DOI:10.3880/j.issn.1006-7647.2019.05.012

基于 FP-growth 的大坝安全监测数据挖掘方法

毛宁宁[1,2],苏怀智[1,2],高建新[1,2]

(1. 河海大学水文水资源与水利工程科学国家重点实验室,江苏 南京 210098;
2. 河海大学水利水电学院,江苏 南京 210098)

摘要:为改善大坝安全监测数据库的数据挖掘方法运行速度慢、占用内存大的问题,提出改进 FP-growth 算法,将已预处理的监测数据剪枝后,生成 Priority 树再进行频繁项挖掘。以此方法挖掘大坝变形量与水温等环境量的相关关系,不仅挖掘速度快、精度高、结果简洁,还能够对比单个因子或分析多个因子耦合与目标变量的关系。实例表明改进后的 FP-growth 算法思想为大坝安全监测数据挖掘提供了一条良好的思路。

关键词:大坝安全监测;大坝变形分析;数据挖掘;关联法则;FP-growth 算法

中图分类号:TV689.1 **文献标志码**:A **文章编号**:1006-7647(2019)05-0078-05

Data mining method for dam safety monitoring based on FP-growth algorithm//MAO Ningning[1,2], SU Huaizhi[1,2], GAO Jianxin[1,2] (1. *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering*, *Hohai University*, *Nanjing* 210098, *China*; 2. *College of Water Conservancy and Hydropower Engineering*, *Hohai University*, *Nanjing* 210098, *China*)

Abstract: In order to improve the current data mining method of the dam safety monitoring database which runs slowly and takes up a lot of computational space, a modified FP-growth algorithm was proposed. The pre-processed monitoring data was pruned, and then frequent item mining was performed after the Priority tree was generated. In the application of exploring the correlation between dam deformation and water temperature and other environmental quantities, the proposed method not only has high mining speed, high precision, and simple results, but also can compare a single factor or analyze the relationship between multiple factor coupling with target variables. The example shows that the improved FP-growth algorithm provides a good idea for dam safety monitoring data mining.

Key words: dam safety monitoring; dam deformation analysis; data mining; association rule; FP-growth algorithm

数据挖掘方法常应用于大坝安全监测数据的分析和处理。工程中常对变形、渗流、应力和裂缝开口等大坝安全监测资料进行整编和分析,挖掘数据之间的相关关系,以此评估大坝运行状态,防止灾害发生[1-2]。基于监测资料和坝工原理的统计模型也可分析其相关关系,但实际监测资料存在不连续、精确度低的情况,复相关系数低,且无法对比相似的影响因子或分析多因子耦合作用的影响[3]。

近年来,学者们对大坝安全监测数据挖掘方法展开了研究。苏振华等[4]采用关联分析方法对温度数据进行深度挖掘,确定了溪洛渡大坝施工过程中冷却通水的最高温度。张海燕[5]基于关联规则决策树法进行了大坝安全监测数据挖掘,并利用上下游水位、降雨量、温度等数据对坝顶沉降进行预测。王伟等[6]将大坝安全监控统计模型的求解转换为多目标函数的优化,利用混合蛙跳算法同步确定调整系数和回归系数,建立基于混合蛙跳算法的混凝土坝加权变形预模型。FP-growth 算法是 Apriori 算法的延伸,阮志毅[7]将 FP-Growth 算法和空间多尺度剖分进行结合,提出一种频繁项集的精确挖掘算法。顾军华等[8]提出一种新的基于 Spark 的并行 FP-Growth 算法—BFPG,以提高算法的执行效率。刘冲等[9]提出的占用内存少、能满足大型数据库挖掘需求的改进的 FP-growth 算法,使挖掘速度大大提高,适合于大型数据库的关联规则挖掘算法。

在大坝安全监测数据中,变形量关乎大坝状态是否稳定。大坝变形量受到包括水位因素、温度因素、时效因素等影响,具有很强的非线性、随机

基金项目:国家重点研发计划(2018YFC0407101,2016YFC0401601);广西重点研发计划(桂科 AB17195074)

