计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900026

KSN:一种基于知识图谱和相似度网络的 Web 服务发现模型

于 扬[1] 邢 镔[2] 曾 骏[1] 文俊浩[1]
1 重庆大学大数据与软件学院 重庆 400000
2 重庆工业大数据创新中心有限公司应用技术国家工程实验室 重庆 400000
(yuyang1996@cqu.edu.cn)

摘 要 服务发现旨在解决服务信息爆炸的问题,找到定位满足服务请求者需求的服务。由于服务描述信息主要由带有噪声的短文本组成,并且具有语义稀疏的特征,因此很难提取服务描述文档的隐含上下文信息,此外,传统的服务发现方法在获取服务的特征表示后,直接进行相似度计算,其使用的度量函数是不符合人类感知的。针对上述两个问题,文中提出了一种基于知识图谱和神经相似网络的服务发现框架(KSN)。它使用知识图谱来连接服务描述和规格中的实体以获得丰富的外部信息,从而增强服务描述的语义信息,使用卷积神经网络(Convolutional Neural Network,CNN)提取服务的特征向量,并将其作为神经相似网络的输入,神经相似网络会学习一个相似度函数,用于计算服务和请求之间的相似度以支持服务发现过程。通过对 ProgrammableWeb 爬取的真实服务数据集的大量实验结果表明,就多种评估指标而言,KSN 优于现有的 Web 服务发现方法。

关键词:Web 服务发现;服务嵌入;卷积神经网络;知识图谱;神经相似网络
中图法分类号 TP319

KSN:A Web Service Discovery Method Based on Knowledge Graph and Similarity Network

YU Yang[1],XING Bin[2],ZENG Jun[1] and WEN Jun-hao[1]
1 School of Big Data & Software Engineering,Chongqing University,Chongqing 400000,China
2 Chongqing Innovation Center of Industrial Big-Data Co. Ltd,National Engineering Laboratory for Industrial Big-Data Application Technology,Chongqing 400000,China

Abstract Service discovery aims to solve the problem of service information explosion,find and locate services that meet the needs of service requesters. Since the service description information is mainly composed of short text with noise and has the feature of sparse semantics,it is difficult to extract the implicit context information of the service description document. In addition,the traditional service discovery method directly obtains the characteristic representation of the service. According to the cosine similarity to calculate the similarity,the used measurement function is not in line with human perception. In response to the above two problems,this paper proposes a service discovery framework (KSN) based on knowledge graphs and neural similar networks. It uses the knowledge graph to connect the entities in the service description and specifications to obtain rich external information,thereby enhancing the semantic information of the service description. And it uses convolutional neural network (CNN) to extract the feature vector of the service as the input of the neural similarity network. The neural similarity network will learn a similarity function to calculate the similarity between the service and the request to support the service discovery process. A large number of experiments on real service data sets crawled by ProgrammableWeb show that KSN is superior to existing Web service discovery methods in terms of multiple evaluation metrics.

Keywords Web service discovery,Service embedding,Convolutional neural network,Knowledge graph,Neural similarity network

1 引言

随着面向服务计算的日益流行,大量以 Web API 的形式存在的 Web 服务正在被越来越多的开发者关注[1],开发者们可以通过简单的 API 来访问海量数据,如类似 Google Maps API 的地理数据、类似 Amazon E-Commerce 的产品数据,为开发者们提供了极大的便利,截至 2020 年 6 月 1 日,全球最大的在线 Web 服务数据库 ProgrammableWeb 已经注册了超过 23000 个 API,极大地丰富了用户和互联网的交互信息,从而广泛地促进了面向服务的应用程序的开发,但是繁多的服务数量可能会让用户难以承受,给用户带来不好的体验。为了减小信息爆炸的负面影响,为用户发现满足自己需求的服务至关重要。

Web 服务发现是根据服务请求者的需求来查找和定位现有 Web 服务的过程。服务提供商在向存储库注册服务时提供了许多相似性的功能描述,如基于自然语言描述的服务

到稿日期:2020-09-03 返修日期:2020-11-08 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划课题(2019YFB1706104)
This work was supported by the National Key Research and Development Program of China(2019YFB1706104).
通信作者:邢镔(xing.bin@hotmail.com)

信息、服务类别、服务提供者名称等。现有的服务发现方法主要依赖于关键字匹配的信息检索技术。但是，由于用户检索的信息中存在语法稀疏问题（关键字稀疏性），搜索引擎可能会返回大量不相关的服务。为了解决关键字稀疏问题，一些研究工作通过扩展搜索查询来实现更好的发现性能，如文献[2-3]。尽管需求扩展提供了查询的更多语义信息，但是这类方法过于依赖被扩展请求的语料质量，并且需求扩展方法并未充分利用服务描述文档中的功能性和非功能性信息。已有大量研究表明，挖掘服务描述信息可以改进服务发现性能[4-9]。基于聚类的服务发现方法旨在将具有类似功能的服务聚类在一起，减少发现空间以改善发现结果[10-14]。Web 服务的聚类方法主要依靠概率主题模型（如潜在狄利克雷分配（LDA））和启发式信息[15-17]来挖掘服务功能特征。然而，当前基于概率主题模型（LDA）的方法主要基于词包统计模型，该模型主要从服务描述中提取显式特征，而忽略了服务描述中隐含的上下文信息和服务之间的深层逻辑关系。换句话说，由于服务描述是由短文本组成的，这将导致相当大的语义稀疏性问题，并对服务的特征表示产生不利影响，不利于进行进一步的基于相似度排序的服务发现方法。此外，对于人类来说，判断两个服务之间是否相似是一项自然的任务。然而，对于算法来说，它并不能够总是成功地匹配包含相似元素的服务。服务由特征向量进行描述，虽然可以使用余弦相似度直接度量相似性，但可能会忽略数据集中可能存在的数据依赖关系，常规的度量方法可能无法捕捉到这种关系。

