doi:10.3969/j.issn.1671-637X.2017.01.008

基于赋权网络优化聚类的服务识别算法研究

李琳琳, 郑燕山, 焦 阳
(火箭军工程大学,西安 710025)

摘 要: 针对服务识别算法中聚合度、耦合度等重要指标的优化问题,基于业务流程间的关联度模型,运用网络拓扑聚类算法,引入聚类邻接参数,以聚合度-耦合度为优化目标函数,提出基于赋权网络优化聚类的服务识别算法。结合具体案例,应用Matlab软件进行仿真分析,根据聚合度-耦合度优化模型,选择不同邻接参数取值下的最优聚类效果,验证了该算法在服务识别上的有效性。

关键词: 服务识别; 关联度模型; 邻接参数; 聚合度-耦合度; 有效性
中图分类号: TP393.02 **文献标志码**: A **文章编号**: 1671-637X(2017)01-0033-04

A Service Identification Algorithm Based on Weighed Network Optimized Clustering

LI Lin-lin, ZHENG Yan-shan, JIAO Yang
(Rocket Force University of Engineering, Xi'an 710025, China)

Abstract: Considering the optimization of such indexes as convergence and coupling degree in service identification algorithm, we proposed a service identification algorithm based on Weighed Network Optimized Clustering (WNOC) by using the relational degree model and the method of network topology clustering, introducing the adjacency parameter of clustering, and taking the coupling-convergence degree as the optimized objective function. Simulation was made with Matlab to a certain example, and the optimal clustering result was selected according to the coupling-convergence degree optimized model. The result proves the validity of the method in service identification.

Key words: service identification; relational degree model; adjacency parameter; convergence-coupling degree; validity

0 引言

服务识别可看作是对一组业务模型进行分解、再合并和抽象操作,得到一组满足特定原则服务的过程,主要分为业务建模、模型分解、服务抽象、粒度设计等方面[1],是实现面向服务系统设计的关键。主要任务是确认系统中有哪些服务,并根据服务设计的原则、设计服务优化算法进行优选,获得满足服务特性的最优候选集。

针对服务识别目的的聚类分析,许多学者提出了不同的聚类和优化方法。文献[2]以业务实体作为样本点集合,以业务实体间的关联强度作为关联值,计算各实体与样本点间的权值,并通过权值聚类实现最终划分,但并未对实体间的关联度计算进行讨论;文献[3]提出了一种基于无向有权网络的聚类算法,该算法将空间点聚类问题转化为网络划分问题,但未提出怎么将其运用于服务识别;文献[4]对经典的基于密度聚类算法进行改进,在建立的网络图基础上进行服务聚类识别,但未考虑服务之间的关联程度。

在已有算法的基础上,本文基于服务识别中的两个重要指标聚合度和耦合度,建立了聚合度-耦合度优化模型,并将此模型应用于服务识别聚类算法。在将业务流程转化为有向赋权网络图的基础上,通过对抽象的网络图模型进行优化聚类分析,达到识别高内聚、松耦合服务的目的。

1 赋权网络图构造过程

根据业务流程各对象间的关系计算对象间的关联

收稿日期:2015-10-20 **修回日期**:2016-09-12
基金项目:国家"八六三"计划重点资助课题(2011701AA221)
作者简介:李琳琳(1974—),女,辽宁营口人,博士,副教授,研究方向为信息服务体系建模与仿真。

度,并以此为基础构建关联度矩阵。而后,将各对象看作是赋权网络图中依序排列的节点,关联度看作是网络图中节点间边的权重,构建赋权网络图。

1.1 相关图理论

根据文献[5]中相关图理论,一个赋权网络图是一个有序的三元组 $<V,E,W>$,记作 D,其中,$V\neq\varnothing$,称为 D 的顶点集,其元素称为顶点,E 为边集,W 为边权重的集合。设 $\boldsymbol{A}$ 为图 D 的邻接矩阵,对于一个赋权网络图,它的邻接矩阵 $\boldsymbol{A}$ 中的元素为 a_{ij},如果节点 i 和 j 之间存在一条赋权边,权值为 w_{ij},则 $a_{ij}=w_{ij}$,否则,$a_{ij}=0$。

1.2 关联度

网络图邻接矩阵元素表示为各业务流程间关联度的大小,对应网络图中赋权边权重。

业务实体间关联度可分为静态关联度和动态关联度。静态关联主要是指对象之间的泛化、组合、聚集、关联、依赖等关系,而动态关联则是指用例或对象之间的使用与调用关系[1]。目前,关于关联度计算方法研究也有很多,本文采用文献[6]中提出的根据 UML 模型类图和用例图,进行业务实体间关联度计算的方法。

