软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(3):489−503 [doi: 10.3724/SP.J.1001.2012.04010]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一个基于流程相似性的自动服务发现框架*

黄子乘+, 怀进鹏, 刘旭东, 李 翔, 朱蒋俊

(北京航空航天大学 计算机学院,北京 100191)

Automatic Service Discovery Framework Based on Business Process Similarity

HUANG Zi-Cheng+, HUAI Jin-Peng, LIU Xu-Dong, LI Xiang, ZHU Jiang-Jun

(School of Computer Science and Engineering, BeiHang University, Beijing 100191, China)

+ Corresponding author: E-mail: huangzc@act.buaa.edu.cn

Huang ZC, Huai JP, Liu XD, Li X, Zhu JJ. Automatic service discovery framework based on business process similarity. *Journal of Software,* 2012,23(3):489−503. http://www.jos.org.cn/1000-9825/4010.htm

Abstract: Service discovery is a critical stage of the development for Internet-scale software produced through service composition. Presently, development efficiency of composite service is confined by a low degree of automation and accuracy of service discovery. This paper propose the AutoDisc (automatic service discovery framework based on business process similarity) schema to improve development efficiency through two aspects. One is to improve the efficiency of discovery by automatic recommendation. The other is to improve the accuracy of service discovery by combining the structural and behavioral factors of services. Through the proposed approach, this paper automatically models the requirements of service discovery and recommends the most appropriate composite services to developers. Finally, the paper illustrates the effectiveness of AutoDisc with a set of experimental evaluations which show that AutoDisc can increase the development efficiency by 75.5% in the evaluating context.

Key words: automatic service discovery; service recommendation; process similarity; service composition

摘 要: 服务发现是面向服务的网络软件开发过程的关键阶段,同时也是影响服务组合效率的关键因素.针对当前服务发现自动化程度低下、准确性不高的现状,从两个方面提高服务组合效率:首先,提出一种自动组合服务发现模式,支持流程粒度的组合服务发现及复用;其次,在发现过程中设计了一种量化的相似性评估算法,综合考虑组合流程的静态结构与动态行为特征,以提高服务发现的准确率;最后,结合以上两方面形成一个基于流程相似性的自动服务发现框架(automatic service discovery framework based on business process similarity,简称 AutoDisc).利用真实数据完成的评估实验结果表明,AutoDisc 的准确性优于单纯考虑结构或行为的发现方法,在所给出的应用案例中,使服务组合效率提高 75.5%,具有较好的可扩展性.

关键词: 自动服务发现;服务推荐;流程相似性;服务组合

中图法分类号: TP311 **文献标识码**: A

在开放网络环境中,软件需求呈现高度动态及不断变化的特征,传统的软件开发方法已很难满足高效网络

* 基金项目: 国家高技术研究发展计划(863)(2007AA010301, 2006AA01A106, 2009AA01Z419)
收稿时间: 2010-04-05; 定稿时间: 2011-03-07

化软件开发的需求.在此背景下,随着服务计算技术的发展,基于服务组合的网络化软件开发方法应运而生,该方法通过组合互联网中丰富的服务资源,高效地开发可以满足复杂、动态业务需求的网络化应用[1].当前,服务组合方法已经得到学术界的广泛重视,其中,流程感知的服务组合方法借鉴传统工作流(workflow)和业务流程管理(business process management)等领域的研究成果,已取得了一定的研究进展并成功应用到工业生产实践中[2].在该方法中,业务流程用于描述服务间的交互和调用关系,发挥着贯穿整个开发过程的关键性作用.流程感知的组合过程首先由最终用户向业务分析人员描述需求,后者使用诸如BPMN[3],UML此类业务层次的建模语言,通过图形符号和图元属性刻画一种抽象的需求流程模型;软件开发人员则根据该模型将服务资源库中提供的服务进行组合,对流程模型进一步编排和精化,形成可执行的组合服务流程.

在网络计算时代,提高开发效率和质量依然是网络软件设计生产的核心问题,流程感知的服务组合在一定程度上为这一问题的解决提供了方法.然而,服务组合效率的进一步提高仍然面临着巨大的挑战:首先,随着服务计算技术的广泛应用及发展,开放网络环境中存在的可访问且可复用的服务资源数目呈爆炸式增长,同时,服务资源的描述方式也越来越复杂且多样化.开发者需要掌握大量服务描述标准、检索技术以及领域相关知识,才能发现合适的可组合服务.这一现状极大地限制了组合的自动化程度,因此,服务组合过程中的自动服务发现是提高组合效率的一个重要因素;其次,现有的服务发现主要针对单一的组件服务,导致服务资源的复用粒度和效率低下.流程感知的服务组合方法为我们提供了一个提高资源复用粒度的基础,同时,在一些应用领域经过长期运营实践,大量业务活动和业务逻辑逐步积累,形成了标准化、流程化的业务操作,这些操作流程对新应用的开发具有及其重要的参考和复用价值.因此,将服务资源的复用粒度提升为流程,针对流程及其片段进行服务发现具有重要意义;最后,现有的服务发现方法主要关注于其描述信息的语法、语义或流程结构关系,缺乏动态行为特征的考虑.而服务的功能与其行为是密切相关的,一个服务在运行时所表现出来的行为特征决定了其所完成的功能,另外,两个在语法、语义和流程结构上十分相似的服务,其运行时的行为也可能截然不同.因此,现有方法缺乏行为特征的考虑严重影响了服务发现的准确性[4,5].

针对上述问题,本文提出了一个基于流程相似性的自动 Web 服务发现框架(automatic Web service discovery framework based on business process similarity,简称 AutoDisc),主要贡献包括:

1) 针对组合服务流程,基于 Petri 网理论,建立可综合描述其结构和行为信息的基础模型 PSM;
2) 在 PSM 模型的基础上,针对当前广泛采用的业务流程建模语言 BPMN,提出一种自动化的模型映射方法,以支持组合服务开发需求的自动建模;
3) 基于 PSM 提出一种自动服务发现机制,以组合服务流程或流程片段作为基本单元,通过自动化的模型映射对功能需求建模,然后分析 PSM 模型可达图之间的模拟关系,综合流程的结构及行为特征实现量化的相似性判定算法,根据抽象需求自动发现可满足功能的组合流程片段,支持需求到组合实现的自动映射.

本文第 1 节分析国内外相关的研究现状.第 2 节通过一个应用场景分析引出 AutoDisc 所解决的几个问题.第 3 节给出组合服务描述模型 PSM 的形式化定义以及 BPMN 到 PSM 的自动映射方法.第 4 节详细阐述如何基于 PSM 模型的可达图进行组合服务流程的相似性判定以及自动匹配.第 5 节对 AutoDisc 框架进行实验分析及评估.最后对全文进行总结并对未来的工作进行展望.

1 相关工作比较

对于自动服务发现及流程相似性判定的相关研究,目前已有工作取得了一定的进展.本节从两个方面对相关工作进行分析比较.

自动服务发现及服务推荐.服务发现是影响服务组合效率的重要因素,针对服务发现的特征及问题,文献[6]讨论了一种候选服务过滤的方法来提高服务发现的效率,其基本思想是,在服务组合过程中,首先采用服务操作接口间的可连接性选取候选服务,然后基于组合上下文(所处的领域等)及用户操作历史信息(用户偏好等)对候选服务进行进一步过滤得到最终结果.另一些工作则通过收集及分析服务调用的历史记录,挖掘服务间的使用

及依赖关系,结合服务发现过程以提高其效率和准确性[7].文献[8]提出了服务推荐的概念,在服务组合过程中,根据开发者的偏好以及服务的非功能属性信息自动生成满足后续组合需求的候选服务并主动向开发者推荐.在对开发者偏好和服务非功能信息建模及分析的过程中,该方法使用了特定领域的本体库及规划技术,导致通用性和可扩展性不高.但是,其提出了服务发现的一种新模式,即由服务资源库或服务组合工具根据组合上下文为开发者提供可满足需求服务的推荐.然而,这些方法均是针对组件服务进行发现,且发现的过程没有结合服务的动态行为特征,导致服务资源使用效率及服务发现准确率均不高.

