面向领域标签辅助的服务聚类方法

田 刚[1,2],何克清[1],王 健[1],孙承爱[2],徐建建[2]
(1.武汉大学软件工程国家重点实验室,武汉大学计算机学院,湖北武汉 430072;2.山东科技大学信息学院,山东青岛 266590)

摘 要: Web服务数量的激增对服务发现提出了更高的要求,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,缺乏考虑服务的领域特性和服务标签的应用.针对这些问题,本文首先使用本体辅助的支持向量机和面向领域的服务特征降维技术建立服务的特征内容向量,然后使用一种标签辅助的主题服务聚类方法T-LDA建立融合标签信息之后的隐含主题表示,并利用归一化方法消除通用主题的影响,综合上述方法建立一个面向领域标签辅助的Web服务聚类方法DTWSC.实验结果表明,该框架能够提高针对不同类型的服务文档的聚类效果.与LDA、K-Means等方法相比,该方法在聚类纯度、熵和F-Measure指标上均具有更好的效果.

关键词: Web服务聚类;面向领域;标签辅助;主题模型
中图分类号: TP311.5 **文献标识码:** A **文章编号:** 0372-2112 (2015)07-1266-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.07.003

Domain-Oriented and Tag-Aided Web Service Clustering Method

TIAN Gang[1,2], HE Ke-qing[1], WANG Jian[1], SUN Cheng-ai[2], XU Jian-jian[2]
(1. *State Key Laboratory of Software Engineering School of Computer, Wuhan University, Wuhan, Hubei* 430072, *China*;
2. *College of Information Science, Shandong University of Science and Technology, Qingdao, Shandong* 266590, *China*)

Abstract: The growing number of web services puts forward higher requirements for searching desired web services and clustering Web services can greatly enhance the discovery of Web service. However, the existing clustering approaches are only for a single type of service documents, and they are lacking of considering the domain characteristic and the tags information of services. To solve these problems, the proposed approach constructs the feature vectors of Web service contents by using ontology empowered SVM and domain oriented feature dimension reduction technology. Then a tag aided service clustering model called T-LDA is proposed to construct the hidden topic representations of Web service and general topical information which has less discriminative power is normalized. Finally all methods mentioned above are combined to form the domain oriented and tag aided Web service clustering (DTWSC). Experimental results show that the proposed approach can improve the effect of clustering. Compared with the approaches of LDA and K-means, the proposed approach achieves better performance of the purity, entropy and F-measure.

Key words: Web service clustering; domain-oriented; tag aided; topical model

1 引言

面向服务的计算(Service-Oriented Computing)吸引了工业界越来越多的注意并且被应用到了多个领域[1].随着面向服务计算技术的快速发展,互联网上的Web服务呈现出快速增长的趋势.如截至2014年03月31日ProgrammableWeb(PWeb)上发布的Web API已有11222个,Mashup的数量也已经有7400个.同时,服务资源描述的多样化特征显著.例如PWeb中服务描述语言有WSDL、OWL-S、WSMO等,而占服务总数量超过60%的RESTful服务采用了自然语言来描述.服务规模的剧增和描述的多样化给用户准确、高效地发现服务资源增加了困难,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.

服务聚类是辅助服务发现的一种重要方法.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[2,3].目前,基于功能相似度的服务聚类方法已有大量研究.例如,文献[3]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5个关键特征,然后基于这些特征计算WSDL级别的服务相似度,

收稿日期:2014-10-14;修回日期:2015-02-03;责任编辑:孙瑶
基金项目:国家973重点基础研究发展计划(No.2014CB340404);国家自然科学基金(No.61373037,No.61202031);重点实验室开放课题(No.SKLSE 2014-10-07)

从而将服务聚类到功能相似的类簇.文献[4]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.

尽管基于 WSDL 的服务聚类方法提高了查询的效率,但是服务搜索的结果仍然不够理想.近年来,为 Web 2.0 的对象进行关键字(标签)标注技术开始变得流行起来.因为标签能够为被标注对象提供有含义的描述,所以它逐渐成为进行信息检索的最常用的文本特征.PWeb 在开发者注册服务的时候允许开发者对所注册的服务进行标注,即根据服务的功能为服务添加相关的标签,经过标签标注后的服务如图 1 所示.图 1 中的服务是一个关于旅游和住宿的服务,它的标签如图所示.如果我们在检索的时候使用标签数据,用户在检索旅游和住宿的时候都可能搜索到这个服务.另一方面,用户倾向于使用相同的标签标注功能类似的服务,这也帮助搜索引擎提供更多查询结果给用户.因此,在服务聚类的时候加入标签信息能够有效的帮助提高聚类的质量.

HotelsCombined:Highlights

Summary	Hotel comparis on search/aggregator
Category	Travel
Tags	travel hotel lodging api accommodation
Protocols	
Data Formats	
API home	http://www.hotescombined.com/Affiliates.aspx

图1　标签标注的服务实例

尽管现有的服务聚类方法在各自情境下取得了不错的效果,但是现有的服务聚类方法在以下三点考虑不足:

(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[3,4]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)没有充分利用 Web2.0 环境中的标签信息.服务描述的自然语言通常比较短而且无法充分提供代表服务功能的关键字信息,标签作为服务功能描述的重要补充,在服务聚类的时候能够起到重要的作用.现有方法缺少同时针对自然语言和标签的服务聚类方法.

(3)没有充分考虑服务的领域特性.现有的服务聚类方法大都是使用服务描述文档直接进行聚类,而没有考虑领域词汇对服务聚类的影响.比如,Financial 领域的"BankCheck"服务与 EMail 领域的"Data8 Email Validation"服务都具有"用户验证"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个类簇中,但是显然用户在检索 Financial 领域服务的时候,不希望得到 Email 领域的信息.

