

基于分布式的语义 Web 服务发现新模型

李文娟, 史维峰

(西北大学 信息科学与技术学院, 陕西 西安 710127)

摘要:为了快速、准确和高效地发现目标服务,提出了一种基于分布式和语义描述的 Web 服务发现新模型。该模型将领域分类的思想与 P2P 网络相结合,构造出一种基于 P2P 的双层拓扑结构,并采用一种层次化的注册管理机制,从而提高了服务发现效率。同时通过使用两阶段搜索算法及二层服务匹配算法对服务进行层层递进查找,使该模型在保证查准率的基础上大大提高了服务发现的查找速度。最后,通过原型系统证明了该模型是合理的和高效的。

关键词:P2P; 语义 Web 服务; 服务发现; 搜索算法

中图分类号:TP309

文献标识码:A

文章编号:1673-629X(2009)12-0108-05

A New Semantic Web Service Discovery Model Based on P2P Network

LI Wen-juan, SHI Wei-feng

(School of Information Science and Technology, Northwest University, Xi'an 710127, China)

Abstract: In order to find the aimed Web Service precisely and efficiently, presents a new service discovery model based on P2P network and semantic description. The improved model combines the idea of field classification with P2P networks, and implements a P2P double-layered topology structure which contains hierarchy registration mechanism, thereby improves the efficiency of the service discovery. It also uses two phase search algorithm and the two-layer service matching algorithm to recursively discover the service which greatly enhances the speed of service discovering on the basis of ensuring precision ratio. At last a prototype system proves that the model has a better performance.

Key words: P2P; semantic Web service; service discovery; search algorithm

0 引言

近年来 Web 服务标准的持续完善使得发布在网络上的服务呈现出爆炸性增长趋势,如何以快速、准确和高效的方式发现目标服务成为一个迫切需要解决的问题,对此国内外学者进行了诸多研究。一方面,为了解决如何存储、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的问题,通过引入 P2P 技术来处理服务元数据的交换^[1],试图克服传统 UDDI^[2]技术中服务元数据集中注册、集中存放对搜索广度带来的限制。另一方面,为了解决如何准确、细致地刻画服务能力,从而支持用户需求与服务描述之间更精确的匹配操作的问题,引入了语义网技术^[3,4],借助于本体和描述逻辑等逻辑

推理系统的使用,加强服务描述信息的机器可理解性,支持用户需求和服务能力之间的逻辑推理匹配。因此提高了服务发现的精度和效率。

文中在分析现有研究成果的基础上^[5~9],提出了一种基于分布式和语义描述的新型 Web 服务发现新模型,将领域分类的思想与 P2P 网络相结合,构造出一种 P2P 的双层拓扑结构,它将服务按照语义形成域和簇,设计了与之相适应的两阶段服务搜索算法。该算法综合考虑了注册节点的邻近性、节点所发布服务在语义上的相似性和发现相应节点的时效性等各种因素。同时通过基于语义的二层服务匹配算法对注册服务与搜索服务进行匹配,使该模型在保证查准率的基础上提高了服务发现的速度。

1 服务发现模型总体结构

本服务发现模型采用双层 P2P 非结构网络拓扑结构并结合语义 Web 的特点,如图 1 所示,其中包括领域代理和簇节点注册中心,在领域内,多个簇节点注册中心以 P2P 结构组织在一起,为整个领域提供完整

收稿日期:2009-04-02;修回日期:2009-07-23

基金项目:国家高技术研究发展计划(863)重点资助项目(2007AA010305)

作者简介:李文娟(1984-),女,陕西西安人,硕士研究生,研究方向为计算机网络与分布式系统、SOA;史维峰,教授,研究方向为计算机网络与分布式系统、SOA 及 CAD/CAM。

的注册与发现服务。同时,各域通过领域代理组织成一个较高层的 P2P 网络以支持跨域的服务匹配与发现。语义上相似的服务都注册在同一个簇节点注册中心上,而具有相同领域特征的簇节点又注册在领域代理内。

