计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(11):132−137 [doi: 10.15888/j.cnki.csa.007130]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

Web 架构驱动的石油工业知识挖掘系统①

耿祖琨[1], 张卫山[1], 王志超[2], 李　博[3]

[1](中国石油大学 (华东) 计算机与通信工程学院, 青岛 266580)
[2](东营市人力资源和社会保障局, 东营 257091)
[3](东营市勘察测绘院, 东营 257000)
通讯作者: 耿祖琨, E-mail: 506149151@qq.com

摘　要: 石油工业大数据具有无限潜力与价值, 将大数据与数据挖掘技术应用其中, 不仅可以提升石油行业工业化水平, 而且对石油行业智慧化发展起到强有力地推动作用. 由此提出了一个 Web 架构驱动的、集成了数据挖掘五大模块的新型工业知识挖掘系统-即石油工业数据挖掘系统, 包含: 数据集管理、预处理算法管理、数据挖掘算法管理以及数据挖掘流程管理和数据结果可视化五大模块. 本系统实现了完全自助式的数据提取、数据预处理、数据分析与知识挖掘和结果可视化展示的完整知识挖掘流程. 通过以 Web 的形式满足油田不同层级的用户在不同场景下的即时使用需求, 极大提高了系统的灵活性. 通过本系统, 油田的技术开发人员可忽略大数据的搭建以及其他复杂构建过程, 更好的服务于油田数据建模和分析.

关键词: 大数据分析; 数据挖掘; 石油工业; Web 架构; 自助式服务

引用格式: 耿祖琨,张卫山,王志超,李博.Web 架构驱动的石油工业知识挖掘系统.计算机系统应用,2019,28(11):132–137. http://www.c-s-a.org.cn/1003-3254/7130.html

Knowledge Mining System for Petroleum Industry Driven by Web Architecture

GENG Zu-Kun[1], ZHANG Wei-Shan[1], WANG Zhi-Chao[2], LI Bo[3]

[1](College of Computer and Communication Engineering, China University of Pertroleum, Qingdao 266580, China)
[2](Dongying City Human Resources and Social Security Bureau, Dongying 257091, China)
[3](Dongying City Survey and Mapping Institute, Dongying 257000, China)

Abstract: Big data of petroleum industry has infinite potential and value. The application of big data and data mining technology can not only improve the industrialization level of petroleum industry, but also play a strong role in promoting the intelligent development of petroleum industry. This paper presents a new industrial knowledge mining system—Petroleum Industry Data Mining System, which is driven by Web architecture and integrates five modules of data mining. The five modules inclide data set management, pre-processing algorithm management, data mining algorithm management, data mining process management, and data result visualization. The system realizes completely self-service data extraction and data pre-processing, and completes knowledge mining process of management, data analysis, knowledge mining, and visualization of results. The flexibility of the system is greatly improved by satisfying the real-time requirements of users at different levels in different scenarios in the form of Web. Through this system, oilfield technicians can neglect the construction of large data and other complex construction processes, and better serve oilfield data modeling and analysis.

Key words: big data analysis; data mining; petroleum industry; Web architecture; self-service

① 基金项目: 国家自然科学基金 (61309024); 山东省自然科学基金 (F020509, F060604)
Foundation item: National Natural Science Foundation of China (61309024); Natural Science Foundation of Shandong Province (F020509, F060604)
收稿时间: 2019-03-29; 修改时间: 2019-04-26; 采用时间: 2019-04-29; csa 在线出版时间: 2019-11-06

随着国家智能制造的大规模发展, 石油行业不断创新发展, 其工业设备也越来越复杂, 传感器、摄像头等的广泛部署使得石油工业设备的运行状态得到有效监控, 由此也产生了大规模的工业数据. 石油工业大数据[1,2]的采集、处理、存储、分析和利用的价值不断提升, 为指导石油探测、开采和企业改革发展的推进提供了重要依据. 目前已经包含抽油机井基本状态信息、地质数据、勘探数据以及生产数据等等, 现有的采油相关数据类别已达 600 多种, 而且相关数据每时每刻不在产生、交互、传回, 石油大数据呈现爆发增长、海量集聚的特点.

