基于 RankClus 算法的机场流程日志活动挖掘

徐 涛[①②] 孟 野*[①] 卢 敏[①②]
[①](中国民航大学计算机科学与技术学院 天津 300300)
[②](中国民航信息技术科研基地 天津 300300)

摘 要：流程挖掘技术可以提取机场流程日志中的有用信息用于流程分析。但机场流程日志处于细节化的低抽象层次，不符合分析者的预期。对机场流程日志挖掘得到的流程模型呈现意面状的复杂结构，流程模型的含义难于理解。解决该问题的一种方法是通过活动挖掘，将低抽象层次活动聚类为流程模型中表征高抽象层次活动的活动类簇。为此提出了一种基于 RankClus 算法的活动挖掘方法，将机场流程日志的活动聚类与活动排序评分计算相结合，从而构建更易理解的活动聚类流程模型。实验结果表明，RankClus 活动聚类流程模型的日志回放一致性与原生日志流程模型大致相当，但在结构复杂度上要显著低于原生日志流程模型。

关键词：流程挖掘；活动挖掘；RankClus；踪迹聚类
中图分类号： TP391 **文献标识码**： A **文章编号**：1009-5896(2016)08-2033-07
DOI: 10.11999/JEIT151137

Activity Mining for Airport Event Logs Based on RankClus Algorithm

XU Tao[①②] MENG Ye[①] LU Min[①②]
[①](*College of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*)
[②](*Information Technology Research Base of Civil Aviation Administration of China, Tianjin* 300300, *China*)

Abstract: Process mining is a technology which can extract non-trivial and useful information from airport event logs. However, the airport event logs are always on a detailed level of abstraction, which may not be in line with the expected abstract level of an analyst. Process models generated by these event logs are always spaghetti-like and too hard to comprehend. An approach to overcome this issue is to group low-level events into clusters, which represent the execution of a higher-level activity in the process model. Therefore, this paper presents a new activity mining method which is based on RankClus algorithm to generate activity clusters integrated with ranking. On this basis, the activity-clustered model which is easier to comprehend can be constructed. The experiment results show that this activity-clustered model, which shares a similar level of conformance with the meta model, is significantly less complex.

Key words: Process mining; Activity mining; RankClus; Trace clustering

1 引言

机场运行过程中时刻有各类事件发生，机场业务信息系统随之生成一系列机场流程日志。对机场流程日志进行流程挖掘可得到相应的业务流程模型[1]，借由业务流程模型指导，机场可开展运行决策支持[2]及业务趋势预测[3]等一系列工作以提高机场运行效率。因此机场流程日志的流程挖掘具有重要意义。

流程挖掘研究通常将流程日志罗列为活动组成的踪迹(trace)，构建目标日志流程模型并分析。流程挖掘研究主要分 3 个方向[1]：(1)流程发现：在无先验知识指导下建立流程日志的流程模型；(2)一致性检测：对比已有流程模型与真实情况以验证模型合理性，常用日志回放实现；(3)模型增强：根据所观测事件信息扩展业务流程模型。国内机场流程日志中活动以工作人员上传的文本描述为主，抽象层次(abstract level)较低。直接对这类日志进行流程发现时，得到的流程模型结构复杂且难以理解。文献[4]提出一种基于全局踪迹分割的活动挖掘方法，该

收稿日期：2015-10-10；改回日期：2016-04-15；网络出版：2016-06-03
*通信作者：孟野 mykonakona@foxmail.com
基金项目：国家自然科学基金(61502499)，中国民航科技创新引导资金项目重大专项(MHRD20140105)，中央高校科研业务费专项资金(3122013C005, 3122014D032, 3122015D015)，中国民航大学科研基金(2013QD18X)，中国民航信息技术科研基地开放课题基金(CAAC-ITRB-201401)
Foundation Items: The National Natural Science Foundation of China (61502499), The Civil Aviation Key Technologies R&D Program of China (MHRD20140105), The Fundamental Research Funds for the Central Universities of China (3122013C005, 3122014D032, 3122015D015), The Scientific Research Foundation from Civil Aviation University of China (2013QD18X), The Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (CAAC-ITRB-201401)

方法设定时间窗口对邻近活动进行层次聚类。但仅考虑事件间的时间邻近度，其聚类结果不能很好反映领域知识。文献[5]采用领域专家手工标记方式为流程日志添加活动类标签，并用标记日志训练活动描述的文本分类器，再对活动分类。该方式所得活动类别较为细碎，专家标记的主观误差对结果影响较大。文献[6]假设事件与活动间存在一对多或多对多关系，采用词干提取等文本挖掘技术挖掘日志文本描述中的领域知识，将所得知识用于匹配事件与活动以合理定位流程日志抽象层次。该方法在中文流程日志中实现较困难。

