计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181102110

基于日志自动机的业务流程混沌活动过滤方法

李 娟 方贤文 王丽丽 刘祥伟
安徽理工大学数学与大数据学院 安徽 淮南 232001
(571785571@qq.com)

摘 要 业务流程事件日志有时包含混沌活动,混沌活动是独立于流程状态且不受流程约束,会随时随地发生的一类活动。混沌活动的存在会严重影响业务流程挖掘的质量,因此过滤混沌活动成为业务流程管理的关键内容之一。目前,混沌活动的过滤方法主要是从事件日志中过滤不频繁行为,以高频优先为基础的过滤方法并不能有效地过滤日志中的混沌活动。为了解决上述问题,提出了一种基于日志自动机和熵的方法来过滤日志中的混沌活动。首先,根据活动的直接前集率和直接后集率计算得到熵值大的可疑混沌活动集;然后,基于事件日志构建日志自动机,利用日志自动机模型计算得到不频繁弧的活动集与日志中熵值大的活动集,对其取交集得到混沌活动集;最后,运用条件发生概率和行为轮廓确定该混沌活动与其他活动之间的依赖关系,从而决定是在日志中完全删除该混沌活动还是保留该混沌活动在日志中的正确位置而删除其他位置的此活动。案例分析验证了该方法的有效性。

关键词:Petri 网;混沌活动;日志自动机;熵;条件发生概率;行为轮廓
中图法分类号 TP391

Chaotic Activity Filter Method for Business Process Based on Log Automaton

LI Juan,FANG Xian-wen,WANG Li-li and LIU Xiang-wei
College of Mathematics and Big Data,Anhui University of Science and Technology,Huainan,Anhui 232001,China

Abstract Business process event logs sometimes contain chaotic activities,which are a kind of activity independent of process state and free from process constraints,and may happen anytime and anywhere. The existence of chaotic activities can seriously affect the quality of business process mining,so filtering chaotic activities becomes one of the key contents of business process management. At present,the filtering method of chaotic activity mainly filters infrequent behavior from the event the log,and the filtering method based on high frequency priority is not effective in filtering chaotic activities in the log. In order to solve the above problems,a method based on log automata and entropy is proposed to filter chaotic activities in logs. Firstly,a suspicious chaotic activity set with high entropy is obtained by calculating the direct preset rate and direct posterior set rate of activity. Then,the log automata is constructed from the event log. From the log automata model,the intersection of the activity set of infrequent arc and the activity set of high entropy in the log is calculated to obtain the chaotic activity set. Finally,the conditional occurrence probability and behavior profile are used to determine the dependence between the chaotic activity and other activities,so as to decide whether to delete the chaotic activity completely in the log or to keep the chaotic activity in the correct position in the log to delete other activities. The effectiveness of the method is verified by case analysis.

Keywords Petir net,Chaotic activity,Log automaton,Entropy,Conditional occurrence probability,Behavioral profile

1 引言

过程发现是过程挖掘[1]的重要内容之一,旨在从存储过程执行数据日志中发现合理的流程模型。Leemans 等[2]介绍了一种用于流程发现的框架,并提出了使用该框架的 3 种算法。此框架在只传递一次日志的情况下能够保证发现模型的质量。挖掘业务流程模型的大多数发现算法假设事件日志中所有的事件日志都符合业务流程模型的正确执行,因此尽可能地把事件日志的所有行为都合并在业务流程模型中,从而导致业务流程模型较为复杂且精确度和合适度极低。事实证明,真实事件日志中往往包含噪音、非频繁行为和混沌活动等。为了提高发现业务流程模型的质量,以更好地帮助业务流程分析人员理解事件数据,在过去十年内学者们提出了许多过程发现技术,如从包含非频繁行为的事件日志中挖掘过

到稿日期:2018-11-16 返修日期:2019-04-29 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61572035,61272153,61402011);安徽省自然科学基金项目(1508085MF111)
This work supported by the National Natural Science Foundation of China (61572035,61272153,61402011) and Natural Science Foundation of Anhui Province,China(1508085MF111).
通信作者:方贤文(280060673@qq.com)

程模型、基于域的挖掘方法和整线性规划挖掘方法等；产生的业务流程模型形式也逐渐多样化，如 Petri 网、过程树[3]和 BPMN 模型。文献[4]提出了一种数据修复方法，试图检测和修复给定事件数据的异常行为，实验表明该方法能够检测和修改事件数据中大多数类型的异常值。文献[5]提出了一种新的过滤非频繁行为的方法，该方法在现有过程发现算法中的应用显著提高了发现过程模型的质量。在业务流程的执行中，混沌活动独立于业务流程状态，可以在任何时间随意发生，破坏事件日志中活动之间的直接后继关系，隐藏事件日志的正确行为，导致运用基于后继关系的过程发现技术无法过滤混沌活动，从而使挖掘得到的业务流程模型繁冗且实用性低。从事件日志中过滤混沌活动已成为当前研究的重点课题。

