

基于 Web 信息搜索的空间信息门户^{*}

李琦 黄丰 涂勇

(北京大学数字地球工作室 北京大学遥感与 GIS 研究所 北京 100871)

摘要 空间信息门户为用户提供一个统一的入口访问分散异构的空间信息资源,是国家空间信息基础设施的重要组成部分。传统的空间信息门户大多采用 OGC 在空间信息门户参考体系架构中推荐的“发布-检索-绑定”模式,本文提出一个新的基于 Web 信息搜索的体系架构,并对其中空间元数据自动提取等关键技术进行了研究,最后通过原型系统进行验证。

关键词 空间信息门户, Web 信息搜索, 信息提取, 元数据, 空间信息共享

Spatial Information Portal Based on Web Information Search

LI Qi HUANG Feng TU Yong

(Cybergis Studio, PKU Institute of Remote Sensing & GIS, PKU, Beijing 100871)

Abstract Spatial Information Portal allows user to access distributed spatial information directly and is one of the most important parts of spatial information infrastructure. A new portal architecture based on Web information search is promoted, which is different from traditional “publish-search-bind” mode recommended in OGC Spatial Information Reference Architecture. Finally, we develop a pilot system to verify the idea.

Keywords Spatial information portal, Web information search, Information extraction, Metadata, Spatial information sharing

1 引言

地理空间信息被广泛用于电子政务、科研、企业应用以及日常生活的方方面面。然而由于地理空间信息的特殊性,如何快速有效地发现、获取和利用地理空间信息,实现信息共享,一直是学术界研究的难点和热点问题。为此,开放地理信息系统联盟(OGC)制定了一系列的标准和指南,并于 2004 年提出空间信息门户参考体系架构^[1],旨在提供一个开放的基

础架构,指导空间信息门户的建设。

另一方面,随着 Web 技术的成熟和广泛应用,越来越多的信息以网页的方式来发布,地理空间信息也不例外。搜索引擎为自动搜集、索引和检索网页提供了一种有效的解决方案。

因此,在研究 OGC 空间信息门户参考体系结构的基础上,本文尝试提出一个新的基于 Web 信息检索技术的体系架构。本文首先对国内外空间信息共享方面的研究进行了回顾和总结,然后指出空间信息门户的概念及其重要性。接下来两部分是本文的重点,分别阐述基于 Web 信息搜索技术的空间信息门户体系架构以及地理空间元数据自动提取和元数据两个关键技术。最后,总结全文,并指出进一步研究的方向和内容。

2 国内外研究现状

地理空间信息是一种基础信息,空间信息门户则是国家空间信息基础设施的重要组成部分,一直是 GIS 领域关注的焦点。李琦^[2~4]研究了空间元数据以及基于元数据的地理信息共享模型和解决方案。各个国家和组织也纷纷推出空间信息门户网站,典型的有哥伦比亚大学建设的地球科学数据中

心,中国地球系统科学数据共享网等等,最具有影响力的则是美国联邦地理数据委员会推动的 Clearinghouse 项目。2002 年美国政府在 24 个电子政务项目计划中正式提出 Geo-One-Stop(GOS)计划。该计划正式提出建设地理空间信息门户,为联邦政府和公众提供一个统一的空间信息资源访问入口,并能支持事务处理和决策支持,是 Clearinghouse 项目的延伸。同时,OGC 总结和制定了一系列的规范,包括目录服务规范^[7]、空间信息门户参考体系架构(Geospatial Portal Reference Architecture, GPRA)^[1]等,对空间信息门户涉及的操作和接口进行了标准化,为空间信息门户建设提供指导。

尽管上述空间信息门户建设的解决方案不尽相同,但是它们都基于“发布-检索-绑定”的模式。本文试图在 GPRA 基础上,结合 Web 信息检索的方法,给出一个新的空间信息门户建设方案。二者最大的不同在于,前者通过注册和发布机制来获得地理空间数据的元数据信息,而后者利用搜索引擎技术,主动去发现 Web 上的各种空间信息。