本文构建二类型网络(bi-type network)描述活动与踪迹关系，视活动与踪迹为不同类型结点，用活动在各踪迹结点类簇的排序评分向量量化表示活动，为此需得到合理踪迹聚类结果以计算该排序评分。文献[7–9]的踪迹聚类方法难用于踪迹聚类的活动排序评分计算，不能很好衔接后续活动聚类工作。文献[10]提出有效结合聚类和排序的 RankClus 算法。该算法主要功能是对二类型网络排序与聚类。应用在机场流程日志活动挖掘能够得到较准确的踪迹结点划分结果，并计算出活动结点在踪迹划分生成子网络的排序评分。在 RankClus 算法基础上，本文将踪迹聚类与活动聚类相结合，设计机场流程日志低抽象层次活动的聚类算法，使基于聚类结果挖掘所得日志流程模型在保持一定日志重现度的同时，有效降低流程模型的结构复杂度。

2 机场流程日志活动挖掘分析

机场流程日志的流程挖掘主要关注提交时间、部门、模块、活动、实例号等属性。表 1 是国内某机场的部分流程日志，类似“新增要客航班：HU7703, CA1321。”，“要客航班更新：CA947，请各单位加强关注。”活动描述的事件大量存在，这类事件可统一视为“要客航班更新”。但流程挖掘时低抽象层次事件与活动间一对一映射的关系[11]及复杂的活动描述语义，使数据预处理合并事件的做法难以实现，挖掘到的流程模型充斥大量低抽象层次活动。为此需将低抽象层次事件通过聚类方式抽象为高抽象层次的活动类簇。将“新增”、“更新”等活动描述标识的事件聚类为表示“要客航班更新”的活动类簇以简化流程模型结构。

可将表 1 中 390962 号实例与 390963 号实例分别表示为踪迹<A,C,D,E,F>与踪迹<B,C,D,E,F>。若将这两条业务响应[12]类似的踪迹聚为一类，形如“新增要客航班”、“要客航班更新”的活动便出现于同类踪迹中。活动即可表示为在不同类踪迹中的分布情况。表 2 的日志结构分析表明机场流程日志活动有较高的绝对数目与事件记录占比，大量低抽象层次活动使流程模型结构呈“意面状”(Spaghetti-like)[4]。以模块或其他属性构建踪迹可简化所发现流程模型的结构，但造成模型抽象层次过高，仅能反映“当前部门开展了某项活动”这类不具体的活动语义，模型丢失大量信息。因此聚类时需为活动指定介于两者间的抽象层次。

可构造如图 1 所示的二类型网络描述活动与踪迹间关系，并区分网络中活动在各类踪迹中重要度以聚类相似的低抽象层次活动。用高抽象层次活动类簇替代原日志活动，构建踪迹集合。该网络由活动结点与踪迹结点组成，网络的实线视为该活动在踪迹中出现了一次，虚线则表示结点间存在相似性。采用二类型网络来描述活动与踪迹间的关系，使得流程日志活动挖掘问题转变为聚类二类型网络活动结点的问题[13,14]。

表 1 国内某大型枢纽机场部分流程日志

提交时间	部门	模块	活动	实例号
08-01-13 22:50:26	TAMCC	指挥协调	新增要客航班：HU7703、CA1321。	390962
08-01-13 22:52:56	安保公司	安保指挥中心	安保公司收到。	390962
08-01-13 22:53:27	安保公司	飞行区安检部	安保飞行区安检部收到。	390962
08-01-13 22:53:50	安保公司	综合安检部	安保综合安检部收到。	390962
08-01-13 22:55:51	TAMCC	指挥协调	要客航班更新：CA947，请各单位加强关注。	390963
08-01-13 22:57:54	安保公司	安保指挥中心	安保公司收到。	390963
08-01-13 22:58:32	安保公司	飞行区安检部	安保飞行区安检部收到。	390963
08-01-13 22:58:38	安保公司	综合安检部	安保综合安检部收到。	390963
08-01-13 23:03:15	安保公司	东区安检部	安保东区安检部收到。	390962
09-01-13 05:35:09	安保公司	东区安检部	安保东区安检部收到。	390963
⋮	⋮	⋮	⋮	⋮

表 2 国内某大型枢纽机场 2013 年流程日志结构分析

时间区间	事件记录总数	踪迹数目	活动数目	活动占记录比例(%)
7 月 1 日至 7 月 15 日	748	222	335	45
8 月 1 日至 8 月 15 日	1528	360	640	42
10 月 1 日至 10 月 15 日	2174	469	636	29

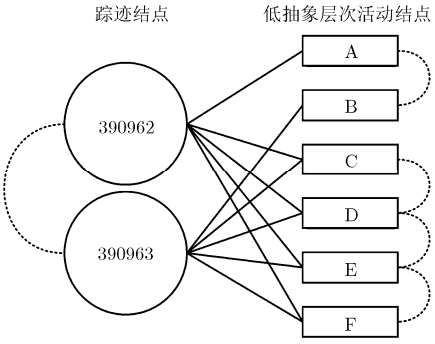


图 1 活动与踪迹的二类型网络

3 基于 RankClus 的机场流程日志活动挖掘算法

3.1 RankClus 混合模型

为聚类活动与踪迹的二类型网络中的活动结点，需划分踪迹结点，将活动结点表示为在各类踪迹上的重要度排序评分的评分向量。为获取踪迹结点的合理划分结果，可引入 RankClus 算法的混合模型(mixture model)，通过模型参数估计得到的踪迹结点表示向量，对踪迹结点进行划分。

