

土地适宜性分析的数据处理技术框架及其构件研究

文/文琴 罗飞 曹亮

摘要

现在对土地适宜性分析的需求日益复杂化、多样化,传统的将数据与GIS技术相结合的土地适应分析方式已经满足不了系统的性能需求。本文针对现有的土地适宜性分析的需求,研究分析了一种将NetCDF存储与Spark分布式计算相结合的高性能数据处理方式。我们将数据处理的各个关键步骤分解成标准的构件,针对土地性适应分析我们将分析过程分解为了三个主要的标准构件分别为:数据存储构件是将数据用NetCDF模型进行存储,以便将多数据源数据统一存储,减少数据检索时间,提高数据的存储效率;数据插值构件是利用分布式计算框架Spark对土地适宜性分析中的数据进行并行插值(离散点数据插值为格点数据);每个格点适宜度的计算构件是利用分布式计算框架Spark并行计算每个格点的适宜性等级,以提高数据处理效率。最后以人工草地建植适宜性综合管理系统为例,给出了一个基于Spark及NetCDF构件的实例,研究结果表明将NetCDF与Spark技术应用于土地适宜性分析以及将其以构件的形式进行开发可大大提高分析的效率。

【关键词】Spark NetCDF 土地适宜性分析 构件

1 引言

由于土地适宜性分析中很多模块的分析数据是标准的数据格式,针对这些数据,我们根据以前的积累和经验建立一些标准的构件,构件可以在以后的分析中进行复用,这样很大程度上节约了重新开发所需要的时间,避免了重新开发可能会导致的错误,从而提高了软件开发的效率和软件的质量。而以构件的形式实现软件复用,又被视为是现实可行的主要方法之一。为了提高土地适宜性分析的开发效率和开发质量,本文重点研究土地适宜性分析中重要的三个过程即:数据的存储、数据的插值和每个格点适应度的计算的高性能构件。