在以往的过程发现研究中，学者们主要从事件日志中过滤噪音，挖掘非频繁行为，并区分有效低频行为和无效非频繁行为。现实中，事件日志中存在异常值或者“噪音”，这些噪音可能是数据输入错误和丢失数据造成的，并且在过程发现中低频事件也会导致复杂模型。文献[6]提出了一种基于最大模式挖掘(Maximal Pattern Mining，MPM)方法，基于频率提出根据事件序列中的可执行的日志构建模式过滤噪音，从而确保挖掘模型的合理性。文献[7]通过在执行的流程实例中识别频繁的常见行为和低频行为来检测偏差，并在 Prom 中对人造日志和真实日志进行评估。但是，非频繁行为在流程中并不都是“噪音”，有的非频繁行为能够优化业务流程，因此直接将其删除并不合适。

已有的过滤技术对过滤事件日志中的混沌活动存在一定的局限性。基于事件的过滤技术[8]都是以事件日志中活动发生的频数为标准，在日志中将发生次数少的活动认定为异常值而过滤掉，保留事件日志中发生频数多的活动作为主体行为。这种以频率为基础的过滤方法不能过滤日志中的混沌活动。例如，在事件日志 L 中，活动 X 是混沌活动，若活动序列 $\langle \cdots A, X \cdots \rangle$ 发生的频率高于活动序列 $\langle \cdots B, X \cdots \rangle$ 的频率，则运用基于事件的过滤技术，活动序列 $\langle \cdots B, X \cdots \rangle$ 被认为是异常日志而被直接删除，而活动序列 $\langle \cdots A, X \cdots \rangle$ 发生的频率高，被当作主体活动序列被保留，最终混沌活动 X 并没有在事件日志中被彻底过滤掉。更糟糕的是，混沌活动 X 被视为活动 A 的直接后继，隐藏了活动 A 的真正直接后继关系，从而严重影响了业务流程的质量。归纳挖掘(Inductive Miner)[9]和启发式挖掘(Heuristics Miner)[10]是基于事件日志活动的直接后继关系图和活动之间的最终后继关系来构建业务流程模型，而混沌活动可以随机发生在任何位置，会破坏直接后继关系和最终后继关系。文献[11]提出了从事件日志中过滤混沌活动的算法，但是没有考虑事件日志的语义，也没有说明如何确定事件日志中混沌活动的数量。

本文从混沌活动不受任何约束，可以随机发生在业务流程任何位置的特性出发，运用日志自动机、熵、条件发生概率提出过滤事件日志中混沌活动的算法，并以业务流程模型合适度为判断标准，将混沌活动分为两种类型。第一种混沌活动的存在隐藏了业务流程真实的后继关系，使得业务流程变得繁冗；应该在日志中彻底删除这类混沌活动。第二种混沌活动是业务流程不可缺少的活动，只是由于发生了故障或其他原因导致此活动在日志中任意发生；对于这种混沌活动，需要运用条件概率和行为轮廓知识来进一步确定其在业务流程中的正确位置。

本文第 2 节给出了一个动机例子；第 3 节介绍了相关基础知识；第 4 节对日志自动机下基于熵的混沌活动挖掘方法进了行分析；第 5 节为部分案例分析；最后总结全文并展望未来。

2　动机例子

急诊是医院的急救窗口，承载着医院各类急、危、重症的首诊和首接抢救任务，还对患者入院时的疾病分类和分流去向起着关键作用。表 1 列出了急诊业务流程中各个字母代表的事件。医院急诊的流程主要包括以下步骤：患者来院就诊、医生病情评估、普通病人专科就诊、危重病人开通绿色通道、检查并确定病情级别、医生抢救、住院治疗等。在整个急诊业务流程中，会不断有 120 急救电话打入、通知家属缴费、通知家属签字等事件发生，而且这些事件在整个急诊业务流程中不受任何约束，可以在任何时间发生。模糊事件日志的直接后继关系使得业务流程变得复杂，因此过滤医院急诊系统中的混沌活动，得到更简练、合适度高的急诊业务流程模型是有意义的。急诊事件日志如表 2 所列。

表 1　急诊业务流程中各个字母代表的事件

Table 1　Activity event represented by each letter in emergency business process

活动	活动名称	活动	活动名称
A	患者来院就诊	L	生命特征不稳定
B	普通病人	M	生命特征稳定
C	预检分科挂号	N	药物治疗
D	专科就诊	O	病情恶化
E	危重病人	P	病情减轻
F	开通绿色通道	Q	医生抢救
G	医生：询问病史、病情评估	R	回家休养
H	护士：吸氧、建立静脉通道	S	住院治疗
I	检查并确定病情级别	T	通知家属
J	病情：重度	U	转院
K	病情：轻微	X	120 急救电话

