一种基于剖分的空间数据存储调度服务模型

杜根远[1] 苗 放[2] 熊德兰[1]
(许昌学院国际教育学院 许昌 461000)[1] (成都理工大学地球物理学院 成都 610059)[2]

摘 要 随着空间数据应用领域深度和广度的扩展,空间数据在组织、存储、更新、应用等方面存在速度、效率等难题。针对上述问题,基于地球剖分组织理论,结合面向客户端聚合服务的G/S模式架构,研究并提出一种空间剖分数据存储调度服务模型。给出了该模型体系架构、数据访问流程,设计了模型的地址编码结构及地址解析过程,形成了一种有效的"数据分散存储,客户端信息汇聚"的空间剖分数据组织管理、按需整合、快捷调度的机制,实现并验证了服务原型系统,其具有一定的理论意义和应用价值。

关键词 空间数据,全球剖分,G/S模式,面片,存储调度
中图法分类号 TP701 **文献标识码** A

Service Model of Spatial Data Storage Scheduling Based on Subdivision Theory

DU Gen-yuan[1] MIAO Fang[2] XIONG De-lan[1]
(International School of Education, Xuchang University, Xuchang 461000, China)[1]
(College of Geophysics, Chengdu University of Technology, Chengdu 610059, China)[2]

Abstract With the spatial data applications to the depth and breadth of expansion, spatial data exist for speed, efficiency and other problems in the organization, storage, update, applications and other aspects. In view of the above problems, based on global subdivision organization theory, combined with the client oriented aggregation service that G/S mode, studied and proposed a space subdivision data storage scheduling service model, and presented the model architecture, data access process, designs the address coding structure and resolution process, forming effective mechanism of space subdivision data organization and management, according to the needs of integration, fast scheduling with data distributed storage and client information aggregation. Finally realized and verified the prototype system, it has certain theory significance and the application value.

Keywords Spatial data, Global subdivision, G/S mode, Facet, Storage scheduling

地理信息,本质上是整合的,每一事物、要素都包含空间位置信息,空间位置是信息之间惟一明显的标识。遥感数据具有实时性高、覆盖范围广、信息丰富客观等优点,已被广泛应用于航空航天、军事侦察、灾害预报及动态监测、资源勘探、土地规划与利用等诸多军事及民用领域。随着传感器、遥感平台、数据通信等相关技术的发展,通过遥感获取的空间数据量急剧膨胀,从而造成"空间数据的生产和传输能力远远大于空间数据解析能力"的局面;同时,应用领域对遥感影像的实时性、精度和可靠性要求也越来越高,处理速度已经成为遥感影像快捷应用的瓶颈,这对以遥感影像为基础的空间数据的存储、组织和服务提出了更高的要求。目前研究热点及问题主要存在以下几个方面:数据组织及存储、存储资源的分配调度、地球数据和地球模型的结合、数据快速更新、数据快捷处理及应用等[1]。

本文基于地球剖分组织理论,结合面向客户端聚合服务的G/S模式架构,借鉴TCP/IP协议体系分层自治原理,针对空间数据访问过程中每一层所涉及的共性问题,建立标准化体系结构;研究并提出一种适用于空间信息领域的地球剖分数据存储调度服务模型,给出了空间剖分数据网络服务体系的架构、数据访问流程,设计了剖分数据存储调度服务模型的地址编码结构及地址解析过程,形成了一种有效的"数据分散存储,客户端信息汇聚"的空间剖分数据组织管理、按需整合、快捷调度的机制;实现并验证了服务原型系统。

1 相关理论研究

1.1 地球剖分组织理论

全球空间数据组织是指按照一定规则对全球空间数据进行有序配置,主要包括空间数据地理空间索引架构及在此基础上的分割、编目、存储及相应的编码、表达与调度体系。数据组织的优劣将直接影响数据的检索效率及应用性能。传统

到稿日期:2011-12-02 返修日期:2012-03-30 本文受国家自然科学基金项目(61071121),河南省科技攻关(重点项目)计划(112102210079),河南省基础与前沿技术研究计划(102300410060),河南省高等学校青年骨干教师资助计划(2010GGJS-177),河南省教育厅自然科学研究计划(2010A520035)资助。