在本模型的领域代理和簇节点注册中心中增加了反馈机制,可以对拓扑中的服务和簇节点的状态进行主动监控,并更新相关信息,从而有效地保障了模型的有效性和容错性,克服了目前注册中心中服务目录被动更新的缺陷。当请求者向某一簇节点注册中心发送请求,若规定时间内未收到簇节点注册中心返回的应答消息,则进行重试,若仍未收到返回消息,则认为该簇节点注册中心已被动退出,并向领域代理发送该反馈信息;反之,认为该簇节点注册中心仍存活,等待返回的结果信息。当服务请求者得到服务的信息后,在与服务进行绑定时,如果无法与服务连接,则进行重试,如果在某次重试中,得到了正确的信息,则认为该服务可用,与此服务进行绑定;反之,则认为该服务已失效,并向领域代理发送该反馈信息,由领域代理向服务所在簇节点注册中心发出删除服务的消息。

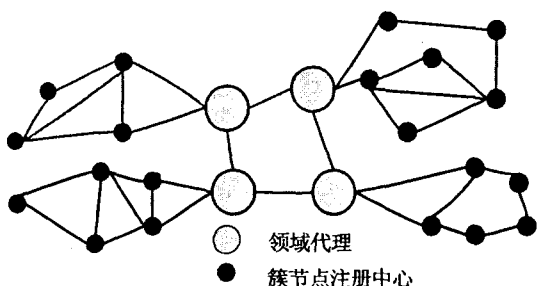


图 1 服务发现模型双层 P2P 拓扑图

1.1 领域代理

领域代理在本模型中是领域内注册中心与其他领域交互的中介。一方面将本领域内的请求按照一定的策略转发给其它领域代理以实现跨域匹配,另一方面接收其它领域代理发来的请求并将请求转发给本领域内的各簇节点注册中心以进行语义服务匹配。领域代理在创建时,根据所创建行业的本体进行语义描述,生成自身的语义描述,并采用成熟的关键词抽取方法和约束方法生成自身领域的关键词候选集,便于进行簇节点注册中心和 Web 服务的匹配。

1.2 簇节点注册中心

簇节点注册中心在创建时,也根据同样的关键词抽取方法和约束方法生成关键词候选集,并向临近的领域代理发出请求,领域代理进行语义匹配。若匹配程度高于一定阈值,簇节点注册中心首先将自身信息注册到领域代理中,领域代理将返回与其语义相似的

节点簇结果列表。簇节点注册中心将返回的列表生成内部的朋友列表,同时向各朋友簇节点注册中心发送自身信息,与其建立朋友关系。

1.3 Web 服务的发布

Web 服务将自己的语义描述信息作为广播向离它最近的领域代理发出匹配请求。领域代理进行语义匹配,若匹配程度高于一定阈值,则把匹配的所有结果返回给该 Web 服务,Web 服务将分别向这些簇节点注册中心发出注册请求,并比较所返回的匹配程度值,选择匹配程度最高的簇节点进行注册,即 Web 服务存储该簇的相关信息,并将自己的相关信息发布到此簇节点注册中心上。若不存在这样的节点簇,领域代理将存储 Web 服务的相关信息,该 Web 服务的节点自行建立一个簇节点注册中心。否则,该领域代理将该 Web 服务的语义描述信息转发给其它领域代理进行匹配,实现注册。

2 服务发现过程

通常情况下,服务请求者向离它最近的簇节点注册中心发出查找请求,该簇节点注册中心在进行查找的同时,也会将请求转发给其它朋友节点进行查找,如果查找不成功,则通过所在领域代理请求转发给其它领域代理,进行查找。

2.1 相关定义

为便于对服务进行语义相似度计算从而进行准确的匹配,对服务、服务请求和语义相似性作如下定义:

定义 1 Web 服务信息描述: $WS_i(\text{ServiceID}, (\text{Key Words}), (\text{Inputs}, \text{Out - puts}, \text{Pre}, \text{Effect}))$;其中, WS 为服务的名称; ServiceID 为服务的标识; (KeyWords) 为多个可以描述服务的关键词; $\text{Inputs}, \text{Outputs}, \text{Pre}, \text{Effect}$ 为针对 OWL - S 对服务的语义描述。

定义 2 Web 服务请求描述: $WSR_i(\text{Class}, (\text{Key Words}_i, \text{Threshold}_i), (\text{IPOEs}))$;其中, WSR_i 表示 Web 服务请求名称; Class 标识了本服务请求的服务属性; $(\text{KeyWords}_i, \text{Threshold}_i)$ 表示所需匹配的关键词和其对应的阈值; (IPOEs) 表示 IPOE 的语义描述。