作者简介:毛宁宁(1995—),女,硕士研究生,主要从事水工结构工程安全监控研究。E-mail: 2644387194@qq.com

通信作者:苏怀智(1973—),男,教授,博士,主要从事涉水工程安全防控与提能延寿研究。E-mail: su_huaizhi@hhu.edu.cn

性[10-11]。本文基于关联规则原理,应用改进的 FP-growth 算法,依据大坝安全监测数据,挖掘变形量与其影响因子之间的相关关系,分析大坝运行性态。实例结合梅山水库的变形和环境量监测资料,利用此方法对数据进行横向和纵向的比对,实现对各个影响因子进行综合评价。

1 改进的 FP-growth 算法

1.1 关联规则

关联规则是数据挖掘方法之一,涉及 2 个基本概念:支持度、置信度。

a. 支持度(support):包含 M 且包含 N 的元组数占总元组数的比例,即项集 A 和项集 B 在数据库中同时出现的概率:

$$S(M \Rightarrow N) = \frac{|W_{M \cup N}|}{|W|} = P(M \cup N) \qquad (1)$$

式中:$W_{M\cup N}$为包含 M 且包含 N 的元组数;W 为总元组数。

b. 置信度(confidence):包含 M 且包含 N 的元组数占包含 M 的元组数的比例,即含有项集 M 的事件中项集 N 同时出现的概率:

$$C(M \Rightarrow N) = \frac{|W_{M \cup N}|}{|W_M|} = P(N/M) \qquad (2)$$

式中:W_M 为包含 M 的元组数。

关联规则即根据给定的最小支持度 S_{min}和最小置信度 C_{min}在事务数据库 W 中找出的事物相关关系[12]。

1.2 改进的 FP-growth 算法

探索不同影响因子在数据库中的重要程度,即是发现频繁项集的过程,亦是对关联规则中“支持度”的应用。FP-Growth 算法建立在 Apriori 算法思想的基础上,在挖掘频繁模式的算法中应用最广。算法采用一种紧凑的数据结构组织构成频繁模式树(Priority 树),通过压缩方式存储数据库中的数据,并直接从 Priority 树中提取频繁项集[12]。

当数据库总量大、事务集数目多或频繁项集的数目大时,重复扫描降低了运行速度,同时在进行数据挖掘时难以抓住主次,找到有用的关联规则。改进的 FP-growth 算法在计算过程中对数据库进行剪枝,以提高计算效率。算法具体过程如下:

输入:事务数据库 W;最小支持度;最小置信度。

步骤 1:扫描数据库 W,以不小于 S_{min}为条件找出频繁项集,并得到其出现的次数计数 m(或支持度),注意区别于 FP-Growth 算法,这一步无须产生候选项集。按照支持度递减排列频繁项集各项,得到频繁项集集合 L。设 $L=\{I_m, I_{m-1}, \cdots, I_1\}$(其中 I_m 的支持度最高,I_1 的支持度最小)。

步骤 2:将支持度小于 S_{min}的项从各事务中删除,再次扫描数据库 W,按照频繁项集支持度递减的次序重新排列各事务中的项,得到数据库 W'。

步骤 3:根据 L 中的各项的支持度大小,按照以下规则由小到大依次构造各项数据库子集,并利用 FP-growth 算法分别对其 Priority 树分支进行约束频繁项挖掘:扫描数据库 W',从中提取所有包含项 $I_i(i=m, m-1, \cdots, 1)$的频繁项集,然后删除这些事务中支持度小于该项的支持度的项集,所得事务集合便为项 I_i 的数据库子集 W_i;其次利用 FP-growth 算法对数据库子集 W'进行频繁项集挖掘;最后构造该项的条件模式基,然后构造其条件 Priority 树,在该条件 Priority 树上挖掘出包含该项的频繁项集 C_L,即完成在数据库子集 W_i 上的约束频繁项集的挖掘。

步骤 4:当 L 中所有的项的约束频繁项集 C_L 被依次挖掘出来后,合并这些约束频繁项集,即取这些约束频繁项集 C_L 的并集,便可得到数据库 W 的所有频繁项集,结束挖掘过程[9,12]。

