

<http://www.ixueshu.com>

论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外,不包含其他人或机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中做了明确的声明并表示了谢意。

作者签名: 王丽魏


日期: 2017.6.7

论文使用授权声明

本人完全了解上海师范大学有关保留、使用学位论文的规定,即:学校有权保留送交论文的复印件,允许论文被查阅和借阅;学校可以公布论文的全部或部分内容,可以采用影印、缩印或其它手段保存论文。保密的论文在解密后遵守此规定。

作者签名: 王丽魏

导师签名:



日期: 2017.6.7

摘要

移动互联网技术的飞速发展以及移动终端设备的广泛普及,人们对空间信息服务的需求日益增大。空间信息服务、无线移动服务和社交网络服务的结合,产生了海量的社交媒体地理数据。社交媒体地理数据作为大数据时代地理信息的重要组成部分,具有自身独特的性质,包含空间、时间、语义等属性信息,数据更新速度快、数据量大,与用户活动息息相关,信息量丰富等,使其相比于传统地理数据,蕴含较大的发展潜力与应用价值。得到用户长期的行为操作数据可以发掘用户的行为模式,认知用户的偏好,真实再现用户在现实世界中的生活轨迹,帮助认知用户的行为模式;从提供基于位置服务的角度来说,对用户活动及出行提供更好的认知,也可以帮助运营商提高服务水平与质量;另一方面可以根据用户的关注度来发现用户关注的城市热点区域,针对不同地点,不同时段的用户出行需求制定基于位置的服务。对政府和社会而言,这些数据是对人类活动的真实采样。从这些数据中可以分析出很多有意义的信息,为政府的决策提供参考依据。

本文基于社交媒体地理数据的特点,从理论研究和实际应用出发,对社交媒体地理数据的获取、存储、处理及应用进行了深入研究和分析,研究内容集中在以下三个方面:

- 1.通过对位置大数据相关理论的研究及对社交媒体地理数据概念及特点的研究,设计实现了对海量社交媒体地理数据获取的方法,利用 API 数据访问接口,采用 Java SDK 开发包,基于 Eclipse 开发平台设计实现了社交媒体地理数据的抓取方法。以社交媒体地理数据中的签到数据的获取为例,通过微博 API 中的位置服务接口和位置地点动态接口获取了上海客源到江苏浙江两省 A 级以上景区出游的签到数据;最终获取到 2013~2015 年在江苏省 A 级以上景区的签到数据达 1,360,011 条记录,在浙江省 A 级以上景区的签到数据达 527,825 条记录。

- 2.其次根据社交媒体地理数据的特点及数据结构,本文采用 SQL Server 数据库和 ArcGIS 的地理空间数据库作为数据的存储媒介,研究了 SQL Server 数据库连接 GeoDatabase 的方法;在数据的预处理过程中,基于 Microsoft Visual Studio 2010 开发平台,采用 C#语言编写了三个签到数据预处理程序,分别为签到数据整合、兴趣点上的签到人数统计和签到时间字段解析,实现海量签到数据的快速处理;最后,通过对 POIID 的清洗、POI 类别筛选、目标数据提取,得到 2013~2015 年上海客源在江苏省 A 级以上景区的签到数据提取量为: 59073 条记录,在浙江省 A 级以上景区的签到数据提取量为: 33176 条记录。

- 3.对获取到的社交媒体地理数据进行应用实例挖掘分析,分别从时间、空间

上对数据进行挖掘分析。从时间角度出发,通过对获取到的签到数据进行筛选、统计后,分别对 2013~2015 年上海客源到江苏浙江两省 A 级以上景区出游的签到数据进行年际变化特征分析,节假日变化特征分析及节假日、周末和工作日的对比变化特征分析。从空间角度出发,利用 ArcGIS 空间分析方法,通过对 2013~2015 年上海客源到江苏浙江两省 A 级以上景区出游的签到数据做核密度分析,探索上海客源感兴趣的旅游热点区域,再通过对节假日出游的空间地理流量、流向挖掘,分析上海客源在黄金周、小长假、双休日的出游模式及出游特点,为人们智慧出行、智慧旅游的开发提供依据。

关键字: 社交媒体地理数据; 空间数据挖掘; API; 核密度分析

<http://www.ixueshu.com>

Abstract

The rapid development of mobile Internet technology and the widespread popularity of mobile terminal equipment, people's growing demand for spatial information services. The combination of spatial information services, wireless mobile services and social networking services resulted in massive social media geography data. Social media geography data as an important part of the era of large-scale geographic information, has its own unique nature, including space, time, semantics and other attribute information, data update speed, large amount of data, and user activities are closely related, rich information, So that compared with the traditional geographical data, contains a greater potential for development and application value. The user's long-term behavioral operation data can discover the user's behavior pattern, the cognitive user's preference, the real reproduction of the user's life trajectory in the real world, and help the cognitive user's behavior pattern; from the perspective of providing the location- User activities and travel to provide better awareness, but also can help operators to improve service levels and quality; the other hand, according to the user's attention to find the user concerned about the city hot spots, for different locations, different times the user travel needs Develop location-based services. For the government and society, these data are the real sampling of human activities. From these data can be analyzed a lot of meaningful information for the government to provide a basis for decision-making.

Based on the characteristics of social media geography data, this paper makes a deep research and analysis on the acquisition, storage, processing and application of social media geography data from theoretical research and practical application. The research contents are concentrated in the following three aspects:

1. Through the study of the theory of large number of positions and the study of the concept and characteristics of social media geography data, the method of obtaining the geographic data of social media is designed and implemented. Using API data access interface, using Java SDK development kit, The development platform design implements the method of crawling social media geography data. Taking the access data of social media geography data as an example, through the location service interface and location location dynamic interface in the microblogging API, the entry

data of Shanghai guest source to Jiangsu and Zhejiang provinces above Grade A and above are obtained. 2013 ~ 2015 in Jiangsu Province above the A-level area of the check-in data of 1,360,011 records, in Zhejiang Province above the A-level area of the check-in data reached 527,825 records.

2. Secondly, according to the characteristics and data structure of social media geography data, this paper uses SQL Server database and ArcGIS geospatial database as the storage medium of data, and studies the method of connecting GeoDatabase to SQL Server database. In the process of data preprocessing, Studio 2010 development platform, the use of C # language prepared three check-in data preprocessing procedures, respectively, for the registration of data integration, the number of points on the attendance statistics and check-in time field analysis, to achieve rapid processing of mass check-in data; Finally, through the POIID Of the cleaning, POI category screening, the target data extraction, from 2013 to 2015 Shanghai tourists in Jiangsu Province above the A level area of the check-in data extraction volume: 59073 records, in Zhejiang Province above the A-level area of the sign data extraction : 33176 records.

3. The application of the social media data to the mining of examples of mining analysis, respectively, from the time and space on the data mining analysis. From the time point of view, through the access to the check-in data to filter, after the statistics, respectively, 2013 ~ 2015 Shanghai tourists to Jiangsu and Zhejiang provinces above the A-level scenic spots to visit the data of the interannual change characteristics of the characteristics of holiday changes Analysis and analysis of changes in holidays, weekends and working days. From the spatial point of view, the use of ArcGIS spatial analysis method, through the 2013 ~ 2015 Shanghai tourists to Jiangsu and Zhejiang provinces above the A-level scenic area travel 92279 records to do nuclear density analysis, to explore the Shanghai tourist source of interest in tourism This paper analyzes the characteristics of Shanghai Jiaoyuan in the Golden Week, small holiday and weekend, and provides the basis for the development of wisdom and wisdom tourism.

Key Words: location-based social network; Spatial data mining; API; nuclear density analysis

目录

摘要.....	III
Abstract.....	III
目录.....	V
第 1 章 绪论.....	1
1.1 研究背景及意义.....	1
1.1.1 研究背景.....	1
1.1.2 研究意义.....	2
1.2 国内外研究现状.....	3
1.2.1 位置大数据获取与分析处理研究进展.....	3
1.2.2 社交媒体地理数据分析处理研究现状.....	5
1.2.3 居民出行时空特征研究现状.....	6
1.2.4 问题的提出.....	7
1.3 研究目标与研究内容.....	8
1.4 论文结构与技术路线.....	8
第 2 章 相关理论研究.....	10
2.1 位置大数据.....	10
2.1.1 位置大数据的定义.....	10
2.1.2 位置大数据的分类.....	10
2.1.3 位置大数据的研究框架.....	12
2.2 社交媒体地理数据.....	13
2.2.1 社交媒体地理数据的概念.....	13
2.2.2 社交媒体地理数据的特点.....	13
2.2.3 社交媒体地理数据的应用.....	14
2.3 空间数据挖掘及空间聚类方法选择.....	15
2.3.1 空间数据挖掘方法.....	15
2.3.2 空间聚类算法选择.....	16
第 3 章 社交媒体地理数据的获取及预处理.....	18
3.1 社交媒体地理数据的获取.....	18
3.1.1 社交媒体地理数据获取来源选择.....	18
3.1.2 社交媒体地理数据获取原理.....	19
3.2 社交媒体地理数据采集方法设计.....	22
3.2.1 基础数据的准备.....	22
3.2.2 API 接口选择.....	24

3.2.3 获取 POIID 时请求参数 range 的范围界定	24
3.2.4 POIID 的获取实现	25
3.2.5 签到数据的获取实现	26
3.3 社交媒体地理数据的存储与管理	28
3.3.1 SQL Server 数据库	28
3.3.2 Geodatabase 数据库	29
3.4 社交媒体地理数据预处理	30
3.4.1 景区签到数据整合	30
3.4.2 兴趣点上的签到人数统计	30
3.4.3 签到时间字段解析	30
3.4.4 数据的清洗和筛选	31
3.4.5 目标数据提取	31
3.5 数据质量验证	32
3.6 数据采集结果可视化	33
第 4 章 社交媒体地理数据挖掘应用实例分析	36
4.1 签到数据的时间特征	36
4.1.1 年际变化特征	36
4.1.2 节假日变化特征	37
4.1.3 节假日、周末和工作日的对比变化特征	38
4.2 签到数据的空间特征	39
4.2.1 旅游目的地的整体热区探索	39
4.2.2 旅游目的地的局部热区探索	40
4.3 节假日出游的空间地理流向流量挖掘	47
4.3.1 黄金周出游的空间地理流量流向分析	47
4.3.2 小长假出游的空间地理流量流向分析	49
4.3.3 双休日出游的空间地理流量流向分析	54
4.3.4 节假日与双休日出游的空间地理流量流向对比分析	56
第 5 章 结论与展望	57
5.1 论文研究成果	57
5.2 论文研究的不足之处及展望	58
参考文献	59
攻读学位期间取得的研究成果	62
致谢	63
论文独创性声明	64
论文使用授权声明	64

第1章 绪论

1.1 研究背景及意义

1.1.1 研究背景

计算机技术在信息领域的应用及其发展,催生了地理信息系统的出现,进一步解决了地理信息在获取、存储、转换、管理、分析与应用中的难题^[1],而网络技术与地理信息系统的结合,极大地扩展了地理信息的应用,改变了地理信息的服务模式,使地理信息应用对象由原来的政府机关、高校、大型企业和科研院所扩展到普通大众。网络中地理信息的应用,已经成为人类日常生活必不可少的一部分。2015年3月5日十二届全国人大三次会议上,李克强总理在政府工作报告中首次提出“互联网+”行动计划。地理信息行业推出“互联网+地理信息”创新驱动发展战略,按照服务化、移动化、融合化、集成化、平台化、开放化的思路,深入开展地理信息技术与移动互联网、物联网、云计算、大数据等技术的融合,推动地理信息资源的开发、开放、共享、服务,持续推动地理信息产业的新发展,立足全社会信息化建设需求,倡导创新、协调、开放、绿色、共享理念,深入挖掘地理信息资源,最终的目的是让人人享有地理信息带来的服务。

Web2.0时代的到来,由用户主导生成内容的各种应用随之发展起来,其中最引人注目的是社交网络服务(Social Network Services, SNS)。从国外的Facebook、Twitter到中国的微博、QQ、微信已成为用户量最大、传播范围最广、商业价值最高的互联网应用。近年来,随着移动定位设备的普及和基于位置服务的兴起,在地理信息系统(Geographic Information System, GIS)、全球定位系统(Global Positioning System, GPS)、基于位置的服务(Location Based Service, LBS)、地图可视化(cartographic visualization)等技术的发展与支持下,传统社交网络、定位技术与移动GIS技术的融合产生了一种新型的在线社交媒体——基于地理位置的社交网络(Location Based Social Network, LBSN)。LBSN是位置和社交的融合,它支持用户随时随地记录并分享自己的地理位置信息。在LBSN中,最重要的单位是用户和位置,用户与位置关联产生社交媒体地理数据。个人的社交媒体地理数据可以表示个人的历史移动轨迹,大量用户的社交媒体地理数据可以揭示人类的移动模式和生活规律。LBSN与传统社交网络最大的不同是它增加了一个新的维度——空间,可以跟踪和共享人的位置信息,使得虚拟世

界中的社交网络回到了现实世界。

移动互联网技术的飞速发展以及移动终端设备的广泛普及,人们对空间信息服务的需求日益增大。空间信息服务、无线移动服务和社交网络服务的结合,使得移动 GIS 技术快速发展。根据美国皮尤研究中心 2012 年 5 月的调查报告显示,美国 74% 的智能手机用户使用手机获取基于地理位置的实时信息服务。报告同时指出,18% 的美国智能手机用户使用 Foursquare 等地理社交服务网站。所谓移动 GIS 是指用户(终端设备)处于移动情况下使用的地理信息系统,它是伴随 PDA、智能手机、车载移动终端等智能移动设备的出现而产生的。另外移动 GIS 的不同用户通过互动交流与信息沟通,已在地图网站形成了自己的虚拟社区。谷歌地图、必应地图、雅虎地图等,成为了吸引数百万用户的虚拟社区。通过用户生成内容(User Generated Content, UGC)机制,在线地图已经逐渐成为一个进行公共信息交流的平台,各地公众公开表达他们的意见,关注实时新闻,共享感兴趣的事件。这也很好地突出了移动 GIS 作为社会化媒体的角色。

随着科技的进步与数据的累积,2008 年,Miller HJ 和 Han J 表示地理信息数据已经走向“海量数据”时代^[2],随后麦肯锡公司提出大数据时代已经到来。在北京城市实验室(BCL)2015 年会上中科院地理科学与资源研究所的王江浩博士首次阐述了社交媒体地理的概念,提出社交媒体地理是互联网时代社交媒体与地理位置服务结合体的扩展与延生,基于社交媒体与地理位置服务的应用产生了海量的社交媒体地理大数据。与此同时,随着 T-GIS(Time GIS)的发展与应用,有效解决了时间、空间和属性三要素之间的相互作用关系,并有利于揭示地理要素的状态和变化过程^[3],可以更好地表达人们行为的时空关系,研究人们在不同时间点的空间分布,揭示群体日常活动的时空特征。

1.1.2 研究意义

社交媒体地理数据作为大数据时代地理信息的重要组成部分,具有自身独特的性质,它包含空间、时间、语义等属性信息,数据更新速度快,数据量大,与用户活动息息相关,信息量丰富等,这些特点使得社交媒体地理数据的挖掘研究及应用发展具有重要的研究意义与实践价值。目前国内外尚处于社交媒体地理数据的挖掘研究与应用发展的初步探索阶段。对于做传统空间分析的人来说,常规的研究方法就是用遥感、统计数据、问卷等得到数据源,然后对他们进行空间化,在 GIS 系统里进行空间数据建模。这种方法有一个弊端,难以获取用户活动信息。针对这一问题,本研究探索了新的数据源——社交媒体地理数据,并设计实现了海量社交媒体地理数据获取的方法及数据的存储方式。

理论意义：社交媒体地理数据的出现为地理信息学科的发展提供了一个新的方向，本研究对社交媒体地理数据的研究充实和丰富了国内位置大数据的研究体系，加深人们对社交媒体地理数据的认识。同时，通过研究游客外显的时空行为，在精确刻画的基础上帮助研究者理解某一类客源的出游模式，丰富以客源为出发点的相关研究视角。

实践价值：社交媒体地理数据作为大数据时代地理信息的重要组成部分，相比于传统地理数据，蕴含较大的发展潜力与应用价值。得到用户长期的行为操作数据可以发掘用户的行为模式，认知用户的偏好，真实再现用户在现实世界中的生活轨迹，帮助认知用户的行为模式。这些数据是对人类活动的真实采样，从这些数据中可以分析出很多有意义的信息，为政府的决策提供参考依据。

1.2 国内外研究现状

1.2.1 位置大数据获取与分析处理研究进展

大数据将在政府公共服务、医疗服务、零售业、制造业以及涉及个人位置服务等领域得到广泛应用，并产生巨大的社会价值和产业空间。TMR（透明度市场研究）最新发布的报告《大数据市场：2012~2018 年全球形势、发展趋势预测》指出，2012 年全球大数据市场产值为 63 亿美元，预计 2018 年该产值将达 483 亿美元。位置大数据（location big data, LBD）是大数据中一个重要的组成部分。随着位置服务应用的不断普及，由地理信息数据、轨迹数据和社交媒体地理数据等构成的位置大数据已经成为当前用来感知人类社群活动规律，分析地理国情和构建智慧城市的重要战略资源。

通过对位置大数据的处理分析，使得传统测绘强调的物理世界的测量结果（即位置）可引申到对人类社会的某些动态情况量测中去，这极大地促进了当代计算机科学技术、数据科学技术与测绘科学技术的融合，用户在任何地点、任何时间为认知自然和社会环境与人的关系而创建和使用地图的活动。它强调人、环境等地理信息的社会属性，强调人的活动与数据融合。通过对位置大数据进行适当的处理以及社会计算，提供个性化的、实时的、动态的位置信息。每个人都可以是位置数据的提供者，同时也可以成为位置服务的受众^[4]。

早期的地理信息应用服务主要是面向行业，用户群体范围已知，特征明显，主要是通过一对一的用户访谈或者是焦点小组的方法，直接获取用户的需求。随着地理信息服务应用的扩大及普及，用户规模扩大，用户群体呈现出多样化趋势，

单纯依靠专家访谈、焦点小组获取用户需求已经无法满足研究需要,覆盖范围更广、成本更低、更容易实施的问卷调查法、现场观察法、卡片排序和用户测试等多种用户研究方法逐渐成为用户数据获取的主要方法。这些方法在网络环境下迅速与网络技术相结合,产生了网络观察法、网络调查法等,进一步扩展了研究的范围,提高了效率。上个世纪 90 年代末,美国地图学者 Mark Harrower 采用传统的用户测试方法,对 16 个专业的地理学家和非专业人员进行网络地图用户测试^[5]。Muki Haklay 采用线上的问卷调查法,共收集到了 385 份有效问卷,评估 GIS 应用程序的可用性的技术和方法^[6]。Jacobs 等人^[7]通过研究户外固定摄像头拍摄的图像与卫星图像之间的关联度来确定摄像头的地理位反之根据摄像头拍摄的图像与卫星图像之间的关联度信息,通过摄像头图像生成卫星图像。文献^[8]通过 Google 的卫星地图和 Google 的街景图片来估计带有 GPS 位置的照片的朝向信息。Cristani 等人的研究工作^[9]利用主题模型对照片的视觉内容进行分析,将照片内容划分成若干与土地覆盖相关的子类,如城市、田地、山川、湖泊等,再结合照片的地理位置信息,就能得到土地覆盖子类的空间分布信息。Leung 等人^[10]利用照片分享网站中带有 GPS 标签的照片,通过抽取边缘信息作为照片的视觉特征,并采用监督学习的方式来预测土地的使用情况,使用情况分为两类:已开发的土地和未被开发的土地。文献^[11]对 GPS 轨迹数据进行处理,提取出用户的停留点。然后对停留点进行层次聚类,得到一个图结构,越高的层次中的聚类包含的停留点越多,代表的空间区域也越大。利用类似于 HITS 的算法可以对这些停留点和区域进行排序,从而得到兴趣点、景点和热门区域。

文献^[12]在从 GPS 轨迹中挖掘兴趣点的基础之上,进一步研究了兴趣点和活动的关联。由于原始的“地点-活动”的标签很稀疏,文献使用了外部网页数据来获取“活动-活动”之间的关联。最后通过协同矩阵分解的方法,利用低维特征重构原始矩阵,以解决原始矩阵稀疏问题。文献^[13]从出租车 GPS 轨迹中挖掘人的日常交通模式,并结合地点的属性信息,构建了一个话题模型来挖掘区域的功能属性。王建华等通过发放问卷的方式,对全国范围内的地图用户进行社会调查,分析了目前地图用户使用地图的基本情况,归纳和总结了地图用户的意见和建议^[14];文献^[15]通过问卷调查研究了用户对主流网络地图的满意程度及其使用偏好,提出了网络地图的改进方案。文献^[16]结合移动电子地图的功能和特点,分析影响移动电子地图使用的因素,并基于模糊综合评测方法构建移动电子地图评价模型;通过对移动电子地图进行评价,发现目前移动电子地图存在的主要问题。游万来等人通过对网络地图的用户界面,导航模式以及缩放和漫游操作方法进行可用性测试,系统科学地获取和分析了用户操作网络地图的行为数据^[17-18]。凌云等人通过电子地图用户界面认知实验的方法获取用户的基本认知特征^[19]。郑束

蕾等人对地图学实验研究科学活动三要素进行解析,介绍了问卷调查、眼动实验、出声思维等目前常用的各种地图学认知实验方法,并对眼动实验在地图学研究中的作用进行了归纳分析^[20]。

1.2.2 社交媒体地理数据分析处理研究现状

以 Twitter、微博为代表的社交应用带动了自媒体的发展,使得每个人都能随时随地发布自己的见闻和感想。由于这些社交内容与时间和地点高度相关,而且实时性非常好,所以社交应用的用户可以看作是分布在各地的传感器,他们源源不断地将真实世界中正在发生的事情采集下来并发布出去。一些研究工作通过挖掘这些“众包的人类传感器”所产生的数据,进行实时的事件检测以及事件的发展趋势预测,例如犯罪预测,地震等突发事件的检测和报告,以及流行性感冒的预测等。

Gerber 的研究工作^[21]通过对带有 GPS 标签的 Twitter 数据进行分析进行犯罪预测。Gerber 利用主题模型对 Twitter 的内容进行分析,并对芝加哥地区的犯罪进行了预测。实验结果表明,相对于传统的基于概率密度估计的预测方法,加入了 Twitter 内容特征之后,在研究的 25 种犯罪中,有 19 种犯罪的预测准确率得到了提升。

