一种基于P2P的两阶段Web服务发现研究

龚世忠[1],唐文忠[2]
(1.北京航空航天大学 计算机学院 网络技术北京市重点实验室,北京 100083;
2.北京航空航天大学 计算机学院,北京 100083)

摘 要:文中针对Web服务发现的两个难题提出了一个新的基于P2P的两阶段Web服务发现机制(TSBP)。本机制在使用标准的WSDL服务描述语言规范的基础上,提出一种结合向量空间算法和结构匹配的两阶段Schema匹配方法度量服务的相似性,同时采用基于P2P的分布式的服务信息交换机制代替原有的UDDI集中存储方式。在第一阶段的服务搜索中,使用IR技术将各种服务进行粗分类得到第二阶段匹配的候选集,在第二阶段匹配中,提取Web服务的WSDL文档的树状结构,使用改进的编辑距离树算法对候选集进行进一步匹配。最后通过实验来和其他的服务搜索方式进行对比,验证本机制的有效性。

关键词:向量空间;Web服务发现;Schema匹配;P2P;编辑距离
中图分类号:TP311.11 **文献标识码**:A **文章编号**:1673-629X(2010)06-0121-05

A Web Services Discovery Research of Two Stages Based on P2P

GONG Shi-zhong[1], TANG Wen-zhong[2]
(1.Beijing Key Laboratory of Network Technology, School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100083, China;
2.School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: The key and most difficult problem in web services discovery is how to exactly describe services and exchange the services meta data, in this paper, a web services discovery research of two stages based on P2P mechanism (TSBP) is discussed in which VSM (Vector Space Model) and structure matching is combined to measure the similarities of web services. At the same time, the services meta data is changed in P2P network instead of the UDDI. At first stage, information retrieve method is adopted to presort the web service. At second stage, an operation included in a web service is modeled as an unordered labeled tree, then a modified edit tree distances algorithm is adopted to further match the candidate web services. At the end of this paper, experiments are made to compare with other web services search method.

Key words: VSM; web services discovery; Schema matching; P2P; edit distance

0 引 言

Web服务发现中面临的主要难题[1]:一个是如何准确细致地刻画服务能力,从而支持用户需求和服务描述之间更精确的匹配操作;另一个是如何存储、索引、交换服务元数据,既保证服务发现的搜索广度,又将搜索时间限定在用户可接受的范围内。

为解决第一个问题,目前研究的热点是将"语义"[2]概念引入Web服务,通过对外服务功能尽可能准确的描述来实现服务的有效查找。但是这对目前已大量存在的Web服务来说是不现实的,不仅要增加OWL-S、WSMO等协议进行服务描述,更重要的是这些技术目前尚不成熟,还不能投入商业运行。

而目前作为解决第二个问题方面的一个方案UDDI,需要服务提供者将自己的服务手动注册到服务注册中心,而服务使用者需要在该中心按照商业分类来定位到该类别下面,然后使用关键字搜索到自己需要的服务。显然,Web服务搜索是一个相当精细的过程,按照商业分类和关键字搜索太粗糙了,而且UDDI集中式存放和搜索有搜索广度和以及可靠性难以保证、维护困难等问题。

文中针对上面提到的两个方面的问题提出一种新的Web服务发现机制TSBP,它结合IR(Information Retrieval)技术和结构匹配两阶段综合匹配方法进行

收稿日期:2009-10-13;**修回日期**:2010-01-19
基金项目:北京市教育委员会共建项目(JD100060630)
作者简介:龚世忠(1979-),男,湖北襄樊人,工程师,硕士研究生,从事电子政务、Web服务的研究;唐文忠,研究员,从事电子政务、虚拟组织研究。

服务相似度度量,使用 P2P 网络作为查询和响应信息的交换,具有以下的特点:

1)服务提供者不再需要去手工注册服务,简化了服务发布工程;

2)充分发掘现有 WSDL 文件的潜力,抽取文本和结构两方面的服务描述信息,克服语义 Web 服务的高代价问题;

3)提出一个可以支持 Web 服务操作级别相似性搜索算法;

4)能够对返回的结果进行相似度综合排序;