随着数字油田以及智慧油田建设的不断深入, 不同应用系统的数据结构呈现多样化发展, 从原先单一数据结构转变为多维化、多元化结构, 数据之间的显性与潜在的分布关系也越来越模糊. 如何将海量的抽油机井状态数据、地质数据以及工作生产数据进行数据预处理、特征选择并进行关联性分析, 以找寻对油气开采有利的生成信息, 是指导项目实际生产、提升油气产量、降低产量递减速率、提升剩余油开采几率的重要数据依据.

与此同时, 如何对石油大数据进行快速、及时的数据挖掘与知识发现, 传统的单机服务器, 需要利用现有的分布式集群以及快速通用的计算引擎, 同时, 需要现有石油工程以及采油工程等相关学科专业知识与通用算法, 而且需要建立石油工业大数据仓库进行辅助存储, 提高平台读写速度以及提升平台执行计算能力.

由此, 石油大数据分析与数据挖掘技术二者结合成为趋势[3], 通过石油工业大数据分析得到的结果可以辅助企业制定出符合工业发展的策略, 并能依据石油工业大数据进行生产状况的及时调整, 以促进国内整体石油工业水平的提升.

1 系统概述

目前大数据挖掘算法已经被应用到油气开采相关领域, 但是相对油田行业众多技术人员而言, 不仅对数据挖掘算法难以掌握, 而且如何编码实现数据采集、存储、调用以及执行和可视化, 和搭建大数据集群也是其中的难点. 与此同时, 各类大数据平台层出不穷, 基于 Python 的 Orange 有较好的可视化编程工具和强大的 Python 脚本, 基于 Java 的 KNIME 集成了基础机器学习组件与数据挖掘算法等等.

如何将石油工业大数据与数据挖掘技术相结合[4–7], 并且与具备可控算法流程的大数据分析平台相融合[8–11]是当前石油工业数据分析领域需要探索的问题.

尽管各种工具都有其优势, 但是针对石油领域的知识挖掘系统而言, 如下主要问题需要解决:

(1) 针对数据采集过程中如何支持多种数据结构的并支持一键选择本地数据源导入到大数据仓库的数据采集操作模块问题;

(2) 针对大数据处理过程中, 如何选择大数据仓库中的数据源构建不同的数据集问题;

(3) 在大数据分析工作流程的创建过程中, 选择单数据集条件下的, 如何通过简单的拖拉拽等操作创建单一算法模型或多个算法模型的数据分析处理流程问题;

(4) 无法通过系统将数据源、数据集或者大数据分析结果进行二维图形或者三维图形的可视化展示.

针对当前石油领域的知识挖掘系统存在未能实现具有可控大数据完整分析工作流程界面的以及大数据信息可视化等问题, 在本文中, 提出了一个 Web 架构驱动的石油工业知识挖掘系统, 来解决此类问题, 包括如下两个部分:

(1) 可控工作流程的知识挖掘系统: 该系统在选择需要进行分析的数据集后, 支持用户采用拖拽操作快速完成数据建模, 支持单数据源单模型算法构建、支持单数据源多模型算法构建, 用户提交数据分析流程后, 系统在大数据分析后台执行模型组建、数据处理、数据分析以及分析结果存储.

(2) 自助式数据挖掘: 该系统提供可视化操作的流程创建和丰富的图表展示分析结果, 比如: 表格、柱状图、雷达图、折线图、散点图等等, 实现灵活、多样的数据分析, 从而可快速发现数据中的规律.

在本节中, 将介绍石油工业知识挖掘系统架构, 主要包含以下 4 大部分: 数据采集层、数据处理层、数据服务层、自助式可视化层. 以图 1 所示将分别介绍各个模块.

(1) 数据采集层

石油大数据采集层包含 3 部分: 数据采集服务器、数据存储服务器和 FTP 服务器集群.