文献^[22]从微博中检测与流行性感冒相关的微博,并通过各种回归模型建立微博中关键词的词频与疾病控制和预防中心发布的官方数据之间的关联性,尝试用该模型检测流行性感冒的发生。文献^[23]通过建立回归模型,利用与感冒症状相关词汇的微博来预测感冒的发病率,该方法在英国不同地区的 H1N1 流感数据上都获得了显著的效果。文献^[24]基于微博数据在更大广度上对公众健康进行了研究。这项研究工作从微博数据中挖掘了过敏、抑郁、癌症、肥胖、流感等多种疾病相关的知识,涉及了疾病的症状、治疗方式、空间分布,时间分布等各个方面,可以应用于疾病的监测,发展趋势预测,风险控制,症状和治疗手段分析等等。

一些研究对含有地理位置信息的媒体中的位置信息的抽取或优化位置的精度,或者利用这些数据对没有地理位置信息的媒体或者用户进行位置的预测和推断。文献^[25]通过分析微博内容中与地点关联的词汇,来推测微博的位置信息。通过对微博中词汇的地理位置分布建立概率模型,并挖掘微博中用户相互关注关系中所隐含的地理位置的远近关系,结合微博内容以及微博用户这两方面因素,可以显著地提高微博位置推测的准确率。文献^[26]利用博客中的照片和地点信息,建立了一个包含“用户—景点—照片”的层次化的图结构。然后在图结构上利用类

似于 HITS 的算法,对景点、用户、照片进行排序。文献^[27]利用签到数据挖掘地点的类别属性,如饮食、购物等等。通过用户的签到历史数据,来计算各个地点之间的相关度。最后,利用从单个地点中提取的特征和从多个地点相关度中提取的特征,为每个地点类别分别建立了分类器,用来对地点的类别进行分类。文献^[28]利用博客中对餐馆的评价,使用命名实体识别的方法来挖掘菜品名称,处理了超过 12,000 条中国餐馆评论,并生成了“美食地图”,在地图视图中显示最受欢迎的餐馆,以帮助用户做出良好的购买决定。Gioni 等人的研究工作^[29]通过挖掘签到类社交应用的数据进行城市内部的地点线路推荐,这项工作提出的线路推荐算法考虑了地点的种类、访问地点的顺序、空间和时间约束、以及访问地点带来的满意度。袁书寒利用开源的 Heritrix 网络爬虫获取了嘀咕网中位置签到数据,计算了基于地理位置坐标的用户相似度,研究了基于语义的位置服务社交网络中用户相似性分析方法。王丽文介绍了利用 Python 调用 API 提取数据和 Python 模拟登陆新浪微博提取热门话题的方法,并对参与热门话题的用户分级聚类,根据用户话题的相似度进行推荐^[30]。

1.2.3 居民出行时空特征研究现状

城市日常生活的许多问题都与居民出行行为相关,近年来居民出行行为的分析在城市道路规划中越发显得重要,并逐渐成为城市道路交通建设的一个重要参考依据。

国外对于居民出行行为相关分析的研究较早。早期的城市居民出行调查主要是为了给城市道路交通的建设提供参考信息,大多采用入户访谈调查的方式^[31]。美国从 1943 年到 1958 年期间就已经对超过 100 个城市的居民出行(origin and destination)情况进行了统计调查,Curran 和 Stegmaier (1958)分析了其中 50 个城市的居民活动调查数据^[32]。

90 年代以后,随着时间地理学的兴起、时间地理学与 GIS 技术的结合以及 GPS 技术在民用方面的大力发展,更是给居民出行行为相关研究带来了新的突破。Murakam 通过征集志愿者,在其私家车上安装 GPS 和 PDA 结合的装置进行家庭出行调查^[33]。Wolf 探讨了利用 GPS 数据记录器取代出行日志的调查方式来研究居民出行行为^[34]。Kthe paperlla 提出了浮动车数据的概念,并利用其构建了道路交通网络通行量预测模型,为今后基于浮动车数据的相关研究做出有益的探索^[35]。Kwan 探讨了 GIS 地理空间计算方法以及三维可视化的方式来展现了居民出行行为模式,并通过实际案例验证了该方法的有效性^[36]。Laube 采用时空数据挖掘方法,对志愿者发放 GPS 装置,对采集到的志愿者出行轨迹点数据进行分

析,挖掘有意义的出行活动时空特征^[37]。Yu 和 Shaw 开发出在时间与空间维度表达和分析个体活动特性,具有探索性数据分析功能的活动模式分析器^[38]。BAZZANI 通过车载 GPS 数据对意大利佛罗伦萨市私家车驾驶者的出行行为进行实证统计,发现驾驶者的出行距离服从指数分布^[39]。

我国对居民出行行为相关研究从上世纪 90 年代开始,如杨涛、王琳等以马鞍山市的城市居民为研究对象,通过大量咨询调查,再结合已有的资料,研究了该城市居民出行行为背后的心理学因素^[40]。本世纪初至今,随着我国 GIS 技术的飞速发展与广泛应用、GIS 技术与时间地理学的结合使得出行行为时空特征相关研究课题逐渐兴起。路培聪、曲大义等,总结了居民出行的基本特征和原因,并为城市交通规划提供意见和建议^[41-42]。白永平、张艳萍通过调查问卷对兰州市居民购物行为时空特征进行了分析^[43]。周素红、邓丽芳利用 T-GIS 技术,通过出行日志数据,研究了广州居民日常活动的时空特征和社会阶层日常活动的时空分异^[44]。申悦,柴彦威结合时间地理学方法与 GIS 三维可视化技术,对深圳居民日常活动的时空特征进行描述^[45]。张艳,柴彦威借鉴国外时间地理学对微观个体日常活动的分析框架,对北京城市中低收入者日常活动的时空特征进行系统分析^[46]。郭文伯利用北京市居民活动 GPS 调查数据,对城市郊区居民日常活动时空特征进行分析和比较^[47]。

1.2.4 问题的提出

1、通过以上对位置大数据的研究分析可以发现对于做传统空间分析的人来说,常规的研究方法就是用遥感、统计数据、问卷等得到数据源,然后对他们进行空间化,在 GIS 系统里进行空间数据建模。这种方法有一个弊端,难以获取人(即用户)的信息。对于大数据热及大数据特点,以往的位置大数据获取只注重量而不注重质,且对于获取到的数据分析挖掘不深入。

2、通过对社交媒体地理数据的研究分析发现,当前网络地理信息中的用户行为数据获取和分析方法仍然以传统的问卷调查和实验室环境中的测试为主,对 Web 挖掘方法的应用较少,获取数据内容简单,体量较小。目前通过获取网络用户信息分析用户行为的挖掘研究逐渐成为主流,但研究领域多集中在网络安全、电子商务化及网络应用等领域,针对网络地理信息应用中的用户行为研究较少,尚不成体系。

3、通过对居民出行时空特征的研究分析发现,以往的研究常常是通过出行问卷调查、分析土地利用、人口密度等社会经济统计数据,该数据收集方法费时费力、无法获取居民具体出行路线、不能完整的反映出行时空信息,或者基于有

限的个人 GPS 与活动日志数据进行分析,但这种方式多集中于对微观个体日常活动进行个体行为时空分析,而很少以宏观的视角,挖掘大量移动对象群体轨迹背后隐藏的信息。

1.3 研究目标与研究内容

本文首先阐述了研究社交媒体地理数据的时代背景和现实意义,针对社交媒体地理数据的特点,详细研究了获取社交媒体地理数据的方法,并结合社交媒体地理数据的数据结构以及运用环境,选择 SQL Server 数据库和 GeoDatabase 作为社交媒体地理数据的存储媒介,实现海量社交媒体地理数据的快速高效存储。

其次,设计实现了三个签到数据预处理中间程序,分别为签到数据整合、兴趣点上的签到人数统计和签到时间字段解析,实现海量签到数据的快速处理。通过多层次数据清洗和数据筛选方法的设计实现,得到较高精度的社交媒体地理数据。根据社交媒体地理数据所含信息,研究了空间大数据独特的筛选方式。

最后,对获取到的社交媒体地理数据进行应用实例挖掘分析,通过提取上海客源到江苏浙江 A 级以上景区出游的社交媒体地理数据,分析上海客源到江苏浙江的旅游热点分布和节假日出游的空间地理流量、流向的变化。为上海客源旅游出行提供辅助决策信息,为社交媒体地理数据的挖掘与应用研究提供参考。

1.4 论文结构与技术路线

根据研究内容及实际情况,本文共分为五个章节,每章内容安排如下:

第一章为绪论,主要介绍研究背景与意义,国内外研究现状,研究内容和目标和论文结构与技术路线。

第二章分别对位置大数据研究框架、社交媒体地理数据和时空数据挖掘的相关理论进行研究和讨论。

第三章分析了社交媒体地理数据获取的原理,设计实现了对海量社交媒体地理数据获取的方法,以社交媒体地理数据中的签到数据的获取为例,通过微博 API 的位置服务接口和位置地点动态接口获取了上海客源到江苏浙江两省 A 级以上景区出游的签到数据。并研究了海量社交媒体地理数据存储的方法;设计实现了三个签到数据预处理中间程序,分别为签到数据整合、兴趣点上的签到人数统计和签到时间字段解析,实现海量签到数据的快速处理;完成了数据的清洗和验证。

第四章分别从时间、空间两方面对获取到的社交媒体地理数据进行应用实例挖掘分析，得出了上海客源到江苏浙江出游的签到数据的年际变化特征，节假日变化特征，及节假日、周末和工作日的对比变化特征。在空间上，利用 ArcGIS 空间分析方法，探索了上海客源感兴趣的热点区域，及黄金周、小长假、双休日期间的出游模式及出游特点，为人们的智慧出行提供依据。

第五章论述本文的研究结论及创新点，并总结论文研究中的不足与展望。

选题背景及意义

位置大数据

社交媒体地理

时空数据挖掘

阅读相关文献了解国内外研究进展

研究内容

社交媒体地理数据采集方法设计实现

社交媒体地理数据存储及预处理

- 数据来源选择
- 基础数据获取
- API 接口选择
- 请求参数 range 范围确定
- 获取目的地 POIID
- 获取社交媒体地理数据

- 数据导入 SQL Server
- 数据导入 Geo Database
- 数据预处理方法设计实现
- 数据清洗和筛选
- 数据验证
- 目标数据提取

社交媒体地理数据挖掘应用实例分析

目标客源出游的时空特征分析

- 签到数据的时序变化特征
- 签到数据的热区探索

GIS

客源节假日出游空间地理流量流向挖掘

- 黄金周出游空间地理流量流向分析
- 小长假出游空间地理流量流向分析
- 双休日出游空间地理流量流向分析

结论

图 1 - 1 技术路线图

第2章 相关理论研究

2.1 位置大数据

近年来,随着移动互联网技术的快速发展、网络地理信息系统和基于位置社的服务技术的应用,产生了类型众多、数据量丰富的各种位置大数据,基础地理数据、轨迹数据和社交媒体地理数据等是构成位置大数据的主要数据类型,此类数据的研究不仅是大数据研究领域中的重要组成部分,还是用来分析人类社会生活规律、出行特点及构建智慧城市的重要组成部分。

2.1.1 位置大数据的定义

现有的位置大数据来源很多,主要来源于移动社交媒体的位置服务、车联网等新兴互联网应用,针对于这一类位置大数据的数据特点及数据结构,郭迟等人对位置大数据定义给出了如下描述^[48]:

定义 1.位置数据集记为: $LBD=\{O,T,P\}$, 其中 $O=\{o_1,o_2,\dots\}$ 表示该数据集中的移动对象集合,包括了 $|O|$ 个产生位置的移动目标; T 为观察数据集的时间; $|T|$ 天内总共获得 $|P|$ 个位置记录。

定义 2.单个位置数据记录 p 主要包含移动目标 o 和位置的地理坐标 (x, y) 和记录时刻 t , 可以用一个四元或五元组表示。如果是车辆轨迹数据,一般还包含车辆的速度 v 以及一组状态信息 $S=(S_1,S_2,\dots)$, 如行驶方向、油耗值、载客状态等, 记为 $p=(o, x, y, t, v, S)$, 其中, 一个具体的状态 S_i 可能有多个状态取值; 如果是用户在社交网络等媒体上主动分享的位置数据, 则还包括与位置相关的媒体信息 I , 可记为 $p=(o, x, y, t, I)$, 一般地, 将移动目标 o_i 的第 j 条位置记录记为, 在不影响理解的情况下也可直接写作。

2.1.2 位置大数据的分类

位置大数据的来源和类型很多,在本文研究中根据数据的特点将位置大数据分为三个类别:地理信息数据、轨迹数据和社交媒体地理数据。各类型实例及数据体量举例如表 2-1 所示。

1、地理信息数据。地理信息是指与地理环境要素有关的物质的属性、数量、空间关系、关联和规律等的总称,它的表现形式和信息的承载方式多种多样,可

以是文字、图形、图像、数字、声音、视频等；地理信息包含了信息所具有的特性外，还有自己的独特性质，包括区域性、多维性、动态性，区域性是指按照特定的经纬网或公里网建立的地理坐标来实现空间位置的识别，并可以按照指定的区域进行信息的并或分；多维性是指地理信息可以是二维空间信息，也能加以其它的数据指标实现地理信息的三维展示；动态性主要是指地理信息在时间序列中的变化特征，不同时间段的地理信息体现的信息价值有所不同，比对不同时间段的地理信息数据，可以及时的更新地理信息数据，挖掘数据中隐藏的规律，也能对未来做出预测和预报。地理信息数据无处不在，已经深入我们生活的方方面面，给我们的工作生活带来了极大的便利，正以电子地图、卫星导航、遥感影像等地理信息产业创造着奇迹^[49]。

2、轨迹数据。轨迹数据是指在一定的时间、空间环境下，对单个或者多个运动对象进行采样所得到的数据，数据信息包括移动对象在采样点的空间位置信息、时间信息和移动速度等，多个采样点信息按照移动的先后顺序形成轨迹数据。数据来源主要有智能手机产生的位置数据；交通轨迹数据，例如公交刷卡数据、出租车轨迹数据等；还有物流大数据等等。

3、社交媒体地理数据。社交媒体地理数据是指由移动社交网络与位置服务技术结合产生的新型网络应用，由社交媒体地理应用产生的社交媒体地理数据包含空间、时间、语义等丰富的用户信息，数据来源众多，只要是安装了具有位置服务的社交应用的智能移动终端设备都能产生社交媒体地理数据；数据结构多样，由于该类型的数据产生设备来源众多，数据产生形式丰富多样，可以是含位置因素的文字、图片、声音、视频和动画等，所以数据结构可以是结构化数据、半结构化数据和非结构化数据等，数据的实时性较强，更新速度快。

随着社交网络应用的快速发展以及大众对社交网络应用的极度青睐，当前市场上出现了各类型基于位置服务的社交网络应用，根据表现位置信息方式的不同，基本上将含位置服务的社交媒体应用分为以下两类：

- 由媒体内容表示位置信息

由媒体内容表示位置信息产生的位置数据主要是基于地理标记的媒体内容，用户通过智能终端设备将现实世界中获取到的带有位置信息的媒体内容以文字、照片或视频的方式上传到社交网络中，用以供其社交圈中的好友进行浏览、评论或点赞其媒体内容。从该类型社交媒体地理应用中提取到的用户位置信息、时间信息和语义信息数据多为半结构化数据或非结构化数据，信息量丰富，但信息提取困难，网络技术服务焦点主要为媒体内容，位置服务只是丰富媒体内容的附加因素。

- 由位置点表示位置信息

由位置点表示位置信息的位置数据可以直接提取到基于位置点的信息,用户通过智能移动终端设备对感兴趣的特定地点进行“签到”或分享位置信息,例如用户在旅游的过程中遇到自己喜欢的景点或餐厅、酒店等兴趣点,就可以实时进行签到,通过这样的签到信息,使用该用户周围的朋友可以发现该用户,该用户也可以发现其周围一定范围内的朋友,从而进行下一步的社交活动,也可以通过该位置点的签到数、评论和点赞数来发现出游热点,从而给其它用户提供建议。从该类社交媒体地理应用中提取到的用户位置信息和时间信息数据多为结构化数据,语义信息数据多为半结构化数据或非结构化数据。

表 2-1 位置大数据实例

位置大数据类型	实例	体量举例
地理信息数据	数字矢量线画地图 (DLG)	全国 1:5 万 DLG 有 250GB, DOM 有 10TB, 1:1 万 DLG 约 5.3TB, DOM 约 350TB
	数字栅格地图 (DRG)	
	数字正射影像地图 (DOM)	
	数字高程模型	
轨迹数据	手机信令数据	北京市 12000 辆出租车 110d 产生 577000000 条轨迹记录
	出租车轨迹数据	
	公交刷卡数据	
社交媒体地理数据	Twitter、微博等的签到数据	Twitter63261 用户 30d 产生 15944084 条位置签到记录

2.1.3 位置大数据的研究框架

位置大数据属于大数据研究范畴,同时也属于空间数据研究领域,除了具有大数据所具有的特点外,还有其自身的特点,自有一套研究框架,主要包含数据的采集、预处理、数据计算、数据存储和数据可视化等一系列过程^[50]。

1) 数据采集

不同类型的位置大数据,数据的采集方法也不同,对于地理信息数据,主要依靠带有各类型传感器的卫星、飞机进行采集,还有各种地面采集仪器,例如全站仪、GPS 数据采集车等进行地理信息数据的采集;对于轨迹数据的采集,主要是通过已经建立的一些交通系统例如地铁收费系统、公交刷卡系统和出租车管理系统等对用户信息进行采集;对于社交媒体地理数据的采集,主要通过信息检索、网络爬虫、调用 API 接口等方法来提取社交媒体中用户的时空信息。

2) 数据处理分析

根据位置大数据的应用需求,选择相应的位置大数据处理分析方法,数据处理方法需要按照一定的标准进行,应用于政府、企事业单位,来源于地理空间信息行业的位置大数据需采用国家标准进行处理,对于近年来新产生的轨迹数据、

社交媒体地理数据等新型位置大数据类型，需采用非标准方法处理。

位置大数据类型多样，数据来源众多，数据采集方法及采集设备存在误差，用户自发上传的数据，根据用户使用终端的不同以及用户个人习惯，所产生的数据也是多种多样，所以不管以何种方式获取到的位置大数据，都存在一定的误差和错误，所以在数据使用前，对数据进行预处理是很重要的步骤，具体包括对数据的清洗、筛选、补全等。

3) 计算和存储

综合利用计算机语言及 Hadoop 等计算框架，建立高效的时空索引，采用分布式计算方法，综合利用多种时空数据库存储技术。

4) 可视化

因为位置大数据所特有的空间特性及其数据体量大，利用常规的数理统计方法及统计图表，无法准确的反应数据所反映的总体趋势，所以需要运用特定的数据可视化方法对数据进行表达，从而挖掘数据中所具有的深层次信息。

2.2 社交媒体地理数据

2.2.1 社交媒体地理数据的概念

社交媒体地理初期的描述称为基于位置的社交网络（Location Based Social Network，简称 LBSN），Mell 等人给出的正式定义如下表述^[51]：将基于位置服务的技术加入以往的社交网络应用中，让虚拟的社交网络回归现实世界，通过位置共享判断好友关系在现实世界中的相关性，通过一定时间内的位置信息累计，可得出用户活动轨迹、兴趣爱好等。

王江浩^[52]在 2015 年的 BCL 年会上首次阐述了社交媒体地理的概念，提出社交媒体地理是互联网时代社交媒体与地理位置服务结合体的扩展与延生，基于社交媒体与地理位置服务的应用产生了海量的社交媒体地理大数据。它支持用户随时随地记录并分享自己的地理位置信息，最重要的单位是用户和位置，用户与位置关联产生社交媒体地理数据。

2.2.2 社交媒体地理数据的特点

社交媒体地理数据包含了丰富的空间信息、时间信息和语义信息等，在社交网络应用中加入位置服务后，用户更愿意随时随地的分享身边的事件或者环境，空间和时间因素的引入，不但改变了用户使用社交网络的习惯，还丰富了社交网

络中的信息内容,在一定时间段内,随着用户位置信息的增加,可以挖掘用户的活动兴趣区域及活动轨迹等,也可根据大量的位置信息累加,推断活动热点区域。所以该类数据有其自身的特点。

1、数据量大、产生速度快

随着智能手机、平板电脑等移动设备的应用越来越普遍,网络用户大量增加,特别是现在大多数用户对于社交网络的依赖,用户可以随时随地进行数据的签到和分享,使得每天大量的网络数据、社交媒体地理数据大量、迅速的产生。

2、数据结构多样

社交媒体地理数据可以是加入了位置要素的文字、图片、声音、视频和动画等媒体数据,又根据地理信息表达和组成方式的不同分为结构化数据、半结构化数据和非结构化数据,结构化数据例如地理信息系统数据库、兴趣点数据库,半结构化数据例如 Twitter、微博等的签到数据,非结构化数据例如游记、百科等。

3、现势性高

相比于传统的政府及专业机构等的数据更新方式,社交媒体地理数据的更新周期很短且时效性强,例如某个地方发生自然灾害或者交通事故等,用户可以及时的上传到网络,随着网络数据的分享、转载等,可以更为迅速的获取到事发地点的情况,为政府工作提供辅助,使问题能够快速高效的得到解决。

4、数据质量参差不齐

社交媒体地理数据来源很多,每个人都是数据的产生者,也是数据的使用者,数据的产生可以是专业人员产生,也可以是非专业人员自发产生,非专业人员产生的数据比较随意和个性化,且数据产生的设备众多,精度差异大,有的数据被编辑的次数较多等,所以数据质量参差不齐,难以控制。

2.2.3 社交媒体地理数据的应用

基于位置的服务的普及以及大众对于该服务的快速接受,使得近几年来基于位置的社交网络应用快速发展,大量的社交媒体地理数据迅速产生。用户产生数据,行业对数据进行统分析后再次服务于用户。社交媒体地理数据的应用大致可以分为两类:基于用户的应用和基于位置的应用^[53]。

1、基于用户的应用

数据分析的初衷就是服务于社会,服务于大众,从用户的角度出发,数据的应用可体现在以下几个方面:(1) 好友推荐。根据用户之间的相似性度量,为用户推荐有共同兴趣爱好的好友或者根据用户运动轨迹的相似性,为共同的用户推荐兴趣点。(2) 群体挖掘。根据用户的活动习惯及兴趣爱好将用户分为不同的群

