第38卷 第24期
Vol.38 No.24

计 算 机 工 程
Computer Engineering

2012年12月
December 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)24—0182—06 文献标识码：A 中图分类号：TP391

基于服务簇的空间信息服务自动发现

陈 科[1]，成 毅[1]，谢明霞[1]，艾 彬[2]

(1. 解放军信息工程大学地理空间信息学院，郑州 450052；2. 78138部队，成都 610036)

摘 要：现有Web服务自动发现方法中存在服务匹配程度区分不明显、服务发现精度不高等问题。为此，提出一种基于服务簇的空间信息服务自动发现算法。对发布的空间信息服务进行聚类分析，计算服务请求与各服务簇中心的相似度，由此确定最优匹配簇，根据服务请求与最优匹配簇中服务的语义相似度，得出服务请求的最优匹配服务。实验结果表明，该算法在实现对Web服务匹配程度定量表示的同时，能有效提高匹配程度的区分度和服务发现的查全率和效率。

关键词：空间信息服务；聚类；语义相似度；服务簇；服务发现；最优匹配簇

Spatial Information Service Automatic Discovery Based on Service Cluster

CHEN Ke[1], CHENG Yi[1], XIE Ming-xia[1], AI Bin[2]

(1. Geospatial Information Institute, PLA Information Engineering University, Zhengzhou 450052, China;
2. 78138 Troup, Chengdu 610036, China)

【Abstract】Aiming at the problem existing in the traditional methods for Web service discovery, such as unobvious distinction of service matching degree and low precision of service discovery, an algorithm for spatial information service automatic discovery based on service cluster is proposed. This algorithm clusters the advertised spatial information service into some clusters and selects the most matching cluster through computing the similarity between the service request and each clustering center. The most matching spatial information service is determined according to the semantic similarity between service request and each matching spatial information service belongs to the most matching cluster. Experimental results show that this algorithm can quantify the Web service matching degree while improving the distinction, service discovery recall and efficiency.

【Key words】spatial information service; clustering; semantic similarity; service cluster; service discovery; optimum matching cluster

DOI: 10.3969/j.issn.1000-3428.2012.24.043

1 概述

所谓 Web 服务发现，就是客户以某种方式在这些不同类型的 Web 服务中找到其所需要的服务，以执行 Web 服务请求[1]。Web服务发现是Web服务系统架构中的一个重要部分。近年来，面向服务的体系架构作为一种新的信息架构，逐渐被引入到科学应用研究中。

网络服务技术作为面向服务体系架构的一种实现，推动了空间信息共享与应用服务的发展。大量的空间数据、强大的计算资源，以及不同类别的空间数据处理功能，以Web服务的形式呈现给科研用户，将极大地提高数据分析和利用的程度。用户可用 Web 服务的形式共享自己的模型和数据产品。通过数据和服务的累积，一个开放的虚拟科研环境将形成并演化成为一个庞大的知识库系统[2]。由于空间数据来源多样、处理种类繁多、要素间关系复杂，造成了对空间信息服务的语义进行准确、全面、统一表达的困难；而且空间信息应用领域广泛，不同的领域对空间信息服务的理解和认知侧重点不同，造成了空间信息服务语义理解的不一致。这些问题最终阻碍了空间信息服务的有效发现和组合。在网络承载的海量信息环境下，如何在Internet 上快速地、智能地发现和使用这些大量的空间信息服务，成为当前国内外空间信息服务的研究热点。

目前，利用统一描述、发现和集成(Universal Description, Discovery, and Integration, UDDI)实现服务发现主要依靠数据的空间范围、服务的简单分类、服务描述关键字和服务接口简单的数据类型的匹配，这些匹配方式存在明显的不足，具体表现为以下5个方面：(1)关键字并不能够完整地体现用户的需求，不能准确描述所查询的目标。(2)遗漏大量与查询关键词相关或同义的信息。(3)不能度

基金项目：国家自然科学基金资助项目(41271392)；数字制图与国土信息应用工程国家测绘局重点实验室开放研究基金资助项目(GCWD201105)

