基于描述语境特征词与改进 GSDMM 模型的服务聚类方法

胡强，沈嘉吉，荆广辉，杜军威
（青岛科技大学信息科学技术学院，山东 青岛 266061）

摘 要：针对现有聚类方法中存在的服务表征向量生成质量较差问题，提出了一种面向描述语境特征词与改进 GSDMM 模型的服务聚类方法。首先，构建了基于语境权重的特征词提取方法，将与服务描述语境契合度高的词语抽取出，构建用于服务表征向量生成的功能特征词集合。然后，建立了带有主题分布概率修正因子的 GSDMM 模型，实现服务表征向量的生成以及非关键主题项概率分布修正。最后，基于修正后的服务表征向量，采用 K-means++算法实现服务聚类。以 Programmable Web 上真实服务进行了多轮次实验，实验结果表明，采用所提方法生成的服务表征向量质量显著高于其他常用主题模型，所构建的服务聚算法性能优于其他常用算法。

关键词：Web 服务；服务聚类；主题模型；GSDMM
中图分类号：TN92
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021150

Service clustering method based on description context feature words and improved GSDMM model

HU Qiang, SHEN Jiaji, JING Guanghui, DU Junwei
School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China

Abstract: To address the problem that current service clustering methods usually faced low quality of service representation vectors, a service clustering method based on description context feature words and improved GSDMM model was proposed. Firstly, a feature word extraction method based on context weight was constructed. The words that fit well with the context of service description were extracted as the set of feature words for each service. Then, an improved GSDMM model with topic distribution probability correction factor was established to generate service representation vectors and achieve distribution probability correction for non-critical topic items. Finally, K-means++ algorithm was employed to cluster Web services based on these service representation vectors. Experiments were conducted on real Web services in Web site of Programmable Web. Experiment results show that the quality of service representation vectors generated by the proposed method is higher than of other topic models. Further, the performance of our clustering method is significantly better than other service clustering methods.

Keywords: Web service, service clustering, topic model, GSDMM

1 引言

随着云计算、大数据以及物联网、移动互联网等新一代信息技术的快速发展，面向服务架构（SOA, service-oriented architecture）的业务系统开发和部署得到广泛应用和推广[1]。企业将自身业务功能或产品封

收稿日期：2021−04−07；**修回日期**：2021−06−29
通信作者：杜军威，djwqd@163.com
基金项目：国家自然科学基金资助项目（No.61973180）；山东省自然科学基金资助项目（No.ZR2019MF033）；山东省重点研发计划基金资助项目（No.2018GGX101052）；国家重点研发计划基金资助项目（No.2018YFB1702902）
Foundation Items: The National Natural Science Foundation of China (No.61973180), The Natural Science Foundation of Shandong Province (No.ZR2019MF033), The Key Research and Development Program of Shandong Province (No.2018GGX101052), The National Key Research and Development Program of China (No.2018YFB1702902)

装为服务，通过互联网进行发布；用户可以通过云平台查找和调用所需服务，并与已有业务系统集成，快速构建各类复杂的增值应用系统，弥补自身业务能力的不足，进而可降低企业运营成本，提高竞争力[2]。

作为 SOA 架构下主流服务实现方式，Web 服务是一种采用规范化协议封装的 Web API，分为 SOAP 和 RESTful。截至 2021 年 3 月，在 Programmable Web 注册的服务已经超过 23 000 条，其中，绝大多数为 RESTful 型 Web API。不同于 SOAP 类型的服务采用 Web 服务描述语言（WSDL, Web services description language）结构化文档进行服务信息描述，RESTful 类型 Web 服务通常采用非结构化的自然语言描述服务功能或使用方法，从而使服务查找与发现难度增大。此外，随着 SOA 架构应用的普及，大量企业不断推出新的 Web 服务，使服务数量飞速增加，进一步增大了服务发现难度[3]。

聚类可以将功能相似的服务划分为一个簇，有效地缩减服务查找空间，提高发现效率[4]。采用 WSDL 文档描述服务信息的 Web 服务中，WSDL 文档显式地设置了多种标签，易于提取服务描述的各类特征信息。此类服务在聚类时通常提取若干能够表示服务功能的关键词，按照标签类别分别计算这些关键词的语义或词频相似度即可实现服务功能相似性的度量[5]。近年来，随着 RESTful 类型 Web 服务数量的增加，越来越多的 Web 服务采用短文本自然语言进行服务描述，因此，如何为此类服务实现高质量的聚类成为新的研究热点[6]。

然而，采用自然语言刻画的 Web 服务描述信息中文本较短，特别是缺乏显式标签对服务描述文本进行语义信息的标识，使服务的各类特征信息提取难度增大。为了有效地从服务描述信息文本中提取可以表达服务功能的关键特征信息，研究者将主题模型应用于服务描述信息的建模和抽取[7]。

主题模型是一类以非监督学习的方式对文本的隐含语义结构进行聚类的统计模型，以概率分布形式生成文本对应的主题向量。利用主题模型可以为服务功能描述生成主题向量，该主题向量表达了服务在若干功能主题上的概率分布。通过计算不同服务表征向量之间的相似度可以实现服务功能相似度的判定。本文将采用主题模型为 Web 服务生成的主题向量称为服务表征向量。

目前，常于生成服务表征向量的主题模型有 LSA（latent Semantic analysis）[8-9]、LDA（latent Dirichlet alloation）[10-11]、BTM（biterm topic model）[12-13]、HDP（hierarchical Dirichlet process）[14]、GSDMM（Gibbs sampling for the Dirichlet multinomial mixture）[15-16]等，其中，LDA 模型应用最广泛。服务描述文本较短，且功能、操作以及 QoS 描述词语混合在一起[17]，导致采用主题模型生成的服务表征向量中通常主题松散和稀疏。文献[15]利用实验进行多项指标对比验证后指出，相比其他主题模型，GSDMM 为服务描述生成主题向量的质量最高。然而，GSDMM 生成主题向量时强化关键主题概率，弱化次要主题概率，这种情形虽有利于短文本分类、提高生成向量的主题区分度，但在一定程度上影响主题语义分布的均衡性。

针对上述问题，为进一步提升服务表征向量的生成质量，改进服务聚类效果，本文提出了一种基于描述语境特征词与改进 GSDMM 的服务聚类方法，主要工作和贡献如下。

1) 提出了一种基于语境权重的服务描述特征词提取方法。该方法将服务描述文本中词语的词频与语境相似度相结合，建立词语的语境权重，筛选出一定数量可以代表服务描述功能的特征词，去除噪声词语，有效缓解服务描述文本中的主题松散问题。

2) 构建了一种基于带有主题分布概率修正因子 GSDMM 服务表征向量生成模型。该模型将通过修正 GSDMM 模型生成的主题向量中非关键主题的概率分布，改善生成服务表征向量主题分布的语义均衡性，提高了服务表征向量的质量。

3) 构建了面向服务表征向量与 K-means++算法的服务聚类方法。以 Programmable Web 上的真实数据开展多轮次聚类实验，验证了所提出的服务描述特征词提取方法、改进的 GSDMM 的服务表征向量生成模型以及聚类算法的有效性与先进性。

2　相关工作

如何从 WSDL 结构化文档中抽取 Web 服务描述信息，并进行服务相似度计算与聚类是早期服务聚类领域开展的主要工作。例如，Liang 等[18]提出一种名为 WCCluster 的聚类方法，该方法以二分图划分的形式同时对 WSDL 文档及抽取出的关键词进行聚类。Wu 等[19]基于标签共现、挖掘和语义相关性度量进行服务标签推荐，进而构建了一种基于混合 Web 服务标签的服务聚类方法，相比其他 WSDL 聚类方法，

