

# 基于 OpenGIS WMS 的空间信息 搜索引擎系统原型

白玉琪 杨崇俊 刘冬林 朱华吉 卢亚辉 芮小平

(中国科学院遥感应用研究所, 北京 100101)

**摘要** 面向在线空间信息的搜索引擎是空间信息共享和互操作领域一个新的研究内容, 为了帮助用户能够自动地找到感兴趣的空间信息, 提出了一种遵循 OpenGIS WMS 规范构建的空间信息搜索引擎系统原型, 并详细介绍了研究背景、设计方案、系统实现以及性能测试结论。实践证明, 该搜索引擎不仅能自动地搜索互联网上的 WMS 服务器, 并能不断更新和维护相关的 WMS 信息数据库, 还能根据用户的查询需求实现动态的服务连接和透明的空间信息获取。

**关键词** 空间信息 网络搜索引擎 网络制图 开放式地理信息系统 网络地图服务

中图法分类号: TP391.3 P208 文献标识码: A 文章编号: 1006-8961(2004)01-0105-07

## OpenGIS WMS Based Prototype of Spatial Information Search Engine

BAI Yu-qi, YANG Chong-jun, LIU Dong-lin, ZHU Hua-ji, LU Ya-hui, RUI Xiao-ping

(Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101)

**Abstract** Spatial Information Search Engine (SISE), aiming to search the World Wide Web for online geographic information to meet the end users needs, is totally a new research area in the field of spatial information sharing and geographical information system interoperability. In this paper we propose a prototype system of SISE based on the OpenGIS WMS specification. The system architecture, working principles, detailed function of each component, implementation strategies and system performance test results of this SISE prototype are introduced. This SISE prototype is capable of discovering available WMS Servers dynamically, saving and updating capabilities information of them into relational databases and maintaining necessary R tree indexes. For each spatial information request from clients, it can choose qualified WMS servers according to the request autonomously and fetching the matched geographical information for them from remote WMS servers transparently.

**Keywords** Spatial information, Search engine, Web Mapping, OpenGIS, Web Map Service

## 引言

借助于互联网制图技术(Web Mapping), 基于 Internet 的空间信息共享已经有了长足的进步。如今互联网上提供数字图像和地图数据等空间信息的站点数量与日俱增, 其所提供的空间数据的种类和数量也越来越多, 从元数据到卫星图像、航空像片、地理基础图和专题图, 几乎无所不包<sup>[1]</sup>, 也就是说, 随着空间信息网络应用的不断深入, 人们可以在线

访问到越来越多的空间信息。

但是由于空间信息具有海量性的特点, 因此至今没有一个单独的网络空间信息系统能够存储和管理所有的空间数据, 以满足最终用户的空间信息查询需求, 此外, 由于已经存在的网络空间信息系统之间缺乏有机的联系, 人们很难直接有效地访问这些空间信息, 而实际上逐步增多的空间信息站点也逐步增加了最终用户查询的困难, 因此, 需要一个能够自动搜索在线空间信息, 以满足用户查询需求的搜索引擎。

互联网上的空间信息一般以作为静态的网页成分

和作为服务器动态的响应内容两种形式存在,其中,前者指网页中的描述性文本或者嵌入的静态图像;后者指由服务器根据用户请求,通过查询后台相应的空间数据库来实时动态地向客户端返回的响应内容。

考虑到近年来用先进的对地观测技术获得的空间信息量迅速增长,为了满足最终用户查询的现势性和准确性、海量空间信息存储和管理的可扩展性和可维护性等要求,空间信息网络发布只能采用在后台专有服务器支持下的实时动态响应方式。在这种情况下,由于互联网上的空间信息提供者主要是网络空间信息系统,即 Web-GIS 系统,因此,如何实现面向这些网络空间信息系统的自动发现、实时连接和动态获取是在线空间信息自动化搜索的主要研究内容。

从 1995 年开始,互联网的信息搜索逐渐成为研究热点。由于 Gudivada、Lawrence 等人详细论述了面向 Internet 的信息搜索的一般概念和关键问题<sup>[2~4]</sup>,而 Brin 和 Page 又提出了可扩展的系统设计<sup>[5]</sup>,并提出了非常独特而有效的用于衡量网页质量的 PageRank 算法<sup>[6]</sup>,从而最终研制出了互联网搜索引擎 Google。

