业务流程模型抽象中基于约束的行为聚类方法研究

王 楠 孙善武
(吉林财经大学管理科学与信息工程学院 长春 130117)
(吉林财经大学物流产业经济与智能物流吉林省重点实验室 长春 130117)

摘 要 将业务流程模型抽象中的行为聚合解释为一个半监督聚类过程,利用基于试探的启发式方法选择合适的行为集合作为初始簇,进而提高抽象的质量。另外,为了同时满足模型转换的保序性需求和子流程的业务语义完整性,在将行为归类到某个簇(候选子流程)时,进一步考虑了流程控制流的影响,设计了由两部分构成的约束函数,即语义距离和控制流顺序冲突。其中,第一部分引入了虚拟文档来表示行为和子流程,计算其之间的语义距离;第二部分利用行为概要文档中的4种行为顺序关系,设计函数来表示行为归类带来的控制流冲突。将该方法应用于真实的流程模型库,与传统的 k-means 行为聚类对比,如随机生成初始簇集和基于语义的距离测量方法,结果表明所提方法生成了更接近于人工设计的流程抽象结果。

关键词 业务流程模型抽象,基于约束的行为聚类,行为概要文档
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.01.048

Constraint-based Activity Clustering in Business Process Model Abstraction

WANG Nan SUN Shan-wu
(College of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China)
(Laboratory of Logistics Industry Economy and Intelligent Logistics, Jilin University of Finance and Economics, Changchun 130117, China)

Abstract This paper interpreted activity aggregation of business process model abstraction as a problem of semi-supervised clustering. It chooses appropriate activity sets as initial clusters based on a heuristic method to improve the quality of abstraction. In order to satisfy the order-preserving requirement of the model transformation and the business semantic integrity of the subprocesses, the control flow is further considered when classifying an activity to a cluster (candidate subprocess). A constraint function was designed with two parts: semantic distance and control flow conflict. The first part computs semantic distance between activities and subprocesses by introducing virtual document to represent them. In the second part, according to four ordering relations of behavioral profiles, a function is designed to show the control flow conflict caused by activity classifying. The proposed method is applied to a process model repository, comparing to the traditional k-means based activity clustering, such as methods of randomly generating the initial clusters and only based on semantics distance measurement, the proposed method is more closely approximating the decisions of the involved modelers to cluster activities.

Keywords Business process model abstraction, Constraint-based activity clustering, Behavioral profiles

1 引言

业务流程模型抽象(Business Process Model Abstraction, BPMA)最重要的用例是构建流程的概要视图以便快速理解复杂的流程模型[1,2]。为了解决该问题,可以将流程模型作为粗粒度行为的部分有序集合,其中每个粗粒度行为与一组较低层的细节行为相对应。聚合行为的方法有很多,从用户的角度来看,将语义上相似的一组行为进行聚合则更具有实际意义[3]。目前行为抽象方法大多是由流程结构驱动的模型转换[4-6],虽然结构聚合可以实现大型流程模型的相当大范围的化简,但是得出的结果模型并不能显示出所有需要的元素,或者将不应该聚合的元素进行了聚合,即这些方法没有考虑用户的抽象目标和抽象结果的业务意义。基于语义的业务流程模型抽象方法[2,7,8]则针对结构抽象方法的局限,提出从聚合行为的语义信息角度生成具有独立业务语义的行为集合。但是,业务流程模型中的行为集合是一类特殊的数据集,行为之间除了具有业务语义的相关性外,还具有流程控制流的顺序约束[3,5,9-11],这部分约束条件在模型抽象运算之前是已知的。目前基于语义的业务流程模型抽象采用的 k-means 聚类[8,19]是一种无监督的学习方法,均没有考虑该约束条件。

本文将行为语义和控制流一致性需求结合,设计相应的约束函数来指导行为聚类,并将该方法应用于真实的流程模

到稿日期:2016-03-19 返修日期:2016-06-20 本文受国家自然科学基金(61402193),吉林省教育厅"十三五"科学技术研究项目(2016105),吉林省教育科学"十二五"规划课题(GH150285)资助。

王 楠(1980—),女,博士,副教授,主要研究方向为业务流程模型抽象、基于模型的诊断、自动推理,E-mail:ctuwangnan@126.com;孙善武(1969—),男,硕士,副教授,主要研究方向为建模与抽象、传感器网络、网络安全。

型库，并与现有的基于 k-means 方法的行为聚类对比（如文献[8]中随机生成初始簇集和仅基于语义的距离测量方法），结果表明提出的方法生成了更接近于人工设计的子流程划分结果。

本文第 2 节给出使用到的一些概念；第 3 节解释了提出的行为聚类方法，同时给出了基于约束的行为聚类算法；第 4 节利用真实的业务流程模型库进行实验验证，并做了对比分析；最后对全文进行总结并展望了后续的研究方向。

