计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20120776
51(2):408-419,2014

一种面向主题的领域服务聚类方法

李 征[1,2] 王 健[2] 张 能[2] 李 昭[2] 何成万[3] 何克清[2]
[1](河南大学计算机与信息工程学院 河南开封 475001)
[2](软件工程国家重点实验室(武汉大学) 武汉 430072)
[3](武汉工程大学计算机科学与工程学院 武汉 430073)
(zhengli_hope@whu.edu.cn)

A Topic-Oriented Clustering Approach for Domain Services

Li Zheng[1,2], Wang Jian[2], Zhang Neng[2], Li Zhao[2], He Chengwan[3], and He Keqing[2]
[1](*School of Computer and Information Engineering, Henan University, Kaifeng, Henan* 475001)
[2](*State Key Laboratory of Software Engineering (Wuhan University), Wuhan* 430072)
[3](*School of Computer Science and Technology, Wuhan Institute of Technology, Wuhan* 430073)

Abstract With the development of SOA and SaaS technologies, the scale of services on the Internet shows a trend of rapid growth. Faced with the abundant and heterogeneous services, how to efficiently and accurately discover user desired services becomes a key issue in service-oriented software engineering. Services clustering is an important technology to facilitate services discovery. However, the existing clustering approaches are only for a single type of service documents, and they do not consider the domain characteristic of services. To avoid these limitations, on the basis of domain-oriented services classification, this paper proposes a services clustering model named as DSCM based on probability and domain characteristic, and then proposes a topic-oriented clustering approach for domain services based on the DSCM model. The proposed clustering approach can cluster services described in WSDL, OWL-S, and text, which can effectively solve the problem of single service document type. Finally, experiments are conducted on real services from ProgrammableWeb to demonstrate the effectiveness of the proposed approach. Experimental results show that the proposed approach can cluster services more accurately. Compared with the approaches of classical latent Dirichlet allocation (LDA) and *K*-means, the proposed approach can achieve better in the purity of cluster and *F*-measure, which can greatly promote on demand services discovery and composition.

Key words services clustering; latent Dirichlet allocation; topic; probability; feature dimension reduction

摘 要 随着互联网上服务资源规模的快速增长,如何高效、准确地发现服务成为一个亟待解决的关键问题.服务聚类是促进服务发现的一种重要技术.但是,现有服务聚类方法只对单一类型的服务文档进行聚类,并且没有考虑服务的领域特性.针对该问题,在对服务进行领域分类的基础上,提出了一种基于概率、融合领域特性的服务聚类模型——领域服务聚类模型(domain service clustering model, DSCM),然后基于该模型提出了一种面向主题的服务聚类方法.最后通过 ProgrammableWeb 网站提供

收稿日期:2012-09-03;**修回日期**:2013-06-19
基金项目:国家自然科学基金项目(61202031,61100017,61100018,61272115);国家科技支撑计划基金项目(2012BAH07B01);国家云计算示范工程项目(中小企业管理云应用研发与产业化);中央高校基本科研业务费专项资金项目(201121102020004,2012211020201)
通信作者:王 健(jianwang@whu.edu.cn)

的真实服务集对提出的方法进行了验证.实验结果表明,该方法可以准确地对不同类型的服务文档进行聚类.与经典的潜在狄利克雷分配(latent Dirichlet allocation, LDA),K-means 等方法相比,该方法在聚类纯度和 F-measure 指标上均具有更好的效果,从而为按需服务发现与服务组合提供更好的支持.

关键词　服务聚类;潜在狄利克雷分配;主题;概率;特征降维

中图法分类号　TP311

随着软件即服务(software as a service, SaaS)与面向服务的架构(service oriented architecture, SOA)技术的快速发展,互联网上的 Web 服务呈现出快速增长的趋势.如截止到 2012 年 7 月 10 日,Web 服务搜索引擎 Seekda① 统计的 Web 服务已超过 2.8 万个,Web 服务编程网站 ProgrammableWeb (PWeb)② 上发布的 Web API 也已超过 6 400 个.此外,通过对企业中遗留软件系统进行碎片化拆分和封装得到的 Web 服务也开始在网络中部署[1].同时,服务资源的异构特征明显.一方面,服务遵循的协议并不单一,编程网站(programing Web, PWeb)的统计数据显示,发布在该网站上遵循简单对象访问协议(Simple Object Access Protocol, SOAP)的服务约占 21%,而遵循表达性状态迁移(representational state transfer, REST)协议的轻量级 RESTful 服务约占 70%.另一方面,服务描述语言也逐步多样化,例如 WSDL,OWL-S,WADL,WSMO 等,还有大量服务(特别是 RESTful 服务)通过自然语言文本进行描述.服务规模的剧增以及服务的异构性给大众用户准确、高效地发现服务增加了困难,也为软件开发者有效发现和重用服务资源带来了极大的挑战.

服务聚类是支持服务发现的一种重要辅助手段.基于功能相似度进行服务聚类能够改善 Web 服务搜索引擎的能力[2].通过将用户请求定位到特定服务类簇,并从中选择具有相似功能的服务,可以有效地降低服务搜索空间.目前,基于功能相似度的服务聚类已有大量研究.例如,文献[2]提出了一种 WSDL 文档挖掘方法,从 WSDL 文档中抽取体现服务功能的关键特征,然后基于这些特征将服务聚类为功能相似的类簇.文献[3]提出一种基于服务和操作联合聚类(co-clustering)的服务社区学习算法,把具有相似功能的服务聚类为同构的服务社区.然而,已有的服务聚类方法在以下两点考虑不足:

1) 进行聚类的服务文档类型比较单一.现有的服务聚类方法大多针对 WSDL 文档[2-4] 或 OWL-S 文档[5-6] 等单一类型的服务描述文档,并且这些服务大都遵循 SOAP 协议,对通过自然语言文本描述的 RESTful 服务的关注相对较少.

2) 没有考虑服务的领域特性.现有的服务聚类方法大都是针对从互联网中收集的服务描述文档直接进行聚类,而没有对服务所属的领域加以区分.比如,Travel 领域的"旅馆预订"服务与 Shopping 领域的"购物"服务都具有"获取价格"和"进行支付"操作,如果使用上述聚类方法,则这两个来自不同领域的服务可能被分到同一类簇中.但是,用户显然不希望在查找"购物"服务时,却找到"旅馆预订"服务.

因此,面对互联网上服务的规模化增长以及服务的异构性,针对已有的服务聚类方法中存在的不足,如何进行准确、高效地服务聚类成为一个极具挑战性的问题.

