Hans Journal of Wireless Communications 无线通信, 2016, 6(3), 73-80
Published Online June 2016 in Hans. http://www.hanspub.org/journal/hjwc
http://dx.doi.org/10.12677/hjwc.2016.63010

Hans汉斯

Research on Relevance between Atmospheric Temperature and Surface Temperature in Forests Based on Hadoop

Bowen Yang, Ziyan Wang, Wenjing Xun, Xiaofeng Liu, Zhengli Zhu

School of Information Science and Technology, Nanjing Forestry University, Nanjing Jiangsu
Email: liuxiaofeng@njfu.edu.cn

Received: Jun. 9th, 2016; accepted: Jun. 25th, 2016; published: Jun. 28th, 2016

Abstract

Base on the big data about the atmospheric temperature and surface temperature in the forest of Purple Mountain in Nanjing, the paper proposes a method to analyze the data of forestry Internet of things by the cloud computing. The paper uses Alphabet's Hadoop cloud computing platform to analyze these data, and studies the relationship between atmospheric temperature and surface temperature. Framework of the MapReduce is used to carry out the data of the sensor data processing. Then, we use MATLAB to analyze the relationship between atmospheric temperature and surface temperature. Land surface temperature is considered of strategic importance in the field of plant growth, climatology, and biological research, and it also has important value in Agricultural Meteorology.

Keywords

Cloud Computing, Big Data, Internet of Things, Hadoop, MapReduce

Hadoop平台下森林大气温度与地表温度关联研究

杨博文，汪子炎，荀文婧，刘晓峰，朱正礼

文章引用: 杨博文, 汪子炎, 荀文婧, 刘晓峰, 朱正礼. Hadoop 平台下森林大气温度与地表温度关联研究[J]. 无线通信, 2016, 6(3): 73-80. http://dx.doi.org/10.12677/hjwc.2016.63010

南京林业大学信息科学与技术学院，江苏 南京
Email: liuxiaofeng@njfu.edu.cn

收稿日期：2016年6月9日；录用日期：2016年6月25日；发布日期：2016年6月28日

摘 要

本文基于南京紫金山地区森林关于大气温度、地表下5 cm处的土壤温度的大数据，对传统的数据分析方法作出改进，提出了用云计算对林业物联网数据进行分析的方法，运用Alphabet公司的Hadoop云计算平台的MapReduce大数据处理框架，研究大气温度和地表下5 cm处的土壤温度间的关系。本文利用Hadoop云计算平台下的MapReduce框架对林业物联网大数据进行数据处理，并利用MATLAB软件对两者进行分析研究，进而研究大气温度对地表温度的影响。地表温度在植物生长、气候学、生物化学研究中具有重要的意义，在农业气象中有重要的应用价值[1]。

关键词

云计算，大数据，物联网，Hadoop，MapReduce

1. 引言

地表温度在地球物理参数中具有非凡的意义，它是全球地表物理过程中的一个非常关键的参数。在一些农业干旱监测模型、森林火灾预测等模型中都把地表温度参数考虑在内[2]。近年来，为研究气温与地表温度的关系，国内外学者发展了直接由气温或土表温度反演土壤温度的多种模式，但是这种情况下，土壤温度必定没有实际测量的精准。传统模式的土壤温度来自于实际测量，这种方法虽然耗时费力，但很精准。本文就是基于南京紫金山地区森林实际测量的大气温度和地表温度的大数据，利用 Hadoop 云计算平台研究它们的关系。传统上讲，地表温度与大气温度、土壤湿度、光照强度、经度纬度等一系列有关。本研究抛弃复杂的因素，利用 Hadoop 平台研究南京紫金山地区森林海拔在 500 米以下的大气温度和地表下 5 cm 处土壤温度的关系。

