Application of Network Crawling Spatial Data based on ArcGIS

Lanlan Ding
Chongqing Jiaotong University, Chongqing, 400074, China

Abstract: Today, with the "big data storm" swept away, web crawler technology is becoming more and more popular and practical at present, and the combination of data mining and extracting information with ArcGIS processing spatial data is in line with the problems that many industries will involve. How to correctly use the crawling data and set the correct coordinate system is the key. Through this research, we can understand the data mining channels, master the process of setting and transforming the network crawling space data and the subsequent processing and analysis. This paper uses the network to crawl the relevant information of the hot pot restaurant in Chongqing, the public comment network, as a data source, analyzes the density distribution of hot pot restaurants in the main city of Chongqing, and produces the heat map of the distribution of hot pot restaurants in Chongqing, and discusses the distribution of hot pot restaurants.The GeoSharp network was used to climb the raw data of the hot pot store in Chongqing. The data was imported into ArcGIS software to set up the coordinate system and convert the projection. The distribution of the hot pot restaurant in the main city of Chongqing was analyzed through density analysis. The results show that the hottest place in the main city is the most densely distributed place in the Yuzhong District, as well as the junction of Shapingba District, Dadukou District, Nan'an District and Jiulongpo District.

Keywords: ArcGIS; Network crawling; Projection conversion

基于 ArcGIS 网络爬取空间数据的应用

丁兰兰
重庆交通大学，重庆市， 400074

摘 要：在“大数据风暴”席卷而来的今天，网络爬虫技术在当下越来越流行和实用，而将数据挖掘和提取信息与 ArcGIS 空间分析功能充分结合起来，满足当下很多行业的需求。如何利用爬取数据并设置正确的坐标系统是关键所在，通过研究了解数据挖掘的渠道，掌握网络爬取空间数据坐标设置与转换以及数据处理与分析的过程。本文利用 GeoSharp 网络爬取重庆市火锅店相关信息作为数据源，在 ArcGIS 中对其进行坐标设置和投影转换，通过密度分析对重庆市主城区火锅店的分布制作热力专题地图。结果表明，主城的火锅店集中分布最密集的地方在渝中区，以及沙坪坝区、大渡口区、南岸区、九龙坡区等区的交界处。

关键词：ArcGIS； 网络爬取； 投影转换

1 引言

随着计算机技术的提高，地理信息系统技术（以下简称 GIS）得到了大力发展，在我国，使用和学习 GIS 的人越来越多，通过 GIS 对数据进行存储、编辑、处理、空间分析、显示等操作，可以提取大量隐含的有用信息，为地图制图、资源分配、环境保护、金融、城市规划、航海、土地利用等领域的决策提供科学的依据。在实际应用当中，时常需将多源数据转换到统一的参考基准面、统一的地图投影和统一的坐标系统中，才能够继续对数据操作和处理[1]。为了实现这个原则，就需要地图投影和坐标系的转换。而在大数据时代的背景下，网络爬虫技术的兴起，为越来越多的行业提供便利，对于测绘专业来讲，如果正确利用网络爬取数据以达到我们想要的目的是关键所在。国内外其他学者对网络爬虫技术和 ArcGIS 中坐标转换的研究大致情况如下：王明军[2]探讨在普通爬虫技术基础上，提出了空间敏感爬虫的思想体系，并从其概念、与普通爬虫异同、工作方式等多个方面对其进行阐述；冯明远[3]为了实现深度网络信息的全自动检对深度，探讨了网络爬虫技术的特征和基本构

架，对网络信息的爬取关键技术进行了研究；韩丽君[4]等介绍了地图投影的概念、分类和不同投影之间的变换，以及 GIS 中地图投影的设置；诸云强[5]系统阐述了地图投影、地图投影变换、高斯投影及其分带投影的基本原理，以地图投影及其变换为基础，采用 Visual Basic6. 0 作为软件平台，开发了 GIS 的地图投影变换软件。

本文涉及网络爬取的空间数据可能面临着坐标系未知或定义不正确等问题，如何对网络爬取数据设置正确的坐标系是后期对数据进行利用的前提条件。实验将利用网络爬取的空间数据，在 ArcGIS 当中进行坐标系的设置和投影转换，将数据挖掘和提取信息与 ArcGIS 处理空间数据进行充分结合起来，符合当下很多行业会涉及的问题，并且通过实验了解数据挖掘和数据处理与分析的过程。