针对该问题,我们提出了一种两阶段的服务聚类方法.首先,利用融合领域特性的支持向量机(support vector machine, SVM)对服务进行分类;然后,对分类得到的领域服务集进行面向主题的聚类.我们的前期工作已经实现了面向领域的服务分类方法[7],本文的工作聚焦在第 2 阶段,即利用前一阶段得到的领域服务集和领域词汇排序表(对特定领域中的词汇按照其与领域的关联度进行排序),将特定领域的服务集进一步聚类为不同的主题类簇.我们把采用 WSDL、自然语言文本等描述的服务看作短文本文档,将其统一转换成向量形式,结合服务分类得到的领域词汇排序表,使用 LDA[8] 主题模型将服务描述文档表征为服务-主题-特征词间的关系,进而通过概率方式对特定领域的服务进行聚类,以得到功能相似的主题类簇.该方法避免了用于聚类的服务文档类型的单一化,有利于对服务资源进行组织管理,从而促进服务发现.同时,该方法还有助于服务组合,从功能相似的主题类簇中可以更容易地发现兼容的服务以及可替代服务.

① http://webservices.seekda.com/browse

② http://www.programmableweb.com/

1 基于概率的领域服务聚类模型

为了实现融合领域特性、面向主题的服务聚类，本文使用 LDA 主题模型构建了一个用于生成服务与主题、主题与特征词间概率分布关系的领域服务聚类模型(domain service clustering model, DSCM)，从服务描述文档中抽取其所包含的主题，通过包含主题的概率将特定领域的服务聚类为不同的主题类簇.

1.1 DSCM 描述

定义 1. DSCM 可以被表征为 $DSCM=(DS, DT, V, Po)$，其中，$DS=\{S\}$是由服务分类得到的领域服务集；$DT=\{T\}$是特定领域内所有服务包含的主题(topic)集合；$V=\{t\}$表示特定领域内所有服务包括的特征词(term)集合. $Po=\{policy\}$表示服务聚类时可以采取的策略集，比如相似度计算策略、概率计算策略等. 本文采用概率计算策略，即根据服务文档包含不同主题的概率确定该服务隶属于哪个主题类簇.

定义 2. 服务 $S\subset V$，即一个服务可以被表示为 V 中多个特征词组成的集合. 这里，服务特指一个 Web 服务的描述文档，包括从自然语言描述文本以及从 WSDL，OWL-S 等文档中抽取的体现服务功能的特征词.

定义 3. 主题 $T\subset V$，即一个主题可以被表示为 V 中多个特征词组成的集合. 注意，不同主题可以包含相同的特征词，但包含该词的概率可能不同.

DSCM 模型通过 LDA 得到服务与主题、主题与特征词间的概率分布关系，进而使用概率方式进行面向主题的服务聚类. LDA 是一种生成概率模型，其最大特点在于可避免概率潜在语义索引与 unigrams 混合等模型中存在的过度拟合问题[8]. LDA 将每个服务描述文档表示为其所包含的主题的集合，每个主题由一系列特征词的分布组成，如图 1 所示.

给定特定领域内的服务集 DS，特征词集合 V，使用 LDA 生成 DS，以得到该领域内服务与主题、主题与特征词间的概率关系. 对于 DS 中的每一个服务 S，LDA 使用如下生成过程.

1) 确定服务 S 包括的特征词数 N：选择 $N\sim Poisson(\xi)$，参数 ξ 不需要计算；

2) 确定服务 S 包含的主题：选择 $\boldsymbol{\theta}\sim Dir(\alpha)$；

3) 确定服务 S 中的特征词：对于每个特征词 t_n $(1\leqslant n\leqslant N)$，选择一个主题 $T_n\sim Multinomial(\boldsymbol{\theta})$，然后根据主题 T_n 条件下的多项式概率 $p(t_n|T_n,\beta)$ 选择特征词 t_n.

其中，$Poisson()$表示泊松分布；$Dir()$表示狄利克雷(Dirichlet)分布；$Multinomial()$表示多项式分布；α，β 为 Dirichlet 分布的超参数；$\boldsymbol{\theta}$ 是一个 $1\times|DT|$ 的向量，表示 $|DT|$ 个主题发生的概率；T_n 为离散型随机变量，在主题集合 DT 中取 $|DT|$ 个离散值；t_n 也是离散型随机变量，在特征词集合 V 中取 $|V|$ 个离散值. 通过上述过程，可以确定 DS 中每个服务 S 包含不同主题的概率，进而根据概率将该服务聚类到相应的主题类簇.

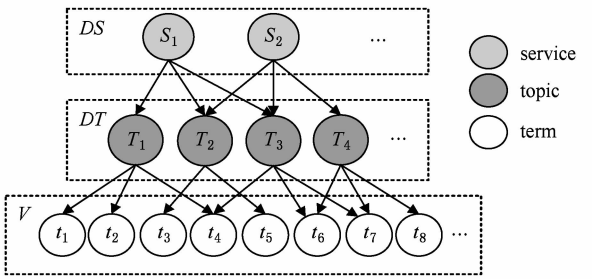


Fig. 1 Service document described by LDA.

图 1 使用 LDA 描述服务文档

1.2 DSCM 模型构建过程

针对 DS 中的服务 S，首先根据 DS 中不同主题的分布情况(由 Dirichlet 分布超参数 α 给出)确定 S 包含不同主题的概率(即参数 $\boldsymbol{\theta}$)，然后根据给定主题的条件下特征词出现的概率(即参数 β)，进而可以根据概率过程生成 DS 中的所有服务，并根据每个服务 S 包含不同主题的概率对服务进行聚类. 但是在实际应用中，只有计算上述参数才能进行概率关系的计算进而得到相应的概率，因此 DSCM 模型的构建过程实质上是确定 LDA 模型参数的过程. 但是，LDA 主题模型很难进行直接推理，通常使用变分推理和 EM 算法[8]以及吉布斯(Gibbs)抽样方法[9]. 本文采用 Gibbs 抽样方法确定 LDA 模型参数并计算所需要的概率，进而实现 DSCM 模型的构建.

Gibbs 抽样是一种从多元概率分布获得随机样本序列的马尔可夫链蒙特卡洛算法①. 对于 LDA 模型，每一步 Gibbs 抽样都服从如下分布：

$$P(T_n=j|T_{-n},V)\propto\frac{N_{-n,j}^{(t_n)}+\beta}{N_{-n,j}^{(\cdot)}+|V|\beta}\frac{N_{-n,j}^{(S_n)}+\alpha}{N_{-n,\cdot}^{(S_n)}+|DT|\alpha},\tag{1}$$

其中，T_{-n}表示除当前指定主题外其他所有主题；$|V|$表示特征词集合中词的个数；$|DT|$表示特定领域内

① http://en.wikipedia.org/wiki/Gibbs_sampling

所有服务包含的主题数;$N_{-n,j}^{(t_n)}$表示特征词 t_n 分配给主题 j 的次数;$N_{-n,j}^{(\cdot)}$表示分配给主题 j 的所有特征词的总数;$N_{-n,j}^{(S_n)}$表示服务文档 S_n 包括的特征词分配给主题 j 的次数;$N_{-n,\cdot}^{(S_n)}$表示服务文档 S_n 包括的特征词的总数[10].注意,此处所有相关特征词的数目都不包括 $T_n=j$ 的分配.进行 Gibbs 抽样时,T_n 使用 1～$|DT|$间的整数值进行初始化,确定马尔可夫链的初始状态.然后根据式(1)得到马尔可夫链的下一状态,如此迭代多次直到该链接近目标分布,记录 T_n 的当前值.当 Gibbs 抽样过程迭代足够多次后,使用抽样结果对参数进行估算.

