小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems

2014 年 9 月 第 9 期
Vol. 35 No. 9 2014

一种领域服务目标知识库的构建方法

张 能[1],王 健[1],李 征[1,2],何克清[1]

[1](武汉大学 软件工程国家重点实验室 计算机学院,武汉 430072)
[2](河南大学 计算机与信息工程学院,河南 开封 475001)
E-mail:jianwang@ whu. edu. cn

摘 要: 随着互联网上服务资源规模的迅速增长,如何根据用户个性化需求进行按需服务发现成为一个亟待解决的问题. 目前,已有大量的方法研究基于语义的服务发现,但这些方法往往并未涉及如何从服务的描述信息特别是自然语言文本描述中进行服务功能语义的挖掘. 针对该问题,本文提出了一种可扩展的服务目标(即功能语义)抽取方法,能够从多种服务描述信息中进行服务目标的抽取,并在此基础上进行领域服务目标知识库的构建. 最后,通过 ProgrammableWeb 网站上真实的服务集验证了所提出方法的有效性.

关 键 词: 服务目标;领域知识库;服务发现
中图分类号: TP311 **文献标识码:** A **文 章 编 号:** 1000-1220(2014)09-1943-06

Approach of Constructing Domain Service Goal Knowledgebase

ZHANG Neng[1], WANG Jian[1], LI Zheng[1,2], HE Ke-qing[1]
[1](State Key Laboratory of Software Engineering, Computer School, Wuhan University, Wuhan 430072, China)
[2](School of Computer and Information Engineering, Henan University, Kaifeng 475001, China)

Abstract: With the rapid increase of service resources in Internet, how to achieve on-demand service discovery according to users' personalized requirements becomes an urgent issue. Many approaches on semantic-based service discovery have been proposed, but few researches focus on mining functional semantics of services from the services' description information, especially from the services described in natural language. To solve this problem, we propose a scalable approach for extracting service goals (functional semantics), which can extract service goals from various types of service description. Furthermore, we investigate how to construct domain-specific knowledgebase of service goals. Finally, the effectiveness of the proposed approach is verified by using the real set of services on the ProgrammableWeb.

Key words: service goal; domain knowledgebase; service discovery.

1 引 言

Web 服务发现是一个根据服务请求者的需求查找合适服务的过程. 随着互联网上 Web 服务资源规模的快速增加,服务发现的重要性也日益凸显. 例如,截止到 2013 年 6 月,Web 服务编程网站 ProgrammableWeb[1] 上发布的 Web 服务或 Web API 已超过 9300 个,并且其增长趋势十分明显. 如何根据用户(包括开发者和终端用户)的个性化需求进行按需服务发现和重用是一个重要的挑战性问题.

已有的服务发现方法大致可分为语法级和语义级两类[1]. 前者基于简单的关键字匹配进行服务搜索,由于缺少语义信息,查全率和查准率都不高,难以满足用户的需求. 为了解决基于关键词搜索的局限性,大量基于语义的服务发现方法被提出. 这些方法大多是通过描述逻辑或一阶逻辑对服务进行精确的形式描述,形成了各种基于本体的语义 Web 服务描述语言如 WSDL-S、OWL-S 和 WSMO 等,并将服务匹配问题转化为逻辑推理问题. 然而,语义 Web 服务发现在很大程度上取决于是否有可用的、良构的领域本体,而构造这种体现领域知识的本体往往非常困难,这也为语义 Web 服务发现方法的实现带来了一定的困难.

在实际的 Web 搜索中,用户通常会使用一些能够准确表达其需求的高层次目标[2],如"规划行程(plan a trip)"、"预订旅馆(book a hotel)"等. 与基于关键字的语法查询相比,这种目标驱动的、能体现用户意图的查询方式可以为用户返回更准确的结果. 但是,还没有相关工作来研究如何从服务的描述信息中进行领域相关的服务目标知识的挖掘. 针对该问题,本文提出了一种从服务描述中进行服务目标抽取的方法,并进一步研究领域服务目标知识库的构造方法,从而为服务发现提供支持.