2 基于改进 FP-growth 算法的大坝运行性态分析方法

2.1 数据预处理

大坝监测数据是不连续且无序的,计算前需对数据进行预处理。影响大坝变形的因素众多,假定因变量 Q 为某大坝运行时期的某处变形量,基于以下自变量建立数据模型:水压因子 A、气温因子 B、水温因子 C、时效因子 D[13-15]。

2.1.1 异常值处理

实际所得的大坝监测数据中,有些是连续的自动化监测值,有些是不连续的人工监测值,同时也存在数据缺失或数据异常的情况,如由于仪器损坏导致某一时间段内的测值缺失,或者测水温或气温时,由于某种原因测值在某一时段出现异常等。本算法中采取忽略元组法,舍去自变量缺失或异常时间段内的所有数据。

2.1.2 数据集成变换

数据集成变换能够消除不同属性的数值因大小不一而造成的计算偏差,将不同量级的数据缩放至同一数值区间,以便比较和处理。以自变量 A 为例,假定某特定时间段内数值的最大值 A_{max}和最小值 A_{min}为规范值,通过公式:

$$A' = \frac{A - A_{min}}{A_{max} - A_{min}} \qquad (3)$$

将自变量测值 A(或者 B、C、D)以及因变量 Q 的测值映射到区间[0, 1]区间。

2.1.3 数据规约

数据规约可将复杂的数据库简化,分析和挖掘也能接近和保持原数据的完整性,并产生相同的分析结果[12]。

a. 采用离散化对数据进行规约:将给定的连续属性值分区,采用式(4)将4个自变量 A、B、C、D 变换后的值再映射到区间[0, 10]区间:

$$A'' = 10 \mid A' - Q' \mid \tag{4}$$

式中:A''的大小反映了自变量与因变量的离散程度。

b. 通过将属性值域划分为区间,用区间的标记替代实际的数据集:由"等距分箱"思想在某一时段对 A''、B''、C''、D''取整数,数值相同的分为一类,则 A 可得到 $A_0 \sim A_9$ 不同项,B、C、D 同理。选取某时段的数据,按照上述步骤进行数据预处理,每个时段即可得到集合$\{A_x, B_x, C_x, D_x\}$,其中 x 为 1 ~ 10 的整数,由此生成数据库 K。

2.2 基于改进 FP-growth 算法的大坝安全监测数据挖掘步骤

具体挖掘流程见图1,步骤如下。

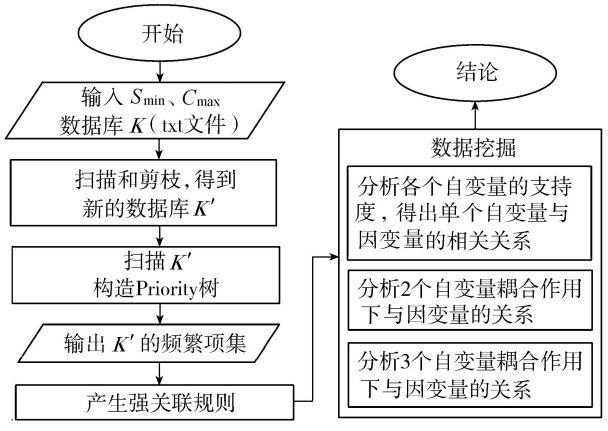

图1 改进的 FP-growth 算法流程

步骤1:参考大坝安全监测的资料,输入合适的最小支持度、最小置信度,将已预处理的大坝监测数据库 K 生成 txt 文件,导入改进 FP-growth 算法 Python 程序。

步骤2:利用 Python 程序对数据库 K 进行扫描,根据不同监测数据与支持度的对比结果进行剪枝,得到新的数据库 K',这一过程剪去出现几率小于置信度的项集以及相关分支,即部分异常的数据。例如在实例中,由于读数误差造成的与平均值偏差较远的水压项 A_9、A_8 等在此步骤中被剪去。

