基于多重关系主题模型的 Web 服务聚类方法

石 敏 刘建勋 周 栋 曹步清 文一凭
(湖南科技大学知识处理与网络化制造湖南省普通高校重点实验室 湖南 湘潭 411201)

摘 要 如何有效地发现合适的 Web 服务是面向服务计算领域需要解决的核心问题之一.随着 Internet 上 Web 服务数量的不断增加,服务的自动发现面临着极大的挑战.将功能相似的 Web 服务进行聚类是一种有效的服务发现与服务管理方法.目前国内外主流的方法为挖掘 Web 服务的隐含功能语义信息,如使用 LDA 主题模型训练提取 Web 服务功能描述文档的主题信息,然后基于某种聚类算法如 *K*-means 将隐含主题分布相似的 Web 服务聚为一类.然而,Web 服务的功能描述文档通常短小,目前大部分主题模型无法对短文本进行良好地建模,从而影响了 Web 服务聚类的效果.针对该问题,文中提出了一种考虑多重 Web 服务关系的概率主题模型 MR-LDA,其可对 Web 服务之间相互组合的关系以及 Web 服务之间共享标签的关系进行建模,能有效提高 Web 服务聚类的精度.同时,基于该 MR-LDA 主题模型进一步提出了一种有效的 Web 服务聚类算法 MR-LDA+,该算法首先利用上述多重 Web 服务关系信息对 Web 服务隐含主题分布概率矩阵进行修正,然后根据这些隐含主题对 Web 服务进行聚类.基于 ProgrammableWeb 收集的真实数据实验表明,文中所提出的方法明显优于其它 Web 服务聚类算法.

关键词 Web 服务;聚类;多重关系网络;先验知识;主题模型
中图法分类号 TP301 **DOI 号** 10.11897/SP.J.1016.2019.00820

Multi-Relational Topic Model-Based Approach for Web Services Clustering

SHI Min LIU Jian-Xun ZHOU Dong CAO Bu-Qing WEN Yi-Ping
(*Key Laboratory of Knowledge Processing & Networked Manufacturing, Hunan University of Science & Technology, Xiangtan, Hunan* 411201)

Abstract Web service discovery is a significant and nontrivial task in the domain of Web service computing. With the rapid growth in the number of Web services on the Internet, e.g., an increasing number of enterprises tend to make public their software and other resources in the form of services within and outside the organizations, locating exactly the desired Web services is becoming increasingly hard for users. It has been shown that clustering Web services according to their functionalities is an efficient way to facilitate Web services discovery as well as services management. The clustering results can help us better understanding the more fine-granted categorically functional features of Web services, and meanwhile significantly reduce the searching space and retrieval time with respect to a given user query. Existing methods on this topic mainly focus on mining the semantic functional information of Web services, etc., adopting LDA to firstly elicit the functional semantics of Web services and then clustering Web services according to their topic distributions based on some clustering methods such as *K*-means. However, the natural description documents of Web services generally contain limit number of words. It is hard for most existing LDA-based methods to model short text documents, which may seriously degrade

收稿日期:2017-12-26;在线出版日期:2018-07-05.本课题得到国家自然科学基金(61872139,61876062,61572187)、湖南省自然科学基金(2018JJ2139)、湖南省教育厅创新平台开放基金项目(17K033)资助.**石 敏**,男,1991 年生,博士研究生,助理研究员,中国计算机学会(CCF)会员,主要研究方向为社会网络与服务计算等.E-mail:toshimin132@gmail.com.**刘建勋**(通信作者),男,1970 年生,博士,教授,博士生导师,主要研究领域为服务计算与云计算、工作流管理的理论与应用等.E-mail:ljx529@gmail.com.**周 栋**,男,1979 年生,博士,教授,主要研究方向为信息检索、自然语言处理、机器学习等.**曹步清**,男,1979 年生,博士,副教授,主要研究方向为服务计算与云计算等.**文一凭**,男,1981 年生,博士,副教授,主要研究方向为工作流管理与服务计算等.

the Web service clustering accuracy. To narrow such negative effect, this paper aims to mitigate the data sparse issue by mining and leveraging some types of auxiliary information that is helpful to the service clustering problem. After a careful exploration of the Web service multi-relational network that is naturally established from users' frequent behaviors (e. g., invoking and annotating) of using Web services, we found that the composition relationships between Web services and the annotation relationships between Web services sharing identical tags could be used to improve the semantics extraction and service clustering processes, i. e., services with annotation relationships tend to share similar functional semantics, whereas services with composition relationships should follow dissimilar latent topic distributions. Based on these observations, we first propose a multi-relational probabilistic topic model, MR-LDA, to simultaneously model the composition relationships as well as the annotation relationships, where services with either composition or annotation relationship will exert an impact on each other during the topic sampling process for each word. Based on the topic model, we further propose an efficient Web service clustering algorithm, MR-LDA+, to firstly revise the obtained topic distribution probabilities of Web services such that above two kinds of relationships information can be explicitly encoded, and then based on it performs the Web services clustering. We extensively evaluate the proposed topic model and clustering approach on a real world dataset crawled from ProgrammableWeb. The experimental comparisons demonstrate that our approach significantly outperforms other state-of-the-art Web services clustering methods. In addition, we also design experiments to verify if the used auxiliary information can help to extract more accurate semantics by conducting service classification and vector visualization tasks based on the Support Vector Machine and t-SNE algorithms, respectively, and both the classification performance and vector visualization results demonstrate the positive impact of the introduced auxiliary relationships information.

Keywords　Web services; clustering; multi-relational network; prior knowledge; topic model

1　引　言

Web 服务是一种依赖于互联网的应用系统，它面向互联网用户提供各种数据计算和资源共享等服务. 随着 Web 2.0、移动互联网、物联网与云计算等技术的迅猛发展，大量基于 SOA(Service Oriented Architecture)架构的互联网应用被创建(例如移动应用 APP 和微信小程序等各类微小服务)，而 Web 服务逐渐成为实现 SOA 架构的主流技术. 为了使用和集成现有的 Web 服务，用户需要在 Web 服务注册平台上从海量的 Web 服务中寻找和匹配满足需求的服务. 然而，随着互联网上 Web 服务数量以及服务功能种类不断增多，从数量庞大且功能属性差异难以界定的服务集合中精确地定位满足用户特定业务需求的服务日益变得困难和费时. 据统计，在 ProgrammableWeb(目前最大、最活跃的 Web 服务发布和共享平台)上每天都会产生数十个新的被称为 API(Application Programming Interface)的 Web 服务，如何辅助用户有效地发现合适的 Web 服务是面向服务计算领域需要解决的核心问题之一. 前人研究表明，将功能相似的 Web 服务进行聚类能够有效地改善 Web 服务发现的过程和效果[1-3]. 首先，同一聚类中的 Web 服务提供的功能相似，利于用户进行垂直服务检索，能极大地减小服务搜索的空间和时间[3]；第二，将 Web 服务进行聚类能帮助用户梳理和理解海量 Web 服务的功能层次结构，方便对它们进行管理和使用，如进行后续的服务发现与选择[4-7]、服务替换[7]、服务组合[1]以及服务推荐[8-10]等过程.

目前已存在一些 Web 服务聚类方法[1-2,4-7]. 其中，基于挖掘 WSDL(Web Service Description Language)文档特征的方法被广泛采用[4,11]，这些方法首先抽取 WSDL 文档的关键特征，如 WSDL 描述内容、WSDL 类型、WSDL 端口以及 Web 服务名称等信息. 接着基于这些提取的特征信息，采用如余弦相似度等方法计算 Web 服务之间的相似度从而对它们

进行聚类.然而,由于 WSDL 文档通常包含很少的描述内容,这些算法通常无法取得较满意的聚类效果[2-3,12].此外,基于 WSDL 文档的方法通常忽略了 Web 服务之间语义信息的关联[2-3].因此,许多方法采用 LDA(Latent Dirichlet Allocation)[13] 主题模型或其扩展主题模型[2-3,12]提取 Web 服务的隐含主题信息,使用低维的主题向量对 Web 服务的功能属性进行编码,然后基于此计算服务之间的相似度和进行服务聚类.这些基于主题模型的方法通常能取得比较好的效果[2-3].然而,Web 服务 WSDL 文档或者描述文档通常篇幅较短、特征稀疏和信息量少,而目前大部分主题模型都无法对缺乏训练语料的短文本进行很好的建模[14].例如 LDA 模型需要基于大量的已知观测样本来推测隐含的后验主题分布概率[15].虽然有些主题模型在训练的过程中引入了辅助信息,如词聚类信息[3]、标签信息[12]等,有利于提取得到更加准确的 Web 服务隐含功能语义信息.但相比传统的 LDA 主题模型,现有改进的主题模型对 Web 服务聚类准确度的提升并不明显.因此,寻找更加优化的聚类方案具有重要的研究意义.

然而,现实世界中 Web 服务并非独立存在,它们之间通常具有某种关联关系并以某种方式相互影响[16],如 Web 服务之间互相组合的关系以及 Web 服务之间共享标签的关系.图 1 所示为一个真实的 Web 服务多重关系网络部分示例,该图中包含三种实体(即 Mashup、API 和标签(Tag))和两种关系(即组合关系(composition relationship,图中虚线箭头)与标注关系(annotation relationship,图中实线箭头)).其中,组合关系表示某个 Mashup 组合了多个 APIs(下文中亦称 Web 服务),如图中名称为“City Tube”的 Mashup 组合了“YouTube”和“Google maps”两个 APIs.标注关系表示某些标签(或关键字)被用于标注 APIs,如图中标签“mapping”被用于标注了“Google maps”和“foursquare”两个 APIs.随着服务不断地被创建以及被用户频繁地调用,上述 Web 服务多重关系网络将日益演化和变得庞大.在复杂的 Web 服务生态系统中,Web 服务除了包含丰富的本体信息外还与网路中其它节点(如标签、API 和 Mashup 等)具有直接或间接的关联关系,因此可从多个角度对 Web 服务的功能属性特征进行刻画[17].例如,图 2 所示为来自 Programmable-Web 的一个真实 Web API,其包含服务名称、标签描述文档和功能类别.可以看出,Web 服务附带的这些元数据信息可以较清楚地传达该服务所具有的功能特征.此外,服务涉及的功能属性一定程度上可基于服务的网络拓扑结构推断出来.如网络中共享标签的服务在某个或某些功能上具有相似性,因此基于拓扑结构可通过功能特征已知的服务推断出未知的服务.有研究表明[18-20],节点的网络拓扑结构信息(Structure Information)和节点的文本内容信息(Textual Content Information)相互独立又互为补充.例如,在一个科技文献数据库中,对文献之间关联关系的挖掘分析既可以从纯文本的视角展开,也可从文献引用网络结构的视角出发,并且两种方式对于知识的发现具有近似等价的效果[16,19-20].尤其在某一方面信息缺乏时(如 Web 服务的功能描述信息较稀疏),这种互为补充的信息(如服务之间组合关系和共享标注关系等)对算法性能的提高具有极大的促进作用[10,19].对于 Web 服务功能聚类问题,也可从 Web 服务网络拓扑结构的角度挖掘一些有利的知识,如可近似地认为具有组合关系的两个 Web 服务功能上相互排斥,因为实际应用中被组合在一起(如 Mashup)的两个 Web 服务通常以协同互补的方式用于解决某个特定的业务问题,因此具有不同的功能特征隶属于不同的功能领域类别.同时,标签通常被用于概括性地传达 Web 服务所具有的功能属性,当两个 Web 服务存在共享标签的关系(共享标注关系)时,则可认为它们在功能上具有一定的相似性,可能隶属于相同或相似的功能领域类别.基于本文所采用地数据集进行统计表明,被组合(如 Mashup)在一起的服务,其属于不同功能类别的概率超过 90%.同时,当服务间共享一个标签时,其属于相同类别的概率超过 30%,当共享 3 个标签时该概率上升至 50%.上述 Web 服务组合关系与共享标注关系能进一步改善 Web 服务聚类的过程,尤其在缺乏充分的训练语料时,可作为 Web 服务聚类算法的重要辅助和补充信息.

