软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(9):2006−2019 [doi: 10.3724/SP.J.1001.2011.03887]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

融合描述文档结构和参引特征的 Web 服务发现*

魏登萍+, 王 挺, 王 戟

(国防科学技术大学 计算机学院 并行与分布处理国家重点实验室,湖南 长沙 410073)

Web Service Discovery by Integrating Structure and Reference Features of Description Documents

WEI Deng-Ping+, WANG Ting, WANG Ji

(National Key Laboratory of Parallel and Distributed Processing, College of Computer, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: E-mail: dpwei@nudt.edu.cn

Wei DP, Wang T, Wang J. Web service discovery by integrating structure and reference features of description documents. *Journal of Software*, 2011,22(9):2006−2019. http://www.jos.org.cn/1000-9825/3887.htm

Abstract: This paper first investigates two main kinds of features of Web service description language (WSDL) documents: the structure features and the reference features. Next, a novel multi-vector model for Web services is introduced, which is distinguished from the general text representation model by the explicit features of Web services. The structure features are represented by multiple vector spaces and the term weighting in the sub-vector is determined by the reference features. A method to compute the similarity between two Web services is proposed and a Web service discovery prototype system based on this new model is implemented. Finally, a Web service discovery test collection is constructed, which has 1576 WSDL documents together with incomplete relevance judgments. The experimental results on this collection show that Web service discovery based on the proposed model is more effective than based on simple vector space model of text with the confidence of 95%.

Key words: Web service; Web service discovery; WSDL; vector space model; incomplete information; average precision

摘 要: 首先分析研究 Web 服务描述文档(WSDL 文档)的两大特征——结构特征和参引特征,然后根据各个特征对 Web 服务功能语义描述的影响,提出相应的 Web 服务表示模型——多向量表示模型.区别于通用文本表示模型,该模型能够显式地表示Web服务描述文档的本质特征.其中,结构特征语义表现在多向量空间的划分上,参引特征语义映射到子向量模型中特征权重的计算上.提出了基于多向量模型的 Web 服务相似度计算方法,并实现了基于该模型的 Web 服务发现原型系统.最后,在真实 Web 服务描述文档集合上构造了一个具有不完全相关性判断且涵盖了1576 个 WSDL 文档的 Web 服务发现测试集,并在该测试集上进行了基于多向量模型的 Web 服务发现实验评估.实验结果表明,基于多向量模型的 Web 服务发现方法的检索效果比基于简单文本向量空间模型发现方法的检索效果在 95%的置信度下有了显著提高.

* 基金项目: 国家自然科学基金(60873097, 90612009); 国家重点基础研究发展计划(973)(2005CB321802); 新世纪优秀人才计划(NCET-06-0926)

收稿时间: 2009-08-04; 定稿时间: 2010-05-05

关键词: Web 服务;Web 服务发现;WSDL;向量空间模型;不完全信息;平均准确率
中图法分类号: TP311 文献标识码: A

自从以 Web 服务为基础的分布式计算模式 SOC(service-oriented computing)产生以来,互联网上的 Web 服务急剧增多.Web 服务发现作为实现 Web 服务体系结构的一个首要任务,已成为能否成功实现面向服务计算模式的关键问题,并得到越来越多研究人员的关注.目前,Web 服务发现的方法主要分为两大类:基于语法的发现方法和基于语义 Web 技术的发现方法.基于语法的方法实现简单,计算复杂性低,但是发现准确率较低.语义 Web 服务技术的核心是利用本体对 Web 服务的功能进行形式化描述,使其功能语义更加明确.该类方法发现准确率较高,但本体推理的时间复杂性过高;同时,准确地标注大量的 Web 服务需要耗费较大的代价,是一项艰巨的工作.因此,在当前互联网上 Web 服务数量与日剧增[1]且语义 Web 服务技术尚未成熟的情况下,如何进一步提高当前大量 Web 服务发现的效果是亟待解决的问题.鉴于此,我们的工作采用两阶段的 Web 服务发现策略:首先,采用基于语法的方法来匹配 Web 服务描述文档,以尽可能准确地选取满足用户需求的候选 Web 服务集合;然后,再利用语义 Web 技术准确地定位满足用户需求的 Web 服务.我们之前的工作[2]着重完成了第 2 阶段的工作,本文的主要工作集中在第 1 阶段,即力图从大量的 Web 服务描述文档中有效地检索出满足用户需求的 Web 服务.

基于语法的 Web 服务发现方法主要有基于关键字的检索、基于 XML Schema 的匹配和基于向量空间模型的检索.基于关键字的检索是当前大多数 Web 服务门户网站(portal)(如 Xmethods,BindingPoint 等)和各大搜索引擎(如 Google,Yahoo!等)采用的主流技术,它们都不支持针对 WSDL 文档特征的特定检索,使得对 WSDL 文档检索的准确率较低.

基于 XML Schema 匹配的发现方法[3−5]主要解决 Web 服务操作的输入/输出参数的匹配问题,通过计算输入/输出参数对应的 XML Schema 数据类型定义的语法和结构的匹配值来度量操作之间的相似度.该类方法主要用于计算两个特定的 Web 服务操作的相似性,进而度量两个 Web 服务的相似性.它对于整个 Web 服务的语义缺乏全面的理解.同时,基于 XML Schema 匹配的计算复杂性较高,其应用于大量 Web 服务的检索具有一定困难.

基于向量空间模型的检索采用词向量来表示 Web 服务描述文档,并通过计算查询向量与 Web 服务描述文档向量之间的相似度来度量查询与 Web 服务的相似度.文献[6]提出了基于向量空间模型的 Web 服务搜索引擎.文献[7]的工作主要集中在研究采用何种相似度度量方法和何种排序算法对于 Web 服务发现有较好的效果.这些方法忽略了 Web 服务描述文档的特性,使得 Web 服务的发现与文本检索没有本质区别.因此,本文的目标是分析 Web 服务描述文档的本质特征,进而研究基于这些特征的 Web 服务发现的特有方法,从根本上与通用的文本检索任务区分开来,以提高 Web 服务发现的性能.

本文从分析 Web 服务描述文档的结构特征和参引特征出发,分析挖掘其不同于一般文本和 XML 文档的本质特征,并根据这些特征对 Web 服务重新建模,使得新模型既能反映一般文本文档的特征又能体现 Web 服务描述文档的特有特征.然后,将 Web 服务的表示模型映射到多向量模型上,使得 Web 服务的发现问题转化为基于多向量模型的检索问题,提出了基于多向量模型的 Web 服务发现方法.最后,在大量真实实验数据集 QWS-wsdls[1]上构造了一个具有不完全相关性判断的 Web 服务发现测试集.实验验证了基于多向量模型的 Web 服务发现效果比基于简单文本向量空间模型的发现效果在 95%的置信度下具有显著提高.

1 Web 服务描述文档特征

Web 服务描述语言 WSDL 作为 Web 服务的标准描述语言,定义了 Web 服务的两部分信息:抽象定义信息和具体描述信息.WSDL1.1 规范[8]定义了 Web 服务抽象信息的 3 个主要组件:(1) PortTypes,定义 Web 服务各个操作的抽象信息,包括操作名、操作相关文本描述、输入消息和输出消息;(2) Message,定义各个操作所需要的消息类型,包括消息名称、数据类型等;(3) Types,独立于平台和语言的类型定义,定义了各个消息类型所参引的元素信息或者数据类型信息.通常,Types 由 XML Schema 实现.具体描述主要用于描述 Web 服务的调用信息,包括访问 Web 服务采用的通信协议以及访问地址,它主要包括绑定(binding)信息和服务(service)信息:绑定信息定义

了 PortTypes 中每一部分操作的绑定实现;Services 信息定义了每一绑定的端口地址.本文主要目标是重点分析 Web 服务描述文档中的抽象定义信息的特征,并研究基于这些特征的 Web 服务发现方法.

