Journal of System Simulation

Volume 30 | Issue 3 Article 9

1-2-2019

# Association Rules Analysis Method of Spatial Data Under MapReduce Framework

Mingzhi Zhang
*1.The Department of Information Operation & Command Training, National Defense University, Beijing 100091, China; ;*

Li Yi
*1.The Department of Information Operation & Command Training, National Defense University, Beijing 100091, China; ;2.The Forces of 66194, Beijing 100012, China;*

# Association Rules Analysis Method of Spatial Data Under MapReduce Framework


## Abstract

Abstract: Spatial data has the characteristic of extensity, timeliness, multidimensional, large amount of data and complex relations. Some non-conventional data screening tool for analysis and mining is required to find out the patterns, rules and characteristics knowledge in the spatial big data for battlefield situation awareness and battle space management. In view that *the existing Apriori algorithm scans the database too frequently, the Apriori algorithm is improved on the basis of working principle of Map Reduce .The fast analysis ideas and technologyframework of spatial data is proposed. An elementary validate prototype is built for the key technology experimentation.*Experimental results show that, the technical route and framework can improve the speed of massive spatial data analysis and processing.



## Keywords

spatial data, big data, association rules, analysis method, parallel computing

第 30 卷第 3 期 系统仿真学报© Vol. 30 No. 3
2018 年 3 月 Journal of System Simulation Mar., 2018

# Map Reduce 框架下空间大数据的关联规则分析方法

张明智[1]，李义[1,2]
(1. 国防大学信息作战与指挥训练教研部，北京 100091；2. 66194 部队，北京 100012)

**摘要：**空间数据具有空间性、时间性、多维性、大数据量、空间关系复杂等特点，如何在空间大数据里面寻找模式、规律和特征知识，用于战场态势感知和战场空间管理，这需要利用和开发一些非传统的数据筛选工具进行分析、挖掘。*针对现有 Apriori 算法扫描数据库过于频繁的问题，结合 Map Reduce 的工作原理，对 Apriori 算法进行了改进，提出基于 Map Reduce 的空间大数据快速分析思路和技术框架，初步搭建了验证原型，对关键技术进行实验。*实例验证结果表明，利用该技术路线和框架，可以提升海量空间数据的处理分析速度。

**关键词：**空间数据；大数据；关联规则；分析方法；并行计算
中图分类号：TP393 文献标识码：A 文章编号：1004-731X (2018) 03-0840-06
DOI: 10.16182/j.issn1004731x.joss.201803009

## Association Rules Analysis Method of Spatial Data Under MapReduce Framework

*Zhang Mingzhi*[1], *Li Yi*[1,2]
(1. The Department of Information Operation & Command Training, National Defense University, Beijing 100091, China;
2. The Forces of 66194, Beijing 100012, China)

**Abstract:** Spatial data has the characteristic of extensity, timeliness, multidimensional, large amount of data and complex relations. Some non-conventional data screening tool for analysis and mining is required to find out the patterns, rules and characteristics knowledge in the spatial big data for battlefield situation awareness and battle space management. In view that *the existing Apriori algorithm scans the database too frequently, the Apriori algorithm is improved on the basis of working principle of Map Reduce .The fast analysis ideas and technologyframework of spatial data is proposed. An elementary validate prototype is built for the key technology experimentation.*Experimental results show that, the technical route and framework can improve the speed of massive spatial data analysis and processing.

**Keywords:** spatial data; big data; association rules; analysis method; parallel computing

## 引言

空间大数据描述的挑战来自于空间大数据来源的多样性和异构性，如何实现 TB 级空间数据的快速分析，是当前海量空间数据分析面临的重要问题[1-3]。目前学术界对大数据的管理和分析进行了深入研究，企业界也提出和研发了多类创新技术。微软提出了 Dryad[4]的数据处理模型，用来构建支持有向无环图类型数据流的并行程序；Google 针对文件类大数据应用需求，提出了 Dreme[5]计算模型，设计开发了 GFS(Google file system)[6]分布式文件系统和 Big table[7]数据库；Facebook 针对海量小文件的文件系统，推出了 Haystack[8]，通过多个逻辑文件共享同一个物理文件、增加缓存层、部分元

