基于本体论和词汇语义相似度的 Web 服务发现

吴 健 吴朝晖 李 莹 邓水光

(浙江大学计算机科学与技术学院 杭州 310027)

摘 要 Web 服务的大量涌现对服务发现提出了挑战,UDDI 上基于关键词和简单分类的服务发现机制已经不能很好满足需要.该文在分析现有相关研究的基础上,提出了基于本体论和词汇语义相似度的 Web 服务发现方法.通过构建 Web 服务本体,给出一个明晰的 Web 服务发现的研究对象,指出可对 Web 服务进行的几种相似度计算,并对其中的词汇语义相似度计算进行详细讨论.文中具体给出两种词汇语义相似度计算方法,其中第一种方法计算词汇语义相似度基于词语间距离度量,第二种方法计算词汇语义相似度则建立在义原相似度基础上.引入本体论和词汇语义相似度,为 Web 服务相似度计算、Web 服务发现提供了一种有效可行的方法.

关键词 本体;Web 服务;词汇语义相似度;Web 服务发现

中图法分类号 TP311

Web Service Discovery Based on Ontology and Similarity of Words

WU Jian WU Zhao-Hui LI Ying DENG Shui-Guang

(*College of Computer Science, Zhejiang University, Hangzhou* 310027)

Abstract Current infrastructure of Web services discovery such as UDDI provides text and taxonomy based search capabilities. With the large growth in number of Web services, the current Web service discovery mechanism becomes inefficient. Ontology and similarity based discovery or matching of services is a promising approach to address this challenge. This paper proposes Web service ontology and indicates several approaches to measuring similarity on this ontology. The method based on similarity of words which can be estimated in two ways is discussed in detail. This approach used in project will significantly improve Web services discovery.

Keywords ontology; Web service; similarity of words; Web service discovery

1 引 言

Web 服务发现是 Web 服务系统架构中的一个重要部分[1,2].UDDI 是众多发展和支持 Web 服务的解决方案中最受瞩目的一个.UDDI 上的 Web 服务发现,是通过对 UDDI 上的服务注册信息进行关键词精确匹配实现的,主要是对服务 ID 或名称、或是服务的有限的属性值进行匹配.但如同使用搜索引擎一样,人们在感谢 UDDI Web 服务注册中心带来的寻找 Web 服务的便捷的同时,也常常为查准率和查全率不高困扰.

主要原因在于,基于关键词匹配的 Web 服务发现具有以下缺陷:(1)对所需查询的目标不能准确描述;(2)不能度量候选者和查询目标间的符合程度;(3)不能使用细化、泛化、平级扩展等语义操作进行

收稿日期:2004-12-06;修改稿收到日期:2005-02-24.本课题得到国家"八六三"高技术研究发展计划项目基金(2001AA414320,2001AA113142)与国家"九七三"重点基础研究发展规划项目基金(2003CB316906)资助.**吴 健**,男,1975 年生,博士,讲师,研究方向为 Web 服务、网格计算、数据挖掘.E-mail:wujian2000@zju.edu.cn.**吴朝晖**,男,1966 年生,教授,博士生导师,研究方向为分布式人工智能、网格计算、生物认证与嵌入式系统.**李 莹**,男,1973 年生,博士,讲师,研究方向为软件体系结构、编译技术、中间件.**邓水光**,男,1979 年生,博士研究生,研究方向为 Web 服务、工作流、中间件.

查询.其中,前两点是影响查准率的重要因素,第三点主要影响查全率.

在中医药网格 DartGrid[3] 的构建过程中,高效准确的服务发现一直是我们的研究重点之一.在 DartGrid 和服务工作流[4~6]等研究工作的基础上,我们通过构建服务本体,采用词汇语义相似度、服务属性相似度、参量相似度、结构相似度等多种相似度度量方法,提供蕴涵语义的检索来解决上述问题.用本体描述 Web 服务可以为服务发现提供一个明晰的研究对象,多种相似度度量方法为比较候选者和查询目标间的符合程度提供可操作性,蕴涵语义的检索通过细化、泛化、平级扩展等语义操作实现语义联想提高查全率.关于服务属性相似度、参量相似度、结构相似度和蕴涵语义的检索,作者已在文献[6]中详细阐述.

在此,本文在文献[3~6]的基础上,简要给出服务本体的构建、Web 服务间相似度度量的几种方法后,重点讨论两种词汇语义相似度的计算方法.