2 研究区与数据来源

2.1 研究区

重庆地处中国西南部，位于东经 105°11'-110°11'、北纬 28°10'-32°13'。由于重庆地处山区，常年多雨多雾，气候潮湿，火锅是当地人民最受欢迎的美食之一，火锅店在大街小巷随处可见。

2.2 数据来源

网络爬取方法很多，可以利用火车采集器等免费软件进行爬取火锅店经纬度等信息。通过火车采集器软件对大众点评网进行网络爬取时，少量个数的网址进行采集测试是可以成功的，但大量采集时就会出错，数据不完整。而利用 GeoSharp 软件爬取百度地图上的火锅店位置等信息，用时最长，但数据量较完整，因此本文的数据源采用 GeoSharp 爬取结果。

3 数据处理

3.1 实验步骤

实验技术流程如图 1 所示，主要利用 ArcGIS 软件对数据进行空间坐标设置和可视化分析。

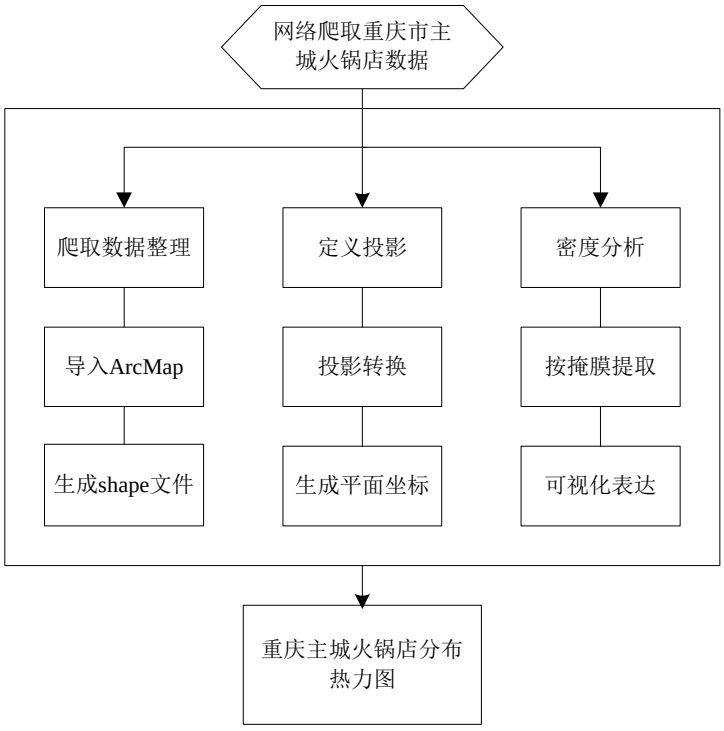

Figure 1. Technical flow char
图 1. 技术流程图

3.1.1 数据整理与导入

对网络爬取后的重庆市主城 9 个区火锅店分布的数据，进行数据清理和整理，由于爬取数据是按矩形

窗口进行拾取数据（详见重庆主城区下载范围图 2），故在一些区边界处会产生重复数据，并且会爬取到其他区县的数据，因此要将数据进行整理，只保留主城 9 个区的数据， 最终保存为“sData.xls”文件。

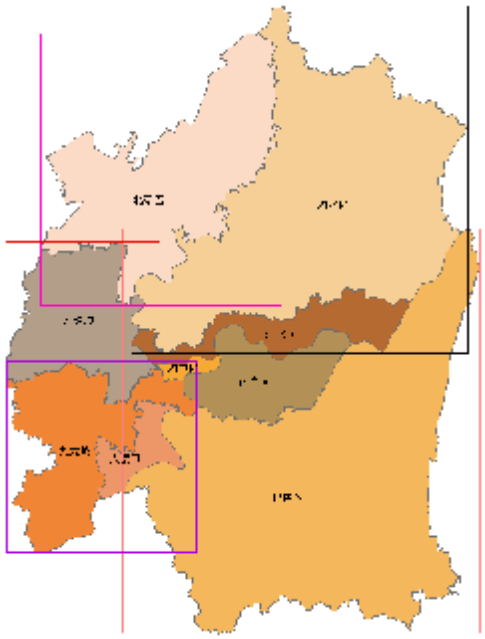

Figure 2. Chongqing main city download range map
图 2. 重庆主城区下载范围图

打开 ArcMap，添加“sData.xls”文件，将 Excel 文件中“sheet1”添加进来。开始添加点，先右击添加的“Sheet1”，然后选择“显示 XY 数据（X）”，打开“显示 XY 数据”的数据框。然后生成 shape 点文件，将数据点击右键，选择导出数据。

