一种基于P2P的语义Web服务发布和发现模型

陈星豪,李陶深
(广西大学 计算机与电子信息学院,广西 南宁 530004)

摘 要:服务发布和发现是Web服务应用中极其重要的环节。随着Web服务数目的增长和对实时性需求的增加,分布式服务发布和发现机制成为Web服务应用的一个新研究方向。对当前Web服务发现问题研究进行了探讨,提出了基于P2P的Web服务发布和发现的工作模型,并以实例说明了其工作过程。该模型采用两层结构:第一层是纯Peer-Peer连接,通过一定的路由机制保证通信的畅通,有效地支持系统的可扩展性;第二层采用集中式的连接方式,各个Peer节点集中连接到指定的Broker,以方便统一管理和维护,提高系统的查询效率。

关键词:语义Web服务;P2P;本体
中图分类号:TP313 **文献标识码**:A **文章编号**:1673-629X(2008)01-0077-03

A P2P-Based Semantic Web Services Publishing and Discovery Model

CHEN Xing-hao, LI Tao-shen
(School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

Abstract: Service publishing and discovery play a very important role in Web service applications. As large growth in number of Web servicesing and the demand for registries, the mechanism of distributed Web services publishing and discovery has become a new research activity in Web service applications. In this paper, a Web services publishing and discovery model based on P2P technology is proposed, and the work process is illustrated by instance. There are two level in this model. The first level is Peer-Peer connection way, and its routing mechanism guarantees the connection smoothly and supports the system extension effectively. The second level is central connection way, in which each Peer connect to the corresponding Broker directly, so that can make the management and maintenance conveniently, and also improve the search efficiency of the system.

Key words: semantic Web service; P2P; ontology

0 引 言

Web服务是对当前Web的一种扩展,是一种独立于平台和实现的软件构件。它能用服务描述语言来描述,在服务注册处发布;它通过标准的机制,在运行或设计时被发现;它可以与其他服务结合,通过一种应用集成的方式快速地发布服务应用。

传统的基于UDDI的Web服务工作机制并不能适应Web服务越来越迅速的发展,主要原因在于:UDDI的查询机制是基于关键字匹配,不能表达相关的语义信息,即相同的关键字结合不同的语境可能会有不同的语义,而不同的关键字也可能会有相同的语义。而且,集中式的UDDI注册机制又使得系统会因服务器出现故障而导致全面瘫痪。这说明随着Web服务数目的增长和对实时性需求的增加,传统的基于关键字查询技术和集中式的注册机制的服务工作机制已经难以满足实际需求,于是分布式的且基于本体的Web服务技术应运而生。为了克服关键字带来的查全率和查准率低的问题,引入语义Web对Web服务进行扩展,已经成为学术界共识[1]。

当前对P2P底层结构的研究基本分为两大类:第一类以FreeNet、Guntella为代表采用非结构化连接的底层结构[2,3]。事实证明这类结构对可扩展性缺乏有效的支持,当节点的数目增大时,系统的性能急剧下降;第二类以Chord, Pastry, Tapstry和CAN为代表,采用结构化底层结构互连[2,3]。这类结构对可扩展性和稳定性提供了可靠的保证,但是其资源发现模型没有对资源进行有效的分类,使得每个查询都会分派到大量无关的节点中,降低了效率[4]。

文中基于P2P的底层架构,提出了一个以服务注册节点分类的Web服务发布和发现的工作模型,并叙

收稿日期:2007-03-30
基金项目:广西自然科学基金项目(桂科自0640026)
作者简介:陈星豪(1980-),男,广西人,硕士研究生,研究方向为Web技术与智能搜索;李陶深,教授,CCF会员,研究方向为分布式数据库、网络信息安全、网络路由算法。

述模型中的服务发布和发现的工作流程。

1 基于 P2P 语义 Web 服务发布和发现模型

1.1 模型的设计

模型的设计思路是:采用集中式与 Peer－to－Peer 结构混合的网络模型,即在网络上分布着很多超级节点,超级节点存储与它直接相连接的节点的各种信息,且各超级节点是按领域本体分类的[5]。即超级节点和与其连接的普通节点构成集中式的结构,而各超级节点之间是纯 Peer－to－Peer 的结构。这里的超级节点是在普通节点的基础上增加了一些机制,使其具有存储、转发等功能。这里把超级节点称为 Broker。

