基于 Spark 的大数据热图可视化方法

张 繁[1,2], 袁兆康[1], 肖凡平[1], 尤 堃[1], 王章野[2]*

[1] (浙江工业大学计算机科学与技术学院 杭州 310023)
[2] (浙江大学 CAD&CG 国家重点实验室 杭州 310058)
(zywang@cad.zju.edu.cn)

摘 要: 针对普通客户端浏览和分析大数据困难的问题, 结合 Spark 和 LOD 技术, 以热图为例提出一种面向大数据可视化技术框架. 首先利用 Spark 平台分层并以瓦片为单位并行计算, 然后将结果分布式存储在 HDFS 上, 最后通过 web 服务器应用 Ajax 技术结合地理信息提供各种时空分析服务. 文中重点解决了数据点位置和地图之间的映射, 以及由于并行计算导致的热图瓦片之间边缘偏差这 2 个问题. 实验结果表明, 该方法将数据交互操作与数据绘制和计算任务分离, 为浏览器端大数据可视化提供了一个新的思路.

关键词: 热图; 并行计算; 大数据; 细节层次
中图法分类号: TP391.41

Research on Heatmap for Big Data Based on Spark

Zhang Fan[1,2], Yuan Zhaokang[1], Xiao Fanping[1], You Kun[1], and Wang Zhangye[2]*

[1] (*College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou* 310023)
[2] (*State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou* 310058)

Abstract: It is important to provide data analysts with effective and efficient exploratory tools via web browsers. However, due to the characteristics of big data, current data visualization approaches can hardly display the whole datasets. This paper proposes a general-purpose visualization framework based on Spark and LOD. Firstly, we implement a tile-based parallel computing algorithm for layered datasets with Spark. Secondly, we store the temporary results on HDFS. Finally, with Ajax and geographic information, we provide all kinds of spatial-temporal analysis services via web. This paper resolves two problems: one is the mapping of data points from heat map to geographic map; the other is the correction of the marginal error in overlapping areas caused by parallel computing. The experiment results suggest that by separating the data display and manipulation from the data rendering and computing tasks, our method provides a new way for big data visualization via web browsers.

Key words: heat map; parallel computing; big data; level of detail

随着大数据时代的到来, 领域科学家面对的数据集规模从 TB 级, 变成了 PB, EB 级, 甚至达到了 ZB 级, 所以在数据分析处理上遭遇各种限制和阻碍. 大数据研究成为一个新的关注热点, 其核心目的是使原本各自孤立的数据得以互相关联、融合, 数据挖掘和可视化正好为大数据研究提供了平台和技术, 数据挖掘是从底层探讨如何解析大数据的方法; 而可视化则是从展示层探究如何表达

收稿日期: 2016-04-30; 修回日期: 2016-07-28. 基金项目: 国家自然科学基金(61303141, 61272302); 浙江省自然科学基金(LY12F02034); 国家“八六三”高技术研究发展计划(2015AA016404). 张 繁(1978—), 男, 博士, 副教授, 硕士生导师, CCF 会员, 主要研究方向为并行计算、可视分析; 袁兆康(1992—), 男, 在校学生; 肖凡平(1991—), 男, 硕士研究生, 主要研究方向为并行计算、数据可视化; 尤 堃(1994—), 男, 在校学生; 王章野(1965—), 男, 博士, 副教授, 硕士生导师, CCF 会员, 论文通讯作者, 主要研究方向为信息可视化、计算机图形学.

大数据的手段，以帮助用户以可视便捷的方式，从几百万条数据中探索出各种复杂关系，从而使大数据变得可理解. 例如，蒲剑苏等[1]研究发现，针对 GPS，RFID 等数据量大、维度高的移动轨迹数据,可视化方法不但可以有效展示数据中包含的时空特征，还可以发现轨迹中多维属性之间的关系，探索数据中隐藏的时空规律. 任磊等[2]通过研究也认为，其可视化和可视分析是大数据分析的重要方法，其能够有效地弥补计算机自动化分析方法的劣势与不足.

