计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(6): 1357－1362, 2012

图形工作流驱动的空间信息服务链研究

张建博[1,2]　刘纪平[2]　王　蓓[3]
[1](武汉大学资源环境与科学学院　武汉　430079)
[2](中国测绘科学研究院　北京　100830)
[3](中国科学院地理科学与资源研究所　北京　100101)
(finecho@163.com)

Spatial Information Services Chaining Based on Graphic-Workflow

Zhang Jianbo[1,2], Liu Jiping[2], and Wang Bei[3]
[1](*School of Resource and Environmental Science, Wuhan University, Wuhan* 430079)
[2](*Chinese Academy of Surveying and Mapping, Beijing* 100830)
[3](*Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing* 100101)

Abstract　Traditional workflow technology can't meet the requierments of spatial information services chaining aggregation because of the particularity of the long time cost of spatial data processing and the interface that workflow calls spatial information services chaining. The main problems lie in the mismatch of interface, and low efficiency. Thence, the model of spatial information service chaining aggregation is established, expending the model of traditional WPS chaining whose insufficient is considered. By analyzing the main problems between the interface and spatial data processing about spatial information services during the workflow scheduling, we present corresponding methods of the interface transformation to make the workflow call spatial information services chaining well, and adopt a strategy of GML-compressed scheduling to guarantee the spatial data flow be lowly transmitted on the network. Finally, the concrete solutions of spatial information services chaining processing are presented in the paper combined Kepler workflow engine. The experimental results show that this scheduling method not only breakthroughs interface bottlenecks through the interface transformation used in the workflow before, but also improves spatial information service chain execution efficiency through the strategy transformation of spatial data flow.

Key words　spatial information services; services chaining; combining; GML data flow strategy; graphic-workflow; Kepler

摘　要　空间信息服务链由于其空间数据操作对时间的要求以及其应用于工作流对接口要求的特殊性，使得传统的工作流很难调度空间信息服务链，主要的问题在于接口不匹配以及运行效率低下. 在传统WPS服务链的基础上，针对其不足提出了基于图形工作流的空间信息服务链聚合模型. 通过分析工作流调度空间信息服务链存在的主要问题，提出相应的接口改进方法，以及基于地址引用、压缩GML的数据调度策略. 并结合Kepler工作流引擎，给出了空间信息服务链聚合的具体过程. 经过实验证明：基于图形工作流环境，不但突破了以往空间信息服务应用于工作流的接口瓶颈，而且改进的数据流调度策略有助于空间信息服务链执行效率的提高.

关键词　空间信息服务；服务链；聚合；GML调度；图形工作流；Kepler

中图法分类号　TP311

收稿日期：2011-03-17；**修回日期**：2012-02-06
基金项目：国家自然科学基金项目(40901195)

基于工作流技术聚合标准的空间信息服务为网络环境下分布式空间信息的共享和互操作提供了有效的解决方案[1]. 但是 OWS(OGC Web service)服务链应用于工作流主要存在两个方面的问题:

1) 接口层面. 面向服务的工作流引擎采用 OASIS[2]标准,而空间信息服务则采用 OGC(open geospatial consortium)标准,导致空间信息服务在工作流中不能被聚合.

2) 数据层面. OGC 标准化的 WFS(Web feature service)和 WPS(Web processing service)服务以 GML 为共享交换格式,密集的 GML 操作使得工作流执行空间信息服务链效率较低.

目前,对基于工作流的空间信息服务链聚合研究主要集中在 2 个方面:从抽象空间信息服务描述中自动派生 BPEL(工作流描述语言),送入工作流引擎执行[3-6],以及采用语义、本体的方法构建服务链从而驱动工作流[7-9]. 这些方法主要是从概念层次上讨论工作流,并没有针对空间信息服务在工作流中如何建模作深入的探讨. 因此,本文主要研究如何改进 OWS 服务接口,以及服务链执行策略,使 OWS 服务链无缝地应用于图形工作流.