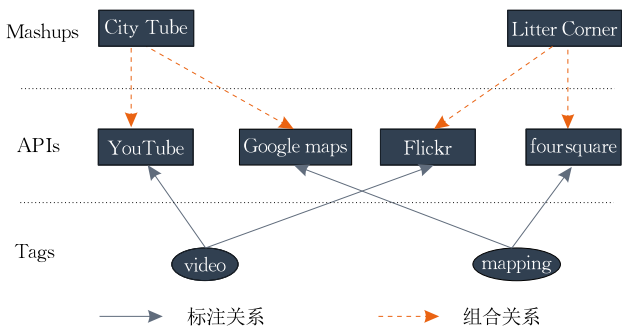


图 1 Web 服务多重关系网络部分示例

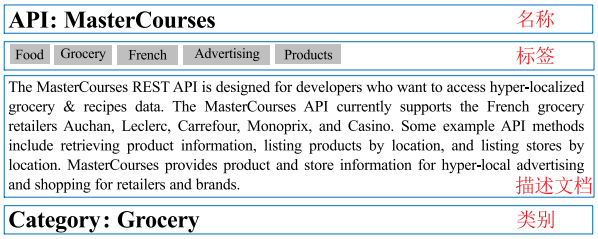


图 2　ProgrammableWeb 上真实的 API 示例

针对目前大部分主题模型无法对缺乏描述文本的 Web 服务进行良好建模，以及仅基于文本信息学习到的隐含功能主题空间无法很好地度量服务之间的相似性等问题，本文提出了一种考虑多重 Web 服务关系的概率主题模型 MR-LDA(Multi-Relational LDA). 该主题模型不仅对 Web 服务描述文本信息进行建模，同时还考虑引入 Web 服务之间的拓扑结构信息(如组合关系与共享标注关系等)作为内容层面建模的补充. 模型认为具有组合关系的 Web 服务的隐含主题分布相似度较小，对应隐含主题空间中的距离较小；而具有共享标注关系的 Web 服务的隐含主题分布相似度较大，对应隐含主题空间中的距离较大. 传统的基于 LDA 的主题模型只考虑对文本内容进行建模，因此学习到的低维主题分布向量只对内容层面的信息进行了编码. 而本文提出的主题模型同时考虑内容和网络拓扑结构层面信息，使得学习到的向量同时编码了 Web 服务的功能文本相似性以及网络拓扑结构相似性. 并且这两种层面的相似性互为补充，共同全局地作用于度量两个服务之间功能的相似程度，因此能明显提升 Web 服务聚类的精度. 接着，为进一步探索上述讨论的 Web 服务网络结构信息对服务聚类效果的影响，基于所提出的 MR-LDA 模型本文提出一种显式的融合 Web 服务多重关系与主题信息的 Web 服务聚类算法，MR-LDA+. 其首先基于主题模型得到 Web 服务描述文档的主题概率分布信息，然后使用 Web 服务组合关系和共享标注关系对 Web 服务的隐含主题概率矩阵进行迭代修正，直至融入关系后的 Web 服务隐含主题分布矩阵趋于稳定或者迭代次数结束. 最后，根据 Web 服务的主题信息对服务进行聚类. 根据从 ProgrammableWeb 收集的真实数据进行的实验表明，使用 Web 服务的多重关系信息对主题概率进行修正后能进一步改善聚类的效果.

本文第 2 节分析目前相关的研究工作；第 3 节介绍本文算法涉及的预备知识；第 4 节具体介绍基于多重关系主题模型的 Web 服务聚类算法；第 5 节分析实验数据集并对提出的方法进行实验评估；第 6 节对全文的工作进行总结.

2　相关工作

服务聚类是一种有效的辅助服务管理与组合的技术[1,17]，其主要目标为依据服务的功能将服务归于不同的类型中，即将所有的服务划分成若干个功能独立的类别，使得同一类中服务的功能具有较高的相似性，而不同类别之间的服务功能差异性较大. 目前国内外众多学者开展了对服务聚类的研究工作[2-5]. 根据其关注点主要划分为两大类[22]：基于功能的(Function-based) Web 服务聚类方法[2-3]和基于非功能的(Non-Function-based) Web 服务聚类方法[8-9].

(1) 基于功能的服务聚类方法

该类方法可从两方面进行归纳：服务聚类所使用的数据信息(如文本和服务网络拓扑结构等)和聚类时所使用的方法(如关键词匹配和主题模型等).

服务聚类所使用的方法一般与其聚类时所使用的数据有关. 从聚类时所使用的数据与信息看，通常的做法为从 Web 服务描述文档(WSDL)[4,23]或者简洁和轻量化的 RESTful 服务(Web API)描述信息[17,24]中提取特征，例如服务名称、端口号、服务类型、服务消息等. 然后采用诸如余弦相似度等方法计算服务之间的相似性，进而采用传统聚类方法，例如 QT(Quality Threshold)聚类算法[4]和 K-means 等对这些服务进行聚类. 由于 Web 服务描述文档中可用的自然语言描述信息较少，以及不同的服务开发者编写的描述文档比较随意和个性化，导致严重的特征稀疏和词汇语言鸿沟问题[25]，使得难以训练得到鲁棒的服务聚类模型[3]. 为缓解该问题，Chen 等人[11-12]提出整合 WSDL 文档和服务标签信息来改进服务聚类的精度，他们认为标签可以表征一个服务的功能特征，并与 WSDL 文档一起可以更加准确地判断服务归属的功能类别. 针对服务功能描述文档词汇稀疏导致聚类效果不佳等问题，Shi 等人[3]提出首先基于 Word2vec[26]工具将所有词汇根据语义进行聚类，然后在训练服务的主题信息时同时考虑与服务描述文档中的词汇属于同一聚类的词汇的辅助作用，可以显著地提高服务聚类的效果. 上述方法均主要集中于使用描述文档数据或者扩展的词汇和标签辅助数据[24]来对服务进行聚类. 然而，除了挖掘文本数据进行服务聚类外，很多方法都尝试挖掘 Web 服务网络的拓扑结构数据信息以辅助加强基于文本的服务聚类方法的效果[2,15,17]. 例如 Cao 等

人[2,17]认为借助 Mashup 服务的网络结构有利于学习得到更加准确的文档主题信息,因此提出一种同时对服务拓扑结构和描述文档进行语义建模的服务聚类模型.然而,他们的方法主要用于对 Mashup 服务进行聚类,未涉及如何对 Web 服务网络进行建模.而本研究聚焦 Web 服务的聚类,提出引入丰富的 Web 服务关系特征(如服务之间的相互组合关系)辅助提取更加精确的服务功能语义信息用于 Web 服务聚类.

过去也有较多研究聚焦于探索更加高效的服务聚类算法[13,21,27-28].例如,为克服传统服务聚类方法需要大量人工标注的训练样本这一问题,Liu 等人[29]提出首先基于小规模人工标注的样本训练得到一个初步的 SVM(Support Vector Machine)分类器,然后基于此为其它未被人工标注的服务推荐潜在的标签,最后基于这些标注样本对服务进行服务聚类.针对传统的基于关键词匹配的方法[4]忽略了服务之间功能语义的关联等问题,很多方法开始采用语义建模工具如 LDA 或其扩展模型提取服务的功能语义信息,使用固定长度的和低维的主题向量对功能特征进行编码,并基于此计算服务之间的相似性进行服务聚类[1,13,21].例如,Cao 等人[17]基于 LDA 主题模型提取 Mashup 服务的隐含主题信息,然后基于此对 Mashup 服务进行聚类;为弥补 WSDL 文档信息匮乏等问题,很多主题模型在训练过程中引入辅助信息[3,11-12,27,30],如标签信息[11-12,23]、词聚类信息[3]、服务组合关系[10,15,31]等,从而提高主题信息提取的准确性和改善服务聚类的效果.例如,Chen 等人[12]提出了一种增强的概率主题模型(WT-LDA),可同时对 Web 服务描述文档和服务标签信息进行建模;Shi 等人[10]改进关系主题模型(RTM)提出了 MAT(Mashup-API-Tag)概率主题模型,同时将服务描述信息、标签信息以及服务之间的组合关系信息进行建模,从而提高主题信息提取的精度.这些措施均可一定程度上缓解数据稀疏问题和改善服务聚类的效果.然而,现有方法大都通过引入辅助信息局部地改进主题模型训练的方式从而使得到的低维向量能够对更多的信息进行编码(如融合了标签信息),并未全面地考虑融合有利于服务聚类任务的信息.本研究通过对 Web 服务生态网络进行较深入的分析,发现了两类有利于服务聚类的客观信息,即 Web 服务组合关系和共享标注关系.与传统的仅仅基于功能文本相似性聚类的角度不同,本文认为基于用户使用 Web 服务的轨迹所形成的 Web 服务网络拓扑结构在某种程度上具有大众服务聚类的效果,如经常组合在一起的两个服务很大可能性属于不同的功能类别.在服务聚类过程中,网络层面的信息与功能文本层面的信息可以互为补充,从而进一步弥补功能文本词汇稀疏问题和改善服务聚类的效果.