1.1 结构特征

WSDL 语言是专为描述 Web 服务而提出的一种特定的 XMLSchema 定义,因此,WSDL 文档作为一种特有的 XML 文档,具有 XML 文档的结构特性.此外,它还存在着明显区别于 XML 文档的特有特征:

(1) 语义特性.WSDL 语言是基于 XML 描述框架的语言,它定义了描述 Web 服务的元语言.WSDL 文档中的各个组件(component)定义了 Web 服务不同方面的信息,如 wsdl:Operation 定义了 Web 服务中的操作信息.即 WSDL 文档中每个组件都描述了 Web 服务不同方面的语义,它相当于是对 Web 服务描述信息进行了语义标注的格式化文档.通用 XML 文档的 Schema 只是定义了 XML 文档的结构信息,部分语义隐含在其结构中,但各个元素自身的语义信息并没有显式定义;

(2) 结构统一属性.WSDL 文档是根据 WSDL 语言来描述 Web 服务的格式化文档.因此,遵循同一 WSDL 规范的所有 WSDL 文档的结构相同.而 XML 文档由于其 Schema 定义的多样化,XML 文档的结构不具有统一属性.

文本的向量空间表示模型对各个词汇出现的位置是不敏感的,即只要文本中出现相应的词汇,不管其出现在什么位置,都具有相同的向量表示.而实际应用中,Web 服务的结构特征使得 WSDL 文档中描述信息的出现位置对 Web 服务的语义影响较大.例如,同时出现在服务名称和 XML Schema 中的一个词,由于其出现在不同的 Web 服务描述组件中,使得该词具有不同的语义角色.采用简单的词向量组成的文本向量空间模型表示 WSDL 文档,可能会使得具有相同向量表示的 WSDL 文档的语义可能截然不同.因此,利用文本向量来表示 Web 服务描述文档存在较大的局限性,它忽略了 Web 服务描述中的结构特征所隐含的语义信息.

1.2 参引特征

在一般信息检索中,文本中的单个词项对文本语义的贡献取决于它在该文本中出现的次数或者位置(新闻文本中的标题、首句等).除了具有较好的结构特征外,WSDL 文档与一般文本文档的本质区别还体现在参引结构上:

(1) WSDL 文档中各部分元素存在相互参引,使得整个文档的描述呈现出一种类似树的知识表示结构.例如,操作描述参引消息定义,消息定义参引数据类型定义.而一般文本中各部分信息不存在相互参引关系,它的信息呈一种平面的结构;

(2) WSDL 文档允许参引外部资源,主要体现在两个方面:(a) WSDL 规范指出,Web 服务的抽象描述信息可以单独组成一个独立的文件,供多个 WSDL 文档参引使用.这就使得单个 WSDL 文档中可能不存在抽象描述信息,需要解析其参引的独立抽象描述信息文档才能获得其抽象描述信息;(b) 对外部 XML Schema 的参引,XML Schema 规范定义使用“import”元素可以参引外部 XML Schema 定义.因此,Web 服务的语义不仅仅体现在 WSDL 文档内部,还体现在 WSDL 文档所参引的外部资源内.而文本通常是一个独立的个体,其信息主要体现在文本内部;

(3) WSDL 文档使用的 XML Schema 中的元素存在相互参引.例如,一个 XML Schema 中的元素 a 是另一个 XML Schema 中的元素 b 的子元素;一个 XML Schema 中的数据类型 t 是同一个或者另一个 XML Schema 中定义的元素 c 的数据类型.

基于上述参引特征,每一个 Web 服务的描述文档应该由其自身的 WSDL 文档和该 WSDL 文档所参引的各个文档组成.同时,通过分析大量 WSDL 文档发现,并不是每个 XML Schema 元素或者数据类型都被该 WSDL 文档中的组件一一参引,会存在有些元素或数据类型被多次参引,而有些元素或数据类型从未被参引的现象.

与语义蕴含在文档内部的文本文档不同,WSDL 文档中没有出现的外部 XML Schema 信息,应该作为 Web 服务语义描述的一部分;同时,显示出现在 XML Schema 中的信息,由于没有被任何该 Web 服务的组件所参引,不能很好地帮助描述 Web 服务的语义.被多个操作参引的元素显然比被单个操作参引的元素更能体现 Web 服

务的数据语义.基于这些特征分析,我们提出如下假设:

假设 1(贡献假设). 对某 Web 服务描述文档 d 中的元素或数据类型定义 a,若参引 a 的组件越多,则它对于该 Web 服务语义的描述越重要.

定义 1(元素集合 E). 元素集合 E 为 WSDL 文档中使用的所有 XML Schema 中出现的以“element”为标签的元素名称的集合.XML Schema 定义中的元素集合 E 可分为顶层子元素集合 E_1 和非顶层子元素集合 E_2.

图 1 所示 Web 服务定义中的元素集合为所有 types 定义中的元素集合,即 $E=\{A,B,C,D,E,F,G,H,I,J,K,L\}$.图中顶层元素用点圈表示,即 $E_1=\{A,D,G,H,L\}$.

定义 2(元素直接引用次数). 定义函数 $DR:E\to\mathcal{N}$ 为从某 Web 服务 XML Schema 定义中所有元素集合 E 到自然数的映射空间,$DR(e)$表示元素 e 被 Web 服务中操作消息直接引用的次数,非顶层子元素的直接引用次数为 0.图 1 中,元素 H 的直接引用次数为 2,即 $R(H)=2$.

定义 3(元素的父元素集合). 定义函数 $P:E\to 2^E$ 为从 Web 服务 XML Schema 定义中所有元素到集合 E 的幂集的映射,$P(e)$表示元素 e 在 XML Schema 定义组成的树型结构中的所有父亲元素的集合.图 1 中,元素 G 的所有父亲元素为$\{D,J\}$,即 $P(G)=\{D,J\}$.

定义 4(元素的引用次数). 定义函数 $R:E\to\mathcal{N}$ 为从某 Web 服务 XML Schema 定义中所有元素集合 E 到自然数的映射空间,$R(e)$表示元素 e 的引用次数,且元素 e 的引用次数为所有其父元素的引用次数与该元素的直接引用次数之和,即 $R(e)=\sum_{e'\in P(e)}R(e')+DR(e)$.

通过元素的引用次数,可以计算各个元素描述信息的使用频率.图 1 中,元素 D 被操作 2 的消息所参引,所以 $DR(D)$的值为 1.元素 D 的父亲元素 A 被操作 1 的消息所参引,因此该元素的 $R(D)$值为 2.元素 C 未被任何操作的消息所参引,所以 $DR(C)$的值为 0,其父亲元素 A 的 $R(D)$值为 1.因此,元素 C 的 $R(C)$值为 1.