定义 3 语义相似性^[10](semantic similarity)。给定两组概念 $C_1 = \{c_{11}, c_{12}, \dots, c_{1n}\}$ 和 $C_2 = \{c_{21}, c_{22}, \dots, c_{2m}\}$, 两组概念之间的语义相似性 $\text{Sim}_S(C_1, C_2)$ 定义为

$$\text{Sim}_S(C_1, C_2) =$$

$$\text{Max} \left(\sum_{i=1}^n \sum_{j=1}^m \text{Sim}_S(c_{1i}, c_{2j}) \right) / \text{Min}(n, m) \quad (1)$$

其中, $\text{Sim}_S(c_{1i}, c_{2j})$ 表示本体两个概念类间的语义

相似性,包括语法层次上的相似性 SimT 、内部结构相似性 SimIS 、外部结构相似 SimES 以及外延相似性 $\text{SimEX}^{[11]}$ 。 $\text{Sim}_S(c_{1i}, c_{2j})$ 按照加权求和计算得到,即 $\text{Sim}_S(c_{1i}, c_{2j}) = \omega_1 \text{SimT}(c_{1i}, c_{2j}) + \omega_2 \text{SimIS}(c_{1i}, c_{2j}) + \omega_3 \text{SimES}(c_{1i}, c_{2j}) + \omega_4 \text{SimEX}(c_{1i}, c_{2j})$, 其中 $\omega_1 + \omega_2 + \omega_3 + \omega_4 = 1$, SimT 为基于最大相同子串得到的相似性; SimIS 通过比较概念的属性来计算概念相似性; SimES 通过比较概念的父概念和子概念来计算概念之间的语义相似性; SimEX 主要比较概念的实例之间的相似性,从而得到概念之间的相似性。

2.2 两阶段搜索算法

两阶段搜索算法使用经典泛洪策略搜索适合的簇节点注册中心,选用泛洪策略是因为查找效果受网络规模影响较小,且对服务领域拓扑的平均路径长度不敏感。一方面,由于本体自身的目的就是为了实现领域的共享,为领域提供一个统一的领域模型,本体所包含的本体模块一般都比较少并且它们之间关系简单,这样,领域内的注册服务器之间的 P2P 网络也比较简单。因此,经典泛洪算法所带来的网络开销在领域内的注册服务器之间的 P2P 网络中并不显得突出。另一方面,在领域内,本体模块之间的关联在一定程度上已经表示了不同注册中心之间的关联性。因此通过经典泛洪算法来定位注册中心实际上是一种非常有效的方法。

(1) 域内搜索的算法如下:

```
TTL = TTL - 1
if sq in {Reg1, Reg2, ..., Regm} //已经收到过消息
if TTL = 0 then //至今为止未找到合适的注册服务器,且消息不能继续传播
[SearchRegister(DAgent, ST, {Reg1, Reg2, ..., Regm}, TTL, flag)]g //消息转发给领域代理
else
Sim = SemSim(sq, ST) //簇节点注册中心和服务请求之间实施简单的相似性匹配
if Sim >  $\xi$  then //该簇节点注册中心满足要求
[SearchRegister(sq', ST, {Reg1, Reg2, ..., Regm}, sq], TTL - 1, True)]g
else
[SearchRegister(sq', ST, {Reg1, Reg2, ..., Regm, sq}, TTL - 1, False)]g
```

(2) 域间搜索算法如下:

```
foreach SearchRegister do
if SearchRegister.flag = false then
NotFound ++
end if
Total ++
```

```
end for
if NotFound/Total = 1 then
foreach neighbour of DAgent do
ST' = Transform(ST, DAgent, neighbour of DAgent)
[SearchRegister(sq, ST',  $\emptyset$ , TTL, False)]g
end for
end if
```