Hadoop 是一个开源的软件平台，它使得编写以及应用于处理大数据的应用或者程序更加简便。它是一个很方便简洁地方便编程人员开发并行处理大规模数据的分布式云平台。它的主要优势在于扩展性好、开源、成本低廉、可靠等。总之，它是分析海量数据的重要工具之一[3]。针对南京紫金山地区森林关于大气温度和地表下 5 cm 处的土壤温度的大数据，为研究它们之间的关系，传统的数据分析方法并不适用，为此本研究选用了 Hadoop 云计算平台。

该研究会应用 Hadoop 平台的 MapReduce 框架对数据进行降噪，去除一部分无效数据。然后会依次按天、按月分析大气温度和地表下 5 cm 处的土壤温度。具体分析方法是，对于日数据，利用 MapReduce 剔除一部分无效数据后，用 MATLAB 软件构建两者间的关系模型；对于月数据，会利用 MapReduce 计算出月平均大气温度、地表下 5 cm 处的月平均土壤温度，再利用 MATLAB 构建出月平均数据模型，进而研究两者间的关系。

2. 数据的采集

在南京紫金山地区森林中，存在多个传感器，以此构建一个林业物联网。每一个传感器每隔一段时间便会向互联网发送一个数据包，其中数据包中会包含传感器编号、记录的时间、当前森林中的大气温

度、地表下 5 cm 处的土壤温度等等。原始的数据会存储在数据表中，但是原始的数据并不一定真实可靠，在原始数据中会发现一些噪声数据，比如发现大气温度在 100 摄氏度以上的数据，这些噪声数据一般是由于传感器损坏、传感器无电等原因造成的。在实际的数据分析中，去除这些噪声数据很有必要。对海量数据的处理，传统方法效率过于低下，难以胜任此工作，为此在本研究中采用了 Hadoop 的 MapReduce 框架进行处理。

3. 数据处理方法

3.1. MapReduce 编程模型

MapReduce 是 Hadoop 的核心设计，它是一种用于并行计算的分布式编程模型框架。MapReduce 借助于函数式的编程思想，把海量数据的处理方式抽象成 Map 和 Reduce 等操作。MapReduce 模型大致分为 3 个阶段，分别是数据初始化阶段，最为核心的 Map 和 Reduce 阶段，以及数据结果汇集阶段。MapReduce 使用的是“分而治之”的处理思想[4]，简而言之，它把一个海量数据的处理任务分解为多个数量多但任务规模小的处理任务，将每个小规模的处理任务分发给多个节点，让多个节点共同完成，最后再将结果汇总[5]。这样能大大地减少单个节点的任务压力，从而达到了提高运算效率的效果。

具体的执行过程如下：

1) 数据预处理：MapReduce 框架将海量数据分成若干片。

2) 任务分配：JobTracker 会将步骤 1 中的数据分片分配给集群中的若干节点，使之执行 Map 或者 Reduce 任务。

3) Map 阶段：节点将输入的数据分片进行转化，转化为<K,V>的键值对，Map 接口对数据分片进行处理，并且以一个新的<K,V>键值对输出到中间接口，再进而处理，最后把处理结果输出到 Reduce 接口中。

4) Reduce 阶段：Reduce 接口会读取步骤 4 的输出结果，并重新计算，将计算结果输出。

5) 所有任务结束，系统通知用户并报告执行结果等信息，从而整个 MapReduce 任务完成[6]。详细过程见图 1。大数据集(big Data)分为了 Split1、Split2 等，然后各个分片被分到了 Map1、Map2 等节点，数据转化为键值对后，经过 Map 过程的处理，处理结果传递到 Reduce1、Reduce2 等节点，处理之后就得到了各个处理结果 Output1、Output2 等。

3.2. 基于 MapReduce 的数据去噪

处于森林之中的传感器十分脆弱，传感器损坏、传感器电量不足时会传回一些噪声数据，处理噪声数据在研究中十分重要。

通过对传感器传回的数据研究发现，存在大气温度在 100℃以上的数据，对于这部分不合法的数据，我们必须舍去。利用 MapReduce 框架进行数据去噪操作十分简单，由于气温在正常情况下具有一定的取值范围，对于南京地区而言，气温一般在−10℃到 45℃之间。