1 OWS 服务链与图形工作流

1.1 空间信息服务选择

OGC 标准化了多种的空间信息服务:地图服务(WMS,WCS)、要素服务(WFS)、处理服务(WPS)等. 本文以 WFS 要素服务为数据源、WPS 处理服务为节点,构建 OWS 服务链.

1.2 图形工作流

工作流引擎按照过程建模的不同可分为 BPEL 过程建模工作流和图形建模工作流. 基于 BPEL 过程建模的工作流原理是采用 BPEL 语言定义服务链,然后交给工作流引擎执行,由于缺乏可视化工具支持,造成 BPEL 谱写困难.

图形建模工作流是通过手动拖拽图形代理的方式定制服务链. 相比 BPEL 过程建模工作流具有以下优点:1)提供了过程建模的可视化工具,用户不需要通过手工谱写过程建模语言;2)客户端应用程序只需要实现 WMS、WFS 要素服务的可视化,执行过程交由图形工作流执行,实现应用程序客户端、图形工作流客户端以及工作流引擎最大程度上的松耦合架构;3)面向服务的图元代理能够动态发现、聚合经过接口改造的 OWS 服务,易于实现 GML 数据调度策略.

比较典型的是美国加利福尼亚大学的 Kepler[10]开源工作流. 因此,本文选择 Kepler 聚合 OWS 服务链.

2 图形工作流驱动 OWS 服务链模型

2.1 传统 WPS 服务链模式

WPS[11]标准定义了简单的服务链. 要求服务链中参与聚合的 WPS 服务实例以 KVP 编码的方式提交 GET 请求. 可以描述为

$$GetURL = req\ (req_1(req_2(req_n(\cdots)))), \quad (1)$$

其中,req 为当前 WPS 服务请求,req_i 为参与服务链的 OWS 服务请求,包括 WFS 的 GetFeature 以及其他 WPS 的 Execute. 从式(1)中可看出,WPS 标准定义的服务链以嵌套 URL 的方式聚合服务请求,由于 URL 编码的字符限制,这种服务链仅仅适合少数 OWS 服务的聚合,并且嵌套的服务聚合给服务监控以及异常捕捉带来很大的困难.

2.2 图形工作流驱动 OWS 服务链模型描述

图形工作流引擎聚合 OWS 服务,不再采用 KVP 编码接口. 如图 1 所示. 假设存在 WFS_1, WFS_2, WPS_1, WPS_2 四个 OWS 服务. 首先对其添加 WSDL 描述,对消息响应采用 SOAP 绑定;接着改进 WFS 接口,使之能够输出 GML 地址引用,同时为 WPS 接口添加接口解析器,便于 WPS 非 GML 数据类型的参数传递(ZIP);然后采用工作流引擎的可视化界面构建 OWS 服务链流程,由图形工作流引擎完成执行、监测、输出工作.

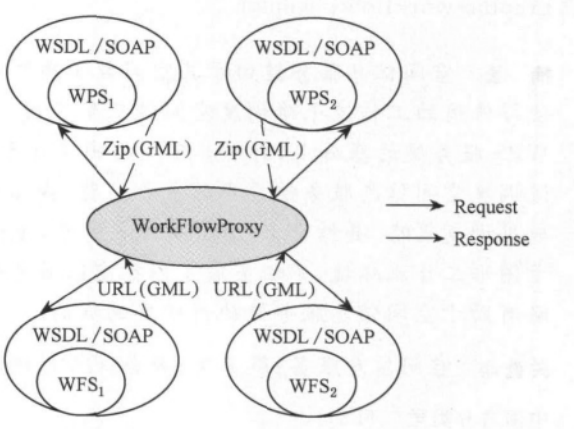

Fig. 1 The description of OWS chaining model with workflow.

图 1 工作流驱动服务链概念模型描述

从图 1 可以看出,图形工作流引擎作为调度中心控制每个 OWS 服务,其服务聚合遵循式(2):

$$V = WorkFlow(OWS_1, OWS_2, \cdots, OWS_n). \quad (2)$$

不再以 WPS 为单独的服务链入口,而是将嵌套的服务拆分为多个 WFS,WPS 服务. 工作流中的 OWS 服务通过多点服务增值的方式聚合,而不是式(1)描述单点 WPS 服务聚合;语义层面的 URL 聚合变为 GML 格式的数据体聚合.

