

# 网格环境下分布式空间离群挖掘体系的设计与应用

姚明经<sup>1</sup>, 林甲祥<sup>1</sup>, 陈崇成<sup>1\*</sup>, 马亨冰<sup>2</sup>

(1. 福州大学福建省空间信息工程研究中心、空间数据挖掘与信息共享教育部重点实验室, 福州 350002;

2. 福建省经济信息中心, 福州 350001)

**摘要:** 空间离群是指空间数据集中那些非空间属性值与邻域中其他空间对象明显不同的空间对象。空间数据一般按地理分布存储具有海量特性, 传统的集中式处理模式不能满足海量数据处理的效率和空间数据本身的安全性等要求。因此, 在研究小组开发的地理知识服务网格平台 GeoKS-Grid 的基础上, 本文针对分布式空间离群挖掘, 提出了一个基于网格的分布式体系框架, 制定了网格环境下分布式空间离群挖掘的策略, 实现了具体的分布式空间离群挖掘算法。另遵循分布式空间数据挖掘的一般过程和网格服务通用、可重用和可组合的原则, 将算法按合理粒度进行分解, 并封装成多个基本的原子服务, 进而以网格工作流的方式进行服务发现与组合, 完成包括局部离群挖掘和全局离群挖掘在内的分布式空间离群挖掘。最后, 通过福建省生态地球化学调查土壤数据离群分析实例, 验证了服务或系统的合理性和有效性。

**关键词:** 空间离群; 分布式挖掘; 知识网格; 原子服务; 服务组合

**DOI:** 10.3724/SP.J.1047.2011.00383

## 1 引言

空间离群是指那些非空间属性值与空间邻域中其他对象显著不同的空间对象<sup>[1]</sup>。空间离群挖掘可为人们提供很多丰富多彩的信息。其在分布式网络应用环境中, 由于空间数据集的地理分布特征, 以及空间离群的自身区域部分特征, 空间离群往往从局部节点和全局数据集中的局部区域两个角度对空间离群进行刻画。因此, 分布式空间离群挖掘中既需局部节点进行空间离群的检测, 也需从全局数据集的角度对局部不稳定特征进行挖掘。局部空间离群是指“局部数据集中非空间属性与邻域中其他空间对象显著不同的空间对象”, 而全局空间离群是指“从全局数据集来看, 非空间属性值与邻域中其他对象显著不同的空间对象”。

网格分布式数据挖掘与传统的分布式数据挖掘相似, 主要对平台的分布式体系结构、算法和挖掘系统等三个方面进行研究。在算法研究方面, 典

型的工作如: Cristian 等<sup>[2]</sup>在网格平台上实现了分布式 Apriori 算法; Rawat 和 Rajamani<sup>[3]</sup>以网格实现了分布式 FP-growth(Frequent Pattern Growth)算法; Ali Meligy 等<sup>[4]</sup>提出网格的分布式支持向量机 SVM 算法, 并予以实现; Yang 等<sup>[5]</sup>提出基于网格的决策树并在网格环境中实现了 SPRINT 算法。在分布式数据挖掘系统方面, Pérez、Khoussainov、Senger、Alia、Talia 等<sup>[6-10]</sup>结合怀卡托智能数据分析与挖掘环境(Weka)和网格, 分别开发了一系列网格的数据挖掘系统, 如 GridWeka、GridWeka2、Inhambu、FAEHIM、Weka4WS 等。而 Brezan 等<sup>[11]</sup>提出的 Grid Miner 主要组成部分是: 调解服务、信息服务、资源代理和 OLAP 立方体管理。其为数据整合、流程管理、数据挖掘和 OLAP 提供基本服务。Stankovski 等提出的 DataMiningGrid<sup>[12]</sup>使网格环境下数据挖掘应用的开发和部署协调一致。然而, 现有的大多分布式数据挖掘算法主要以关联规则和分类预测为重点, 以 Weka 软件包为核

**收稿日期:** 2010-11-01; **修回日期:** 2011-03-29.

**基金项目:** 国家自然科学基金项目(30972299); 中-匈政府间科技合作项目(国科外字[2008]333号); 欧盟第七框架计划项目(FP7-2009-People-IRSES, No. 247608); 福建省重点科技项目(2010I0008)。