目前大数据可视化面临的主要问题包括:

1) 数据复杂散乱. 经常发生数据缺失、数据值不对、结构化程度不高.

2) 迭代式分析成本高. 在初次查询后如果发现结果不对，改变查询条件重新查询代价高.

3) 构建复杂工作流困难. 从多数据源取得包含各种不同特征的原始数据，然后执行机器学习算法或者复杂查询，探索过程漫长.

4) 受到原有技术限制，对小规模数据分析很难直接扩展到大数据分析.

5) 数据点的规模超过普通显示器可能提供的有效像素点.

Hadoop①和 Spark②先后成为大数据分析工业界的研究热点，前者是一个能够对大量数据提供分布式处理的软件框架和文件系统(hadoop distributed file system, HDFS)；后者是一个通用大数据计算平台，可以解决大数据计算中的批处理、 交互查询及流式计算等核心问题. Zeppelin③可以作为 Spark 的解释器，进一步提供基于 Web 页面的数据分析和可视化协作，可以输出表格、柱状图、折线图、饼状图、点图等，但是无法提供更为复杂的交互分析手段.

1 相关工作

随着用户对浏览器端数据可视化的需求日益增长，工业界出现了许多面向 web 的轻量级数据可视化工具. D3[3]是一个常用动态图形显示数据的 JavaScript 库，其基于 HTML，SVG(矢量图形)和 CSS 将任意数据绑定到一个 DOM(文档对象模型)元素，并对 DOM 实施基于数据的变换，通过图形变换和放大缩小等交互操作展示数据. D3 具有良好的可移植性，缺点是使用 SVG 不能支持十亿个点以上的数据. 然而使用 canvas 画布绘图的 heatmap.js 在面对大数据量时也无能为力. 与 D3 相似的 JavaScript 库还有百度公司的 ECharts④，Python 语言库 Matplotlib⑤和 Bokeh⑥等.

热图是一种常用的基本数据可视化技术，通常用颜色编码数值大小， 并以矩阵或方格形式整齐排列，在二维平面或者地图上呈现数据空间分布,被广泛应用在许多领域. 近年来，许多研究者成功地将热图应用在眼动数据可视分析上，有效地概括并表达用户视觉注意力的累计分布[4-5]. Ma 等[6]利用热图展示基于基站数据的人口流动量；Röthlisberger 等[7]利用可配置的热图为开发人员在软件集成开发环境中基于任务导航提供有效支持；Gove 等[8]针对动态社交网络统计和内容数据，设计并实现了一款基于热图和矩阵可视化工具——NetVisia，将社交网络节点属性的变化展示给用户，从而支持用户探索社交网络随时间演进信息. Kopp 等[9]发现，在对多步集合模拟数据分析中，应用热图可视化方法可以协助用户有效识别高活跃区域和低活跃区域，从而提高利用神经网络模型进行决策分析的准确率.

针对数据可视化绘制速度慢、效率低等问题，孙敏等[10]提出基于格网划分的 LOD(levels of detail)分层方法，实现对大数据集 DEM 数据的实时漫游. 巴振宇等[11]在研究流场可视化时，为了更有效地可视化流场特征结构，利用 LOD 思想建立了一个基于图像空间的交互系统，通过缩放操作控制流场特征的不同细节层次显示效果. 聂俊岚等[12]利用边绑定技术展现更加整齐抽象的可视化效果，增强可读性，并利用四叉树结构对地域信息进行划分，将热点影响域与地理位置挂接，大大加快了热图计算的速度，实现交通数据信息与人口迁徙信息的可视化. 贺群等[13]对 WebGIS 大数据可视化