3 图形工作流驱动 OWS 服务链实现

本文根据 Kepler 工作流以及标准空间信息服务的特点,通过分析 OWS 服务链面向图形工作流的接口以及数据调度问题,提出了图形工作流引擎聚合 OWS 服务链的执行策略.

3.1 OWS 服务接口封装

OWS 服务与图形工作流在内部接口和外部接口层面都存在较大的差异.

1) 外部接口. ①Kepler 工作流面向标准 Web 服务. 节点映射的服务以标准的 WSDL 描述、节点之间以 SOAP 为通信协议进行互操作. 但是 OGC 并没有给出相关规范;②虽然 OGC 在 OWS-5 中提出了 WPS 要支持 WSDL,SOAP 协议的需求,但是没有明确提出支持 WFS 的 WSDL 封装,因此图形工作流不能协调 WFS,WPS 的服务聚合.

针对这种外部接口层面的不一致问题,需要针对 WFS,WPS 进行接口封装. 以 WPS 为例,包含 3 种接口 GetCapabilities, Execute, DescribeProcess. 仅需要将 GetCapabilities 请求通过 WSDL 封装,同时改造 Execute 请求,使之支持 SOAP 协议,如图 2 所示. Kepler 引擎驱动 WPS 服务能够从封装后的 GetCapabilities 中获取 Execute 的 SOAP 请求.

2) 内部接口. ①虽然为 WPS 添加了 WSDL 描述,但是 WPS 的输入参数类型并不在 GetCapabilities 接口中描述,而是在 DescribeProcess(WFS 为 describeFeatureType)接口中定义,Kepler 工作流不能够从 GetCapabilities 接口中提取 SOAP 请求所必备的参数类型;②OGC 标准没有针对 WPS 的输入输出参数做强制规范,可以是单独参数也可以是复杂类型 complexType 的多参数输入输出,WPS 入口无法动态识别复杂类型参数.

例如图 1 描述的 OWS 服务,WPS_1 的输出接口 $Response_1$ 可能为 GML 实体数据类型,也可能是包含在 GML 中的 URL 地址引用类型. 在 Kepler 驱动的这个 OWS 服务链中,WPS_2 输入接口无法确认是 GML 实体数据,还是 URL 地址引用,因为两者的参数类型都是 AnyURL 这一弱类型语言. 这一问题的实质是图形 Kepler 工作流无法自动解析非实体化 GML 数据.

```
〈wsdl:definitions name="SpatialAnalysis"〉
〈wsdl:types〉
〈xsd:schema elementFormDefault="qualified"〉
〈xsd:element name="ExecuteProcess_GMLBufferResponse"〉
〈xsd:complexType〉
〈xsd:sequence〉
〈xsd:element maxOccurs="1" minOccurs="0" name=
    "ExecuteProcess_GMLBufferResult"〉
〈/xsd:sequence〉
〈/xsd:complexType〉
〈/xsd:element〉
〈/xsd:schema〉
〈/wsdl:types〉
〈wsdl:message name="ExecuteProcess_GMLBufferSoapRequest"〉
〈wsdl:part name="parameters" element="tns:ExecuteProcess_
    GMLBuffer/"〉
〈/wsdl:message〉
```

Fig. 2 WPS with WSDL.

图 2 WPS WSDL 描述

解决这个问题需要在 WFS,WPS 服务的输出端添加 GML 解析器. 如图 3 所示:假设 WPS_1 输出 GML,非 LineString 类型,而是包含遥感图片的 URL 地址,那么 GML 解析器就为下一节点 WPS_2 解析出正确的 URL 地址及其类型,而不是 GML 数据体. 需要指出的是,OWS 服务链构建者不需要另行编码设计,Kepler 提供了 XML 解析组件,用户需要利用这种组件提取所需的参数以及类型.