(2)基于非功能的服务聚类方法

基于非功能的服务聚类方法通常首先将服务根据功能属性进行聚类,然后在每个对应的功能类别中再依据不同方面的 QoS(Quality of Service)属性(如价格、可用性、响应时间、可靠度和声誉等)进行聚类[32].目前有很多研究关注非功能的服务聚类[6,8-9].例如,Liu 等人[33]提出一种基于服务本体信息的服务聚类算法,挖掘的信息包括服务名称、性能、接口和 QoS 属性.Zhou 等人[34]使用遗传算法基于 QoS 实现对服务的聚类.为避免算法仅返回局部最优解,他们在遗传算法中引入熵的概念用来度量和改进种群多样性.Chen 等人[35]提出了一种基于历史 QoS 数据的物理特征相似的服务聚类算法,保证同一类别中的服务具有相似的物理环境特征.虽然仅考虑非功能属性的算法通常具有相对小的执行复杂度[28],但由于服务的非功能属性特征通常很难获取且很不稳定(如响应时间动态变化等),因此该类算法通常不具有很好的扩展性.

3 预备知识

3.1 问题定义

定义 1(Web 服务功能描述文档). 一个 Web 服务 h 的功能描述文档可表示为 $D_h=\{w_1,w_2,\cdots,w_{|D_h|}\}$,其中 w_i 表示文档中的第 i 个词汇,$|D_h|$ 表示文档包含的词汇总数.

定义 2(Web 服务组合关系). Web 服务可能被用户(如 Mashup 开发者)调用一次或多次.如果两个服务 A1 和 A2 至少一次同时被某个用户调用,则认为服务 A1 与 A2 具有组合关系.与 Web 服务 h 具有组合关系的集合可表示为 $C_h=\{cd_1^h,cd_2^h,cd_3^h,\cdots,cd_{|C_h|}^h\}$,其中 $|C_h|$ 表示与服务 h 具有组合关系的服务数目.

定义 3(Web 服务共享标注关系). 每个服务包含若干个标签,如 Web 服务 h 的标签集可表示为 $T_h=\{t_1,t_2,\cdots,t_{|T_h|}\}$,其中 t_i 表示 Web 服务的第 i 个标签并称 t_i 与 Web 服务 h 具有标注关系,$|T_h|$ 是该服务包含的标签的总数.不同 Web 服务可能包含一个或多个相同的标签,当两个服务 A1 和 A2 共享至少一个标签时,则认为它们具有共享标注关系.与

Web 服务 h 具有共享标注关系的 Web 服务集合可表示为 $A_h=\{ad_1^h,ad_2^h,ad_3^h,\cdots,ad_{|A_h|}^h\}$，其中 $|A_h|$ 表示与 h 具有共享标注关系的 Web 服务数目.

定义 4(Web 服务多重关系网络). 一个多重关系网络(Multi-relational Network)可以定义为图 $G=(V,E)$. 其中 V 为所有顶点(vertex)的集合，$E=\{E_1,E_2,\cdots,E_w\}$ 为所有边(Edge)的集合，$E_q\subseteq(V\times V):1\leq q\leq w$ 表示具有某类关系的边的集合，w 表示边关系的种类数. 在 Web 服务多重关系网络中，顶点 V 为所有 Web 服务和标签(Tag)的集合，边 $E=\{A,C\}$ 为所有标注关系和组合关系的集合.

定义 5(Web 服务聚类任务). Web 服务聚类的任务是根据 Web 服务所提供的功能，将它们聚到不同的功能类别领域中. 如图 2 所示为 ProgrammableWeb 上一个真实名称为“MasterCourses”的 API. 可以看到，该 API 有 5 个不同的标签和一个简短的功能描述文档. 同时可以看出该 API 所属的功能类别名称为“Grocery”.

3.2 LDA 概率主题模型

LDA[24]是一种非监督的机器学习算法，能够从大规模的文档集合 D 中发现隐含的主题 Z. LDA 假设每个文档 D_h 都包含若干隐含主题并服从一个概率分布 θ_d，其中每个隐含主题 z 都服从所有词汇上的一个概率分布 φ_z. 图 3 所示为 LDA 模型的图形表示. $\boldsymbol{\theta}$ 和 $\boldsymbol{\varphi}$ 分别表示文档的主题分布和主题下词汇的主题分布. α 和 β 是它们的先验参数. 对于文档中每个词 W_i，可以通过式(1)为其抽样一个主题 Z_i，同时通过式(2)选择该词.

$$\boldsymbol{\theta}\sim\text{Dirichlet}(\alpha),\ Z_i\sim Multinomial(\theta_{D_h}) \quad (1)$$

$$\varphi_z\sim\text{Dirichlet}(\alpha),\ W_i\sim Multinomial(\varphi_z) \quad (2)$$

重复迭代上述式(1)和式(2)$|D_h|$次即可生产文档 D_h，其中 $|D_h|$ 是文档 D_h 中词汇的数目. 采用吉布斯抽样(Gibbs Sampling)方法[37]，$\boldsymbol{\theta}$ 和 $\boldsymbol{\varphi}$ 分别可被推断出来，从而得到每个文档和词汇的主题分布向量.

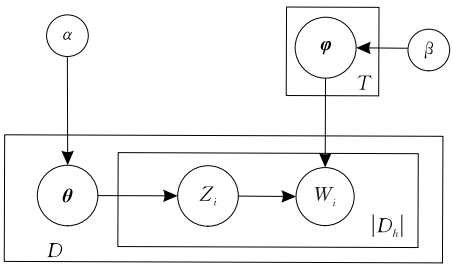


图 3　LDA 主题模型

虽然，LDA 可以有效地提取文档内容的主题信息，但它认为参与训练的文档是互相独立的. 很多情况下，文档之间并非独立存在，如 Web 服务之间存在组合关系[2,10]以及科技文献之间存在互相引用的关系[16]. 这种文档之间相互关联的关系体现在它们的隐含主题信息之间也存在特定的关联并相互影响[2,16,19-20]. 因此，针对 Web 服务聚类问题，本文提出一种考虑了多重 Web 服务关系的概率主题模型用于对 Web 服务描述文档进行建模，使得最终学习到的低维向量同时对服务的功能特征、服务组合关系和共享标注关系进行统一的编码，从而提高 Web 服务聚类的精度.

4　Web 服务聚类方法

本节首先提出一种考虑多重 Web 服务关系的概率主题模型 MR-LDA. 该模型在训练的过程中同时融入了 Web 服务组合关系以及 Web 服务共享标注关系. 将上述 Web 服务的多重关系作为模型训练过程中文本层面的辅助和补充信息，以影响 Web 服务描述文档主题分布概率的大小，最终使得具有组合关系的 Web 服务隐含主题分布相似度较小，而具有共享标注关系的 Web 服务隐含主题分布相似度较大. 然后基于提出的主题模型，本节继续提出一种融合多重 Web 服务关系和主题信息的 Web 服务聚类算法. 下面分别详细介绍. 为了阅读方便，表 1 总结了本文所使用的常用符号及其含义说明.

表 1　常用符号及含义说明

符号	含义说明
D	Web 服务文档集合
D_h	表示一个 Web 文档集合中第 h 个文档
$\|D_h\|$	文档 D_h 中词汇的数目
D_{hi}	表示 D_h 中第 i 个词汇
A_h	与服务 h 具有共享标注关系的 Web 服务集合
C_h	与 h 具有组合关系的 Web 服务集合
A	表示数据集中所有标注关系的集合
C	表示数据集中所有组合关系的集合
$\boldsymbol{W}$	Web 服务语料库中所有词汇的集合
N	Web 服务语料库中词汇的总数目
T	模型训练的隐含主题数目
w_i	文档 D_h 中当前正进行主题抽样的第 i 个词汇
Z_i	当前词汇 w_i 抽样的对应主题编号
$\boldsymbol{Z}$	$\boldsymbol{W}$ 中所有词汇的主题编号集合，$\boldsymbol{Z}=\{Z_i\}$
$\boldsymbol{\theta}$	维度为 T 的一维向量，表示文档 D_h 的主题概率分布向量
$\boldsymbol{\varphi}$	维度为 T 的一维向量，表示对应主题 Z_i 下所有词汇的概率分布
ad_p^h	A_h 中的第 p 个 Web 服务功能描述文档
cd_q^h	C_h 中的第 q 个 Web 服务功能描述文档
G	Web 服务文档的主题概率分布矩阵
$\delta_{i,j}$	表征集合 TW 中描述文档 d 对抽样的主题 Z_i 施加的影响
$\gamma_{i,j}$	表征集合 MW 中描述文档 d 对抽样的主题 Z_i 施加的影响

4.1 MR-LDA 概率主题模型

MR-LDA 模型如图 4 所示. 假设模型训练的主题个数为 T,Web 服务描述文档的所有词汇数目为 N. 则 $\boldsymbol{\theta}$ 是一个长度为 T 的向量,表示 Web 服务描述文档 D_h 是服从所有主题上的一个概率分布. $\boldsymbol{\varphi}$ 是一个长度为 N 的向量,表示某个主题服从所有词汇的一个概率分布. $\gamma_{i,j}$ 和 $\delta_{i,j}$ 分别刻画了上述两种关系在训练过程中对文档 D_h 第 i 个词汇的第 j 个主题的抽样概率上施加的影响. 模型所有参数和变量的联合概率表示为

$$\begin{cases} p(\boldsymbol{W},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\varphi}|\alpha,\beta,\lambda,\mu)= \\ \quad \prod_{h=1}^{D}\left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{j=1}^{T}\theta_{hj}^{\alpha-1+n_{hj}\cdot}\cdot I(A_h,C_h,\lambda,\mu)\right)\cdot \\ \quad \prod_{j=1}^{T}\left(\frac{\Gamma(N\beta)}{\Gamma(\beta)^N}\prod_{i=1}^{N}\varphi_{ji}^{\beta-1+n_{\cdot ji}}\right) \\ n_{ht\cdot}=\sum_{i=1}^{N}n_{hji},\ n_{\cdot ji}=\sum_{h=1}^{D}n_{hji} \\ n_{hji}=\#\{i:D_{hi}=w_i,z_i=j\} \end{cases} \tag{3}$$