杜根远(1974-),男,博士,副教授,CCF会员,主要研究领域为空间信息技术、图像智能处理,E-mail:genyuan_du@sina.com;苗 放(1958-),男,博士,教授,博士生导师,主要研究领域为空间信息技术;熊德兰(1980-),女,硕士,讲师,主要研究领域为空间信息技术。

的基于地图的空间信息表达、组织、管理和发布方式不能满足全球空间数据管理的需要[2]。由于较大的浏览跨度下地球的曲面特征非常显著，传统的平面坐标系统以及相关的理论与算法均难以适应，因此多年来地球剖分数据模型技术是地学及空间信息等学科的研究重点。同时，矢量数据的传统分幅存储模式不利于全球空间数据的统一表达、管理和应用。因此，构建一个新的基于全球的、多尺度、融合空间索引机制、无缝、开放的层次性空间数据管理框架，并基于此框架实现各类空间数据的表达和组织成为实际应用中亟待解决的问题[3]。

针对异构海量空间数据如何高效管理和快捷应用，地球剖分格网理论(Global Subdivision Grid，GSG)是解决该问题的一个研究热点。该理论是一种多层次、多尺度的基于全球格网划分的数据组织方式，具备独特的性质及在空间信息表达与管理上的优势。GSG 研究如何将地球(或球面)剖分为形状规则、变形较小的层状面片(称为剖分面片或面片，Subdivision facet or facet)，是一种新的基于全球、支持多分辨率、多尺度变换，空间位置分布均匀、融合空间索引机制，无缝、开放的层次性空间数据管理框架，能够实现在全球范围内的海量数据存储、提取和分析，解决传统数据模型在全球范围内多尺度、海量数据和层次数据上存在的局限性，保证全球空间数据的空间表达是全球的、连续的、层次的和动态的模型[4-6]。

地球剖分组织理论采用面片格网剖分的思路，把地球表面剖分为无缝、层次的网格单元，每个单元均有全球唯一编码，为全球空间信息建立多级索引体系，用于解决空间信息的组织与管理难题。球面剖分数据模型是一种多层次、多尺度的基于全球格网划分的数据组织方式[5,7]，其直接决定了离散格网数据的存储方式和索引方式，影响数据的调度效率。球面网格模型可分为基于地理坐标系的球面网格和基于多面体剖分的球面网格系统，典型的剖分模型有：基于正八面体球面剖分的 QTM(Quaternary Triangular Mesh)模型[8]、基于正二十面体球面剖分的 EARPIH(sphere triangle quadtree model based on EARP IcosaHedron projection)模型[9,10]以及基于等经纬差剖分的 SIMG(Spatial Information Multi-Grid)模型[11]等。剖分模型一般采用四叉树结构和剖分编码来组织剖分面片，从而实现不同剖分层级之间，以及同一剖分层级中不同剖分面片之间相互关联的全球遥感影像剖分组织体系。

1.2 空间数据汇聚技术

网络服务是用于访问 Internet、并被其它应用使用的软件架构。空间数据网络服务是指在 Internet 上提供空间数据和地理功能服务，用户通过网络访问空间数据和功能，并把它们集成在自己的系统和应用中，而不需要额外开发特定的 GIS 工具或数据[12]。Shao 和 Li[13]提出面向服务的空间信息共享框架，对理论模型和技术特性进行了分析，并实现了其原型平台；李德仁等[14]提出一种新的基于可量测实景影像(Digital Measurable Image，DMI)的空间信息服务模式。

针对原有服务模式在处理空间数据时的不足，苗放等[15]提出了一种新的空间信息网络服务 G/S(Geo-information browser/Spatial data servers)模式，并定义了其概念和内涵，建立了相关理论体系。G/S 模式是以“数据分散、信息汇聚、服务聚合”为原则的架构体系，采用自适应和负载均衡的分布式服务器集群存储、管理海量空间数据，对分布式网络环境下的各种类型、格式的数据(集)进行组织、存储和管理，同时客户端对分布在网络上的数据和服务进行聚合。该模式按照“请求-聚合-服务”的客户端聚合服务工作机制，在客户端完成数据和功能的聚合，最终生成并实现各种空间信息服务，能有效地解决海量空间数据的组织管理和高效访问等难题。