因此,面对互联网上服务的规模化增长,针对现有服务聚类方法中存在的不足,如何进行准确、高效的服务聚类成为一个极具挑战性的问题.针对上述问题,提出一种面向领域标签辅助的主题模型服务聚类方法.

2　标签辅助的服务主题聚类方法

为了实现面向领域、标签辅助的服务聚类,本文使用面向领域的特征降维方法建立服务描述的内容特征向量,利用扩展后的主题模型 ATM (Author Topic Model)[7,8]建立融合标签信息和服务描述特征向量的服务的隐含主题表示,并基于隐含主题表示实现了服务聚类.面向领域、标签辅助的服务聚类模型 DTWSC (Domain Oriented and Tag Aided Web Service Clustering)如图 2 所示.

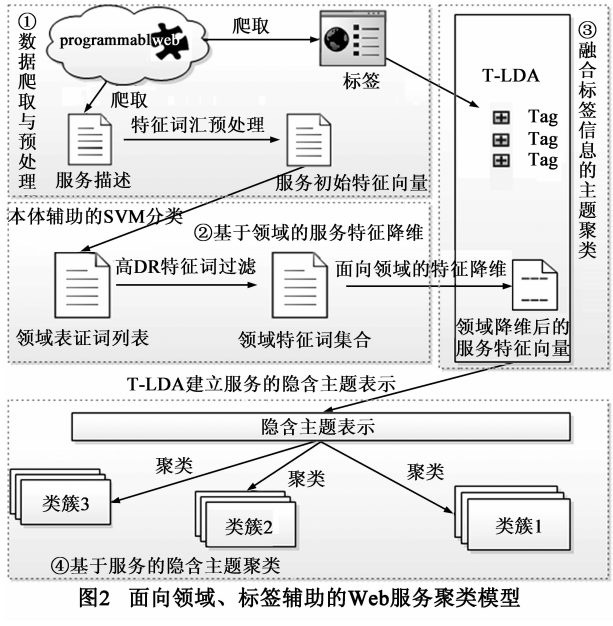


图2　面向领域、标签辅助的Web服务聚类模型

DTWSC 模型是一种四阶段的服务聚类模型:第一阶段为服务爬取和数据预处理阶段.首先使用爬虫技术从互联网上将服务的短文本描述和服务标签爬取下来,然后使用文本特征抽取方法预处理数据,建立能够表征服务内容的特征向量.第二阶段为面向领域的服务特征降维.使用文献[9]中的本体辅助式支持向量机(Ontology-Empowered SVM)进行面向领域的服务分类方法和特征降维方法建立服务的内容特征向量.第三阶段使用融合了标签信息的 LDA 模型即 T-LDA 基于内容特征向量和服务标签信息做模型学习并最终建立服务的隐含主题表示.第四阶段利用上一步骤建立的服务的隐含主题表示进行服务聚类,将服务文档组织成不同的类簇.这些类簇将会被用来促进 Web 服务搜索引擎的查询性能.因此,当一个用户使用 Web 服务搜索引擎进行查询的时候,融合了聚类信息之后的搜索引擎

会返回更加精确的查询结果.从图 2 中可以看出,在生成类簇之前的所有处理都是离线进行的,因此性能是可以保证的,在本文后续的讨论中,我们重点关注方法的准确性.

2.1 数据爬取与处理

在 DTWSC 的数据处理阶段,首先需要从互联网上爬取服务文本描述和服务的标签信息.在获得这些信息之后,需要按照如下的步骤抽取能够表征服务内容的特征向量.

(1)建立初始向量.在这一步中,我们采用自然语言处理工具包 NLTK* 将 Web 服务描述文档断词之后建立初始文档特征向量.

(2)词干还原.具有相同词干的单词往往具有相同的含义,例如 Searched 和 Searching 具有相同的词干 Search.因此,我们采用 NLTK 提供的 PorterStemmer 对步骤 1 中获得初始文档特征向量中的特征词进行词干还原.

(3)移除功能词.本步骤主要从文档特征向量中移除功能词.能够表征文档内容的词汇通常是名词、动词和形容词,而功能词如"a"、"the"等词汇对于表征文档内容基本上没有帮助,所以需要将这些功能词汇从文档内容向量中移除.我们使用 NLTK 的功能词处理方法,将相关的功能词从文档特征向量中移除.

2.2 面向领域特征降维

对服务文档进行面向领域的特征降维在服务聚类中具有重要作用:(1)通过降维能够加快 Gibbs 抽样的收敛速度,提高算法执行速度;(2)约减对表征服务内容贡献度不大的特征词汇,可以提高聚类准确度.因此,在得到服务内容表征向量之后,我们对服务内容表征向量进行面向领域的特征降维.

文献[9]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.使用本体辅助式 SVM 对服务进行分类后,进一步使用 KF-IRF (Keyword Frequency-Inverse Repository Frequency)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,得到分类后的领域服务集和相应的领域词汇排序表.

建立领域词汇排序表之后,进一步使用词汇的领域的表征度 DR(Degree of Representation)来筛选领域特征词汇.DR 指一个特征词 t_i对一个领域 d 的表征程度,其公式为:

$$DR(t_i,d)=\frac{|\{S_j|t_i\in S_j\wedge S_j\in DS_d\}|}{|DS_s|} \tag{1}$$

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.因此 $DR(t_i,d)$越大,说明 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.例如 Weather 领域中的"weather"在该领域的大部分服务中都会出现.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,所以进行领域服务聚类的时候反而意义不大.因此,我们利用经验阈值 th (0~1 之间的百分数)将 DR 大于 th 的特征词去除,从而实现了面向领域的服务特征降维.

