 牛晓光[2], 聂昌龙[1]

[1](山东建筑大学 信息与电气工程学院, 济南 250101)
[2](山东电力工程咨询院有限公司 清洁能源中心, 济南 250013)

摘 要: 针对建筑能耗数据无法有效利用这一问题, 提出利用 Hadoop 分布式架构, 结合建筑基本信息对公共建筑能耗数据进行数据挖掘的方法. 对基于 Hadoop 的公共建筑能耗数据挖掘系统进行了初步设计, 并对系统的基本架构和各模块的功能进行了设计和说明. 同时, 对 Apriori 算法和 C4.5 算法实现 MapReduce 分布式设计. 以山东省 100 栋办公建筑制冷期的空调系统耗电量为例进行实验分析, 得到 6 类建筑信息属性对空调系统能耗的影响规律, 并生成空调系统耗电量判定树, 可判别建筑空调系统耗电量等级, 并对样本建筑的节能改造提供具有针对性的建议.

关键词: 建筑能耗; Hadoop; Apriori 算法; C4.5 算法; MapReduce

Data Mining Method for Public Buildings Energy Consumption Based on Hadoop

WANG Lei[1], ZHANG Yong-Jian[1], JIA Ji-Peng[1], NIU Xiao-Guang[2], NIE Chang-Long[1]

[1](School of information & Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China)
[2](Clean Energy Project Centre, Shandong Electric Power Engineering Consulting Institute Corr, Ltd., Jinan 250013, China)

Abstract: The utilization of building energy consumption data is still inefficient. According to this problem, in this paper, a new method based on Hadoop for data mining of public buildings energy consumption combining with building information is proposed. The paper designs the data mining system of public building energy consumption based on Hadoop, and performs designs and illustrations to the basic framework and functional modules. Apriori algorithm and C4.5 algorithm are implemented distributively using MapReduce programming model. The paper takes 100 office buildings in Shandong Province as examples to analyse the data of air conditioning system energy consumption. The experimental conclusions are the influence rules of 6 kinds of building information on air conditioning system energy consumption. Moreover, the experiment obtains the decision tree of air conditioning system energy consumption. According to the decision tree, we can distinguish the energy consumption level of air conditioning system, and offer targeted advice on energy saving renovation of sample buildings.

Key words: buildings energy consumption; Hadoop; Apriori algorithm; C4.5 algorithm; MapReduce

据住建部公开信息, 截至 2013 年底, 全国累计完成公共建筑能源审计 10000 余栋, 对 5000 余栋建筑进行了能耗动态监测. 目前, 山东省对公共建筑进行节能监测的数量累计达到 1000 余栋. 真实准确的公共建筑能耗统计数据对进一步推进建筑节能工作具有重大意义, 但是对能耗数据的有效利用还存在明显不足. 随着数据量不断增加, 数据的分析工作也面临巨大的挑战. 这些能耗数据背后蕴涵着丰富的知识, 且数量巨大, 常规分析方法难以发现和总结这些数据中所蕴涵的知识.

本文借助于大数据概念和处理模式[1], 利用大数据处理技术的 Hadoop 分布式架构, 将数据挖掘算法进行 MapReduce 编程, 以办公建筑为例把建筑基本信息与建筑电量能耗数据相结合进行数据挖掘研究, 得到

① 基金项目:山东省住房和城乡建设厅项目(2013-HT-01)
收稿时间:2015-06-04;收到修改稿时间:2015-08-31

这些不同类型的数据之间的关系和规律, 完善了公共建筑节能监测信息管理的数据分析功能. 可为建筑能耗分析与节能决策提供一种新的思路和借鉴, 使做出的决策行为基于数据分析, 而不是像过去更多凭借经验和直觉. 同时也弥补了以往的数据分析软件只适用于单机运行, 对海量数据的分析, 出现成本高、效率低的缺陷.

1 关键技术

1.1 Hadoop 技术

1.1.1 Hadoop 概述

Hadoop 是 Apache 软件基金会旗下基于 Java 语言开发的一个开源分布式计算平台. 以 HDFS (Hadoop Distributed File System, Hadoop 分布式文件系统)和 MapReduce 为核心的 Hadoop 为用户提供了系统底层细节透明的分布式基础架构. HDFS 的高容错性、高伸缩性等优点允许用户将 Hadoop 部署在低廉的硬件上, 形成分布式系统; MapReduce 分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序[2]. 所以用户可以利用 Hadoop 组织计算机资源, 从而搭建自己的分布式计算平台, 并且可以充分利用集群的计算和存储能力, 完成海量数据的处理.

1.1.2 Hadoop 的特点

Hadoop 是一个开源的分布式计算平台, 用户可以在 Hadoop 上开发和运行处理海量数据的应用程序. 它主要有以下几个优点[3]:

(1) 可靠性: Hadoop 能自动地维护数据的多份副本, 并且在任务失败后能自动地重新部署计算任务.

(2) 可扩展性: Hadoop 在可用的计算机集群间分配数据并完成计算任务, 这些集群可以方便地扩展到数以千计的节点中.