业务实体 c_i 和 c_j 的静态关联度值为

$$R_{Sij}=\sum_{y\in Y}(U_{yi}\cdot U_{yj}\cdot W_y) \tag{1}$$

式中:Y 为实体间关系的集合;U_{yi} 为二值变量,如果关系 y 中包含实体(样本点)c_i,则 U_{yi} 设为 1,反之,设为 0;U_{yj} 与 U_{yi} 的取值方法相同;W_y 为关系 y 所占的权重,根据关系 y 的紧密程度,即 UML 实体类之间的关系种类而设定。

UML 中各实体类之间的关系强弱顺序为泛化 > 组合 > 聚合 > 关联 > 依赖。因此,可设定泛化关系 $W_y=5$,组合关系 $W_y=4$,聚合关系 $W_y=3$,关联关系 $W_y=2$,依赖关系 $W_y=1$。

业务实体 c_i 和 c_j 的动态关联度值为

$$R_{Dij}=\sum_{z\in Z}(V_{zi}\cdot V_{zj}\cdot W_z) \tag{2}$$

式中:Z 为用例的集合;V_{zi} 为二值变量,如果用例 z 使用了实体 c_i,则 V_{zi} 设为 1,反之,设为 0;V_{zj} 与 V_{zi} 的取值方法相同;W_z 为用例 z 所占的权重,可根据重要程度进行设定。

将 c_i 与 c_j 的静态关联度值与动态关联度值相加,即得到总关联度值为

$$R_{ij}=R_{Sij}+R_{Dij}\ 。 \tag{3}$$

1.3 网络图构造

对任何业务流程,根据各对象间的关联关系,可建立如图 1 所示的赋权网络图表现形式。图中,a_1,a_2,…,a_9 分别代表各业务实体,具体数值代表业务流程间关联度值大小。

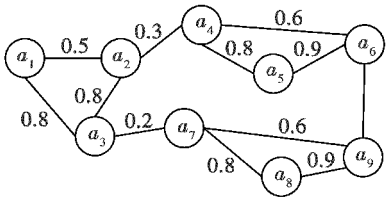

图 1 赋权网络图

Fig. 1 The figure of weighed network

2 基于赋权网络聚类的服务识别算法

2.1 算法流程设计

基于赋权网络聚类优化的服务识别算法的流程设计如图 2 所示。

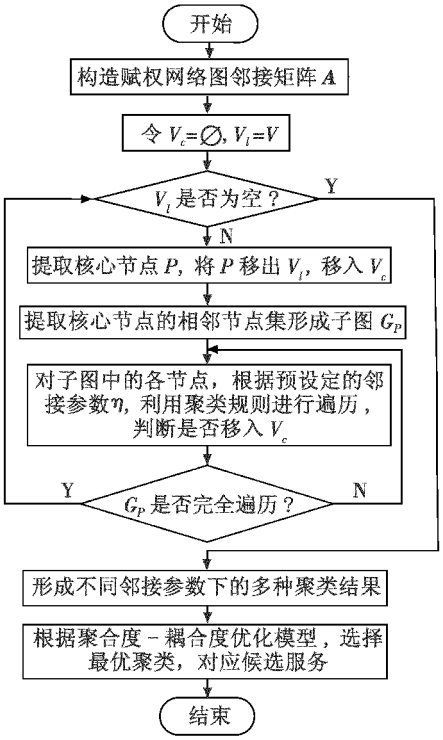

图 2 服务识别算法流程图

Fig. 2 The flow chart of service identification algorithm

2.2 算法描述

为了便于对算法进行描述,先进行以下定义。

定义 1 网络节点度定义为网络图中与该节点连接的其他节点数目的总和。

定义 2 网络节点 c_i 关联度强度 $\boldsymbol{S}_{(G,i)}$ 定义为此节点与其余所有相邻节点关联度大小的总和。对应于图的邻接矩阵表示为

$$\boldsymbol{S}_{(G,i)}=\sum_{i=1}^{j}a_{ij}\ 。 \tag{4}$$

定义 3 定义 η 为网络图拓扑聚类过程中的邻接参数,η 值的大小决定了从核心节点开始扩展,能够聚类得到的完整节点簇的大小。

定义 4　网络节点 p 聚合度 d_p 定义为节点 p 和其相邻节点组成的子图 G_p 中，链路边数 $|E_p|$ 与节点总数 $|V_p|$ 的比值[7]，即

$$d_p = \frac{|E_p|}{|V_p|} 。 \tag{5}$$

在本文研究中，聚合度主要指的是聚类形成的各节点簇内部节点间的联系程度，反映的是节点簇内部结构的稳定性。结合赋权网络知识，将节点簇聚合度 D 定义为簇中链路边权值的平均值，即

