文章编号:1009-3087(2015)05-0116-07 DOI:10.15961/j.jsuese.2015.05.017

迁移知识辅助的语义稀疏服务聚类方法

田 刚[1,2],何克清[1],高 莹[2],黄 颖[1,3]
(1.武汉大学 软件工程国家重点实验室 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;3.赣南师范学院 数计学院,江西 赣州 341000)

摘 要:现有服务聚类方法缺乏对服务描述语义稀疏情境下的研究,因此将迁移学习技术应用到服务聚类领域,尝试解决语义稀疏服务聚类的问题。通过对偶PLSA模型将目标领域和辅助领域语料知识进行融合,利用无监督的方式迁移辅助领域知识,从而提高目标领域语义稀疏服务聚类的能力。实验结果表明,该方法能够提高语义稀疏服务的聚类效果。与*K*-Means、Agglomerative和PLSA等方法相比,该方法在聚类纯度、熵上均具有更好的性能。

关键词:Web服务聚类;迁移学习;语义稀疏
中图分类号:TP311 **文献标志码**:A

Transferred Knowledge Aided Semantic Sparse Service Clustering

TIAN Gang[1,2], *HE Keqing*[1], *GAO Ying*[2], *HUANG Ying*[1,3]
(1. State Key Lab. of Software Eng., School of Computer, Wuhan Univ., Wuhan 430072, China;
2. College of Info. Sci. and Eng., Shandong Univ. of Sci. and Technol., Qingdao 266590, China;
3. Inst. of Mathematical and Computer Sci., Gannan Normal Univ., Ganzhou 341000, China)

Abstract: The existing clustering approaches are lacking of researching on clustering Web services whose descriptions are semantic sparse. Therefore, a new approach was proposed to makes an attempt to solve the problem of clustering sematic sparse Web service by applying transfer learning method in the domain of Web service clustering. A dual PLSA model was introduced to integrate knowledge of target domain and auxiliary domain which can transfer knowledge using a unsupervised mode to facilitate the process of semantic sparse Web service clustering. Experimental results showed that the proposed method can improve the performance of semantic sparse Web service clustering. Compared with the approaches of *K*-means, Agglomerative and PLSA, the proposed approach achieves better performance of the purity and the Entropy.

Key words: Web service clustering; transfer learning; semantic sparse

随着软件即服务(software as a service, SaaS)、云计算和SOA(service oriented architecture)的飞速发展,面向服务的软件开发技术正逐步成为互联网上的软件开发技术的主流。在这种趋势下,互联网上服务资源的规模呈现出快速增长的趋势,如截止2014年3月31日,著名的服务注册网站ProgrammableWeb上发布的Web API已有11 222个,Mashup的数量也已经有7 400个,其Web服务的数目相比于2011年9月份的3 815涨幅达到了196%。服务规模的剧增给用户准确、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。

将相似功能的服务聚类到一起能够显著改善Web服务搜索引擎的能力[1-2]。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[1]解析WSDL文档,从中抽取表达服务功能的5个关键特征,利用这些特征建立服务的特征表达并计算服务相似度,从而将服务组织到到功能相似的

收稿日期:2014-12-16
基金项目:国家重点基础研究发展计划资助项目(2014CB340404);国家自然科学基金资助项目(61202031);江西省自然科学基金资助项目(20142BAB217028);软件工程国家重点实验室开放课题资助项目(SKLSE 2014-10-07)
作者简介:田 刚(1982—),男,博士生,讲师.研究方向:服务计算;知识工程;机器学习. E-mail:tiangang@whu.edu.cn

类簇。实验表明,服务聚类提升了服务发现的效率。

尽管现有的服务聚类方法在各自情景下提高了服务发现的能力,但是在面对服务的描述文件语义稀疏的时候仍然表现出一定的不足。其主要原因是当文本资料语义稀疏的时候,很多传统的聚类算法会因为缺乏足够的统计信息而无法得到满意的结果[3]。例如 ProgrammableWeb 提供了针对各类服务的自然语言描述信息,但是这些自然语言描述常常都是短文本,因此基于这些短文本描述的查询效果并不好。主要原因就是这些短文本描述的语义太过稀疏(包含服务最多的前 10 个分类中服务描述的平均长度仅为 72 个单词)。