2 相关定义

业务流程模型抽象通常是指行为的抽象，要求从低层步骤转换为高层任务[5]，本节引入文献[3]中的一些相关概念以便后续章节使用。

定义 1（业务流程模型）[3] 一个元组 $PM=(A,G,F,t,s,e)$ 称为一个业务流程模型，其中：A 是行为的有限非空集合；G 是网关的有限集合；$N=A\cup G$ 是节点的有限集合，且 $A\cap G=\emptyset$；$F\subseteq N\times N$ 表示流关系，使得 (N,F) 是一个连接图；每个行为至多有一个入边，至多有一个出边；s 是唯一没有入边的行为，即起始行为，e 是唯一没有出边的行为，即终止行为；$t:N_G\rightarrow\{and,xor\}$ 是为每个网关分配控制流构件的函数；每个网关表示 *split* 或 *join*，其中 *splits* 只有一个入边和至少两个出边，*joins* 有至少两个入边和唯一一个出边。

引入文献[3]中根据系统中的行为定义的保序抽象概念，这些行为用行为概要文档（Behavioral Profiles，BP）[12]来描述。令 $\mathcal{T}_{PM}$ 是流程模型 PM 的完整的流程轨迹集合，其中包含形式为 $sA*e$ 的列表，由行为的执行顺序组成。令 $a\in\sigma$ 表示行为 a 是一个完整流程轨迹的一部分，其中 $\sigma\in\mathcal{T}_{PM}$。

一个行为概要文档通过成对行为节点之间的 3 个关系来描述流程模型的行为特征，这些关系基于弱序的概念。如果流程模型中存在一个轨迹使得一个行为在另一个行为之后发生，那么称这两个行为是弱序关系。

定义 2（弱序关系）[3] 令 $PM=(A,G,F,t,s,e)$ 是一个流程模型，$\mathcal{T}_{PM}$ 是它的轨迹集合。弱序关系 $\succ_{PM}\subseteq(A\times A)$ 包含所有这样的行为对 (a,b)，使得在 $\mathcal{T}_{PM}$ 中存在一个轨迹 $\sigma=n_1,\cdots,n_l$，有 $j\in\{1,\cdots,l-1\}$，$j<k\leqslant l$，$n_j=a$ 且 $n_k=b$ 成立。

基于弱序关系定义，将行为概要文档定义如下。

定义 3（行为概要文档）[3] 令 $PM=(A,G,F,t,s,e)$ 是一个流程模型，行为对 $(a,b)\in(A\times A)$ 是以下关系之一：

strict 顺序关系 $\leadsto_{PM}$，如果 $a\succ_{PM}$ 且 $a\not\succ_{PM}b$；

exclusiveness 关系 $+_{PM}$，如果 $a\not\succ_{PM}b$ 且 $b\not\succ_{PM}a$；

interleaving 顺序关系 $\parallel_{PM}$，如果 $a\succ_{PM}b$ 且 $b\succ_{PM}a$。

所有这 3 种关系 $BP=\{\leadsto_{PM},+_{PM},\parallel_{PM}\}$ 的集合即是 PM 的行为概要文档。$a\not\succ_{PM}b$ 表示从 a 到 b 没有弱序关系。行为概要文档中的 3 种关系与 strict 顺序的逆关系：$\leadsto^{-1}=\{(a,b)\in(A\times A)\mid(b,a)\in\leadsto\}$ 共同划分行为集合的笛卡尔积。

定义 4（聚合函数）[3] 令 $PM=(A,G,F,t,s,e)$ 是一个流程模型，$PM_a=(A_a,G_a,F_a,t_a,s_a,e_a)$ 是 PM 对应的抽象模型。函数 $aggregate:A_a\rightarrow(P(A)\backslash\emptyset)$ 确定了 PM_a 中的一个行为与 PM 中的行为集合之间的对应关系。

定义 5（保序的业务流程模型抽象）[3] 令 $PM=(A,G,F,t,s,e)$ 是一个流程模型，业务流程模型抽象 α 将 PM 映射到 $PM_a=(A_a,G_a,F_a,t_a,s_a,e_a)$，即 $\alpha:(PM,activitygroups)\rightarrow PM_a$，使得 PM_a 中的行为是抽象对象。令函数 *aggregate* 用于建立 PM 与 PM_a 中的行为之间的关联关系，则运算 α 是保序业务流程模型抽象，当且仅当对于 $\forall x,y\in A_a,x\neq y,\forall a,b\in A,a\in aggregate(x)$ 且 $b\in aggregate(y)$，有以下结果成立：

$$a\leadsto_{PM}b\Rightarrow x\leadsto_{PM_a}y$$

$$a\leadsto_{PM}^{-1}b\Rightarrow x\leadsto_{PM_a}^{-1}y$$

$$a+_{PM}b\Rightarrow x+_{PM_a}y$$

$$-a\parallel_{PM}b\Rightarrow x\parallel_{PM_a}y$$

3 基于约束的半监督行为聚类

BPMA 的行为聚合过程可以依据流程结构标准，如基于模式的方法[4,13-15]和基于分解的方法[5,6,16-18]。根据结构方法发现的流程片段不能保证语义上的完整性，其中包含的行为在业务语义上有可能并不应该属于同一子流程。行为聚类也可以解释为行为业务语义的聚类分析问题[8]，待聚类的对象集合即为行为集合。对象根据某个聚类测量标准实现聚类：相互距离较“近”的对象被归类到一个集合。距离测量标准的语义部分可以根据行为业务语义的各种表示方法进行计算，向量空间方法[8]对行为属性值有较强的假设，只根据行为标签描述的测量方法[19]由于提供太少的信息而导致出现过多的错误聚类结果。因此考虑利用与行为相关联的尽可能多的信息表示行为，即引入虚拟文档[20]。另外，将行为聚合解释为一个半监督聚类分析[21]问题，并且在考虑行为的业务语义相似性的同时考虑模型转换的保序性要求。