2　基于 DSCM 模型的领域服务聚类方法

本节将重点介绍在 DSCM 模型的基础上进行领域服务聚类的方法.首先介绍了服务特征降维策略以及领域服务聚类策略,然后给出了面向主题的领域服务聚类算法.

2.1　服务特征降维策略

对服务文档的特征向量进行降维在面向主题的服务聚类中具有重要作用.一方面,可以降低计算量并加快 Gibbs 抽样的收敛速度,从而提高聚类效率.另一方面,去除那些对服务聚类贡献度不大的特征词,有助于提高聚类准确度.因此,在确保不影响服务文档特征提取的前提下,应最大限度地降低服务特征向量的维度.首先,根据面向领域的服务分类得到的领域词汇排序表的前 k 个(Top-k)特征词,计算词对该领域的表征度,将表征度超过给定阈值的词过滤掉,得到领域特征词集,然后对分类后得到的领域服务集进行特征降维,如图 2 所示.下面将详细阐述特征降维流程.

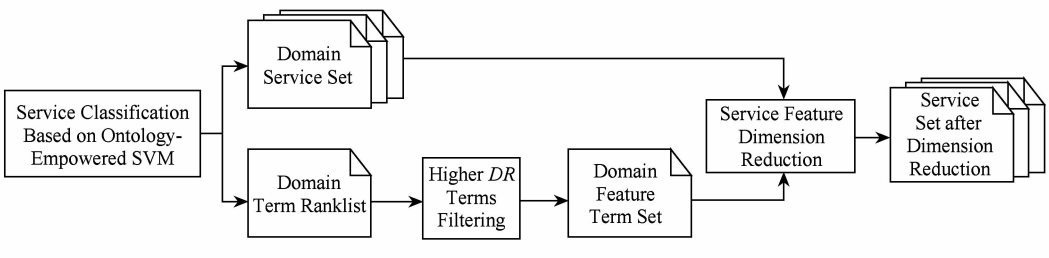


Fig. 2　Service feature dimension reduction process.

图 2　服务特征降维流程

我们的前期研究[7]提出了一种基于本体辅助式支持向量机进行面向领域的服务分类方法.其基本流程如下:首先对服务的自然语言描述文本进行预处理,具体包括文档解析(分词处理)、动词和名词抽取、停词处理、单词还原(stemming)以及词频统计.除了自然语言描述文本外,对于某些服务所具有的 WSDL 文档,从中抽取 contents,types,messages,ports,service name 等体现服务功能的信息;对于某些服务所具有的 OWL-S 文档,从中抽取 service name,input,output,precondition,effects 等信息,并且采用类似的方式对这些抽取的内容进行预处理,从而得到描述服务核心功能的特征词集.然后从预处理后得到的服务文档集中选取训练集和测试集.由于 TF-IDF 是用于评估一个词对一个文档集或语料库中一份文档的重要程度①,而没有考虑该词对于领域的重要度.因此,考虑到服务所具有的领域特性,我们将经典的 TF-IDF 改造成两部分:KF-IDF-DF(keyword frequency-inverse document frequency-domain frequency)用来度量一个服务与其所属领域间的关系;KF-IRF(keyword frequency-inverse repository frequency)用来度量一个领域与整个服务储存库(service repository)间的关系.利用 KF-IDF-DF 构造向量空间,将不同类型的服务描述文档统一转换成向量形式,使用本体辅助式 SVM 对服务进行分类后,使用 KF-IRF 对分类后得到的该领域中所有服务文档包括的词汇进行排序.根据设定的迭代条件与前一次领域词汇排序表进行比较,用以决定是否进行进一步迭代.当连续两次得到的领域词汇排序表的前 m 个词的排序相同时终止迭代,得到分类后的领域服务集和相应的领域词汇排序表.

定义 4. 词对领域的表征度(degree of representation, DR).指一个特征词 t_i 对一个领域 d 的表征程度,用式(2)表示:

① http://en.wikipedia.org/wiki/Tf%E2%80%93idf

$$DR(t_i,d)=\frac{|\{S_j|t_i\in S_j\wedge S_j\in DS_d\}|}{|DS_d|},\quad(2)$$

其中,$|\{S_j\}|$表示领域 d 中包含特征词 t_i 的服务文档数,$|DS_d|$表示领域 d 中的服务文档总数.不难看出,$DR(t_i,d)$越大,特征词 t_i 在领域 d 的服务文档中出现的越频繁,对该领域的表征度也越强.如 Travel 领域中的"travel"这个特征词,在该领域的大部分服务文档中都会出现.这样的特征词在面向领域的服务分类中非常重要,但由于在该领域大多数服务文档中都会出现,所以,在对特定领域进行服务聚类时反而意义不大.因此,我们给定阈值 th(0~1 之间的百分数),把 DR 大于 th 的特征词(即可以高度代表一个领域的特征词)去除.

定义 5. 领域特征词集(domain feature term set, $DFTS$).选择领域词汇排序表中的前 k 个词,并从中去除 DR 高于给定阈值 th 的特征词后得到的领域词汇构成领域特征词集.

使用 $DFTS$ 对分类后得到的服务文档进行特征降维,即把每个服务文档中存在的不属于 $DFTS$ 的特征词过滤掉,得到降维后的服务文档集,为服务聚类做准备.

2.2 领域服务聚类策略

基于 DSCM 模型,使用 Gibbs 抽样得到的模型参数计算所需概率,然后根据给定的主题集以及每个服务包含的不同主题的概率对服务进行聚类.

假设服务 S_i 包含 r 个主题 $T_1,T_2,\cdots,T_r$,$P(S_i,T_j)$表示服务 S_i 包含主题 T_j 的概率.如果一个服务包含某个主题的概率越大,则该服务隶属于该主题的可能性就越大.因此,服务 S_i 被分到 $P(S_i,T_j)$值最大的主题类簇(topic cluster, TC),用式(3)表示:

$$TC(S_i)=T_K\wedge\forall j((j\neq k)\rightarrow P(S_i,T_j)<P(S_i,T_k)),\ 1\leqslant j,k\leqslant r.\quad(3)$$

因此,我们采用的领域服务聚类策略是,如果一个服务包含某个主题的概率最大,则该服务就隶属于相应的主题类簇.