流程的相似性判定.文献[9]提出了流程模型变体(variant)的概念,给出了一种基于流程活动描述间语言学相似性的模型规约方法,以此进行流程变体的判定并作为流程间相似性的衡量.文献[10]讨论了一种结合流程结构信息来衡量流程相似性的方法,该方法基于经典的图匹配理论,首先将流程转换成有向图,通过编辑距离衡量有向图间的相似性,属于一种在语法层次上进行相似性比较的方法.文献[20]更进一步借鉴文本处理方法研究,给出了结合流程结构及语义的综合评估方法,具有较大的使用价值及参考意义.但是,上述方法均缺乏对流程动态行为特征的综合考虑.文献[4,5,11]则初步讨论了行为特征对流程相似判定的影响.文献[4]中,流程的行为特征使用其所表示的活动序列及消息交换序列描述,服务发现过程针对该序列进行,同时,结合活动的前序条件和执行结果作为相似性判定的准则;文献[5]则使用有向图来表示流程,将流程相似性判定转换为有向图的子图同构问题,并采用有向图间的编辑距离来衡量其相似程度,以此综合评估两个流程的行为相似性;文献[11]综合分析了轨迹等价性的实现,基于标记转移系统的研究成果,通过因果图轨迹向量描述流程的动态行为特征,并使用轨迹向量间的夹角余弦值来评估两个流程的相似性.上述方法要求所比较流程中的相似活动必须严格相同,即拥有同样的名称和描述,使其不适用于大多数服务发现场景;另一方面,上述方法的相似性评估过程仅考虑结构或动态行为因素的影响,缺乏综合评价的机制,而且以上基于行为特征的判定方法均只考虑了系统的某个状态、活动的行为,无法从整体上全面衡量其相似程度,因此相似性评估的准确性不高.本文将综合考虑流程的静态结构及动态行为信息,结合整个流程行为活动序列的影响给出流程状态间的量化相似性衡量以及状态的一对多关系评估,实现其模糊的综合度量,同时提高其准确性.

综上所述,Web 服务的自动发现及推荐技术研究已经受到广泛的关注,并取得了一定的成果.但目前的大多数工作侧重于组件服务资源的发现与推荐,缺乏一种针对更大粒度资源的发现机制以提高资源使用效率.另一方面,在大粒度的流程发现及兼容性检测研究领域,流程的相似性判定方法研究已得到广泛关注,然而大多数研究工作仅关注于流程的静态结构或动态行为中的某个方面的精确判定,缺乏综合的量化评估方法以提高流程发现的准确性.

2 AutoDisc 框架概述

在流程感知服务组合方法中,开发者通常使用BPMN,BPEL等流程建模语言进行组合服务编排,由业务分析人员形成的抽象需求模型或开发者在精化过程中形成的一系列组合流程实现均可作为服务发现的参考条件.如图 1 所示,图中上半部分为服务组合的上下文环境,包括编排中的组合流程结构、事件、活动等信息(图 1 中P_1部分),这些信息可以通过流程建模工具实时地抽取出来.图 1 的下半部分显示了一个维护服务描述信息的服务资源库及其中的组合流程信息.

在本文中,我们将讨论一个可复用组合服务流程及其片段的自动发现框架.图 1 中的CS_1,CS_2,CS_3等表示服务资源库中积累的组合流程实现;而M_1,M_2,M_3等则是其对应的描述模型,该模型用于屏蔽不同业务流程建模语言之间的差异,同时简化流程建模信息,以降低相似性判定的复杂度.在流程感知的服务组合过程中,AutoDisc框架将从流程建模工具实时抽取组合上下文,将其自动建模为服务描述模型,与服务资源库中的候选服务描述模型进行相似性匹配,可以最大程度满足组合需求的候选服务向开发者推荐,以指导其完成精化和开发过程,提高服务组合效率.图 1 中的M_3与M_2即AutoDisc所推荐的候选服务,开发者可根据M_3的实现直接完成图中P_2部分的编排.与完全人工地完成该部分的编排或者单一的组件服务发现方法相比,服务组合的效率得到了明显的提高.

为实现 AutoDisc 框架的上述功能,首要问题是定义一个适于描述组合服务流程结构及行为特征的基础描

述模型.在此基础上给出业务流程建模语言到该模型的自动映射方法,以支持服务的自动发现.另一方面,针对开发过程可获取信息不足及开发上下文中流程信息不完整的特征,服务发现过程中的流程相似性判定算法应支持两个组合流程间结构和行为特征相似程度的可量化判定,且支持流程与流程片段之间的相似性衡量.本文的第 3 节、第 4 节将对以上两个问题进行详细讨论.

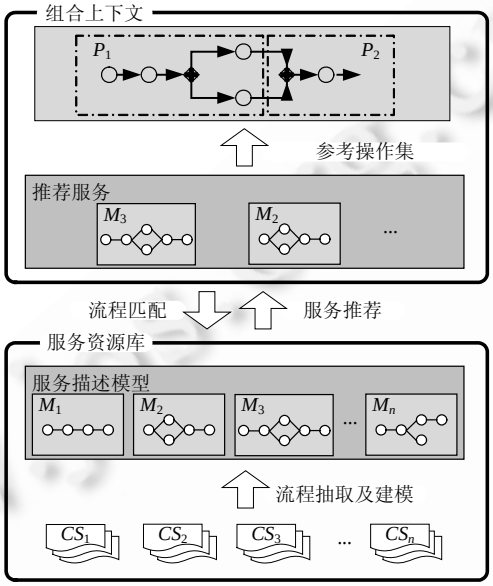


Fig.1 Scenarios of automatic service discovery in the process-aware service composition

图 1 流程感知服务组合中的自动服务发现应用场景

3 组合服务流程描述

一个组合流程的静态结构特征包括活动、事件、控制结构以及连接三者的顺序流等信息,而动态行为特征则通常包括其内部的控制流、数据流、交互协议以及状态迁移等方面[12].本文的前序工作[13]使用基于非确定有穷自动机(NFA)的描述模型来综合描述组合流程的结构和行为信息,在进一步的深入研究中我们发现,基于NFA的描述方法存在如下的缺点:(1) 由BPMN到该模型的映射算法在最坏情况下的时间复杂度为$O(n!)$,即在BPMN流程中并发流程分支数目较多的情况下,映射算法所消耗时间随分支数目呈阶乘级数增长,缺乏良好的扩展性;(2) 基于有穷状态机的组合服务流程描述模型仅能够描述系统处于某个执行点时的状态及动态行为特征,而无法从整体的角度全面刻画组合服务所处的系统状态及行为序列等特征,对其行为描述缺乏准确性.因此,本文采用基于Petri网的组合服务描述模型,Petri网模型支持并发语义且可有效地描述一个系统的静态控制结构及动态行为特征,在一定程度上克服了NFA模型的缺点.同时,为了从整体上刻画BPMN流程的内部状态及行为活动序列,本文借助Petri网的一种重要的分析模型——可达图(reachability graph)作为模型间相似性比较的基础.下面将针对组合服务的描述模型及其可达图给出形式化的定义.

3.1 基于Petri网的组合服务描述模型

基于 Petri 网的标识转移系统具有标识和转移两大特征,标识可以很好地描述组合服务的活动、事件、控制结构等静态属性,而转移则可支持组合服务动态行为的有效描述;同时,该系统也可支持模型间相似性的高效判定.因此,本文基于 Petri 网定义组合服务描述模型.