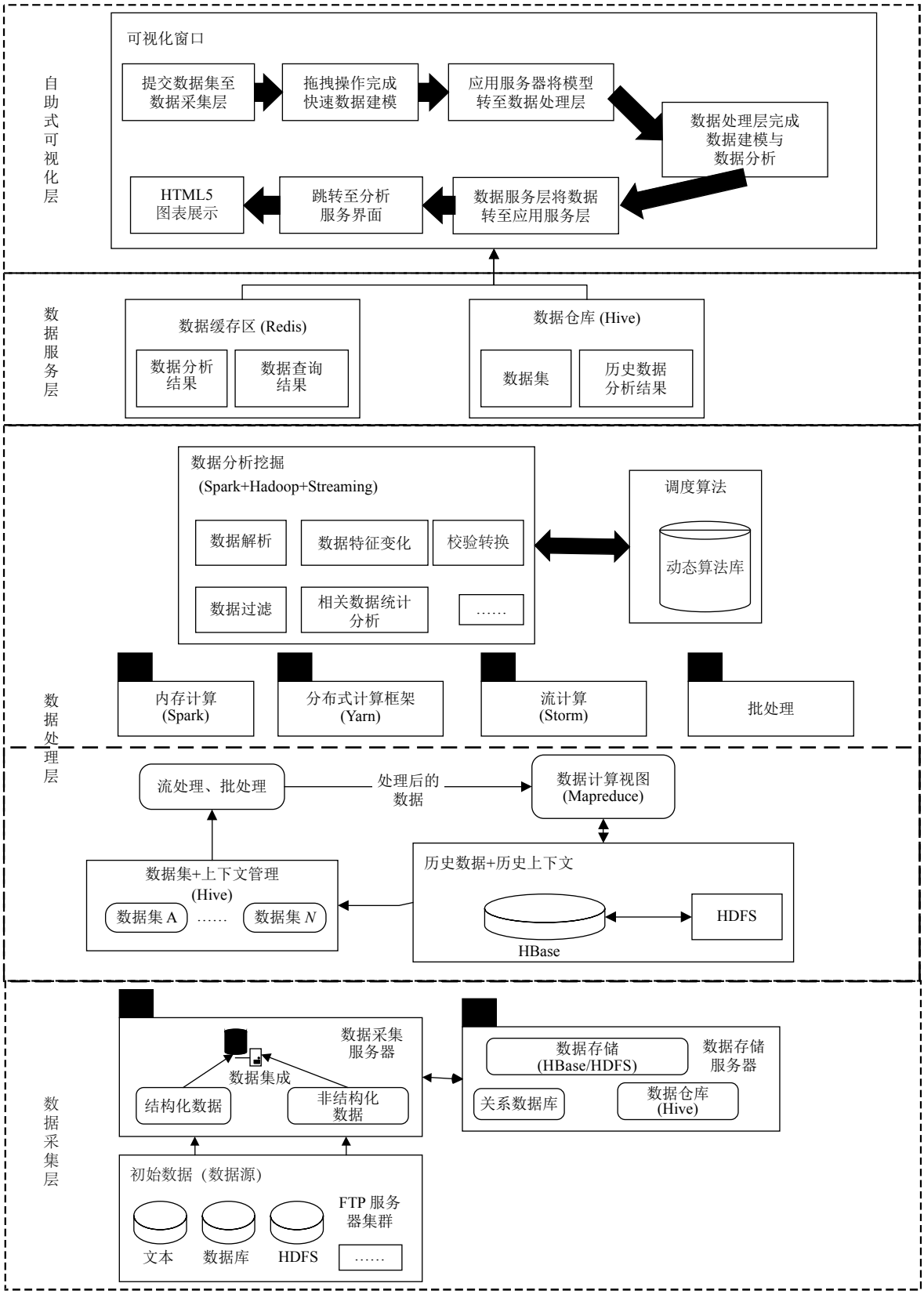

图 1　石油工业知识挖掘系统架构

原始数据 (如采油领域相关文本数据、A1/A2 等数据库中存储的数据以及 HDFS 存储的数据) 视为存储在各个 FTP 服务器中, 通过数据采集服务器, 将来自不同数据源的数据进行结构化和非结构化构建[12], 数据集成后统一存储到数据存储服务器中. 采用 Hive 搭建数据仓库. 同时, HBase[13]作为面向列存储的数据库, 不仅可以存储结构化数据, 而且弥补了 Hive[14]在分析查询和实时查询的不足. 并且, 将非结构化数据存储在 HDFS 中[15], 由此, 数据采集层完成了数据分析的重要底层部分-数据源和数据集存储.