收稿日期：2016-04-01 修回日期：2016-06-11；
基金项目：国家自然科学基金(61174156, 61273189)，中国博士后科学基金特别资助(2017T100791)；
作者简介：张明智(1962-)，男，陕西，博士后，教授，博导，研究方向为战争复杂系统建模与仿真、数据与规则；李义(1981-)，男，河北，博士后，助工，研究方向为数据与规则、计算机战争模拟。



 

数据加载到内存等方式有效地实现了海量图片的存储；Amazon 针对大数据时代的数据存储，推出了 Dynamo[9]，通过综合键/值存储、改进的分布式哈希表、向量时钟等技术实现了一个完全分布式、去中心化的高可用系统;开源界提出了类似Google的 Hadoop[10]技术框架体系，并在各个行业得到广泛应用。虽然目前已经形成了大量的大数据存储、管理和分析技术，但是这些技术大多数是针对文本数据、属性数据、图像图形数据的，在空间大数据的处理方面往往不能够直接应用。本文从空间大数据的特点出发，结合空间数据分析的具体需求，提出空间大数据关联规则分析思路，并基于此思路构建了基于 Hadoop 框架的验证原型，对关键技术进行仿真实验，验证了所提方法的有效性。

# 1 概述

## 1.1 空间数据

空间数据描述信息世界中的空间对象在现实世界内的具体地理方位和空间分布，包括空间实体的属性、数量、位置及其相互关系等，涵盖从宏观、中观到微观的整个层次，可以是点的高程、道路的长度、多边形的面积、建筑物的体积、像元的灰度等数值，地名、注记等字符串，图形、图像等多媒体成分，空间关系等拓扑结构[11-12]。与一般的数据相比，空间数据具有空间性、时间性、多维性、大数据量、空间关系复杂等特点[13]。

当今的空间大数据，其内容非常宽泛(涵盖与空间位置、属性等相关的一切信息)，形式纷繁(包括遥感图像、GPS 定位数据、数字地图、数字高程模型、三维地理模型、主题图等)。这些数据分布广泛(跨地域、跨行业、跨部门)、格式异构(有非空间数据，也有空间数据，空间数据又分矢量数据与栅格数据)，共享极为困难，利用率很低，无法得到有效的统一综合应用。如何自动化、智能化地在空间大数据里面寻找模式、规律和特征知识，用于战场态势感知和战场空间管理，并协助指挥员和指挥机关做出更好的决策，这就需要利用和开发一些非传统的数据筛选工具进行分析、挖掘，最终提供有用的数据洞察为决策者所用。应用事件搜索算法，可以得到关联规则挖掘所需要的事件项集，进而可以挖掘空间数据相关参数间的因果规律。

## 1.2 关联规则

关联规则是描述数据库中数据项之间存在的潜在关系的规则，即根据一个事物中某些项的出现，可推导出另一些项在同一事物中也出现。关联规则的挖掘问题可形式化描述如下，设 $I=\{i_1,i_2,\cdots i_m\}$ 是项的集合。$D$ 是数据库所有事务的集合，其中每个事务 $T$ 是 $I$ 中一组项目的集合，即 $T\subseteq I$，每个事务有唯一的标识符 $TID$。关联规则在 $D$ 中的支持度(Support)是 $D$ 中事务同时包含 $A$、$B$ 的百分比，即 $P(A\cup B)$；置信度(Confidence)是 $D$ 中事务已经包含 $A$ 的情况下，包含 $B$ 的百分比，即 $P(B|A)$。如果满足最小支持度阈值和最小置信度阈值，则认为关联规则是有趣的[14]。关联规则就是支持度和信任度分别满足用户给定阈值的规则。