2 相关研究

针对 Web 服务发现使用关键词匹配,导致服务查准率查全率不高的问题,业界已有不少相关研究:文献[7]提出了语义 Web 描述语言 DAML 描述服务,以 Prolog 语言为推理语言的服务发现系统,服务发现的依据是预先定义的服务属性本体属性值.文献[8]提出以 DAML-S 语言描述服务,通过服务的属性和接口的输入输出概念匹配,得到匹配结果.文献[9]提出了一种基于过程本体论(process ontologies)的 Web 服务发现技术,通过描述服务过程的匹配来提高服务匹配的查准率和查全率,服务的查找和发现过程就是过程模型中的过程本体论与查询中所用到的过程本体论相匹配的过程.文献[10]提出以 RDF 和 DAML 描述服务,用本体上下位关系进行服务发现.文献[11]用 DAML+OIL 描述 Web 服务,并给出相应相似度计算方法,该方法未定义输入输出和服务质量等参数,在服务发现时存在明显缺陷.文献[12]使用 LARKS 语言定义 Web 服务,并用自定义的权重网络计算 Web 服务之间的相似度,该算法中自定义的权重网络需要人工干预,面对海量的 Web 服务,构造权重网络的工作量将会成为 Web 服务发现的瓶颈.

综观上述研究,在提出各自 Web 服务发现技术时都考虑有效利用语义信息和本体论,从而在理论上给出合理的服务发现技术.但是,从实践上来看,上述方法还需要进行完善.目前,各行业的领域本体库正在构建过程中,到其完备并成为该领域的工业标准还需要很长时间.而且,即使形成比较完备的领域本体库,库中的概念实体也难以覆盖服务本体定义中所有出现的信息,如服务名称、服务提供者、服务功能等.而《WordNet》、《HowNet》、《EuroWordNet》等词汇库比领域本体库相对成熟完备一些.因此,可以用词汇语义相似度作为度量 Web 服务相似度的手段之一.使用词汇语义相似度的 Web 服务发现有两大优势:(1)服务发现基于词汇语义而不仅仅是关键词匹配;(2)能够定义服务间相似度,有利于 Web 服务查询结果的排序.

除了词汇语义相似度,还可结合服务属性相似度、参量相似度、结构相似度等多种方法对 Web 服务进行相似度度量,使用蕴涵语义的检索来解决目前 Web 服务发现中存在的问题.

3 Web 服务本体

在给出词汇语义相似度算法前,需要先给出 Web 服务本体,以明确词汇语义相似度可以操作的对象.针对不同行业不同应用各自的特点,Web 服务本体的构建会有较大的差别.但归纳来看,一个通用的 Web 服务本体大致可以分为以下四个部分:服务公共属性、服务专有属性、输入输出接口、服务质量.

定义 1. Web 服务本体模型:

$$WS=\langle CP,SP,Is/Os,QoS\rangle,$$

其中,

CP 为服务公共属性.即所有服务都必须有的属性描述,如服务 ID、服务提供者 ID、服务名称、服务提供者名称、服务功能、服务分类、服务提供者分类、联系方式、版本等.服务公共属性包括 UDDI 白页和黄页中的所有信息,同时包含指向基于文件的 URL 的不同发现机制的指针.

SP 为服务专有属性.指具体的服务所特有的属性,如药用藻类查询服务中会有产地属性(淡水或者是海水).

Is/Os 为输入输出接口集合,$Is=\bigcup Input(WS)$.Web 服务将能被更广范围的用户接受作为目标,因此其输入输出参数所使用的数据大都采用常见数据类型,如整数、字符串、日期等.Is 可表示为输入参数和数据类型的集合,如某单味药的查询服

务:$Is=\langle\{Name,int\}\cup\{Category,int\}\cup\{Useage,string\}\rangle$.但常见数据类型在一定程度上可以相互转换,并且从输入参数名称的语义上往往可以估计其采用的数据类型.而且对常见数据类型相似性度量不是本文的关键所在,因此可将 Is 简单表示为[13]:

$Is=\langle\{Name\}\cup\{Category\}\cup\{Useage\}\rangle$输出接口集合 Os 的定义与 Is 类似.

QoS 为服务质量.服务质量(QoS)是 Web 服务可用性可靠性评价的重要指标.因特网的动态性和不可预知性引起通信模式的变化、拒绝服务攻击、基础构造失效、Web 协议的低性能,迫使应用程序争用不足的网络资源,产生了对 QoS 标准的需求.Web 服务中服务质量[6,14]主要是指 Web 服务的非功能属性,主要是可用性、可访问性、完整性、性能、可靠性、安全性.

$QoS=\langle Time,Cost,Reliability,Fidelity,Security\rangle$.

在此 Web 服务本体中,输入输出接口和服务质量的相似度由参量相似度计算.服务公共属性和服务专有属性中已经在 Web 服务本体的概念实体、概念属性、概念间关系上定义的可用服务属性相似度计算,其余即可使用词汇语义相似度来计算.