对文献[8]中的传统 k-means 聚类方法进行扩展。首先根据业务流程模型的连接性结构特征，利用启发式方法选择初始聚类；然后将流程的控制流一致性与同组行为语义相似性相结合，构造指导聚类过程的实例层约束条件；最后给出 BPMA 的基于约束的 k-means 聚类算法，并用真实的流程进行实验验证。

3.1 初始簇中心确定

构成业务流程模型子流程的行为，除了在语义上具有较大的相似性外，在结构上通常也连接紧密，即空间上距离较近的行为比距离较远的行为更容易聚合成同一个子流程（不考虑流程的方向）。为了利用子流程的结构紧密特征，首先，将流程模型 PM 表示成一个无向图 $G(V,E)$，其中 V 表示流程中的节点（即定义 1 中的 N 集合），E 表示节点之间的边，设 $P(|V|\times|V|)$ 为图 G 对应的二维矩阵表示，若 $P(V_1,V_2)=1$，则表示节点 V_1 和 V_2 之间有边；若 $P(V_1,V_2)=0$，则表示节点 V_1 和 V_2 之间无边。

然后，构建行为距离矩阵 $D(|A|\times|A|)$（A 表示 PM 中的行为，$A\subseteq V$），并对其值进行初始化，其中 $D(a_2,a_1)(a_1,a_2\in A)$ 的值表示行为 a_1 与 a_2 之间的最短路径，采用 Dijistra 算法（复杂度为 $O(|V|^4)$），利用矩阵 P 计算得到。

接下来，采用文献[22,23]中的算法选取初始聚类中心在业务流程模型中的位置，该方法是模式识别领域中一种基于试探的启发式方法，其基本思想是取尽可能离得远的对象作为聚类中心，避免了初始选取时可能出现的初始聚类中心过于临近的情况。与之不同的是，根据矩阵 D，优先选择两个距离最远的行为作为初始聚类中心，而不是随机选择一个行为。具体算法描述如下。

1. 初始化 k//聚类(子流程)个数
2. 选择 D 矩阵中最大值对应的两个行为 a^1,a^2,并且令 S←{a^1,a^2},j←2
3. j<k 时,反复执行:
   3.1. j←j+1
   3.2. 选择与 S 距离最远的行为 a^j,S←S+a^j

其中,在第 3.2 步中,通过求解如下最优化问题来确定与行为集合 S 距离最远的行为 a^j,即最优函数:$\max\limits_{a^j\in A-S}\ \min\limits_{a^i\in S} D(a^j,a^i)$。

该最优函数表示:对集合 $A-S$ 中的每一个行为 a^t($1\leqslant t\leqslant|A-S|$,$|A-S|$ 表示集合 $A-S$ 中的行为个数),求出 a^t 到 S 中所有行为的最近距离 d_t,则 a^j 是与集合 S 距离最远的行为,当 $d_j=\max\limits_{1\leqslant t\leqslant|A-S|}\{d_t\}$。

假设流程模型 PM 中共有 n 个行为,则求解该最优函数所需要的渐进时间表达式为:$T=|A-S|\cdot|S|\cdot a+|A-S|\cdot b$,其中 a 和 b 为正常数。由于在算法第 2 步执行的过程中 $|S|\leqslant k<n$,且 $|A-S|\leqslant n$,则得到 $T\leqslant an\cdot k+nb<an^2+nb\in O(n^2)$。但是在真实的流程模型中,子流程的个数往往远远小于流程总的行为个数,即 $k\ll n$,因此实际应用中,求解最优函数的时间 $T\in O(n)$,即可以在线性时间内选出 k 个初始簇中心。

3.2　约束函数设计

为了同时满足模型转换的保序性需求和业务语义完整性需求,在将行为归类到某个簇(候选子流程)时进一步考虑流程控制流的影响,设计由两部分构成的约束函数,即语义距离和控制流顺序冲突。

令 $A=\{a_1,\cdots,a_n\}$ 是业务流程模型 PM 的行为集合,$D=\{d_1,\cdots,d_n\}$ 是行为对应的虚拟文档集合。$\{\mu_1,\cdots\mu_k\}$ 表示 3.1 节中初始化的簇集合 $\{S_1,\cdots,S_k\}$ 对应的 k 个划分中心。对于每个 $a\in A$,当将其分配到簇 S_i 时,不仅需要考虑 a 和 μ_i 之间的语义相似性(距离),同时也要考虑 a 加入到 S_i 可能产生的控制流冲突(约束函数的第二部分)。因此,将语义相似性和控制流顺序相结合来设计约束函数限制簇的选择方案,即当将行为 a 分配到某一个簇时,选择使得以下目标函数最小化的簇 S_i:

$$objective(S_i,a)=dist(d,\mu_i)+conflicts*(S_i\cup\{a\}) \quad (1)$$

对于约束函数的第一部分,引入虚拟文档[19]表示行为。

一个行为的虚拟文档由一些词构成,这些词来自于与该节点相关联的所有文本的信息。在业务流程模型上下文中,一个行为的虚拟文档由一些术语的集合构成,这些术语由行为标签、执行角色标签(如果该信息可用)、输入/输出数据以及行为的文本描述生成[24]。一组行为的虚拟文档则通过合并所有行为的文档生成。虚拟文档的生成包含了术语的标准化、连接词过滤以及词干提取等[23]。给定两个虚拟文档,可以基于它们向量空间的距离计算其相似性,其中维度就是出现在文档中的术语,各个维度的值使用术语出现的频率计算得到[25]。