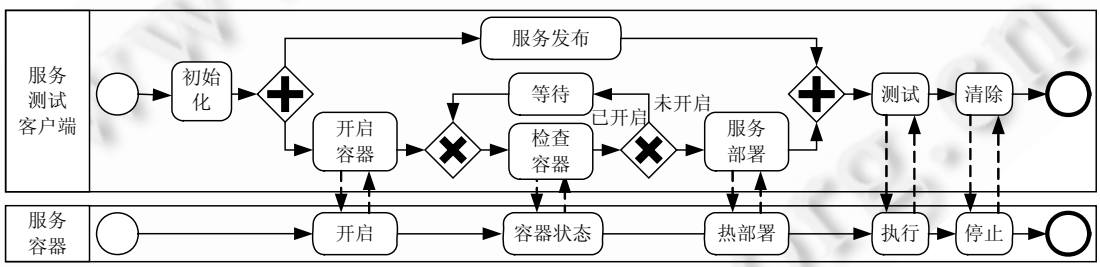
为便于本文的理解,我们回顾几个Petri网相关的概念.对于一个普通Petri网$N=(P,T)$,映射$M:P\to\{0,1,2,\ldots\}$称为N的一个标识,二元组(N,M)称为一个标识网.初始标识表示为m_0,终止标识集表示为M_F.令x为任意库所$p\in P$或

变迁$t\in T$,$^{\bullet}x$表示x的前驱集合,$x^{\bullet}$表示x的后继集合.通常,如果$^{\bullet}p=\varnothing$,称p为开始库所;如果$p^{\bullet}=\varnothing$,称p为终止库所.对于任意属于M的m,$m(p)$表示库所p中的标记数目.对于一个标识m,变迁$t\in T$是使能的当且仅当$\forall p\in{}^{\bullet}t:m(p)>0$ 成立,记为$m[t\rangle$.如果t是使能的,t可以触发并导致标识变迁到m',记为$m[t\rangle m'$.当存在一个触发序列$t_1t_2\ldots t_n$使得$m[t_1\rangle m_1[t_2\rangle\ldots m_{n-1}[t_n\rangle m'$,则称标识$m'$是从$m$可达的.$R(N,m)$表示从$m$可达标识的集合.在此基础上,我们给出组合服务描述模型的形式化定义.

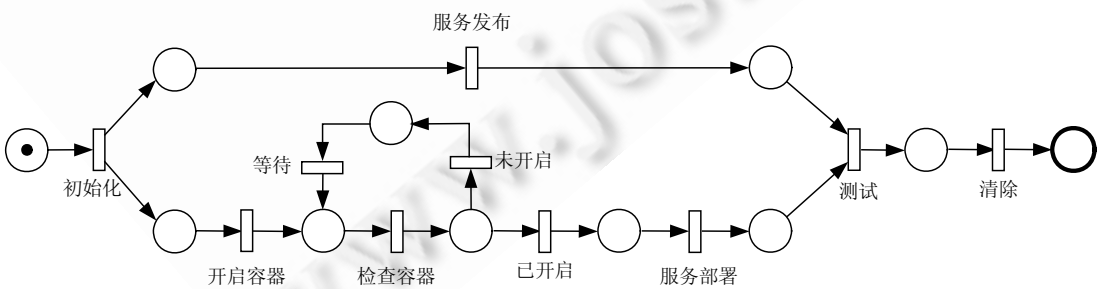
定义 1(基于 Petri 网的服务描述模型,简称 PSM). 基于 Petri 网的服务描述模型 PSM 是一个五元组$N=(P,T,F,I,O,\delta)$,其中:

- P 是库所的有限集合;
- T 是变迁的有限集合;
- $F\subseteq(P\times T)\cup(T\times P)$是库所与变迁之间弧的集合;
- $I\subseteq P$是输入库所集合,对于任意库所$p\subseteq I$有$^{\bullet}p=\varnothing$;
- $O\subseteq P$是输出库所集合,对于任意库所$p\subseteq O$有$p^{\bullet}=\varnothing$;
- $\delta:T\to\Sigma$是一个映射函数,其中,Σ是字母表,代表了组合服务运行期所表现的外在行为集合.针对每个变迁$t\in T$,δ对其进行标注,将其映射为原子服务的行为;将δ进行扩展定义为$T^*\to\Sigma^*$,则是将一个变迁序列映射为一个服务行为序列的函数.

图 2 给出了一个组合服务流程及其对应的 PSM 模型表示.其中,图 2(a)是一个用 BPMN 来建模的 Web 服务测试应用的流程.该流程包括两个参与方:图 2(a)上部描述了测试客户端的流程,拥有“初始化”、“服务部署”、“服务测试”等一系列业务活动;图 2(a)下部则描述了一个 Web 服务容器的内部流程特征.图 2(b)是图 2(a)中测试客户端的对应 PSM 模型表示.



(a) 一个 Web 服务测试应用的内部处理流程



(b) 测试客户端的对应 PSM 模型

Fig.2 Inner process of composite service and the corresponding PSM model

图 2 组合服务业务流程及其对应的 PSM 模型表示

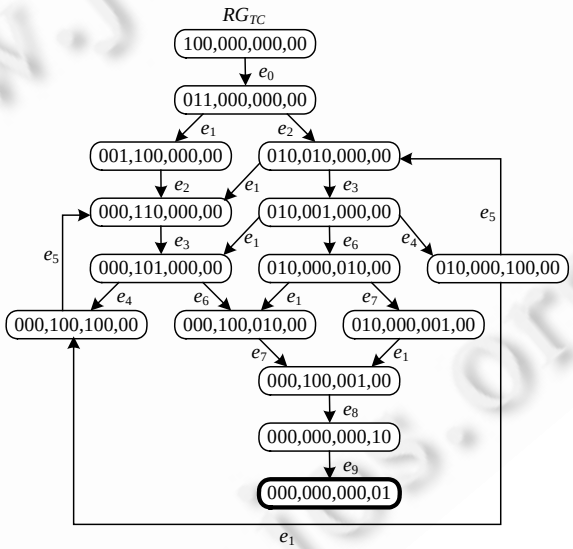
基于 PSM 模型的定义及基本 Petri 网相关概念及属性,PSM 的可达图形式化定义见定义 2.

定义 2(PSM的可达图,简称RG). 令$N=(P,T,F,I,O,\delta)$为一个组合服务的PSM,则其可达图$RG=(V,E,v_0,V_F,s)$为

一个有向标记图,其中:

- $V=R(N,m_0)$是节点的集合,其中每个节点对应N中从初始标识可达的一个标识;
- $E\subseteq V\times\Sigma\times V$ 是边的集合,其中每条边代表 N 从一个标识经过一个变迁到达另一个标识的过程;
- v_0是RG的根节点,对应N的初始标识m_0;
- V_F代表RG的终止节点集合,对应N的终止标识集合M_F;
- $s:E\to\Sigma$是边上的变迁标记函数,返回所对应的变迁名称.

在 RG 中,每个节点所对应的标识可以表示成一个$|P|$维的正整数向量,代表了系统当前所处的状态信息.其构造过程首先为 PSM 所包含的库所编号,标识向量每一位的取值等于当前状态时对应编号库所中所拥有的标记数目,如图 2(b)所示 PSM 状态仅有编号为 1 的库所拥有一个标记,因此所对应的 RG 标识节点取值应为(100,000,000,00).RG 中的边对应 PSM 中的每一个变迁,可通过对 PSM 模型的遍历分析来完成对应 RG 图的构造过程.图 3 给出了图 2 中测试客户端流程 PSM 模型所对应的可达图表示.这里,为了避免 BPMN 流程中具有相同名称任务对构造可达图时产生的非确定性影响,引入了可达图中边上的变迁标记函数.可达图中的每条边对应 PSM 模型中的一个变迁,使用可达图中特定的标记来描述,而其对应的变迁名称则通过变迁标记函数来获取.由图 3 可看出,可达图给出了一个 PSM 模型的所有可能运行状态及变迁发生序列的完整描述,同时也在一定程度上反映了模型的结构信息,如并发分支的切分与归并、循环等控制结构均得到了合理的描述.因此,通过对应可达图的比较来综合衡量两个组合服务流程间的相似性是合理且有效的.



$s(e_0)$=初始化; $s(e_1)$=服务发布; $s(e_2)$=开启容器; $s(e_3)$=检查容器; $s(e_4)$=未开启; $s(e_5)$=等待; $s(e_6)$=已开启; $s(e_7)$=服务部署; $s(e_8)$=测试; $s(e_9)$=清除.