关联规则的挖掘分为两个步骤：

(1)找出所有满足最小支持度的频繁项集；

(2)由频繁项集产生满足最少支持度和最少置信度的强关联规则。

## 1.3 Apriori 算法

Apriori 算法是最具影响的一类关联规则挖掘算法，它运用逐层搜索迭代方式来通过 $K$ 项集进行($K$+1)项集的探索，需对数据集进行一次扫描，进而生成频繁 1–项集 $L_1$，利用 $L_1$ 进行频繁项集 $L_2$ 的探索，以不断迭代的方式持续到频繁项集为空集。这种算法的不足之处是需要多次扫描数据库才可探索出所有的频繁项集，显然具有海量数据的空间数据应用并不适合这一算法，多次扫描会耗损大量内存及时间。因此，本文借鉴 Hadoop 分布式并行计算性质，将该性质移植在 Apriori 算法上，建立 Map Reduce 架构以存储扫描数据库，查找频繁项集所获得的关联规则，扫描处理将在各个 Date



https://dc-china-simulation.researchcommons.org/journal/vol30/iss3/9
DOI: 10.16182/j.issn1004731x.joss.201803009

第 30 卷第 3 期 系统仿真学报 Vol. 30 No. 3
2018 年 3 月 Journal of System Simulation Mar., 2018

Node 节点中并行操作，由此获得各计算节点上的局部频繁项集。利用 Master 将实际的全局的支持度、频繁项集统计与确定出来，以此来节省系统的时间与内存消耗，实现数据挖掘效率的大大提高。

### 1.4 MapReduce 研究现状

2004 年 Google 研究员 Jeffrey Dean 和 Sanjey Ghemawat 通过对网页数据存储和并行分析处理研究后，提出 Map Reduce 计算模型，为大数据分析处理问题提供了一个新的有效解决方法和途径。自 MapReduce 发表以来，业界和学术界掀起了 MapReduce 的研究热潮。Tyson 等对数据传输进行了改进，使 MapReduce 内部数据流水线传输，提高了系统的可用性；Jaliya 等利用 MapReduce 对高能物理数据进行分析，实现了基于 MapReduce 的 Kmeans 聚类，提出了数据流式处理模型，有利于基于 MapReduce 的数据密集型科学计算；Michal 等在语义标注解决方案 Ontea 的基础上，引入 MapReduce 架构对大规模网页进行自动语义标注，实验取得很好的性能表现；Abouzeid 等提出的 HadoopDB 通过结合 MapReduce 和并行数据库系统用于大规模数据处理，利用 MapReduce 的可扩展性、容错性和灵活性的优势，再利用并行数据库性能和效率上的优势；在数据挖掘方面，Liu 等提出了基于 MapReduce 的 BP 神经网络 MBNN 用来处理大规模移动设备数据的聚类，还引入 Adaboosting 机制改善聚类性能；White 等把 MapReduce 用于图像处理，包括分类器训练、背景提取和图像配准等，为 MapReduce 用于可视化数据处理奠定了基础；复旦大学把 MapReduce 用于科学计算，并对科学计算中遇到的上些典型问题提出相应的优化方案；中国科学技术大学提出的 HPMR(High Performance MapReduce)是专为高性能计算，尤其是并行科学计算而设计开发的，通过 MapReduce 模型隐藏了与并行有关的繁琐技术细节，程序员只需编写计算代码，所有与数据组织、任务并行和通信有关的功能都交给 HPMR 的运行时系统去完成；浙江大学和中山大学等对基于 MapReduce 的数据挖掘算法进行了研究；上海交通大学和北京邮电大学等对 MapReduce 框架进行了改进，实现 MapReduce 框架的通用性或用于某一方面的数据处理。

## 2 基于 Map Reduce 的 Apriori 算法

Map Reduce 的工作原理是将大规模数据集分解为许多小数据集，集群中每个节点处理一个小数据集，主节点进行控制和管理集群中的所有节点，集群中每个节点都完成任务后，主节点把输出结果汇总起来。所有节点每隔一段时间通过心跳向主节点汇报自己的工作状态，主节点接收这些状态信息并进行记录，如果发现某一节点发生故障，主节点会将该节点上的任务分配给其它正常的节点重新执行，如图 1 所示。