以机场流程日志活动-踪迹二类型网络为例，X 表示机场日志踪迹结点集合，Y 表示机场低抽象层次活动结点集合，则可表示机场日志踪迹结点与机场低抽象层次活动构成的二类型网络，$\boldsymbol{W}$ 为网络的邻接矩阵，分块可得：

$$\mathrm{G}=\left\langle\{\mathrm{X}\cup\mathrm{Y}\},\boldsymbol{W}\right\rangle,\boldsymbol{W}=\begin{pmatrix}\boldsymbol{W}_{\mathrm{XX}} & \boldsymbol{W}_{\mathrm{XY}}\\ \boldsymbol{W}_{\mathrm{YX}} & \boldsymbol{W}_{\mathrm{YY}}\end{pmatrix} \tag{1}$$

其中 $\boldsymbol{W}_{\mathrm{XY}}(i,j)$ 表示包含活动 j 的踪迹 i 的个数，$i=1,2,\cdots,m;j=1,2,\cdots,n$，$\boldsymbol{W}_{\mathrm{YX}}=\boldsymbol{W}_{\mathrm{XY}}^{\mathrm{T}}$。活动与活动间、踪迹与踪迹间的关系暂不考虑，即令 $\boldsymbol{W}_{\mathrm{XX}}=0$, $\boldsymbol{W}_{\mathrm{YY}}=0$。假设网络中集合 X 的聚类结果已知，为 $\mathrm{X}_1,\mathrm{X}_2,\cdots,\mathrm{X}_K$。给定某一排序函数，可得类 X_k 的条件排序 $\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}_k}$ 与 $\boldsymbol{r}_{\mathrm{X}|\mathrm{X}_k}$，$(k=1,2,\cdots,K)$。记 $p_k(\mathrm{Y})=\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}_k}$，$p_k(\mathrm{X})=\boldsymbol{r}_{\mathrm{X}|\mathrm{X}_k}$。

将 x_i 与 Y 中结点有边相连的概率表示为 $p_{x_i}(\mathrm{Y})=p(\mathrm{Y}\mid x_i)$，则 X 中的所有结点 x_i $(i=1,2,\cdots,m)$ 均满足这一分布。记 $\pi_{i,k}$ 为 x_i 属于第 k 类的后验概率，可对 $p(\mathrm{Y}\mid x_i)$ 建立如式(2)的 RankClus 混合模型：

$$p_{x_i}(\mathrm{Y})=\sum_{k=1}^{K}\pi_{i,k}p_k(\mathrm{Y}),\quad\sum_{k=1}^{K}\pi_{i,k}=1 \tag{2}$$

这里，$\pi_{i,k}=p(k\mid x_i)$，因为 $p(k\mid x_i)=p(x_i\mid k)p(k)$，而 x_i 在第 k 类中的条件排序值 $p(x_i\mid k)$ 已知，则需要对先验概率 $p(k)$ 进行估计。

用 EM 算法估计混合模型参数 $\boldsymbol{\Theta}$，$\boldsymbol{\Theta}$ 为 $\pi_{i,k}$ 组成矩阵：$\boldsymbol{\Theta}_{m\times K}=\{\pi_{i,k}\},(i=1,2,\cdots,m;k=1,2,\cdots,K)$。E 步中，引入隐藏变量 $z\in\{1,2,\cdots,k\}$ 标记每条边表示边 $\langle x,y\rangle$ 的类簇归属。对数似然函数可写为

$$\ln L(\boldsymbol{\Theta}\mid\boldsymbol{W}_{\mathrm{XY}},z)=\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(i,j)\ln\left(p_z\left(x_i,y_j\right)p\left(z\mid\boldsymbol{\Theta}\right)\right) \tag{3}$$

其中 $p_z\left(x_i,y_j\right)$ 表示所生成的边 $\left\langle x_i,y_j\right\rangle$ 来自类簇 z 的概率。将 x_i 与 y_j 各自的条件排序值视为其在网络中被访问到的概率，可假设 x 与 y 是相互独立的，即

$$p_z\left(x_i,y_j\right)=p_z\left(x_i\right)p_z\left(y_j\right) \tag{4}$$

设置模型初始参数为 $\boldsymbol{\Theta}=\boldsymbol{\Theta}^0$，条件分布 $p\left(z=k\mid x_i,y_j,\boldsymbol{\Theta}^0\right)$ 满足贝叶斯公式：

$$p\left(z=k\mid x_i,y_j,\boldsymbol{\Theta}^0\right)=p_k^0\left(x_i\right)p_k^0\left(y_j\right)p^0\left(z=k\right) \tag{5}$$

M 步中，为了对 $p(z=k)$ 进行估计，需要最大化 $Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^0)$。引入拉格朗日乘子后计算 $p(z=k)$ 估计值为

$$p(z=k)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(i,j)\,p\left(z=k\mid x_i,y_j,\boldsymbol{\Theta}^0\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(i,j)} \tag{6}$$

