第32卷 第20期 计 算 机 工 程 2006年10月
Vol.32 № 20 Computer Engineering October 2006
·博士论文· 文章编号：1000—3428(2006)20—0033—02 文献标识码：A 中图分类号：TP391

基于领域的Web服务查找方法

许 斌
（清华大学计算机系，北京 100084）

摘 要：在构建面向服务的Web应用时，往往需要按照领域进行Web服务查找。现有的UDDI的Web服务查找方式是通过tModel分类信息和关键字匹配来进行的，不便于按照领域进行查找。通过直接在互联网上搜索WSDL文件，并利用支持向量机来构建基于领域的WSDL文件分类器，实现了按照领域进行Web服务查找的方法。实验证明该方法具有较高的精确度。

关键词：Web服务查找；领域；支持向量机；WSDL

Web Services Search Method Based on Domain

XU Bin
（Department of Computer Science, Tsinghua University, Beijing 100084）

【Abstract】 While building service oriented Web application, it is necessary to search domain related Web services. Current search in UDDI servers is based on taxonomy of tModel and keyword-matching, which is not convenient to find domain related Web services. By crawling WSDL files directly from the Internet, and utilizing support vector machine to build WSDL file classifier, a method is proposed to realize domain based Web services search. Experiment shows that the method has a good accuracy.

【Key words】 Web service search; Domain; Support vector machine; WSDL

1 Web服务查找介绍

Web服务是新一代的Web应用，是通过Web来进行发布、查找和调用的自包含、自描述的模块化应用。一个Web服务是一个定义明确、自包含、并与其他服务的上下文和状态无关的函数。Web服务技术的兴起为Web的使用提出了新的思路，通过Web服务，用户不仅可以共享信息，还可以共享计算资源，通过Web服务之间的互操作，可以轻松构建基于SOA(Service Oriented Architecture)的复杂应用。

Web服务查找是整个SOA系统中非常关键的一部分，因为在构建Web应用的过程中，Web服务查找是进行后续的服务组合、调用和绑定的基础。传统的Web服务查找是在UDDI服务器中通过tModel分类信息、采用关键字匹配来进行的。这种查找存在两点不足：一是查找的范围有限，一般只能在微软和IBM提供的UDDI服务器上进行查找；实际上有很多Web服务并未在这些服务器上注册，而是直接放在互联网上，通过UDDI服务器未必查找得到；二是查找过程只能按照tModel分类信息和关键字进行，非常不便于查找某个领域的Web服务。

但是在构建SOA应用的时候，往往需要调用同一个领域中的Web服务。例如在开发一个旅游方面的Web应用时，开发人员往往会查找旅游领域的Web服务，涉及到有关机票订购、旅馆订房、汽车租赁、天气预报等方面；这就需要在UDDI服务器中依次通过“旅游”、“机票”、“宾馆”、“汽车”、“天气预报”等关键字来进行查询，查询的召回率和查准率都不高，也非常不方便。

实际上当前的Web服务都是通过WSDL文件进行描述的，只要得到WSDL文件，就可以访问和运行该Web服务。WSDL文件描述了Web服务的功能和接口，尤其是每个接口的名字（在标记 <operation name>中）往往包含了该Web服务是属于哪个领域的信息。例如一个Web服务“Global Weather”，其接口名“GetWeather”就非常清晰地表明它是关于天气预报的服务。

因此，本文提供一种方法直接从互联网上获得WSDL文件，并从中提取特征、确定该文件的所属领域，达到按照领域进行Web服务查找的目标。

2 WSDL文件爬行器

WSDL文件爬行器可以直接从互联网上获取WSDL文件。该爬行器通过3个来源：XMethods，Google和百度来搜索。XMethods是一个专门收集Web服务的网站，它列出Web服务的发布者、风格、服务名字、描述和实现。通过它的Web服务完全列表页面可以链接到每个服务的WSDL文件。

由于互联网上的WSDL文件在链接中往往包含“wsdl”字符串，因此在Google界面中输入“inurl:wsdl”，就可以得到所有包含“wsdl”字符串的网络链接。另外，包含“asmx”字符串的网络链接也可以得到WSDL文件，也可以在Google中查找“inurl:asmx”。然后通过分析上述搜索获得的网页，进而获取WSDL文件。

百度是中文搜索引擎，分别输入“inurl:(wsdl)”和“inurl:(asmx)”就可以得到包含字符串“wsdl”和“asmx”的网络链接。然后通过分析上述搜索获得的网页，进而获取WSDL文件。

另外，在Google和百度中搜索到的链接有许多是重复的，因此，最后能够得到的WSDL文件是：从XMethods中得到422个，从Google中得到511个，从百度中得到43个，总共976个。

作者简介：许 斌(1973－)，男，讲师、博士生，主研方向：Web服务，P2P，语义Web

收稿日期：2005-11-15 **E-mail：**xubin@tsinghua.edu.cn

3 WSDL 文件分类器

支持向量机(SVM)是基于Vapnik-Chervonenkis (VC) 理论[1]的学习方法，已经被广泛应用到文本分类和图像处理等领域。下面利用支持向量机构造出某个领域的WSDL文件分类器。

3.1 领域特征

为了区分不同的领域，每个领域通过一个领域关键字集合来表示其领域特征：

$DomainKeyword=[keyword_1, keyword_2, \dots, keyword_i]$

例如旅游领域涉及到机票订购、旅馆订房、汽车租赁、天气预报等方面，因此可以用下面集合表示：

$DomainKeyword_{travel}$=[Travel, Car, Weather, Hotel, Airport, Airline, Flight, Ticket, Aircraft, Reservation, Room, Restaurant,…]