因此,面对互联网上不断增长的服务规模,针对现有服务聚类方法中在语义稀疏情境下存在的不足,如何更好地提高服务发现的效果成为一个极具挑战性的问题。

服务聚类能够有效地辅助服务发现和服务推荐[1-2],在这方面,国内外近年来已有大量研究。现有工作常常抽取 WSDL 文件中的内容特征如服务名、服务功能、服务消息、服务运行时 binding 信息来建立服务的特征向量[1-2]。文献[2]基于抽取的服务特征提出一种 WTCluster 方法,融合 WSDL 相似和标签(tag)相似性,利用 *K*-Means 聚类算法对服务进行聚类。Cao 等[4]提出一种应用于 Info-*K*means 的增量学习方法,从而避免了传统 *K*-Means 算法中因为中心点选取 0 值特征向量导致 *KL* 散度值无穷大的问题。同时 V-SAIL 算法和多线程模式的 PV - SAIL 算法,提高了聚类的效果和速度。

Richi 等[5]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。Platzer 等[6]也采用层次聚类算法对服务进行功能聚类,并使用多维度的相似度量方法度量服务的相似程度。层次模型聚类的结果可以帮助以层次的方式组织服务,能够有效地提高服务发现的性能。而 Dasgupta 等更是建立一种自组织的分类(taxonomic)聚类算法,针对语义 Web 服务进行服务聚类和分类组织,从而有效地促进了语义服务发现的性能[7]。

目前,基于潜在主题模型进行服务聚类的研究并不是很多。其中,Cassar 等[8]使用 PLSA 和 LDA (latent dirichlet allocation)[9]从服务描述中发现潜在主题,然后利用主题对服务的 Profile 描述和功能描述 2 个方面进行聚类。Chen 等[10]利用文献[2]中的方法抽取 WSDL 文件特征,用 LDA 方法将文件特征组织为层次结构化文本文档,建立主题对应的关键词分布,然后基于主题对服务进行检索。李征等[11]利用领域本体在对服务进行领域分类的基础上,提出一种基于 LDA、融合领域特性的服务聚类模型 DSCM,然后基于该模型提出了一种面向主题的服务聚类方法。

利用统计模型来聚类服务的方法还有很多,例如孙萍等[12]从服务功能相似和过程相似 2 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。Skoutas 等[13]提出支配服务(dominance)的概念,并根据服务之间的支配关系建立一种服务聚类方法,通过支配关系发现服务与用户请求的相关性。

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)上述方法都假设所获得的服务描述信息是语义丰富的,然而根据对 ProgrammableWeb 网站的调查发现该假设在现实中并不总是正确。2)尽管已有方法在考虑使用三方的数据来提高服务聚类的效率,例如文献[2,10]引入标签信息进行服务聚类。但是三方数据的引入仍然是作为服务聚类的一种补充,并没有明确提出针对语义稀疏情境下的解决办法。

针对上述问题,作者在服务发现过程中,针对服务描述文件语义稀疏的情境,基于迁移学习的方法无监督地抽取辅助语料的语义知识,进而丰富目标领域服务语义,利用对偶 PLSA 模型在潜在特征空间建立新的特征映射从而提供服务与潜在特征对应关系,实现基于迁移学习的语义稀疏服务聚类。

1　语义稀疏服务聚类方法

针对上述问题,作者提出一种两阶段的迁移知识辅助的语义稀疏服务聚类方法。如图 1 所示,该方法主要分成 2 步:1)数据爬取与预处理;2)迁移知识辅助的语义稀疏服务聚类。第 1 步,利用爬虫程序从服务注册网站中爬取相关的服务描述信息。通过自然语言处理的相关方法建立服务的初始特征向量。这些服务描述信息将作为对偶(probabilistic latent semantic analysis,PLSA)方法的目标领域(target domain)聚类的输入。在第 1 步同样会建立对偶 PLSA 的辅助领域(auxiliary domain)语料信息。利用 Wiki 词条对服务名称中核心词汇的解释作为辅助领域的语料信息,相关的语料会被预处理从而建立辅助特征向量。第 2 步,服务初始特征向量和服务辅助特征向量分别作为对偶 PLSA 模型中目标领域聚类和辅助领域聚类的输入。2 种语料将联合在

