序列模式挖掘在通信网络告警预测中的应用

张光兰 杨秋辉 程雪梅 姜 科 王 帅 谭武坤
(四川大学计算机学院(软件学院) 成都 610000)

摘 要 告警预测是保证整个网络的稳定性和可靠性的技术之一。现有的告警预测技术存在未考虑告警数据的时间顺序、难以获取先验知识等缺陷。由此,提出了一种基于拓扑约束的序列模式挖掘方法以发现有意义的告警序列模式。该方法主要考虑网络节点之间的拓扑连接关系,将其作为告警序列模式挖掘的约束条件;并且为了发现非频繁重大告警模式,改进了序列模式挖掘的剪枝操作,将包含重大告警的序列模式直接保留。实验结果表明,采用基于拓扑约束的序列模式挖掘方法挖掘出的告警序列模式可以提高网络告警预测的精度和效率,并能较准确地预测非频繁的"重大"告警。

关键词 通信网络,告警预测,序列模式挖掘,网络拓扑结构
中图法分类号 TP311 **文献标识码** A

Application of Sequence Pattern Mining in Communication Network Alarm Prediction

ZHANG Guang-lan YANG Qiu-hui CHENG Xue-mei JIANG Ke WANG Shuai TAN Wu-kun
(School of Software, Sichuan University, Chengdu 610000, China)

Abstract Alarm prediction is one of the techniques that ensures the stability and reliability of the entire network. Existing alarm forecasting technologies have defects such as not considering the time sequence of warning data and difficult to obtain the priori knowledge. Therefore, this paper proposed a sequence pattern mining method based on topological constraints to find a meaningful alarm sequence pattern. This algorithm mainly considers the topological connections between network nodes and takes them as constraints for mining the alarm sequence pattern. In order to find non-frequent major alarm mode, it improves pruning of sequential pattern mining, preserves sequence patterns containing major alarms directly. Experiments show that the alarm sequence mode mined by the sequential pattern mining method based on topological constraints can improve the accuracy and efficiency of the network alarm prediction and predict the infrequent "major" alarms more accurately.

Keywords Communication network, Alarm prediction, Sequence pattern mining, Network topology architecture

1 引言

随着人们对网络的需求越来越大,网络规模逐渐扩大,网络环境也随之变得越来越复杂,导致网络中的告警信息具有海量、冗余、时序相关等特点。利用数据挖掘技术从这些告警数据中挖掘出有意义的知识,是目前一个重要的研究方向[1-2]。网络中的故障往往会引发大量的告警,这些告警是通过网络设备进行传播的,使用数据挖掘技术从这些具有时序相关的告警信息中发现模式可以进行告警预测[3-5]。

网络告警预测是通过对告警的分析,来预测未来网络中可能会出现的告警,帮助网络管理员监控和预测整个网络的状态可能发生的失效,提前做好保护措施,以降低损失。本文对网络告警预测问题进行了研究。通过使用序列模式挖掘技术发现历史告警数据中的告警序列模式,构建告警预测模型,并进行实时告警预测。

2 相关工作

Agrawal 和 Srikant 首次提出了序列模式挖掘。序列模式挖掘的目标是发现事件发生顺序之间的关系,从序列数据库中找出频繁出现的具有时间顺序的子序列。Hatonen 等[4-7]首先将序列模式挖掘方法引入到通信网络告警数据库中,提出了 TASA 系统,用来发现和浏览来自告警数据库中的告警序列模式。Pei-Hsin 等[2,8]研究了一种从 GSM 系统告警数据中挖掘告警序列模式的方法,该方法根据告警数据的特点设计和实现了数据清理的操作,然后利用时间约束来限制告警之间的时间差,最后使用了一种新的序列模式挖掘算法 MSAP 来发现有用的告警序列模式。Garcia 等[9]在开源监测软件 Osmius 上增加了一个预测模块,用来对事件进行预测,通过频繁模式挖掘和序列模式挖掘两种方式实现了对事件的预测,指出序列模式挖掘更加适合于对网络告警的预测。

但目前的序列模式挖掘在通信网络告警中的运用并没有考虑产生告警的设备之间拓扑结构的连接关系,且使用固定的时间窗口将同一时间窗口内的告警视为一个事务,导致挖掘出不完整的模式。其次,根据告警出现的频次进行挖掘使结果未包含频次较低的"重大"告警。