体或者通过计算用户在地理空间的相似性将用户分成不同群体。(3) 行为分析。根据用户的活动规律及活动轨迹, 挖掘用户的行为模式, 进行热点区域分析。

2、基于位置的应用

从面向位置的角度出发, 可以将基于位置的应用概括为以下几个方面:(1) 路径发现。可从大量不确定性的轨迹中挖掘两个位置间最可能的路径, 在用户指定的查询下选择出最好的几个路径返回给用户。(2) 商店位置选择。对收集到的社交媒体地理数据进行热点探测和流行性挖掘。(3) 区域功能发现。根据用户一定时间段内的活动轨迹, 区分活动区域中的生活区、工作区、商业区等。(4) 流行位置和流行路径推荐。在旅游活动中, 可根据以往游客的签到兴趣点数量及点评内容提取当地最受欢迎的出游地点或旅游路线。(5) 行程规划。用户可以在给定的起点位置、终点位置、时间等指定条件下找出最合适出行方式等。

2.3 空间数据挖掘及空间聚类方法选择

2.3.1 空间数据挖掘方法

20 世纪 90 年代, “从数据库中发现知识”(KDD) 及其核心技术——数据挖掘应运而生。作为地理信息的载体, 地理空间数据具有不同于一般数据的特质, 即数据的空间和时间上的多尺度性、属性数据的多维结构、数据表达的不确定性等等, 使得基于地理空间数据库的数据挖掘和知识发现特别需要综合知识的支持和结合。空间数据挖掘和知识发现作为数据挖掘的一个新的研究分支, 是指从空间数据库中提取隐含的、用户感兴趣的空间和非空间的模式和具有普遍特征的知识的过程^[54]。

空间数据挖掘和知识发现的过程大致可分为以下多个步骤: 数据准备、数据选择、数据预处理、数据缩减或者数据变换、确定数据挖掘目标、确定知识发现算法、数据挖掘、模式解释、知识评价等。空间数据挖掘汇集了人工智能、机器学习、数据库技术、模式识别、统计学、GIS、可视化等领域的相关技术, 因而空间数据挖掘的方法很多^[55]。下面介绍几种常用的方法。

1、空间分析方法 (Spatial Analysis Approach)

利用 GIS 的各种空间分析模型和空间操作对空间数据库中的数据进行深加工, 从而产生新的信息和知识。目前常用的空间分析方法有综合属性数据分析、拓扑分析、缓冲区分析、密度分析、距离分析、叠置分析、网络分析、地形分析、趋势面分析、预测分析等, 可发现目标在空间上的相连、相邻和共生等关联规则,

或发现目标之间的最短路径、最优路径等辅助决策的知识。空间分析方法常作为预处理和特征提取方法与其它数据挖掘方法结合使用。

2、统计分析方法 (Statistical Analysis Approach)

统计方法一直是分析空间数据的常用方法,着重于空间物体和现象的非空间特性的分析。在运用统计方法进行数据挖掘时,一般并不将数据的空间特性作为限制因子加以考虑,空间数据所描述的事物的具体空间位置在这类挖掘中也并不起制约作用。尽管此种挖掘方式与一般的数据挖掘并无本质的差别,但其挖掘后发现的结果都是以地图形式来描述的,对发现结果的解释也必然要依托地理空间进行,挖掘的结果揭示和反映的必然是空间规律。

3、聚类方法 (Clustering Approach)

聚类是按一定的距离或相似性系数将数据分成一系列相互区分的组,根据定义可以把其分为四类:基于层次的聚类方法;分区聚类算法;基于密度的聚类算法;网格的聚类算法。常用的经典聚类方法有 K-mean, K-medoids, ISODATA 等。

4、数据可视化方法 (Data Visualization Approach)

人类的可视化能力,允许人类对大量抽象的数据进行分析。人的创造性不仅取决于人的逻辑思维,而且取决于人的形象思维。人脑的空间认知分析能力目前尚无法全部用计算机代替,因此可视化技术为知识发现提供了有力的帮助。为了了解数据之间的相互关系及发展趋势,人们可以求助于可视化技术。海量的数据只有通过可视化技术变成图形或图像,才能激发人的形象思维——从表面上看来是杂乱无章的海量数据中找出其中隐藏的规律。数据可视化技术将大量数据以多种形式表示出来,帮助人们寻找数据中的结构、特征、模式、趋势、异常现象或相关关系等。从这个角度讲,数据可视化技术不仅仅是一种计算方法,更是看见不可见事物或现象的一种重要手段和方法。

2.3.2 空间聚类算法选择

空间聚类分析是空间数据挖掘与知识发现的主要手段之一,是指将空间数据集中的对象分成由相似对象组成的类。同类中的对象间具有较高的相似度,而不同类中的对象间差异较大。作为一种无监督的学习方法,空间聚类不需要任何先验知识。空间聚类的主要方法有五大类:划分聚类算法、层次聚类算法、基于密度的方法、基于网格的方法和基于模型的聚类方法。

1、划分聚类算法

主要包括: K-means、K-medoids、IPAM、CLARA、K-模、K-原型、EM 和 CLARANS 等。基本思想: 给定一个包含 n 个对象或数据的集合, 将数据集划分

为 k 个子集，其中每个子集均代表一个聚类 ($k \leq n$)，划分方法首先创建一个初始划分，然后利用循环再定位技术，即通过移动不同划分中的对象来改变划分内容。

2、层次聚类算法

层次聚类方法是通过将数据组织为若干组并形成一个相应的树来进行聚类的，层次聚类方法又可分为自顶向下的分裂算法和自底向上的凝聚算法两种。

分裂聚类算法，首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到每个对象自成一簇，或者达了某个终结条件，这里的终结条件可以是簇的数目，或者是进行合并的阈值。而凝聚聚类算法正好相反，首先将每个对象作为一个簇，然后将相互邻近的合并为一个大簇，直到所有的对象都在一个簇中，或者某个终结条件被满足。

3、基于密度的方法

绝大多数划分方法基于对象之间的距离进行聚类，这样的方法只能发现球状的类。因此，出现了基于密度的聚类方法，其主要思想是：只要邻近区域的密度（对象或数据点的数目）超过某个阈值，就继续聚类，这样的方法可以过滤噪声数据，发现任意形状的类。从而克服基于距离的方法只能发现类圆形聚类的缺点。代表性算法有：DBSCAN 算法、OPTICS 算法、DENCLUE 算法等。

4、基于网格法

主要思想是将空间区域划分若干个具有层次结构的矩形单元，不同层次的单元对应于不同的分辨率网格，把数据集中的所有数据都映射到不同的单元网格中，算法所有的处理都是以单个单元网格为对象，其处理速度要远比以元组为处理对象的效率要高的多。代表性算法有：STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法等。

5、基于模型法

给每一个聚类假定一个模型，然后去寻找能够很好地满足这个模型的数据集。常用的模型主要有两种：一种是统计学的方法，代表性算法是 COBWEB 算法；另一种是神经网络的方法，代表性的算法是竞争学习算法。COBWEB 算法是一种增量概念聚类算法。这种算法不同于传统的聚类方法，它的聚类过程分为两步：首先进行聚类，然后给出特征描述。因此，分类质量不再是单个对象的函数，而且也加入了对聚类结果的特征性描述。竞争学习算法属于神经网络聚类。它采用若干个单元的层次结构，以一种“胜者全取”的方式对系统当前所处理的对象进行竞争。

第3章 社交媒体地理数据的获取及预处理

社交媒体地理数据是位置大数据的重要组成部分,也是近年来产生的新型的大数据类型,具有其鲜明的特点和数据结构,来源众多,普通大众是数据的产生者,也是数据的使用者,对此类数据进行获取、处理、分析及应用,可为大众提供更好的智慧化服务。但是,此类数据涉及用户的隐私安全及数据质量的问题,数据获取来源有限且数据获取方法困难,让此类数据的使用受到一定的限制。因此,选择合适的数据来源并尝试设计实现此类数据的快速采集、存储、处理的方法显得非常重要,这也是本文的研究重点,本章节首先阐述此次选择微博签到数据作为社交媒体地理数据获取代表的缘由,然后再针对微博签到数据的特点设计实现此类社交媒体地理数据获取的方法,最后说明数据的采集流程、存储方式及预处理流程。

3.1 社交媒体地理数据的获取

3.1.1 社交媒体地理数据获取来源选择

签到功能是基于位置的服务中的重要服务模式之一,签到实际上是一种将信息和地理位置串联起来的动作,从而达到将信息按照地理位置重新组织的方式,提供基于地理信息科学、位置地点信息(包含该位置地点的空间信息和属性信息)以及 GNSS 定位技术的实际空间位置信息,实现基于移动 App 应用和移动互联网的定位行为与地理空间的整合;用户需要主动签到以记录自己所在的位置,或通过绑定用户的其它社会化工具,以同步分享用户的地理位置信息,所以,签到数据是社交媒体地理数据的主要数据类型。国外的位置签到服务主要有 Foursquare、Twitter、Gowalla 等,国内则有嘀咕、玩转四方、街旁以及微博等几十家。

本文根据国内用户数量较多、受欢迎程度较大且数据接口开放的微博社交应用作为社交媒体地理数据获取的来源,也以此应用中的签到数据为例来说明社交媒体地理数据获取方法。微博是全球 2015 年活跃用户最多的十大社交网络服务平台之一,和美国的 Twitter 一样都提供相应的 API 接口以支持对其用户数据的访问,微博提供的用户数据访问接口为“微博开放平台”(Weibo Open Platform),其提供的公共 API 接口可以用来访问微博的用户数据。在本研究中我们通过这

一技术手段来实现对社交媒体地理数据的获取——以获取微博的地理位置签到数据，来阐述通过 API 接口获取社交媒体地理数据的方法及详细过程。

微博开放平台是基于微博海量用户和强大的传播能力，接入第三方合作伙伴服务，向用户提供丰富应用和完善服务的开放平台。该平台具有海量用户资源、丰富的接口资源、完善的服务支持等核心能力。截止至 2013 年 3 月底，微博用户数已达 5.56 亿，活跃用户数高达 5000 万。接入微博，实现用户的快速回流和拓展。超过 200 个数据接口，包括微博内容、评论、用户、关系、话题等信息，API 日均调用量超过 330 亿次。不限语言、不限平台的自由接入，不收取任何费用。多种 SDK，包括 C++、PHP、JAVA、Action Script、Python、JS、iOS、Android、WP7 等流行语言的软件开发工具包。发微博、读取微博等功能实例代码，可以帮助你快速掌握调用 API 方法，降低开发门槛。开发者数量同期也增长了近 75%。微博移动端的用户数增长迅猛，从移动端登录微博的用户比例已超过 73%^[56]。

3.1.2 社交媒体地理数据获取原理

总结以往的研究发现，国内对海量社交媒体地理数据的获取多采用网络爬虫技术进行获取。目前国外学者对社交媒体地理数据进行挖掘时已经很少网络爬虫技术，而是采用授权方式的 API 接口进行数据获取。国内很多研究者对于网络数据采集依然使用网络爬虫，该方法获取到的数据并没有取得合法授权，存在很大的安全隐患。

虽然微博开放平台的主要目的是方便第三方开发应用，但由于微博开放平台提供了公共访问接口来访问海量用户数据，所以通过微博开放平台可获取大量用户的社交媒体地理数据。通过微博开放平台的开放 API 接口访问微博用户数据，需要注册、申请、开发、审核等一系列前期准备，以便获取授权。

1、创建应用并获取 App Key 和 App Secret

首先需要注册一个微博账号，注册成功后，该账号可以同时登陆微博和微博开放平台。登陆微博开放平台后，在应用开发页面中创建一个应用，微博提供了客户端应用、网页应用、浏览器插件三种应用类型，开发者可以根据自己的需要创建某一类型的应用，根据页面提示的创建流程创建即可。创建成功后需要完善应用信息及开发者信息。接口的调用需要个人实名认证审核通过后才能使用，同时需要通过邮箱认证和手机认证，审核时间一般为三到五天，审核通过后，在“我的应用”中，身份认证一栏会变成“已认证”。此时，开发者就可以开发自己的应用，即可以访问 API 接口获取用户数据。

在此过程中需要记住应用生成的 App Key 和 App Secret，App Key 是微博应

用的唯一识别标志，微博开放平台通过 App Key 鉴别应用的身份，而 App Secret 是给应用分配的密钥，从而保证应用来源的可靠性。

2、微博开放平台 OAuth2.0 授权获取 Access Token

在开放平台创建了一个应用之后，就获得了识别该应用的一组标识 App Key 和 App Secret。App Key 是 API 接口验证序号，是用于验证 API 接口合法性的。App Secret 是 API 接口密钥。目前微博使用的 OAuth 认证服务的 2.0 版本。OAuth2.0 认证流程图如图 3-1 所示：

OAuth2.0 授权设置：应用创建完后可以在“应用信息”的“高级信息”中设置网站的授权回调页和取消授权回调页。如果没有授权回调页，是无法获取到 Access Token 的，而且必须保证回调页的网址是可以访问的。授权回调页会在用户授权成功后被回调，同时传回一个“code”参数，开发者可以用 code 获取 Access Token 的值。注意 code 的值每次都是不一样的。

获取 Access Token：本文中采用的是微博开放平台提供的 SDK 包，选择了 Java SDK 版本，配置 SDK 开发包，在获取程序代码中填写 client_ID、client_SECRET、redirect_URI，client_ID 即应用的 App Key，client_SECRET 即应用的 Access Token，redirect_URI 为授权回调页地址。填好后运行程序，获取 code 的值，将 code 值复制到程序运行的控制台，即可获得 access token 值和 uid 的值。

access token 为用户授权信息，access token 的值获取成功后，即用户的 OAuth2.0 授权成功。Access token 的值有一定的时效性，access token 失效后需要重新申请。

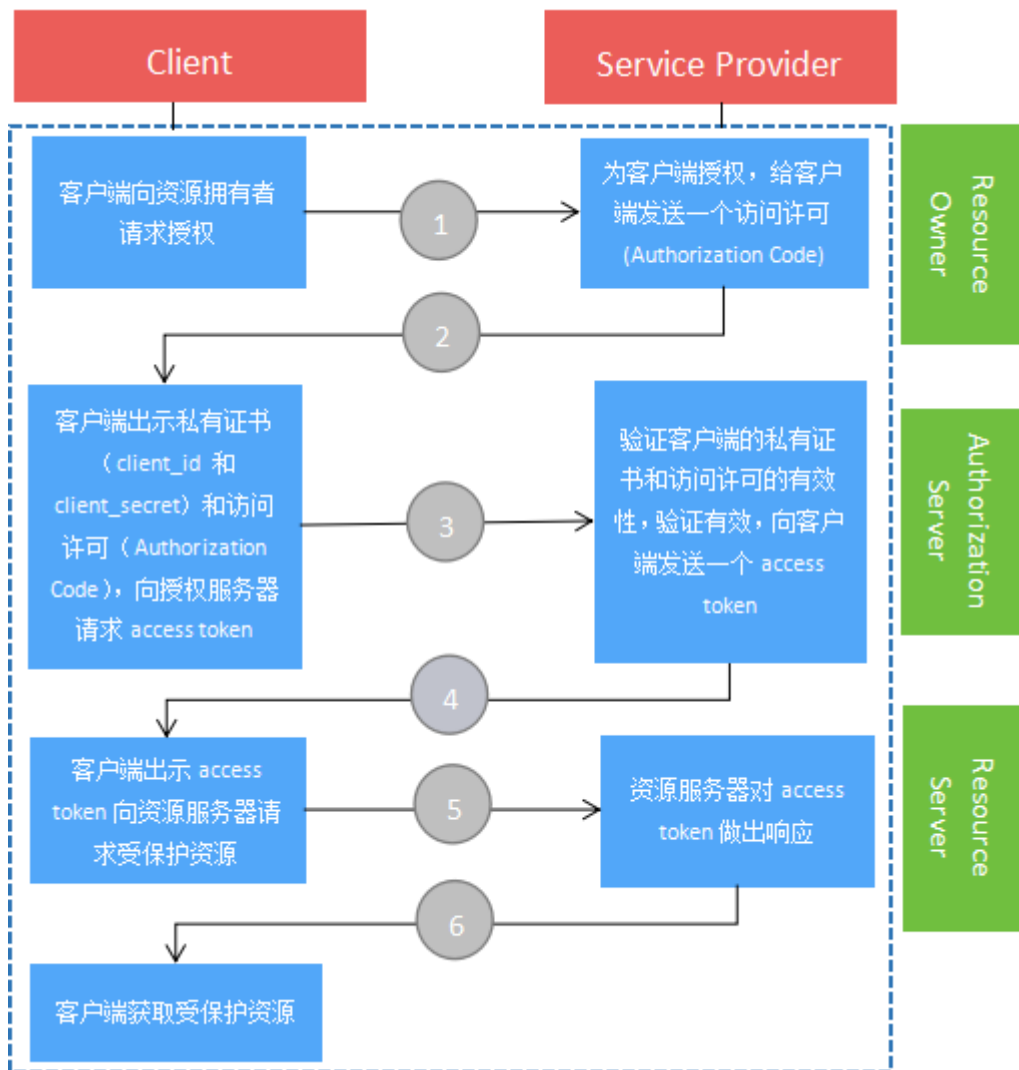


图 3-1 Oauth2.0 认证流程

Oauth2.0 认证流程图中各个步骤分别表示以下内容^[58]:

(1) 客户端 (Client) 从资源所有者 (Resource Owner) 那里请求授权。授权请求能够直接发送给资源所有者, 或者间接的通过授权服务器 (Authorization Server) 发送请求;

(2) 资源所有者为客户端授权, 给客户端发送一个访问许可 (Authorization Code);

(3) 客户端出示自己的私有证书 (client_id 和 client_secret) 和上一步拿到的访问许可 (Authorization Code), 来向授权服务器 (Authorization Server) 请求一个访问令牌 (Access Token);

(4) 授权服务器 (Authorization Server) 验证客户端的私有证书和访问许可的有效性, 如果验证有效, 则向客户端发送一个访问令牌 (Access Token), 访问令牌包括许可的作用域、有效时间和一些其他属性信息;

(5) 客户端出示访问令牌 (Access Token) 向资源服务器 (Resource Server) 请求受保护资源;

(6) 资源服务器对访问令牌 (Access Token) 做出响应。

3、微博开放平台 API 接口的调用

通过 OAuth2.0 授权后, 我们就获得了访问微博 API 数据接口的用户身份认证。在新浪微博公共 API 接口定义中, 用户可以选择返回什么格式类型的微博数据载体文件。本文实验程序采用的是多个微博账号轮流授权应用的策略, 实现了不间断的调用 API 接口获取数据, 以保证不会被接口调用规则所限制。

3.2 社交媒体地理数据采集方法设计

对于本文的实验程序, 考虑到开发程序的时间较短, 所以使用了 Java 版本的 SDK 开发包, 由于 Java 版本的 SDK 编写的较为出色, 程序的性能仍然很好, 但 Java 版本的 SDK 并未提供说明文档, 使用起来需要一定时间的摸索和尝试。

社交媒体地理数据的获取、存储及处理是本文的重点, 数据获取之后的重点就是数据的应用, 本研究的另一个研究重点是基于数据挖掘的游客时空行为的实例研究, 微博开放平台并没有提供直接获取旅游类签到行为数据的接口, 因此, 本文通过间接的方法设计实现数据的采集方法。具体实现方法如下说明。

3.2.1 基础数据的准备

本文的研究区域选择了上海、江苏、浙江三个省市, 这三个省市位于长三角地区, 该地区的社交网络与 LBS 应用活跃程度较高, 所产生的社交媒体地理数据比较丰富。

本文以上海游客在江苏浙江两省所有 A 级以上景区的签到数据为研究对象, 首先需要获取到上海市、及江苏省浙江省的基础地理信息数据, 数据来源为国家基础地理信息中心, 研究区域的地理区位如图 3-2 所示; 将上海市及江苏浙江两省作为研究区域, 是因为该三省市一直以来都是长三角地区重要的旅游核心地区, 目前, 中国 (长三角) 高铁旅游联盟成立, 这是第一家高铁与旅游部门联手成立的联盟, 通过 “高铁+旅游” “旅客+游客” “快旅+慢游”, 推进长三角旅游一体化的实质性发展。另外, 长江旅游联盟、京杭大运河城市旅游推广联盟等助推了区域合作, 在前不久举行的上海国际旅游交易会上, 上海市旅游局牵头举办 “长三角之夜” 联合推介会, 搭建了三省一市直接面向国际旅行商的大平台。根据《杭州方案》的约定, 三省一市将打造以 “城市+乡村” 旅游产品为特色的长

三角旅游目的地，共同设计并推出长三角旅游形象标识。同时，加快长三角旅游供给侧结构性改革，以旅游产业集群为导向，旅游风情小镇、旅游度假区、高等级旅游景区为平台，重点优化旅游新业态布局，全面提升乡村旅游、生态旅游、海洋海岛旅游质量，推动旅游大数据库建设，发展智慧旅游，实现数据实时共享，以及加快旅游集散咨询公共服务体系及标准化建设，引导高速公路服务区设立旅游驿站，实现旅游信息化网络的全覆盖^[57]。

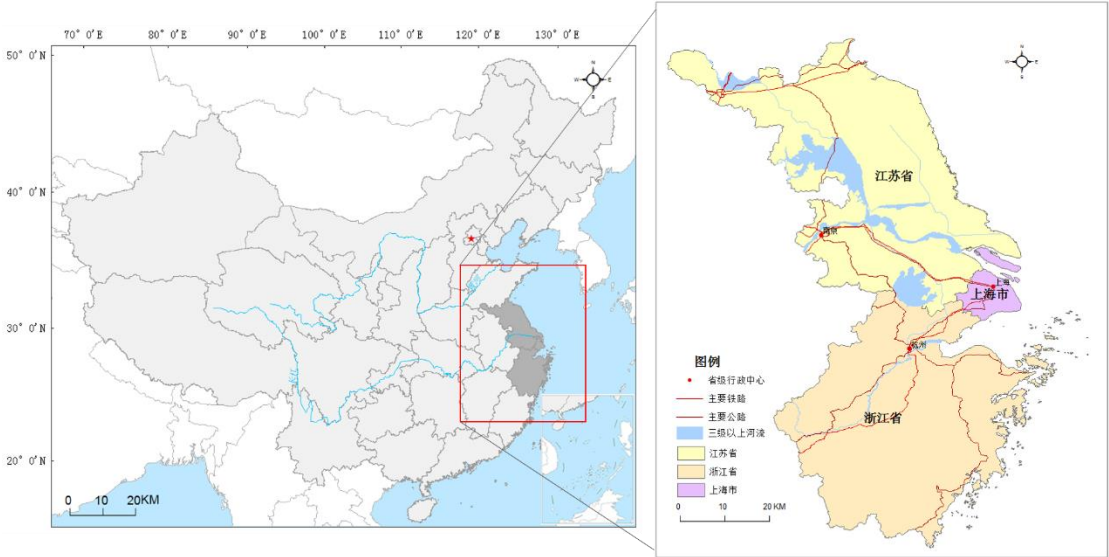


图 3-2 研究区域地理区位

其次还需获取江苏浙江两省 A 级以上景区的景区信息，为了保证数据的完整性和准确性，景区信息均来自两省旅游局官方网站统计结果，获取到的信息字段为景区所在省份、景区所属市县、景区级别、景区名称及景区位置等，获得江苏省 A 级以上景区名录 490 个，浙江省 A 级以上景区名录 337 个。