① http://hadoop.apache.org

② http://spark.apache.org

③ https://zeppelin.incubator.apache.org

④ http://echarts.baidu.com

⑤ http://www.matplotlib.org

⑥ http://bokeh.pydata.org

研究进行了详细的探讨，通过选取部分数据点来缩减绘制的时间开销，但是在对海量数据绘制过程中，每次拖动数据地图都需要重新实时绘制．杨微等[14]提出一种基于热图的地理对象空间分布热度计算方法，通过多级网格空间聚类方法虽然减少了热度计算量，但是会造成整体数据分析的不准确．Heatmap.js①是一个轻量级的专用 JavaScript 库，通过 canvas 来绘制热图，可以支持多达 4 万个数据点的热图绘制．但同样受到浏览器绘制能力的限制，当绘制更多数据时这个工具就无能为力了．因此，在面向大数据全样本分析研究中，如何有效地实现基于 web 浏览器的可视化仍然是一个巨大挑战．

2　并行计算大数据热图

为了将热图和地图叠加，呈现数据的时空特性，我们首先将数据中的经纬度坐标与地图位置坐标互相对应；然后由 Spark 集群计算热图，解决并行处理导致的边缘问题；最后把热图与实际地图叠加，利用 Ajax(Asynchronous JavaScript and XML)技术实现在客户浏览器端的大数据热图可视化．

2.1　经纬度换算

对于地理上的同一片区域，用户通过放大或者缩小操作看到精度更高的地图．此时的热图也应该与放大后的地图匹配，因此我们首先要根据用户缩放要求绘制与不同精度地图匹配的热图．

根据 LOD 思想，本文设计了一种多层次绘制方法，以离线计算的方式预先绘制不同层级的热图．如图 1 所示，编号为(0,0)的地图块为 $n\times n$ 个像素，把它叫作第 0 层．将第 0 层放大一倍后，分割成 4 个子区间{(0,0)(0,1)(1,0)(1,1)}，得到更高精度的地图，把它叫作第 1 层．对第 1 层重复上面的操作,得到第 2 层；重复这个过程，直到满足所需要的精度为止．在本文实验中，取 5 层作为默认值，也可以根据用户需要设置更多层．

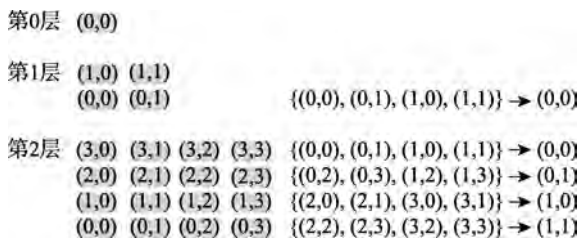


图 1　分层并行计算示意图

为了确定热点在不同层级瓦片上所在的像素位置，我们采用墨卡托投影[15](正轴等角圆柱投影)来获得绘制点转换后的位置，将经纬度坐标转换到瓦片上位置坐标的转换公式为

$$x=\left[2^{z-1}\cdot\left(\frac{\lambda}{180}+1\right)\right] \tag{1}$$

$$y=\left[2^{z-1}\cdot\left(1-\frac{\ln\left[\tan\left(\frac{\pi\cdot\varphi}{180}\right)+\sec\left(\frac{\pi\cdot\varphi}{180}\right)\right]}{\pi}\right)\right] \tag{2}$$

$$m=\left[\left(\frac{\lambda}{180}+1\right)/2^{1-z}-x\right]\cdot 256 \tag{3}$$

$$n=\left[\left[1-\ln\left[\tan(\varphi+90)\cdot\pi/360\right]/\pi\right]/2^{1-z}-y\right]\cdot 256 \tag{4}$$

在式(1)(2)中，λ 和 ψ 为经纬度；z 为所在层级，x 和 y 为所在层级的瓦片位置，再由式(3)(4)计算得到该瓦片上需要绘制的像素位置(m, n).