该方法将聚类精度提升约 14%。Agarwal 等[6]提出一种基于长度特征权重的方法 LFW（length feature weight）对 WSDL 文档向量化，并使用 K-means 聚类算法完成聚类，其效果优于使用 TF-IDF（term frequency-inverse document frequency）方法生成 WSDL 表示向量的方法。上述方法通常是从 WSDL 文档中提取各类标签描述的特征词，然后通过计算特征词的语义或者借助于词频转化为向量，进而计算服务相似度并实现聚类。

随着采用短文本自然语言进行功能描述的 Web 服务数量的增加，已有从结构化文档中提取服务特征信息的方法不再适用于此类服务[20]。研究者通常采用主题模型或者神经网络模型对服务描述进行特征向量提取，通过计算特征向量的相似度完成服务发现或聚类。

Cao 等[21]利用 Mashup 服务之间的关系构建了 Mashup 服务网，使用二级主题模型来挖掘潜在主题，并设计了一种基于协同过滤（CF, collaborative filtering）的 Web API 推荐算法。文献[11]提出了一种利用高斯 LDA 模型处理文本词嵌入的方法，在此基础上。Lizarralde 等[22]将服务描述加入高斯 LDA 模型中以获取服务描述表示，再通过用户查询与服务描述表示之间的相关性对服务进行排名。Zhang 等[23]使用 LDA 主题模型将服务进行分组聚类，然后基于服务集群进行服务发现，设计了推荐语义相关服务的机制帮助用户优化初始查询。刘建勋等[24]提出一种基于主题模型的 Mashup 标签推荐方法，通过联合使用 LDA 与 RTM（relation topic model），使推荐精度显著提高。石敏等[25]考虑多重 Web 服务关系的概率构建主题模型 MR-LDA（multi-relational-LDA），对 Web 服务之间的组合关系以及 Web 服务之间共享标签的关系进行建模，利用这些关系对主题分布矩阵进行修正，并对其进行高效服务聚类。

Baskara 等[13]将 Web 服务结构建模为加权有向无环图（WDAG, weighted directed acyclic graph），然后使用 BTM 主题模型在已建模的 WDAG 上挖掘主题，并通过计算主题相似度进行服务发现。为了解决描述特征稀疏的问题，Shi 等[26]提出了一种基于概率主题模型的句子扩展方法以及基于深度学习 LSTM 模型的服务推荐方法，效果优于基础 LSTM 方法。Ye 等[27]提出一种 WSC-GCN（Web services classification based on graph neural network）模型，将单词−文档关系与单词−单词关系作为边、单词与文档作为点构建出一张无向图，并使用 TF-IDF 值作为边的权重，放入 GCN 中获得文档向量进行聚类。

相比以 LDA 为代表的上述主题模型，Yin 等[16]提出的 GSDMM 更适合为短文本进行主题建模。虽然 GSDMM 模型生成的服务表征向量质量优于其他主题模型，但在生成服务表征向量的概率分布完备性和均衡性层面还有一定的提升空间。此外，在服务表征向量生成时，通过对服务描述信息中的特征词进行筛选，可以进一步提高服务表征向量的质量。基于上述考虑，本文开展了基于描述语境特征词与改进 GSDMM 模型的服务表征向量生成与聚类研究，本文的研究流程如图 1 所示。

3 基于语境权重的服务描述特征词提取

本文中参与聚类的服务是采用短文本自然语言进行功能描述的 Web 服务。该类服务以 Programmable Web 上的 Web API 为典型代表，图 2 给出一个具体的 Web API 服务描述。

服务标签是该服务提供功能所隶属的类别标志，服务描述文本则是以短文本自然语言的形式给出的有关服务功能、使用以及质量评价方面的文字说明。这两类文本中的词语是服务聚类时进行功能相似度度量的主要依据。为了后文描述方便，首先给出服务的形式化定义。

定义 1 服务

服务定义为一个四元组，$s=(\mathrm{Id},n,l,d)$，其中，Id 为服务的标识 ID 号，n 为服务的名称，l 为服务标签集合，d 为服务描述信息。

定义 2 词语的语义相似度

w_i 和 w_j 为文本 T 中的 2 个单词，$\boldsymbol{V}_{w_i}$ 和 $\boldsymbol{V}_{w_j}$ 分别为词语 w_i 和 w_j 对应的词向量，则 w_i 和 w_j 的语义相似度为

$$\mathrm{SemSim}(w_i,w_j)=\frac{\boldsymbol{V}_{w_i}\boldsymbol{V}_{w_j}}{|\boldsymbol{V}_{w_i}|\cdot|\boldsymbol{V}_{w_j}|}$$

定义 3 词语的语境相似度

T_{im} 为一段包含 m 个词语的文本，w_{ij} 为文本 T_{im} 中的一个词语，w_{ij} 在文本 T_{im} 中的语境相似度为

$$\mathrm{Context\_SemSim}(w_{ij},T_{\mathrm{im}})=\frac{\sum_{k=1}^{j-1}\mathrm{SemSim}\left(w_{ij},w_{ik}\right)+\sum_{k=j+1}^{j-1}\mathrm{SemSim}\left(w_{ij},w_{ik}\right)}{m-1}$$

