第 39 卷 第 7 期
Vol.39 No.7

计 算 机 工 程
Computer Engineering

2013 年 7 月
July 2013

• 人工智能及识别技术 •　文章编号：1000—3428(2013)07—0219—05　文献标识码：A　中图分类号：TP391

基于模糊聚类优化的语义 Web 服务发现

王永明，张英俊，谢斌红，潘理虎，陈立潮

(太原科技大学计算机科学与技术学院，太原 030024)

摘　要：语义 Web 服务发现机制在发现服务时的准确率较低。为解决该问题，提出一种基于模糊聚类优化的语义 Web 服务发现方法。采用改进的模糊 C-均值(FCM)聚类算法，实现对服务聚类预处理，在模糊聚类时，综合考虑服务的输入、输出、前提、效果 4 个功能性参数，并扩展已有的服务匹配机制，在匹配时，将服务的 4 个功能性参数全部作为服务相似度的计算因子。实验结果表明，在模糊聚类稳定的条件下，该方法的服务平均查全率为 79.6%，平均查准率为 85.9%，均高于未采用聚类处理和只采用输入/输出参数的 FCM 聚类处理方法。

关键词：领域本体；本体描述语言；本体距离；模糊聚类；语义 Web 服务；服务发现

Semantic Web Service Discovery Based on Fuzzy Clustering Optimization

WANG Yong-ming, ZHANG Ying-jun, XIE Bin-hong, PAN Li-hu, CHEN Li-chao

(Institute of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

【Abstract】Aiming at the problem of low efficiency for semantic Web service discovery mechanism in finding service, this paper proposes a novel method based on fuzzy clustering for optimizing semantic Web service discovery. It adopts the modified Fuzzy C-means(FCM) clustering algorithm to realize the cluster preprocessing of services. When clustering services, it can comprehensively consider the input, output, premise and the effect of service as the clustering parameters. This paper expands existing services matching mechanism. When matching services, it can take four functional parameters of service as its factors for similarity calculation. Experimental results show that under in fuzzy clustering stable conditions, the method of service average recall rate of 79.6%, and the average prospective rate of 85.9%, higher than the clustering process and only using Input/Output(I/O) parameters FCM method of clustering processing.

【Key words】domain ontology; ontology description language; ontology distance; fuzzy clustering; semantic Web service; service discovery

DOI: 10.3969/j.issn.1000-3428.2013.07.049

1　概述

随着语义 Web 服务应用的普及，Web 服务数量急剧增加，服务注册信息库不断膨胀。在判断一个服务是否能够满足请求时，顺序查找的方法要求服务注册库中的所有服务都逐一地与服务请求进行匹配比较。当服务库中有成千上万条记录时，就会产生服务查找费时、结果不准确等一系列问题。虽然统一描述、发现与集成(Universal Description, Discovery and Integration, UDDI)提供一些分类方法系统，以便于提高查询效率，但这些分类系统并不统一而且比较粗糙，大都仅仅依靠服务的领域进行划分，不能保证所用的分类方法所提供的功能。因此，如何快速、准确和高效地发现目标服务成为一个迫切需要解决的问题。

目前，在解决以上问题上最常用的方法是增强 Web 服务语义信息，提高服务可解释、可推理能力，从而提高服务查找匹配的精度，同时，利用服务服务类的思想提供服务发现的效率。如文献[1]利用领域本体和图论聚类方法对服务进行聚类，通过提取服务描述中的语言信息，将其作为划分类别的依据来提高聚类的准确性和服务发现效率。同样，文献[2]应用 OWL-S 描述服务，采用凝聚的层次聚类算法 Single-Link 实现相似的 Web 服务聚类。文献[3]基于 Web 服务描述语言(Web Services Description Language, WSDL)的服务描述完成对服务的聚类，从而提高服务发现效率。文献[4]采用聚类的方法在提高 Web 服务的发现效率上有所提高。同时，文献[5-6]在搜索引擎和文档的检索中，采用聚类来提高检索效率，实验结果表明其检索效率较高。