如前文所述,互联网上的空间信息主要由网络空间信息系统通过动态响应方式提供,但 Google 等提出的通用的搜索引擎不能够直接访问到这些系统后台专有空间数据库中的空间信息。此外,尽管 Google 等搜索引擎目前提供了静态图像的查询功能,并在一定程度上可以查到图形化表示的空间信息,但由于这种查询缺乏语义支持,因此查询的准确度较低。事实上,由于缺乏详细数据的支持,致使这种仅仅依靠从互连网上获取零散图像的方式是不能够满足用户无级缩放和进一步查询的需求,因此 Google 这类以互联网上的文本和图像等静态的、无结构的信息为搜索对象的通用搜索引擎 GPSE (General Purpose Search Engine) 的构建技术路线,并不适用于构建面向在线空间信息的自动化搜索引擎 SISE (Spatial Information Search Engine)。

为了能够实现面向网络空间信息系统的空间信息自动化搜索,人们进行了大量有益的探索,从技术方案的角度来看,主要是以下 3 种:设计地理标注<sup>[7]</sup>、建立空间元数据交换站<sup>[8]</sup>和实现地理信息网络<sup>[9]</sup>。

但是,地理标注只能适用于以静态的网页成分存在的空间信息形式;空间元数据交换站虽然能在很大的程度上帮助最终用户找到所需要的空间数据的提供者,但是却不能够帮助他们完成在线的、实时

的、透明的数据获取;美国环境系统研究所 ESRI 独立构建的地理信息网络 (Geography Network) 虽然可以帮助用户查询并访问到多种多样的空间信息 (包括动态地图、可下载的数据以及空间信息服务),但它仅仅支持 ESRI ArcIMS 地图服务器,所以仍不能够成为普适的解决方案。

由此可见,这些已有的技术路线都不能够满足面向在线空间信息的动态获取和实时响应的基本要求。经笔者在中国科技信息研究所查新发现:到目前为止,面向互联网的空间信息搜索引擎的研究和实现尚未见相关报道。

在“十五 863”前期研究的基础上,笔者提出了基于 OpenGIS WMS (Web Map Service)<sup>[10]</sup> 规范实现空间信息搜索引擎的技术方案。本文在简单介绍该规范后,将集中介绍基于该方案的空间信息搜索系统原型设计和实现细节。

## 2 网络地图服务规范 (OpenGIS WMS)

OpenGIS WMS 规范作为 OpenGIS 的一个实现规范,它的适用对象是以向客户程序提供动态生成的静态图像文件 (GIF, JPG, PNG) 的方式提供地理信息服务的地图服务器。

该规范定义了这些地图服务器应该遵循的服务描述方式规范以及应该提供的接口参数规范和对请求的预定义响应规范。

具体地说,它主要定义了这种地图服务器应当支持的操作接口: GetCapabilities、GetMap 和 GetFeatureInfo。这些操作目前仅支持基于超文本传送协议 (HTTP) 的分布式计算平台,后续规范的版本中将支持更多的分布式计算平台。

### 2.1 GetCapabilities 操作

GetCapabilities 操作用于向客户端提供当前地图服务器可以提供的空间信息类型和范围、具体的图层信息和显示样式、支持的查询方式、没有查询结果时 (异常) 的缺省返回信息格式等服务描述信息。

这些信息通常先组织成一个引用了固定的文档类型定义 (DTD) 的 XML 实例文件,再由服务器通过 application/vnd. ogc. wms\_ xml 的 MIME 类型发送给客户端程序。

### 2.2 GetMap 操作

GetMap 操作的目的在于请求服务器生成一幅具有确定地理位置坐标范围的地图图像,但按照

WMS 规范,这个操作需要明确地指定出操作本身遵循的 WMS 规范的版本号以及需要显示的具体图层、对应的坐标范围、返回图像的大小和格式等。

### 2.3 GetFeatureInfo 操作

GetFeatureInfo 操作向 WMS 的客户端程序提供了进一步查询特定空间实体信息的能力。这种操作往往是由客户程序在 WMS 服务器先前返回的地图上指定了一个空间实体,进而提交查询而形成。

