

文章编号:1671-251X(2018)11-0080-04

DOI:10.13272/j.issn.1671-251x.2018040090

煤矿综采设备运行状态大数据清洗建模

马宏伟, 吴少杰, 曹现刚, 徐博远, 张鑫媛

(西安科技大学 机械工程学院, 陕西 西安 710054)

摘要:针对煤矿综采设备运行状态数据量大、数据存在噪声和缺失值等问题,建立了一种基于MapReduce的煤矿综采设备运行状态大数据清洗模型。该模型采用双MapReduce协同工作:通过第1个MapReduce对数据中的噪声点和缺失值进行修正,输出多个清洗后的数据文件;通过第2个MapReduce对多个清洗后的数据文件按采集时间及日期进行排序,并合并成单个数据文件输出。实验结果表明,该模型能有效剔除噪声数据和补全缺失数据,具有较好的数据清洗效果。

关键词:综采设备;设备运行状态;大数据;数据清洗;MapReduce

中图分类号:TD67 文献标志码:A

网络出版地址:<http://kns.cnki.net/kcms/detail/32.1627.TP.20181026.1500.008.html>

Big data cleaning modeling of operation status of coal mine fully-mechanized coal mining equipment

MA Hongwei, WU Shaojie, CAO Xiangang, XU Boyuan, ZHANG Xinyuan

(College of Mechanical Engineering, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract:In view of problems of large amount of data and noise and missing values existed in data of operation status of coal mine fully-mechanized coal mining equipment, a big data cleaning model of operation status of coal mine fully-mechanized coal mining equipment based on MapReduce was established. The model is composed of dual MapReduce. Noise points and missing values in data are corrected and multiple cleaned data files are output through the first MapReduce. The multiple cleaned data files are sorted according to collection time and date and combined into a single data file through the second MapReduce. The experimental results show that the model can effectively eliminate noise data and complement missing data with good data cleaning effect.

Key words: fully-mechanized coal mining equipment; equipment operation status; big data; data cleaning; MapReduce

0 引言

在煤炭行业,随着机器智能化程度的不断提高,管理方式由粗放型转变为精细化,中国大中型煤矿企业由人产生的数据规模一般在TB级别,由设备产生的数据规模达PB级别^[1-3]。煤矿综采设备工

作环境复杂、恶劣,设备运行状态数据在采集、传输过程中会受到影响,使得噪声和缺失值存在于数据中^[4-5],导致数据质量不能达到后续数据分析的要求。因此,对原始数据进行清洗是必不可少的^[6-8]。

杨东华等^[9]针对基于MapReduce的数据清洗过程中存在计算冗余的问题,提出了一种任务合并

收稿日期:2018-04-28;修回日期:2018-10-12;责任编辑:盛男。

基金项目:国家自然科学基金面上项目(51875451)。

作者简介:马宏伟(1957—),男,陕西兴平人,教授,博士,博士研究生导师,主要研究方向为智能检测与控制,煤矿机电设备及其自动化、智能化等,E-mail:mahw@xust.edu.cn。

引用格式:马宏伟,吴少杰,曹现刚,等.煤矿综采设备运行状态大数据清洗建模[J].工矿自动化,2018,44(11):80-83.

MA Hongwei, WU Shaojie, CAO Xiangang, et al. Big data cleaning modeling of operation status of coal mine fully-mechanized coal mining equipment[J]. Industry and Mine Automation, 2018, 44(11): 80-83.

的优化技术,将数据中的冗余计算和利用统一输入文件的简单计算进行合并,从而减少数据清洗时间。陈少龙^[10]建立了基于 Hadoop 平台的煤矿设备数据清洗系统,利用 HDFS(Hadoop Distributed File System, Hadoop 分布式文件系统)和 MapReduce 实现对煤矿设备数据的海量存储和异常点清洗。在上述研究的基础上,本文提出了一种基于 MapReduce 的煤矿综采设备运行状态大数据清洗模型,旨在提高数据清洗效果和清洗效率,为综采设备健康状态评估及故障诊断提供有效数据支撑。