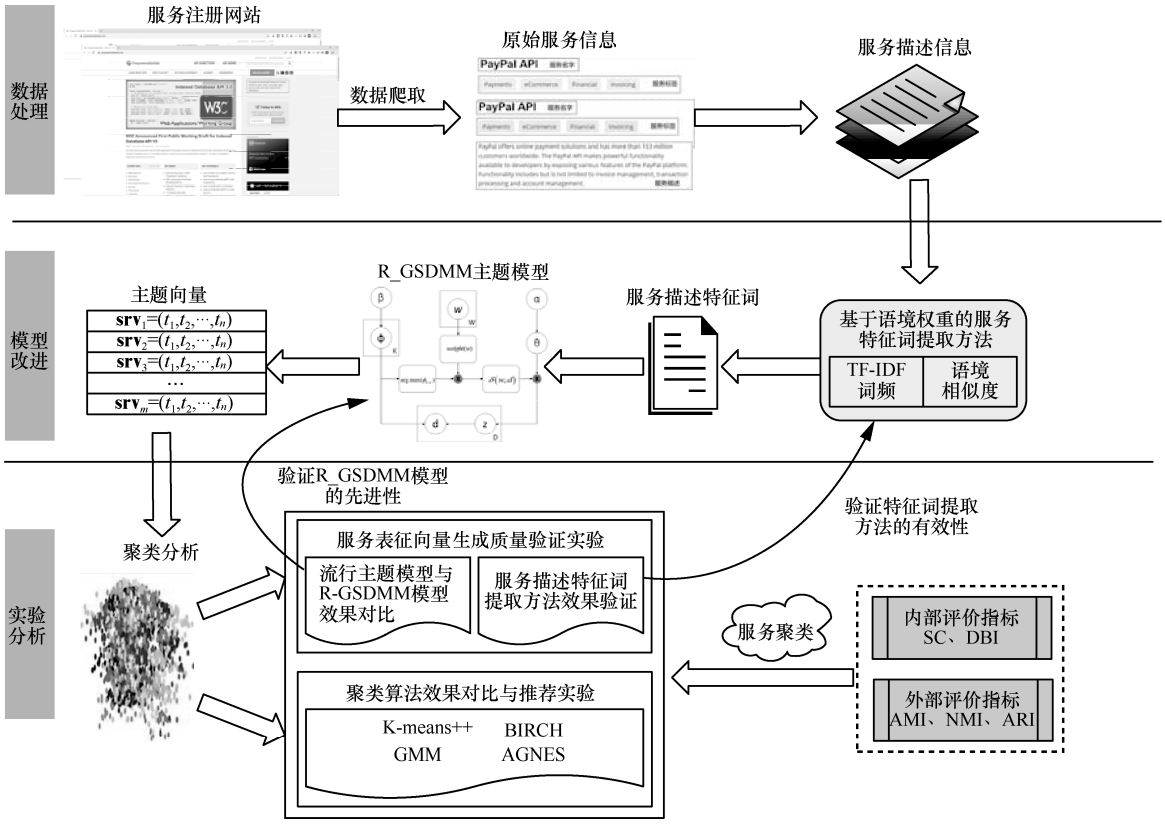

图 1　研究流程

定义 4　词语的 TF-IDF

d 为文档集合 D 中的一个文档，w_i 是 d 中的一个词语。w_i 在 d 中的 TF-IDF 定义为

$$\mathrm{TF-IDF}(w_i,d,D)=\mathrm{TF}_{w_i}\mathrm{IDF}_{w_i}$$

其中，$\mathrm{TF}_{w_i}=N_{w_i}/N_w$，$N_{w_i}$ 与 N_w 分别为文档 d 中包含词语 w_i 的数量以及 d 中的单词总数；$\mathrm{IDF}_{w_i}=\lg\dfrac{|D|}{\left|\{j:w_i\in d_j\}\right|}$，$|D|$ 为 D 中文档数量，j 为包含词语 w_i 的文档数量。

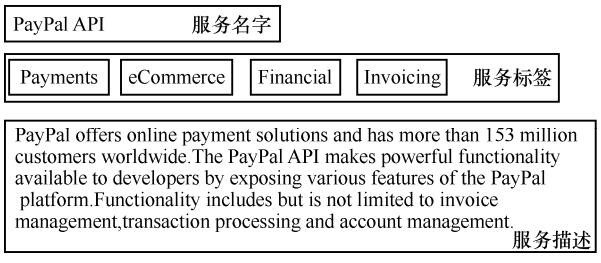

图 2　Programmable Web 服务示例

定义 5　词语的语境权重

T_{im} 为语料库 T_Corpus 中一段包含 m 个词语的文本，w_{ij} 为文本 T_{im} 中的一个词语，w_{ij} 在文本 T_{im} 中的语境权重定义为

$$\mathrm{Context\_Weight}(w_{ij},T_{\mathrm{im}})=\mathrm{Context\_SemSim}(w_{ij},T_{\mathrm{im}})\mathrm{TF-IDF}(w_{ij},T_{\mathrm{im}},T\_\mathrm{Corpus})$$

由定义 2 可知，词语的语义相似度为词语所对应词向量之间的余弦夹角值，语境相似度是该词语与其所在文本中的其他词语的语义相似度的平均值。词语的语境权重则需要基于词语所在的文本以及语料库计算出该词语相对于所在文本的 TF-IDF 词频，再乘以该词语的语境相似度。因此，语境权重综合考虑了词频以及词语之间的语义相似度，更好地反映了词语的重要性。

现有采用主题模型进行服务聚类的方法中，通常是将服务描述删除虚词后的全部词语均作为服务表征向量生成的语料，这种做法存在主题稀疏和分散的问题。也有研究工作仅从服务描述中提取若干个关键词，通过 Word2Vec 等工具将这些词语转化为向量，用加权的方式生成表征服务的最终向量，这类方法存在关键词提取难度大的问题，而且容易因加权操作造成语义信息丢失。

为此，本文提出一种基于语境权重的服务描述特征词提取方法。该方法按照语境权重排名筛选出

一定数量的词语作为服务描述的特征词。这种特征词筛选的方式有效降低了噪声词语的数量，提高了服务表征向量生成文本的质量。对于服务 s，令 $S=\{s_i\}$，$1\leqslant i\leqslant n$ 为待聚类服务构成的服务集合，算法 1 给出了基于语境权重的服务描述特征词提取方法。

算法 1 FeatureWord_Extract

输入 the set of service S

输出 the set of feature word FW_S for Web services in S

1) Corpus_w=∅

2) FW_S=∅

3) for $\forall s\in S$

4) Corpus_w=Corpus_$w\cup s.d$

5) end for

6) train the vector $\boldsymbol{V}(w)$ for each word w in Corpus_w by Word2Vec

7) for each $s\in S$

8) fw_s=∅

9) for each word w in $s.d$

10) compute Context_Weight(w, $s.d$)

11) end for

12) s.fw=$s.l\cup$ Rank($s.d$, ContextWeight(w, $s.d$),α)

13)FW_S= FW_$S\cup$ {s.fw}

14)end for

15) return FW_S

算法 1 第 1)～2)行初始化 2 个空集合，Corpus_w 为语料库集合，用于存储服务集合 S 中所有服务的服务描述文本；FW_S 为服务特征词集合，存储经算法 1 处理后得到的集合 S 中所有服务的特征词。第 3)～6)行将 S 中包含的所有服务的服务描述文本加入语料库 Corpus_w，然后利用 Word2Vec 为语料库中的每个词语 w 训练一个向量 $\boldsymbol{V}(w)$。

在生成服务 s 的特征词集合时，针对服务描述 $s.d$ 中的每个词 w，算法 1 第 9)～11)行计算 w 在 $s.d$ 中对应的语境权重；第 12)行构建了服务特征词集合 s.fw。服务标签是平台服务分类、存储和查找的重要类别依据，文献[24]指出不同服务共享标签的数量越多，它们同属一个类别的可能性越大，当服务之间共享 3 个以上标签时，隶属于同一个类别的概率上升至 50%，因此，服务标签是聚类所需考虑的重要因素。在构建 s.fw 时，首先将服务标签集合 $s.l$ 中的所有单词加入 s.fw；然后对服务描述中词语依照服务描述语境权重 ContextWeight(w)进行排序，分别取占比为前α的词语加入 s.fw。通过实验验证，当α为 75%～90%时，所提取的特征词生成服务表征向量质量最佳。循环处理可得服务集合 S 中所有服务的特征词集合 FW_S。

4 基于带有主题分布概率修正因子 GSDMM 模型的服务表征向量生成

本文提出一种面向主题分布概率修正的 GSDMM 模型，引入概率分布修正因子，修正服务表征向量中主题对应的概率分布值，提高服务表征向量对服务功能刻画的完备性与均衡性。

GSDMM 模型是一种概率生成式无监督模型，基于狄利克雷混合（DMM, Dirichlet multinomial mixtures）模型生成文档，然后使用吉布斯采样算法近似求解模型。在 DMM 模型中，由主题 k 得到文档 d 的概率为

$$p(d\mid z=k)=\prod_{w\in d}p(w\mid z=k) \tag{1}$$

为获得描述中的单词−主题分布，依据文献[16]，假设主题在单词上是多项式分布，则有

$$p(w\mid z=k)=p(w\mid z=k,\boldsymbol{\Phi})=\phi_{k,w} \tag{2}$$

其中，$\boldsymbol{\Phi}$为单词−主题分布矩阵，刻画了单词 w 属于第 k 个主题的概率；$\phi_{k,w}$ 为单词 w 在主题 k 上的概率分布，同一篇文档中所有单词的主题分布之和为 1，即 $\sum_{w=1}^{V}\phi_{k,w}=1$。