2.3 聚类算法

算法 1 给出了面向主题的领域服务聚类算法(domain service clustering algorithm, DSCA).

算法 1. 面向主题的领域服务聚类算法 DSCA.

输入:服务集 SS,特征词数 k,表征度阈值 th,超参数 α,β,聚类主题数 $Tnum$;

输出:聚类结果.

① $DS,DTR\leftarrow ontology\text{-}empowered SVM(SS)$;

② for each term $t\in DTR$

③ if($t\in DTR(k)$ && $t.dr<th$) /* dr 为词对领域的表征度 */

④ $DFTS\leftarrow t$; /* 将 t 加入领域特征词集 $DFTS$ */

⑤ $InputDataFile\leftarrow filter(DS,DFTS)$; /* 服务特征降维得到每个服务与其包含不同主题的概率分布 */

⑥ $ST\leftarrow LDAGibbs(InputDataFile,\alpha,\beta,TNum)$; /* 得到服务主题类簇 */

⑦ $TCs\leftarrow parse(ST)$;

⑧ return TCs.

其中,第①行利用本体辅助式 SVM 进行面向领域的服务分类,得到分类后的领域服务集 DS 和领域词汇排序表(domain term ranklist, DTR);第②~④行对 DTR 中的前 k 个特征词根据其对领域的表征度阈值 th 进行过滤,得到领域特征词集 $DFTS$;第⑤行利用 $DFTS$ 对 DS 进行特征降维,构建 DSCM 模型的输入;第⑥行利用基于 Gibbs 抽样的 LDA 方法得到每个服务与其包含主题的概率分布;第⑦~⑧行对得到的服务-主题概率分布进行解析,将每个服务聚类到其所包含主题概率最大的主题类簇,最终将特定领域内的所有服务聚类到不同的主题类簇 TCs.

3 实验分析

3.1 实验准备

文中所有的实验和算法均通过 Java 实现,基于 DSCM 模型的领域服务聚类方法主要基于 JGibbLDA① 工具.所有实验运行在一台具有 Intel Core™ i5 CPU 760 @ 2.80GHz,4GB 内存,操作系统为 Windows7 的 PC 上.

实验数据来源于 PWeb,该网站提供的服务中,70%为 REST 服务,21%为 SOAP 服务,并且提供服务的文本描述信息.图 3 所示为 PWeb 上".tel" API② 的相关 Profile 信息,包括 API 的名字、描述、标签等.我们使用爬虫并结合该网站提供的开放

① http://jgibblda.sourceforge.net/

② http://www.programmableweb.com/api/.tel

API 将图 3 矩形框中的信息爬取下来,存储在文本文档中,然后进行预处理操作. 对于 PWeb 上遵循 SOAP 协议的服务,除了收集图 3 所示的文本描述信息,还进一步收集了相应的 WSDL 文档,然后结合文献[2]中的预处理方法,从中抽取相应的核心参数,比如服务名、操作名、输入、输出等. 我们一共收集了 PWeb 上来自 63 个领域的 6400 多个服务的描述文档.

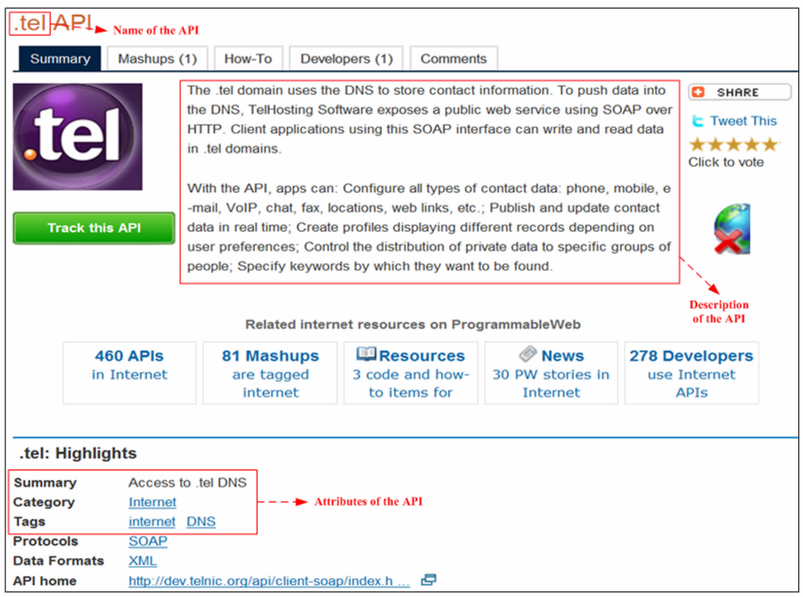


Fig. 3　Related information of ".tel" API Profile.

图 3　".tel" API Profile 的相关信息

由于 PWeb 上各个领域所含的服务数差异较大(Internet 领域最多(413 个),Dating 领域最少(2 个)),因此,使用本体辅助式 SVM 进行面向领域的分类时,很难为这些领域统一指定训练集大小. 同时,如果训练集的规模较小,那么使用根据小规模训练集得到的训练模型进行分类时将得不到较高的分类准确率. 所以,我们目前只关注服务数大于 100 的除"Other"领域(PWeb 中所有未分类的服务所属的领域)外的 24 个领域,对这些领域使用本体辅助式 SVM 进行分类,表 1 所示为随机选择的 5 个领域的分类结果. 本文将使用这 5 个领域分类后的 1811 个服务文档进行聚类实验.

从表 1 可以看出,使用本体辅助式 SVM 进行分类得到的每个领域的服务数(表 1 黑体部分)比 PWeb 上原始的服务数多,主要因为 PWeb 上的每个服务只属于一个领域,并没有考虑一个服务可能会跨越多个领域的情形. 而我们的服务分类方法则支持一个服务属于 2 个或多个不同领域.