Fig.3 Reachability graph of the testing client's PSM model

图 3 测试客户端流程 PSM 模型的对应可达图表示

3.2 BPMN到PSM的映射

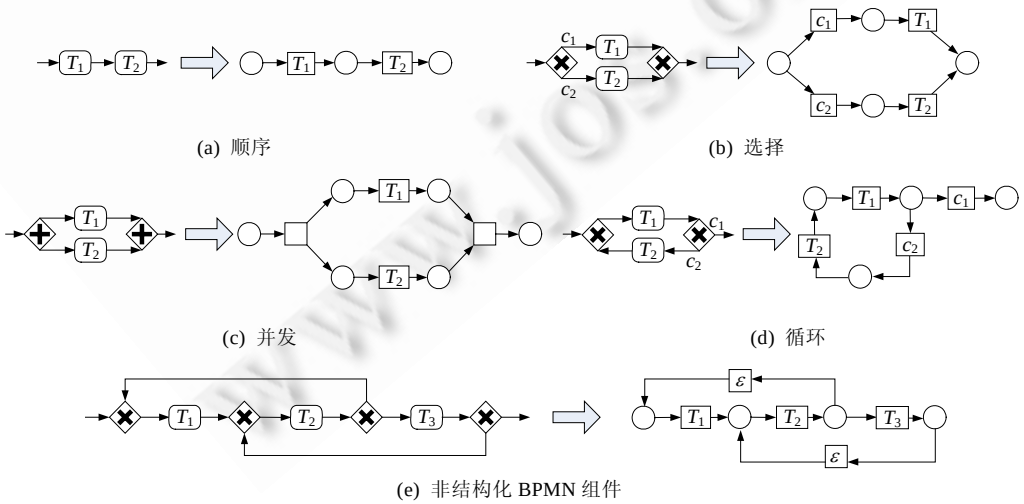
在面向服务的软件开发领域,业务流程建模标注语言(BPMN)已经成为一种通用的图形化描述语言,被广泛应用于组合服务的建模.与 BPEL 相比,BPMN 由于其具有图形化的特征与高度的抽象性,特别是能够描述多方参与的应用场景,因此更适合进行服务编排.在本文中,我们使用 BPMN 作为组合服务业务流程的建模语言,并提出一种 BPMN 到 PSM 模型的自动映射机制以支持组合服务流程相似性的判定.

为提高映射过程的效率,本文工作主要针对文献[14]所述的一个 BPMN 核心子集,该子集包括 BPMN 规范中用于描述流程控制的主要元素,可满足大部分组合流程建模的需求,同时,可以更有效地支持模型映射及相似

性判定算法.而对于使用该子集以外的元素进行建模的 BPMN 流程,则可以通过文献[15]所述的 BPMN 元素使用特征及建议来制定过滤和转换方法,以得到满足子集约束的流程.另一方面,由于 BPMI(BPMN 制定委员会)并未提供严格的 BPMN 规范理论基础,BPMN 到目前为止仍然没有确定的执行语义,致使通过不同 BPMN 建模工具开发的组合服务业务流程在结构和行为语义上存在较大的差别,任意的流程结构和语义表示导致由 BPMN 到 PSM 模型映射的复杂度难以满足实时服务发现框架的要求.为此,本文采用文献[14]所给出的良构 BPMN 核心模型对流程进一步约束,与 PSM 模型的映射以及整个 AutoDisc 框架将基于良构流程进行.

由良构 BPMN 核心流程到 PSM 模型的映射过程可分为两个部分:基础元素映射和控制结构映射.基础元素映射包括事件、任务、子流程、控制网关以及顺序流等元素的映射,其中:1) 任务或中间事件元素映射为 PSM 模型中连接输入库所和输出库所间的变迁,此变迁将被标记为该任务或事件的名称,用于描述任务或事件的执行;2) 开始或结束事件元素映射为 PSM 模型中的相同结构,而变迁将被标记为“Start”或“End”,用于标识流程的开始和结束;3) 子流程元素在映射过程中将首先被看做一个整体,其内部则展开成基本元素进行直接映射;4) 控制网关将映射为 PSM 模型中的库所;5) 一般顺序流元素映射为 PSM 模型中的空库所,而对于带有条件表达式的顺序流元素,则将映射为 PSM 模型中的条件判断变迁.

现有的业务流程建模语言通常支持 4 类基本的控制流模式:顺序(sequence)、选择(switch)、并发(parallel)以及循环(loop),图 4 给出了这几类基本控制流模式的 BPMN 表示到 PSM 模型之间的映射关系.其中,BPMN 中的顺序模式通过若干先后执行的任务来描述顺序的行为控制模型,这一模式映射为 PSM 模型中的一系列顺序的变迁及连接变迁的库所的集合,如图 4(a)所示;BPMN 中的选择模式则通过一对对应的选择切分/归并网关及其之间的任务来描述选择的行为控制模型,这一模式映射为 PSM 模型中由一个空库所触发,且开始于相应条件判断变迁的若干选择顺序流的集合,其中的每条顺序流包含对应选择分支中的任务变迁,如图 4(b)所示;BPMN 中的并发模式通过一对对应的并行切分/归并网关及其之间的任务来描述并发的行为控制模型,这一模式映射为 PSM 模型中由一个空变迁触发的若干并发顺序流的集合,其中的每条顺序流包含对应并发分支中的任务变迁,如图 4(c)所示;类似于选择模式的映射方法,图 4(d)中的 BPMN 循环模式映射为 PSM 模型由循环条件判断变迁切分的两条选择顺序流的集合.对于更一般的没有采用标准流程模式嵌套设计的非结构化 BPMN 流程,基于以上基本元素及控制结构的映射过程可能会出现库所直接相连的情况,这是 PSM 模型所不允许的.因此,本文进一步引入空变迁(ε变迁),当出现以上情况时,则在连接库所的弧上插入ε变迁,图 4(e)是该情况的一个实例及对应的 PSM 模型表示.



(a) 顺序 (b) 选择

(c) 并发 (d) 循环

(e) 非结构化 BPMN 组件

Fig.4 Mappings from basic control patterns of BPMN to PSM

图 4 基本控制流模式 BPMN 和 PSM 模型的映射

4 流程相似性判定算法

基于组合服务的 PSM 模型及其可达图表示,本文提出了一种新的组合流程相似性判定算法,该算法综合考虑流程的静态结构及运行时动态行为信息,通过可达图间的模拟关系来衡量两个组合流程的相似程度,并可有效地处理存在无限循环结构及非结构化 BPMN 流程的相似性判定问题.

4.1 流程相似性判定的准则

一个组合服务的行为特征是其在运行期所表现出来的动态功能集合,包括其所执行的任务序列及其中每个任务对整个序列的影响,是反映一个服务功能的重要方面.传统的流程相似性判定方法主要针对流程结构信息进行,通过流程中任务的类型、任务节点的出入度、任务的描述以及流程间的编辑距离等信息来综合度量流程的相似性[10,20].然而,由于组合服务所处的开放、动态网络化运行环境,导致其运行时行为具有很强的动态特征,单纯使用静态流程结构无法准确刻画一个组合服务的功能.我们通过一个实例来说明流程的结构与行为相似间的关系.图 5 显示了 3 个组合流程片段对应的PSM模型,其中,片段(a)与片段(b)完成了相同的任务序列,具有相似的功能,但是它们结构上并不相似:片段(a)拥有 4 个库所,而片段(b)则拥有 5 个库所;另一方面,片段(a)与片段(c)在结构上是相似的,拥有相等的库所数、变迁数及对应的库所出入度特征,然而它们分别执行了不同的任务序列,在行为特征上相差较大,具有不同的功能实现.由此可见,动态行为信息可更准确地描述一个组合服务的功能,组合流程的相似性判定方法应结合其结构与行为信息进行综合评定.

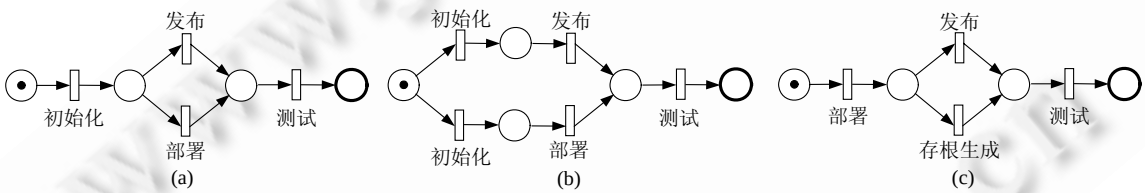


Fig.5 Structural and behavioral similarity between processes

图 5 流程的结构相似与行为相似间的关系

传统的进程代数研究提出了一系列评价标识转移系统等价性的方法,这些方法可划分为两大类别:1) 执行轨迹等价性(trace equivalence),如果两个系统可以执行相同序列的任务则是等价的.轨迹等价方法将系统作为一个黑盒来看待,考虑其所表现的外部行为的等价性;2) 模拟等价性(simulation equivalence)的基本思想是,如果两个系统所执行的任务序列可按逐个任务的执行来匹配,则这两个系统是等价的.模拟等价方法除了考虑系统外在的行为特征,还结合了系统运行时的内部状态信息,因此更能反映两个系统间的功能等价关系[16,17].基于以上分析,本文采用PSM模型可达图间的模拟关系衡量两个组合流程的功能相似性,与传统模拟等价方法不同,组合流程的相似性判定具有两个特殊要求:1) 判定结果应该是可量化的.AutoDisc框架面向组合服务的开发过程,服务资源库中的候选服务很有可能无法与待开发的组合服务完全匹配,此时,量化的相似性度量值对自动服务发现具有更大的意义.因此,本文提出了一种新的PSM模型可达图模拟关系判定算法,支持组合流程相似性的量化评估;2) 判定方法应支持两个组合流程间的“弱”模拟(weak simulation)关系,即执行序列中任务间的一对多匹配关系.在服务组合过程中,候选服务与待开发服务可能处于不同的抽象级别,任务间的严格一对一关系可能导致具有相似功能的服务匹配失败.因此,本文引入“停顿”(即空变迁ε)的概念来模拟评估过程中的一对多匹配情况,支持组合流程相似性的模糊评估.