(3) 高效性: Hadoop 能够在节点之间动态地移动数据, 并保证各个节点的动态平衡, 因此并行处理速度非常快.

(4) 高容错性: Hadoop 能够自动保存数据的多个副本, 并且能够自动将失败的任务重新分配

(5)低成本: Hadoop 集群可以由普通机器组成, 集群可达数千个节点. 并且, Hadoop 是开源平台, 软件成本因此会大大降低.

1.2 数据挖掘技术

1.2.1 数据挖掘的概念

数据挖掘(Data Mining, DM), 是数据库中的知识发现(Knowledge Discover in Database, KDD)的一个步骤, 是目前人工智能和数据库领域研究的热点. 所谓数据挖掘是指从数据库的大量数据中揭示出隐含的、先前未知的并有潜在价值的信息的非平凡过程. 数据挖掘是一种决策支持过程, 它主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等, 高度自动化地分析数据, 做出归纳性的推理, 从中挖掘出潜在的模式, 帮助决策者调整策略, 减少风险, 做出正确的决策[4].

1.2.2 知识发现的过程

知识发现(KDD)的过程, 如图 1 所示, 由数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估、知识表示 7 个步骤迭代组成[5].

图 1 知识发现(KDD)的过程

2 系统设计

2.1 体系结构设计

公共建筑的特点之一是耗电密度高, 开展对公共建筑节能研究是我国建筑节能工作的重点. 山东省公共建筑节能监测平台于 2012 年底投入运行, 2013 年实现了全省联网, 同时, 节约型高校和节约型医院节能监测平台也接入到省级数据中心. 目前, 全省 17 个城市和省直机关均已建立能耗动态监测平台, 累计对 1000 余栋公共建筑进行节能监测. 建筑物节能监测子系统②数据采集周期为平均 5-15 分钟, 进行数据处理后, 每小时向上级数据中心实时报送总能耗、各分项

② 建筑物节能监测子系统的任务是实时记录建筑物用能状况, 自动进行能耗数据处理, 完成建筑能耗结构、建筑用能效率以及建筑节能潜力数据分析, 并将相关统计数据报送上一级数据中心.

能耗、单位面积能耗等统计分析数据，形成建筑物24h/日以及月度、年度统计分析数据[6].

随着时间的推移，越来越多的建筑能耗数据被采集并存储. 但是，大量的能耗数据却带来了“数据灾难”，形成“信息孤岛”. 收集数据不是建立能耗监测平台的目的，目的是发现能耗数据中潜在的、有用的知识，对节能工作提供直接性的决策支持和指导，从而达到节能的目的. 本文提出基于Hadoop分布式计算框架构建公共建筑能耗数据分析系统的方法，集群客户端通过Internet与省级数据中心相连接，读取建筑能耗数据和建筑基本信息表中的数据，并存入 HDFS，实现利用Hadoop集群进行建筑能耗数据挖掘. 本系统可完善山东省公共建筑节能监测信息管理系统的数据分析功能. 系统结构如图 2 所示.