表 2　急诊事件日志

Table 2　Emergency event logs

案例	事件日志	实例数
1	XAETFGHIJLQS	2985
2	AEFTGHIJXLQTU	856
3	AEFXHGITJMNOQS	1023
4	AEFHXGIJMTNPR	523
5	AEXFGHIJMNOQTU	1080
6	ABXCDTIJLQS	1880
7	ABCDIJMNOTQS	1985
8	ABCTDXIJMNOQTU	327
9	ABCDIKTNXPR	1734
10	ABTCDIKNOQXS	847
11	ABCDIKXNOQTU	1090
12	AEFGHIXKNPTR	1350
13	ATEFGXHIKNPR	420
14	AEFHGIKNOQTU	1100
15	AEFHGIKNXOTQS	2530
16	AEQSTX	120
17	AEQXTU	150

根据表 2 所列日志信息直接构建业务流程模型 M_0，如图 1 所示。可以看出，混沌活动的存在，使得事件日志中的行为被过度泛化，破坏了活动之间的直接后继关系；此外，模型 M_0 中存在许多沉默变迁，使得日志中有些活动被直接跳过而未执行，导致急诊业务流程模型 M_0 的合适度 $fitnes_{M_0}=0.23\ll 1$。针对混沌活动使得业务流程变得繁冗、可解释性低的问题，本文提出了基于日志自动机的业务流程混沌活动的过滤方法。

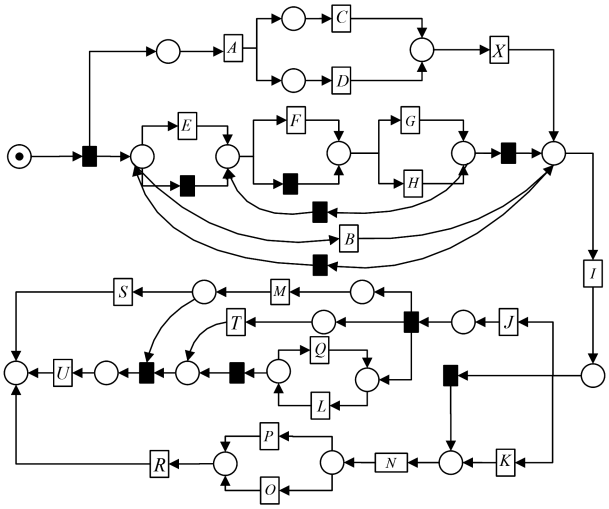

图 1 带混沌活动的急诊业务流程模型 M_0

Fig. 1 Emergency business process model with chaotic activities M_0

3 基础知识

本文主要以日志自动机、熵的概念为基础，提出了有效的过滤日志中混沌活动的方法。下面介绍贯穿全文的主要基本概念。

定义 1[12]（标签 Petri 网） 标签 Petri 网 $N=\langle P,T,F,M,I,O,L\rangle$。

(1) P 是有限的库所集，T 是有限的变迁集，且 $P\cap T=\Phi$；

(2) $F\subseteq(P\times T)\cup(T\times P)$ 是库所和变迁之间弧线的集合，称为流关系；

(3) $M=(m(p_1),m(p_2),m(p_3),\cdots,m(p_n))$ 是网标识；

(4) I 是输入库所集，且 $\cdot I=\{x\in T|(x,I)\in F\}=\varphi$；

(5) O 是输出库所集，且 $O^{\cdot}=\{x\in T|(O,x)\in F\}=\varphi$；

(6) $L:T\to\Sigma$ 是一个标签函数，主要作用是分配一个标签给变迁。

定义 2[12]（日志自动机） 事件日志 L 的日志自动机定义为有向图 $A^*=(\Gamma,\to)$。

(1) 初始状态集：$\uparrow A=\{x\in\Gamma|\nexists\, y\in\Gamma[y\to x]\}$；

(2) 终止状态集：$\downarrow A=\{x\in\Gamma|\nexists\, y\in\Gamma[x\to y]\}$；

(3) 活动发生频数函数：$\#_\Gamma(x)=|z\in\varepsilon|T(z)=x|$；

(4) 直接后继依赖函数：$\#_\to(x,y)=|\{(e_1,e_2)\in\varepsilon\times\varepsilon|T(e_1)=x\wedge T(e_2)=y\wedge e_1\to e_2\}|$。

定义 3[13]（依赖度量） 设 $(\to,||,+)$ 是行为轮廓，给定不同的活动 $a,b\in\Sigma$，定义 $a\Rightarrow b:\Sigma\times\Sigma\to[-1,1]$ 为从 a 到 b 的因果依赖强度。其中：

$$a\Rightarrow b=\begin{cases}\dfrac{|a\to b|-|b\to a|}{|a\to b|+|b\to a|+0.1}, & a\to b\\ \dfrac{|a+b|-|b+a|}{2(|a+b|+|b+a|)+0.1}, & a+b\\ \dfrac{|a\parallel b|+|b\parallel a|}{|a\parallel b|+|b\parallel a|+0.1}, & a\parallel b\end{cases}$$

定义 4[14]（条件发生概率） L 为事件日志，A 为事件日志上的活动集，σ' 为一个子序列，活动 $a\in A$，条件发生概率的定义如下：

$$COP=(a,\sigma',L)=\begin{cases}\dfrac{freq(\sigma'\cdot\langle a\rangle,L)}{freq(\sigma',L)}, & \text{if } freq(\sigma',L)\neq 0\\ 0, & \text{otherwise}\end{cases}$$