模型的结构如图 1 所示。Broker 之间采用非结构化的连接方式连接;而 Broker 与其相关的 Peer 节点采用集中式方式相连接[6,7]。

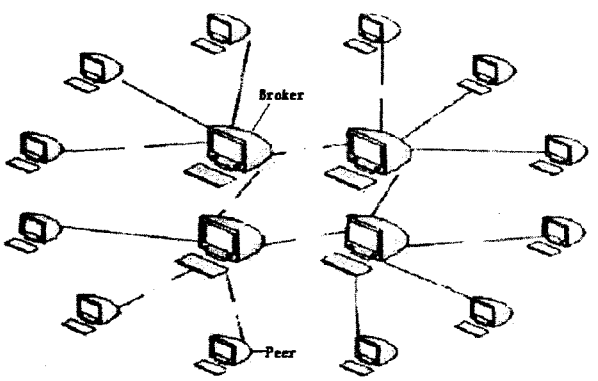

图 1 基于 P2P 的语义 Web 服务发布和发现模型结构

在该模型中,每个注册节点根据需要和兴趣提供某个方面的服务描述信息服务,节点使用一定的机制来组织和管理,并按领域本体进行分类。每一个服务注册节点隶属于某个服务领域,并提供该领域的服务信息的存储和检索服务,所有注册节点共同提供注册中心的功能。

1.2 Broker 模型结构

对于模型中的 Broker,希望它能声明自己属于哪一类领域本体,且能够存储属于这一类领域本体的 Web 服务的有关信息,以方便用户查询和易于管理。Broker 能接收查询请求,且对查询请求进行规范化,当查询失效时,本地的 Broker 能向邻接 Broker 转发。为了实现向邻接 Broker 转发功能,Broker 应具有获取邻接 Broker 地址的功能。为了防止消息的混乱,为每个 Broker 设计了一个转发内容缓存表,用于存储最近接收到的匹配请求。当一个匹配请求到达时,首先与缓存表中存储的所有匹配请求进行比较,如有相同的,则拒绝接收,且返回拒绝回应。Broker 模型的结构如图 2 所示。

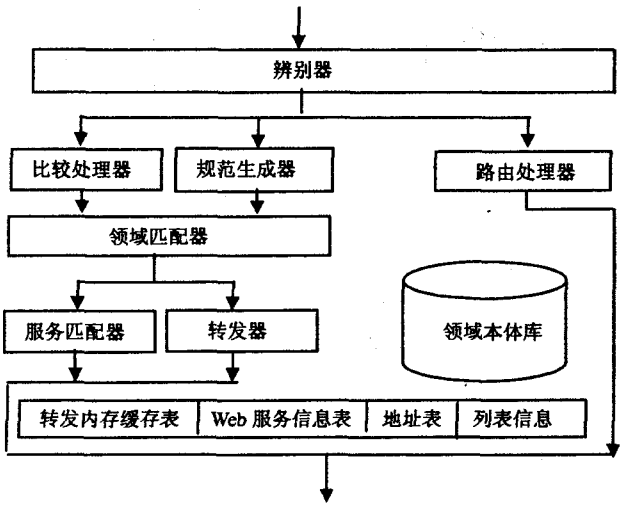

图 2 Broker 模型结构示意图

模型中各模块的功能如下:

(1) 辨别器:对接收到的消息进行判断,并发送至相应的处理模块。

(2) 比较处理器:专门处理转发请求,将转发请求与转发内容缓存表中的内容进行一一匹配,如有相同的,则拒绝继续处理,且按原路返回拒绝回应。如果没有找到完全匹配的内容,则将转发请求的内容存入转发内容缓存表,且把此转发请求发送至领域匹配器。

(3) 规范生成器:接收来自用户发来的加入请求和查询请求,并对请求进行规范化。这里要求用户按照一定的格式输入加入请求和查询请求。规范化完后,将其发送至领域匹配器。

(4) 路由处理器:这是 Broker 最先自动启动的模块,其任务是提取本地地址向周围发送交换地址广播,但仅限是邻接的 Broker;对接收到的交换地址广播请求进行验证,合法则将其存入地址表,不合法则返回拒绝回应。这样一来,每个 Broker 都知道自己相邻的 Broker 的位置。

(5) 领域匹配器:将接收到的匹配请求与领域本体库进行匹配。如果匹配失败,则将接收到的请求按原格式保持不变发送至转发器;如果匹配成功,则要区分是加入请求还是查询请求,若是加入请求则将信息存入 Web 服务信息表,查询请求则发送至服务匹配器。

(6) 服务匹配器:结合领域本体库按照匹配算法,将接收到的请求与 Web 服务信息表的内容进行一一匹配。

(7) 转发器:将接收到的转发消息打上转发烙印,然后除了上一级转发节点外,按照地址表里的地址全部转发,并给接收到的消息的源地址发送回一个本地

地址,以作为该信息源地址的路径节点。

(8) 转发内容缓存表:暂时存储曾经转发过的消息,作为转发器的参照依据。

(9) Web服务信息表:存储此领域各个Web服务的具体信息和具体地址,服务匹配器将会遍历它。该表的表头是对整个领域本体的概括,供领域匹配器访问。表头内容用WS_ T 表示,表中存储的Web服务用WS_ $i(i = 1,\cdots,n)$ 表示。