作者简介：陈 科(1983－)，男，讲师、博士，主研方向：地理信息系统，空间信息服务；成 毅，副教授、硕士；谢明霞，硕士；艾 彬，高级工程师

收稿日期：2012-02-21 **修回日期：**2012-04-30 **E-mail：**chenk626@yahoo.cn

量待选服务同查询目标之间的匹配程度。(4)以 UDDI 为代表的传统服务发现技术是通过精确匹配实现的，不能使用细化、泛化和平级扩展等基于语义约束的模糊查询。(5)同一服务在不同站点上的不同接口导致的语义冲突问题。这些导致了服务发现的查准率和查全率不高，使得服务执行的整个过程受到影响。随着服务数量的日益膨胀，服务发现的难度也随之增加。

现有 Web 服务自动发现方法可以归纳为 2 类：(1)基于语义关系的 Web 服务发现：文献[1,3]提出的服务匹配方法仅考虑了服务和服务请求的输入输出参数之间的匹配，而未考虑其他信息对匹配度的影响，服务匹配准确性低；文献[2]仅通过语义关系实现服务匹配，不能对服务匹配度进行定量表示，服务匹配程度区分不明显；文献[4]提出一种融合了距离模型和信息量模型的增强语义精确度的 Web 服务发现方法，该方法不需要采用外在的文本集，但在语义相似度计算过程中，仅考虑服务输入输出概念间的包含关系。(2)基于分类的 Web 服务发现：文献[5]提出服务虚拟化的概念，但该虚拟化是通过手工构建的，由于不同的人对同一服务的认识不同，对服务的分类结果将千差万别，因此无法对其进行统一描述；文献[6-7]分别采用 Quality Threshold 和凝聚嵌套聚类算法实现对 WSDL 文档和 Web 服务的自动分类，但在 2 类方法中，都存在聚类数难以确定，且计算复杂度高的问题；文献[8]利用核 Batch SOM 神经网络聚类实现了 Web 服务的自动分类，从而有效减少了分类过程中的主观因素干扰。但是由于神经网络方法本身固有的缺点，因此严重影响了基于神经网络的 Web 服务分类规则的获取，以及分类算法的可用性和效率。为此，本文提出一种基于服务簇的空间信息服务自动发现算法。

2　基于服务簇的空间信息服务自动发现关键技术

定义 1(服务簇) 将 Web 服务集 $S=\{s_1,s_2,\cdots,s_n\}$ 进行聚类操作，获得 k 个聚类簇 $C=\{c_1,c_2,\cdots,c_k\}$，每个聚类簇即代表一个服务簇。

定义 2(最优匹配簇) 假设有 n 个 Web 服务 $S=\{s_1,s_2,\cdots,s_n\}$，服务请求为 r，将 Web 服务集 S 聚类成 k 个服务簇 $C=\{c_1,c_2,\cdots,c_k\}$，则服务请求 r 的最优匹配簇为 c_l，若：

$$\|c_l-r\|=\max_{1\le i\le k}\|c_i-r\| \tag{1}$$

其中，c_i 为各服务聚类簇的中心。

2.1　本体层次树描述

定义 3(概念树) 领域本体中包含空间信息服务概念集的子树 T。

设置领域本体层次树顶点的概率值为 1，即其信息量为 0，根据其包含的子节点个数获取子节点的概率和包含的信息量，以此类推，计算领域本体层次树中各节点的信息量[4]，根据式(2)计算树 T 中各边的权值：

$$w(c,p)=\frac{1}{m}\frac{IC(c)-IC(p)}{IC(c)+IC(p)} \tag{2}$$

其中，$IC(c)$ 为子节点的信息量；$IC(p)$ 为父节点的信息量；m 为父节点的子节点数量。由于子节点要比父节点更具体，因此其含有的信息量要比父节点多，则 $IC(c)\ge IC(p)$，保证了权值计算的非负性。各节点信息量根据下式计算：

$$IC(c)=-\mathrm{lb}P(c) \tag{3}$$

本体层次树中各概念信息量的计算如图 1 所示，其中，各框边数字表示对应节点的概率值；各连接边上的数字表示边权值。由图 1 可知，随着层次结构树中概念层次的深入，对应连接边的语义距离权重逐步减少，而随着概念密度的增加，语义距离也逐渐减小，这也符合基于网络距离模型方法总结出的概念语义距离和概念层次结构树之间的一般规律。根据获取的空间信息服务描述和输入输出概念集，分别获取领域本体中包含描述概念集和输入输出概念集中各概念的子树[9]，即描述概念树 T_D 和输入输出概念树 T_{IO}。同领域本体层次树中各节点信息量和各边权值的计算方法，获取各概念树中的节点信息量和边权值。