系统的用户数据信息以及提供模型搭建流程等系统信息存储在结构数据库 MariaDB 数据库中, 它有着更好的子查询优化与线程池等优势.

(2) 数据处理层

石油工业知识挖掘系统以 Hadoop 生态系统[16]作为底层基础, 系统将数据采集层中的历史数据存储于 HBase 与 HDFS 中, 数据集数据存储于 Hive 数据仓库中, 通过数据流处理与批处理提供更快的速度给 MapReduce, 进而快速得到数据计算视图. 在此, Spark[17]平台提供内存计算服务, Yarn 提供分布式计算框架, Storm 提供流计算服务与批处理服务. 数据挖掘模块由 Spark 平台和 Hadoop 平台搭建, 通过连接应用服务器的建模方案和动态算法库的算法信息, 进行数据集引用、模型搭建和数据分析, 可以提供数据解析、数据过滤、数据特征变化、数据统计分析以及校验转换等等数据预处理操作.

(3) 数据服务层

由于数据处理层的处理结果需提供给应用服务器供用户查询, 系统提供基于内存计算的 Redis 数据库作为数据缓存区, 提供查询数据分析结果与数据执行结果. 它基于内存执行缓存存储, 不仅可以提升数据查询效率, 而且, 支持数据持久化操作, 支持异步操作将内存中的数据写到硬盘中, 且不中断服务. 所以, Redis[18]数据库提升了系统公共缓存能力, 降低了系统存储数据库的负载.

数据仓库提供给用户允许有较低延时查询数据的服务, 包含大数据量的数据集查询与历史数据分析结果查询等等.

以上保证系统不仅可提供实时查询当前任务处理结果, 而且可提供有延迟的历史任务处理过程与结果.

(4) 自助式可视化服务层

石油工业知识挖掘系统提供了自助式可视化层作为用户访问的窗口, 有以下几个功能:

1) 提交数据集至数据采集层

该数据集管理模块为用户提供多种数据源提交模式, 用户可根据数据源格式选择提交模式, 系统将数据源导入到数据采集服务器中, 进行结构化数据与非结构化数据转换并进行数据集成, 将分别存储到 Hive、HBase 与 HDFS 中.

2) 拖拽操作完成快速数据建模

数据建模模块在用户选择数据集后通过应用服务器向动态算法库模块发送请求, 服务器提供给用户数据建模模块, 展示数据预处理、数据集成、数据挖掘算法等等各种算法, 用户采用拖拽方式将算法拖到编辑区, 用户按照要求输入算法不定项的参数, 并选择连接新的算法, 以此循环至模型搭建完成.

3) 应用服务器将模型转至数据处理层

应用服务器将数据集 ID 以及模型信息转至数据处理层, 数据处理层在各个组成部分配合下, 根据数据集 ID 导入数据源并执行数据模型流程.

4) 数据处理层完成数据建模与数据分析

数据处理层根据数据集与模型信息, 调用动态算法库中算法 jar 文件, 并进行基于 Spark 平台的分布式数据分析.

5) 数据服务层将数据转至应用服务层

数据服务层将处理结果与之前数据集信息和模型信息回执到数据服务层, 数据服务层进行快速缓存存储[18], 准备提供结果给应用服务器.

6) 跳转至分析服务界面

应用服务器从数据服务层获取实时分析结果与延时数据信息, 通过可视化展示分析数据结果, 并依据结果进行知识发现.

7) HTML5 图标展示

系统提供了多种图形化技术, 帮助来理解数据间的关键性联系, 指导以最便捷有效的途径找到问题的最可能的解决办法. 它融合了图形、表格等多种可视化技术来处理多维数据, 使得数据所表现出的特性、类别、模式和关联性等信息一目了然, 在结果输出时可方便快捷的进行多种统计结果演示, 支持散点图、分布图、折线图、饼图等.

2 实验分析

为了验证系统的有效性, 通过研究抽油机井采油

系统效率影响因素[19–21]的关联性的实验进行分析[22]. 包含数据集选定、模型构建、模型执行以及结果可视化展示四个过程.