张光兰(1994-),女,硕士,主要研究方向为软件自动化测试,E-mail:214608304@qq.com;**杨秋辉**(1970-),女,副教授,主要研究方向为软件自动化测试框架和平台、自动化单元测试工具、数据挖掘等;**程雪梅**(1991-),女,硕士,主要研究方向为数据挖掘;**姜 科**(1994-),男,硕士,主要研究方向为软件自动化测试;**王 帅**(1992-),男,硕士,主要研究方向为软件自动化测试;**谭武坤**(1990-),男,硕士,主要研究方向为软件自动化测试。

3 整体方案

3.1 基本思想

本文的整体方案如图 1 所示，包括 3 个步骤：1)对历史告警数据进行预处理。预处理之后的历史告警数据不包括冗余告警、不完整告警、非法告警和闪段式告警等。2)将网络拓扑约束数据库和预处理后的历史告警数据作为输入，使用基于拓扑约束的序列模式挖掘从告警数据中挖掘出告警序列模式。3)利用基于拓扑约束的告警序列模式挖掘出的告警序列模式来构建告警预测模型，用于网络告警预测。

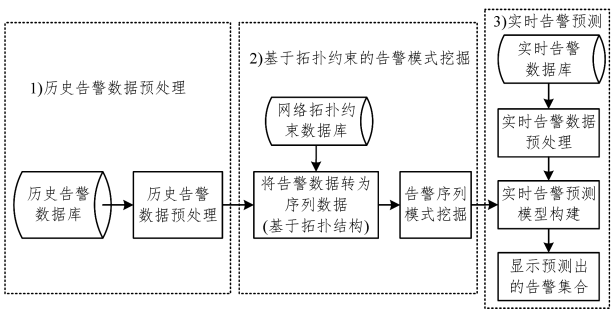

图 1 方案的整体流程

3.2 基于拓扑约束的告警序列模式挖掘

首先面向拓扑结构对通信网络进行建模，通过该模型判断网络设备之间是否满足拓扑连接的约束。然后将告警数据基于拓扑约束转换为序列数据；最后使用序列模式挖掘算法从序列数据中挖掘出告警序列模式。

(1)面向拓扑约束的通信网络建模

通过对通信网络结构和告警传播的分析可以发现，告警是基于网络拓扑进行传播的，告警可能从低层传播到高层，也可能从高层传播到低层。但无论告警怎么传播，传播的路径都是被限制在同一个网络元素簇(简称网元簇)中，网元簇被定义为由许多拓扑相连的网络元素组成的集合，属于同一网元簇的网络元素视为是网络拓扑相连的[10-11]。告警之间是否满足拓扑约束，不需要考虑告警对应的网络元素之间是否存在连接边，只需要判断告警对应的网络元素是否属于同一网元簇即可。本文对网络进行了抽象，如图 2 所示，将实际的网络结构抽象为一般的拓扑模型图，将通信网络建模为一个有向无环图，图中的每个节点表示一个网络元素，每条边表示节点之间的功能依赖关系[11-12]。

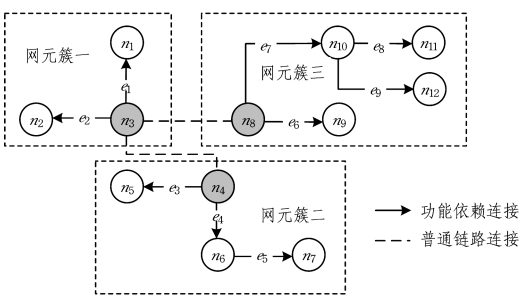

图 2 通信网络拓扑结构图示例

(2)基于拓扑约束的数据转换

序列模式挖掘的输入是序列数据，因此必须先将告警数据转换为事务数据，然后转换为序列数据。针对此问题，本文提出了一种基于拓扑约束的重叠滑动时间窗口：为了包括任意有可能的传播序列的起点，窗口的点沿着时间线以一小步的时间跨度滑动，窗口的终点不再是通过固定时间窗口的长度决定的，而是通过相邻告警之间的拓扑连接关系决定的，新的被覆盖的告警被包含到当前事务中当且仅当它们与之前已经被包含到事务中的告警之间是拓扑相连的。如果满足拓扑约束，窗口继续增长直至达到一个特定的阈值，否则当前窗口结束，新窗口开始。最终的结果是每个事务中的告警都是拓扑相连的。事务数据转换为序列数据也是一样的原理。

