文章编号：1674—8247(2020)06—0045—04
DOI:10.12098/j.issn.1674－8247.2020.06.009

基于分布式计算模型的施工工序关联分析

杨 科 唐朝国 袁 焦 伏 坤 邹文露

（中铁二院工程集团有限责任公司， 成都 610031）

摘 要：在利用计算机处理铁路工程施工日志庞大数据集合的过程中，计算时间长、内存消耗大、最终导致整个计算系统崩溃的现象时有发生，因此，亟需寻求一种可靠、高效的数据处理方式。本文利用分布式架构，采用成熟的数据分析 R 语言，通过对铁路工程现场施工数据进行抽取、存储和分析，以分布式计算方法实现了 Apriori 关联分析算法，完成了对海量现场数据的高效分析，提升了分析速度，降低了计算资源的占用。经实际工程应用验证，分析结果真实清晰地反映了现场实际，能够为施工活动提供有力的数据支撑，有效规避和降低过程风险，对铁路工程建设活动的管理具有重要参考意义。

关键词：分布式；施工工序；关联分析；R 语言；Apriori；EBS
中图分类号：TP311.1 **文献标志码**：A

Association Analysis of Construction Process Based on Distributed Computing Model

YANG Ke TANG Chaoguo YUAN Jiao FU Kun ZOU Wenlu

(China Railway Eryuan Engineering Group Co., Ltd., Chengdu 610031, China)

Abstract: In the process of using computer to process the huge data set of railway engineering construction log, phenomena such as long calculation time, large memory consumption and collapse of the whole computing system finally caused often occur. Therefore, it is urgent to find a reliable and efficient data processing method. In this paper, the Apriori association analysis algorithm is realized with distributed computing method by using distributed architecture and mature data analysis R language through extraction, storage and analysis of railway engineering site construction data, which has been used to complete the analysis of large amounts of site data efficiently, improve the analysis speed and reduce the occupation of computing resources. It has been verified by practical engineering application that the analysis results truly and clearly reflect the actual site condition, which can provide powerful data support for construction activities, effectively avoid and reduce process risks, and have important reference significance for the management of railway engineering construction activities.

Key words: distributed; construction process; association analysis; R language; Apriori; EBS

在铁路工程建设期间，施工质量问题时有发生，分析这类问题发现，其中相当一部分问题是由施工工序安排不当造成的。其原因在于施工安排在很大程度上依赖于现场管理人员的知识和经验，人为因素影响较

收稿日期：2020-09-11
作者简介：杨科(1982-)，男，工程师。
引文格式：杨科，唐朝国，袁焦，等. 基于分布式计算模型的施工工序关联分析[J]. 高速铁路技术，2020，11(6)：45－48.
YANG Ke, TANG Chaoguo, YUAN Jiao, et al. Association Analysis of Construction Process Based on Distributed Computing Model [J]. High Speed Railway Technology, 2020, 11(6): 45－48.

大,在进度压力下,往往易忽视施工工序的规律性。

施工工序为铁路工程风险管理提供了基础支撑,是定位风险事件的最小单位,寻找工序之间的逻辑关系,再根据现场实际选择合理的工序安排,能帮助现场施工人员缓解和预防风险的发生[1]。

施工工序的安排具备客观性,若能通过分析现场施工数据,找到工序之间的规律,势必能够改善风险控制、成本监控、进度管理等过程。随着铁路工程电子施工日志在全国铁路建设项目中的普及,施工现场数据逐渐被统一、规范,为上述问题的解决提供了数据基础。本文将利用分布式计算框架,借助铁路工程实体分解结构(EBS)[2],从海量施工日志数据中寻找铁路工程施工工序之间的规律。

1　现状分析

1.1　电子施工日志和 EBS

电子施工日志系统利用信息化的方式将以往纸质的施工日志统一规划,将施工现场的技术情况、安全检查情况、质量检查情况等管理起来,以桌面应用、移动应用的方式呈现,方便现场用户进行填报。

EBS 是铁路工程实体分解结构的缩写,由中国铁路 BIM 联盟于 2014 年发布。不同于基于工作分解结构(WBS)[3],EBS 结构按照专业将工程系统分解为树形结构,更满足工程系统的特点,更利于铁路工程管理。电子施工日志提供的数据采集功能就是以 EBS 为纽带串联起来。桥涵专业的部分 EBS 项如表 1 所示。