**作者简介:** 姚明经(1984-), 男, 湖北宜都人, 硕士研究生, 主要研究方向为空间数据挖掘与信息可视化。

E-mail: mingjingyao@yahoo.com.cn

\* **通讯作者:** 陈崇成(1968-), 男, 福建闽清县人, 博士、教授, 主要研究方向为空间数据挖掘与知识网格、地学可视化与虚拟地理环境。E-mail: chencec@fzu.edu.cn

心的数据挖掘套件本身不具备专门的离群挖掘能力,对空间数据进行挖掘处理的针对性不强,且大多数研究工作从算法设计的角度出发,不能实现算法的“一次部署,多次使用”的目标。因此,本文以现有的地理空间知识服务网格平台 GeoKS-Grid<sup>[13]</sup>为支撑平台,充分利用网格环境中的各种数据、软件和硬件资源,对面向服务的分布式空间离群挖掘,开展网格的分布式空间离群挖掘体系框架和相适应的分布式空间离群挖掘算法进行研究。同时通过福建省生态地球化学调查土壤数据离群分析,对服务或系统的有效性进行了验证。

2 网格分布式空间离群挖掘体系框架的设计

2.1 地理知识服务网格 GeoKS-Grid

GeoKS-Grid 平台是福建省空间信息工程研究中心历经 6 年时间研发形成的一个通用的地理知识服务网格系统<sup>[14]</sup>。系统采用开放网格体系结构,将各种网格资源(包括计算资源、存储资源、网络、

程序、数据库等)抽象为服务,并通过统一的标准接口来管理和使用网格。其中,网格服务(Grid Service)是一种 Web Service,该服务提供了一组相对统一的接口,这些接口的定义明确并且遵守特定的管理,所有的网格服务都基于这些接口实现,可以方便地进行扩展获得网格服务的集合,也可以很容易地构造出具有层次结构、更高级别的服务,这些服务可以跨越不同的抽象层次,以一种统一的方式来看待。GeoKS-Grid 平台由资源层、网格通用基础层、核心知识网格层、高级知识网格层和网格门户等多层结构组成。

2.2 网格分布式空间离群挖掘体系框架的设计

以地理知识服务网格 GeoKS-Grid 平台为依托环境,结合网格平台自身的体系结构,设计如图 1 所示分布式空间离群挖掘的体系框架。在该体系结构中,分布式离群数据挖掘作为一种知识服务,处于网格平台的知识网格层,通过平台提供的一系列具有统一接口的网格服务,实现分布式离群挖掘的各个过程、步骤和组件之间的通讯与合作,进而