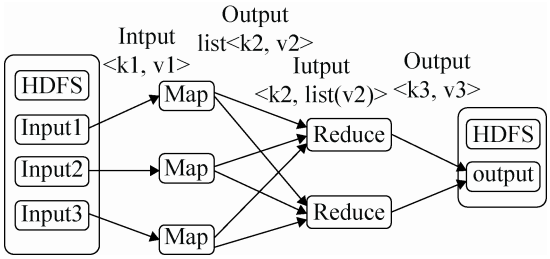


图 1 Map Reduce 工作原理示意图
Fig. 1 The working principle diagram of Map Reduce

依据以上工作原理，基于 Map Reduce 的 Apriori 算法的具体实现流程如下：

(1) 采用 Map Reduce 思想，依据参数将输入文件分割成大小在 16~64M 范围的 M 块；用户设置关联规则需要的最小支持度、置信度，请求挖掘服务；

(2) 执行程序主要包括主控程序 Master、分工作机 Worker，其中 Map 操作有 M 个，Reduce 操作有 R 个，空闲 Worker 会接收到 Master 分配的 Map 或 Reduce 处理任务；接收到用户请求的 Master 需向主存储节点(NameNode)申请相关的 PML 文件，对空闲节点列表进行访问，分配任务给空闲的计算存储节点(DataNode)，将各个 DataNode 所需的存储算法节点的算法进行调度与并行处理；



 

(3) Worker在处理Map任务时会对处理数据进行读取，将每个计算存储节点(DataNode)利用 Map 函数进行<Key, Value>对映射与新键值的处理，生成一个局部候选频繁 $K$ 项集，用 $C_k^n$ 来表示，每一 $C_k^n$ 的支持度用 1 表示；而后 Worker 将<Key, Value>传送给 Map 函数并产生中间结果，将其缓存在内存中，定时向本地硬盘传送缓存的中间结果，以分区函数将其划分为 R 个块区，将本地硬盘所接收数据的位置信息通过 Master 向 Reduce 函数传送；

(4) 依据 Master 传送的文件信息，Reduce Worker 通过远程读取方式来寻找与之相对应的本地文件，利用 Reduce 函数进行调用计算，累加每个计算存储节点(DataNode)上相同的候选项集的支持度，以生成一个实际的支持度，对比用户申请时设置的最小支持度，以产生局部频繁 $K$ 项集的集合，用 $L_k^n$ 表示，通过有序排列文件中的中间 Key，远程向具体执行的 Reduce 发送信息；

(5) 根据 Key 排序后的中间数据，Reduce Worker 将 Key 与相应的中间结果集传送给 Reduce 函数，合并所有的处理结果，以产生全局频繁 $K$ 项集 $L_k$，并以最终输出文件来编写最后的结果。其挖掘算法实现流程如图 2 所示。

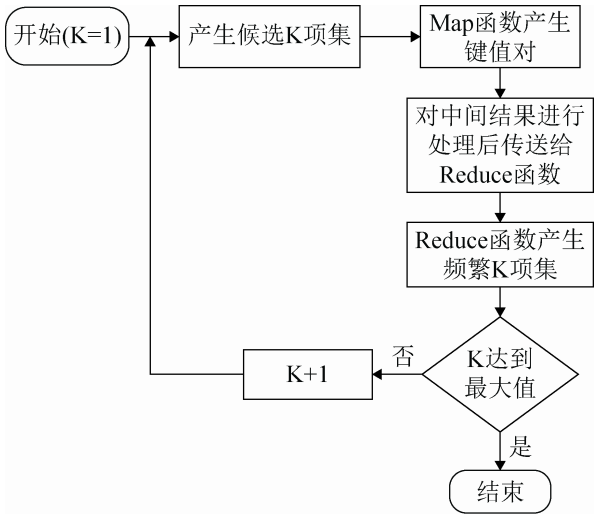


图 2 基于 Map Reduce 的 Apriori 算法实现流程图
Fig.2 The implementation flow chart of Apriori algorithm based on Map Reduce

## 3 验证原型设计

基于上述框架，本文搭建了验证原型，对本文提出的关键技术进行实验验证。原型结构[15]如图 3 所示。

原型结构主要包括数据存储层、数据挖掘算法层、挖掘任务处理层，其主控节点是 Master，任务是交互用户、调度与管理所有节点。其 Map Reduce 化的数据挖掘算法被存储在一部分的节点中，利于实现挖掘的高效性。在 HDFS 分布式存储系统中，主要由 1 个主存储节点(NameNode)、若干计算存储节点(DataNode)构成，其中计算存储节点(DataNode)负责接收用户的请求，同时向用户返回存储数据的存储节点的 IP，并向其它接收副本的计算存储节点(DataNode)发送通知。