Table 1　Examples of the Classification Results

表 1　分类结果示例

Domain	Number of Services		Domain Term Ranklist(Top 10,ranked by KF-IRF)
	PWeb	Ontology-empowered SVM	
Financial	217	**255**	finance, stock, trade, invoice, currency, tax, exchange, money, quote, payment
Internet	412	**532**	internet, url, cloud, ip, host, monitor, link, utility, website, server
Mapping	229	**309**	map, location, address, place, direction, gp, latitude, longitude, gi, route
Shopping	210	**243**	shopping, deal, product, coupon, affiliate, merchant, payment, cart, order, price
Social	342	**472**	social, twitter, network, friend, share, medium, post, tweet, photo, platform

下面为表 1 中的 5 个领域确定用于聚类的最优主题数. 首先根据文献[11]中的方法得到 5 个领域的最优主题数,接着使用该主题数对 5 个领域的服务集进行聚类,得到每个领域的主题类簇集. 然后根据每个主题类簇包括的服务数对每个领域的主题类簇集进行过滤. 即把每个领域中主题类簇包括的服务数

大于 10 的类簇总数作为该领域聚类的最优主题数，结果如表 2 所示. 此处针对 5 个领域而不是单独针对每个领域进行最优主题数选择是因为目前每个领域包括的服务数较少，Gibbs 抽样过程很难达到稳定状态.

Table 2 Optimal Number of Topics in Each Domain

表 2 每个领域的最优主题数

Domain	Optimal Number of Topics
Financial	7
Internet	14
Mapping	9
Shopping	5
Social	10

3.2 评估指标

本文使用纯度和 F-measure 对聚类结果进行评估分析. 纯度越高表明聚类效果越好. F-measure 则是通过综合考虑准确率和召回率对聚类结果进行评估.

定义 6. 聚类纯度(purity of cluster, PC)[12]. 设特定领域内服务集 DS 的标准聚类结果 $SC=\{C_1,C_2,\cdots,C_x\}$，由人工判断得到. 实验得到的主题聚类结果为 $TC=\{TC_1,TC_2,\cdots,TC_y\}$，则每个主题类簇的聚类纯度定义为

$$PC(TC_j)=\frac{1}{|TC_j|}\max_i(n_j^i),\ 1\leqslant i\leqslant x,1\leqslant j\leqslant y, \tag{4}$$

其中，$|TC_j|$ 表示主题类簇 TC_j 中的服务数，n_j^i 表示第 i 个标准类簇中的服务被分到第 j 个主题类簇中的服务数. 在此基础上，特定领域内聚类结果的整体纯度定义为

$$PC(TC)=\sum_{j=1}^{n}\frac{|TC_j|}{|DS|}PC(TC_j). \tag{5}$$

由于服务文档数较多，标准聚类结果单独依靠人工进行判断耗时费力，并且完全依靠人工定义往往有失偏颇. 为此，借助于 Cosine 相似度计算公式，用式(6)计算分类后特定领域内每个服务与该领域包含的每个主题间的相似度，把该服务划分到相似度最大的主题类簇，在此基础上进行人工筛选，确定最终的标准聚类结果.

$$sim(S_a,T_b)=\frac{\boldsymbol{WT}(S_a)\cdot\boldsymbol{WT}(T_b)}{|\boldsymbol{WT}(S_a)||\boldsymbol{WT}(T_b)|}=\sum_{i=j=1}^{n}\frac{w_{t_i}\times w_{ft_j}}{\sqrt{\sum_{i=1}^{n}w_{t_i}^2}\times\sqrt{\sum_{j=1}^{n}w_{ft_j}^2}}, \tag{6}$$

其中，$\boldsymbol{WT}(S_a)=(t_i,w_{t_i}):t_i\in DFTS$，$w_{t_i}=f_{t_i}\times p_{t_i}$，$f_{t_i}$ 为服务 S_a 中特征词 t_i 的词频，p_{t_i} 为该词在主题 T_b 中出现的概率. $\boldsymbol{WT}(T_b)=(ft_j,w_{ft_j}):ft_j\in FTS$，$ft_j$ 为主题特征词，依靠人工从每个主题包括的词汇集中按概率从高到低进行筛选得到，主题特征词的集合构成 FTS. w_{ft_j} 为主题包含该词的概率. 注意：计算相似度时，如果服务中的特征词在主题特征词集合 FTS 中不存在，则相应的权重 $w_{t_i}=0$，如果存在，则主题向量中相应的特征词权重 $w_{ft_j}=w_{t_i}$.

F-measure 是准确率 P(precision)与召回率 R(recall)的调和平均数，用式(7)计算：

$$F_i=\frac{2P_iR_i}{P_i+R_i},\ P_i=\frac{|TC_i\cap C_i|}{|TC_i|},\ R_i=\frac{|TC_i\cap C_i|}{|C_i|}. \tag{7}$$

3.3 服务特征降维参数对聚类效果影响

实验中先确定 DSCA 算法中用于服务特征降维的参数 k 与表征度阈值 th 的不同取值对服务聚类的影响. 首先确定 k 的不同取值对服务聚类效果的影响，此时设置阈值 $th=100\%$，即聚类时考虑领域词汇表中的所有特征词. 为了客观体现 k 的取值对服务聚类的影响，本文针对 5 个领域进行实验，然后对聚类结果取平均值，结果如表 3 所示. 实验结果表明，选取领域词汇排序表的前 200 个词，可以取得较好的聚类效果(表 3 中黑体部分). 当少于 200 个词时，由于选取的领域词汇数较少，导致得到的服务特征较少而使得聚类效果不好；当超过 200 个词时，由于选取的领域词汇数较多，一些不相关的词可能导致服务特征不明显，使得聚类效果不好.

Table 3 The Effect of Values of k to Clustering

表 3 k 的取值对聚类效果的影响

Top-k	Purity	F-measure
50	0.508	0.458
100	0.534	0.487
150	0.495	0.440
200	**0.535**	**0.489**
300	0.533	0.488

下面确定表征度阈值 th 对服务聚类的影响. 针对 5 个领域，表 4 给出了选取领域词汇表的前 200 个词，表征度阈值 th 的不同取值对服务聚类效果的影响. 从表 4 可以看出，当阈值 th 取 90%时，聚类效果较好(表 4 中黑体部分)，因为去除了在每个领域的大多数服务文档中都出现而对聚类意义不大的词. 如果 th 小于 90%，由于过滤掉了很多领域核心词汇，反而不能很好地体现服务的特征导致聚类效果不好.

Table 4　The Effect of Threshold *th* to Clustering

表 4　表征度阈值 *th* 对聚类效果的影响

Threshold *th*/%	Financial		Internet		Mapping		Shopping		Social	
	Purity	*F*-measure	Purity	*F*-measure	Purity	*F*-measure	Purity	*F*-measure	Purity	*F*-measure
50	0.647	0.638	0.652	0.576	0.553	0.446	0.444	0.442	0.532	0.496
60	0.557	0.529	0.624	0.537	0.544	0.493	0.576	0.586	0.540	0.497
70	0.537	0.519	0.673	0.599	0.586	0.529	0.588	0.591	0.551	0.514
80	0.522	0.497	0.618	0.536	0.573	0.514	0.531	0.528	0.578	0.560
90	**0.784**	**0.785**	**0.782**	**0.772**	**0.706**	**0.698**	**0.737**	**0.740**	**0.731**	**0.743**
100	0.561	0.528	0.620	0.538	0.495	0.418	0.539	0.547	0.462	0.412

3.4　实验结果分析

取每个领域词汇排序表中的前 200 个词,设置词对相应领域的表征度阈值 *th* 为 90%,使用 DSCA 算法对每个领域的服务集针对表 2 中相应的最优主题数进行服务聚类.结果除了得到每个主题类簇包括的服务外,还可以得到每个服务-主题以及主题-特征词的概率分布.根据服务-主题概率分布以及 2.2 节所述的领域服务聚类策略,将服务聚类到相应的主题类簇.

