地图兴趣点分布式空间分析服务

郭庆胜[1]，王　勇[1,2]，蓝振家[1]，周贺杰[1]，刘纪平[2]，纪莹莹[3]
（1. 武汉大学 资源与环境科学学院，武汉　430079；2. 中国测绘科学研究院，北京　100830；
3. 北京四维图新科技股份有限公司，北京　100102）

摘　要：针对互联网地图上海量兴趣点的应用分析需要提高效率的问题，该文利用 MongoDB 设计并搭建了一个分布式集群，对这些互联网兴趣点数据进行了储存；然后通过 MapReduce 机制改进并实现了适用于海量兴趣点数据的空间同位模式挖掘的 Apriori 算法和几个常用的空间分布特征值计算方法；最后依据开放地理信息系统协会的 Web 处理服务规范，设计并实现了一个互联网兴趣点分布式分析服务实验系统。该文所提出的改进后的算法在数据吞吐量和计算效率上有优越性，且计算效率比传统空间分析工具和传统 Apriori 算法有所提高。

关键词：数据库；地图兴趣点；分布式计算；空间同位模式
【中图分类号】 P211　　**【文献标志码】** A　　**【文章编号】** 1009-2307(2018)01-0089-04
DOI：10.16251/j.cnki.1009-2307.2018.01.016

Research on distributed spatial analysis service of map interest points

Abstract: Aiming at the problem of the application analysis of the mass interest points on the internet map, this paper uses MongoDB to design and build a distributed cluster, and stores these internet interest point data. Then, through the MapReduce mechanism to improve and implement the spatial co-location pattern mining Apriori algorithm which is applicable to the mass of interest point data, and several common spatial distribution eigenvalues calculation methods. Finally, according to the web processing service specification of the open geographic information system association, an experimental system of internet interest point distributed analysis service is designed and implemented. The improved algorithm is superior in data throughput and efficiency, and the computational efficiency is better than traditional spatial analysis tools and traditional Apriori algorithm。

Keywords: data base; map interest points; distributed computing; spatial co-location pattern

GUO Qingsheng[1], *WANG Yong*[1,2], *LAN Zhenjia*[1], *ZHOU Hejie*[1], *LIU Jiping*[2], *JI Yingying*[3]
(1. School of Resource and Environment Science, Wuhan University, Wuhan 430079, China; 2. Chinese Academy of Surveying and Mapping, Beijing 100830, China; 3. Beijing NavInfo Polytron Technologies Inc, Beijing 100102, China)

0　引言

随着大数据时代的来临，大规模的分布式集群架构、地理信息深加工和地球空间信息服务大众化已成为地理信息领域的研究热点[1-2]。传统的关系型空间数据库性能瓶颈日益凸显，主要表现在水平扩展能力差，并发访问速度低等缺点。文献［3］在 1998 年首次提出了 NoSQL 的概念，随后，很多公司推出了自己的 NoSQL 数据库产品。在这些 NoSQL 数据库中，以 MongoDB 数据库对于点要素的支持性能最好。近年来，国内外的众多学者也开始对 MongoDB 的空间应用进行了研究[4-5]。以开放地理信息联盟 OGC 所制定的 WPS 网络处理服务规范也被越来越多地应用到了地理信息相关的行业中[6-10]。

互联网兴趣点(points of interest，POI)数据的应用已经受到越来越多的学者关注，其中包括 POI

作者简介：郭庆胜（1965—），男，湖北阳新人，教授，博士，主要研究方向为地理信息智能化处理和地图自动综合。
E-mail：guoqingsheng@whu.edu.cn

收稿日期：2016-05-18
基金项目：国家自然科学基金项目(41471384)；国家“863”计划项目(2013AA12A403)

分类标准的建立，POI的空间特征分析等[11-12]。为了满足海量互联网POI数据的空间分析的需要，本文试图利用MongoDB数据库，结合OGC的WPS规范，实现一个针对互联网POI数据的分布式空间分析服务系统，并充分利用分布式计算和并行计算技术改进一些基础空间分析方法，提高计算效率。