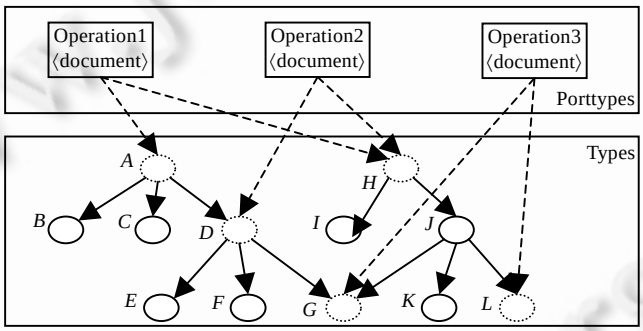


Fig.1 A snippet of a Web service abstract description

图 1 一个 Web 服务抽象描述片断

2 多向量表示模型

2.1 Web服务的抽象描述

在不区分 WSDL 规范版本的情况下,Web 服务的抽象描述主要包括操作(operation)、消息定义(message)、类型定义(types)这 3 个部分,且每部分都可能包含该元素的文本描述.通过分析当前互联网上存在的大量 WSDL 文档发现,高达 98%的消息名称在去掉 SoapIn,HttpIn 等对 Web 服务功能语义描述无贡献的词后,与操作名称完全相同,约占 2%的消息名称由于字母缩写等拼写习惯而与操作名称不相同.例如,名称为 AccountRequest 的操作对应的消息名称为 AR.根据此现象,本文分析利用的 Web 服务抽象信息如图 2 所示(不包括消息名称).

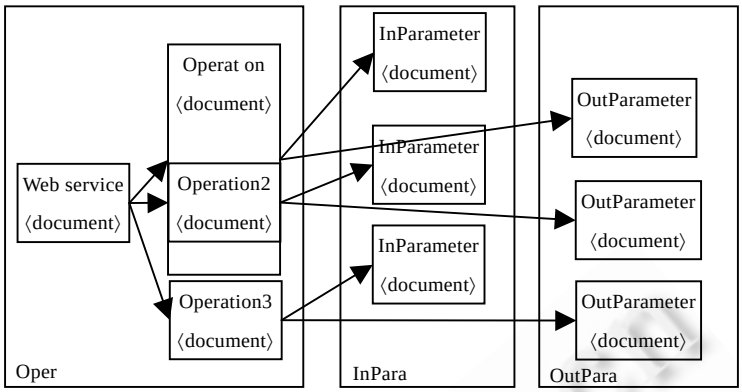


Fig.2 Partition of Web service abstract description

图 2 Web 服务抽象描述划分

定义 5(Web 服务抽象描述). Web 服务 $WS=\langle Name_ws,Document_ws,O,E\rangle$主要由 4 部分信息组成: *Name_ws* 为该 Web 服务的名称;*Document_ws* 为该服务的标注文本,对应 WSDL 中 *Definition* 下的 *document* 标签所定义的内容;$O=\{Operation_1,Operation_2,\ldots,Operation_m\}$表示该 Web 服务的 m 个抽象操作描述信息,对应 WSDL 文档中 Interface(WSDL2.0)或者 porttypes(WSDL1.1)中定义的 m 个抽象操作描述信息;E 为 Web 服务所使用的 XML Schema 中的元素集合.

定义 6(元素描述). 元素 $e=\langle Name_e,Document_e,Offspring,type\rangle$由 4 部分信息组成:*Name_e* 表示该元素的名称;*Document_e* 表示该元素的标注文本,对应 Element 下 *document* 标签的内容;函数 $Offspring:E\rightarrow 2^E$ 为元素到元素集合的映射,表示元素 e 的所有子孙元素(若元素 e 为简单类型,则它没有子孙元素;若该元素为复杂类型,则该元素的子孙元素为所有组成复杂类型元素的子孙元素构成);函数 $type:E\rightarrow\{in,out,null\}$为元素到类型的映射,$type(e)$表示元素 e 被操作参引的类型,in 表示元素 e 为该操作的输入参数,out 表示元素 e 为该操作的输出参数,null 表示该元素不被任何操作所引用.

定义 7. 函数 $Para:O\rightarrow 2^E$ 为操作到元素集合的映射空间,$Para(o)$表示操作 o 所参引的元素集合.若元素 $e\in Para(o)$,则对$\forall e'\in Offspring(e)$,有 $e'\in Para(o)$.

定义 8(操作描述).操作 $o=\langle Name_o,Document_o,E_o\rangle$主要由 3 部分信息构成:*Name_o* 表示操作 o 的名称;*Document_o* 表示操作 o 的标注文本,对应 Operation 组件描述中 document 标签的内容;$E_o=Para(o)$表示操作 o 的消息所参引的所有元素集合.

2.2 多向量模型

2.2.1 单向量模型

定义 9(Web 服务的文本描述集合). 一个 Web 服务所有的描述文本的集合 d 构成了该 Web 服务的文本描述,即 $WS_d=\{ServiceDes,OperationDes,SchemaDes\}$由服务描述信息集合、操作描述信息集合和数据类型描述信息集合组成.其中,$ServiceDes=\{ServiceName,ServiceDoc\}$由服务名称和服务描述文本组成,$OperationDes=\{OperationName,OperationDoc\}$由操作名称和操作描述文本组成,$SchemaDes=\{ElementName,ElementDoc\}$由元素名称和元素描述文本组成.

定义 10(单向量模型). 根据 Web 服务的文本描述集合组成的文档 d,建立文档 d 的单向量模型 $WS_S=(w_1(f_1,d),\ldots,w_n(f_n,d))$来表示 Web 服务,其中,$w_i(f_i,d)$表示项 f_i 在文档 d 中的特征权重,n 表示向量的维度.

单向量模型的向量空间由所有出现在文本描述信息中的词项组成.建立 Web 服务的单向量模型时,将 Web 服务描述文档看作普通文本,各个部分词项的权重由该词在整个抽象描述文档中的出现频率与其倒排文档频率来决定.

2.2.2 多向量模型

定义 11(*n* 元划分). 每一个文档 d 可以通过某些划分规则 $Rule$ 划分为一系列子文档集合$\{d_1,d_2,\ldots,d_n\}$,记为 $d \xrightarrow{Rule} \{d_1,d_2,...,d_n\}$,$n$ 为该次划分的粒度.

任意一个 Web 服务 ws 的抽象描述信息(如图 2 所示)组成该 Web 服务对应的描述文本 d,我们根据 Web 服务的结构特征定义相应的划分规则对文档 d 进行划分,得到子文档集合$\{d_1,d_2,\ldots,d_n\}$.其中:n 为子文档集合的势,表示划分粒度;文档 d_i 表示从该 Web 服务描述文档 d 中抽取出来的某部分信息组成的子文档.

定义 12(多向量模型). Web 服务抽象描述信息 d 的多向量模型可以表示成 $WS_m=\langle V_1,V_2,\ldots,V_m\rangle$,其中,$m$ 表示该 Web 服务抽象描述信息对应的向量个数,它是 Web 服务抽象描述信息划分的粒度;

$V_i=(w_1(f_i,d_i),...,w_{n_i}(f_{n_i},d_i))$ 表示子文档 d_i 对应的向量.其中,$w_j(f_j,d_i)$表示项 f_j 在子文档 d_i 中的特征权重,n_i 表示子向量 V_i 的维度.

通过上述定义可以看出,来自不同组件的同一个词在不同的子文档向量中得以区分,且向量中的项权重由该词在子文档中的重要程度来决定.单向量模型由于不能表达词出现的位置,无法区分一个词到底是出现在 Web 服务操作描述中,还是出现在 Web 服务的类型系统定义中.因此,多向量模型比单向量模型能够更准确地反映 Web 服务的结构特征.

本文根据 Web 服务的数据语义的准确程度对其抽象描述信息实现了两种划分:二元划分与三元划分.

定义 13(二元划分). Web 服务 $WS=\{Operation,Element\}$由操作信息和元素信息组成,其中:$Operation=\bigcup_{i=1}^{\#O}\{Name\_o_i,Document\_o_i\}$ 为该 WSDL 文档中所有操作的名称和相应的文本标注信息的集合,其中,$\#O$ 表示该 Web 服务中操作的个数; $Element=\bigcup_{i=1}^{\#O}\{e\in E\wedge e\in Para(o_i)\}$ 为所有被操作参引的元素集合,单个元素信息包括元素名称和元素的文本标识.