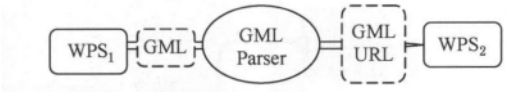

Fig. 3 WPS output interface define.

图 3 WPS 输出接口解析器

3.2 图形工作流调度 OWS 服务链策略

标准 OWS 服务链以 GML 为共享数据交换格式,密集的 GML 输入输出给工作流驱动 OWS 服务链带来效率上的问题,需要额外的网络传输时间开销,而且网络传输大数据量的 GML 速度慢.

图 4 表达了数据调度的两种策略. 细实线为 Kepler 代理的服务请求,虚线和粗实线表示数据传输的 2 种方式,数据请求、传输由各个图元节点控制.

传统的数据调度方法如虚线所示,所有的 OWS 服务都返回 GML 数据,是一种高密度 GML 调度模

式. 模拟数据传输过程可用以下方法:

$$GML_1 = KeplerProxy(WFS_1: GetFeature); \quad (3)$$

$$GML_2 = KeplerProxy(WPS_1: Execute(GML_1)); \quad (4)$$

$$GML_3 = KeplerProxy(WPS_2: Execute(GML_2)). \quad (5)$$

由于从一个 OWS 服务到下一 OWS 服务包含服务响应和数据传输两个环节,其总体时间消耗为

$$T = 2O(N_1) + 2O(N_2) + O(N_3), \quad (6)$$

其中,N_1, N_2, N_3 分别为 GML_1, GML_2, GML_3 的数据量.

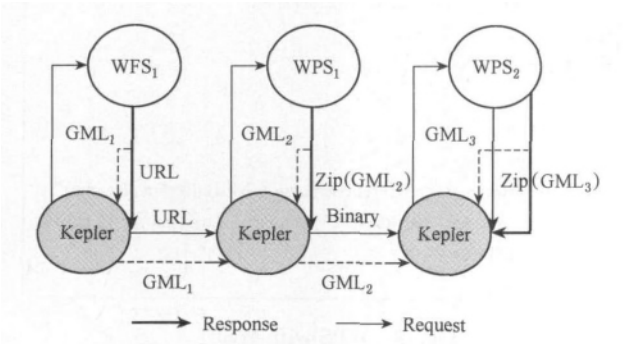

Fig. 4 Data flow of OWS services chaining.

图 4 图形工作流模式下服务链数据流图

为了减少 GML 在网络的传输次数,本文改进了 OWS 服务链的调度策略,如图 4 粗实线所示:①将 WFS_1 的 GML 实体输出类型改造为包含 URL 的 complexType 类型;②将 WPS 的 GML 实体输出类系改为压缩包的参数类型,并通过二进制流的方式在网络上传输;③为 WFS, WPS 添加接口解析器(3.1 节),保证改进后的参数类型能够被图形工作流正确解析. 其过程可用下面公式表达.

$$URL = KeplerProxy(WFS_1: GetFeature); \quad (7)$$

$$Zip(GML_2) = KeplerProxy(WPS_1: Execute(URL)); \quad (8)$$

$$Zip(GML_3) = KeplerProxy(WPS_2: Execute(ZIP(GML_2))). \quad (9)$$

在数据流调度中,由于为 Kepler 添加接口解析器(3.1 节),WPS_1 可以接受 WFS_1 传递来的 URL. 工作流引擎就不必先将 GML 下载到本地,然后传输给 WPS_1,而是将响应结果作为 URL 直接交给 WPS_1 处理,避免了 GML 到代理客户端的响应传输(URL 的传输代价忽略不计);同时,WPS_1 可以以 Zip(GML)的格式输出 GML,最大限度地压缩 GML. 整个服务链的传输时间消耗为

$$T = O(N_1) + 2O(K \times N_2) + O(K \times N_3), \quad (10)$$

其中,K 为 GML 的压缩率. 由于 $K \times N < N (0 < K < 1)$,相比式(6)其时间消耗明显减少.