5)借助自组织的分布式 P2P 查找网络,克服集中式 UDDI 的信息一致性维护的难题。

1 系统总体结构

文中所述的 TSBP 服务发现机制是借助 P2P 网络的自组织能力构建的分布式查找网络实现的,服务发布和服务查找均采用统一的客户端。该客户端嵌入了 P2P 通信模块、服务匹配模块、特征抽取模块、服务监测模块等四大部分功能。客户端之间使用 P2P 通信模块组成 P2P 网络并进行消息通信和数据传输。特征抽取模块将该服务描述文件进行预处理。服务匹配模块使用两阶段查找法找到相似度最高的 Web 服务集,并对相似程度进行排序,返回给用户一个经过排序的结果集。查找成功后,通过 SOAP 调用完成对服务的使用。系统结构如图 1 所示。

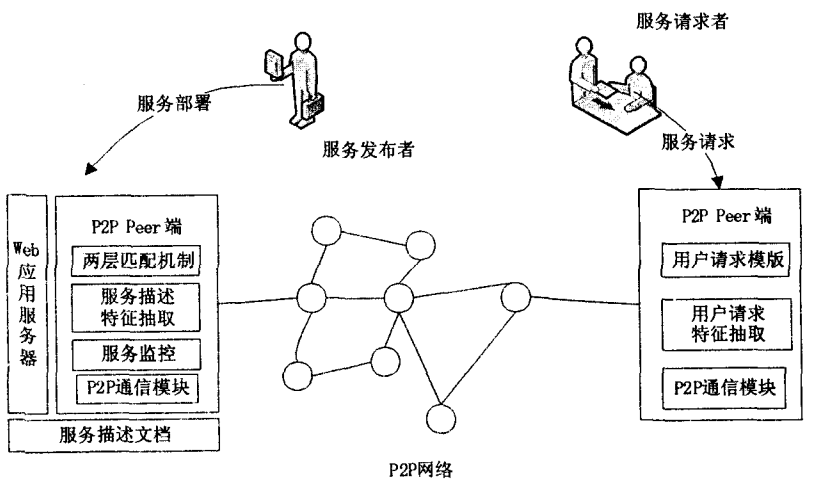

图 1 系统总体结构图

2 Web 服务匹配

文中所述的 Web 服务匹配是为了最大限度地利用符合 WSDL 规范的 Web 服务描述文件来更精确、更自动化地实现 Web 服务发现。为了这个目的,需要对 WSDL 服务描述文档进行解析,从中抽取本文结合了信息检索技术的向量空间模型和结构匹配算法来实现 Web 服务的匹配。

2.1 Web 服务描述文档分析

每个服务都有一个相应的 WSDL 文件来描述服务的功能和接口。每个 Web 服务都由一系列的操作组成。每个操作形成一个逻辑上的树状模型,其中根节点代表该操作,第二层的节点代表操作调用的消息,第三层节点代表消息调用的参数。根节点的子树包含两类子节点:一类是输入消息的子节点,一类是输出消息的子节点。第三层 Data type 不是叶子节点,而是基于 XML Schema 构建的,具有规范良好的结构信息的模式树,如图 2 所示。可以将 Web 服务定义为一个三元组 WS = (DtSet, MsgSet, Opset)[3]。DtSet 是 data type 的集合,MsgSet 是用 data type 定义的消息集合,Opset = {Opi(Inputi, Outputi)}是操作的集合,$Input_i$ 和 $Output_i$ 是用来在操作间进行交换的数据消息参数。每个 Web 服务操作是一个多输入多输出的函数。

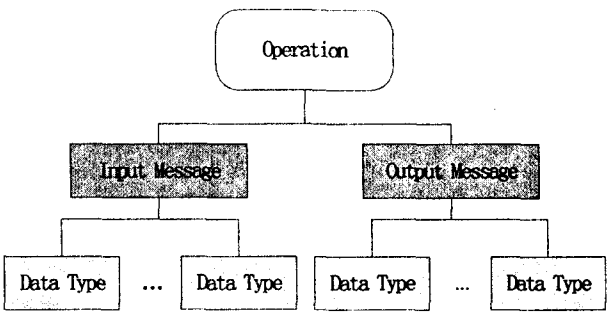

图 2 Web 服务操作树结构

从上面的分析可以看出,Web 服务描述文档可以分为两个部分:一部分是对 Web 服务及其操作的说明,包括文本描述,操作、输入和输出参数的名称;另外一部分是操作及其参数的层次结构。