4　Web 服务相似度

我们在上节定义的 Web 服务本体基础上给出 Web 服务间相似度度量的几种方式:词汇语义相似度、服务属性相似度、参量相似度、结构相似度.

词汇语义相似度 $Sim_{\text{lexical}}(s_1,s_2)$:使用《WordNet》、《HowNet》、《同义词词林》等语义知识词典,计算词语间的距离.词语间距离可以通过词汇在树状结构中相应节点间的距离或描述词汇的义原间的距离来计算.

服务属性相似度 $Sim_{\text{properties}}(s_1,s_2)$:参照建立 Web 服务本体时定义的概念实体、概念属性、概念间关系等,给出服务属性间的相似程度.上下位关系、整体部分关系、同义、反义关系等语义关系是语义细化、语义泛化、语义平级扩展等操作的基础.使语义外延和语义联想等检索算法得以实现.因此,概念实体、概念属性、概念间关系的定义是服务属性相似度计算的关键.在此基础上还可考虑不同本体系统间的本体相似度[15].

参量相似度 $Sim_{\text{parameter}}(s_1,s_2)$:对可供度量的参量,如输入参数、输出参数、服务质量等参数进行相似度计算.输入输出参数的相似程度可以从语义、类型和数值考查,服务质量参数一般从数值上进行比较,目前 UDDI 尚未支持服务质量,对此我们在文献[6]中已提出相应的解决方案.

结构相似度 $Sim_{\text{structure}}(s_1,s_2)$:对 Web 服务本体的结构进行比较,用层次递归比较或相互映射等方法给出相似度[16].

定义 2.　Web 服务间的相似度

$$SimWS(S_1,S_2)=\frac{1}{5}(SimCP(S_1,S_2)+SimSP(S_1,S_2)+SimIs/Os(S_1,S_2)+SimQoS(S_1,S_2)+Sim_{\text{structure}}(S_1,S_2)).$$

其中,

$$SimCP(S_1,S_2)=\frac{\max\sum_{i=1}^{m}\sum_{j=1}^{n}Sim_{\text{lexical}}(s_{1i},s_{2j})+\max\sum_{i=1}^{p}\sum_{j=1}^{q}Sim_{\text{properties}}(s_{1i},s_{2j})}{\min(m,n)+\min(p,q)},$$

$$SimSP(S_1,S_2)=\frac{\max\sum_{i=1}^{m}\sum_{j=1}^{n}Sim_{\text{lexical}}(s_{1i},s_{2j})+\max\sum_{i=1}^{p}\sum_{j=1}^{q}Sim_{\text{properties}}(s_{1i},s_{2j})}{\min(m,n)+\min(p,q)},$$

$$SimIs/Os(S_1,S_2)=\left(\max\sum_{i=1}^{u}\sum_{j=1}^{v}Sim_{\text{parameter}}(s_{1i},s_{2i})\right)/(\min(u,v)),$$

$$SimQoS(S_1,S_2)=\left(\max\sum_{i=1}^{k}\sum_{j=1}^{l}Sim_{\text{parameter}}(s_{1i},s_{2i})\right)/(\min(k,l)).$$

即分别计算服务公共属性、服务专有属性、输入输出接口、服务质量的相似度,再加上服务间的结构相似度.计算 $SimCP$ 时,服务 S_1,S_2 的公共属性按照已经定义的概念实体和未定义的概念实体分为两部分.在每个部分中找 S_1 和 S_2 的属性两两配对计算相似度,并取相似度和最大的配对组合,因 S_1S_2 公共属性的数量不一定相等,有未能配对的属性不参加计算.已定义的概念实体对用 $Sim_{\text{properties}}(s_1,s_2)$ 计算相似度,未定义的概念实体对用 $Sim_{\text{lexical}}(s_1,s_2)$ 即词汇语义相似度计算.$SimSP$ 的计算与 $SimCP$ 计算相似.

词汇语义相似度 $Sim_{\text{lexical}}(s_1,s_2)$ 在下一节详细

讨论.服务相似度 $Sim_{\text{properties}}(s_1,s_2)$、参量相似度 $Sim_{\text{parameter}}(s_1,s_2)$、结构相似度 $Sim_{\text{structure}}(s_1,s_2)$在文献[6]有详细的定义和讨论.

5 词汇语义相似度

词汇语义相似度在 Web 服务发现中的应用曾有文献[13,14]提及,但是都未提出具体算法,而且因为两者都只涉及英文词汇语义,所以未提《HowNet》和中文语义计算问题.本文提出两个词汇语义相似度算法,分别建立在《WordNet》和《HowNet》基础上.

《WordNet》、《同义词词林》是比较详尽的词语语义知识词典,在《WordNet》和《同义词词林》中所有同类的语义项,如《WordNet》的 Synset 或《同义词词林》中的词群构成一个树状结构,如图 1 所示.