其中算法描述中 ST 表示服务请求; $\text{SearchRegister}(sq, \text{ST}, \text{rp}, \text{TTL}, \text{Flag})$ 表示服务请求者向簇节点注册中心 sq 发送查询服务 ST 的请求, rp 为已经请求的节点列表, TTL 为消息的存活周期, Flag 标志目前为止是否找到匹配服务(找到为 True 反之为 False); $[\text{SearchRegister}(sq, \text{ST}, \{\text{Reg}_1, \text{Reg}_2, \dots, \text{Reg}_m\}, \text{TTL} - 1), \text{True}]_g$ 表示按照泛洪策略转发请求, 当没有邻居节点时转发给领域代理。

2.3 两次服务匹配

虽然基于语义描述的推理机制匹配精确度高,弥补了传统语法级 Web 服务发现技术精确度低的不足^[12]。但其最大的缺点在于匹配推理过程耗时巨大,所以在规模巨大的网络系统处采用基于语义的匹配方法是不合适的。因此文中基于文献^[13]提出了一种两次服务匹配算法,它第一次匹配使用基于多关键字粗糙匹配;第二次则进行基于语义的精确匹配。该方法在不降低查准率的前提下,可以在很大程度上减少匹配耗时。

搜索算法中 SearchRegister 的具体服务匹配流程如图 2 所示。

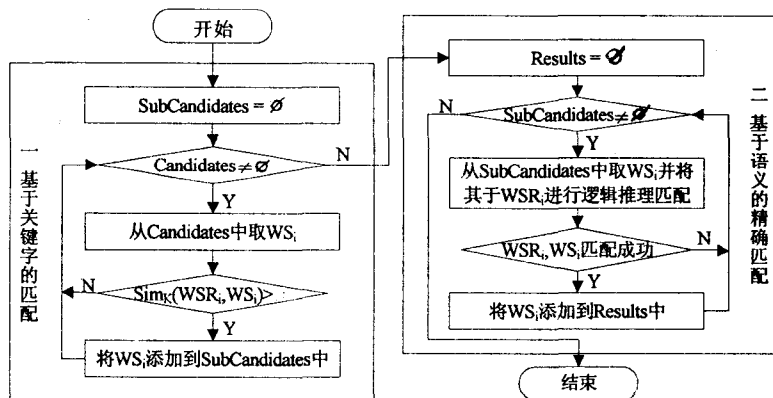


图 2 服务匹配流程图

(1) 基于关键字的匹配。当服务请求者定位到符合其服务属性的簇节点注册中心后,由该簇节点注册服务器进行基于关键字的匹配,将用户的不同关键字与服务关键字进行相似度计算,得到关键字相似度 Sim_K 并与用户给定的相应阈值 θ 进行比较,将查找到的相似度大于 θ 的服务保存到查询服务候选集中,作为第二次基于语义推理匹配的候选服务集。

该阶段算法简单,计算量主要在于相似度计算。相似度计算可以采用比较成熟的算法,如向量空间模型、TFIDF 等。经过本次粗糙匹配,可以过滤掉大部分不匹配的服务,缩小了下一次精确匹配的候选范围,从而缩短了查询服务的时间。而且服务关键字的匹配方式也不会漏掉相关服务,保证了服务查找的查准率^[14]。

(2) 基于语义的精确匹配。第二次匹配利用服务描述中的语义信息、描述逻辑和本体语言 OWL-S 对 Web 服务功能和用户目标描述进行精确匹配。由簇节点注册中心在其已发布服务的存储库中读取候选子集中的服务描述信息,与用户查询服务的需求描述进行逻辑匹配推理,若存在匹配成功的注册服务,则保存到最终的匹配服务子集中。当完成所有匹配后,由簇节点注册中心将匹配服务子集中服务的信息和匹配程度发送给服务请求者,以供其选择调用。

3 实验结果及分析

目前 Web 服务发现模型还没有公认的评价标准,在多数论文中,仍然借用信息检索中使用的查准率和召回率指标作为评价标准,查准率和查全率越高,服务匹配算法越好^[5]。

就查准率来说,由于文中提出的服务发现模型是基于语义的,相比于传统的基于关键字的服务发现过程而言,可以保证较高的查准率,而相对于仅依靠语义进行的服务发现过程,不会降低查准率。因为在第一阶段搜索中仅过滤掉不符合的服务,而查询返回集合是在第二阶段搜索时决定的,因此,影响查准率的是基于语义和逻辑推理的匹配过程,是否使用了第一阶段搜索方法来缩小候选服务子集对系统查准率不造成影响。