1.3 空间数据剖分存储集群

空间数据剖分存储集群是以地球剖分组织理论为基础，由剖分面片区域所对应的唯一存储单元编码生成存储地址，剖分面片和单元地址之间相互映射的空时一体化存储管理架构[1]。其具有按照空间位置进行存储、全球多尺度、实时存储更新、按需动态扩展等特点。

2 剖分数据存储调度协议原型体系

2.1 协议原型体系

空间剖分数据存储调度协议体系(Geospatial Information Protocols，GeoIP)处于多级物理存储体系的高带宽链路之上、地球空间信息剖分组织系统之下，为空间信息剖分组织系统存储和获取空间剖分面片数据提供寻址、路由、传输控制等服务。该模型是用于描述空间信息剖分面片数据的存储、调度及管理的规则总和，是在物理存储实体和逻辑应用之间标识、定位和访问空间面片数据的协议体系，其为空间数据提供灵活多变、编目统一的存储、调度、分发等应用奠定了坚实的基础。

根据空间信息剖分组织理论，参考 TCP/IP 协议簇，对其划分层次，定义功能，确定基本架构。协议簇被划分为 5 层，分别为剖分数据应用层、剖分数据访问服务层、剖分数据逻辑组织层、剖分数据表示层、剖分数据存储层。其中，存储层和表示层体现存储调度，逻辑组织层体现数据调度，访问服务层体现服务调度。网中各节点具有相同的层次，各层中包含所必需的协议，各层对其它层而言是透明的，不同节点的同等层次具有相同的功能，同一节点相邻层之间通过接口通信。每一层使用下层提供的服务，并向其上层提供服务，不同节点的对等层按照协议实现它们之间的通信。

2.2 原型体系架构

图 1 为空间剖分数据网络服务体系架构，该架构共分为 5 层，即剖分数据应用层(application layer)、剖分数据访问服务层(access service layer)、剖分数据逻辑组织层(logical organization layer)、剖分数据表达层(presentation layer)、剖分数据物理存储层(physics storage layer)。

应用层在低层协议的基础上解决面向各个特定领域的实际应用问题，解决如何扩展日益丰富的空间信息应用的问题，在低层的基础上处理特定应用程序的细节问题。

访问服务层针对空间信息应用丰富多样、不同应用对数据有多样化要求的特点，建立统一、易用的剖分数据访问界面。这是该层的主要任务，解决的主要问题包括：剖分存储集群系统数据统一访问视图、服务资源访问控制、服务多等级管理和基于通用剖分服务的资源实时调度等。涉及的主要功能和协议包括：地理坐标与剖分模型转换协议、剖分解析/聚合协议 Geo-DNS(Geo-Domain Name Server)、通用服务访问界

面、二次开发接口等。

逻辑组织层针对剖分数据超大规模数据量和相对集中的数据处理要求，研究适应剖分存储集群特点的剖分数据逻辑组织机理，重点解决剖分数据海量存储和管理热点数据快速响应等问题。涉及的主要功能和协议包括：热点数据快速响应机制、海量数据管理、并行加载协议 Geo-DPP(Geo-Digital Parallel Processing)、动态索引机制 Geo-ARP(Geo-Address Resolution Protocol)等。

表达层解决空间信息多层次、多尺度、多属性、多比例尺的数据表示问题，通过数据模型的建立实现地球表面任意范围、任意尺度快速应用、快速访问。该层将重点解决剖分数据资源的调度问题，建立剖分存储资源与剖分面片间的逻辑映射关系，利用空间信息的区域化访问特征，在地球统一剖分编码、编址的基础上，实现空间信息存储资源的有序存储、按需扩展、绿色存储、即插即用，提高可靠性、高可用性存储调度，为上层实现数据索引和虚拟全在线管理提供有效的支撑。涉及的主要功能和协议包括：多层次、多尺度表示、空间属性访问协议、编码协议、表达协议等。