4 日志自动机下基于熵的混沌活动挖掘方法分析

混沌活动可以发生在日志中的任意位置，且没有固定的顺序，这增加了过滤混沌活动的难度。现有研究较少涉及混沌活动的过滤方法，且没有考虑混沌活动与其他活动之间的依赖关系。下面基于日志自动机和熵的概念提出了一种有效过滤业务流程混沌活动的方法。

基于熵的混沌活动过滤分析如下。

定义 5[15]（直接前集率、直接后集率） 直接后集率记为 $dfr(a,b,L)$，表示在日志中活动 b 是活动 a 的直接后集所占的比率，即 $dfr(a,b,L)=\dfrac{\#(\langle a,b\rangle,L)}{\#(a,L)}$。其中，$\#(\langle a,b\rangle,L)$ 表示在日志 L 中 b 是 a 的直接后集的数量，$\#(a,L)$ 表示在日志 L 中包含活动 a 的日志数量。同理，直接前集率 $dpr(a,b,L)=\dfrac{\#(\langle b,a\rangle,L)}{\#(a,L)}$。

定义 6[12]（熵） 引入分类概率分布函数 $H(X)=-\sum_{x\in X}x\log_2(X)$，利用直接前集率和直接后集率在日志 L 中定义活动 $a\in Activities(L)$ 的熵值为：

$$H(a.L)=H(dfr(a,L))+H(dpr(a,L))$$

当 $0\in dfr(a,L)\vee 0\in dpr(a,L)$ 时，对应的对数函数值 $0\log_2(0)=0$，其与极限值 $\lim_{p\in 0^+}p\log_2(p)=0$ 一致。

基于熵的可疑混沌活动发现方法如算法 1 所示。

算法 1 基于熵的可疑混沌活动发现方法

输入：事件日志 L_0

输出：可疑混沌活动集 SC′

步骤 1 从事件日志中舍去不完善轨迹，得到可执行事件轨迹 L_i。

步骤 2 根据定义 6，由直接前集率公式计算每个活动的直接前集率：

$DPR=\{dpr(a_1,L),dpr(a_2,L),\cdots,dpr(a_k,L)\}$，$a_k\in L$，$k\in N^+$

根据直接后集率公式计算每个活动的直接后集率：

$DFR=\{dfr(a_1,L),dfr(a_2,L)\cdots dfr(a_k,L)\}$，$a_k\in L$，$k\in N^+$

根据定义 7，计算每个活动对应的熵值：

$H(a_k,L)=H(dfr(a_k,L))+H(dpr(a_k,L))$，$a_k\in L$，$k\in N^+$

步骤 3 根据步骤 2 的计算结果，构建每个活动的直接前集率 DPR、直接后集率 DFR、熵值 H 表。

步骤 4 根据熵值表，将最大熵值对应的活动赋值给集合 SC′，$SC'=\underset{a\in acts}{\operatorname{argmax}}H(a,L_i)$。

步骤 5　在日志 L_i 中删除熵值最大的活动，得到新的日志 $L_{i+1}=L_i\uparrow_{acts\setminus sc'}$。

步骤 6　判断日志 L_{i+1} 中活动的个数。

若 $|Activities(L_{i+1})|>2$，将新日志 L_{i+1} 代入步骤 2。

若 $|Activities(L_{i+1})|\leqslant 2$，由于从活动个数少于 2 的事件日志中不能发现任何活动关系，因此转入步骤 7。

步骤 7　输出可疑混沌活动集 SC'。

利用上述基于熵的活动过滤方法得到可疑混沌活动集 SC' 以后，需要进一步研究确定该活动是否为混沌活动。若为混沌活动，如何在日志中过滤混沌活动以提高业务流程的合适度？为了解决上述问题，本文提出算法 2 来对混沌活动进行进一步的研究。

算法 2　日志自动机下精确定位并过滤混沌活动的方法

输入：日志 L，弧频率阈值 θ_1，条件发生概率阈值 θ_2，依赖度阈值 θ_3

输出：过滤混沌活动后的业务流程事件日志 L_T

步骤 1　根据日志自动机的定义，将源日志 L 转化为日志自动机 $A^*=(\Gamma,\rightarrow)$。

步骤 2　根据构建的日志自动机，运用公式 $c(x,y)=\dfrac{2\times\#_{\rightarrow}(x,y)}{\#_{\Gamma}(x)+\#_{\Gamma}(y)}$ 计算每条弧的频率，将其记为 $C_f=\{c_1,c_2,\cdots,c_n\}$，$n\in N^+$。

步骤 3　将所得到的弧的频率 $C_f=\{c_1,c_2,\cdots,c_n\}$ 与弧频率阈值 θ_1 做比较，若 $c_i<\theta_1$，$i\in[1,n]$，则 c_i 为不频繁弧，记为 $C_{uf}=\{c_1,c_2,\cdots,c_k\}$，$k\in N^+$。在日志自动机中用红线将不频繁弧标注出来。