考虑到上述问题，如何通过外部信息丰富服务描述信息和建立一个更好的相似度度量方法已成为服务发现的关键。服务存储库为每个服务提供各种功能性信息，如“提供者”“类别”和“描述”等，其中包含许多外部信息。作为揭示实体之间关系的语义网络，知识图谱可以提取服务之间的深层逻辑联系并丰富服务的特征表示。知识图谱是一种具有对应节点和对应边的有向异构图。近年来，一些知识图谱（如 DBpedia，AceKG，Freebase，Google 知识图和 Microsoft Satori）已成功用于诸如推荐系统和信息检索之类的应用场景中。此外，本文基于从大量数据中学习相似度估计可以提高服务检索任务的性能这个事实，设计了一个可以学习服务相似度函数的神经网络，同时考虑到知识图谱的强大语义表示，本文提出了一种基于知识图谱和相似性网络的服务发现框架（KSN）。KSN 是一种端到端的服务发现模型，它使用候选服务和用户查询服务作为输入，KSN 首先将服务描述中的实体与知识图谱中的服务提供者实体连接起来，并搜索其一跳内邻居实体以丰富服务信息，再使用 LDA 模型学习服务描述中的主题以获取主题向量。受图像检索领域的启发，本文使用 CNN 提取了 k 维服务表示，其中本文将词嵌入、主题嵌入和实体嵌入 3 个矩阵模拟图像的三通道组合作为 CNN 的输入[18]。输出的 K 维向量使用相似度神经网络为服务发现过程计算服务之间的相似度。总之，本文工作的主要贡献如下：

（1）针对服务描述文档语义稀疏问题，本文提出了一种基于知识图谱的发现模型，该模型通过链接知识图谱来获取大量外部信息以丰富服务描述文档，以服务描述文档中的词嵌入矩阵、实体嵌入矩阵和主题嵌入矩阵为 CNN 输入，获得了高质量的服务特征向量。

（2）针对相似度感知问题，与传统的相似性度量（余弦相似度和直接对提取向量的欧几里得距离度量）不同，本文提出了一种基于神经网络的监督相似性网络。它可以通过大量数据学习到一个相似度函数，相比传统的相似度度量方法，检索性能有所提升。

（3）在真实世界数据集上的大量实验证明，本文提出的 KSN 模型优于最新的服务发现方法。

本文第 2 节介绍了 Web 服务发现的相关工作；第 3 节介绍了本文方法的总体框架；第 4 节展示了实验细节；最后总结全文。

2　相关工作

近年来，云计算和面向服务的计算领域发展迅速，服务发现作为其核心任务已引起研究人员的广泛关注。

服务发现的最初方法是使用传统的信息检索技术（IR）。2016 年的 WSQBE＋方法[3]使用 TF-IDF 和 K-means 算法将相似的服务聚类在一起，从而在服务发现阶段过滤掉不相关的服务。上述方法虽然在一定程度上改进了服务发现性能，但是 TF-IDF 并没有挖掘服务描述文档的隐式信息。此外，经典概率统计模型已被证明适用于服务发现。如 2016 年的 ICN-MSC 方法[19]使用 LDA 模型学习服务描述文件的主题分布，同时将标签信息合并到具有相同最大主题得分的服务集群中。2018 年，LDA-PK 方法[20]使用先验知识以半监督的方式增强聚类过程。此方法首先使用概率主题模型（LDA），从 Web 服务描述文档中提取潜在的主题向量，然后将 K-means ＋＋算法与先验知识结合起来进行聚类。但是，LDA 模型在服务描述的短文本的主题提取方面有一些限制，如无法获取高质量的服务表示向量。2017 年，WE-LDA[21]方法使用 Word2Vec 获得了比使用 LDA 模型更高质量的词向量，并将其集成到 LDA 中进行聚类，从而提高了服务发现的性能。考虑到概率主题模型无法提取服务描述的隐藏上下文信息，2019 年提出的 DeepWSC 方法介绍了一种深度神经网络框架 DeepWSC，该框架结合了递归神经网络和卷积神经网络的优点，在服务描述中自动提取上下文信息，以获得更高质量的特征向量。

语义稀疏性是服务描述文档中的常见问题。现阶段已经提出了大量的语义增强方法。2017 年 Chen 等[22]提出了一种增强语义的方法，将 WordNet 关系中的多重概念集成到 Web 服务发现中，将服务描述文档中的单词对应到 WordNet 中，在 WordNet 中描述了不同概念相对应的单词（如 Is-a，Has-a 和 Ant）关系。同时与用户交互信息（I/O）结合使用，将各个概念之间的相似性加权以得到服务之间的相似性。2012 年，Upadhyaya 等[23]使用 K-means 对服务进行功能聚类，结合 WordNet 从服务描述中提取概念和语义关系，同时扩展用户请求以获得更丰富的语义。