(10) 地址表:存储邻接Broker的地址,与路由处理器交互。

1.3 用户界面

用户界面如图3所示,界面包括Web服务加入请求界面和查询请求界面。界面还附带一个结束判断器,用于判断所发出的请求是否已经经过所有的Broker。无论是加入请求还是查询请求,都要输入具体的信息。因为Web服务采用OWL-S语言描述,则用户要输入的查询请求定义为:

WSR=<Sc,Sn,Si,So>

其中,Sc(Server category)代表该Web服务属于哪个领域的,Sn代表服务名称,Si代表服务的输入,So代表服务的输出。Si, So也可表示为以下的集合:

Si=(In1,In2,$\cdots$,In n)

So=(Out1,Out2,$\cdots$,Out n)

当加入请求通过检查时,用户输入的信息就被当成一个Web服务的信息存储在相应领域的Broker的Web服务信息表里。所以用户的加入请求界面内容必须要包括Web服务具体信息和具体地址。

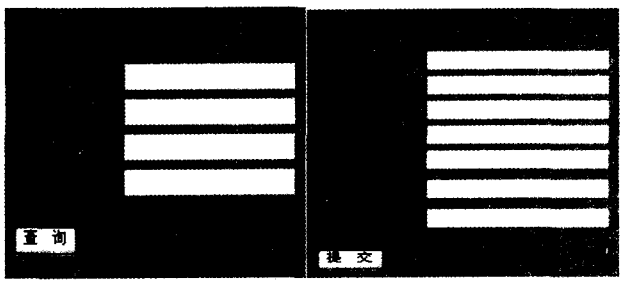

图3　用户界面的示意图

2 工作机制

2.1 Web服务的发布

Web服务发布的工作流程大致如下:

Step1 用户输入的加入请求信息;

Step2 离用户加入请求信息最近的Broker抽取信息中的Sc,由匹配算法Match1(WSR-Sc,WS_T)联合本身的领域本体库,将此Sc与Web服务信息表头的内容进行匹配;

(1)如果匹配成功,则将此Web服务存储到本Broker的Web服务信息表中;

(2)如果匹配失败,则将此信息送到转发器,由转发器对照转发内容缓存表;如果此信息跟表中的某个信息内容一致,说明曾经转发过此信息,则转发器拒绝转发此信息,否则将信息发送到邻接的Broker,并向信息源地址发送一个当地地址。

Step3 邻接的Broker将重复以上的工作程序,直到发布行为完成。

2.2 Web服务的发现

假设用户输入的查询信息为WSR=<汽车,价格查询,(品牌、型号),价格>。Web服务发现的工作流程大致如下:

Step1 查询信息首先到达离它最近的Broker;

Step2 Broker抽取信息中Sc=汽车,将之与Broker中Web服务信息表头的内容进行。

(1)如果匹配成功,则调用匹配算法Match2(WSR,WS_ i)。由该算法联合本身领域本体库,将此WSR与本Broker的Web服务信息表里的WS_ $i(i = 1,\cdots,n)$ 进行匹配:

①如果找到相符的服务,则给用户返回服务的具体信息和具体地址,由用户直接调用服务;

②如果找不到相符的服务,则给用户返回查找失败的信息。

(2)如果匹配失败,则将此信息WSR=<汽车,价格查询,(品牌、型号),价格>送到转发器,由转发器对照转发内容缓存表:

①如果信息跟表里某一信息内容一致,代表曾经转发过此信息,则转发器拒绝转发此信息;

②否则将信息发送到邻接的Broker,并向信息源地址发送一个当地地址。

Step3 邻接的Broker将重复以上的工作程序,直到发现行为完成。

对于用户来说,附带的结束判断器如果长时间收不到转发器发回来的地址,则意味着信息走完所有Broker都找不到相符合的Web服务。

2.3 模型的性能分析

在Broker模型中,各模块的功能确保Web服务能快速、准确地找到所属领域进行发布行为;同时,为Web服务发现机制提供了保障,确保发现请求不会产生环路等死循环情况。

实验结果表明,在执行效率上,文中提出的服务发布和发现模型利用领域本体分类思想将服务分类,可以有效地减少检索目标服务的数量,提高查询的效率;在服务发现性能方面,由于本模型从服务信息的发布

(下转第83页)