表 1　桥涵专业部分 EBS 项列表

EBS 名称	EBS 编号
桥涵专业	030101010212
特大桥	03010101021201
复杂特大桥	0301010102120105
建筑工程费	030101010201
上部	03010101020102
支座	030101010212
(1)金属支座	03010101021201
⑤球型钢支座	0301010102120105
预应力混凝土简支箱梁	030101010201
架设	03010101020102

1.2　分布式系统

作为一款采集现场数据的信息系统,施工日志具备面向事务系统(OLTP)的所有特征,其数据通过关系型数据库进行存储,根据业务的不同,将数据分散到不同的数据表中。采集端在确认用户信息后,在本地将用户每日填报的现场环境、施工进度、材料使用情况、人员投入情况等打包上传,通过互联网集中传输到应用服务器,并最终进入到数据库中。这样的方式可使业务系统快速响应各个增删查改的需求,但面对耗时、海量查询的统计分析需求时,就显得力不从心,强行在 OLTP 系统上执行统计分析,反而会使业务的处理率下降,甚至造成数据丢失,严重影响系统的推广使用。

目前,已有很多研究对铁路工程的“大数据”统计分析进行了实施和应用[4]。随着分布式系统逐步成为大数据的核心技术,以 Hadoop 生态圈为代表的平台将为工程的建设业务提供解决方案,为项目的成功落地提供技术保障[5]。

1.3　关联分析和 R 语言

关联分析是数据挖掘的一种主要方法,用于查找隐藏在数据集合中的频繁模式,即集合中各项之间存在的关联规则。利用支持度和置信度两个指标来保证找到有意义的规则,避免某些偶尔出现规则对于整个模型的影响。Apriori 算法是关联分析的常用算法,解决了很多潜在频繁模式的挖掘问题(如经典的“啤酒尿布”案例等)。

R 语言是一门用于统计分析的编程语言和开发环境,包括了丰富的统计算法和制图函数,更加贴近统计学家的使用习惯,在大数据前沿科学的研究中使用更加广泛。如 RHadoop[6]利用 MapRecude[7] API 集成了常用的 R 函数,使数据分析人员能够利用 R 语言进行 HDFS、HBase 的连接和访问,再配合 R 语言强大的数据分析能力,快速实现分析目标。但这类软件也存在缺点,如 RHadoop 对于数据库的支持就较弱,没有 API 直接访问 Hive,通过 RJDBC 访问需对数据进行转换,比较耗能。

1.4　施工工序的关联分析

施工工序是项目管理的基础,要实施施工组织管理信息系统,就需建立可靠的施工工序活动。传统常利用需求工程方法,从现场环境、人员、设备、材料到场情况等维度,向项目管理人员、施工技术人员收集原始信息,依靠他们的丰富经验和知识来保证系统的可靠性。但从信息系统的角度来看,领域知识的获取最为困难,易导致后期模型的不确定性。另一个方法是依赖现场数据,从实际工作中寻找客观规律。施工日志为我们提供了数据来源,但面对海量的数据,传统的分析方法在时间耗费、资源占用上都已满足不了实际需要。面对用户需求和分析方法的矛盾,采用分布式架构,解决数据的大容量可靠存储以及分析的并行计算,势必会成为施工日志数据向应用转换的唯一途径。

2　关联模式分析

2.1　数据及数据模式

要构建海量数据基础上的关联分析,必须首先分析现有数据的内在关系,构建稳定、成熟的数据模式。在施工日志中,工序相关的数据分散在多张数据表中,具备典型的多维数据特点。为便于分析,且保证分析过程不影响业务系统的正常运行,需对多维数据进行抽取、转换和加载(ETL 过程),最终进入分布式存储系统之中。

具体而言就是将多源数据平面化,这些平面化数据称为“现场数据”(如表 2 所示),数量已达到亿级,需要借助 Sqoop 工具,用命令行的方式执行相关 SQL 并导入 HDFS。根据不同的分析维度(如每天、每周、每月等),灵活地执行不同的 SQL 语句,抽取相应的数据结果。

表 2　数据分析原始数据表

EBS 编号	日期	工点编号	天气
0301020102130102	20170818	100404	多云
0301020102130102	20170820	100404	晴
030102010211	20170818	22060	阵雨

2.2　关联分析的分布式设计

R 语言在 Hadoop 上已有成熟的应用,统计分析人员也更习惯于采用 RHadoop 进行数据分析,虽然已有其他编程语言(平台)实现甚至改进了分布式系统上的关联分析算法[8],但可惜的是 R 语言提供的 arule 包的对应算法并不支持分布式计算,在面对海量数据时,无法保证处理效率。因此需寻求一种基于 MapReduce模型的算法来支持大数据的分析计算。