比如,两个虚拟文档 d_1 和 d_2 分别用向量 $\overrightarrow{v_{d_1}}$ 和 $\overrightarrow{v_{d_2}}$ 表示,则它们的相似度用这两个向量夹角的余弦值计算,即:

$$sim(d_1,d_2)=\cos(\overrightarrow{v_{d_1}},\overrightarrow{v_{d_2}})=\frac{\overrightarrow{v_{d_1}}\cdot\overrightarrow{v_{d_2}}}{|\overrightarrow{v_{d_1}}||\overrightarrow{v_{d_2}}|}$$

两个虚拟文档 d_1 和 d_2 之间的距离为:

$$dist(d_1,d_2)=1-sim(d_1,d_2)$$

在式(1)中,d 是 a 的虚拟文档表示,$dist(d,\mu_i)$ 表示根据以上距离测量方法计算的行为 a 与簇 S_i 中心的距离。

抽象模型中的每个行为映射为原始模型中的一组细节行为,两个抽象行为之间的最终控制流关系可能导致原始模型中对应的细节行为之间的控制流顺序不一致。设有原始模型 PM 中的行为 a,b,c,其中 a 和 c 之间的关系为 r_1,b 和 c 之间的关系为 r_2。PM_a 为 PM 对应的抽象模型,PM_a 中的抽象行为为 x,y,且 x 和 y 之间的关系为 r,其中行为 a 和 b 分别映射到抽象行为 x,c 映射到抽象行为 y。观察到,抽象行为 x 和 y 在抽象 PM_a 模型中的关系 r 导致与其各自对应的 PM 中的行为关系也变更为 r,即 a 与 c 及 b 与 c 之间的关系变为 r,因此,若原始模型 PM 中 $r_1\neq r_2$,则 a,b,c 的归类产生了控制流冲突。

显然地,可以通过生成抽象模型中各个抽象行为之间的控制流关系,推导出该抽象模型引起的原始模型中细节行为的控制流冲突情况,并以此判断抽象结果与原始模型的一致性程度。但是,一方面,如何生成抽象行为之间的控制流关系并非本文研究的范围(详见文献[26]);另一方面,对最终结果模型的评估需要对所有行为完成聚类,无法在抽象过程中对某一行为的归类进行指导。

因此,设计式(1)中的第二部分 $conflicts*(S_i\cup\{a\})$,用来表示将行为 a 归类到 S_i 引起的可能的控制流顺序冲突,冲突的计算利用了行为概要文档(定义详见第 2 节)。

行为概要文档定义了行为之间的 4 种顺序关系:$R=\{\rightsquigarrow_{PM},\rightsquigarrow_{PM}^{-1},+PM,\|_{PM}\}$,设 $PM=(A,G,F,t,s,e)$ 是一个流程模型,$PM_a=(A_a,G_a,F_a,t_a,s_a,e_a)$ 是其对应的抽象模型,BP 是 PM 的行为概要文档。对于行为 $a,b,c\in A$,设 $\exists z\in A_a$,使得 $a,b\in aggregate(z)$,$c\notin aggregate(z)$,$BP(a,c)=r_i$ 并且 $BP(b,c)=r_j$($r_i,r_j\in R$),则 w_{r_i,r_j} 表示将 a 和 b 聚合到 z 导致的冲突权值,定义如下:

$$w_{r_i,r_j}=\begin{cases}1, & r_i\neq r_j\text{//控制流发生冲突而不聚合}\\ 0, & r_i=r_j\text{//控制流不冲突而可以聚合}\end{cases}$$

令 $S\subset A$ 是 A 的子集,对于每个行为 $a_k\in A\backslash S$,S 与 a_k(将 a_k 归类至 S)的冲突值可以通过式(2)进行计算:

$$conflicts(S,a_k)=\frac{1}{|S|(|S|-1)}\sum_{\substack{a_i,a_j\in S\\ 1\leqslant i\leqslant j\leqslant|S|}} w_{BP(a_i,a_k),BP(a_j,a_k)} \quad (2)$$

其中,$|S|$ 表示集合 S 中行为的个数。

进一步地,集合 S 的控制流冲突值用式(3)计算:

$$conflicts*(S)=\frac{1}{|A\backslash S|}\sum_{a_k\in A\backslash S} conflicts(S,a_k) \quad (3)$$

3.3　算法描述

基于 seeded-KMeans 算法[27],利用 3.1 节生成的初始簇集合和式(1)所示的目标函数 objective 作为输入参数,给出 BPMA 基于约束的聚类算法描述,如算法 Constrained-clustering-for-BPMA 所示。