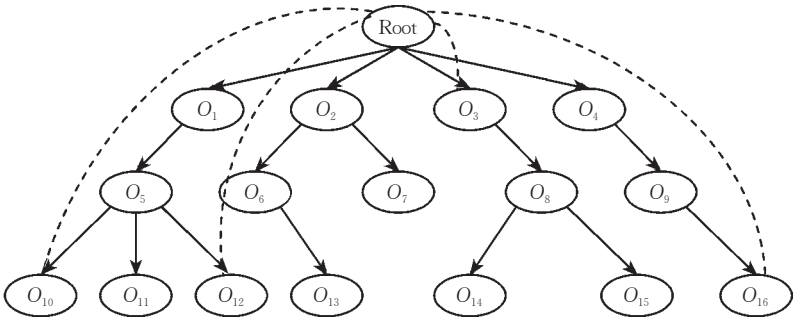

图 1 语义分类树形图

由此,主观性相当强的概念词汇语义相似度就可转换为对词语间距离的度量.两个词语的距离越大,其相似度越低;反之,两个词语距离越小,其相似程度越大.定义词语距离为 0 时,其相似度为 1;词语距离为无穷大时,其相似度为 0;相似度为词语距离的单调递减函数.

两词语 O_1 和 O_2 的相似度记为 $Sim_{\text{lexical}}(O_1,O_2)$,其词语间距离记为 $Dis(O_1,O_2)$.可得以下定义.

定义 3. 词汇语义相似度

$$Sim_{\text{lexical}}(O_1,O_2)=\frac{\alpha\times(l_1+l_2)}{(Dis(O_1,O_2)+\alpha)\times\max(|l_1-l_2|,1)},$$

其中 l_1,l_2 是 O_1,O_2 分别所处的层次,α 是相似度为 0.5 时 O_1,O_2 之间的距离,α 是一个可调节的参数,一般 $\alpha>0$.

词语的距离越大,其相似度越低,如图 1 所示,O_{10} 与 O_{16} 间的距离为 $Dis(O_{10},O_{16})=6,l_{10}=3,l_{16}=3$,因此相似度为 $Sim_{\text{lexical}}(O_{10},O_{16})=\frac{6\times\alpha}{6+\alpha}$,而 O_{12} 与 O_3 间的距离为 $Dis(O_{12},O_3)=4,l_{12}=3,l_3=1$,由定义 3 相似度为 $Sim_{\text{lexical}}(O_{12},O_3)=\frac{4\times\alpha}{(4+\alpha)\times2}$.因 $\alpha>0$,所以 $Sim_{\text{lexcial}}(O_{10},O_{16})>Sim_{\text{lexical}}(O_{12},O_3)$.

在考虑节点间的路径长度的同时还应该考虑到,词语所处的节点的深度和该深度上节点密度对相似度计算也有影响.同样距离的两个词语,词语相似度随着他们所处层次的总和的增加而增加,随着他们之间层次差的增加而减小.假设根节点为“药用植物”,O_1 为“藻类”,O_2 为“菌类”,O_{10} 为“药用海藻”,O_{12} 为“药用淡水藻”.虽然,$Dis(O_{10},O_{12})Dis(O_1,O_2)$同为 2,但 O_{10} 和 O_{12} 间的相似程度比 O_1 和 O_2 间的相似程度大,即 $Sim_{\text{lexical}}(O_{10},O_{12})=\frac{6\times\alpha}{2+\alpha}>Sim_{\text{lexcial}}(O_1,O_2)=\frac{2\times\alpha}{2+\alpha}$.因为,层次总和的增加意味着分类趋向细致,和同样词语距离的层次总和较小的词语对比较,其相似程度就越高.从实际的意义上来看也比较符合,药用海藻和药用淡水藻同藻类,只是在产地上稍有不同,而藻类和菌类各方面相差就比较大.同样可得:路径长度相同的两个节点,在节点密度高的区域比在节点密度低的区域,语义相似程度高.

词汇语义相似度除了可以通过词汇间距离来计算,还可用描述词汇的义原间相似程度来计算.《HowNet》中强调词语可以表达为一个或多个义项(概念),义项由称为“义原”(sememe)的知识表示语言来描述,义原是用于描述概念的最小意义单位.董振东先生反复强调,《HowNet》不是一个在线的词汇数据库,《HowNet》也不是一部语义词典.《HowNet》跟一般语义词典,如《同义词词林》、《WordNet》相比,最大的不同在于:《HowNet》并不是简单的将所有的“概念”归结到一个树状的概念层次体系中,而是用一系列的“义原”来对每一个“概

念”进行描述. 可以将《HowNet》看成词汇本体库,义原是概念实体,义原间的 8 种关系是概念间关系. 由这 8 种关系义原之间可以组成一个复杂的网状结构,而不是一个单纯的树状结构. 8 种义原关系中最重要的是上下位关系,上下位关系将所有的“基本义原”组成一个义原层次体系. 这个义原层次体系是一个树状结构,这也是我们进行语义相似度计算的基础.