物理存储层具体解决剖分数据物理存储组织问题，针对具体文件系统和物理存储设备的特点，根据剖分数据编码方式和剖分数据表示方式将剖分数据、各种属性数据存储到相应的存储单元和对象中，建立剖分数据物理存储系统。主要涉及物理访问协议、物理存储协议、传输控制协议、资源调度协议等。

图 1　空间剖分数据网络服务体系架构

2.3　协议支持下的空间数据访问流程

当进行空间数据访问时，从上层到下层的数据流程为：首先通过区域解析协议，将地理坐标信息转换成统一的剖分面片编码，单一面片可能解析成一定剖分细节的剖分面片集合；其次，根据剖分面片编码进行剖分数据索引，查询得到相应的存储结点；再次，根据其属性信息(空间位置、分辨率、尺度等)进行属性访问寻址，查到剖分面片集合特定属性的剖分数据；最后，通过数据存储层的相关路由寻址和传输协议完成数据的定位。

从下到上的数据流程是一个反向流动的过程。不同节点的对等层按照协议实现之间的透明通信，通信双方在对等层次上进行，在不对称层次上不能进行通信。

3　剖分数据存储调度模型研究

3.1　调度模型

空间剖分数据存储调度服务模型描述地球空间剖分数据和分布式空间剖分数据存储服务器群的存储单元之间的映射关系。地球空间数据进入剖分体系后，首先进行数据预处理，然后根据空间数据元数据中所提供的分辨率/比例尺信息，来确定该数据在剖分模型中的级数，即剖分层次；根据数据中心点的经纬度坐标确定所对应的中心剖分面片位置；之后，根据待处理空间数据的左上、右下角经纬度坐标，按照剖分面片的大小对空间数据进行剖分处理，确定该数据所包含的剖分面片集合；按照剖分组织理论中的面片地址码对剖分面片进行编码；按照剖分编码对数据进行组织，得到按剖分面片组织的空间数据。

全球空间信息剖分编码模型是以全球空间信息剖分模型为基础，结合剖分面片的地址码和数据属性信息，对全球空间信息进行编码的一种编码模型。剖分面片的层次性决定了剖分面片编码之间的信息传递性，即根据上一级的剖分面片编码可以得到下一级子面片的地址码，并且下一级剖分面片的地址码包含了上一级剖分面片的地址码。剖分面片编码由剖分面片的地址码和属性码两部分组成。

空间剖分数据存储调度服务模型是在物理存储实体和逻辑应用之间标识、定位和访问地球空间信息剖分面片数据的协议体系。其中，GeoIP 地址生成算法根据剖分面片编码建立剖分面片逻辑地址到物理存储地址的映射关系，进而生成 GeoIP 地址，用于标识剖分面片的存储位置。最终由 GeoIP 地址解析算法得到主机地址，根据地址映射表获得与地理特征区域对应的物理地址。

3.2　地址编码结构

空间信息物理存储实体通过 GeoIP 网卡进行组织和管理。GeoIP 网卡是具有数据处理功能的网络接口卡，是连接网络与存储介质的硬件设备，其内部建立了地理调度协议栈与文件系统，通过地理剖分数据调度协议地址编码与解析实现区域存储单元的定位与其内部存储体的访问。

一个完整的地理剖分数据调度协议地址编码共设 $m+n$ 位，由剖分编码(m 位)与主机地址编码(n 位)生成，用于标识剖分面片的存储位置。其中，主机地址码(n 位)通过地址映射表(Address Mapping Table，AMT)实现地理剖分数据调度协议地址与 GeoIP 网卡物理地址的一一对应，如图 2 所示。