2.2.1　购物篮化过程

Apriori 所需的事务数据集以某一属性为维度,聚合该维度下的数据。现场数据是每日施工情况的流水账,应当以“工点”+“日期”为维度,聚合多道工序,该过程可称为“购物篮化”。其数据结构如表 3 所示,分布式流程如图 1 所示。

表 3　事务数据集(购物篮化)表

序号	施工工序
100404_20170818	030101010214,0301010103
100404_20170820	0301010101010202,03010101020101,030102010211
22060_20170818	03010101020101,03010101020101,0301020102040405,030105,030102010211,030207010105

输入数据是平面化的数据,每一行包含一组施工信息,处理逻辑可分为两个子流程。首先是将同一天、同一工点的数据找出来的分组逻辑,然后是将同一分组内的工序(EBS 编码)拼接并以逗号分隔的合并逻辑,最终输出到 HDFS 之中。对应到 RHadoop 的相关开发包,需在 Map 任务实现第一个子流程,在 Reduce 任务实现第二个子流程。鉴于第一个子流程产生的中间结果比较庞大,利用 RHadoop 提供的Combine任务,可对结果进行合并,只需实现和 Reduce 的相同逻辑,即可减轻网络负载,减小 Reduce 任务执行期压力。Reduce 任务还可过滤聚合后只有一道工序的记录,降低下一过程的计算量。

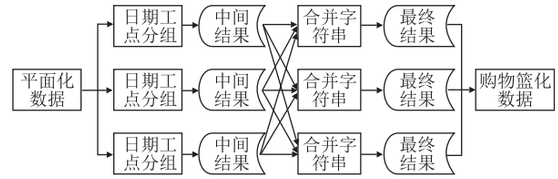

图 1　购物篮化现场数据 MapReduce 流程图

2.2.2　初始化频繁 1 – 项集过程

频繁 1 – 项集是从购物篮化的数据中,抽取只包含 1 条 EBS 编码的作为项。该过程需从购物篮化后的数据中生成,由于数据量预期都会很大,因此还需借助 MapReduce 来实现。其具体流程如图 2 所示。

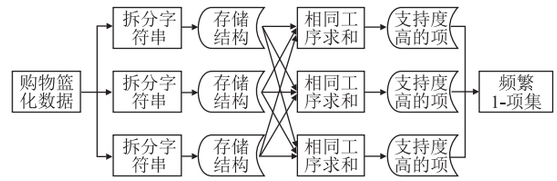

图 2　生成频繁 1 – 项集 MapReduce 流程图

2.2.3　生成候选 k – 项集过程

该过程在一个循环体中。候选项集的生成有很多种算法,主流的算法是对频繁($k-1$) – 项集自身做组合,生成候选 k – 项集。一个专业的 EBS 编码大概在 1 000 ~ 20 000区间内,组合以后满足支持度阈值的更少,因此该过程可在内存中进行计算,以提高整体分析的速度。候选集的算法过程如图 3 所示。

图 3 是一个频繁 3 – 项集生成候选 4 – 项集的过程,A – E 代表了 1 条 EBS 编码,由于 BC、CE 开头的只有一项,根据 Apriori 定理,不会产生频繁项,因此直接淘汰,从 ABC、ABD、ABE 中生成。

2.2.4　生成频繁 k – 项集过程

该过程和上一个过程在同一个循环体内,利用上一过程产生的候选 k – 项集,在购物篮中遍历查找是

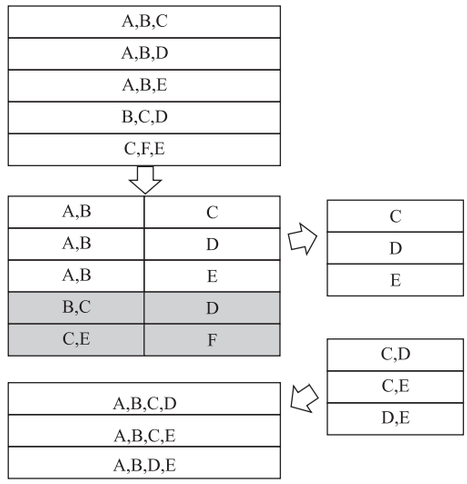

图 3　生成候选 k – 项集流程图

否存在这样的项,找到 1 次计数加 1。当数据量较大时,在购物篮中遍历查找也是一个非常消耗性能的操作,必须把该过程放到 Map 中,实现数据的分而治之,随着 Map 计算节点的增加,遍历的时间会得到有效的控制。其具体流程如图 4 所示。