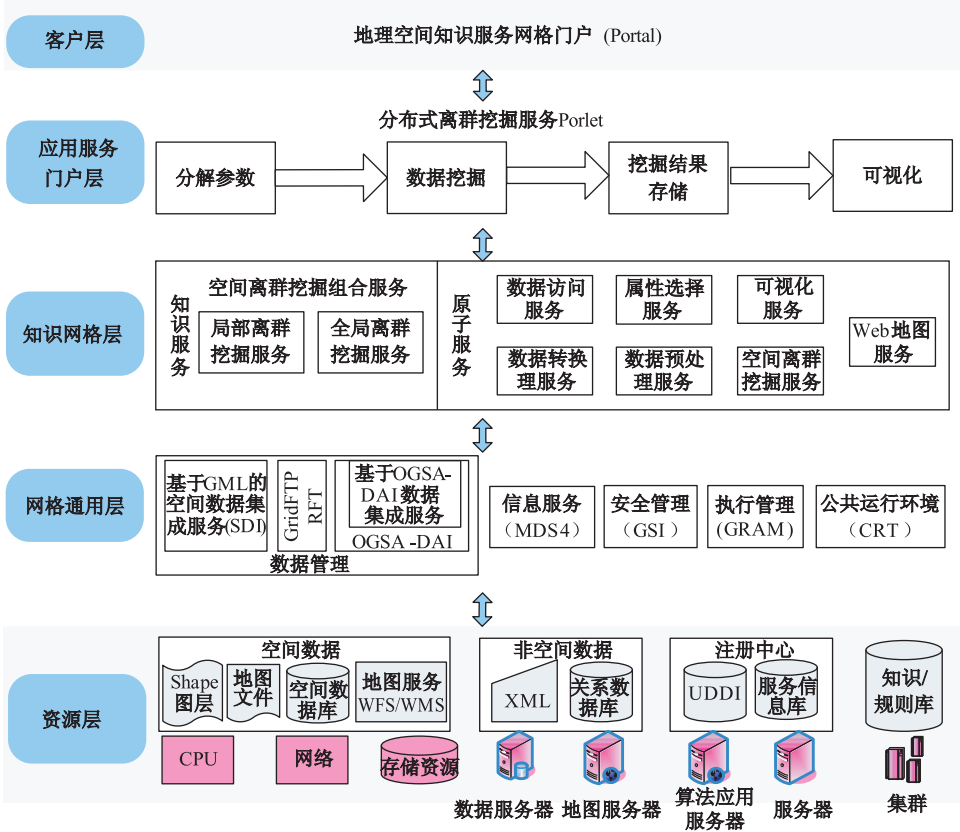


图 1 网格分布式空间离群挖掘体系框架

Fig. 1 The architecture of distributed spatial outlier mining in grid environment

实现分布式空间离群挖掘的功能,为网格用户提供持续、稳定的空间离群挖掘服务。

从系统的组成结构来看,与传统的数据管理、局部挖掘、全局挖掘三个核心模块的分布式数据挖掘系统不同,以地理知识服务网格 GeoKS-Grid 平台的分布式空间离群挖掘系统主要通过架设于底层的数据资源、存储资源、计算资源、知识服务资源之上的信息服务中心、资源管理中心、执行管理中心和知识服务中心,实现具体的分布式空间离群挖掘任务。

信息服务中心对网格平台中的各种软硬件资源进行监控,为分布式空间离群挖掘任务提供基本的服务检索、发现、选择等功能,为分布式空间离群挖掘选择各种可用的数据资源、存储资源、计算资源、服务资源等提供服务。

资源管理中心对数据、存储和计算等资源进行统一的管理,对待挖掘的数据(包括数据文件、关系数据、空间数据集等)进行管理、集成、分块与分配,而不仅仅局限于纯粹的数据管理,从而为分布式空间数据挖掘任务的执行提供了重要的环境,其中的

关系数据集成服务(OGSA-DAI)将异构操作系统中的数据文件、XML 文档等、不同数据库服务器(如 MS SQLServer、Oracle 等)中数据表进行集成,它将数据资源以服务的形式向外发布;地理空间数据集成服务(SDI)支持不同地理空间数据资源的集成,采用 GML 格式向外提供标准的数据;

执行管理中心引用特定的数据资源、计算资源和存储资源,在多个局部网格节点进行分布式空间离群挖掘,而全局节点对局部数据进行汇总与全局挖掘,从而最终实现分布式空间数据挖掘。

知识服务中心提供具体的挖掘服务,包括原子服务和组合服务,实现具体的局部挖掘和全局挖掘功能,并提供了数据挖掘过程中必要的的数据预处理、数据转换、数据归约等服务。

根据图 1 所示的分布式空间离群挖掘体系框架,分布式空间离群挖掘策略如图 2 所示。分布式空间离群挖掘的实现主要由待挖掘空间数据准备、局部空间离群挖掘、全局空间离群挖掘、挖掘结果确认等 4 个部分构成:

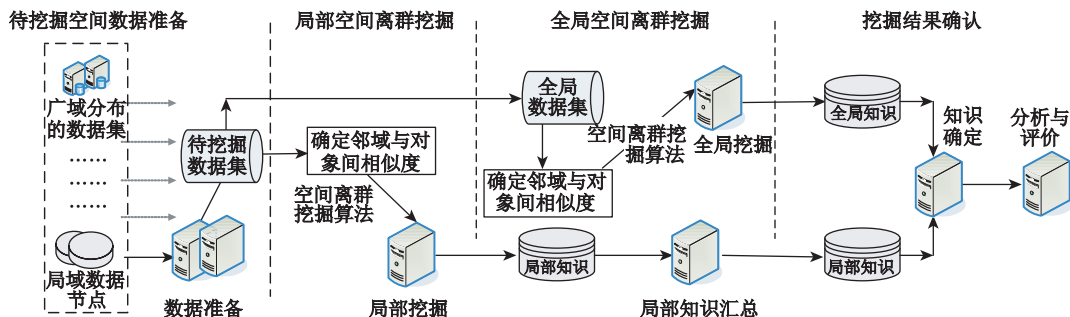


图 2 分布式空间离群挖掘策略

Fig. 2 The implementation strategy of distributed spatial outlier mining

(1)待挖掘空间数据准备:主要完成空间数据的预处理(如,数理清理、特征选择、维归约、离散化等),为空间数据挖掘提供基础数据集。

(2)局部空间离群挖掘:对各局部节点上的待挖掘数据,确定邻域与对象相似度,主要完成空间邻近关系与邻域定义,以及对象相似度计算,为空间对象的离群因子定义与计算奠定基础,进而调用空间离群挖掘算法完成局部离群挖掘。