基金项目：山西省自然科学基金资助项目(2009011022-1)；太原科技大学研究生创新基金资助项目(20111025)

作者简介：王永明(1985－)，男，硕士研究生，主研方向：智能控制，自动推理；张英俊，教授级高级工程师；谢斌红，副教授、硕士；潘理虎，副教授、博士；陈立潮，教授

收稿日期：2012-07-09　**修回日期：**2012-09-17　**E-mail：**zyj4131829@sina.com

上述研究从不同角度对 Web 服务进行聚类，一定程度上提高了服务发现的效率。但由于所采用的聚类方法把每个待处理的服务严格地归属于某个类，而现实中大多数的服务并没有严格的属性，这样就致使查全率、查准率降低。

本文综合语义信息和聚类等因素对 Web 服务发现的作用，提出一种在领域本体基础上对语义 Web 服务进行模糊聚类预处理的发现方法。利用 Web 服务本体语言(Ontology Web Language for Service, OWL-S)描述 Web 服务，并提取服务中的语义信息将其作为模糊聚类算法的输入数据，实现服务的聚类预处理。

2 基本概念

2.1 领域本体

领域本体给出了特定领域中概念和概念间的关系，用明确的方式表达概念的内涵，最终目标是精确地表示隐含的或不明确的信息。

定义 1 领域本体可用五元组表示[7]：$O=\{C,R,H^c,rel,A^o\}$，其中，C 表示概念的集合；R 表示关系的集合；H^c 表示概念层次；rel 表示概念间的关系；A^o 表示本体公理。

2.2 Web 服务本体语言

为了支持语义 Web，本文采用 OWL-S 来描述 Web 服务模型。OWL-S 是在本体描述语言 OWL 基础上提出的一个 Web 服务本体描述语言。它是基于本体的、用于语义 Web 服务描述的一个规范语言，其包含一整套本体，提供描述 Web 服务的词汇表，描述 Web 服务的语义，能够进行适当的推理，使得 Web 服务具有机器可理解性，以支持自动化 Web 服务发现、组合、调用、互操作和执行监控的能力。基于该服务本体描述语言，本文将 Web 服务表示为如下形式。

定义 2 一个 Web 服务可以用下面的表达式来描述[7]：$WS_i=\{D_i,I_i,O_i,P_i,E_i\}$，其中，$D_i$ 表示该服务的功能；I_i 表示该服务的输入集合；O_i 表示该服务的输出集合；P_i 表示该服务执行的前提条件；E_i 表示该服务执行后的效果。

2.3 模糊 C-均值聚类算法

模糊 C-均值(Fuzzy C-means, FCM)算法是基于目标的模糊聚类算法，是目前比较常用的模糊聚类算法，它有着完善的理论和深厚的数学基础。该算法是相对可伸缩和高效率的，因为算法的复杂度是 $O\{ncb\}$，其中，n 是数据对象的个数；c 是聚类的数目；b 是迭代的次数。

设样本空间 $X=\{x_1,x_2,\cdots,x_m\}$ 中元素有 n 个特征，要把 X 分为 c 类 $(2\leqslant c\leqslant n)$，设有 c 个聚类中心 $V=(v_1,v_2,\cdots,v_c)$。

目标函数为[8]：

$$F(\boldsymbol{U},V)=\sum_{j=1}^{n}\sum_{i=1}^{c}\left(u_{ij}^m\left\|x_i-v_j\right\|^2\right) \tag{1}$$