1 基于MongoDB的POI数据存储

在MongoDB中，数据均以文档的形式存储。一个文档就是一个JSON格式的数据结构。文档即对象，它是任意的，甚至有复杂的嵌套层次。POI数据作为最简单的空间数据，在MongoDB中以GeoJSON的格式进行存储[13]。GeoJSON是JSON格式的扩展，作为地理空间信息数据交换格式被广泛地应用于GIS中。一般来说，POI数据包含有4个方面的信息，即POI名称、坐标位置(经纬度)、详细描述和类别。其中，坐标位置信息是必须要有的。GeoJSON格式可以包含这些信息。

分片(sharding)是MongDB能够进行横向扩展的重要体现。对于一个集合来说，如果使用了分片技术，那么其数据就会以数据块(chunk)的形式由MongoDB内置的均衡器(balancer)分配到各个节点中。所有数据存储在主分片的主节点上，MongoDB还需要依据片键(shard key)对庞大的数据集进行分片操作。片键的选择一方面影响读写负载的平衡，另一方面影响数据块的大小。因此，这里使用POI的“ID”属性的哈希值作为片键。以哈希值作为片键可使数据块在各个分片节上点较均匀分布，易于读写负载的均衡，并且能够保证每个文档都有不同的片键，因此数据块能够很精细。通过命令开启对整个数据集合的分片操作后，MongoDB中的均衡器会自动地将数据块移动到各个分片上，最后达到各个分片的平衡。

在完成数据分片后，还需要对入库的POI数据建立空间索引。MongoDB原生支持GeoHash结构的球面空间索引2dsphere，因此，可在数据入库后进入路由服务器，通过ensureIndex命令建立。

2 基于MongoDB的POI数据分布式计算

2.1 MongoDB的MapReduce机制

MongoDB内部提供了一个MapReduce机制。不同于Hadoop，MongoDB的MapReduce仅能用于实现数据的统计，类似于传统关系型数据库中的Group By操作。集群中的各个节点先通过Map函数将拥有同一ID号的数据进行分组映射，将分组后的键值对分配给集群中的各个节点；再通过Reduce函数进行累加聚合；最后将结果合并生成新的键值对。MapReduce编程模型充分利用了集群分布式的特性，使得集群在进行大规模数据计算时能够有很高的效率。

2.2 POI数据分析算法的MapReduce设计

MongoDB的MapReduce计算功能有限，POI数据分析中需要灵活应用这些功能，这里以计算海量POI的平均中心算法为例说明应用方法，其Map函数伪代码如下。

function() {emit(分组键， {count：1，totalX：X坐标，totalY：Y坐标});}

Map函数将具有相同指定键值的数据分在了一起，指定的分组键可以是ID，也可以是其他任意属性。通过emit函数将存有POI数据的X、Y坐标的total X和total Y变量以及计数的count变量发射给Reduce函数。Reduce函数将Map函数发射过来的键值对进行累加求和。因为对于同一个键值，Reduce函数可能会被多次调用，所以Reduce函数所返回的对象result结构应和Map函数所发射的对象结构一致。最后还需要调用Finalize函数将Reduce函数聚合的结果进行均值化处理。Finalize函数处理后的结果将以新的键值对形式返回。

为了进一步说明基于MapReduce的POI数据分析算法的设计，这里以空间同位模式挖掘的Apriori算法为例。本文提出基于MapReduce和多线程并行处理的空间同位模式挖掘算法，实现大规模互联网POI数据同位模式和同位规则挖掘。该算法的基本思想为：利用MapReduce并行编程模型特性对原始数据集进行Map分块，处理生成分块的数据集，主进程将分块的数据集分配到Hadoop集群中的每台计算机上，然后在每台计算机上对分块的数据集进行多线程处理。在计算项目集支持度计数时可以利用多线程并行计算的方法来实现，基本思想是：在获取k-项集时，主程序不直接计算项集的个数，而是把计算的任务分配给多个线程，在不同线程中计算，最后把各个线程的计算结果返回给主程序。该算法的计算过程如下：①利用MongoDB实现大规模地理数据的存储；②建立空间索引。根据用户设置的邻近距离，利用MongoDB的空间查询功能获取同位模式挖掘