二元划分后的 Web 服务描述信息可以由二元向量模型 $WS_2=\langle Vec_ops,Vec_Ps\rangle$表示,其中,$Vec\_ops$ 为操作描述子文档的向量表示,Vec_Ps 为元素描述子文档的向量表示.

定义 14(三元划分). 一个 Web 服务 $WS=\{Operation,InElement,OutElement\}$由操作信息、输入元素和输出元素信息组成,其中: $Operation=\bigcup_{i=1}^{\#O}\{Name\_o_i,Document\_o_i\}$ 为 WSDL 文档中所有操作的名称和相应的文本标注信息的集合; $InElement=\bigcup_{i=1}^{\#O}\{e\in E\wedge e\in Para(o_i)\wedge type(e)="in"\}$ 为所有操作的输入消息所参引的元素集合; $OutElement=\bigcup_{i=0}^{\#O}\{e\in E\wedge e\in Para(o_i)\wedge type(e)="out"\}$ 为所有操作的输出消息所参引的元素集合.其对应的三元向量模型为 $WS_3=\langle Vec\_ops,Vec\_inPs,Vec\_outPs\rangle$,其中,$Vec\_ops$ 为操作子文档的向量表示,Vec_inPs 为输入参数子文档的向量表示,Vec_outPs 为输出参数子文档的向量表示.

从上述定义可以看出,三元划分是在二元划分的基础上,根据操作所参引的元素类型进一步区分输入元素与输出元素.该划分的依据是区分输入/输出参数能够很好地描述 Web 服务的数据语义,有利于进一步准确地描述 Web 服务.

2.3 特征权重计算

通过分析我们发现,WSDL 文档中组件名称和数据类型定义名称大多采用复合词的形式命名.例如,Google 搜索服务的名称为“GoogleSearch”,它由两个单词“Google”和“Search”组成.对于 Web 服务描述文档 d,若直接以复合词为特征建立向量,则可能会导致具有不同词法表示但词义相同的复合词映射到文档向量的不同特征维,造成匹配率较低.例如,“BookRoom”与“RoomBooking”都表示预定房间,它们对应于特征向量中的不同特征,分别由这两词组成的文档的向量将匹配失败.因此,有必要将复合词进行分解,以单词作为特征来建立向量模型.Web 服务描述文档中的复合词除了以单词的首字母大写直接拼接而成的形式外,大量复合词的组成很不规则.它们要么直接以单词的小写形式拼接而成,要么包含某些单词缩写、数字或者一些特殊符号.这些特性使得复合词的分解存在一定的困难,为此,我们设计并实现了一种专门用于分解软件文档中复合词的基于词典的双向交叠匹配复合词分解算法.

定义 15(出现). 若文档 d 中任意一个复合词 c 可分解为若干单词的序列 $w_1,w_2,\ldots,w_n$,则称复合词 c 为单词 w_i 的一次出现,记为 $c\in Occ(w_i)$.分解后的单词 w_i 在序列中的出现次数称为单词 w_i 在复合词 c 中的出现次数,记为 $ON(c,w_i)$.例如,复合词“GoogleSearch”为单词“Google”的一次出现,且出现次数为 1,即

“GoogleSearch”$\in Occ$(“*Google*”),且 ON(“*GoogleSearch*”,“*Google*”)=1.

若某个单词 w 为不可分单词,则它为其自身的一次出现.

定义 16(项频率). Web 服务抽象描述信息组成的文档 d 中,单词项 f_i 的频率 $tf(f_i,d)$为该项 f_i 在文档 d 中所有出现的频率之和;每次出现的频率为该次出现的引用次数与该项在此次出现中出现次数的积,记为

$$tf(f_i,d)=\sum_{e\in Occ(f_i)} R(e)\times ON(e,f_i).$$

标准化为

$$tf'(f_i,d)=(0.5+0.5\times tf(f_i,d))/\max_{f_j\in d}(tf(f_j,d)).$$

该项频率的计算方法考虑了 WSDL 文档中的参引特征,即被引用次数越多的元素,其频率越高.这与通常文本向量空间模型中词频的计算方法是不同的.

定义 17(倒排文档频率). 每个项 f_i 在 Web 服务描述文档集合 D 中的倒排文档频率定义为

$$idf(f_i)=\ln(N/m),$$

其中,N 表示 Web 服务描述文档集合 D 的势,m 表示包含有项 f_i 的 Web 服务描述文档的个数.

向量模型的关键是设置向量中每个特征项的权重,本文采用 $tf\times idf$ 来计算词项 f_i 的特征值 $w_i(f_i,d)$:

$$w_i(f_i,d)=tf'(f_i,d)\times idf(f_i)/\sqrt{w_1(f_1,d)^2+\ldots+w_n(f_n,d)^2},$$

其中,n 为向量空间的维度,即文档 d 中出现的词的个数.

3 基于多向量模型的 Web 服务发现

3.1 Web服务发现原型系统

本文基于多向量模型的 Web 服务发现系统框架如图 3 所示,主要由 Web 服务与查询向量的生成和基于多向量的相似度计算两部分组成.Web 服务与查询向量的生成包括解析 WSDL 文档(解析用户查询请求)、词项识别、去根、停用词过滤等操作.首先,通过 WSDL4J(http://sourceforge.net/projects/wsdl4j)对 WSDL 文档进行解析,提取文档中各个组件的名称、文本描述并对这些描述按照组件进行划分.采用复合词分解算法对复合词进行分解,分别建立二维词向量空间和三维词向量空间.同时利用 Castor(http://www.castor.org)对 WSDL 文档中使用的 XML Schema 进行分析,获取各个复合词的被引用次数并计算各个词项在各子向量中的特征值.

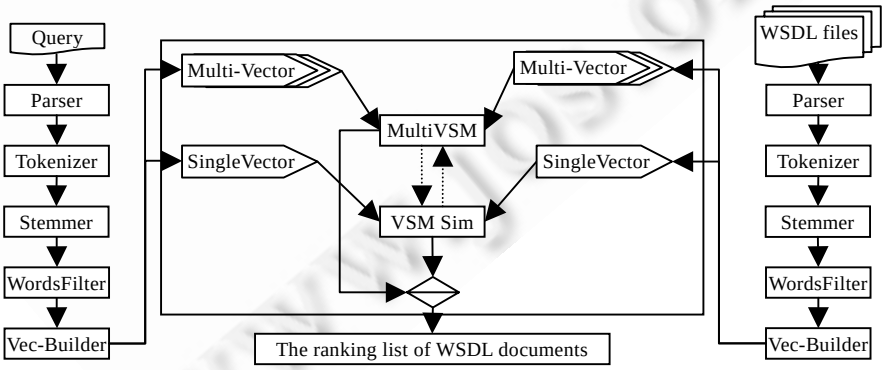


Fig.3 Multi-Vector based Web service discovery framework

图 3 基于多向量模型的 Web 服务发现框架

Web 服务描述的抽象信息大多是由描述文本和各个组件的名称组成,前者可以直接利用 GATE[9]中 ANNIE

的 Tokenizer 组件进行识别,后者利用基于词典的正逆向交叠匹配的分解方法得到.

最后,对各个单词进行去词根操作,并过滤掉停用词和一些 Web 服务描述文档中高频出现且对 Web 服务功能描述没有实质意义的词,例如 service,operation,request 等.系统建立了一个关于这些特殊词汇的词库,专门用于对去根后的非停用词进行过滤,以避免这些特殊词汇影响向量之间的相似度计算.