本文方法利用知识图谱来丰富服务描述的语义信息，同时使用 Word2Vec 增强 LDA 模型来挖掘服务描述文本的主题信息。将词嵌入矩阵、主题嵌入传递矩阵和实体嵌入传递矩阵作为 CNN 输入，提取丰富的上下文信息，有效解决了语义稀疏性问题，并具有良好的服务发现性能。

3　基于知识图谱和相似度网络的服务发现方法

本节将详细介绍本文提出的 KSN 模型。

3.1 服务发现的问题定义

形式上，Web 服务表示为 $S=\{s_1,s_2,s_3,\cdots,s_n\}$，Web 服务 s_i 的描述文档表示为 $D_{s_i}=\{W_1,W_2,W_3,\cdots,W_T\}$，其中 T 是文档中包含的单词数目，W_t 代表文档中的第 t 个单词。服务发现的目的是找到满足用户需求的前 n 个服务的列表。

3.2 KSN 的整体架构

本文提出的 KSN 方法的结构如图 1 所示，其整体可以分为 3 个部分：嵌入模块、特征提取模块和相似度计算模块。对于每个需求服务，使用 LDA、Word2Vec 和知识图谱嵌入技术获得服务的主题分布向量、词向量和实体向量，进行矩阵对齐后将其作为 CNN 的输入以提取深层服务描述信息，然后将 CNN 处理的向量作为神经相似网络的输入来计算与服务存储库中服务生成的特征向量的相似性，为后续的服务发现过程做准备。

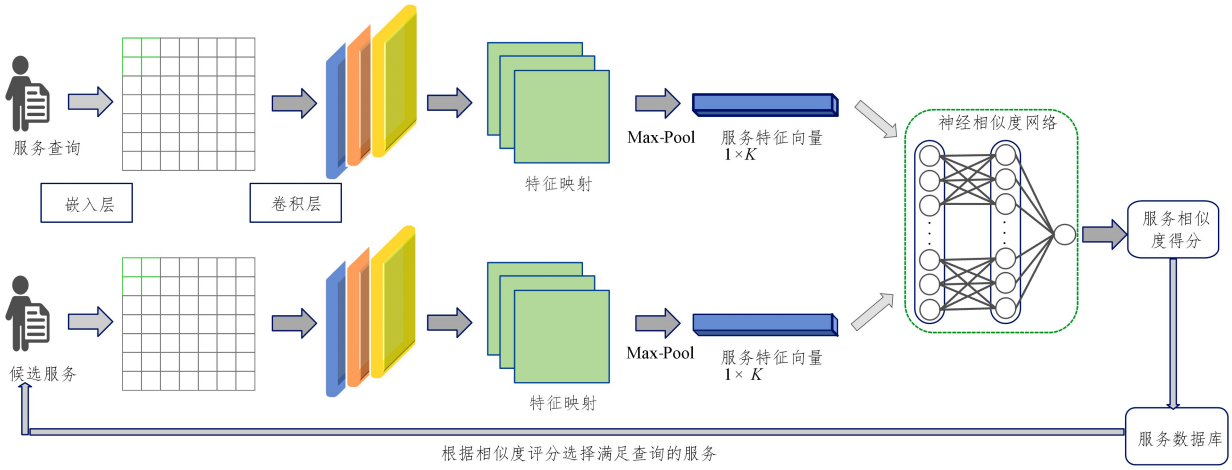

图 1 方法的总体框架

Fig. 1 Overall framework of the proposed method

3.3 基于 Word2Vec 的词嵌入

嵌入模块进行词嵌入、主题嵌入和实体嵌入的工作，为了获得服务描述的嵌入向量，首先对服务描述的短文本进行预处理，包括分词、大小写转化、去除生僻字符等，经过预处理后，我们将服务描述文档中的每个单词通过训练好的 Word2Vec 模型将其投影为固定长度的向量 $\boldsymbol{e}(\boldsymbol{W}_i)\in\mathbb{R}^{d\times 1}$。因此，$(\boldsymbol{e}(\boldsymbol{W}_1)\ \boldsymbol{e}(\boldsymbol{W}_2)\ \cdots\boldsymbol{e}(\boldsymbol{W}_n))\in\mathbb{R}^{d\times n}$ 就是服务描述的单词嵌入矩阵，其中 d 是词嵌入的维数。

3.4 基于 TransD 的知识图谱嵌入

知识图谱是由实体和关系组成的多关系图。每个边代表两个实体，它们之间通过特定的关系来连接，具体来说是三元组的格式 (h,r,t)。h，r 和 t 分别代表头实体、关系和尾实体，如三元组(Google，CEO，Sundar Pichai)表示为 Google 的 CEO 是 Sundar Pichai。知识图谱嵌入的思想是将实体和关系嵌入到一个低纬空间，并将其转换为低维向量，同时保留知识图谱的原始结构。近年来，许多知识嵌入方法被提出，如 2013 年提出的 TransE 方法[24]、2014 年提出的 TransH 方法[25]、2015 年提出的 TransR 方法[26]和 TransD 方法[27]。在实体嵌入阶段，首先使用实体链接技术将服务描述中的实体与提供商实体和知识图谱中的实体进行匹配，然后在知识图谱中构造一个包含该实体的子图。使用 TransD 对实体执行表示学习，并获取该实体嵌入向量 $\boldsymbol{e}(\boldsymbol{E}_j)\in\mathbb{R}^{K\times 1}$。图 2 给出了对服务描述中实体和提供商实体进行嵌入的过程，本文选择实体的一跳内的所有实体一同进行嵌入作为服务实体嵌入矩阵 $(\boldsymbol{e}(\boldsymbol{E}_1)\boldsymbol{e}(\boldsymbol{E}_2)\cdots\boldsymbol{e}(\boldsymbol{E}_n))\in\mathbb{R}^{K\times n}$。其中，$K$ 是向量的维数。