2.3 融合标签信息的服务主题模型 T-LDA

T-LDA 是基于 ATM 模型的可融合标签和服务内容向量进行主题聚类的方法,其概率图模型如图 3 所示.

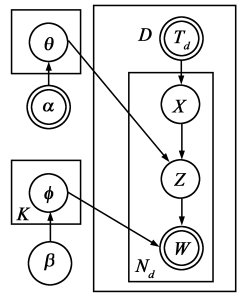


图3 LDA模型和T-LDA模型

T-LDA 同[10]中的 WT-LDA 不同,在 T-LDA 中标签被作为了文档级别的观测量,而 WT-LDA 是全局的.如图 3 所示,Tag_d是在每一个服务中都出现的观测量.隐变量 x 表示从标签集合 Tag_d中选取并且同给定单词相关的标签,每个标签都和一个在主题上分布 θ 相关,θ 是从一个先验参数为 α 的狄利克雷分布中选择出的.x 和 θ 一起用来选择主题 z,然后和 z 相关的分布 φ 被用来产生单词 w,分布 φ 是从一个先验参数为 β 的狄利克雷分布中得出的.图 3 对应的 T-LDA 的生成过程为:

(1)为每一个标签 $t=1,\cdots,T$ 选择 $\theta_t\sim$ Dirichlet(α);

(2)为每一个主题 $k=1,\cdots,K$ 选择 $\varphi_k\sim$ Dirichlet(β);

(3)对于每一个服务 $d=1,\cdots,D$

(a)给定每个文档的标签向量 t_d,

(b)对于服务 d 中每个单词 $w_{di}, i=1,\cdots,N_d$

(ⅰ)选择一个标签 $x_{di}, x_{di}\sim$ Uniform(t_d)

(ⅱ)选择一个主题 $z_{di}, z_{di}\sim$ Multi($\theta_{x_{di}}$)

(ⅲ)选择一个单词 $w_{di}, w_{di}\sim$ Multi($\varphi_{z_{di}}$)

其中,Θ 为从一个先验参数为 α 的狄利克雷分布产生的标签在主题上的分布.θ_t为给定标签 t 之后主题的概率分布,Φ 为从一个先验参数为 β 的狄利克雷分布产生的主题在单词上的分布.φ_k为给定主题 k 之后主题在单词上的概率分布.Dirichlet(·)为狄利克雷分布,Uniform(·)为均匀分布,Multi(·)为多项分布.

在上述生成过程中,每个主题都是给定 Θ 之后独

* http://www.nltk.org/

立获得,每个单词是在给定 Φ 和 z 之后对获得.那么给定 Θ,Φ,t_d 之后的所有服务的条件概率为:

$$\mathrm{P}(w \mid \Theta,\Phi,T) = \prod_{d=1}^{D}\mathrm{P}(w_d \mid \Theta,\Phi,t_d) \tag{2}$$

在给定 Θ,Φ,t_d 之后,每个服务中单词的条件概率可以通过累加隐变量 x 和 z 取得:

$$\begin{aligned}&\mathrm{P}(w_d \mid \Theta,\Phi,t_d)\\ &= \prod_{i=1}^{N_d}\mathrm{P}(w_{di} \mid \Theta,\Phi,t_d)\\ &= \prod_{i=1}^{N_d}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{P}(w_{di},z_{di}=k,x_{di}=t \mid \Theta,\Phi,t_d)\\ &= \prod_{i=1}^{N_d}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{P}(w_{di} \mid z_{di}=k,\Theta)\\ &\cdot \mathrm{P}(z_{di}=k \mid x_{di}=t,\Theta)\cdot \mathrm{P}(x_{di}=t \mid t_d)\\ &= \prod_{i=1}^{N_d}\frac{1}{T_d}\sum_{t\in t_d}\sum_{k=1}^{K}\varphi_{w_{di}k}\theta_{kt}\end{aligned} \tag{3}$$

等式(2)和(3)可以用来计算服务在给定 Θ,Φ,t_d 之后概率,Θ,Φ 是 T-LDA 的参数,则式(3)即为似然函数.通过极大似然估计或者最大后验概率的方法可以进行参数估计.然而 T-LDA 主题模型很难进行直接推理,通常使用变分推理和 EM 算法以及吉布斯(Gibbs)抽样方法[11].本文采用 Gibbs 抽样方法估计 T-LDA 模型参数.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.对于 T-LDA 模型,每一步 Gibbs 抽样都服从如下分布:

$$\begin{aligned}&\mathrm{P}(z_{di}=k,x_{di}=k \mid w_{di}=w,z_{-di},x_{-di},w_{-di},\alpha,\beta,t_d)\\ &\propto \frac{C^{WK}_{uk,-di}+\beta}{C^{WK}_{w'k,-di}+w\beta}\times\frac{C^{KT}_{kt,-di}+\alpha}{C^{KT}_{k't,-di}+K\alpha}\end{aligned} \tag{4}$$

$C^{KT}_{\cdot\cdot}$表示主题 – 标签矩阵,$C^{KT}_{kt,-di}$表示排除掉单词 w_{di}后标签 t 分配到主题 k 下单词的总数.$C^{WK}_{\cdot\cdot}$表示单词-主题矩阵,$C^{WK}_{uk,-di}$表示排除掉单词 w_{di}后主题 k 对应单词的总数.x_{di}为每个单词标签指派,z_{di}为每个单词主题指派.