本文第二节介绍相关研究工作;第三节介绍如何从服务描述信息中进行服务目标抽取;第四节阐述领域服务目标知识库的构建方法;第五节给出实验结果及分析;最后是对本文

1. http://www. programmableweb. com/

收稿日期:2013-10-01 收修改稿日期:2013-08-14 基金项目:国家自然科学基金项目(61202031)资助;国家科技支撑计划项目(2012BAH07B01)资助;国家云计算示范工程项目资助;高等学校学科创新引智计划项目(B07037)资助. 作者简介:张 能,男,1990 年生,硕士研究生,研究方向为软件工程和服务计算;王 健,男,1980 年生,博士,讲师,研究方向为软件工程和服务计算;李 征,女,1984 年生,博士,研究方向为软件工程和服务计算;何克清,男,1947 年生,教授,博士生导师,研究方向为软件工程和服务计算.

工作的总结与展望.

2 相关工作

本文的相关研究包括服务发现和文本特征挖掘两个方面. 就服务发现而言,国内外已有大量研究. 在文献[3]中,利用自定义的服务功能描述模型对服务进行标注,提出了一种基于功能语义的服务发现方法,但未涉及如何从已有的服务描述中进行功能语义的获取. 文献[4,5]利用输入/输出参数、前置/后置条件等信息来完善对服务功能的表达,但未考虑用户需求的表达方式. 同时,现有的服务发现研究大都是针对规范化描述的 Web 服务,对以自然语言文本描述的 Web 服务的发现则较少涉及.

从自然语言文本中进行特征挖掘在其它领域已经得到广泛研究. Ghose 等人实现了一个称为"R-BPD"的工具包,通过使用语法分析器来识别文本描述中包含的业务过程[6]. Friedrich 等人在文献[7]中提出了一种基于自然语言处理工具如 Stanford Parser[2],FrameNet[3] 和 WordNet[4],进行 BPMN 模型自动生成的方法,阐述了如何从文本描述中进行流程知识抽取. 文献[8]中,Dumitru 等人利用文本挖掘和一种自定义的增量式聚类算法可以从产品的文本描述中发现领域的特征集合. 与这些工作相比,本文着重研究如何从服务的文本描述中进行服务目标抽取,并且在抽取过程中,针对服务描述信息的特点,采用了不同的抽取策略.

在文献[9]中,我们介绍了服务目标的抽取和推荐方法以及基于服务目标的服务发现框架,并通过实验分析了服务目标在服务发现和推荐中的重要作用. 本文侧重于阐述领域服务目标知识库的生成,并对服务目标的抽取方式进行了细化和改进,并增加了对 WSDL 服务描述的处理.

3 服务目标抽取

服务目标(Service Goal)是用来展现一个服务所具备的从用户需求角度出发的功能特征. 在抽取服务目标前,需要先确定服务目标的表示形式,然而,目前并没有标准的服务目标描述方式. 在文献[10]中,用户目标被表示为一个动词和一组参数形成的短语. target 参数是其中最重要的参数,用来表示目标所影响的实体(包括对象或结果),可以用一个名词或名词短语来描述. 因此,目标通常可用一种动名词对(verb-noun pair)的形式来刻画. 事实上,很多 WSDL 操作名的命名也是推荐采用动名词对的形式[11],如 GetWeatherByZip. 因此,在文献[9]中,我们采用如下形式对服务目标进行定义.

定义 1. 一个服务目标 sg 可用如下三元组表示:< sgv, sgn, sgp >,其中,sgv 是动词或动词短语,表示 sg 要执行的动作;sgn 为名词或名词短语,代表 sg 的操作对象. sgv 和 sgn 是 sg 的必要组成部分. sgp 是可选部分,用于对 sg 进行补充说明,比如 sg 的操作方式和约束等.

确定服务目标的表示形式后,下面考虑如何从服务的描述信息中进行服务目标抽取. 虽然服务的描述方式较多,但大致可分为结构化与非结构化两种,前者包括 WSDL、WADL 等规范化描述;后者主要是自然语言文本. 图 1 给出了服务目标的抽取流程. 针对文本和 WSDL 两种主要的描述方式,我们分别提供了相应的抽取流程,然后将这两种方式得到的服务目标集进行合并,形成服务的服务目标集.