其中，m 为模糊指标；$\boldsymbol{U}=\{u_{ij}\}$ 是隶属度矩阵，表示第 i 个点属于第 j 个聚类中心的隶属度；$\|x_i-v_j\|$ 是样本 x_i 和聚类中心 v_j 之间的欧式距离。聚类中心值 v_j 的迭代公式为[8]：

$$v_j=\frac{\sum_{j=1}^{n}x_ju_{ij}^m}{\sum_{j=1}^{n}u_{ij}^m} \tag{2}$$

FCM 聚类算法的原理就是求模糊划分矩阵 $\boldsymbol{U}$ 以及聚类中心 V 使得目标函数达到最小值。

3 基于 FCM 优化的服务发现

3.1 语义 Web 服务相似度计算

本文采用 W3C 推荐的本体描述语言(Ontology Web Language, OWL)来描述本体，采用基于OWL扩展的OWL-S来描述服务。OWL-S 对服务的描述如定义 2。由于服务提供者在描述服务时是相互独立的，即使使用相同的本体库，在概念的使用上也会有差异，因此本文采用基于功能属性(基本描述、输入参数、输出参数、前提参数、效果参数)来计算服务相似度。

文献[9]在计算服务相似性时，只是基于服务的输入/输出(Input/Output, I/O)来计算的。最近几年，很多研究者为了提高服务相似性计算的准确性都把服务的前提/效果(Premise/Effect, P/E)也作为服务相似计算的 2 个很重要的指标。因此，本文对服务相似性的计算也同时考虑了服务的 4 个功能性参数即 IOPE。

定义 3 在一个本体分类体系中，概念 A 和概念 B 之间的本体距离 $dis(A,B)$ 定义为在本体树中连接它们的最短路径的边数[9]。

已知如下 2 个服务：

$$WS_i=\{D_i,I_i,O_i,P_i,E_i\},WS_j=\{D_j,I_j,O_j,P_j,E_j\}$$

(1)计算基本描述的相似性

基于 2 个概念之间的本体距离计算其相似度[9]：

$$sim(D_i,D_j)=\frac{dis_{\max}-dis(D_i,D_j)}{dis_{\max}-dis_{\min}}$$

其中，$dis(D_i,D_j)$ 是两者之间的本体距离；$dis_{\max}$、$dis_{\min}$ 则是本体树中所有概念间距离的最大值及最小值。以下类似。

(2)计算基于输入参数的相似度

结合语义距离以及参数之间的匹配程度，可以采用以下公式计算其相似度[9]：

$$sim(I_i,I_j)=\frac{dis_{\max}-dis(I_i,I_j)}{dis_{\max}-dis_{\min}}\times\frac{1}{|I_i|+|I_j|-1}$$

(3)计算基于输出参数的相似性

类似的，基于输出参数相似度[9]为：

$$sim\left(O_i,O_j\right)=\frac{dis_{\max}-dis\left(O_i,O_j\right)}{dis_{\max}-dis_{\min}}\times\frac{\left|O_i,O_j\right|}{\left|O_i\right|+\left|O_j\right|-\left|O_i,O_j\right|}$$

(4)计算基于前提参数的相似性

类似的，基于前提参数的相似度为：

$$sim\left(P_i,P_j\right)=\frac{dis_{\max}-dis\left(P_i,P_j\right)}{dis_{\max}-dis_{\min}}\times\frac{\left|P_i,P_j\right|}{\left|P_i\right|+\left|P_j\right|-\left|P_i,P_j\right|}$$

(5)计算基于结果参数的相似性

类似的，基于结果参数的相似度为：

$$sim\left(E_i,E_j\right)=\frac{dis_{\max}-dis\left(E_i,E_j\right)}{dis_{\max}-dis_{\min}}\times\frac{\left|E_i,E_j\right|}{\left|E_i\right|+\left|E_j\right|-\left|E_i,E_j\right|}$$

(6)计算加权相似度

综合考虑上述各因素，则服务 WS_i、WS_j 之间基于功能描述的相似度可表示为：

$$\begin{aligned}sim\left(DIOPE_i,DIOPE_j\right)=&X_1sim\left(D_i,D_j\right)+\\&X_2sim\left(I_i,I_j\right)+\\&X_3sim\left(O_i,O_j\right)+\\&X_4sim\left(P_i,P_j\right)+\\&X_5sim\left(E_i,E_j\right)\end{aligned}$$