2.2 使用向量空间的第一阶段匹配

向量空间模型[4]将文档分割成一组关键字,每个关键字表示向量空间中的一维,这样,一个文档就可以表示成这个"项空间"中的一个向量。将 Web 服务定义为一组操作的集合,尽管这些操作都被 WS 所支持,它们所完成的任务有时并不属于同类,因此,在操作级上对服务进行相似性计算级上将更有意义。WSDL 文档项集的抽取与无结构的文档有所不同,根据 WSDL 的标签含义解析文档,抽取 Service 和 Operation 的<documentation>标签下的内容,以及服务名称、操作名称和参数名称。将获得的信息作分词处理,并做去除停词(stop words)和取词干(Stemming)等预处理,最终构成文档特征词汇集合。然后参照传统 VSM 技术的

统计词频,构造向量,以及计算向量夹角余弦值,将 Web 服务按照统计语义的相似程度进行粗分类。

2.3 基于 WSDL 文档结构匹配的第二阶段匹配

在图 2 所示的模型中,Data Type 子树的节点根据它们的标签分成两类:一类是原始类型,像 Int 和 String 类型;另外一类是复合类型,复合类型往往携带了结构和内容的信息,称其为 Constraints[5]节点,包括 Sequence 节点、Union 节点和 Multiplicity 节点。直观地说,两个 Web 服务操作相似是通过它们有相似的输入/输出的 Data Type,因此,Web 服务操作匹配的问题就转换为 schema 树的匹配问题。在度量树的相似程度的研究成果中,树编辑距离[6]是一种有效的方法。树的编辑一般有三种操作:i)插入;ii)删除;iii)改写。一棵树可以通过若干步这样的操作变成另外一棵树,每种操作都有一定的代价,所有的这些操作的代价总和就是两棵树的距离,代价越小的两棵树代表着它们有更高的相似度。但是 WSDL 文档有它的特殊性,刚才提到的 Constraint 节点,传统的树编辑距离算法无法处理这类节点,必须对这类节点变形。

定义 1　编辑距离映射[6]。假设 $t[i]$ 是树 T 上的第 i 个节点,定义三元组 (M, T_1, T_2) 为从 T_1 到 T_2 的满足约束的编辑距离映射,其中 M 是有序整数对 (i, j) 的集合,$|T_1|$,$|T_2|$ 分别是树 T_1 和 T_2 的节点数目。满足如下条件:

1) $1 \le i \le |T_1|, 1 \le j \le |T_2|$;

2) M 是一个编辑距离映射;

3) 对于 M 中的任何三对 (i_1, j_1),(i_2, j_2) 和 (i_3, j_3),令 $t_1[I]$ 为 $t_1[i_1]$ 与 $t_2[i_2]$ 的最小共同祖先,$t_2[J]$ 为 $t_2[j_1]$ 与 $t_2[j_2]$ 的最小共同祖先,则 $t_1[I]$ 是 $t_1[i_3]$ 的真祖先(即不包括本身)当且仅当 $t_2[J]$ 是 $t_2[j_3]$ 的真祖先。

定义 2　编辑距离映射的代价[7]。令 λ 为空节点,I 和 J 表示 T_1 和 T_2 的节点集合,集合内部的节点都未被包含在映射 M 中,则 M 的代价定义为:

$$\gamma(M) = \sum_{(i,j)\in M} \gamma(t_1[i] \to t_2[j]) + \sum_{i\in I} \gamma(t_1[i] \to \lambda) + \sum_{j\in J} \gamma(\lambda \to t_2[j])$$

定理 1　两棵树的编辑距离等于这两棵树的最小编辑映射距离的代价,即

$$D_e(T_1, T_2) = \min_{M_e} \{\gamma(M_e) \mid M_e \text{ 是 } T_1 \text{ 到 } T_2 \text{ 的编辑距离映射}\}$$

传统的树编辑距离算法不能适用于 XML schema 树,因为它不能处理约束节点。提出三种转换方法[8]来处理这个问题。将转换 Sequence, Union 和 Multiplicity 节点为标签节点。