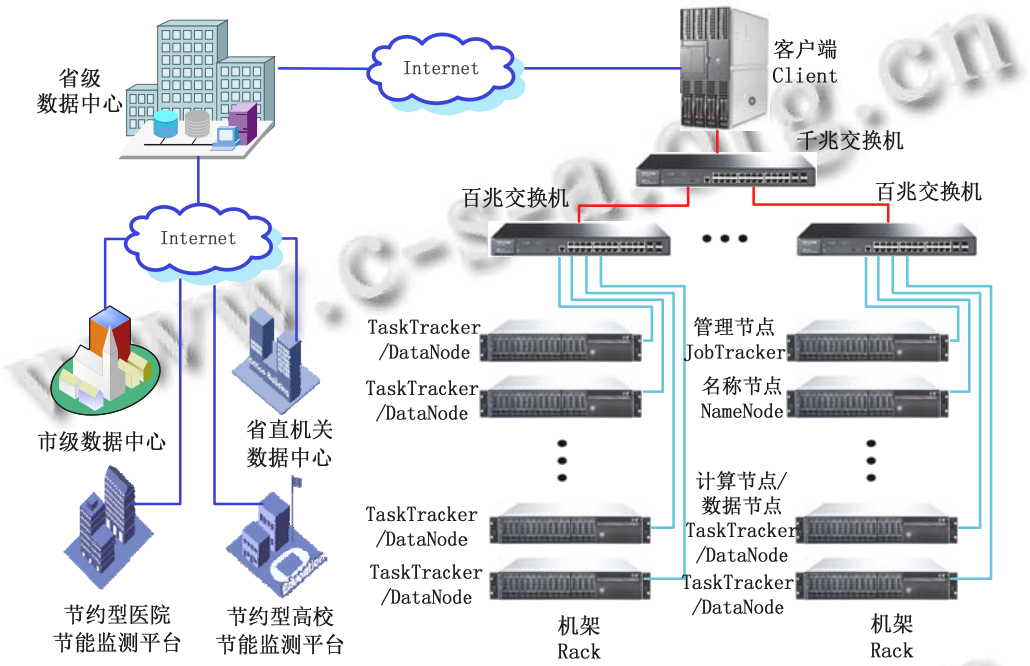

图 2　基于 Hadoop 的公共建筑能耗数据挖掘系统结构

系统结构图右侧是由多个机架组成的集群，为一个典型 Hadoop 集群的构造. 名称节点(NameNode)和管理节点(JobTracker)独占一个服务器，这两类节点作为集群的主节点(Master)，负责管理分布式数据、资源监控和作业调度. 作为主节点的服务器只有少量磁盘储存器，但具有更快的 CPU(中央处理器)及更大的DRAM(动态随机存取存储器)，其他服务器作为从节点(Slaver)，主要负责分布式数据存储以及任务的执行. 各个节点所扮演的角色本质都是 Java 进程，这些进程进行相互调用来实现各自的功能，主节点与从节点一般运行在不同的 Java 虚拟机之中，因此节点之间的通信是跨虚拟机的通信. 为了实现高速通信，一般使用局域网，在内网中使用千兆网卡、高频交换机、光纤等.

2.2 系统硬件结构平台的搭建

本文硬件结构平台的搭建在实验室环境中完成，由四台计算机组成 Hadoop 集群，形成一个 Hadoop 最小规模机架，计算机软、硬件配置要求如下:

计算机配置:　CPU: Intel Pentium E5800(3.2GHz)
内存:4G
硬盘可用容量:500G
计算机系统版本: Linux 系统: Ubuntu 13.04
软件版本:　JDK 版本: jdk-6u31-linux-i586.bin
Hadoop 版本: Hadoop-1.0.0.tar.gz

平台通过 Internet 实现远程数据通信，为保证信息安全，通信系统采用 VPN 技术在集群与省级数据中心之间建立虚拟专用加密通道，保障数据传输安全. 搭建的 Hadoop 实验室集群硬件结构如图 3 所示.

Hadoop 集群采用主从(Master/Slave)体系结构. 四台计算机名称分别定义为 Master、Slaver1、Slaver2、Slaver3，均安装 Linux 系统，通过交换机组成局域网. 同时，确保各台机器之间网络畅通，机器名与 IP 地

址之间解析正确, 从任一台机器都可以 ping 通其它机器的机器名.

将 Master 定义为分布式文件系统 HDFS 的 NameNode(名称结点)及 MapReduce 运行过程中的 JobTracker 结点(任务管理结点), 将 Master 称之为主结点, 主要负责总管分布式数据和分解任务的执行. 其它三台机器 Slaver1、Slaver2、Slaver3 作为 HDFS 的 DataNode(数据结点)以及 MapReduce 运行过程中的 TaskTracker 结点(任务执行结点), 这些结点可统称为从结点, 主要负责分布式数据存储以及任务的执行. 如需要部署更多的机器, 将新加入的机器作为 DataNode 以及 TaskTracker 结点即可. 搭建的 Hadoop 集群节点 IP 地址分布如表 1 所示.

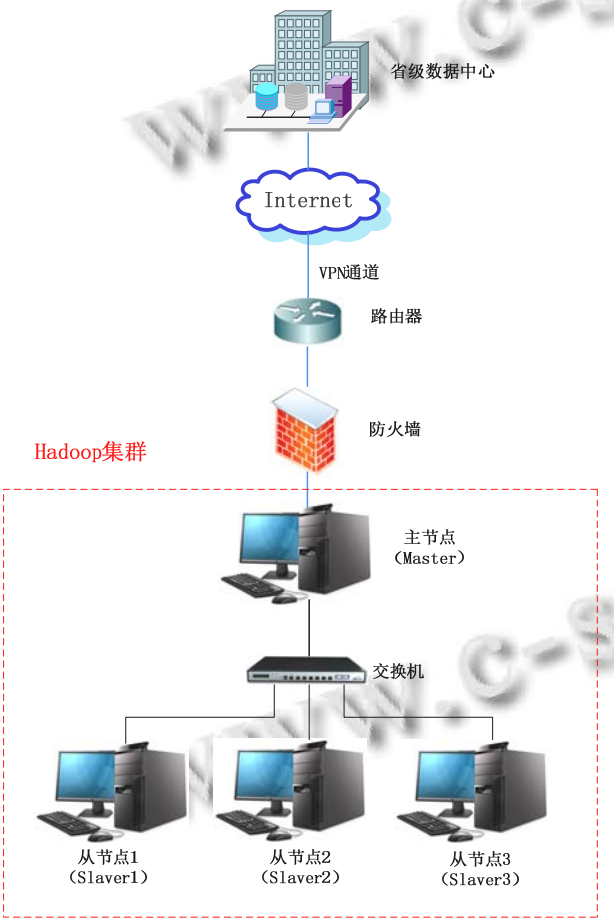

图 3 实验室环境下搭建的 Hadoop 集群

表 1 Hadoop 集群部署规划

序号	机器名称	IP 地址	作 用
1	Master	192.168.211.12	JobTracker NameNode
2	Slaver1	192.168.211.13	TaskTracker DataNode
3	Slaver2	192.168.211.14	TaskTracker DataNode
4	Slaver3	192.168.211.15	TaskTracker DataNode

2.3 系统软件架构及模块功能的说明

根据建筑能耗数据挖掘的特点以及系统的功能, 基于 Hadoop 的公共建筑能耗数据挖掘系统的软件架构模型如图 4 所示. 本架构自上而下依次为交互层、应用层、数据挖掘层、分布式计算层.