获取江苏浙江两省 A 级以上景区的景区信息是为获取景区空间位置信息做准备，因为微博开放平台所提供的获取签到数据的接口，需要通过微博开放平台本身数据库中位置地点的编号 (POIID)，然而获取这个位置地点的编号需要传入这个地点的空间位置信息^[59]，所以，本文通过江苏浙江两省旅游局官网获取到的 A 级以上景区信息，通过百度地图 API 中的地图拾取工具获取景区的空间信息。百度地图 API 是为开发者免费提供的一套基于百度地图服务的应用接口^[60]，它所提供的地图拾取工具支持地址的精确/模糊查询；支持 POI 点坐标显示和复制；支持坐标鼠标跟随显示；支持坐标查询等；具体的使用方式是有两种：获取地点名称坐标和坐标反查，本文中我用到的是获取地点坐标名称的方式，在搜索框中搜索关键词后，显示列表中会显示该点的坐标，点击该条信息或地图上该点，都会将坐标信息显示在地图右上角的 Input 框中，然后点击收集整理即可，当获取

数据较大时,也可以使用百度地图 API 所提供的空间位置信息调用接口。

3.2.2 API 接口选择

微博开放平台有超过 200 个的数据接口,包括微博、评论、用户、关系、位置服务、地理信息等,可以根据应用需求选择相应的接口类型。在本文中需要获取的是用户的位置地点签到数据,但微博开放平台没有提供直接获取用户签到数据的接口,需要通过微博开放平台本身数据库中位置地点的编号(POIID)得到,所以,首先需要确定获取位置地点编号 POIID 的数据接口。

在数据准备阶段,获取到江苏浙江两省 A 级以上景区的空间信息,作为获取 POIID 的必选请求参数,通过微博 API 中的位置服务接口获取位置地点的编号(POIID),具体调用的接口 URL 为 <https://api.weibo.com/2/place/nearby/pois.json>,传入的请求参数包括 access-token、lat、long、range、q、category、count、page、sort 和 offset^[59]。其中,access-token 为 oauth 授权后获得的密钥;lat 为 API 请求查询的有效纬度范围;long 为 API 请求查询的有效经度范围;range 为 API 请求查询的查询范围半径,默认为 2000m,最大为 10000m,range 参数值的确定可以影响数据获取质量,本文中 range 参数的确定方法将在下面的叙述中进行说明;category 为 API 请求查询的分类代码;page 为 API 请求的页数;count 为 API 请求的单页显示的数量等。

根据获取到的位置地点的编号(POIID),通过调用微博 API 中获取位置地点的动态接口获取旅游签到信息,具体调用的接口 URL 为 https://api.weibo.com/2/place/poi_timeline.json,传入的请求参数包括 access-token、poiid、since_id、max_id、count、page、base_app。access-token 是采用 OAuth 授权方式的必填参数,通过 OAuth 授权后获得;poiid 是需要查询的兴趣点的编号;count 为 API 请求的单页显示的数量;page 为 API 请求的页数;base_app 询问是否只获取当前应用的数据,0 为否(所有数据),1 为是(仅为当前数据)。

3.2.3 获取 POIID 时请求参数 range 的范围界定

本文中所获取的社交媒体地理数据是以获取上海客源在江苏浙江 A 级以上景区旅游的签到数据为例,API 中 range 参数查询范围半径默认为 2000m,但不同级别景区的景区规模和旅游设施有所不同,为了提高数据质量,不能使用相同的 range 参数,所以在本文中,设计了景区分级界定 range 参数的方法。假设将景区面积近似看作一个圆形区域,再给出景区面积,求算出该景区的影响半径,最后统计某一级别景区的总数,算出该类级别景区的平均影响范围,此平均影响

范围的值就可以作为该类景区级别的 **range** 参数值。从江苏浙江旅游局官网、百科获取两省 A 级以上旅游景区面积并进行景区面积单位换算。江苏浙江两省 A 级以上旅游景区面积从各省旅游局官网、百科获取，最终求算得到的结果为：5A 级景区的 **range** 参数查询范围半径为 3000m；4A 级景区的 **range** 参数查询范围半径为 2000m；3A 级景区的 **range** 参数查询范围半径为 1500m；2A 级景区的 **range** 参数查询范围半径为 1000m。

3.2.4 POIID 的获取实现

本文利用微博 API 提供的 Java SDK 开发包，基于 Eclipse 开发平台设计实现了景区 POIID 及签到数据的抓取方法。获取 POIID 时传入的参数文件如图 3-3 所示，主要传入景区名称、景区空间位置的经纬度坐标及 POI 查询范围的 **range** 值，运行程序，程序根据给定的 **range** 值的范围获取该范围内景区的所有兴趣点信息，程序部分代码如下文所示。

返回值的字段信息包括 POIID、兴趣点名称、兴趣点地址、兴趣点经纬度信息、兴趣点类别代码和兴趣点所属景区名称等。

文件(F)	编辑(E)	格式(O)	查看(V)	帮助(H)
1	夫子庙秦淮风光带	32.028287	118.806106	
3000	0			
2	钟山风景名胜区－中山陵园风景区	32.064347		
118.859411	3000	0		
3	沙家浜·虞山尚湖旅游区	31.652177	120.693665	
3000	0			
4	苏州拙政园	31.330275	120.635554	3000 0
5	同里古镇游览区	31.161831	120.725909	3000 0

图 3-3 获取 POIID 导入参数文件示例

```
public static void main(String[] args) {
    int temp = 0;
    String access_token = "";
    for (int i = SpotId; i < allScenicSpots.size(); i++) {
        for (;;) {
            access_token = "2.00_wVRPGknwfSB3fdcf73efe0aKB4g";
            if (access_token.equals("NULL")) {
            } else {
                try {
                    Place p = new Place(access_token);
                    List<Places> list = new ArrayList<Places>();
                    list = p.nearbyPois(allScenicSpots.get(i).Lat,
```

```

        allScenicSpots.get(i).Lon, Page,
        allScenicSpots.get(i).Range);
    for (Places pl : list) {
        try {
            poi p1 = new poi(pl.getPoiid(), pl.getTitle(),
                String.valueOf(pl.getLat()),
                String.valueOf(pl.getLon()),
                pl.getAddress(), pl.getCategorys(),
                pl.getCity());
            allPoi.add(p1);
        }
    }
    Page++;
} catch (WeiboException e) {
    return "";
}
}

```

3.2.5 签到数据的获取实现

在获取到 POIID 之后，将获取到的景区兴趣点的 POIID 当作获取签到数据的必要请求参数来获取签到数据，获取签到数据时传入的参数文件如图 3-4 所示，传入的请求参数即为获取 POIID 时的返回值，包括 POIID、兴趣点名称、兴趣点地址、兴趣点经纬度信息、兴趣点类别代码和兴趣点所属景区名称等，程序部分代码如下文所示。

返回值的字段信息包括微博 ID、微博创建时间、微博信息内容、微博来源、发送这条微博的用户名、城市代码、用户性别、签到地点信息等内容。由于权限级别受限，所能获取的数据量有限，所以本文中采用了分布式计算的思想，申请多个账号，多次授权，轮流交替使用 access_token。最终获取到 2013~2015 年在江苏省 A 级以上景区的签到数据达 1,360,011 条记录，在浙江省 A 级以上景区的签到数据达 527,825 条记录。

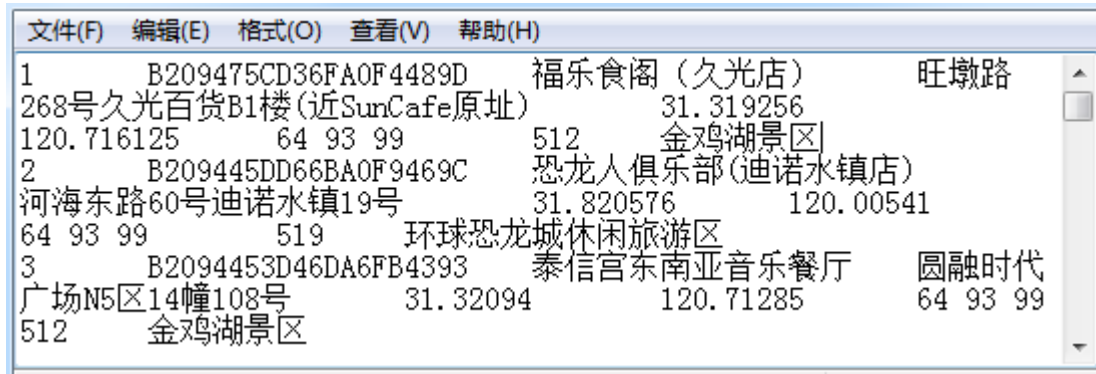


图 3-4 获取签到数据导入参数文件示例

```

public class gettl {
    public static void main(String[] args) {
        String access_token = "2.00_wVRPGOLVIEr8a1ff0a0cfPKjQ_B";
        for (int j = 0; j < 1651; j++) {
            savePathString = "E:\\Java\\getweibopoi\\output 签到江苏 5A\\";
            poiid = poi_List.get(j);
            String Start = timeList.get(0);
            String End = timeList.get(1);
            Place p = new Place(access_token);
            try {
                StatusWapper sw = p.poisTimeLine(poiid, Start, End, 1);
                contentString += poiid + '\t'
                    + String.valueOf(sw.getTotalNumber()) + "\r\n";
            } catch (WeiboException e) {
                contentString += poiid + '\t' + "0" + "\r\n";
            }
            savePathString += poiid + ".txt";
            creatTxtFile(new File(savePathString));
            page = 1;
            int judge = 0;
            for (;;) {
                try {
                    StatusWapper sw = p.poisTimeLine(poiid, Start, End, page);
                    for (int i = 0; i < sw.getStatuses().size(); i++) {
                        data_List.add(sw.getStatuses().get(i));
                    }
                    page++;
                    judge = 0;
                } catch (WeiboException e) {
                    if (judge == 0) {
                        judge++;
                        continue;
                    }
                }
            }
        }
    }
}

```

```
        } else {
            System.out.println("单个结束!!! ");
            break;
        }
    }
}
```

3.3 社交媒体地理数据的存储与管理

根据前面章节总结的社交媒体地理数据的特点及数据结构，本文采用 SQL Server 数据库作为社交媒体地理数据的存储媒介。此外，在后续章节的应用实例分析中要进行游客时空行为分析和热区探索，需要用到 ArcGIS Desktop 软件，因此还需将 SQL Server 中的签到数据导入到 ArcGIS 的地理空间数据库中。

3.3.1 SQL Server 数据库

本文之所以选择 SQL Server 数据库作为签到数据的存储媒介除了它在存储海量数据方面有众多优势之外，最重要的还是 SQL Server 数据库可直接连接到 ArcGIS Desktop。本文将上述获取到的 POI 数据及用户签到数据一一存储到 SQL Server 数据库中，每条数据记录都有唯一标识编号，数据存储的表结构见表 3-1 及表 3-2 所示：

表 3-1 POI 信息字段

字段	说明	示例	备注
poi_id	兴趣点编号	B209475CD36FA0F4489D	唯一标识
poi_name	兴趣点名称	灵山胜境	
address	所在地址	无锡市滨湖区群灵路	
category	属性编号	510	
latitude	纬度	31.42233	
longitude	经度	120.1048	
Scenic_area	所属景区	灵山景区	

表 3-2 签到数据信息字段

字段	说明	示例	备注
status_id	状态编号	3.69759E+15	唯一标识
create_time	微博发布时间	Mon Sep 07 01:02:14 CST 2015	
status_text	微博内容	怀旧风的一个地方，这幅照片我都想带走了?? http://t.cn/RwufRtN	
user_id	用户编号	1787320723	
user_name	用户名	Qian 钱珩	
city	所属城市编号	31	
gender	性别	f	
poi_id	兴趣点编号	B2094451D16AA0F4439D	
latitude	纬度	32.03851	
longitude	经度	118.8657	

3.3.2 Geodatabase 数据库

Geodatabase 是一种采用标准关系数据库技术来表现地理信息的数据模型。Geodatabase 支持在标准的数据库管理系统表中存储和管理地理信息，支持多种 DBMS 结构和多用户访问，且大小可伸缩。目前有两种结构：个人 Geodatabase 和多用户 Geodatabase。个人 Geodatabase，对于 ArcGIS 用户是免费的，它使用 Microsoft Jet Engine 数据文件结构，将 GIS 数据存储在小数据库。个人 Geodatabase 更像基于文件的工作空间，数据库存储量最大为 2GB。可以实现以下功能：

- 支持海量的，连续的 GIS 数据库；
- 多用户的并发访问；
- 长事务和版本管理的工作流。

因为 ArcGIS Desktop 支持直接连接 SQL Server 数据库，因此将存储在 SQL Server 数据库中的签到数据导入 Geodatabase 数据库很方便，本文中用到的 SQL Server 数据库版本为 SQL Server 2008 R2，具体的实现方法如下：

1、在 Arccatalog 中创建地理数据库；

配置环境，选择直连数据库的方式，有 SQL_Server、Oracle、Postgresql 三种方式，选择 SQL_Server；输入授权文件 Authorization File，即安装 ArcGIS server 时的破解文件即可。

2、建立数据库连接。在 Database Connection 目录中创建一个数据库直连方式，输入相关的信息；

3、在直连数据库方式里面创建要素集以及导入相关的地理坐标系；

4、最后导入要素类数据或者新建要素类。

3.4 社交媒体地理数据预处理

在前面章节中论述社交媒体地理数据的特点时说过数据质量参差不齐是该类数据的一大特点之一，数据来源很多，数据的产生可以是专业人员产生，也可以是非专业人员自发产生，非专业人员产生的数据比较随意和个性化，且数据产生的设备众多，精度差异大等。所以需要获取到的数据进行进一步的处理、清洗和筛选，以提高数据挖掘质量。主要是达到如下目标：格式标准化，异常数据清除，错误纠正，重复数据的清除等。

3.4.1 景区签到数据整合

在数据的获取过程中，数据获取程序设计实现时采用 txt 文档格式输出，一个兴趣点信息输出一个 txt 文档，但在后面对签到数据挖掘的应用实例分析中是以景区为单位的，一个景区内包含多个兴趣点，也就包含多条兴趣点信息，一个兴趣点上又包含多条签到记录，所以需要实现将单个兴趣点上的签到记录统计为同一类级别的景区的签到记录总和，我在本文中的具体实现方法是：基于 Microsoft Visual Studio 2010 开发平台，采用 C#语言编写签到数据整合预处理程序，实现海量签到数据的快速处理。

3.4.2 兴趣点上的签到人数统计

同样的，一个兴趣点上的签到数据以 txt 文档输出时，并没有自动统计在一个兴趣点上相应时间段内共有多少条签到记录，只给出了文件大小，只有统计出相应兴趣点上的签到数据量，才能进行后期的旅游热区分析，所以还需要设计实现快速统计兴趣点上签到数据量的中间程序，实现方法同样是：基于 Microsoft Visual Studio 2010 开发平台，采用 C#语言编写兴趣点上签到记录统计预处理程序。

3.4.3 签到时间字段解析

签到数据的返回值字段中，签到数据的时间格式是以字符串形式返回，例如“Tue Apr 21 23:25:45 CST 2015”，这样的数据格式并不利于后面对签到数据挖掘的应用实例分析的研究，所以需要进行时间字段解析，将字符串形式的时间格式解析成年、月、日、星期、时段的字段格式，所以在本文中我同样是基于 Microsoft Visual Studio 2010 开发平台，采用 C#语言编写签到时间字段解析的预处理程序。

3.4.4 数据的清洗和筛选

由于社交媒体本身就是一个自由、开放的平台，所产生的社交媒体地理数据也就带有这样的特点，用户就有相应的权限，可注册多个账号信息、也可删除、撤销自己的数据记录等，所以所产生数据有可能会有重复、错误和为空的情况，此时就需要清洗数据，剔除重复、错误和为空的数据记录，以提高数据质量。在本文中根据研究的需要及获取到的数据的特点，需要进行以下两个主要的数据清洗和筛选工作：

1、POIID 的清洗

提取 POIID 字段信息，进行比对统计，删除重复、错误的 POIID 信息记录。

2、POI 类别筛选

获取 POI 类别标准，参照标准，删除与研究目标无关的 POI 类别，本文是以游客出游行为为例来说明签到数据挖掘应用的，旅游活动大致分为吃、住、行、游、购、娱六大类，所以应当删除与旅游活动无关的 POI 类别，例如删除医院、住宅区等与旅游活动无直接相关的嫌疑数据，以提高数据质量。

3.4.5 目标数据提取

通过前期一系列的数据获取、数据存储和数据预处理工作后，得到了 2013~2015 年的在江苏省 A 级以上景区的签到记录达 1,360,011 条，在浙江省 A 级以上景区的签到记录达 527,825 条，此次对于社交媒体地理数据挖掘的应用实例分析目标主要是针对上海客源在江苏浙江 A 级以上景区出游的出游情况分析，所以需要在所获取的这些签到记录中提取出上海客源的旅游活动签到记录，在以往的研究中，多是通过在旅游目的地发放问卷、访谈或者通过旅游官方统计数据进行旅游者的行为研究，但是此方法仅限于统计旅游目的地的所有客源信息，无法提取出大量某一类客源的旅游活动信息，对于这个问题，本文获取到的旅游签到数据可以很好的解决这个问题。

针对于本文社交媒体地理数据挖掘的应用实例分析的目标，需要在将近两百万条的目标旅游目的地签到数据中提取出上海客源的旅游签到数据信息，首先通过微博 API 接口获取中国省份代码，然后查询上海市的城市代码，编号值为 31；其次在签到数据的返回值中提取出城市代码为 31 的签到记录，2013~2015 年上海客源在江苏省 A 级以上景区的签到数据提取量为：59073 条记录，在浙江省 A 级以上景区的签到数据提取量为：33176 条记录。

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)			
省份代码	名称	省份代码	名称
11	北京市	43	湖南省
12	天津市	44	广东省
13	河北省	45	广西壮族自治区
14	山西省	46	海南省
15	内蒙古自治区	50	重庆市
21	辽宁省	51	福建省
22	吉林省	52	贵州省
23	黑龙江省	53	云南省
31	上海市	54	西藏自治区
32	江苏省	61	陕西省
33	浙江省	62	甘肃省
34	安徽省	63	青海省
35	河南省	64	宁夏回族自治区
36	江西省	65	新疆维吾尔自治区
37	山东省	71	台湾省
41	湖北省	81	香港特别行政区
42	四川省	82	澳门特别行政区

图 3-5 中国各省（市）代码返回值

3.5 数据质量验证

根据社交媒体地理数据的特性我们知道，其数据来源很多，数据产生的设备多样，精度差异大等，所以需要获取到的数据进行进一步的处理、清洗和筛选，以提高数据质量，达到数据格式标准化，异常数据清除，错误纠正，重复数据的清除等。

数据质量不高会导致数据不能有效的被利用，甚至出现严重的决策失误，检验数据的可靠性是数据利用的前提。所以在数据清洗之后需要对所获取的数据进行数据质量检验。本研究采取的数据质量检验方法是统计对比方法，通过查询 2014 年中国旅游年鉴获得江苏省 2014 年国内旅游市场分布数据：江苏本省是省内旅游最大客源地，当年接待本省游客 21154.77 万人次，八大客源地省（市）依次为上海、北京、浙江、安徽、广东、山东、河南和湖北。

统计 2014 年到江苏省内进行旅游的各客源地游客的签到数量，江苏本省的省内旅游签到数据最多，为 257861 条记录，是省内旅游最大客源地。除江苏本省外，其余省（市）旅游签到次数最多的十大客源地依次为上海、北京、浙江、安徽、广东、山东、河南、湖北、福建和四川，将排名前八的客源地省（市）与旅游年鉴统计数据作对比，结果一致则通过数据质量检验。所以采用本研究方法所获取的社交媒体地理数据可用于游客空间行为分析。

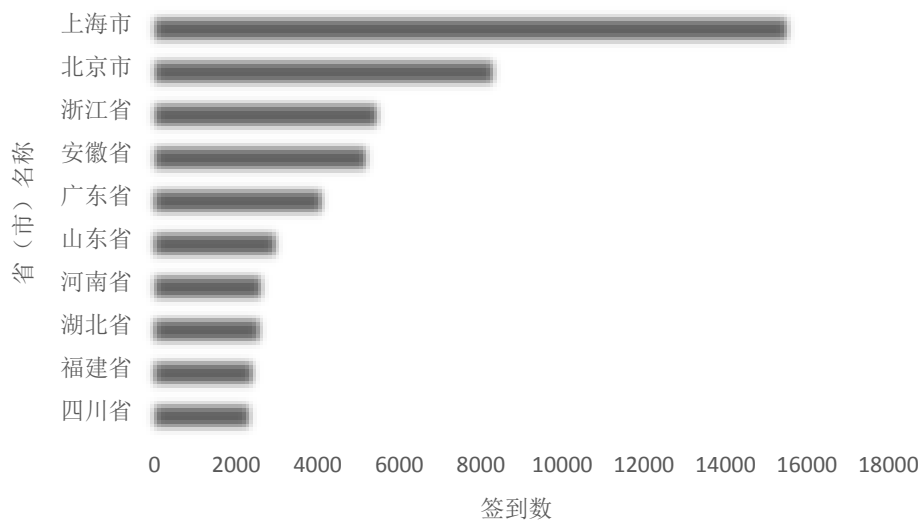


图 3-6 江苏省 2014 年旅游签到次数前十的省（市）

3.6 数据采集结果可视化

根据前面章节对社交媒体地理数据采集方法的设计实现以及对数据存储及预处理的设计实现，基于微博开放平台提供的开放数据接口，基于上海市、江苏省、浙江省的行政区划，共获取到了江苏浙江两省 827 个 A 级以上景区内上海客源的签到数据 92249 条记录，其江苏浙江两省 A 以上景区的空间分布结果和上海客源的旅游签到数据采集结果的可视化图如图 3-7 和图 3-8 所示。

