

基于语义相似度的 Web 服务发现研究

刘克非,王 红,王卫玲

(山东师范大学 信息科学与工程学院,山东 济南 250014)

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。文中在分析现有相关研究的基础上,给出了一种基于语义相似度的 Web 服务发现方法。该方法充分利用服务中存在的语义信息,针对服务请求和广告服务中描述的功能进行匹配,并通过语义相似度来衡量两者匹配的程度。文中具体给出了服务间语义相似度的计算方法并通过示例说明了服务匹配的过程。

关键词 Web 服务;本体;语义相似度;Web 服务发现

中图分类号 TP18;TP301.2

文献标识码 A

文章编号 1673-629X(2007)02-0016-03

Research on Web Service Discovery Based on Semantic Similarity

LIU Ke-fei, WANG Hong, WANG Wei-ling

(College of Information Science and Engineering, Shandong Normal University Jinan 250014, China)

Abstract Current infrastructure of Web services discovery such as UDDI provides text and taxonomy based search capabilities. With the large growth of Web services, the current discovery mechanism becomes inefficient. Ontology and semantic based discovery of services is a promising approach to address this challenge. Proposes a semantic similarity based service discovery approach which makes good use of semantic information and concentrates on the matching of services on the basis of the capabilities in terms of inputs, outputs, preconditions and effects. The method to measure semantic similarity among services is discussed in this paper in detail.

Key words Web service; ontology; semantic similarity; Web service discovery

0 引 言

Web 服务发现是 Web 服务系统架构中的一个重要组成部分,是制约 Web 服务发展的关键技术。目前的 Web 服务架构中主要采用 UDDI(Universal Discovery, Description and Integration)技术来实现 Web 服务发现。UDDI 是一套实现 Web 服务注册中心的标准规范,它允许企业描述并注册自己的服务并提供服务的查询功能。UDDI 上的 Web 服务发现是一种基于关键词匹配的发现技术,其主要是对服务 ID、名称或服务的有限属性值进行匹配。虽然 UDDI 为服务注册和发现提供了一种有效的方法,但其技术上的局限性影响了服务发现的准确率(precision)和查全率(recall)。基于关键词匹配的 Web 服务发现具有以下局限性^[1,2]:

(1)服务描述缺乏语义信息,不能准确地表达服务的功能,不支持服务功能细节的匹配;

(2)不支持服务间灵活匹配,不能度量广告服务和
服务请求间的符合程度;

(3)不支持概念间的推理匹配,不能使用细化、泛化、平级扩展等语义操作提高查全率。

为了克服基于关键词匹配的局限性,在服务发现中引入语义描述和服务本体论进行语义层次的匹配是一个有效的解决途径。

1 语义 Web 服务描述语言 OWL-S

OWL-S^[3]是基于 Web 本体语言 OWL 的,用于语义 Web 服务描述的一个规范语言。它为 Web 服务提供了核心的标记语言结构,用于精确描述 Web 服务的功能和属性。图 1 给出了 OWL-S 服务本体的上层结构。OWL-S 服务本体由三部分组成,它们都是关于服务最本质的描述。

(1)ServiceProfile 即服务轮廓又称服务能力广告,用来描述服务是做什么的。ServiceProfile 提供了搜索服务主体所必须的信息和服务的功能描述,从而使服务请求者能够决定这个服务是否满足其需要。ServiceProfile 中最本质的部分是关于服务功能的描述,它

收稿日期 2006-05-14

基金项目:山东省优秀中青年科学家奖励基金资助项目(03bs009)

作者简介:刘克非(1980-)男,山东济宁人,硕士研究生,主要研究方向为 Web 服务发现、P2P 技术;王 红,副教授,博士,主要研究方向为 Web 服务和移动 Agent 等。