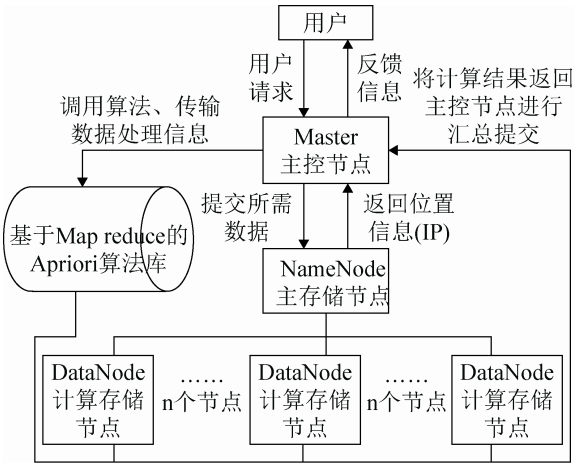


图 3 验证原型基本结构
Fig. 3 The basic structure of Prototype verification

结合参考文献[16]，上述验证原型系统所采用的具体技术如图 4 所示。

通过数据抽取工具，将空间数据库里的文件数据导入关系数据库中，之后利用数据转换工具，将关系数据库表数据转换为 GIS 文件，并将文件存储到 HDFS 分布式文件系统之中。之后运用基于格网的并行空间数据叠加分析工具进行分析，分析结果通过并行程序存储到结果数据库，供其它分析引擎和显示引擎进行分析和展示。




DOI: 10.16182/j.issn1004731x.joss.201803009

第 30 卷第 3 期 系统仿真学报 Vol. 30 No. 3
2018 年 3 月 Journal of System Simulation Mar., 2018