其中,Γ 为 Gamma 函数,函数 I 表示与服务 h 具有组合关系的服务集(C_h)和共享标注关系的服务集(A_h)中的服务在当前第 j 个主题概率($\theta_{hj}^{\alpha-1+n_{hj}\cdot}$)施加的影响. 将所有的观测样本词汇 $\boldsymbol{W}$ 作为输入,MR-LDA 模型将会迭代训练得到后验参数 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}$ 以及 $\boldsymbol{z}$. 其详细的生成过程如算法 1 所示,其中 ψ 和 τ 分别代表 Web 服务组合关系与 Web 服务共享标注关系对当前抽样文档的词汇 w_i 施加的影响. α、β、λ 和 μ 是 MR-LDA 模型的先验参数,用于平滑模型. 将所有 Web 服务描述文档以及构建好的 Web 服务多重关系网络作为模型的输入,$\boldsymbol{\theta}$、$\boldsymbol{\varphi}$ 以及 $\boldsymbol{z}$ 等隐含参数即可以通过 Gibbs 抽样方法估计出来[37]. 其中,Web 服务多重关系网络基于本文所使用的数据集建立. 构建过程中,某个服务包含多个标签则对应网络中多条边,表示它们之间的标注关系. 另外,某个 Mashup 包含的多个 Web 服务中任一两个服务之间将对应网络中的一条边,表示它们之间具有组合关系. Web 服务多重关系网络详细统计数据如表 2 所示. 模型训练过程中会建立一条马尔科夫链,算法不断地迭代更新链的状态,直至词汇隶属的主题 $\boldsymbol{z}$ 不在变化或迭代次数结束. 迭代过程中 Web 服务描述文档中词汇的更新规则如式(4)所示:

$$p(z_i=j|A_h,C_h,\alpha,\beta,\mu,\lambda,z_{\neg i})\propto \frac{n_{j,\neg i}^{(w_i)}+\beta}{n_{j,\neg i}^{(\cdot)}+N\beta}\times\frac{v_{j,\neg i}^{(D_h)}+\alpha}{v_{\cdot,\neg i}^{(D_h)}+T\alpha}\times\exp(\tau(\mathrm{MR2})-\psi(\mathrm{MR1})) \tag{4}$$

其中

$$\psi(\mathrm{MR1})=\frac{\lambda}{|C_h|}\sum_{q\in C_h}z_j^{(cd_q^h)} \tag{5}$$

$$\tau(\mathrm{MR2})=\frac{\mu}{|A_h|}\sum_{p\in A_h}z_j^{(ad_p^h)} \tag{6}$$

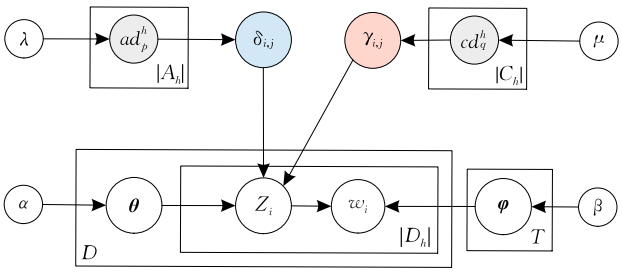


图 4 MR-LDA 主题模型

算法 1. MR-LDA 模型的生成过程.

输入:Web 服务描述文档集合 D,Web 服务组合和共享标注关系网络

输出:$\boldsymbol{\theta}$、$\boldsymbol{\varphi}$ 以及 $\boldsymbol{z}$ 等隐含参数

procedure begin:

1. for each $j=1:T$ do
2. 根据 Dirichlet 分布 Dirichlet(β)抽样得到该主题上的一个多项式分布向量 $\boldsymbol{\varphi}|\beta\sim$Dirichlet($\beta$)
3. end for
4. for each $D_h\in D$ do
5. 根据 Dirichlet 分布 Dirichlet(β)抽样得到该主题上的一个多项式分布向量 $\boldsymbol{\theta}|\alpha\sim$Dirichlet($\alpha$);
6. for each $w_i\in D_h$ do
7. 从主题分布概率向量 $\boldsymbol{\theta}$ 中为该词汇抽样一个主题 $Z_i|\boldsymbol{\theta}\sim Multinomial(\boldsymbol{\theta})$;
8. 从主题 $Z_i=j$ 中抽样得到一个词汇 $w_i|Z_i\sim Multinomial(Z_i)$
9. 基于 Web 服务多重关系网络,找到与文档 D_h 有组合关系的所有 Web 服务 C_h
10. for each $cd_q^h\in C_h$ do
    根据式(5)刻画一个 $\delta_{i,j}\sim\psi(Z_i,\lambda)$
11. end for
12. 基于 Web 服务多重关系网络,找到与文档有共享标注关系的所有 Web 服务 A_h
13. for each $ad_p^h\in A_h$ do
14. 根据式(6)刻画一个 $\gamma_{i,j}\sim\tau(Z_i,\mu)$
15. end for
16. end for
17. end for

end procedure

表 2 Web 服务多重关系网络统计信息

信息类别	数量
网络顶点总数	17412
网络边总数	65307
Web 服务顶点数	16013
标签顶点数	1399
标注关系数	52894
Web 服务组合关系数	12413

与传统的 LDA 相比，MR-LDA 模型的更新规则引入了 Web 服务的拓扑结构信息，式(4)中表示为 $\exp(\tau(\mathrm{MR2})-\psi(\mathrm{MR1}))$，其中 $\tau(\mathrm{MR2})$ 表征了与当前训练服务 D_h 具有共享标注关系的服务对 D_h 的 i 个词汇的第 j 个主题的抽样概率上施加的影响，$\psi(\mathrm{MR1})$ 表征了与当前训练服务 D_h 具有组合关系的服务对 D_h 的 i 个词汇的第 j 个主题的抽样概率上施加的影响. 从形式上可以看出基于 Web 服务结构信息，模型要尽量使得具有共享标注关系的 Web 服务主题概率比较接近，而具有组合关系的服务主题分布相差较大. 对于每一个训练的 Web 服务描述文档 $D_h\in D$，$n_{j,\neg i}^{(w_i)}$ 和 $v_{j,\neg i}^{(h)}$ 分别记录了词语 w_i 归入主题 j 的次数和文档 D_h 中归入主题 j 的词汇的数目；$n_{j,\neg i}^{(\cdot)}$ 表示归入主题 j 的词汇数量；$v_{\cdot,\neg i}^{(D_h)}$ 表示文档 D_h 中所有词汇的数量；$z_{\neg i}$ 表示除当前主题外的主题向量. 通过上述的迭代抽样过程后，根据式(7)可计算得出 Web 服务描述文档的隐含主题概率：

$$p(z_i=j)\propto\frac{v_j^{(D_h)}+\alpha}{v_{\cdot}^{(D_h)}+T\alpha} \tag{7}$$

$p(z_i=j)$ 表示文档 D_h 属于第 j 个主题的概率. 文档的主题分布向量和语料库的主题概率分布矩阵则可由下列式(8)和式(9)计算得出：

$$\boldsymbol{Z}(D_h)=\left\{\sum_{j=1}^{T}\frac{v_j^{(D_h)}+\alpha}{v_{\cdot}^{(D_h)}+T\alpha}\right\} \tag{8}$$

$$\boldsymbol{G}=\left\{\sum_{h=1}^{D}\left\{\sum_{j=1}^{T}\frac{v_j^{(D_h)}+\alpha}{v_{\cdot}^{(D_h)}+T\alpha}\right\}\right\} \tag{9}$$

基于上述模型训练可得到 Web 服务描述文档的主题分布向量. 和传统的 LDA 主题模型相比，MR-LDA 模型训练过程中，Web 服务文档之间不再互相独立，每个 Web 服务描述文档均具有双重身份，即 Web 服务描述文档内容本身以及 Web 服务之间基于组合关系与共享标注关系形成的关联网络中的节点，分别对应文档内容层次和链接层次信息的建模. 这两个层次对应的隐含主题信息相互独立，又通过某种方式互相影响[4,25,30]，也即某个 Web 服务的功能语义信息可以从内容层面通过挖掘其附带的文本描述得到，也可以隐式地基于其所在的 Web 服务网络拓扑结构借助其它互补(共享标注关系)或互逆(组合关系)服务的功能语义信息推测得到. 基于 MR-LDA 模型，具有组合关系 Web 服务隐含主题分布相似度较小，而共享标签的 Web 服务隐含主题分布较接近. 假设第 i 个 Web 服务描述文档的主题分布向量为 $\boldsymbol{Z}^i=\{k_1^i,k_2^i,\cdots,k_T^i\}$. 基于主题模型的 Web 服务聚类算法中，每个服务属于概率值最大的某个主题，最后归属于同一个主题的 Web 服务被聚为一类，代表一个 Web 服务功能领域类别. 下节将采用多种评价指标，从不同角度比较分析基于本文所提出的 MR-LDA 主题模型与传统主题模型的 Web 服务聚类效果.

虽然上述主题模型 MR-LDA 训练过程中隐式地考虑了 Web 服务之间的多重关系，但由于 Web 服务描述文档的篇幅普遍较短，根据对本文所采用的 16013 个 Web 服务进行统计知 Web 服务描述文本平均仅包含 42 个词汇. 由于缺乏充分的训练样本，大部分主题模型无法很好地对短文本进行建模[4,17]，可能会影响 Web 服务聚类的效果. 这是因为，虽然 MR-LDA 模型训练过程中已同时对 Web 服务功能文本信息和拓扑结构信息进行了编码，但依然是基于文本观测样本数据进行训练，可能会因为数据稀疏等问题影响聚类的效果. 因此，本文继续探索将 Web 服务网络拓扑结构信息显式地作用于聚类的结果，从而进一步改善上述基于文本和隐式结构信息的聚类方法的效果. 本文将改方法命名为 MR-LDA+. 下面将详细介绍该种显式融合 Web 服务网络结构信息(如 Web 服务组合关系和共享标注关系)和主题信息的聚类算法.

4.2　Web 服务聚类算法 MR-LDA+

假设 Web 服务语料库表示为 $D=\{d_1,d_2,d_3,\cdots,d_{|D|}\}$，$|D|$ 表示 Web 服务描述文档数目，主题模型训练得到的主题集合表示为 $Z=\{K_1,K_2,\cdots,K_T\}$，每个文档都是主题集合 Z 上的一个概率分布 $\boldsymbol{Z}^i=\{k_1^i,k_2^i,\cdots,k_T^i\}$，则整个语料库的 Web 服务文档主题分布向量可表示成二维矩阵的形式 $\boldsymbol{G}=\langle D,Z\rangle=\{Z^1,Z^2,Z^3,\cdots,Z^{|D|}\}$. 如图 5 所示为整个

Web服务描述文档主题分布向量矩阵					
	K_1	K_2	K_3	…	K_T
d_1	0.21	0.45	0.10	…	0.05
d_2	0.15	0.20	0.008	…	0.23
d_3	0.13	0.04	0.16	…	0.14
d_4	0.35	0.05	0.34	…	0.24
…	0.12	0.35	0.26	…	0.01
$d_{\lvert D\rvert}$	0.48	0.14	0.14	…	0.08