通过输入(inputs) 输出(outputs) 前置条件(preconditions) 和结果(effects) 来描述服务功能的核心内容。

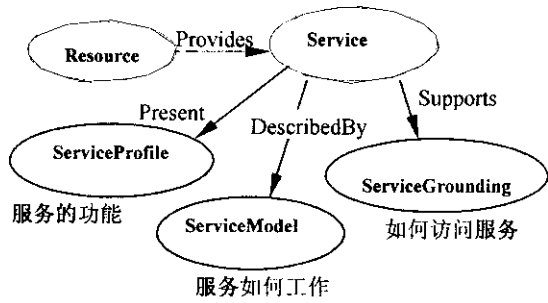


图 1 OWL-S 服务本体的主要结构

(2) ServiceModel 描述服务是如何工作的。它使得服务请求者能够进一步深入地分析以判断服务是否满足需求,并在服务执行时协调各参与者的活动。

(3) ServiceGrounding 描述如何访问服务,包括网络协议、消息格式、串行化、传输和编址等。

可以看出 OWL-S 使用 IOPEs(Inputs ,Outputs , Preconditions , Effect s) 来描述服务核心功能,这为基于功能语义相似度的服务发现提供了必要的信息。文中介绍的基于语义相似度的匹配方法主要是从 IOPE 四个方面对广告服务和请求的功能进行匹配。

2 基于语义相似度的服务匹配

文中所讲的匹配方法是对广告服务和请求中语义描述的服务功能进行匹配,并通过语义相似度来衡量两者匹配的程度。具体地说,首先分别计算广告服务和请求在输入参数、输出参数、前置条件和结果(IOPE)四个功能方面的语义相似度,例如输入参数的语义相似度为 $Sim_c(RI, AI)$ 其中 RI 指服务请求的输入参数集合, AI 指广告服务的输入参数集合;然后综合这四个方面得到两者整体上的功能语义相似度 $Sim_c(A, R)$ 根据这个衡量指标来选择最佳的 Web 服务。

由于服务的输入参数、输出参数、前置条件和结果是通过 OWL 定义的领域本体中的类进行描述的,例如输入参数集合中的某成员和本体中的某个类相关联,所以通过计算本体类集合的相似度可以得到服务功能各方面的语义相似度从而得到整体上的功能语义相似度。下面首先给出两个本体类之间的语义相似度定义,在此基础上给出服务功能语义相似度的定义。

2.1 类语义相似度

在两个类的语义相似度计算方面,路径相似度计算方法被广泛采用。路径相似度计算方法通过计算语义网络中节点之间的语义距离来得到语义相似度。由于计算节点间语义距离复杂度较高,文中参考 Tversky

的基于特征的相似性模型^[4-6],给出类之间的相似度定义。

设两个领域本体类:类 X 与类 Y 则 X 与 Y 的语义相似度定义如下:

$$Sim_c(X, Y) = \begin{cases} 1 & X = Y \\ 1 & Y \text{ 是 } X \text{ 的父类} \\ \frac{|\mu(X)|}{|\mu(Y)|} & X \text{ 是 } Y \text{ 的父类} \\ 1 & Y \text{ 是 } X \text{ 的属性} \\ \frac{1}{|\mu(Y)|} & X \text{ 是 } Y \text{ 的属性} \\ \sin(X, Y) & \text{其他} \end{cases} \quad (1)$$

(1) 两个类相同或存在同义关系,则类 X 与类 Y 的相似度为 1;

(2) Y 是 X 的父类或祖先类,类 X 通过继承最少也会具有类 Y 的所有属性,此时 X 与 Y 语义相似度为 1;

(3) X 是 Y 的父类或祖先类,类 X 的属性中可能不存在与类 Y 的某些属性对应的属性,此时 X 与 Y 的语义相似度为 X 的属性数量 $(|\mu(X)|)$ 和 Y 的属性数量 $(|\mu(Y)|)$ 的比率;

(4) 类 Y 是类 X 的属性,类 X 囊括了类 Y 能提供的全部数据, X 与 Y 的语义相似度为 1;

(5) 类 X 是类 Y 的属性,类 X 只能提供类 Y 的部分数据, X 与 Y 的语义相似度为 1 和类 Y 的属性数量 $(|\mu(Y)|)$ 的比率,这个比率表示 Y 的属性只有一个得到满足;