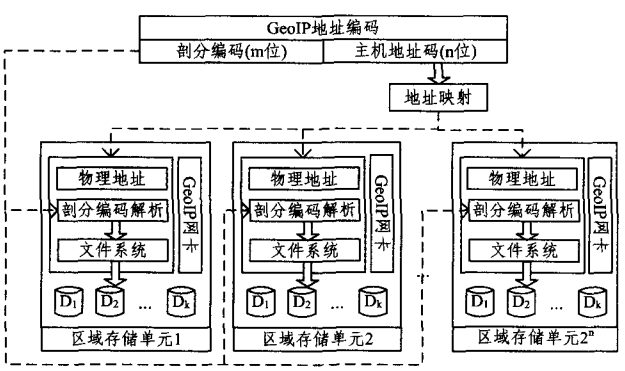

图 2　地址编码结构图(1)

基于地球剖分模型的剖分编码(m 位)结合了剖分面片的地址信息和空间实体属性信息,通过地理剖分数据调度协议体系对 GeoIP 地址编码中的剖分编码部分进行解析,获得需要访问的剖分面片的逻辑地址索引,进而通过集成在 GeoIP 网卡中的文件系统访问相应的物理存储体。

剖分面片数据的组织通过地理特征区及其包含的区域存储单元实现。区域存储单元由单个 GeoIP 网卡实现定位,其包含了若干物理存储实体。地理特征区由若干区域存储单元组织构成,可实现按广域地理特征的定位,如图 3 所示。

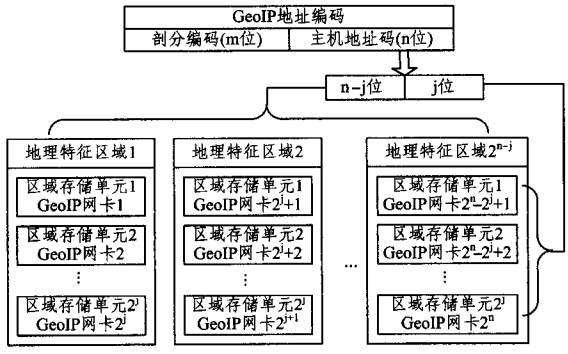

图 3 地址编码结构图(2)

其中,主机地址码分为 j 位和 $n-j$ 位两个编码段,其中 $0<j<n$。

$n-j$ 位段用于地理特征区寻址,共可组织构建 2^{n-j} 个地理特征区。j 位段用于实现地理特征区内的区域存储单元定位,每个地理特征区内可构建 2^j 个区域存储单元。

3.3 寻址流程

地理剖分数据调度协议寻址流程如图 4 所示。

① 通过屏蔽码 A,获取 GeoIP 地址编码中的主机地址码;

② 通过地址映射表,获得 GeoIP 网卡的物理地址,将其定位到区域存储单元;

③ 通过屏蔽码 B,获取逻辑地址索引;

④ 逻辑地址索引通过已定位区域存储单元中的文件系统,实现对物理存储体的数据访问。

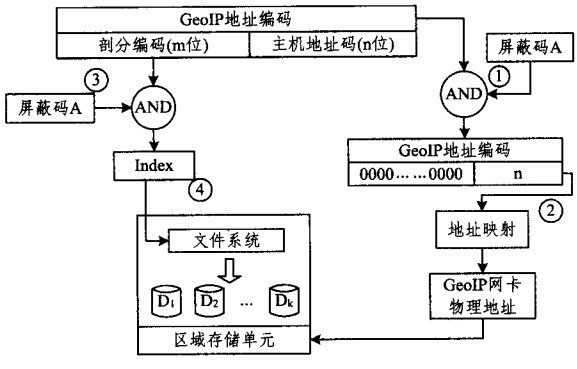

图 4 地理剖分数据调度协议寻址流程

4 服务应用

目前,空间数据应用正在经历从行业到公众的革命,业界推出了多个虚拟数字地球系统,用于无缝集成、表现和分析大范围乃至全球的多尺度、多类型的海量空间数据。课题组成员采用 WorldWind 的 Java 版 SDK 作为基础开发包,界面和容器采用 Eclipse RCP 应用程序,自主开发了一个数字地球平台原型系统。该平台系统已经在数字旅游、数字园区、虚拟/数字月球数据共享服务等项目中提供数据支持。图 5 为数字地球平台原型系统结构功能框架,图 6 对数字九寨进行了展示。