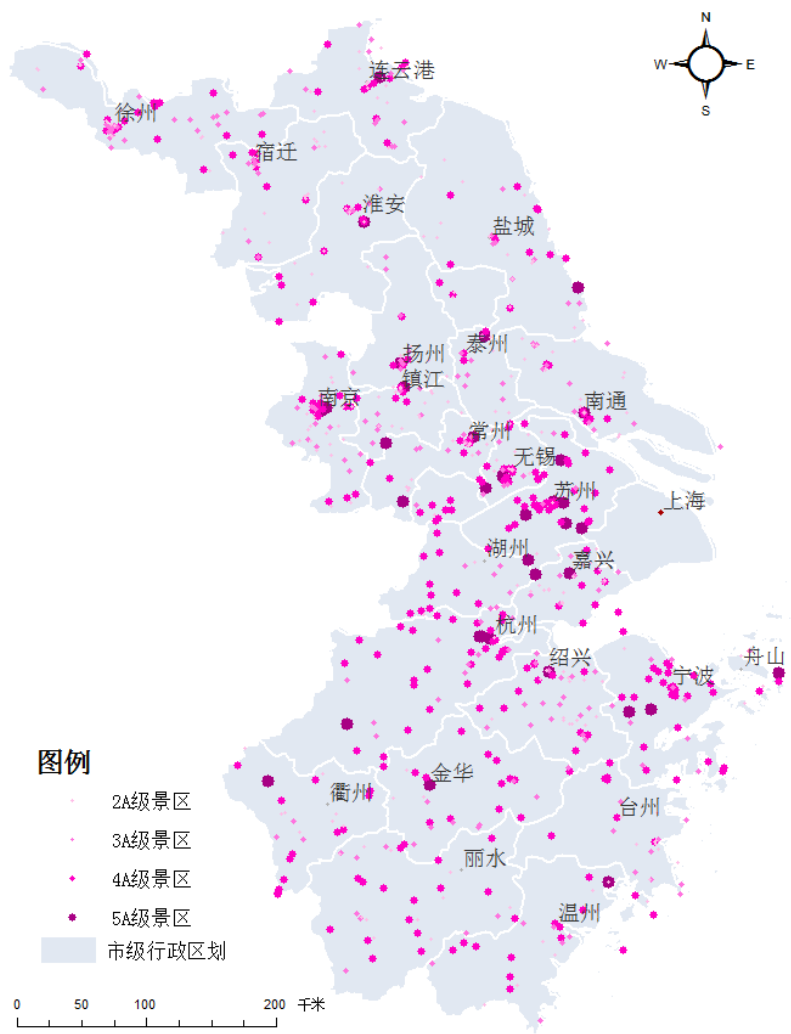


图 3-7 江苏浙江两省 A 以上景区的空间分布图

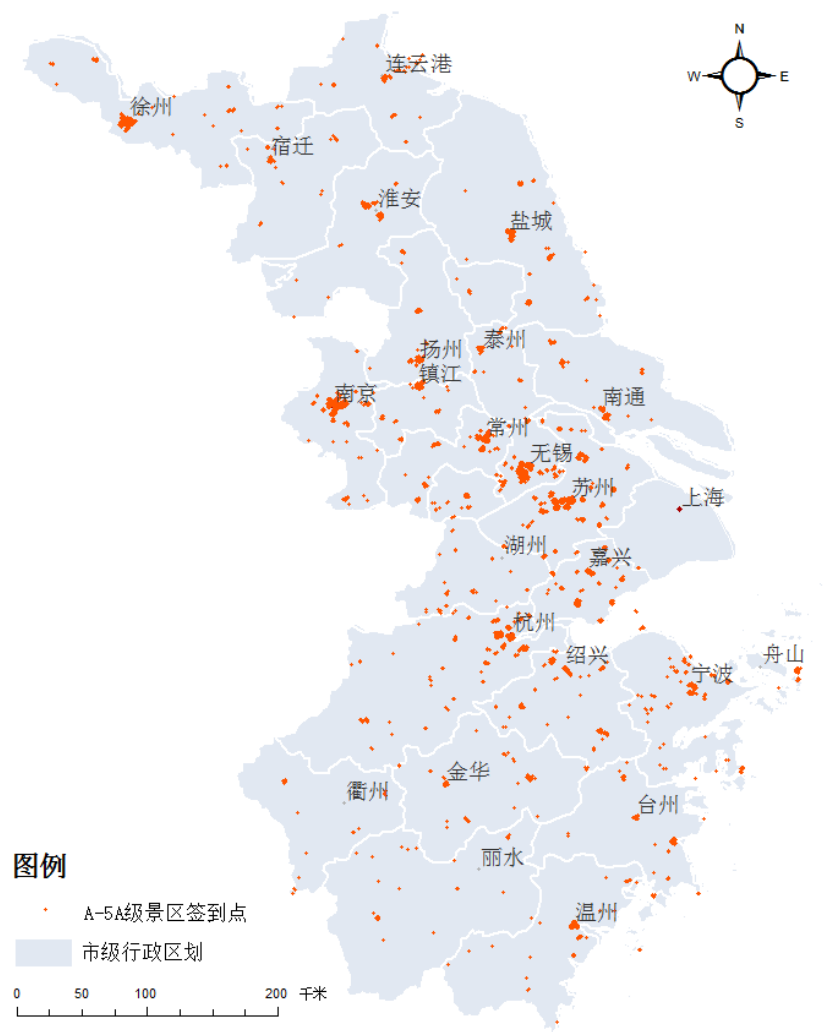


图 3-8 上海客源在江苏浙江 A 级以上景区出游的签到数据空间分布图

第 4 章 社交媒体地理数据挖掘应用实例分析

数据获取及数据处理的最终目的是得到目标数据，并将目标数据应用到实际中解决实际问题，本章就针对采集而来的上海客源的出游签到数据进行社交媒体地理数据挖掘的应用实例分析。分别从时间、空间两方面展开分析，总结上海客源到江苏浙江出游的签到数据的年际变化特征，节假日变化特征，及节假日、周末和工作日的对比变化特征。在空间上，利用 ArcGIS 空间分析方法，探索上海客源感兴趣的出游热点区域，及黄金周、小长假、双休日期间的出游模式及出游特点，为人们的智慧出行提供依据。

4.1 签到数据的时间特征

4.1.1 年际变化特征

从年际变化角度出发，首先将三年的签到数据按不同年份不同景区级别进行统计，2013 年上海客源在江苏浙江两省 A 级以上旅游景区签到的数据记录是 25115 条，2014 年为 24130 条记录，2015 年为 43003 条记录，然后得出如图 4-1 的 2013~2015 年各年份上海客源在江苏浙江 A 级以上景区签到数对比统计图，由图可以看出：2015 年上海客源在江苏浙江 A 级旅游景区签到记录相较于 2013 年和 2014 年明显增多，且上海客源对 4A 级景区的兴趣程度最大。

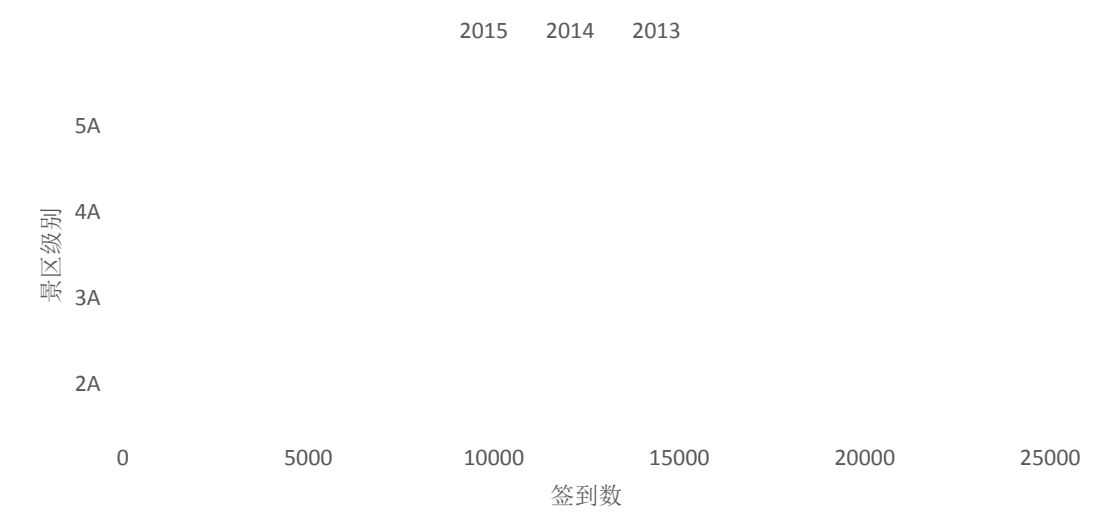


图 4-1 2013~2015 年各年份上海客源在江苏浙江 A 级以上景区签到数据对比图

数据采集从 2013 年 1 月开始, 2015 年 12 月结束, 对比三年数据, 发现数据量在 2013 年和 2014 年之间相差不大, 而且还有 2013 年超过 2014 年的趋势, 而 2015 年的数据在 2A~5A 景区都大大超过 2013 年和 2014 年, 这源于在 2013 年微博这个社交应用热潮达到顶峰, 之后伴随着其他多类社交应用的产生, 此应用于 2014 年进入平稳区, 然而, 在 2015 年 1 月, 微博开放微博 140 字的发布限制, 微博的使用热潮再次进入上升期。另一个原因是一系列长三角旅游行动纲领的推出, 例如 2014 年底, 苏、浙、皖、沪旅游管理部门在上海签署了《长三角地区率先实现旅游一体化行动纲领》。2014 年 8 月, 长三角地区旅游业还依托苏、浙、皖、沪快速便捷的高铁网络, 整合地区旅游资源, 推出“高铁+景区门票”、“高铁+酒店”旅游线路和产品, 着力打造“铁字头”旅游品牌, 率先实现长三角旅游业一体化发展。再者, 国民经济的提高, 推动大众进行越来越多的旅游活动, 旅游成为大众工作之余释放压力、放松身心的主要方式, 加上智能手机的普遍使用和数据网络的发展, 使得人们在旅游活动中记录了大量的旅游数据, 并上传到社交媒体应用的虚拟空间, 例如: 出游的照片、文字、视频、音频等, 并伴随着大量的签到记录, 以此形式分享旅途中的美好与精彩。

4.1.2 节假日变化特征

从节假日变化角度出发, 首先将三年的签到数据进行汇总, 再统计出 2013 年~2015 年所有的节假日放假时间, 对比三年的节假日时间分别提取出元旦节、春节、清明节、劳动节、端午节、中秋节和国庆节在不同级别景区的出游签到记录, 得出 2013~2015 年节假日期间上海客源在江苏浙江 A 级以上景区的签到数统计对比图, 如图 4-2 所示。由图可以看出: 国庆节、春节是出游高峰, 其次是劳动节、端午节和元旦节, 清明节和中秋节的出游率相对较低。



图 4-2 2013~2015 年节假日上海客源在江苏浙江 A 级以上景区的签到数据对比图

结果分析：图中结果显示，国庆节和春节的出游签到记录最多，这是因为国庆节和春节有 7 天长假，称之为黄金周，相较于其它五个节假日来说游客有更多的出游时间；同为 7 天黄金周假期，春节的签到记录要少于国庆节很多，这是因为春节是中国最重视的传统节日，意味着与家人团聚之意，大多数人都会选择回家过年，但毕竟是一年之中难得的 7 天假期，也有越来越多的人会选择抓住机会外出游玩。其次是劳动节、端午节和元旦节，对于这三个小长假而言，刚好是一年之中的中间时段，工作了半年难得的休息机会，而且刚好气候适宜，也是很多出游爱好者最佳出游机会。出游签到记录最少的节日就是清明和中秋了，这两个节日的节日特征明显，清明节是人们回乡祭祖和扫墓的日子，而中秋是人们回乡和情人团聚的日子，大多数人都会选择回家陪伴亲人，所以出游活动减少。总结出游签到数据的节假日特点，符合我们日常生活中的出游习惯和规律，也证明了数据的可用性。

4.1.3 节假日、周末和工作日的对比变化特征

从节假日、周末和工作日对比的角度出发，首先分年份统计出各年中节假日、周末和工作日的签到记录，在分别统计出 2013~2015 年各年中节假日、周末和工作日各有多少天，然后算出游客在小长假、周末和工作日中每天的平均签到记录，得出 2013~2015 年各时段上海客源到江苏浙江 A 级以上景区的签到记录对比统计图，如图 4-3 所示。对比三年数据，出游签到数据都是节假日最多、周末次之、工作日最少，但在 2015 年的数据中周末和工作日的出游签到数据比例大幅上升，且周末上海客源到江苏浙江出游的签到记录已快赶上小长假出游签到记录数。

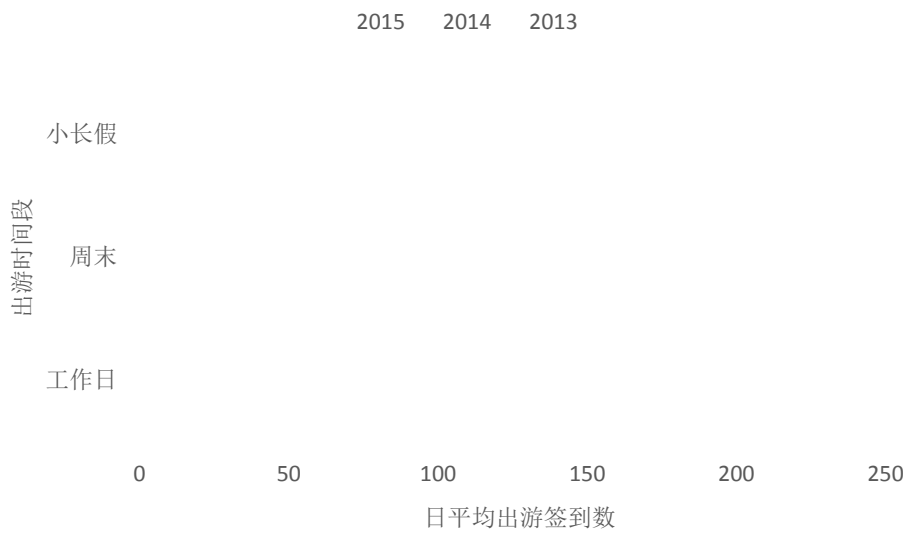


图 4-3 2013~2015 年各时段上海客源到江苏浙江 A 级以上景区的签到记录对比图

结果分析：从节假日、周末和工作日的对比统计图中可以看出，出游签到数据节假日最多、周末次之、工作日最少，这符合人们日常作息时间安排，而在 2015 年的数据中周末和工作日的出游签到数据比例大幅上升，且周末上海客源到江苏浙江出游的签到记录已快赶上小长假出游签到记录数。出现这样的结果并不奇怪，这在很大程度上源于人们出游意识的提高，旅游是城市生活中越来越不可少的休闲娱乐方式，江苏浙江位于长三角区域，拥有丰富的旅游资源，风景名胜点星罗棋布，长江、钱塘江、黄浦江浩浩荡荡，淮河、京杭大运河奔腾不息，西湖、太湖、天目湖、巢湖烟波浩渺。且在 2014 年 9 月，上海铁路局增开 15 趟“旅游专列”，助推长三角旅游热，大大提高了上海到江苏浙江出游的几率，也缩短了交通上花费的时间，这将让短距离出游不在局限于节假日期间，所以周末出游率也大大提高。

4.2 签到数据的空间特征

将获得的上海客源在江苏浙江两省 A 级以上景区出游的签到数据，经过数据的处理、清洗与筛选后，获得有效的签到数据信息。在本章节中将通过核密度分析方法探索上海客源在江苏浙江两省的旅游热点区域，并进一步分析挖掘结果隐含的信息。

4.2.1 旅游目的地整体热区探索

利用 ArcGIS 10.2 对 2013~2015 年上海客源到江苏浙江两省 A 级以上景区出

游的签到数据 92249 条记录做核密度分析，可视化结果如图 4-4 所示，签到数据的密度高峰值出现在南京、杭州、苏州、无锡，集中度远高于其他地区，其次是常州、盐城、扬州和嘉兴，南京、无锡、苏州、杭州为出游签到数据密度的峰值地区是许多因素综合导致的，南京和杭州是江苏、浙江的省会城市，南京素有“江南佳丽地，金陵帝王州”之称，苏州和无锡虽不是省会城市，但同样旅游资源丰富，城市文化休闲宜人。另一个重要的原因就是这几个密度高值区域都距离上海不远，且交通便利，方便上海客源的短距离出游。

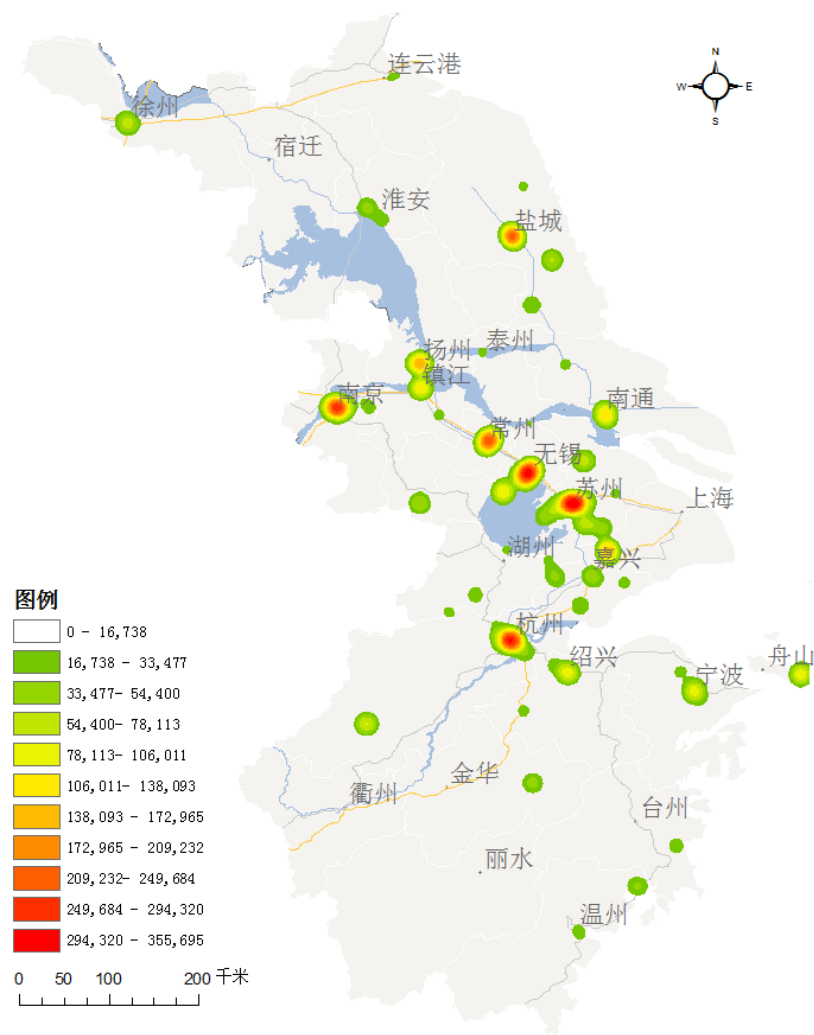


图 4-4 上海客源在江苏浙江 A 级以上景区签到点的核密度分布图

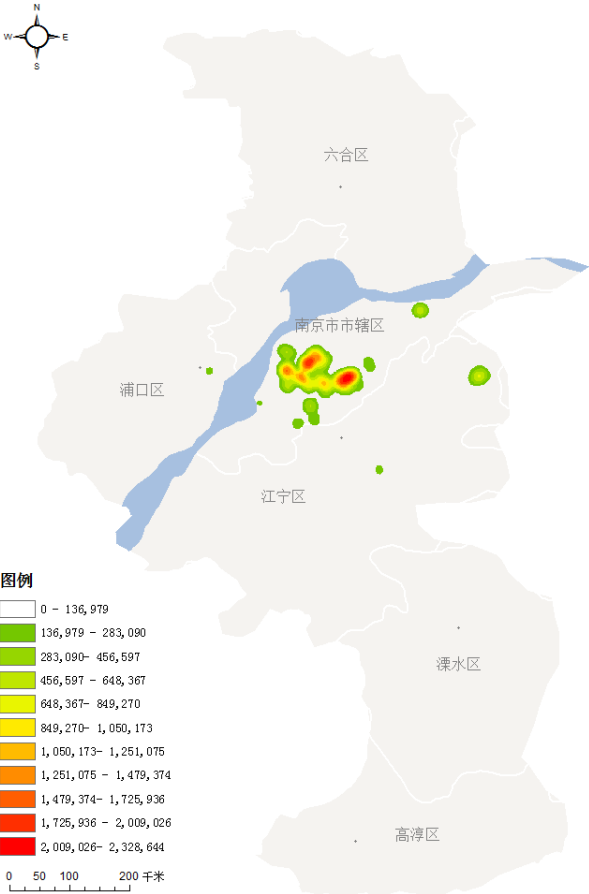
4.2.2 旅游目的地的局部热区探索

在出游目的地的整体热区探索可视化结果中，南京、杭州、苏州、无锡四地的出游签到数据密度出现高峰值，在本章节中将这四个地区的数据再一次进行提

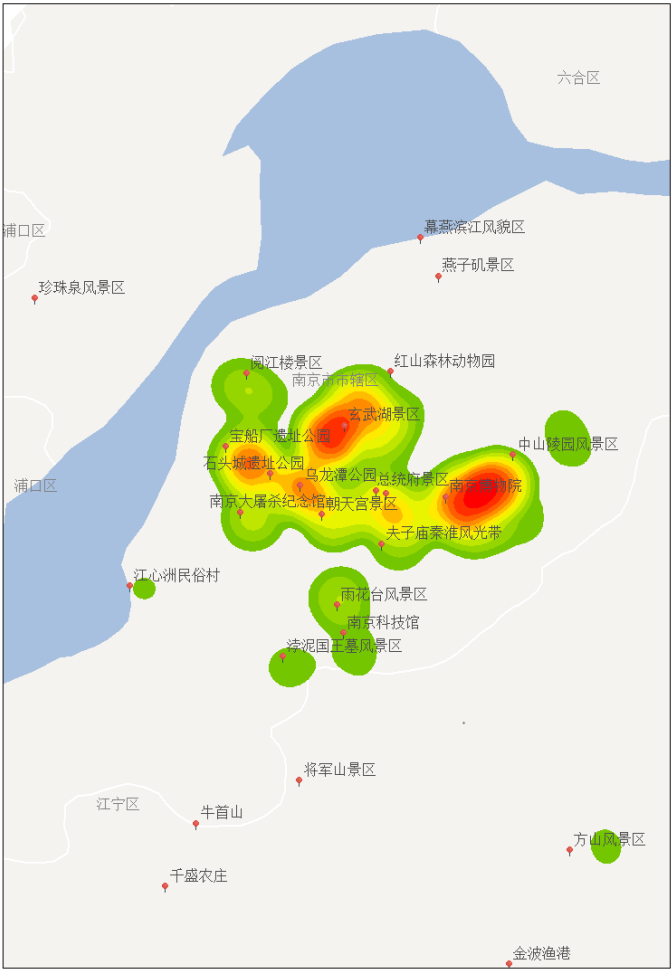
取，分别对这四个城市的 A 级以上景区的签到数据做核密度分析，意在探测这四个城市的哪些景区最吸引上海客源。