(3)全局空间离群挖掘:汇总并可视化局部离群挖掘结果,同时对各局部节点的待挖掘数据集集成,形成全局空间数据集,对整体数据集中对象之间的空间邻近关系描述与表示,以及对象相似度计

算,进而采用空间离群挖掘算法进行全局离群挖掘,获得全局离群。

(4)离群挖掘与结果评价:完成具体的空间离群挖掘,并对空间离群挖掘结果的意义、有效性和可靠性进行确认,并结合应用领域的背景知识对离群的意义以及产生的原因进行分析获得最终的离群模式或知识。

### 3 算法及其服务的设计与实现

#### 3.1 分布式空间离群挖掘算法

本文采用如上所述的体系结构和挖掘策略来

设计分布式离群挖掘算法,完成分布式空间离群挖掘。本文提出的分布式离群挖掘算法主要分为三部分:空间离群挖掘算法、局部离群和全局离群。

### (1)空间离群挖掘算法

离群挖掘的关键在于离群的定义,离群挖掘和离群的解释。离群挖掘包括对象之间的相似性度量定义、邻域的确定。其中,空间离群定义、对象之间的相似性度量定义、邻域的确定作为关键技术,对于空间离群挖掘算法的成败具有重要的制约作用。算法设计与实现详见<sup>[14]</sup>。

本文采用基于密度的思想将空间离群定义为空间离群因子值最大的若干个目标对象,空间离群因子分别采用改进的 SLOF<sup>[15]</sup>和 SLOM<sup>[16]</sup>,依据空间属性,分别采用 K 邻域和 Delaunay 三角网对空间对象之间的邻近性进行定义与描述<sup>[17]</sup>;依据非空间属性,采用欧氏距离对空间目标之间的相似度进行计算。在空间离群定义的基础上,算法首先由对象包含的空间属性信息和空间关系信息确定对象的邻域,然后依据空间对象的专题属性计算对象和邻域的差异来实现离群挖掘。挖掘过程采用各种算法和技术,经过一系列的综合分析与处理,从空间数据集中提取潜在有用的、偏离数据集主体模式的知识。

### (2)局部离群

在各个局部节点上调用空间离群挖掘算法进行分布式空间离群挖掘,获得相对于各个局部节点的局部空间离群,核心挖掘步骤包括空间数据预处理(如,数理清理、特征选择、维归约、离散化等)、Delaunay 三角网构建、空间离群挖掘。然后,使用 GeoKS-Grid 平台提供的空间数据集成服务,汇总局部离群挖掘结果。

### (3)全局离群

对局部各节点空间数据进行集成,形成整体数据集,构成整体 Delaunay 三角网,对整体数据集中对象之间的空间邻近关系描述与表示,进而采用空间离群挖掘算法进行全局离群检测,获得全局离群。

## 3.2 分布式空间离群挖掘服务

为了使算法能被更多的用户使用,且与研究小组的工作更好地结合,本文把分布式空间离群挖掘算法封装成网格服务,发布在研究小组开发的地理知识服务网格(GeoKS-Grid)平台上,将之作为网

格应用服务为类似的研究提供参考,也为实际项目应用提供快速的挖掘结果,减少算法的重复开发。分布式空间离群挖掘的实现,具体包括地理空间数据集成服务、Delaunay 三角网构网服务、GML 文件解析服务、属性选择服务、空间离群挖掘服务和 GML 文件生成服务。

### 3.2.1 原子服务

根据上文设计的分布式空间离群挖掘算法,遵循通用、可重用和可组合原则,采用数据分布并且算法分布思想,将算法按一定粒度分解,并封装成网格服务。依据分布式空间离群挖掘的过程,将分布式空间离群挖掘算法分解为以下 6 个原子服务。

#### (1)地理空间数据集成服务

空间数据集成服务<sup>[18]</sup>运用网格的高性能计算能力,根据用户对空间数据的需求,同时调用管理空间数据的 Web Service,访问多个地理分布的数据节点,并将不同数据节点上获取的空间数据进行整合,成为 GML 格式的数据流返回给用户。