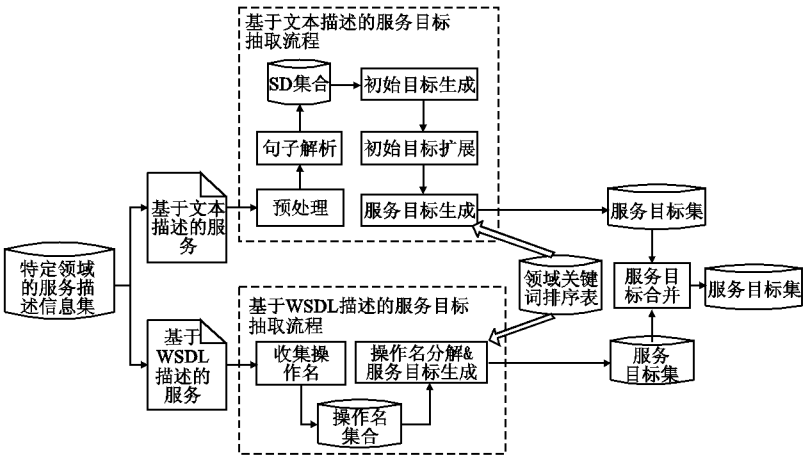

图 1 服务目标抽取

Fig. 1 The Process of service goal extraction

由于服务目标抽取的主要目的是对服务的领域语义特征进行抽取,因此,需要先对服务进行面向领域的分类. 在文献[12]中我们提出了一种基于支撑向量机的服务分类方法,该方法在进行服务分类的同时可以得到领域关键词排序表. 下面将对文本描述和 WSDL 描述的服务目标抽取流程进行详细说明.

3.1 基于文本描述的服务目标抽取

为了实现从文本描述中进行服务目标抽取,必须对文本描述中语句的语法结构进行分析. 目前,已有许多开源的工具可以实现这一功能,Stanford parser 是使用最广泛的一种. 利用 Stanford parser 不仅可以得到句子的语法结构,而且还能解

2. http://nlp. standford. edu:8080/parser/
3. http://framenet. icsi. berkeley. edu/
4. http://wordnet. princeton. edu/wordnet/

析出体现句中词汇间语法依赖关系的 Stanford Dependency(简称 SD)集合. 一条 SD 可以表示为:relName(w1,w2),表示词 w1 与词 w2 间具有 relName 关系,例如,"dobj(create,graph)"表示 create 的直接宾语是 graph. 在 3.1.1 和 3.1.2 小节中将具体介绍本文用到的 relName.

本文采用 Stanford parser 对描述语句进行解析,然后基于生成的 SD 集合进行服务目标抽取. 服务目标的抽取过程包括以下三个阶段:初始目标生成、初始目标扩展和服务目标生成.

3.1.1　初始目标生成

通过对多种句子结构的解析结果进行观察,我们发现,尽管一个句子可能会包含多个服务目标,但总存在一个或几个"初始目标(Initial Goal)",只有基于这些初始目标才能进一步发现其它的服务目标. 我们主要从以下 3 种情形进行初始目标集 *IG* 的获取:

情形 1:nsubjpass(w1,w2),出现于被动语态的句子中,表示名词 w2 是该句子的主语,且与该句子的主要动词 w1 具有 nsubjpass 关系. 在这种情况下,w1 和 w2 将分别作为初始目标的动词和名词部分;

情形 2:dobj(w1,w2),出现于主动语态的句子中,表示名词 w2 是动词 w1 的直接宾语. 在这种情况下,w1 和 w2 将分别作为初始目标的动词和名词部分;

情形 3:prep(w1,w2)和 nsubj(w1,w3):情形 2 无法直接处理诸如"search for"和"deal with"之类的动词短语,这种情形的初始目标包含在介词关系 prep(w1,w2)中,可以从中识别出初始目标的动词 w1 和名词 w2,但并非所有介词关系都可以识别出初始目标所需的动词和名词,如 prep_of(information,hotels). 因此,还需要通过 nsubj(w1,w3)来确定 w1 是句子中的主要动词. 只有当 w1 出现在该关系中时才可以确定 w1 是初始目标所需的动词. 例如,"This API can search for the newest travel information."包含 SD 关系 prep_for(search-4,information-9)和 nsubj(search-4,API-2),从中可以识别出 *IG* = {search information}. 注意,prep 往往具有多种不同的表现形式,这是根据其包含的介词部分不同而导致的,如 prep_for 和 prep_with 等.

3.1.2　初始目标扩展

上面生成的初始目标还存在一些不足. 一方面,初始目标往往过于简单或抽象,缺失了体现领域特征的语义信息,如上一小节例句中的初始目标 search information,我们希望得到的目标是 search newest travel information. 另一方面,除了初始目标外,还有一些潜在的服务目标有待发现. 为了解决上述问题,需要进一步利用其他相关的 SD 关系对初始目标进行扩张. 表 1 列举了进行初始目标扩展需要考虑的 SD 关系.