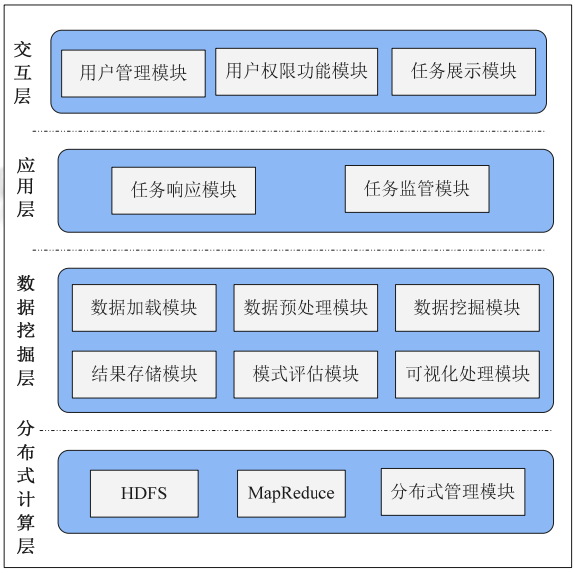

图 4 基于 Hadoop 的建筑能耗数据挖掘系统架构模型

系统架构模型中各层及其模块功能说明如下:

1) 交互层: 系统和用户之间的接口. 通过图形界面, 使得用户可以进入系统查看或者下载能耗信息.

①用户管理模块: 对用户的身份进行识别, 保存用户的注册信息, 以及保存用户的操作记录. 用户身份可分为: 系统管理者、各级政府主管部门、公众用户等.

②用户权限功能模块: 根据用户身份设定相应权限, 不同的用户调用不同的功能模块, 进行权限内的操作.

③展示模块: 根据不同用户的操作请求, 将用户的请求结果以报表、图形等形式反馈给用户.

2) 应用层: 实现对任务的控制和调度. 用户提交的任务在应用层被处理、控制和调度. 应用层通过调用数据分析层的模块来完成交互层提交的任务, 并将结果返回到交互层.

①任务响应模块: 用户提交的操作任务通过交互层传输到应用层, 根据该任务调用数据分析层的相关

模块, 以完成操作.

②任务监管模块: 对任务的执行状态进行监督和管理, 若任务执行过程中进入死循环可终止执行.

3) 数据分析层: 这一层是整个系统的核心, 数据分析层的主要任务是实现各种任务过程中算法的并行化, 并将任务提交到 Hadoop 分布式计算层进行运算, 将结果返回给应用层.

①数据加载模块: 将需要处理的数据加载到系统的分布式文件系统(HDFS)中.

②数据预处理模块:将分布式文件系统中的处理数据进行清理、集成、变换、归约等操作, 以便满足数据分析模块的分析和处理要求.

③数据挖掘模块: 将数据挖掘的各种模式算法封装在此模块中, 如分类模式算法(Classification)、关联模式算法(Association)、序列模式算法(Sequence)等.

④结果存储模块: 存放数据分析模块的分析结果, 即数据挖掘的知识库.

⑤模式评估模块: 将结果模块中的存储结果, 通过人工或结论数据库进行评价.

⑥可视化处理模块: 将通过评估的结论以图形或表格的直观形式呈现出来, 以便能够借此分析或报告数据的特征和数据属性之间的关系.

4) 分布式计算层: 这一层利用 Hadoop 框架来实现集群存储、计算. Hadoop 提供了分布式文件系统和并行运行模式, 同时实现了对分布式系统的管理.

①HDFS: 负责系统所有数据的分布式存储.

②MapReduce: 通过 MapReduce 编程模型实现数据挖掘算法.

③分布式管理模块: 实现对分布式文件系统和并行运行模式管理.

3 基于MapReduce的数据挖掘算法设计

3.1 关联分析

3.1.1 Apriori 算法

Apriori 算法为布尔型关联规则挖掘频繁项集算法, 使用逐层搜索的迭代方法, 利用 k 项集探索 k+1 项集. 首先, 通过扫描数据库, 累计每个项的支持度计数, 并收集满足最小支持度的项, 找出频繁 1 项集的集合记为 L_1. 如果 L_1 非空, 连接 $L_1 \bowtie L_1$ 产生长度为 2 的候选项集合 C_2, 对事务数据库中进行全局扫描, 累加 C_2 中的每个候选项集的支持度计数, 筛选出 C_2 中所有支持度计数满足最小支持度的项集组成长度为 2 的频繁项集 L_2. 用以上步骤重复处理得到新的频繁项集 L_k, 直到没有新的频繁项集产生[7].

3.1.2 算法改进

Apriori 算法找出每个频繁项集 L_k 需要对数据库进行多次全局扫描. 在处理海量数据时, 将会耗费大量的时间和内存[8,9]. 本文通过划分技术, 对 Apriori 算法进行改进, 只需要对所有数据事务进行 2 次全局扫描, 即可挖掘出频繁项集 L_k, 从而提高海量数据挖掘的效率.