$\boldsymbol{\Theta}$ 中的 $\pi_{i,k}$ 可由式(8)求解。重复上述过程更新参数，最终 $\boldsymbol{\Theta}$ 将收敛至一个局部极大值。

$$\pi_{i,k}=p\left(z=k\mid x_i\right)=\frac{p_k\left(x_i\right)p\left(z=k\right)}{\sum_{l=1}^{K}p_l\left(x_i\right)p\left(z=l\right)} \tag{7}$$

3.2 排序评分的计算

给定网络 $\mathrm{G}=\left\langle\{\mathrm{X}\cup\mathrm{Y}\},\boldsymbol{W}\right\rangle$ 后，可计算集合 X 与集合 Y 中各结点的排序得分函数 $f:\mathrm{G}\rightarrow$

$\{\boldsymbol{r}_{\mathrm{X}},\boldsymbol{r}_{\mathrm{Y}}\}$，$\boldsymbol{r}_{\mathrm{X}}$ 与 $\boldsymbol{r}_{\mathrm{Y}}$ 分别表示 X 与 Y 中各结点排序评分，$\mathrm{G}=\langle\{\mathrm{X}\cup\mathrm{Y}\},\boldsymbol{W}\rangle$ 的排序函数采用 SimpleRank 函数[10]，表示为

$$\left.\begin{aligned}\boldsymbol{r}_{\mathrm{X}}(x)&=\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(x,j)\Big/\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(i,j)\\ \boldsymbol{r}_{\mathrm{Y}}(y)&=\sum_{i=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(i,y)\Big/\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(i,j)\end{aligned}\right\}\tag{8}$$

令 $\mathrm{X}'\subseteq\mathrm{X},\mathrm{G}'=\langle\{\mathrm{X}'\cup\mathrm{Y}\},\boldsymbol{W}'\rangle$ 表示顶点集 $\mathrm{X}'\cup\mathrm{Y}$ 诱导所得子网络，$\boldsymbol{W}'$ 为该子网络邻接矩阵。子网络 G' 上的排序函数 $(\boldsymbol{r}_{\mathrm{X}'},\boldsymbol{r}_{\mathrm{Y}})=f(\mathrm{G}')$ 的排序结果可表示为

$$\left.\begin{aligned}\boldsymbol{r}_{\mathrm{X}'}(x)&=\sum_{j=1}^{n}\boldsymbol{W}'_{\mathrm{XY}}(x,j)\Big/\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}'_{\mathrm{XY}}(i,j)\\ \boldsymbol{r}_{\mathrm{Y}}(y)&=\sum_{i=1}^{n}\boldsymbol{W}'_{\mathrm{XY}}(i,y)\Big/\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}'_{\mathrm{XY}}(i,j)\end{aligned}\right\}\tag{9}$$

$\boldsymbol{r}_{\mathrm{X}'}=\boldsymbol{r}_{\mathrm{X}'|\mathrm{X}'}$ 为对 X 聚类时 X' 的类内排序评分，$\boldsymbol{r}_{\mathrm{Y}}=\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}'}$ 为对 X 聚类时 Y 的条件排序评分，分别反映一类相似踪迹中某踪迹出现频繁程度和各活动参与情况。$\boldsymbol{r}_{\mathrm{X}|\mathrm{X}'}$ 为 $\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}'}$ 在网络 G 上所得传递得分，可定义为

$$\boldsymbol{r}_{\mathrm{X}|\mathrm{X}'}(x)=\frac{\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(x,j)\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}'}(j)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{\mathrm{XY}}(i,j)\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}'}(j)}\tag{10}$$

3.3 聚类中心和距离的计算

每个 x_i 对应一 K 维向量 $\mathrm{q}_i=\left[\pi_{i,1}\ \pi_{i,2}\ \cdots\ \pi_{i,K}\right]$，如令 $\varpi_{j,k}=p_k\left(y_j\right)$，则每个 y_j 可对应一 K 维向量 $\boldsymbol{\eta}_j=\left[\varpi_{j,1}\ \varpi_{j,2}\ \cdots\ \varpi_{j,K}\right]$；计算 X 类簇或 Y 类簇中所有结点对应向量的平均值，得到每个类簇的类簇中心：

$$\left.\begin{aligned}\mathrm{l}_k&=\sum_{x_i\in\mathrm{X}_k}\boldsymbol{\theta}_i\Big/\left|\mathrm{X}_k\right|\\ \mathrm{m}_l&=\sum_{y_j\in\mathrm{Y}_l}\boldsymbol{\eta}_j\Big/\left|\mathrm{Y}_l\right|\end{aligned}\right\}\tag{11}$$

$\left|\mathrm{X}_k\right|$，$\left|\mathrm{Y}_l\right|$ 分别为类 k，类 l 所包含的结点数。而结点到类簇的距离可定义为

$$\left.\begin{aligned}D\left(x_i,\mathrm{X}_k\right)&=1-\frac{\sum_{m=1}^{K}\boldsymbol{\theta}_i(m)\boldsymbol{\lambda}_k(m)}{\sqrt{\sum_{m=1}^{K}\left(\boldsymbol{\theta}_i(m)\right)^2}\sqrt{\sum_{m=1}^{K}\left(\boldsymbol{\lambda}_k(m)\right)^2}}\\ D\left(y_j,\mathrm{Y}_l\right)&=1-\frac{\sum_{m=1}^{K}\boldsymbol{\eta}_j(m)\boldsymbol{\mu}_l(m)}{\sqrt{\sum_{m=1}^{K}\left(\boldsymbol{\eta}_j(m)\right)^2}\sqrt{\sum_{m=1}^{K}\left(\boldsymbol{\mu}_l(m)\right)^2}}\end{aligned}\right\}\tag{12}$$