图 5　语料库中 Web 服务描述文档的主题分布向量矩阵

语料库 Web 服务文档主题分布向量(或概率)矩阵 $\boldsymbol{G}$ 的示例图,从图中可以看出,每个 Web 服务均归属于概率最大的主题(图中阴影填充位置),比如描述文档 d_1 属于主题为 K_2 的类别. 基于主题模型的 Web 服务聚类中,属于同一个主题的 Web 服务属于同一类,聚类后的结果可以表示为 $ESC=\{EC_1, EC_2, \cdots, EC_T\}$,其中 T 表示主题模型训练的主题个数. 基于上文对 Web 服务之间多重关系的分析,如果聚类过程中尽可能使得存在组合关系的 Web 服务被聚到不同类别,而存在共享标注关系的 Web 服务聚为同一类,则认为有利于提高 Web 服务聚类的准确率. 基于本文对 Web 服务组合关系和共享标注关系的分析,从 Web 服务网络拓扑结构的角度出发认为如下函数值越大越好:

$$\Omega=\mathrm{Max}(Sup(\mathrm{MR1})+Sup(\mathrm{MR2})) \quad (10)$$

其中

$$Sup(\mathrm{MR1})=\sum_{C_i\in ESC}\sum_{C_{j_{\neg i}}\in ESC}\frac{Comp(C_i,C_{j_{\neg i}})}{count(\mathrm{MR1})} \quad (11)$$

$$Sup(\mathrm{MR2})=\sum_{C_i\in ESC}\frac{Anno(C_i)}{count(\mathrm{MR2})} \quad (12)$$

Ω 函数将 Web 服务拓扑结构信息在聚类结果中的作用进行了量化. $Sup(\mathrm{MR1})$ 是聚类结果中 Web 服务之间组合关系的支持率,表示聚类结果中具有组合关系的 Web 服务被聚到不同类别的比例,$count(\mathrm{MR1})$ 表示数据集中 Web 服务之间组合的总次数,$Comp(C_i, C_{j_{\neg i}})$ 表示聚类 C_i 中的 Web 服务与聚类 $C_{j_{\neg i}}(j\neq i)$ 中的 Web 服务存在组合关系的总数. $Sup(\mathrm{MR2})$ 是聚类结果中 Web 服务之间共享标注关系的支持率,$count(\mathrm{MR2})$ 表示数据集中 Web 服务之间共享标签的总次数,$Anno(C_i)$ 表示聚类 C_i 中 Web 服务之间共享标签的总次数. 优化函数 Ω 值越大说明 Web 服务之间的多重关系在聚类过程中发挥的作用越大. 本文将对所有基线方法计算 Ω 值,以此来比较 Web 聚类效果与 Ω 值之间的趋势关系. 如基于本文收集的数据,通过实验表明基于传统的 LDA 主题模型进行聚类其优化函数值 $\Omega(\mathrm{LDA})=1.4273$,基于 MR-LDA 主题模型进行聚类其优化函数值 $\Omega(\mathrm{MR\text{-}LDA})=1.5953$. 为了使 Ω 达到最优化,同时满足属于同一聚类中的 Web 服务功能语义信息尽可能相似这一条件,这里提出一种利用 Web 服务之间的多重关系对 Web 服务主题分布概率进行修正的方法,具体实施过程如算法 2 所示.

算法 2. 结合先验知识的文档主题概率矩阵修正算法.

输入:Web 服务文档主题分布向量(或概率)矩阵 $\boldsymbol{G}=\langle D,Z\rangle$、Web 服务多重关系网络、临时矩阵 $\boldsymbol{G}'$ 并初始化为 $\boldsymbol{G}$

输出:修正后的主题分布向量(或概率)矩阵 $\boldsymbol{G}=\langle D,Z\rangle$

start:

repeat:

1. for $K_i\in Z$, $i=1,2,\cdots,T$ do
2. for $d_j\in D\cap cluster(d_j)=K_i, j=1,2,\cdots,|D|$ do
3. $Collection(K_i)\leftarrow d_j$
4. end for
5. $Count(K_i)\leftarrow Size(Collection(K_i))$
6. for $d_j\in Collection(K_i), j=1,2,\cdots,Count(K_i)$ do
7. $p_{j,i}\leftarrow \exp\left(\frac{TCount(d_j)\times a-MCount(d_j)\times b}{count(K_i)}\right)\times \boldsymbol{G}\langle d_j,K_i\rangle$
8. $\boldsymbol{G}'\langle d_j,K_i\rangle\leftarrow p_{j,i}$
9. end for
10. 对 $\boldsymbol{G}'$ 每一行执行归一化操作
11. $\boldsymbol{G}\leftarrow\boldsymbol{G}'$
12. end for

until

$\boldsymbol{G}=\langle D,Z\rangle$ 收敛或迭代结束

end

算法 2 基于先验知识对已训练完成的 Web 服务文档主题概率分布矩阵进行 $\boldsymbol{G}=\langle D,Z\rangle$ 迭代修正. 算法在主题概率矩阵趋于稳定或者迭代次数完成后结束. 第 2 行 $cluster(d_j)=K_i$ 表示文档 d_j 所属的主题(或聚类)类别为 K_i. 第 2 到 4 行将属于类别 K_i 的 Web 服务加入到类别集合 $Collection(K_i)$ 中. 第 5 行计算 $Collection(K_i)$ 中的 Web 服务数目 $Count(K_i)$. 第 6 到 8 行考虑了 Web 服务组合关系与共享标注关系对描述文档 d_j 属于主题 K_i 的概率 $p_{j,i}$ 进行更新并保存在临时矩阵 $\boldsymbol{G}'$ 的对应位置中,其中 $TCount(d_j)$ 表示 $Collection(K_i)$ 中与 d_j 具有共享标注关系的 Web 服务数量,$MCount(d_j)$ 表示 $Collection(K_i)$ 中与 d_j 具有组合关系的 Web 服务数量,参数 a 和 b 分别用于控制共享标注关系和组合关系的影响力度. 该行对概率进行更新的主要动机为当某个服务与其所在的类别中多个服务具有共享标注关系时,则倾向增强其属于该类别的概率. 相反,当某个服务与其所在的类别中多个服务具有组合关系时,则倾向削弱其属于该类别的概率. 第 10 行对主题概率矩阵进行归一化操作,确保每个 Web 服务的主题概率之和为 1. 该主题信息修正算法的时间复杂度为 $Iter\times T\times|D|^2$,其中 $Iter$ 表示算法

迭代的次数.完成 Web 服务主题概率的修正之后，基于这些概率值将 Web 服务进行聚类.下面，本文设计了多组实验，采用多种评价指标来检验所提出的融合多重关系与主题信息的 Web 服务聚类算法的有效性.

5　实验评估

5.1　数据集描述

截止至 2016 年 10 月，我们从 ProgrammableWeb 平台上爬取了 6208 个 Mashup 服务，12 920 个 API 服务，13 613 次 Mashup 调用 API 的关系，以及其它相关的信息如标签等.此外，由于该数据集中很多被 Mashup 调用的 API 无法在 API 数据库中找到，因此将爬取的数据集与另外一个 Web 服务数据集(该数据集为 2013 年从 ProgrammableWeb 上爬取，并被广泛应用于服务推荐[17]和标签推荐等任务[10])进行了合并，合并后的详细数据集统计信息如表 3 所示.爬取的数据中共包含 384 个功能领域类别，平均每个类别包含 33.73 个 APIs，同时每个类别中包含的 API 数量极其不均匀.本文选取包含 API 服务最多的前 20 个类别作为实验数据集，共包含 6916 个 API 服务.表 4 所示为前 20 类别以及其包含的 API 数量.

表 3　Web 服务数据集相关统计信息

参数类别	数量
API 总数	16 012
Mashup 总数	7816
Mashup 调用 API 的总次数	16 449
平均每个 Mashup 调用 API 的数目	2.1
平均每个 API 包含的标签数	3.6
平均每个 API 描述文档包含的词汇数目	44
API 功能领域类别数	384
平均每个领域类别包含的 API 数	33.73

表 4　Top-20 类别及 API 服务的分布情况

类别	API 数量	类别	API 数量
Tools	961	Telephone	368
Financial	713	Science	366
Social	604	Search	319
Enterprise	539	Video	300
Messaging	483	Internet	292
Mapping	446	Advertising	288
eCommerce	435	Travel	286
Government	401	Email	285
Reference	384	Security	271
Payments	374	Photos	270

5.2　聚类评价指标

首先，和文献[2]针对聚类问题采取的评估方法类似，本文采用四种评价标准对所提出的方法进行评估，分别是准确率(*Precision*)、召回率(*Recall*)、纯度(*Purity*)和熵(*Entropy*).假设 Top-M 个服务类别中标准的 Web 服务分类表示为 $RSC=\{RC_1, RC_2, \cdots, RC_M\}$，可以从真实数据集中获取.实验的聚类结果表示为 $ESC=\{EC_1, EC_2, \cdots, EC_v\}$.则 *Precision* 和 *Recall* 的定义如式(13)和式(14)所示：

$$Recall(EC_i)=\frac{|EC_i \cap RC_i|}{|RC_i|} \tag{13}$$

$$Precision(EC_i)=\frac{|EC_i \cap RC_i|}{|EC_i|} \tag{14}$$

其中，$|EC_i|$ 表示类别 EC_i 中的 Web 服务数目，$|RC_i|$ 表示类别 RC_i 中的 Web 服务数目.$|EC_i \cap RC_i|$ 表示正确聚类的 Web 服务数目.

第 i 个类 EC_i 的 *Purity* 和聚类结果的平均 *Purity* 的定义如式(15)和式(16)所示：

$$Purity(EC_i)=\frac{\max\limits_{j}|EC_i \cap RC_j|}{|EC_i|},\ 1\leq j\leq M \tag{15}$$

$$Purity(ESC)=\sum_{i=1}^{TK}\frac{|EC_i|}{S}\times Purity(EC_i) \tag{16}$$

其中，S 表示 RSC 中 Web 服务的总数，TK 表示选定的 Top k($1\leq k\leq V$)实验聚类数据.类似地，EC_i 的 *Entropy* 和聚类结果的平均 *Entropy* 的定义如式(17)和式(18)所示：

$$Entropy(EC_i)=-\sum_{j=1}^{M}\frac{|EC_i \cap RC_j|}{|EC_i|}\times \log_2\frac{|EC_i \cap RC_j|}{|EC_i|} \tag{17}$$

$$Entropy(ESC)=\sum_{i=1}^{TK}\frac{|EC_i|}{N}\times Entropy(EC_i) \tag{18}$$

对参与实验的每一个 Web 服务类别聚类结果，都分别计算 *Recall*、*Precision*、*Purity* 和 *Entropy*，然后对所有类别的聚类结果求上述评价指标的平均值.

其次，和文献[20]类似，基于服务的低维向量表示，我们采用一种监督的分类算法(如使用决策树、支持向量机等分类器)来评估本文提出算法的效果，并采用 Micro-$F1$ 和 Macro-$F1$ 得分作为衡量指标.其具体计算公式定义为

$$\text{Micro-}F1=\frac{\sum\limits_{i=1}^{M}2TP^i}{\sum\limits_{i=1}^{M}(2TP^i+FP^i+FN^i)} \tag{19}$$

$$\text{Macro-}F1=\frac{1}{M}\sum_{i=1}^{M}\frac{2TP^i}{(2TP^i+FP^i+FN^i)} \tag{20}$$

其中，TP^i，FP^i 和 FN^i 分别表示分类结果中对应第 i 个服务功能类别预测为正的正样本数(True Positive)，预测为正的负样本数(False Positive)和

预测为负的负样本数(False Negative). Micro-$F1$ 和 Macro-$F1$ 得分越高表明服务分类效果越好.