步骤3:利用 Python 程序对数据库 K' 进行扫描,构造 Priority 树。

步骤4:对 Priority 树进行关联规则分析运算,挖掘大坝监测数据库 K'中频繁项集以及强关联规则。

从3个角度对运算结果进行分析论证:分析$\{A_x\}$、$\{B_x\}$、$\{C_x\}$、$\{D_x\}$的支持度,得出监测数据与单个环境量之间的关系;分析含$\{A_x, B_x\}$、$\{A_x, C_x\}$、$\{A_x, D_x\}$、$\{B_x, C_x\}$、$\{B_x, D_x\}$、$\{C_x, D_x\}$、$\{A_x, B_x, C_x\}$、$\{A_x, B_x, D_x\}$、$\{A_x, C_x, D_x\}$、$\{A_x, B_x, C_x, D_x\}$的支持度,得出监测数据与2个环境量耦合作用下的关系;分析含$\{A_x, B_x, C_x\}$、$\{A_x, B_x, D_x\}$、$\{A_x, C_x, D_x\}$、$\{A_x, B_x, C_x, D_x\}$的支持度,得出监测数据与3个环境量耦合作用下的关系。

3 工程实例

梅山水库位于淮河支流史河上游的安徽省金寨县境内,流域面积 1 970 km^2,于 1954 年 3 月动工,1956 年 4 月竣工,是一座以防洪为主、结合灌溉、发电、航运、水产养殖等效益的多年运行的老坝。选取梅山水库 2012—2014 年间 3 组环境量监测资料(上游水位、平均气温、平均水温)以及位于坝顶的 PL2 测点(如图2)在大坝横截面方向(即上下游方向,简称 X 方向)的测值为基本资料进行分析。梅山大坝坝体水平位移监测采用垂线法,其测点的具体布置见图2。

3.1 数据预处理

现有资料中测点 PL2 在 X 方向上的位移测读频率为一周一测,其余环境量测值为一天一测。其中上游水位缺少 2012 年 3 月 22 日—2012 年 4 月 4 日 2 周的测值,PL2 测点的位移监测数据缺少 2013 年 9 月 26 日和 2014 年 3 月 27 日 2 次测值。将相关时间段内的所有数据剔除,取余下时段中上游水位、平均气温、平均水温一周测值的平均值与当周的位移测值,过程线见图3 和图4。

3.2 数据集成变换与规约

按照前述步骤对数据进行数据预处理,得到数据集 K,部分结果见表1(全部数据共 153 组)。表1中下标 0 ~ 9 表示各个自变量与因变量的离散程度。数字越小,表示此自变量与因变量有关的可能性较大;数字越大,表示自变量与因变量有关的可能性较小。

3.3 数据集扫描与剪枝

将数据集 K 生成 txt 文件,导入 Python 中,经过查阅相关资料和多次试验,选取 $S_{min} = 0.9$, $C_{min} = 0.1$,对数据集 K 进行剪枝,生成新的书库的 K'。此时数据集 K 中的项目,如 $A_5 \sim A_9$、$B_3 \sim B_9$、$C_3 \sim C_9$ 被剪去,得到最终的事务集 K',部分结果见表2(全部数据共 153 组)。