1) 拆分:对于 Sequence 节点 $t = [t_1, t_2, \cdots, t_s]$ 将 t 节点拆分成 s 个节点 $t_1, t_2, \cdots, t_s$,原来的 t 节点被这 s 个节点代替。

2) 合并:对于 Union 节点,将所有的子节点合并成一个子节点,并用这个子节点代替原来的节点。

3)删除:对于 Multiplicity 节点,删除原来的节点,用子节点代替原来的节点。

由于在 XML Schema 的定义中,complex 类型的节点能够互相嵌套,因此这三种变形规则需要从底向上一层一层地递归地实施。

3 服务查找网络的构造及查找过程

传统的 UDDI 是集中式拓扑,依赖于多个数据库定期复制来保持数据的一致性,集中式拓扑可扩展性差、容易单点失效并易受攻击。TSBP 引入 P2P 环境来处理服务的部署与查询,根据用户的服务描述和 P2P[9]的自组织能力对服务进行自动分类,既提高了效率和便利性,也克服了集中式结构的局限性。

TSBP 采用一种半分布式 P2P 结构[10],吸取了中心化结构和全分布式非机构化拓扑的优点。具体做法是服务相似度较高的节点自行组织为一个组,同时选举一个组长,组长负责组员的管理、组内的服务监控、组内查询和组间的查询转发。所有的组长组成 P2P 网络进行交换,组员并不加入 P2P 网络,组长和组员组成一个集中式的管理体制,至此,整个查找网络分为两层(如图 3 所示)。

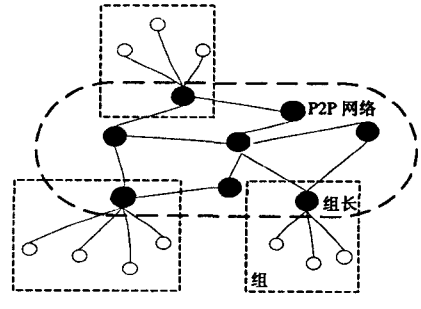

图 3　P2P 网络层次结构

搜索的范围从整个网络缩小到某个或某几个组,从而达到更高的搜索效能。

3.1 P2P 查找网络的构造

在 TSBP,客户端嵌入了 P2P 模块,但不是每个客户端都会加入 P2P 网络,而是根据选举算法和相似度计算决定是否加入。

组网步骤如下:

1)当有一个新的服务提供者部署服务后,使用 2.2 节描述的向量空间方法将服务描述抽取成向量。

2)此服务提供者作为一个 peer(假设为 A)查找并加入现有的 P2P 网络。

3)如果查找不成功,说明 P2P 网络还没有构建,将自己作为第一个 P2P 节点并自举为组长。

4)如果查找并加入成功,将自己的服务描述向量和全 P2P 网络的节点做第一阶段的相似度衡量,如果有相似度大于阈值的节点,找到相似度最高的节点(假设为 B),然后 A 退出 P2P 网络并将自己作为 B 的组员。将自己的位置信息、服务描述向量信息注册到 B 的信息库里。如果找不到相似度大于阈值的节点,转到第 5 步执行。

5)如果找不到相似度大于阈值的节点,A 作为一个新的 P2P 节点加入网络,并自举为组长,然后建立路由表,建立和其他节点的关系。

从以上构造过程可以看出,本服务查找机制将查找过程的一部分工作提前到服务部署的时候进行,即在服务部署时就将服务做相似性分组,查找服务时直接定位到找到相似度较高的分组里面查找。

组长负责对本组成员的服务监控,一旦发现组员失效,则将其从自己的注册信息库中删除。每个组再选举一个副组长,作为组长的备份,组长的注册信息库定期向副组长复制,当组长失效时,副组长将接替组长的职能。

3.2 服务查找

服务请求者和服务提供者使用同样的客户端程序,查找过程如下:

1) 首先将用户的请求通过模版规范化后抽象成向量。

2) 检查自己是否在查找网络中,如果是,将查找请求提交给所在组的组长,并再转到第 3 步执行;如果不是,转到第 5 步执行。

3) 组长收到查找请求后,将查找请求和自己的特征向量做比较,如果相似度大于某一阈值,则在自己的注册信息库内查找.同时将查询请求转发到路由表里和查询请求相似的若干个组里进行第二阶段的查找,并将查找到的服务信息返回给请求者;否则,转到第 4 步执行。