具体实现方式是利用 MapReduce 框架，在 Map 过程中，如果大气温度或者突然温度超出了该范围，则在 Map 过程直接舍弃该数据即可，不必输出键值对[7] [8]。经过处理后的数据我们再进行下一步的操作。具体实现代码如图 2。

3.3. 基于 MapReduce 的双重排序

在对原始数据进行初步去噪后，要研究森林中大气温度和地表 5 cm 下的土壤温度的日关系，需要把相同传感器传回的数据放在一起，并且按时间排序。由于不同传感器处于森林中不同的地理位置，每两

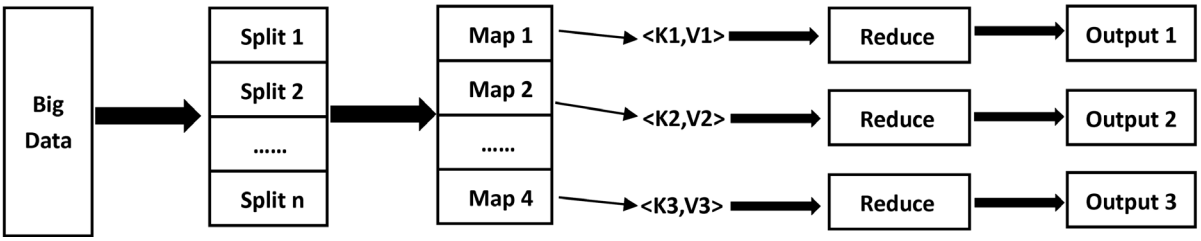

Figure 1. Task processing flow diagram in MapReduce framework
图 1. MapReduc 任务处理流程示意图

```
Map 过程代码
Protected void map(LongWritable key,Text value,Context context)
{
//value 数据格式为:传感器编号+"\t"+大气温度+"\t"+土壤温度+"\t"+时间
String[] data=value.toString().split("\t");
//大气温度或者土壤温度不在合理范围时则直接丢掉该数据
if(data[1]>-10 and data[1]<45 and data[2] >-10 and data[2]<45)
{
        line=value;
        context.write (line, "" );
}
}
Reduce 过程代码
Protected void reduce(LongWritable key,Iterable< LongWritable> value,Context context)
{
context.write (key, "" );
}
```

Figure 2. Code of data denoising in MapReduce framework
图 2. MapReduce 的数据去噪代码示意图

个传感器记录的温度可能有不同，为了结果的精确，而研究同一个传感器大气温度和土壤温度的关系。这就要求对数据进行双重排序。

基于 MapReduce 双重排序，同样需要把数据转化为<K,V>的键值对，其核心思想是把两个数据组合抽象成一个数据类型，通俗来讲，就是把两个数据当成一个 K 值来处理，这就需要用户自定义此类型的排序方法[9]。在双重排序的 map 过程中，键值对的表现形式为<<key1,key2>,value1>，在该 map 过程中，要先对 key1 的值进行排序。接着按照 key1 的值，JobTracker 会把数据分为若干分片，每一分片会被分派到集群中的一个 Reduce 节点。接着，每一个 Reduce 节点会对输入的键值对继续分组，对于 key1 相同的数据则分为一组。最后，再对 key2 进行排序。其详细算法流程图见图 3。首先数据按第一列进行排序，然后根据第一列的值把数据分到不同的节点，比如图中第一列为 2，4 的数据分到了一个节点，而第一列为 5 的数据分到了另一个节点；然后每一个节点根据第二列的值再进行从小到达排序，最后得到结果。在该流程图中，第一个参数相当于传感器编号，第二个参数相当于时间，第三个参数相当于不参与排序的大气温度和地表温度。