3.4 构建 Priority 树

由事务集 K'带入 Python 中运行算法,把 A、B、C、D 作为项,生成 Priority 树,根据计算结果绘制的 Priority 树见图5。

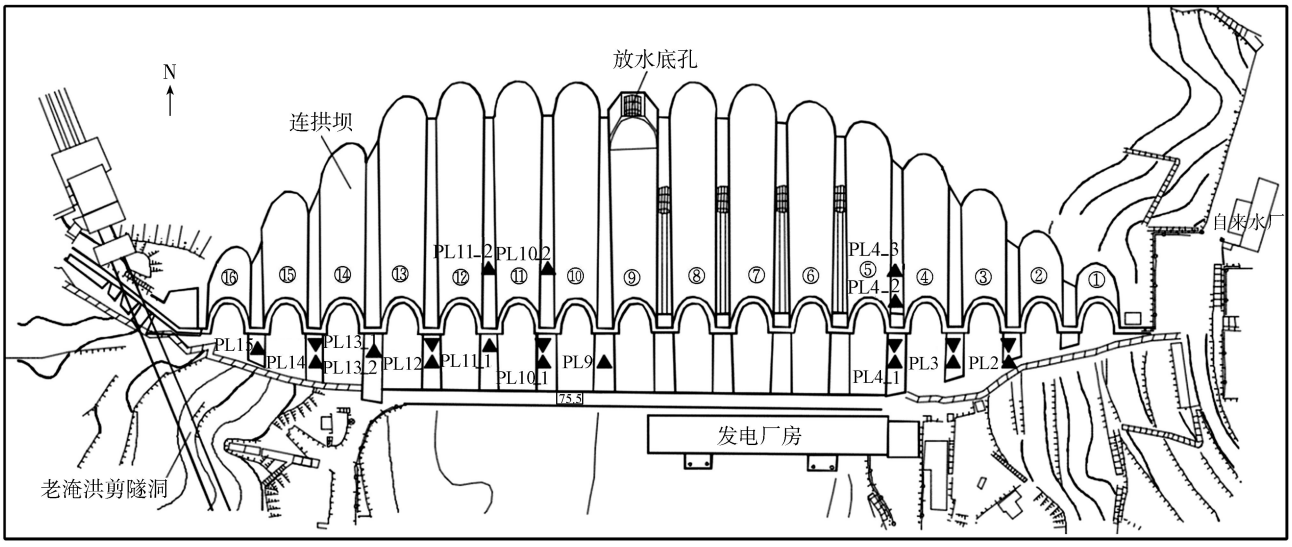

图 2　梅山水库水平位移正倒垂线(自动化和人工)测点布置

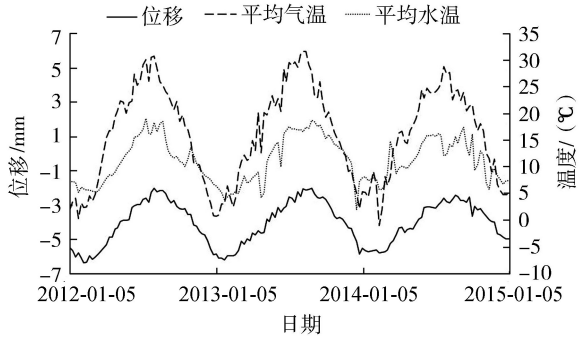

图 3　梅山水库 PL2 测点位移与温度过程线

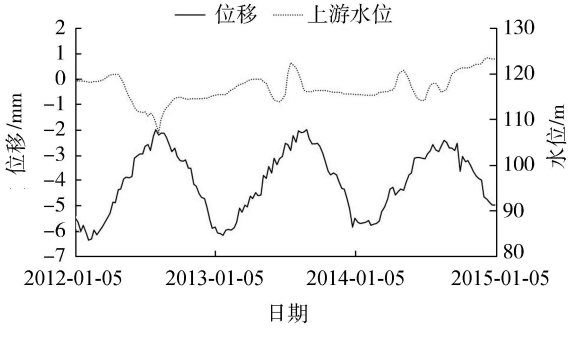

图 4　梅山水库 PL2 测点位移与上游水位过程线

表 1　梅山水库数据集 K(部分)

时间	上游水位	平均气温	平均水温	时效
2012-01-05	A_5	B_1	C_1	D_2
2012-01-12	A_5	B_0	C_1	D_2
2012-01-19	A_6	B_1	C_2	D_1
2012-01-26	A_6	B_1	C_1	D_1
2012-02-02	A_6	B_1	C_2	D_0
2012-02-09	A_7	B_1	C_2	D_0
2012-02-16	A_7	B_1	C_2	D_0
2012-02-23	A_6	B_1	C_1	D_0

Priority 树将上游水位、平均气温、平均水温、时效 4 组复杂的数据与位移之间的关系梳理的简明清晰。由 Priority 树可知，B 为 Priority 树的主枝干，且

表 2　梅山水库数据集 K'(部分)

