软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(6):1335−1349 [doi: 10.3724/SP.J.1001.2012.04088]

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

基于网络信息搜索的 Web Service 文本描述信息扩充方法*

王立杰[1,2], 李 萌[1,2], 蔡斯博[1,2], 李 戈[1,2+], 谢 冰[1,2], 杨芙清[1,2]

[1](北京大学 信息科学技术学院 软件研究所,北京 100871)

[2](高可信软件技术教育部重点实验室,北京 100871)

Internet Information Search Based Approach to Enriching Textual Descriptions for Public Web Services

WANG Li-Jie[1,2], LI Meng[1,2], CAI Si-Bo[1,2], LI Ge[1,2+], XIE Bing[1,2], YANG Fu-Qing[1,2]

[1](Software Institute, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

[2](Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing 100871, China)

+ Corresponding author: E-mail: lige@sei.pku.edu.cn

Wang LJ, Li M, Cai SB, Li G, Xie B, Yang FQ. Internet information search based approach to enriching textual descriptions for public Web services. *Journal of Software*, 2012,23(6):1335−1349. http://www.jos.org.cn/1000-9825/4088.htm

Abstract: With the development of Web services technologies, more and more public Web services have been published on the Internet. During the searching and utilizing of these public services, services' textual descriptions (such as introduction and user manual), which are generally expressed in natural language, provide great help for service consumers to locate, understand, and utilize proper Web services. Existing methods for services discovery usually try to obtain such descriptions only from services' WSDL files. However, according to this investigation, lots of Web services do not contain enough textual descriptions in their WSDL files. This paper proposes an approach to enriching textual descriptions for public Web services on the Internet using the information sources outside of WSDL files. Given a Web service, the study collects related Web pages containing its features from the Internet. Then, the enriched descriptions for the service are identified from the Web pages using information retrieval technologies. Experiments conducted on real data indicate that our approach can enrich descriptions for about half of the public services on the Internet effectively. The collected data is publicly available on the Internet.

Key words: Web service; service textual description; service discovery; Internet information searching

摘 要: 随着 Web 服务技术的不断成熟和发展,互联网上出现了大量的公共 Web 服务.在使用 Web 服务开发软件系统的过程中,其文本描述信息(例如简介和使用说明等)可以帮助服务消费者直观有效地识别和理解 Web 服务并加以利用.已有的研究工作大多关注于从 Web 服务的 WSDL 文件中获取此类信息进行 Web 服务的发现或检索,调研发现,互联网上大部分 Web 服务的 WSDL 文件中普遍缺少甚至没有此类信息.为此,提出一种基于网络信息搜索的从 WSDL 文件之外的信息源为 Web 服务扩充文本描述信息的方法.从互联网上收集包含目标 Web 服务特征标识的相关网页,基于从网页中抽取出的信息片段,利用信息检索技术计算信息片段与目标 Web 服务的相关度,并选取相

* 基金项目: 国家自然科学基金(60803010); 国家高技术研究发展计划(863)(2007AA010301)
收稿时间: 2011-03-11; 定稿时间: 2011-07-04

关度较高的文本片段为Web服务扩充文本描述信息.基于互联网上的真实数据进行的实验,其结果表明,可为约51%的互联网上的 Web 服务获取到相关网页,并为这些 Web 服务中约 88%扩充文本描述信息.收集到的 Web 服务及其文本描述信息数据均已公开发布.

关键词: Web 服务;服务描述信息;服务发现;互联网信息搜索

中图法分类号: TP311 **文献标识码:** A

近年来,随着 Web 服务技术的不断成熟和发展,互联网上出现了大量的公共 Web 服务[1−3],这些服务以SOAP[4]等平台无关的通信协议为服务消费者(service consumer)提供功能,构成了未来大规模面向服务计算的基础.Web 服务由按照 Web 服务描述语言(Web service description language,简称 WSDL[5])编写的 WSDL 文件进行描述,服务提供者(service provider)除了可以对 Web 服务的接口进行结构化描述,还可以在 WSDL 文件的 wsdl:documentation 元素中提供关于 Web 服务的介绍、说明等描述信息,该信息可以是任意形式的文本内容,一般使用自然语言表达(本文使用“文本描述信息”指代此类信息).文本描述信息类似于代码中的注释,提供了关于 Web 服务的说明和解释,帮助服务消费者方便、直观地了解 Web 服务.借助于文本描述信息,服务消费者可以更容易地检索、了解 Web 服务,进而决定该服务能否满足需求,并最终学习使用该服务,并将其集成到软件系统中.Web 服务领域目前的工作也广泛使用分析和索引 WSDL 文件中文本描述信息的方法,进行 Web 服务的发现和检索[2,3,6−11].然而相关研究表明,目前互联网上只有很少一部分 Web 服务的 WSDL 文件中包含详细的文本描述信息[1−3,12].本文对目前互联网上Web服务的文本描述信息状况进行的统计分析发现,其WSDL文件中此类信息的平均长度仅为 36.4 个单词,长度少于平均值的比例高达 85%,有超过 67%的 Web 服务的 WSDL 文件中不包含任何文本描述信息(详细调研结果请参见本文第 1.2 节).如此短的描述信息往往难以准确说明一个 Web 服务,这在一定程度上限制了 Web 服务的发现和检索,不利于 Web 服务的共享和使用.

本文对互联网上 Web 服务的调查分析发现,对于一个 Web 服务来说,除了其 WSDL 文件以外,互联网上许多网页中存在与其相关的信息,例如包含该 Web 服务发布信息的网页、包含该 Web 服务使用方法讨论的网页、对该 Web 服务进行宣传介绍的网页等,本文将此类网页称作 Web 服务的相关网页.这些网页中,关于 Web 服务使用方法的讨论、宣传材料等信息有助于服务消费者了解该服务,尤其是那些 WSDL 文件中不包含详细文本描述信息的 Web 服务.尽管互联网上的网页中包含很多关于 Web 服务的文本描述信息,服务消费者手工从互联网上检索这些信息比较费时费力,而且如果服务消费者花费了很长时间从互联网上了解了关于某个 Web 服务的信息后发现该服务并不能满足自己的需求,也会造成精力的浪费.

基于上述考虑,本文提出了一种利用分布在互联网上的 Web 服务的相关网页为 Web 服务扩充文本描述信息的方法.首先,从互联网上收集包含目标 Web 服务标识的相关网页;经过解析,将网页内容分割成若干个文本片段;利用信息检索技术识别与目标 Web 服务相关度较高的文本片段,用其内容为目标 Web 服务扩充文本描述信息.相关度由两方面因素决定:文本片段与目标Web服务特征的相似程度;文本片段和目标Web服务在原始网页中出现位置的相对距离.本文基于互联网上的真实数据进行了实验研究,实验结果表明,本方法可以比较有效地为 Web 服务扩充文本描述信息.

我们曾在文献[13]中针对此问题展开了研究,相比原有工作,本文对互联网上 Web 服务描述信息的现状进行了更全面的调研,基于调研结果对网页内容切分、相关度计算以及结果选取等方法进行了改进,并对方法作了更加全面的验证和讨论.

本文的主要工作包括:

(1) 对互联网上的公共 Web 服务的文本描述信息状况进行了统计分析;
(2) 提出了一种为互联网上 Web 服务收集相关网页的方法;利用对网页内容的分析,提出一种基于信息检索技术的 Web 服务文本描述信息扩充方法;
(3) 本方法为从互联网获取的样本数据中 51%的 Web 服务收集到了相关网页,并为有相关网页的 Web 服务中的 88%扩充了文本描述信息,同时对数据的有效性进行了验证;

(4) 本文收集到的 Web 服务数据以及扩充的文本描述信息已发布到国家高技术研究发展计划(863)课题“可信的国家软件资源共享与协同生产环境”支持的“可信软件资源库(http://tsr.trustie.net)”中.