同样地，每个主题的概率服从式(3)所示的多项式分布

$$p(d\mid z=k)=p(d\mid z=k,\boldsymbol{\Theta})=\theta_{k,d} \tag{3}$$

其中，$\boldsymbol{\Theta}$为文档−主题分布矩阵，刻画了文档 d 在主题 k 上的概率分布；$\theta_{k,d}$ 为文档 d 在主题 k 上的概率分布，在一篇文档描述中遵循 $\sum_{k=1}^{K}\theta_{k,d}=1$。

吉布斯采样算法在所有主题上不断采样，最终得到文档−主题矩阵$\boldsymbol{\Theta}$和单词−主题矩阵$\boldsymbol{\Phi}$。吉布斯采样中描述属于某个主题的概率计算式为

$$p(z_d=z\mid\vec{z}_{\neg d},\vec{d})\propto\frac{m_{z,\neg d}+\alpha}{D-1+K\alpha}\cdot\frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(n_{z,\neg d}^{w}+\beta+j-1)}{\prod_{i=1}^{N_d}(n_{z,\neg d}+V\beta+i-1)} \tag{4}$$

其中，K 为初始主题个数，α、β 为狄利克雷先验参数，D 为语料库中文档总数，$\vec{z}$ 为文档的主题标签，$\vec{d}$ 为语料库中的文档，m_z 为主题 z 下的文档数，n_z 为主题 z 下的单词数，n_z^w 为主题 z 下的单词 w 出现的次数，$\neg d$ 为去除当前文档，V 为词语表中的词语数量，N_d 为文档 d 中的单词数量，N_d^w 为文档 d 中单词 w 出现的次数。

本文将服务表征向量中具有最大分布概率值的主题称为关键主题，其他主题称为次要（非关键）主题。在 GSDMM 中引入主题分布概率修正因子δ，通过δ修正生成服务表征向量中的各个次要主题分布概率，有效地提高服务表征向量的完备性，改进的 GSDMM 模型如图 3 所示。

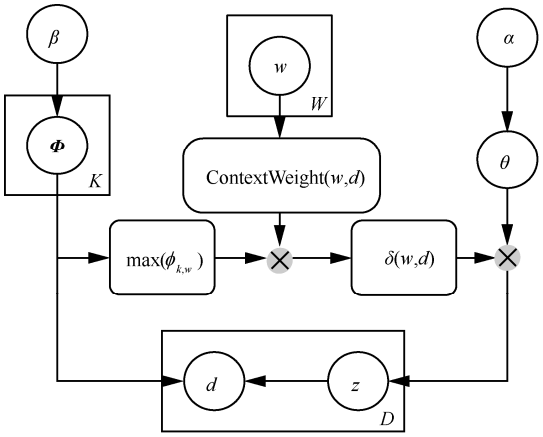

图 3　带有主题概率分布修正因子的 GSDMM 模型

修正因子为

$$\delta(w,d)=1+\max(\phi_{k,w})\ \mathrm{ContextWeight}(w,d)\lambda$$

其中，$\max(\phi_{k,w})$为单词 w 在所有主题 k 下的最大概率分布值，$\mathrm{ContextWeight}(w,d)$ 为单词 w 在文档 d 中的语境权重，λ 为调节修正因子影响力的超参数。算法 2 给出面向主题概率分布修正 GSDMM 的服务表征向量生成方法。

算法 2　SRV_RGSDMM

输入　the set of service S

输出　the set of service representation vectorSRV

1) obtain ContextWeight(w,s) and s_fw for each service s in S by 算法 1

2) feed each s_fw to GSDMM for ten rounds and get

3)　{ the initial service representation vector srv(s)

4)　the topic-word matrix $\boldsymbol{\Phi}$

5)　the topic-service matrix $\boldsymbol{\Theta}$}

6) for each $s\in S$

7)　for each word w in s_fw

8)　k-max-w=argmax($\phi_{k,w}$)

9)　$\delta(w,\ s.d)$=1+ max($\phi_{k,w}$)ContextWeight(w, $s.d$)λ

10)　k-max-s=argmax($\theta_{k,s}$)

11)　if (k-max-s!= k-max-w)

12)　update **srv**(s) by $\theta_{k-\max-w,s}=\theta_{k-\max-w,s}*\delta(w,s)$

13)　end if

14)　end for

15) SRV=SRV∪**srv**(s)

16) end for

17) return(SRV)

算法 2 中，第 1)行调用算法 1，为服务 s 的服务描述文本中的每一个词语 w 计算语境权重 ContextWeight(w,s)，并筛选出服务特征词集合 s_fw；第 2)～5)行将服务集合中所有服务的特征词集合依次输入 GSDMM 模型中，进行 10 轮训练后得到主题−词语矩阵 $\boldsymbol{\Phi}$ 和服务−主题矩阵 $\boldsymbol{\Theta}$，以及每个服务 s 对应的初始服务表征向量 **srv**(s)。

算法 2 第 6)～14)行实现修正因子的计算，并完成对初始服务表征向量 **srv**(s)的概率修正。首先，针对服务 s 的特征词集合 s_fw 中的每一个单词 w，在单词−主题矩阵 $\boldsymbol{\Phi}$ 找到 w 对应的分布概率值最大的主题 k-max-w。将单词 w 的最大主题分布概率值 max($\phi_{k,w}$)与其对应的语境权重 ContextWeight(w, $s.d$)相乘，并乘以超参数修正率 λ=0.1，得到服务 s 表征主题中该单词 w 对应的修正因子$\delta(w,\ s.d)$。算法 2 第 10)～13)行判定单词 w 对应的最大概率分布主题 k-max-w 是否为服务 s 表征向量中的次要主题，如果是次要主题，则将修正因子$\delta(w, s.d)$与表征向量中该主题已有分布概率值相乘，完成基于单词 w 的主题概率分布修正。将服务 s 的所有特征词集合中的单词均完成修正后即可得到最终的服务表征向量。算法 2 第 17)行返回服务表征向量集合 SRV。

5　基于 K-means++的服务聚类算法

本节基于 K-means++与生成的服务表征向量，构建服务聚类算法，以算法 2 生成的服务表征向量作为聚类数据，如算法 3 所示。

算法 3 ServiceCluster_KM

输入 the set of service represent vector SRV，the number of clusters k

输出 k service clusters

1) select a **srv** randomly in SRV as the first center point cp1

2) repeat

3) for each **srv** and center point cp

4) compute Euclidean distance ED (**srv**, cp)

5) end for

6) $D(\mathbf{srv})=\min(\mathrm{ED}(\mathbf{srv},cp))$

7) $P(\mathbf{srv})=\dfrac{D(\mathbf{srv})^2}{\sum\limits_{\mathbf{srv}\in\mathrm{SRV}} D(\mathbf{srv})^2}$

8) select a **srv** as a center point by roulette method

9) until k center points is generated

10) repeat

11) for each **srv** and center point cp

12) compute Euclidean distance ED(**srv**,cp)

13) end for

14) add **srv** into the cluster with the min (ED (**srv**, cp))

