第 35 卷 第 23 期 计 算 机 工 程 2009 年 12 月
Vol.35 No.23 Computer Engineering December 2009
·博士论文· 文章编号：1000—3428(2009)23—0021—03 文献标识码：A 中图分类号：TP311

基于分布式服务模型的遥感影像数据挖掘系统

李广水[1,2]，郑 滔[3]，宋丁全[2]
(1. 南京林业大学森林资源与环境学院，南京 210037；2. 金陵科技学院，南京 211100；3. 南京大学软件学院，南京 210093)

摘 要：依据遥感影像的特点及 WCS 标准，研究面向服务的遥感数据挖掘模式及其基于工作流的分布式系统架构。在此基础上，采用.Net 体系及 WSBPEL 流程建模语言，设计一个遥感影像纹理关联规则的挖掘系统，并研究对大数据集的访问、网络资源占用、程序的可伸缩性等方面的实现技术，为 Internet 环境下遥感影像的处理提供了依据。

关键词：遥感影像数据挖掘；Web 服务；WSBPEL 建模语言

Data Mining System for Remote Sensor Image Based on Distributed Service Model

LI Guang-shui[1,2], ZHENG Tao[3], SONG Ding-quan[2]
(1. College of Forest Resource and Environment, Nanjing Forest University, Nanjing 210037; 2. Jinling Institute of Technology, Nanjing 211100; 3. Software Institute, Nanjing University, Nanjing 210093)

【Abstract】According to the characteristic of remote sensor image and the open geospatial consortium Web Coverage Service (WCS) specification, this paper studies the model of data mining from remote sensor image and the architecture for integrating distributed SOA by work flow. Basing on the .Net framework and WSBPEL, an association rule mining system from the value of feature texture of remote sensor image is designed, which studies on aspects of accessing to a large data source, occupying the network resources, process scalability. It can provide the reference to the analysis for remote sensor image in Internet.

【Key words】data mining for remote sensor image; Web service; WSBPEL

1 概述

面向服务的架构是设计一个松散耦合、支持功能模块重用且容易扩展的系统平台，基于工作流的分布式系统是构建大型信息系统的关键。随着空间对地观测技术的发展，多传感器、多分辨率、多光谱遥感影像广泛地应用于相关行业，遥感数据已成为 GIS 的一个重要数据来源。而对遥感影像进行数据分析和数据挖掘的工作也得到了充分的重视。国际空间联盟(OGC)颁布了 WCS 标准(Web Coverage Service Implementation Standard)，以此来规范和推广基于 Web 服务的遥感数据处理[1]。

本文遵循 WCS 标准，提出一个基于分布式 Web 服务的遥感影像数据挖掘架构，并利用.Net 开发体系及 WSBPEL 流程建模给予具体实现。

2 遥感影像数据挖掘

基于遥感影像的数据挖掘研究，是采用数据挖掘技术从遥感影像中提取隐含未知的、潜在有用的并最终可理解的空间或非空间的一般知识或规则的过程，是依据一定的度量值和临界值从原始数据中抽取知识以及与之相关联的预处理、抽样和数据变换的一个多步骤、反复进行的交互过程。

文献[2]探讨了构建遥感影像的事务型数据矩阵及关联规则挖掘研究。分析对遥感数据的挖掘，总体上可将其归纳为以下几个步骤：(1)对遥感影像中包含的地物目标，地学现象和过程等进行描述、识别、分类、特征提取，并构建多源空间数据库；(2)依据遥感影像的像元亮度构建信息决策表等基础挖掘数据，在此基础上依据不同目的采用相应规则进行数据挖掘；(3)融合专家知识库、空间数据的信息挖掘模型等，对挖掘精度进行评定。

3 面向服务的遥感数据挖掘框架

构建面向服务的遥感数据挖掘框架是为了实现一个基于网络环境、松散耦合、方便集成的遥感数据处理平台。

文献[3]采用 C++, JAVA 语言开发了一个服务于地球物理领域的 Web 集成系统，该系统以 Web 服务的形式提供 3 个应用模块，分别服务于图像数据分析和数据挖掘服务、人机交互等功能，并针对具体实现分析了远程调用、可视化与 Web service 等相关技术。

依据面向服务的数据挖掘特点，可以将面向服务的遥感数据挖掘归类为 2 种实现方法，一种是基于 3 层 B/S 模式的结构，由 Web 服务器访问局域网内 Web 数据库实现的数据挖掘，并将挖掘过程以服务形式提供给服务消费者，其模式如图 1 所示，本文称为本地数据挖掘，它常常出现在业务相关部门之间的数据分析。一般而言，该模式在综合图像资源共享和人机交互操作方面得到一定的平衡，但也需要读取 Web 数据库服务器中的数据至 Web 服务器上进行数据处理，因而存在对数据库的频繁访问和大量的数据传输，另一个明显缺