一起基于迁移学习的方法共同训练该对偶 PLSA 模型。通过设定相关调节参数,辅助领域的语料信息将被用来丰富目标领域的语义信息。最终,各个服务特征表达将映射入潜在特征空间,根据各个服务在潜在特征空间中的分布,建立服务的聚类结果。该方法通过引入辅助语料提高了目标领域语义丰富的程度,从而有效地缓解语义稀疏对服务聚类的影响。

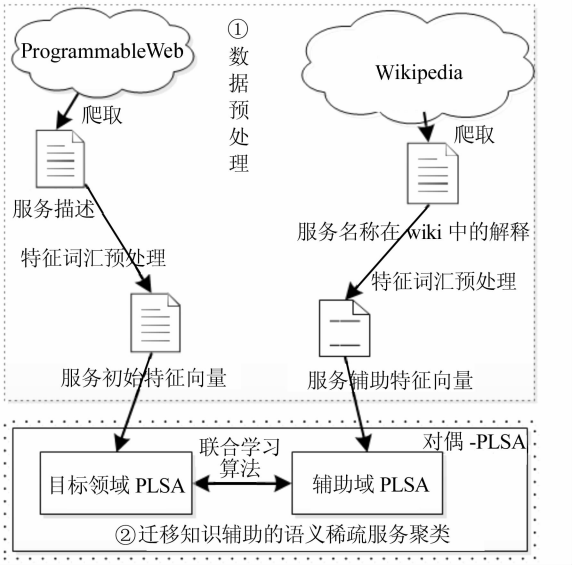

图 1 迁移知识辅助的语义稀疏服务聚类框架

Fig. 1 Framework of transferred knowledge aided semantic sparse Web service clustering

1.1 数据预处理

在数据预处理阶段,首先需要从互联网上爬取服务文本描述和服务名称所包含的关键词在 Wiki 中的词条信息。在获得这些信息之后,需要按照如下步骤进行数据预处理。

1)建立初始向量:在这一步中,作者采用自然语言处理工具包 NLTK (http://www.nltk.org/)将 Web 服务描述文档和服务名称 Wiki 词条信息处理之后建立初始文档特征向量。

2)移除功能词:功能词如“a”、“the”等词汇对于表征文档内容没有帮助,所以需要将其从文档向量中移除。

3)词干还原:具有相同词干的单词往往具有相同的含义,例如 go 和 going 具有相同的词干 go。因此,采用 PorterStemmer 对步骤 2 中的特征词进行词干还原。

1.2 迁移知识辅助的语义稀疏服务聚类

通过引入辅助领域 PLSA 模型,将 PLSA 模型扩展为对偶 PLSA 模型(dual probabilistic latent semantic analysis,D-PLSA)。该模型能够从辅助语料中无监督地抽取隐含语义从而辅助目标领域的服务聚类过程。如图 2 所示,对偶 PLSA 可以简单分成 2 个共享相同潜在特征 z 的 PLSA 模型。2 个 PLSA 模型在模型学习的过程中分别无监督发现各自语料库中的潜在模式分布,在联合学习的过程中通过共享的潜在变量 z 共享知识,从而实现从辅助领域迁移语义丰富目标领域语义的目的。