图 4 所示为来自 5 个领域不同主题包括的具有较高出现概率的 10 个特征词,词右边的数字表示该词在相应主题中出现的概率.从图 4 可以看出,每个主题所包括的特征词的出现概率不同,并且相同的特征词在不同主题中出现的概率也不同(图 4 中黑体).比如,Internet 领域包含的主题 8、Mapping 领域包含的主题 4 均包括特征词"location",但包含该词的概率不同,分别为 0.029,0.164.

文献[13]根据定义的服务功能描述模型构建领域功能本体,而我们可以从每个主题中选取出现概率较高的特征词,用来扩充由面向领域的服务分类得到的相应领域的核心词汇,以增量的方式逐步构造结构良好的领域本体(如果是来自同一主题的特征词映射到本体中的概念,则这些概念间的关联关系较密切,有助于本体中概念间关联关系的定义),从而为基于语义的按需服务发现奠定基础.

Financial-Topic 1

Term	prob.
currency	0.203
money	0.127
rate	0.073
exchange	0.070
Social	0.067
enterprise	0.042
license	0.040
platform	**0.033**
game	0.027
transaction	0.019

Internet-Topic 8

Term	prob.
tool	0.534
ip	0.163
address	0.070
location	**0.029**
return	0.026
response	0.017
lookup	0.017
ping	0.016
format	0.014
call	0.012

Shopping-Topic 1

Term	prob.
payment	0.084
store	0.042
product	0.041
mobile	0.039
cart	0.037
order	0.035
customer	0.031
auction	0.028
platform	**0.024**
marketplace	0.023

Mapping-Topic 4

Term	prob.
location	**0.164**
search	0.150
mobile	**0.129**
gps	0.101
advertisig	0.045
navigation	0.041
local	0.039
map	0.033
content	0.030
telephony	0.029

Social-Topic 7

Term	prob.
network	0.114
friend	0.043
profile	0.032
method	0.030
account	0.029
support	0.027
platform	**0.026**
community	0.024
create	0.024
comment	0.023

Fig. 4　Examples of topic-term distribution from different domains.

图 4　不同领域中的主题-特征词分布示例

实验 1. 聚类效果分析

使用 DSCA 算法、Direct-LDA 算法(直接使用 LDA 进行聚类)以及 *K*-means 算法的聚类结果对比如图 5、图 6 所示.从图中可以看出,针对 5 个领域采用 DSCA 算法得到的聚类结果在纯度和 *F*-measure 上都比 Direct-LDA 算法以及 *K*-means 算法好.就平均值而言,采用 DSCA 算法得到的 5 个领域聚类纯度以及 *F*-measure 的平均值均为 0.748;Direct-LDA 算法得到的聚类纯度、*F*-measure 的平均值分别为 0.580,0.451;而 *K*-means 算法得到的聚类纯度以及 *F*-measure 的平均值最低,分别为 0.373,0.301.

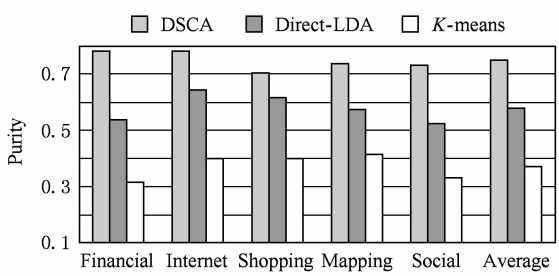


Fig. 5　Comparisions of purity.

图 5　服务聚类纯度对比

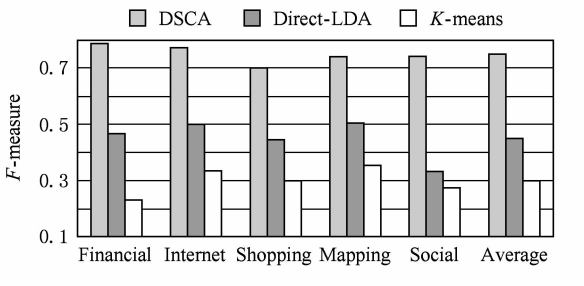


Fig. 6 Comparisions of *F*-measure.

图 6 服务聚类 *F*-measure 对比

实验结果表明采取的服务特征降维策略以及聚类策略可以有效地改进聚类效果. *K*-means 算法的聚类效果最差的原因可能主要有以下 2 点:1)计算服务间相似度时仅使用词频将服务文档表示为向量形式;2)使用该算法得到的每个类簇所包括的服务数相差较大. 每个领域中都有 1～2 个类簇包括了该领域的绝大多数服务,而这 1～2 个类簇的纯度反而较低. 根据式(5)中整个聚类纯度的定义,导致整体聚类纯度较低. 由于这种服务数分布导致其余类簇的召回率较低,根据式(7)中 *F*-measure 的定义,使得 *F*-measure 值也较低.

实验 2. 聚类时间分析

我们还对比了 3 种算法针对 5 个领域的聚类时间,如图 7 所示:

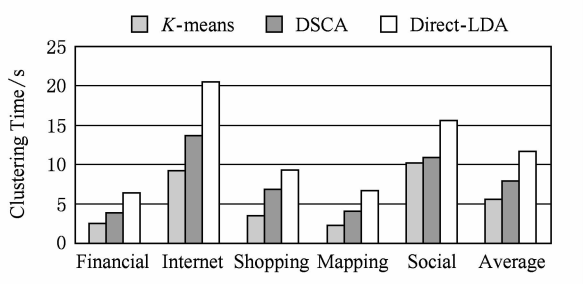


Fig. 7 Comparisions of clustering time.

图 7 服务聚类时间对比

从图 7 可以看出,针对 5 个领域中的每个领域,*K*-means 算法所耗时间最少,DSCA 算法次之,Direct-LDA 算法花费时间最长. 就 5 个领域的平均值而言,3 种算法所用时间分别为:5.623 s,7.915 s,11.766 s. DSCA 算法比 *K*-means 算法所耗时间较长是因为 Gibbs 抽样过程需要迭代多次才能得到收敛于目标分布的马尔可夫链. DSCA 算法比 Direct-LDA 算法花费时间较少在于该算法使用 *DFTS* 对服务集进行特征降维,加快了 Gibbs 抽样过程.