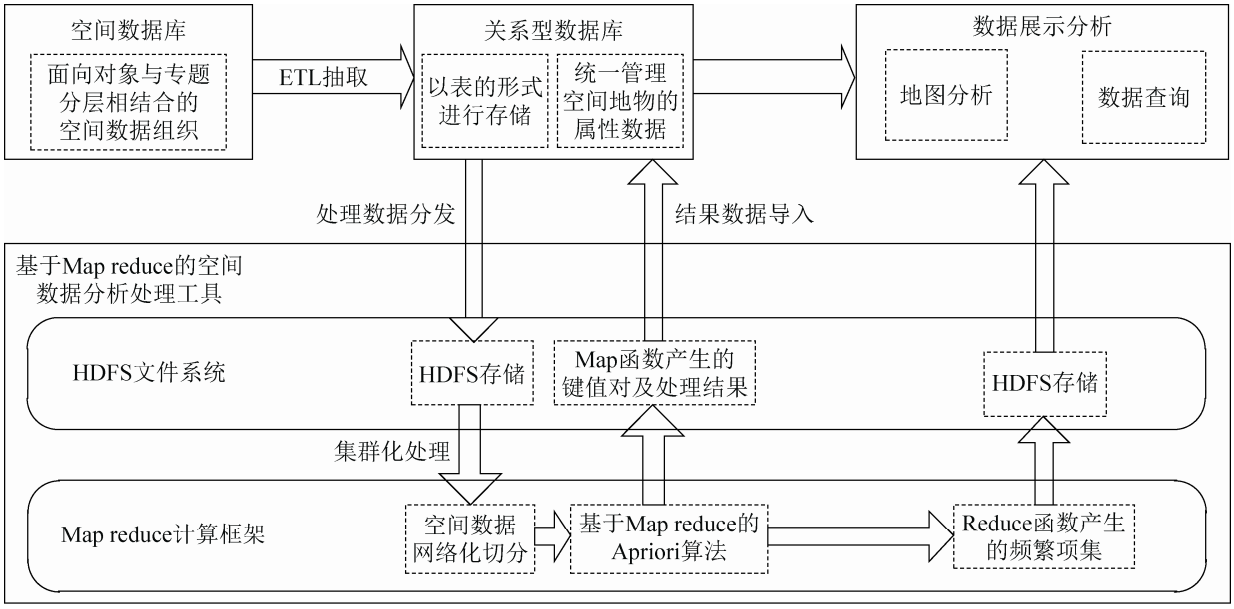


图 4 验证原型的技术架构
Fig. 4 The technical architecture of Prototype verification

## 4 仿真实验分析

基于 Hadoop 搭建 1 个主节点、3 个数据节点的 Hadoop 集群环境，选取缺省的 HDFS 文件块大小为 64MB。每个数据节点配置 1 颗 4 核 CPU、8G 内存、1TB 硬盘和千兆网卡。实验平台为：Ubuntu32 位，eclipse Mars4.5，Hadoop-2.5.2，jdk1.7。在虚拟机上安装 3 个 Ubuntu32 的系统，搭建分布式 Hadoop 平台，其中一个为 Master，另外两个作为 Slaves。Ubuntu32 的 3 个系统的配置完全相同。算法是由 Java 实现的。选取的实验的数据组成分为三大数据节点：测绘部门以文件形式存储的矢量地图数据节点、水文部门存储于 SQL Server 数据库中的雨量值属性数据节点和气象部门存储的云图影像数据节点，共 98 956 条记录，每条记录平均有 10 个属性。分别观察在 0slave, 1slave, 2slaves 和串行算法在不同的支持度下的执行速度。

首先实验当支持度为 1 时，在同一台设备上串行 Apriori 算法的运行时间，其运行结果见图 5。

由图 5 可以看出，串行 Apriori 算法在数据量很小的情况下，就已经显示出了较高的运行时间。随着记录数的增加，运行时间呈线性增长。等数据记录到达 50 条时则已无法完成时间的计算。因此可知，串行 Apriori 算法的执行效率较低。

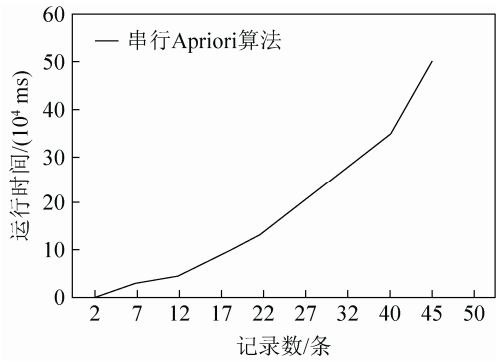


图 5 串行 Apriori 运行时间
Fig. 5 Serial Apriori running time

其次，在相同实验环境下，对并行 Apriori 算法的运行时间进行测试，分别记录不同节点(0、1、2)在不同支持度下执行实验数据的运行时间，其结果如图 6 所示。

由图 6 可知，在不同支持度下，随着 Hadoop 节点数的增加，运行的时间是在呈线性减小的。所以增加 Hadoop 节点的个数，有利于提高算法的执行速度。另外，与串行 Apriori 算法进行比较，并行算法在大量数据情况下的运行时间则明显低于串行算法的运行时间。



 

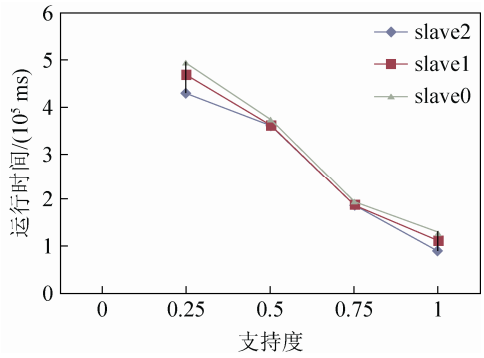


图 6 并行 Apriori 算法运行时间比较

Fig.6 Parallel Apriori algorithm runing time comparison

## 5 结论

通过仿真实验可以得出以下结论：基于 Map Reduce 并行的关联规则算法相较串行的关联规则算法有着更好的执行效率。其原因在于，基于 Map Reduce 的 Apriori 算法优化了传统算法需要多次扫描数据库和产生大量候选项集的缺点。新算法不需要多次扫描数据库，只是从 HDFS 中读取上一个频繁项目集的数据即可完成相应处理。基于 Map Reduce 的 Apriori 算法还可减少候选项集的产生。

## 参考文献:

[1] 张晓祥. 大数据时代的空间分析[J]. 武汉大学学报(信息科学版), 2014, 39(6): 655-659.
ZHANG Xiaoxiang. Spatial Analysis in the Era of Big Data[J]. Geomatics and Information Science of Wuhan University, 2014, 39(6): 655-659.