算法 1 给出了初始目标的生成及扩展算法. 该算法以一条描述语句的 SD 集合 *SenSD* 作为输入,得到该语句中包含的候选服务目标集合 *CSG*.

算法 1. 初始目标生成及扩展算法

输入:一条语句的 SD 集合 *SenSD*;

输出:该语句的候选服务目标集合 *CSG*.

算法描述:

```
1. begin
2. IG ←nil;          //初始化初始目标集合
3. CSG←nil;          //初始化候选服务目标集合
4. IG←generateIG(SenSD);   //根据 3.1.1 节中 3 种情形初始目标生成
5. for ig ∈ IG
   //对 ig 进行扩展,得到 ig 对应的候选服务目标集合
6. ig_csgs←extendIG(ig);
7.    add ig_csgs to CSG;
8. end for
9. end
```

表 1　支持初始目标扩展的 SD 关系

Table 1　SD relationships for the extension of initial goals

relName	作用	句子示例	扩充结果示例
conj	识别与初始目标的 *sgv* 或 *sgn* 并列的动词或名词,发现潜在服务目标	search or reserve hotel and flight	*IG* = {search hotel} *IG'* = {search hotel, search flight, reserve hotel, reserve flight}
prep	对初始目标的 *sgv* 或 *sgn* 进行扩充	find articles about travel	*IG* = {find articles} *IG'* = {find articles about travel}
appos	识别初始目标的 *sgn* 的介词短语中的并列名词,发现潜在的服务目标	get information of travel, hotel	*IG* = {get information} *IG'* = {get information of travel, get information of hotel}
prt	用于对初始目标的 *sgv* 进行扩充	look up hotels	*IG* = {look hotels} *IG'* = {look up hotels}
nn	用于对初始目标的 *sgn* 进行扩充	create travel network	*IG* = {create network} *IG'* = {create travel network}
amod	用于对初始目标的 *sgn* 进行扩充	retrieve historical weather	*IG* = {retrieve weather} *IG'* = {retrieve historical weather}

3.1.3　服务目标生成

在扩展后的候选服务目标集中,某些候选服务目标可能包含对服务功能无实际意义的动词和名词,如 allow、let 等动词以及 information、more 等名词,且存在词形不一致的问题,如 retrieve、retrieving 和 retrieved. 另外,由于语言表达方式的多样性,候选服务目标集中可能存在一些冗余的服务目标,如 get hotel information 和 get information of hotel,虽然这两个服务目标的表达方式不同,但它们的语义是一致的. 针对上述问题,对候选服务目标集需要依次进行如下处理:

1)词形还原:利用 WordNet 的 Stemming 算法对动词和名词进行还原处理,将它们的多种形式统一化为基本原型,例如"retrieving""retrieved"都会变换成"retrieve";

2)服务目标替换:对 sgn 的名词中仅包含 information、functionality 等抽象词的服务目标,利用 sgp 中与 sgn 紧邻的介词结构中的名词对抽象词进行替换,并将使用到的介词结构从 sgp 中去掉;

3)停词过滤:由于无意义的动词不多,可以通过自定义停词表 StopVerbs 进行过滤,将 sgv ∈ StopVerbs 的候选服务目标删除;对于名词,可利用领域关键词排序表进行处理,如只保留领域关键词排序表中位于前 100 的名词.

3.2　基于 WSDL 描述的服务目标抽取

如前所述,由于很多 WSDL 文件的操作名命名方式采用了类似于动名词对的形式,因此,从 WSDL 描述中进行服务

目标抽取较为容易,其主要任务就是对操作名进行分解,从中抽取相应的部分构建服务目标,具体流程如图 1 中的基于 WSDL 描述的服务目标抽取流程所示.

具体而言,在抽取过程中,充分利用了大多数 WSDL 文件中操作名命名方式的如下特点:1) 命名符合 Pascal 标记法,即各单词的首字母大写,在对操作名进行分词时可以充分利用这个特点. 但也有一些特殊情况,如"GetWeatherByWMOID"中的"WMOID". 对这种特殊情况,首先判断这组连续的大写字母后面是否还有小写字母,即是否是操作名的结尾,若是,则整个截取这段大写字母,否则,取至倒数第二个大写字母结束;2) 在操作名中,服务目标各组成部分的位置是相对固定的,通常是以 sgv-sgn-sgp 的形式出现,如 Get[sgv]Weather[sgn]ByCityState[sgp],其中,"[ ]"内的标记指示了它前面的词汇或短语在服务目标中的成分.