改进方法的基本思路如下:

1)假设有 n 个执行 Map 任务的节点, 将数据库中待分析的数据事务平均分为 n 个数据事务子集;

2)每个节点对其数据事务子集进行扫描, 产生该子集的候选 k 项集的集合 C_k^n, 其支持度计数为 1;

3)将每个节点上相同的候选 k 项集的支持度计数累加, 得到候选 k 项集在该节点上的支持度计数 sup_kn;

4)利用 hash()函数将 C_k^n 分成 r 个不同的分区分配到指定的节点上, 同时将其支持度计数 sup_kn 发送到相应节点;

5)r 个节点把具有相同 k 项集的支持度计数累加, 得到最后的实际支持度 sup_k, 当 sup_k 大于等于最小支持度阈值 sup_min 时, 则将此节点上的频繁 k 项集确定为 L_{kr};

6)把 r 各节点产生的所有频繁 k 项集 L_{k1}-L_{kr} 合并, 即可得到全部的频繁 k 项集的集合 L_k.

直到不再产生新的 L_k 算法结束.

基于 MapReduce 编程模型的改进 Apriori 算法流程图如图 5 所示.

3.2 分类分析

3.2.1 决策树的概念

决策树代表对象属性与对象值之间的一种映射关系, 是一个分类、预测模型. 算法首先对数据进行处理, 利用归纳算法生成可读的规则和决策树, 然后使用决策对新数据进行分析. 树中每个节点表示某个对象, 而每个分叉路径则代表对象的某个可能的属性值, 而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值. 常用的决策树算法有: ID3、C4.5 和 CART[5].

3.2.2 C4.5 算法基本思路

由于建筑信息数据的每类信息(属性)有 3-6 种不同的取值, 数据量较大, 并且数据上传存在不完整或者错误, 而C4.5算法能够对不完整训练样本进行处理, 且分类规则易于理解, 准确率较高[10], 因此本课题采用 C4.5 算法, 对建筑信息数据建立分类模型.

运用 C4.5 算法建立决策树分为两个部分, 第一部分是生成决策树, 第二部分是对生成的决策树进行剪枝, 以得到一个精确度较高的、完整的决策树.

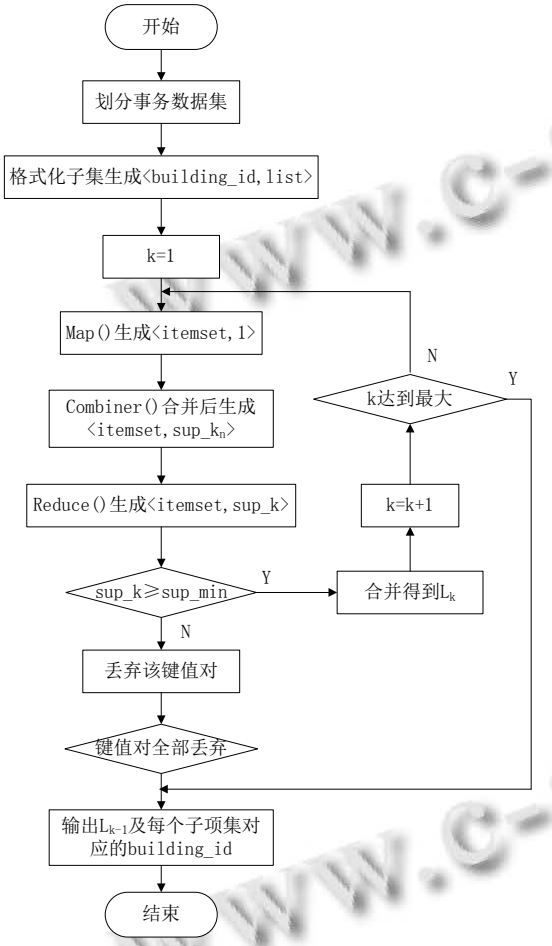

图 5 基于 MapReduce 编程模型的改进 Apriori 算法流程图

C4.5 算法建立决策树步骤[5]:

1) 预处理样本数据集, 计算给定的训练数据集分类的期望信息

$$Info(D) = -\sum_{i=1}^{m} -p_i \log_2(p_i) \tag{1}$$

其中, p_i 是 D 中任意元组属于类 C_i 的非零概率, 并用 $|C_{i,D}|/|D|$ 估计;

2) 计算每个属性的信息增益 $Gain(A)$

基于按 A 划分对 D 的元组分类所需要的期望信息

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{2}$$

信息增益

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

3) 计算每个属性的信息增益率 $GainRatio(A)$ 分裂信息值

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|}) \tag{4}$$

信息增益率

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \tag{5}$$

选择信息增益率最大的属性作为当前的属性节点, 得到决策树的根节点.

4) 根节点属性的每一个取值对应一个子集, 对样本子集递归执行步骤(2), 直到划分的每个子集中的数据在分类或者属性取值相同, 或者没有剩余属性可以进一步划分数据(终止条件), 则停止划分, 生成决策树.