3.1.2 设置坐标系

重新打开导出数据，打开数据图层属性发现其坐标系未定义。故先定义一个地理坐标系西安 80 坐标系。在目录中选择“sData.xls”，右键点击属性，找到右上角图标添加坐标系，选择地理坐标系西安 80。

要将地理坐标系的经纬度转换为平面坐标系，必须经过投影这一步。考虑重庆地区的经纬度范围（105 度—108 度），选择高斯投影 6 度带的 18 带。选择“数据管理工具”中选择“投影”工具。选择确定后，新建文档打开，将投影转换后的文件添加至图层显示。可以明显看见坐标显示已经变成平面坐标系，以米为单位。将重庆市主城区划图添加至图层，先定义投影到地理坐标系西安 80，然后投影到相同的坐标系下，结果如图 3 所示。

最后生成（x,y）平面坐标。利用 Feature 的几何字段来获取平面坐标的数据。打开属性表，分别添加字段 X，Y。选中 X 点击计算几何。同理对 Y 进行计算几何设置。最终计算出平面坐标。

3.1.3 密度分析

要制作热力图需要对重庆市主城区域的火锅店分布数据进行点密度分析。打开工具里面的空间分析工具中的密度分析，选择点密度分析，得到密度图。

3.1.4 可视化

由于我们只要重庆市主城区域的密度分析图，故需要对密度分析后的数据提取重庆市主城区域的结果，选择空间分析工具中的“按掩膜提取”工具。对提取后的图进行调节色阶，选择合适的颜色表达。最终制作成为一幅专题地图，如图 4 所示。