假定服务 S_1 的服务名称 O_1 有 m 个义项 C_{11}, $C_{12},\cdots,C_{1m}$,服务 S_2 的服务名称 O_2 有 n 个义项 $C_{21},C_{22},\cdots,C_{2n}$. 假定义项 C_1 由 m 个义原描述 s_{11}, $s_{12},\cdots,s_{1m}$,义项 C_2 由 n 个义原描述 $s_{21},s_{22},\cdots,s_{2n}$. 基于义原的词汇语义相似度可以定义如下.

定义 4.　基于义原的词汇语义相似度

$$Sim_{\text{lexical}}(O_1,O_2)=\begin{cases}Sim_{\text{comcept}}(C_1,C_2), & O_1O_2 \text{ 都只包含一个义项}\\ \max\sum_{i=1}^{m}\sum_{j=1}^{n}Sim_{\text{comcept}}(C_{1i},C_{2j})x_{ij}/\max(|O_1|,|O_2|), & \text{其它}\end{cases};$$

$$Sim_{\text{concept}}(C_1,C_2)=\begin{cases}Sim_{\text{sememe}}(s_1,s_2), & C_1C_2 \text{ 都只由一个义原描述}\\ \max\sum_{i=1}^{m}\sum_{j=1}^{n}Sim_{\text{sememe}}(s_{1i},s_{2j})y_{ij}/\max(|C_1|,|C_2|), & \text{其它}\end{cases};$$

$$Sim_{\text{sememe}}(s_1,s_2)=\frac{\alpha\times(l_1+l_2)}{(Dis(s_1,s_2)+\alpha)\times\max(|l_1-l_2|,1)}.$$

其中,

$$x_{ij}=\begin{cases}1, & C_{1i}\text{和}C_{2j}\text{配对}\\ 0, & \text{其它}\end{cases};\ \sum_{j=1}^{n}x_{ij}=1,i=1,2,\cdots,m;$$

$$\sum_{i=1}^{m}x_{ij}=1,j=1,2,\cdots,n;$$

$$y_{ij}=\begin{cases}1, & s_{1i}\text{和}s_{2j}\text{配对}\\ 0, & \text{其它}\end{cases};\ \sum_{j=1}^{n}y_{ij}=1,i=1,2,\cdots,m;$$

$$\sum_{i=1}^{m}y_{ij}=1,j=1,2,\cdots,n.$$

两词汇分别有一个或多个义项,计算相似度即将两词汇义项两两配对计算相似度,取其相似度和的最大值作为词汇间相似度,即得 $Sim_{\text{lexical}}(O_1,O_2)$. 如两词汇包含义项数不等,则会出现某些义项未能找到匹配义项,计算中忽略这些义项.

义项由一个或多个义原描述,计算义项间相似度即将两义项的义原两两配对计算义原相似度,取其相似度和的最大值作为义项相似度,即得 $Sim_{\text{concept}}(C_1,C_2)$. 同样的,如两义项包含的义原数不等,会有某些义原未能找到匹配义原的,计算中忽略这些剩余的义原.

义原相似度计算建立在由上下位关系构建的树状义原层次体系结构基础上,用和词汇距离类似的义原间距离来表示义原的相似程度. 因此,义原相似度计算 $Sim_{\text{sememe}}(s_1,s_2)$ 与本节开始介绍的词汇相似度计算类似. 不同之处在于,树状层次结构上的节点一个是具体的词语,另一个是描述词语的最基本单位——义原.

引入变量 x_{ij} 是为约束词汇 O_1 中第 i 个义项 C_{1i} 只能和 O_2 中的义项匹配一次,O_2 中的第 j 个义项 C_{2j} 只能和 O_1 中的义项匹配一次. 同样,引入变量 y_{ij} 时是为约束义项 C_1C_2 中的义原只能配对一次.

在 $Sim_{\text{lexical}}(O_1,O_2)$ 和 $Sim_{\text{concept}}(C_1,C_2)$ 中都会遇到对象两两配对计算相似度并求相似度和的最大值的问题,这比较类似于组合优化问题中的指派问题(assignment problem). 目前,求指派问题的最好方法是匈牙利算法,其时间复杂度为 $O(n^3)$. 通常指派问题是求最小值,而此处配对问题需要求最大值. 因此,我们在匈牙利算法的基础上给出配对算法.