4.2 可达图标识间的量化模拟评估

基于上述分析,PSM 模型可达图的模拟关系评估主要包含结构相似及行为状态模拟两个部分.为衡量两个可达图的结构相似特征,我们针对其节点和边分别定义相似性函数,其中,节点相似性函数 $N:V\times V\rightarrow[0,1]$通过入度与出度之间的差异来衡量两个节点的结构相似程度,定义如下:

$$N(s,t)=1-\frac{\|in(s)|-|in(t)\|+\|out(s)|-|out(t)\|}{|in(s)|+|in(t)|+|out(s)|+|out(t)|},s,t\in V \tag{1}$$

可达图中边的相似性函数$L:\Sigma\times\Sigma\to[0,1]$通过边上对应变迁标记间的语言学关系来衡量其相似程度.现有的衡量两个词或词组相似性的方法可以分为两大类,其中一类基于编辑距离,该方法主要关注两个词之间的语法差异,无法针对其所表达的功能相似性进行判定;而另一类则基于词典,该方法根据两个词之间的语义距离,即其在词典中的距离信息来衡量相似性,更适合于组合流程相似性判定的需求.本文借鉴广泛采用的WordNet[18]方法作为可达图中边相似性函数的基础,给出$L(a,b)$的形式化定义如下:

$$L(a,b)=-\log\frac{[\min_{c_1\in sen(a),c_2\in sen(b)}len(c_1,c_2)]}{2d_{\max}},a,b\in\Sigma \tag{2}$$

其中,$sen(a)$返回标记a在词典中所有可能释义的集合,$d_{\max}$表示词典类别树的最大深度,而$len(c_1,c_2)$则表示在词典类别树中释义c_1到c_2间的最短路径长度.这 3 个值均可通过WordNet提供的开放接口来获取.

在可达图节点及边的结构相似性函数基础上,结合可达图所表达的动态行为信息,本文提出一种量化的模拟关系衡量方法来综合评估可达图标识间的相似关系.该方法主要借鉴状态机模拟关系的定义及评估[11]:状态机中不同状态间的模拟是一种递归关系,两个状态是模拟的当且仅当它们可以通过相同的变迁转移到两个存在模拟关系的状态上,而两个终止状态间则可直接模拟.这是一个相对精确的匹配关系,为了实现流程相似性的可量化及模糊评估,本文采用标记相似性函数L作为变迁相似程度的度量,而针对终止状态,则采用节点相似性函数N来作为其相似性度量指标,最终给出两个可达图间标识对的量化模拟关系,定义如下.

定义 3(可达图中两个标识间的量化模拟关系Q_S). 令$RG_1=(V_1,E_1,v_{10},V_{1F},s_1)$和$RG_2=(V_2,E_2,v_{20},V_{2F},s_2)$为两个PSM模型对应的可达图,对于两个标识$v_1\in V_1,v_2\in V_2$,量化模拟关系$Q_S(v_1,v_2)$描述了$v_2$可模拟$v_1$的程度,其取值决定于它们所使能的变迁之间的相似程度以及经过该变迁所到达的后继标识之间的模拟关系.特别地,若$v_1\in V_{1F}$或$v_2\in V_{2F}$,那么$Q_S(v_1,v_2)=N(v_1,v_2)$.

根据Q_S的定义,我们给出其计算方法如下:

$$Q_S(v_1,v_2)=\begin{cases}N(v_1,v_2), & \text{if } v_1\in v_{1F} \text{ or } v_2\in v_{2F}\\ (1-p)\cdot N(v_1,v_2)+\dfrac{p}{n}\cdot\sum_{v_1[a\rangle v_1'}\max_{v_2[b\rangle v_2'}(L(a,b)\cdot Q_S(v_1',v_2')), & \text{otherwise}\end{cases} \tag{3}$$

其中,参数p用于衡量两个标识本身及其后继标识对Q_S影响的权重,即系统的整体行为状态特征对当前标识的影响;n则是v_1所使能的所有变迁的数目.由上式可以看出,在可达图中包含v_1的路径上,与v_1距离越远的标识和变迁对其量化模拟关系判定的影响越小,这也符合实际应用对流程相似性判定的要求.

4.3 可达图标识间的“弱”模拟关系

通过公式(3)可评估可达图中两个标识间的模拟程度,其不足是要求所比较的两个可达图标识和变迁之间满足一对一的匹配关系,当两个可达图中标识和变迁的数量不匹配时,即使完成相似的功能,通过公式(3)计算的Q_S值依然不高.考虑图 6 中的情况,图 6(b)中可达图所表示的流程片段是图 6(a)的一个精化.在开发过程中,抽象任务“编译”被进一步精化为两个具体的任务“创建目录”及“javac”.由于两个流程在结构上极为相似,无法根据结构信息有效衡量其功能相似性,因此在综合评价中,我们给予行为相似性判断更大的权重,将p取值为 0.8.此时,如果标记相似性函数取值L(“编译”,“创建目录”)=0,那么$Q_S(S_0,Q_0)$的计算结果为 0.36.很明显,这个结果与我们的直观理解相差甚远.然而,这种跨抽象级别间的流程相似性判定需求在实际应用中却是很重要的一个方面.我们知道,流程感知的服务组合方法通常包括业务建模及服务编排两个开发阶段.在这一过程中,开发人员需要面对从业务模型到执行流程的跨抽象级别精化,因此,支持从抽象功能到具体流程任务的自动发现是提高服务组合效率的重要方面.

在图 6 所描述的场景中,为进一步支持不同抽象级别间流程相似性的判定,当可达图的标识模拟关系评估进行到S_1与Q_1之间的比较时,应使S_1停顿一次,即不进行任何变迁,而使Q_1变迁到标识Q_2,尝试评估S_1与Q_2之间的模拟关系.以此类推进行后续标识的匹配,最终通过各种情况下模拟关系的综合评估来决定S_1与图 6(b)中哪些

标识之间存在模拟关系.图 6 中,计算终止时,S_2与Q_3的模拟衡量值较高,确定了两者间的模拟关系;同时,进一步得出S_1与Q_1、S_1与Q_2的模拟衡量值均较高,即S_1与Q_1,Q_2组成的片段间存在模拟关系.为实现这种特殊情况下的评估,本文参考经典状态机间模糊模拟的方法[19],引入"停顿"(即空变迁ε)的概念模拟评估过程中的一对多匹配情况.ε变迁是可达图变迁的一个扩展,当系统触发一个ε变迁时,不执行任何的动作,标识不发生转移,且满足$\varepsilon\cap\Sigma=\varnothing$.同时,我们扩展标记相似性函数来满足$\varepsilon$变迁与非空变迁之间的相似性衡量$L:(\Sigma\cup\{\varepsilon\})\times(\Sigma\cup\{\varepsilon\})\to[0,1]$.借助$\varepsilon$变迁,在每次进行标识模拟关系评估时,我们分别评估两个标识同时发生变迁或其中一个发生空变迁等情况,取其中的最大相似值作为标识间模拟关系的衡量.在评估过程中,ε变迁可以发生在所比较的两个可达图中,以支持可达图标识间的一对多及多对一匹配情况.同时,为了防止两个可达图中的标识同时触发ε变迁且导致相似性衡量值增加的情况,标记函数$L(\varepsilon,\varepsilon)$的取值应设为 0.综合以上扩展,我们基于公式(3)给出Q_S计算方法的形式化描述:

$$Q_S(v_1,v_2)=\begin{cases}N(v_1,v_2), & \text{if } v_1\in V_{1F} \text{ or } v_2\in V_{2F}\\ W_1+\max\left(W_2,\dfrac{p}{n}\cdot W_3\right), & \text{otherwise}\end{cases} \tag{4}$$