2.2　并行计算热图

在 Spark 平台上实现热图的绘制，首先根据 2.1 节所描述的方法将经纬度坐标转换为对应不同瓦片上的像素坐标．每个基站的辐射范围可近似认为相同，即每个基站的初始影响力近似相同，因此可采用影响力叠加法[14]将数据点绘制到画布上，然后做径向渐变，叠加出每个位置的影响大小，得到初始灰度图，如图 2a 所示．然后将每一个像素点着色，根据每个像素的灰度值大小，以及调色板将灰度值映射成相对应的颜色．图 2b 是一个透明的 PNG 格式图片，调色板如图 2c 所示．本文中出现的热图均采用图 2c 调色板．

将计算出的热图结果存储在 HDFS 上，并与经纬度以及层级建立索引关系方便以后读取，拼接后的热图绘制效果如图 3 所示．

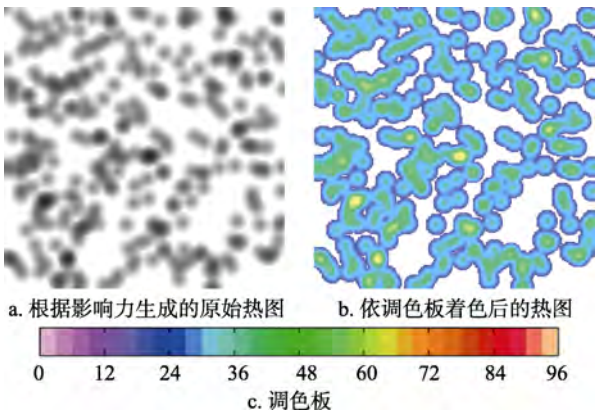


图 2　热图灰度图通过映射转为彩色图

① http://www.patrick-wied.at/static/heat-mapjs

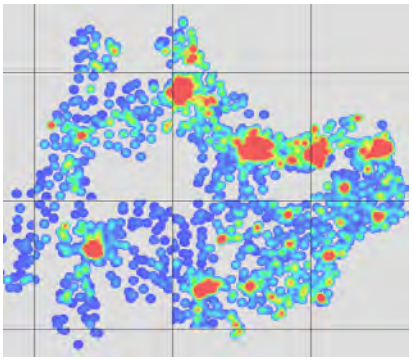


图 3 拼接后的热图绘制效果

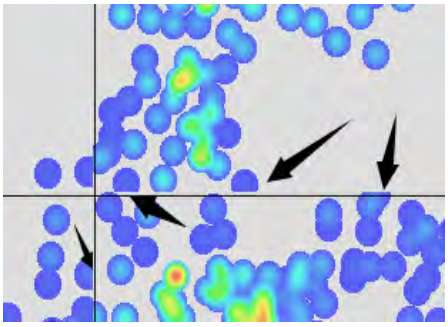


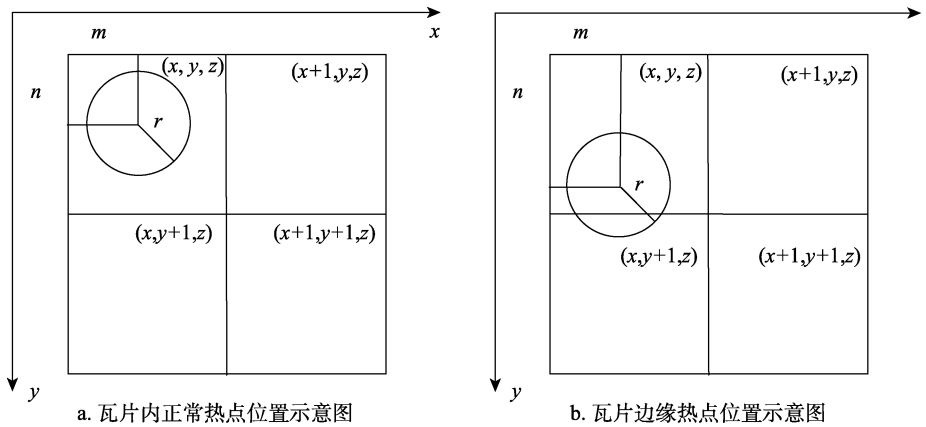
图 4 瓦片间热图边缘问题

3 瓦片边缘问题

在以瓦片为单位并行计算热图时，某些热点中心到相邻瓦片的距离会小于热点半径，这就会出现瓦片边缘热点计算不完整的情况，图 4 所示箭头所指位置即为存在边缘问题的区域.