本文第 1 节介绍本文工作的主要背景.第 2 节详细描述本文方法.第 3 节对本文方法进行验证并对实验结果进行介绍和分析.第 4 节介绍本文的相关工作.第 5 节对本文工作中涉及的问题和未来工作展开讨论.第 6 节对本文进行总结.

1 工作背景

本节通过一个实例说明本文的工作动机,并对目前互联网上的 Web 服务进行了相关调研,调研内容包括:

1) 目前互联网上 Web 服务的 WSDL 文件中,文本描述信息的整体状况;
2) Web 服务相关网页的数量分布状况.

1.1 动机例证

以 Web 服务‘SendSMS(http://www.aswinanand.com/sendsms.php?wsdl)’为例,其 WSDL 文件中的文本描述信息为:‘Sends the same SMS to multiple phone numbers. Give your 10 digit phone number for user ID. Separate each phone number with a semicolon (‘;’).’.该信息表明,这个 Web 服务可以用来向移动电话发送短消息,并且需要 10 位的手机号作为输入.然而,服务消费者在决定是否使用该服务时往往还有其他方面的考虑,例如:‘这个 Web 服务是不是免费的?’’、‘这个 Web 服务是否可以用来向世界各地的移动电话发送短消息?’.该服务的 WSDL 文件中所提供的文本描述信息显然无法提供这些问题的解答.实际上,在互联网上可以找到很多关于该服务的信息,例如,从一个用来进一步介绍该服务的网页(http://www.aswinanand.com/2008/07/send-free-sms-Web-service/)中有:‘Send Free SMS—Web Service’,‘send SMS alerts to various Indian mobile numbers, an account at www.way2sms.com would be used for this service’.从一些技术论坛**上还可以找到很多讨论信息:‘First signup to www.way2sms.com then use the following Webservice in your Web application http://www.aswinana-nd.com/sendsms.php?wsdl’,‘Add this Link as Web Reference: http://www.aswinanand.com/ sendsms.php And Use the Way2SMs Site Account to send sms’.利用这些信息可以很容易地了解到关于该服务进一步的信息,例如,可以免费向印度移动手机上发送短信、使用该服务之前需要先在网站 www.way2sms.com 上注册.

从该实例可以看出,尽管有些 Web 服务的 WSDL 文件提供了关于该服务的文本描述信息,但这些信息往往不足以帮助服务消费者很好地了解并使用该服务,那些 WSDL 文件中没有任何文本描述信息的 Web 服务则更难以理解.在互联网上存在的很多关于 Web 服务的信息可以为服务消费者提供帮助.因此,可以提供一种有效的方法利用这些信息为 Web 服务扩充文本描述信息.

1.2 现状统计

为了对互联网上的 Web 服务进行调研,需要构造 Web 服务样本集.本文从一些优秀的 Web 服务门户站点上收集 Web 服务,包括 XMethods.net,Seekda[7]和 WebServiceList.com 等,这些站点也被一些相关工作采纳为 Web 服务数据来源[1,2,14].本文共收集了超过 13 000 个 Web 服务,为了保证调研结果能够反映目前互联网上公开的 Web 服务的真实状况,并保证样本集中的 Web 服务可用,且对服务消费者来说具有使用价值,本文对收集到的 Web 服务进行了验证,并将未通过验证的服务排除出样本集.验证内容包括 WSDL 文件是否符合 WSDL 规范以及 Web 服务是否可以访问等.经过验证和过滤,一共保留了 10 756 个 Web 服务,这些 Web 服务也将作为本文实验部分的数据.本文对这些 Web 服务的 WSDL 文件中的文本描述信息长度分布状况以及互联网上相关网页的数量分布状况进行了统计分析.

** http://forums.asp.net/p/1458598/3348650.aspx#3348650,
http://www.dotnetfunda.com/forums/thread369-how-to-send-sms-in-net-csharp.aspx

1.2.1 WSDL 文件中文本描述信息长度分布状况

Web 服务的 WSDL 文件中的文本描述信息是指包含在 wsdl:documentation 元素中的文本内容.本文从 WSDL 文件中抽取此类信息,并对信息长度(单词的个数)进行统计分析.

统计分析结果如图 1 所示.经统计,所有 Web 服务 WSDL 文件中的文本描述信息的平均长度仅为 36.4,其中,大约有 85.73%的 Web 服务的 WSDL 文件中的此类信息长度低于该平均值.另外,有 7 210 个 Web 服务的 WSDL 文件中没有任何文本描述信息,比例超过 67%.文本描述信息长度大于 30 的 Web 服务所占的比例仅为 15.5%.

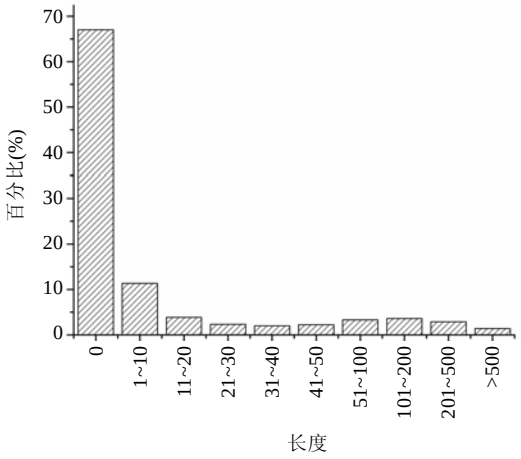


Fig.1 Distribution of length of textual descriptions in WSDL files

图 1 WSDL 文件中描述信息的长度分布状况

当然,如果文本描述信息偏少的 Web 服务中大部分可用性较差,那么对这些 Web 服务扩充描述信息便没有太大意义.经过统计发现,在没有提供任何文本描述信息的 7 210 个 Web 服务中,有 92.6%的 Web 服务的可访问率***高于 90%;在文本描述信息长度少于 10 的 8 430 个 Web 服务中,有超过 91.4%的 Web 服务的可访问率高于 90%.这表明,在统计的 Web 服务中有相当数量可访问率较高的 Web 服务没有或者只提供了非常少的文本描述信息,为这些 Web 服务扩充文本描述信息将有助于这些服务的使用.

虽然信息长度不能作为衡量文本描述信息状况的唯一标准,但仅仅几十个单词长度的信息即使非常短小精辟也难以解释清楚一个 Web 服务,Web 服务消费者仅仅利用 WSDL 文件中的文本描述信息难以很好地理解相应的 Web 服务,所以需要从 WSDL 文件以外的信息源为 Web 服务扩充更详细的文本描述信息.

1.2.2 互联网上 Web 服务的相关网页的数量分布状况

本文提出了一种为互联网上的 Web 服务收集相关网页的方法,基于该方法为样本集中的 Web 服务收集其相关网页.本节对相关网页数量分布状况进行统计分析,并基于该结果阐述本文工作的可行性.相关网页的具体收集方法将在第 2.2 节的方法说明中详细介绍.

样本集中所有 Web 服务的相关网页的数目分布状况统计结果见表 1.本文能够为超过 50%的 Web 服务收集到相关网页,超过 21.5%的 Web 服务至少有 2 个相关网页.另外,有 11%的 Web 服务有 5 个或 5 个以上相关网页.统计结果表明,本文能够为约半数的 Web 服务有效收集到相关网页.此类网页中往往包含关于相关联 Web 服务的文本描述信息,以此为信息源为 Web 服务扩充文本描述信息有利于提高服务消费者检索、选择、使用互联网上 Web 服务时的工作效率.