其中，X_1、X_2、X_3、X_4、X_5 表示权重，值的大小由基本描述、输入、输出、前提及效果在服务功能描述中所起的重要性所决定。

3.2　语义 Web 服务聚类预处理

文献[10]采用模糊聚类算法对服务进行聚类预处理，但该文献在聚类时只将服务的输入、输出作为聚类参数。本文将服务的输入、输出、前提、效果全部作为 FCM 算法的聚类参数，这样对语义 Web 服务进行模糊聚类预处理时，对服务发现的准确率会更高。由于 FCM 算法的数据集都是在一个向量空间中来计算的，而语义 Web 服务无法映射到一个向量空间，只能计算它们之间的相似度，然后通过相似度来判断算法的终止以及聚类中心的更新。本文根据语义 Web 服务的特点，终止准则函数采用如下函数：

$$F\left(\boldsymbol{U},V\right)=\sum_{j=1}^{n}\sum_{i=1}^{c}\left(u_{ij}^{m}\left(DIOPE_i,DIOPE_j\right)^2\right)\tag{3}$$

其中，m 为模糊指标；$\boldsymbol{U}=\left\{u_{ij}\right\}$ 是隶属度矩阵，表示第 i 个点属于第 j 个聚类中心的隶属度。

$$u_{ij}=\sum_{i=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/m-1}\tag{4}$$

其中，$d_{ij}=sim\left(DIOPE_i,DIOPE_j\right)$ 是样本 x_i 和聚类中心 v_j 之间的相似度。并且，$0<u_{ij}<1$，约束为：$\sum_{i=1}^{c}u_{ij}=1;\ i=1,2\cdots,c;j=1,2,\cdots,n$。

聚类中心迭代函数还用原始公式，如式(2)所示，其中，u_{ij} 用式(4)来计算。

因此，其服务聚类算法描述如下：

(1)给定聚类数目 c，迭代误差 ε，以及模糊指标 m。

(2)随机从当前的语义 Web 服务集中选取 c 个 Web 服务作为聚类中心。

(3)其余的 Web 服务按照式(4)计算与这些聚类中心的隶属度，得到新的隶属度矩阵，并根据隶属度，把这些服务加入到相应的聚类中。

(4)根据式(2)，更新每个聚类的聚类中心。

(5)根据式(3)计算目标函数，将结果存入矩阵 $\boldsymbol{F}$ 中。

(6)若目标函数的差值小于阈值 ε，即 $F\left[i\right]-\left[i-1\right]<\varepsilon$，则算法停止；否则，转回步骤(2)继续迭代。

3.3　语义 Web 服务发现

语义 Web 服务发现就是根据请求者描述的服务需求，在已有的服务中查找最满足需求的服务的过程。在此过程中，匹配器需要对服务请求和服务注册中心的服务进行一一比较，由于注册中心的服务规模和数量都相当庞大，而每个服务的 OWL-S 服务描述也是比较复杂的，再加上本体库中概念之间描述关系的复杂性，服务的匹配效率成为语义 Web 服务发现中的一个瓶颈，因此本文在服务匹配之前先进行服务聚类预处理，过滤掉与服务请求完全不同类别的服务。从而避免在相似度较低甚至无相似的匹配计算上浪费时间，以提高服务匹配、发现效率。

基于模糊聚类优化的语义 Web 服务发现算法如下：

输入 服务请求 $R_{ws}=\left(R_I,R_O,R_P,R_E\right)$；基于功能属性相似的服务聚类集合 $D_1,D_2,\cdots,D_i$