5.3 方法比较

本文采用 6 种方法作为基线系统进行评估和对比分析,分别介绍如下:

(1) TFIDF-K[11]. 该方法采用 K-mean 算法对服务进行聚类. 首先将 Web 服务描述文档表示成长度为 N 的一维向量. 向量中的每个元素取值为对应词汇的词频和逆文档频率(TF-IDF). 然后基于 Web 服务的向量表示使用 K-means 算法进行聚类;

(2) LDA[36]. 基于 LDA 主题模型进行聚类. 首先基于 LDA 提取 Web 服务的功能语义信息,并用低维的长度为 T 的主题向量进行表示. 然后对向量的主题概率进行排序,每个服务属于主题概率最大的类别. 相同主题的 Web 服务被聚为一个类. 与本文提出的 MR-LDA 方法相比,LDA 仅对 Web 服务的文本信息进行建模,未考虑 Web 服务拓扑结构层面的信息;

(3) WT-LDA[12]. 基于一种增强的(augmented) LDA 主题模型进行 Web 服务聚类. 与 LDA 主题模型不同的是,该模型在训练过程中同时考虑了描述文档和服务标签等信息,其认为与 Web 服务描述文档一样,服务的标签信息也可以用于决定服务的功能主题分布. 该方法聚类过程与 LDA 类似,每个服务隶属概率最大的主题,同一主题下的所有服务为一类;

(4) LDA-K[2]. 该方法采用 K-mean 算法对服务进行聚类. 首先利用 LDA 将每个服务表示成低维的向量,然后基于 K-means 对这些向量进行聚类. 与上述仅依据最大主题概率进行聚类的 LDA 方法相比,该方法更注重主题分布向量整体的相似性;

(5) MR-LDA. 本文所提出的方法. 基于 MR-LDA 主题模型从文本和网络结构两个层面提取 Web 服务的功能语义信息. 每个服务属于主题概率最大的类别. 相同主题的 Web 服务被聚为一个类;

(6) MR-LDA+. 本文所提出的方法. 首先基于 MR-LDA 主题模型提取 Web 服务的功能语义信息. 然后基于 Web 服务之间的组合关系与共享标注关系对文档主题概率矩阵进行修正. 与 MR-LDA 类似,最后每个服务隶属概率最大的主题,相同主题的 Web 服务被聚为一类. 与 MR-LDA 方法相比,MR-LDA+显式地将 Web 服务地拓扑结构信息作用于 Web 服务聚类的过程.

模型的超参数通常很难确定,针对不同的问题可能设置不同. 本文采取如下简单的方式来设置聚类模型的先验参数. 由于实验只选取包含 Web 服务最多的前 20 类别用于评估所提出的算法,实验设置主题模型的主题个数为 20($T=20$),使得每一个主题近似地对应一个服务领域类别(如每个服务属于主题概率最大的类别). 对所有的主题模型,先验参数 α 和 β 分别设置为 2.0 和 0.1[3,10]. MR-LDA 模型的先验参数 μ 和 λ 分别设置为 3.0. 算法 2 中主题概率矩阵修正的迭代次数设置为 50($Iter=50$). 此外,本文对算法中的平滑参数 a 和 b 进行了多次实验设置,以选取对聚类结果最有利的取值.

此外,针对基于监督的服务分类方法,我们采用线性的(Linear)支持向量机(SVM)作为分类器,基于 sklearn.svm.SVC 工具实现. 实验中,我们将服务的低维向量作为输入特征(Feature)和功能类别作为对应特征的标签(Label)训练一个服务分类器,然后基于该分类器预测验证样本集中服务的功能类别并与真实结果比较进行评估. 其中 $p\%$ 随机选择的服务作为训练集(Training set),剩余数据作为测试集(Test set),p 将取不同值以评估不同稀疏数据环境下各基线方法的性能.

5.4 实验结果

基于 ProgrammableWeb 数据集,本节首先介绍不同聚类算法(5.3 节中介绍)的效果. 针对每种聚类算法,分别采用 *Recall*、*Precision*、*Purity* 和 *Entropy* 四种指标对结果进行评估,其中 *Recall*、*Precision*、*Purity* 越大以及 *Entropy* 越小则表明聚类效果越好. 然后分析了 Web 服务之间的多重关系及相关参数对聚类结果的影响. 接着本节继续采用 SVM 基于 Web 服务向量训练一个分类器,然后对 Web 服务进行功能分类和评估,以说明主题模型训练过程中对 Web 服务不同信息(如文本信息和 Web 服务拓扑结构信息等)进行编码所产生的效果.

5.4.1 Web 服务聚类结果分析

表 5 所示为不同方法的聚类评估结果. 从表中可以得出如下结论:

表 5 Web 服务 Top 20 类别聚类实验结果

Methods	*Recall*	*Precision*	*Purity*	*Entropy*	Ω 值
TFIDF-K	0.0735	0.2692	0.1446	2.8582	0.8424
LDA	0.5206	0.4806	0.5351	1.5755	1.4273
WT-LDA	0.5420	0.4978	0.5368	1.5944	1.4312
LDA-K	0.4695	0.4700	0.5289	1.6236	1.4159
MR-LDA	0.5982	0.5357	0.5864	1.4491	1.5953
MR-LDA+	0.6678	0.6220	0.6357	1.2525	1.7686

(1)四种评估标准的结果，基于主题模型的聚类方法(MR-LDA＋、MR-LDA、LDA-K、WT-LDA 和 LDA)效果均明显好于基于关键词匹配的方法(TF-IDF)，这是由于 TF-IDF 忽略了 Web 服务之间功能语义的关联，而通常 Web 服务描述文档较不规范，不同的 Web 服务开发者采用不同的方式来描述 Web 服务所提供的功能，这时挖掘 Web 服务的隐含主题信息十分必要[4-5,15]；

(2)我们提出的融合 Web 服务多重关系的主题模型(MR-LDA)比其它的主题模型(LDA 和 WT-LDA)表现要好，这是由于 Web 服务的描述文档篇幅较短，传统的主题模型不能较好地对缺乏语料的短文本进行建模，导致提取的主题信息不够准确. 此外，传统基于 LDA 的主题模型得到的向量主要对 Web 服务的文本信息进行编码，忽略了 Web 服务之间的拓扑结构信息. 前人研究表明[16,19-20]，网络的拓扑结构信息对学习顶点的低维向量表达形式十分重要，并与基于文本的学习方式互为影响和补充. 如 Web 服务的功能信息可以从功能文本中挖掘得到，也可以借助结构信息从其它相同或相似的服务中推断出来. 尤其在文本数据比较稀疏时，结构信息能作为很强的补充知识改善算法的性能. 在 MR-LDA 模型中，每个 Web 服务在模型训练过程中具有双重作用，即 Web 服务内容本身以及与其它服务之间的链接关系. 链接层面上的主题信息对 Web 服务内容的主题概率的分布起到了调节作用，使得学习到的主题向量同时编码了 Web 服务文本信息和拓扑结构信息，全面地对一个服务的功能特征进行了表达. 此外，实验结果表明我们提出的融合多重关系与主题信息的 Web 服务聚类算法(MR-LDA＋)能够显著地提升聚类的准确度，进一步验证了结构信息对 Web 服务聚类地促进作用，表明提出的显式地利用 Web 服务组合关系及共享标注关系对 Web 服务主题概率矩阵进行修正算法的有效性，其平均推荐性能(综合四种指标)相比 MR-LDA、LDA-K、WT-LDA、DLA 和 TFIDF-K 分别提升了 12%、29%、22%、24%和 333%；

(3)比较表 5 中所有算法的聚类效果可以得出一个普遍的现象，即 Web 服务聚类效果越好，Ω 值越大. 本文聚类算法优化函数 Ω 由式(10)定义，用于解释 Web 服务之间的组合关系以及共享标注关系在 Web 服务聚类中所起的作用. 这一正向变化的趋势表明上述 Web 服务之间的关系对改善 Web 服务聚类的效果有积极的作用.

本文设计了多组实验对算法 2 中涉及的平滑参数 a 和 b 进行调整参数，使聚类效果达到最好. 其中参数 a 刻画了 Web 服务组合关系对聚类效果的影响. 从图 6 中可以看出，基于四种评估标准，a 值的

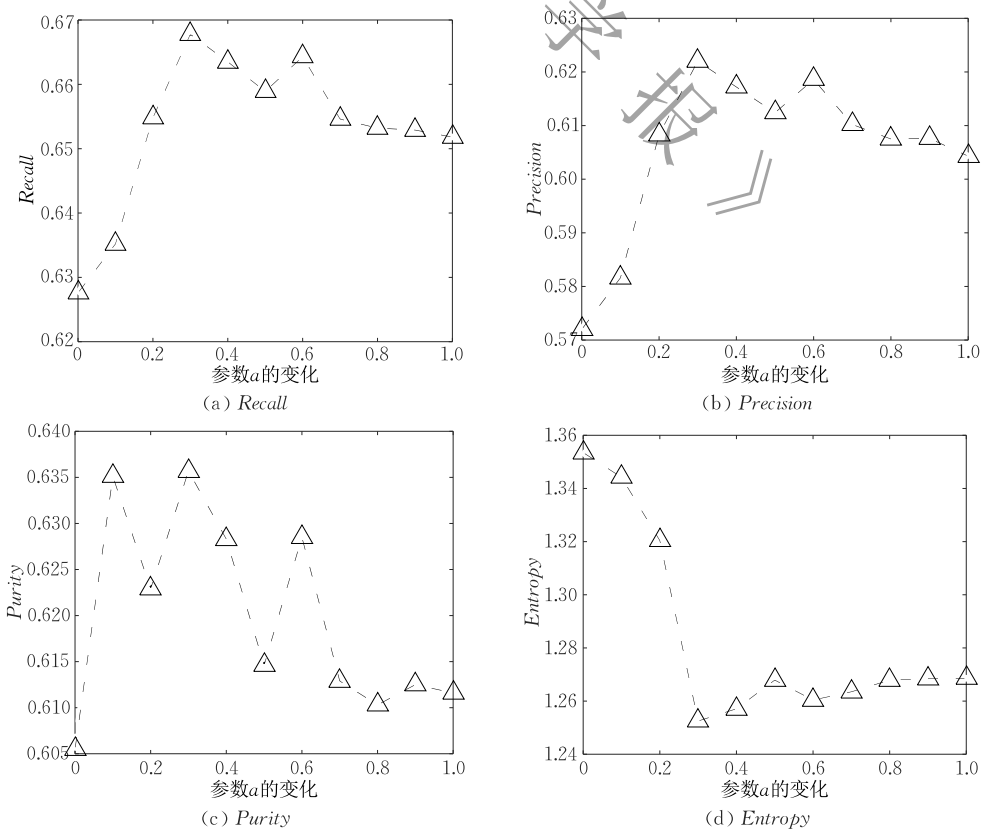


图 6　平滑参数 a 的取值对聚类结果的影响($b=0.5$)