3.2 相似度计算方法

用户查询请求可以根据用户喜好提供不同程度的信息.若用户的查询请求是一段自然语言,则采用基于单向量表示模型的匹配方法来计算各个 Web 服务与查询向量的相似度;若用户的查询请求能够区分哪些属于操作信息,哪些属于参数信息,则可以采用多向量模型的匹配来计算各 Web 服务与查询向量的相似度.

两个 Web 服务 WS_1,WS_2 的多向量模型分别表示为 $WS_m^1=\langle Vector_1^1,...,Vector_m^1\rangle$ 与 $WS_m^2=\langle Vector_1^2,...,Vector_m^2\rangle$,它们的相似度计算函数定义为

$$Sim(WS_1,WS_2)=\sum_{i=1}^{m}\lambda_i\times Cos(Vector_i^1,Vector_i^2),$$

其中,$0\leqslant\lambda_i\leqslant 1,\sum_{i=1}^{m}\lambda_i=1$,$m$ 为多维向量的维度.每一维子向量的相似度通过经典的余弦相似度度量来计算,向量 $V_1=\{(f_{1i},w_{1i})\}$ 与向量 $V_2=\{(f_{2i},w_{2i})\}(i\in\{1,\ldots,n\})$ 的夹角余弦定义为

$$Cos(V_1,V_2)=\sum_{i=1}^{n}w_{1i}\times w_{2i}.$$

各子向量的维度取决于该子向量空间中单词项集合的个数.

4 实验与分析

4.1 测试集

目前,可用于 Web 服务发现的公开测试集主要有:OWL-TC(http://www.semwebcentral.org/frs/frs/?group_id=89),SWS Discovery Data Set 1.0(https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Projects/xmedia/dl-tree.htm)和 SAWSDL-TC1(http://projects.semwebcentral.org/frs/?group_id=156&release_id=331).OWL-TC 是基于 OWLS 1.1 标准的语义 Web 服务集合,最新版本 3.0 的数据集个数已达到 1 007 个,查询 29 个.SWS Discovery Data Set 1.0 将语义服务存储在基于描述逻辑的知识库中,并包含 96 个复杂概念用于描述 Web 服务.这两种数据集中的 Web 服务除了简单的文本描述外,不包含任何 WSDL 文档中的数据类型等信息.SAWSDL-TC1 是在 OWL-TC 的基础上采用半自动的方法构建的基于 SAWSDL 的语义服务发现测试集.它含有 894 个采用 SAWSDL(针对 WSDL 1.1)描述的语义 Web 服务.由于是人工构建的测试集,其中每个服务都只具有一个接口,每个接口也只有一个操作,大多数的数据类型定义就是简单的一系列数据元素的列举.因此,它不具有现实 Web 服务描述文档的特征.本文的主要工作是为现有互联网上的 Web 服务发现提供有效的发现方法,因此这些数据集均不适用于验证本文的研究工作,我们将自己构建基于 WSDL 文档的发现测试集.

4.1.1 Web 服务集合

实验采用的 Web 服务集合 QWS-wsdls 来源于加拿大 Guelph 大学 Eyhab AI-Masri 等人收集的 WSDL 文档集.该集合包含 2 417 个 WSDL 文档,且每个 Web 服务都被标注有相应的服务质量信息,用于研究基于服务质量的发现技术.该数据集在很大程度上反映了当前互联网上的 Web 服务现状,具体参见文献[1].

通过分析整理各 WSDL 文档,删除一些可能会影响实验结果客观性的 Web 服务,我们的实验数据集大小为 1 576 个.删除的 WSDL 文档主要包括:(1) WSDL4J 不能成功解析的 WSDL 文档(大约 295 个);(2) 具有相同内容的 WSDL 文档(大约 440 个);(3) 引用了外部 XML Schema 类型系统的 WSDL 文档.

4.1.2 查询构造与相关集合

目前,大部分的信息检索系统采用 TREC 的测试集构造方法——Cranfield 范型来构造其测试集.该类方法

包含多个假设,最重要的一个假设是相关集合判断是完全的.即对于每一个请求,文档集合中的所有相关文档都要被正确标注.当文档集合的势较大时,获取完全的相关性判断是不太可能的.因此,我们只对一部分 WSDL 文档进行相关性判断,构成一个具有不完全相关性判断的测试集.构建步骤如下:

(1) 确定需要考虑的查询领域.该测试集中的查询主要涉及与天气和地理位置信息有关的领域.通过与领域最可能相关的关键词,选取出可能属于该领域的 Web 服务 137 个,组成候选集合;

(2) 根据人们在各个领域的感兴趣话题,定义了 12 个查询,如查询某一地区的天气情况;

(3) 人工对各个查询的候选集合中的 Web 服务进行相关性判断.

除去不存在相关文档的 4 个查询,还剩 8 个查询.各个查询的相关集合数量情况见表 1.可以看出,由于 Web 服务数据集分布的不均衡性,各个查询的相关文档集合的大小不一.

Table 1 The size of relevant documents for each query

表 1 查询相关文档集合大小

QueryID	1	2	3	4	5	6	7	8
RelevanetSets	7	24	14	2	2	5	5	2

4.1.3 检索模型

为了评价基于多向量模型的检索性能,我们实现了如图 4 所示的 5 种检索模型.图 4 中横坐标表示多向量模型中子向量空间的划分粒度,纵坐标表示是否考虑 WSDL 文档中元素的参引特征,即考虑参引特征时,元素的出现与否取决于该元素是否被某个操作作为参数引用,元素在该 WSDL 文档中的频率取决于该元素被引用的次数.图 4 中,SingleFlat 表示单向量模型,向量空间中每个词项的频率取决于该项在文档中出现的次数.该模型类似于通用文本检索的单向量空间模型.SingleRef 也表示单向量模型,但是属于数据类型定义信息的词项的频率取决于该数据类型被参引的次数.TwoFlat 与 TwoRef 都表示二维向量空间模型,对 WSDL 文档中的操作信息和参数信息分别建立子向量空间,在操作信息子向量空间中词项的频率取决于文档中出现的频率. TwoFlat 中参数子向量空间中词项的频率取决于其在文档中出现的频率;而 TwoRef 中的参数子向量空间中词项的频率则取决于其被操作参引的次数.ThreeRef 表示三维向量空间模型,即分别对操作信息、输入参数信息和输出参数信息建立子向量空间,在输入输出参数子向量空间中词项的频率计算与 TwoRef 相同.由于在三元模型中,若不考虑元素的参引消息,则很难区分哪些元素为输入参数,哪些元素为输出参数,因此,本文没有实现 ThreeFlat 检索模型.