给定 $x,z,D^{\text{train}},\alpha,\beta$,根据狄利克雷分布和多项分布是共轭分布,可以直接计算 φ_k,θ_t,得到:

$$\varphi_k \mid z,D^{\text{train}},\beta \sim \text{Dirichlet}(C^{WK}_{\cdot K}+\beta) \tag{5}$$

$$\theta_t \mid x,z,D^{\text{train}},\alpha \sim \text{Dirichlet}(C^{KT}_{\cdot t}+\alpha) \tag{6}$$

其中 $C^{WK}_{\cdot\cdot}$是每个分配给主题 K 的单词出现的次数组成的向量.$C^{KT}_{\cdot\cdot}$为每个分配给标签 T 的主题出现的次数组成的向量.根据式(5)(6),计算对应参数的期望为:

$$E[\varphi_{wk} \mid Z^S,D^{\text{train}},\beta] = \frac{(C^{WK}_{uk,-di})^s+\beta}{(\sum_{w'}C^{WK}_{w'k,-di})^s+w\beta} \tag{7}$$

$$E[\theta_{kt} \mid X^S,Z^S,D^{\text{train}},\alpha] = \frac{(C^{KT}_{kt,-di})^s+\alpha}{(\sum_{k}C^{KT}_{kt,-di})^s+K\alpha} \tag{8}$$

2.4　建立服务的隐主题表示

T-LDA 可以采用一种矩阵因子分解的方式来描述[7,8].如图 4 所示在 T-LDA 中,代表服务和单词的矩阵 $\boldsymbol{P}$ 可以分解成三个矩阵的乘积 $\boldsymbol{\Phi}\times\boldsymbol{\Theta}\times\boldsymbol{T}$.其中 $\boldsymbol{P}\in W\times D$ 为单词在文档中分布,$\boldsymbol{\Phi}\in W\times T$ 为主题下的单词分布,$\boldsymbol{\Theta}\in K\times T$ 为标签在主题下的分布.$\boldsymbol{T}\in T\times D$ 为标签在文档中的分布.其中矩阵 $\boldsymbol{T}$ 表示了每个服务描述中标签的均匀分布.每一个 $\boldsymbol{T}_{td}$的取值如式(9)所示:

$$\boldsymbol{T}_{td}=\begin{cases}\dfrac{1}{|\boldsymbol{T}_d|}, & t\in\boldsymbol{T}_d\\ 0, & \text{其他情况}\end{cases} \tag{9}$$

因此,在 T-LDA 模型的学习中,根据式(9)能够获得标签主题分布 $\boldsymbol{\Theta}$,由式(9)能够取得文档标签分布 $\boldsymbol{T}$,令 $\boldsymbol{\Theta}'=\boldsymbol{\Theta}\times\boldsymbol{T}$,计算的结果即为服务在主题上的分布,其中 $\boldsymbol{\Theta}'\in K\times D$,下一步将利用 $\boldsymbol{\Theta}'$进行服务聚类.

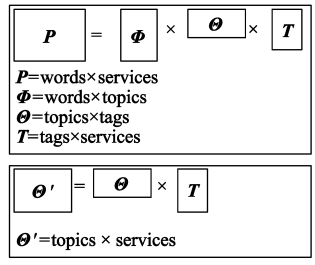


图4　T-LDA的矩阵因子分解解释

2.5　利用服务的隐主题表示聚类

根据上文的分析,计算得到服务在主题上的分布 $\boldsymbol{\Theta}'\in K\times D$.根据服务在主题上的分布进行聚类的方法有很多,其中文献[12]认为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(Topic Cluster),可以通过式(10)获得服务所属的主题类簇:

$$TC(S_i)=T_k\cap\forall j((j\neq k)\rightarrow \mathrm{P}(S_i,T_j)<\mathrm{P}(S_i,T_k)) \tag{10}$$

其中,$1\leqslant j,k\leqslant K$,$\mathrm{P}(S_i,T_j)$表示服务 S_i包含主题 T_j的概率,服务 S_i包含 K 个主题 $T_1,T_2,\cdots,T_K$.

文献[10]提供了另外一种对服务聚类的方法,将 WT-LDA 学习得到的参数 Θ 进行累计从而得到文档的主题分布,其中 $x\in t_d$,表示 x 属于文档 d 所包含的标签.

为了取得更好的性能,经过试验比较,在本文中采用不同于文献[10]的方法.在我们的方法中,Θ'是服务在主题上的分布,因此每个服务可以表示成式(11)的形式:

$$WS_d = \left[\frac{\theta_{d,1}}{\sum_i\theta_{i,1}},\frac{\theta_{d,2}}{\sum_i\theta_{i,2}},\cdots,\frac{\theta_{d,K}}{\sum_i\theta_{i,K}}\right] \tag{11}$$

其中 $\sum_i\theta_{i,j}$是归一化参数,目的是为了降低某些太泛化的主题的重要性.因为有些泛化的主题包含的单词

分布为功能词或者公共词条,但是却缺乏表征服务内容的能力.

在得到使用主题表征的文档向量之后,采用传统的 KMeans 聚类算法计算文档的类簇.面向领域标签辅助的服务聚类算法如算法 1 所示.第 1 步使用本体辅助的支持向量机得到领域服务集和领域词汇列表.第 2 ~ 4 步使用经验阈值对服务特征词汇过滤,第 5 步完成面向领域的服务特征降维,建立服务的内容特征向量.第 6 ~ 8 步利用 T-LDA 融合标签和服务特征向量,通过模型学习建立服务的隐式主题表示.

算法 1 面向领域标签辅助的服务聚类算法

输入:服务集 SS,标签集合 Tags,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←ontology-empoweredSVM(SS);
2 FOREACH 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 ← t;//将 t 加入领域特征词集 DFTS
5 WSContentFeature←filter(DS,DFTS);//服务特征降维
6 θ←T-LDA(WSContentFeature,Tags,α,β,TNum);//得到服务主题类簇
7 $\Theta' = \Theta \times T$;
8 RETURN Θ'.