表1　分词速度统计表

文档类型	分词速度(字/秒)
政治	2300
经济	1980
文化	2820
体育	2032
科技	2400

经过多次测试,如果使用传统的逆向最大机械分词算法,MAXL取14个字节,即7个汉字,平均分词速度为280字/秒;使用改进后的分词词典,逆向最大机械分词算法分词速度达到2530字/秒。

通过分析以上的测试结果,可以看出该中文分词系统将来所要继续改进的方向是词典识别专有名词的能力和其动态性增容性方面。可以建立不同的分类词典,对专门的领域使用专门的分词词典,并且词典容量也可以动态变化以适应不同领域中出现的新词汇。

5　结束语

自动分词是汉语自然语言处理的第一步。目前,汉语自然语言处理的应用系统处理对象越来越多的是大规模语料(如Internet信息搜索引擎,各种全文检索系统等),因此分词的速度和分词算法的易实现性变得相当关键。在多种分词算法中,正向最大匹配分词算法简洁、易于实现,在实际工程中应用最为广泛。但基于统计的分词算法和基于理解的分词算法都是对基于规则分词算法扩充和完善,一般的分词系统都是将其中几种结合起来一起使用,很少单纯使用一种分词算法。基于理解的分词算法实现起来复杂,但其分词精度相当高,适合于要求分词精度高的场合;而基于统计分词算法对识别未登录词和专有名词有着自己的优势。三者的有机结合将是未来的发展方向。

参考文献:

[1] 冯书晓,徐　新,杨春梅.国内中文分词技术研究新进展[J].情报杂志,2002(11):29-30.

[2] JieeSoft.OFBiz简单介绍[EB/OL].2004-04-16/2004-06-05.http://www.jieesoft.com/modules.php.

[3] 文庭孝,邱均平,侯经川.汉语自动分词研究展望[J].现代图书情报技术,2004,112(7):6-10.

[4] 湛　燕,陈　昊,袁　方,等.基于中文文本分类的分词方法研究[J].计算机工程与应用,2003(23):87-91.

[5] 文庭孝.汉语自动分词研究进展[J].图书情报,2005(5):54-63.

[6] 邹海山,吴　勇,吴月珠,等.中文搜索引擎中的中文信息处理技术[J].计算机应用研究,2000(12):21-24.

[7] 吴　岩,李秀坤,刘　挺,等.中文自动校对系统的研究与实现[J].哈尔滨工业大学学报,2001,33(1):60-64.

[8] 吕学强.机器翻译概述[J].辽宁师专学报,2002,4(1):8-11.

[9] 郭　辉,苏中义,王　文,等.一种改进的MM分词算法[J].微型电脑应用,2002,18(1):13-15.

[10] 李家福,张亚非.一种基于概率模型的分词系统[J].系统仿真学报,2002,14(5):544-546.

[11] 彭希鸿.基于WEB内容挖掘的网页分类与过滤研究与应用[D].长沙:中南大学,2003.

(上接第79页)

到服务的查找都利用基于本体的匹配算法,因此同传统的关键字匹配机制比较可以获得更准确的语义信息,进而能更精确地定位服务,提高查准率并改善Web服务发现性能。

3　结束语

基于P2P的底层架构,文中提出一个语义Web服务发布和发现模型。该模型的底层采用两层结构:第一层节点采用非结构化的方式连接,通过一定的路由机制保证通信的畅通,避免回路,同时有效支持系统的可扩展性;第二层采用集中式的结构组织节点,有效地提高查询效率。在此模型的基础上,设计具体高效的匹配算法将是下一步研究的目标。

参考文献:

[1] 宋　炜,张　铭.语义网简明教程[M].北京:高等教育出版社,2004.

[2] 方馨馨,熊齐邦.基于P2P网络的语义Web服务发现机制[J].计算机工程,2005,31(17):115-117.

[3] 尹晓璐,李广军.基于语义的Web服务查询[J].实验科学与技术,2005,24(1):31-34.

[4] Cerami E.Web服务精髓[M].陈　逸译.北京:中国电力出版社,2003.

[5] Raman R,Solomon L M.MatchMaking:an extensible Framework for distributed resource management[J].Cluster Computing,1999(2):129-138.

[6] Li Lei,Horrocks L.A Software Framework for Matchmaking Based on Semantic Web Technology[C]//Proceedings International WWW Conference.Budapest,Hungary:[s.n.],2003:20-24.

[7] Le-Hung Vu,Hauswirth M,Aberer K.Towards P2P-based Semantic Web Service Discovery with QoS Support[D].Lausanne,Switzerland:School of Computer and Communication Sections,Ecole Polytechnique Federale de Lausanne(EPFL),2005.