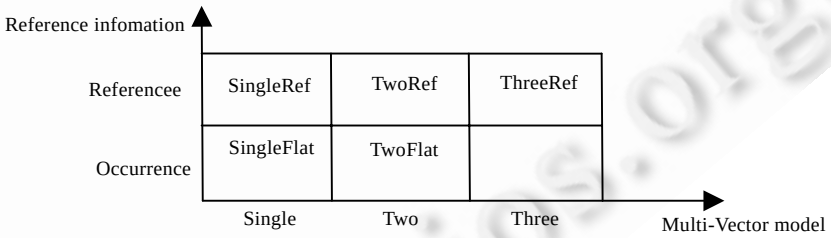


Fig.4 Features based Web service discovery models

图 4 基于特征的 Web 服务发现模型

4.1.4 评价标准

本文的相关集合判断是在一定的候选集合中的判断,对于非候选集合中的元素未进行判断.在判断过的集合中,我们能够保证结果的正确性,但不能保证未判断的数据集都与查询请求不相关.本文的相关性判断在整个 Web 服务集合中存在不完全性(incomplete).因此,我们必须采用不完全相关性判断的检索评价指标.本文采用 inferredAP[10]标准来衡量检索的性能.

inferredAP 采用不同于准确率、召回率、P@10 等标准的评价依据,inferredAP 是对平均准确率的一个近似

估计,它并不直接考虑返回的相关结果的准确率和召回率,而是通过人工判断过的标准答案(相关以及不相关的文档)在返回结果中相对位置的关系,即相关的文档排在不相关的文档之前的程度来估算其准确率.inferredAP 值为平均准确率的期望值,即在每一个相关文档处的期望准确率的平均值,计算公式如下:

$$inferredAP=\frac{1}{R}\sum_{i\in n}E[precision\ at\ rank\ k_i],$$

其中,R 为相关集合的个数,i 为检索返回列表中判断为相关的第 i 个文档,k_i 为第 i 个相关文档在返回列表中的位置,n 为所有已判断相关集合的大小.

每个相关文档处的期望准确率为

$$E[precision\ \text{at}\ rank\ k]=\frac{1}{k}\cdot 1+\frac{(k-1)}{k}\cdot\frac{r+\varepsilon}{r+n+2\varepsilon},$$

其中,k 为该相关文档在返回列表中的位置,r 为排在该相关文档之前的相关文档个数,n 为排在该相关文档之前的不相关文档个数.ε为避免除数为 0 的平滑因子,本实验对该变量的取值与其在 TREC 评测中的默认取值相同,为 0.000 01.从上式可以看出,对于某个查询的返回列表,若排在相关文档前面的不相关文档越少且相关文档在返回列表中的位置越靠前,则 inferrredAP 值越高,检索效果越好.

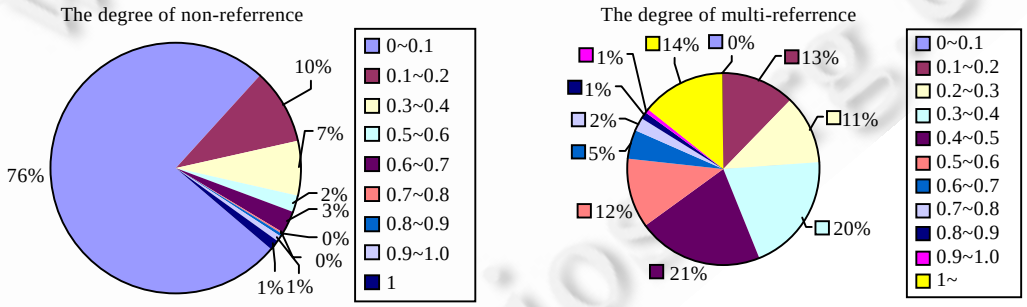
4.2 参引特征对发现的影响

4.2.1 参引特征分析

首先,定量分析测试集中元素的参引情况.通过解析测试集中各 WSDL 文档,计算出每个 Web 服务中各个元素 e 被引用的次数 $R(e)$.

定义 18(未被引用比率). Web 服务中未被该服务的操作所引用的元素占该服务中所有元素的比重定义为 $NonRef(w)=\#\{e|e\in E\wedge R(e)=0\}/\#E$,$\#E$ 表示 XML Schema 定义中所有元素集合 E 的势.

统计结果显示,实验数据集中有 693 个(约占 43.94%)Web 服务存在元素未被引用的情况,其中未被引用元素比率的分布如图 5(a)所示.可以看出,存在未被引用元素的 Web 服务中,约占 76%的 Web 服务元素未被引用率低于 10%,大约 10%的 Web 服务的 $NonRef(w)$值在 0.1~0.2 之间,甚至存在一小部分(约 1%)Web 服务中的元素全部未被引用.通过分析这些服务的 WSDL 文档发现,在这些服务描述文档中,元素定义没有被任何消息定义所参引.



(a) Distribution of services with non-reference elements
(a) 存在未被引用元素的服务分布

(b) Distribution of services with multi-reference elements
(b) 存在多次参引元素的服务分布

Fig.5 Referred percentages of elements

图 5 元素被引用情况图

定义 19(多次参引程度). 若 Web 服务中的元素被多个操作所参引,其被多次参引的程度定义为

$$RefDegree(w)=\sum_{e\in E\wedge R(e)>1}(R(e)-1)\Big/\#E.$$

$\#E$ 表示 XML Schema 定义中所有元素集合 E 的势.

通过分析可以发现,大约 63.35%的 Web 服务存在单个元素被多次引用的情况,即满足 *RefDegree*(*w*)>0 的 Web 服务.图 5(b)展示了 WSDL 文档中元素被多次参引的情况.从图 5(b)可以看出,大约 34%的 Web 服务中,元素的平均被参引次数超过 1(*RefDegree*(*w*)≥0.5).

通过上述对元素参引情况的分析发现,一部分 Web 服务描述文档中的元素未被引用,一部分 Web 服务描述文档中的某些元素被多次引用.这说明在真实 Web 服务描述文档,并不是任何描述信息都对描述 Web 服务的功能的描述有用,不同的元素对描述 Web 服务功能的贡献不相同.因此,在对 Web 服务描述信息进行建模时,考虑参引结构特征是非常必要的.同时,测试集中不包括含有参引外部 XML Schema 文档的 Web 服务.这类服务参引的元素大多是外部 XML Schema 文档中的一部分元素,若加上这部分描述文档,Web 服务参引特征将更加明显.

4.2.2 参引特征对发现的影响

表 2(a)展现了一元表示模型中,参引特征对检索性能的影响.从表 2(a)可以看出:查询请求 6 在不考虑参引结构时,检索的 inferredAP 值比考虑参引结构时的 inferredAP 值高;查询请求 1~查询请求 5 和查询请求 7 中,考虑参引结构时的 inferredAP 值要比不考虑参引结构的 inferredAP 值高;查询请求 8 中,两种检索模型的 inferredAP 值相当.造成这种现象的原因可能在于,多次被参引的元素的频率比操作描述的频率高,因而其特征的权重过大,使得元素信息比操作描述信息对检索性能的影响更大.这与 Web 服务描述的结构特征是相违背的,因此我们在考虑参引结构时,一般不使用一元表示模型,这会使得检索性能在很大程度上取决于元素信息的匹配.

因此,在考察参引结构对 Web 服务发现性能的影响时,我们通常使用二元表示模型.表 2(b)展示了二元表示模型中,元素子向量的特征权重计算在考虑参引结构和不考虑参引结构时对 Web 服务性能的影响.可以看出:考虑参引结构的二元表示模型上,除查询 8 外,各个查询的 inferredAP 值都略高于不考虑参引结构的情况;查询 8 在两种情况下的 inferredAP 值相等.总体来看,在二元检索模型下使用参引特征作为计算特征权重的方法,能够使得检索性能在置信度为 95%以上具有显著提高.