3.3 Kepler 调度 OWS 服务链具体过程

在提出图形工作流中 OWS 服务链的接口改进方法以及调度策略的前提下,本文提出面向 Kepler 工作流的 OWS 服务链聚合过程.

图 5 描述了采用 Kepler 引擎聚合 OWS 服务链的控制流模式图. 在 Kepler 中,面向 Web 服务的图元节点称为 Actors,实质上是服务代理客户端,通过简单的插件机制执行任何 WSDL 定义的 Web 服务,并通过其端口链接到其他 Actors.

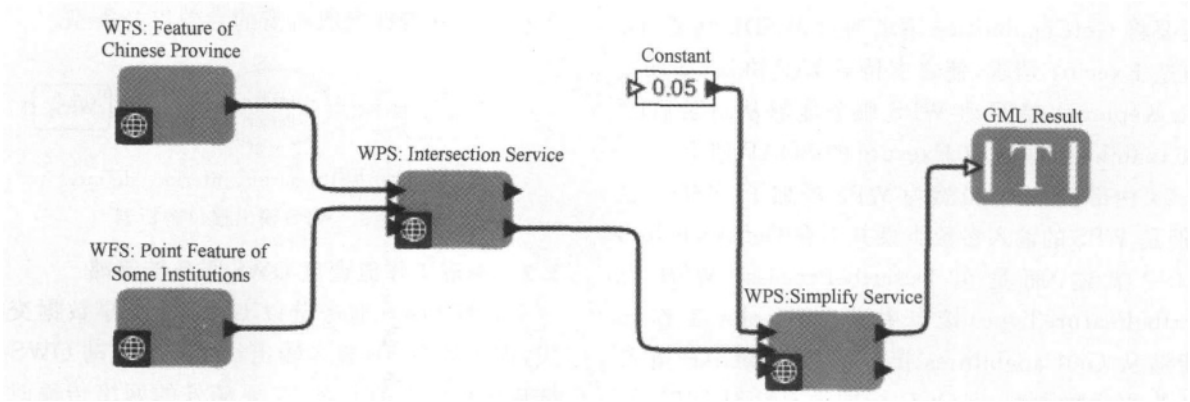

Fig. 5 Kepler workflow model.

图 5 Kepler 过程建模

利用 Kepler 的图形界面定制工作流,我们首先初始化所需要的服务代理 Actors,按照 WFS, WPS 服务的参数要求分别配置代理端口,然后按照 OWS 服务链的业务执行逻辑连接 Actors,组成工作流. 其数据流向按照图 4 中的描述进行映射.

1) 接口改造. ①首先在服务端按照 3.1 节的方法对上述 OGC 服务进行接口改造,使之满足图形工作流对 OWS 服务链采用 WSDL 描述的要求. 此时的

WFS 以及 WPS 的 GetCapabilities 接口作为 Kepler 节点映射服务的 URL 地址,如:http://ServerUrl?wsdl.②改造 WFS 的输出接口,输出的 GML 文件不再表达矢量对象,而是包含 AnyURL 类型的 URL 地址引用.实质上是将 WFS 逻辑对象转换成地址引用,减少 GML 操作.③为 WPS 接口添加接口解析器,解析 WFS 的输出文件中包含的 GetFeature 请求返回的 URL 引用地址.然后交给下一节点的 WPS 执行.

2) 服务发现. Kepler 引擎提供 Web 服务搜索工具——图元代理 Actor,可通过拖拽的方式将参与 OWS 服务的 Actor 代理工具置于工作流,actor 代理按照 WSDL 描述自动发送 SOAP 协议,并捕获输出以及异常.

3) 服务聚合.①配置图元代理 $Actor_1$,$Actor_2$,以发现 WFS_1,WFS_2 要素服务,发送 GetFeature 请求(SOAP 协议),返回 GML_1,GML_2,但包含的是 GML 数据的 URL 地址引用;②将 GML 通过解析器解析出要素地址引用;③图元代理 $Actor_3$ 接收 URL 引用,并发送 Execute 请求,内部调用 WFS 的 GetFeature 请求,获取参与求交运算的 GML 数据;然后通过矢量求交运算处理,获取求交后的多边形混合数据 GML_3;最后交由面状要素化简处理服务化简数据,得到 GML_4,交由客户端验证显示.