算法　Constrained-clustering-for-BPMA

输入:待处理业务流程模型的行为集合 A={a_1,…,a_n}对应的虚拟文档集合 D={d_1,…,d_n};子流程(簇)数 k;初始簇(种子)集合 $S=\bigcup_{i=1}^{k}S_i$;行为概要文档 BP(n×n)

输出:使得目标函数最小化的 A 的 k 个不相交划分 $\{C_i\}_{i=1}^{k}$

1. 初始化:初始簇中心 $\mu_i^{(0)}\leftarrow\frac{1}{|S_i|}\sum_{d\in S_i}d$,初始划分 $C_i^{(0)}\leftarrow S_i$(i=1,…,k);t←0

2. 重复执行以下步骤，直到所有行为不能再分配：

2.1. 行为再分配：将每个行为 a 归类到簇 h^*（即集合 $C_h^{(t+1)}$），使得

$h^* = h \mid \min(\text{objective}(C_h, a))$

2.2. 估算新的簇中心：$\mu_h^{(t+1)} \leftarrow \frac{1}{|C_h^{(t+1)}|} \sum_{d \in C_h^{(t+1)}} d$

2.3. $t \leftarrow (t+1)$

3. 输出所有 k 个划分 $\{C_i\}_{i=1}^k$

算法 Constrained-clustering-for-BPMA 的时间复杂度主要集中在对行为的再分配，需要对每个行为求解其与 k 个簇的目标函数值（步骤 2.1），如式(1)所示。设模型中行为个数为 n，根据式(2)和式(3)可以估算该步骤的时间渐进表示为 $T \approx \sum_{i=1}^{k} |A - C_i| \cdot |C_i|^2 \cdot a + b$，其中 a 和 b 为常数，b 表示求解语义距离 $dist$ 所用的时间。进一步地，$T < a \cdot k \cdot n^3 + b \in O(n^3)$，由此得到整个算法的时间复杂性属于 $O(n^4)$。

4 实验验证

为了验证提出的行为聚类方法对业务流程模型的抽象结果与人工抽象结果的相似程度，本节对真实的业务流程模型集合进行了实验验证，并对验证结果进行了解释。

4.1 实验构架

(1)选择业务流程模型集合

本文基于笔者所在的省重点实验室开放项目，从项目的合作单位（某大型合资汽车生产商及其伙伴物流公司）获取了流程模型集合作为研究对象。选取了 40 个行为标签描述规范、行为属性来描述完整的流程模型集合，其中均包含人工设计子流程的结构。为了利用出现在行为及其属性标签中的词，将其以向量空间的形式表示为虚拟文档，与相关工作人员一起对术语进行了重新规范，并达成共识。另外，为了获取尽量多的信息，进一步考虑了流程中的控制流信息，并将提取的词加入到相邻行为的虚拟文档中，单数字信息根据含义转化为变量名。由此，流程模型转化为了虚拟文档集合，其中保留子流程的层次关系。

需要指出的是，由于该公司的全局生产线的生产流程很复杂，流程模型涉及的域也比较广，包括业务流程、生产流程、物流流程，并且生产流程中经常包含了各种零部件的物流子流程，因此，为了验证提出的方法不依赖于行为的具体域，保证实验结果的有效性，对于具有子流程层次的模型，按照以下原则对模型进行选取：1)展开后规模适中；2)包含尽量多的其他域子流程；3)尽量不选取包含多于两层子流程的流程模型。表 1 给出了所选流程模型的相关属性。

表 1 实验流程模型 $M_1 - M_{40}$ 的相关属性

	Activities	Subprocesses	Activities of Subprocesses
Average	94.10	7.97	7.52
Maximum	127.00	20.00	10.50
Minimum	59.00	3.00	4.20

(2)评估 BPMA 的基于约束的聚类算法

对 BPMA 同时应用 2.3 节提出的基于约束的聚类算法（Constrained_Clustering_for_BPMA）和传统的无监督 k-means 聚类过程（本文称为 K-Means_for_BPMA）。第二种算法与文献[8]类似，通过计算行为和簇之间的距离自动获取细节，展开流程模型的子流程分解，其中距离的计算只根据行为的业务语义（如本文的“dist”），并采用随机选择行为的方法进行簇中心的初始化。对这两种方法生成的抽象结果进行比较，将包含人工设计子流程的流程模型 $M_1 - M_{40}$ 转换成对应的展开模型，分别应用算法 Constrained_Clustering_for_BPMA 和算法 K-Means_for_BPMA 生成簇（子流程），比较它们与初始人工设计子流程的相似度。