步骤 4　计算包含活动 x 的所有弧中不频繁弧的占比率 $p_x=\dfrac{\#_{c_{uf}}(x)}{\#_{c_f}(x)}$，$\#_{C_{uf}}(x)$ 表示在不频繁弧 C_{uf} 中活动 x 出现的频数，$\#_{C_f}(x)$ 表示在所有的弧 C_f 中活动 x 出现的频数。若 $p_x=\dfrac{\#_{c_{uf}}(x)}{\#_{c_f}(x)}\approx 1$，则将活动 x 放入集合 SC''。

步骤 5　将算法 1 输出的可疑混沌活动集 SC' 与步骤 4 所得活动集 SC'' 做比较。若 $SC'\cap SC''=\Phi$，则源日志 L 无混沌活动，直接输出事件日志 L，算法结束；若 $SC'\cap SC''=CA\neq\Phi$，则源日志 L 存在混沌活动，转入步骤 6。

步骤 6　在源日志 L 中删除活动集 CA 中的活动，得到新日志 $L'=L\uparrow_{act\setminus\{x\in AC\}}$。

步骤 7　根据日志 L'，运用 α 算法得出业务流程模型 M_1，计算模型的合适度 fitness。若 fitness=1，则所得到的业务流程模型 M_1 即为过滤混沌活动后的最优模型，输出过滤混沌活动后的事件日志 L'，算法结束。若 fitness<1，则所得到的业务流程模型 M_1 不是最优模型，$\exists y\in CA$ 的混沌活动不应该在源日志中完全删除，需要进一步确定该活动与其他活动的关系，具体定位该活动在业务流程模型中的位置。

步骤 8　在事件日志 L 中，对所有 $x\in CA$ 的活动构建行为轮廓关系表，得出活动 $x\in CA$ 与日志 L 中其他活动之间的关系。

步骤 9　运用条件发生概率

$$COP=(a,\sigma',L)=\begin{cases}\dfrac{freq(\sigma'\cdot\langle a\rangle,L)}{freq(\sigma',L)}, & \text{if } freq(\sigma',L)\neq 0\\ 0, & \text{otherwise}\end{cases}$$

计算在日志中其他活动 $b\notin CA$ 发生的条件下混沌活动 $a\in CA$ 作为直接前集或直接后集发生的条件概率。找出条件发生概率 $COP>\theta_2$ 的活动对，转入步骤 11 进一步确定活动对之间的具体关系。

步骤 10　运用定义 3 的依赖度量公式计算出活动 $x\in CA$ 与其他活动是严格、排他、交错顺序的度量值。若 $val(a\rightarrow b)=\dfrac{|a\rightarrow b|-|b\rightarrow a|}{|a\rightarrow b|+|b\rightarrow a|+0.1}>\theta_3$，其中 $a\in CA$，$b\notin CA$，则在流程模型中 a 和 b 是严格序关系，在源日志 L 中保留活动 a 是活动 b 的直接前集位置关系，将其他位置活动 a 过滤掉，得到过滤后的日志 L''。若 $val(a+b)=\dfrac{|a+b|-|b+a|}{2(|a+b|+|b+a|)+0.1}>\theta_3$，其中 $a\in CA$，$b\notin CA$，则在流程模型中 a 和 b 是排他序关系，在源日志中直接过滤活动 $a\in CA$，得到新的事件日志 L''。若 $val(a\parallel b)=\dfrac{|a\parallel b|+|b\parallel a|}{|a\parallel b|+|b\parallel a|+0.1}>\theta_3$，其中 $a\in CA$，$b\notin CA$，则在流程模型中 a 和 b 是交叉序关系，在源日志 L 中保留活动 a 在活动 b 直接前集的位置和活动 a 在活动 b 的直接后集的位置关系，将其他位置活动 a 过滤掉，得到过滤后的日志 L''。

步骤 11　输出过滤混沌活动后的事件日志 L''。

5　案例分析

本节以动机例子中的急诊业务流程为例来验证算法的有效性。给定合理性判定阈值 $fitness=1$，不频繁弧的阈值 $\theta_1=0.1$，条件发生概率阈值 $\theta_3=0.9$，依赖度量阈值 $\theta_2=0.9$。

合适度表示事件日志中的活动是否能够在业务流程模型中执行，当合适度为 1 时，说明事件日志中的活动能够完全在业务流程模型中重演。通过多次实验保证日志自动机的连通性，取不频繁弧阈值 $\theta_1=0.1$ 能够有效防止频繁弧被删除而过滤了主要活动。条件发生概率和依赖度量用于判断两个活动之间的紧密度关系，选取 $\theta_3=0.9$ 和 $\theta_2=0.9$ 较为合理。

动机案例图 1 中已经给出了带混沌活动的急诊业务流程模型 M_0，但合适度 $fitnes_{M_0}=0.23\ll 1$。下面运用本文所提出的算法来定位并过滤混沌活动，以挖掘合理性高的急诊业务流程模型。