4 实验与分析

实验展示了 Kelper 工作流聚合 OWS 服务链的效果,并对聚合策略进行了性能测试,将采用地址引用、压缩 GML 的策略和密集 GML 操作的策略作了比较分析.

4.1 实验数据与环境

首先提出参与实验的 OWS 服务链实例:获取全国某机构办事处的点状要素,与所在省份的多边形要素求交,经过 Douglas-Pcucker 化简算法多边形,返回客户端显示.涉及的空间信息服务均为 OGC 标准的空间信息服务,具体如下:

1) 中国省级行政区划要素服务(WFS_1);

2) 某机构办事处的点状要素服务(WFS_2);

3) 要素求交服务(WPS_1);

4) 要素化简服务(WPS_2).

在实验中,参与聚合的要素服务(WFS)采用 Geoserver(开源地图服务引擎)发布,处理服务(WPS)由开源软件 52°North 的空间信息处理组件发布.聚合生成的 GML 数据由客户端 Openlayers 展示.采用 Kepler 图形工作流引擎聚合 OWS 服务.

4.2 结果与分析

实际的工作流执行过程如图 5 所示.图 6 为聚合的最终结果.图 6(a)(b)(c)分别为 Openlayers 客户端展示的原子级别的服务,图 6(d)为 Kepler 工作流调用 WPS(douglas-pcucker 化简算法)化简多边形,输出 GML 的最终展示结果.

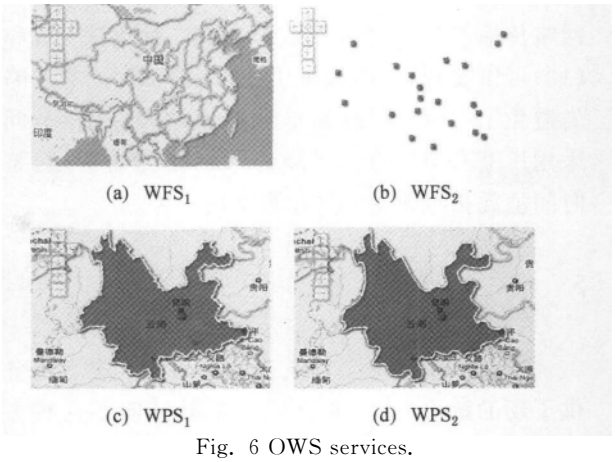

Fig. 6 OWS services.

图 6 OWS 服务

为了验证地址引用、压缩 GML 策略和密集 GML 策略针对不同的 GML 数据量的执行效果,提供 3 种比例尺的省级行政区划数据:1:4000000,1:1000000和 1:250000 矢量数据(主要是对行政区划进行化简).由 Geoserver 发布后的 GML 数据量分别为 1.47 MB,11.2 MB 和 23.5 MB.表 1、表 2 分别为不同策略在不同数据量环境下的时间费用.实验结果表明,当 GML 数据量较小时,优化的调度策略的时间优势不明显;但是当服务链处理的 GML 数据量越大时,对聚合调度的优化效果越明显.

Table 1 Time Cost of Strategies with Larger GML

表 1 密集 GML 调度策略时间消耗 s

GML/MB	T_{WFS}	T_{WPS1}	T_{WPS2}
1.47	1.2	2	3.5
11.2	17	19	11
23.5	52	61	29

Table 2 Time Cost of Optimizing Strategies

表 2 优化的调度策略时间消耗 s

GML/MB	T_{WFS}	T_{WPS1}	T_{WPS2}
1.47	1	2.5	1
11.2	9	16	5
23.5	28	37	10

整个服务链聚合执行的时间费用可以分为3个部分:①WFS_1产生的GML传输费用;②WPS_1解析、处理GML费用;③WPS_2解析、处理GML费用.其中,空间处理算法依赖于服务端算法的性能优略,两种策略使用相同的WPS处理算法,在该部分所耗费的时间没有差异.因此,整体的性能取决于GML的传输以及对GML的序列化和解析的能力.