3.2 从 WSDL 文件中抽取特征

WSDL 文件中的标记<wsdl:operation name="……"> 描述了 Web 服务的接口名字，该名字往往能够表明 Web 服务的作用。例如 Web 服务“Global Weather”的 WSDL 文件中：

```
<?xml version="1.0" encoding="utf-8"?>
…
<wsdl:portType name="GlobalWeatherSoap">
<wsdl:operation name="GetWeather">
<documentation xmlns="http://schemas.xmlsoap.org/wsdl/">Get weather report for all major cities around the world
</documentation>
<wsdl:input message="tns:GetWeatherSoapIn" />
<wsdl:output message="tns:GetWeatherSoapOut"/>
</wsdl:operation>
<wsdl:operation name="GetCitiesByCountry">
<documentation xmlns="http://schemas.xmlsoap.org/wsdl/">Get all major cities by country name(full / part)
</documentation>
<wsdl:input message="tns:GetCitiesByCountrySoapIn" />
<wsdl:output message="tns:GetCitiesByCountrySoapOut" />
</wsdl:operation>
</wsdl:portType>
…
```

第 1 个标记<wsdl:operation name="…">中的接口名字是“ GetWeather ”，通过大写字母分词以后可以得到“ Get ”和“ Weather ”两个词；第 2 个标记<wsdl:operation name=“ … ” >中的接口名字是“ GetCitiesByCountry ”，分词后能够得到“ Get ”、“ Cities ”、“ By ”、“ Country ”4 个词。所有<wsdl:operation name=“ … ” >标记中的接口名字以上述方式处理后，再用停词表（stop words list）进行处理，得到表示该 WSDL 文件特征的关键字集合：

WSDLKeyword=[Get, Weather, Get, Cities, Country]

在停词表的作用下，“ By ”就无法列入关键字集合。

在 WSDL 文件关键字集合的基础上，可以定义 WSDL 文件的向量：

$Vector_{WSDL}=[w1, w2, w3,\dots,w_i]$

$$w_i = \frac{TFi}{\sum_{i} TFi}$$

其中w_i是对第i个关键字按照词频[2]统计，并进行归一化处理得到的结果。

3.3 形成 SVM 训练集

为了构造某个领域的 WSDL 文件分类器，必须在支持向量机中提供训练集，包括正例集和反例集。正例集和反例集的筛选过程如图 1 所示。

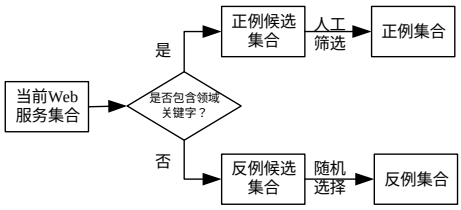

图 1 正例集和反例集的筛选过程

正例集的产生包括两个步骤：关键字筛选和人工筛选。首先从现有的 976 个WSDL文件集合WS_{full}中，凡是在任何一个WSDL文件的关键字集合中，包含了领域关键字集合中任意一个关键字的Web服务，都可以列入正例候选集合$WS_{filtered}$，否则列入反例候选集合。但是$WS_{filtered}$中的WSDL文件不一定是都是属于该领域的Web服务，因此还需要人工从中选择确实属于该领域的Web服务，形成正例集合$WS_{positive}$。从反例候选集合中只需要随机选择一定数量的WSDL文件，就可以形成反例集合$WS_{positive}$。

利用正例集合和反例集合中的所有 WSDL 文件的向量，就形成了支持向量机的训练集。支持向量机通过该训练集的训练以后，就可以形成对该领域的 WSDL 文件的判定规则。针对某一特定领域的 WSDL 文件分类器就形成了。

WSDL 文件爬行器与领域的 WSDL 文件分类器组合在一起，就实现了基于领域的 Web 服务查找功能。当发出该领域的 Web 服务查找请求时，WSDL 文件爬行器从互联网上得到新的 WSDL 文件，并提取该 WSDL 文件的向量，交给该领域的 WSDL 文件分类器判断，就可知道该 WSDL 文件是否属于该领域。但是，针对不同的领域需要分别构造该领域的 WSDL 文件分类器。

4 实验

为了验证WSDL文件分类器的精度，在正例候选集合$WS_{filtered}$中挑选 32 个WSDL文件形成正例集合$WS_{positive}$，从反例候选集合中选择 48 个WSDL文件形成反例集合$WS_{positive}$。然后，将正例集合和反例集合中的WSDL文件分别分成 5 组，每次挑选其中 4 组形成支持向量机的训练集，剩下的一组形成测试集。测试的结果如表 1 所示。

表 1 分类器精度测试

训练集	组 2,3,4,5	组 1,3,4,5	组 1,2,4,5	组 1,2,3,5	组 1,2,3,4
测试集	组 1	组 2	组 3	组 4	组 5
精度	82.35%	87.50%	87.50%	100.0%	94.12%

表 1 中的每组包括正例和反例的 WSDL 文件 14~17 个。从测试的精度结果来看，最高的精度是 100%，最低的精度是 82.35 %，平均精度是 90.29%。实验结果表明基于支持向量机的 WSDL 分类器的精度是比较高的。

5 结论和展望

在构建基于 SOA 的应用的时候，往往需要基于领域来查找 Web 服务；但是现有的 UDDI 查找方式并不能够很好地进行领域 Web 服务的查找，而且有许多 Web 服务并未在 UDDI 服务器中注册。本文提供了一种方法直接从互联网上进行基于领域的 Web 服务的查找。该方法通过网络爬行器直接从互联网上获得大量 WSDL 文件，利用关键词列表筛选出可能是该领域的 WSDL 文件，然后抽取每个 WSDL 文件的标记<operation name>的信息，形成表征该 Web 服务的关键词向

(下转第 88 页)