实验采用 FPGrowth 算法[23]进行影响抽油机井采油系统效率影响因素的关联性分析. 抽油机井系统效率不仅反映当前抽油机的采油质量与效益, 而且综合反映了油田的技术水平和装备水平, 因此研究抽油机井系统效率提升是提高油田工作质量的重要方向.

操作步骤如下: 通过选择华北油田 2016–2017 年抽油机井某区块某单口采油井生产数据的本地数据源进行数据源导入, 实现将本地数据源转入到 HDFS 和 HBase 的大数据仓库中. 选定该生产数据集后, 针对该数据集进行数据预处理, 首先进行筛选 5 个有用列, 包含日产液 (t)、泵深 (m)、动液面 (m)、冲程 (m)、冲次 (n/min); 其次针对数据集中的缺失值通过取该条数据集前 5 个和后 5 个数据的平均值进行数据填充; 针对严重离群的数据进行该列均值填充法进行修正; 最后, 为了消除各特征的量纲影响, 进行各个因素的标准化处理, 对各列数据进行零中心归一化操作, 将数据归一到同一数量级. 数据清洗完成后, 执行 FPGrowth 算法对每个项进行挖掘, 在界面设置算法支持度为 0.62, 可得到各个因素因素对抽油杆机采油系统效率的影响程度, 即可得到整个频繁项集. 流程创建提交完成后, 大数据分析后台进行基于 Spark 平台的分布式数据分析, 分析完成后, 数据结果存储于 Hive 数据仓库仓库中, 用户通过查看该数据流程分析调取最终数据分析结果并进行可视化展示. 流程如图 2 所示, 以此得到各个影响因素对抽油机井采油系统效率的关联程度, 如图 3 所示. 饼图效果图如图 4 所示.

通过石油工业知识挖掘系统分析的抽油机井采油系统效率与影响因素的关联性分析, 华北油田的专家与工程师根据经验对结果满意, 为接下来的抽油机井采油系统效率预测奠定了良好的基础.

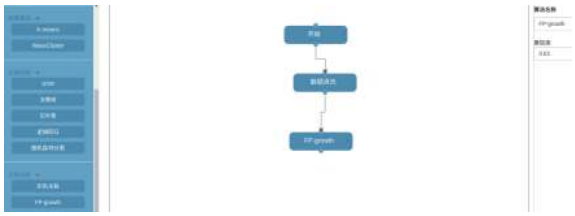

图 2 系统流程操作图

影响因素	关联度
日产液 (t)	0.863
泵深 (m)	0.725
动液面 (m)	0.603
冲程 (m)	0.720
冲次 (n/min)	0.561

图 3 FPGrowth 算法关联性分析结果图

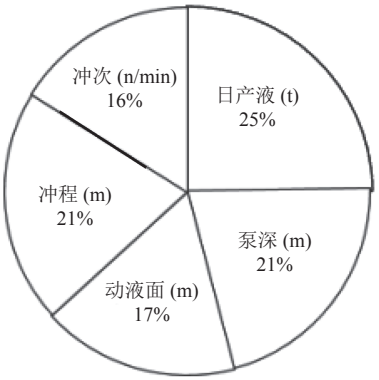

图 4 FPGrowth 算法关联性分析结果饼图

3 总结

本文提出了一种 Web 架构驱动的石油工业知识挖掘系统, 用于以石油工业数据为基础, 结合采油工程数据挖掘领域相关算法和大数据平台进行知识挖掘. 包含了可控工作流程的知识挖掘系统, 实现了通过简单的拖拽操作完成模型构建并进行数据分析; 包含了自助式数据挖掘模块, 通过可视化操作流程与丰富图表展示结果, 帮助用户发现数据中的规律. 用户可以直接忽略大数据底层搭建与编辑大数据算法等工作, 直接通过本系统进行数据收集、数据提取、模型建模、模型执行以及结果可视化展示, 目前已经在华北油田部署并运行超过 1 年, 为该单位的石油大数据知识挖掘发挥了重要作用, 通过发现石油大数据之间显性与隐性关系, 指导实际项目生产, 已经成为该单位提升油气产量、降低产量递减速率、提升剩余油开采几率和尽可能解决储采失衡问题的重要数据支撑与理论依据.