引用文献[18]中定义的部分相关指标来比较人工设计的子流程分解与自动生成行为簇之间的各种特征，具体说明如下。

1)subprocesses：模型中子流程的数目。

2)avg activities per subprocess：每个子流程中行为的平均数目。

3)max activities each subprocess：子流程中行为的最大数。

4)min activities per subprocess：子流程中行为的最小数。

5)Precision：自动生成的并且同时也是人工定义的子流程数除以自动生成的子流程数。

6)Recall：自动生成的并且同时也是人工定义的子流程数除以人工定义的子流程数。

7)Overshoot：自动生成的候选子流程的节点中不属于与该子流程匹配的人工定义的子流程的节点比例。

8)Undershoot：应该属于某个人工定义的子流程但是在自动生成的候选子流程中没有生成的节点比例。

对于 Precision 和 Recall，采用节点匹配法，而不是流程的整体匹配法。根据文献[18]，各个测量指标的计算定义如下。

令 N 是一个流程中的所有节点的集合（包括其子流程），$P_M \subseteq PN$ 是人工确定的子流程集合，$P_A \subseteq PN$ 是自动确定的候选子流程集合。$P_M \in P_M$ 是一个人工确定的子流程，$P_A \in P_A$ 是一个自动确定的候选子流程。P_A 与 P_M 之间的覆盖 Overlap 定义为：

$$Overlap = \frac{|P_A \cap P_M|}{\max(|P_A|, |P_M|)}$$

如果 P_A 与 P_M 之间的覆盖 $Overlap > 0$，并且不存在其他自动生成的子流程 $P_A' \in P_A$ 与 P_M 之间的覆盖 $Overlap' > Overlap$，则称 P_A 是 P_M 的一个最相关匹配。令函数 $match$：$P_M \rightarrow PN$ 返回每个人工定义的子流程的最相关匹配，如果不存在，则返回空集。

Precision 和 Recall 定义如下：

$$Precision = \frac{\sum_{P_M \in P_M} |P_M \cap match(P_M)|}{\sum_{P_A \in P_A} |P_A|}$$

$$Recall = \frac{\sum_{P_M \in P_M} |P_M \cap match(P_M)|}{\sum_{P_M \in P_M} |P_M|}$$

F 值定义为 Precision 和 Recall 两个值的调和平均数：

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Overshoot 和 Undershoot 分别定义如下：

$$Overshoot = \frac{\sum_{P_M \in P_M} |\text{match}(P_M) - P_M|}{\sum_{P_A \in P_A} |P_A|}$$

$$Undershoot = \frac{\sum_{P_M \in P_M} |P_M - \text{match}(P_M)|}{\sum_{P_M \in P_M} |P_M|}$$

4.2 结果与分析

表 2 列出了对流程模型 $M_1 - M_{40}$ 运行两种算法后的验证结果，为了简化，只给出 40 个模型对于 4.1 节中各个指标值的平均值。

表 2　模型 M_1-M_{40} 中各种指标的平均值

Metric	Constrained_Clustering_for_BPMA	K-Means_for_BPMA	Original
subprocesses	8.40	8.40	8.40
avg activities per subprocess	12.30	13.59	8.31
max activities each subprocess	23.50	33.80	15.80
min activities per subprocess	4.80	2.10	4.10
Precision	0.54	0.30	—
Recall	0.60	0.35	—
F	0.58	0.34	—
Overshoot	0.35	0.57	—
Undershoot	0.40	0.65	—

F 值是一个很重要的指标，它给出了自动子流程分解与人工设计的子流程分解之间的接近程度[19]。根据表 2 的结果可以看出，算法 Constrained_Clustering_for_BPMA 的抽象过程比 K-Means_for_BPMA 算法的划分方法更加接近于人工划分结果。

由于使用行为语义和流程结构相结合的约束函数指导聚类过程，使得每个自动生成的子流程中行为的最大数大大减少，更接近于人工设计的子流程，这表明对归类行为进行了一个相对有效的控制。

当然，也发现 Overshoot 和 Undershoot 的值仍然偏高。两种行为聚类算法都是一种“硬聚类”方法，即每个行为必须属于一个并且仅属于一个子流程。但是在实际的流程模型抽象中，人工划分过程使得存在不属于任何人工子流程的行为。对 M_1-M_{40} 中每个流程的行为总数以及那些包含在人工构造子流程中的行为数进行了统计比较，发现在平均情况下，有 10%的行为不属于任何子流程，但在运行本文的两种算法时，这些行为都会自动归类到某个子流程中。这些行为是导致 Overshoot和 Undershoot 值偏高的一个主要原因。

导致 Overshoot 和 Undershoot 值偏高的另一个原因是，一些行为即使语义上或结构上与子流程 S_2 距离较近（如本文提出的 *objective* 函数或其它函数计算的距离值），但根据设计者的标准，它们仍然被归类到了另外一个距离较远的子流程 S_1 中。这种情况下，仅仅根据行为与子流程之间的相似性值对行为进行归类已经不能满足实际的需求。

结束语　本文提出了一种新的业务流程模型抽象方法，该方法根据行为语义和控制流顺序，采用基于约束的半监督聚类算法发现互相关联的行为集合，其中每个行为集合对应抽象流程模型中的一个粗粒度行为，基于真实的流程模型库进行的实验结果验证了提出方法的适用性及有效性。

本文提出的方法有以下局限和假设：首先，该方法假设子流程数 k 可以事先确定，但是在实际应用中发现，它很难根据建模者的经验准确获得。其次，k-means 聚类是一个数据集的硬划分方法，即每个行为必须归类至某一个子流程。但是实际上，在建模者手工进行子流程划分时，存在大量不属于任何子流程的行为，有些与某个子流程距离较近的行为甚至会被人工分配到其他距离较远的子流程中。

这些局限为我们提供了未来研究的方向，比如最直接的方向就是设计恰当的评价指标来评估流程抽象结果，生成最优子流程数。另外，也可以应用和改进软聚类技术，如用 FCM(Fuzzy C-Means)聚类代替 k-means 聚类，从而更灵活地对行为进行子流程归类。

参 考 文 献

[1] SMIRNOV S, REIJERS H A, WESKE M H, et al. Business process model abstraction: a definition, catalog, and survey[J]. Distributed and Parallel Databases, 2012, 30(1): 63-99.