4) 组长向路由表里面和查询请求相似度最低的 n 个组转发查找请求,当有组长收到请求后,转到第 3 步执行。查询在达到转发最大次数时终止。

5) 如果查找用户没有在网络中,则找到离自己最近的组中,并加入该组,并返回第 2 步继续执行。

组长之间的 P2P 网络的通信、组织、节点的加入、退出等功能可以采用成熟的 P2P 算法,在此不做详细说明。

4 性能评价与分析

为检验 TSBP 的有效性,设计一个原型系统进行实验,该系统使用 J2SE 1.5 实现,使用开源 wsdl4j 对 WSDL 文件进行解析,抽取需要的数据项并构造向量空间进行相似度比对,使用 JXTA[11]规范进行 P2P 网络的构建。

目前还没有对服务发现机制的一个公认评价标准,在此借用传统信息检索使用的查全率[12]和查准率作为评价标准。

$$\text{查准率}=\frac{\text{发现的满足查询请求的服务数}}{\text{发现的服务总数}}\times 100\%$$

$$\text{查全率}=\frac{\text{发现的满足查询请求的服务数}}{\text{与查询相关的服务总数}}\times 100\%$$

通过在 Google, Strikeiron 和 Xmethods 收集到的 142 的服务,共 623 个操作,将这些操作分为三大类,在这三个类别中进行查全率和查准率的实验,通过与基于操作向量空间算法和基于编辑距离树的两种算法比对,可以得到如图 4 和图 5 的结果。

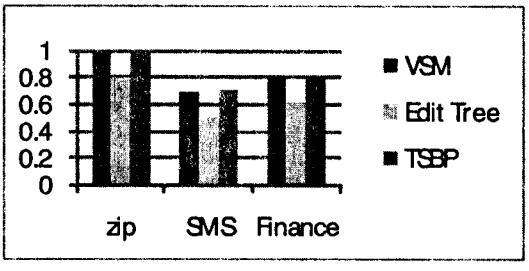

图 4 查全率比较

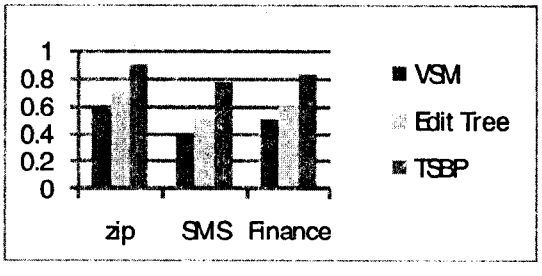

图 5 查准率比较

从图 4 中可以看出,在查全率方面,向量空间算法由于算法简单,因而检索范围大,有较高的查全率,而文中提出的方法在第一阶段匹配中正是基于向量空间算法,因而有着类似的查全率。从图 5 中可以看出,编辑距离树算法由于融合结构的特征因而比向量空间算法有较高的查准率,而文中的方法因为结合了这两种方法,有着更高的查准率。在本系统中服务部署阶段生成查找网络结构的过程并不影响查全率,P2P 网络的路由机制可以保证在一定的函数量级时间内遍历所有的节点,查准率依赖于服务匹配算法,实验表明,经过改进的向量空间算法可以有效地提高服务发现性能。在查询时间方面,本系统已经预先将相似的服务

组织在相同的组里面，而这种预先组织是在服务部署的时候进行的，这几乎不会增加服务查询时间。在鲁棒性方面，组长集成了服务监控模块，可以定期探查组员的状态，同时每个组有副组长，在组长失效时，可以接管组长的职能，提高了可靠性。

5　结束语

文中针对现有的 Web 服务发现机制的问题提出了一种基于向量空间和 P2P 架构的服务搜索机制，该机制改进了传统的通过关键字查找服务的精确性，在查找结构上，提出了将查找过程先行准备的观点，在保证查准率的基础上提高了查找速度。基于 P2P 的分布式结构有着集中式结构所没有的优点，但 P2P 机构的“非中心化”、非权威性同样会带来信息不可控、安全机制有待加强的特点。下一步研究的重点应在服务的自动组合和执行以及 Web 服务发现机制的安全问题上。

参考文献：

[1] 陈德伟，许　斌，蔡月茹，等. 服务部署与发布绑定的基于 P2P 网络的 Web 服务发现机制[J]. 计算机学报，2005，28(4)：615－626.