输出 满足功能需求的服务集合 S

(1)初始化集合 S 为空。

(2)随机选择一个新的服务聚类集合 D_i，如果所有聚类集合匹配完毕，转步骤(4)。

(3)在集合 D_i 中任选一个服务 W_i 与请求服务 R_{ws} 进行功能匹配[11]，匹配完毕转步骤(2)。

(4)选出请求服务 R_{ws} 与所有聚类集合匹配度最大集合 D_i，这样可以过滤掉很多无用服务。

(5)将请求服务 R_{ws} 与选出的集合 D_i 中的服务进行功能匹配，按匹配度从大到小排序放入集合 S 中。

(6)输出满足功能需求的服务集合 S。

(7)从服务集合 S 中按用户关心的非功能属性进行选

择，从而选出最佳服务。

4 仿真实验

4.1 实验环境

该实验环境是在OWLS-MX[12]进行扩展。OWLS-MX[12]推理和信息检索混合的语义 Web 服务匹配器，但其匹配器只是针对服务的 I/O 进行匹配。本文实验在此基础上扩展了服务的 PE 匹配，使服务的发现能够同时基于服务的 IOPE 功能参数进行匹配。同时，考虑使用软计算技术来提高服务发现性能。因此，该实验在此基础上还扩展了模糊聚类即 FCM。在聚类时将服务的 IOPE 功能性参数同时作为聚类参数。实验条件：Java 编程语言，Eclipse 平台，OWLS-MX 匹配器，Weka 实验平台，MySQL 数据库，Pellet 推理工具。实验对象：1 000 条基于 OWL-S 语言描述的语义 Web 服务。

4.2 实验验证及分析

在该实验中，本体概念个数为 30~100，在服务注册中心注册了 1 000 个用 OWL-S 描述的语义 Web 服务。首先将服务注册中心的服务导入到数据库中。其次在 Weka 平台中扩展 FCM 模糊聚类算法，对数据库中的服务进行聚类预处理。在聚类算法中，将普通 FCM 算法中相似性计算的欧式距离算法替换为 OWLS-MX 匹配器中关于服务相似度计算的算法。将服务的 IOPE 功能性参数全部作为 FCM 的聚类参数，如果只把 I/O 作为聚类参数进行聚类，其聚类效果会与实际用户想要请求的服务有所偏差，甚至相差很远。I/O 只代表服务的输入/输出，但是用户的请求服务的前提/效果条件即 P/E 没有给出，没有充分考虑用户请求服务的前提及要到达的效果的聚类，会降低服务发现的效率。因此，该实验综合考虑了服务的输入、输出、效果、前提 4 个功能性参数，对聚类效果及服务发现准确性的影响，将 IOPE 全部作为模糊聚类参数。

由于模糊聚类预处理是在服务发现之前进行的，因此聚类本身并没有增加服务发现的时间。采用本文模糊聚类预处理方法既能提高服务发现效率，又能提高服务发现的准确性。

最后，对 OWLS-MX 进行扩展。之前的 OWLS-MX 匹配器只是按照服务的输入/输出即 I/O 进行匹配发现，考虑到服务匹配的准确性，便在其基础上扩展了服务前提/效果即 P/E 匹配。扩展后的 OWLS-MX 匹配器如图 1 所示。