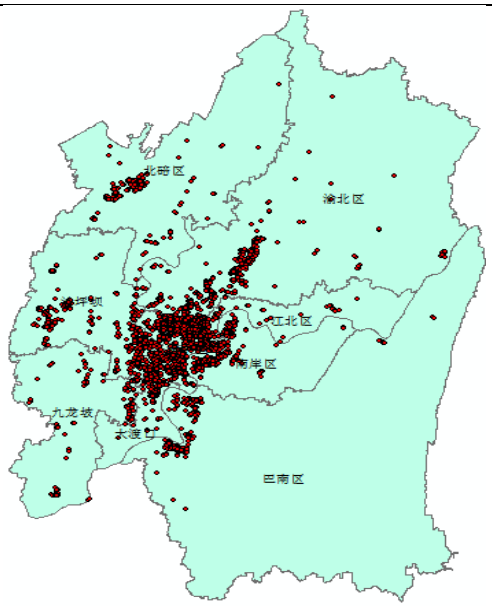

Figure 3. Hot pot shop distribution map
图 3. 火锅店分布图

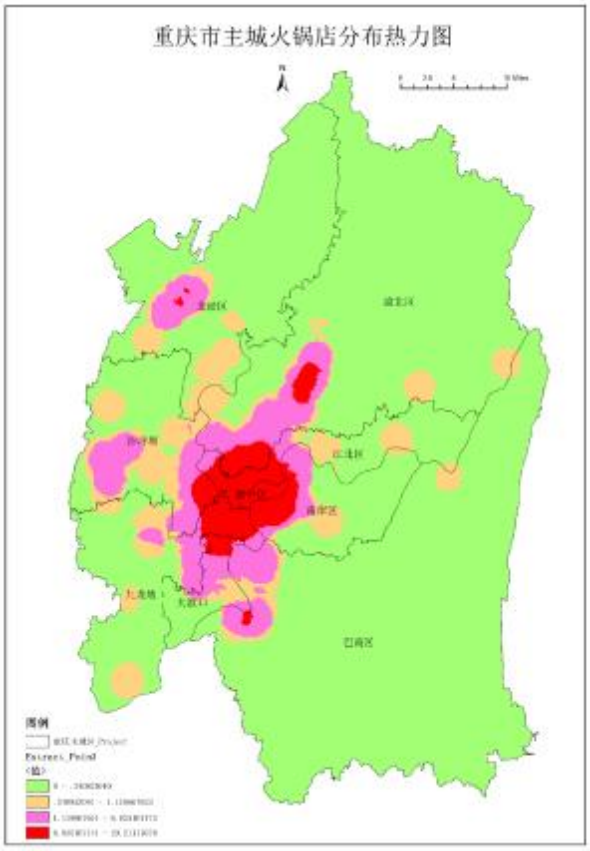

Figure 4. Hotpot shop thematic map
图 4. 火锅店分布专题地图

3.2 结果分析

由结果可知，主城的火锅店集中分布最密集的地方在渝中区，以及沙坪坝区、大渡口区、南岸区、九龙坡区等区的交界处，而各大商圈、景区、居民区都是火锅馆最密集的地方。

4 结束语

各行各业卷入大数据风暴，了解和应用大数据时代带来的优势，这是一个潮流，也是一种使命。但是网络爬取的空间数据可能面临着坐标系未知或定义不正确等问题，如何对网络爬取数据定义和设置正确的坐标系是后期对数据进行利用的前提条件。本文将利用网络爬取的空间数据，在 ArcGIS 当中进行坐标系的设置和投影转换，将数据挖掘和提取信息与 ArcGIS 处理空间数据进行充分结合起来，符合当下很多行业会涉及的问题，并了解数据挖掘和数据处理与分析过程。

致 谢

要诚挚感谢我的老师和同学。论文的完成离不开老师的细心指导和耐心负责，要一直保持着实事求是的态度，尊重科学事实，尊重理论依据，掌握基本的研究方法，用心写好论文。同时，非常感谢我的同学，这篇论文实验中的每一个实验步骤和细节，都离不开同学的热心帮忙和耐心解答。

References（参考文献）

[1] Qi Yu. Research on Distributed Spatial Data Integration System Based on Mediator/Wrapper System. Master's Thesis of National University of Defense Technology.
祁羽. 基于 Mediator/Wrapper 体系的分布式空间数据集成系统研究. 国防科学技术大学硕士论文.

[2] Wang Mingjun. Research on Spatial Data Crawling and Measurement Based on Web. Wuhan University. 2013.
王明军. 基于 Web 的空间数据爬取与度量研究. 武汉大学, 2013.

[3] Feng Mingyuan. Research and Implementation of Key Tech-nologies for Deep Network Information Crawling. Zhejiang University. 2010.
冯明远. 深度网络信息爬取关键技术研究与实现. 浙江大学. 2010

[4] Han Lijun, An Jiancheng. Map Projection and Its Application in GIS. Library and Information Guide. 2009, 19(8), 136-138.
韩丽君, 安建成. 地图投影及其在 GIS 中的应用. 图书情报导刊. 2009, 19(8), 136-138.

[5] Zhu Yunqiang, Gong Huili, Xu Huiping. Map Projection Transformation in GIS. Journal of Capital Normal University(Natural Science Edition). 2001, 22(3) , 88-94.
诸云强, 宫辉力, 许惠平. GIS 中的地图投影变换. 首都师范大学学报(自然科学版). 2001, 22(3), 88-94.

[6] Zhang Dengjun, Wang Baoshan. Discussion on the Selection and Setting of Map Projection in GIS. Surveying and Spatial Geographic Information.
张灯军，王宝山. 浅谈 GIS 中地图投影的选择与设置. 测绘与空间地理信息.

[7] Xu Yuying. A Brief Introduction to the Method of Making thematic Maps by ArcGIS. Modern Surveying and Mapping.
许玉英. 简述 ArcGIS 制作专题地图的方法. 现代测绘.

[8] Berthoud, M.G. An Equal-area Map Projection for Irregular Objects. Icarus 2005, 175, 382-389.

[9] Snyder, J.P. Map Projections-A Working Manual. United States Government Printing: Washington, DC: USA, 1987.

[10] Olson, J.M. Map Projections and the Visual Detective: How to Tell if a Map is Equal-Area, Conformal, or Neither. J. Geogr. 2006, 105, 13-32.