其中,

$$W_1=(1-p)\cdot N(v_1,v_2) \tag{5}$$

$$W_2=\max_{v_2[b\rangle v_2'} L(b,\varepsilon)\cdot Q_S(v_1,v_2') \tag{6}$$

$$W_3=\sum_{v_1[a\rangle v_1'}\max(\max_{v_2[b\rangle v_2'}(L(a,b)\cdot Q_S(v_1',v_2')),L(a,\varepsilon)\cdot Q_S(v_1',v_2)) \tag{7}$$

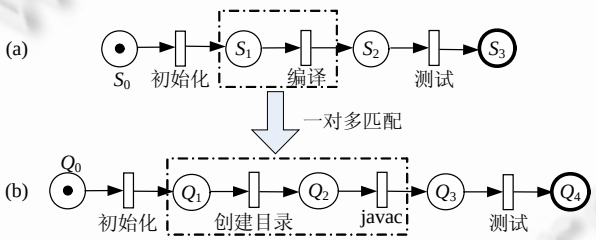


Fig.6 One-to-Many relationship between markings and transitions in two RGs

图 6 可达图标识及变迁之间的一对多匹配关系

4.4 可达图的模拟关系评估算法

通过公式(4),给定两个PSM模型的可达图并选定其中之一(称为"请求可达图")的某个标识,我们可以找到另一个可达图中与其存在最大模拟关系的标识.对于本文的应用场景,只需找出一个可最大程度模拟给定组合流程的候选服务,便可以对待开发流程提供进一步精化的参考.由于公式(4)的计算过程是沿着可达图中变迁执行序列的嵌套计算,因此从给定可达图的终止节点开始,按照节点与根节点距离从大到小的顺序计算节点间的Q_S值,可以有效地重用已有结果,提高计算效率.然而,当流程中存在循环回路时,上述计算过程由于存在递归调用,会产生相互引用而导致无法停机.例如,图 7 中所示的两个可达图,根据公式(4),$Q_S(v_0,v_0')$的计算过程依赖于$Q_S(v_1,v_1')$的取值,而$Q_S(v_1,v_1')$的计算过程由于存在变迁e_1及e_1',使其同时依赖$Q_S(v_2,v_2')$和$Q_S(v_0,v_0')$的取值,这样的循环过程导致递归调用无限执行而无法停机.为了处理上述情况,在评估可达图模拟关系前,应首先对其进行一个预处理,删除造成环路的有向边,计算所有标识对之间的模拟关系衡量值.然后,将所删除的边重新添加到原可达图中,对这些有向边所影响的计算结果进行修正,以此取得标识模拟评估的近似值.通过比较评估结果,可以得出与请求可达图中每个标识可取的最大Q_S值,该取值对应的标识则是另一个可达图中与其存在最大模拟关系的标识.最后,我们使用这一系列最大Q_S值的归一化线性均值来综合度量两个可达图间的模拟程度.计算方法见公式(8),其中综合考虑了流程内标识数目差别的影响.

$$RGSim(RG_1,RG_2)=\frac{\sum_{v\in V_1}\max_{v'\in V_2}Q_S(v,v')}{|V_1|}\cdot\left(1-\frac{\|V_1|-|V_2\|}{|V_1|+|V_2|}\right) \tag{8}$$

针对图 5 中的流程示例,我们初步比较文献[20]所述的结构化方法与本文的方法.其中,结构化相似性的计算结果为:*Structural_Sim*(*a*,*b*)=0.91,*Structural_Sim*(*a*,*c*)=0.922,流程 *a* 与流程 *c* 更相似;而通过可达图间标识相似性的计算及结构相差程度比较,公式(8)的计算结果为 *RGSim*(*a*,*b*)=0.937,*RGSim*(*a*,*c*)=0.855,流程 *a* 与流程 *b* 更相似,与经验分析的结果相符.通过以上比较可看出,本文综合结构及行为特征的算法可以更准确地衡量流程间的相似关系.与相关流程相似性评估方法的进一步比较将在本文实验与评估部分进行深入讨论.

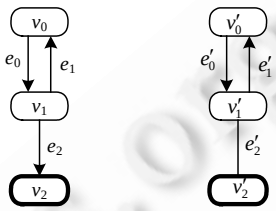


Fig.7 Simulation evaluation between two RGs with loop path

图 7 存在循环回路情况下的模拟关系评估

综合以上步骤,我们给出对可达图模拟关系评估的 ProSim 算法(见算法 1).该算法主要包括 6 个步骤:

1) 通过深度优先搜索得到可达图的深度优先生成树(第 2 行);
2) 对可达图中不在深度优先生成树上的边进行分析,若该有向边由一个树节点指向其祖先节点,则构成了可达图中的一个回环,将其从原可达图中删除(第 3 行~第 8 行);
3) 针对“去环”之后所得的请求可达图进行拓扑排序,得到标识模拟关系计算过程的依赖关系序列 *dependStack*,该序列使用堆栈式存储方式(第 9 行);
4) 按照拓扑排序的结果将标识顺序出栈,即从序列*dependStack*的尾部向头部逆向处理各标识,分别计算其与另一可达图中所有标识之间的Q_S值(第 10 行~第 15 行);
5) 将步骤 2)中删除的边重新添加到可达图中,对受其影响的标识Q_S值进行修正.修正的准则是:针对每条重新添加的有向边e_i,令其源标识与目标标识分别为v_i和v_j,固定v_j的所有Q_S值不变,重新计算v_i的Q_S值.并以此为基础,逐步沿可达图的拓扑排序序列向深度优先生成树的根节点递推,重新计算v_i的所有祖先节点对应标识与另一可达图中所有标识间的Q_S值(第 16 行、第 17 行及第 20 行~第 34 行);
6) 取出对应各标识的最大Q_S值,通过公式(8)计算两个可达图模拟关系的衡量值(第 18 行).

算法 1. 可达图模拟关系评估算法 ProSim.

输入:请求可达图RG_1及目标可达图RG_2;

输出:RG_2对RG_1的模拟程度衡量值$RGSim(RG_1,RG_2)$.

```
1   ProSim(RG1,RG2){
2       dfsTree1:=DFS(RG1), dfsTree2:=DFS(RG2);
3       for each e∈edgeSet(RG1)∧e∉edgeSet(dfsTree1) do
4           acyclicRG1:=RG1−e;
5           deletedEdges1:=deletedEdges1∪e;
6       for each e∈edgeSet(RG2)∧e∉edgeSet(dfsTree2) do
7           acyclicRG2:=RG2−e;
8           deletedEdges2:=deletedEdges2∪e;
9       dependStack1:=toplogySort(acyclicRG1), dependStack2:=toplogySort(acyclicRG2);
10      tmpStack1:=dependStack1, tmpStack2:=dependStack2;
11      for (i:=|vertexSet(RG1)|−1; i≥0; i−−) do
12          v1i:=tmpStack1.pop();
13          for (j:=|vertexSet(RG2)|−1; j≥0; j−−) do
```

```
14                  v2j:=tmpStack2.pop();
15                  verSim_{v1i,v2j} := Q_S(v1i,v2j) ;
16          ReviseRGSim(acyclicRG1,acyclicRG2,deletedEdges1,dependStack1,dependStack2);
17          ReviseRGSim(acyclicRG2,acyclicRG1,deletedEdges2,dependStack2,dependStack1);
18          return RGSim(RG1,RG2);
19    }
20    ReviseRGSim(revRG,passiveRevRG,deletedEdges,depStack_RevRG,depStack_PassiveRevRG){
21          for each e∈deletedEdges do
22                revRG:=revRG+e;
23                revStartVer:=e.sourceVertex(), revStarted:=false, tmpStack1:=depStack_RevRG;
24                for (i:=|vertexSet(revRG)|−1; i≥0; i−−) do
25                      vi:=tmpStack1.pop();
26                      if (vi!=revStartVer & revStarted==false) then continue;
27                      revStarted:=true;
28                      if (vi==revStartVer||vi∈ancestorVers(revStartVer)) then
29                            tmpStack2:=depStack_PassiveRevRG;
30                            for (j:=|vertexSet(passiveRevRG)|−1; j≥0; j−−) do
31                                  vj:=tmpStack2.pop();
32                                  if (e∈edgeSet(RG1)) then  verSim_{v1i,v2j} := Q_S(vi,vj) ;
33                                  else  verSim_{v1j,v2i} := Q_S(vj,vi) ;
34    }
```

定理 1. 算法ProSim是可终结的,其时间复杂度为$O(|V_1|\times|V_2|\times|E_1|\times|E_2|)$.