首先，根据表 2 的源日志 L 构建日志自动机，如图 2 所示。根据弧的频率公式 $c(x,y)=\dfrac{2\times\#_{\rightarrow}(x,y)}{\#_{\Gamma}(x)+\#_{\Gamma}(y)}$ 计算日志自动机中 77 条弧的频率，其中 $c_i<0.1$ 的不频繁弧有 34 条，在日志自动机中用加粗黑箭头标注，其频率如表 3 所列。

根据公式 $p_Z=\dfrac{\#_{c_{uf}}(Z)}{\#_{c_f}(Z)}\approx 1$，得到活动 $p_T=\dfrac{16}{17}=0.94\approx 1$，$p_X=\dfrac{16}{18}=0.96\approx 1$，则活动集 $SC''=\{T,X\}$。由算法 1 可以计算日志中活动的直接前集率、直接后集率、熵值。具体计算结果如表 4 所列，可得可疑混沌活动集 $SC'=\{T,X,R,G,Q\}$，则 $SC'\cap SC''=\{T,X\}$，即在源日志 L 中活动 T 和 X 为可疑混沌活动。

在源日志中过滤活动 T 和 X，得到新的日志 L'。根据新日志 L'，运用 α 算法得出业务流程模型 M_1，如图 3 所示。计算可得 $fitnes_{M_1}=0.84<1$，过滤混沌活动 T 和 X 后，虽然合适度有了明显提高，但其仍然小于 1。由于沉默变迁的存在不能在模型 M_1 中很好地重演，即日志 $\langle A,E,F,G,H,I,J,M,N,O,Q,U\rangle^{10803}$，$\langle A,B,C,D,I,K,N,O,Q,U\rangle^{1090}$，$\langle A,E,Q,U\rangle^{150}$ 不能在业务流程模型 M_1 中执行，因此需要进一步研究混沌活动与其他活动之间的行为轮廓关系。

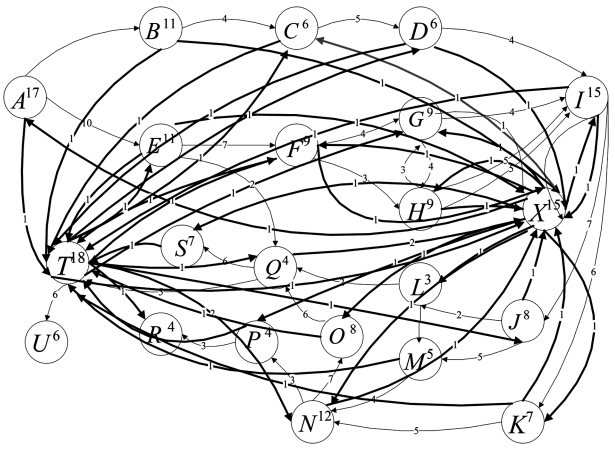

图 2 急诊日志自动机

Fig. 2 Emergency log automat on

表 3 急诊业务流程非频繁弧的频率列表

Table 3 Frequency list of infrequent arcs in emergency business

$C(X,A)=0.063$	$C(M,T)=0.087$	$C(K,T)=0.08$	$C(T,R)=0.09$
$C(E,T)=0.069$	$C(E,X)=0.077$	$C(T,N)=0.067$	$C(A,T)=0.057$
$C(T,F)=0.074$	$C(X,F)=0.083$	$C(B,T)=0.083$	$C(T,E)=0.069$
$C(T,G)=0.074$	$C(B,X)=0.095$	$C(T,C)=0.083$	$C(G,X)=0.083$
$C(J,X)=0.087$	$C(X,C)=0.095$	$C(X,S)=0.091$	$C(P,T)=0.09$
$C(F,X)=0.083$	$C(D,T)=0.083$	$C(K,X)=0.091$	$C(X,I)=0.067$
$C(I,T)=0.061$	$C(T,I)=0.061$	$C(X,N)=0.074$	$C(X,G)=0.083$
$C(T,J)=0.077$	$C(C,T)=0.095$	$C(I,X)=0.067$	
$C(H,X)=0.083$	$C(D,X)=0.095$	$C(X,K)=0.09$	

表 4 熵值表

Table 4 Entropy value table

迭代	直接前集率 DPR	直接后集率 DFR	熵值 H
1	$dpr(T,L)=4.91$	$dfr(T,L)=3.54$	$H(T)=8.45$
2	$dpr(X,L)=3.25$	$dfr(X,L)=3.49$	$H(X)=6.73$
3	$dpr(I,L)=1.97$	$dfr(I,L)=1.52$	$H(I)=3.49$
4	$dpr(G,L)=1.51$	$dfr(G,L)=1.49$	$H(G)=3.00$
5	$dpr(Q,L)=1.67$	$dfr(Q,L)=1.16$	$H(F)=2.83$