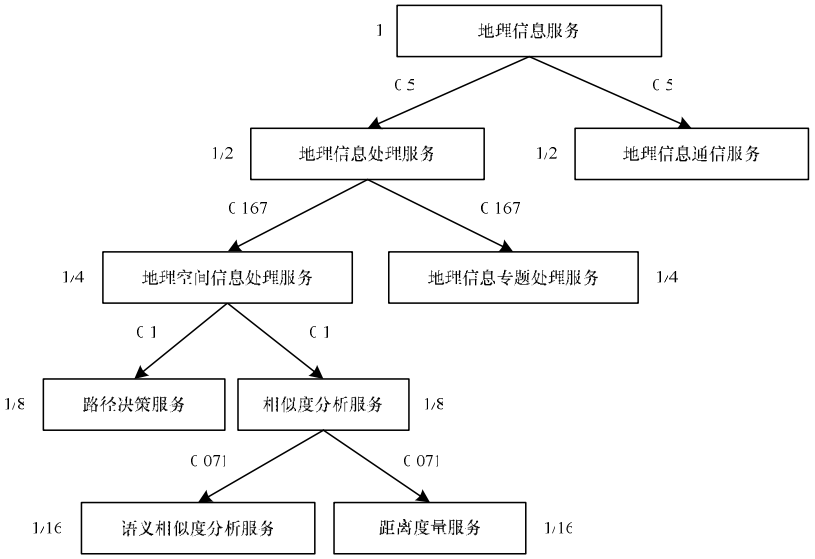

图 1　本体层次树中各概念信息量的计算

2.2 空间信息服务语义相似性度量

定义 4(概念关系) 设置概念间关系主要包括 *Exact*、*PlugIn*、*Subsume* 和 *Disjoint* 4 种，其中，各关系的语义相似性比较为 *Exact*>*PlugIn*>*Subsume*>*Disjoint*，各关系相对于概念“地理空间信息处理服务”而言，概念间的语义关系如图 2 所示。

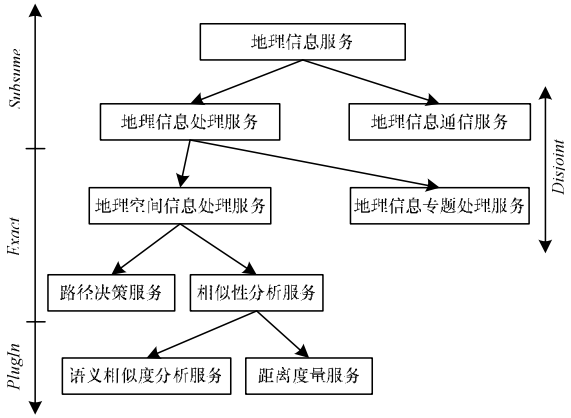

图 2 概念间的语义关系

计算概念集中各概念在概念子树中的相似度，概念间相似度计算方法如下：

(1)若 2 个概念 *Exact*，设置 $sim=1$。

(2)若 2 个概念 *PlugIn*，计算领域本体层次树中，连接 2 个概念节点的最短路径中各边的权值和 d，利用下式将权值和 d 转换为相似度 sim：

$$sim=1/(1+d) \tag{4}$$

(3)若 2 个概念 *Subsume*，概念间语义相似度为：

$$sim=w_s/(1+d) \tag{5}$$

其中，$0<w_s<1$，一般设置 $w_s=0.8$。

(4)若 2 个概念 *Disjoint*，$sim=0$。

对于空间信息服务的分类和匹配而言，定义服务的语义相似度是至关重要的一步，相似度定义得恰当与否直接影响空间信息服务分类和服务匹配的效果。应用于空间信息服务匹配的语义相似性度量方法的设计原则不同于应用于服务聚类的相似性度量方法，前者需要对服务的各属性信息进行全面分析[10]，从而获得服务之间的细划分；后者的目的仅是将功能大致相同的服务聚在一起，从而获得服务集的内在结构。前者针对的是单个空间信息服务，而后者则是空间信息服务类，即前者具体，而后者抽象。