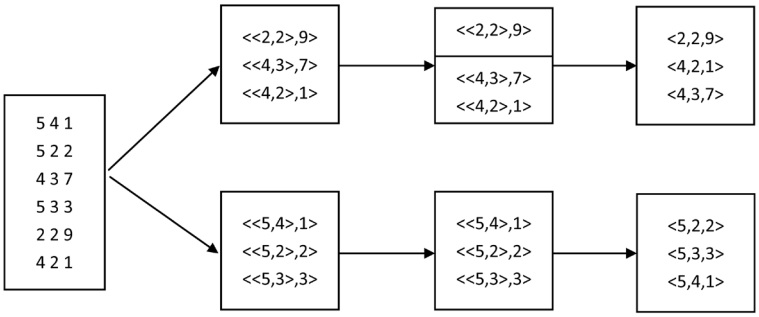

Figure 3. Secondary Sort flow diagram in MapReduce framework
图 3. MapReduce 的双重排序流程图

3.4. 基于 MapReduce 的均值计算

为研究南京紫金山地区森林下的月平均大气温度和地表温度之间的关系，需要对大气温度和地表温度进行求均值操作。对于经过排序操作后的数据，我们对每个节点的大气温度、地表温度按月进行求均值。

均值计算的算法思路是，在 map 过程中，将不需要被求均值的字段当作 map 过程的 K 值输出，将需要被平均的字段当作 V 值输出；在 reduce 过程中，计算出要被平均字段的综合，以及被除数，然后计算出平均数即可，Reduce 输出的 K 值为 map 过程的 K 值，而 V 值自然就是平均值[10]。图 4 给出了计算均值的伪代码。

4. 数据的分析

4.1. 按天分析大气温度和地表温度的关系

该研究分析了南京 2014 年紫金山地区森林的气温和地表下 5 cm 处的温度每日数据，并在 MATLAB 中构建了数据分析模型。利用 MapReduce 框架处理得到的按小时、按天、按月得到的大气平均温度和地表平均温度数据，进行图像拟合等，用 MatLab 画出二维图像，从而建立了该数据分析模型。通过构建的图像发现气温和土壤温度存在明显的季节特征，并且土壤温度的变化曲线和气温变化曲线在总体上一致，但时间先后顺序上却不一致。由于数据量大，该研究选取了几幅具有代表意义的气温和地表 5 cm 处的关系图。详见图 5 到图 8。

从日变化曲线上看，气温和地表 5 cm 下的温度存在明显的季节特征，在夏季，地表温度在大部分时间都低于大气温度；而在春、秋、冬季，地表温度一般都高于气温。另外，地表温度变化相较于气温变化，曲线上更加平滑，说明了土壤具有一定的保温效果。地表温度的变化趋势和大气温度基本一致，但是从时间上来看，变化趋势一般晚于气温 2~4 小时。

4.2. 按月分析大气温度和地表温度的关系

利用 MapReduce 求出月平均气温和地表温度后，进一步地研究发现，地表温度在除了 6~8 月外低于气温，在其它月份地表温度均高于气温。这可能是因为在森林里，地表接触不到日照，所以温度较低。另外，地表温度和气温的温差一般在 6℃以内。2015 年南京地区某森林关于气温、地表温度的关系图，见图 9。

5. 结论

1) 在该研究区内地表下 5 cm 处的土壤温度和气温密切相关，并且土壤温度的变化曲线和气温变化曲线在总体上一致，但时间先后顺序上却不一致，土壤温度变化趋势一般晚于气温 2~4 小时。

```
Map 过程代码
Protected void map(LongWritable key,Text value,Context context)
{       key1=编号+","+月份;
value1=大气温度+","+土壤温度;
context.write (key1,value1 );
}
Reduce 过程代码
Protected void reduce(LongWritable key,Iterable< LongWritable> value,Context context)
{
While(value.hasnext())
{       count++;//统计总共有多少行
        sum1+=大气温度;
        sum2+=土壤温度;
}
avg1=sum1/count;avg2=sum2/count;value2=avg1+","+avg2;//value2 包含两个平均数
context.write (key, value2 );
}
```