[2] Sivashanmugam K，Verma K，Sheth A，et al. Adding semantics to Web services standards[C]//In：Proceedings of the 1st International Conference on Web Services (ICWS'03). Las Vegas，Nevada：[s. n.]，2003：395－401.

[3] Wang Y，Stroulia E. Flexible interface matching for Web－service discovery[C]//Proceedings of Fourth International Conference on Web Information Systems Engineering. Roma，Italy：[s. n.]，2003：10－12.

[4] 陈江锋，于建军. 基于扩展 VSM 的 Web 服务发现[J]. 计算机工程，2008(12)：25－27.

[5] Shvaiko P，Euzenat J. A Survey of Schema－based Matching Approaches[J]. Journal on Data Semantics，2005(4)：146－171.

[6] Zhang K. A constrained edit distance between unordered labeled trees[J]. Algorithmica 1996，15：205－222.

[7] Zhang Kaizhong，Shasha D. Simple fast algorithms for the editing distance between trees and related problems[J]. Siam J. Comput，1989，18(6)：1245－1262.

[8] 何玲娟，刘连臣，吴　澄. 一种改进的基于 WSDL 描述的操作相似性度量方法[J]. 计算机学报，2008(8)：1331－1339.

[9] 胡放明，李俊兵，贺贵明. 对 P2P 网中发现机制的研究[J]. 计算机应用，2004，24(2)：521－524.

[10] 江武汉，叶从欢，孙世新. P2P－Grid 结构模型研究与设计[J]. 计算机技术与发展，2006，16(2)：135－138.

[11] 张　智，李瑞轩. 基于 JXTA 的 Web 服务发现模型研究[J]. 计算机工程与应用，2005，41(9)：137－139.

[12] Gudivada N，Raghavan V，Grosky W，et al. Information Retrieval on the World Wide Web[J]. IEEE Internet Computing，1997，1(5)：58－68.

(上接第 120 页)

粒子数目不超过 10000 个时，该方法能达到较好的实时显示效果。

表 1　绘制速率对比表

粒子数目(个)	3000	10000	20000
帧频率(帧/s)	108	65	31

5　结束语

基于粒子系统的不规则物体模拟是计算机仿真中的一项复杂课题。文中以实时性、逼真度出发提出了一种模拟陨石爆炸的简单方法。该方法取得了较好的实验结果，在调整陨石粒子数目的时候可以看到陨石爆炸的变化。在今后的研究中，可以考虑陨石粒子间相互碰撞的效果，在真实光照条件下建立模型，实现更加逼真的模拟效果。

参考文献：

[1] Reeves W T. Particle Systems－a Technique for Modeling a Class of Fuzzy Objects[J]. Computer Graphics，1983，17(3)：359－376.

[2] 张　芹，张　健，闵建平. 提高粒子系统实时性的方法研究[J]. 计算机工程，2003，29(18)：46－48.

[3] 万华根，金小刚，彭群生. 基于物理模型的实时喷泉水流运动模拟[J]. 计算机学报，1998，21(9)：774－779.

[4] 陈应松，胡汉春，肖世德. 基于 OpenGL 纹理映射技术实现动态图像的应用[J]. 计算机仿真，2004，21(2)：130－132.

[5] Shreiner D，Woo M，Neider J，et al. OpenGL 编程指南[M]. 徐　波，等译. 北京：机械工业出版社，2006.

[6] 管　宇，邹林灿，陈　为，等. 基于粒子系统的实时瀑布模拟[J]. 系统仿真学报，2005，16(11)：2471－2474.

[7] Sims K. Particle Animation and Rendering Using Data Parallel Computation[J]. Computer Graphics，1990，24(4)：405－413.

[8] Fosts N，Metaxas D. Realistic Animation of Liquids[J]. Graphical Models and Image Processing，1996，58(5)：471－483.

[9] 赵静谧，张　慧，郑国勤. 基于粒子系统的喷泉模拟[J]. 计算机应用研究，2006(1)：244－246.

[10] Angel E. OpenGL 程序设计指南[M]. 张文祥，李桂琼，译. 北京：清华大学出版社，2005.