就查全率而言,由于第二阶段搜索是在组内以广播的方式进行,对候选服务子集进行了穷举查询。所以,在第二阶段搜索过程中对查全率产生影响的是查询匹配方法,与搜索机制无关。虽然第二阶段搜索不影响系统的查全率,但由于在第一阶段搜索时没有对候选服务集进行穷举,系统过滤掉了认为最不可能属于标准集的服务,但不能完全排除过滤掉的服务中含有标准集中的元素的可能,所以,两层搜索机制的第一层搜索会导致系统召回率的降低。总的来讲,两阶段搜索机制是通过牺牲系统召回率来达到缩短查询时间的目的。

对于平均查询时间,有如下估算公式:

$$T = T_r + T_i + T_q$$

其中 T_r 为簇节点注册中心路由操作的平均时耗; T_i

为平均网络延时; T_q 为簇节点注册中心上二次匹配查询的平均时耗。由于两阶段搜索方法中的第一阶段搜索过滤掉了很大一部分候选服务,从而节约了匹配操作的计算时间。相对集中式的服务发现模型而言,文中所述 Web 服务发现模型构建在 P2P 环境之上,某次搜索分散到了各个不同的 Peer 上,达到了将作业并行处理的效果,因此也可以缩短查询处理时间。但不足的是,基于 P2P 的查询处理过程相对来讲有更多的网络通信,因此网络的延时比较大。

文中根据模型的体系结构开发了原型系统,系统在 J2SE V1.5 平台实现,其中领域本体库和服务库采用 Tomcat5.0 与 Axis1.1 进行发布;采用 Racer1.7 作为描述逻辑的推理机,本体的描述语言采用 OWL-DL,Web 服务的描述采用 OWL-S1.0 的 Profile,采用 OWLS-Editor 对 OWL-S 描述文件进行处理。实验使用 10 个不同领域的本体文件(OWL 格式),以及 20 多个与领域相关的 WSDL 文件,并用相关领域本体标注这些 WSDL 文件,将标注过的 WSDL 文件发布到 UDDI 库中,这样每个 Web 服务都有一个与存储在 UDDI 库中相对应的服务描述。

实验比较了:a)基于关键字匹配的 Web 服务发现方法;b)UDDI 方法;c)文中方法。具体测试结果如图 3 所示。

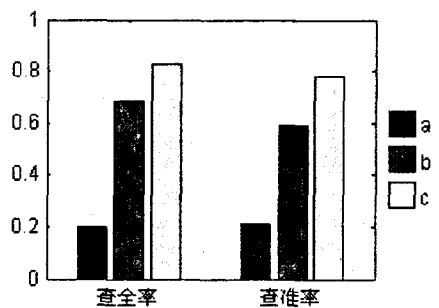


图 3 测试结果图

通过实验发现文中方法的查全率和查准率都比基于关键字的 Web 服务发现方法和 UDDI 方法高,这里假设各个阶段的相似度阈值均为 50%。

4 结束语

在基于语义的 Web 服务发现研究的基础上,提出了一种基于分布式和语义描述的新型 Web 服务发现模型,将领域分类的思想与 P2P 网络相结合,构造出一种 P2P 的双层拓扑结构,并提出了两阶段搜索算法与之适应,在服务发现的时间开销和精度约束之间取得了一定的平衡。通过两层搜索机制使 P2P 网络的消息转发复杂度和搜索广度得到了兼顾。同时在搜索过程中使用了两次服务匹配算法进行服务匹配,最后

分析了该模型的查准率、查全率和时间性能,该模型在保证查准率的基础上提高了服务发现的查找速度。

参考文献:

- [1] 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.
- [2] Luc C, Andrew H, von Claus R, et al. UDDI Version 3.0.2 [EB/OL]. 2004-10. <http://www.uddi.org/pubs/uddi-v3.htm>.
- [3] 付燕宁,金英,刘磊,等.基于语义的 Web 服务体系结构[J].计算机技术与发展,2008,18(3):28-31.
- [4] Banaei-Kashani F, Chen C C, Shahabi C. WSPDS: web services peer-to-peer discovery service[C]//International Symposium on Web Services and Applications. Nevada, USA: [s. n.], 2004: 733-743.
- [5] 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):595-602.
- [6] Liu Jie, Zhuge Hai. A semantic-link-based infrastructure for web service discovery in P2P networks[C]//Proceedings of International World Wide Web Conference. Chiba, Japan:

[s. n.], 2005: 940-941.