Table 2 Retrieval performance based on reference features

表 2 参引特征对检索性能的影响

(a) The inferredAP of single vector model

(a) 单向量检索模型的 inferredAP 值

	Query1	Query2	Query3	Query4	Query5	Query6	Query7	Query8
SingleFlat	0.142 9	0.015 7	0.139 4	0.045 5	0.033 2	0.246 3	0.181 9	0.625
SingleRef	0.209 3	0.017 6	0.140 1	0.054 4	0.034 8	0.216 1	0.270 3	0.625

(b) The inferredAP of two-vector model

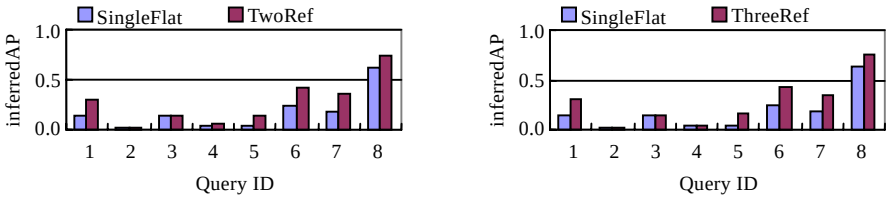
(b) 二维向量表示模型的 inferredAP 值

	Query1	Query2	Query3	Query4	Query5	Query6	Query7	Query8
TwoFlat	0.280 6	0.024 5	0.141 8	0.048 1	0.114 6	0.421 3	0.349 9	0.75
TwoRef	0.304 1	0.025 4	0.142 0	0.054 4	0.146 7	0.421 7	0.354 6	0.75

4.3 结构特征对发现的影响

第 4.2 节讨论了参引特征对 Web 服务语义表示的影响,本节主要讨论在考虑参引结构特征的同时,结合 Web 服务结构特征的 Web 服务检索性能.图 6(a)展示了单向量表示模型和二维向量表示模型对 Web 服务发现性能的影响,除在查询 3 上二维向量模型的检索性能略高于单向量模型外,二维向量模型的检索模型的性能明显高于单向量模型.总体来看,二维向量模型的检索性能比单向量模型的检索性能在置信度 95%以上具有显著提高.

图 6(b)展示了单向量表示模型和三维向量表示模型对 Web 服务发现性能的影响,除在查询 4 上三维向量的检索性能略高于单向量表示模型外,三维向量模型的检索性能明显高于单向量模型.总体来看,三维向量模型的检索性能比单向量的检索性能置信度在 95%以上具有显著提高.



(a) Retrival performance of two-vector model
(a) 二维向量模型检索性能

(b) Retrieval performance of three-vector model
(b) 三维向量模型检索性能

Fig.6 Retrieval performance based on the structure features
图 6 结构特征对 Web 服务检索性能影响

表 3 展示了二维向量模型和三维向量模型在本文提出的测试集上的检索性能比较.查询 1、查询 4、查询 7 在二维向量模型下的 inferredAP 值略高于三维向量模型下的 inferredAP 值;查询 2、查询 3、查询 5、查询 6 在二维模型下的 inferredAP 值略低于三维模型下的 inferredAP 值;查询 8 在两种检索模型下的 inferredAP 值相等.除查询 5 在三维向量模型下 inferredAP 值比二维向量下 inferredAP 值高 0.0158 外,其余查询在两种检索模型上产生的 inferredAP 值的误差不超过 0.01.

通过分析发现,多维向量表示模型总是比单向量表示模型更能清楚地表达 Web 服务的语义,因而检索性能有显著提高.是使用二维向量模型还是三维向量模型,我们需要根据用户请求的特定类型来进行判断.若用户请求是二维向量模型,则使用二维检索模型;若用户请求是三维向量模型,则可使用三维检索模型.

Table 3 Comparison results of performance between two- and three-vector model
表 3 二/三维向量模型检索性能比较结果

	Query1	Query2	Query3	Query4	Query5	Query6	Query7	Query8
TwoRef	0.304 1	0.025 4	0.142 0	0.054 4	0.146 7	0.421 7	0.354 6	0.75
ThreeRef	0.303 6	0.026 5	0.143 3	0.049 5	0.162 5	0.424 7	0.348 5	0.75

5 相关工作

Web 服务发现作为服务计算的一个关键问题,已被众多国内外研究者高度关注.该问题的研究主要分为两个方面:关于发现时效的研究和关于发现性能的研究.前者主要研究如何快速地发现 Web 服务,后者主要研究如何准确地发现相关的 Web 服务.为了快速地发现 Web 服务,大部分方法[11,12]采用缓存机制将发现产生的中间结果保存起来,后续发现工作将利用这些发现结果,以避免重新计算.Stollberg 等人[12]采用缓存机制将一些先验知识和之前的发现结果保存起来,以减小后续发现操作的搜索空间,节省计算代价.

以提高性能为目标的 Web 服务发现,旨在准确地发现 Web 服务.根据发现的依据不同,除了少数基于过程模型的 Web 服务发现方法外[13,14],该类研究工作又可以分为两大类:基于非功能属性的发现和基于功能的 Web 服务发现[15].基于非功能属性的发现工作主要为 Web 服务建立合理的非功能属性模型,再根据各项非功能属性评价指标选取最优的 Web 服务.Ran[16]提出了基于 Qos 的发现模型.AI-Masri[17]等人提出了一种 Web 服务质量的评价模型,并提出了基于该评价模型 Web 服务发现方法.他们标注了大规模真实的 WSDL 文档集合的服务质量,为基于 QoS 的 Web 服务发现工作提供了评价语料.文献[18]则根据反馈的信任形式和决策机制来选取 Web 服务.文献[19]则提出了一种基于情境和推理规则的 Web 服务发现方法.

基于功能的发现主要根据 Web 服务的功能描述发现满足用户需求的 Web 服务.当前的主要方法有基于语法的方法、基于语义 Web 服务的方法和混合的方法.文献[15]对这些方法进行了详细且全面的综述.总的来说,基于语法的发现方法计算开销小,准确率相对较低;基于语义 Web 服务的方法计算开销大但准确率较高.很多研究工作者结合了两种方法,使其计算开销相对较低,准确率相对较高.OWLS-MX[20],WSMO-MX[21]和 SAWSDL-MX[22]就是混合方法的典型代表.Keifer 等人[23]提出了 iSPARQL 查询语言,基于语义的 Web 服务发现过程同时

支持基于各种语法相似度的计算.

在互联网技术飞速发展的同时,Web服务的数量也在迅速增长.同时,语义Web技术尚不成熟,现阶段发布的Web服务大都没有进行语义标注.在这种形势下,如何利用有效的基于语法的方法来快速地发现Web服务已成为迫切需要解决的问题.文献[6]提出了基于向量空间模型的Web服务搜索引擎.文献[7]的工作主要集中在研究采用何种相似度度量方法和何种排序算法对于Web服务发现有较好的效果.Woogle[4]作为一个专业的Web服务搜索引擎,除支持基于关键字的简单搜索外,还支持基于相似度的搜索.该方法主要是采用精化的(refinement)聚合聚类将Web服务的参数集合聚类到语义概念上,然后通过比较结果概念计算相似度.该方法的搜索粒度是操作级的,即相似度计算的对象是操作对,本文的方法则是基于服务级的相似度计算.何玲娟等人[5]提出了一种基于WSDL的改进的操作相似度度量方法(MOSM),该方法通过计算操作的无序标签树的编辑距离来度量操作之间的相似度.本文的方法在利用结构特征的同时,也利用了参引特征,并将这两种特征融合到多向量模型中,因而计算量相对较小.

当前基于语法的发现方法大多是利用了Web服务的文本描述以及服务参数名称等语法信息,没有深入研究如何利用Web服务的本质特征以提高Web服务的检索性能.本文的工作在分析了大量Web服务描述文档特征的基础上,提出了基于结构特征和参引特征的高效的Web服务发现方法.该方法旨在建立能够体现Web服务描述文档本质特征的Web服务模型,尽可能地提高基于这种模型的Web服务发现准确率,以弥补传统的基于语法Web服务发现方法的不足.