在得到使用主题表征的文档向量之后,采用传统的 KMeans 聚类算法计算文档的类簇.聚类得到的类簇需要通过算法与标准分类进行比对以确定算法的性能,例如标准分类 Tools 可能和聚类类簇 1,2,3 都有交集,但是 Tools 和类簇 3 的 $\frac{|p \cap c|}{|p|}$ 值最大,那么就认为类簇 3 即为 Tools 分类,采用算法 1 实现类簇与标准分类的匹配.在算法 1 中,P 为标准分类集合,C 为聚类类簇集合.步骤 1 ~ 4 用来计算标准分类和聚类类簇之间的 $\frac{|p \cap c|}{|p|}$ 的取值.其中 $|p \cap c|$ 为标准分类和类簇的交集中包含的服务数目,$|p|$ 为标准分类中服务的数目.算法的 5 ~ 8 步骤会根据 $\frac{|p \cap c|}{|p|}$ 取值计算得到 p 和 c 的匹配.首先获得 $\frac{|p \cap c|}{|p|}$ 取值最大的 $(p_{\max}, c_{\max})$ 对,然后设置 $p_{\max}$ 和 $c_{\max}$ 为一对匹配,然后将 $p_{\max}$ 和 $c_{\max}$ 分别从标准分类和类簇集合中删除,重复上述过程直至全部标准分类和类簇得到匹配.

算法 2 类簇与标准分类的匹配

输入:实际分类 P,聚类类簇 C
输出:实际分类与聚类类簇匹配结果 match

1 FOREACH $p \in P$ DO
2 match(p)←$\varnothing$
3 FOREACH $c \in C$ DO
4 overlap(p,c)←$\frac{|p \cap c|}{|p|}$
5 WHILE overlap≠$\varnothing$ DO
6 $(p_{\max}, c_{\max})$←GetMaxOverlap(overlap)
7 match($p_{\max}$)← $c_{\max}$
8 overlap←overlap − {overlap($p_{\max,*}$),overlap($*$,$c_{\max}$)}

3 实验评价

3.1 实验方法

文中涉及的算法都是通过 Java 编程语言实现,我们的实验环境是:Intel(R) Core(TM)i5 M460@2.53GHz,内存 4G,myEclipse8.6.

我们从 PWeb 中爬取到了 10050 个 Web 服务信息,因为不同领域的服务数目差别很大(Tools 有 761 个,Portal 只有 1 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.所以选取包含 Web 服务数目前十的领域作为实验数据一共包含 4402 个 Web 服务.每类服务包含的服务数量和每类服务下面包含的标签数量如表 1 所示,对这些领域使用本体辅助式 SVM 进行分类.

表 1 服务数量前十的分类以及其包含的服务和标签数量

服务分类	服务数量	标签数量
Tools	761	2879
Internet	600	1943
Social	491	1660
Financial	465	1662
Enterprise	444	1647
Mapping	349	1293
Reference	337	1143
Shopping	331	1093
Government	315	984
Science	309	979

3.2 评价指标

本文使用 F-measure,纯度和熵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.其中准确率(Precision)和召回率(Recall)的计算公式如下:

$$\text{Precision}_{c_i} = \frac{\text{hit}(c_i)}{\text{hit}(c_i) + \text{mishit}(c_i)} \tag{12}$$

$$\text{Recall}_{c_i} = \frac{\text{hit}(c_i)}{\text{hit}(c_i) + \text{missed}(c_i)} \tag{13}$$

其中 hit(c_i)指的是正确分入 c_i中的服务数目,mishit(c_i)表示错误的分入 c_i中的服务. missed(c_i)指的是应该分入 c_i却被分配到其他类簇中的服务. F-measure 是通过综合考虑 Precision 和 Recall 的调和平均数,其计算公式如下:

$$\text{F-measure}_{c_i} = \frac{2 \times \text{Precision}_{c_i} \times \text{Recall}_{c_i}}{\text{Precision}_{c_i} + \text{Recall}_{c_i}} \tag{14}$$

聚类纯度 PC(Purity of Cluster)用来评价聚类的质量,设类簇 c_i中元素个数为 n_i,那么每个主题类簇的聚类纯度和所有类簇的平均聚类纯度定义为:

$$Pu(c_i) = \frac{1}{n_i} \times \max_j(n_i^j) \tag{15}$$

$$\text{purity} = \sum_{i=1}^{k} \frac{n_i}{n} Pu(c_i) \tag{16}$$

n_i表示主题类簇 c_i中包含的服务数目,n_i^j 指的 j 个分类中成功分入类簇 c_i中的服务数目.使用相似的表示符号,每个类簇的熵和所有类簇的熵表示为:

$$E(c_i) = -\frac{1}{\log(q)} \sum_{j=1}^{q} \frac{n_i^j}{n_i} \log\left(\frac{n_i^j}{n_i}\right) \tag{17}$$

$$\text{entropy} = \sum_{i=1}^{k} \frac{n_i}{n} E(c_i) \tag{18}$$

3.3　结果与分析

3.3.1　服务聚类的三种方法比较

因为 K-Means 和 LDA 算法聚类结果具有不稳定性,我们对相关方法进行了 100 次聚类实验,最后绘制盒线图来进行对比分析.在这一部分,把 DTWSC 方法同现有研究中四种常用的服务聚类方法进行比较,这几种方法的分别包括:

(1)K-Means 聚类算法. K-Means 算法是应用最广泛的一种基于划分的聚类算法.文献[10,13]中都采用该方法实现了对服务的聚类.

(2)Agglomerative 层次聚类算法. Agglomerative 是一种自底向上的层次聚类算法,文献[14]都采用了该方法进行服务聚类分析.