*** 本文对样本数据集中的 Web 服务的可访问率进行了 1 个月的监控统计,每 2 小时访问一次 Web 服务,可访问率=可访问次数/总访问次数.

Table 1 Distribution of the number of related Web pages for each Web service

表 1 Web 服务的相关网页数目的分布状况

相关网页数目	比例(%)
≥10	4.86
9	0.95
8	1.22
7	1.46
6	1.43
5	1.09
4	1.58
3	2.57
2	6.37
1	29.09
Total	**50.61**

2 Web 服务文本描述信息扩充方法

图 2 展示了本文方法的整体结构,包括 5 个步骤:1) 从 WSDL 文件中提取 Web 服务特征信息;2) 利用特征信息从互联网上收集 Web 服务的相关网页;3) 解析相关网页,将网页内容划分为若干文本片段;4) 为文本片段和目标Web服务建立特征向量;5) 计算每个文本片段与目标Web服务的相关度,选取相关度最高的若干个网页文本信息片段作为目标 Web 服务的文本描述信息.

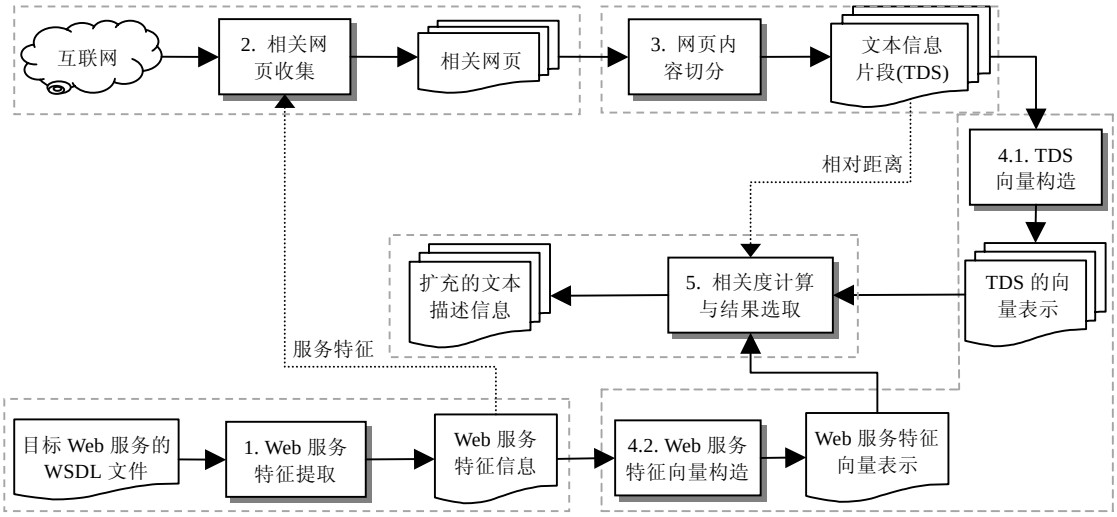


Fig.2 Overview of our approach

图 2 方法概览

2.1 Web服务特征提取

基于 WSDL 规范解析 Web 服务的 WSDL 文件,提取 Web 服务特征信息,包括服务名称、操作名称、消息名称、参数名、Endpoint URL 以及 WSDL 文件中的文本描述信息(如果存在的话).Web 服务特征信息的用途包括标识 Web 服务、为收集 Web 服务相关网页提供依据、用来计算 Web 服务与网页中文本信息片段的相关度.

2.2 相关网页收集

经分析发现,Web 服务的相关网页通常具有一个特征:此类网页中一般包含所关联 Web 服务的访问地址,即 WSDL 文件的 URL 或者 Endpoint(Endpoint 声明了一个 Web 服务对外提供服务的位置,访问该位置即可访问到该 Web 服务[5])的 URL,如图 3 所示.本文将 Web 服务的访问地址作为其在网页中出现的标识,也以此作为收集相关网页的依据.该标识可能在网页中以超链接方式出现,也可能是普通文本.本文采用如下方法从互联网

上收集符合该特征的相关网页:1) 使用反向链接(backlink)搜索功能收集包含超链接方式 Web 服务标识的相关网页.反向链接搜索指利用互联网上网页的链接关系获取指向特定目标网页或实体的网页,例如,网页 *a* 中存在指向网页 *b* 的超链接,那么网页 *a* 是网页 *b* 的反向链接[15];2) 使用文本搜索功能收集包含文本形式 Web 服务标识的相关网页,即通过检测网页中的文本内容寻找包含特定文本内容的网页.在方法实现中,本文利用通用搜索引擎 Google 的搜索接口收集相关网页.即针对上述两种情况分别构造并向 Google 提交适当的查询条件,利用 Google 获得满足需要的页面:

1) Google 的反向链接搜索功能查询格式为 link:〈target_URL〉,其中,搜索限定词'link'表示搜索包含指向特定目标实体的网页,目标实体由〈target_URL〉指定.为了收集符合第 1 种情况的 Web 服务相关网页,构造的查询条件为 link:〈WSDL URL〉和 link:〈Endpoint URL〉;
2) 为了收集符合第 2 种情况的 Web 服务相关网页,构造的查询条件为"〈WSDL URL〉"和"〈Endpoint URL〉",即分别直接将 Web 服务的 WSDL 文件 URL 和 Endpoint URL 作为查询条件进行检索,查询条件两侧的引号则用于要求搜索引擎将查询条件作为一个短语而不需要切词.

Fig.3 An example of content of Web service's related Web pages

图 3 Web 服务相关网页内容示例

本文利用以上方法为样本集中的每个 Web 服务从 Google 获取检索结果,再从每个查询条件返回的结果列表中选取前 *N* 个结果构造候选页面集(基于对实验数据的观察和分析,方法实现中 *N* 取值 20),最后对候选页面集进行过滤以获得 Web 服务的相关网页:1) 如果相同网页在多个查询条件的结果中出现,需要进行结果去重;2) 检索结果中可能包含 Web 服务的 Endpoint 页面,但由于 Endpoint 页面一般是在发布 Web 服务时由一些标准模板自动生成,其内容一般与 Web 服务 WSDL 文件的内容重复,不具备参考价值,因此需要去除这类页面.

2.3 网页内容切分

此步骤对相关网页的文本内容进行切分,网页的文本内容会被切分成若干文本信息片段(textual description segment,以下简称 TDS).网页内容切分方法以伪代码描述如图 4 所示.

方法说明:

1) 为网页文档建立对应的 DOM 树并删除 DOM 树中的属性节点,如图 4 中第 2 行、第 3 行所示.DOM (document object model,文档对象模型)提供了一个平台和语言无关的接口,允许程序可以动态访问或修改文档的内容或结构[16].HTML DOM 定义了访问 HTML 文档的标准方法,它将一个 HTML 文档表示为一个带有元素(HTML 标签)、属性(标签的属性内容)和文本(标签中的文本内容)这 3 种节点的树状结构.删除属性节点是为了清理数据,以便于后续步骤的处理.但是,超链接标签〈A〉的 HREF 属性内容(即超链接目标地址)是目标 Web 服务时需要进行相应标记以记录 Web 服务在网页中出现的位置;

2) 从 DOM 树叶节点开始遍历,逐层提取 TDS,如图 4 中第 4 行~第 12 行所示.如果一个叶节点的文本内容长度(包含的单词数)过短(即小于给定阈值 threshold,本文方法中 threshold 设定为 3),则将该节点直接删除(如图 4 中第 5 行所示),因为此类节点内容往往无意义(例如导航条的内容等).如果遍历到的叶节点所对应的标签是一个段落标签[****],则将该节点的文本内容提取为一个 TDS;否则,将该节点的文本内容保留至其父节点中,即将其内容与其父节点内容合并,并将该节点从 DOM 树中删除.