配对算法(计算 $Sim_{\text{lexical}}(O_1,O_2)$时).

1. 构造相似矩阵. 计算 O_1 中第 i 个义项 C_{1i} 和 O_2 中第 j 个义项 C_{2j} 之间的相似度,并赋值到相似矩阵的第 i 行第 j 列.

2. 用各行元素的最大值减去各行元素.

3. 将每列元素减去本列中的最小值,此时每行及每列中必然都含有零元素.

4. 从包含零最少的行(或列)开始,取一个零做上标记,划去其所在行和列.

5. 重复步 4,直到所有的零元素被标记或被划去. 如果被标记的零元素有 $\min(m,n)$个则已经求得最优解,按每个零元素所在的行列分别取 O_1O_2 的义项配对即可. 如零元素个数不够 $\min(m,n)$个,则执行步 6.

6. 对没有标记零元素的行标 $*$,对标了 $*$ 的行上所有零元素对应的列标 $*$,在对标了 $*$ 的列上所有标记了零元素的行标 $*$,直到不能再增加 $*$ 为止.

7. 将没有标 $*$ 的行与标了 $*$ 的列划上直线. 这样,我们用最少的直线覆盖了所有的零元素.

8. 找出没有被直线覆盖的所有元素中的最小元素记为 x_{ij}.

9. 对标记了 * 的行减去 x_{ij}，对标记了 * 的列加上 x_{ij}，返回步 4.

配对问题还可看成一个定义如下的特殊匹配问题：构造一个二分图 $G=(V,E)$，其中 V 分划为 S，T；S 表示 O_1（或 C_1）中的所有元素，即 $\{C_{11},C_{12},\cdots,C_{1m}\}$（或 $\{s_{11},s_{12},\cdots,s_{1m}\}$），$T$ 表示 O_2（或 C_2）中的所有元素，即 $\{C_{21},C_{22},\cdots,C_{2n}\}$（或 $\{s_{21},s_{22},\cdots,s_{2n}\}$）. 给 S 和 T 中每个元素都连一条边. 边 $\{C_{1i},C_{2j}\}$（或 $\{s_{1i},s_{2j}\}$）的权就是 C_{1i}，和 C_{2j}（或 s_{1i} 和 s_{2j}）的相似度. 这样，配对问题就等价于求二分图 G 上使总权值最大的最大权匹配问题.

6 仿真性能测试

目前，Web 服务发现性能的度量还没有统一标准：众多的 Web 服务定义方法，自由的 Web 服务比较策略、缺少专用的 Web 服务测试样本集，为 Web 服务发现方法之间的比较带来了一定的困难.

文献[7]中服务发现的依据是预先定义的服务属性本体属性值，需要定义服务属性. 文献[9,10]的服务发现基于服务本体相匹配的过程，需要定义服务本体，尤其是本体中的上下位关系. 文献[12] 用自定义的权重网络计算 Web 服务之间的相似度，需要构建权重网络. 上述方法考虑到当前公共 UDDI 上注册的 Web 服务缺少语义描述，于是都各自增加了对 Web 服务的语义描述. 上述方法中增加的语义信息对 Web 服务发现影响很大，因此我们不作比较.

文献[11] 中的服务发现方法未考虑输入输出和服务质量等参数，在服务发现时存在明显缺陷，我们也不进行比较.

由此，我们只比较当前 UDDI 上使用的基于关键词匹配的 Web 服务发现方法、文献[8]中的 Web 服务发现方法和本文中提到的两种基于词汇语义相似度的 Web 服务发现方法.

DartGrid 中包括中药方剂数据库、现代方剂数据库、中国医药产品数据库、中国药学文献数据库、中药材数据库、中药药理数据库、中药毒理数据库、中药临床药理数据库等 30 多个各种类型的异质异构分布数据库. 国内外许多中医药研究机构针对各自应用需要开发了相应数据库的数据清理、数据查询、数据分析、数据挖掘、数据整合等数据库管理 Web 服务和药理分析、毒理分析、材料分析、化学分析、方剂匹配、自动方剂生成等 Web 服务. 各 Web 服务用 DAML-S 描述，同时创建描述各服务的本体实例，并注册到 DartGrid 的 UDDI 中. 我们选取其中 110 个服务：数据清理(5)、数据查询(8 个)、数据分析(6 个)、数据挖掘(6 个)、数据整合(10 个)、药理分析(15)、毒理分析(12)、材料分析(18)、化学分析(11 个)、方剂匹配(16 个)、自动方剂生成(3 个)，作为 Web 服务发现仿真测试样本集.

我们用查准率和查全率作为度量 Web 服务发现性能的指标，查准率是指查询返回符合查询条件的 Web 服务数量与查询返回 Web 服务总数量的比率，查全率是指查询返回符合查询条件的 Web 服务与测试样本集中符合查询条件的 Web 服务的比率.