5) 对决策树进行剪枝.

3.2.3 基于 MapReduce 编程模型的 C4.5 算法

由于生成决策树的根节点和每个内部节点需要多次计算, 并且每次计算涉及多种建筑信息属性, 因此将运算过程实现 MapReduce 并行化, 将会减少运算时间, 提高数据挖掘效率.

将生成决策树的过程分为 5 个 MapReduce 作业:

1) 计算三个节能效果(叶节点)的期望信息作业: InfoMapReduce;

2) 计算每个属性(如结构形式、外墙材料类型、保温形式等)的期望信息需求作业: $Info_A$MapReduce;

3) 计算每个属性的信息增益作业: GainMapReduce;

4) 计算每个属性的信息增益率作业: GainRatioMapReduce;

5) 排序作业: RankMapReduce.

InfoMapReduce 作业设计描述如下:

Map()

{统计分配到本节点的数据集 D'的叶节点 low, medium, high 的个数, 生成<key,value>键值对; /key 表示叶节点的取值, value 表示该取值的个数

}

合并函数 combiner( )

{将本地所有 key 值相同的键值对 value 值合并，生成<key,value>键值对;

}

Reduce( )

{将 key 值相同的 value 值相加;

计算对应的 $-p_i \log_2(p_i)$，生成三个<key,$-p_i \log_2(p_i)$>键值对;

将三个键值对的 value 值相加生成< D, info(D)>;

}

其他 4 个 MapReduce 作业设计与 InfoMapReduce 作业类似.

基于 MapReduce 编程模型的 C4.5 算法流程图如图 6 所示. 图 6 中终止条件是划分的每个子集中的数据在分类或者属性取值相同，或者没有剩余属性可以进一步划分.

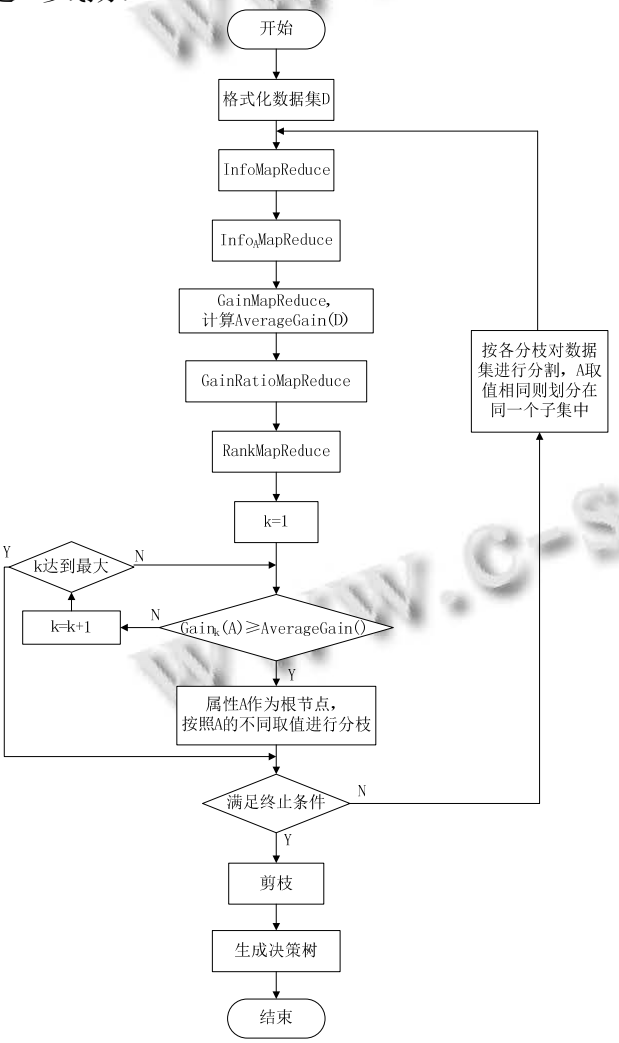

图 6　基于 MapReduce 编程模型的 C4.5 算法流程图

4　实验结果与分析

4.1 数据预处理

办公建筑约占全部监测公共建筑数量的 30%，空调系统能耗在总耗电量中约占 35%，本实验选取山东省 100 栋采用“风机盘管＋新风系统”空调形式的办公建筑，以 2014 年 6 月—9 月四个月的空调制冷期的空调系统耗电量为例进行数据分析. 数据样本主要由建筑基本信息和建筑用电能耗分项数据构成，由数据加载模块将数据资源导入到系统的 HDFS 中，然后数据预处理模块对原始数据进行预处理.

根据 2015 年 5 月 1 日开始实施的《公共建筑能耗远程监测系统技术规程》中对建筑基本信息和建筑附加信息的要求，建筑信息应包括：建筑名称、建筑地址等信息[11]. 该信息由业主单位以.xml 格式文件的形式上传至上级数据中心，经上级监测平台系统解析后生成建筑基本信息报表. 系统通过省级数据中心读取到解析后的某建筑基本信息如图 7 所示.