时间	事务集	日期	事务集
2012-01-05	$\{B_1, C_1, D_2\}$	2012-02-02	$\{B_1, C_2, D_0\}$
2012-01-12	$\{B_0, C_1, D_2\}$	2012-02-09	$\{B_1, C_2, D_0\}$
2012-01-19	$\{B_1, C_2, D_1\}$	2012-02-16	$\{B_1, C_2, D_0\}$
2012-01-26	$\{B_1, C_1, D_1\}$	2012-02-23	$\{B_1, C_1, D_0\}$

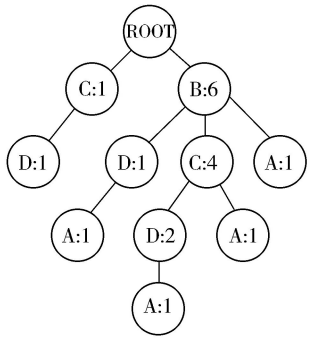

图 5　梅山水库运算结果 Priority 树

B 的计数比 C 多，说明与大坝变形量关系最密切的是 B 平均气温因子，其次是 C 平均水温因子。再次，Priority 树中 A 上游水位因子与 D 时效因子计数相当，说明与 B 和 C 相比，其与位移之间的关系较不密切。

3.5　频繁项集及其支持度

得到 Priority 树之后，利用 python 程序计算各个项集的支持度，得到的所有项集及其支持度见表 3。

由表 3 可知，在含有 2 个因子的项集中，频繁项集 $\{B, C\}$ 支持度为 82%，其次是频繁项集 $\{A, B\}$ 支持度为 78%，再次之是频繁项集 $\{C, D\}$ 支持度为 70%，说明当 2 项因子耦合作用的情况下，气温因子和水温因子的耦合作用，是导致大坝变形的主要原因，而气温因子和上游水位交互作用，以及水温因子与时效因子交互作用对大坝变形也有很大的影响。同样，不经过频率分析，通过对 Priority 树枝干进行

分析(即图 5 中数字比)也可得到上述结论。在含有 3 个因子的项集中,频繁项集$\{A,B,D\}$的支持度为 84%,比率最高;其次,频繁项集$\{A,B,C\}$的支持度为 71%;再次,频繁项集$\{B,C,D\}$的支持度为 69%。说明 3 项因子耦合作用的情况下,上游水位因子、气温因子和时效因子的耦合作用,对大坝变形的影响最大,而上游水位因子、气温因子和水温因子交互作用,以及气温因子、水温因子与时效因子交互作用对大坝变形也有影响。

表 3　挖掘数据库 W' 所得到的项集

频繁项集	支持度/%	频繁项集	支持度/%
$\{A,B,C,D\}$	59	$\{B,C\}$	82
$\{A,B,C\}$	71	A	78
$\{A,B,D\}$	84	B	99
$\{B,C,D\}$	69	C	93
$\{A,B\}$	78	D	75
$\{C,D\}$	70		

4　结　语

a. 关联规则挖掘大坝安全监测数据时,对监测资料的完整性要求较低,并且可以比较同类的影响因子重要程度,如比较“气温因子”与“水温因子”的重要程度,也可将不同因子耦合作用的情况进行对比。以上通过统计模型则很难实现。

b. 改进的 FP-growth 算法新颖、思路清晰、结果简约、实用性高,可通过多种编程软件来实现。利用改进的 FP-growth 算法挖掘大坝变形监测数据库,能够很好地建立大坝变形量和影响因子之间的相关关系,得出简明结论。实例表明,改进后的 FP-growth 算法为大坝安全监测数据挖掘提供了一条良好的思路。

参考文献:

[1] 顾冲时,张晶梅. 大坝服役非概率可靠性分析方法[J]. 水利水电科技进展,2018,38(5):1-9. (GU Chongshi, ZHANG Jingmei. Non-probabilistic reliability analysis methods of dam service performance[J]. Advances in Science and Technology of Water Resources, 2018, 38(5):1-9. (in Chinese))

[2] SU H Z, WU Z R, WEN Z P. Identification model for dam behavior based on wavelet network[J]. Computer-Aided Civil and Infrastructure Engineering, 2007, 22(6): 438-448.