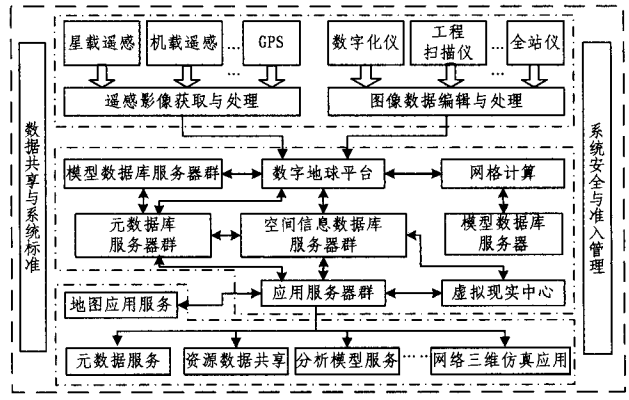

图 5 数字地球平台原型系统结构框架

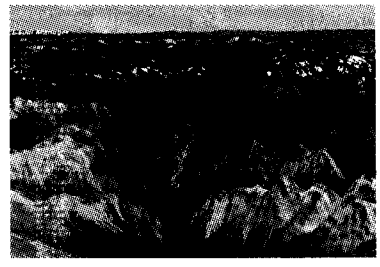

图 6 数字九寨

结束语 目前空间数据存在组织效率瓶颈问题和快捷应用难题,如查询检索速度慢、存取速度慢、整合应用慢。针对该问题,在地球剖分组织理论、面向客户端聚合服务的 G/S 模式研究基础上,研究并提出一种适用于空间信息领域的地球剖分数据存储调度服务模型。描述了该模型的体系架构、数据访问流程,设计了模型的地址编码结构及地址解析过程,形成了一种有效的"数据分散存储,客户端信息汇聚"的空间剖分数据组织管理、按需整合、快捷调度的机制。利用原型测试对上述思路进行的部分验证表明,该原型系统数据访问应用速度快、存储更新容易、对大数据适应,具有一定的理论意义和应用价值。

参 考 文 献

[1] 程承旗,吕雪锋,关丽. 空间数据剖分集群存储系统架构初探[J]. 北京大学学报:自然科学版,2011,47(1):103-108

[2] Goodchild M F. Discrete global grids for digital earth[C]// Proceedings of 1st International Conference on Discrete Global Grids. Santa Barbara, California, USA, 2000

[3] 关丽,程承旗,吕雪锋. 基于球面剖分格网的矢量数据组织模型研究[J]. 地理与地理信息科学,2009,25(3):23-27

[4] 宋树华,程承旗,关丽,等. 全球空间数据剖分模型分析[J]. 地理与地理信息科学,2008,24(4):11-15

[5] 程承旗,郭辉. 基于剖分数据模型的影像信息表达研究[J]. 测绘通报,2009(10):12-14,17

[6] 程承旗,宋树华,万元嵬,等. 基于全球剖分模型的空间信息编码模型初探[J]. 地理与地理信息科学,2009,25(4):8-11

[7] 袁文,程承旗,马蔼乃,等. 球面三角区域四叉树 L 空间填充曲线[J]. 中国科学 E 辑,2004,34(5):584-600

[8] Dutton G H. A hierarchical coordinate system for geoprocessing and cartography[J]. Lecture Notes in Earth Science, Berlin: Springer-Verlag,1999,79

[9] 袁文,马蔼乃,管晓静. 一种新的球面三角投影:等角比投影(EARP)[J]. 测绘学报,2005,34(1):78-84

[10] 袁文,庄大方,袁武,等. 基于等角比例投影的球面三角四叉树剖分模型[J]. 遥感学报,2009,13(1):103-111

[11] 李德仁. 论广义空间信息网格和狭义空间信息网格[J]. 遥感学报,2005,9(5):513-520

[12] 邓淑明,胡思仁. 地理信息网络服务与应用[M]. 曾杉,译. 北京:科学出版社,2004

[13] Shao Zhen-feng, Li De-ren. Design and implementation of service-oriented spatial information sharing framework in digital city [J]. Geo-Spatial Information Science,2009,12(2):104-109