3.4 算法流程

文献[10]为控制聚类数及得到更具意义聚类结果，指定算法聚类结点数较少类型的结点，未提供聚类网络中结点数较多类型结点的相应解决方案，不能直接聚类多于踪迹的活动。表 3 算法流程输出准确的基于踪迹聚类的流程日志活动排序评分后继续迭代计算活动排序评分。这一评分可为活动聚类提供足够信息。

4 实验与分析

对原生日志添加活动聚类标签后，可生成活动聚类流程日志 (activity-clustered event log) 挖掘流程模型。比对各流程模型的日志重现度以验证聚类结果合理性；分析各流程模型的结构复杂度以验证活动聚类日志能在保持回放准确度的同时有效降低模型结构复杂度。本文实验数据集为表 2 中 3 组不同时间区间的流程日志，并选用文献[15]的 Inductive Miner 方法挖掘流程日志的 Petri 网流程模型，噪声参数设置为 0.1。

4.1 机场流程日志活动聚类实验

机场等大型机构数据聚类分析的参数设定多依赖于领域专家知识[16]。结合机场运行专家知识[16,17]及数据源机场实际运行情况归纳得 15 类业务流程及 20 类业务活动，分别作为踪迹结点聚类数与活动结点聚类数。文献[4]总结低抽象层次活动与业务流程间关系为两类：(1)业务流程由被单一活动类簇覆盖的低抽象层次活动组成；(2)业务流程由分散在不同活动类簇中的低抽象层次活动组成。图 2 是算法稳定时踪迹结点各类簇的活动结点评分，图 3 是活动结点聚类结果。数据集 1 结点数最多的类簇 15 主要为重点保障航班保障活动，活动描述以“CZ390 有旅客要下机，需客梯车到现场。”、“MU5714 航班滑回，需客梯车。”等居多。数据集 2 活动结点数最多的类簇 2 主要由活动描述为“安保公司收到，转飞行区安检部。”的机场安检公司业务响应活动组成。这些同类簇的低抽象层次活动间有较强相关性且满足第 1 类关系，直接分析原生日志流程模型也能得到类似结果。

数据集 3 活动结点较多的类簇为 1, 11, 12。类簇 12 的活动描述以航班计划、共享航班等信息更新活动为主，活动间关系与数据集 1 的类簇 15、数据集 2 的类簇 2 相似。类簇 1 与类簇 11 的活动描述由机场地服公司开展的业务活动组成，但侧重不同；类簇 1 与数据集 1 中类簇 4 的活动结点描述相仿，侧重于机位作业业务，而类簇 11 则侧重于开展重点航班保障相关活动。类簇 1 与类簇 11 的低抽象层次活动间相关性较弱，直接分析原生日志流程模型易

表 3 基于 RankClus 算法的流程日志活动挖掘算法流程

输入：二类型网络 $\mathrm{G}=\langle\{\mathrm{X}\cup\mathrm{Y}\},\boldsymbol{W}\rangle$，排序函数 f, X 聚类数 K, Y 聚类数 L，最大迭代次数 iterNum，最小变动类簇数 espi

输出：$\left\{\mathrm{X}_1^{(t)},\mathrm{X}_2^{(t)},\cdots,\mathrm{X}_K^{(t)}\right\}$, $\left\{\mathrm{Y}_1^{(t)},\mathrm{Y}_2^{(t)},\cdots,\mathrm{Y}_L^{(t)}\right\}$

(1) $t=0,\ e=0$；

(2) $\left\{\mathrm{X}_1^{(t)},\mathrm{X}_2^{(t)},\cdots,\mathrm{X}_K^{(t)}\right\}=\mathrm{X}$ 的随机划分，$\left\{\mathrm{Y}_1^{(t)},\mathrm{Y}_2^{(t)},\cdots,\mathrm{Y}_L^{(t)}\right\}=\mathrm{Y}$ 的随机划分；

(3) while $t<$ iterNum && $e>$ espi do

(4) if $\left\{\mathrm{X}_k^{(t)}\right\}$ 或 $\left\{\mathrm{Y}_l^{(t)}\right\}$ 中存在空类；

(5) goto 步骤(2)；

(6) else

(7) for $k=1$ to K

(8) $\mathrm{G}_k^{(t)}=$ 用 $\mathrm{X}_k^{(t)},\mathrm{Y}$ 生成的 G 的子网络；

(9) $\left(\boldsymbol{r}_{\mathrm{X}_k|\mathrm{X}_k}^{(t)},\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}_k}^{(t)}\right)=f\left(\mathrm{G}_k^{(t)}\right)$；