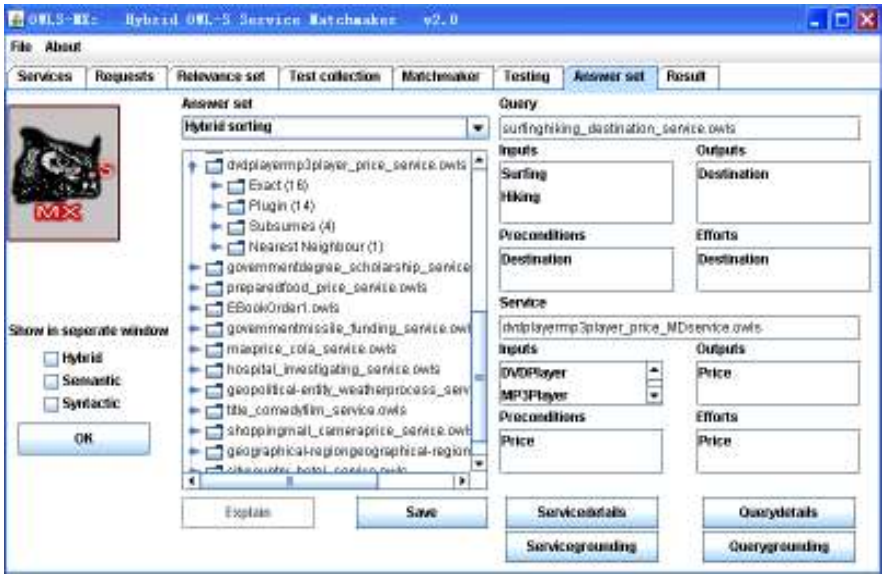

图 1 扩展后的 OWLS-MX 匹配器

以下 2 组实验结果是在此仿真实验中运行得出的。

实验 1 首先在服务注册中心注册不同数量的服务，然后对某一给定的服务请求，比较其在未采用聚类处理、采用加权图聚类处理和采用模糊聚类处理时能够发现的可行服务的效率及数量。结果如表 1 所示。

表 1 3 种情况服务发现数与运行时间比较

服务中心的服务数	未采用聚类处理		加权图聚类处理		模糊聚类处理	
	发现服务数	运行时间/ms	发现服务数	运行时间/ms	发现服务数	运行时间/ms
100	15	10 063	15	3 980	13	3 997
200	31	20 102	29	4 974	30	5 000
400	59	41 031	51	6 981	54	7 010
800	101	83 350	63	9 472	75	9 721
1 000	124	100 578	83	10 735	109	10 983

从表 1 可以看出，模糊聚类的发现效率比加权图聚类略低一点，但其服务发现的数量有所增加；与未采用聚类处理相比，发现效率明显变快，而且服务数量越多，效果越明显。同时，考虑服务发现的数量与效率，本文模糊聚类方法较适合。

实验 2 考虑在未采用聚类处理、采用带 2 个服务参数(I/O)的模糊聚类处理和采用带 4 个服务参数(IOPE)的模糊聚类处理的 3 种情况下，针对某一个给定服务请求，在有相同的候选服务时，比较其服务发现的查全和查准的比率。结果如图 2 所示。从图 2 可以看出，本文由于在服务发现之前，采用模糊聚类对注册中心的服务进行了聚类预处理，并且在模糊聚类时，将服务的 4 个功能性参数全部作为聚

类参数，从而过滤掉了很多不相关的服务，使得服务的平均查全率达到 79.6%的同时，服务的平均查准率达到了 85.9%，查全率、查准率优于其他 2 种方法。

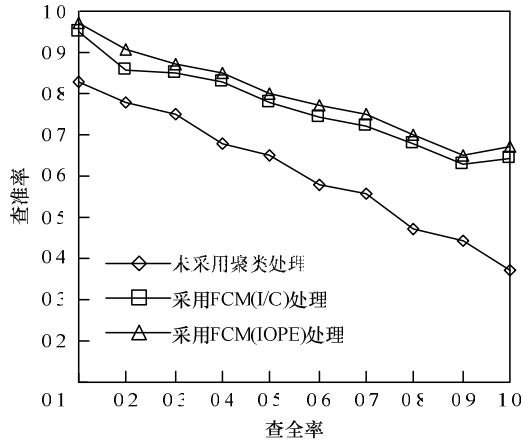

图 2 3 种情况查全率、查准率比较

5 结束语

本文提出一种基于模糊聚类优化的语义 Web 服务发现方法。一般的聚类预处理都只考虑服务的输入、输出，没有考虑服务的前提条件及调用服务后产生的效果，而本文综合考虑了服务发现的效率和准确率，在进行模糊聚类时，将语义 Web 服务的功能属性(IOPE)全部作为聚类参数，这样既能提高服务发现的效率，又能提高服务发现的准确性，而且采用模糊聚类又能扩大发现范围。实验结果表明，该方法在语义 Web 服务发现上能提高服务发现效率，还能在保证一定查全率的基础上提高查准率。但是，该方法在对语义 Web 服务进行聚类时，需要事先指定聚类，而且聚类数目直接影响着聚类结果，需要经过多次实验才能选出较好的聚类数目。因此，下一步将研究聚类数目，以提高对语义 Web 服务发现的实用性。

参考文献

[1] 徐小良, 陈金奎, 吴 优. 基于聚类优化的 Web 服务发现方法[J]. 计算机工程, 2011, 37(9): 68-70.