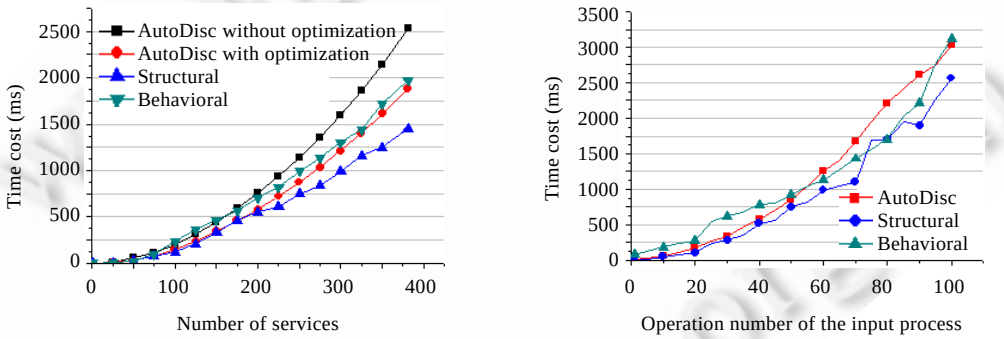
证明:算法ProSim用于评估两个PSM模型可达图间的模拟关系,返回衡量其模拟程度的量化值.显然,算法ProSim是可以终结的:首先,算法的前两个步骤是基于有向图的深度优先搜索算法,针对两个可达图的处理可以分别在$O(|V_1|+|E_1|)$和$O(|V_2|+|E_2|)$时间内终结;由于进行了“去环”处理,无环可达图的拓扑排序也可分别用$O(|V_1|+|E_1|)$和$O(|V_2|+|E_2|)$时间完成;同时,无环可达图中的标识对间模拟关系计算已不存在循环依赖,故算法的第 4)步中每个值的计算将在其中某个可达图达到最终状态时终止,需要针对两个可达图中的所有标识对分别进行处理,时间复杂度为$O(|V_1|\times|V_2|)$;算法第 5)步中,针对造成回环的各有向边进行修正处理,修正算法基于各标识间的已有模拟衡量值,不再进行递归计算,因此该步骤可在多项式时间内完成.在最坏情况下,需要对$|E_1|\times|E_2|$条边进行修正,每条边需要综合考虑$|V_1|\times|V_2|$个标识对的修正情况,因此,该步骤在最坏情况下的时间复杂度为$O(|V_1|\times |V_2|\times|E_1|\times|E_2|)$;算法第 6)步的计算过程可在常数时间复杂度内完成.综上所述,ProSim算法是可终结的,其时间消耗取决于第 5)步的时间复杂度,即$O(|V_1|\times|V_2|\times|E_1|\times|E_2|)$,可在多项式时间复杂度内进行两个可达图间模拟关系的量化评估.

5 实验分析与评估

本节通过采用真实数据集的仿真实验对AutoDisc框架的性能和有效性进行评估.为提高实验的可信度,我们采用从生物计算研究社区myexperiment(www.myexperiment.org)收集而来的381个生物信息学流程作为实验的数据集,该社区是供全球生物计算研究者们发布及共享生物计算流程及实验计划等信息的一个协同环境.为了与本文的研究前提和上下文一致,我们首先将收集到的生物计算流程使用BPMN语言进行建模,生成 381 个组合流程保存到服务资源库中,其中每个生物信息处理函数的调用均认为是一次组件服务调用过程.同时,为说明本文工作的有效性,在性能及准确性评估实验中,本文方法将分别与基于结构信息或行为信息的流程相似性判定方法中具有代表性的工作[11,20]进行比较,并给出直观的评估结果.仿真程序使用JDK1.6 和Eclipse3.4 开发,运行在一台CPU为Intel Xeon 3GHz,内存为 4G的Windows操作系统PC机上.

实验 1 和实验 2 对 AutoDisc 框架的时间开销进行了评估.AutoDisc 的性能主要受两个因素的影响:资源库中组合流程总数以及组合上下文中流程模型的规模,其中,实验 1 说明了 AutoDisc 的时间开销与资源库中组合

流程总数的关系(如图 8(a)所示).在该实验中,我们随机选取一个组合服务作为开发上下文中的组合流程,然后逐次抽取数目从 1~381 个的流程集合作为资源库中的侯选流程,针对每个集合大小分别进行 100 次随机抽取,共完成 38 100 次实验,取时间开销的平均值.由实验结果可看出,AutoDisc 的时间开销与服务资源库中流程总数呈线性增长关系.由于 AutoDisc 是针对组合上下文对资源库中的服务进行实时的自动匹配,考虑到资源库中组合流程的变化不大,AutoDisc 框架可以进行适当的优化:由后台模块预先完成 BPMN 到 PSM 模型的转换及其可达图的求解.优化后重新进行实验,得到图 8(a)中对应的圆点曲线.可以看出,优化后,AutoDisc 的整体时间性能得到明显的提高,且与相关基于结构或行为的方法性能处于同一量级.实验 2 说明了 AutoDisc 的时间开销与组合上下文中流程模型规模的关系.我们首先定义对组合流程所作的每一次编排为一个编辑操作,包括 BPMN 流程中事件、任务、控制网关和顺序流的插入、删除及修改等操作.为简化过程,在本文的实验中,以各元素的插入操作作为流程开发消耗的度量,删除及修改操作仅用于保证编辑操作定义的完整性,在此不作考虑.实验 2 中,流程模型的规模定义为完成该模型开发的编辑操作总数.实验过程使用 381 个流程作为候选流程集合,并使待开发流程规模从 1~100 逐步增长,每次实验选取具有相应规模的组合流程作为上下文流程,使用 3 种方法分别完成自动服务发现过程,每组实验做 100 次取时间开销的平均值.实验结果如图 8(b)所示.从图中可以看出,AutoDisc 的时间开销随着开发上下文中流程规模的增加呈线性增长,完成针对 100 次编辑操作规模的自动服务发现最大耗时为 3 035ms,这种量级的时间开销对于大多数服务发现场景是可以接受的.同时也可以看到,虽然综合了静态结构与动态行为的处理过程,AutoDisc 的时间消耗仍然保持了与结构或行为特征独立处理方法相近的水平,具有较好的可扩展性.



(a) 组合服务流程总数的影响 (b) 输入流程规模的影响

Fig.8 Performance of AutoDisc

图 8 AutoDisc 框架性能分析

实验 3 对比了是否采用AutoDisc情况下的流程开发效率.实验过程中,推荐服务的模拟关系阈值取值为 0.8.每次实验取出服务资源库中的一个组合流程,仿真程序依照该流程一步步模拟其开发过程,每一步进行一个元素的插入操作,完成每步操作后通过AutoDisc框架给出推荐的侯选服务列表.当所选出的组合流程第 1 次出现在候选服务列表中时,认为开发已完成,结束本次实验,记录当前所进行的编辑操作数;同时,计算该流程的完整操作数,作为未使用AutoDisc的情况下完成所选出组合流程开发所需的插入操作总数.图 9 显示了对 381 个流程进行实验并抽取操作数从 8~98 的 20 个流程样本的统计结果.由图可看出,没有使用AutoDisc框架的情况下,完成 381 个流程开发所需的平均编辑操作数为 44.5,而通过AutoDisc框架的自动服务发现,完成所有流程开发所需的平均编辑操作数为 10.9,开发效率提高了 75.5%.另一方面,仅考虑结构信息[20]或行为信息[11]的相似性判定方法的实验结果如图 9 中部的两条曲线所示,其完成所有流程开发所需的平均编辑操作数分别为 27.1 和 31.5.可见,它们需要依赖于更多的开发上下文信息来完成自动服务发现,准确性及发现效率均低于本文所提出的综合评估方法.最后,我们对AutoDisc框架完成了原型系统实现,图 10 展示了AutoDisc与BPMN建模工具集成后的运行效果.其中:上部分是该建模工具的主要操作区域,图中显示了一个开发中的组合流程;图的左下角则显示了

推荐服的务列表,该列表通过AutoDisc框架生成;当选中其中某个服务后,在主操作区域的下部将显示所选服务的流程结构,底部则是对应的详细属性及实现信息.开发者根据这些信息,可以直接重用所选服务的流程实现,迅速完成后续的开发过程,以进一步提高开发效率.