基金项目：国家“863”计划基金资助项目“基于程序意识的软件安全性研究”(863/2007AA01Z448)；江苏省林业三项工程基金资助项目(lysx[2007]08)

作者简介：李广水(1965－)，男，副教授、博士研究生，主研方向：数据挖掘，系统集成；郑 滔，教授；宋丁全，教授、博士生导师

收稿日期：2009-06-06 **E-mail：**yz_lgs@126.com

点是因为挖掘服务是与特定数据库紧密相关的，不易于用户对算法的扩展。

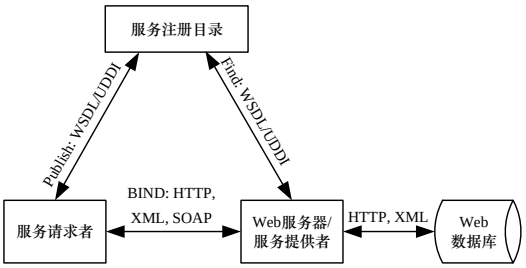

图 1　本地数据挖掘服务模式

另一种挖掘模式是构建一个独立的挖掘算法集成包，并提供给服务消费者。消费者调用该包中的某一服务并参数传输需挖掘的数据集以及相关参数，即可得到挖掘结果，其模式如图 2 所示，本文称为远程数据挖掘。该方法体现了算法的集中，可以方便地扩展挖掘算法，形成逐步完善的挖掘算法中心，但是在调用服务过程中将有大量数据的远程传输。

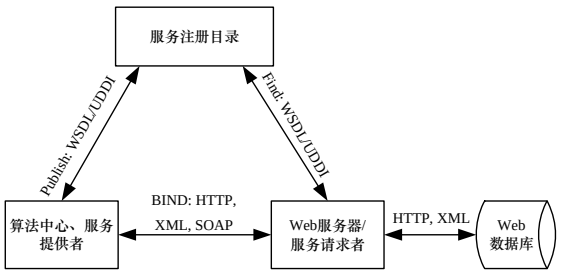

图 2　远程数据挖掘服务模式

遥感数据挖掘要在挖掘过程中进行反复叠代并产生大量的中间结果，对主机资源提出了更高的要求，因此，必须保证算法的可伸缩性。而需要远程调用的 Web 服务在数据挖掘中面临的另一主要问题是如何优化对数据库的访问以及控制网络数据传输流量，这也是基于 Web 服务的数据挖掘的主要瓶颈，一般而言，在现有的条件下，除了在程序设计方面需要考虑以上的问题外，从系统架构出发，采用分布式系统是解决问题的关键。

4　基于BPEL工作流的分布式系统

将业务过程分解为独立的任务，不同的任务实现一个具体的功能，通过工作流组合多个任务形成一个分布式系统。由于直接抽象于业务过程，因此工作流能充分表征复杂业务逻辑，并能适应业务过程的变化。

为了使分布式的遥感图像挖掘协同工作，须将功能模块进行流程建模，再由执行引擎按照工作流程规范定义来执行。基于 XML 的 BPEL (Business Process Execution Language)规范是 IBM 的 WSFL 和 Microsoft 的 XLANG 相融合的产物，它提供了一套 XML 语法来对业务流程交互中的 Web 服务行为及控制逻辑进行描述，可以将孤立的、无状态的 Web 服务进行整合，从而发挥 Web 服务技术作为应用集成平台的全部潜力。BPEL 不仅实现了 Web 服务间的交流和流程编排，而且将流程自身也暴露为 Web 服务，从而为编排 Web 服务提供了强大的优势。

文献[4]针对面向服务的体系结构探讨了基于 BPEL 工作流执行体的访问控制模型，结合遥感图像挖掘的特点，可以构建基于工作流的、分布式 Web 服务的遥感图像挖掘系统，其结构如图 3 所示，主要步骤如下：

(1)考虑到遥感图像数据量大，其基础数据不能直接进行数据挖掘的特性，为得到易于挖掘的事务数据集，可能存在多次反复的人机交互操作，因此，应将包括构建信息决策表等数据预处理功能开发成为本地数据挖掘服务。