基于上述特点,可以实现操作名的分解和服务目标的生成. 对得到的候选服务目标集合,需要进行与 3.1.3 节相同的处理过程.

3.3 服务目标合并

如果一个服务包含多种描述方式,则需要对多种描述的抽取结果进行合并,从而得到最终的服务目标集. 在合并时,对存在包含关系的多个服务目标,仅保留语义信息最丰富的服务目标.

4 领域服务目标知识库的构建

为了促进服务目标的共享和重用,利用领域所有服务的服务目标集,可以进一步抽取领域服务目标集合,从而辅助领域专家进行领域服务目标知识库的构建或完善,为后续的服务提供者对服务功能进行描述提供指导. 另外,用户也可以利用领域服务目标知识库进行服务发现. 领域服务目标知识库构建的基本框架如图 2 所示.

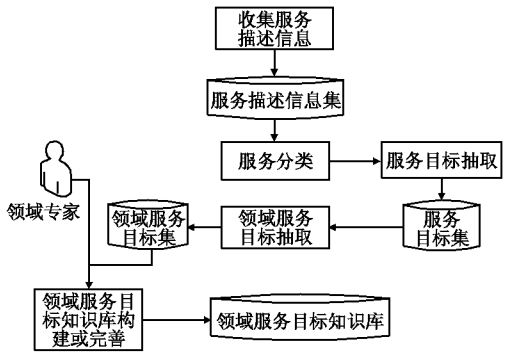

图 2 领域服务目标知识库的构建框架

Fig. 2 The framework for constructing domain-specific service goal knowledgebase

领域服务目标(Domain Service Goal)是指在一个领域中出现次数较多、且能体现一定领域共性特征的服务目标. 因此,领域服务目标抽取就是要将存在包含(包括部分包含)关系的多个服务目标(如 search airline 与 search airline ticket)合并到能够体现其共性特征的形式.

由于在领域所有服务的服务目标集中可能会存在等价的服务目标,如 get hotel 和 retrieve hotel,因此,在领域服务目标抽取前,需要利用 WordNet 的同义词集将存在同义关系的动词或名词统一成某个指定的动词或名词,从而将等价的服务目标变换成一个代表性的服务目标. 例如,通过将同义动词集{get、retrieve、find、fetch、obtain}中的动词都统一成 get,可以将 retrieve hotel 变换成 get hotel.

定义 2. 一个领域服务目标 dsg 可用如下五元组表示:$<sgv,sgn,NL,NR,SGP>$,其中,sgv 代表其动词部分,sgn 代表其名词部分;*NL* 是位于 sgn 左侧的可选属性集合;*NR* 是位于 sgn 右侧的可选属性集合;*SGP* 是 dsg 的补充说明集合.

基于定义 2,进一步定义领域服务目标抽取规则.

定义 3. 领域服务目标抽取规则,规定了任意给定一个新的服务目标 sg,如何对领域服务目标集 *DSG* 进行更新或演化.

首先,确定领域服务目标集中与 sg 相关的子集 $DSG_{sg}=\{dsg \mid dsg \in DSG \wedge dsg.sgv = sg.sgv\}$,即定位到所有与 sg 中动词部分相同的领域服务目标.

然后,按照如下情形处理:

1) 若 DSG_{sg} 中存在非空子集

$S_1=\{dsg \mid (dsg \in DSG_{sg}) \wedge sg.\text{sgn} = l_{sg-dsg} \circ dsg.\text{sgn} \circ r_{sg-dsg}\}$

则对 $\forall dsg \in S_1$:

$dsg.NL = dsg.NL \cup \{l_{sg-dsg}\}$

$dsg.NR = dsg.NR \cup \{r_{sg-dsg}\}$

$dsg.SGP = dsg.SGP \cup sg.sgp$

$N(dsg) = N(dsg) + N(sg)$

其中,l_{sg-dsg}、r_{sg-dsg} 分别为 sg. sgn 相对于 dsg. sgn 的左侧部分和右侧部分,为字符串连接操作,N(x)代表 x 在领域中出现的次数,即在领域的多少个服务中出现.