Figure 4. Mean value algorithm in MapReduce framework
图 4. MapReduce 的均值算法示意图

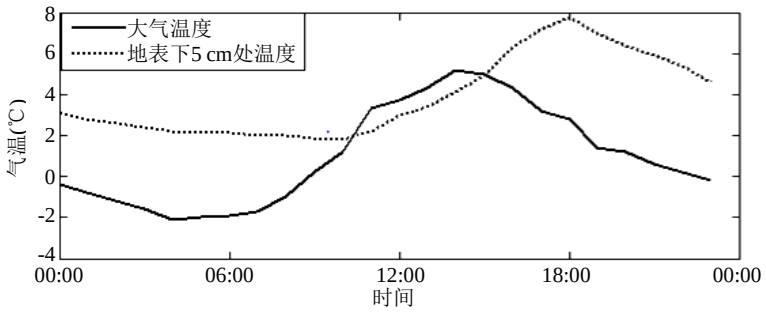

Figure 5. Relationship diagram between atmospheric temperature and soil temperature in Nanjing's forestry in January 2015
图 5. 2015 年 1 月某日南京地区某森林关于气温、地表温度的关系图

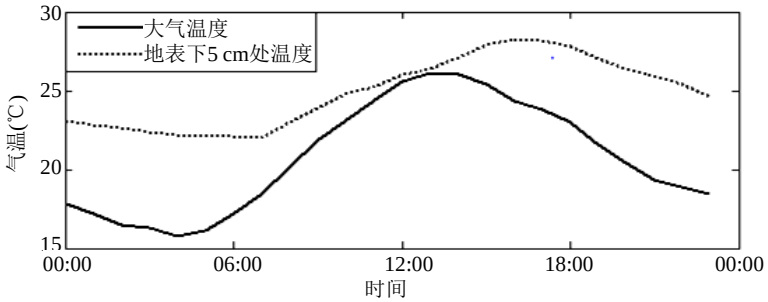

Figure 6. Relationship diagram between atmospheric temperature and soil temperature in Nanjing's forestry in April 2015
图 6. 2015 年 4 月某日南京地区某森林关于气温、地表温度的关系图

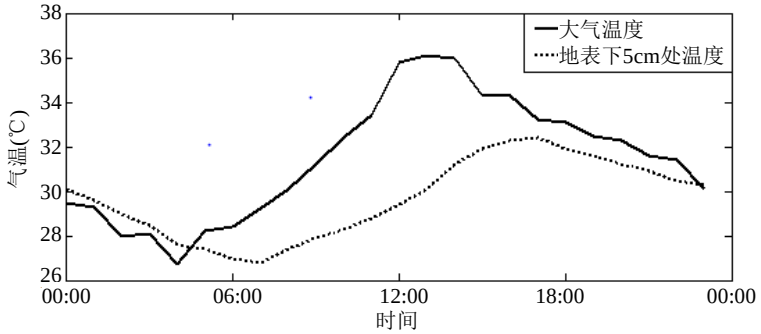

Figure 7. Relationship diagram between atmospheric temperature and soil temperature in Nanjing's forestry in July 2015
图 7. 2015 年 7 月某日南京地区某森林关于气温、地表温度的关系图

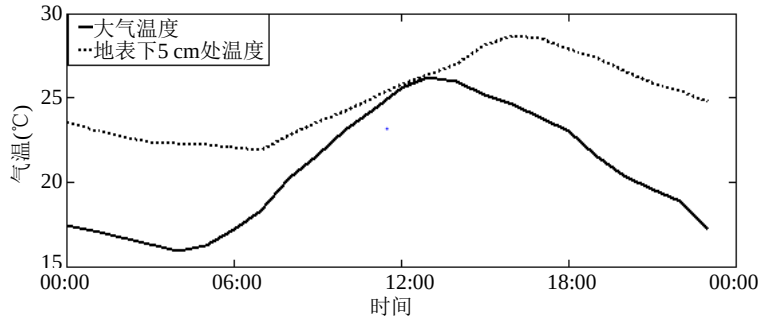

Figure 8. Relationship diagram between atmospheric temperature and soil temperature in Nanjing's forestry in October 2015
图 8. 2015 年 10 月某日南京地区某森林关于气温、地表温度的关系图