#### (2)Delaunay 三角网构网服务

Delaunay 三角网构网服务为空间离群挖掘提供空间邻域的定义与描述,已实现的 Delaunay 三角网构网方法包括分治法、逐点插入法、混合法和凸壳法。

#### (3)GML 文件解析服务

在 GeoKS-Grid 平台上,地理空间数据集成服务主要集成了分布在各节点的 GeoServer 数据,以 WFS/WMS 服务的形式共享,而空间离群挖掘算法一般针对表格形式的数据,GML 文件解析服务负责解析请求 WFS 返回的 GML 格式的数据,并转换为二维字符串数组。

#### (4)属性选择服务

GML 格式的数据经过 GML 文件解析服务转换成二维字符串数组后,需要区分出 ID、空间坐标和属性数据,以适应空间离群挖掘算法的需要。属性选择服务提供了各种各样的接口,可以从二维字符串数组中选择出是否包含 ID,包含空间坐标或属性数据的数据。

#### (5)空间离群挖掘服务

空间离群挖掘服务根据不同的邻域和离群因子,提供空间离群挖掘方法。包括 K 邻域和改进的 SLOF 离群因子的空间离群挖掘算法 KnnSLOF、Delaunay 图的  $k$  阶邻近和改进的 SLOF 离群因子的空间离群挖掘算法 DtinSLOF、K 邻域和 SLOM

离群因子的空间离群挖掘算法 KnnSLOM、Delaunay 图的  $k$  阶邻近和 SLOM 离群因子的空间离群挖掘算法 DtinSLOM。

#### (6) GML 文件生成服务

GML 文件生成服务将空间离群挖掘的结果转换成 GML 格式的数据, 以便可视化。

#### 3.2.2 局部离群挖掘

通过调用上述原子服务完成局部离群挖掘, 调用过程如图 3(a-b-c-d-e-f-g-h-i-j-k-l-m-n) 所示:

图中, (a-b) 步骤对应地理空间数据集成服务, 对从不同数据节点上获取的空间数据进行整合, 成为 GML 格式的数据流返回给用户; (c-d) 步骤对应 GML 文件解析服务, 主要完成从服务申请信息到待处理 GML 数据转换成二维字符串数组; (f-g) 步骤对应属性选择服务, 从各局部节点二维字符串数组中, 提取待挖掘点对象的空间属性和非空间属性; (h) 步骤对应 Delaunay 三角网构网服务, 主要是根据提取的空间属性构建 Delaunay 三角网; (i-j) 步骤对应空间数据离群挖掘服务则是根据各局部节点 Delaunay 三角网构网信息、非空间属性, 以及用户提供的各种参数调用空间离群挖掘算法来进行离群检测, 并将挖掘结果存储并返

回; (k-l) 步骤对应 GML 文件生成服务, 将各节点局部空间离群挖掘的结果, 转换成 GML 格式的数据; (m-n) 是对局部离群结果进行确认与分析, 并调用地理空间数据实现服务集成并可视化局部离群挖掘结果。

#### 3.2.3 全局离群挖掘

全局离群挖掘其调用过程如图 3(a-b-c-d-e'-f'-g'-h'-i'-j'-k'-l'-m-n) 所示:

其中, a-d 步骤同上, (e') 步骤将各节点解析后的二维字符串数组进行合并, 生成总的二维字符串数组; (f'-g') 步骤对应属性选择服务, 从合并后总的二维字符串数组中, 提取待挖掘点对象的空间属性和非空间属性; (h') 步骤对应 Delaunay 三角网构网服务, 主要是根据提取的总的空间属性构建 Delaunay 三角网; (i'-j') 步骤对应空间数据离群挖掘服务则是根据全局 Delaunay 三角网构网信息、全局非空间属性, 以及用户提供的各种参数调用空间离群挖掘算法来进行离群检测, 并将挖掘结果存储并返回; (k'-l') 步骤对应 GML 文件生成服务, 将全局空间离群挖掘的结果转换成 GML 格式的数据; (m-n) 是对全局离群结果进行确认与分析, 并调用地理空间数据实现服务集成并可视化全局离群挖掘结果。