2) 若 DSG_{sg} 中存在非空子集

$S_2=\{dsg \mid (dsg \in DSG_{sg}) \wedge dsg.\text{sgn} = l_{dsg-sg} \circ sg.\text{sgn} \circ r_{dsg-sg}\}$

则对 DSG_{sg} 作如下处理:添加 newdsg,删去 S_2. 其中,newdsg 为:

$newdsg.sgv = sg.sgv$

$newdsg.\text{sgn} = sg.\text{sgn}$

$newdsg.NL = \cup_{dsg \in S_2}(dsg.NL \cup \{l_{dsg-sg}\})$

$newdsg.NR = \cup_{dsg \in S_2}(dsg.NR \cup \{r_{dsg-sg}\})$

$newdsg.SGP = \cup_{dsg \in S_2} dsg.SGP \cup sg.sgp$

$N(newdsg) = \sum_{dsg \in S_2} N(dsg) + N(sg)$

其中,l_{dsg-sg}、r_{dsg-sg} 分别为 dsg. sgn 相对于 sg. sgn 的左侧部分和右侧部分.

3) 若不存在 S_1、S_2,则将 sg 添加到 DSG_{sg} 中.

最后,得到的领域服务目标集可以帮助领域专家进行领域服务目标知识库的构建或完善.

5 实验分析

本节将通过实验对本文方法的有效性进行评估. 所有算法均采用 Java 实现,运行在一台具有 Intel Core(TM) i5 CPU 760 @ 2.80GHz,4GB 内存,操作系统为 Windows7 的 PC 上.

5.1 数据集

实验数据集来源于 ProgrammableWeb,该网站上提供的所有 Web 服务均包含短文本的描述信息.我们使用该网站提供的开放 API 并结合网络爬虫的方法将服务的概要信息(包括名称、概述、描述、类别、标签等)提取出来,保存在文本文档中作为初始数据集.此外,对存在 WSDL 描述的服务,还收集了其 WSDL 文件.

在实验前,先利用我们前期研究[12]中提出的基于支撑向量机的服务分类方法对初始数据集进行面向领域的分类并得到领域关键词排序表.

5.2 实验结果及分析

为验证服务目标抽取方法的有效性,我们以服务数较多的三个领域:Travel(190)、Music(190)和 Financial(408)为例进行实验,括号中的数字表示该领域的服务数目.

5.2.1 文本描述的服务目标抽取

由于初始获取的文本描述中会包含一些特殊字符,如 user creation/authorilization,会影响句子解析的准确度.因此,在对文本描述进行服务目标抽取前,需要进行一定的预处理,尽可能消除或替换这些特殊字符,如可以将"/"替换为"and".然后,对预处理后三个领域内的服务利用基于文本描述的服务目标抽取方法进行处理,得到每个服务的服务目标集.

为了对实验结果进行评估,我们从三个领域中分别随机选取 10 个文本描述长度超过 70 个词的服务作为实验数据集,让 5 名研究生用手工方式对实验数据集中的每个服务进行服务目标抽取,从而对每个服务得到 5 个不同的服务目标集.然后根据每个用户的抽取结果,使用准确率(Precision)和召回率(Recall)对自动抽取的结果进行评估.关于 Precision 和 Recall 的定义如式(1)、(2)所示.

$$Precision = \frac{|SG_{Ui,s} \cap SG_{M,s}|}{|SG_{M,s}|} \quad (1)$$

$$Recall = \frac{|SG_{Ui,s} \cap SG_{M,s}|}{|SG_{Ui,s}|} \quad (2)$$

其中,$SG_{M,s}$代表利用本文方法进行自动抽取得到的服务 s 的服务目标集,$SG_{Ui,s}$为用户 Ui 对 s 手工抽取的服务目标集.

表 2 是 Travel 领域的实验结果,其中,SG_{U1} ~ SG_{U5}列分别是 5 个用户的抽取结果,列记录"m(n)"代表该列用户从服务 s 中抽取的 m 个服务目标中有 n 个出现在 $SG_{M,s}$中.