建筑基本信息

建筑代码：	370303A	建筑类型：	办公建筑	竣工时间：	2006年
建筑名称：	市 大厦		所属单位：	市 大厦	
所属行政区划：	370	建筑地址：	市 区		
总建筑面积：	120000	建筑监测状态：	1	建筑业主：	市 大厦
建筑坐标-经度：	3.14	建筑坐标-纬度：	.45	地上建筑层数：	36
地下建筑层数：	2	空调面积：	100000	采暖面积：	100000
上传日期：	2014/7/9				

建筑围护结构

建筑结构形式：	框架结构	玻璃类型：	普通玻璃	外墙材料形式：	空心黏土砖
窗框材料类型：	铝合金窗	外墙保温形式：	外保温	外窗类型：	中空双层玻璃窗

空调系统

建筑空调形式：	分体式空调或VRV的局部式机组系统	建筑采暖形式：	散热器采暖

图 7　某建筑基本信息表

实验选取与数据挖掘任务相关的 6 类建筑信息属性：建筑结构形式、建筑外墙材料类型、外墙保温形式、外窗类型、玻璃类型、窗框材料类型，并对这 6 类建筑信息进行编码. 对建筑结构形式的编码如表 2 所示，其他 5 类信息编码与表 2 类似.

表 2　建筑结构形式编码

名称	砖混结构	混凝土结构	钢结构	木结构	其他
编码	1A	1B	1C	1D	1E

根据上述编码形式，某建筑信息可表示为：building_56: [1E,2B,3B,4D,5A,6B].

通过数据编码格式化，可将建筑信息统一成适合数据挖掘的形式. 对于建筑信息缺失的能耗数据，直接删除；缺失的空调系统能耗数据，用本月空调系统能耗平均值填补. 在不影响数据挖掘结果质量的前提下，尽可能保持了原数据的完整性.

4.2 关联数据挖掘

100 栋办公建筑能耗数据样本经预处理后, 剩余 97 栋, 其中建筑面积在 10000~20000m^2 的建筑数量占样本总数的 35%, 20000~40000 m^2 占样本总数的 32%. 对数据样本 2014 年空调制冷期(6 月~9 月)的空调系统耗电量进行统计, 得到单位面积(月)耗电量柱状图如图 8 所示, 可将耗电等级划分为: 耗电低、耗电正常、耗电高三个等级.

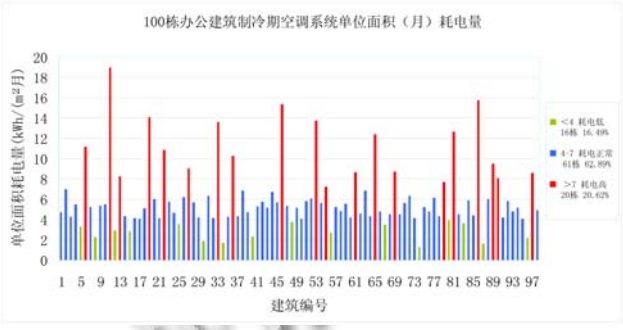

图 8 制冷期空调系统单位面积(月)耗电量柱状图

对经过编码的 97 栋办公建筑的基本信息应用改进的 Apriori 算法进行频繁项挖掘, 可得到若干频繁项集及对应的建筑编号如表 3 所示.

表 3 频繁项集

频繁项集	支持度	建筑编号
<1B,2A,3B,4C,5C,6D>	8	3,5,13,21,32,62,78,93
<1B,2A,3B,4C,5A,6B>	12	2,4,7,17,23,27,42,45,51,58,75,88
<1B,2B,3B,4C,5B>	13	7,9,15,18,20,26,28,31,39,47,56,61,85
<1B,2D,3B,4E,6D>	7	8,12,29,44,71,87,92
<1C,2E,3B,5C,6B>	4	6,33,69,81
……	……	……

数据样本中有 93 栋办公建筑采用混凝土结构和外保温形式. 将 Apriori 算法得到的结果与制冷期空调系统单位面积(月)耗电量数据相结合可得到以下结论: 在其他建筑信息相同的情况下, 采用普通玻璃的建筑空调系统单位面积能耗略高于 Low-E 玻璃建筑; 窗框材料采用断热窗的建筑空调系统单位面积能耗低于普通铝合金窗建筑; 采用玻璃幕墙的建筑空调系统单位面积能耗较高; 窗墙比越大的建筑空调系统单位面积能耗越高. 该结论与国内建筑能耗影响因素研究结论基本一致[12-15].

4.3 分类(决策树)数据挖掘

结合 Apriori 算法的结论, 对经过编码的 97 栋办公建筑数据样本应用基于 MapReduce 编程模型的 C4.5 算法进行分类, 生成的空调系统耗电量判定树, 如图 9 所示. 将数据样本中 60 组数据作为训练集, 其余 37 组数据作为检验集, 经检验分类错误率为 31.58%. 影响建筑能耗的因素之间存在耦合关系, 并且本实验考察的因素只有 6 类建筑信息, 因此分类错误率较高.