Input: D_i, HTML document of related page p_i for target service s; *F_TAG*, set of fragment tags; *threshold*, threshold for data filtering;
Output: T_i, set of TDS extracted from D_i.
Begin
1: $T_i \leftarrow \Phi$;
2: $DOMTree_i \leftarrow buildDOMTree(D_i)$; //Build DOM tree for HTML document
3: *removeAttributeNodes*($DOMTree_i$); //Remove attribute nodes from DOM tree
4: **repeat do** from leaf nodes *leaf* //Traverse DOM tree from leaf nodes to root
5: **if** *contentLength*(*leaf*)>*threshold* **then do** /*Only the nodes whose content is longer than given threshold are considered as possible candidates*/
6: **if** *node.tag* is in *F_TAG* **then do** //If the encountered tag is a fragment tag
7: *add*(*leaf*,T_i); //Identify this node as a TDS
8: **else do**
9: reserve *leaf.content* to *leaf.parent*; /*Reserve the content of this tag to its parent tag*/
10: **end if**
11: **end if**
12: *remove*(*leaf*,$DOMTree_i$); //Remove this node from DOM tree
13: **until** $DOMTree_i$ is empty
14: **return** T_i;
End

Fig.4 Algorithm for Web page segmentation

图 4 网页内容切分方法

一般来说,在网页中文本信息片段与目标实体的距离越近,意味着两者之间的关系越紧密[17].在识别出一个 TDS 时,方法会记录其与目标 Web 服务在原始网页中的距离(即间隔文本内容的单词个数),本文根据 WSDL URL 或 Endpoint URL 来判定目标 Web 服务的出现位置.如果目标 Web 服务在网页中出现多次,则以较短的距离为准.这个距离将被用于计算 TDS 与目标 Web 服务的相关度.

2.4 特征向量构造

经过了之前的 3 步(Web 服务特征提取、相关网页收集、网页内容切分),我们得到了从 WSDL 文件中抽取到的 Web 服务特征信息以及从相关网页中获取到的文本描述片段 TDS 等文本文档.本文利用信息检索技术来处理这些文本文档,用于计算每个 TDS 与 Web 服务特征信息的相似性,该相似性是衡量 TDS 与该服务相关度的一方面因素.

向量空间模型(vector space model,简称 VSM)[18]是一个被广泛使用的信息检索领域的建模技术.在向量空间模型中,每个文本文档被表示为一个 n 维的向量$\langle w_{d,1},w_{d,2},\ldots,w_{d,n}\rangle$,每一维对应一个词汇(indexing term),词汇从文本文档中获得.其中,n 是在所有文档中出现的不同词汇的个数,$w_{d,i}$($1\leqslant i\leqslant n$)表示第 i 个词汇(用 $term_i$ 表示)在文档 d 中的权重.$w_{d,i}$ 通常使用公式(1)计算:

$$w_{d,i}=\frac{tf_{d,i}}{L_d}\times idf_i \tag{1}$$

其中,$tf_{d,i}$ 是指 $term_i$ 在文档 d 中出现的次数,L_d 是文档 d 的长度,idf_i 由如下公式定义:

$$idf_i=\log(N/df_i) \tag{2}$$

其中,N 是所有文档的数目(N 在本文的值是“TDS 的数目加 1”),df_i 是在文档集合中包含 $term_i$ 的文档的数目.

**** 段落标签(fragment tag)是指通常用于在 HTML 文档中展示文本段落内容的标签.本文定义的段落标签为:TABLE,TR,TD,P,DIV,PRE.

向量空间模型将每个文本文档表示为一个向量,然后通过计算两个向量之间夹角的余弦来计算两个文本文档的相似度.对于两个文本文档 d_1 和 d_2,假设其对应的向量表示分别为 $V_1=\langle w_{1,1},w_{1,2},\ldots,w_{1,n}\rangle$ 和 $V_2=\langle w_{2,1},w_{2,2},\ldots,w_{2,n}\rangle$,那么两个文档的相似度一般利用如下公式计算:

$$sim(V_1,V_2)=\frac{\sum_{i=1}^{n}(w_{1i}\times w_{2i})}{\sqrt{\sum_{i=1}^{n}(w_{1i})^2\times\sum_{i=1}^{n}(w_{2i})^2}} \tag{3}$$

本文对获得的 Web 服务特征信息以及 TDS 根据向量空间模型建立相应的向量表示.如前所述,建立空间向量模型时的每一维对应一个词汇,因此,首先需要从文档中获取词汇.经过观察和分析获取到的文本文档,本文采用如下方法从中获取词汇:

1) 根据大写字母以及诸如‘_’,‘-’的分隔符对组合词进行分词;
2) 提取词干并去除停用词*****;
3) 由于在 Web 服务领域经常出现的一些通用词汇(例如 service,soap,response,request,set,get 等)对于 Web 服务来说区分度并不大,而且容易引入干扰,这些词汇也被当作停用词去除.举例来说,‘GetProductSellingPages’经过分词、取词干、去除停用词后,我们可以得到‘product’,‘sell’,‘page’这 3 个词汇.

由于 Web 服务特征信息和 TDS 内容性质的差异,本文采用不同方法为这两种文本文档构造向量.

2.4.1 TDS 向量构造

对于每个文本信息片段 TDS,本文直接采用公式(1)和公式(2)为其构造向量.

2.4.2 Web 服务特征信息向量构造

在公式(1)、公式(2)的定义中,词汇在文档中出现的频率是计算其权重的一个重要因素,但对于 Web 服务特征信息来说,除了词频以外,词汇在 WSDL 文件中出现的位置也很重要.例如,由于出现在 Web 服务名称处的词汇的区分度相对来说要比其他位置的词汇大,这些词汇的重要性一般要比其他词汇高.因此,在构造 Web 服务特征信息的向量时,应该根据词汇出现位置的不同给予不同的权重.本文对公式(1)做了适当改进得到公式(4),并利用公式(2)和公式(4)为 Web 服务的特征信息构造向量.公式(4)中的 $num_{k,d}$ 是词汇 $term_k$ 在文本信息段 d 中出现的次数,n 是不同词汇的个数,$pos(term_i,j)$ 是根据词汇 $term_i$ 在特征信息中第 j 次出现的位置给予的相应权重,服务名称中词汇的权重是 ω_{s_name},操作或者消息名称中词汇的权重是 ω_{om_name},出现在其他位置的文本内容中词汇的权重是 ω_{other}.本方法实现中,ω_{s_name},ω_{om_name} 和 ω_{other} 分别取值为 2.0,1.5,1.2.

$$\left.\begin{aligned} w_{d,i}&=\frac{\sum_{j=1}^{num_{i,d}}pos(term_i,j)}{\sum_{k=1}^{n}\left(\sum_{j=1}^{num_{k,d}}pos(term_k,j\right)}\times idf_i\\ pos(term_i,j)&=\begin{cases}\omega_{s\_name}, & \text{if the } j\text{th occurence of } term_i \text{ is in service name}\\ \omega_{om\_name}, & \text{if the } j\text{th occurence of } term_i \text{ is in operation or message name}\\ \omega_{other}, & \text{other areas}\end{cases}\end{aligned}\right\} \tag{4}$$

2.5 相关度计算和结果选取

本文从两个方面计算 TDS 与目标 Web 服务的相关度:(1) TDS 与 Web 服务特征信息的文本相似度;(2) TDS 与目标 Web 服务在原始网页中出现位置的相对距离.一般来说,一段对 Web 服务进行描述或讨论的文本信息会与该 Web 服务共享某些词汇或术语,而且一般会提到一些该服务的其他信息,例如服务的名称、操作等.基于该

***** 停用词(stop word)是指那些在文本中普遍出现但没有含义的功能词,例如‘the’,‘a’,‘an’等限定词以及‘over’,‘under’等介词等.取词干(word stemming)将一个词由于变形(如复数、时态)或者派生(如动词后加后缀-ation 得到对应名词)产生的不同形式简化为一个共同的词干.去除停用词和取词干是信息检索领域常用的数据预处理工作之一.