$$D = \frac{\sum w}{n} \tag{6}$$

式中，n 为簇中链路边数。D 值越大，表示簇内部各业务实体间的联系越紧密，内部结构越稳定。

考虑服务识别聚类结果的两个重要衡量指标，即聚合度和耦合度，好的聚类效果应该是同类之间具有更高的聚合度（即内聚度），类和类之间具有较低的耦合度。基于此，本文建立如下的优化模型

$$P = \alpha \frac{\sum D}{m} - \beta Y \tag{7}$$

式中：m 为聚类簇的个数；α，β 分别为聚合度、耦合度在聚类效果衡量中所占的权重，具体权重大小可结合层次分析法等确定（限于篇幅原因，在这里不详细介绍）。P 值表示综合的聚类效果，值越大，效果越理想。

基于上述相关定义和聚合度－耦合度优化模型，算法流程为：

1）构造赋权网络图的邻接矩阵 $\boldsymbol{A}$，邻接矩阵元素值为对应各业务流程间关联度的归一化值，即 $a_{ij} = R'_{ij}$，令 $V_c = \varnothing$，$V_l = V$；

2）计算 V_l 中所有节点的关联度强度 S_i，访问关联度强度最大的节点作为第一个特征节点，以它作为初始类，若强度最大的节点不唯一，则依次抽取各最大值节点为核心节点，将核心节点 p 移出 V_l，移入 V_c；

3）提取核心节点的相邻节点集 V_p 和边 E_p，形成子图 G_p；

4）分别计算子图 G_p 中各节点新的关联度强度 $S_{(c_p,i)}$，判断若对于同一个节点 c_i，存在 $1 - \dfrac{S_{(c_p,i)}}{S_{(c,i)}} \geqslant \eta$（$\eta$ 进行多值设定），则将该节点移入 V_c，重复 4），直至遍历 G_p 中所有节点；

5）转 2），直到遍历图中所有节点，即 $V_l = \varnothing$；

6）根据聚合度－耦合度优化模型，对不同 η 取值下的多种聚类结果进行优化选择；

7）将优化选择后的网络拓扑聚类节点簇收缩为一个节点，对应于一个高内聚、松耦合的候选服务。

2.3　算法时间复杂度

算法的时间复杂度，也就是算法的时间度量，记作 $T(l) = O(f(l))$，它表示随问题规模 l 的增大，算法执行时间的增长率与 $f(l)$ 的增长率相同，其中 $f(l)$ 为问题规模 l 的某个函数。

在本文提出的服务识别算法中，执行一次邻接参数，最好的情况下，即核心节点的相邻节点均与其处于同一个节点簇，算法的执行次数为网络图中节点数，复杂度为 $O(l)$；最坏的情况下，即网络图中所有节点均为孤立节点，算法的执行次数为节点数的平方，复杂度为 $O(l^2)$。

3　实验仿真与分析

为了检验算法的有效性，结合图 1 中构造的赋权网络图进行实验仿真分析。

3.1　参数选取

邻接参数 η 设定范围为 $[0,1]$，在程序运行时根据实际需求进行选取，本例中选取 1/6，1/9 两个邻接参数进行仿真运算；聚合度－耦合度优化模型中的权重参数均取值 0.5，即认为在衡量聚类效果中，聚合度和耦合度具有相当权重。

3.2　算法运用

根据选取的参数，运用 Matlab 软件进行仿真运算，得到的聚类结果如表 1 所示，其中，(a_i, a_j, a_k) $(i, j, k \in 1, 2, \cdots, 9)$ 表示一个聚类簇。

表 1　不同参数下的聚类结果

Table 1　The clustering results of different parameters

聚类结果	聚类簇 1	聚类簇 2	聚类簇 3
1	(a_9, a_8, a_7)	(a_6, a_5, a_4)	(a_3, a_2, a_1)
2	(a_9, a_8, a_7, a_6)	(a_3, a_2, a_1)	(a_5, a_4)
3	(a_6, a_5, a_4, a_9)	(a_3, a_2, a_1)	(a_8, a_7)
4	(a_3, a_2, a_1, a_7)	(a_6, a_5, a_4, a_9)	(a_8)
5	(a_3, a_2, a_1, a_7)	(a_9, a_8, a_6)	(a_5, a_4)