的实例集；③根据用户设置的支持度、MapReduce和多线程并行处理技术，利用 Apriori 算法生成同位模式；④根据用户设置的置信度，从同位模式中利用 Apriori 算法生成关联规则的方法生成同位规则；⑤对生成的同位模式和同位规则进行分析与评价。

3　POI 数据分析服务的设计

Web 服务(web service)是面向服务架构(service-oriented architecture，SOA)中的一种技术。地理信息服务是 Web 服务在地理信息领域内的应用。最新提出的 WPS(web processing service)网络处理服务有 3 个接口，它们分别是 GetCapabilities、DescribeProcess 和 Execute。GetCapabitities 返回服务级元数据；DescribeProcess 返回指定的一个或多个算法的详细描述信息，包含输入、输出参数和格式等；Execute 返回指定执行算法的处理结果。

WPS 规范并没有具体过程的实现，它只是制定了一套接口的标准，开发者需要依据这套标准去实现其所描述的框架。与传统的 Web 服务类似，一个完整的 WPS 服务流程也有发布、发现、绑定、调用等操作，同时也需要具备 3 个要素：客户端、实现 WPS 的服务器、实现 WPS 接口的分析算法。一个完整的分析服务流程，分为 4 步：①算法的注册发布。各个分析的算法首先要注册发布到 WPS 服务器上。算法的注册实际上就是实现 WPS 规范里的 DescribeProcess 和 Execute 两个接口的过程。②WPS 服务器登记和客户端发现，算法发布好后，WPS 规范中 GetCapabilities 接口的 WPS 服务器会将注册的算法加入到节点列表中。③客户端绑定算法。在查询了所有 WPS 服务器支持的算法后，客户端可以根据任务的需要选择一个算法。通过 DescribeProcess 请求查询该算法的相关参数信息，实现与算法的绑定。④客户端调用算法。在按照 DescribeProcess 响应返回的算法描述文档，编辑好算法所需的参数后，客户端通过 Execute 请求将相关参数编码成 XML 并以 POST 方式传递到算法的逻辑模块进行计算处理，最终客户端会得到一个计算的返回结果。

4　实验与分析

本文以 B/S 三层架构设计了一个互联网 POI 数据在线分析的实验系统，算法计算结束后将结果以符合 WPS 规范的 XML 编码的形式返回给客户端，如图 1 所示。考虑到实验室里计算机数量的限制，本实验采用 5 台数据服务器、1 台分析服务器和 1 台客户机。MongDB 集群使用了 5 台内存为 4 GB 双核的 64 位 PC 机；WPS 服务器使用了 1 台内存为 2 GB 的双核的 32 位 PC 机；客户机与之相同。客户端的开发使用了 OpenLayers 3.0 开发包，其主要功能是显示底图数据和分析结果，以及生成分析参数的面板。WPS 服务器采用的是第三方开源类库 WPS.NET。WPS.NET 是基于 .NET 平台，按照 OGC WPS 1.0.0 规范开发的 Web Service 类库，可以部署到 IIS 上；MongoDB 数据库采用的版本是 2.6.3 版，由于需要在算法中操作数据库，因此软件方面还用到了 MongoDB 提供的官方 C＃驱动。