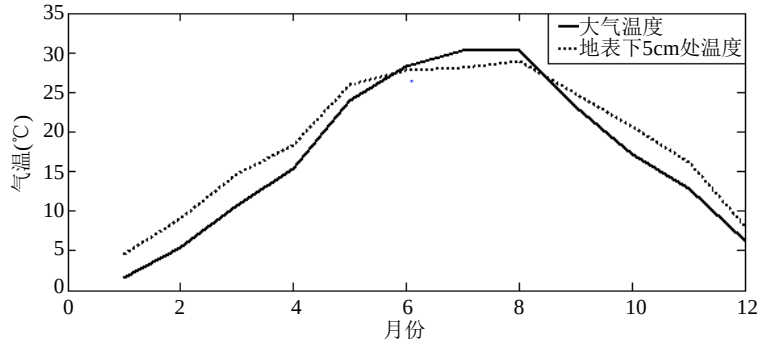

Figure 9. Relationship diagram between atmospheric temperature and soil temperature in Nanjing's forestry in 2015
图 9. 2015 年南京地区某森林关于气温、地表温度的关系图

2) 在研究区内，气温和土壤温度存在明显的季节特征，一般来说，夏季土壤温度比气温低，而春、秋、冬季土壤温度比气温要高。

3) 在该研究区域内，地表温度和气温的温差一般在 6℃以内。

该研究利用 Hadoop 平台进行大数据的相关数处理，研究了气温和地表下 5 cm 处温度的关系。但是此研究仍然有一些不足之处，一方面，没有考虑到天气情况对地表温度的影响；另一方面，该研究是否具有普遍性，是否同样适用于其它地区的森林不可知。因此，利用 Hadoop 平台讨论其它森林气温和地表温度之间的关系是将来研究工作的方向之一。

致 谢

感谢国家自然科学基金(31300472)，江苏省大学生实践创新训练计划重点项目(201410298030Z, 201510298046Z)，江苏高校品牌专业建设工程资助项目(TAPP: Top-notch Academic Programs Projects of Jiangsu Higher Education Institutions)。

参考文献 (References)

[1] 张芳. 东亚地区地表温度与大气温度场的相关关系[J]. 青海科技, 2009, 16(6): 67-70.

[2] 张树誉, 杜继稳, 景毅刚. 基于 MODIS 资料的遥感干旱监测业务化方法研究[J]. 干旱地区农业研究, 2006, 24(3): 1-6.

[3] 刘鹏. 实战 Hadoop [M]. 北京: 电子工业出版社, 2011: 4-5, 60-61.

[4] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, **51**, 107-112. http://dx.doi.org/10.1145/1327452.1327492

[5] 杨博文. Hadoop 平台下森林大气温度与地表温度关联研究[J]. 电脑知识与技术, 2015,(30): 2-3.

[6] 朱珠. 基于 Hadoop 的海量数据处理模型研究和应用[D]: [硕士学位论文]. 北京: 北京邮电大学, 2008: 33-37.

[7] 徐文龙. 基于 Hadoop 分布式系统的重复数据监测技术研究与应用[D]: [硕士学位论文]. 湖南: 湖南大学, 2013: 16-17.

[8] 卢永菁. 一种高性能重复数据删除系统设计及研究[D] : [硕士学位论文]. 湖南: 湖南大学, 2013: 17-20.

[9] 路秋瑞. 基于 Hadoop 的大规模数据排序算法的研究[J]. 信息与电脑, 2015(17): 110-112.

[10] MapReduce Tutorial. http://hadoop.apache.Org/docs/rl.2.l/mapred_tutorial.html