在利用聚类分析对空间信息服务按照服务功能进行分类过程中，仅对服务的描述信息进行分词处理，获取服务的描述信息概念集及其相应的描述概念树，根据描述概念集中包含的各概念的信息量和概念间的语义相似度进行加权平均计算服务聚类相似度。在空间信息服务匹配过程中，服务语义相似性度量采用融合语义推理、距离模型和信息量模型，综合考虑了空间信息服务的名称、分类、描述、质量和输入输出等信息，兼容 *Exact*、*PlugIn*、*Subsume* 和 *Disjoint* 关系的空间信息服务全属性语义相似性度量方法。即首先获取空间信息服务的名称、分类、描述、质量和输入输出等各属性信息的概念集，然后判断名称和分类概念间的语义关系，若满足预设的关系阈值，分别获取描述概念树和输入输出概念树，根据相应的概念树对描述和输入输出概念集根据概念间语义关系进行概念映射(将 2 个概念集中语义关系最强的概念进行成组映射)，计算成组映射后的概念集的语义相似度，根据相似度阈值获取服务匹配的候选服务，最后计算各候选服务的质量符合程度，根据质量符合程度对候选服务进行排序。其中，概念集间的语义相似度根据下式进行计算：

$$Sim(i,j)=\frac{\sum_m\sum_n IC(m)IC(n)sim(m,n)}{\sum_m\sum_n IC(m)IC(n)} \tag{6}$$

其中，$m\in S_i$，$n\in S_j$，$IC(m)$ 是概念 m 在概念树 T 中的信息量；$IC(n)$ 是概念 n 的信息量；$sim(m,n)$ 是概念 m 和 n 在概念树 T 中的相似度。

2.3 空间信息服务分类

在同一领域中，采用不同分类标准形成的分类体系之间存在语义异构，遵循不同分类体系的系统之间不能直接进行语义映射，难以进行语义互操作，而基于空间信息服务聚类的分类方法可以通过对聚类中心的相似性比较，实现在不同的系统之间的类别映射。利用聚类技术对空间信息服务进行分类能够有效地提高服务自动发现的效率和精度。在空间信息服务分类中，首先根据多维尺度分析(Multidimensional Scaling, MDS)降维方法对服务进行二维映射，针对服务降维后的二维数据进行 K-means 聚类获取空间信息服务聚类簇[11]，将各服务簇中距离簇中心最近的二维坐标点对应的空间信息服务作为空间信息服务聚类簇的中心。利用服务聚类簇中心信息自动获取未标识的空间信息服务的分类信息的框架，如图 3 所示。

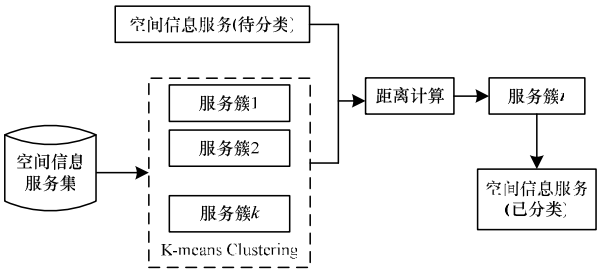

图 3 空间信息服务的自动分类框架

2.4 空间信息服务语义注册

对于服务注册，UDDI 采用按照标准分类的方法，由 Web 服务提供者将 Web 服务注册到 UDDI 适当的目录位置。对于 Web 服务提供者，必须了解服务注册中心的分类标准，以使得 Web 服务注册到合适的位置，从而便于后续的 Web 服务查找和发现。因此，UDDI 中的 Web 服务分类方法对 Web 服务提供者提出了要求。

对于基于服务簇的空间信息服务自动发现方法，在服务注册过程中，管理人员建立服务簇，并将各服务簇的簇中心服务在 UDDI 中进行标识，当服务提供商在注册新的空间信息服务时，通过服务聚类结果将服务自动注册在对应的服务簇中。基于语义描述和聚类的空间信息服务注册

框架如图 4 所示。

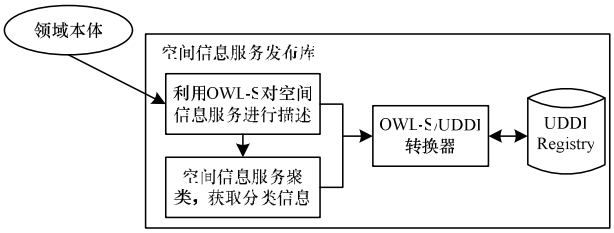

图 4　基于语义描述和聚类的空间信息服务注册框架

3　基于服务簇的空间信息服务自动发现

随着 Web 服务数量的增加，服务注册信息库不断膨胀，在判定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，很多时间会浪费在匹配比较一些毫不相关的服务上。因此，传统的 UDDI 服务发现机制已经不能满足日新月异的需求变化。