以聚类结果 1 为例，说明算法中聚合度－耦合度优化模型的运用。聚类效果如图 3 所示。

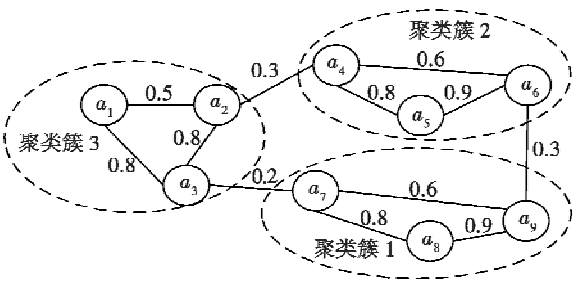

图 3　聚类效果图

Fig. 3　The clustering results

聚类簇1中，a_7，a_8，a_9 三节点的链路边权值分别为0.8，0.9，0.6，三者的平均值为0.77，因此聚类簇1的聚合度为0.77，可知，聚类簇2、聚类簇3的聚合度分别为0.77，0.70。可得，此种聚类情况下，整体聚合度为0.747。由聚类簇1、聚类簇2、聚类簇3构成的新的网络连接图的耦合度为3条连接边权值的平均值，即0.3，0.3，0.2的平均值0.27。总体效果衡量值为聚合度、耦合度值的加权平均值0.238。

同理，可得其他聚类情况下的聚合度、耦合度、总体效果衡量值，如表2所示。

表2 不同聚类结果比较

Table 2 Comparison of different clustering results

聚类结果	聚合度	耦合度	总体效果衡量值
聚类结果1	0.747	0.27	0.238
聚类结果2	0.717	0.42	0.148
聚类结果3	0.717	0.42	0.148
聚类结果4	0.408	0.72	-0.156
聚类结果5	0.658	0.58	0.039

由表2可知，聚类结果1的聚类效果最好，分别将聚类结果1中的聚类簇1、聚类簇2、聚类簇3抽象简化，即得到图4所示业务流程服务识别后的表示图。

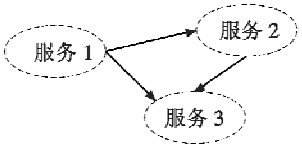

图4 服务识别表示图

Fig. 4 The representing of service clustering

3.3 算法运用

对使用赋权网络聚类优化算法前后的网络聚合度、耦合度进行比较，结果如图5、图6所示。

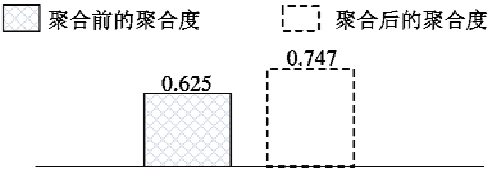

图5 聚合前后聚合度比较

Fig. 5 The covergence degree before and after clustering

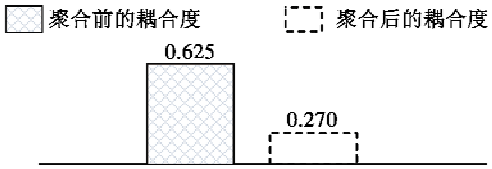

图6 聚合前后耦合度比较

Fig. 6 The coupling degree before and after clustering

可以看出，聚合后网络的聚合度明显增强，耦合度明显降低，即节点簇内部联系更加紧密，而簇与簇之间耦合性有明显改善。

4 结束语

本文提出了一种可用于服务识别的赋权网络聚类优化算法，首先将复杂业务流程转化为赋权网络图，通过聚合度－耦合度优化模型，对网络图聚类后的多种聚类结果进行优化选择，达到识别高聚合－松耦合服务的目的，并结合具体应用案例，进行了仿真分析，为解决服务识别问题提供了一种新的研究思路，具有一定的实用性。但是，从算法复杂度、实验效果等方面考虑，还有较大的进步与完善空间。

参 考 文 献

[1] 王庸豪. 面向业务层次的服务识别方法[D]. 合肥：合肥工业大学，2010.

[2] HEMANT J, CHALIMEDA N, IVATURI N, et al. Business component identification——A formal approach[C]//Proceedings of the 5th International Enterprise Distributed Object Computing Conference, 2001：183-187.

[3] 金敏. 一类基于无向有权网络的聚类算法研究[D]. 杭州：浙江理工大学，2013.

[4] 管清波，冯书兴. 天基信息服务体系与作战应用[M]. 北京：国防工业出版社，2014.

[5] 耿素云. 集合论与图论(二分册)[M]. 北京：北京大学出版社，1998.

[6] 徐玮，尹宝林，李昭原. 企业信息系统业务构件设计研究[J]. 软件学报，2003，14(7)：1213-1220.

[7] 杜胜永，柴乔林，王华. 基于节点聚合度的生成簇算法[J]. 计算机应用，2006，26(4)：948-950.