## 3 空间信息搜索引擎系统原型设计

由于 OpenGIS WMS 规范从语义和语法两个角度详细定义了 WMS 服务器向客户程序提供的操作接口,所以它提供了规范化网络空间信息系统的功能。如以遵循该规范的 WMS 服务器为直接的搜索对象,并以它们所提供的空间信息为间接的搜索对象,则可以实现自动化的在线空间信息搜索。下面给出基于该规范的空间信息搜索引擎系统原型的设计方案。

如图 1 所示,该系统的搜索对象是互联网上独立运行的 WMS 服务器,它们都遵循 OpenGIS WMS 规范;它的服务对象是图中右侧所示的空间信息查询客户程序。这些客户程序分为如下两类:一类是通过系统提供的查询网页提交查询的普通客户程序;另一类是笔者开发的一种可以和客户端工作环境紧密集成的空间信息搜索入口程序——文图自通。

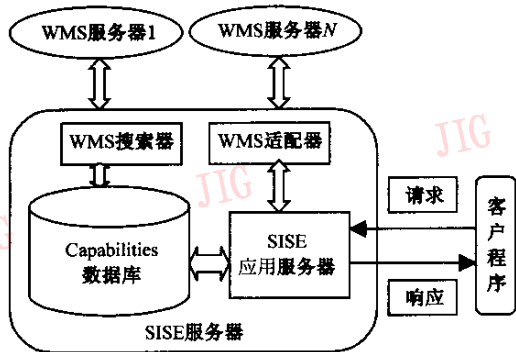


图 1 空间信息搜索引擎系统原型系统结构

按照本文的设计,空间信息搜索引擎原型包括 WMS 搜索器,Capabilities 数据库,WMS 适配器和 SISE 应用服务器 4 个模块。该 4 个模块紧密集成,其中 WMS 搜索器主要工作在后台,7×24 小时运行,而 WMS 适配器和 SISE 应用服务器则负责实时地处理客户程序提交的空间信息查询请求。

### 3.1 WMS 搜索器

WMS 搜索器的主要任务在于通过大规模并行的互联网搜索发现处于工作状态的 WMS 服务器。它的搜索策略就在于利用 WMS 服务器对 GetCapabilities 操作进行预定义响应,它的工作原理是基于站点爬行和网页遍历。对于新发现的 WMS 服务器,该模块负责获取其服务描述文件,并以结构化的形式将其存储在 SISE 服务器的 Capabilities 数据库中。同时,该模块还能够定期更新这些服务描述信息,以最大程度地保持现势性。