假设,本文将文本相似度作为衡量两者相关度的标准之一.另外,在网页中,文本片段与目标实体的距离越近,意味着该文本片段与目标实体存在关联的可能性越大.抽取目标实体附近的文本片段对目标进行标注,也是被普遍采用的方法[17].

本文利用公式(3)计算 Web 服务和 TDS 的文本相似度,两者的距离使用步骤 3)获取的数据.计算相关度的方法见公式(5),其中,s 表示目标 Web 服务,t 表示 TDS,$sim(s,t)$表示两者之间的文本相似度,$dis(s,t)$表示两者在网页中的距离,$relevance(s,t)$表示两者的相关度.

$$relevance(s,t)=\frac{sim(s,t)}{\log_{10}(dis(s,t)+1)+1} \tag{5}$$

在计算了每个 TDS 与目标 Web 服务的相关度之后,需要从很多候选的 TDS 中选取若干个作为结果返回.从排序列表中选取若干个结果的方法大致有两种:一种是阈值法(threshold based method),即选择分值大于某个给定阈值的元素,阈值需要事先指定;另一种则是固定个数法(cut point N),即选择排序结果中位置最靠前的 N 个元素,选取的数目则需要事先指定.考虑到本文工作的应用场景,本文采用一种混合式的结果选取方法.首先设置一个相关度阈值α以及结果数目限制 N,按照相关度由高到低遍历目标 Web 服务的所有 TDS,如果一个 TDS 的相关度高于α,则将其加入结果列表,直至结果数目达到限制 N 或者遍历完所有 TDS.通过对实验数据的分析和经验判断,α设置为 0.5/($\log_{10}$(10000+1)+1=0.099999,即文本相似度为 0.5,距离为 10 000 情况下的相关度,结果数目限制 N 设置为 10.

3 方法评估

本文使用从互联网上获取的真实数据对本文方法的有效性进行验证和评估.在第 1.2 节介绍的收集到的 10 756 个 Web 服务中,本文为 5 444 个 Web 服务收集到了相关网页,本节将对这 5 444 个 Web 服务扩充文本描述信息的效果进行分析和评估.

3.1 扩充效果概览

本文单独对这 5 444 个 Web 服务的 WSDL 文件中文本描述信息的长度进行了统计,平均长度 58.35 个单词,57.5%的 Web 服务的 WSDL 文件中没有任何文本描述信息,有 75%的 Web 服务的 WSDL 文件中的文本描述信息少于 20 个单词.本文方法共为其中 88.5%的 Web 服务(4 815 个)扩充了文本描述信息.在这 4 815 个 Web 服务中,有 2 754 个(57.2%)Web 服务的 WSDL 文件中没有任何文本描述信息,WSDL 文件中文本描述信息不足 10 个单词的有 70%.扩充信息之后,Web 服务的文本描述信息平均增加了 4.5 段,Web 服务的文本描述信息的平均长度从 58.35 增长到了 249.5,有文本描述信息的 Web 服务比例由 42.5%增加至 93.2%,长度大于 30 的比例由 25%升至 84%.扩充前和扩充后的文本描述信息长度分布状况如图 5 所示.

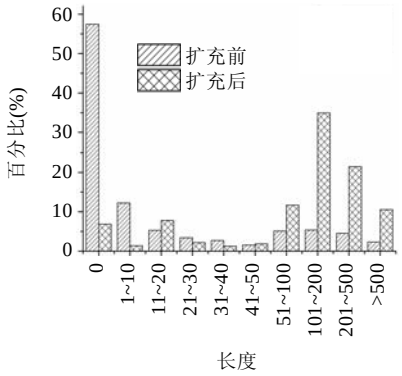


Fig.5 Distribution of length of textual descriptions before/after enriching

图 5 扩充前/后文本描述信息长度分布状况

3.2 效果示例

本节以第 1.1 节提到的 Web 服务 SendSMS 为例,介绍本文方法为该服务扩充文本描述信息的效果.本文方法为该服务扩充的文本描述信息见表 2.不难发现,服务消费者使用这些信息可以了解更多与该 Web 服务相关的信息,从而帮助其更好地做出是否要对该服务做进一步尝试的决定.然而从表 2 中也可以看出,本文方法扩充的文本描述信息中存在一些冗余信息,本文未来将对如何进一步精化信息展开深入研究.

Table 2 Enriched textual descriptions for SendSMS service

表 2 SendSMS 服务的扩充文本描述信息

扩充的文本描述信息 序号|内容		所属的相关网页 网页标题,原始 URL
1	Hello, I am trying to send a text message to my phone in C#. I have added a Webservice reference to my solution. It is described as: vijju311985 RPC SMS services you can send Sms for free the WSDL is at: http://www.aswinanand.com/sendsms.php?wsdl	topic:it http://pubsub.com/Getting-parameters-topic-it-Assembly-cclUQw8QRKeS
2	Hi friend u know how to use Web service .... U know means use this wsdl .... U can send sms. http://www.aswinanand.com/sendsms.php?wsdl but one thing is u must have a account in way2sms.com then only u able to send sms to other mobile. Try it. All the best.	Sending SMS http://social.msdn.microsoft.com/Forums/en/vssmartdevicesvbcs/thread/881300ba-9e18-402b-816e-ab7b5537f830
3	There are few articles there which was working earlier like this: http://www.codeproject.com/KB/aspnet/SendingSMS.aspx http://www.codeproject.com/KB/cpp/SendSmsThroughWS.aspx But now they are just a piece of crap. If your working area is India, you can use way2sms to do this. Try http://www.aswinanand.com/sendsms.php?wsdl But the only thing is that you need to have an account in Way2sms. www.abhisheksur.com	How to send SMS using asp.net application ...—ASP.NET Forum http://www.dotnetfunda.com/forums/thread1612-how-to-send-sms-using-aspnet-application.aspx
4	Hi Aswin, actually i used ur Webservice url in .net, but now its not working (means, code is not showing any errors, it is executing, but sms is not going). 3 weeks before i have used this url: http://www.aswinanand.com/sendsms.php?wsdl	Free SMS Web service updated—2|Waves http://www.aswinanand.com/2009/12/free-sms-Web-service-updated-2
5	Example in first URL: http://coeservice.en.kku.ac.th:8080/TemperatureConvertor/TemperatureConvertorService?xsd=1.com this link is not working .... so dont try ... try for only working WSDL (means Webservice file). Try following Link, copy all the data from browser to notepad remove first line and “-” and space from each line which i have suggested earlier and then save it as WSDL file and try. http://www.aswinanand.com/sendsms.php?wsdl Hope this helps.	SAP community network forums: Client proxy through WSDL ... http://forums.sdn.sap.com/message.jspa?messageID=7526306
6	First signup to www.way2sms.com then use the following Webservice in your Web application http://www.aswinanand.com/sendsms.php?wsdl	How to send sms in .net c# ...—ASP.NET forum http://www.dotnetfunda.com/forums/thread369-how-to-send-sms-in-net-csharp.aspx
7	Ivan Krizsan wrote: Hi! I just tested to create a new Web service from your WSDL in NetBeans 6.8 and it works flawlessly. Note that the URL of the WSDL must be prefixed by “http://”, that is: http://www.aswinanand.com/sendsms.php?wsdl Best wishes!	Error while creating the proxy? (Web Services forum at JavaRanch) http://www.coderanch.com/t/490116/Web-Services/java/Error-while-creating-proxy