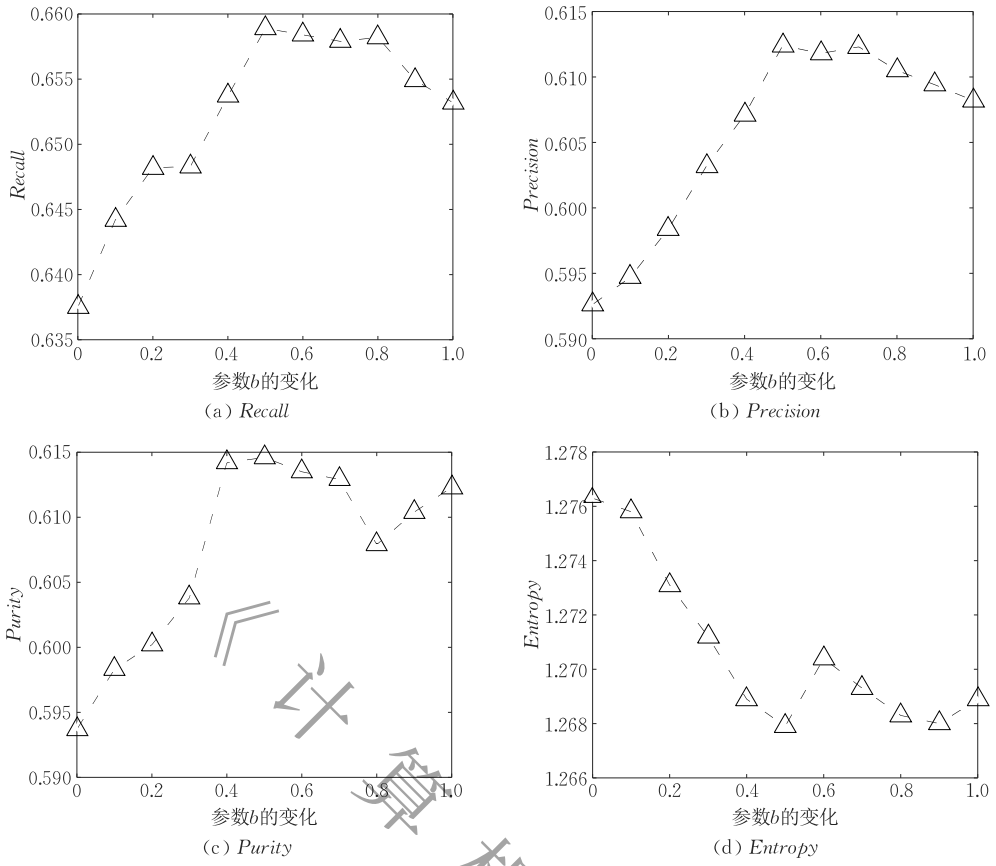


图 7 平滑参数 b 的取值对聚类结果的影响($a=0.5$)

变化将引起聚类效果的变化,当取值 $a=0.3$ 时,算法的综合性能达到最好.类似地,参数 b 用于控制 Web 服务共享标注关系对聚类效果的影响程度.图 7 所示为对 b 设置不同的参数值的聚类效果,当设置 $b=0.5$ 时,四种评估标准的综合指标达到最佳.图 6 与图 7 的变化趋势表明了算法中参数 a 和 b 设置的合理性.

5.4.2 Web 服务分类结果分析

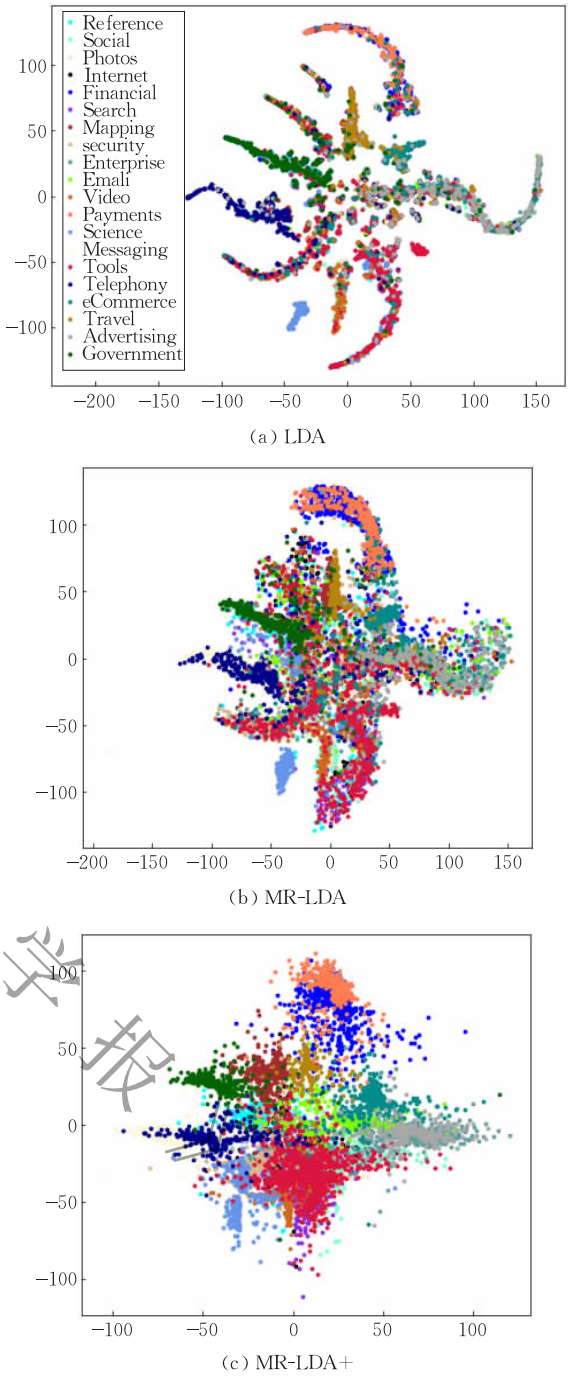
主题向量表征了 Web 服务的功能特征,其质量的好坏(是否能全面地表达一个服务的特征)将直接影响服务聚类(如基于 K-means 等聚类方法)和分类的效果[20].5.3 节列举的基线方法采用了不同的方式将服务表示成低维的向量特征,如 TFIDF-K 和 LDA 等得到的向量仅对服务的功能文本信息进行了编码,而 MR-LDA 等同时编码了文本信息和网络结构信息.基于这些向量进行监督地训练(如 $p\%$的数据用于训练,其余用于测试)和分类的结果如表 6 所示.可以看出基于本文采用的两种评价指标,随着 p 值的增大,所有基线方法的分类性能大体上逐渐提升.如基于 Micro-$F1$ 指标,p 从 10 变化至 90,TFIDF-K 的效果也随之从 0.1363 增长至 0.1521.众所周知,对于监督学习算法(如 SVM 等),训练样本越丰富,得到的模型泛化能力越强,因此对未知数据的预测效果越好.同时还可以观察到,基于

表 6 基于 Web 服务向量表示采用 SVM 进行 Web 服务分类的结果

	$p\%$	10%	20%	30%	40%	50%	60%	70%	80%	90%
Micro-$F1$	TFIDF-K	0.1363	0.1381	0.1402	0.1431	0.1446	0.1448	0.1505	0.1518	0.1521
	LDA	0.3754	0.3772	0.3823	0.3877	0.3886	0.3811	0.3888	0.3895	0.3955
	WT-LDA	0.3758	0.3769	0.3831	0.3878	0.3899	0.3921	0.3927	0.3939	0.3987
	MR-LDA	0.3731	0.3849	0.3881	0.3919	0.3950	0.4058	0.4091	0.4103	0.4132
	MR-LDA+	0.3968	0.4079	0.4105	0.4174	0.4193	0.4231	0.4278	0.4337	0.4383
Macro-$F1$	TFIDF-K	0.1162	0.1174	0.1195	0.1204	0.1237	0.1245	0.1238	0.1247	0.1259
	LDA	0.2765	0.2867	0.2869	0.2927	0.2984	0.2932	0.2990	0.3091	0.3199
	WT-LDA	0.2767	0.2793	0.2871	0.2936	0.2987	0.3012	0.3127	0.3133	0.3148
	MR-LDA	0.2764	0.2893	0.292	0.2952	0.2991	0.3027	0.3183	0.3191	0.3196
	MR-LDA+	0.2769	0.2878	0.3012	0.3067	0.3128	0.3149	0.3210	0.3263	0.3291

本文提出的 MR-LDA 模型所得到的服务功能特征向量，SVM 分类器性的综合性能比其它方法表现好（p 设置为 20 到 90 时效果均最优）. 这是因为与其它方法比较，MR-LDA 同时对服务的结构信息进行了编码，并且有利于学习得到更优的特征向量. 最后，MR-LDA＋比 MR-LDA 性能更优，说明了 Web 服务的网络拓扑结构信息对分类效果具有促进作用，并在文本数据缺乏时可作为补充信息提升算法的性能.

为直观地比较仅对文本信息进行编码的特征向量（如 LDA 方法）和同时编码文本信息和结构信息的特征向量（如 MR-LDA 与 MR-LDA＋）的效果，针对表 4 中列出的 Top-20 功能类别中的所有 Web 服务，本文借助 t-SNE 工具[38] 基于 PCA 降维技术对这三种方法的特征向量进行了可视化，其效果如图 8 所示. 可以看出，基于 LDA 方法很多不同类别的服务空间位置重合度较高，表明它们具有很相似的特征向量，导致了很多错误的服务分类. 而基于 MR-LDA 模型，具有共享标注关系的服务特征向量较相近，同时具有组合关系的服务特征向量相差较大. 从图 8(b)可看出 MR-LDA 比 LDA 更容易不同类别的服务区分开. 例如，“Reference”和“Government”类别在图 8(a)中的空间位置基本重合，算法很难将他们正确区分和分类. 而图 8(b)中这两个功能类别已经具有较好的区分度. 此外，可以清楚地观察到显式地将 Web 服务地拓扑结构信息作用于服务分类将显著地改善效果（如图 8(c)所示）.