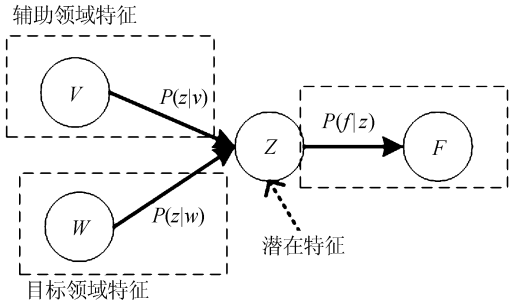

图 2 对偶 PLSA 的概率图模型

Fig. 2 Probabilistic graphical model of dual PLSA

在目标 PLSA 模型中,目标领域的服务 – 特征共现矩阵可以通过给定潜在变量 Z 之后分解成不同的独立分布,如式(1)所示。其中,W 表示服务,Z 为潜在变量,F 为服务特征。

$$P(f \mid w) = \sum_{z \in Z} P(f \mid z) P(z \mid w) \tag{1}$$

同目标领域类似,利用式(2)表示辅助领域的特征和服务特征之间的相互关联,其中,v 为辅助领域服务特征。

$$P(f \mid v) = \sum_{z \in Z} P(f \mid z) P(z \mid v) \tag{2}$$

根据图 2,式(1)和(2)共享相同的条件概率。因此,结合式(1)、(2),建立一种联合的概率模型。在该模型中,潜在特征变量 Z 不仅依赖服务和特征的关系,而且依赖于辅助特征和服务特征之间的关系。那么,基于这种共享关系,辅助语料中的语义信息将可以被用来帮助目标领域的服务聚类过程。

基于图 2 中的概率图模型,可以建立目标函数的 lg 似然函数,如式(3)所示:

$$L = \sum_j \left[\lambda \sum_i \frac{A_{ij}}{\sum_{j'} A_{ij'}} \lg P(f_j \mid v_i) + (1 - \lambda) \sum_l \frac{B_{lj}}{\sum_{j'} B_{lj'}} \lg P(f_j \mid w_l) \right] \tag{3}$$

其中:$\boldsymbol{A}^{|V| \times |F|} \in \mathbb{R}^{|V| \times |F|}$ 为辅助词和服务特征的共现矩阵;$\boldsymbol{B}^{|W| \times |F|} \in \mathbb{R}^{|W| \times |F|}$ 为服务与服务特征的共现矩阵;$\frac{A_{ij}}{\sum_{j'} A_{ij'}}$ 和 $\frac{B_{lj}}{\sum_{j'} B_{lj'}}$ 为归一化参数;λ 为 2

个共现矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的调解参数,其控制了参与联合学习辅助语料信息的数目。

为了最大化式(3)的似然函数,使用 EM(expectation maximization)算法来估计 3 种条件概率 $P(f \mid z)$、$P(z \mid w)$ 和 $P(z \mid v)$。

E 步骤:基于现有估计值 $P(f \mid z)$、$P(z \mid w)$ 和 $P(z \mid v)$,在给定目标领域服务特征 f,辅助领域文本特征 v 和目标领域服务 w 的情况下,利用式(4)、(5)计算每个潜在变量 z 的后验概率为:

$$P(z_k \mid v_i, f_j) = \frac{P(f_j \mid z_k)P(z_k \mid v_i)}{\sum_{k'} P(f_j \mid z_k')P(z_k' \mid v_i)} \tag{4}$$

$$P(z_k \mid w_l, f_j) = \frac{P(f_j \mid z_k)P(z_k \mid w_l)}{\sum_{k'} P(f_j \mid z_k')P(z_k' \mid w_l)} \tag{5}$$

M 步骤:利用式(6) ~ (8)重新计算条件概 $P(z_k \mid v_i)$、$P(z_k \mid w_l)$ 和 $P(f_j \mid z_k)$:

$$P(z_k \mid v_i) = \sum_j \frac{A_{ij}}{\sum_{j'} A_{ij'}} P(z_k \mid v_i, f_j) \tag{6}$$

$$P(z_k \mid w_l) = \sum_j \frac{B_{lj}}{\sum_{j'} B_{lj'}} P(z_k \mid w_l, f_j) \tag{7}$$

$$P(f_j \mid z_k) \propto \sum_j \left[\lambda \sum_i \frac{A_{ij}}{\sum_{j'} A_{ij'}} P(z_k \mid v_i, f_j) + (1-\lambda) \sum_l \frac{B_{lj}}{\sum_{j'} B_{lj'}} P(z_k \mid w_l, f_j)\right] \tag{8}$$