1、南京旅游热区探索

对南京 A 级以上景区的上海客源签到数据做核密度分析，可视化结果如图 4-5（a）、（b）两图所示，4-5（a）是对南京整体的旅游热区进行探索，图中签到密度高峰值出现在主城区内，空间极化现象明显，南京市市辖区中心位置集中度远高于其他地区，在外围区域呈现出分散但局部聚集的特点，聚集点一般出现在该区域的中心；南京市市辖区中心位置出现峰值是很多因素综合导致的，该区域公共服务设施十分完善，还有丰富的休闲娱乐设施，南京市大量的旅游资源均集中在该区域。4-5（b）是对南京旅游热区的局部探索，结果显示上海客源对南京市较感兴趣的旅游景区为玄武湖景区、宝船厂遗址公园、石头城遗址公园、乌龙潭公园、总统府景区、南京大屠杀纪念馆、朝天宫景区、南京博物馆、南京夫子庙及秦淮风光带等景区，其次中山陵园风景区、江心满民俗村、花雨台风景区、南京科技馆和方山风景区等也是大部分上海客源青睐的地方。



(a)

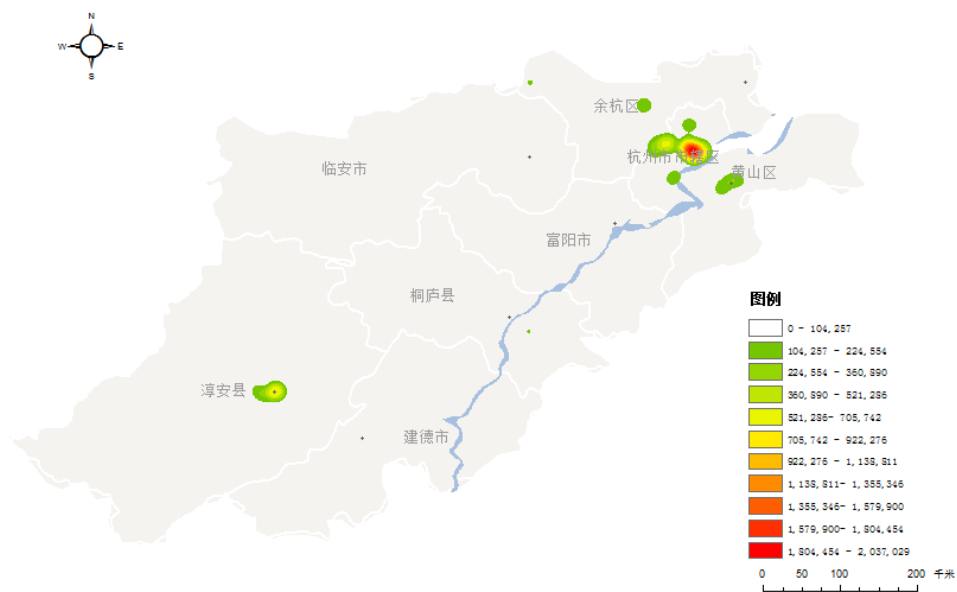


(b)

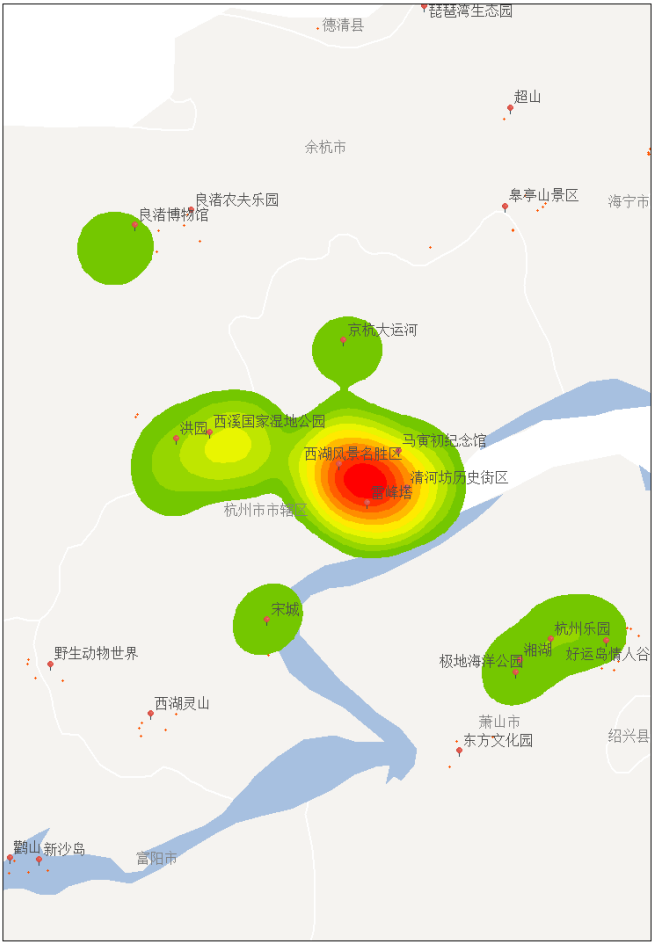
图 4-5 上海客源在南京出游的热点区域分布图（a 为整体热点区位、b 为局部热点区位）

2、杭州旅游热区探索

对杭州 A 级以上景区的上海客源签到数据做核密度分析，可视化结果如图 4-6（a）、（b）两图所示，4-6（a）是对杭州整体的旅游热区进行探索，图中签到密度峰值出现在杭州市市辖区及淳安县，空间极化现象明显，杭州市市辖区中心位置集中度远高于其他地区，在外围区域同样呈现出分散但局部聚集的特点，聚集点一般出现在该区域的中心；杭州市市辖区中心位置出现峰值是很多因素综合导致的，该区域公共服务设施十分完善，还有丰富的休闲娱乐设施，杭州市大量的旅游资源均集中在该区域。4-6（b）是对杭州旅游热区的局部探索，结果显示上海客源对杭州市较感兴趣的旅游景区为西湖风景名胜区、雷锋塔、清河坊历史街区、马寅初纪念馆、西溪国家湿地公园、洪园等，其次为京杭大运河、宋城、良渚博物馆、杭州乐园、好运岛情人谷、极地海洋公园等也是大部分上海客源青睐的地方。



(a)

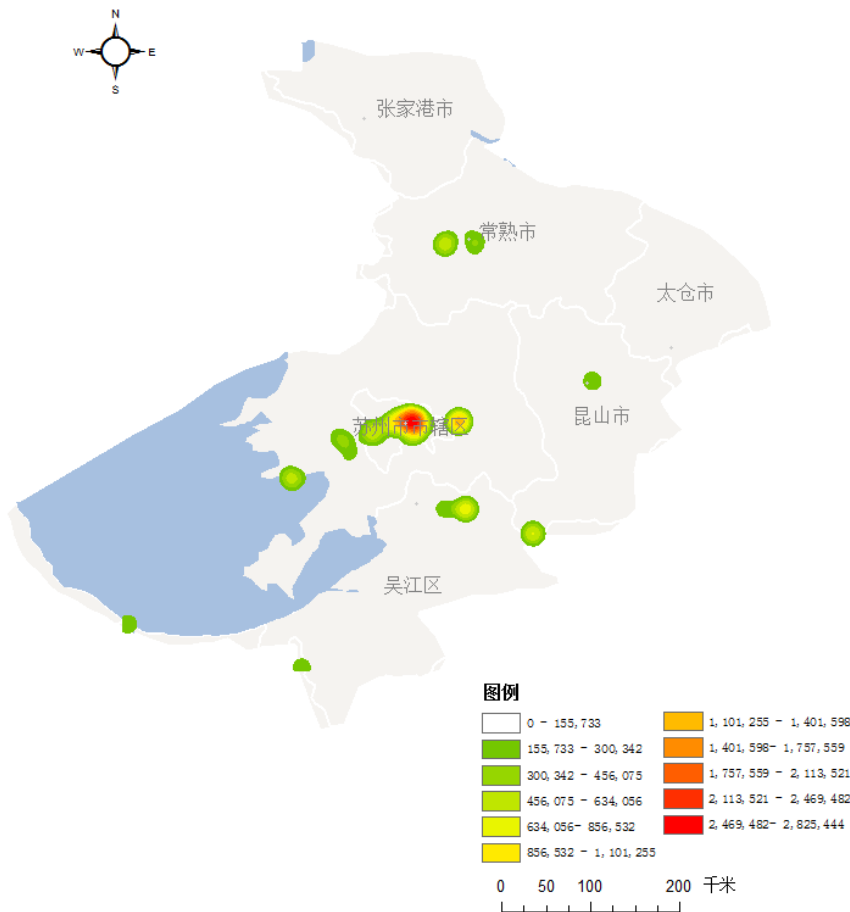


(b)

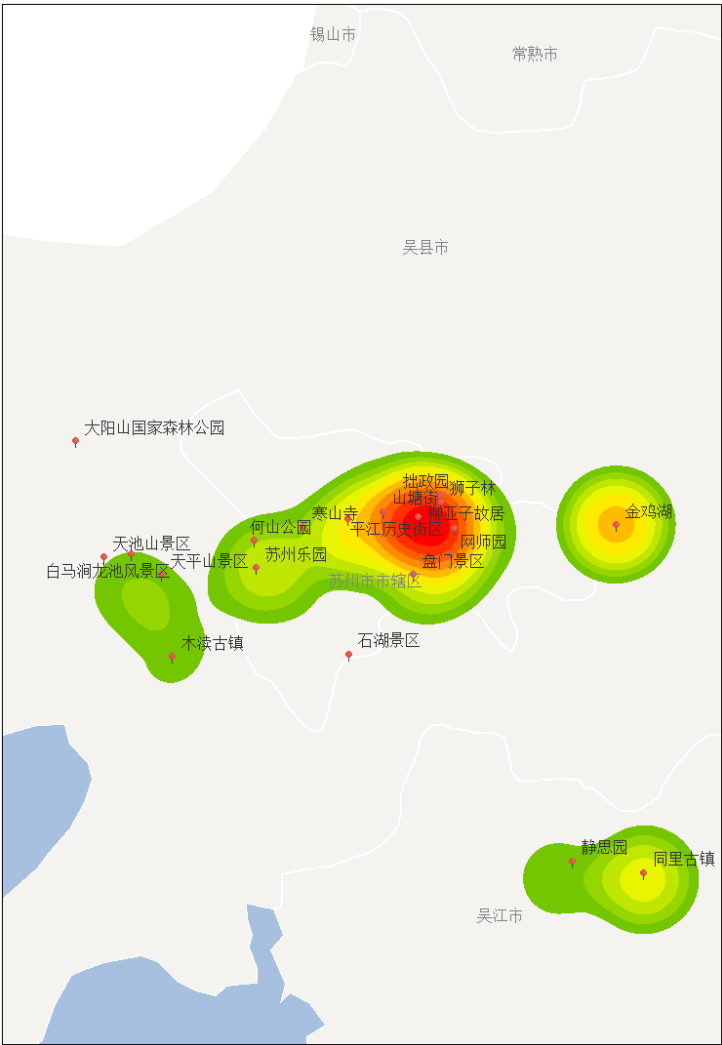
图 4-6 上海客源在杭州出游的热点区域分布图 (a 为整体热点区位、b 为局部热点区位)

3.苏州旅游热区探索

对苏州 A 级以上景区的上海客源签到数据做核密度分析，可视化结果如图 4-7（a）、（b）两图所示，4-7（a）是对苏州整体的旅游热区进行探索，图中签到密度高峰值出现在苏州市市辖区，吴江区和常熟市也有部分热区，空间极化现象明显，苏州市市辖区中心位置集中度远高于其他地区，在外围区域同样呈现出分散但局部聚集的特点，聚集点一般出现在该区域的中心；苏州市市辖区中心位置出现峰值是很多因素综合导致的，该区域公共服务设施十分完善，还有丰富的休闲娱乐设施，苏州市大量的旅游资源均集中在该区域。4-7（b）是对苏州旅游热区的局部探索，结果显示上海客源对苏州市较感兴趣的旅游景区为拙政园、狮子林、山塘街、柳亚子故居、平江历史街区、网师园、盘门景区、寒山寺、金鸡湖、同里古镇、天平山景区、苏州乐园等，其次为天池山景区、白马涧龙池风景区、木渎古镇、静思园等也是大部分上海客源青睐的地方。



(a)

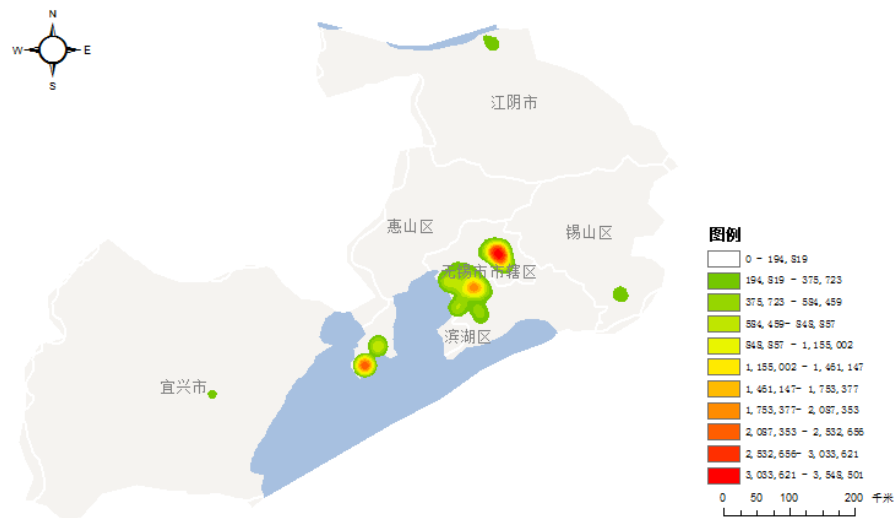


(b)

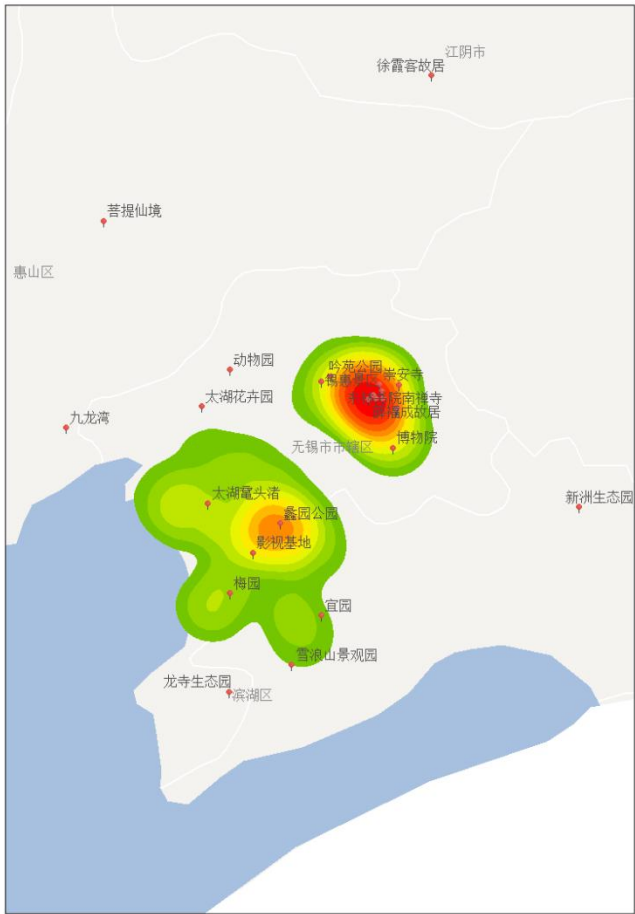
图 4-7 上海客源在苏州出游的热点区域分布图（a 为整体热点区位、b 为局部热点区位）

4.无锡旅游热区探索

对无锡 A 级以上景区的上海客源签到数据做核密度分析，可视化结果如图 4-8（a）、（b）两图所示，4-8（a）是对无锡整体的旅游热区进行探索，图中签到密度高峰值出现在无锡市市辖区和惠山区局部，空间极化现象明显，无锡市市辖区中心位置集中度远高于其他地区，在外围区域同样呈现出分散但局部聚集的特点，聚集点一般出现在该区域的中心；无锡市市辖区中心位置出现峰值是很多因素综合导致的，该区域公共服务设施十分完善，还有丰富的休闲娱乐设施，无锡市大量的旅游资源均集中在该区域。4-8（b）是对无锡旅游热区的局部探索，结果显示上海客源对无锡市较感兴趣的旅游景区为吟苑公园、崇安寺、锡惠景区、东林书院、南禅寺、薛福成故居、无锡博物馆、太湖鼋头渚景区、影视基地等，其次为梅园、宜园、雪浪山景观园等也是大部分上海客源青睐的地方。



(a)



(b)

图 4-8 上海客源在无锡出游的热点区域分布图 (a 为整体热点区位、b 为局部热点区位)

4.3 节假日出游的空间地理流向流量挖掘

针对第三章获取到的签到数据的空间属性,引入基于距离的 O-D (Origin-Destination) 网络的 GIS 空间分析方法,探索上海客源到江苏浙江 A 级以上景区出游模式,通过可视化方法得出了各节假日在空间地理流量、流向上的出游特征,具体分析如下。

4.3.1 黄金周出游的空间地理流量流向分析

黄金周是指法定节假日中的 7 天假期,分别为国庆节和春节,因为这两个节日放假时间长,且时间集中,其出游模式和出游特点有别于其他时间段的出游特点,分析上海客源在国庆节和春节期间出游江苏浙江的地理空间模式,对于游客而言,可以给游客在国庆和春节期间出游江苏浙江时提供参考,对于旅游规划部门而言,分析出的结果具有一定的参考价值,积极发展热门旅游城市,对于有旅游发展潜力且旅游资源丰富的冷门旅游城市,找到问题的主要原因,可及时解决并进行开发。

1、国庆节出游的空间地理流向流量分析

国庆期间上海客源到江苏浙江 A 级以上景区出游的签到数据空间地理流量、流向分析如图 4-9 所示,本文利用 ArcGIS 中基于距离的 O-D (Origin-Destination) 网络空间分析方法将签到数据流量依次分为三个等级,一级出游线路表示最受上海客源喜欢的出游线路,二级出游线路次之,三级出游线路最少。国庆期间上海客源到江苏浙江 A 级以上景区出游的一级出游线路有 5 条,流向分别是上海-苏州、上海-杭州、上海-嘉兴、上海-无锡、上海-盐城;二级出游线路有 6 条,流向分别是:上海-南京、上海-扬州、上海-常州、上海-绍兴、上海-丽水、上海-宁波;三级出游线路有 13 条,流向分别是上海到连云港、徐州、宿迁、淮安、南通、泰州、镇江、湖州、舟山、衢州、金华、温州和台州。

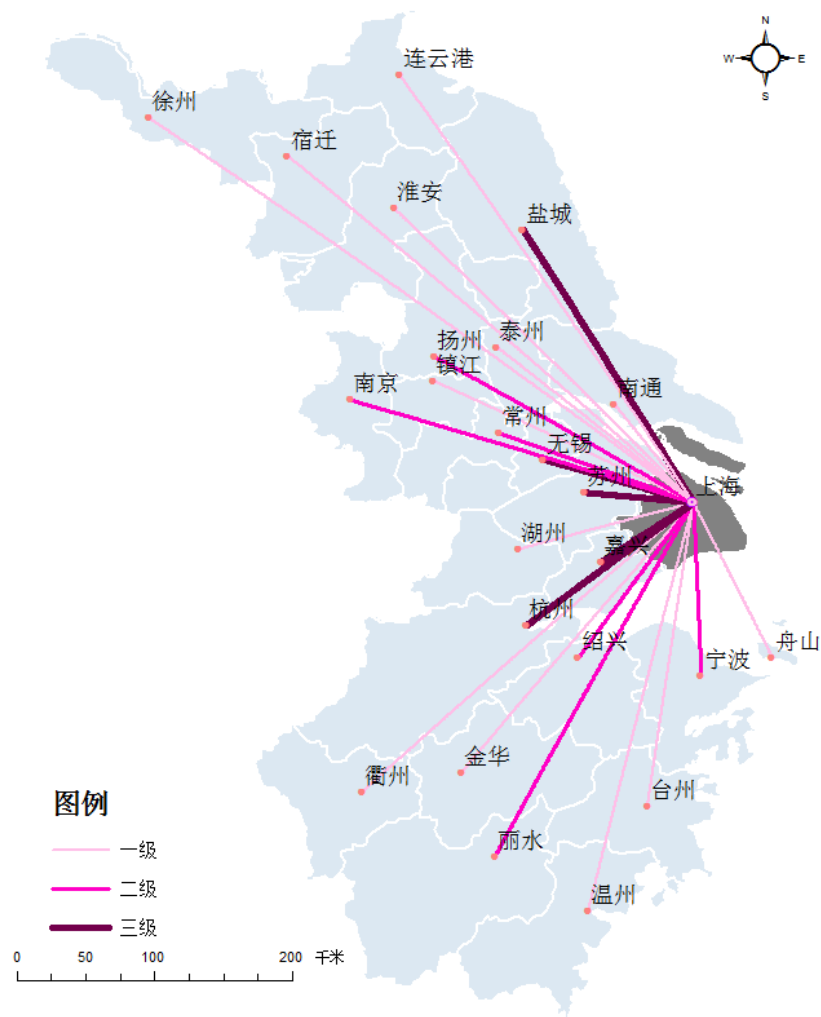


图 4-9 国庆节出游的空间地理流向

2.春节出游的空间地理流向流量分析

春节期间上海客源到江苏浙江 A 级以上景区的出游签到数据空间地理流量、流向分析如图 4-10 所示，在出游流量上，同样将出游签到数据分为三个等级，春节期间上海客源到江苏浙江 A 级以上景区出游的一级出游线路有 5 条，流向分别是上海-苏州、上海-杭州、上海-常州、上海-无锡、上海-扬州；二级出游线路有 6 条，分别是：上海-南京、上海-淮安、上海-盐城、上海-绍兴、上海-嘉兴、上海-宁波；三级出游线路有 13 条，分别是上海到连云港、徐州、宿迁、南通、泰州、镇江、湖州、舟山、衢州、金华、丽水、温州和台州。一级出游线路代表上海客源春节期间到江苏浙江出游最喜欢的线路，二级出游线路次之，三级出游线路只有很少一部分人喜欢。