- [7] Koller D, Sahami M. Toward optimal feature selection[C]//Proceedings of International Conference on Machine Learning. Bari: Morgan Kaufmann, 1996: 284-292.
- [8] 胡建强.服务发现若干关键技术研究[D].长沙:国防科学技术大学,2005.
- [9] 张孝国,黄广君,郭洪涛.基于本体的 Web 服务描述与发现机制研究[J].计算机工程与应用,2008,44(16):148-150.
- [10] 刘志忠,王怀民,周斌.一种双层 P2P 结构的语义服务发现模型[J].软件学报,2007,18(8):1922-1932.
- [11] Euzenat J, Bach T L, Barrasa J, et al. D2.2.3: State of the art on ontology alignment[EB/OL]. 2004. <http://www.starlab.vub.ac.be/publications/kweb-223.pdf>.
- [12] Paolucci M, Kawamura T, Payne T R, et al. Importing the semanticweb in UDDI[C]//Computer Science. Proceedings of Web Services, E-Business and Semantic Web Workshop. London: Springer-Verlag, 2002: 225-236.
- [13] 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.
- [14] 郭得科,任彦,陈洪辉,等.一种 QoS 有保障的 Web 服务分布式发现模型[J].软件学报,2006,17(11):2324-2334.

(上接第 107 页)

成熟的软件模块并对它进行修改,使之符合已经指定的规范,在系统的开发工程中,工作人员需要广泛的参考其他比较成熟的模型资料,一般硬件和软件的厂商都会提供几个开发样板资料,还有一些开源项目也为极具参考价值,工作人员需要对这些资源进行合理的利用;其次,指定标准的 API 接口函数,符合统一的编程规范,制作的过程同上面一样,需要广泛的参考既定的标准,这样就可以使系统开发工作具有更好的可操作性;再次,需要适当地扩充自己的软件模块,并与前面的标准相互补充,在系统中,不可行别人的所有的一切都可以拿来直接使用,这样需要编写自己的模块。

4 结束语

文中的重点工作包含两项,一是如何更好地进行嵌入式系统开发,嵌入式系统开发流程是首先对微处理器进行选型,接下来对软件部分和硬件部分协调设计与实现,完成后需要综合测试,直到嵌入式系统能稳定运行为止。二是如何进行嵌入式系统开发和设计,嵌入式系统设计一般分为项目立项调研、用户需求分析、系统需求分析、系统设计、系统实现、系统测试和运行维护这几个阶段。文中提出的工程化思想对实际工

程研发具有很好的参考和借鉴价值。

参考文献:

- [1] McUmber W E, Cheng B H C. UML Based Analysis of Embedded Systems Using a Mapping to VHDL[J]. IEEE High Assurance Software Engineering, 1999(11):56-63.
- [2] 李炜,张义超,卢英,等.基于 GPRS 环境与安全监测终端设计与实现[J].计算机技术与发展,2008,18(9):232-234.
- [3] 怯肇乾.嵌入式系统硬件体系设计[M].北京:北京航空航天大学出版社,2007.
- [4] 李光成,褚伟.基于 $\mu\text{C}/\text{OS}-\text{II}$ 嵌入式实时系统的优先级倒置分析[J].计算机技术与发展,2007,17(7):98-101.
- [5] Labrosse J J. 嵌入式实时操作系统 $\mu\text{C}/\text{OS}-\text{II}$ [M]. 第 2 版.邵贝贝等译.北京:北京航空航天大学出版社,2003.
- [6] 程广河,郝凤琦,张让勇,等.嵌入式环境中的软件构件化研究[J].计算机技术与发展,2007,17(9):139-141.
- [7] Konrad S, Cheng B H C, Campbell L A. Object Analysis Patterns for Embedded Systems[J]. IEEE Trans on Software Engineering, 2004,30(12):970-992.
- [8] Gajski D D. 嵌入式系统的描述与设计[M].边计年,吴为民,等译.北京:机械工业出版社,2005.