(2)将基于关联分析、聚类分析等数据挖掘算法包装成算法中心的 Web 服务。

(3)流程执行引擎依据 BPEL 调用相关服务并监控流程执行情况，实现 Web 服务的流程调度。

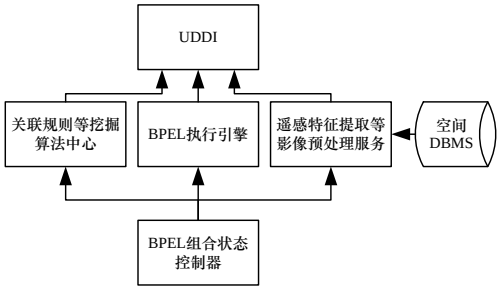

图3　基于BPEL工作流的分布式遥感图像挖掘系统

5　基于.Net体系及WSBPEL的系统实现

本节采用.Net 开发体系，实现一个基于 WSBPEL 工作流的、分布式的 Web 服务，进行遥感影像纹理关联规则的挖掘。

5.1　.Net处理数据密集型系统的特点

.Net 体系依据 ADO.Net 创建分布的、数据共享的应用程序。ADO.Net 使用 XML 来交换不同对象的数据调用，保证了数据在 Internet 上的畅行无阻。ADO.Net 采用连接池实现对数据库的访问，依据可靠的、分层的非连接数据缓存技术，以脱机方式处理数据，有效地控制了对网络资源的占用。

ADO.Net 可以访问包括 Oracle 等多个主流数据库管理系统，而其与 SQL_server 的集成从一开始就保持了独特的优势，一方面 ADO.Net 包装了 SQL_server 提供的底层接口组件，另一方面 SQL_server 本身也积极拓展其与 ADO.Net 的接口，在其 2005 版中的一个最大变化是数据库与.Net 的公共语言运行库的良好结合：SQL_server 托管执行.Net 程序进一步实现了两者的集成，为基于网络环境下的数据操作开启了一个有效的通道。

5.2　遥感影像纹理关联规则挖掘原理

图像中像素的光谱特征，构成了纹理图像的各个像元数据，各个纹理基元之间都具有关联关系，这是关联规则挖掘能够应用于图像的前提。挖掘遥感图像纹理之间的关联规则，需要将一个像素及其邻域看作一个事务，从中找出在图像中重复出现的模式，一般在实际分析中，可依据图幅大小、分辨率及研究目的进行邻域距离区划的多次设定，以获得更好的效果。

对遥感影像的纹理关联规则挖掘，主要包括以下几步：

(1)遥感像素中项的建立：将每一个非边界的像素设为根像素，对其所在的邻域每间隔 45° 角共 8 个方向通过一个关系(X, Y, I)来定义一个项，其中，X 和 Y 分别是邻域相对于根像素的偏移量；I 是像素的灰度值。

(2)事务表的建立：同一根像素下所有项集组成一个事务，所有事务构成事务表。

(3)关联规则挖掘：依据事务表进行关联规则挖掘，根据给定的支持度和可信度得到基于位置、灰度值的关联规则。

另外，考虑到太多的灰度值会导致巨大的项集，且也不利于纹理的表达，因此，依据灰度分布直方图进行灰度合并，

可以采用均等像素个数或均等距离为一段，取该段的灰度均值作为该段的灰度值，从而实现灰度降阶。

5.3 本地遥感影像的预处理服务

采用与 Web 数据库同一局域网内的 Web 服务器来对遥感数据进行数据预处理，可以减少网络远程传输流量，提高数据处理效率，该服务主要包括对遥感图像进行事务分析，构造易于挖掘的事务集。

数据挖掘特别是图像数据挖掘将来回访问数据库，且有大量的数据在服务器和数据库间传输，因此，数据挖掘系统的结构设计变得非常重要。传统的提高数据库访问效率的手段包括通过 SQL 游标、调用数据库的存储过程、调用自定义函数、扩展 SQL 至 DMM 等几种方法，但其编程基于数据库内部命令或自定义函数局限了数据挖掘中对算法的拓展，甚而一些功能的无法实现，且没有充分利用数据库对内存、线程、同步方面的管理优势，这也极大地限制了强调算法的数据挖掘在数据库中的集成。

利用.Net 的托管存储过程来实现对指定数据的挖掘，可以克服以上提及的缺点，这种与数据库紧密集成的挖掘服务一方面由于对数据的操作使用存储过程来进行，避免了数据库中的数据与 Web 服务器之间的来回交互，同时，托管的代码都是依靠 SQL_server 的管理机制实现对数据的访问，这就充分发挥了数据库对数据的管理优势，使得算法的可伸缩问题完全拓展到数据库管理系统的能力[5]，且利用.Net 的高级语言编码克服 T-SQL 语言编码所固有的柔性低、相对功能简单的缺陷。文献[5]利用实证的方法进行两者速度比较，认为 SQL 托管存储过程远远优于 T_SQL 存储过程。