为了能够快速有效地对服务请求进行定位，首先对已有的服务进行聚类分析，通过判断服务请求与各服务簇的相似程度，将服务的搜索范围缩至最相似服务簇中，获取搜索服务的候选服务群，即服务请求最优匹配簇。根据 Web 服务聚类结果和提供的 Web 服务请求信息计算服务请求的最优匹配聚类簇过程算法如下：

Algorithm: SelectingMatchmakingServiceCluster
1. ServiceClusterSelection(r,C,k,e){
2. Input:requested service(r);service clusters(C={$c_1, c_2, \cdots, c_k$});
3. k:number_of_cluster; e:const number
4. Output:a selected service cluster MC;
5. $e \to +\infty$
6. Begin
7. for each cluster $c_i \in C=\{c_1, c_2, \cdots, c_k\}$ do
8. if $\|c_i,r\|<e$ then do
9. MC=c_i,e=$\|c_i,r\|$
10. end if
11. end for
12. Return MC
13. End
14. }

利用聚类技术进行空间信息服务匹配的框架设计，如图 5 所示，其中，序号表示处理顺序。

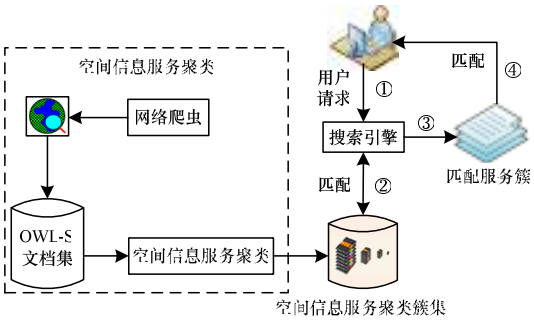

图 5　基于聚类的空间信息服务匹配框架

通过对服务进行聚类，可以方便地获取各空间信息服务的候选服务，基于服务簇的服务发现，对空间信息服务的动态变化性提供支持，提高空间信息服务应用的容错能力。基于服务簇的空间信息服务自动发现流程如图 6 所示。

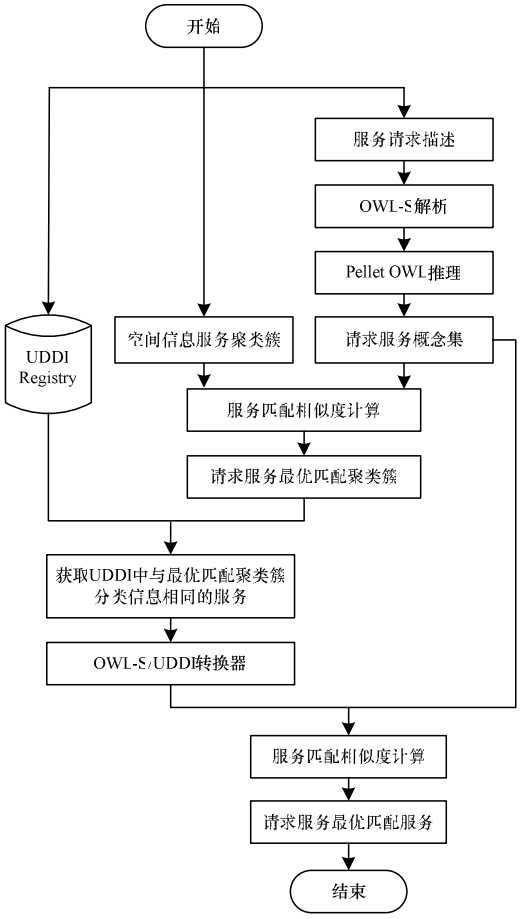

图 6　基于服务簇的空间信息服务自动发现流程

基于服务簇的空间信息服务自动发现具体步骤如下：

Step1　对空间信息服务进行聚类，获取服务聚类簇。

Step2　使用 Pellet OWL 推理机对服务请求信息进行语义推理，提取服务请求的描述概念集。

Step3　计算服务请求与各服务簇中心的语义相似度。

Step4　请求服务定位于最优匹配簇。

Step5　分别计算服务请求与最优匹配簇中各服务的语义相似度，根据相似度对最优匹配簇中的服务进行排序，最优匹配服务即为簇中相似度最高的服务。

在计算服务请求与各服务簇中心的语义相似度时，首先获取服务请求的描述概念集，然后根据 2.2 节介绍的概念映射方法对服务请求和各簇中心的描述概念集进行概念映射，最后计算各映射概念间的语义相似度，将其带入式(6)获取服务请求与各服务簇中心的语义相似度。