15) recompute the center points for each cluster

16) until no alteration in center points

17) return k service clusters

算法 3 首先随机选择一个服务表征向量作为初始聚类中心点 cp_1。第 2)～9)行确定 k 个初始化聚类中心点，其方法是计算每个服务表征向量与已经产生的聚类中心点之间的欧氏距离 ED(**srv**,cp)，将服务表征向量 **srv** 与已产生中心点的最近距离定义为 $D(\mathbf{srv})$；然后通过第 7)行中的 $P(\mathbf{srv})$ 计算 **srv** 被选中为下一个聚类中心点的概率；最后，通过轮盘法选出下一个聚类中心点。

算法 3 第 10)～16)行实现服务聚类。在聚类过程中，通过计算待聚类服务表征向量与聚类中心点之间的欧氏距离，将待聚类服务划归到具有最短欧氏距离的聚类中心点所在的簇。完成一轮次聚类后，重新计算每个服务簇的中心点，实现新一轮次的聚类，直至各个服务簇的聚类中心点不再变动，最终聚类完成，输出 k 个服务簇。

6 实验与分析

6.1 评价指标

聚类质量评价指标可分为外部和内部评价指标两类。外部评价指标使用样本标签来评价聚类是否合理。内部评价指标是通过刻画聚类质量的参数对聚类效果进行评价。本文选取以下常用指标对聚类质量进行。

6.1.1 内部评价指标

1) 轮廓系数（SC, silhouette coefficient）

对于单个样本 x，设 a 是与它同类别中其他样本的平均距离，b 是与它距离最近不同类别中样本的平均距离，其轮廓系数为

$$\mathrm{SC}(x)=\frac{b-a}{\max(a,b)} \tag{5}$$

SC 分数的取值范围是[−1, 1]，分数越高代表聚类效果越好。

2) 戴维森堡丁指数（DBI, Davies-Bouldin index）

DBI 表示为任意 2 个类别的类内平均距离之和除以 2 个聚类中心距离求最大值。该指标计算式为

$$\mathrm{DBI}(x_i,x_j)=\frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d\left(c_i,c_j\right)}\right) \tag{6}$$

其中，x_i 和 x_j 为 2 个聚类，n 为聚类的类别个数，c_i 为第 i 个类别中心，σ_i 为 i 类中所有的点到中心点的平均距离，$d(c_i,c_j)$为 2 个聚类中心点 c_i 与 c_j 之间的距离。类内距离越小、类间距离越大，则 DBI 越小，分类效果越好。

6.1.2 外部评价指标

1) 标准化互信息（NMI, normalized mutual information）

NMI 是互信息（MI）分数的归一化，其计算式为

$$\mathrm{MNI}\left(X,Y\right)=\frac{\mathrm{MI}\left(X,Y\right)}{F\left(H\left(X\right),H(Y)\right)} \tag{7}$$

其中，$X=\{x_1,x_2,\cdots,x_K\}$ 为聚类后的样本划分，$Y=\{y_1,y_2,\cdots,y_K\}$ 为真实类别划分，MI 可计算出 X 与 Y 的相关程度，$H(X)$与 $H(Y)$分别为 X 与 Y 的熵，F 为归一化函数，本文选取算术平均的方法。NMI 分数区间为[0,1]，值越大代表聚类效果越好。

3) 调整互信息（AMI, adjusted mutual information）

AMI 在 MI 基础上对其进行调整，已知聚类标签与真实标签，MI 能够度量两者之间的相关性，并忽略标签的排序。该指标的计算式为

$$\mathrm{AMI}\left(X,Y\right)=\frac{\mathrm{MI}\left(X,Y\right)-\mathrm{E}\left\{\mathrm{MI}\left(X,Y\right)\right\}}{F\left(H\left(X\right),H(Y)\right)-\mathrm{E}\left\{\mathrm{MI}\left(X,Y\right)\right\}} \tag{8}$$

其中，E 为互信息 $\mathrm{MI}(X,Y)$ 的期望。AMI 取值范围是[−1,1]，值越大代表聚类结果与真实情况越吻合。

3) 调整兰德系数（ARI, adjusted Rand index）

ARI 计算式为

$$\mathrm{ARI}(X,Y)=\frac{\mathrm{RI}(X,Y)-\mathrm{E}\{\mathrm{RI}(X,Y)\}}{\max(\mathrm{RI}(X,Y))-\mathrm{E}\{\mathrm{RI}(X,Y)\}} \tag{9}$$

RI 的作用是确定正确决策的占比，其计算式为 $\mathrm{RI}(X,Y)=$

$$\frac{\mathrm{TP}(X,Y)+\mathrm{TN}(X,Y)}{\mathrm{TP}(X,Y)+\mathrm{FP}(X,Y)+\mathrm{TN}(X,Y)+\mathrm{FN}(X,Y)} \tag{10}$$

ARI 取值范围是[−1,1]，反映了真实标签与聚类结果之间的吻合程度，ARI 越大表示聚类结果与真实情况吻合度越高。

6.2 实验数据集

爬取 Programmable Web 网站上的 Web API 服务作为实验对象。删除描述文本过短、重复注册的服务，保留 18 439 条有效服务，利用 Python 工具包对实验数据进行以下处理。

1) 文本分词，将服务描述中单词按空格分开。

2) 大小写转化，将服务描述中存在的大写字母单词全部转化为小写字母单词。

3) 去停用词，删除 a、an、the、of 等无效词语。

4) 词性还原，服务描述中单词存在时态、语态的变化，还原为最初的词态。

为了更好地进行聚类效果的内部评价指标和外部评价指标的对比，本文划分了 6 个数据集，即 DS_1～DS_6。其中，DS_1～DS_3 用于内部评价指标的对比，采取增量式构建，包含的服务类别如表 1 所示。DS_4～DS_6 用于外部评价指标对比，采取无交叉式构建，包含的服务类别如表 2 所示。

6.3 实验对比验证

实验机器配置为 i7-8750h，16 GB 内存，Windows10 操作系统，采用 Python3.6+Java 编写程序。外部指标对比时，簇粒度分别设置为 40、60、80、100、200、400，内部指标评价则采用服务标签作为聚类的类别标记，DS_4、DS_5、DS_6 的聚类数目分别设置为 10、10、14。服务表征向量主题个数依次设置为 20、40、60、80、100、150 和 200。所有实验数据均为不同主题、簇粒度数目下的数据均值。

6.3.1 服务表征向量生成质量验证实验

实验采用聚类算法为 K-means++，将本文所提出的带有主题分布概率修正因子的 GSDMM 模型命名为 R_GSDMM。服务表征向量生成质量实验对比数据如表 3～表 6 所示。

1) 内部评价指标比对

数据集 DS_1～DS_3 对应的内部指标评价数据如表 3 所示。表 3 中，行数据表示不同主题模型在各数据集中生成的服务表征向量所构建服务聚类对应的 SC 和 DBI。表 3 中每一列 SC 和 DBI 均由两部分值组成，其中 SC_1 和 SC_2 分别表示未采用与采用本文基于描述语境权重特征词提取的情况下生成服务表征向量聚类得到的 SC 指标值。DBI 列各数据值的含义与 SCl 列设置情况类似。

从 SC_1 与 SC_2 以及 DBI_1 与 DBI_2 的对比值可以

表 1　内部评价数据集

数据集	标签	数量/条
DS_1	Financial\Tools\Messaging\Payments\eCommerce\Social\Mapping\Government\Data\Science\Security\Email\Telephony\ Transpor- tation\ Reference\Enterprise	8 180
DS_2	DS1\Tools\Messaging\eCommerce\Science\SecurityData\Transportation\Sports\Education\Travel\Video\Advertising\Banking\ Cloud\ Music\Photos\Weather\Cryptocurrency\Stocks\Shipping\Games\Telephony\	12 043
DS_3	all	18 439