如果传输的GML数据量小,网路传输和解析GML的费用没有太大的区别.然而当数据量大时,网络传输GML的费用成倍增加,对比式(6)和式(10)可以发现,采用地址引用、压缩GML的策略首先避免了GML下载到本地造成的时间消耗,同时采用压缩GML的方式降低了网络传输费用,整体时间消耗接近密集GML策略的一半.

5 结 论

本文利用图形工作流为聚合标准OWS服务提供了新的解决方案,并且通过实验证明了这种方案不仅突破了OWS服务应用于工作流的接口瓶颈,而且通过改进的数据流调度策略提高了OWS服务链的执行效率.

参 考 文 献

[1] Alameh N. Chaining geographic information Web services [J]. IEEE Internet Computing, 2003, 7(5): 22-29

[2] Nickull D. Reference model for service oriented architecture, version 2.2 [S]. Billerica: Organization for the Advancement of Structured Information Standards (OASIS), 2006

[3] Jia Wenyu, Li Bin, Gong Jianya. Research on dynamic GIS chain based on workflow technology [J]. Geomatics and Information Science of Wuhan University, 2005, 30(11): 982-985 (in Chinese)
(贾文珏, 李斌, 龚健雅. 基于工作流技术的动态GIS服务链研究[J]. 武汉大学学报: 信息科学版, 2005, 30(11): 982-985)

[4] Granell C, Gould M, Ramos F. Service composition for SDIs: Integrated components creation [C] //Proc of the 2nd Int Workshop on Geographic Information Management (GIM'05). Piscataway, NJ: IEEE, 2005: 475-479

[5] Kiehle C, Greve K, Heier C. Standardized geoprocessing-taking spatial data infrastructures one step further [C] //Proc of the 9th AGILE Int Conf on Geographic Information Science. Piscataway, NJ: IEEE, 2006: 239-251

[6] Kiehle C, Greve K, Heier C. Requirements for next generation spatial data infrastructures-standardized Web based geoprocessing and Web service orchestration [J]. Transactions in GIS, 2007, 11(6): 819-834

[7] Granell C, Lemmens R, Gould M, et al. Integrating semantic and syntactic descriptions to chain geographic services [J]. IEEE Internet Computing, 2006, 10(5): 42-52

[8] Friis-Christensen A, Ostlander N, Lutz M, et al. Designing service architectures for distributed geoprocessing: Challenges and future directions [J]. Transactions in GIS, 2007, 11(6): 799-818

[9] Hu Naijing, Gu Ning, Shi Baile. Correctness of concur-rency based on semantic constraint resource workflow [J]. Journal of Computer Research and Development, 2003, 40(5): 712-719 (in Chinese)
(胡乃静, 顾宁, 施伯乐. 基于语义约束资源工作流并发正确性保证[J]. 计算机研究与发展, 2003, 40(5): 712-719)

[10] Ludäscher B, Altintas I, Berkley C, et al. Scientific workflow management and the kepler system [J]. Concurrency and Computation: Practice and Experience, 2006, 18(10): 1039-1065

[11] Schut P. OpenGIS Web Processing Service [M/OL]. Open Geospatial Consortium Inc (OGC), 2007. [2007-06-08]. http://www.opengeospatial.org

Zhang Jianbo, born in 1982. PhD candidate of Wuhan University. His main research interests include intelligent spatial information service.

Liu Jiping, born in 1967. Professor, PhD supervisor of the Chinese Academy of Surveying and Mapping. His main research interests include government geographic information system research and application(liujp@casm.ac.cn).

Wang Bei, born in 1983. PhD candidate in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. Her main research interests include economic geography and regional planning(wangb.09b@igsnrr.ac.cn).