表 2 Travel 领域的服务目标抽取结果

Table 2 Result of service goal extraction for Travel

API Name	$\|SG_M\|$	SG_{U1}	SG_{U2}	SG_{U3}	SG_{U4}	SG_{U5}	Average Precision	Average Recall
Active. com Camping	5	3(3)	4(4)	4(4)	4(4)	4(4)	0.76	1.0
Add To Trip	14	10(10)	12(11)	13(12)	10(10)	12(12)	0.7911	0.9680
AdventureLink	9	8(7)	8(8)	10(8)	8(8)	10(8)	0.8667	0.8950
Amadeus Cruise	11	10(10)	9(9)	11(11)	10(10)	11(11)	0.9273	1.0
Belair	5	4(4)	6(4)	6(4)	5(4)	6(4)	0.8	0.760
BookingMarkets	10	8(8)	8(8)	10(10)	10(10)	9(9)	0.9	1.0
BookingSync	10	10(10)	10(10)	10(10)	10(10)	10(10)	1.0	1.0
Budget Your Trip	7	5(5)	6(6)	6(6)	5(5)	6(6)	0.8	1.0
Castilla-La Mancha	4	4(4)	4(4)	4(4)	4(4)	4(4)	1.0	1.0
Cleartrip Hotel	10	9(9)	9(9)	9(9)	9(9)	9(9)	0.9	1.0

从表 2 可以看出,5 个用户对同一个服务的抽取结果有所不同,主要是因为不同用户对服务重要功能的理解存在一定差异,例如,对 Add To Trip 服务而言,只有用户 U1 的抽取结果中不含 collaborate with friends on trip 和 collaborate with family on trip.此外,每个服务的平均召回率大多都接近 1.0,而平均准确率略低.表 3 列出的三个领域的平均 Precision 和 Recall 也反映了类似的结果.这说明尽管本文方法的抽取结果中存在一些无意义的目标需要进一步过滤,但整体而言可以很好地覆盖不同用户的抽取结果.由于在服务发现时,用户通常更关心得到的服务能否满足需求,而不在乎该服务是否包含其它的服务目标,所以,本文方法能够有效地抽取出满足用户需求的服务目标.对于表 2 中部分服务的平均召回率小于 1.0 的情况,主要有以下原因:

1) Stanford parser 在某些动词的识别上存在不足,例如,integrate trips and book reservations 中的"book"及 access mobile apps and view mileage balances 中的"view";

2) 目前的服务目标抽取策略不够完善,对一些动词的名词化和现在分词形式难以识别,例如,retrieval of market data 及 clearing of completed orders.这将是我们下一步有待改进的地方.

表 3 三个领域的 Precision 与 Recall 均值

Table 3 Three domain's average precision and recall

Domain	*Precision*	*Recall*
Travel	0.8745	0.9623
Music	0.8513	0.9428
Financial	0.8531	0.8816

5.2.2 WSDL 描述的服务目标抽取

ProgrammableWeb 网站中 SOAP 服务约占服务总数的 23%,且大多数 SOAP 服务都包含相应的 WSDL 描述文件.就 Financial、Travel 和 Music 这三个领域而言,175 个服务包含相应的 WSDL 链接,其中有效的 WSDL 文件数为 123.

为了验证 WSDL 描述的服务目标抽取效果,先对三个领域的 WSDL 文件集利用基于 WSDL 描述的服务目标抽取方法进行处理.然后,从三个领域的 WSDL 文件集中分别随机选取 10 个 WSDL 文件作为实验数据集,采用与 5.2.1 中相同的评估方法对实验数据集进行处理.

所得实验结果的平均准确率和平均召回率分别为

89.4% 和 83.9%,说明该方法对 WSDL 描述方式同样具有比较好的处理效果.但是该方法也存在一些不足,这种不足主要是因为某些 WSDL 操作的命名方式不符合 sgv-sgn-sgp 形式,包括以下情形:

1)操作名中缺少服务目标的必要组成部分,难以进行自动抽取,如 ProgrammableWeb 站点上服务 Villarenters 的 WSDL 文件中的操作名 GoogleMapsLocators 和 VillaSearch;

2)尽管操作名中服务目标的必要组成部分都具备,但各部分间的次序不符合 sgv-sgn-sgp 形式,如 ProgrammableWeb 站点上服务 RadioTime 的 WSDL 文件中的操作名 Account_UserAuthenticate 和 Search_CategoryGetByType 所对应的服务目标应分别是 authenticate account user 和 get search category by type.这将是我们下一步有待改进的地方.