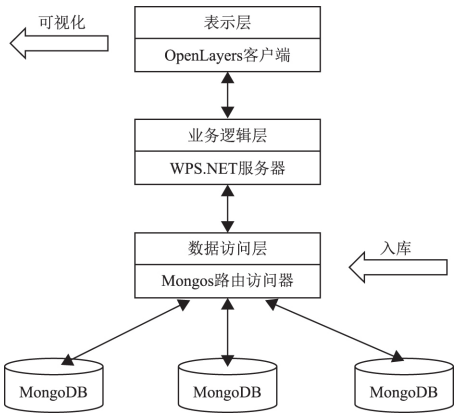

图 1　POI 数据分析服务的系统架构

Fig 1　System Architecture of POI Data Analysis Service

为了验证 POI 数据分析中采用了分布式技术和并行计算技术后的优势，本实验将传统 ArcToolBox 分析工具(ArcGIS 10.0 桌面平台上)和传统的 Aprior 算法进行了对比分析。

实验 1 使用了“863”项目组提供的西藏(334 160 条数据)和福建(403 739 条数据)两个数据集。考虑到计算机不同运行状态对计算耗时的影响，实验中进行了多种不同的计算，将平均时间作为最后的结果，实验结果如表 1 所示。从表 1 可以看出，当数据量超过 40 万条时，ArcToolbox 工具已出现内存溢出错误，而实验系统却没有这个瓶颈。

实验 2 选取了“863”项目组提供的百万条 POI 数据对本文改进的 Apriori 算法进行验证。传统 Apriori 算法[14]与改进后的 Apriori 算法在计算 1-项集支持度计数时的运行时间对比如图 2 所示。由实验 1 和实验 2 表明，采用分布式技术的 POI 数据分析系统在数据吞吐量和计算效率上具有明显优势。

表 1 实验 1 的计算耗时统计

Tab. 1 Computational Time Statistics of Experiment 1

名称	西藏数据计算耗时/s		福建数据计算耗时/s	
	ArcToolBox 工具	分布式分析系统	ArcToolBox 工具	分布式分析系统
平均中心	53.58	6.42	MemoryError	5.67
中位数中心	112.80	58.22	MemoryError	47.73
标准距离	45.12	10.44	MemoryError	12.41
标准差椭圆	49.18	15.40	MemoryError	18.98

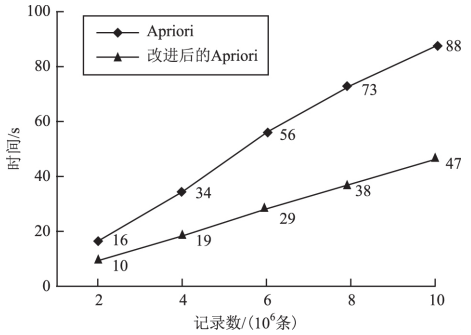

图 2 实验 2 的运行时间对比

Fig. 2 Comparison of Runtime of Experiment 2

5 结束语

互联网 POI 数据与其他传统的大数据一样，因其体量巨大，越来越需要扩展能力强的数据储存系统和计算效率高的处理方式。以 MongoDB 为代表的 NoSQL 数据库因其水平扩展能力强、数据吞吐量大、并发访问速度快等特点而备受人们青睐。本文以空间统计中的 4 种方法以及同位模式挖掘中的 Apriori 算法为例，研究了基于 MongoDB 的 POI 数据存储、POI 数据分布式分析计算以及基于 OGC 的 WPS 规范的 POI 分析服务。结合这 3 种技术，利用 B/S 架构设计并实现了互联网 POI 数据的分布式分析服务系统。通过和 ArcToolBox 工具以及传统 Apriori 算法进行对比实验发现，本文所设计开发的 POI 数据分布式分析服务实验软件取得了良好效果。

参考文献

[1] 王家耀. 地图制图学与地理信息工程学科发展趋势[J]. 测绘学报，2010，39(2)：115-119.（WANG Jiayao. Development trends of cartography and geographic information engineering[J]. Acta Geodaetica et Cartographica Sinica，2010，39(2)：115-119.）

[2] 李德仁. 多学科交叉中的大测绘科学[J]. 测绘学报，2007，36(4)：363-365.（LI Deren. On geomatics in multi-discipline integration[J]. Acta Geodaetica et Cartographica Sinica，2007，36(4)：363-365.）