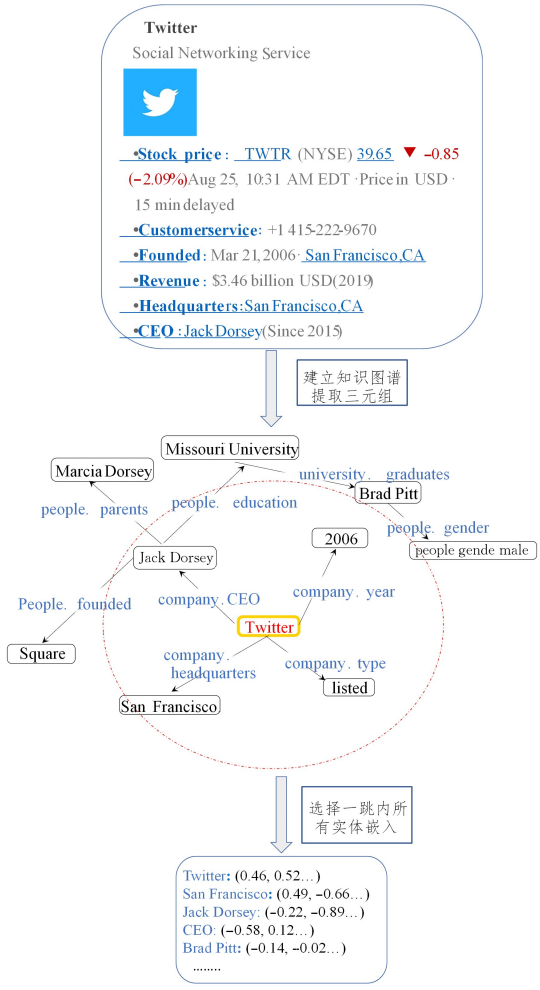

图 2 实体嵌入的过程

Fig. 2 Process of entity embedding

TransD 方法定义了两个向量空间——关系空间和实体空间。每个实体和关系由两个向量表示,第一个向量获取实体/关系的含义,另一个用于构造映射矩阵。例如,给定一个三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$,它的向量是$\boldsymbol{h},\boldsymbol{h}_p,\boldsymbol{r},\boldsymbol{r}_p,\boldsymbol{t},\boldsymbol{t}_p$,其中下标 p 表示该向量为投影向量。其中$\boldsymbol{h},\boldsymbol{h}_p,\boldsymbol{t},\boldsymbol{t}_p\in R^n$和$\boldsymbol{r},\boldsymbol{r}_p\in R^m$。对于每个三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$,设置两个映射矩阵$\boldsymbol{M}_{rh},\boldsymbol{M}_{rt}\in R^{m\times n}$将实体从实体空间投影到关系空间。它们的定义如下:

$$\boldsymbol{M}_{rh}=\boldsymbol{r}_p\boldsymbol{h}_p^{\mathrm{T}}+I^{m\times n} \tag{1}$$

$$\boldsymbol{M}_{rt}=\boldsymbol{r}_p\boldsymbol{t}_p^{\mathrm{T}}+I^{m\times n} \tag{2}$$

得分函数定义为:

$$f_r(\boldsymbol{h},\boldsymbol{t})=-\parallel\boldsymbol{h}_\perp+\boldsymbol{r}-\boldsymbol{t}_\perp\parallel_2^2 \tag{3}$$

其中,$\boldsymbol{h}_\perp=(\boldsymbol{r}_p\boldsymbol{h}_p^{\mathrm{T}}+I^{m\times n})\boldsymbol{h}$,$\boldsymbol{t}_\perp=(\boldsymbol{r}_p\boldsymbol{t}_p^{\mathrm{T}}+I^{m\times n})\boldsymbol{t}$,$\boldsymbol{h}_{ip}$,$\boldsymbol{t}_{ip}$($i=1,2,3$)和关系$\boldsymbol{r}_p$是投影向量,$\boldsymbol{h}_{i\perp}$和$\boldsymbol{t}_{i\perp}$分别是头实体和尾实体的投影向量。所使用的损失函数定义为:

$$L=\sum_{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})\in S}\sum_{(\boldsymbol{h}',\boldsymbol{r},\boldsymbol{t}')\in S'}[\gamma+f_r(\boldsymbol{h},\boldsymbol{t})-f_r(\boldsymbol{h}',\boldsymbol{t}')]_+ \tag{4}$$

其中,γ是超参数,S和S'是正确三元组和不正确三元组的集合。

3.5　基于 LDA 的主题嵌入

主题嵌入采用潜在狄利克雷分配(LDA)模型[28]对服务描述文档建模,以获得服务描述的主题向量$\boldsymbol{e}(\boldsymbol{T}_r)\in\mathbb{R}^{Y\times1}$,$Y$为实体向量的维数。LDA 是一种概率主题模型,可以识别文档中的主题并挖掘语料库中的隐藏信息。LDA 模型的生成过程如下:

(1)给定一个 Web 服务描述文档D_i。

(2)从狄利克雷分布β中取样生成文档D_i的主题分布θ_i。

(3)从主题的多项式分布θ_i中取样生成文档D_i第j个词的主题$Z_{i,j}$。

(4)从狄利克雷分布α中取样生成主题$Z_{i,j}$对应的词语分布$\varphi_{z_{i,j}}$。

(5)从词语的多项式分布$\varphi_{z_{i,j}}$中采样最终生成词语$W_{i,j}$。

最后,当 LAD 算法流程结束时,我们收集 LDA 算法中 Web 服务描述文档D_i的主题分布,作为该 Web 服务描述文档D_i的主题向量。

由于服务描述文档中主题的数量有限,并且主题向量的生成质量高度依赖于主题数目T的选择,因此在实验中,本文将实体嵌入的维度设置为等同主题向量的维度,以确保实验的可靠性,即$Y=K=T$。其中,T为主题的数目,Web 服务语料库中的单词数目为N,$\boldsymbol{\zeta}$是一个长度为T的向量,表示描述文档D_i中所有主题的比例,$\boldsymbol{\psi}$是一个长度为N的向量,表示所有单词的分布,α和β是先验参数,LDA 以所有的描述文档为输入,利用 Gibbs 取样方法可以近似估计潜在变量$\boldsymbol{\zeta}$,$\boldsymbol{\psi}$和Z_i的后验分布。在训练过程中,建立马尔可夫链,并从中提取主题样本,改变链的状态并进行更新。经过 LDA 对描述文档建模后,我们可以得到第i个文档的主题分布,表示为$Topic_i=(z_1^i,z_2^i,\cdots,z_T^i)$。

3.6　服务描述的特征提取

考虑到主题嵌入向量、实体向量和词向量的维数可能不同,不利于后续的研究工作,本文采用可训练的转移矩阵$\boldsymbol{M}\in\mathbb{R}^{d\times k}$进行矩阵对齐,使特征向量的维度相同。因此,可以获得转移的实体嵌入矩阵$\mathbb{R}_{\mathrm{entity}}^{d\times n}$和主题嵌入矩阵$\mathbb{R}_{\mathrm{topic}}^{d\times n}$,其中:

$$\mathbb{R}_{\mathrm{entity}}^{d\times n}=[\boldsymbol{Me}(\boldsymbol{E}_1)\boldsymbol{Me}(\boldsymbol{E}_2)\cdots\boldsymbol{Me}(\boldsymbol{E}_n)] \tag{5}$$

$$\mathbb{R}_{\mathrm{topic}}^{d\times n}=[\boldsymbol{Me}(\boldsymbol{T}_1)\boldsymbol{Me}(\boldsymbol{T}_2)\cdots\boldsymbol{Me}(\boldsymbol{T}_n)] \tag{6}$$

实验中将 3 个对齐的矩阵作为 CNN 的输入,滤波器的大小为$f\in\mathbb{R}^{d\times m\times3}$,并且卷积过程可以写为:

$$\tilde{\boldsymbol{\omega}}^{R\times l\times D}=\sum_{p=1}^{3}\boldsymbol{I}^{U\times V\times C}*\boldsymbol{f}_D^{d\times m\times p} \tag{7}$$

其中,$\boldsymbol{I}$为输入,表示为特征图,大小为$U\times V$,通道数为C,$\tilde{\omega}$为输出,大小为$R\times l\times D$,经过最大池化操作后,服务的特征向量为:

$$\boldsymbol{e}(\boldsymbol{S}_I)=\text{Max-Pool}[\tilde{\boldsymbol{\omega}}_i] \tag{8}$$

3.7　神经相似网络

考虑到神经网络更符合人类对事物的认知,本文使用神经网络来学习服务相似性,该神经网络由一组完全连接的层组成,激活函数设置为 Sigmoid,该神经网络的输入为经 CNN 处理过的一对服务特征向量,KSN 用神经网络学习一个相似函数$H(\cdot)$,以学习服务之间的相似度,并输出两个服务之间的相似度得分$S_{i,j}$。

$$S_{i,j}=H(f(S_i,\boldsymbol{W}_f),f(S_j,\boldsymbol{W}_f),\boldsymbol{W}_H) \tag{9}$$

其中,$\boldsymbol{W}_H$是可学习参数,根据相似度得分,可以找到最类似于需求的前N个服务,为了利用方程(5)优化相似函数$H(\cdot)$的权重W_H,定义下面的损失函数:

$$L(I_i,I_j)=\begin{cases}\dfrac{1}{2}(s_{x_i,s_j}-\mathrm{sim}(I_i,I_j))^2,\\ \qquad\text{if }\mid s_{x_i,s_j}-\mathrm{sim}(I_i,I_j)\mid\leqslant\delta\\ \delta\mid s_{x_ix_j}-\mathrm{sim}(I_i,I_j)\mid-\dfrac{1}{2}\delta^2,\\ \qquad\text{otherwise}\end{cases} \tag{10}$$

其中,$\mathrm{sim}(I_i,I_j)$表示两个服务使用余弦相似度计算出的相似度得分,s_{x_i,s_j}是学习到的相似度得分,而δ是可优化的参数。注意,为每一个可能的训练服务提供相似性标签是不可行的,受图像检索的启发,本文完全遵循文献[29]中的对神经相似网络的训练方法。