(3)基于改进剪枝策略的告警序列模式挖掘

使用序列模式挖掘算法从历史告警数据中挖掘出告警序列模式，如果挖掘算法的效率不高，则会导致挖掘的过程非常耗时。序列模式挖掘算法有 Apriori-like 算法、GSP 算法、SPADE 算法等，其中 SPADE 算法本身的效率较高且考虑了时间约束特点。结合告警数据的时序特性，本文选取 SPADE 算法作为基础。同时考虑到“重大”告警相关模式的非频繁性，提出了 PruneSPADE 算法，修改了 SPADE 算法的剪枝策略，在算法的剪枝阶段直接将包含“重大”告警的序列保留，不会因为不满足最小支持度而被剪掉，保证最后结果中包含“重大”告警相关模式。

如图 3 所示，PruneSPADE 算法主要包括计算频繁 1-序列、计算频繁 2-序列作为父类、磁盘扫描、临时连接(针对每个等价类通过深度优先搜索和广度优先搜索计算出所有其他的频繁序列)、修剪序列 5 个步骤。

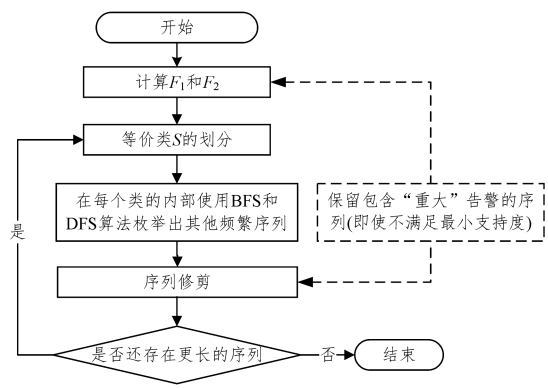

图 3 PruneSPADE 算法的流程图

3.3 实时告警预测

对实时告警数据进行预处理后，根据告警序列模式和实时告警数据，基于告警序列模式匹配进行实时告警预测。假设告警序列模式 P 包含 n 项，在指定大小的监测窗口内，模式 P 的前 $n-1$ 项和实时告警匹配成功，则可以认为模式 P 的第 n 项告警将在未来一段时间内发生。匹配成功包括 3 个条件：1)告警序列模式中除最后一项的每一项都在实时告警中有一个对应的匹配告警；2)实时告警中匹配到的告警满足告警序列模式中告警的顺序规定；3)实时告警中匹配到的告警都在告警序列模式的窗口之中。

将实时告警预测主要分为 5 个步骤：

(1)对于新来的一段时间 t 内的告警数据，首先进行预处理。

(2)从告警序列模式数据库中读取告警序列模式，并存放到数据结构中，这个数据结构包含了告警序列号的告警发生时间、告警属于的事务标号、标志位。

(3)针对时间段 t 内的每一条实时告警，遍历所有的告警序列模式，针对每个告警序列模式找出告警序列号与实时告警的序列号相同的项，然后判断该项与其前一项的事务标号

是否相同。如果不相同，则表示不属于同一事务；如果相同，则表示属于同一事务。然后需要当前读入的告警的发生时间与该事务的第一项的发生时间差小于或等于属于同一事务的最小时间窗口，并且若前一项的标志位已经被置为 1，则将该项的标志位置为 1。

(4)遍历完成以后，将数据结构中标志位全为 1 的告警序列模式输出，无需再进行预测，将该告警序列模式的所有项的标志位重置为 0，返回读下一条告警，直到遍历完时间段 t 内所有的告警数据。

(5)最后遍历告警序列模式的数据结构，找出满足前 $n-1$ 项标志位都被置为 1，第 n 项标志位为 0 的告警序列模式，则第 n 项为即将发生的告警，将第 n 项对应的网络告警加入到预测告警集合中。输出预测告警集合，预测结束。