[3] 申德荣，于戈，王习特，等. 支持大数据管理的 NoSQL 系统研究综述[J]. 软件学报，2013，24(8)：1786-1803.（SHENG Derong，YU Ge，WANG Xite，et al. Survey on NoSQL for management of big data[J]. Journal of Software，2013，24(8)：1786-1803.）

[4] 郜庆林，陈柏垒. 使用文档型数据库 MongoDB 深度挖掘天气现象数据内在价值[J]. 计算机与网络，2012(16)：67-70.（GAO Qinglin，CHEN Bokun. Making a full use of meteorological data by document database mongoDB[J]. Computer & Network，2012，(16)：67-70.）

[5] 陈超，王亮，闫浩文，等. 一种基于 NoSQL 的地图瓦片数据存储技术[J]. 测绘科学，2013，38(1)：142-143，159.（CHEN Chao，WANG Liang，YAN Haowen，et al. A map titles data storage technology based on NoSQL[J]. Science of Surveying and Mapping，2013，38(1)：142-143，159.）

[6] 张登荣，俞乐，邓超，等. 基于 OGC WPS 的 Web 环境遥感图像处理技术研究[J]. 浙江大学学报(工学版)，2008，42(7)：1184-1188.（ZHANG Dengrong，YU Le，DENG Chao，et al. OGC WPS-based remote sending image processing in web environment[J]. Journal of Zhejiang University (Engineering Science)，2008，42(7)：1184-1188.）

[7] 宋全红. 基于 OGC WPS 标准的空间统计 PSE 研究及 PSE-SDBI 实现[D]. 阜新：辽宁工程技术大学，2009.（SONG Quanhong. The Study on PSE for spatial statistics and realization of PSE-SDBI based on OGC WPS[D]. Fuxin：Liaoning Technical University，2009.）

[8] 吴楠，何洪林，张黎，等. 基于 OGC WPS 的碳循环模型服务平台的设计与实现[J]. 地球信息科学学报，2012，14(3)：320-326.（WU Nan，HE Honglin，ZHANG Li，et al. Design and implement an online carbon cycle model service platform based on OGC Web processing service[J]. Journal of Geo-information Science，2012，14(3)：320-326.）

[9] 佘星星，兰茹，覃燕. 基于 WFS 和 WPS 服务的气象信息 WebGIS 动态显示技术实现[J]. 农业网络信息，2014(12)：58-60.（SHE Xingxing，LAN Ru，QIN Yan. The implementation on the WebGIS dynamic display of meteorology information using WFS and WPS service[J]. Acriculture Network Information，2014，(12)：58-60.）

[10] SCHÄFFER B，BARANSKI B，FOERSTER T. Towards spatial data infrastructures in the clouds[M]. Springer Berlin Heidelberg：Geospatial Thinking，2010：399-418.

（下转第 100 页）

[4] HELD R T,COOPER E A,O'BRIEN J F,et al. Using blur to affect perceived distance and size[J]. ACM transactions on graphics,2010,29(2):19.

[5] 冯文灏.近景摄影测量:物体外形与运动状态的摄影法测定[M].武汉:武汉大学出版社,2002:116-142.(FENG Wenhao. Close range photogrammetry:shape and motion measurement[M]. WuHan:Wuhan University Press,2002:116-142.)

[6] 张永军,张祖勋,张剑清.利用二维 DLT 及光束法平差进行数字摄像机标定[J].武汉大学学报(信息科学版),2002,27(6):566-571.(ZHANG Yongjun, ZHANG Zuxun,ZHANG Jianqing. Camera calibration using 2D-DLT and bundle adjustment with planar scenes[J]. Editoral Board of Geomatics and Information Science of Wuhan University,2002,27(6):566-571.)

[7] 孙韬,陈晓宁,李莹.基于 DLT 的数字天顶摄影测量仪相机检校[J].测绘科学,2011,36(2):146-148.(SUN Tao,CHEN Xiaoning,LI Ying. The DLT-based camera calibration of digital zenith photogrammetry[J]. Science of Surveying and Mapping,2011,36(2):146-148.)