4 方法应用

CloudCRM① 云服务超市是课题组与金蝶软件(中国)有限公司合作开发的 SaaS 服务管理与定制平台. 该平台对开源客户关系管理系统 SugarCRM 进行服务化封装[1],将遗留软件系统通过开放的服务方式进行部署和运营,并且注册了大量开放的 Web 服务和 Web API,通过多租户技术为中小企业用户提供按需服务定制. 文中提出的方法已经应用于该平台.

对 CloudCRM 云服务超市中的服务,首先根据 2.3 节所述的聚类算法从服务所属的领域视角进行面向主题的聚类. 具体说来,对平台中采用不同语言(比如 WSDL、自然语言)描述的服务,首先抽取体现其核心功能的特征词,然后使用 2.1 节中的预处理方法对由特征词组成的服务文档进行预处理,进而将不同类型的服务文档统一转换成向量形式,以使用本体辅助式 SVM 进行面向领域的服务分类. 对分类后得到的领域词汇排序表的前 200 个词,计算词对领域的表征度,将表征度大于 90% 的特征词过滤掉,得到领域特征词集,对分类后得到的领域服务集进行特征降维. 接着使用 JGibbLDA 工具得到每个降维后的服务文档与其包含的不同主题的概率分布,然后对概率分布文档进行解析,即找到每个服务所包含的主题概率的最大值,将其聚类到该主题对应的主题类簇中. 在此基础上,进一步对特定主题类簇中的服务利用 RGPS(role-goal-process-service)[14] 从需求视角实现服务聚类. 这样,有助于对服务进行组织管理,促进按需服务发现与服务组合,从而有助于给用户带来高质量的服务体验.

目前,互联网上可公开访问的服务注册库主要包括:PWeb,Seekda,WebServiceList②,Xmethods③ 等. 其中,PWeb 按服务所属分类(category)对服务进行组织;Seekda 首先根据服务提供者所属的国家对服务提供者进行分类,然后按照特定国家的服务提供者对服务进行组织;WebServiceList 按照服务所属的领域进行组织;Xmethods 提供的服务列表

① http://202.114.107.230:8080/CloudCrm/login.jsp

② http://www.webservicelist.com/

③ http://www.xmethods.com/ve2/index.po

按照服务的提交时间进行排序.而本文方法可以提供比按领域分类更加细化的描述能力(不仅可以提供服务所属的领域信息,还可以进一步提供特定领域的服务所属的主题信息),有利于促进服务资源的管理和查询.

此外,本文方法还有助于进行服务组合.从功能相似的主题类簇中可以更容易地发现组合服务中的可兼容服务以及可替代服务.根据已有的服务组合模式或使用模式,建立服务主题类簇间的关联关系,从而为服务组合提供支持.

因此,本文方法在服务发现和服务组合中具有实际的应用价值.

5　相关工作

服务聚类是一种有效地促进服务发现的技术[15],在这方面国内外近年来已有大量研究.文献[16]使用 Agglomerative 层次式聚类算法对相似的服务进行聚类,以改进服务发现效率.文献[4]也使用 Agglomerative 算法对服务进行聚类,以提高搜索引擎的能力.文献[17]基于相似度计算从服务功能属性和过程模型两个层面对服务进行聚类,根据聚类结果过滤掉大量无关服务,从而提高服务发现效率.文献[18]提出一种通过服务本体引导服务进行聚类的方法.该方法从 ServiceName,Interface,Capability,QoS(quality of service)4 个方面计算服务与同一服务本体间的相似度,将大于给定相似度阈值的服务进行聚类,从而降低搜索空间,提高服务查找效率.文献[15]提出一种自组织的分类(taxonomic)聚类算法,对语义 Web 服务进行分类组织,用于促进服务发现.文献[19]基于支配(dominance)服务概念提出一种服务聚类方法,通过服务支配关系决定服务与用户请求的相关性.文献[20]使用 K-means 算法对服务和需求进行聚类,有效地降低了服务搜索空间.文献[2]从 WSDL 文档中抽取 5 个关键特征,然后基于这些特征将服务聚类为功能相似的类簇.文献[21]提出一种 WTCluster 方法,使用 K-means 算法基于 WSDL 相似性以及标签(tag)相似性对服务进行聚类.

目前,就我们的知识所及,有 2 篇文献将 LDA 应用到服务计算领域.其中,文献[22]直接用 LDA 将 WSDL 文档建模为结构化(文档-主题-词)文本文档,得到每个主题的关键词分布,然后基于主题对服务进行检索.文献[6]使用概率潜在语义分析(probabilistic latent semantic analysis, PLSA)和 LDA 从服务描述中抽取潜在主题,然后根据主题基于 OWL-S 服务的 Profile 描述和功能属性两个方面进行服务聚类.

但是,上述方法在如下方面有所欠缺:1)大多数服务聚类方法基于相似性度量,相似度计算时不同方面权重参数的设置会在一定程度上影响聚类性能[2,17-18,21];2)现有服务聚类方法的一个重要局限是聚类的服务文档类型比较单一,比如文献[4,16]针对 WSDL,文献[15,18]针对 OWL-S,还有些针对随机生成的服务[17];3)现有方法大都是从服务的功能、流程、QoS 等方面直接进行聚类,而没有考虑服务的领域特性.

本文在面向领域的服务分类基础上,提出了一种基于 DSCM 模型、融合领域特性的服务聚类方法.该方法通过概率的方式对服务进行聚类,不需要使用权重机制进行相似度计算,并且可以对 WSDL、OWL-S、文本等服务描述文档进行聚类,不再局限于单一文档类型.

LDA 通常用于文档降维,即将文档-词形成的向量空间降为文档-主题形成的向量空间,但该方法得到的文档特征粒度往往较大,而本文使用 DFTS 对服务文档进行降维,去除那些对聚类效果影响不大的词,不会改变文档特征的粒度.

6　结束语

在面向领域的服务分类的基础上,本文提出了一种领域服务聚类模型 DSCM,然后基于该模型对服务进行面向主题的聚类,把特定领域内具有相似功能的服务组织为主题类簇.最后,以 PWeb 上真实的服务集进行实验,验证了基于 DSCM 模型的服务聚类方法的可行性和有效性.与 Direct-LDA 和 K-means 的对比实验分析表明,本文提出的面向主题的服务聚类方法在纯度、F-measure 方面均具有更好的效果.而且,本文方法有助于服务资源的组织管理,从而促进服务发现与服务组合,具有较好的实际应用价值.