3 OGC 空间信息门户参考体系架构

2004 年 7 月,OGC 发布了空间信息门户参考体系架构技术规范讨论版^[1],旨在为空间信息门户建设提供一个可操作的、更容易实施的框架,如图 1 所示。这是一个基于 SOA 的体系架构,空间信息门户中涉及的操作被抽象为四种类型的服务,即门户服务(Portal Services)、目录服务(Catalog Services)、数据表现服务(Portrayal Services)和数据服务(Data Services)。

门户服务是指用来接受来自个人、企业和政府的请求的服务,提供单点登录信息访问,同时还包括门户系统的管理与

^{*} 本文得到国家 973 项目“网络环境下空间信息智能服务及应用示范”(课题编号:2006CB701306)的资助。李琦 教授,博士生导师,研究领域:空间信息科学与技术、数字地球、数字城市等;黄丰 硕士研究生,研究领域:空间信息搜索、智能 GIS、数字城市;涂勇 硕士研究生,研究领域:数字城市、电子政务等。

监控服务;目录服务则提供一个通用的机制,利用元数据对数据分类、描述、注册、发布、维护和查询的服务;数据表现服务可以理解为提供空间信息可视化功能的服务,例如 WMS 服务;数据服务提供对空间数据内容的直接访问服务,例如 WFS 和 WCS 服务。其中,门户服务和目录服务由空间信息门户本身所提供的,而数据表现服务和数据服务则是由各空间信息资源提供者来提供。

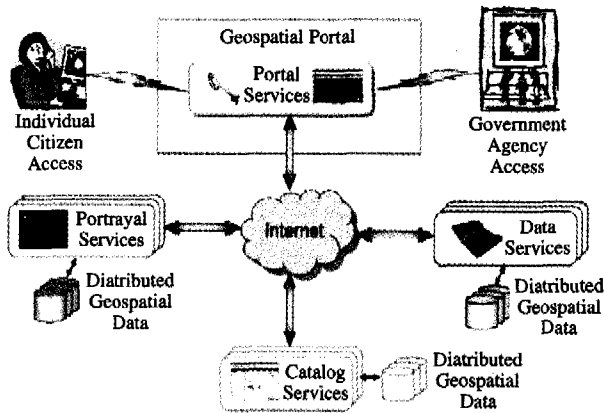


图1 OGC空间信息门户参考体系架构^[1]

这一体系架构可以进一步抽象为图2所示的“发布-检索-绑定”模式。分散异构的地理空间数据和服务的元数据被注册并发布到一个中心节点,形成元数据目录。用户直接通过中心节点,就可以检索元数据并了解到如何进一步获取该数据。然后,用户再向具体的资源提供者请求数据服务或者数据表现服务,或者下载和购买,这称为资源绑定。

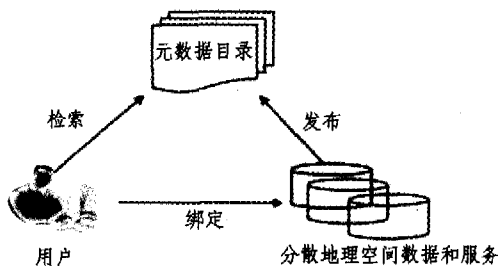


图2 发布-检索-绑定模式

GPRA 还通过 OGC 目录服务规范^[7]对元数据注册、发布、更新和检索过程中涉及的数据、操作和接口进行了规定,通过一系列应用指南^[8]说明了如何在门户系统中使用 ISO19115 或者 CSDGM 元数据标准,等等。

总之,OGC 制定的 GPRA 及其相关的技术规范,给出了一个值得参考的基于 SOA 的空间信息门户参考体系架构。但是,同时它也存在不足之处。首先,GPRA 基于一种信息注册和发布的机制来获得地理空间信息,而不是主动搜索信息。其次,GPRA 没有考虑现有系统的集成问题。以中国的情况为例,在空间信息门户推出之前,许多部门已经采用行业元数据标准,国家也先后投入了巨资建立许多地理空间信息的共享网站。如果采用 GPRA 的模式来建设 NSII 空间信息门户,就要按照门户规定的元数据标准、接口和协议把空间信息注册到中心节点的目录中,这对分节点造成很大的工作量,也将直接阻碍空间信息门户的建设。