经过仿真性能测试，我们发现基于关键词匹配的 Web 服务发现方法、文献[8]中的 Web 服务发现方法和本文中提到的两种基于词汇语义相似度的 Web 服务发现方法的平均查全率为 21%，69%，89%，88%，平均查准率为 17%，61%，83%，81%. 具体测试结果如图 2，图 3 所示. 其中，基于词汇语义相似度的服务发现方法 I 计算词汇语义相似度基于词语间的距离度量，基于词汇语义相似度的服务发现方法 II 计算词汇语义相似度建立在义原相似度基础上. 在上述两种基于词汇语义相似度的服务发现方法中，我们设相似度阀值为 50%，即查询中只返回与查询条件相似度在 50%以上的 Web 服务. 相似度阀值是经验值，针对不同应用会有所调整.

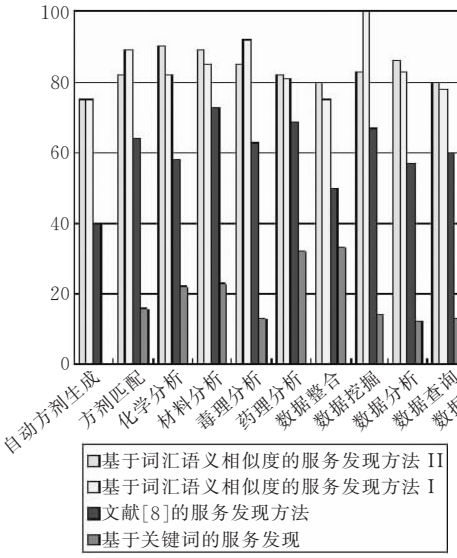

图 2 仿真性能测试查准率比较图

测试结果表明基于关键词的服务发现方法查准率和查全率都比较低，文献[8]服务发现方法的性能较基于关键词的服务发现方法有明显的改善. 基于

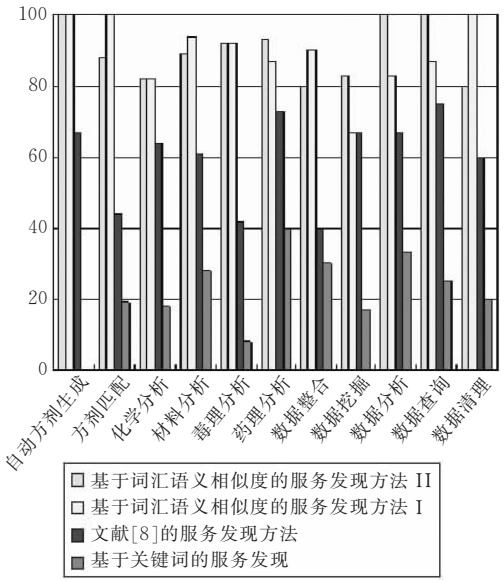

图 3　仿真性能测试查全率比较图

词汇语义相似度的两种服务发现方法查准率和查全率都达到 80% 以上,与上述两种服务发现方法相比,性能有较大的提高.

7　结　语

本文构建了 Web 服务本体,并给出两种语义相似度计算方法,以此为基础的 Web 服务发现与现有 UDDI 上基于关键词匹配的服务发现机制相比有以下优点:(1)为 Web 服务发现提供明晰的研究对象;(2)提供了基于词汇语义的 Web 服务相似度度量方法.

在国家"八六三"高技术研究发展计划资助项目(2001AA414320,2001AA113142)支持下,我们进行服务语义化描述、服务相似度计算、服务自动发现、服务合成的研究. 基于本体论的 Web 服务词汇语义相似度计算、Web 服务服务属性相似度计算、Web 服务参量相似度计算、Web 服务结构相似度计算等方法已经在 DartGrid 项目中得到应用,并取得明显的效果.

本文的研究只是一个开始,还有许多问题有待进一步研究. 如《HowNet》中定义的义原间的相互关系有 8 种,本文的义原相似度计算只利用了义原的上下位关系,如果在计算中能够充分利用其它关系,可能会得到更精细的义原相似度度量. 文中提到的 Web 服务之间的相似度是五个特征的相似度(公共属性、服务专有属性、输入输出接口、服务质量、服务间的结构相似度)的平均值,下一步我们将通过训练集取权值的方法得到各特征相似度对应的权值,来取代目前所用的平均值,以获得更精确的服务间相似度.