4　实验验证

网络告警预测的主要目的是增强网络系统的可靠性和稳定性。当由于一些故障或者配置更新而使操作条件发生改变时，网络设备经常会生成告警，这些告警通过监控过程被收集起来[13]。本文将某通信公司提供的数据和仿真数据作为实验数据，对本文提出的基于拓扑约束的序列模式挖掘在告警预测中的应用方案进行实验，并对实验结果进行了分析。

4.1　实验目的和数据

(1)实验目的

验证加入拓扑约束以后是否能提高网络告警预测的精度和效率，此外，针对数据量很少的“重大”告警，是否能够发现其告警序列模式并成功进行预测。

(2)实验数据

本文的实验数据为某通信公司提供的告警数据和按照告警数据格式仿真的数据，其中包含连续 8 天的网络告警数据，大约 5 万条告警，225 个网络元素，256 个不同的告警类型，告警数据包含 4 个等级：提示、次要、重要、紧急。过滤掉告警数据中存在的一些非法告警，例如含有非法属性值的告警、冗余的告警、闪断式告警和属性缺失的告警，最终原始告警数据剩余 19543 条，对应 165 个网络元素，178 个不同的告警类型，将其作为实验数据。

4.2　实验步骤

(1)告警数据预处理

首先，将原始告警数据分为两部分，取前一周的数据进行预处理，将过滤后的数据用于挖掘告警序列模式，取第 8 天的数据作为实时告警数据，用于告警预测。

(2)转换告警数据

序列模式挖掘算法的输入是序列数据，使用固定滑动时间窗口方法和基于拓扑约束的滑动时间窗口分别将过滤后的告警数据转换成为事务数据。然后再使用滑动时间窗口机制将事务型数据转换为序列数据。按照不同的事务划分方式，数据预处理部分共得到两组条件告警数据，分别为按照固定滑动时间窗口划分的一周告警数据和按照基于拓扑约束的滑动时间窗口划分的一周告警数据。

(3)上载网络拓扑结构信息

在将告警数据转换为序列数据以后，不能立即进行告警序列模式的挖掘，因为本文提出的方案需要结合网络节点之间的拓扑连接关系来优化挖掘出的告警序列模式，所以需要将网络拓扑结构中的节点与边，以及它们的连接关系存入到数据库中，用于告警序列模式的挖掘。

(4)挖掘告警序列模式

在经过转换的两组条件告警数据上，使用序列模式挖掘算法(SPADE 算法)，加入拓扑约束的序列模式挖掘算法(T-PruneSPADE 算法)。设置不同的支持度和预测窗口大小，得到对应的告警序列模式。

(5)预测网络告警

将第 8 天的告警数据作为实时告警数据与告警序列模式一起作为输入，设置不同的预测窗口大小，得到最终的告警预测结果，包含已经出现的告警和即将出现的告警。

5　实验结果及分析

本文在对网络告警预测结果进行对比评估时，主要采用了告警预测中的准确率、召回率和 F-值作为告警预测效果的评价指标，以响应时间作为告警预测性能的评价指标[14-16]。下面分别从支持度、预测时间及重大告警等方面进行实验，比较 SPADE 算法和 T-PruneSPADE 算法对应的告警预测的准确率、召回率和 F-值。

5.1　不同支持度下的告警预测结果

以支持度为变量，设置预测窗口大小为 6 h，考查两种算法在不同支持度下的告警预测的性能。从实验结果可以看出，随着支持度的减小，告警预测的准确率和 F-值先增大后减小，这是因为随着支持度的减小，挖掘出的告警序列模式增多。正确的告警预测数量小幅度增加，而总共预测的告警数量大幅度增多，由于准确率是正确的告警预测数量与总共预测的告警数量的比值，因此当支持度很低时准确率反而会下降。召回率随着支持度的减小而增大，因为支持度越小，挖掘出的告警序列模式越多，正确的告警预测数量越多，而预测窗内实际发生的告警数不变，召回率是二者的比值，所以召回率会变高。响应时间随着支持度的减小而增加，特别是当支持度小于 0.3 时，响应时间迅速增加，因为支持度越小，挖掘过程越耗时，挖掘出的告警序列模式越多，导致告警预测的响应时间越长。T-PruneSPADE 算法与 SPADE 算法相比，随着支持度的降低，T-PruneSPADE 算法具有更好的效率。这是因为 T-PruneSPADE 算法将不满足拓扑约束的序列模式过滤掉了，减少了挖掘出的告警序列模式的数量，虽然 T-PruneSPADE 算法改进了 SPADE 算法的剪枝阶段，保留了所有包含“重大”告警的序列，会增加挖掘出的告警序列模式的数量，但是这部分模式数量相对来说非常少。因此从整体上讲，T-PruneSPADE 算法对应的告警预测精度和效率都是最优的。不同支持度下两种算法对应的准确率结果如图 4 所示，召回率结果如图 5 所示，F-值结果如图 6 所示，响应时间的结果如图 7 所示。