表 2　外部评价数据集

数据集	标签	数量/条
DS_4	Internet of Things\Database\Analytics\Backend\News Services\Medical\Events\Entertainment\Location\Media	1 402
DS_5	Banking\Cloud\Music\Photos\Weather\Cryptocurrency\Stocks\Shipping\Bitcoin\Project Management	2 171
DS_6	Tools\Messaging\eCommerce\Science\Security\Telephony\Data\Transportation\Sports\Education\Travel\Video\Games\Advertising	5 751

看出，参与实验的 7 种主题模型中，除了 HDP 模型外，其他 6 种模型均满足 $SC_2>SC_1$ 和 $DBI_2<DBI_1$。这说明本文所提出的服务特征词提取方法能够有效地将表达服务功能主题的关键词抽取出，减少了噪声词语，提高了服务表征向量的质量。同时，从有效模型数量可以看出，该方法具备较好的普适性，适合在绝大多数主题模型中使用。

表 3　服务表征向量质量验证实验数据

数据集	主题模型	SC_1	SC_2	DBI_1	DBI_2
DS_1	R_GSDMM	0.882	**0.908**	0.637	**0.546**
	GSDMM	0.857	0.892	0.765	0.621
	LDA	0.230	0.395	1.579	1.401
	LSA	0.073	0.145	1.952	1.493
	LDA_W2V	0.225	0.358	1.106	0.781
	BTM	0.227	0.272	1.391	1.308
	HDP	0.331	0.247	0.840	1.123
DS_2	R_GSDMM	0.870	**0.905**	0.696	**0.547**
	GSDMM	0.851	0.893	0.774	0.695
	LDA	0.192	0.326	1.660	1.499
	LSA	0.063	0.146	2.002	1.482
	LDA_W2V	0.184	0.215	1.235	1.156
	BTM	0.195	0.239	1.472	1.393
	HDP	0.361	0.300	0.770	1.079
DS_3	R_GSDMM	0.842	**0.896**	0.812	**0.557**
	GSDMM	0.826	0.863	0.843	0.808
	LDA	0.142	0.182	1.907	1.841
	LSA	0.042	0.130	2.218	1.564
	LDA_W2V	0.116	0.119	1.554	1.560
	BTM	0.150	0.189	1.679	1.565
	HDP	0.385	0.335	0.788	0.957

从表 3 中各个模型的 SC 与 DBI 对比数据可知，GSDMM 模型的指标值优于 LDA、LSA、LDA_W2V、BTM、HDP 等模型，而本文所提 R_GSDMM 模型的指标值优于 GSDMM 模型。这说明在相同服务表征向量生成语料情况下，R_GSDMM 模型生成服务表征向量质量最高，即本文所提带有主题概率分布修正因子的 GSDMM 模型有效提升了服务表征向量的生成质量。

通过 3 个数据集的平均值数据项，计算 R_GSDMM 对比 GSDMM 模型的性能提升，结果如图 4 所示。在图 4 中，SC_WM、DBI_WM 分别表示采用本文方法以及带有主题概率分布修正因子的 GSDMM 模型的性能提升值。相比传统 GSDMM 模型，在 3 个数据集上，本文方法将 SC 和 DBI 指标平均提升了约 6.9%和 30.6%。

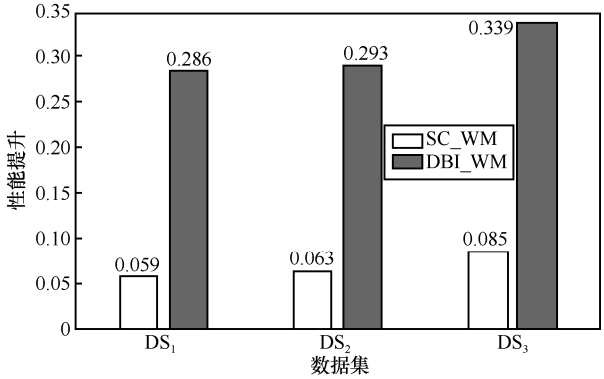

图 4　本文方法对 SC 与 DBI 指标提升对比

2) 外部评价指标对比

数据集 DS_4～DS_6 中进行的外部指标评价对比数据如表 4～表 6 所示。其中，AMI_1、NMI_1、ARI_1 和 AMI_2、NMI_2、ARI_2 分别表示未采用和采用本文所提基于语境权重特征词提取方法的聚类质量评价数据。

表 4　服务表征向量质量验证 DS_4 实验数据

主题模型	AMI_1	AMI_2	NMI_1	NMI_2	ARI_1	ARI_2
R_GSDMM	0.349	**0.443**	0.371	**0.465**	0.219	**0.274**
GSDMM	0.317	0.364	0.334	0.389	0.206	0.183
BTM	0.266	0.345	0.294	0.384	0.145	0.168
LDA	0.106	0.186	0.123	0.209	0.072	0.101
LSA	0.292	0.317	0.366	0.385	0.113	0.118
LDAW2V	0.018	0.022	0.031	0.036	0.007	0.009
HDP	0.033	0.066	0.048	0.107	0.015	0.012

表 5　服务表征向量质量验证 DS_5 实验数据

主题模型	AMI_1	AMI_2	NMI_1	NMI_2	ARI_1	ARI_2
R_GSDMM	0.669	**0.688**	0.682	**0.704**	0.520	**0.524**
GSDMM	0.634	0.659	0.653	0.686	0.453	0.471
BTM	0.286	0.335	0.408	0.436	0.089	0.132
LDA	0.368	0.499	0.378	0.520	0.267	0.384
LSA	0.548	0.628	0.596	0.665	0.279	0.402
LDAW2V	0.055	0.070	0.066	0.080	0.020	0.028
HDP	0.127	0.223	0.141	0.262	0.053	0.108

从表 4～表 6 可知，7 个主题模型生成的服务表征向量所构建的服务聚类，均满足 $AMI_2>AMI_1$、$NMI_2>NMI_1$ 和 $ARI_2>ARI_1$。这表明，采用本文所提方法构建的服务聚类质量比未采用此方法显著提高。因此，证明本文的特征词提取方法是有效的。此外，所有的模型在采用本文方法后，外部评价指标 AMI、NIMI 和 ARI 均得到提高，因此从外部评

价指标看，本文的特征词提取方法适用于主题模型范围非常广泛。

表 6　服务表征向量质量验证 DS_6 实验数据

主题模型	AMI_1	AMI_2	NMI_1	NMI_2	ARI_1	ARI_2
R_GSDMM	0.441	**0.536**	0.450	**0.551**	0.315	**0.342**
GSDMM	0.421	0.484	0.433	0.501	0.260	0.274
BTM	0.426	0.503	0.460	0.546	0.176	0.274
LDA	0.201	0.343	0.209	0.357	0.103	0.192
LSA	0.338	0.387	0.418	0.443	0.107	0.181
LDAW2V	0.130	0.218	0.138	0.225	0.052	0.088
HDP	0.031	0.146	0.039	0.152	0.010	0.063