参　考　文　献

1 Yue Kun, Wang Xiao-Ling, Zhou Ao-Ying. Underlying techniques for Web services: A survey. Journal of Software, 2004, 15(3): 428～442(in Chinese)
(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): 428～442)

2 Paolucci M., Kawamura T., Payne T. R., Sycara K.. Semantic matching of Web services capabilities. In: Proceedings of the 1st International Semantic Web Conference (ISWC), Sardinia, Italia, 2002, 34～43

3 Wu Zhao-Hui, Chen Hua-Jun *et al*.. DartGrid: Semantic-based database grid. Lecture Notes in Computer Science 3036, 2004, 59～66

4 Deng Shui-Guang, Wu Zhao-Hui, Kuang Li, Lin Chuan, Jin Yue-Ping, Chen Zhi-Wei, Yan Shi-Feng, Li Ying. Management of service flow in a flexible way. Lecture Notes in Computer Science 3036, 2004, 428～438

5 Deng Shui-Guang, Yu Zhen, Wu Zhao-Hui, Huang Lican. Enhancement of workflow flexibility by composing activities at run-time. SAC, 2004, 667～673

6 Wu Jian. Research of Web services based product configuration technology under internet based manufacturing[Ph. D. dissertation]. Hangzhou: Zhejiang University, 2004(in Chinese)
(吴　健. 基于 Web 服务的网络化产品配置技术研究[博士学位论文]. 杭州: 浙江大学, 2004)

7 Dipanjan Chakraborty, Filip Perich, Sasikanth Avancha *et al*. DReggie. Semantic service discovery for M-Commerce applications. In: Proceedings of the Workshop on Reliable and Secure Applications in Mobile Environment, 20th Symposium on Reliable Distributed Systems, New Orleans, USA, 2001, 10: 28～31

8 Terry R Payne, Massimo Paolucci, Katia Sycara. Advertising and matching DAML-S service descriptions. In: Proceedings of the International Semantic Web Working Symposium (SWWS). Amsterdam: IOS Press, 2001, 411～430

9 Klein M., Bernstein A.. Searching services on the semantic Web using process ontologies. In: Proceedings of the International Semantic Web Working Symposium (SWWS). Amsterdam: IOS Press, 2001, 159～172

10 David Trastour, Claudio Bartolini, Javier Gonzalez-Castillo. A semantic Web approach to service description for matchmaking of services. In: Proceedings of the International Semantic Web Working Symposium (SWWS). Amsterdam: IOS Press, 2001, 447～461

11 Gonzalez-Castillo J., Trastour D., Bartolini C.. Description logics for matchmaking of services. In: Proceedings of Work-

shop on Application of Description Logics (KI 2001), Vienna, Austria, 2001, 582～586

12 Sycara K., Klusch M., Widoff S., Lu J.. Dynamic service matchmaking among agents in open information environments. SIGMOD Record. A. Ouksel and A. Sheth, 1999, 47～53

13 Cardoso J., Sheth A.. Semantic e-workflow composition. Journal of Intelligent Information Systems, 2003, 21(3): 191～225

14 Wu Jian, Dong Jin-Xiang. Qos based Web service fuzzy sort approach in internet-based manufacturing. Journal of Computer Aided Design & Computer Graphics, 17(5)(in Chinese)
(吴 健,董金祥.网络制造中 Web Service 的服务质量模糊排序方法. 计算机辅助设计与图形学学报, 2005, 17(5))

15 Rodriguez A., Egenhofer M.. Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(2): 442～456

16 Wang Y., Stroulia E.. Semantic structure matching for assessing Web-service similarity. In: Proceedings of the 1st International Conference on Service Oriented Computing, Trento, Italy, 2003, 194～207

WU Jian, born in 1975, Ph. D.. His main research interests include Web service grid computing, data mining.

WU Zhao-Hui, born in 1966, professor, Ph. D. supervisor. His main research interests include artificial intelligence, grid computing, embedded system and so on.

LI Ying, born in 1973, Ph. D., lecturer. His research interests include software architecture, compiler technology and middle ware.

DENG Shui-Guang, born in 1979, Ph. D. candidate. His research interests include Web service, workflow, middle ware.

Background

DartGrid is developed by grid computing lab of Zhejiang University, which is built upon Globus toolkit and based on OGSA/WSRF standards, and has been successfully applied in data sharing for Traditional Chinese Medicine in China. With the increasing growth in popularity of Web services and the emergence of a service-oriented view of computation on the grid, matchmaking of relevant Web services becomes a significant challenge. To cope with the limitations in current service matchmaking, the authors have developed a suite of methods which assesses the similarity of Web services, which helps users discover an appropriate service from tremendous ones more efficiently and more effectively. This paper proposes Web service ontology and indicates one similarity measurement to measuring similarity on this ontology. The method based on similarity of words which can be estimated in two ways is discussed is detail. We believe that this approach used in project will significantly improve Web services discovery.