[3] CHEN B, WU Z R, LIANG J C, et al. Time-varying identification model for crack monitoring data from concrete dams based on support vector regression and the Bayesian framework [J]. Mathematical Problems in Engineering, 2017(5):1-11.

[4] 苏振华,周宜红,赵春菊,等. 基于数据挖掘技术的溪洛渡大坝施工期温度监测数据分析[J]. 水电能源科学,2016,34(3):70-73. (SU Zhenhua, ZHOU Yihong, ZHAO Chunju, et al. Analysis of temperature monitoring data of Xiluodu Dam during construction based on data mining[J]. Water Resources and Power, 2016, 34(3): 70-73. (in Chinese))

[5] 张海燕. 数据挖掘技术在大坝安全监测系中的研究与应用[D]. 兰州:兰州理工大学,2013.

[6] 王伟,沈振中,钟启明. 基于混合蛙跳算法的混凝土坝加权变形预报模型[J]. 水利水电科技进展,2013,33(2):37-41. (WANG Wei, SHEN Zhenzhong, ZHONG Qiming. Weighted deformation forecast for concrete dams based on shuffled frog leaping algorithm[J]. Advances in Science and Technology of Water Resources, 2013, 33(2):37-41. (in Chinese))

[7] 阮志毅. 多尺度 FP-Growth 算法及其在规律路径挖掘中的应用[J]. 自动化学报,2019,45(2):1-17. (RUAN Zhiyi. Multiscale FP-Growth algorithm and its application on regular route mining[J]. Acta Automatica Sinica, 2019,45(2):1-17. (in Chinese))

[8] 顾军华,武君艳,许馨匀. 基于 Spark 的并行 FP-Growth 算法优化及实现[J]. 计算机应用,2018,38(11):3069-3074. (GU Junhua, WU Junyan, XU Xinyun. Optimization and implementation of parallel FP-Growth algorithm based on Spark [J]. Journal of Computer Applications, 2018, 38(11):3069-3074. (in Chinese))

[9] 刘冲,陈晓辉,宋小小. 关联规则中 FP 树算法的研究与改进[J]. 网络安全技术与应用,2012(10):53-55. (LIU Chong, CHEN Xiaohui, SONG Xiaoxiao. Research and improvement on FP-tree algorithm of association rule[J]. Network Security Technology & Application, 2012(10): 53-55. (in Chinese))

[10] SU H Z, WEN Z P, CHEN Z X, et al. Dam safety prediction model considering chaotic characteristics in prototype monitoring data series[J]. Structural Health Monitoring, 2016, 15(6):639-649.

[11] HE Jinping, JIANG Zhenxiang, ZHAO Cheng, PENG Zhengquan, SHI Yuqun. Cloud-Verhulst hybrid prediction model for dam deformation under uncertain conditions[J]. Water Science and Engineering, 2018, 11(1): 61-67.

[12] 王振武. 数据挖掘算法原理与实现[M]. 北京:清华大学出版社,2017.

[13] SU H Z, LI X, YANG B B, et al. Wavelet support vector machine-based prediction model of dam deformation[J]. Mechanical Systems and Signal Processing, 2018, 110(15): 412-427.

[14] 李家田,苏怀智,赵海超,等. 基于蒙特卡罗模拟的混凝土坝渗流性态区间综合评价[J]. 水利水电科技进展,2018,38(3):32-35. (LI Jiatian, SU Huaizhi, ZHAO Haichao, et al. Comprehensive interval assessment of seepage behavior for concrete dams based on Monte Carlo simulation[J]. Advances in Science and Technology of Water Resources, 2018, 38(3):32-35. (in Chinese))

[15] SU H Z, WEN Z P, WU Z R. Study on an intelligent inference engine in early-warning system of dam health[J]. Water Resources Management, 2011, 25(6):1545-1563.

(收稿日期:2018-10-23　编辑:郑孝宇)