[2] SMIRNOV S, DIJKMAN R, MENDLING J, et al. Meronymy-based aggregation of activities in business process models[C]// Conceptual Modeling-ER 2010. Lecture Notes in Computer Science, Volume 6412, 2010: 1-14.

[3] SMIRNOV S. Business Process Model Abstraction, [Doctor Dissertation]. Germany: University of Potsdam[OL]. http://opus.kobv.de/ubp/volltexte/2012/6025/pdf/smirnov_diss.pdf.

[4] POLYVYANYY A, SMIRNOV S, Weske M. Reducing Complexity of Large EPCs[OL]. http://polyvyanyy.com/pdf/psw_MOBIS_EPK_2008-PSO TPRINT.pdf.

[5] POLYVYANYY A, SMIRNOV S, Weske M. On Application of Structural Decomposition for Process Model Abstraction[C]// Proceedings of the BPSC 2009. Leipzig, 2009: 110-122.

[6] VANHATALO J, VÖLZER H, KOEHLER J. The refined process structure tree[J]. Data & Knowledge Engineering, 2009, 68(9): 793-818.

[7] FRANCESCOMARINO C D, MARCHETTO A, TONELLA P. Cluster-based Modularization of Processes Recovered from Web Applications[J]. Journal of Software Maintenance and Evolution Research and Practice, 2013, 25(2): 113-138.

[8] SMIRNOV S, REIJERS H A, WESKE M. A Semantic Approach for Business Process Model Abstraction[C]// Proceedings of the CAiSE 2011, Vol. 6741 of LNCS. Springer, 2011: 497-511.

[9] BOBRIK R, REICHERT M, BAUER T. View-Based Process Visualization[C]// BPM 2007. Berlin, Vol. 4714, 2007: 88-95.

[10] ESHUIS R, GREFEN P. Constructing Customized Process Views [J]. Data and Knowledge Engineering, 2008, 64(2): 419-438.

[11] LIU D, SHEN M. Workflow Modeling for Virtual Processes: an Order- Preserving Process-View Approach[J]. Information Systems, 2003, 28(6): 505-532.

[12] WEIDLICH M, MENDLING J, WESKE M. Efficient Consistency Measurement based on Behavioural Profiles of Process Models[J]. IEEE Transactions on Software Engineering, 2011, 37(3): 410-429.

[13] SMIRNOV S, WEIDLICH M, MENDLING J, et al. Object-Sensitive Action Patterns in Process Model Repositories[M]// Business Process Management Workshops. vol. 66, 2010: 251-263.

[14] LOHRMANN M, REICHERT M. Effective application of process improvement patterns to business processes[M]// Software & Systems Modeling. Springer, 2015.

[15] SMIRNOV S, WEIDLICH M, MENDLING J, et al. Action Patterns in Business Process Models [C] // ICSOC/ServiceWave 2009. vol. 5900, 2009: 115-129.

[16] POLYVYANYY A, SMIRNOV S, WESKE M. The Triconnected Abstraction of Process Models[C]// BPM 2009. Germany, vol. 5701, 2009: 229-244.

[17] VÖLZER V H, LEYMANN F. Faster and More Focused Control-Flow Analysis for Business Process Models Through SESE Decomposition[J]. Lecture Notes in Computer Science, 2007, 4749: 43-55.

（下转第 294 页）

[4] GE Pan-pan, CHEN Qiang, GU Yi-he. Based on Harris corner and surf feature of remote sensing image matching algorithm [J]. Application Research of Computers, 2014, 31(7): 2205-2208. (in Chinese)
葛盼盼,陈强,顾一禾. 基于 Harris 角点和 SURF 特征的遥感图像匹配算法[J]. 计算机应用研究,2014,31(7):2205-2208.

[5] MOBLEY C D, SUNDMAN L K, DAVIS C O, et al. Interpretation of hyperspectral remote-sensing imagery by spectrum matching and look-up tables[J]. Applied Optics, 2005, 44(17): 3576-3592.

[6] STERNLICHT D D, DE MOUSTIER C P. Remote sensing of sediment characteristics by optimized echo-envelope matching [J]. The Journal of the Acoustical Society of America, 2003, 114(5): 2727-2743.

[7] LOWE D G. Object recognition from local scale-invariant features[C]// The Proceedings of the Seventh IEEE International Conference on. IEEE, Computer Vision, 1999, 2: 1150-1157.

[8] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[9] MIKOLAJCZYK K, SCHMID C. Scale & affine invariant interest point detectors[J]. International Journal of Computer Vision, 2004, 60(1): 63-86.