(10) $\boldsymbol{r}_{\mathrm{X}|\mathrm{X}_k}^{(t)}=\boldsymbol{W}_{\mathrm{XY}}\boldsymbol{r}_{\mathrm{Y}|\mathrm{X}_k}^{(t)}$；

(11) end for

(12) 用混合模型估计参数 $\boldsymbol{\Theta}$，得到 X 每个节点 x_i 对应概率向量 $\boldsymbol{\theta}_i$；

(13) for $k=1$ to K

(14) 计算第 k 类 $\mathrm{X}_k^{(t)}$ 中心 $\boldsymbol{\lambda}_k^{(t)}$；

(15) end for

(16) foreach $x\in\mathrm{X}$

(17) for $k=1$ to K

(18) 计算 x 到第 k 类 $\mathrm{X}_k^{(t)}$ 中心距离 $D\left(x,\mathrm{X}_k^{(t)}\right)$；

(19) end for

(20) x 归入类 $\mathrm{X}_{k_0}^{(t+1)}$，$k_0=\arg\min_k D\left(x,\mathrm{X}_k^{(t)}\right)$；

(21) end for

(22) for $l=1$ to L

(23) 计算第 l 类 $\mathrm{Y}_l^{(t)}$ 中心 $\boldsymbol{\mu}_l^{(t)}$；

(24) end for

(25) foreach $y\in\mathrm{Y}$

(26) for $l=1$ to L

(27) 计算 y 到第 l 类 $\mathrm{Y}_{l_0}^{(t)}$ 中心距离 $D\left(y,\mathrm{Y}_l^{(t)}\right)$；

(28) end for

(29) y 归入类 $\mathrm{Y}_{l_0}^{t+1}$，$l_0=\arg\min_l D\left(y,\mathrm{Y}_l^{(t)}\right)$；

(30) end for

(31) end if

(32) t=t+1,e=$\left\{\mathrm{Y}_1^{(t)},\mathrm{Y}_2^{(t)},\cdots,\mathrm{Y}_L^{(t)}\right\}$ 中相比于 $\left\{\mathrm{Y}_1^{(t-1)},\mathrm{Y}_2^{(t-1)},\cdots,\mathrm{Y}_L^{(t-1)}\right\}$ 发生变动的类簇数目；

(33) end while

混淆这两类低抽象层次活动，影响流程发现准确性。只有通过活动聚类结果反映低抽象层次活动与业务流程的第 2 类关系，才可合理地区分低抽象层次活动。

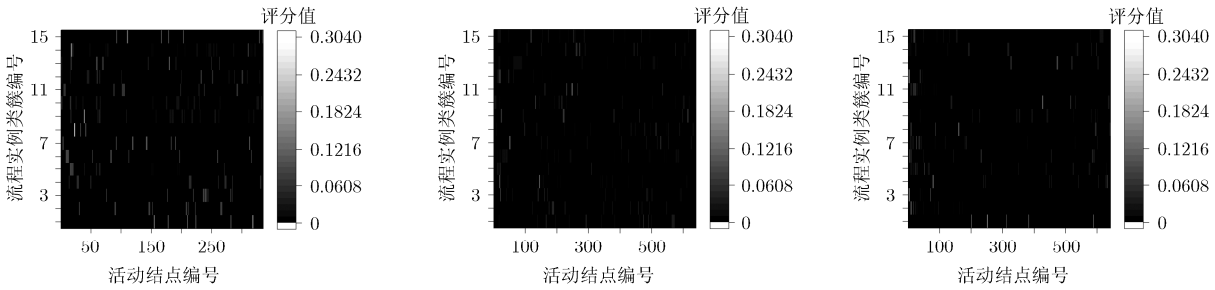
4.2 机场流程日志一致性检测实验

日志回放含 3 种情况[1]：(1)流程模型活动与当前踪迹活动匹配；(2)踪迹中活动与流程模型活动不匹配，模型预期活动未在踪迹中观测到时，回放算法可不移动踪迹中活动，前移流程模型中活动以进行匹配；(3)踪迹中活动与流程模型活动不匹配时，回放算法可不移动流程模型中活动，前移踪迹中活动以进行匹配。上述 3 种情况的日志回放准确度分别对应踪迹重现度(trace fitness)、模型移动重现度(move-model fitness)和日志移动重现度(move-log fitness) 3 项指标，取值范围均为 0 到 1。为 1 时意味着该情况下模型可完全回放日志。日志回放选用文献[18]基于代价的 A*算法。采用文献[17]中基于离散实例仿真系统分析的 DTW (Dynamic Time Warping)聚类算法作为对比算法。该方法运用离散实例仿真(Discrete Event Simulation, DES)技术将机场行李托运系统的运行建模为离散实例序列。采用 DTW 算法度量特定时刻用于标记系统状态变化的实例序列间的相似性并聚类。根据实例序列类簇特征分析系统行为(如是否存在瓶颈等)。实验结果如表 4 所示。