参考文献

1 李金诺. 浅谈石油行业大数据的发展趋势. 价值工程, 2013, 32(29): 172–174. [doi: 10.3969/j.issn.1006-4311.2013.29.090]

2 路宽一, 张庆霖. 关于架构石油行业大数据的探讨. 信息系统工程, 2016, (1): 121–122. [doi: 10.3969/j.issn.1001-2362.2016.01.084]

3 韦博. 互联网背景下的石油行业大数据的信息化应用. 电脑知识与技术, 2017, 13(24): 25–26.
4 段新民, 陈清华, 李琴. 石油勘探开发数据信息与数据挖掘. 情报杂志, 2004, 23(7): 111–112. [doi: 10.3969/j.issn.1002-1965.2004.07.046]
5 谭锋奇, 李洪奇, 孟照旭, 等. 数据挖掘方法在石油勘探开发中的应用研究. 石油地球物理勘探, 2010, 45(1): 85–91.
6 檀朝东, 张恒汝, 马永忠, 等. 油气生产大数据挖掘系统的研究及应用. 数码设计, 2016, 5(1): 49–52, 5.
7 张尘. 数据挖掘技术在石油勘探中的应用研究. 中国石油和化工标准与质量, 2014, (6): 49. [doi: 10.3969/j.issn.1673-4076.2014.06.060]
8 Hansen KM, Zhang WS, Fernandes J. Flexible generation of pervasive web services using OSGi declarative services and OWL ontologies. Proceedings of 2008 15th Asia-Pacific Software Engineering Conference. Beijing. 2008. 135–142.
9 单康康, 王佶, 常晓洁, 等. 高校网络日志大数据分析平台研究. 计算机时代, 2017, (4): 86–88.
10 Zhang WS, Xu L, Li ZW, *et al*. A deep-intelligence framework for online video processing. IEEE Software, 2016, 33(2): 44–51. [doi: 10.1109/MS.2016.31]
11 Zhang WS, Lv H, Xu L, *et al*. An online-offline combined big data mining platform. Proceedings of 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress. Orlando, FL, USA. 2017. 1220–1225.
12 Shidaganti G, Prakash S. Feedback analysis of unstructured data from collabrative networking a BigData analytics approach. Proceedings of International Conference on Circuits, Communication, Control and Computing. Bangalore, India. 2014. 343–347.
13 Apache HBase. https://hbase.apache.org/.
14 Apache Hive. http://hive.apache.org/.
15 李亮, 聂瑞华. 高性能计算平台的 IO 性能测试与分析. 计算机与现代化, 2011, (6): 160–164. [doi: 10.3969/j.issn.1006-2475.2011.06.045]
16 Apache Hadoop.http://hadoop.apache.org/.
17 Apache Spark. http://spark.apache.org/.
18 陈晓旭, 吴恒, 吴悦文, 等. 基于最小费用最大流的大规模资源调度方法. 软件学报, 2017, 28(3): 598–610. [doi: 10.13328/j.cnki.jos.005167]
19 甘庆明, 郭方元, 韩二涛, 等. 抽油机井系统效率影响因素的灰色关联分析. 石油矿场机械, 2009, 38(11): 82–84. [doi: 10.3969/j.issn.1001-3482.2009.11.023]
20 刘波, 尹俊禄, 陈沥, 等. 影响抽油机井系统效率因素分析及对策. 石油石化节能, 2013, 3(1): 19–22. [doi: 10.3969/j.issn.2095-1493.2013.001.006]
21 陈强. 影响抽油机井机采系统效率因素浅析. 中国石油和化工标准与质量, 2011, 31(12): 296. [doi: 10.3969/j.issn.1673-4076.2011.12.253]
22 申彦. 大规模数据集高效数据挖掘算法研究[博士学位论文]. 镇江: 江苏大学, 2013. 2–5.
23 Deng LL, Lou YS. Improvement and research of FP-growth algorithm based on distributed spark. Proceedings of 2015 International Conference on Cloud Computing and Big Data. Shanghai, China. 2015. 105–108.