3.8　Web 服务发现

经过上述介绍,我们可以定义服务发现过程:给定一个查询服务,从数据库中随机选择任意一个服务计算与查询之间的相似度得分。如果相似度得分高于定义的阈值(设定为 0.5),本文将其称为正样本。继续评估,每次进行选择和重新排名,直到只有最好的样本在列表顶部。从收集到的最佳样本中选择最佳的 top-k 作为发现结果。

4　实验

本节将展示实验结果并讨论参数的调整。

4.1　数据集

本文的数据集是在 ProgrammableWeb 上爬取的 13 884

个服务 API，包括 API 名称、API 提供商信息、服务描述和所属类别。此外，在处理数据阶段，搜索数据集中所有发生的实体以及 Microsoft Satori 知识图谱中它们的一跳内的实体，并以高于 0.9 的置信度提取其中的三元组，同时还删除了 448 个不包含任何实体的服务。服务处理后的数据集描述如表 1 所列。最后，本文随机选择 70% 的服务作为训练集，30% 的服务作为测试集。在实验中，本文发现通过该划分获得的结果是最佳的。

表 1 预处理后的数据集统计

Table 1 Data set statistics after preprocessing

# API	13436
# description	13436
# entities	15220
# relations	24
# triples	59071

注："#"表示"数量"

4.2 评价指标

本文使用 Precision，Recall，F-Measure 来评估本文模型，这些指标定义为：

$$Precision=\frac{|R(c)\bigcap T(c)|}{|T(c)|} \tag{11}$$

$$Recall=\frac{|R(c)\bigcap T(c)|}{|R(c)|} \tag{12}$$

$$F\text{-}Measure=2\frac{Precision*Recall}{Precision+Recall} \tag{13}$$

其中，$|R(c)|$ 是与服务 C 相似的相关服务列表中的服务数量，$|T(c)|$ 是与测试查询 C 关联的排名服务的前 n 个列表中的服务数量。

4.3 对比实验

为了证明 KSN 模型的性能，本文将 KSN 与以下方法进行了比较。

(1)WE-LDA 使用从 Word2vec 模型中学到的词向量来增强 LDA 模型，使用 K-means ++算法对服务进行聚类，并将聚类结果用于服务发现。

(2)T-CNN 使用卷积神经网络提取文本的特征向量，并使用学习的特征向量直接计算文本之间的相似度。

(3)DeepWSC[30]是一个深度神经网络框架，它结合了递归神经网络和卷积神经网络来提取服务描述中的特征，并使用 K-means++算法对服务进行聚类。我们使用 DeepWSC 所学习的特征向量作为相似网络的输入与本文方法进行比较。

(4)WSC-GCN[31]首先以 Web 服务的名称、描述文字、标签为基本语料，根据单词共现和单词来构建"Words 和 Web 服务描述文档"的异构图形网络，利用图卷积神经网络提取特征。

本文方法结合知识图谱获取服务描述的外部信息，使用卷积神经网络提取服务特征向量，并设计用于服务发现的监督相似网络。

4.4 实验设置

为了证明知识图谱嵌入对于服务发现的有效性以及本文方法的可行性，本文进行了对比实验。由于 T-CNN 和 DeepWSC 都是深度学习方法，为了比较的公平性，本文将词向量维度统一设置为 128，过滤器统一设置为 50，使用 SGD 训练模型，对于 WE-LDA，本文将主题数 K 设置为 50，这与本文的 KSN 的主题设置数量一致。此外，采用随机分区工具 Sklearn，将数据集分为 70%的训练集和 30%的测试集。KSN 模型的一些重要参数包括：$Learning\ rate=0.02$，$Epochs=20$，$Dropout=0.5$。所有实验都在 Intel Core i5-6200U CPU 上运行。

4.5 参数影响

本文模型有两个可以学习的参数：主题数目 K 和 δ，主题数目本文设定为 50，参数 δ 的不同取值对性能的影响如图 3 所示，本文报告了 δ 在[0，0.8]范围内对服务发现 F 值的影响。当 δ 为 0.6 时实验结果最好，因此在后续的对比实验中，本文方法默认 δ 为 0.6。

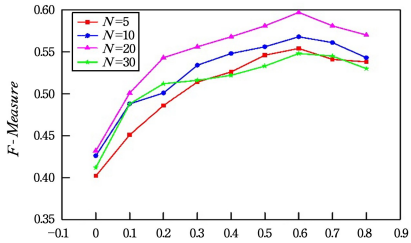

图 3 参数 δ 的影响

Fig. 3 Impact of parameter δ

4.6 实验结果

本文首先进行了消融实验，以显示融合知识图谱信息的重要性。本文报告了服务发现的 *F-Measure*，结果如表 2 所列。

表 2 不同 KG 嵌入和 KG 嵌入去除的结果比较

Table 2 Comparison of the results of different KG embedding and KG embedding removal

	F-Measure			
	$N=5$	$N=10$	$N=20$	$N=30$
KSN with KG($k=1$)	0.554	0.568	0.597	0.548
KSN with KG($k=2$)	0.537	0.549	0.553	0.503
KSN without KG	0.402	0.446	0.443	0.399