4 基于 Web 信息搜索技术的空间信息门户

随着 Web 技术的成熟和广泛应用,越来越多的信息以

Web 方式来发布,众多学者对如何自动搜集、索引、提取和检索 Web 信息也进行了非常多的研究,并取得了卓有成效的成果^[9]。Web 信息搜索的一般模式可以用图3来表示。

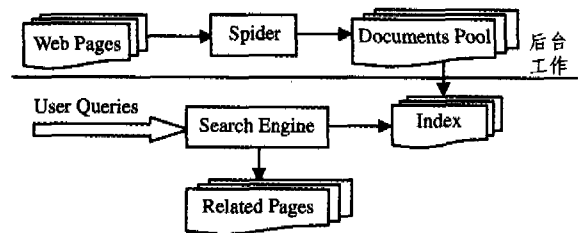


图3 Web信息搜索的一般模式

在搜索引擎的模式下,其原理可以简述如下。系统后台有一个被称为 Spider 的网页抓取程序,它利用网页的链接自动搜集网页信息。同时后台运行着一个文本索引的程序,对文本进行自然语言分词,然后建立高效的索引,例如倒排索引,以支持高效的查询。当用户请求来临时,系统采用某种信息检索模型,例如经典的向量检索模型,得到相关文档,最后使用预定义的排序算法对文档排序,返回结果给用户。

这是一种完全不同于“发布-检索-绑定”的信息搜索模式。其最大的不同是,搜索引擎主动搜集分散的 Web 信息,而不是通过一种注册的机制来被动接受信息。为此,本文试图把 Web 信息搜索技术引入空间信息门户,对 GPRA 进行改进,以解决上一节提出 GPRA 的两个不足之处。这一新的体系架构,我们称其为基于 Web 信息搜索技术的空间信息门户体系架构(WIR-GPRA),如图4所示。

在 WIR-GPRA 中,中心节点目录元数据信息通过两种方式来得到。一种是与 GPRA 相同,采用信息注册的方式。个人或者一些小的部门,可以把拥有的空间信息注册到中心节点。另一种,则是 Web 搜索和提取的方式。在这种方式下,中心节点主动到现有的空间信息网站上去抓取网页。在这里,我们把这些空间信息网站称为“源站点”。然后通过信息提取模块,提取网页中的空间信息元数据,加入到中心节点的元数据目录库,然后就可以利用目录服务,为用户提供信息检索服务。最后,还需要一个更新模块,当源站点的内容更新时,对中心节点的元数据目录进行增量更新。

上述过程与普通的 Web 信息搜索存在三个很大的不同。

(1)首先,crawler 或者说网页抓取器,集中抓取发布空间信息的网站下的网页,而不是抓取所有的网页。

(2)其次,在空间信息门户中,用户的查询请求往往具有明显的语义,例如“查询地理覆盖范围在((112,56), (113,57))的空间信息”,而在搜索引擎中,所有的语义都被理解为“词”的相关性。

(3)这直接带来了信息提取的问题。在空间信门户中,为了支持用户带有语义的查询请求,需要提取网页中的元数据记录。而一般情况下,搜索引擎并不做信息提取的工作,返回匹配查询请求的整个 Web 页面,而不是具体的结果。

上述三个方面构成了 WIR-GPRA 体系架构技术上的两个难点。第三个难点则是消除信息的重复性,即如果在两个源站点发布了同一个空间数据的元数据,那么在中心节点如何消除信息的重复性,保证同样一条记录不会出现多次。本文对第二个难点,即 Web 地理空间元数据的自动提取进行了研究,在下一节进一步讨论。