重复上述步骤直到算法收敛,使用式(9)计算给定服务 w 所属潜在特征的概率:

$$g(w) = \arg\max_{z \in Z} P(z \mid w) \tag{9}$$

根据上述分析,使用算法 1 来迭代完成 EM 算法从而建立式(3)中的目标函数 L 的局部最优解。第 1) ~2)步初始化各种变量,3) ~5)步使用 EM 算法计算各个概率,7) ~10)步最终获得类簇分布。

算法 1　迁移知识辅助的语义稀疏服务聚类算法

输入:服务 - 特征共现矩阵 $\boldsymbol{A}$,辅助语料特征和服务特征共现矩阵 $\boldsymbol{B}$。

输出:聚类结果:聚类函数为 $w \rightarrow z$,即服务 $w \in W$ 属于类簇,$z \in Z$。

1) 初始化 Z,并使得 $\mid Z \mid$ 等于聚类数目

2) 随机初始化 $P(z \mid w)$、$P(z \mid v)$ 和 $P(f \mid z)$

3) WHILE 算法未收敛 DO

4) E 步骤:根据式(4)、(5),更新 $P(z \mid w, f)$、$P(z \mid v, f)$,计算最大似然估计

5) M 步骤:根据式(6) ~ (8),更新 $P(z \mid v)$、$P(z \mid w)$ 和 $P(f \mid z)$,重新计算各参数值

6) END WHILE

7) FOR w in W DO

8) $g(w) = \arg\max_{z \in Z} P(z \mid w)$

9) END FOR

10) RETURN g。

2　实验评价

2.1　实验方法

文中涉及的算法都是通过 Python 编程语言实现,实验环境是:Intel(R) Core(TM)i5 M460@2.53 GHz,内存为 4 G,Python2.7 和 MyEclipse8.6。

实验数据来源于 ProgrammableWeb,该网站提供了下载服务信息的 API,可以利用这些 API 爬取所需要的服务描述短文本信息。实验选取包含最多服务的前 10 个分类,一共 4 402 个服务,短文本描述的平均长度为 72 个单词。同时,利用 Wiki 词条为每个服务建立服务辅助描述,辅助服务描述的平均长度为 973 个单词。

表 1　数据集相关信息

Tab. 1　Statistical information of the data set

服务类别	该类别下服务数目	服务类别	该类别下服务数目
Tools	761	Mapping	349
Internet	600	Reference	337
Social	491	Shopping	331
Financial	465	Government	315
Enterprise	444	Science	309

2.2　评价指标

为了评测算法的性能,引入纯度(purity)和熵(entropy)对聚类结果进行评测。纯度越高,表明聚类效果越好。与之相对的是,熵值越低,聚类效果越好。

单个类簇和全部类簇聚类的纯度分别定义如式(10)和(11)所示。设类簇 c_i 中包含元素 n_i 个,那么每个类簇的聚类纯度和所有类簇的平均聚类纯度分别定义为:

$$Pu(c_i) = \frac{1}{n_i} \times \max_j(n_i^j) \tag{10}$$

$$purity = \sum_{i=1}^{k} \frac{n_i}{n} Pu(c_i) \tag{11}$$

其中,n_i 为类簇 c_i 中包含的服务数目,n_i^j 为 j 个分类中正确划分入类簇 c_j 中的服务数目。

单个类簇的熵和全部类簇的平均熵使用式(12)、

(13)表示,其中,各个标号的含义同上定义。

$$E(c_i) = -\frac{1}{\lg(q)}\sum_{j=1}^{q}\frac{n_i^j}{n_i}\lg\left(\frac{n_i^j}{n_i}\right) \quad (12)$$

$$entropy = \sum_{i=1}^{k}\frac{n_i}{n}E(c_i) \quad (13)$$

2.3 结果与分析

2.3.1 服务聚类的4种方法比较

把对偶PLSA方法同4种的服务聚类方法进行比较,这几种方法分别包括:

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[2]采用了该方法实现了对服务的聚类。在下文的测试中,*K*-Means方法主要基于短文本服务描述。