为了能绘制完整热图，本文提出了一种对边缘热点多次重叠计算的方法，以解决边缘偏差问题.

对瓦片相邻的边缘热点进行多次重叠计算，设(x, y, z)为第 z 层的序号为(x, y)的瓦片，(m, n)为热点相对当前所在瓦片的像素，热点位置可定义为(x, y, z, m, n). 如图 5 所示，其中半径 r 为热点的影响力范围.



a. 瓦片内正常热点位置示意图　　b. 瓦片边缘热点位置示意图

图 5 瓦片内部及边缘热点位置示意图

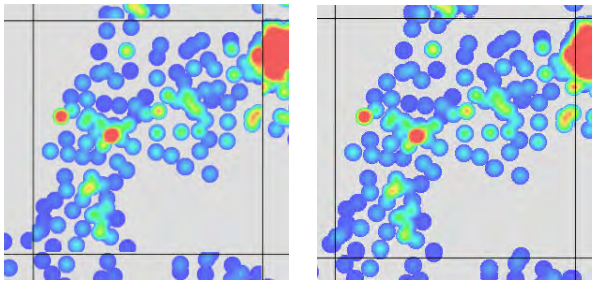
我们假设瓦片大小为 $k\times k$ 像素范围，若 $k-n<r$，即瓦片$(x, y+1, z)$的上边缘存在缺失. 因此我们可以在瓦片$(x, y+1, z)$的相对位置$(m, -(k-n))$再次计算半径为 r 的热图，这样即可补全热图，从而解决边缘偏差.

现给出具体流程：假设瓦片的大小为 $k\times k$ 像素，首先在(x, y, z)上的(m, n)处绘制半径为 r 的热图，然后判断热点所在位置是否处于瓦片边缘. 伪码如下：

```
//对数据集中的每一个点，判断是否处于瓦片边缘并需要重叠计算
for each (x, y, z, m, n) in DataSet:
//在第 z 层编号为(x, y)的瓦片上，计算圆心为(m, n)、半径为 r 的热点区域
calculate(x, y, z, m, n)
//以 4 个象限为基准判断是否处于瓦片边缘，重复 8 次
if(n < r)
    calculate (x, y−1, z, m, k+n);
if(m < r)
    calculate (x−1, y, z, k+m, n);
if(m < r && n < r)
    calculate (x−1, y−1, z, m+k, k+n);
if(k − n < r)
    calculate (x, y+1, z, m, − (k−n));
if(k − m < r)
    calculate (x+1, y, z, − (k−m), n);
if(k − n < r && k − m < r)
    calculate (x+1, y+1, z, − (k−m), − (k−n));
if(k − n < r && m < r)
    calculate (x−1, y+1, z, k+m, − (k−n));
```

if($n < r$ && $k - m < r$)
calculate (x+1, y–1, z, – (k–m) , k+n);
end for

从以上伪码可以看出，边缘热点可能处于 2 片或者 4 片瓦片之间，因此需要通过 2 次或者 4 次重复计算．通过本文提出的重叠计算方法可以解决热图分片计算的边缘问题，图 6a 为未解决边缘问题的热图，图 6b 为解决边缘偏差后的热图．