[10] SEDAGHAT A, EBADI H. Remote Sensing Image Matching Based on Adaptive Binning SIFT Descriptor[J]. IEEE Transactions on Geosience & Remote Sensing, 2015, 53(10): 5283-5292.

[11] BENEDIKTSSON J A, PESARESI M, AMASON K. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations [J]. IEEE Transactions on Geoscience and Remote Sensing, 2003, 41(9): 1940-1949.

[12] KOENDERINK J J. Scale-time[J]. Biological Cybernetics, 1988, 58(3): 159-162.

[13] LINDEBERG T. Scale-space theory: A basic tool for analyzing structures at different scales[J]. Journal of Applied Statistics, 1994, 21(1/2): 225-270.

[14] DU B, ZHANG L. A discriminative metric learning based anomaly detection method[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(11): 6844-6857.

[15] DU B, ZHANG L. Target detection based on a dynamic subspace[J]. Pattern Recognition, 2014, 47(1): 344-358.

[16] ZHANG F, DU B, ZHANG L. Saliency-guided unsupervised feature learning for scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(4): 2175-2184.

[17] ADELSON E H, ANDERSON C H, BERGEN J R, et al. Pyramid methods in image processing[J]. RCA Engineer, 1984, 29(6): 33-41.

[18] FERGUSON T S. A Bayesian analysis of some nonparametric problems[J]. The annals of statistics, 1973: 209-230.

[19] MEI Su-yu, WANG Fei, ZHOU Shui-geng. Dirichlet process mixture model, the extended model and application [J]. Science Bulletin, 2013, 57(34): 3243-3257. (in Chinese)
梅素玉,王飞,周水庚. 狄利克雷过程混合模型,扩展模型及应用[J]. 科学通报,2013,57(34):3243-3257.

[20] ZHANG Lin, LIU Hui. The clustering algorithm of mixed model in Dirichlet process [J]. Journal of China University of Mining and Technology, 2012, 41(1): 159-163. (in Chinese)
张林,刘辉. Dirichlet 过程混合模型的聚类算法[J]. 中国矿业大学学报,2012,41(1):159-163.

[21] NEUBERT M, HEROLD H, MEINEL G. Evaluation of remote sensing image segmentation quality-further results and concepts [C]// Proceedings of the 1 st International Conference on Object-Based Image Analysis. 2006.

[22] LIU Y, BIAN L, MENG Y, et al. Discrepancy measures for selecting optimal combination of parameter values in object-based image analysis[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2012, 68(1): 144-156.

[23] ZHANG X, XIAO P, FENG X. An unsupervised evaluation method for remotely sensed imagery segmentation[J]. IEEE Geoscience & Remote Sensing Letters, 2012, 9(2): 156-160.

[24] NEUBERT M, HEROLD H. Assessment of remote sensing image segmentation quality[C]// Proceedings of Geobia Pixels Objects Intelligence-geographic Object Based Image Analysis for the St Century. 2008.

[25] CARDOSO J S, CORTE-REAL L. Toward a generic evaluation of image segmentation[J]. IEEE Transactions on Image Processing, 2005, 14(11): 1773-1782.

(上接第 263 页)

[18] ANUGRAH I G, SARNO R, ANGGRAINI R N E. Decomposition using Refined Process Structure Tree (RPST) and control flow complexity metrics[C]// 2015 International Conference on Information & Communication Technology and Systems (ICTS). Surabaya, 2015: 203-208.

[19] REIJERS H A, MENDLING J, DIJKMAN R M. On the Usefulness of Subprocesses in Business Process Models[OL]. http://bpmcenter. org/wp-content/uploads/reports/2010/BPM-10-03. pdf.

[20] QU Y, HU W, CHENG G. Constructing virtual documents for ontology matching[M]// The Semantic Web. Springen Berlin Heidelberg, 2012: 23-31.

[21] ZHAO Wei-zhong, MA Hui-fang, LI Zhi-qing, et al. Efficiently active learning for semi-supervised document clustering [J]. Journal of Software, 2012, 23(6): 1486-1499. (in Chinese)
赵卫中,马慧芳,李志清,等. 一种结合主动学习的半监督文档聚类算法[J]. 软件学报,2012,23(6):1486-1499.

[22] BELIAKOV G, KING M. Density based fuzzy c-means clustering of non-convex patterns[J]. European Jounal of Operational Research, 2006, 173(3): 717-728.

[23] PORTER M F. An algorithm for suffix stripping [J]. Program, 1980, 14(3): 130-137.

[24] WEIDLICH M, DIJKMAN R, MENDLNG J. The ICoP framework-Identification of correspondences between process models [J]. Advanced Information Systems Engineering, Lecture Notes in Computer Science Volume 6051, 2010: 483-498.

[25] EUZENAT J, SHVAIKO P. Ontology matching [M]. Springer-Verlag, 2007.

[26] SMIRNOV S, WEIDLICH M, MENDLING J. Business Process Model Abstraction Based on Behavioral Profiles[J]. Service-Oriented Computing, Lecture Notes in Computer Science, 2010, 6470: 1-16.

[27] BASU S, BANERJEE A, MOONEY R J. Semi-Supervised clustering by seeding[C]// Proc. of the 9th Int'l Conf. on Machine Learning. 2002: 19-26.