(a) LDA

(b) MR-LDA

(c) MR-LDA+

图 8　Web 服务低维功能特征向量的可视化结果

6　总　结

针对 Web 服务描述文档篇幅较短，现有大部分主题模型不能很好地对短文进行建模从而影响了 Web 服务聚类的效果这一问题，本文提出了一个考虑多重 Web 服务关系的主题模型，建模过程中，每个 Web 服务具有双重身份，即 Web 服务内容本身以及与其它 Web 服务形成的链接关系. 充分考虑链接层面的隐含主题信息能够提高 Web 服务描述文档主题信息提取的精度. 实验结果表明与传统的主题模型相比，考虑 Web 服务网络结构信息作为文本建模的辅助信息能提高算法的性能. 同时，从利用结构信息进行聚类的角度，本文继续提出了一种融合 Web 服务多重关系与主题信息的 Web 服务聚类算法，聚类结果说明利用 Web 服务组合关系及共享标注关系对 Web 服务主题信息进行修正能进一步改善聚类的效果. 本文也设计了丰富的实验，如基于服务功能特征向量进行 Web 服务多分类任务和向量距离的可视化，结果表明，Web 服务网络拓扑结构信息对提取高质量的功能特征具有促进作用.

致　谢　在此，对本文工作给予帮助与建议的老师和同学以及提出宝贵评审意见的审稿专家表示衷心的感谢！

参 考 文 献

[1] Xia B, Fan Y, Tan W, et al. Category-aware API clustering and distributed recommendation for automatic Mashup creation. IEEE Transactions on Services Computing, 2015, 8(5): 674-687

[2] Cao B, et al. Mashup service clustering based on an integration of service content and network via exploiting a two-level topic model//Proceedings of the IEEE International Conference on Web Services(ICWS). San Francisco, USA, 2016: 212-219

[3] Shi M, et al. WE-LDA: A word embeddings augmented LDA model for Web services clustering//Proceedings of the IEEE International Conference on Web Services (ICWS). Honolulu, USA, 2017: 9-16

[4] Elgazzar K, Hassan A, Martin P. Clustering WSDL documents to bootstrap the discovery of Web services//Proceedings of the IEEE International Conference on Web Services (ICWS). Miami, USA, 2010: 147-154

[5] Wen T, Sheng G, Li Y, Guo Q. Research on Web service discovery with semantics and clustering//Proceedings of the IEEE Joint International Information Technology and Artificial Intelligence Conference (IJCAI). Menlo, California, 2011: 62-67

[6] Xia Y, Chen P, Bao L, et al. A QoS-aware Web service selection algorithm based on clustering//Proceedings of the IEEE International Conference on Web Services (ICWS). Washington, USA, 2011: 428-435

[7] Zhou Z, Sellami M, Gaaloul W, et al. Data providing services clustering and management for facilitating service discovery and replacement. IEEE Transactions on Automation Science and Engineering, 2013, 10(4): 1131-1146

[8] Zhang M, Liu X, Zhang R, Sun H. A Web service recommendation approach based on QoS prediction using fuzzy clustering//Proceedings of the IEEE International Conference on Services Computing (ICSOC). Shanghai, China, 2012: 138-145

[9] Zhu J, Kang Y, Zheng Z, Lyu M R. A clustering-based QoS prediction approach for Web service recommendation//Proceedings of the IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops (ISORCW). Shenzhen, China, 2012: 93-98

[10] Shi M, Liu J X, Zhou D, et al. A probabilistic topic model for Mashup tag recommendation//Proceedings of the IEEE International Conference on Web Services(ICWS). San Francisco, USA, 2016: 444-451

[11] Chen L, Hu L, Wu J, et al. WTCluster: Utilizing tags for Web service clustering//Proceedings of the International Conference on Service Oriented Computing(ICSOC). Paphos, Cyprus, 2011: 204-218

[12] Chen L, Wang Y, Yu Q, et al. WT-LDA: User tagging augmented LDA for Web service clustering//Proceedings of the International Conference on Service Oriented Computing (ICSOC). Berlin, Germany, 2013: 162-176

[13] Yu Q, Wang H, Chen L. Learning sparse functional factors for large scale service clustering//Proceedings of the IEEE International Conference on Web Services (ICWS). New York, USA, 2015: 201-208

[14] Hu W, Tsujii J. A latent concept topic model for robust topic inference using word embeddings//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany, 2016: 380-386

[15] Shi M, Liu J X, Zhou D, Tang Y. A topic-sensitive method for Mashup tag recommendation utilizing multi-relational service data. IEEE Transactions on Services Computing, 2018, 13(1): 1-14

[16] Guo Z, Zhang Z M, Zhu S, et al. A two-level topic model towards knowledge discovery from citation networks. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(4): 780-794

[17] Cao B, Liu B X, Rahman M M, et al. Integrated content and network-based service clustering and Web APIs recommendation for Mashup development. IEEE Transactions on Services Computing, 2017, 3(22): 1-14

[18] Cui P, Wang X, Pei J, Zhu W. A survey on network embedding. 2017, arXiv preprint arXiv:1711.08752

[19] Le T M, Lauw H W. Probabilistic latent document network embedding//Proceedings of the IEEE International Conference on Data Mining (ICDM). Shenzhen, China, 2014: 270-279

[20] Sun X, Guo J, Ding X, Liu T. A general framework for content-enhanced network representation learning. 2016, arXiv preprint arXiv:1610.02906

[21] Cao B, Liu X F, Liu J, Tang M. Domain-aware Mashup service clustering based on LDA topic model from multiple data sources. Information and Software Technology, 2017, 90(2): 40-54

[22] Hasan M H, Jaafar J, Hassan M F. Fuzzy-based clustering of Web services' quality of service: A review. Journal of Communications, 2014, 9(1): 81-90

[23] Tian Gang, He Ke-Qing, Wang Jian, et al. Domain-oriented and tag-aided Web service clustering method. Acta Electronica Sinica, 2015, 43(7): 1266-1274(in Chinese)
(田刚，何克清，王健等. 面向领域标签辅助的服务聚类方法. 电子学报，2015，43(7)：1266-1274)

[24] Zhao Guo-Dong, Zhou Ying, Song Li-Ya. Semi-supervised ISHC hierarchy description based Mashup service clustering. Journal of Jilin University(Natured Science), 2015, 53(4): 698-704(in Chinese)
(赵国栋，周莹，宋丽亚. 基于半监督 ISHC 层次描述的 Mashup 服务聚类. 吉林大学学报：理学版，2015，53(4)：698-704)

[25] Hao Y, Fan Y, Tan W, Zhang J. Service recommendation based on targeted reconstruction of service descriptions//Proceedings of the IEEE International Conference on Web Services (ICWS). Honolulu, USA, 2017: 285-292

[26] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. 2013, arXiv preprint arXiv:1301.3781

[27] Ji X, Xu W. Document clustering with prior knowledge//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). New York, USA, 2006: 405-412

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

[29] Liu X, Agarwal S, Ding C, Yu Q. An LDA-SVM active learning framework for Web service classification//Proceedings of the IEEE International Conference on Web Services (ICWS). San Francisco, USA, 2016: 49-56

[30] Tian G, Wang J, He K. Leveraging auxiliary knowledge for Web service clustering. Chinese Journal of Electronics, 2016, 25(5): 858-865

[31] Sukumar A S, Loganathan J, Geetha T. Clustering Web services based on multi-criteria service dominance relationship using Peano Space filling curve//Proceedings of the International Conference on Data Science and Engineering (ICDSE). Washington, USA, 2012: 13-18

[32] Kumara B T, Paik I, Siriweera T H A S, Koswatte K R. QoS aware service clustering to bootstrap the Web service selection//Proceedings of the IEEE International Conference on Services Computing (SCC). Honolulu, USA, 2017: 233-240

[33] Liu J X, He K Q, Wang J, Ning D. A clustering method for Web service discovery//Proceedings of the International Conference on Services Computing (SCC). Washington, USA, 2011: 729-730

[34] Zhou J, Li S. Semantic Web service discovery approach using service clustering//Proceedings of the International Conference on Information Engineering and Computer Science (ICIECS). Wuhan, China, 2009: 1-5

[35] Chen F, Yuan S, Mu B. User-QoS-based Web service clustering for QoS prediction//Proceedings of the IEEE International Conference on Web Services (ICWS). New York, USA, 2015: 583-590

[36] Blei D, Ng A, Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022

[37] Heinrich G. Parameter estimation for text analysis. Technical Report, University of Leipzig, Leipzig, Germany, Version 2.4, 2008

[38] Maaten L V, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(9): 2579-605

SHI Min, born in 1991, Ph. D. candidate, assistant researcher. His research interests include social network, service computing.

LIU Jian-Xun, born in 1970, Ph. D., professor, Ph. D. supervisor. His research interests include service computing, cloud computing, theory and application of workflow management, etc.

ZHOU Dong, born in 1979, Ph. D., professor. His research interests include information retrieval, natural language processing and machine learning, etc.

CAO Bu-Qing, born in 1979, Ph. D., associate professor. His research interests include service computing and cloud computing, etc.

WEN Yi-Ping, born in 1981, Ph. D., associate professor. His research interests include workflow management and service computing, etc.

Background

This paper studies the problem of Web service clustering. The rapid growth in both the number and diversity of Web services raises new requirement of clustering techniques. As an important tool for exploratory data analysis, Web services clustering can help us capture and understand the hierarchy of the huge number of Web services within a services repository. It then becomes an enabling technique for a wide range of Web services management tasks such as Web services discovery or selection, Web services match-making and retrieval, services replication, Web services composition/Mashup, and Web services recommendation of large volume of Web services.

In the past, majority of clustering approaches have been proposed. Most of these methods focus on the functional-based clustering techniques by mining the Web service descriptions only. While the results are somehow unsatisfactory

due to the quite limit number of words contained in descriptions, which poses a challenge on traditional machine learning methods like LDA and its extensions. More efficient clustering methods are required in today's big service data environment.

However, apart from the textual information that is already being used, we can perform the service clustering with the aid of Web service network structure, since services are not independent but frequently linked with other services by composition relationships and annotation sharing relationships. Sometimes, topological structure information provides equally important knowledge as the functional textual information in service clustering task. For example, services sharing identical tags are most probably coming from the same or similar functional groups, and composited services are highly likely to belong to different clusters. Based on these assumptions, we propose to perform service clustering by leveraging both service textual and structure information and propose a novel probabilistic topic model to incorporate these two types of knowledge.

The group's research interests are currently focused on the theories and methods of service computing, service management, the discovery and composition of services, etc.

The research was supported by the National Natural Science Foundation of China under Grant Nos. 61572187, 61872139, 61876062, the project Sponsored by the Hunan Provincial Natural Science Foundation under Grant No. 2018JJ2139, and the project Sponsored by Innovation Platform Open Foundation of Hunan Provincial Education Department under Grant No. 17K033.