具体地说,WMS 搜索器模块包含主控线程和搜索线程,其负责执行具体 WMS 服务器搜索任务的搜索线程的程序流程图如图 2 所示。

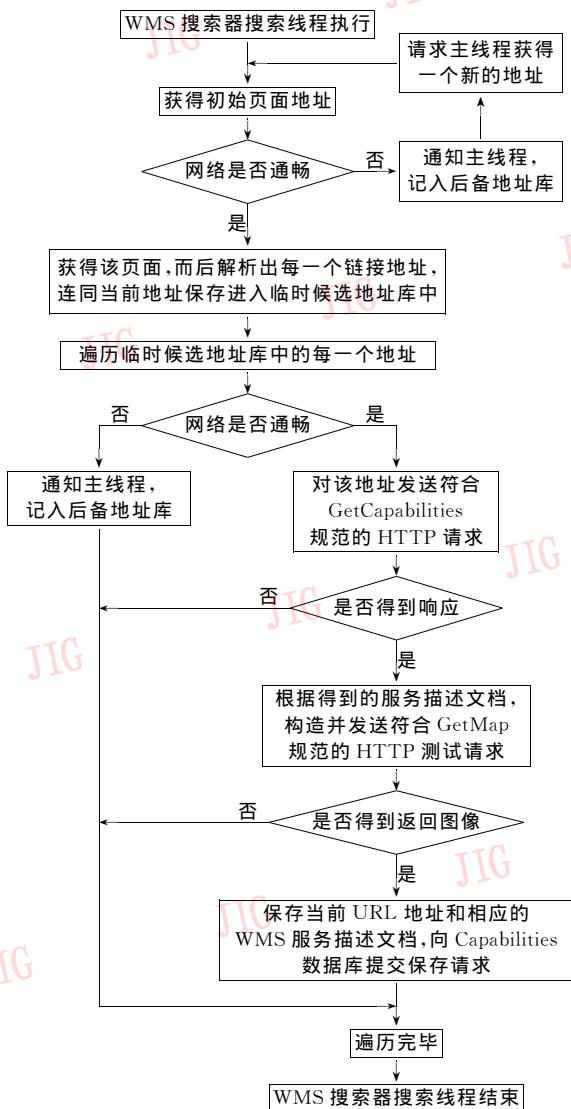


图 2 WMS 搜索器搜索线程的程序流程图

### 3.2 SISE 应用服务器

SISE 应用服务器负责处理客户程序的请求,即

首先根据客户查询请求所对应的空间范围,从 Capabilities 数据库中筛选出候选的 WMS 服务器,然后通知 WMS 适配器启动独立进程动态地连接这些 WMS 服务器,并提交相应的 WMS 查询请求。在得到 WMS 适配器的响应后,它则根据原始请求的类型,把响应消息组织后反馈给客户程序。

### 3.3 WMS 适配器

WMS 适配器接受 SISE 服务器的调度,并首先根据其指定的目标 WMS 服务器的服务描述信息进行动态连接,然后实时生成符合相应 WMS 规范的查询请求后发送给这些 WMS 服务器,并在分别解析完它们的全部回应消息后,再按照约定的内部格式回送给 SISE 应用服务器。

### 3.4 Capabilities 数据库

Capabilities 数据库负责以结构化的形式把每一个 WMS 服务器的服务描述信息存储起来,以备查询。同时,该模块还将根据每一个 WMS 服务器提供的图层信息的坐标范围,组织成一个 R 树索引结构。这个索引结构提供给 SISE 服务器,用于支持其根据客户程序请求的空间范围,快速地确定出候选的 WMS 服务器。

Capabilities 数据库中主要定义了 48 个表,表名如下: Service、GetLegendGraphic、UserDefinedSymbolization、ContactInformation、Capability、Dimension、LegendURL、AuthorityURL、Post、GetFeatureInfo、LatLonBoundingBox、LogoURL、StyleSheetURL、

GetStyles、Extent、HTTP、Exception、Style、Get、MetadataURL、PutStyles、Identifier、KeywordList、FeatureListURL、GetCapabilities、Attribution、DCPType、StyleURL、OnlineResource、ScaleHint、ContactAddress、BoundingBox、GetMap、Layer、DescribeLayer、DataURL、WMT\_MS\_Capabilities、ContactPersonPrimary、Request、Format7、Format6、Format5、Format4、Format3、Format2、Format1、Keyword、SRS。这些表都对应于 WMS 规范中服务描述文档 Capabilities.xml 中的相应元素,同时依据元素之间的约束关系定义这些表之间的关系。此外,Capabilities 数据库中还存储了相应的 R 树索引结构。

## 4 空间信息搜索引擎系统原型实现

为了验证上述设计方案,笔者建立了一个原型示范系统(其中引用了利用 WMS 搜索器技术搜索到的美国科罗拉多州 i-cubed LLC 公司和荷兰代夫特市 Demis 公司分别建立的 WMS 服务器)。这个系统的网络结构如图 3 所示。

其中,考虑到运行平台的特殊要求和较好的可移植性,空间信息搜索引擎系统采用 Java2 平台实现。

该系统已经实现了基于 WMS 1.1.0 和 1.0.0 规范的服务器 Capabilities 服务描述信息的数据库存储,并实现了基于服务描述信息中不同图层信息范围建立的 R 树索引结构的子模块。该模块位于内