3.3 准确性评估

我们从北京大学软件工程实验室挑选了 10 位研究生作为评估者进行实验,他们均具有一年以上的 Web 服务使用经验.由于要判定一条描述信息是否与目标 Web 服务相关,需要评估者对目标 Web 服务有较深入的了解,考虑到工作量的因素,并为了保证实验的准确性,我们首先让每位评估者从具有扩充文本描述信息的 4 815 个 Web 服务中任意选取 20 个 Web 服务作为其评估对象,在选取评估对象时,我们要求以其能够较准确给出判定结果为原则,同时,我们允许他们使用其他工具(例如 Google)作为参考.然后由评估者分别对其选择的 20 个 Web 服务的扩充文本描述信息是否与对应的 Web 服务相关做出判定,相关程度是 0~10 之间的整数,0 表示完全不相关,10 表示非常相关.我们要求评估者从文本描述信息能否为其提供关于了解或使用该 Web 服务有价值信息的角度进行判定,一般从如下几个方面考虑:介绍了该 Web 服务的基本信息,例如服务由谁提供、是否收费、QoS

状况等;讨论了该 Web 服务的某些功能,例如提供了何种功能、可以在什么范围内使用等;讨论了如何使用该 Web 服务或者在交流该 Web 服务的使用经验或评价,例如是否需要特定账户、使用过程中出现的一些问题等.

本文使用公式(6)计算准确率,公式中 $S=S_1\cup S_2\cup S_3\cup...\cup S_{10}$,其中,$S_i$(1<=$i$<=10)是第 i 位同学评估的 Web 服务集合,S 是所有被评估的 Web 服务集合,T_s 表示 Web 服务 s 的扩充文本描述信息集合,$score(s,t)$表示评估者对 s 的扩充文本描述信息 t 的打分(如果有多位用户对该条目打过分,则取所有用户打分的平均值作为最终得分).

$$Precision=\frac{\sum_{s\in S}\left(\sum_{t\in T_s}score(s,t)\right)}{10\times\sum_{s\in S}|T_s|}\times 100\% \tag{6}$$

最终,10 位评估者一共对 174 个 Web 服务的扩充文本描述信息的相关性进行了评估,这 174 个 Web 服务一共有 538 条扩充的文本描述信息.评估者判定的扩充文本描述信息与 Web 服务的相关性分布状况见表 3,总体准确率 Precision 为 71%.另外,我们分别计算了每位评估者对所评估的 Web 服务的文本描述信息相关性打分的平均值,分布状况见表 4.

Table 3 Results of relevance evaluation

表 3 相关度评估结果

打分	比例(%)	打分	比例(%)
0	1.1	1	3.30
2	2.20	3	4.40
4	3.30	5	5.49
6	18.68	7	10.99
8	14.29	9	19.78
10	16.48		

Table 4 Average score of each assessor

表 4 评估者平均打分统计

Id	平均打分	Id	平均打分
1	7.0	2	8
3	6.7	4	7.1
5	7.9	6	8.0
7	7.6	8	6.9
9	5.8	10	6.8

评估扩充结果的相关性是一个相对主观的过程,这既受样本数据的影响,也受评估者自身的经验、偏好的影响.在实验过程中,我们尽量选择具有较丰富 Web 服务使用经验的人员作为评估者参与实验,并由评估者自己选择评估的 Web 服务对象,希望以此来减少主观性给实验带来的影响.当然,要更加充分地验证工作的效果,我们未来还需要开展更加细致的实验设计、分析工作,例如扩大实验规模、基于选定样本数据集进行分组实验、基于评估者的打分偏好对其打分结果进行归一化处理等.

4 相关工作

4.1 面向互联网的Web服务发现

在丰富的 Web 服务中发现合适的 Web 服务,是复用 Web 服务的基本前提.随着近些年 Web 服务技术的发展和应用,Web 服务发现在学术界和产业界得到了大量的关注.相对于特定上下文环境(例如组织内部)中的自动服务发现与绑定,本文主要关注于面向互联网的大规模 Web 服务发现.这其中也有一些工作从 WSDL 文件外部为 Web 服务收集文本描述信息,本文将在此与这些工作进行对比.

在传统的 Web 服务技术框架中,UDDI(universal discovery description integration) Registry 负责服务的发布与发现.但实践验证,统一的 UDDI Business Registry(UBRs)并不能很好地适用于面向互联网的大规模服务发布与发现[19−22],以 uddi.ibm.com 为代表的 UBRs 因为无法为复用者提供数量足够、具有质量保证的 Web 服务被陆续关闭[23].

除了 UDDI 注册中心以外,较早的一批提供 Web 服务发现功能的工作包括 bindingpoint,grandcentral, salcentral 和 Webservicelist.这些站点依靠服务提供者主动向其提交 Web 服务,利用服务发布时提供的描述信息,通过基于关键字的匹配进行服务检索[8].但目前这些站点要么已经无法访问,要么长期处于停止维护更新的状态.

已有的研究工作表明,目前互联网上的 Web 服务信息大量分散在各个普通的 Web 站点中[1−3,12].在面向互联网的 Web 服务发现的工作中,华盛顿大学开发的 Web 服务搜索引擎 Woogle[8]是较早的工作之一.Woogle 从特

定站点收集 Web 服务,通过对从 WSDL 文件中抽取到的接口的输入、输出信息的名称进行聚类,发现 Web 服务之间的相似关系.在对 WSDL 文件中文本描述信息检索的基础上,利用 Web 服务之间的相似关系改善查询的效果.维也纳技术大学的 Platzer 等人提出的 Web 服务搜索引擎 VSSE4WS(vector space search engine for Web service)[6]从 UDDI 站点中自动获取关于 Web 服务的相关信息(WSDL 文件以及发布 Web 服务时提供的信息),对获取的信息进行解析、切词和取词干等处理,并采用向量空间模型来支持用户查询.在 Fan 等人 2005 年针对互联网上 Web 服务的状况进行的统计分析工作中[12],他们从一些当时规模较大且具代表性的 Web 服务发布站点(如 bindingpoint.com, XMethods.net 等)收集 Web 服务数据.在对描述信息进行统计分析时,他们考虑了 WSDL 文件中的文本描述信息以及在发布站点上提供的信息.北京大学的 Li 等人在 2007 年利用通用搜索引擎和特定站点抓取相结合的方法,对互联网上 Web 服务的状况进行了一次更为全面的统计分析,在收集 Web 服务的描述信息时,他们也仅处理了 WSDL 文件中的文本描述信息[1].Eyhab 等人提出了一个面向大规模 Web 服务发现的爬虫引擎 WSCE(Web service crawler engine)[2,3],WSCE 使用 UDDI 注册库、Web 服务信息门户站点以及通用搜索引擎索引的数据作为数据来源,采集的信息包括 Web 服务的元信息(例如服务提供者、发布时间、发布时提供的描述信息)、WSDL 文件.WSCE 从 WSDL 文件中获取 Web 服务的描述信息,利用该信息建立 Web 服务信息库以进行服务发现.另外一个代表性的工作是 Seekda[7],Seekda 是一个针对互联网上 Web 服务的搜索引擎,它从整个互联网上爬取 Web 服务并提供检索,能够检索的信息包括从 WSDL 文件提取到的 Web 服务的信息(名称、方法列表等)、Web 服务的提供者的地理位置、Web 服务的可用率等 QoS 信息.ServiceFinder[9]是一个关于 Web 服务的综合信息整合网站,基于 Seekda 系统获取到的数据,ServiceFinder 对 Web 服务进行语义标注、自动分类,进而基于这些信息进行服务检索、推荐等工作.国内关于 Web 服务发现与检索方面的相关工作主要有天津大学的 ServiceNetwork 系统[24]和北京航空航天大学的 ServiceXchange 系统[10],基于收集到的 Web 服务数据(Web 服务的 WSDL 文件、提供者等信息),这两个系统都提供对互联网上公共 Web 服务的发现和检索等服务.以上所介绍的工作中,所处理的 Web 服务文本描述信息均来自于 Web 服务 WSDL 文件中的文本描述信息、或者 UDDI 注册中心和特定站点中 Web 服务发布者主动提交的信息,而且从这些站点中获取文本描述信息业均是基于网页结构已知的前提下实现的.