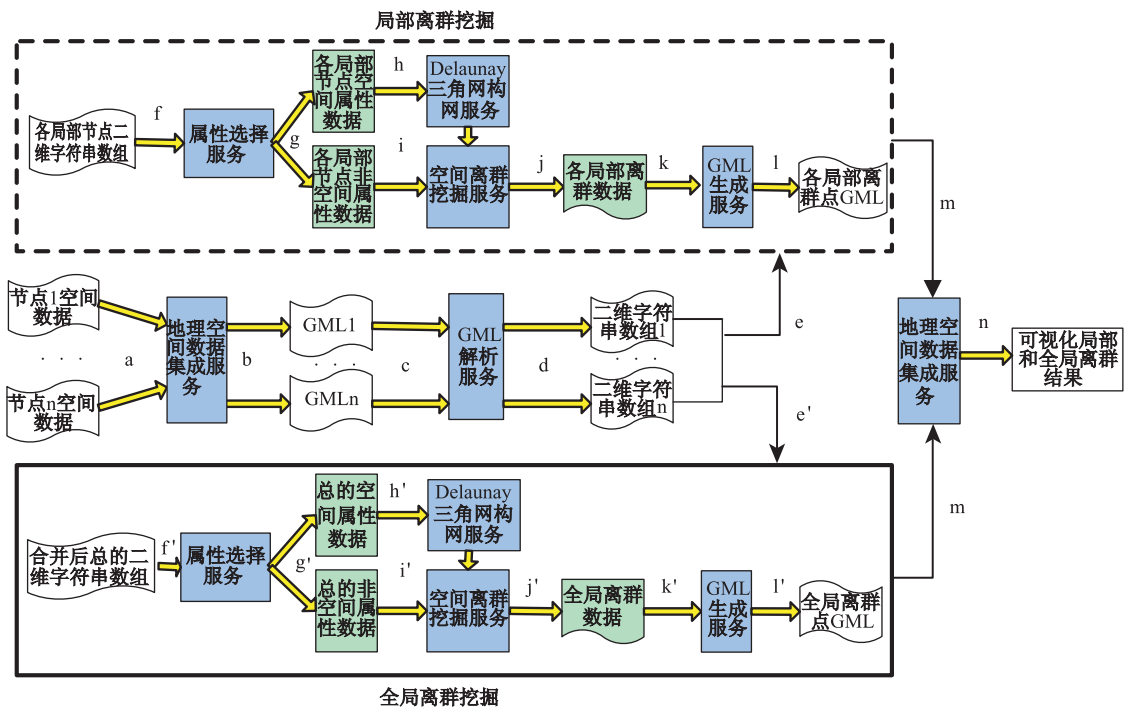


图 3 局部离群挖掘和全局离群挖掘过程

Fig. 3 The process of global and local spatial outlier mining



4 系统在土壤数据离群分析中的应用

示范应用数据主要来源于中国地质调查局开展的福建省沿海经济带生态地球化学调查项目。土壤数据采样区主要来自于福建东部沿海、濒临台湾海峡的福州和泉州两个地区沿海部分,福州地区范围涵盖福州城区、福清、长乐、平潭、连江、罗源、闽侯大部 and 闽清县大部,泉州地区沿海主要包含惠安和石狮两个县级行政单元。土壤数据的采样时间为 2003 - 2004 年。

以市级行政区划为待挖掘数据集划分的边界,将实验区数据集划分为独立的两个数据子集,即泉州地区和福州地区两个数据子集,并由 GeoKS-

Grid 平台中两个具有空间离群挖掘能力的网格服务节点,进行分布式与并行空间离群挖掘。数据子集分别分布在节点 211.80.198.41 和 211.80.198.168,空间离群挖掘网格服务分布在节点 211.80.198.162 和 211.80.198.41,采用 1 阶邻近为地理对象的空间邻域,最终获得两个地区的土壤采样离群点(以空间离群因子最大的 16 个对象为例),如图 4 所示。

服务执行完毕后,点击“显示结果”按钮,获得的离群因子最大的 16 个对象如图 5 所示。其中,图 5(a)为网格节点“211.80.198.41 和 211.80.198.168”数据子集局部离群挖掘结果,图 5(b)为福州地区和泉州地区全局离群挖掘。从图可以看出,全局离群挖掘结果既包括福州地区也包含泉州地区。