a．存在边缘问题的热图　　b．解决了边缘问题的热图

图 6　存在边缘偏差与解决边缘问题后的热图对比

4　实　验

本文采用的数据集是浙江某运营商提供的湖州地区手机移动数据，包含基站数据和通话记录数据．基站数据集的记录由地理位置标识，基站 ID，经纬度，区域类型组成．通话记录数据集的记录由主叫用户 ID，被叫用户 ID，时间戳，主叫用户所在基站 ID，被叫用户所在基站 ID 组成．本文采用的地图数据源自开源 OpenStreetMap，所用计算集群一共有 16 个节点，CPU 型号是 Intel Xeon E7540 (主频为 2.4 GHz)，16 GB 内存，Hadoop 版本是 2.4.0，Spark 版本是 1.6.1，使用普通微机作为客户端，CPU 型号是 Intel i5-3337U(主频为 2.7 GHz)，8 GB 内存，浏览器为 Google Chrome v42.0.

首先对 2013-12-20 数据进行清洗，将字段缺失的记录去除，得到 6 793 495 条记录．然后利用 Spark 对每个整点时间并行计算 5 层热图，着色绘制后将热图结果以 PNG 格式存储在 HDFS 上．当用户在 web 客户端进行交互时，通过 Ajax 的方式加载与当前地图位置对应的不同层级热图，使热图与地图叠加．我们计算并绘制了当天 10:00 565 976 条记录的热图，显示效果如图 7 所示．图 7a 展示层级为第 0 层，用户通过鼠标交互逐级放大显示区域，得到图 7b, 7c 效果．用户通过鼠标拖曳实现地图浏览，通过鼠标滑轮实现放大和缩小，交互操作流畅，无迟滞停顿现象．

进一步计算并比较 2013-12-20 10:00 和 2013-12-21 10:00 的基站通话数据热图，从图 8a, 8b 中可以发现，图 8a 中黑框区域的热度明显高于图 8b 的热度．进一步查阅地图得知，此区域是湖州市政府所在地，这 2 日分别是周五和周六，说明图 8 正确揭示了移动通话数量在工作日和周末的差异．

a．第 0 层热图与地图叠加显示　　b．第 1 层热图与地图叠加显示　　c．第 2 层热图与地图叠加显示

图 7　热图与地图逐级叠加示意图

a. 2013-12-20 10:00 热图　　b. 2013-12-21 10:00 热图

图 8　同一区域不同时间热图比较

5　结　语

本文提出的大数据热图可视化方法能够有效地解决前端绘制计算量大的问题，通过在 Spark 平台上以瓦片为单位分层次并行计算热图，将生成的热图存储在 HDFS 上，然后通过 web 服务器提供浏览器交互服务，用户可以通过在地图上拖动鼠标或放大/缩小等操作选择感兴趣区域，再分析不同

时间点用户行为差异或渐变过程. 通过解决热图数据点和地图映射关系问题以及瓦片热图之间的边缘问题, 提供大数据热图绘制方法, 以满足用户交互、协同和共享等多方面需求. 该方法可以拓展到其他常用可视化方法, 如 ScatterPlot, Bar Chart, 平行坐标等. 但绘制过程是基于 Spark 计算后得到的离线数据, 在实时性上还不能得到保证, 在下一步工作中, 我们将着手利用 Spark Streaming 库来解决这一问题.

参考文献(References):

[1] Pu Jiansu, Qu Huamin, Ni Lionel. Survey on visualization of trajectory data[J]. Journal of Computer-Aided Design & Computer Graphics, 2012, 24(10): 1273-1282(in Chinese)
(蒲剑苏, 屈华民, 倪明选. 移动轨迹数据的可视化[J]. 计算机辅助设计与图形学学报, 2012, 24(10): 1273-1282)

[2] Ren Lei, Du Yi, Ma Shuai, *et al*. Visual analytics towards big data[J]. Journal of Software, 2014, 25(9): 1909-1936(in Chinese)
(任 磊, 杜 一, 马 帅, 等. 大数据可视分析综述[J]. 软件学报, 2014, 25(9): 1909-1936)

[3] Bostock M, Ogievetsky V, Heer J. D^3: data-driven documents[J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2301-2309