RankClus 活动挖掘算法活动聚类结果较为准确，活动类簇反映语义清晰，因此 RankClus 活动聚类流程模型的重现度指标与原生日志流程模型大致相当。DTW 活动挖掘算法聚类的实例序列与活动发生时刻相关性较强，所得流程模型中活动精确到时刻级别，模型过于精密，不能很好适应噪声数据。RankClus 活动挖掘算法所得的基于踪迹聚类的活动排序评分在反映当前流程日志活动信息的同时，包含更具意义的踪迹信息。若流程日志因条目更新等原因掺杂噪声，此时踪迹聚类结果不会急剧变化，模型通过日志移动仍可较好地重现流程日志。因此 RankClus 活动聚类模型的踪迹重现度与日志移动重现度要显著高于 DTW 活动聚类流程模型，而模型移动重现度与 DTW 活动聚类流程模型相当。整体而言，RankClus 活动聚类模型的鲁棒性要优于 DTW 活动聚类流程模型。

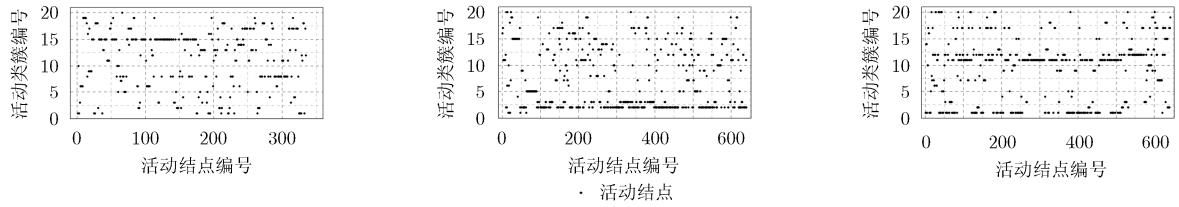
4.3 流程模型结构复杂度对比实验

Petri 网流程模型的结构复杂度可用 Petri 网中的与连接(AND-Joins)、与分歧(AND-Splits)、异或连接(XOR-Joins)、异或分歧(XOR-Splits)数评估。表 5 是对 3 个数据集添加活动类标签前后挖掘所得流程模型的结构复杂度分析结果。流程模型的结构复杂度主要决定于流程日志自身的内容而非所使用的流程挖掘算法[6]。基于 RankClus 的流程日志活动挖掘算法将数量较多的低抽象层次活动聚类为高抽象层次活动类簇，减少了 Petri 网变迁数，所得活动聚类流程模型结构复杂度相较于原生日志流程模型明显下降，且优于 DTW 活动聚类流程模型。



(a)数据集1活动结点评分值 (b)数据集2活动结点评分值 (c)数据集3活动结点评分值

图 2 各数据集下的活动评分



(a)数据集1聚类结果 (b)数据集2聚类结果 (c)数据集3聚类结果

图 3 各数据集下的活动聚类结果

表 4 流程模型一致性检测实验结果

数据集	评价指标	原生日志流程模型	DTW 活动聚类流程模型	RankClus 活动聚类流程模型
1	踪迹重现度	0.9758	0.8898	0.9826
	模型移动重现度	0.9953	1.0000	1.0000
	日志移动重现度	0.9852	0.8898	0.9826
2	踪迹重现度	1.0000	0.8961	0.9581
	模型移动重现度	1.0000	1.0000	1.0000
	日志移动重现度	1.0000	0.8961	0.9581
3	踪迹重现度	1.0000	0.8663	0.9061
	模型移动重现度	1.0000	1.0000	0.9913
	日志移动重现度	1.0000	0.8663	0.9189

表 5 流程模型结构复杂度实验结果

数据集	指标	原生日志流程模型	DTW 活动聚类流程模型	DTW 活动聚类流程模型结构复杂度下降比例(%)	RankClus 活动聚类流程模型	RankClus 活动聚类流程模型结构复杂度下降比例(%)
1	与连接	12	6	50	1	92
	与分歧	12	6	50	1	92
	异或连接	34	18	47	9	74
	异或分歧	35	19	46	11	69
2	与连接	13	4	69	4	69
	与分歧	13	4	69	4	69
	异或连接	51	19	63	19	63
	异或分歧	51	20	60	22	57
3	与连接	23	3	87	4	83
	与分歧	23	3	87	4	83
	异或连接	67	20	70	17	75
	异或分歧	59	19	68	17	71

5　结束语

本文针对非结构化的机场流程日志活动信息，提出基于 RankClus 算法的机场流程日志活动挖掘算法，构建二类型网络描述机场流程日志中活动与踪迹的关系，聚类日志中低抽象层次活动并得到 RankClus 活动聚类机场流程日志。实验表明，对该活动聚类流程日志挖掘所得 RankClus 活动聚类流程模型保持了较高日志重现度，同时显著降低流程模型结构复杂度，使流程模型更易于理解。对低抽象层次流程日志的流程挖掘有较大帮助。

参 考 文 献

[1] VAN DER AALST W M P. Process mining: Overview and opportunities[J]. *ACM Transactions on Management Information Systems*, 2012, 3(2): 1–17. doi: 10.1145/2229156.2229157.

[2] LANZ A, WEBER B, and REICHERT M. Time patterns for process-aware information systems[J]. *Requirements Engineering*, 2014, 19(2): 113–141. doi: 10.1007/s00766-012-0162-3.