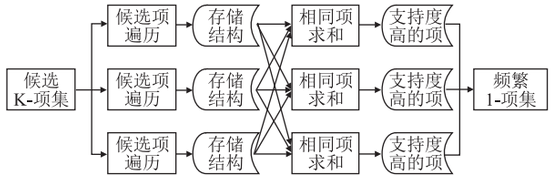

图 4　生成频繁 k – 项集 MapReduce 流程图

每次循环产生的频繁 k – 项集加入到频繁项集集合 F 中,若没有产生频繁 k – 项集,则终止循环体。

2.2.5　从频繁项集生成关联规则

从寻找工序间的关联关系而言,频繁项集 F 已经够用,但从寻找工序的规则而言,还需对频繁项集 F 进一步加工,该过程在内存中进行计算。

规则的生成依赖于置信度,以频繁项集 t:{0301010101010108,03030101010102,030301010211}为例,其有 6 个规则,如{0301010101010108}→{03030101010102,030301010211}、{03030101010102}→{0301010101010108,030301010211}、{03030101010102,030301010211}→{0301010101010108}等作为候选,表示为{Left}→{Right},通过在频繁项集集合 F 中计算该频繁项集支持度 $\delta(t)$ 与箭头左边的项集支持度 δ(Left)的商,即得到 confidence(Left→Right)。无论{Left}还是 t 肯定存在于 F 中,支持度已知,整个计算过程并不会消耗太多资源。

3　分析结果

3.1　模型好坏的评价指标的制定

从上述的设计、实现过程可知,施工工序的关联分析依赖于“支持度阈值”的制定。“支持度阈值”是整个分析的关键指标,该指标决定了模型的好坏以及计算的时间、空间复杂度。因此,采用均衡计算量 + 用户评价的方式,先实验一个较低的支持度阈值(如 0.55),将获取到的规则交给现场施工人员、技术专家评价,再增大或降低支持度阈值[9],力求达到计算量与可信模型之间的平衡。

3.2　实验结果

本文通过分析百万级桥涵专业施工日志的数据,得到实验结果如表 4 所示。

表 4　桥涵专业工序频繁集表(百万级)

规则	支持度
0301020101010201,0301020101010202	0.026
030102010101010401,030102010101010402	0.016

根据 EBS 代码反查工序名称,发现桥墩地基承台的“混凝土”和“钢筋”频繁出现在施工工序安排中,这与施工现场的实际情况相吻合。利用同样的算法,再选取施工日志填报优秀的桥涵工点进行分析,获取更加精准的实验结果,如表 5 所示。

表 5　桥涵专业工序频繁集表(十万级)

规则	支持度
03010101020102,03010101102120105	0.157
03010201020102,0301020102120105	0.290

根据 EBS 代码反查可知,“预应力混凝土简支箱梁架设”与“球型钢支座”的关联程度较高,在日常工作安排中,可考虑两者先后施工。

4　结束语

在调用传统的 R 函数进行关联分析时,内存占用会明显增长,易导致整个分析系统崩溃。采用分布式的 Apriori 算法,在处理施工日志的海量数据时,内存消耗低,且并行的处理方式也降低了分析时间。在此基础上,本文分析了铁路工程施工工序,获得了现场工序安排的规律,为管理人员把握施工进度、合理安排现场工作提供了一种智能化的解决方式。可以预见,随着电子施工日志的逐步普及,越来越多的项目会采用电子日志进行填报,未来的数据质量、数量会进一步增加,最终提高分析结果的精度和覆盖范围。

(下转第 57 页)

Amplification Combined with QPCR (PAA-qPCR) in Spiked Chicken Meat Samples[J]. Food Control, 2019, 99: 79 - 83.

[24] 李朗. 川藏线货物列车牵引重量主要影响因素分析[J]. 铁道经济研究,2019(2):44 - 47.
LI Lang. Analysis on Main Influencing Factors of Traction Weight of Freight Trains on Sichuan-Tibet Railway [J]. Railway Economics Research,2019(2):44 - 47.

[25] 段艳军, 张颖. 复杂艰险山区高速铁路限制坡度的选择[J]. 山西建筑, 2019, 45(1): 126 - 129.
DUAN Yanjun, ZHANG Ying. Selection of Restricted Slopes for High Speed Railways in Complex and Dangerous Mountainous Areas[J]. Shanxi Architecture, 2019, 45(1): 126 - 129.

[26] 矫岩峻, 文艳晖, 刘少克. 中低速磁浮列车上坡牵引策略优化[J]. 机车电传动, 2016(2): 37 - 39.
JIAO Yanjun, WEN Yanhui, LIU Shaoke. Uphill Traction Strategy Optimization of Middle/Low-speed Maglev Train[J]. Electric Drive for Locomotives, 2016(2): 37 - 39.