在已有的土地适宜性分析研究中,随着土地适宜性分析数据的增多,数据资料存储技术和数据处理技术存有一定的缺陷,在这些研

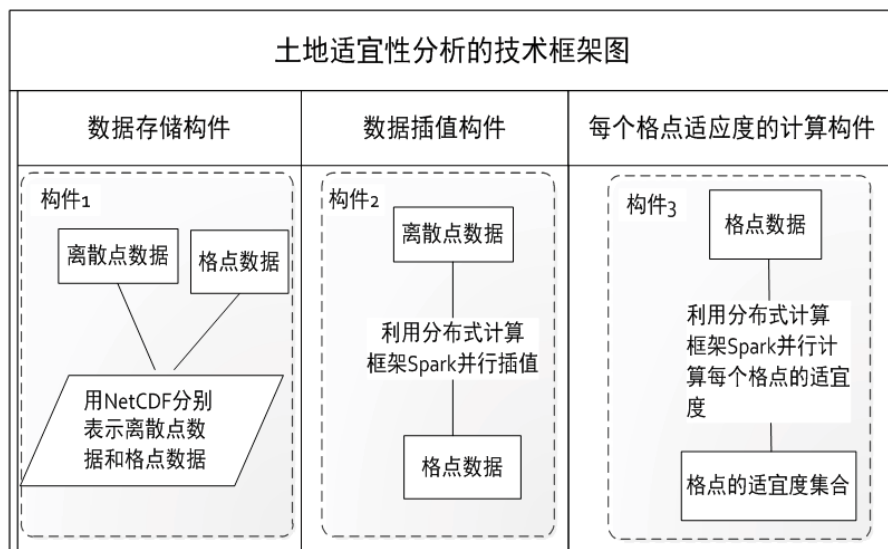


图1：基于NetCDF数据格式与分布式计算框架Spark的土地适宜性分析的技术架构

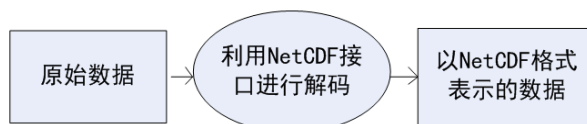


图2：数据存储构件

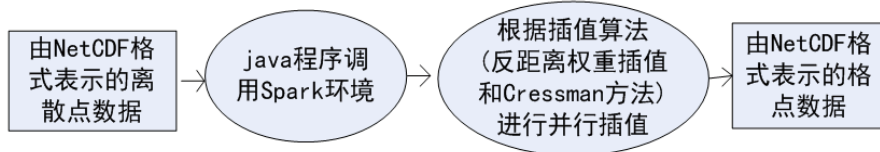


图3：数据插值构件

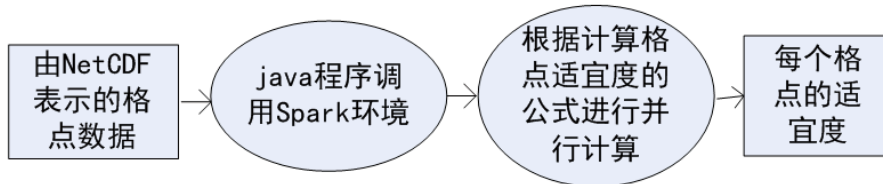


图4：每个格点适宜度的计算构件

究中基本上利用的都是GIS空间数据存储模型,对多样性数据资料进行存储,当数据量增多时,对多样性数据资料进行存取时,存取速度会成为系统性能的瓶颈。对数据进行处理时,数据的执行时间增加且数据执行的效率降低。所以土地适宜性分析的数据存储效率及数据的处理效率较低。因此本文提出将数据处理的各个关键步骤分解成标准的构件,其中在数据存储构件中,利用NetCDF数据模型来存储土地适宜性分析中的数据,以提高数据存储效率;在数