4　实验结果与分析

选取德国人工智能研究中心发布的基于 OWL-S 的 Web 服务语义匹配测试集 OWLS-TC4 中 geography 领域的 60 个 Web 服务，以及相关的 geographydataset、SUMO、protonu 和 protont 领域本体对服务的描述信息进行语义概念标注。根据 Web 服务的描述信息计算相似度矩阵，对相似度矩阵进行变换，利用 MDS 获取 Web 服务的二维映射坐标，在欧几里德二维空间中对其进行显示，测试服务

集 MDS 的降维后坐标如图 7 所示，其中，降维后坐标 1、降维后坐标 2 分别简称为维 1、维 2。根据二维可视化结果，选取初始聚类数为 $k=8$ 进行 K-means 聚类，测试服务集聚类后各聚类簇中心的二维坐标如表 1 所示。

选取测试集 geography 领域中的一个有关寻址的 Web 服务请求的 OWL-S 描述，其中，有语义关联的服务有 17 个，包括根据编码获取相关地址和经纬度坐标；根据位置的相关信息获取具体地名和经纬度坐标等。并利用相关领域本体对空间信息服务各属性信息进行概念标注。由于具有相同输入输出参数的服务不一定提供相同的功能，因此若仅根据服务的输入输出参数计算备选相关服务与服务请求间的匹配度时，则无法对实验中选取的备选相关服务进行区分，导致计算机无法自动确定哪种寻址服务更适合于请求 Web 服务。将语义相关的 Web 服务按照语义关联程度利用专家系统进行人工打分，如图 8 所示。

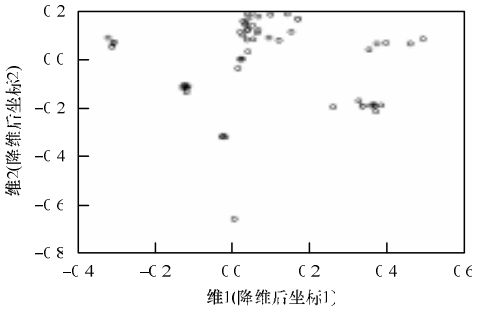

图 7 测试服务集 MDS 的降维后坐标

表 1 测试服务集聚类后各簇中心坐标

分类	簇 1 中心	簇 2 中心	簇 3 中心	簇 4 中心	簇 5 中心	簇 6 中心	簇 7 中心	簇 8 中心
维 1	0.001 7	0.068 3	−0.025 8	0.341 8	−0.123 8	0.020 4	0.411 6	−0.314 0
维 2	−0.652 5	0.139 7	−0.313 8	−0.187 7	−0.112 1	0.003 4	0.067 5	0.072 3

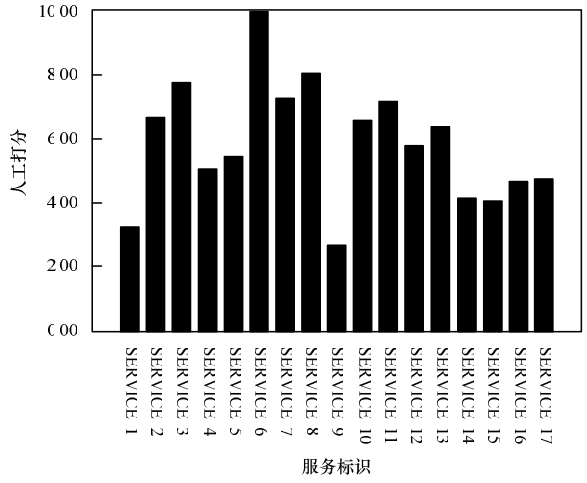

图 8 人工打分结果

计算服务请求与各聚类簇中心的距离，结果如表 2 所示。其中，距离是服务请求的二维坐标与各聚类簇中心的二维坐标之间的欧氏距离值。由于文中算法所涉及的数值都是相对值，没有实际的意义，因此距离值没有单位。根据表 2 获得最优匹配簇为簇 3，簇 3 中包含的服务即为根据空间信息服务的全属性语义相似性度量方法所获取的服务请求的候选服务。