春节期间的各级出游线路在流量上与国庆节相同，但是流向上有所差别。也就是说黄金周期间上海客源到江苏浙江出游的线路均比较丰富，但是不同黄金周

期间所关注的旅游目的地却又有各自的差别。

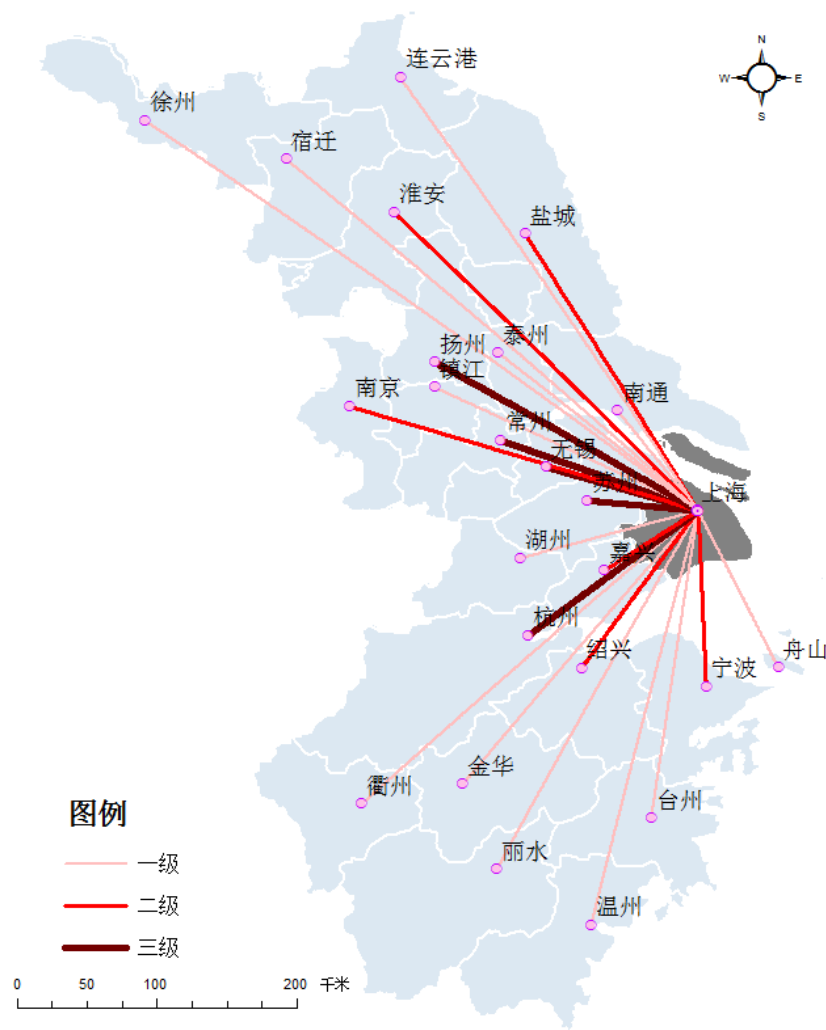


图 4-10 春节出游的空间地理流向

4.3.2 小长假出游的空间地理流向分析

小长假是指法定节假日中的 3 天假期，分别为元旦节、清明节、劳动节、端午节和中秋节，五个小长假在一年中的不同时段，且其节假日特征明显，比如清明和中秋这两个特殊的传统节日，游客的出游动机和出游目的明确，出游模式和出游特点也有别于其他时间段的出游特点，分析上海客源在小长假期间出游江苏浙江的空间地理模式，对于游客而言，可以给游客在小长假期间出游江苏时提供参考，对于旅游规划部门而言，可根据各节日特点各节日出游热点城市推动主题文化旅游模式。

1、元旦节出游的空间地理流量流向分析

元旦节期间上海客源到江苏浙江 A 级以上景区的出游签到数据空间地理流量、流向分析如图 4-11 所示，在出游流量上，同样将出游签到数据分为三个等级，元旦节期间上海客源到江苏浙江 A 级以上景区出游的一级出游线路有 4 条，流向分别是上海-苏州、上海-杭州、上海-常州、上海-无锡；二级出游线路有 6 条，分别是：上海-南京、上海-南通、上海-扬州、上海-镇江、上海-嘉兴、上海-绍兴；三级出游线路有 9 条，分别是：从上海到连云港、徐州、淮安、盐城、湖州、金华、温州、宁波和舟山。而从上海到宿迁、泰州、徐州和丽水则没有出游签到记录。一级出游线路代表上海客源元旦节期间到江苏浙江出游最喜欢的线路，二级出游线路次之，三级出游线路吸引力最小。

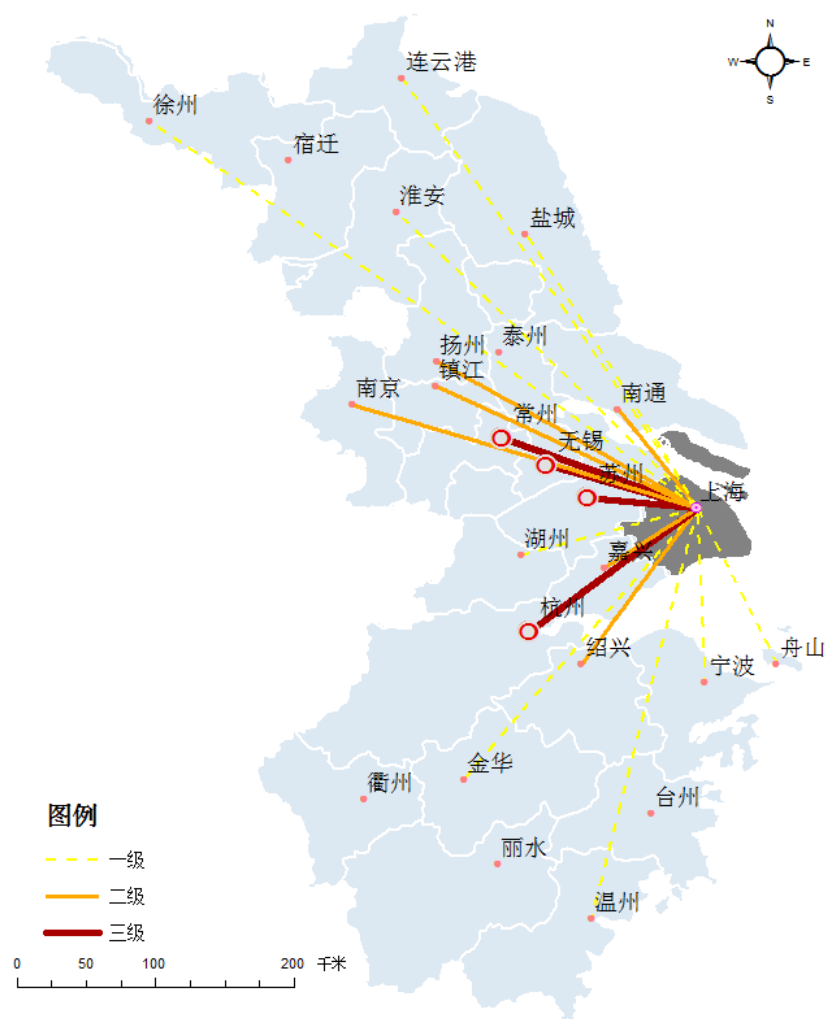


图 4-11 元旦节出游的空间地理流向

2、清明节出游的空间地理流量流向分析

清明节期间上海客源到江苏浙江 A 级以上景区的出游签到数据空间地理流

量、流向分析如图 4-12 所示，在出游流量上，同样将出游签到数据分为三个等级，清明节期间上海客源到江苏浙江 A 级以上景区出游的一级出游线路只有 2 条，流向分别是上海-苏州、上海-杭州；二级出游线路有 7 条，分别是：上海-南京、上海-扬州、上海-常州、上海-镇江、上海-无锡、上海-嘉兴、上海-绍兴；三级出游线路有 8 条，分别是上海到淮安、徐州、泰州、南通、湖州、金华、台州和宁波。而从上海到连云港、宿迁、盐城、衢州、丽水、温州和舟山则没有出游签到记录。一级出游线路代表上海客源清明节期间到江苏浙江出游最喜欢的线路，二级出游线路次之，三级出游线路微乎其微。

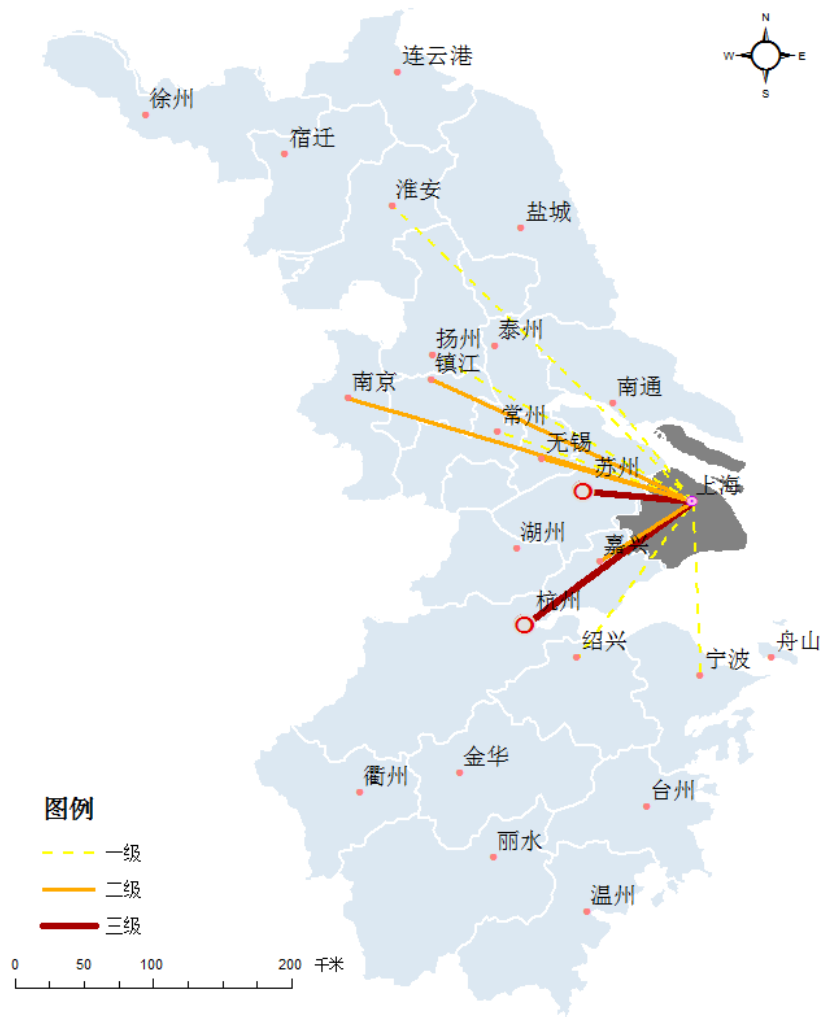


图 4-12 清明节出游的空间地理流向

3.劳动节出游的空间地理流量流向分析

劳动节期间上海客源到江苏浙江 A 级以上景区的出游签到数据空间地理流量、流向分析如图 4-13 所示，在出游流量上，同样将出游签到数据分为三个等

级，劳动节期间上海客源到江苏浙江 A 级以上景区出游的一级出游线路只有 3 条，流向分别是上海-苏州、上海-杭州和上海-无锡；二级出游线路有 5 条，分别是：上海-南京、上海-扬州、上海-常州、上海-嘉兴、上海-绍兴；三级出游线路有 10 条，分别是上海到徐州、连云港、盐城、南通、镇江、湖州、丽水、温州、台州和宁波。而从上海到宿迁、淮安、泰州、衢州、金华、舟山则没有出游签到记录。一级出游线路代表上海客源劳动节期间到江苏浙江出游最喜欢的线路，二级出游线路次之，三级出游线路只有很少一部分人喜欢。

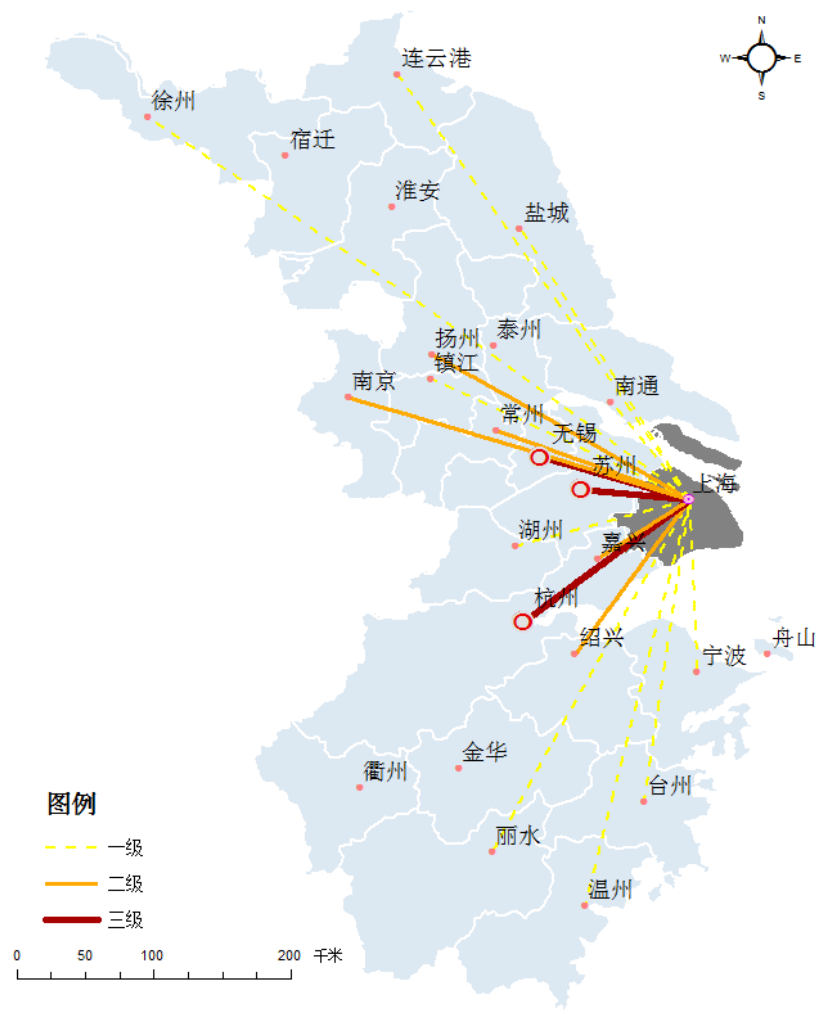


图 4-13 劳动节出游的空间地理流向

4.端午节出游的空间地理流量流向分析

端午节期间上海客源到江苏浙江 A 级以上景区的出游签到数据空间地理流向和流向分析如图 4-14 所示，在出游流量上，同样将出游签到数据分为三个等级，端午节期间上海客源到江苏浙江 A 级以上景区出游的一级出游线路只有 2

条，流向分别是上海-苏州、上海-无锡；二级出游线路有 4 条，分别是：上海-常州、上海-嘉兴、上海-杭州、上海-宁波；三级出游线路有 11 条，分别是：从上海到徐州、盐城、南通、镇江、扬州、南京、湖州、金华、绍兴、温州、舟山。而从上海到连云港、宿迁、淮安、泰州、衢州、丽水、台州则没有出游签到记录。一级出游线路代表上海客源端午节期间到江苏浙江出游最喜欢的线路，二级出游线路次之，三级出游线路只有很少一部分人喜欢。

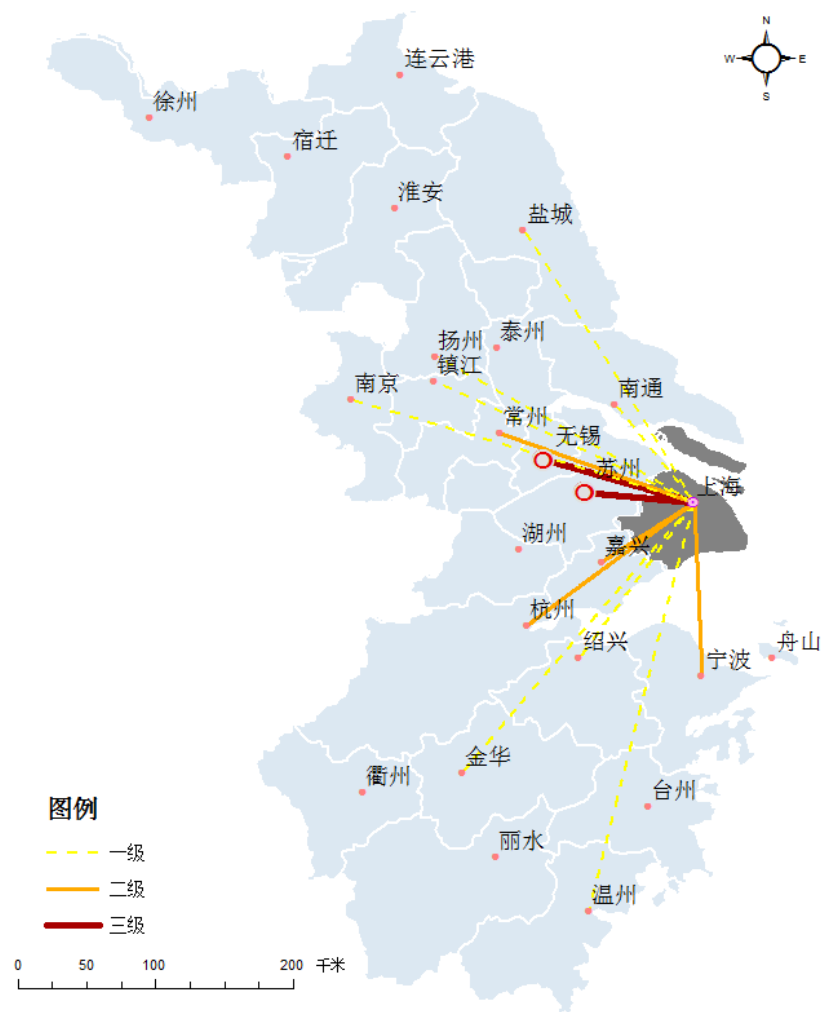


图 4-14 端午节出游的空间地理流向

5.中秋节出游的空间地理流量流向分析

中秋节期间上海客源到江苏浙江 A 级以上景区的出游签到数据空间地理流量、流向分析如图 4-15 所示，在出游流量上，同样将出游签到数据分为三个等级，中秋节期间上海客源到江苏浙江 A 级以上景区出游的一级出游线路只有 2 条，流向分别是上海-苏州、上海-盐城；二级出游线路有 4 条，分别是：上海-南

京、上海-无锡、上海-杭州、上海-嘉兴；三级出游线路有 7 条，分别是上海到徐州、镇江、常州、湖州、金华、绍兴、宁波。而从上海到连云港、宿迁、淮安、泰州、扬州、衢州、丽水、温州、台州、舟山则没有出游签到记录。一级出游线路代表上海客源中秋节期间到江苏浙江出游最喜欢的线路，二级出游线路次之，三级出游线路几乎可忽略不计。

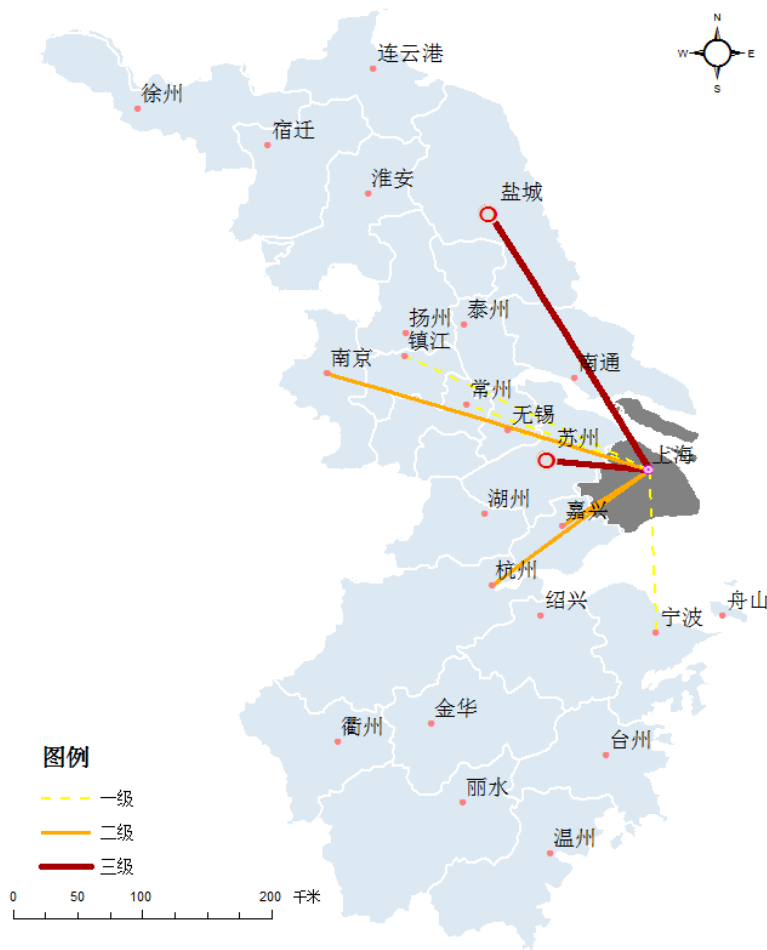


图 4-15 中秋节出游的空间地理流向

4.3.3 双休日出游的空间地理流量流向分析

双休日出游即利用周六周日的休息时间外出游玩，放松身心，在前面章节中已经论述到近年来双休日出游态势增长明显，在未来双休日休闲旅游将是一个新的增长点，主要原因是人们的物质生活水平提高，精神需求也在提高，出游意识和出游欲望增加，人们不再局限于节假日才外出旅游，周末近郊游也慢慢成为人们青睐的旅游方式，同时，因为双休日特点：时间短，只有两天的出游时间；出游时间多，每个周都有双休日出游时间，所以双休日的出游模式和出游特点也有