[2] 张景雨, 余雪丽, 付丰科. 利用聚类优化语义 Web 服务发现[J]. 计算机工程与应用, 2009, 45(34): 139-143.

[3] Ram S, Hwang Y, Zhao Huimin. A Clustering-based Approach for Facilitating Semantic Web Service Discovery[C]//Proc. of the 15th Annual Workshop on Information Technologies & System. Las Vegas, USA: [s. n.], 2006.

[4] Richi N, Bryan L. Web Service Discovery with Additional Semantics and Clustering[C]//Proc. of IEEE/WIC/ACM International Conference on Web Intelligence. Silicon Valley, USA: IEEE Press, 2007.

[5] Maria V, Castellanos Y, Miltiadis D L. CONQUIRO: A Cluster-based Meta-search Engine[J]. Computers in Human Behavior, 2011, 27(4): 1303-1309.

[6] Shariq B. Improving Retrievability with Improved Cluster-based Pseudo-relevance Feedback Selection[J]. Expert Systems with Applications, 2012, 39(8): 7495-7502.

[7] 李 曼, 王大治, 杜小勇, 等. 基于领域本体的 Web 服务动态组合[J]. 计算机学报, 2005, 28(4): 644-650.

[8] 吴 佳, 罗 可. 改进的模糊 C 均值的增量聚类算法[J]. 计算机工程与应用, 2011, 47(23): 141-143.

[9] 孙 萍, 蒋昌俊. 利用服务聚类优化面向过程模型的语义 Web 服务发现[J]. 计算机学报, 2008, 31(8): 1340-1353.

[10] Giuseppe F, Vincenzo L, Sabrina S. A Hybrid Approach to Semantic Web Services Matchmaking[J]. International Journal of Approximate Reasoning, 2008, 48(3): 808-828.

[11] Massimo P, Kawamura T, Terry R P, et al. Semantic Matching of Web Services Capabilities[C]//Proc. of the 1st International Semantic Web Conference. Sardinia, Italy: [s. n.], 2002.

[12] Mztthias K, Benedikt F, Karia S. OWLS-MX: A Hybrid Semantic Web Service Matchmaker for OWL-S Services[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(2): 121-133.

编辑 刘 冰

(上接第 218 页)

[4] 徐 德, 谭 民, 李 原. 机器人视觉测量与控制[M]. 北京: 国防工业出版社, 2011.

[5] 魏雅娟. DM6446 视频处理系统的硬件电路设计[J]. 光机电信息, 2011, 28(11): 80-83.

[6] 周建平, 刘歆浏. 基于 DM3730 平台的智能数字视频监控系统[J]. 兵工自动化, 2012, 31(5): 59-60.

[7] 杨 琤, 周富强. 镜像式单摄像机双目视觉传感器的结构设计[J]. 机械工程学报, 2011, 47(22): 119-123.

[8] 魏占国, 阚江明, 刘晋浩. 基于三维虚拟样机技术大型林业装备设计与研究[J]. 微计算机信息, 2012, 26(28): 8-10.

[9] 杨 磊, 李文彬, 阚江明. 基于数学形态学的立木树枝信息提取方法[J]. 森林工程, 2008, 24(3): 3-5.

[10] 张起贵, 张 胜, 张 刚, 等. 最新 DSP 技术——“达芬奇”系统、框架和组件[M]. 北京: 国防工业出版社, 2009.

[11] 彭启琮. 达芬奇技术-数字图像/视频信号处理新平台[M]. 北京: 电子工业出版社, 2008.

编辑 刘 冰