5.2.3 领域服务目标抽取

表 4 所示的是 Travel 领域中排名前 10 的领域服务目标,该排名是按照领域服务目标在领域服务文档中出现的次数进行 计算的.从表4可以看出,大部分的领域服务目标能够代表

表 4 Travel 领域的领域目标抽取结果(排名前 10)

Table 2 Result of domain goal extraction for Travel (TOP 10)

sgv	[NL] + *sgn* + [NR] + {SGP}
provide	***travel*** [industry, insurance, agency, tip, booking]
provide	[travel, hotel] ***booking*** [platform]
include	[activity, tour] ***listing*** [destination, property]
search	***travel*** [mode, website, experience, activity, tour]
search	[airline, cheap] ***hotel*** [website, room, image, accomodation, list] {by destination, by locale, by region, by specified arrival date, by number of guest}
search	***flight*** [time, delay]
get	[real-time] ***hotel*** [photo, pricing data] {for particular city}
get	***trip*** [plan]
get	***airport*** [delay]
share	[travel] ***experience***

Travel 领域的重要功能特征,说明本文的领域服务目标抽取方法可以在领域专家构造领域服务目标知识库时发挥较好的辅助作用.虽然抽取的结果中存在一些质量不高的服务目标,如 include [activity, tour] listing [destination, property],但是领域专家可以比较容易地对不合适的服务目标进行过滤或改进.

6 总结及展望

为促进基于功能语义的服务发现,本文提出了一种可扩展的服务目标抽取方法,能够从服务的多种描述信息中获取服务的功能语义信息,并以 WSDL 描述和文本描述为例对服务目标的抽取流程进行了说明.然后,基于抽取的服务目标,给出了一种领域服务目标知识库的构建方法,为服务目标的共享和重用奠定了基础.最后,通过实验对上述方法的有效性进行了验证.

本文的后续研究工作包括:1)对服务目标抽取方法进行优化,进一步提高服务目标抽取的质量;2)利用构建的领域服务目标知识库,进一步研究服务发现和推荐方法.

References:

[1] Hu jian-qiang, Zou Peng, Wang Huai-min, et al. Reasearch on web service description language QWSAL and service matching model [J]. Chinese Journal of Computers, 2005, 28(4): 505-513.

[2] Markus Strohmaier, Mathias Lux, Michael Granitzer, et al. How do users express goals on the web? -an exploration of intentional structures in web search [C]. In: Proceedings of the International Conference on Web Information Systems Engineering, Nancy, France, 2007, 67-78.

[3] Ye Lei, Zhang Bin. A method of service discovery based on functional semantics [J]. Journal of Computer Research and Development, 2007, 44(8): 1357-1364.

[4] Massimo Paolucci, Takahiro Kawmura, Terry R. Payne, et al. Semantic matching of web services capbilities [C]. In: Proceedings of the International Semantic Web Conference, Sardinia, Italy, 2002: 333-347.

[5] Kaarthik Sivashanmugam, Kunal Verma, Amit Sheth, et al. Adding semantics to Web services standards [C]. In: Proceedings of the International Conference on Web Services, Nevada, USA, 2003: 395-401.

[6] Aditya Ghose, George Koliadis, Arthur Chueng. Rapid business process discovery (R-BPD) [C]. In: Proceedings of Conceptual Modeling-ER, 2007: 391-406.

[7] Fabian Friedrich, Jan Mendling, Frank Puhlmann. Process model generation from natural language text [C]. In: Proceedings of the International Conference on Advanced Information Systems Engineering, 2011: LNCS 6741: 482-496.

[8] Horatiu Dumitru, Marek Gibiec, Negar Hariri, et al. On-demand feature recommendations derived from mining public product descriptions [C]. In: Proceedings of the International Conference on Software Engineering, Zürich, Switzerland, 2011, 181-190.

[9] Wang Jian, Zhang Neng, Zeng Ceng, et al. Towards services discovery based on service goal extraction and recommendation [C]. In: Proceedings of the International Conference on Services Computing, 2013: 65-72.

[10] Colette Rolland, Carine Souveyet, Camille Ben Achour. Guiding goal modeling using scenarios [J]. IEEE Transactions on Software Engineering, 1998, 24(12): 1055-1071.

[11] David J N Artus. SOA realization: service design principles service design to enable IT flexibility [EB/OL]. http://www.ibm.com/developerworks/webservices/library/ws-soa-design/, 2006.

[12] Zhang Jia, Wang Jian, Patrick C K Hung, et al. Leveraging incrementally enriched domain knowledge to Enhance service categorization [J]. International Journal of Web Services Research (IJWSR), 2012, 9(3): 43-66.

附中文参考文献:

[1] 胡建强,邹 鹏,王怀民,等. Web 服务描述语言 QWSAL 和服务匹配模型研究[J]. 计算机学报,2005,28(4):505-513.

[3] 叶 蕾,张 斌. 基于功能语义的 Web 服务发现方法[J]. 计算机研究与发展,2007,44(8):1357-1364.