(6) 两个类没有直接的关联,不存在彼此包含和从属关系。根据集合理论和 Tversky 的基于特征的相似模型,采用下面的函数 $\sin(X, Y)$ 来计算 X 与 Y 的相似度:

$$\sin(X, Y) = \sqrt{\frac{|\mu(X) \cap \mu(Y)|}{|\mu(X) \cup \mu(Y)|} \times \frac{|\mu(X) \cap \mu(Y)|}{|\mu(Y)|}} \quad (2)$$

其中, $\mu(X)$ 表示类 X 的属性集, $|\mu(X)|$ 表示属性集中元素的个数。

2.2 输入、输出、前提条件和结果各集合语义相似度

接下来给出一种集合间语义相似度的计算方法,通过该算法分别计算广告服务和请求在输入、输出、前提条件和结果各功能方面的语义相似度。

计算集合 A 与集合 B 的语义相似度可归结为计算集合 A 包含集合 B 的“能力”,即计算集合 A 匹配集合 B 的“能力”。其基本思想为:对于集合 B 中的每一个元素 b ,遍历集合 A 搜索到与元素 b 的语义相似度最大的元素 a ,并记录最大语义相似度,可以认为是集合 A 与元素 b 的语义相似度。最后将集合 A 与集合 B 中

的各个元素的语义相似度按该元素的重要程度进行加权,得到 A 与 B 的相似匹配结果。详见以下算法:

```
double match( A , B )
{
    double maxSin[ ] = 0 0 ... 0 ;
    // maxSin[ ]记录 B 中各元素最大语义相似度
    double w[ ] ; // 集合 B 中各元素权重
    int n = 0 ; // 集合 B 中元素的编号
    double result ; // A 与 B 的匹配结果
    for each of [ 集合 B 中的元素 b ]
    {
        for each of [ 集合 A 中的元素 a ]
        {
            if ( semSin( a , b ) > maxSin[ n ] )
            // semSin( a , b ) 计算 a 与 b 语义相似度
            maxSin[ n ] = semSin( a , b ) ;
        }
        n ++ ;
    }
    for each of [ 集合 B 中的元素 b ]
    {
        result + = w[ n ] * maxSin[ n ] ;
        n ++ ;
    }
    return result ;
}
```

使用以上算法就可以将一个广告服务与服务请求在输入、输出、前置条件和结果四个功能方面进行语义相似匹配,得到两者在这四个方面的语义相似度。

函数 $\text{Sim}_i(RI, AI)$, $\text{Sim}_o(AO, RO)$, $\text{Sim}_p(RP, AP)$, $\text{Sim}_e(AE, RE)$ 分别用来计算广告服务与服务请求在输入、输出、前提条件和结果各方面语义相似度。先给出 $\text{Sim}_i(RI, AI)$ 的定义如公式(3)所示:

$$\text{Sim}_i(RI, AI) = \begin{cases} \sum_{j=1}^n w(AI_j) \times [\max_{k=1 \dots m} (\text{Sim}_c(C, RI_k), C, AI_j))] & AI \neq \emptyset, RI \neq \emptyset \\ 0 & AI \neq \emptyset, RI = \emptyset \\ 1 & AI = \emptyset \end{cases} \quad (3)$$

式中, RI 表示服务请求的输入集合; AI 表示广告服务的输入集合; 函数 w 是指对输入参数取对应的权值; 函数 C 是对输入参数取其关联的本体中的类; n, m 分别表示 AI, RI 中元素个数。

函数 $\text{Sim}_o(AO, RO)$, $\text{Sim}_p(RP, AP)$, $\text{Sim}_e(AE, RE)$ 与函数 $\text{Sim}_i(RI, AI)$ 的定义类似, 其中 AO, AP, AE, RO, RP, RE 分别表示广告服务的输出集合、前提条件集合、结果集合和服务请求的输出集合、前提条件集合、结果集合。