1 MapReduce 工作流程

MapReduce 是 Hadoop 的核心之一,是面向海量数据并行处理的计算模型、框架和平台^[11-12]。MapReduce 的核心思想是 Map(映射)和 Reduce(归约),在 MapReduce 执行过程中包含 Map 阶段和 Reduce 阶段^[13],数据以键值对(key, value)形式进行传输。用户通过 Map 函数实现对键值对的高度并行化处理,生成中间键值对进行输出,Reduce 函数接收这些中间键值对,并按照键的相同与否进行适当合并,最终输出用户期望的结果。

在 Map 阶段,将数据划分成 m 块,并将数据块合理分配给各节点,通过节点的并行处理完成 Map 任务;在 Reduce 阶段,将数据划分为 r 块,同样通过节点的并行化处理完成 Reduce 任务。MapReduce 工作流程^[14-15]如图 1 所示,具体步骤如下。

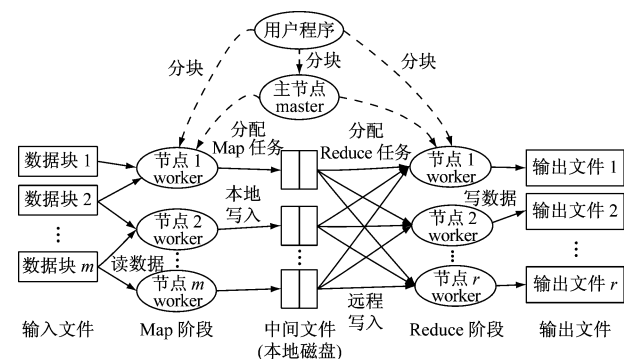


图 1 MapReduce 工作流程
Fig. 1 Work flow of MapReduce

(1) 用户程序中的 MapReduce 函数库自动将数据分割成 m 块,一般每个数据块为 16~64 MB。随后,计算机集群复制执行程序,并在集群内的各计算机上启动程序副本。

(2) 在这些程序副本中,有特殊的程序 master 和普通的程序 worker, master 负责为 worker 分配工作—— m 个 Map 任务和 r 个 Reduce 任务。master 选取空闲的 worker,分配单个 Map 任务或 Reduce 任务。

(3) worker 接收到 Map 任务时,读取相应数据

块,从中解析出若干键值对并传递给用户提供的 Map 函数。Map 函数进行处理后,将生成的中间键值对存储在内存中。

(4) MapReduce 函数库周期性地将内存中的数据写入本地磁盘,并利用分区函数将数据分别写入 r 个不同的区域,最后将数据在磁盘上的位置发送给 master。

(5) master 将位置信息通知给执行 Reduce 函数的 worker,后者利用远程过程调用(Remote Procedure Call, RPC)从 Map 阶段 worker 的本地磁盘读取数据。数据读取完毕后,按照键进行排序,使相同键的数据排列在一起。

(6) Reduce 阶段 worker 对排序后的数据展开迭代,将每个键及其对应的值传递给用户提供的 Reduce 函数。然后将 Reduce 函数的输出附加到本分区的最终输出文件内。

当所有 Map/Reduce 阶段全部完成后,能够获得 r 个输出文件(每个 Reduce 任务输出 1 个文件)。

2 基于 MapReduce 的煤矿综采设备运行状态大数据清洗建模

基于 MapReduce 的煤矿综采设备运行状态大数据清洗建模流程如图 2 所示,具体步骤如下。

(1) 用户程序中的 MapReduce 函数库将数据划分成 m 块,对需要清洗的数据进行初始化,生成能够被 MapReduce 处理的键值对(k_1, v_1),其中 k_1 为数据采集时间, v_1 为采集时间对应的状态量等信息,将键值对以文件的形式存储在 HDFS 中。

(2) 第 1 个 MapReduce 中的 Map 函数接收步骤(1)产生的键值对(k_1, v_1),建立不含异常值的时间序列模型,根据时间序列模型计算残差:

$$X = \frac{\theta(B)}{\varphi(B)\nabla^d} a \quad (1)$$

式中: B 为延迟算子; $\theta(B)$ 为没有公共因子的可逆算子; $\varphi(B)$ 为没有公共因子的平稳算子; $\nabla = 1 - B$; d 为差分次数; a 为白噪声序列。