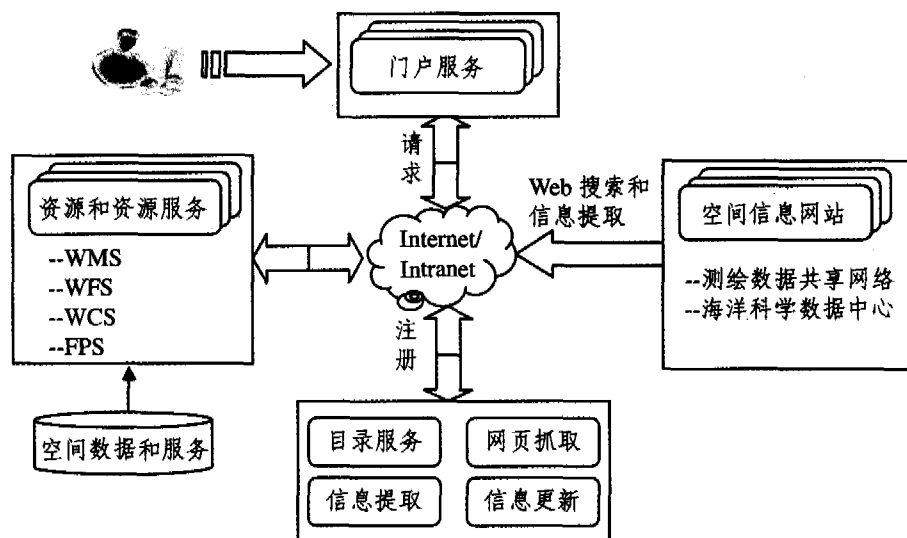


图 4 基于 Web 信息搜索的空间信息门户体系架构

5 Web 地理空间元数据自动提取

所谓 Web 地理空间元数据的自动提取是指,通过自动识别网页的结构,把地理空间元数据自动提取出来。这里的网页并不是一般的网页,而是指一个发布地理空间信息的网页,如图 5 所示。目标就是要将图 5 中的多条查询结果区分开来,提取元数据项。这通常称为半结构化信息提取。

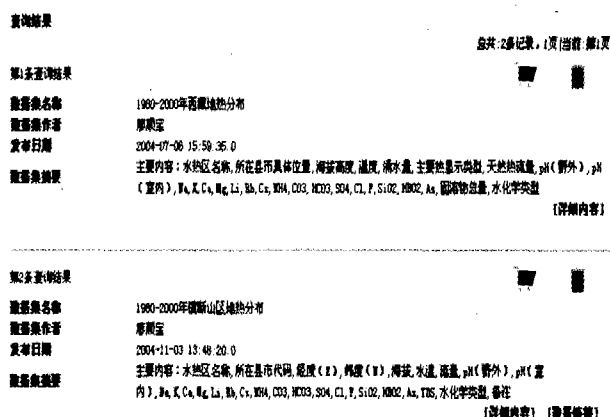


图 5 发布地理空间信息的网页的样例

半结构化 Web 信息提取方法可以归纳为下面三种类型:

- (1) 手动构造 Wrapper: 通常提供一种脚本语言,由专家针对具体网页和具体应用来编写复杂的提取规则。
- (2) 自动生成 Wrapper: 基于机器学习理论来自动生成规则,例如 Kushmerick^[10]。给出若干个经过人工注释的 Web 页面作为训练样例,通过机器学习方法,系统自动或者半自动生成提取器。其优点是能够得到准确的提取规则,缺点是对一个新的 Web 站点,必须重新建立一个训练集。
- (3) 基于模式识别的自动发现方法 (IEPAD): Chang^[11]。这是一种完全不同的信息提取方法,本文采用该方法来进行 Web 地理空间元数据的自动提取,取得了很好的效果。下面对此做一个介绍。

Chang^[11]认为,半结构化网页的数据通常是有固定的格式,并且以规则和连续的布局在网页中展现。这种布局或者格式可以认为是一种模式。通过分析 Web 页面中的 HTML 标签的重复规律,进行模式挖掘并自动生成数据抽取规则,然