2.3 整体功能语义相似度

在有了输入、输出、前提条件和结果各方面语义相似度定义后, 通过综合这四个方面来评估广告服务和请求两者整体功能上的语义相似度。常用的综合方法是求加权平均值或几何平均值, 这里采用几何平均来计算最后的功能语义相似度。定义如(4)所示:

$$\text{Sim}_f(A, R) = \frac{1}{4} [\text{Sim}_i(RI, AI) + \text{Sim}_o(AO, RO) + \text{Sim}_p(RP, AP) + \text{Sim}_e(AE, RE)] \quad (4)$$

式中, A, R 分别表示广告服务和请求。

2.4 示例

下面举一个简单的例子介绍服务请求和广告服务之间功能语义相似度的计算过程。为了简化计算, 在例子中只对广告服务和请求在输入和输出两个功能方面进行语义相似度计算。假设查找一个销售个人电脑的服务, 服务请求用 Req 表示, 它的输入为 Price , 输出为 PC 。现在服务注册中心存在一个销售笔记本电脑的服务广告, 用 Adv 表示, 它的输入为 Price , Processor , Memory , 输出为 Laptop 。如图2所示, 这里的输入输出指的是与服务输入输出参数相关联的领域本体中的类, 下面来计算 Adv 和 Req 的功能语义相似度。

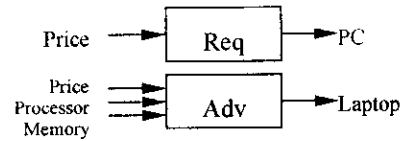


图2 服务 Req 和 Adv 的简单描述

领域本体是实现语义发现的基础, 这里给出计算机领域本体的一部分, 如图3所示。

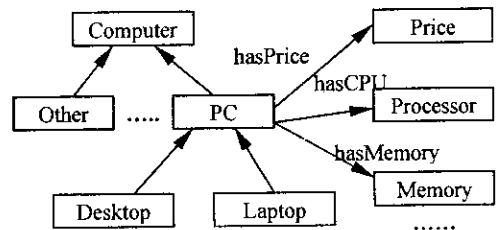


图3 计算机领域部分本体图

下面介绍 Adv 与 Req 之间的服务功能语义相似度的计算过程, 这里默认所有参数的权重均相同。

第一步, 根据式(1)得到 Adv 与 Req 输入/输出参数关联类的语义相似度:

$$\begin{aligned} \text{Sim}_i(\text{Price}, \text{Price}) &= 1; \\ \text{Sim}_i(\text{Price}, \text{Processor}) &= 0; \\ \text{Sim}_i(\text{Price}, \text{Memory}) &= 0; \\ \text{Sim}_o(\text{Laptop}, \text{PC}) &= 1. \end{aligned}$$

第二步, 根据 $\text{Sim}_i(RI, AI)$ (式(3)中已给出) 和 $\text{Sim}_o(AO, RO)$ 得到 Adv 与 Req 在输入集和输出集的语义相似度: (下转第22页)