6 结 论

本文针对当前互联网上大规模Web服务描述文档的有效检索问题,通过深入分析Web服务描述文档特征,提出了一种新的融合结构特征和参引特征的Web服务多向量表示模型.其中,结构特征体现在多向量空间模型的划分上,参引特征体现在子向量模型中的特征权重的计算方法上.提出了基于多向量模型的Web服务发现方法.此外,构建了一个真实WSDL文档集合的Web服务发现测试集,该测试集的构造与完善,为Web服务发现缺乏测试集提供了新的、有益的思路与探索.最后的实验结果表明,融合了结构与参引特征的多向量模型的Web服务发现比基于单向量模型的Web服务发现具有更好的检索性能.

References:

[1] AI-Masri E, Mahmoud QH. Investigating Web services on the World Wide Web. In: Huai JP, ed. Proc. of the 17th Int'l World Wide Web Conf. New York: ACM, 2008. 795−804. [doi: 10.1145/1367497.1367605]

[2] Wei DP, Wang T, Wang J, Chen YD. Extracting semantic constraint from description text for semantic Web service discovery. In: Sheth AP, ed. Proc. of the 7th Int'l Semantic Web Conf. Heidelberg: Springer-Verlag, 2008. 146−161. [doi: 10.1007/978-3-540-88564-1_10]

[3] Hao YN, Zhang YC. Web services discovery based on schema matching. In: Dobbie G, ed. Proc. of the 30th Australasian Computer Science Conf. (ACSC). Ballarat: ACS, 2007. 107−113.

[4] Dong X, Halevy A, Madhavan J, Nemes E, Zhang J. Similarity search for Web services. In: Nascimento MA, ed. Proc. of the 30th VLDB Conf. Toronto: Morgan Kaufmann Publishers, 2004. 372−383.

[5] He LJ, Liu LC, Wu C. A modified operation similarity measure method based on WSDL description. Chinese Journal of Computers, 2008,31(8):1331−1339 (in Chinese with English abstract).

[6] Platzer C, Dustdar S. A vector space search engine for Web services. In: Proc. of the 3rd IEEE European Conf. on Web Services. America: IEEE Computer Society, 2005. 62−71. [doi: 10.1109/ECOWS.2005.5]

[7] Ma JG, Zhang YC, He J. Efficiently finding Web services using a clustering semantic approach. In: Sheng QZ, ed. Proc. of the Int'l Workshop on Context Enabled Source and Service Selection Integration and Adaption (CSSSIA). New York: ACM, 2008. [doi: 10.1145/1361482.1361487]

[8] Christensen E, Curbera F, Meredith G, Veerawarana S. Web services description language (WSDL) 1.1. http://www.w3.org/wsdl

[9] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics. 2002. 168−175.

[10] Yilmaz E, Aslam JA. Estimating average precision when judgments are incomplete. Knowledge and Information Systems, 2008,16(2):173−211. [doi: 10.1007/s10115-007-0101-7]

[11] Ren KJ, Liu X, Chen JJ, Xiao N, Song JQ, Zhang WM. A QSQL-based efficient planning algorithm for fully-automated service composition in dynamic service environments. In: Proc. of the IEEE SCC. Washington: IEEE Computer Society, 2008. 301−308. [doi: 10.1109/SCC.2008.26]

[12] Stollberg M, Hepp M, Hoffmann J. A caching mechanism for semantic Web service discovery. In: Aberer K, Cudré-Mauroux P, Choi KS, Noy N, Allemang D, Lee KI, Nixon L, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber, eds. Proc. of the 6th Int'l and 2nd Asian Semantic Web Conf. Berlin: Springer-Verlag, 2007. 480−493.

[13] Sun P, Jiang CJ. Using service clustering to facilitate process-oriented semantic Web service discovery. Chinese Journal of Computers, 2008,31(8):1340−1353 (in Chinese with English abstract).

[14] Vaculín R, Sycara K. Towards automatic mediation of OWL-S process models. In: Proc. of the IEEE Int'l Conf. on Web Services (ICWS). IEEE Computer Society, 2007. 1032−1039.

[15] Klusch M. Semantic Web service coordination. In: Schumacher M, Helin H, Scheuldt H, eds. Proc. of the CASCOM—Intelligent Service Coordination in the Semantic Web. Birkhaeuser: Springer-Verlag, 2008. 59−104.

[16] Ran S. A model for Web services discovery with QoS. ACM SIGecom Exchanges, 2003,4(1):1−10. [doi: 10.1145/844357.844360]

[17] AI-Masri E, Mahmoud QH. Discovering the best Web service. In: Williamson CL, *et al*., eds. Proc. of the 16th Int'l Conf. on World Wide Web. American: ACM, 2007. 1257−1258. [doi: 10.1145/1242572.1242795]

[18] Wang Y, Lv J, Xu F, Zhang L. An iternetware-software-architecture-oriented trust-driven mechanism for selecting services. Journal of Software, 2008,19(6):1350−1362 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/1350.htm [doi: 10.3724/SP.J.1001.2008.01350]

[19] Feng ZW, He KQ, Li B, Gong P, He YF, Liu W. A method for semantic Web service discovery based on context inference. Chinese Journal of Computers, 2008,31(8):1354−1363 (in Chinese with English abstract).

[20] Klusch M, Fries B, Khalid M, Sycara K. OWLS-MX: Hybrid OWL-S service matchmaking. In: Proc. of the 1st Int'l AAAI Fall Symp. on Agents and the Semantic Web. 2005.

[21] Kaufer F, Klusch M. WSMO-MX: A logic programming based hybrid service matchmaker. In: Proc. of the European Conf. on Web Services. Washington: IEEE Computer Society, 2006. 161−170.

[22] Klusch M, Kapahnke P, Zinnikus I. Hybrid adaptive Web service selection with SAWSDL-MX and WSDL-Analyzer. In: Aroyo L, *et al*., eds. LNCS 5554, Berlin: Springer-Verlag, 2009. 550−564. [doi: 10.1007/978-3-642-02121-3_41]

[23] Kiefer C, Bernstein A. The creation and evaluation of iSPARQL strategies for matchmaking. In: Bechhofer S, Hauswirth M, Hoffmann J, Koubarakis M, eds. Proc. of the ESWC 2008. Heidelberg: Springer-Verlag, 2008. 463−477.

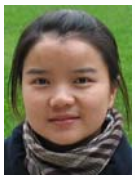
附中文参考文献:

[5] 何玲娟,刘连臣,吴澄.一种改进的基于 WSDL 描述的操作相似性度量方法.计算机学报,2008,31(8):1331−1339.

[13] 孙萍,蒋昌俊.利用服务聚类优化面向过程模型的语义 Web 服务发现.计算机学报,2008,31(8):1340−1353.

[18] 王远,吕建,徐峰,张林.一种面向网构软件体系结构的信任驱动服务选取机制.软件学报,2008,19(6):1350−1362. http://www.jos.org.cn/1000-9825/19/1350.htm [doi: 10.3724/SP.J.1001.2008.01350]

[19] 冯在文,何克清,李兵,龚平,何扬帆,刘玮.一种基于情景推理的语义 Web 服务发现方法.计算机学报,2008,31(8):1354−1363.



魏登萍(1981－),女,四川雅安人,博士生,CCF 会员,主要研究领域为语义 Web,服务发现,信息检索.

王戟(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为高可信软件技术,软件方法学,软件工程.

王挺(1970－),男,博士,教授,博士生导师,主要研究领域为语义 Web,语义 Web 服务,信息抽取,信息检索.