后利用模式匹配的方法提取数据。该方法的精确性和准确性与 Wrapper 相当,但是避免了构造训练样本,同时还支持模糊模式匹配。

利用模式识别的思路和方法,Web 地理空间元数据提取由五个过程组成:去噪和规则化→编码→模式识别→模式过滤→信息提取。这一过程可以用图 6 来说明。

首先对网页内容进行规则化处理,即去掉其内容标签,例如,而保留结构标签,例如<TR><TD><TABLE>,得到如图 6(a)所示的网页。然后对网页标签编码,例如采用图 6(c)所示的编码方案,则可以得到图 6(d)所示的结果。二进制表示的字符串更容易进行处理。然后得到该二进制字符串的后缀字符串序列,建立后缀字符串序列的压缩 PAT 树结构,如图 6(e)所示。基于 PAT 树来识别网页中存在的可能模式,是整个方法的关键。在 PAT 树中,内部节点保存着当前比较的字符串的位置,叶节点保存着后缀字符串的标号。从根节点到每一个内部节点的路径就代表着内部节点所包含的子树的一串重复出现的字符,因此是一个可能的模式。最后通过检查模式的规则性和连续性,进行模式过滤。

6 原型系统

在 WIR-GPRA 体系架构的指导下,作者实现了一个基于 Web 信息搜索和提取的空间信息门户的原型系统。该系统的体系架构如图 7 所示。其中 Web 地理空间信息元数据的自动提取也得到了满意的结果。本文对哥伦比亚大学地球科学信息中心、中国地球系统科学数据共享网等多个网站进行了测试,都可以迅速准确地找到正确的模式,并进行元数据提取。图 8 是一个信息提取的结果。

结论与展望 空间信息门户为用户提供一个统一的入口访问分散异构的空间信息资源,是国家空间信息基础设施的重要组成部分。本文在 OGC 空间信息门户参考体系架构的基础上,提出一种新的基于 Web 信息搜索技术的空间信息门户体系架构 WIR-GPRA,并对其中的关键问题之一,即 Web 地理空间元数据提取进行了研究,最后通过原型系统验证了该体系架构的可行性。尽管该体系架构还存在一些没有解决的问题,例如,Web 地理空间元数据提取的鲁棒性还需要进一步提高,但是,同 GPRA 相比,个人认为,它更符合技术发展的趋势,更具有可扩展性和兼容性,也更有发展的潜力。

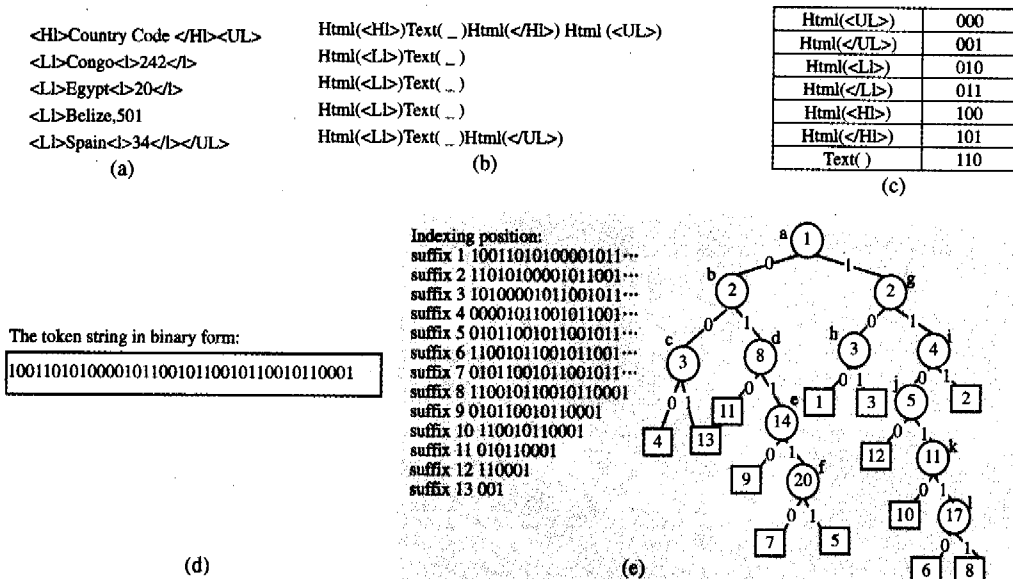


图 6 基于模式识别的 Web 信息提取^[11]