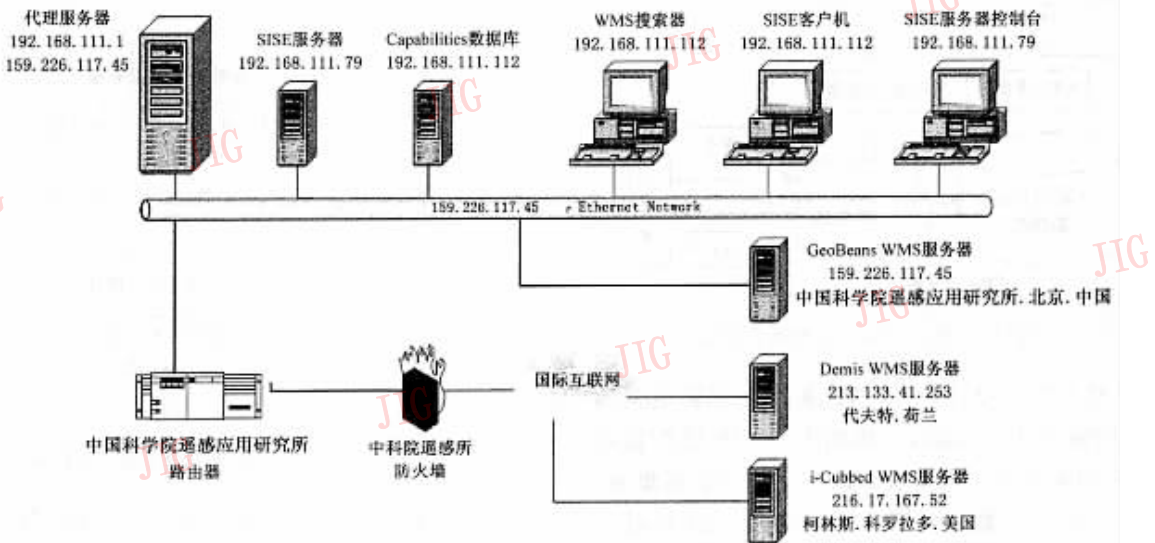


图 3 SISE 原型系统物理部署图

部网 192.168.111.70 机器上。

在基于 Java 的通用搜索器 Spider 程序基础上改进的 WMS 搜索器已经实现,该模块部署于内部网 192.168.111.30 机器上;基于 Java2 开发的 SISE 服务器程序可以向客户端提供搜索服务,该模块部署在内部网 192.168.111.60 机器上。

目前,该系统已连接了 GeoBeans、i-cubed LLC 和 Demis 3 个 WMS 服务器,对应的 IP 地址分别是



图 4 SISE 服务器控制台界面图(I)



图 5 SISE 控制台界面图(II)



图 6 SISE 控制台界面图(III)

图 4 是控制台初始界面图,其中,左侧显示了 SISE 服务器从 Capabilities 数据库中得到的各个 WMS 服务器及其可以提供的空间信息图层的具体情况;右上侧是控制区域,其提供了放大、缩小、全图、移动和查询功能;右侧中间区域显示了全球背景图,它提供给用户用于输入的查询区域;右下侧是与当前坐标位置对应的地理经纬度坐标。

作为演示,用户在右侧中间区域选择了中国区域后,控制台通过判断后就连接了 GeoBeans 和 Demis 两个 WMS 服务器,并最终得到了来自这两个 WMS 服务器的相应的查询结果(因为 i-cubed 服务器提供的数据不包括中国,所以对于本次空间信息查询请求,SISE 服务器通过判断后,不尝试动态连接该 WMS 服务器)。其中,来自 Demis WMS 服务器的结果如图 5 所示,来自 GeoBeans WMS 服务器的结果如图 6 所示。

基于图 5 和图 6,用户还可以进一步查询空间实体的信息,并能够进一步定制面向多个 WMS 服务器的组合查询。

## 5 空间信息搜索引擎系统原型性能测试

原型系统的测试环境为:SISE 服务器采用联想万全 2200C,配置为 2×P III 850, RAM512M, 2×18G SCSI 硬盘。测试机器配置为 P III 667, RAM256M, 80GIDE 硬盘。两者通过 10M 局域网连接。SISE 服务器经由科学院北郊光纤交换机通过科

159.226.117.45, 216.17.167.52 和 213.133.41.253。通过 SISE 客户机上的搜索引擎客户端程序,用户可以访问到由 SISE 服务器透明提供的来自上述 3 个 WMS 服务器动态提供的空间信息。

为了监控服务器运行状态,笔者开发了 SISE 服务器控制台程序,可以用于监控 SISE 服务器的运行状态。作为演示,图 4~图 6 展示了该控制台程序的工作状态。