[27] 宋锴,牛会想. 回转质量系数对高速列车牵引电算的影响[J]. 铁道机车车辆,2010,30(3):56 - 59.
SONG Kai, NIU Huixiang. Influence of Rotary Mass Coefficient on Traction Computer Calculation of High-speed Train [J]. Railway Locomotive & Car,2010,30(3):56 - 59.

[28] 高邓波. 列车制动距离计算的程序设计[J]. 中国科技信息, 2009(3): 30.
GAO Dengbo. Program Design of Train Braking Distance Calculation [J]. China Science and Technology Information, 2009(3): 30.

[29] 曾剑群. 动车组牵引计算仿真系统的研究[D]. 北京:北京交通大学,2009.
ZENG Jianqun. The Research on Simulation System of EMU Traction [D]. Beijing: Beijing Jiaotong University,2019.

[30] 廖勇. 列车运动方程近似积分研究[J]. 电力机车与城轨车辆, 2007, 30(5): 30 - 32.
LIAO Yong. Research about Approximate Integral of Train Kinematics Equation[J]. Electric Locomotives & Mass Transit Vehicles, 2007, 30(5): 30 - 32.

[31] 曾宇清, 于卫东, 扈海军, 等. 高速铁路牵引计算层次约束方法[J]. 中国铁道科学, 2009, 30(6): 97 - 103.
ZENG Yuqing, YU Weidong, HU Haijun, et al. Hierarchical Restriction Method for the Traction Calculation of High-speed Railway [J]. China Railway Science, 2009, 30(6): 97 - 103.

(上接第 48 页)

参考文献:

[1] 魏永幸. 路基工程风险识别与管理研究[J]. 铁道工程学报, 2013, 30(3): 91 - 96.
WEI Yongxing. Research on the Risk Identification and Control of Subgrade Engineeriong[J]. Journal of Railway Engineering Society, 2013, 30(3): 91 - 96.

[2] 佚名. 铁路工程实体结构分解指南(1.0 版)[J]. 铁路技术创新, 2014(6): 5 - 334.
YI Ming. Railway Engineering Entity Structural Decomposition Guide (Version 1.0)[J]. Railway Technical Innovation, 2014(6): 5 - 334.

[3] 刘竹君, 陈伟志, 胡超, 等. 高速铁路工程 WBS 的实践探讨[J]. 高速铁路技术, 2018, 9(4): 20 - 24.
LIU Zhujun, CHEN Weizhi, HU Chao, et al. Discussion and Practice of WBS on High-speed Railway[J]. High Speed Railway Technology, 2018, 9(4): 20 - 24.

[4] 邱永平, 张东卿, 刘菀茹. “大数据”在铁路路基工程设计中的应用探讨[J]. 高速铁路技术, 2017, 8(3): 16 - 19.
QIU Yongping, ZHANG Dongqing, LIU Wanru. Discussion on Application of Big Data in Railway Subgrade Engineering Design[J]. High Speed Railway Technology, 2017, 8(3): 16 - 19.

[5] 杨科. 大数据在铁路工程施工日志中的应用研究[D]. 电子科技大学, 2018.
Yang Ke. Applied Big Data Research On Electronic Construction Diary of Railway Engineering [D]. University of Electronic Science and Technology of China, 2018.

[6] CAI Lijun, GUAN Xiangqing, CHI Peng, et al. Big Data Visualization Collaborative Filtering Algorithm Based on RHadoop [J]. International Journal of Distributed Sensor Networks,2015:1 - 9.

[7] Telmo Silva Morais. Survey on Frameworks for Distributed Computing: Hadoop, Spark and Storm [C]//Dsie15 Doctoral Symposium in Informatics Engineering, 2015.

[8] 魏玲, 魏永江, 高长元. 基于 Bigtable 与 MapReduce 的 Apriori 算法改进[J]. 计算机科学, 2015, 42(10): 208 - 210.
WEI Ling, WEI Yongjiang, GAO Changyuan. Improved Apriori Algorithm Based on Bigtable and MapReduce[J]. Computer Science, 2015, 42(10): 208 - 210.

[9] 余绍黔. Apriori 算法改进及在超市数据挖掘中应用[J]. 微计算机信息, 2011, 27(11): 165 - 167.
YU Shaoqian. Improved of Apriori Algorithm and Appl Ication of Data Mining in Supermarket[J]. Control & Automation, 2011, 27(11): 165 - 167.