[2] Tong D, Mrrray A T. Spatial optimization in geography [J]. Annals of the Association of American Geographers (S1467-8306), 2012, 102(6): 1290-1309.

[3] 尹芳, 冯敏, 诸云强, 等. 基于开源 Hadoop 的矢量空间数据分布式处理研究[J]. 计算机工程与应用, 2013, 49(16): 25-29.
YIN Fang, FENG Min, ZHU Yunqiang, et al. Research on vector spatial data distributed computing using Hadoop projects[J]. Computer Engineering and Applications, 2013, 49(16): 25-29.

[4] Isard M, Budiu M, Yu Y, et al. Dryad: Distributed dataparallel programs from sequential building blocks[J]. ACM SIGOPS Operating Systems Review(S0163-5980), 2007, 41(3): 59-72.

[5] Zhao Dongjie, Yang Haitao, Jiang Jian, et al. Modeling and simulation of the invulnerability of space information network [C]// Internet Technology and Applications, 2010 International Conference on, 2010: 1-5.

[6] Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data [J]. ACM Transaction Computer Systems(S0734-2071), 2008, 26(2): 4-5.

[7] Melnik S, Gybarev A, Long J J, et al. Dremel: Interactive analysis of web scale datasets [J]. Proceedings of the VLDB Endowment(S2150-8097), 2010, 3(1/2): 330-339.

[8] Beaver D, Kumar S, Li H C, et al. Finding a needle in Haystack: facebook's photo storage[C]// Usenix Symposium on Operating Systems Design and Implementation, OSDI, Vancouver, Bc, Canada, Proceedings, 2010: 47-60.

[9] DeCandia Giuseppe. Dynamo: Amazon’s highly available key value store [J]. ACM SIGOPS Operating Systems Review(S0163-5980), 2007, 41(6): 205-220.

[10] Shvachko Konstantin. The hadoop distributed file system [C]//IEEE 26th Symposium on Mass Storage Systems and Technologies. IEEE, 2010: 15-16.

[11] 王树良, 丁刚毅, 钟鸣. 大数据下的空间数据挖掘思考[J]. 中国电子科学研究院学报, 2013, 8(1): 8-17.
WANG Shu-liang, DING Gang-yi, ZHONG Ming. On Spatial Data Mining under Big Data[J]. Journal of China Academy of Electronics and Information Technology, 2013, 8(1): 8-17.

[12] SHEKAR S, XIONG H(Eds.) Encyclopedia of GIS [M]. New York: Springer, 2007.

[13] 李德仁, 王树良, 李德毅. 空间数据挖掘理论与应用[M]. 北京: 科学出版社, 2006.
LI De-ren, WANG Shu-liang, Li De-yi. Spatial Data Mining: Theory and Application[M]. Beijing: Science Press, 2006.

[14] LIN Xueyan. MR-Apriori: Association rules algorithm based on Map Reduce [C]//IEEE Beijing Section. Proceedings of 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing: IEEE Beijing Section, 2014.

[15] 周芳. 基于关联规则 Apriori 算法的物联网海量数据挖掘系统研究[J]. 河北北方学院学报(自然科学版), 2015(31): 15-18.
ZHOU Fang. Mass Data Mining System for Internet of Things Based on Association Rules Apriori Algorithm[J]. Journal of Hebei North University(Natural Science Edition), 2015(31): 15-18.

[16] 贾文珏, 安琼. 基于“分治网格”的空间大数据快速分析方法[J]. 计算机工程与设计, 2015, 36(8): 2317-2321.
JIA Wen-jue, AN Qiong. Fast method for analysis of big spatial data using divide and conquer grids[J]. Computer Engineering and Design, 2015, 36(8): 2317-2321.



https://dc-china-simulation.researchcommons.org/journal/vol30/iss3/9
DOI: 10.16182/j.issn1004731x.joss.201803009