学院网接入国际互联网,该服务器同时兼做内部网的代理服务器、数字地球 Email 服务器,同时也是数字地球中国网站的主 Web 服务器。

原型系统的性能测试采用并发用户数递增算法。测试的约束条件为以下 3 个:请求平均响应时间小于 1 s,上下文切换次数不大于 15 000 次/s,CPU 利用率最大值为 80%。性能测试结果数据表和曲线图分别如表 2 和图 7 所示。测试结果说明,部署在上述典型生产环境下的 SISE 统原型可以支持的最大的并发用户数目为 15 个。

表 2 SISE 原型系统性能测试结果

$N_{User}$ (个)	$N_{Pass}$ (个)	$N_{Fail}$ (个)	$t_{min}$ (s)	$t_{max}$ (s)	$\bar{t}$ (s)	$t_{mean}$ (s)
1	538	0	0.080	1.092	0.109	0.098
2	579	0	0.090	1.062	0.204	0.189
3	588	0	0.170	1.111	0.304	0.287
4	587	0	0.270	1.052	0.407	0.389
5	575	0	0.220	1.292	0.520	0.498
6	565	0	0.230	1.582	0.636	0.613
7	571	0	0.421	2.444	0.734	0.699
8	613	0	0.551	2.163	0.779	0.741
9	630	0	0.501	2.503	0.871	0.834
10	617	0	0.561	2.544	0.972	0.937
11	629	0	0.671	2.824	1.049	1.012
12	737	0	0.580	3.204	0.973	0.934
13	775	0	0.601	2.815	1.011	0.971
14	797	0	0.061	3.165	1.048	1.016
15	812	0	0.600	2.654	1.108	1.082

由图 7 和表 2 可见,在满足前述 3 项约束条件的前提下,性能测试在 15 min 内逐步增加并发的测试用户数目,其中表 2 的第 1 列~第 7 列分别表示了当前测试的并发用户(User)数  $N_{User}$ 、发送请求的总次数  $N_{Pass}$ 、失败的请求次数  $N_{Fail}$ 、单个请求响应的最短

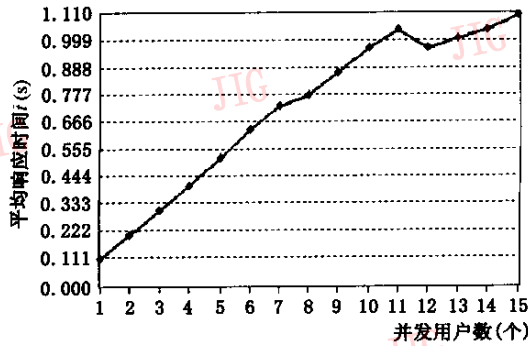


图7 SISE原型性能测试平均响应时间-并发用户数曲线

耗时  $t_{\min}$ 、单个请求响应的最长耗时  $t_{\max}$ 、请求响应的平均耗时  $\bar{t}$  和请求响应中用于实际传输的平均耗时  $t_{\text{mean}}$ 。该测试结果说明,SISE原型系统在单服务器配置的情况下,能够满足一般性业务需求。

## 6 讨论

从前述的 SISE 原型的工作原理,结合图 1 所示的系统结构,不难看出,SISE 系统对用户查询请求的响应过程从逻辑上可以分为 2 个步骤:SISE 确定候选 WMS 服务器,WMS 适配器动态连接获取空间信息。测试表明,第 2 个步骤的耗时(表 2 中  $t_{\text{mean}}$  列所示)是用户查询请求响应平均总耗时(表 2 中  $\bar{t}$  列所示)的主体部分。如果远程 WMS 服务器性能就越好,则它与 SISE 系统之间的网络连接性能越好,且该部分耗时也越短,即 SISE 系统整体响应时间就越短,曲线的走势就越平缓。

考虑到这两个性能因素都是 SISE 系统本身不可控制的部分,因此,在实际生产环境部署时,需要着重从这个角度考虑提高 SISE 系统的整体性能。

在 OpenGIS 的实现规范中,涉及到在线空间信息发布和共享的另一个规范是 OpenGIS Web Feature Service (OpenGIS WFS) 规范。OpenGIS WFS 规范的适用对象是以通过 HTTP 协议向客户端提供创建、查询、删除和更新地理特征数据服务的网络地理特征服务器(Web Feature Server)。为了实现上述功能,该规范详细规定了如下网络地理特征服务器操作接口标准的语义和语法定义: GetCapabilities(用于提供自身可以提供的地理特征的类型以及每一种类型上支持的操作)、DescribeFeatureType(用于描述用户指定的地理特征的结构)、GetFeature(用于提供用户指定的地理