(3)LDA 是一种无监督的机器学习方法,通过学习能够建立一个三层贝叶斯概率模型,即词、主题和文档三层结构.通过该三层结构实现单词或文档的聚类.文献[2,12]采用这种方法实现了对服务的聚类分析.

使用 K-Means 算法、Agglomerative 算法,LDA 算法和本文方法聚类结果对比如图 5 所示.从图中可以看出,针对三类评测指标采用 DTWSC 算法得到的聚类结果都比另外三种方法好.

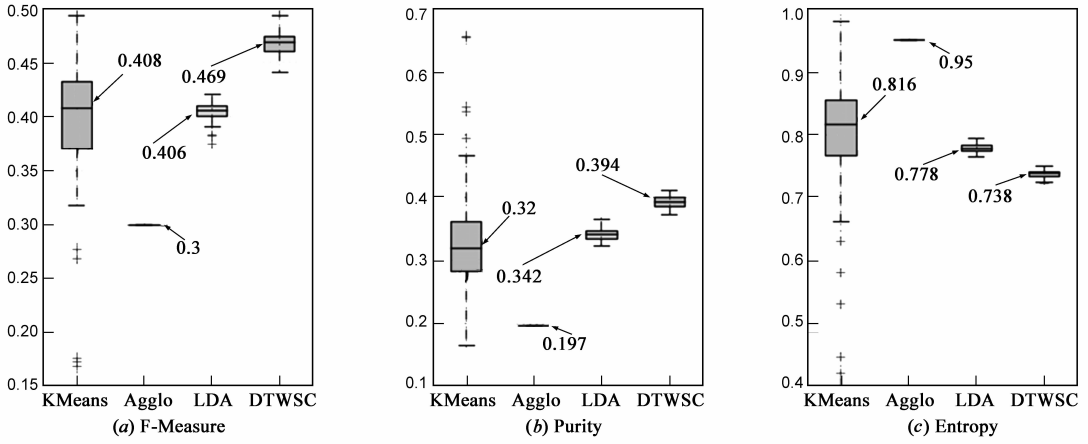


图5　算法性能比较

就 F-Measure 而言,采用 DTWSC 算法比聚类效果第二好的 K-Means 方法的 F-Measure 值提高了 7 个百分点,然而从统计特征上看,DTWSC 得到的 F-Measure 值分布更集中,奇异值更少,偏差更小,因此算法更加稳定.就聚类纯度而言,DTWSC 比聚类效果第二好的 LDA 方法提高了 5 个百分点,这说明引入标签信息和领域信息能够提供服务的聚类效果.就信息熵而言,DTWSC 方法比聚类效果第二好的 LDA 方法提高了 4 个百分点.对比四种方法,发现单纯采用 K-Means 方法,聚类效果波动较大,会产生很多异常值.而 Agglomerative 在本文所采用的数据集上表现并不好,这可能与服务描述文档的特征词的划分关系并不好有关. LDA 方法和 DTWSC 方法在本文的数据集上表现要更加稳定,聚类结果的分布要更加集中,而通过对 LDA 方法添加标签和领域信息而得到的 DTWSC 方法取得了最好的聚类结果.

3.3.2　主题分布聚类算法比较

使用 LDA 方法对服务描述降维之后再进行聚类的

方法在 2.5 节已经介绍过,不同文献采用了不同的方法.针对文献[10,12]和本文采用的方法,我们做了相关实验进行评测.文献[12]认为如果一个服务隶属于它包含的概率最大的主题.文献[10]将 WT-LDA 的参数 Θ 进行累加从而得到文档的分布表示.本文采用一种归一化的服务-主题分布表示形式,三种方法聚类效果如图 6 所示.从图中可以看出,文献[12]的方法在三种评价指标下的表现最差,主要表现在得到评测值分布较分散,中位数也小于另外两种方法.因为 LDA 是一种混合模型,一个文档可以表示成若干主题的混合,因此只用一个主题来代表一个文档是不够全面的,这也解释了文献[12]中的方法在三种方法中表现最差的原因.同文献[10]中的方法相比,DTWSC 利用矩阵分解的方法将标签-主题分布转化成文档主题分布,并且进行了归一化处理,消除了异常主题的影响.因此从图 6 中可以看出,该方法较文献[10]的方法有了进一步的提升.这种聚类效果的提升主要来自于归一化对异常主题的控制.

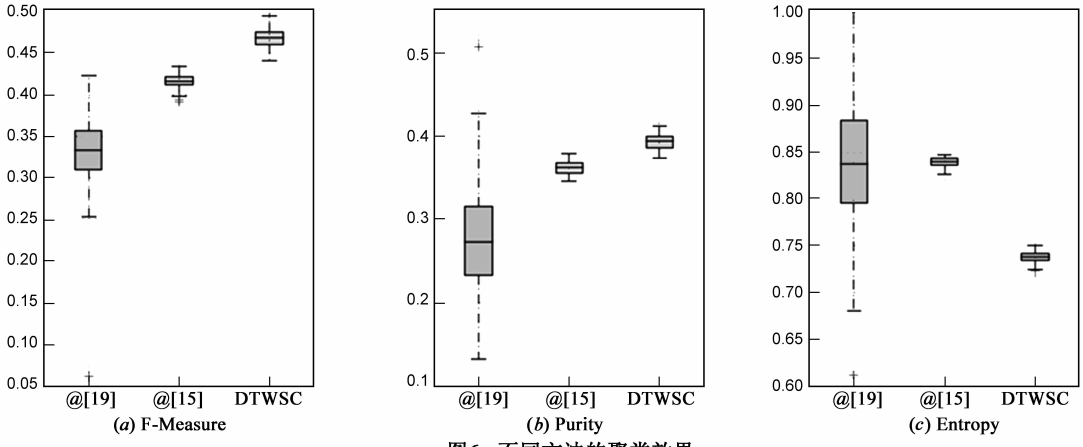


图6 不同方法的聚类效果

3.3.3 参数影响

(1)主题数

基于贝叶斯模型[15]选择的方法,通过计算给定观察数据之后模型的后验概率可以用来寻找合适的参数.首先,通过实验 K 取不同的值分别运行 Gibbs 抽样算法,检测 $\log(P(w|K))$的值的变化.实验结果如图 7 所示.