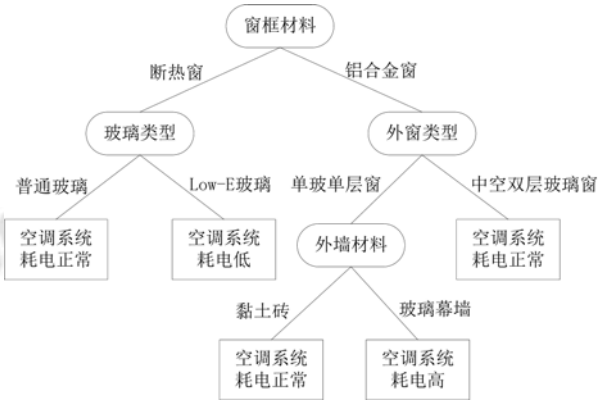

图 9 空调系统耗电量判定树

根据判定树及图 8 统计数据可对样本建筑节能改造工作提供改造建议, 如 26 号办公建筑制冷期空调系统单位面积(月)耗电量为 6.19kWh/(m^2 月)属于正常耗电范围, 若进行节能改造, 可将普通玻璃更换为 Low-E 玻璃; 36 号办公建筑制冷期空调系统单位面积(月)耗电量为 10.25kWh/(m^2 月)属于高耗电, 若进行节能改造, 可将铝合金窗更换为断热窗, 单玻单层窗更换为中空双层玻璃窗. 在新建建筑设计中, 尽量采用中空玻璃窗、Low-E 玻璃、断热窗框, 而尽可能少用玻璃幕墙作为外墙材料.

5 结语

本文在对 Hadoop 分布式计算平台深入分析研究的基础上, 针对大型公共建筑能耗数据分析工作的迫切需求, 提出利用大数据处理技术的 Hadoop 分布式架构, 将建筑基本信息与建筑电量能耗数据相结合进行数据挖掘研究的方法, 对基于 Hadoop 的公共建筑能耗数据挖掘系统进行了初步设计, 并对系统的基本架构和各模块的功能进行了设计和说明. 以山东省 100 栋办公建筑制冷期的空调系统耗电量为例, 运用基于 MapReduce 编程模型的 Apriori 算法和 C4.5 算法对样本数据进行分布式处理, 得到 6 类建筑信息属性对空调系统能耗的影响规则, 并生成空调系统耗电量判定

树, 根据统计数据和判定树可判断建筑空调系统耗电量等级, 并对样本建筑的节能改造提供具有针对性的建议. 若加入更多影响因素进行考察, 系统将得到更多的影响规则, 生成的决策树精度更高, 更加完整, 也是本课题需要改进之处.

本系统弥补了传统数据挖掘方法对海量数据处理的缺陷, 通过分布式处理提高了数据挖掘效率, 对山东省公共建筑节能监测信息管理系统的数据分析功能进行了完善. 从数据的角度客观地分析能耗情况并提供节能决策支持, 从而达到节能的目的. 课题所涉及的思想和方法可推广应用到各类建筑能耗数据挖掘工作中, 对既有建筑的节能改造和新建建筑的节能设计提供参考和借鉴.

参考文献

1 Bryant RE, Katz RH, Lazowska ED. Big data computing: Creating revolutionary breakthroughs in commerce, science, and society. http://www.cra.org/ccc/docs/init/Big_Data.pdf. [2012-10-02].

2 White T. Hadoop: The Definitive Guide.3rd ed., USA: O'Reilly Media, Inc, 2012: 3–6.

3 陆嘉恒.Hadoop 实战.北京:机械工业出版社,2012.

4 互动百科.数据挖掘.http://www.baike.com/wiki/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98.[2015-5-22].

5 Han JW, Kamber M.数据挖掘:概念与技术.第 3 版.范明,孟小峰译.北京:机械工业出版社,2012.

6 山东省建设发展研究院.山东建筑大学.公共建筑节能监测系统技术规范.DBJ/T14-071-2010.

7 Hand D, Mannila H, Smyth P.数据挖掘原理,张银奎,廖丽,宋俊,等译.北京:机械工业出版社,2004.

8 Mohammed AM, Bassam A. An Improved Apriori Algorithm for Association Rules. e-print arXiv:1403.3948, 2014.

9 李玲娟,张敏.云计算环境下关联规则挖掘算法的研究.计算机技术与发展,2011,21(2):43–46.

10 Taherkhani A. Recognizing sorting algorithms with the C4.5 decision tree classifier. ICPC, International Conference on Program Comprehension. 2010. 72–75.

11 深圳市建筑科学研究院股份有限公司.中国建筑科学研究院.公共建筑能耗远程监测系统技术规程.JGJ/T 285–2014.

12 韩保华,马秀力,王成霞,刘婷.国家机关办公建筑能耗现状与节能对策研究.建筑科学,2010,26(2):59–61.

13 陈高峰,张欢,由世均,叶天震,谢真辉.天津市办公建筑能耗调研及分析.暖通空调,2012,42(7):125–128.

14 王春雷.夏热冬暖地区大型办公建筑能耗影响因素研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2010.

15 郝明慧.济南地区办公建筑能耗模拟与节能分析[硕士学位论文].济南:山东建筑大学,2011.