将不同时刻的残差组成残差序列,若残差序列从某时间点开始远大于之前值(出现水平迁移),拟合干预模型;否则计算残差序列中最大值所对应时刻的检验统计量 $|\lambda|$, $\lambda = \max\{|T_{AO}|, |T_{IO}|\}$,其中 T_{IO} 为附加异常值, T_{AO} 为新息异常值,判断 $|\lambda|$ 是否大于常数 C ($3 < C < 4$)。当 $|\lambda| < C$ 时,输出键值对(k_1, v_1);当 $|\lambda| \geq C$ 时,若 $|T_{AO}| > C$,表明存在一个附加异常值,根据式(2)修正数据,若 $|T_{IO}| > C$,表明存在一个新息异常值,根据式(3)修正数据。

$$Z = X + \omega I = \frac{\theta(B)}{\varphi(B)\nabla^d} a + \omega I \quad (2)$$

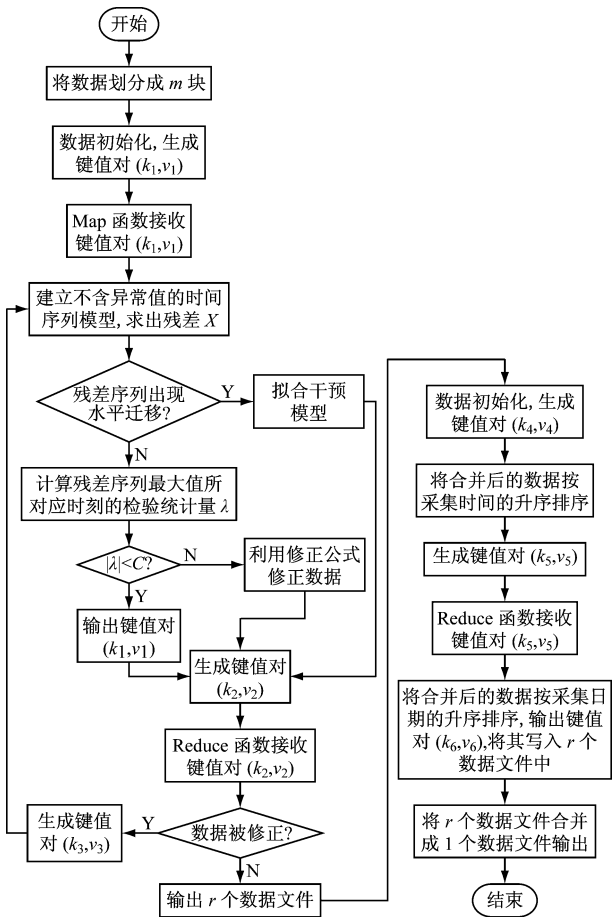


图 2 基于 MapReduce 的煤矿综采设备运行状态
大数据清洗建模流程

Fig. 2 Big data cleaning modeling flow of operation status of
coal mine fully-mechanized coal mining equipment
based on MapReduce