利用托管存储过程对本地遥感影像的预处理包括读入图像文件，将像元数据转化为灰度值，并依据灰度值及其出现频次进行灰度降阶，并映射生成每一个根像素的事务项集，从而得到包含所有项集的事务表。

Web 服务器端的服务程序调用该托管存储过程并发布为 Web 服务，简写为 RSPreAnalyseServer。

5.4 算法中心的关联规则挖掘

Apriori 算法是经典的关联规则挖掘算法，也是数据挖掘领域研究热点之一，一种改进的 Apriori 算法只需对数据库扫描一次，就可以得到所有频繁项集，其主要思想是读取事务数据库中记录，并对每个事务的项进行本事务内部的组合得到 *x*-项集，对扫描过程中出现的相同项集进行次数累加，结束扫描后，就得到事务数据库中所有项集出现的频次。

作为独立的算法服务，用户调用时需要传输被挖掘的数据库地址、源表名称、相关阈值等参数，Web 服务连接上该数据库并访问事务表进行数据处理。利用.Net 的 DateSet 对像可以动态提取远程数据库中的数据并构造易于挖掘的本地内存数据库，这样，在释放数据连接资源后经过对本地内存数据的操作实现数据挖掘。

对关联规则挖掘的具体实现方法和步骤如下：

(1)利用 ADO.Net 组件中具有只读最快速度的 SqlData Reade 对象[5]读入 Web 数据库中的事务数据，并保存在本地内存的 DateSet 的分析源表中，考虑到事务数据中的数据量巨大，并且已经处理过的记录将不再进行扫描处理，可以采用有限多次的覆盖读入方法，这实现了该算法的可伸缩性。

(2)在 DateSet 中创建一个临时表，保存一个项的组合、该组合包含项的个数及重复次数，在对分析源表进行扫描并进行项集所有组合之后，更新该临时表记录值。

(3)依据临时表及最小置信度找出强关联规则，并返回频繁项集。

完成以上步骤后，命名并发布此服务为 AprioriServer。

5.5 BPEL流程的创建

WSBPEL 流程用 3 种活动与外界伙伴进行交互：<invoke>, <reply>和<receive>，通过指定接口类型、操作以及伙伴，可以标识出它所调用的 Web 服务。在结构化的操作方面，WSBPEL 用<sequence>, <flow>, <switch>, <while>等定义顺序、并行、分支、循环等调用，实现流程建模。

在遥感图像纹理关联规则的挖掘服务中首先调用本地遥感影像的预处理服务，生成事务表，算法中心的关联规则挖掘服务通过分析该事务表，找出并返回频繁项集，从而实现遥感影像的挖掘服务。这是一个顺序结构，其具体调用过程如图 4 所示。

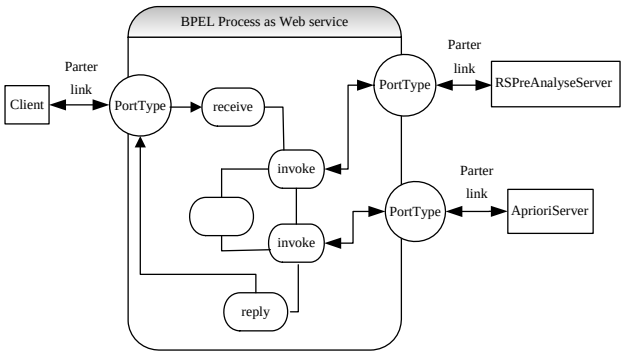

图 4 BPEL 调用 Web 服务的过程

代码片段如下所示：

```
…
<InvokepartnerLink=” RSPreAnalyseServer”
  portType=”RSPreAnalyseServerPT”
  operation=”RSPreAna” />
 <sequence>
  <invokepartnerLink=” AprioriServer”
  portType=” AprioriServerPT”
  operation=” ApriAna”
  inputVariable="SourceAddress"
  …
  outputVariable=”FrequencyIsets" />
</sequence>
…
```

6 结束语

本文探讨了Web服务下分布式遥感影像关联规则的挖掘处理及流程建模的具体实现，如何构建一个鲁棒的、安全的分布式遥感数据处理系统是下一步的研究重点，主要包含以下 2 个方面：

(1)构建算法中心的远程数据挖掘服务，将不可避免地进行大量数据的网络传输，批量读入源数据，构建主存中易于挖掘的数据模式是一个重要的应对策略，而如何依据系统运行时的资源状况及算法要求，动态地确定将要处理的数据块大小是提高系统效率的重点。

(2)在遥感图像数据挖掘系统中，降低时延是非常重要的一步，为降低网络流量如何鉴别将被处理的数据，即提高传输子集的命中率是问题的关键。

(下转第 26 页)