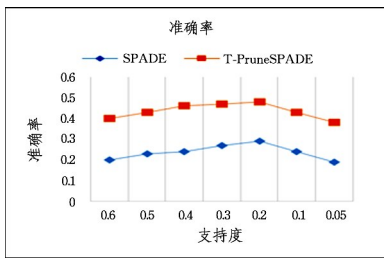

图 4　不同支持度下的准确率告警预测性能

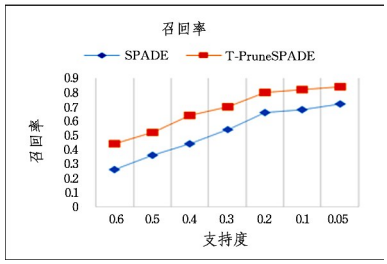

图 5 不同支持度下的召回率告警预测性能

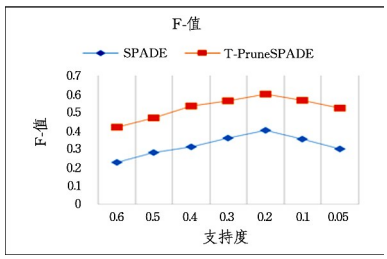

图 6 不同支持度下的 F-值告警预测性能

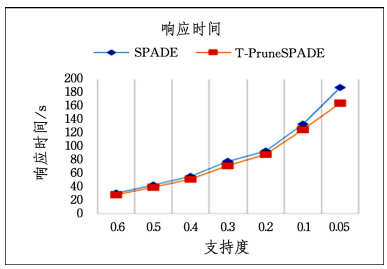

图 7 不同支持度下的响应时间告警预测性能

5.2 不同预测时间窗口下的告警预测结果

以预测时间窗口的大小为变量，设定支持度为 0.3，从实验结果可以看出随着预测时间窗口的增大，告警预测的精度也在提高，特别是当时间窗大于 4 h 后，本文提出的 T-PruneSPADE 算法具有更好的告警预测精度，在整体上都是最优的，其整体精度也是最高的。T-PruneSPADE 算法相比于 SPADE 算法，其精度有了明显的提升。不同预测时间窗口下两种算法对应的准确率告警预测精度如图 8 所示，召回率告警预测精度如图 9 所示，F-值告警预测精度如图 10 所示。

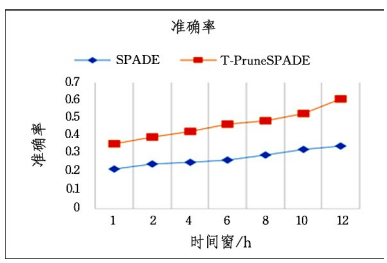

图 8 不同预测时间窗口下的准确率告警预测精度

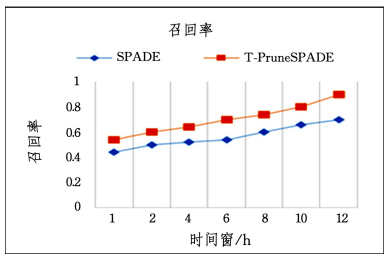

图 9 不同预测时间窗口下的召回率告警预测精度

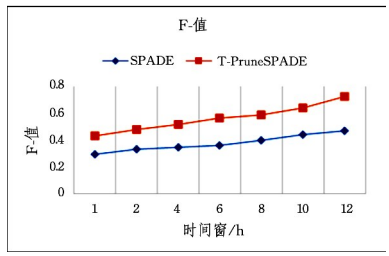

图 10 不同预测时间窗口下的 F-值告警预测精度

5.3 重大告警的预测结果

为了验证方案能预测到“重大”告警，在实验数据中加入了 10 条“重大”告警。从实验结果可以看出，文中提出的方案能够挖掘出这 10 条“重大”告警相关模式，但是只能够预测到 5 条“重大”告警。这是因为有些重大告警只出现极少次或者一次，其对应的告警序列模式具有很强的偶然性，使用这部分模式只能够预测到部分重大告警。