[4] Bojko A A. Informative or misleading? Heatmaps deconstructed [M] //Lecture Notes in Computer Science. Heidelberg: Spring, 2009, 5610: 30-39

[5] Cheng Shiwei, Sun Lingyun. A survey on visualization for eye tracking data[J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26(5): 698-707(in Chinese)
(程时伟, 孙凌云. 眼动数据可视化综述[J]. 计算机辅助设计与图形学学报, 2014, 26(5): 698-707)

[6] Ma Y X, Lin T, Cao Z D, *et al*. Mobility viewer: an eulerian approach for studying urban crowd flow[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(9): 2627-2636

[7] Röthlisberger D, Nierstrasz O, Ducasse S, *et al*. Supporting task-oriented navigation in IDEs with configurable HeatMaps [C] //Proceedings of the 17th IEEE International Conference on ProgramComprehension. Los Alamitos: IEEE Computer Society Press, 2009: 253-257

[8] Gove R, Gramsky N, Kirby R, *et al*. NetVisia: heat map & matrix visualization of dynamics social network statistics & content [C] //Proceedings of the 3rd IEEE International Conference on Privacy, Security, Risk and Trust, and the 3rd IEEE International Conference on Social Computing. Los Alamitos: IEEE Computer Society Press, 2011: 19-26

[9] Kopp C, von Mettenheim H J, Breitner M H. Decision analytics with heatmap visualization for multi-step ensemble data[J]. Business & Information Systems Engineering, 2014, 6(3): 131-140

[10] Sun Min, Xue Yong, Ma Ainai. 3D visualization of large DEM data set based on grid division[J]. Journal of Computer-Aided Design & Computer Graphics, 2002, 14(6): 566-570(in Chinese)
(孙 敏, 薛 勇, 马蔼乃. 基于格网划分的大数据集 DEM 三维可视化[J]. 计算机辅助设计与图形学学报, 2002, 14(6): 566-570)

[11] Ba Zhenyu, Shan Guihua, Liu Jun, *et al*. A feature-based seeding method for multi-level flow visualization[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(1): 32-40(in Chinese)
(巴振宇, 单桂华, 刘 俊, 等. 一种基于特征信息种子点选取的多层次流线可视化[J]. 计算机辅助设计与图形学学报, 2016, 28(1): 32-40)

[12] Nie Junlan, Xin Meiyue, Zhang Jikai, *et al*. An improved method of geographic traffic information based on heatmap visualization[J]. Journal of Sichuan University: Engineering Science Edition, 2015, 47(4): 118-124(in Chinese)
(聂俊岚, 辛妹悦, 张继凯, 等. 一种改进的地理交通信息热图可视化方法[J]. 四川大学学报: 工程科学版, 2015, 47(4): 118-124)

[13] He Qun, Yang Mingchuan. WebGIS-based big data visualization and optimization[J]. Telecommunications Technology, 2015(6): 37-40(in Chinese)
(贺 群, 杨明川. 基于 WebGIS 的大数据可视化研究与优化[J]. 电信技术, 2015(6): 37-40)

[14] Yang Wei, Liu Jiping, Wang Yong. Computation of Heat distribution for geographic object space[J]. Bulletin of Surveying and Mapping, 2012(s1): 391-393+398(in Chinese)
(杨 微, 刘纪平, 王 勇. 基于 Heatmap 的地理对象空间分布热度计算方法[J]. 测绘通报, 2012(s1): 391-393+398)

[15] Li Changchun, Cai Baigen, Shangguan Wei, *et al*. Research and implementation of the map algorithm based on Web mercator[J]. Application Research of Computers, 2012, 29(12): 4793-4796(in Chinese)
(李长春, 蔡伯根, 上官伟, 等. 基于 Web 墨卡托投影的地图算法研究与实现[J]. 计算机应用研究, 2012, 29(12): 4793-4796)