2)Agglomerative层次聚类算法:该方法是一种自底向上的层次聚类算法,文献[5-6]都采用了该方法进行服务聚类分析。在下文的测试中,Agglomerative方法主要基于短文本服务描述。

3)PLSA-S,PLSA-L:在实验中一共获取到了2类服务介绍的语料:来自ProgrammableWeb的短文本服务描述和来自Wiki词条的长文本服务描述。将传统的PLSA算法分别应用到这2类语料库上进行聚类,分别为PLSA-S和PLSA-L。

上述方法运行结果如图3所示,从图中得出如下结论:1)D-PLSA算法无论是在聚类纯度还是在熵上的表现都优于其他算法,特别是同PLSA-S相比,算法有了明显的提升,这表明了所提算法的有效性。这也说明通过使用迁移学习的方法,利用辅助语料提供的语义可以辅助提高目标领域服务聚类的效果。2) *K*-Means、Agglomerative 和PLSA-S在基于短文档聚类的时候表现并不好,这表明传统算法在高维度和高语义稀疏的情境下表现并不好,这也进一步说明了文献[3]结论的正确性。3)传统的PLSA方法分别应用的目标语料(PLSA-S)和辅助语料(PLSA-L)的结果表明,作用于长文档上的方法表现更好,尽管长文档可能会有更多的噪声数据。这一部分实验的结果进一步说明了传统方法在语义丰富的情境下的优势,也进一步说明研究基于语义稀疏情景下的聚类算法的必要性。

2.3.2 辅助语料的影响

辅助语料对算法的影响如图4所示。将每个服务的辅助语料按照一定比例进行模型训练。如图4所示,当辅助数据的比例为0.1,表示每个服务的辅助语料中只有10%的单词会被选中进行模型训练从而辅助目标领域的聚类过程。