式中: Z 为修正量; ω 为异常值影响因子; I 为描述是否存在异常值的示性函数。

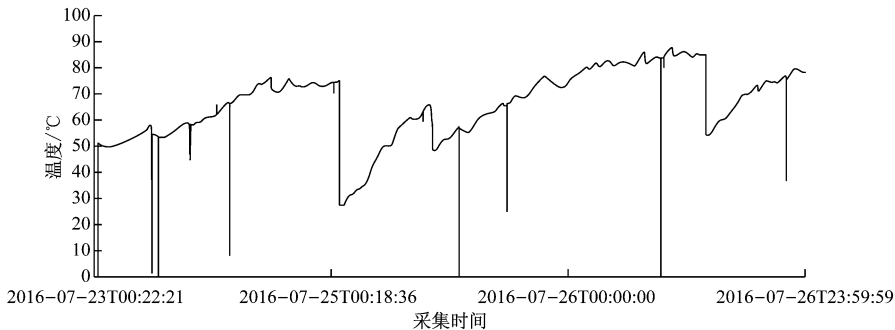


图 3 采煤机牵引变频器温度原始数据

Fig. 3 Raw temperature data of shearer traction inverter

利用基于 MapReduce 的煤矿综采设备运行状态大数据清洗模型对采煤机牵引变频器温度原始数据进行清洗, 结果如图 4 所示, 可看出原始数据中的噪声数据被剔除, 缺失值被补全。

$$Z = X + \frac{\theta(B)}{\varphi(B)\nabla^d} \omega I = \frac{\theta(B)}{\varphi(B)\nabla^d} (a + \omega I) \quad (3)$$

在键值对 (k_1, v_1) 的基础上叠加修正量, 得到修正后的键值对 (k_2, v_2) 。

(3) 第 1 个 MapReduce 中的 Reduce 函数接收键值对 (k_2, v_2) , 判断数据是否被修正。若数据被修正, 生成新的键值对 (k_3, v_3) , 返回步骤 (2), 将键值对 (k_3, v_3) 作为输入进行循环计算; 若数据未修正, 则说明数据清洗完毕, 由 r 个 Reduce 任务输出 r 个数据文件。

(4) 假设步骤 (3) 完成后, 数据中的异常值都被识别并修复。第 2 个 MapReduce 接收 r 个数据文件并初始化后, 将其转化成键值对 (k_4, v_4) , Map 函数负责合并具有相同键的数据, 并将合并后的数据按照采集时间的升序进行排列, 输出键值对 (k_5, v_5) 。

(5) 第 2 个 MapReduce 中的 Reduce 函数接收键值对 (k_5, v_5) , 按照数据采集日期的升序进行排序, 输出键值对 (k_6, v_6) , 并将其写入对应的 r 个数据文件中。

(6) 将 r 个数据文件合并为 1 个数据文件输出。

3 实验验证

选取某煤矿企业采煤机牵引变频器 20 万条历史温度监测数据 (监测时间为 2016-07-23T00:22:21—2016-07-26T23:59:59), 采煤机中途长时间停机 3 次, 2016-07-24 采煤机未运行, 数据中含有噪声点和缺失值。采煤机牵引变频器温度原始数据如图 3 所示。

4 结语

煤矿综采设备运行状态大数据清洗模型采用双 MapReduce 协同工作: 通过第 1 个 MapReduce 对

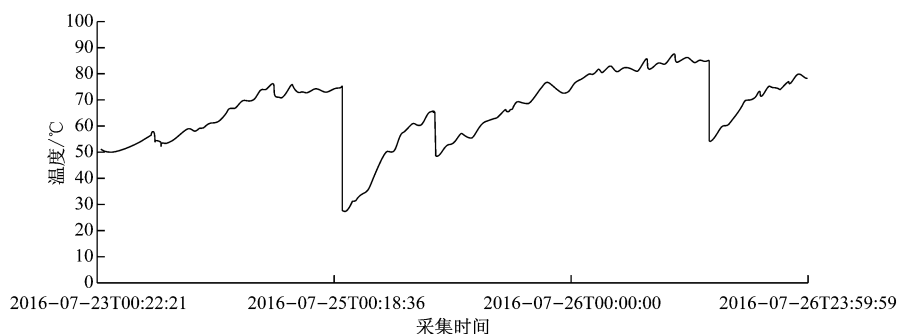


图4 采煤机牵引变频器温度原始数据清洗结果

Fig. 4 Raw temperature data cleaning result of shearer traction inverter

数据中的噪声点和缺失值进行修正,输出多个清洗后的数据文件;通过第2个MapReduce对多个清洗后的数据文件按数据采集时间及日期进行排序,并合并成单个数据文件输出。实验结果表明,该模型对原始数据中存在的噪声点和缺失值具有较好的清洗效果。

参考文献(References):