可以看出，该模型的 *F-Measure* 在删除知识嵌入后得到显著降低。值得注意的是，在移除知识嵌入后，本文将服务描述的主题向量和词向量进行了串联操作并输入神经相似网络中。服务实体的知识图谱嵌入使 *F-Measure* 提高了14.5%。另外，当选择对子图 2 跳内实体进行嵌入时，性能有所下降，这可能是因为 2 跳内实体的嵌入引入了部分与服务不相关的实体，导致知识噪声。

图 4－图 6 分别给出了通过 5 种方法获得的精度、召回率和 F-Measure。

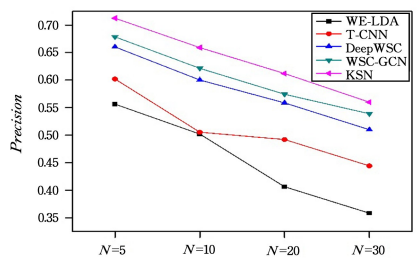

图 4 不同 N 值对应的精度

Fig. 4 Recall with different N

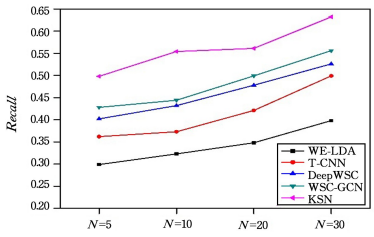

图 5　不同 N 值对应的召回率

Fig. 5　*Recall* with different N

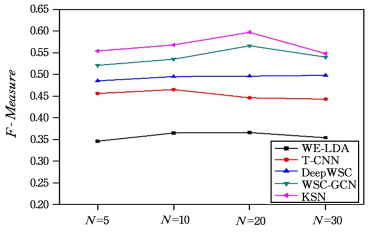

图 6　不同 N 值对应的 *F-Measure*

Fig. 6　*F-Measure* with different N

通过分析可知，随着 N 值的增加，每种方法的召回率和 *F-Measure* 逐渐增加，而精度逐渐降低。这是因为随着 N 值增加，意味着将有更多服务无法与候选服务进行比较。我们可以观察到，深度学习方法 T-CNN，DeepWSC，WSC-GCN 和本文的 KSN 均优于 WE-LDA，表明基于深度学习的方法得到的服务特征向量有助于提高服务发现性能，原因是它可以捕获服务描述文档的隐式上下文信息，并且可以更准确地表示服务之间的关系。并且通过神经网络处理后的向量的语义程度远高于 LDA 模型的服务主体概率分布。WSC-GCN 在 3 个评价指标上都优于 T-CNN 和 DeepWSC，表示将服务看作为网络的这种思想有助于提高服务发现性能，当 N 取 5 时，在 F-Measure 指标上，WSC-GCN 相比 T-CNN 和 DeepWSC 分别提升了 14%和 7%；本文方法 KSN 结合了知识图谱，以扩展服务描述文档的外部知识，相比 WSC-GCN，当 $N=5$ 时，有接近 5%的提升，这是因为本文实验所用的数据集比较小，WSC-GCN 将服务建模为图的形式，会导致图的稀疏性。而 KSN 对小规模数据集有着更好的适应性。当 $N=5$ 时，服务发现的精度达到了 72%，远高于 WE-LDA 的44.6%、T-CNN 的 58.5%和 DeepWSC 的 62%。

虽然本文方法在性能上优于基线方法，但是仍有一定的缺陷。图 7 给出了 5 种方法的计算成本，对于典型的聚类方法——WE-LDA，其计算成本较低，将聚类于深度学习融合的方法——DeepWSC 的计算成本也优于 KSN，这给了本文一个直观的启发，聚类服务将会是提升服务发现性能的一个重要途径。

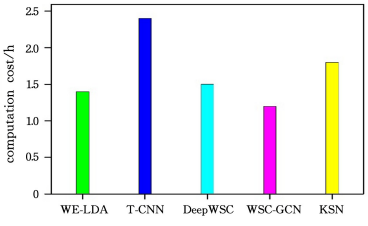

图 7　实验的计算成本比较

Fig. 7　Comparison of computational cost of Baselines

为了降低计算成本，本文建议采用缓存的机制，存储所有查询的 3 种嵌入表示，当执行相同查询时，Web 服务发现系统自动调取缓存与服务库中的服务进行相似度计算，而不是重复地进行服务嵌入，这将大幅降低计算成本，从而改进服务发现方法。

结束语　本文提出了一种基于知识图和相似网络的服务发现方法(KSN)，KSN 通过将服务描述中的实体和提供商实体链接到知识图谱，利用丰富的外部知识扩展服务描述文档的信息，同时结合 Word2Vec 模型和 LDA 模型获得高质量的特征向量，最后使用 CNN 提取服务描述的隐藏上下文信息来获得信息丰富的服务向量表示。为了更好地适应人类的认知，本文设计了一种相似度网络来学习一个相似度函数，用于计算服务之间的相似度，从而进行服务发现过程。在从 ProgrammableWeb 上爬取的真实数据集上进行的大量实验表明，KSN 与基准方法相比有显著改进。

但是，KSN 仍然有一定的局限性，计算成本相比基线方法偏高，这会导致其在大规模服务检索任务上不占优势，未来的工作一是尝试提取并利用更多描述文本功能来实现 Web 服务发现，二是考虑融合聚类方法到本文模型中以降低模型的计算成本。

参 考 文 献

[1] MCILRAITH S A，SHEILA A，TRAN C S，et al. Semantic web services[C]// IEEE Intelligent Systems. 2001：46-53.