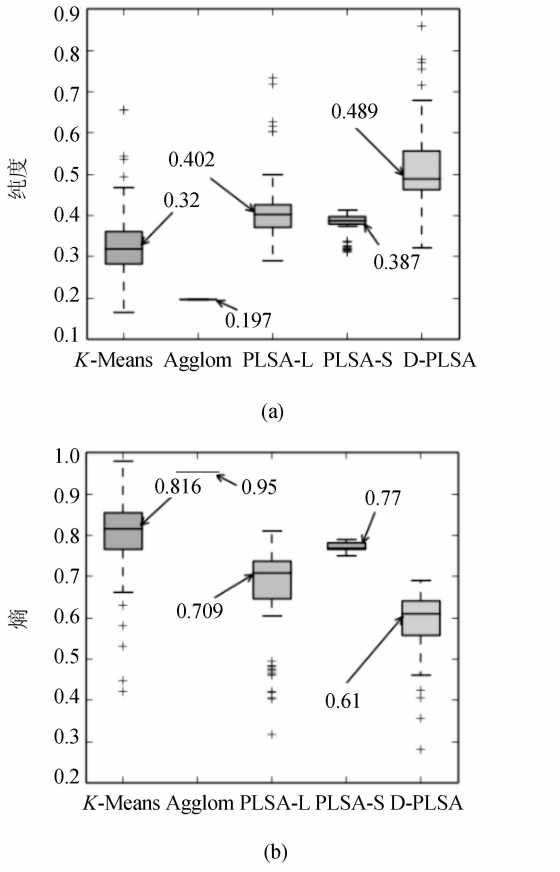

图3 算法性能

Fig. 3 Performance of the algorithm

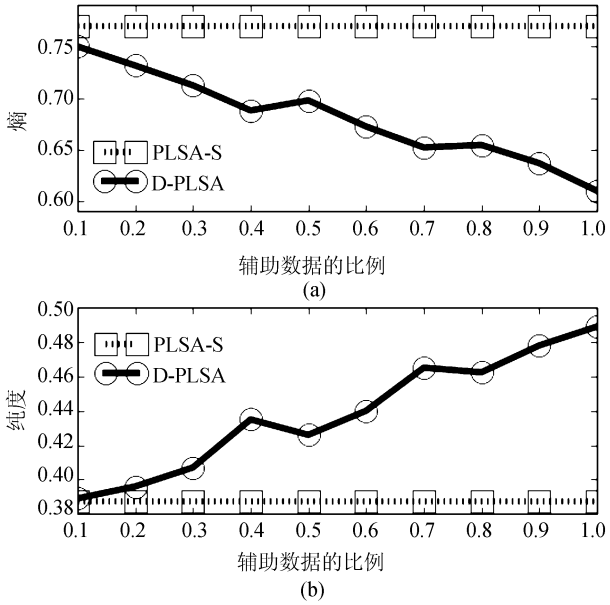

图4 辅助数据的影响

Fig. 4 Influence of auxiliary data

从图4中可以看出,随着辅助语料比例的逐步增加,模型的聚类效果逐步变好。当仅有很少的辅助数据的时候,模型的聚类效果退化如PLSA-S。而随着参与训练的辅助数据的增加,模型的性能也随着提升。这个实验表明,辅助数据的引入能够提高

语义稀疏服务的聚类性能。

2.3.3　噪声数据的影响

在利用 Wiki 词条建立辅助语料库的过程中,不可避免地会引入部分噪声数据。其主要原因有 2 个:1)每个服务的关键词语会出现同形异义的情况,例如 light 既有“光线”(名词)的含义又有“轻的”(形容词)的含义;不同含义对应的词条的是不一样的,因此会引入噪声数据。2)词条内部解释也可能包含大量和原服务描述语义关联不大的内容。因此设计如下实验测试噪声数据的影响:当选择 Tools 分类进行聚类的时候,另外 9 个分类的服务描述将作为噪声数据。如图 5 所示,当噪声数据比例为 0.1 的时候表示辅助领域的数据中有 0.1 的数据是来自于其他 9 个分类的服务描述。

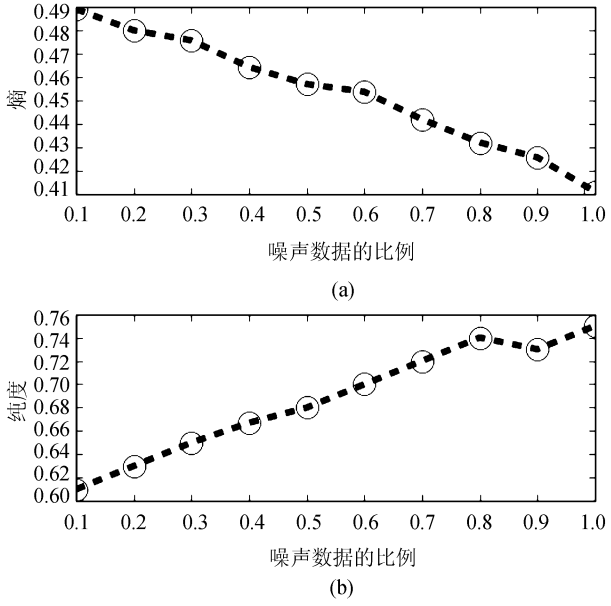

图 5　噪声数据的影响

Fig. 5　Influence of the noisy data

从图 5 中可以看出,随着噪声数据的增加,方法聚类的效果在逐渐下降。当噪声数据的比例等于 1 的时候,方法退化接近 PLSA-S,但是仍然要比 PLSA-S 的聚类效果略好,表明即使辅助语料与目标领域的关联并不密切,作者方法相对传统方法的聚类效果仍然有一定的提升,这也说明了该方法具有一定的鲁棒性。

2.3.4　参数 λ 的影响

在式(3)中,λ 是 2 个共现矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的调解参数,它控制了参与联合学习的辅助语料信息的数目。不同 λ 取值对聚类纯度的影响如图 6 所示。从图 6 中可以看出,当 λ 的取值等于 0.3 的时候,算法取得最优的聚类效果,即参与学习的辅助语料的比例为 0.3。