特征的空间数据信息和属性数据信息)、Transaction(支持面向地理特征数据修改的事务管理)、LockFeature(用于支持事务过程中对特定地理特征数据的锁定)等。

由于 OpenGIS WFS 与 OpenGIS WMS 一样,都是基于 HTTP 协议构建的,两者的 GetCapabilities 操作接口标准有很大的相似性,因此 SISE 系统的站点爬行、网页遍历、路径提取、请求测试的技术路线也适用于 OpenGIS WFS 规范,且 SISE 系统可以平滑地过渡到支持这两种 OpenGIS 实现规范。

## 7 结论

本文论述了笔者在国家 863“十五”攻关课题中的阶段性成果——面向 OpenGIS WMS 的空间信息搜索引擎原型系统 GeoBeans SISE。由于它的创新意义在于基于最新的技术规范,展现了互联网环境下,在线空间信息的自动化搜索的可行性,因此该系统有很高的应用价值,其性能测试的结果表明,该原型系统可以满足典型生产环境下的一般性业务需求。

## 参考文献

- 1 杨崇俊. 在互联网上能找到地理空间数据吗?[A] 见:陈述彭. 数字地球百问[M]. 北京:科学出版社,1999:176~177.
- 2 Gudivada V N, Raghavan V V, Grosky W I, et al. Information retrieval on the world wide web[J]. IEEE Internet Computing, 1997, 1(5):58~68.
- 3 Lawrence S, Giles C L. Accessibility and distribution of information on the Web[J]. Nature, 1999, 400:107~109.
- 4 Lawrence S, Giles C L. Searching the world wide web[J]. Science, 1998, 280(5360):98~100.
- 5 Brin S, Lawrence P. The anatomy of a large-scale hypertextual web search engine[A]. In: 7th International World Wide Web Conference Proceedings [C], Brisbane, Australia, 1998: 107~117.
- 6 Larry P, Sergey B, Motwani R, et al. The PageRank citation ranking:Bringing order to the web[EB/OL]. <http://citeseer.nj.nec.com/page98pagerank.html>,1998.
- 7 Vretblad O. GIR-geographic information retrieval [EB/OL]. <http://www.ssc.se/gir>,1995.
- 8 FGDC. CSGSM-Content standard for digital geospatial metadata [EB/OL]. <http://www.fgdc.gov/metadata/contstan.html>, 1998.
- 9 ESRI. An overview of the geography network[EB/OL]. <http://www.esri.com/library/whitepapers/pdfs/gn-overview.pdf>, 2001

10 OpenGIS. Web map server interface specification [EB/OL].  
http://www.opengis.org/techno/specs/01-068r3.pdf, 2002.



白玉琪 1976 年生, 2000 年获中国地质大学(武汉)计算机应用专业硕士学位, 现为中国科学院遥感应用研究所博士研究生。主要研究方向为互联网地理信息系统、空间信息共享和语义网。



杨崇俊 1954 年生, 1990 年获法国遥感博士学位, 现为中国科学院遥感应用研究所研究员, 博士生导师。主要专业兴趣为遥感、互联网地理信息系统、计算机图像处理、空间数据库和物理等。



刘冬林 1971 年生, 1998 年获中国矿业大学应用地球物理专业硕士学位, 现为中国科学院遥感应用研究所副研究员。主要研究兴趣为互联网地理信息系统。



朱华吉 1975 年生, 2001 年获河海大学工程测量专业硕士学位, 现为中国科学院遥感应用研究所博士研究生。主要研究方向为空间数据模型。



卢亚辉 1976 年生, 2000 年获中国科学院遥感应用研究所地图学与地理信息系统专业硕士学位, 现为清华大学计算机系博士研究生。研究方向为软件工程和互联网地理信息系统。



芮小平 1975 年生, 2001 年获中国矿业大学硕士学位, 现为中国科学院遥感应用研究所博士研究生。主要研究方向为网络三维地理信息系统。发表论文 20 余篇。