5.4 实验结论

从整个实验结果可以看出，从告警预测的精度上看，整体性能最好的为 T-PruneSPADE 算法，并且在告警预测的准确率、召回率和 F-值上都优于 SPADE 算法。除此之外，T-PruneSPADE算法能够发现“重大”告警相关模式并进行预测。

结束语 本文以某通信公司提供的数据为基础，对网络告警预测技术进行了研究，提出了一种基于拓扑约束的序列模式挖掘方法，并对本文提出的方案进行了实验，结果表明该方案具有较好的告警预测精度和较短的响应时间，并且能够较准确地对“重大”告警进行预测。

参 考 文 献

[1] GAO Z,CHEN Z,FENG Y,et al. Mining Sequential Patterns of Predicates for Fault Localization and Understanding[C]// 2013 IEEE 7th International Conference on Software Security and Reliability (SERE). IEEE,2013:109-118.

[2] WU P H,PENG W C,CHEN M S. Mining sequential alarm patterns in a telecommunication database [C] // International Workshop on Databases in Telecommunications. Springer,2001:37-51.

[3] NUNEZ M,MORALES R,TRIGUERO F. Automatic discovery of rules for predicting network management events [J]. IEEE Journal on Selected Areas in Communications,2002,20(4):736-745.

[4] KLEMETTINEN M,MANNILA H,TOIVONEN H. Rule discovery in telecommunication alarm data [J]. Journal of Network and Systems Management,1999,7(4):395-423.

[5] HATONEN K,KLEMETTINEN M,MANNILA H,et al. Knowledge discovery from telecommunication network alarm databases [C]// Proceedings of the Twelfth International Conference on Data Engineering,1996. IEEE,1996:115-122.

[6] MANNILA H,TOIVONEN H,VERKAMO A I. Discovery of frequent episodes in event sequences [J]. Data Mining and Knowledge Discovery,1997,1(3):259-289.

[7] HATONEN K,KLEMETTINEN M,MANNILA H,et al. TASA:Telecommunication alarm sequence analyzer or how to enjoy faults in your network [C]// Network Operations and Management Symposium,1996,IEEE. IEEE,1996:520-529.

(下转第 563 页)

性分析的结果如图 14 所示，其中：

(1)有 3 个过程模型存在死活动；

(2)有 4 个过程模型存在死锁或活锁；

(3)有 2 个过程模型存在路由活动不匹配；

(4)有 91 个过程模型是合理的。

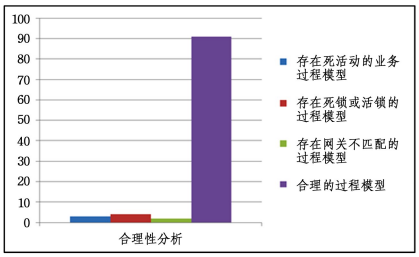

图 14　合理性分析的结果

结束语　BPMN 2.0 过程缺少形式化的语义和分析技术，使得建模者无法确保过程模型的正确性。首先，通过建立 BPMN 2.0 过程到工作流网的映射，使用 Petri 网准确定义过程模型的语义；其次，借助 Petri 网的分析技术，使用这种定义的语义，对 BPMN 2.0 过程模型进行了合理性分析；最后，通过实验表明，这种形式化可以识别 BPMN 2.0 过程模型中存在的语义错误。

本文未定义相容网关、链接事件、补偿事件、错误事件等元素的语义，这是下一步工作的重点。

参 考 文 献

[1] OMG. Business Process Model and Notation (BPMN) Version 2.0[EB/OL]. http://www.omg.org/spec/BPMN/2.0.

[2] PALMER N. XML Process Definition Language[M]. Springer US, 2009.

[3] WOHED P, AALST W M P V D, DUMAS M, et al. Pattern-Based Analysis of the Control-Flow Perspective of UML Activity Diagrams[M]// Conceptual Modeling — ER 2005. Springer Berlin Heidelberg, 2005: 63-78.