[8] 冯文灏,商浩亮,侯文广.影像的数字畸变模型[J].武汉大学学报(信息科学版),2006,2(3):99-103.(FENG Wenhao,SHANG Haoliang,HOU Wenguang. A digital distortion model for all kinds of imaging systems [J]. Editoral Board of Geomatics and Information Science of Wuhan University,2006,2(3):99-103.)

[9] ANONYMOUS. Tilt-Shift photography [EB/OL]. (2016-06-06)[216-07-07]. https://en. wikipedia. org/wiki/Tilt-shift_photography.

[10] ZEITLER W,DOERSTEL C,JACOBSEN K. Geometric calibration of the DMC:method and results[J]. International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences,2002,34(1):324-332.

[11] 李健,刘先林,刘凤德,等. SWDC-4 大面阵数码航空相机拼接模型与立体测图精度分析[J].测绘科学,2008,33(2):104-106.(LI Jian, LIU Xianlin, LIU Fengde,et al. Mosaic model of SWDC-4 large format aerial digital camera and accuracy analysis of stereo mapping[J]. Science of Surveying and Mapping,2008,33(2):104-106.)

[12] GREEN K,TUKMAN M,FINKBEINER M. Comparison of DMC,Ultra Cam,and ADS40 imagery for benthic habitat and propeller scar mapping[J]. Photogrammetric Engineering & Remote Sensing,2011,77(6):589-599.

[13] 卫征.多模态 CCD 相机系统(MADC)构像方式和数据处理研究[D].北京:中国科学院遥感所,2006:136-145.(WEI Zheng. Imaging mode and data processing study of multi-mode aerial digital camera(MADC)[D]. Beijng:Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences. 2006:136-145.)

[14] 张建霞,王留召,刘先林,等.数字航空摄影测量的相机检校[J].测绘通报,2005(11):44-65.(ZHANG Jianxia,WANG Liuzhao,LIU Xianlin,et al. Camera calibration in aerial digital photogrammetry[J]. Bulletin of Surveying and Mapping,2005(11):44-65.)

[15] SUN T,FANG J,ZHAO D,et al. A novel multi-digital camera system based on Tilt-Shift photography technology[J]. Sensors,2015,15(4):7823-7843.

[16] TSAI R Y. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses[J]. Robotics and Automation,IEEE Journal,1987,3(4):323-344.

(上接第 92 页)

[11] 禹文豪,艾廷华.核密度估计法支持下的网络空间 POI 点可视化与分析[J].测绘学报,2015,44(1):82-90.(YU Wenhao,AI Tinghua. The visualization and analysis of POI features under network space supported by kernel density estimation[J]. Acta Geodaetica et Cartographica Sinica,2015,44(1):82-90.)

[12] 张志杰,彭文祥,周艺彪,等.空间点模式分析中离散趋势的描述研究及应用[J].中国卫生统计,2008,25(5):470-473.(ZHANG Zhijie, PENG Wenxiang, ZHOU Yibiao,et al. Study and application on the statistical indices to describe the tendency of dispersion in the spatial point pattern analysis [J]. Chinese Journal of Health Statistics,2008,25(5):470-473.)

[13] SHEKHAR S,CHAWIS S. 空间数据库[M].谢昆青,马修军,杨冬青,译.北京:机械工业出版社,2004.(SHEKHAR S, CHAWIS S. Spatial Database[M]. XIE Kunqing, MA Xiujun, YANG Dongqing, translated. Beijing:China Machine Press,2004.)

[14] 边馥苓,万幼. K-邻近空间关系下的空间同位模式挖掘算法[J].武汉大学学报(信息科学版),2009,34(3):331-334.(BIAN Fuling,WAN You. A novel spatial co-location pattern mining algorithm based on K-nearest feature relationship [J]. Geomatics and Information Science of Wuhan University,2009,34(3):331-334.)