纵向对比表 4～表 6 中各个主题模型的 AMI、NMI 以及 ARI 数据可以看出，不论是否采用本文所提特征词提取方法，R_GSDMM 模型所对应的服务聚类指标均优于 GSDMM 以及其他主题模型。综合数据集 DS_4～DS_6 中的指标数据，采用 GSDMM 模型、引入本文特征词提取后的 GSDMM 模型（用 GSDMM_W 表示）、同时引入本文特征词提取方法与带有主题概率分布修正因子的 GSDMM 模型（用 R_GSDMM_W 表示）所对应的 AMI、NMI 以及 ARI 指标平均值如图 5 所示。相比传统 GSDMM 模型，在同一服务描述文本语料前提下，本文方法将生成聚类的 AMI、NMI 以及 ARI 平均值分别提升 21.7%、21.1%以及 23.9%。

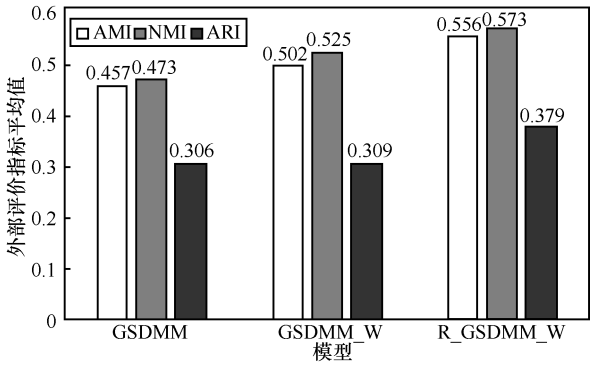

图 5　不同模型外部评价指标平均值对比

6.3.2　聚类算法效果对比实验

实验选取 K-means++、AGNES、BIRCH 以及 GMM 这 4 种常用算法，采用本文方法构建的服务表征向量，各聚类算法生成的聚类质量评价指标如表 7～表 9 所示。

1) 内部评价指标对比

数据集 DS_1～DS_3 中，K-means++、AGNES、BIRCH 以及 GMM 这 4 种聚类算法的 SC 与 DBI 得分平均值如表 7 所示。从表 7 可知，K-means++算法生成的服务聚类质量在 SC 与 DBI 评价指标的占优数量上显著高于其他算法。

表 7　聚类效果质量验证实验数据

聚类数目	聚类方法	SC	DBI
DS_1	K-means++	**0.908**	0.546
	AGNES	0.906	**0.540**
	BIRCH	0.898	0.644
	GMM	0.847	0.720
DS_2	K-means++	**0.905**	**0.547**
	AGNES	0.903	0.587
	BIRCH	0.902	0.606
	GMM	0.840	0.816
DS_3	K-means++	**0.896**	**0.557**
	AGNES	0.891	0.579
	BIRCH	0.889	0.687
	GMM	0.794	1.087

数据集 DS_1～DS_3 对应不同聚类算法的 SC 和 DBI 得分平均值如图 6 所示。从图 6 可以看出，K-means++算法的 SC 得分平均值略优于其他算法，DBI 得分平均值则显著优于其他算法。

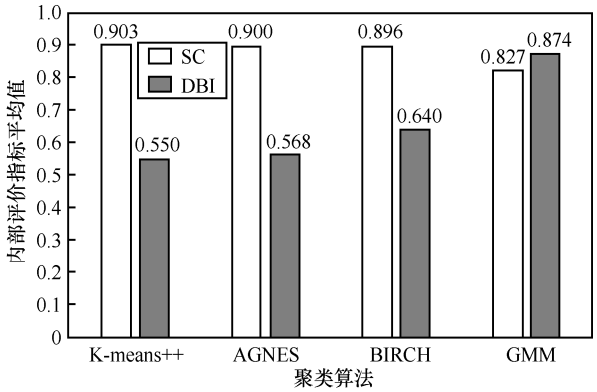

图 6　不同聚类算法内部评价指标平均值对比

综上，在 SC 和 DBI 平均值占优数量和平均值大小 2 个层面，K-means++算法均优于其他算法。

2) 外部评价指标对比

数据集 DS_4～DS_6 中，K-means++、AGNES、BIRCH 以及 GMM 这 4 种聚类算法的 AMI、NMI 与 ARI 评分值如表 8 所示。从表 8 可以看出，K-means++在数据集 DS_5 和 DS_6 上的评分具有领先优势。

4 种算法在数据集 DS_4～DS_6 中外部指标 AMI、

NMI 与 ARI 评分平均值如图 7 所示，K-means++在 3 项外部指标的平均得分中均领先于其他 3 种算法。因此，不论是从外部指标 AMI、NMI 与 ARI 评分平均值的领先数量还是平均值大小，K-means++算法均优于其他算法。

表 8　聚类效果质量验证实验数据

数据集	聚类算法	AMI	NMI	ARI
DS_4	K-means++	0.443	0.465	0.274
	GMM	0.428	0.460	0.255
	BIRCH	0.329	0.373	0.188
	AGNES	**0.445**	**0.467**	**0.277**
DS_5	K-means++	**0.688**	**0.704**	**0.520**
	GMM	0.678	0.702	0.522
	BIRCH	0.496	0.568	0.279
	AGNES	0.687	0.703	0.519
DS_6	K-means++	**0.536**	**0.551**	**0.342**
	GMM	0.534	0.541	0.354
	BIRCH	0.456	0.485	0.253
	AGNES	0.527	0.539	0.323

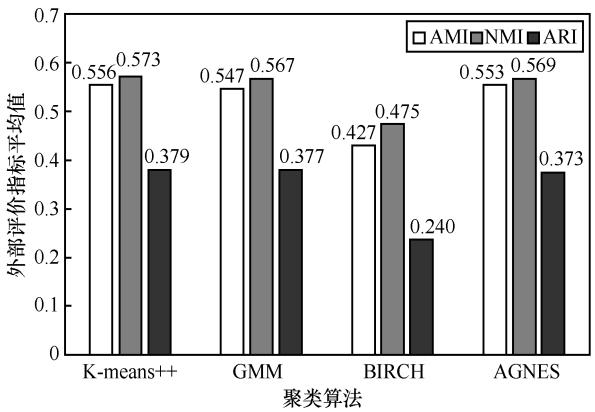

图 7　不同聚类算法外部评价指标平均值对比

本节实验中，内部指标与外部指标最优值次数进行了分类汇总如表 9 所示。在共计 39 轮次的指标统计中，K-means++获得 30 次最优值，其中包含一次 K-means++与 AGNES 并列最优值。因此，本文选用 K-means++算法作为最终的服务聚类算法是合理和有效的。

相比传统 GSDMM 生成服务表征向量，本文方法因引入 Word2Vec 生成词向量而略增加耗时。由于 Word2Vec 词向量生成耗时显著低于 GSDMM 词向量生成，且在提取特征词后，语料词语数量减少和词频相对集中进一步降低 GSDMM 生成向量的时间，因此，本文方法总体计算复杂度未明显升高。

表 9　聚类算法最优指标值统计结果

聚类算法	SC	DBI	AMI	NMI	ARI	SUM
K-means++	13	11	2	2	2	30
AGNES	1	4	1	1	1	8
BIRCH	2	0	0	0	0	2
GMM	0	0	0	0	0	0

7　结束语

为了提高采用短文本自然语言进行服务功能描述的 Web 服务聚类效果，本文构建了一种高质量的服务表征向量生成以及聚类方法，采用适于短文本聚类的 GSDMM 模型作为生成服务表征向量的主题模型，从词汇语料和主题概率分布 2 个维度对服务表征质量生成方法进行了改进；建立的特征词提取算法有效去除服务描述信息中的噪声，构建的概率修正因子平衡了 GSDMM 模型生成服务表征向量时的关键主题与次要主题之间的概率；构建了基于 K-means++的服务聚类算法。采用 5 种评价指标对本文提出的聚类方法效果进行评估，多轮次实验结果显示本文所提方法提高了服务表征向量的生成质量和服务聚类效果。

未来工作将进一步优化服务描述文本中词语的语境权重计算方法以及 GSDMM 模型的主题概率分布修正因子，提升服务表征向量的质量。此外，将对服务之间的协作关系进行建模、量化，并融入服务聚类过程中，以期进一步提高聚类效果。

参考文献：

[1] NIKNEJAD N, ISMAIL W, GHANI I, et al. Understanding service-oriented architecture (SOA): a systematic literature review and directions for further investigation[J]. Information Systems, 2020, 91: 101491.