- [1] XU S, LU B, BALDEA M, et al. Data cleaning in the process industries [J]. Reviews in Chemical Engineering, 2015, 31(5): 453-490.
- [2] 谭章禄, 马营营. 煤炭大数据研究及发展方向[J]. 工矿自动化, 2018, 44(3): 49-52.
TAN Zhanglu, MA Yingying. Research on coal big data and its developing direction [J]. Industry and Mine Automation, 2018, 44(3): 49-52.
- [3] 马小平, 代伟. 大数据技术在煤炭工业中的研究现状与应用展望[J]. 工矿自动化, 2018, 44(1): 50-54.
MA Xiaoping, DAI Wei. Research status and application prospect of big data technology in coal industry [J]. Industry and Mine Automation, 2018, 44(1): 50-54.
- [4] ZHANG Haikuo, LU Zhonghua, LIU Fang, et al. Design and implementation of mass alarm data oriented parallel processing system [J]. Computer Engineering and Design, 2018, 39(2): 407-413.
- [5] BATINI C, CAPPIELLO C, FRANICALANCI C, et al. Methodologies for data quality assessment and improvement [J]. ACM Computing Surveys, 2009, 41(3): 1-52.
- [6] LIU Y, LIU Y Z, ZHANG H, et al. An RFID data cleaning strategy based on maximum entropy feature selection [J]. Journal of Digital Information Management, 2015, 14(2): 86-91.
- [7] YANG Xue, TANG Luliang, ZHANG Xia, et al. A data cleaning method for big trace data using movement consistency [J]. Sensors, 2018, 18(3): 824.
- [8] 娄建楼, 胥佳, 陆恒, 等. 基于功率曲线的风电机组数据清洗算法[J]. 电力系统自动化, 2016, 40(10): 116-121.
LOU Jianlou, XU Jia, LU Heng, et al. Wind turbine

data-cleaning algorithm based on power curve [J]. Automation of Electric Power Systems, 2016, 40(10): 116-121.

- [9] 杨东华, 李宁宁, 王宏志, 等. 基于任务合并的并行大数据清洗过程优化 [J]. 计算机学报, 2016, 39(1): 97-108.
YANG Donghua, LI Ningning, WANG Hongzhi, et al. The optimization of the big data cleaning based on task merging [J]. Chinese Journal of Computers, 2016, 39(1): 97-108.
- [10] 陈少龙. 基于Hadoop的煤矿设备数据规范化和清洗的研究 [D]. 西安: 西安科技大学, 2016.
- [11] ZHOU W, ZHONG Y, WANG Y. A data balance algorithm based on histogram in MapReduce [J]. Journal of Northwestern Polytechnical University, 2018, 36(3): 480-486.
- [12] BENDRE M, MANTHALKAR R. Time series decomposition and predictive analytics using MapReduce framework [J]. Expert Systems with Applications, 2018, 116: 108-120.
- [13] 齐玉东, 何诚, 司维超. MapReduce框架下的Skyline云资源选择算法 [J]. 计算机科学, 2018, 45(增刊1): 411-414.
QI Yudong, HE Cheng, SI Weichao. Cloud resource selection algorithm by Skyline under MapReduce frame [J]. Computer Science, 2018, 45(S1): 411-414.
- [14] 沈小军, 付雪姣, 周冲成, 等. 风电机组风速-功率异常运行数据特征及清洗方法 [J]. 电工技术学报, 2018, 33(14): 3353-3361.
SHEN Xiaojun, FU Xuejiao, ZHOU Chongcheng, et al. Characteristics of outliers in wind speed-power operation data of wind turbines and its cleaning method [J]. Transactions of China Electrotechnical Society, 2018, 33(14): 3353-3361.
- [15] 聂巧平, 冯蕾. 考虑结构突变的单位根检验程序研究——基于“新息异常值模型”的Perron检验分析 [J]. 数量经济技术经济研究, 2008(9): 139-151.
NIE Qiaoping, FENG Lei. The research on the unit root test procedure with structural break [J]. The Journal of Quantitative & Technical Economics, 2008(9): 139-151.