下一步,我们将从如下方面展开深入研究:1)在服务聚类的基础上进一步研究按需服务发现;2)根据服务的组合模式和使用模式,建立服务主题类簇间的关联关系,为服务组合和推荐提供支持.

参 考 文 献

[1] Yu Dunhui, Wang Jian, Hu Bo, et al. A practical architecture of cloudification of legacy applications [C] //Proc of the 7th IEEE World Congress on Services. Piscataway, NJ: IEEE, 2011: 17-24

[2] Elgazzar K, Hassan A E, Martin P. Clustering WSDL documents to bootstrap the discovery of Web services [C] // Proc of the 8th IEEE Int Conf on Web Services. Piscataway, NJ: IEEE, 2010: 147-154

[3] Yu Q, Rege M. On service community learning: A co-clustering approach [C] //Proc of the 8th IEEE Int Conf on Web Services. Piscataway, NJ: IEEE, 2010: 283-290

[4] Platzer C, Rosenberg F, Dustdar S. Web service clustering using multidimensional angles as proximity measures [J]. ACM Trans on Internet Technology, 2009, 9(3): 1-26

[5] Liu Jianxiao, He Keqing, Wang Jian, et al. A clustering method for Web service discovery [C] //Proc of the 8th IEEE Int Conf on Services Computing. Piscataway, NJ: IEEE, 2011: 729-730

[6] Cassar G, Barnaghi P, Moessner K. Probabilistic methods for service clustering [C/OL] //Proc of the 4th Int Workshop on Semantic Web Service Matchmaking and Resource Retrieval, Organised in Conjunction with the Int Semantic Web Conf. 2010: 4-20 [2012-07-19]. http://people.csail.mit.edu/pcm/tempISWC/workshops/SMR22010/SMR2Proceedings.pdf

[7] Wang Jian, Zhang Jia, Hung P C K, et al. Leveraging fragmental semantic data to enhance services discovery [C] // Proc of the 13th Int Conf on High Performance Computing and Communications. Piscataway, NJ: IEEE, 2011: 687-694

[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[9] Griffiths T. Gibbs sampling in the generative model of latent dirichlet allocation [R]. Stanford: Stanford University, 2002

[10] Steyvers M, Griffiths T. Probabilistic topic models [M] // Landauer T K, McNamara D S, Dennis S, et al. Handbook of Latent Semantic Analysis. Mahwah, NJ: Lawrence Erlbaum Associates, 2007: 427-446

[11] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences, 2004, 101 (Suppl1): 5228-5235

[12] Zhao Y, Karypis G. Criterion functions for document clustering-experiments and analysis [R]. Minneapolis: Department of Computer Science/Army HPC Research Center, University of Minnesota, 2002

[13] Ye Lei, Zhang Bin. A method of Web service discovery based on functional semantics [J]. Journal of Computer Research and Development, 2007, 44(8): 1357-1364 (in Chinese)
(叶蕾, 张斌. 基于功能语义的 Web 服务发现方法[J]. 计算机研究与发展, 2007, 44(8): 1357-1364)

[14] He Keqing, Peng Rong, Liu Wei, et al. Networked Software [M]. Beijing: Science Press, 2008 (in Chinese)
(何克清, 彭蓉, 刘玮, 等. 网络式软件[M]. 北京: 科学出版社, 2008)

[15] Dasgupta S, Bhat S, Lee Y. Taxonomic clustering and query matching for efficient service discovery [C] //Proc of IEEE 9th Int Conf on Web Services. Piscataway, NJ: IEEE, 2011: 363-370

[16] Richi N, Bryan L. Web service discovery with additional semantics and clustering [C] //Proc of IEEE/WIC/ACM Int Conf on Web Intelligence. Piscataway, NJ: IEEE, 2007: 555-558

[17] Sun Ping, Jiang Changjun. Using service clustering to facilitate process-oriented semantic Web service discovery [J]. Chinese Journal of Computers, 2008, 31(8): 1340-1353 (in Chinese)
(孙萍, 蒋昌俊. 利用服务聚类优化面向过程模型的语义 Web 服务发现[J]. 计算机学报, 2008, 31(8): 1340-1353)

[18] Liu Jianxiao, He Keqing, Wang Jian, et al. Semantic interoperability oriented method of service aggregation [J]. Journal of Software, 2011, 22(Suppl2): 27-40 (in Chinese)
(刘建晓, 何克清, 王健, 等. 一种面向语义互操作性的服务聚合方法[J]. 软件学报, 2011, 22(增刊 2): 27-40)

[19] Skoutas D, Sacharidis D, Simitsis A, et al. Ranking and clustering Web services using multicriteria dominance relationships [J]. IEEE Trans on Services Computing, 2010, 3(3): 163-177

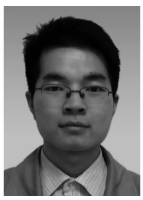
[20] Wang Xianzhi, Wang Zhongjie, Xu Xiaofei. Semi-empirical service composition: A clustering based approach [C] //Proc of the 9th IEEE Int Conf on Web Services. Piscataway, NJ: IEEE, 2011: 219-226

[21] Chen Liang, Hu Liukai, Zheng Zibin, et al. WTCluster: Utilizing tags for Web services clustering [C] //Proc of the 9th Int Conf on Service-Oriented Computing. Berlin: Springer, 2011: 204-218

[22] Chen Jiangfeng, Yu Jianjun. Topic model based structural Web services discovery [J]. Journal of Beijing University of Aeronautics and Astronautics, 2008, 34(6): 734-738 (in Chinese)
(陈江锋, 于建军. 基于主题模型的结构化 Web 服务发现机制[J]. 北京航空航天大学学报, 2008, 34(6): 734-738)

Li Zheng, born in 1984. PhD candidate. Student member of China Computer Federation. Her current research interests include software engineering and service computing.

Wang Jian, born in 1980. PhD and lecturer. Member of China Computer Federation. His current research interests include requirement engineering and service computing.



Zhang Neng, born in 1990. Master candidate. His research interests include software engineering and service computing.

Li Zhao, born in 1986. PhD candidate. Student member of China Computer Federation. His research interests include software engineering and service computing.

He Chengwan, born in 1967. PhD, professor. Senior member of China Computer Federation. His research interests include software engineering, knowledge discovery and data mining.

He Keqing, born in 1947. PhD, professor and PhD supervisor. Senior member of China Computer Federation. His current research interests include software engineering, service computing.

勘误启事

本刊 2013 年第 8 期发表的“连续数据存储中面向 RAID5 的写操作优化设计”(第 1604～1612 页)一文中,因我们工作失误,将该文第 2 作者“张全新”所属单位标错. 应将“张全新[1,2]”改为“张全新[1]”.

在此谨向作者和读者致歉!

《计算机研究与发展》编辑部

2014 年 2 月