据插值构件和每个格点适宜度的计算构件中,利用分布式计算框架Spark将土地适宜性分析中的离散点数据并行插值及并行计算每个格点的适宜性等级,以提高数据的处理效率。

2 土地适宜性分析

2.1 技术架构

采用J2EE体系框架,面向SOA(Service Oriented Architecture,面向服务的体系架构),

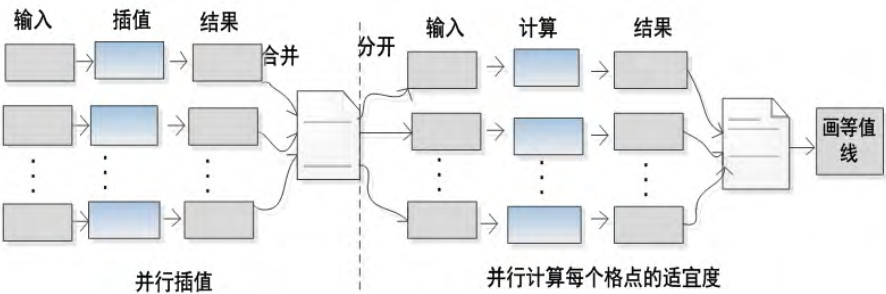


图 5: 数据处理流程图两次利用 Spark 框架的过程

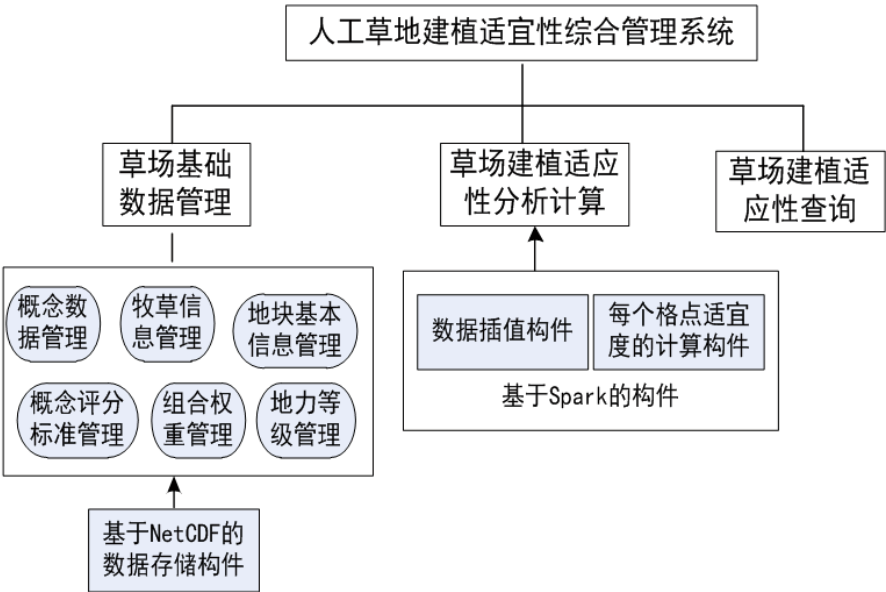


图 6: 人工草地建植适宜性综合管理系统开发模型

利用装配的形式搭建系统功能模块,实现常用构件。将土地适宜性分析系统中关键的三个过程:数据的存储、数据的插值和每个格点适宜度的计算分别做成构件。由于土地适宜性分析系统中,需要分析的数据是多样的、复杂的。同时通过对 GIS 空间数据存储模型、关系数据库模型以及 NetCDF 模型的数据存储效率进行分析,得知 NetCDF 的数据存储效率相比于其他两个存储模型具有明显的优势,因此在数据存储构件中选用 NetCDF 数据模型来表示土地适宜性分析中的数据,使用统一的格式对其进行表示。

由于分布式计算框架 Spark 是一个基于内存计算的集群计算系统,可将中间结果直接保存在内存中,无需进行 IO 操作,这加快了数据的分析。因此,在数据插值构件和每个格点适宜度的计算构件中利用分布式计算框架 Spark 对土地适宜性分析系统中的离散点数据插值为格点数据过程和每个格点适宜性等级的计算过程进行并行优化,加快土地适宜性系统的数据处理速度。基于 NetCDF 数据格式与分布式计算框架 Spark 的土地适宜性分析的技术架构如图 1 土地适宜性分析的技术框架图所示。

图 1 的数据存储过程中的离散点数据和

格点数据分别由 NetCDF 数据模型的表示如图 2 数据存储构件。数据插值过程的详细描述如图 3 数据插值构件以及图 5 中的并行插值部分所示。图 1 中每个格点适宜度的计算过程的详细分析如图 4 每个格点适宜度的计算构件及图 5 的中并行计算每个格点的适宜度所示。

2.2 数据存储构件

土地适宜性分析系统中,主要涉及到离散点数据和格点数据这两类数据,且这两类数据在进行数据分析时均利用 NetCDF 数据模型进行表示。离散点数据指的是在研究区域内分布不均匀的观测数据。格点数据指的是在研究区域内经过离散点数据插值后的,分布均匀的数据。数据存储构件的功能主要是将离散点数据和格点数据分别用 NetCDF 数据模型进行表示。构件的基本思路如下:

首先传入需要用 NetCDF 数据模型表示的离散点数据或者是格点数据,再利用 NetCDF 接口进行解码,最后得到以 NetCDF 数据模型表示的离散点数据或者格点数据,图 2 为原始数据以 NetCDF 数据格式表示的数据格式转化过程。

2.3 数据插值构件

在进行土地适宜性分析时,为了能更加准确的计算出研究区域内每个网格点的适宜度,需要将研究区域内的所有离散点数据插值为研究区域内的格点数据,该过程的数据模型转化如图 3 所示,此过程称为数据插值过程。为了提高土地适宜性分析的效率,利用分布式计算框架 Spark 对土地适宜性分析中的数据进行并行插值同时将该过程做成一个常规的构件如图 3 所示,在使用该构件之前需要搭建好该构件所需的分布式计算框架 Spark 集群环境,构件的基本思路如下:

首先传入需要插值的离散点数据,该离散点数据由 NetCDF 表示,再根据插值算法,在分布式计算框架 Spark 集群环境下将离散点数据并行插值为格点数据,其中插值算法是通过外界传入的参数或预置规则进行选择的,该土地适宜性分析系统中利用到的插值算法主要有反距离权重插值和 Cressman 方法,由于需要处理的数据空间分布的不同,反距离权重插值 (IDW) 算法提供了两个分支, IDW_Neighbour 用于空间分布连续的数据, IDW_Radius 运用于空间分布不连续的数据。插值后的中间结果可保存在内存中。

2.4 每个格点适宜度的计算构件

将离散点数据并行插值为研究区域内的格点数据后,需要计算每个格点的适宜度。本文将该过程做成了一个常规构件,在使用该构件之前需要搭建该构件所需的分布式计算框架 Spark 集群环境,利用分布式计算框架 Spark 并行计算每个格点的适宜性等级,每个格点适宜度的计算构件 (如图 4 所示) 的基本思路如下:

首先传入需要计算的格点数据,格点数据由 NetCDF 表示,再将每个格点根据事先预定好的适宜性公式,在分布式计算框架 Spark 集群环境下并行计算每个格点适宜性等级,最终得到每个格点的适宜性等级,画出等值线。整个数据处理流程如图 5 所示。

如图 5 所示,首先将计算任务进行分解,然后再将分解的任务分配到不同的处理器上,再在分布式计算框架 Spark 集群环境下对数据进行并行插值过程和并行计算每个格点适宜度过程。

3 实例研究

本节结合人工草地建植适宜性综合管理系统,给出基于 Spark 及 NetCDF 构件的土地适宜性分析研究的实例。人工草地建植适宜性综合管理系统旨在为牧民用户提供特定地段上特定牧草的适宜性分析。

人工草地建植适宜性综合管理系统的开发模型如图 6 所示。该开发模型是基于 Spark

计算机网络信息与防御方案设计

文/王阳明

摘要

针对计算机网络信息与防御方案设计进行分析,提出了计算机网络安全管理中存在的问题,其内容有:计算机网络用户信息安全意识淡薄,不注重网络安全管理,计算机网络设计缺陷等。结合这些内容,阐述了计算机网络信息的防御技术应用实践,主要有安全扫描,系统增强,学习、自适应,实施响应和黑客诱骗技术几方面内容。最后,阐述了计算机网络信息安全防御应注意的问题,内容有:合理规划,建设安全防御体系,提高计算机网络信息防毒和杀毒功能,对关键信息进行备份,对网络防护设备进行科学设置等。

【关键词】计算机网络 信息防御 安全意识 安全管理

当下,信息技术不断发展,促使计算机网络覆盖面积逐渐增大。但是,对计算机网络进行实际应用过程中,存在一些不法分子对网络信息进行恶意侵害,导致计算机网络信息安全面临一定安全隐患。这种情况下,要求计算机网络安全管理人员对这一行为进行科学防御,通过科学手段,保障相应网络信息不受侵害。

1 计算机网络安全管理中存在的问题

1.1 计算机网络用户信息安全意识淡薄

相应计算机网络用户,实施相应操作过程中,认为内部网络安全性较好,因此便放松了安全警惕。此外,用户自身安全意识淡薄,对相应密码等进行设置过程中,安全级别较低,甚至有的用户根本不设置密码。用户对移动介质进行使用过程中,并不注重安全检查,在不同计算机上进行U盘等移动介质的随意使用,有些用户在不同计算机上随意拷贝文件。还有一些用户,为了方便起见,直接对插入计算机的内网网线随意切换。这些情况的存在,均为病毒和木马的转换提供了条件,进而为不同单

位和企业等计算就网络安全带来安全性威胁。

1.2 不注重网络安全管理

企业和相应事业机关单位等,内部网络管理人员管理水平参差不齐,部分规模较小的单位,存在一些网络管理人员身兼数职,其自身经历和技术能力等的限制,导致网络管理水平不高。对本单位计算机操作系统进行安装过程中,没有及时对系统安装相应的补丁和杀毒软件等,导致计算机出现漏洞。计算机人员对相应软件进行安装过程中,没有对相应的应用软件做出安全检查,这就导致安装出现隐患,最终导致计算机网络信息安全面临威胁。此外,服务器和交换机等设备在日常运行过程中,缺乏完善的维护措施,导致网络故障出现,进而导致整个网络安全应用受到一定影响。