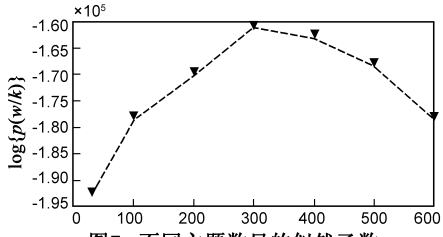


图7 不同主题数目的似然函数

当主题数目取 300 的时候,后验概率取得最好的结果,LDA 模型对于语料库数据取得最佳的拟合度,此时生成文档的能力也最强.因此在本文的算法中,K 的取值为 300. 在实验中发现,K 的取值对文本语料集很敏感,K 所代表的是主题的个数,语料集的选取对其影响会比较大,不同的语料集上会有不同的最佳 K 值,如果所选取的语料集包含的领域知识范围很广,那么就需要相对较多的主题值来对其进行表达,这样文本在不同主题上的分布才会有更好的表达.

(2)超参数

在使用 LDA 模型进行无监督学习的时候,超参数 α 和 β 最常用的经验值为 $\alpha = 50/K$, $\beta = 0.01$[15].在本文的实验中,也将超参数设置如上.

(3)表征度阈值 th 对聚类效果的影响

表征度阈值 th 能够过滤表征一个领域的特征词,不同的阈值会影响聚类的精读.图 8 给出了表征度阈值 th 的不同取值对服务聚类效果的影响.从图 8 可以看出,当阈值 th 取 80%时,聚类效果较好,因为去除了在每个领域的大多数服务描述中都出现但对聚类意义不大的词.如果 th 小于 80%,大量领域核心词汇进入筛选列表,导致很多体现服务特征的词汇被过滤掉,因此服务聚类的效果并不好.

4 相关工作

服务聚类是一种有效地辅助服务发现和服务推荐

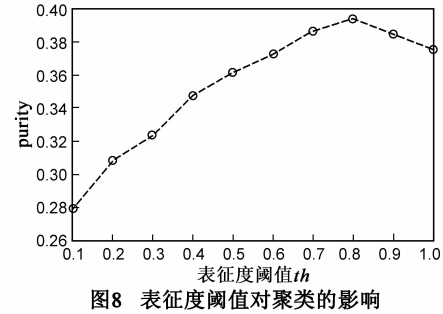


图8 表征度阈值对聚类的影响

的技术[3,16],在这方面,国内外近年来已有大量研究.文献[16]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[17]建立一种统计模型来动态计算相似服务从而促进搜索引擎发现服务的能力.文献[14]使用层次 Agglomerative 算法对功能相似的服务进行聚类,以改进服务发现效率.文献[13]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[10]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.文献[18]从服务功能相似和过程相似两个层面对服务进行聚类,从而降低服务发现的搜索空间,提高服务发现效率.文献[19]基于支配(Dominance)服务概念提出一种服务聚类方法,通过服务支配关系发现服务与用户请求的相关性.文献[20]将服务库中服务聚类形成服务簇,并建立相应的服务簇网元及其矩阵模型,提出一种基于服务簇的服务替换方法.

目前,使用 LDA 进行聚类的研究有很多,例如文献[21]使用 Label-LDA 探讨了用户兴趣挖掘的新方法然,而基于 LDA 进行服务聚类的研究并不是很多.其中,文献[6]使用 PLSA(Probabilistic Latent Semantic Analysis)和 LDA 从服务描述中发现潜在主题,然后利用主题对基于 OWL-S 服务的 Profile 描述和功能描述两个方面进行聚类.文献[22]直接用 LDA 方法将 WSDL 描述文档建模为层次结构化文本文档,得到每个主题对应的关键词分布,然后基于主题对服务进行检索.

文献[2]将 WSDL 文档预处理获得表征服务功能的特征向量,利用 WT-LDA 模型中引入标签信息实现对服务的聚类.该文献的方法同本文方法有相似的地方,但是存在如下几个重要区别:(1)本文方法基于 Web 服务的自然语言描述,而文献[2]的方法主要针对 WSDL 文档;(2)本文的方法在聚类的时候考虑了领域特征,而文献[2]中并没有针对服务的领域特性做任何处理;(3)本文在基于隐含主题做服务聚类的时候采用同文献[2]不同的文档表达方法,通过引入矩阵分解解释和归一化的处理方法,消除了异常主题对文档聚类的影响,实验也表明了本文方法的有效性.

上述方法在各自的情境取得了不错的效果,但是在如下方面仍然有所欠缺:(1)参与聚类的服务文档类型单一,比如文献[10,14,17]针对 WSDL,文献[19]针对 OWL-S,还有些针对仿真服务[18];(2)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;(3)现有方法在聚类过程中缺少对领域特征的考虑;(4)尽管已有方法已经在考虑使用标签信息提高服务发现的效率[2],但是对标签数据在服务聚类中的应用研究仍然不够.

针对上述问题,本文在服务发现过程中,引入面向领域的特征约减方法提高参与聚类的特征向量的领域特性,引入标签信息辅助提高服务聚类效果.同时,本文使用 T-LDA 模型将服务描述和标签信息结合共同参与聚类,并对降维之后的结果使用矩阵分解技术表达文档,然后利用归一化方法处理异常主题从而提升聚类效果.