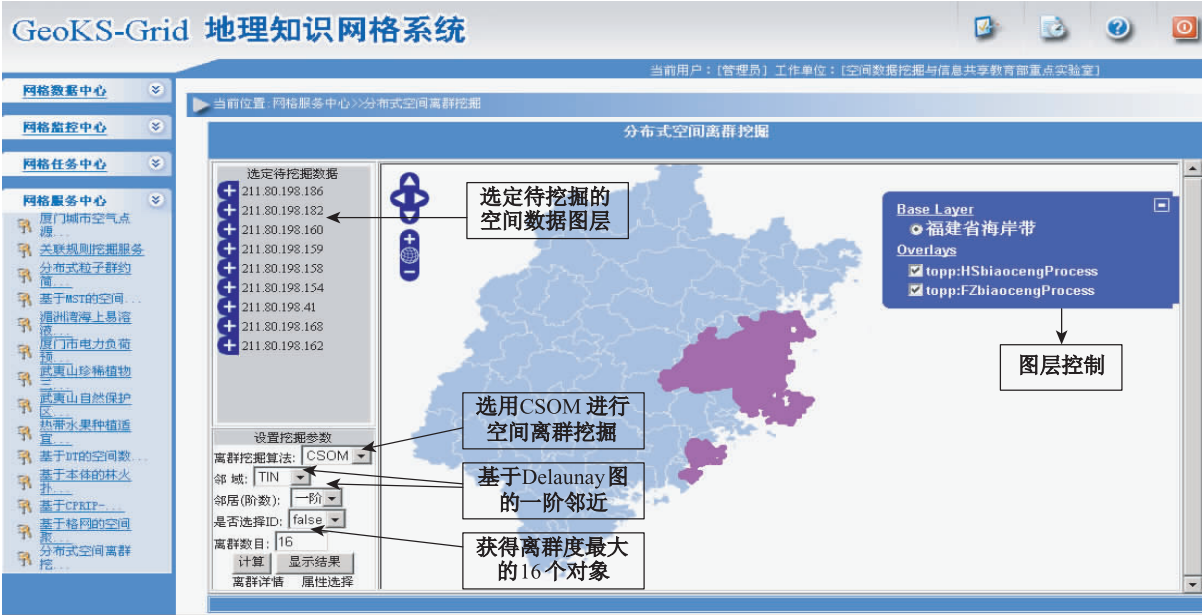


图 4 GeoKS-Grid 平台中分布式空间离群挖掘门户页面(Portlet)

Fig. 4 The distributed spatial outlier mining portlet in GeoKS-Grid



图 5 分布式空间离群挖掘之显示结果

Fig. 5 Results of distributed spatial outlier mining

## 5 结论

本文设计了网格分布式空间离群挖掘体系的框架, 提出了包括空间数据集成与分发、局部挖掘、全局挖掘、结果确认等在内的分布式空间离群挖掘策略。遵循通用、重用原则, 将分布式空间离群挖掘算法按一定粒度分解并封装成多个网格原子服务, 通过采用网格工作流的方式进行服务发现与组合, 完成包括局部离群挖掘和全局离群挖掘在内的分布式空间离群挖掘。并以研究小组开发的地理空间知识服务网格 GeoKS-Grid 为依托环境, 开发了分布式空间离群挖掘门户页面 (Portlet)。通过福建省生态地球化学调查土壤数据离群分析实例, 验证了网格分布式空间离群挖掘服务或系统设计的合理性和有效性。