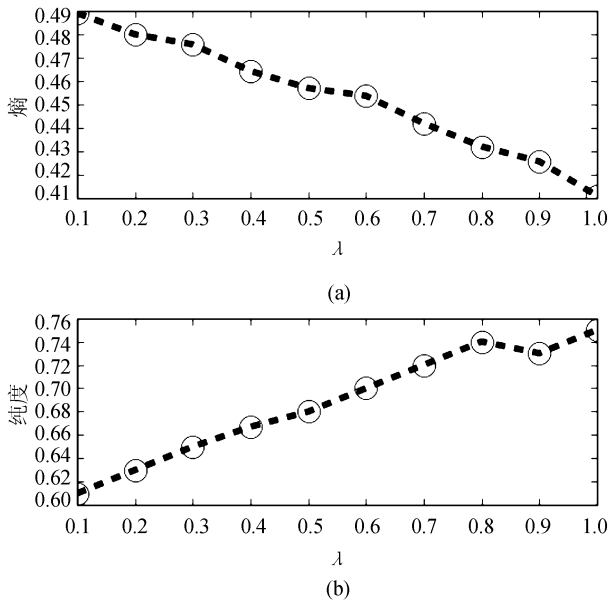

图 6　调解参数 λ 的影响

Fig. 6　Influence of tunning parameter λ

3　总结和下一步工作

基于服务自然语言描述和从 Wiki 中获取的辅助语料信息,利用 PLSA 方法方法,提出了基于迁移学习、面向语义稀疏情景的服务聚类模型。然后基于该模型对服务进行离线的聚类,从而将具有相似功能的服务组织为类簇,提高服务发现的效率。最后,以 ProgrammabelWeb 上真实的数据集进行实验,验证了对偶 PLSA 服务聚类方法的可行性和有效性。性能对比实验分析表明,提出的方法在纯度、熵方面均具有更好的聚类效果。而且,该方法有助于解决在语义稀疏情境下的服务聚类问题,从而促进语义稀疏的服务发现,具有较好的实际应用价值。

下一步,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)研究能够无监督地自动选择参与迁移学习的辅助语料数量的模型,减少人工参与的程度,降低噪声数据对算法的影响,提高算法的鲁棒性。

参考文献:

[1] Elgazzar K, Hassan A E, Martin P. Clustering wsdl documents to bootstrap the discovery of web services[C]//International Conference on Web Services. Los Angeles, USA, 2009: 147 - 154.

[2] Chen Liang, Hu Liukai, Zheng Zibin, et al. WTCluster: Utilizing tags for web services clustering[C]//Proceedings of Int Conference on Service-Oriented Computing. Berlin:

Springer,2011:204 - 218.

[3] Tang J, Wang X, Gao H. Enriching short text representation in microblog for clustering[J]. Frontiers of Computer Science,2012,6(1):88 - 101.

[4] Cao Jie, Wu Zhiang, Wu Junjie, et al. SAIL: Summation-based incremental learning for information-theoretic text clustering[J]. IEEE Transactions on Cybernetics,2013,43(2):570 - 584.

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

[6] 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.

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

[8] Cassar G, Barnaghi P, Moessner K. Probabilistic methods for service clustering[C]//Proceedings of the 4th Int Workshop on Semantic Web Service Matchmaking and Resource Retrieval, Organised in Conjunction with the Int Semantic Web Conference. 2010:4 - 20.

[9] Ge Lin, Ji Xinsheng, Wei Hongquan, et al. Event classification of on-line network information content security incidents based on LDA model[J]. Journal of Sichuan University: Engineering Science Edition,2014,46(3):70 - 79. [葛琳,季新生,卫红权,等. 基于LDA模型的在线网络信息内容安全事件分类[J]. 四川大学学报:工程科学版,2014,46(3):70 - 79.]

[10] Chen Liang, Wang Yilun, Yu Qi, et al. WT-LDA: User tagging augmented LDA for web service clustering[M]// Service-oriented Computing. Berlin Heidelberg: Springer,2013:162 - 176.

[11] Li Zheng, Wang Jian, Zhang Neng, et al. A topic-oriented clustering approach for domain services[J]. Journal of Computer Research and Development,2014,51(2):408 - 419. [李征,王健,张能,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.]

[12] 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. [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.]

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

(编辑 张 琼)