[14] Li De-ren, Shen Xin. Geospatial information service based on digital measurable image-Take Image City Wuhan as an example [J]. Geo-Spatial Information Science,2010,13(2):79-84

[15] 苗放,叶成名,刘瑞,等. 新一代数字地球平台与"数字中国"技术体系架构探讨[J]. 测绘科学,2007,32(6):157-158,168

(上接第 238 页)

文献[10,11]所提多主题分类算法的比较结果。实验中使用的核函数为径向基函数 $K(x,y)=e^{-\gamma\|x-y\|^2}$。本文算法参数 $C_+=C_-=C$,M,文献[10]所提算法参数 C,文献[11]所提算法参数 v 以及径向基函数参数 γ 均通过交叉确认得出,其中折取 3;而本文算法经验参数 κ,文献[10]所提算法经验参数 θ,文献[11]所提算法经验参数 τ,则需通过大量实验得出,如表 2 所列。

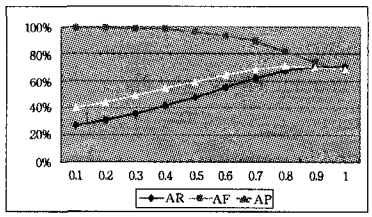

图 1 不同 κ 下 AP、AR、AF 的变化情况

表 1 本文所提多主题分类算法与其他算法的比较结果

算法	平均准确率	平均召回率	平均 F1 值
文献[10]多主题算法	43.13%	46.82%	43.62%
文献[11]多主题算法	49.45%	90.93%	58.30%
本文多主题算法	67.52%	80.94%	70.35%

表 2 参数设置

参数 (交叉确认范围,步长)	文献[4]多主题算法	文献[6]多主题算法	本文多主题算法
$C(2^0\sim2^5,2^1)$	2	——	16
$v(2^{-5}\sim2^{-1},2^1)$	——	0.0625	——
$\gamma(2^{-5}\sim2^5,2^1)$	0.03125	0.125	0.25
$M(2^1\sim2^5,2^1)$	——	——	32
θ	0.7	——	——
τ	——	0.7	——
κ	——	——	0.8

从表 1 可以看出,与文献[10,11]所提算法相比,本文所提多主题分类算法其平均准确率、平均召回率和平均 $F1$ 值都有明显的提高。

结束语 本文提出了一种多主题分类算法,它结合最大间隔最小体积超球支持向量机和模糊理论实现了多主题分类。实验结果表明,该算法与传统算法相比,具有更好的准确率、召回率和 $F1$ 值,是一种更为实用的处理多主题分类的算法。

参 考 文 献

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

[2] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32-42

[3] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Feature[C]// Proceedings of ECML-98,10th European Conference on Machine Learning. Berlin: Springer,1998:137-142

[4] 马金娜,田大钢. 基于支持向量机的中文文本自动分类研究[J]. 系统工程与电子技术,2007,29(3):475-478

[5] 崔国勤,高文. 基于双层虚拟视图和支持向量的人脸识别方法[J]. 计算机学报,2005,28(3):368-375

[6] 谢赛琴,沈福明,邱雪娜. 基于支持向量机的人脸识别方法[J]. 计算机工程,2009,35(16):186-188

[7] Krebel U G. Pairwise Classification and Support Vector Machines[C]// Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT press,1999:255-268

[8] Bennett K P. Combining Support Vector and Mathematical Programming Methods for Classification[C]// Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT press,1999:307-326

[9] Platt J C, Cristianini N, Shawe-Taylor J. Large Margin DAGs for multiclass classification[C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 547-553

[10] 王晔,黄上滕. 基于支持向量机的文本兼类标注[J]. 计算机工程与应用,2006,42(2):182-185

[11] 艾青,秦玉平,李迎春. 基于超球支持向量机的多主题文本分类算法[J]. 计算机工程与设计,2010,31(10):2273-2275

[12] 文传军,詹永照,陈长军. 最大间隔最小体积球形支持向量机[J]. 控制与决策,2010,25(1):79-83