## 参考文献:

- [1] Shekhar S, Lu C T, Zhang P. A Unified Approach to Detecting Spatial Outliers[J]. *GeoInformatica*, 2003, 7 (2): 139 - 166.
- [2] Aflori C, Craus M. Grid Implementation of the Apriori Algorithm [J]. *Advances in Engineering Software*, 2007, 38(5): 295 - 300.
- [3] Rawat S S, Rajamani L. Performance of Distributed Apriori Algorithms on a Computational Grid[C]. *Services Computing Conference. APSCC 2009. IEEE Asia-Pacific*, 2009, 163 - 167.
- [4] Meligy A, Al-Khatib M. A Grid-based Distributed SVM Data Mining Algorithm[J]. *European Journal of Scientific Research*, 2009, 27(3): 313 - 321.
- [5] Yang C T, Tsai S T, Li K C. Decision Tree Construction for Data Mining on Grid Computing Environments [C]. *19th International Conference on Advanced Information Networking and Applications, AINA 2005, Taipei, Taiwan: Institute of Electrical and Electronics Engineers Inc.*, 2005, 421 - 425.
- [6] Pérez M S, Sánchez A, Robles V, *et al.* Design and Implementation of a Data Mining Grid-aware Architecture[J]. *Future Generation Computer Systems*, 2007, 23(1): 42 - 47.
- [7] Khossainov R, Zuo X, Kushmerick N. Grid-enabled Weka: A Toolkit for Machine Learning on the Grid[J]. *ERCIM News*, 2004, 59: 47 - 48.
- [8] Senger H, Hruschka E R, Silva F a B, *et al.* Inhambu: Data Mining Using Idle Cycles in Clusters of PCs [M]. *Network and Parallel Computing. Springer Berlin / Heidelberg*, 2004, 213 - 220.
- [9] Ali A S, Rana O F, Taylor I J. Web Services Composition for Distributed Data Mining[C]. *ICPPW '05 Proceedings of the 2005 International Conference on Parallel Processing Workshops, IEEE Computer Society, Washington DC, USA*, 2005, 11 - 18.
- [10] Talia D, Trunfio P, Verta O. Weka4WS: AWSRF-enabled Weka Toolkit for Distributed Data Mining on Grids [C]. *Proc. PKDD 2005, Porto, Portugal: Springer-Verlag*, 2005, 309 - 320.
- [11] Brezany P, Hofer J, Tjoa A, *et al.* Gridminer: An Infrastructure for Data Mining on Computational Grids [C]. In *APAC Conference and Exhibition on Advanced Computing, Grid Applications and eResearch, PAC, Australia*, 2003.
- [12] Stankovski V, Swain M, Kravtsov V, *et al.* Grid-enabling Data Mining Applications with DataMiningGrid: An Architectural Perspective[J]. *Future Gener. Comput. Syst.*, 2008, 24(4): 259 - 279.
- [13] Wu X, Chen C. The Design, Development and Application of Geographical Knowledge Service Grid Portal [C]. *Proc. of 17th International Conference on Geoinformatics, Fairfax, USA*, 2009.
- [14] 林甲祥. 考虑约束条件的分布式空间离群挖掘及其应用研究[D]. 福州大学博士学位论文, 2010.
- [15] 薛安荣, 鞠时光, 何伟华, 等. 局部离群点挖掘算法研究[J]. *计算机学报*, 2007, 30(8): 1455 - 1463.
- [16] Chawla S, Sun P. SLOM: A New Measure for Local Spatial Outliers[J]. *Knowledge and Information Systems*, 2006, 9(4): 412 - 429.
- [17] 郑旻琦, 陈崇成, 樊明辉, 等. 基于 Delaunay 三角网的空间离群挖掘[J]. *微计算机应用*, 2008, 29(6): 76 - 82.
- [18] 刘丰富. 基于网格的地理空间知识服务技术与原型系统开发[D]. 福州大学硕士学位论文, 2007.

# Service and Application of Grid Based Distributed Spatial Outliers Mining

YAO Minjing<sup>1</sup>, LIN Jiaxiang<sup>1</sup>, CHEN Chongcheng<sup>1\*</sup>, MA Hengbing<sup>2</sup>

(1. *Key Lab of Spatial Data Mining and Information Sharing of MOE, Spatial Information Research Center of Fujian, Fuzhou University, Fuzhou 350002, China*; 2. *Fujian Economic Information Center, Fuzhou 350001, China* )

**Abstract:** A spatial outlier is a spatial object whose non-spatial attribute values are significantly deviated from the other data's in the dataset. The identification of spatial outliers can lead to the discovery of some unexpected knowledge, and it has a number of practical applications. There are massive spatial data maintained over geographically distributed sites in WAN. It's necessary to analyse and process the data by using the high-performance distributed parallel processing system. Grid is one of the most effective approaches to meet this requirement. The geographical knowledge grid platform (GeoKS-Grid) established by our research group is the application of knowledge grid in geo-information science, which integrate technologies of grid computing, web service, WebGIS, data mining, information visualization, knowledge base of ontology and knowledge reasoning, online analytical processing, decision analysis, data warehouse and workflow, to form a geographical problem solving environment. In this paper, a grid based distributed framework and the corresponding strategy for distributed spatial data mining system are discussed, and a distributed algorithm for spatial outlier mining is designed and implemented. In general, the process of distributed spatial outlier mining can be seen to be a series of services including atomic services and composite services. Furthermore, according to the principle of web service reuse and compositionality, the distributed spatial outlier mining algorithm is decomposed into several grid atomic services. Distributed spatial outlier mining including local spatial outlier mining and global spatial outlier mining is realized by grid workflow approach to discovery and composition of knowledge atomic grid services provided by knowledge grid. Finally, demonstration application is carried out on the basis of soil geochemistry data inspected by the Ecological Geochemistry Survey of Fujian Coastal Economic Belt, the efficiency and the validity of the distributed spatial outlier mining service and system are verified and confirmed.

**Key words:** spatial outlier; distributed data mining; knowledge grid; atomic service; service composition