AbuJarour 等人也曾针对从互联网上获取 Web 服务描述信息的问题展开研究[11],在给定 Web 服务的前提下,他们对 Web 服务的提供者站点进行定点抓取,从服务的提供者站点内部获取 Web 服务的相关网页,然后利用启发式规则将 Web 服务 WSDL URL 所在的 HTML 节点的父节点和兄弟节点中的文本内容作为 Web 服务的文本描述信息.与该工作相比,本文从更广的范围内获取了更多的相关网页,例如技术论坛、博客中的网页,这些网页中的内容对 Web 服务消费者来说具有很高的参考价值;另外,本文在整个网页范围内,利用文本相似度和相对距离相结合的方法抽取 Web 服务文本描述信息,这可以获得更多、更丰富的结果.

通过与以上相关工作的对比可以看出,已有的工作仅关注于从 WSDL 文件、UDDI 站点、特定 Web 服务发布站点或者是 Web 服务提供者的站点等固定站点中获取 Web 服务的文本描述信息,这在很大程度上限制了其信息来源,进而影响了获取到的文本描述信息的分布范围和丰富程度.本文提出了一种从互联网上收集 Web 服务的各类相关网页,并利用从相关网页中获取到的信息为 Web 服务扩充描述信息的方法,从更广的范围内收集更丰富的信息.我们相信,本文工作可以很好地弥补已有的服务发现工作在获取 Web 服务描述信息方面的不足,而且也较容易和已有的工作结合.

4.2 语义Web服务

语义 Web 是基于资源描述框架(resource description framework,简称 RDF)和元数据(metadata)对 WWW 上数据的抽象表示,是本体论的具体表示和应用实例[19].对于 Web 服务技术而言,利用本体描述语言对 Web 服务进行描述,从而发现概念的语义信息以及概念之间潜在的关系,使得语义的匹配成为可能,可以实现自动化的 Web 服务发布、发现、绑定与切换等操作.比较著名的工作有基于本体 DAML 的非轻量级语言 DAML-S[25]和 W3C 组织制定的基于本体 OWL 的 Web 服务描述语言 OWL-S[26]等.较有代表性的 Web 服务语义匹配的工作有卡耐基梅隆大学的 Paolucci 等人的 Augment UDDI Registry 系统[27]等.语义 Web 服务的应用依赖于领域本体的

识别和制定,但由于缺乏普遍公认的本体,在很大程度上限制了语义 Web 服务的发展与普及.本文通过从互联网上获取与 Web 服务关联的网页并从中抽取信息为 Web 服务扩充描述信息,扩充的描述信息可以作为抽取领域本体的信息来源.

4.3 图片检索

从互联网上检索图片一直受到在学术界和产业界的关注[28],基于文本的图片检索(text based image retrieval)是目前图片检索采用的主要方法之一,其主要思想是,利用从包含目标图片的网页中抽取到的文本信息对图片进行标注,然后根据这些文本信息,利用基于文本的信息检索技术进行图片检索,网页中在图片周围的文本片段是图片描述信息的主要来源.

基于文本的图片检索为本文的工作提供了借鉴意义,但是两个工作具有如下差别:(1) 目标实体不同.本文工作的目标实体是 Web 服务,图片检索的目标实体是图片.Web 服务不同于图片是其有 WSDL 文件,利用 WSDL 文件可以获得关于 Web 服务的一些特征信息.对于这些特征信息,可以也应该进行一些特别的处理,以取得良好的效果;(2) 描述信息分布范围不同.Web 服务的描述信息可能分布于若干个相关网页中,为 Web 服务扩充描述信息,需要综合考虑如何从多个相关网页中抽取关于 Web 服务的描述信息.而基于文本的图片检索则多关注于从一个网页中获取目标实体的描述信息;(3) 除了判定文本信息与目标实体的相关度以外,本文还需要处理如何从互联网上获取 Web 服务的相关网页等问题.

5 问题讨论和未来工作

5.1 关于Web服务的文档记录

有人可能会认为,如果一个服务提供者想让他(她)的 Web 服务能够被很多人使用,那么他(她)肯定会主动在 Web 服务的 WSDL 文件中提供详细的说明信息.也就是说,如果一个 Web 服务的 WSDL 文件中没有详细的描述信息,则说明其提供者本身就没有期望该 Web 服务可以被很多人使用,那么为这些 Web 服务扩充描述信息则没有太大意义.然而,本文对互联网上 Web 服务的调查结果表明,大部分 Web 服务的 WSDL 文件中并不包含详细的描述信息,其中不乏提供有用功能且质量较高的 Web 服务.WSDL 中文本描述信息缺失的原因可归结为:首先,很多服务开发者习惯于在发布了 Web 服务之后,在另外一个位置提供关于这个 Web 服务详细的介绍说明,而只在 WSDL 文件中提供一些关于 Web 服务的简要说明和概要介绍;其次,很多 Web 服务开发和发布工具对 Web 服务文档化的自动支持不足;另外,WSDL 文件中包含的文本信息一般只是关于 Web 服务本身的功能性介绍内容,而关于 Web 服务使用方面的内容(例如使用过程中遇到问题的讨论、一些特殊情况的处理)多分布在一些论坛、教程等网页中.

5.2 关于相关度的计算方法

本文采用文本片段和 Web 服务特征信息之间的文本相似度作为计算两者相关度的一个因素,该方法在本质上更倾向于那些和 Web 服务的 WSDL 文件具有较多相同词汇的文本信息.但有些文本描述信息虽然在文本上和 Web 服务的特征信息重合不多,但在意义上却和 Web 服务比较相关,例如 Web 服务的某些功能性描述或者某些关于 Web 服务的 QoS 描述等.利用文本相似性判定相关度的方法在识别此类信息方面存在一些不足,为了改善这个问题,我们在构造 Web 服务特征信息的向量时,为 Web 服务名称、操作名称赋予了更高的权重,并在计算相关度时考虑了 Web 服务和文本片段的相对距离.在后续工作中,我们将对方法中各方面因素对方法效果的影响进行更加全面的分析和验证.另外,应用其他相关技术(例如自然语言分析、基于语义的相似度判定)来进一步改进识别那些和 Web 服务相关但文本相似度较低的文本描述信息,也是本文未来要开展的工作.

5.3 关于结果选取方法

从排序列表中选取若干个元素作为结果返回是一项很重要的工作,也是一项很难做到完美的工作.具体采用何种结果选取方法,往往依赖于具体的使用场景.本文中采用的是阈值和结果数目限制相结合的结果选取方

法,相关度阈值和结果数目限制的设置则是基于对实验数据的观察和经验判断.当然,在将本文方法所扩充的文本描述信息应用于具体场景时,应该对结果选取方法做适当调整.例如:将扩充信息用于改善 Web 服务检索效果的情况下,需要尽可能保证信息的准确度;在将扩充信息展示给用户以帮助用户快速了解 Web 服务基本情况的场景下,则可以先向用户展示一部分信息,然后由用户决定是否需要浏览更多的信息,即增量式的结果展示.在后续工作中,本文将对把扩充信息应用在不同场景下的结果选取方法进行研究和评估.