3.2 实验结果

实验结果如图 4 所示。

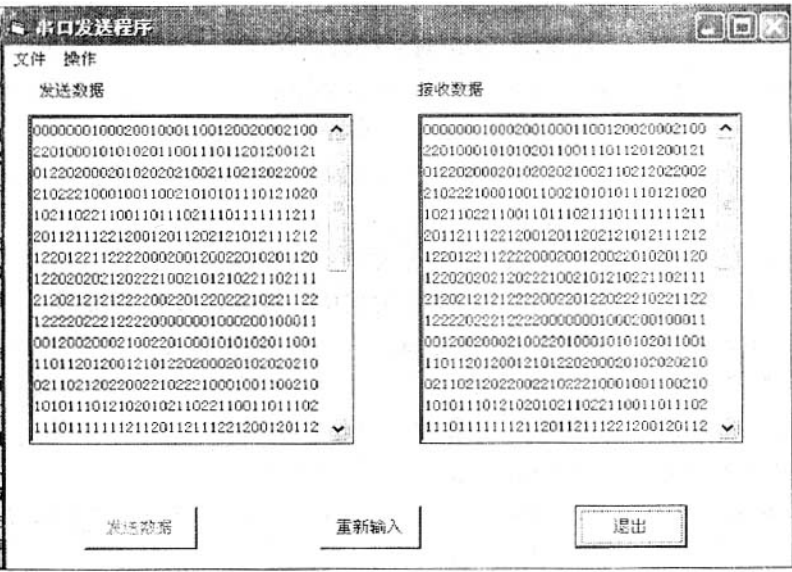


图 4 试验结果

在图 4 中“发送数据”栏中的 0,1,2 是存储在半导体存储器中的数据,共有 1024 个数据将分别用来控制编码器形成 1024 位三态光信号,这些三态光信号被百位三态光解码器(这项工作由课题组其他成员完成)接收,译码后形成输出数据,这些数据显示在图 4 的“接收数据”栏中。图 4 显示的实验结果表明:编码前的数据和解码后的数据一致,证明了三值光计算机百位量级编码器实验成功。

(上接第 18 页)

$$\text{Sim}_I(\text{Req}I, \text{Adv}I) = 0.33;$$

$$\text{Sim}_O(\text{Adv}O, \text{Req}O) = 1.$$

第三步,根据式(4),得到 Adv 与 Req 整体上的功能语义相似度:

$$\text{Sim}_f(\text{Adv}, \text{Req}) = 0.665.$$

例子中为简化计算只涉及到输入和输出两个功能方面,这和涉及四个功能方面的匹配原理是一样的。

3 结束语

随着 Web 上服务数量的日益增多,迫切要求更准确更完善的服务发现方法。文中提出了一种基于语义相似度的 Web 服务发现方法,该方法在服务匹配过程中充分利用了服务中存在的语义信息,通过结合本体技术针对服务的功能进行语义相似度计算来定位服务。这种在语义层上对服务功能进行匹配的方法避免了关键词匹配技术中的缺陷,很大程度上提高了服务检索的查准率和查全率。

4 结 论

实验结果验证了百位量级的三值光信号的编码器和解码器原理,证实了用液晶阵列和偏振片可以构成三值光计算机的数百位编码器。为下一步的百位量级逻辑运算器、百位量级半加器的实现打下了坚实的基础。

参考文献:

[1] Mikats P A, Betzos G A, Irakliotis L J. Optical processing paradigms for electronic computers[J]. Computer, 1998, 31(2): 45-51.

[2] 金 翊,何华灿,吕养天.三值光计算机基本原理[J].中国科学 E 辑, 2003, 33(2): 111-115.

[3] 金 翊,何华灿,吕养天. Ternary Optical Computer Principle[J]. Science in China (Series F) 2003, 40(2): 145-150.

[4] 严军勇,金 翊,孙 浩.三值光计算机多位编码器与解码器的可行性实验研究[J].计算机工程, 2004, 30(14): 175-177.

[5] 孙 浩,金 翊,严军勇.三值光计算机编码器试验原理的试验研究[J].计算机工程与应用, 2004, 40(16): 82-83.

[6] 黄伟刚,金 翊,严军勇,等.三值光计算机百位编码器的设计与构造[J].计算机工程与科学, 2006(4): 139-142.

参考文献:

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

[2] 吴 健,吴朝晖,李 莹,等.基于本体论和词汇语义相似度的 Web 服务发现[J].计算机学报, 2005, 28(4): 595-601.

[3] Martin D. OWL-S: Semantic Markup for Web Services [EB/OL]. 2004. <http://www.daml.org/services/owl-s/1.0/owl-s.pdf>.

[4] Amos T. Feature of similarity[J]. Psychological Review, 1977, 84(4): 327-352.

[5] 钟福金.语义 Web 服务发现及其应用研究[D].合肥:合肥工业大学, 2005.

[6] Rodriguez A, Egenhofer M. Determining Semantic Similarity Among Entity Classes from Different Ontologies[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(2): 442-456.