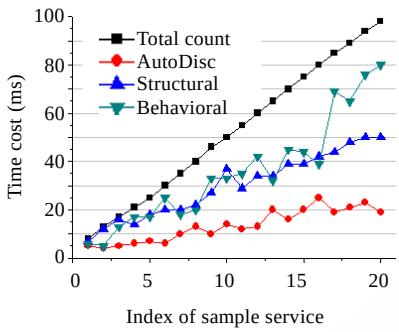


Fig.9 Effectiveness of AutoDisc

图 9 AutoDisc 框架有效性分析

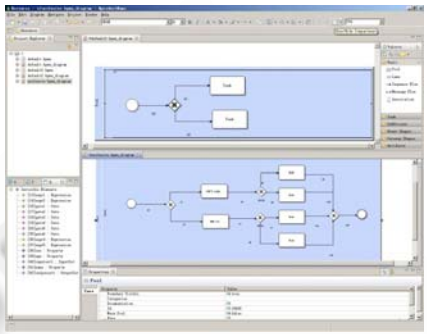


Fig.10 BPMN modoler with AutoDisc

图 10 集成 AutoDisc 框架的 BPMN 建模工具

6 结 论

本文重点讨论了通过自动的组合服务发现机制提高服务组合效率的问题.首先,为了避免服务发现过程要求开发人员过多参与导致自动化程度低下的问题,提出了一种自动化需求建模和服务推荐方法,即自动服务发现框架 AutoDisc.该框架针对广泛采用的流程建模语言 BPMN,定义了一个基础描述模型 PSM 并提出了由 BPMN 到该模型的自动映射算法,以支持开发过程中服务发现需求的自动获取,并作为流程粒度服务发现的基础.其次,针对服务发现的准确性问题,给出了一种综合流程静态结构及动态行为信息的可量化相似性评估方法,该方法通过 PSM 模型的可达图来描述流程的结构及整体行为状态,使用可达图标识的模拟关系来综合度量两个流程的相似程度,并给出量化的衡量指标以及流程相似性评估算法,以支持服务发现中具有相似功能流程的模糊匹配.最后,利用真实应用数据集进行相关的评估实验,证实了本文方法的有效性及实用性.在下一步的工作中,我们将从如何根据流程模型间的对应关系为开发者提供进一步开发的参考操作等方面继续开展研究.

致谢 在此,我们向对全体参与"可信的国家软件资源共享与协同生产环境"课题研发工作的科研人员,尤其是北京航空航天大学课题组参与面向服务软件生产线子课题研发的师生表示感谢.同时,感谢审稿专家对本文工作提出的宝贵意见.

References:

[1] Zhang LJ, Zhang J, Cai H. Services Computing. Beijing: Tsinghua University Press, 2007. 3−18.

[2] Damodaran S. B2B integration over the Internet with XML-rosettanet successes and challenges. In: Proc. of the World Wide Web Conf. (WWW). 2004. 188−195. [doi: 10.1145/1013367.1013398]

[3] OMG Group. Business process modeling notation. Vol.1.1. 2008. http://www.omg.org/spec/BPMN/1.1/

[4] Shen ZN, Su JW. Web service discovery based on behavior signatures. In: Proc. of the Int'l Conf. on Services Computing (SCC). 2005. 279−286. [doi: 10.1109/SCC.2005.107]

[5] Grigori D, Corrales JC, Bouzeghoub M. Behavioral matchmaking for service retrieval. In: Proc. of the Int'l Conf. on Web Services (ICWS). 2006. 145−152. [doi: 10.1109/ICWS.2006.37]

[6] Sirin E, Parsia B, Hendler J. Composition-Driven filtering and selection of semantic Web services. In: Proc. of the American Association for Artificial Intelligence Spring Symp. on Semantic Web Services. 2004. 42−49.

[7] Birukou A, Blanzieri E, D'Andrea V, Giorigini P, Kokash N. Improving Web service discovery with usage data. In: Proc. of the IEEE Software. 2007. 47−54. [doi: 10.1109/MS.2007.169]

[8] Bova R, Paik HY, Hassas S, Benbernou S, Boualem B. WS-Advisor: A task memory for service composition frameworks. In: Proc. of the Computer Communications and Networks. 2007. 535−540. [doi: 10.1109/ICCCN.2007.4317874]

[9] Lu RP, Sadiq S. On managing process variants as an information resource. In: Proc. of the Business Process Management (BPM). 2006. 426−431.

[10] Corrales JC, Grigori D, Bouzeghoub M. BPEL processes matchmaking for service discovery. In: Proc. of the Int'l Conf. on Cooperative Information Systems (CoopIS). LNCS 4275, Springer-Verlag, 2006. 237−254.

[11] Mendling J, van Dongen B, van der Aalst W. On the degree of behavioral similarity between business process models. In: Proc. of the CEUR-Workshop. 2007. 39−58.

[12] Bultan T, Fu X, Hull R, Su JW. Conversation specification: A new approach to design and analysis of e-service composition. In: Proc. of the World Wide Web Conf. (WWW). 2003. 403−410. [doi: 10.1145/775152.775210]

[13] Huang ZC, Huai JP, Sun HL, Liu XD, Li X. BestRec: A behavior similarity based approach to services recommendation. In: Proc. of the IEEE World Congress on Services-I (SERVICES-I). 2009. 46−53. [doi: 10.1109/SERVICES-I.2009.45]

[14] Ouyang C, Dumas M, Hofstede AHM ter, van der Aalst WMP. From BPMN process models to BPEL Web services. In: Proc. of the Int'l Conf. on Web Services (ICWS). Chicago: IEEE Computer Society, 2006. 285−292. [doi: 10.1109/ICWS.2006.67]

[15] Muehlen M zur, Recker J. How much language is enough? Theoretical and practical use of the business process modeling notation. In: Proc. of the Conf. of Advanced Information Systems Engineering (CAiSE). Montpellier: Springer-Verlag, 2008. 465−479. [doi: 10.1007/978-3-540-69534-9_35]

[16] Nejati S, Sabetzadeh M, Chechik M, Easterbrook SM, Zave P. Matching and merging of statecharts specifications. In: Proc. of the Int'l Conf. on Software Engineering (ICSE). 2007. 54−64. [doi: 10.1109/ICSE.2007.50]

[17] Grahne G, Kiricenko V. Process mediation in an extended roman model. In: Proc. of the Int'l Workshop on Mediation in Semantic Web Services (MEDIATE). 2005. 17−33.

[18] Pedersen T, Patwardhan S, Michelizzi J. WordNet: Similarity-measuring the relatedness of concepts. In: Proc. of the Int'l Conf. on Artificial Intelligence (AAAI). 2004. 1024−1025.

[19] Namjoshi KS. A simple characterization of stuttering bisimulation. In: Proc. of the Int'l Conf. on Foundations of Software Technology and Theoretical Computer Science. 1997. 284−296.

[20] Günay A, Yolum P. Structural and semantic similarity metrics for Web service matchmaking. In: Proc. of the Int'l Conf. on Electronic Commerce and Web Technologies (EC-Web). LNCS 4655, Springer-Verlag, 2007. 129−138. [doi: 10.1007/978-3-540-74563-1_13]

黄子乘(1981－),男,广西贵港人,博士,主要研究领域为服务计算,软件设计与生产.

李翔(1978－),男,博士,主要研究领域为服务计算,软件设计与生产.

怀进鹏(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机软件与理论,网络计算技术,信息安全.

朱蒋俊(1986－),男,硕士,主要研究领域为服务计算,软件设计与生产.

刘旭东(1965－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为可信网络计算技术,中间件技术.