6 结束语

互联网上 Web 服务 WSDL 文件中文本描述信息普遍过少,不利于服务消费者发现、理解和使用 Web 服务.基于这一问题,本文提出了一种利用互联网信息检索技术为 Web 服务扩充文本描述信息的方法.基于互联网真实数据进行的实验表明,本文方法可为样本集中 51%的 Web 服务收集到相关网页,并为 45%的 Web 服务扩充了文本描述信息,准确率达到 71%.经过文本描述信息扩充以后,Web 服务的文本描述信息平均长度由 58.35 增至 249.5.本文未来将对扩充的文本描述信息在具体场景中的应用展开研究,例如,利用扩充后相对丰富的信息改善服务发现和选择的效果,从扩充的描述信息中抽取语义信息以支持语义 Web 服务等.

References:

[1] Li Y, Liu Y, Zhang LJ, Li G, Xie B, Sun JS. An exploratory study of Web services on the Internet. In: Proc. of the IEEE Int'l Conf. on Web Services. Salt Lake City: Institute of Electrical and Electronic Engineers, 2007. 380−387. [doi: 10.1109/ICWS.2007.37]

[2] Al-Masri E, Mahmoud QH. WSCE: A crawler engine for large-scale discovery of Web services. In: Proc. of the IEEE Int'l Conf. on Web Services. Salt Lake City: Institute of Electrical and Electronic Engineers, 2007. 1104−1111. [doi: 10.1109/ICWS.2007.197]

[3] Al-Masri E, Mahmoud QH. Investigating Web services on the World Wide Web. In: Proc. of the Int'l Conf. on World Wide Web. Beijing: Association for Computing Machinery, 2008. 795−804. [doi: 10.1145/1367497.1367605]

[4] W3C. SOAP specification. http://www.w3.org/TR/soap/

[5] Web services description language (WSDL) Version 2.0. http://www.w3.org/TR/wsdl20/

[6] Platzer C, Dustdar S. A vector space search engine for Web service. In: Proc. of the 3rd IEEE European Conf. on Web Services. Institute of Electrical and Electronic Engineers, 2005. 62−71 [doi: 10.1109/ECOWS.2005.5]

[7] Seekda Web service search engine. http://Webservices.seekda.com/

[8] Xin D, Alon H, Jayant M, Ema N. Similarity search for Web services. In: Proc. of the 30th Int'l Conf. on Very Large Data Bases. Toronto: Association for Computing Machinery, 2004. 372−383.

[9] ServiceFinder. http://www.cisco-servicefinder.com/

[10] ServiceXchange. http://www.trademarkia.com/servicexchange-75413403.html

[11] Mohammed A, Felix N, Mircea C. Collecting, annotating, and classifying public Web services. In: Proc. of the Confederated Int'l Conf. on the Move to Meaningful Internet Systems. Hersonissos: Springer-Verlag, 2010. 256−272.

[12] Fan J, Kambhampati S. A snapshot of public Web services. Newsletter ACM SIGMOD Record, 2005,34(1):24−32. [doi: 10.1145/1058150.1058156]

[13] Wang L, Liu F, Zhang LJ, Li G, Xie B. Enriching descriptions for public Web services using information captured from related Web pages on the Internet. In: Proc. of the 5th IEEE Int'l Symp. on Service Oriented System Engineering. Nanjing: Institute of Electrical and Electronic Engineers, 2010. 141−150 [doi: 10.1109/SOSE.2010.28]

[14] Vieira M, Antunes N, Madeira H. Using Web security scanners to detect vulnerabilities in Web services. In: Proc. of the IEEE/IFIP Int'l Conf. on Dependable Systems and Networks. Lisbon: Institute of Electrical and Electronic Engineers, 2009. 566−571 [doi: 10.1109/DSN.2009.5270294]

[15] Backlink. http://en.wikipedia.org/wiki/Backlink

[16] DOM. Document object model.

[17] Cai D, He XF, Li ZW, Ma WY, Wen JR. Hierarchical clustering of WWW image search results using visual, textual and link information. In: Proc. of the 12th Annual ACM Int'l Conf. on Multimedia. New York: Association for Computing Machinery, 2004. 952−959. [doi: 10.1145/1027527.1027747]

[18] Christopher DM, Prabhakar R, Hinrich S. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[19] Yue K, Wang XL, Zhou AY. Underlying techniques for Web services: A survey. Journal of Software, 2004,15(3):428−462 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/428.htm

[20] Hu JQ, Zou P, Wang HM, Zhou B. Research on Web service description language QWSDL and service matching model. Chinese Journal of Computers, 2005,28(4):505−513 (in Chinese with English abstract).

[21] Al-Masri E, Mahmoud OH. Discovering Web services in search engines. IEEE Internet Computing, 2008,12(3):74−77. [doi: 10.1109/MIC.2008.53]

[22] Du ZX, Huai JP. Research and implementation of an active distributed Web service registry. Journal of Software, 2006,17(3): 454−462 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/454.htm [doi: 10.1360/jos170454]

[23] Wu C, Chang E. AtomServ architecture: Towards Internet-scaled service publish, subscription, and discovery. In: Proc. of the IEEE Int'l Conf. on e-Business Engineering. Shanghai: Institute of Electrical and Electronic Engineers, 2006. 571−578 [doi: 10.1109/ICEBE.2006.29]

[24] Wang H, Feng ZY, Sui Y, Chen SZ. Service network: An infrastructure of Web services. In: Proc. of the IEEE Int'l Conf. on Intelligent Computing and Intelligent Systems. Shanghai: Institute of Electrical and Electronic Engineers, 2009. 303−308. [doi: 10.1109/ICICISYS.2009.5358185]

[25] Burstein MH, Hobbs JR, Lassila O, Martin D, McDermott DV, McIlraith SA, Narayanan S, Paolucci M, Payne T, Sycara K. DAML-S: Web service description for the semantic Web. In: Proc. of the Int'l Semantic Web Conf. Sardinia: Springer-Verlag, 2002. 348−363. [doi: 10.1007/3-540-48005-6_27]

[26] OWL-S: Semantic markup for Web services. http://www.w3.org/Submission/OWL-S/

[27] Paolucci M, Kawamura T, Payne TR, Sycara K. Semantic matching of Web services capabilities. In: Proc. of the Int'l Semantic Web Conf. Sardinia: Springer-Verlag, 2002. 333−347. [doi: 10.1007/3-540-48005-6_26]

[28] Ronny L, Aya S. PicASHOW: Pictorial authority search by hyperlinks on the Web. In: Proc. of the Int'l Conf. on World Wide Web. New York: Association for Computing Machinery, 2001. 438−448 [doi: 10.1145/371920.372098]

附中文参考文献:

[19] 岳昆,王晓玲,周傲英.Web 服务核心支撑技术:研究综述.软件学报,2004,15(3):428−462. http://www.jos.org.cn/1000-9825/15/428.htm

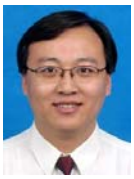
[20] 胡建强,邹鹏,王怀民,周斌.Web 服务描述语言 QWSDL 和服务匹配模型研究.计算机学报,2005,28(4):505−513.

[22] 杜宗霞,怀进鹏.主动分布式 Web 服务注册机制研究与实现.软件学报,2006,17(3):454−462. http://www.jos.org.cn/1000-9825/17/454.htm [doi: 10.1360/jos170454]

王立杰(1986－),男,吉林四平人,博士生,主要研究领域为软件复用,软件构件管理技术,Web 服务技术.

李萌(1988－),男,博士生,主要研究领域为软件复用,软件构件管理技术,Web 服务技术.

蔡斯博(1984－),男,博士生,主要研究领域为软件复用,软件构件管理技术.



李戈(1977－),男,博士,副教授,主要研究领域为软件工程,软件复用.

谢冰(1970－),男,博士,教授,博士生导师,主要研究领域为软件工程,计算机科学理论.

杨芙清(1932－),女,教授,博士生导师,主要研究领域为系统软件,软件工程,软件工业化生产技术和系统.