[3] BOSE R P J C, VAN DER AALST W M P, ZLIOBAITE I, *et al*. Dealing with concept drifts in process mining[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, 25(1): 154–171. doi: 10.1109/TNNLS.2013.2278313.

[4] GÜNTHER C W, ROZINAT A, and VAN DER AALST W M P. Activity mining by global trace segmentation[C]. Proceedings of the 8th International Conference on Business Process Management, Hoboken, 2010: 128–139. doi: 10.1007/978-3-642-12186-9_13.

[5] DESAI N, BHAMIDIPATY A, SHARMA B, *et al*. Process trace identification from unstructured execution logs[C]. Proceedings of the 7th International Conference on Services Computing, Miami, 2010: 17–24. doi: 10.1109/SCC.2010.86.

[6] BAIER T, MENDLING J, and WESKE M. Bridging abstraction layers in process mining[J]. *Information Systems*, 2014, 46(12): 123–139. doi: 10.1016/j.is.2014.04.004.

[7] SONG M, GÜNTHER C W, and VAN DER AALST W M P. Trace clustering in process mining[C]. Proceedings of the 7th International Conference on Business Process Management, Ulm, 2009: 109–120. doi: 10.1007/978-3-642-00328-8_11.

[8] BOSE R P J C and VAN DER AALST W M P. Context aware trace clustering: towards improving process mining results[C]. Proceedings of the 2009 SIAM Data Mining Conference, Sparks, 2009: 401–412. doi: 10.1137/1.9781611972795.35.

[9] BOSE R P J C and VAN DER AALST W M P. Trace clustering based on conserved patterns: Towards achieving better process models[C]. Proceedings of the 8th International Conference on Business Process Management, Hoboken, 2010: 170–181. doi: 10.1007/978-3-642-12186-9_16.

[10] SUN Y, HAN J, ZHAO P, *et al*. Rankclus: integrating clustering with ranking for heterogeneous information network analysis[C]. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, Saint-Petersburg, 2009: 565–576. doi: 10.1145/1516360.1516426.

[11] FERREIRA D R, SZIMANSKI F, and RALHA C G. Improving process models by mining mappings of low-level events to high-level activities[J]. *Journal of Intelligent Information Systems*, 2014, 43(2): 379–407. doi: 10.1007/s10844-014-0327-2.

[12] SHAN S, WANG L, and LI L. Modeling of emergency response decision-making process using stochastic Petri net: an e-service perspective[J]. *Information Technology and Management*, 2012, 13(4): 363–376. doi: 10.1007/s10799-012-0128-7.

[13] 陈季梦, 陈佳俊, 刘杰, 等. 基于结构相似度的大规模社交网络聚类算法[J]. 电子与信息学报, 2015, 37(2): 449–454. doi: 10.11999/JEIT140512.
CHEN Jimeng, CHEN Jiajun, LIU Jie, *et al*. Clustering algorithms for large-scale social networks based on structural similarity[J]. *Journal of Electronics & Information Technology*, 2015, 37(2): 449–454. doi: 10.11999/JEIT140512.

[14] 陈丽敏, 杨静, 张健沛. 一种基于嵌入技术的异构信息网络的快速聚类算法[J]. 电子与信息学报, 2015, 37(11): 2634–2641. doi: 10.11999/JEIT150106.
CHEN Limin, YANG Jing, and ZHANG Jianpei. A fast clustering algorithm based on embedding technology for heterogeneous information networks[J]. *Journal of Electronics & Information Technology*, 2015, 37(11): 2634–2641. doi: 10.11999/JEIT150106.

[15] LEEMANS S J J, FAHLAND D, and VAN DER AALST W M P. Discovering block-structured process models from event logs containing infrequent behaviour[C]. Proceedings of the 11th International Conference on Business Process Management, Eindhoven, 2014: 66–78. doi: 10.1007/978-3-319-06257-0_6.

[16] GRABBE S R, SRIDHAR B, and MUKHERJEE A. Clustering days with similar airport weather conditions[C]. Proceedings of the 14th AIAA Aviation Technology, Integration, and Operations Conference, Atlanta, 2014: 2014–2712. doi: 10.2514/6.2014-2712.

[17] JOHNSTONE M, LE V T, ZHANG J, *et al*. A dynamic time warped clustering technique for discrete event simulation-based system analysis[J]. *Expert Systems with Applications*, 2015, 42(21): 8078–8085. doi: 10.1016/j.eswa.2015.06.040.

[18] ADRIANSYAH A, SIDOROVA N, and VAN DONGEN B F. Cost-based fitness in conformance checking[C]. Proceedings of the 11th International Conference on Application of Concurrency to System Design, Kanazawa, 2011: 57–66. doi: 10.1109/ACSD.2011.19.

徐　涛：男，1962 年生，教授，研究方向为数据挖掘、智能信息处理研究.

孟　野：男，1990 年生，硕士生，研究方向为机器学习、数据挖掘等.

卢　敏：男，1985 年生，助理研究员，研究方向为信息检索、文本挖掘等.