[2] 赵晨阳, 王俊岭. 基于隐含上下文支持向量机的服务推荐方法[J]. 通信学报, 2019, 40(9): 61-73.
ZHAO C Y, WANG J L. Service recommendation method based on context-embedded support vector machine[J]. Journal on Communications, 2019, 40(9): 61-73.

[3] HALILI F, RAMADANI E . Web services: a comparison of soap and rest services[J]. Modern Applied Science, 2018, 12(3):175-183.

[4] 贾春福, 李瑞琪, 王雅飞. 基于同态加密的 DBSCAN 聚类隐私保护方案[J]. 通信学报, 2021, 42(2): 1-11.
JIA C F, LI R Q, WANG Y F. Privacy protection scheme of DBSCAN clustering based on homomorphic encryption[J]. Journal on Communications, 2021, 42(2): 1-11.

[5] 曹步清, 肖巧翔, 张祥平, 等. 融合 SOM 功能聚类与 DeepFM 质量预测的 API 服务推荐方法[J]. 计算机学报, 2019, 42(6): 1367-1383.
CAO B Q, XIAO Q X, ZHANG X P, et al. An API service recommendation method via combining self-organization map-based functionality clustering and deep factorization machine-based quality prediction[J]. Chinese Journal of Computers, 2019, 42(6): 1367-1383.

[6] AGARWAL N, SIKKA G, AWASTHI L K. Enhancing Web service clustering using length feature weight method for service description document vector space representation[J]. Expert Systems with Applications, 2020, 161: 113682.

[7] NABLI H, BEN D R, BEN A I A. Efficient cloud service discovery approach based on LDA topic modeling[J]. Journal of Systems and Software, 2018, 146: 233-248.

[8] VADIVELOU G, ILAVARASAN E. Performance evaluation of semantic approaches for automatic clustering of similar Web services[C]//2014 World Congress on Computing and Communication Technologies. Los Alamitos: IEEE Computer Society, 2014: 237-242.

[9] KIM S, PARK H, LEE J. Word2Vec-based latent semantic analysis (W_2V-LSA) for topic modeling: a study on blockchain technology trend analysis[J]. Expert Systems With Applications, 2020, 152: 113401.

[10] CAO B Q, LIU X, LIU J X, et al. Domain-aware Mashup service clustering based on LDA topic model from multiple data sources[J]. Information and Software Technology, 2017, 90: 40-54.

[11] DAS R , ZAHEER M , DYER C. Gaussian LDA for topic models with word embeddings[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: ACL Press, 2015:795-804.

[12] CHENG X Q, YAN X H, LAN Y Y, et al. BTM: topic modeling over short texts[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2928-2941.

[13] BASKARA A R, SARNO R. Web service discovery using combined bi-term topic model and WDAG similarity[C]//2017 11th International Conference on Information & Communication Technology and System. Piscataway: IEEE Press, 2017: 235-240

[14] JIANG Y C, TAO D D, LIU Y Z, et al. Cloud service recommendation based on unstructured textual information[J]. Future Generation Computer Systems, 2019, 97: 387-396.

[15] AGARWAL N, SIKKA G, AWASTHI L K. Evaluation of Web service clustering using Dirichlet multinomial mixture model based approach for dimensionality reduction in service representation[J]. Information Processing & Management, 2020, 57(4): 102238.

[16] YIN J, WANG J. A Dirichlet multinomial mixture model-based approach for short text clustering[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. New York: ACM Press, 2014: 233-242.

[17] 谢晓兰, 曾兰英, 翟青海. 制造云服务组合中支持服务关联的 QoS 感知评估模型[J]. 通信学报, 2021, 42(1): 118-129.
XIE X L, ZENG L Y, ZHAI Q H. QoS aware evaluation model supporting service correlation in manufacturing cloud service composition[J]. Journal on Communications, 2021, 42(1): 118-129.

[18] LIANG T, CHEN L, YING H, et al. Co-clustering WSDL documents to bootstrap service discovery[C]//2014 IEEE 7th International Conference on Service-Oriented Computing and Applications. Piscataway: IEEE Press, 2014: 215-222.

[19] WU J, CHEN L, ZHENG Z B, et al. Clustering Web services to facilitate service discovery[J]. Knowledge and Information Systems, 2014, 38(1): 207-229.

[20] 张键红, 武梦龙, 王晶, 等. 云环境下安全的可验证多关键词搜索加密方案[J]. 通信学报, 2021, 42(4): 139-149.
ZHANG J H, WU M L, WANG J, et al. Secure and verifiable multi-keyword searchable encryption scheme in cloud[J]. Journal on Communications, 2021, 42(4): 139-149.

[21] CAO B Q, LIU X F, RAHMAN M M, et al. Integrated content and network-based service clustering and Web APIs recommendation for mashup development[J]. IEEE Transactions on Services Computing, 2020, 13(1): 99-113.

[22] LIZARRALDE I, MATEOS C, ZUNINO A, et al. Discovering Web services in social Web service repositories using deep variational autoencoders[J]. Information Processing & Management, 2020, 57(4): 102231.

[23] ZHANG N, WANG J, HE K, et al. Mining and clustering service goals for RESTful service discovery[J]. Knowledge and Information Systems, 2019, 58(3): 669-700

[24] 刘建勋, 石敏, 周栋, 等. 基于主题模型的 Mashup 标签推荐方法[J]. 计算机学报, 2017, 40(2): 520-534.
LIU J X, SHI M, ZHOU D, et al. Topic model based tag recommendation method for Mashups[J]. Chinese Journal of Computers, 2017, 40(2): 520-534.

[25] 石敏, 刘建勋, 周栋, 等. 基于多重关系主题模型的 Web 服务聚类方法[J]. 计算机学报, 2019, 42(4): 820-836.
SHI M, LIU J X, ZHOU D, et al. Multi-relational topic model-based approach for Web services clustering[J]. Chinese Journal of Computers, 2019, 42(4): 820-836.

[26] SHI M, TANG Y F, LIU J X. Functional and contextual attention-based LSTM for service recommendation in Mashup creation[J]. IEEE Transactions on Parallel and Distributed Systems, 2019, 30(5): 1077-1090.

[27] YE H, CAO B, CHEN J, et al. A Web services classification method based on GCN[C]//2019 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking. Piscataway: IEEE Press, 2019: 1107-1114.

[作者简介]

胡强（1980− ），男，山东邹城人，青岛科技大学副教授、硕士生导师，主要研究方向为服务计算、人工智能。

沈嘉吉（1997− ），男，上海人，青岛科技大学硕士生，主要研究方向为服务计算。

荆广辉（1996− ），男，山东日照人，青岛科技大学硕士生，主要研究方向为文本挖掘、推荐系统。

杜军威（1974− ），男，山东文登人，青岛科技大学教授、博士生导师，主要研究方向为软件工程、人工智能。