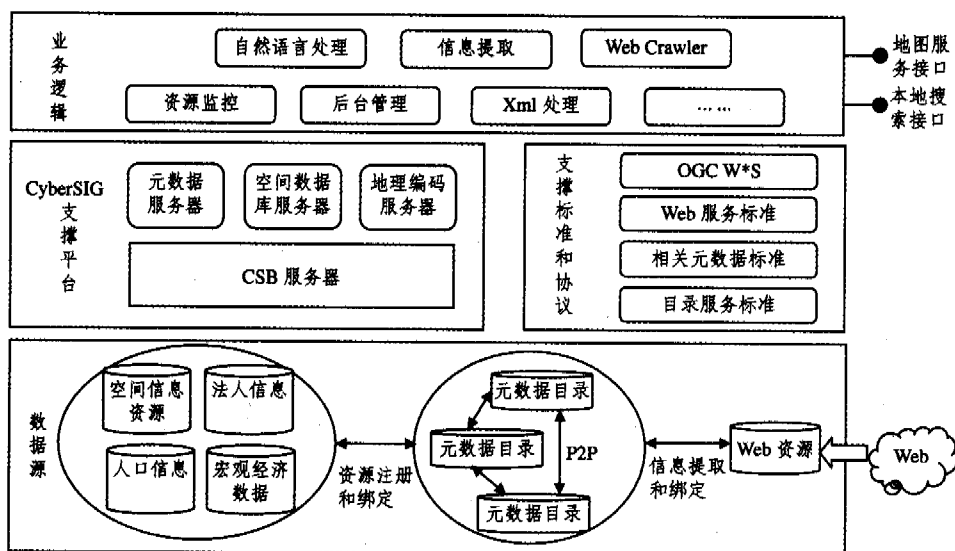


图 7 空间信息门户原型系统体系架构图

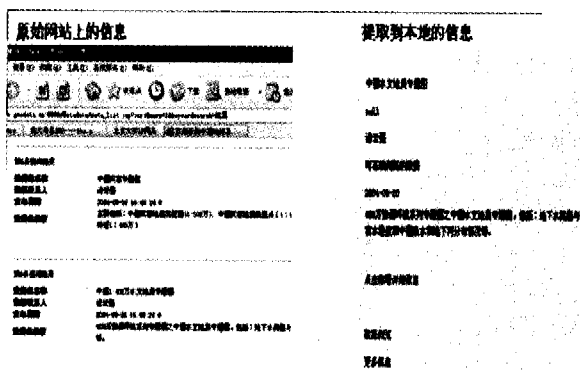


图 8

参考文献

- OGC. Geospatial Portal Reference Architecture, a Community Guide to Implementing Standards-Based Geospatial Portals, 2004
- 陈爱军, 李琦, 等. 数字地球建设中的地理空间 Metadata 研究. 遥

感学报, 2002, 6(4)

- 易善桢, 李琦, 等. 互操作 GIS 模型及其在空间信息基础设施体系结构中的实现途径. 遥感学报, 1999, 4A(11)
- 李琦, 陈爱军, 等. 地理空间信息共享理论基础及其解决方案. 电子科技导报, 1999, 5
- Nativi S. Metadata in the Framework of Geospatial Information Interoperability. Unidata/UCAR, 2002, 1
- 张颖. 基于 XML 的地理信息元数据系统: [学位论文]. 北京大学: 北京大学图书馆(中心馆), 2001
- OGC. OpenGIS Catalogue Service Specification, V2.0, 2004. 05. 11
- OGC. OpenGIS Catalogue Service Specification-ISO19115/ISO19119 Application Profile for CSW 2.0, V0.9.2, 2004. 07. 12
- Baeza-Yates R, 等著. 现代信息检索(英文版). 北京: 机械工业出版社, 2004. 2
- Kushmerick N, et al. Wrapper Induction for Information Extraction. Intl. Joint Conference on Artificial Intelligence, 1997(1): 729~737
- Chang C-H, et al. Automatic information extraction from semi-structured Web pages by pattern discovery. Decision Support Systems, 2003, 35(1)