[4] DIJKMAN R M, DUMAS M, OUYANG C. Formal semantics and analysis of BPMN process models using Petri nets[J]. Information & Software Technology, 2007, 50(12): 1281-1294.

[5] JIN T, WANG J, YANG Y, et al. Refactor Business Process Models with Maximized Parallelism[J]. IEEE Transactions on Services Computing, 2017, 9(3): 456-468.

[6] AALST W. The application of Petri nets to workflow management[J]. Journal of Circuits System & Computers, 1998, 8(1): 21-66.

[7] DIJKMAN R, GORP P V. BPMN 2.0 Execution Semantics Formalized as Graph Rewrite Rules[J]. Lecture Notes in Business Information Processing, 2010, 67: 16-30.

[8] WONG P Y H, GIBBONS J. A Process Semantics for BPMN[C]// Proceedings of the International Conference on Formal Engineering Methods. Berlin, Germany: Springer-Verlag, 2008: 355-374.

[9] WONG P Y H, GIBBONS J. Formalisations and applications of BPMN[J]. Science of Computer Programming, 2011, 76(8): 633-650.

[10] YE J H, SUN S X, SONG W, et al. Formal Semantics of BPMN Process Models Using YAWL[C]// Proceedings of the International Symposium on Intelligent Information Technology Application. Washiongton, D. C., IEEE, 2008: 70-74.

[11] PRANDI D, QUAGLIA P, ZANNONE N. Formal Analysis of BPMN Via a Translation into COWS[C]// Proceedings the 10th International Conference on Coordination Models and Languages. Berlin, Germany: Springer-Verlag, 2008: 249-263.

[12] LAM V S W. A Precise Execution Semantics for BPMN[J]. Iaeng International Journal of Computer Science, 2012, 39(1): 20-33.

[13] ECKLEDER A, FREYTAG T. WoPeD 2.0 goes BPEL 2.0[C]// German Workshop on Algorithms and TOOLS for Petri Nets, Algorithmen Und Werkzeuge Für Petrinetze (Awpn 2008). Rostock, Germany, 2008: 75-80.

[14] DUMAS M. 过程感知的信息系统[M]. 王建民，等译. 北京：清华大学出版社，2009.

(上接第 538 页)

[8] JAIN-ZHI O, PEI-HSIN W, MING-SYAN C. Experimental results on a constraint based sequential pattern mining for telecommunication alarm data [C]// Proceedings of the Second International Conference on Web Information Systems Engineering. IEEE, 2001: 186-193.

[9] GARCIA R, LLANA L, MALAGON C, et al. Event Prediction in Network Monitoring Systems: Performing Sequential Pattern Mining in Osmius Monitoring Tool [M]// Advances in Data Mining: Applications and Theoretical Aspects. Berlin: Springer-Verlag Berlin, 2010: 632-642.

[10] WANG Z, ZHANG B, LI G. A Topological Constraints Based Sequential Data Mining Approach on Telecom Networks Alarm Data [C]// International Joint Conference on Computational Sciences and Optimization, 2009 (CSO 2009). 2009: 750-754.

[11] MEIRA D M, NOVEMBER B H. A Model For Alarm Correlation in Telecommunications Networks [J]. Federal University of Minas Gerais, 1998.

[12] EIRA D M, NOGUEIRA J M S. Modelling a telecommunication network for fault management applications [C]// Network Operations and Management Symposium, 1998 (NOMS 98). IEEE. IEEE, 1998: 723-732.

[13] JAUDET M, HUSSAIN A, SHARIF K. Temporal classification for fault-prediction in a real-world telecommunications network [C]// Proceedings of the IEEE Symposium on Emerging Technologies, 2005. IEEE, 2005: 209-214.

[14] ZAKI M J, LESH N, OGIHARA M. Predicting failures in event sequences [M]// Data Mining for Scientific and Engineering Applications. Springer, 2001: 515-539.

[15] SALFNER F, LENK M, MALEK M. A survey of online failure prediction methods [J]. Acm Computing Surveys, 2010, 42(3): 10.

[16] ZHONG J, GUO W, WANG Z. Study on network failure prediction based on alarm logs [C]// 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC). 2016: 1-7.