1.3 计算机网络设计缺陷

一些企业和机关事业单位的网络所涉及的内容有内网、外网两部分内容,部分单位存在健康网络和涉密网等多套网络,网络环境十分复杂,这些网络的建设时间不同,建设标准

<< 上接 208 页

及 NetCDF 构件方式实现的,系统分为草场基础数据管理、草场建植适宜性分析计算、草场建植适宜性询三大模块,其中草场基础数据管理中的概念数据存储利用基于 NetCDF 的数据存储构件实现,草场建植适宜性分析计算利用基于 Spark 的数据插值构件和每个格点适宜度的计算构件实现。基于此构件模型的人工草地建植适宜性综合管理系统在实际的开发和使用中发现此构件技术大大提高了软件开发的效率、缩短软件开发的周期和提高软件开发的数量。

4 结束语

为了降低软件开发成本,缩短软件开发周期,提高软件开发效率和软件质量,本文提出将土地适宜性分析中的三个关键的过程即数据的存储、数据的插值和每个格点适应度的计算分别做成三个常规的构件,同时为了优化土地适宜性分析中的数据存储技术和数据处理技术,本文利用 NetCDF 数据模型及分布式计算框架 Spark 分别对土地适宜性分析进行优化处理,以提高土地适宜性分析的效率。研究 NetCDF 数据模型和分布式计算框架 Spark 为土地适宜性分析系统的数据存储技术和数据处理技术提供较好的理论及实践基础。

参考文献

- [1] 王想红,刘纪平等.基于 NetCDF 数据模型的海洋环境数据三维可视化研究[J].测绘科学,2013:59-61.
- [2] 孙建伟,孙昭晨,陈轩等.NetCDF 格式数据的创建及应用[J].交通标准化,2010(226):31-34.
- [3] Matei Zaharia 著.Spark 大型集群上的快速和通用数据处理架构[D].加州大学伯克利分校,2014:20-24.
- [4] 陈虹君.基于 Hadoop 平台的 Spark 框架研究.[C]四川:电脑知识与技术,2014:8407-8408.
- [5] 黎文阳.大数据处理模型 Apache Spark 研究[D].现代计算机(普及版),2015:55-60.
- [6] Spark[J/OL].http://spark.apache.org.
- [7] 许鹏著.Apache Spark 源码剖析[M].电子工业出版社,2015:23-30.
- [8] 夏俊鸾,黄洁等.Spark 大数据处理技术[M].电子工业出版社,2015.
- [9] 刘文军,李靖等.基于 NetCDF 数据模型的气象资料存储设计[D].安徽农业科学,2009:1370-1371.
- [10] 李亚萍著.土地适宜性评价方法的研究[J].现代化农业,2009:30-32.
- [11] Lastovetsky A.Parallel Testing

- of Distributed Software[J]. Information and Software Technology,2009,47(10):657-662.
- [12] Parveen P,Tilley T.Towards a Distributed Execution Framework for JUnit Test Cases[C]//Proc.of IEEE International Conference on Software Maintenance. [S.l.]: IEEE Press,2009: 425-428.

作者简介

文琴(1991-),女,四川省成都市人。研究生学历。现为宿迁市科技局局长、高级工程师。主要研究方向为气象信息化技术。
罗飞(1977-),女,四川省成都市人。现为成都信息工程大学软件工程学院副教授。主要研究方向为气象信息化技术、数据集成与可视化研究方向的研究。
曹亮(1985-),男,四川省成都市人。现为成都信息工程大学软件工程学院讲师。研究方向为数据库技术、数据挖掘技术。

作者单位

成都信息工程大学软件工程学院 四川省成都市 610225