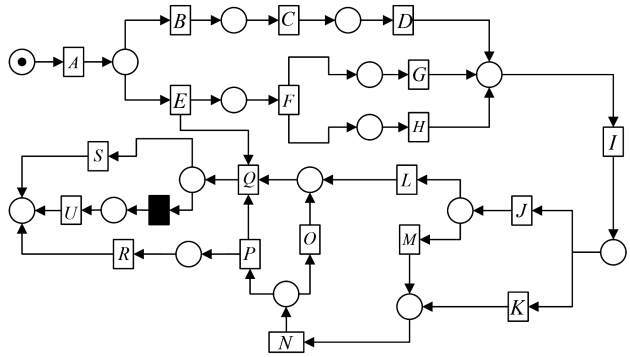

图 3 急诊业务流程模型 M_1

Fig. 3 Emergency business process models M_1

为了定位混沌活动在业务流程中的具体位置，需要计算当其他活动发生时活动 T 和 X 作为直接前集或直接后集发生的条件概率。从表 5 所列的条件发生概率中可以明显看出 $P(T|U)=1$，即活动 U 发生的条件下活动 T 一定发生。需要通过行为轮廓确定 U 和 T 之间的具体关系，其他活动之间的条件发生概率很低，不予考虑。

表 5 急诊业务流程中事件条件发生概率表

Table 5 Probability table of conditional occurrence in emergency business

$P(T\|A)=0.02$	$P(T\|E)=0.28$	$P(T\|I)=0.14$	$P(T\|M)=0.1$	$P(T\|M)=0.1$
$P(T\|B)=0.11$	$P(T\|F)=0.32$	$P(T\|J)=0.09$	$P(T\|N)=0.13$	$P(T\|N)=0.13$
$P(T\|C)=0.15$	$P(T\|G)=0.07$	$P(T\|K)=0.19$	$P(T\|O)=0.45$	$P(T\|O)=0.45$
$P(T\|D)=0.28$	$P(T\|H)=0$	$P(T\|L)=0$	$P(T\|P)=0.54$	$P(T\|P)=0.54$
$P(X\|E)=0.09$	$P(X\|I)=0.09$	$P(X\|M)=0$	$P(X\|Q)=0.06$	$P(X\|A)=0.15$
$P(X\|F)=0.18$	$P(X\|J)=0.08$	$P(X\|N)=0.31$	$P(X\|R)=0$	$P(T\|B)=0$
$P(X\|G)=0.08$	$P(X\|K)=0.27$	$P(X\|O)=0.25$	$P(X\|S)=0.07$	$P(X\|C)=0.24$
$P(X\|H)=0.17$	$P(X\|L)=0.15$	$P(X\|P)=0.43$	$P(X\|U)=0$	$P(X\|D)=0.04$
$P(T\|Q)=0.28$	$P(T\|R)=0.33$	$P(T\|S)=0.01$	$P(T\|U)=1$	

构建混沌活动 T 和 X 与其他活动之间的行为轮廓，如表 6 所列。表 6 中，1 代表严格序→，2 代表交叉序||，3 代表严格逆序$\rightarrow^{-1}$。$T\Rightarrow U=\frac{856+1080+327+1090+1100+150}{856+1080+327+1090+1100+150+0.1}=0.99>0.9$，即 T 和 U 是严格序关系，$T\rightarrow U$；且 S 和 U 是排他关系，$S+U$ 表示活动 S 和 T 是排他关系。至此，便确定了活动 T 在事件日志中的正确位置是在活动 U 的直接前集，而在其他位置上活动 T 应该作为异常值被直接过滤，从而得到新的事件日志，过滤混沌活动后的事件日志如表 7 所列。根据过滤后的事件日志构建业务流程模型 M_T，如图 4 所示。由计算可得，急诊业务流程模型 M_T 的合适度 $fitnes_{M_T}=1$。

表 6 行为轮廓关系表

Table 6 Probability table of conditional occurrence behavior profile

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	U
T	3	3	2	3	2	2	1		2	1			3	1	3	3	1	1	3	1
X	1	3	1	3	3	2	2	1	2	3	2	1		2	1	1	3		1	

表 7 过滤混沌活动后的急诊事件日志

Table 7 Emergency event log after filtering chaotic activity

案例	事件日志	实例数
1	AEFGHIJLQS	2985
2	AEFGHIJXLQTU	856
3	AEFHGIJMNOQS	1023
4	AEFHGIJMNPR	523
5	AEFGHIJMNOQTU	1080
6	ABCDIJLQS	1880
7	ABCDIJMNOTQS	1985
8	ABCDIJMNOQTU	327
9	ABCDIKNPR	1734
10	ABCDIKNOQS	847
11	ABCDIKNOQTU	1090
12	AEFGHIKNPR	1350
13	AEFGHIKNPR	420
14	AEFHGIKNOQTU	1100
15	AEFHGIKNOQS	2530
16	AEQS	120
17	AEQTU	150