[2] CRASSO M，ZUNINO A，CAMPO M. Combining query-by-example and query expansion for simplifying web service discovery[C]// Information systems frontiers. 2011：407-428.

[3] ZHANG N，WANG J，MA Y，et al. Web service discovery based on goal-oriented query expansion[J]. Journal of Systems and Software，2016，1(4)：73-91.

[4] WEN T，SHENG G，LI Y，et al. Research on Web service discovery with semantics and clustering[C]// 2011 IEEE 6th Joint International Information Technology and Artificial Intelligence Conference. 2011：62-67.

[5] YU Q，WANG H，CHEN L. Learning sparse functional factors for large-scale service clustering[C]// IEEE International Conference on Web Services. 2015：201-208.

[6] CAO B Q，FRAN K，LIU X Q，et al. Integrated Content and Network-Based Service Clustering and Web APIs Recommendation for Mashup Development[J]. IEEE Transactions on Services Computing，2020(13)：99-113.

[7] LIU Y S，YANG Y C. Semantic web service discovery based on text clustering and similarity of concepts[J]. Computer Science，2013，40(11)：211-214.

[8] QIU T，LI P F，LIN P. A Web Service Matching Algorithm Based on Semantic Similarity of Concepts[J]. Chinese Journal of Electronics，2009，37(2)：429-432.

[9] YE H F，CAO B Q，PENG Z L，et al. Web Services Classification Based on Wide & Bi-LSTM Model[J]. Access IEEE，2019(7)：43697-43706.

[10] GAO Z，FAN Y，WU C，et al. Seco-lda：Mining service co-occurrence topics for composition recommendation[J]. IEEE Transac-

tions on Services Computing, 2018, 12(3): 446-459.
[11] LIANG T, CHEN L, WU J, et al. SMS: A framework for service discovery by incorporating social media information[J]. IEEE Transactions on Services Computing, 2016, 12(3): 384-397.
[12] LIU M, SHEN W, HAO Q, et al. A weighted ontology-based semantic similarity algorithm for web service[J]. Expert Systems with Applications, 2009, 36(10): 12480-12490.
[13] RUPASINGHA R, PAIK I, KUMARA B. Improving Web Service Clustering through a Novel Ontology Generation Method by Domain Specificity[C]// 2017 IEEE International Conference on Web Services (ICWS). 2017: 744-751.
[14] YANG D J, HE D. Web Service Clustering Method Based on Word Vector and Biterm Topic Model[C]// 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA). 2021: 299-304.
[15] CAO B, LIU J, WEN Y, et al. QoS-aware service recommendation based on relational topic model and factorization machines for IoT Mashup applications[J]. Journal of Parallel & Distributed Computing, 2018, 132(OCT.): 177-189.
[16] DENGPING W, TING W, JI W. Web Service Discovery Combining Description Document Structure and Reference Features [J]. Journal of Software, 2011, 22(9): 2006-2019.
[17] CAO B, LIU J, SHI M, et al. Mashup service clustering based on an integration of service content and network via exploiting a two-level topic model[C]// IEEE International Conference on Web Services. 2016: 212-219.
[18] WANG H, ZHANG F, XIE X, et al. DKN: Deep knowledge-aware network for news recommendation[C]// Proceedings of the 2018 World Wide Web Conference. 2018: 1835-1844.
[19] MA X J, CHEN J D, LI K. Semantic Web service discovery based on IO and information content [J]. Computer Systems & Applications, 2016(2): 141-145.
[20] SHI M, LIU J, CAO B, et al. A prior knowledge-based approach to improving accuracy of Web services clustering[C]// IEEE Conference Services Computing. 2018: 1-8.
[21] SHI M, LIU J, ZHOU D, et al. WE-LDA: A word embeddings augmented LDA model for web services clustering[C]// IEEE International Conference on Web Services. 2017: 9-16.
[22] CHEN F, LU C H, WU H, et al. A semantic similarity measure integrating multiple conceptual relationships for web service discovery[J]. Expert Systems with Applications, 2017, 67: 19-31.
[23] UPADHYAYA B, KHOMH F, ZOU Y, et al. A concept analysis approach for guiding users in service discovery[C]// 2012 Fifth IEEE International Conference on Service-oriented Computing and Applications (SOCA). IEEE, 2012: 1-8.
[24] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Neural Information Processing Systems (NIPS). 2013: 1-9.
[25] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2014, 28(1): 1112-1119.
[26] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2015, 29(1): 2181-2187.
[27] JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 1: Long papers). 2015: 687-696.
[28] BLEI D M, GN A Y, JORDANM I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
[29] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]// International Conference on Learning Representations. 2015: 9-16.
[30] ZOU G, QIN Z, HE Q, et al. Deepwsc: A novel framework with deep neural network for web service clustering[C]// 2019 IEEE International Conference on Web Services. IEEE, 2019: 434-436.
[31] YE H, CAO B, CHEN J, et al. A Web Services Classification Method Based on GCN[C]// 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking. IEEE, 2019: 1107-1114.

YU Yang, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include service computing and recommendation systems.

XING Bin, born in 1962, professor senior engineer, postgraduate. His main research interests include application of industrial bigdata technology.