表 2 服务请求与各聚类簇中心的距离

簇 1	簇 2	簇 3	簇 4	簇 5	簇 6	簇 7	簇 8
0.853 4	0.529 8	0.086 3	0.467 6	0.428 7	0.400 0	0.220 8	0.299 1

计算候选服务与服务请求的语义相似度，并根据语义相似度对候选服务进行排序。将人工判断选取的服务请求的语义关联服务按照打分结果进行排序，并分别与基于服务簇的空间信息服务自动发现、基于距离模型的顺序查找获取的前 17 个候选服务进行对比，结果如表 3 所示。从表 3 可以看出，语义支持的空间信息服务发现算法获取的服务请求的候选服务基本上完全包含了人工选取的语义关联服务，除了语义相似性关系最弱的 SERVICE_9 服务，且两者排序结果大体一致；基于距离模型的顺序查找方式获取的候选服务仅包含了部分人工选取的语义关联服务，且候选服务间的语义相似程度区分不明显，验证了空间信息服务全属性语义相似性度量方法的有效性，且在服务聚类分析和全属性语义相似性度量基础上提出的基于聚类簇的空间信息服务自动发现算法较顺序查找方式具有更高的查全率。使用顺序查找的方法将服务测试集中的所有服务都逐一地与服务请求进行匹配比较，则语义相似度的计算次数为 60；若利用本文算法的计算次数为 8+23=31。基于聚类簇的空间信息服务发现算法有效提高了服务发现的效率和精度。

表 3 请求服务的候选服务与人工选取的语义关联服务对比

人工选取的语义关联服务	基于距离模型的顺序查找		基于聚类簇的空间信息服务自动发现	
	候选服务	语义相似度	候选服务	语义相似度
SERVICE_6	√	1.000 0	√	0.944 2
SERVICE_8	√	0.750 0	√	0.784 0
SERVICE_3	√	0.583 3	√	0.780 7
SERVICE_7	√	0.583 3	√	0.701 6
SERVICE_11	√	0.500 0	√	0.688 5
SERVICE_2	√	0.392 9	√	0.610 6
SERVICE_10	√	0.400 0	√	0.632 7
SERVICE_13	√	0.450 0	√	0.622 1
SERVICE_12	×	0.250 0	√	0.648 2
SERVICE_5	×	0.250 0	√	0.574 3
SERVICE_4	√	0.375 0	√	0.594 0
SERVICE_17	×	0.250 0	√	0.533 2
SERVICE_16	×	0.250 0	√	0.544 7
SERVICE_14	×	0.250 0	√	0.581 4
SERVICE_15	×	0.250 0	√	0.470 4
SERVICE_1	×	0.250 0	√	0.548 0
SERVICE_9	×	0.250 0	×	0.392 7

5 结束语

为快速有效地提供服务请求者所需的 Web 服务，本文提出一种基于服务簇的空间信息服务自动发现算法。该算法在 Web 服务匹配过程中采用了有效融合语义推理、距离模型和信息量模型的空间信息服务语义相似性度量方法，通过多层次的匹配与推理，从服务的分类信息、服务描述、输入输出以及服务的质量信息等多方面，渐进获得更加准确的空间信息服务匹配信息，兼容了 Exact、PlugIn、Subsume 和 Disjoint 语义关系推理，通过服务簇来描述一类 Web 服务，利用服务簇来辅助 Web 服务的注册、查找。实验结果表明，该算法细化了 Web 服务间的匹配程度，使服务的区分能力得到增强，提高了服务发现的精度。

对于空间信息服务自动发现而言，定义服务匹配度是至关重要的一步，匹配度定义的恰当与否直接影响服务自动发现的精度和效率。空间信息服务是面向空间信息领域的一类特殊 Web 服务，在后续工作中将对空间信息服务的匹配评价指标进行深入研究。

参考文献

[1] 卢刘明. 基于语义的 Web 服务发现与组合关键技术研究[D]. 上海: 东华大学, 2006.

[2] 乐 鹏. 语义支持的空间信息智能服务关键技术研究[D]. 武汉: 武汉大学, 2007.