实例分析表明，本文提出的基于日志自动机的业务流程混沌活动过滤方法显著提高了业务流程模型的合适度，合适度如图 5 所示。

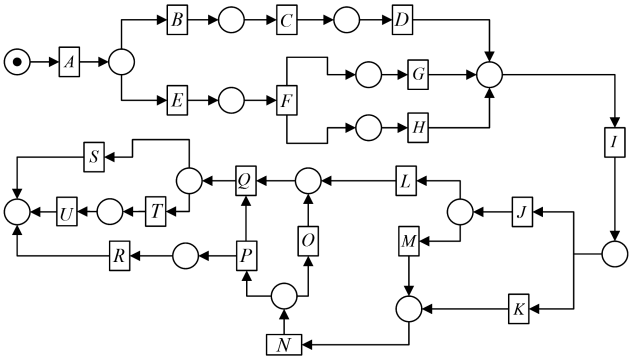

图 4　急诊业务流程模型 M_T

Fig. 4　Emergency business process model M_T

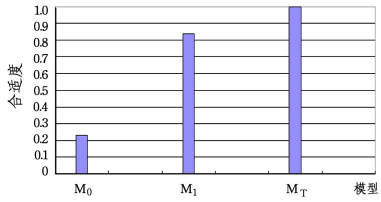

图 5　模型合适度

Fig. 5　Model suitability

结束语　本文在已有研究成果的基础上，提出了过滤事件日志中混沌活动的新方法，并提出了两种算法，分别为基于日志自动机和熵发现可疑混沌活动的方法和条件发生概率下精确定位并过滤混沌活动的方法。该算法不仅能够检验业务流程模型中是否存在混沌活动，以及存在几个混沌活动；还能够识别该混沌活动是使模型变得繁冗且解释性低的混沌活动，还是该混沌活动原本是业务流程的一部分，但由于发生故障或其他原因导致该活动在日志中无序、随机地发生，并且能够确定该混沌活动原本的正确位置。实例分析证明，基于自动机的业务流程过滤混沌活动方法能够有效过滤日志中的混沌活动，大大提高业务流程的合适度。本文提出的混沌活动过滤方法适用于频繁发生的混沌活动，不频繁发生的混沌活动过滤技术将是未来的主要研究方向。

参 考 文 献

[1] WIL V D A. Process Mining: Data Science in Action [M]. Springer Publishing Company, Incorporated, 2016.

[2] LEEMANS S J J, FAHLAND D, AALST W M P V D. Scalable process discovery and conformance checking[J]. Software & Systems Modeling, 2018, 17(2): 599-631.

[3] CHABROL M, DALMAS B, NORRE S, et al. A process tree-based algorithm for the detection of implicit dependencies[C]// IEEE Tenth International Conference on Research Challenges in Information Science. IEEE, 2016: 1-11.

[4] SANI M F, ZELST S J V, AALST W M P V D. Repairing Outlier Behaviour in Event Logs [C]// International Conference on Business Information Systems. Cham: Springer, 2018.

[5] HUANG Y, WANG Y, HUANG Y. Filtering Out Infrequent Events by Expectation from Business Process Event Logs[C]// 2018 14th International Conference on Computational Intelligence and Security (CIS). IEEE Computer Society, 2018.

[6] LIESAPUTRA V, YONGCHAREON S, CHAISIRI S. Efficient Process Model Discovery Using Maximal Pattern Mining[C]// International Conference on Business Process Management. Cham, 2015: 441-456.

[7] LU X, FAHLAND D, BIGGELAAR, et al. Detecting Deviating Behaviors Without Models[C]// International Conference on Business Process Management. Cham: 2015: 126-139.

[8] ROJAS E, MUNOZ-GAMA J, SEPÚLVEDA M, et al. Process mining in healthcare: A literature review[J]. Journal of Biomedical Informatics, 2016, 61: 224-236.

[9] PULSANONG W, POROUHAN P, TUMSWADI S, et al. Using inductive miner to find the most optimized path of workflow process[C]// International Conference on ICT and Knowledge Engineering. IEEE, 2017: 1-5.

[10] BURATTIN A. Heuristics Miner for Time Interval[C]// Esann 2010, European Symposium on Artificial Neural Networks. Bruges, Belgium: DBLP, 2015: 85-95.

[11] LINGALA N, SRI NAMACHCHIVAYA N, PERKOWSKI N, et al. Particle filtering in high-dimensional chaotic systems[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2012, 22(4): 047509.

[12] CONFORTI R, ROSA M L, HOFSTEDE A H M T. Filtering Out Infrequent Behavior from Business Process Event Logs[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(2): 300-314.

[13] MANNHARDT F, DE LEONI M, REIJERS H A, et al. Data-Driven Process Discovery-Revealing Conditional Infrequent Behavior from Event Logs[C]// International Conference on Advanced Information Systems Engineering. Cham: Springer, 2017: 545-560.

[14] SANI M F, ZELST S J V, AALST W M P V D. Improving Process Discovery Results by Filtering Outliers Using Conditional Behavioural Probabilities [C]// International Conference on Business Process Management. Cham: Springer, 2017: 216-229.

[15] TAX N, SIDOROVA N, AALST W M P V D. Discovering more precise process models from event logs by filtering out chaotic activities[J]. Journal of Intelligent Information Systems, 2019, 52(1): 107-139.

LI Juan, born in 1992, postgraduate. Her main research interests include Pet net and Business process management.

FANG Xian-wen, born in 1975, Ph. D, professor, Ph. D supervisor, is member of China Computer Federation (CCF). His main research interests include Petri net and trusted software.