5　总结

本文基于服务自然语言描述和标签信息,利用面向领域的服务特征过滤方法,提出了一种面向领域,标签辅助的服务发现模型 DTWSC,然后基于该模型对服务进行离线的面向主题的聚类,从而具有相似功能描述的服务组织为类簇.最后,以 PWeb 上真实的服务集进行实验,验证了 DTWSC 服务聚类方法的可行性和有效性.性能对比实验分析表明,本文提出的面向领域标签辅助的方法在纯度、熵、F-measure 方面均具有更好的效果.而且,本文方法有助于异构服务资源的组织管理,从而促进异构服务发现,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)在短文本描述对服务的内容表征能力不够的情况下,如何服务发现的能力.

参考文献

[1] L-J Zhang, J Zhang, H Cai. Services Computing[M]. Beijing: Tsinghua University, 2007.

[2] Chen Liang, Hu Liukai, Zheng Zibin, et al. WTCluster: Utilizing tags for Web services clustering[A]. Proceedings of International Conference on Service-Oriented Computing [C]. Berlin: Springer, 2011. 204 – 218.

[3] Elgazzar K, Hassan A E, Martin P. Clustering WSDL documents to bootstrap the discovery of web services[A]. Proceedings of International Conference on Web Services[C]. USA: Piscataway, 2010. 147 – 154.

[4] Yu Q, Rege M. On service community learning: A co-clustering approach[A]. Proceedings of IEEE International Conference on Web Services[C]. USA: Piscataway, 2010. 283 – 290.

[5] Liu Jianxiao, He Keqing, Wang Jian, et al. A clustering method for web service discovery [A]. Proceedings of International Conference on Services Computing [C]. USA: Piscataway, 2011. 729 – 730.

[6] Cassar G, Barnaghi P, Moessner K. Probabilistic methods for service clustering[A]. Proceedings of International Workshop on Semantic Web Service Matchmaking and Resource Retrieval [C]. Shanghai: SRI, 2010. 4 – 20.

[7] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(2): 993 – 1022.

[8] Rosen-Zvi M, Griths T, Steyvers M, Smyth P. The author-topic model for authors and documents[A]. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence[C]. USA: UAI, 2004. 487 – 494.

[9] Wang Jian, Zhang Jia, Hung P C K, et al. Leveraging fragmental semantic data to enhance services discovery[A]. Proceedings of the 13th International Conference on High Performance Computing and Communications[C]. Piscataway, NJ: IEEE, 2011. 687 – 694.

[10] Chen L, Wang Y, Yu Q, et al. WT-LDA: User Tagging Augmented LDA for Web Service Clustering[M]. USA: Service-Oriented Computing, 2013. 162 – 176.

[11] Canny J. GaP: a factor model for discrete data[A]. In SIGIR′04: Proceedings of International Conference on Research and Development in Information Retrieval[C]. New York, NY: ACM Press, 2004. 122 – 129.

[12] 李征,王健,等.一种面向主题的领域服务聚类方法[J].计算机研究与发展,2014,51(2):408 – 419.
Li Zheng, Wang Jian, et al. A topic-oriented clustering approach for domain services[J]. Journal of Computer Research and Development, 2014, 51(2): 408 – 419. (in Chinese)

[13] Wang Xianzhi, Wang Zhongjie, Xu Xiaofei. Semi-empirical service composition: a clustering based approach[A]. Proceedings of International Conference on Web Services[C]. Piscataway, NJ: IEEE, 2011. 219 – 226.

[14] Richi N, Bryan L. Web service discovery with additional semantics and clustering[A]. Proceedings of International Conference on Web Intelligence[C]. Piscataway, NJ: IEEE, 2007. 555 – 558.

[15] Griffiths T L, Steyvers M. Finding scientific topics[A]. Proceedings of the National Academy of Sciences of the United States of America[C]. USA: NCBI, 2004, Vol. 101. 5228 – 5235.

[16] Dasgupta S, Bhat S, Lee Y. Taxonomic clustering and query matching for efficient service discovery[A]. Proceedings of Conference on Web Services[C]. Piscataway, NJ: IEEE, 2011. 363 – 370.

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

[18] 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.
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)

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

[20] 杜玉越,薛洁,李彦成.基于服务簇的服务组合替换与分析[J].电子学报,2014,42(11):2231 – 2238.
DU Yu-yue, XUE Jie, LI Yan-cheng. Substitution andanalysis of service composition based on service clusters[J]. Acta Electronica Sinica, 2014, 42(11): 2231 – 2238. (in Cineses)

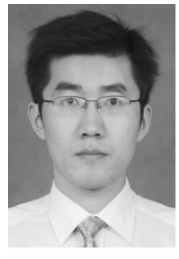
[21] 江雨燕,李平,王清.基于共享背景主题的 Labeled LDA 模型[J].电子学报,2013,41(9):1794 – 1799.
JIANG Yu-yan, LI Ping, WANG Qing. Labeled LDA model based on shared background topics[J]. Acta Electronica Sinica, 2013, 41(9): 1794 – 1799. (in Cineses)

[22] 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.
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)

作者简介

田 刚 男,1982 年 1 月出生,山东青州人,武汉大学在读博士研究生,CCF 会员,主要研究领域为服务计算、知识工程、机器学习等.

何克清 男,武汉大学计算机学院教授,CCF 高级会员,主要研究领域为服务计算、软件工程等.



王 健(通讯作者) 男,武汉大学计算机学院讲师,CCF 会员,主要研究领域为服务计算、需求工程等.
E-mail: jianwang@whu.edu.cn