别于其他时间段，分析上海客源在双休日期间出游江苏浙江的地理空间模式，对于游客而言，可以给游客在双休日出游江苏时提供参考，对于旅游规划部门而言，可根据双休日特点开发相应的双休日旅游产品。

双休日上海客源到江苏浙江 A 级以上景区的出游签到数据空间地理流量、流向分析如图 4-16 所示，在出游流量上，同样将出游签到数据分为三个等级，双休日上海客源到江苏浙江 A 级以上景区出游的一级出游线路有 3 条，流向分别是上海-杭州、上海-苏州、上海-无锡；二级出游线路有 3 条，分别是：上海-南京、上海-常州、上海-嘉兴；三级出游线路有 8 条，分别是上海到徐州、南通、扬州、镇江、湖州、绍兴、宁波、舟山。而从上海到连云港、宿迁、淮安、盐城、泰州、衢州、金华、丽水、台州和温州则没有出游签到记录。一级出游线路代表上海客源双休日期间到江苏浙江出游最喜欢的线路，二级出游线路次之，三级出游线路只有很少一部分人喜欢。

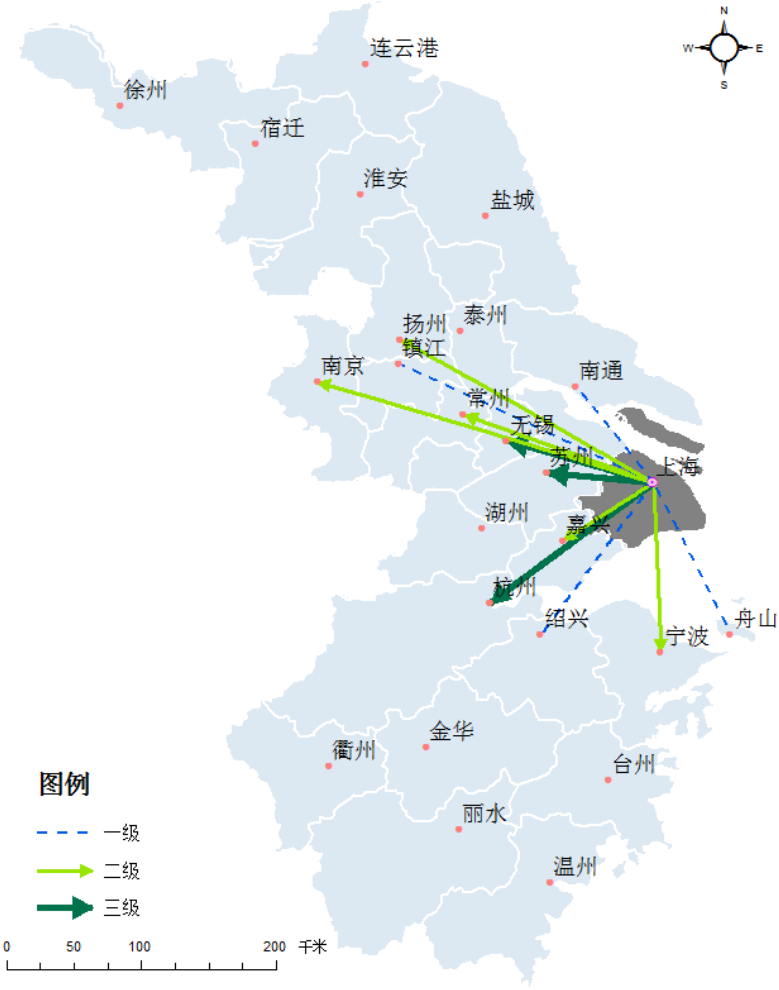


图 4-16 双休日出游的空间地理流向

4.3.4 节假日与双休日出游的空间地理流量流向对比分析

对比上海客源在黄金周、小长假和双休日期间出游江苏浙江 A 级以上旅游景区的空间地理流向流量可视化结果,可以总结出一定的出游规律,黄金周期间的一级出游路线、二级出游路线和三级出游路线包含江苏浙江两省 24 各市均有出游签到记录,小长假的出游线路没有黄金周出游线路丰富,主要以一二级出游线路为主。对比 5 各小长假的地理空间流量,元旦节、劳动节和端午节的出游线路要比清明节和中秋节丰富,出游流向上,以苏州、杭州、无锡、常州、嘉兴居多。中秋节和清明节两个传统节日特征明显,各级出游线路相比于其它三个小长假要少得多。双休日出游的各级出游线路均产生在上海市周边,距离上海市较远的连云港、宿迁、淮安、盐城、泰州、衢州、金华、丽水、台州和温州等城市均无签到记录。出游距离的远近依赖于出游时间的长短,假日时间长且集中的黄金周,大多数人会选择远距离出行,时间相对短的小长假,大多数人会根据时间、交通便利情况,预算等因素制定中长距离的出游,而对于时间较短且每个周末都有的双休日,介于工作、生活的安排,更青睐于周边游。

第 5 章 结论与展望

5.1 论文研究成果

对社交媒体地理数据的研究是近年来在位置大数据领域的研究热点,本文首先从位置大数据的分类、研究框架和社交媒体地理数据的概念、特点及应用等方面进行了相关理论研究,并根据其数据特点设计实现了社交媒体地理数据获取程序及该类数据存储、管理及预处理的方式,通过对数据进行验证,目标数据提取,最后对获取到的社交媒体地理数据进行应用实例挖掘分析。

主要研究成果如下:

1、通过对位置大数据相关理论的研究及对社交媒体地理数据概念及特点的研究,设计实现了对海量社交媒体地理数据获取的方法,利用 API 数据访问接口,采用 Java SDK 开发包,基于 Eclipse 开发平台设计实现了社交媒体地理数据的抓取程序。以社交媒体地理数据中的签到数据的获取为例,通过微博 API 中的位置服务接口和位置地点动态接口获取了上海客源到江苏浙江两省 A 级以上景区出游的签到数据;最终获取到 2013~2015 年在江苏浙江 A 级以上景区的签到数据达 1,887,836 条记录;其中提取上海客源的签到数据量为 92249 条记录。

2、研究了海量社交媒体地理数据存储的方法;设计实现了三个签到数据预处理中间程序,分别为签到数据整合、兴趣点上的签到人数统计和签到时间字段解析,实现海量签到数据的快速处理;完成了数据的清洗和验证,获取到目标数据。

3、分别从时间、空间两方面对获取到的社交媒体地理数据进行应用实例挖掘分析,得出了上海客源到江苏浙江出游的签到数据的年际变化特征,节假日变化特征,及节假日、周末和工作日的对比变化特征,发现上海客源的签到数据峰值出现在 2015 年;国庆节和春节的出游签到记录最多,其次是元旦节、劳动节和端午节,出游签到记录最少的是清明和中秋;对比节假日、周末和工作日各阶段的出游签到数据都发现节假日最多、周末次之、工作日最少,且在 2015 年的数据中周末和工作日的出游签到数据比例大幅上升。在空间上,利用 ArcGIS 空间分析方法,探索了上海客源感兴趣的热点区域,发现签到数据的密度高峰值出现在南京、杭州、苏州、无锡,集中度远高于其他地区;最后对黄金周、小长假、双休日期间的出游模式及出游特点进行分析发现,黄金周期间出游线路丰富,春节期间的各级出游线路在流量上与国庆节相同,但是流向上有所差别;小长假的

出游线路没有黄金周出游线路丰富，主要以一二级出游线路为主。对比 5 各小长假的地理空间流量，元旦节、劳动节和端午节的出游线路要比清明节和中秋节丰富，出游流向上，以苏州、杭州、无锡、常州、嘉兴居多。双休日出游的一二三级线路均集中在上海周边城市，距离上海较远的地区双休日几乎没有出游签到记录。出游流向上，更青睐于以苏州、杭州和无锡等地。

5.2 论文创新点

论文研究在以下几个方面提出了新的见解和思路：

- 1、本研究基于应用程序编程接口 API 开发实现了快速获取海量社交媒体地理数据的方法，并根据最终获取到数据的应用目的改进了接口中请求参数的值，提高了目标数据的质量和精度。
- 2、有机的将 GIS 空间分析方法与社交地理大数据相结合，来分析大量移动对象群体轨迹背后隐藏的信息，真实再现用户在现实世界中的生活轨迹。
- 3、本研究从客源的角度出发，对某一类客源的时空行为模式的研究进行了一个全新的尝试，打破了以往客源数据获取困难，只能研究出游目的地的瓶颈，为某一类客源旅游产品的开发提供了新的思路。

5.3 论文研究的不足之处及展望

现对论文中存在的不足和待改进之处总结如下：

- 1、数据获取来源单一，本次研究区域大，时间短，仅选择了微博签到数据来说明社交媒体地理数据获取的方法，在今后的研究工作中，考虑选择更多的数据源进行尝试，来研究不同类型的社交媒体地理数据获取方法，更多的数据源进行挖掘研究可相互验证，来说明数据的可信度。
- 2、此次研究工作仅选择了上海市、江苏省、浙江省作为研究区域，在今后的工作中考虑扩大研究区域，数据的挖掘应用研究除应用于旅游行业之外，还将应扩展到其它领域。
- 3、此次数据获取仅调用了微博两个 API 接口，数据返回值有限，在今后的工作中，会考虑同时选取多个数据接口，对用户行为进行深入挖掘。

参考文献

- [1] 欧其健, 徐永书, 夏定辉. 地理信息服务的思考与探究[J]. 地理空间信息, 2011, 09(3):79-80.
- [2] Miller H J. Geographic data mining and knowledge discovery[M]. Handbook of Geographic Information Science, 2008: 352-366.
- [3] 王贺封. 时空数据模型及TGIS研究[J]. 测绘与空间地理信息, 2006, 29(4):11-13.
- [4] 刘经南. 泛在测绘与泛在定位的概念与发展[J]. 数字通信世界, 2011(S1):28-30.
- [5] Harrower M, Keller C P, Hocking D. Cartography on the Internet: Thoughts and a Preliminary User Survey[J]. Cartographic Perspectives, 1997, 49(26):1037-1038.
- [6] Haklay M, Zafiri A. Usability Engineering for GIS: Learning from a Screenshot[J]. The Cartographic Journal, 2008, 45(2):87-97(11).
- [7] Jacobs N, Satkin S, Roman N, et al. Geolocating Static Cameras[C]// IEEE, International Conference on Computer Vision. IEEE, 2007:1-6.
- [8] Park M, Luo J, Collins R T, et al. Estimating the camera direction of a geotagged image using reference images[J]. Pattern Recognition, 2014, 47(9):2880-2893.
- [9] Cristani M, Perina A, Castellani U, et al. Geo-located image analysis using latent representations[C]// Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008:1-8.
- [10] Leung D, Newsam S. Proximate sensing: Inferring what-is-where from georeferenced photo collections[C]// Computer Vision and Pattern Recognition. IEEE, 2010:2955-2962.
- [11] Zheng Y, Zhang L, Xie X, et al. Mining interesting locations and travel sequences from GPS trajectories[C]// International Conference on World Wide Web, WWW 2009, Madrid, Spain, April. DBLP, 2009:791-800.
- [12] Zheng V W, Zheng Y, Xie X, et al. Collaborative location and activity recommendations with GPS history data[C]// International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, Usa, April. 2010:1029-1038.
- [13] . Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012:186-194.
- [14] 王建华, 祝国瑞. 地图用户调查及结果分析[J]. 测绘通报, 1996(4):32-37.
- [15] 晏晓红. 基于空间认知的网络地图设计与评价研究[D]. 武汉大学, 2013.
- [16] 李响, 张晶, 江南, 等. 基于模糊综合评测法的移动电子地图分析研究[J]. 测绘通报, 2014(6):43-47.
- [17] You M, Chen C W, Liu H, et al. A usability evaluation of web map zoom and pan functions[J]. International Journal of Design, 2007, 1(1):15-25.
- [18] You M, Chen C W, Lin H. A usability evaluation of navigation modes in interactive maps[J]. 2009.
- [19] 凌云, 陈毓芬. 以用户为中心的电子地图集用户界面设计与实现[C]// 解放军信息工程

- 大学测绘学院博士生论坛. 2009.
- [20] 郑束蕾, 邓毅博, 陈毓芬, 等. 从眼动实验看地图学实验的哲学意义及作用[J]. 测绘科学, 2015, 40(1):28-32.
- [21] Gerber M S. Predicting crime using Twitter and kernel density estimation[J]. Decision Support Systems, 2014, 61(1):115-125.
- [22] Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages[C]// Soma. 2010:115-122.
- [23] Lamos V, Cristianini N. Tracking the flu pandemic by monitoring the social web[C]// International Workshop on Cognitive Information Processing. IEEE, 2010:411-416.
- [24] Paul M J, Dredze M. You Are What You Tweet : Analyzing Twitter for Public Health[J]. In: ICWSM (2011, 2011, 38:265-272.
- [25] Cheng Z, Caverlee J, Lee K. A content-driven framework for geolocating microblog users[J]. Acm Transactions on Intelligent Systems & Technology, 2013, 4(1):1-27.
- [26] Ji, Rongrong, Xie, et al. Mining city landmarks from blogs by graph modeling[J]. 2009:105-114.
- [27] Ye M, Shou D, Lee W C, et al. On the semantic annotation of places in location-based social networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August. DBLP, 2011:520-528.
- [28] Peng T C, Shih C C. Mining Chinese Restaurant Reviews for Cuisine Name Extraction: An Application to Cuisine Guide Service[C]// International Conference on Information Engineering and Computer Science. IEEE, 2009:1-4.
- [29] Gionis A, Lappas T, Pelechrinis K, et al. Customized tour recommendations in urban areas[C]// ACM International Conference on Web Search and Data Mining. ACM, 2014:313-322.
- [30] 王丽文. 基于社交网络的数据挖掘研究[D]. 西安电子科技大学, 2014.
- [31] 王瑞. 城市居民出行调查若干问题研究[D]. 长安大学, 2006.
- [32] Curran F B, Stegmaier J T. Travel patterns in 50 cities[J]. Highway Research Board Bulletin, 1958.
- [33] Murakami E, Wagner D P. Can using global positioning system (GPS) improve trip reporting?[J]. Transportation Research Part C Emerging Technologies, 1999, 7(2-3):149-165.
- [34] Wolf J L. Using GPS data loggers to replace travel diaries in the collection of travel data[C]// Dissertation, Georgia Institute of Technology, School of Civil and Environmental Engineering. 2000:58--65.
- [35] Kwella B, Lehmann H. Floating Car Data Analysis of Urban Road Networks[C]// Computer Aided Systems Theory - EUROCAST'99, Vienna, Austria, September 29 - October 2, 1999, Proceedings. DBLP, 1999:357-367.
- [36] Kwan M P. GIS Methods in Time-Geographic Research:Geocomputation and Geovisualization of Human Activity Patterns[J]. Geografiska Annaler: Series B, Human Geography, 2004, 86(4):267-280.
- [37] Patrick Laube Corresponding author, Stephan Imfeld, Robert Weibel. Discovering relative motion patterns in groups of moving point objects[J]. International Journal of Geographical Information Science, 2005, 19(6):639-668.
- [38] Yu H, Shaw S L. Exploring potential human activities in physical and virtual spaces: a spatio-

- temporal GIS approach.[J]. International Journal of Geographical Information Science, 2008, 22(4):409-430.
- [39] Bazzani A, Giorgini B, Rambaldi S, et al. Statistical Laws in Urban Mobility from microscopic GPS data in the area of Florence[J]. Journal of Statistical Mechanics Theory & Experiment, 2009, 2010(5):823-831.
- [40] 杨涛, 周征舫. 马鞍山市居民出行选择决策心理研究[J]. 城市规划, 1994(4):39-45.
- [41] 骆培聪. 福州市居民出行特征分析与城市交通发展对策研究[J]. 福建师大学报(自然科学版), 2002, 18(2):99-103.
- [42] 曲大义, 于仲臣, 庄劲松, 等. 苏州市居民出行特征分析及交通发展对策研究[J]. 东南大学学报自然科学版, 2001, 31(3):94-97.
- [43] 白永平, 张艳萍. 河谷型城市兰州市居民购物行为时空特征研究[J]. 西北师范大学学报(自然科学版), 2009, 45(6):111-115.
- [44] 周素红, 邓丽芳. 基于T-GIS的广州市居民日常活动时空关系[J]. 地理学报, 2010, 65(12):1454-1463.
- [45] 申悦, 柴彦威. 转型期深圳居民日常活动的时空特征及其变化[J]. 地域研究与开发, 2010, 29(4):67-71.
- [46] 张艳, 柴彦威. 北京城市中低收入者日常活动时空间特征分析[J]. 地理科学, 2011, 29(9):1056-1064.
- [47] 郭文伯, 张艳, 柴彦威, 等. 基于GPS数据的城市郊区居民日常活动时空特征——以北京天通苑、亦庄为例[J]. 地域研究与开发, 2013, 32(6):159-164.
- [48] 郭迟, 刘经南, 方媛, 等. 位置大数据的价值提取与协同挖掘方法[J]. 软件学报, 2014, 25(4):713-730.
- [49] Baidu. 地理信息_百度百科[EB/OL].: <http://baike.so.com/doc/486869-515558.html>, [2012].
- [50] 刘经南, 方媛, 郭迟, 等. 位置大数据的分析处理研究进展[J]. 武汉大学学报信息科学版, 2014, 39(4):379-385.
- [51] Mell P, Grance T. The NIST definition of cloud computing[J]. Communications of the Acm, 2009, 53(6):50-50.
- [52] 城市数据派[EB/OL].: <http://www.udparty.com/topic/1487.html>, [2015].
- [53] 朱立超, 李治军, 姜守旭. 基于位置的社交网络研究综述[J]. 数码世界, 2015(8):70-72.
- [54] 马荣华, 蒲英霞, 马晓冬. GIS空间关联模式发现[M]. 科学出版社, 2007.
- [55] Wang W, Yang J, Muntz R. An Approach to Active Spatial Data Mining Based on Statistical Information[J]. IEEE Transactions on Knowledge & Data Engineering, 2000, 12(5):715-728.
- [56] SINA. Weibo[EB/OL].: <http://open.weibo.com/wiki/微博API>, [2015].
- [57] 杭州新闻网[EB/OL].: <http://hznews.hangzhou.com.cn/>, [2014]
- [58] 时子庆, 刘金兰, 谭晓华. 基于OAuth2.0的认证授权技术[J]. 计算机系统应用, 2012, 21(3):260-264.
- [59] 沈霖. 基于众源地理数据的上海市旅游目的地关注度研究[D]. 上海师范大学, 2015.
- [60] 蒲文栋. 基于移动网络的新型汽车定位装置的设计与实现[D]. 电子科技大学, 2014.

攻读学位期间取得的科研成果

参与项目：

1. 参与导师主持完成课题组上海旅游高等专科学校项目：“旅游特色数据库系统”（2014--2015），主要承担数据收集、处理和文档编写；
2. 参与导师主持完成课题组项目：“智慧社区助老关爱平台”（2016--2017），主要承担数据收集和预处理。

参加学术会议：

1. 参加同济大学举办的国际会议 “The Second International Conference On Location-Based Social Media Data”，摘要收录，并在会议上发表汇报 2016,08。
2. 参加杭州“2016 世界休闲与旅游研究峰会”，摘要收录，并在会议上发表汇报，2016,11。
3. 参加香港大学举办的 “The Asia GIS Conference 2017” 会议，摘要收录 2017,01。

致谢

转眼间，近三年的研究生学习生活就要结束了，而入学仿佛还是昨天的事，刚进校园的场景还历历在目。回忆三年的点点滴滴，感慨不已，欣慰之余又庆幸无比。值得欣慰的是，三年时间里，在汗水和拼搏中度过，我学会了很多，也成长了很多，学到了许多受益终生的东西；庆幸的是我来到了一个无比温馨的大家庭，遇到了很多良师益友，给了我很多的指引和帮助，使我能够顺利的完成学业，在此谨向他们表示衷心的感谢！

最深的谢意要献给我的导师陈能教授。在学术上，您的严谨、认真、耐心、踏实的学术作风深深教导着我。您不仅仅是我学术上的导师，更是我人生道路上的一盏明灯，不管是生活中还是工作上遇到问题遇到瓶颈，您都会耐心的引导我去解决，从一点一滴的小事教育我做人做事的态度和方法，或许我很难说您的哪一句话让我幡然醒悟，但在您的教导下我开始学会，开始明白，开始真切的体悟怎样去做一件事情，怎样去做好一件事情。您的教导是我一生中最宝贵的财富。

其次，要感谢我的大师姐施蓓琦副教授，科研能力强，对待工作极其认真负责，她积极向上的态度深刻的影响了我，从论文撰写和项目实践两方面给我很大的悉心指导和帮助。还要感谢田冬迪师姐，她温柔细致，工作认真负责的态度深深影响了我，在我们一起工作的日子里，传授了我很多学习和工作的技能技巧。同时要感谢温家洪教授、尹占娥教授、林文鹏教授和陈家治副教授等各位我的老师，衷心感谢每一位老师，在我的研究生生涯里的每一节精彩的课程以及对每个问题的解答，教导我们，启发我们。研究生三年的日常生活学习，感谢周琦老师、贺球红老师为我解疑答惑、组织安排各种课余生活。

在研究生生涯里，除了学术知识，同时让我恋恋不舍的更是与同门师兄姐妹的情谊，沈霖师兄、唐律轩师兄、钟杰师兄、许丹苑师姐、林晓宁、陈旭师弟、戈丽师妹、曾麟婷师妹、王兵兵师弟和方俊睿师弟，感谢你们，三年里的每一个点滴，彼此共同带来的快乐与帮助。感谢我的同学：王从笑、曾颖、孙钰科、韩逸宁等，谢谢你们，我会永远铭记和你们在一起的日子。

最后，感谢我的父母，感谢你们的养育之恩，感谢你们经济和生活上对我的支持和帮助，以及给予我精神上的激励，让我能够顺顺利利走完这3年。同时，感谢我的兄弟姐妹，感谢我的男朋友，我知道如果没有你们的支持就没有我的今天，我也知道我永远都无法完全回报你们的爱，所以我希望顺利地完成学业，争取更大的成功，给你们一点欣慰。



知网查重限时 7折 最高可优惠 120元

本科定稿，硕博定稿，查重结果与学校一致

立即检测

免费论文查重: <http://www.paperyy.com>

3亿免费文献下载: <http://www.ixueshu.com>

超值论文自动降重: http://www.paperyy.com/reduce_repetition

PPT免费模版下载: <http://ppt.ixueshu.com>