[3] Massimo P, Kawamura T, Terry R, et al. Semantic Matching of Web Services Capabilities[C]//Proc. of the lst International Semantic Web Conference. Seattle, USA: IEEE Computer Society, 2002.

[4] 白东伟. 基于语义的 Web 服务匹配与发现技术研究[D]. 北京: 北京邮电大学, 2007.

[5] Elgazzar K, Hassan A E, Martin P. Clustering WSDL Documents to Bootstrap the Discovery of Web Services[C]//Proc. of IEEE International Conference on Web Services. Miami, USA: [s. n.], 2010.

[6] 陈 蕾, 杨 庚, 张迎周, 等. 基于核 Batch SOM 聚类优化的语义 Web 服务发现机制研究[J]. 电子与信息学报, 2011, 33(6): 1307-1313.

[7] Guo Deke, Chen Honghui, Zhao Liang. Formalized Model and Implementation of Service Virtualization[C]//Proc. of the IEEE International Conference on Web Service. Orlando, USA: IEEE Computer Society, 2005.

[8] 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.

[9] 刘 丹, 卫金茂, 张 杰. GO 术语间语义相似性度量方法[J]. 东北师范大学: 自然科学版, 2010, 42(1): 36-40.

[10] 魏登萍, 王 挺, 王 戟. 融合描述文档结构和参引特征的 Web 服务发现[J]. 软件学报, 2011, 22(9): 2006-2019.

[11] 王家耀, 谢明霞, 郭建忠, 等. 基于相似性保持和特征变换的高维数据聚类改进算法[J]. 测绘学报, 2011, 40(3): 269-275.

编辑 刘 冰

(上接第 181 页)

白血病数据集本身比较好分类，分类的准确率要好于结肠癌的准确率，从真实数据实验结果来看，确实是这样的。但总体来看，DRSVM 算法效果好于 L_1-SVM 和 L_2-SVM。而经过修正的 IDRSVM 比文献[6]中提出的 DRSVM 算法效果要好，实验的执行步骤也要比 DRSVM 简洁得多。

6 结束语

L_1-SVM 和 L_2-SVM 是支持向量机中比较重要的 2 种算法，在对高维数、小样本的数据分类分析中表现出很多好的特性，但也有局限性。本文提出的 DRSVM 弥补了上述 2 种算法的缺憾。但该优化算法具有不等式约束且不可微的特点，使得数据的运算比较繁琐。运用正号函数等价改变约束条件，用二次多项式损失函数改进算法的可微性，使 DRSVM 算法改进成可微的、无约束的凸优化问题，便于运用更多的优化计算算法，如 BFGS 算法等来求解。模拟实验和真实数据证明，该改进算法可取得较好的分类准确率，且执行步骤简单。

参考文献

[1] Cortes C, Vapnik V. Support-vector Networks[J]. Machine Learning, 1995, 20(3): 273-297.

[2] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York, USA: Springer, 1995.

[3] Mangasarian O L, Musicant D R. Lagrangian Support Vector Machines[J]. Journal of Machine Learning Research, 2001, 22(1): 161-177.

[4] Zhu Ji, Rosset S, Hastie T, et al. 1-norm Support Vector Machines[M]//Neural Information Processing Systems. [S. l.]: MIT Press, 2003.

[5] Wang Li, Zhu Ji, Zou Hui. The Doubly Regularized Support Vector Machine[J]. Statistica Sinica, 2006, 16: 589-615.

[6] Wang Li, Zhu Ji, Zou Hui. Hybrid Huberized Support Vector Machines for Microarray Classification and Gene Selection[J]. Bioinformatics, 2008, 24(3): 412-419.

[7] Li Juntao, Jia Yingmin. An Improved Elastic Net for Cancer Classification and Gene Select[J]. Acta Automatica Sinica, 2010, 36(7): 976-981.

[8] 袁玉波, 严 杰, 徐成贤. 多项式光滑的支撑向量机[J]. 计算机学报, 2005, 28(1): 9-17.

[9] 袁亚湘. 最优优化理论与方法[M]. 北京: 科学出版社, 1997.

[10] 李光明, 田 捷, 赵明昌, 等. 基于 Hessian 矩阵的中心路径提取算法[J]. 软件学报, 2003, 14(12): 2074-2081.

编辑 顾逸斐