

本科生毕业论文

|  |  |
| --- | --- |
| 题目： | 基于社交媒体的旅游  时空行为数据库的建立 |

姓 名： 孙 奇

学 号： 1300012436

院 系： 地球与空间科学学院

专 业： 地理信息科学

指导教师： 高 勇

二〇一七 年 六 月

北京大学本科毕业论文导师评阅表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学 号 | 1300012436 | 学生姓名 | 孙奇 | 论文成绩 |  |
| 学院（系） | 北京大学地球与空间科学学院 | | | 专 业 | 地理信息科学 |
| 导师姓名 | 高勇 | 导师单位 | 地空学院 | 职称 | 副教授 |
| 论文题目 | | 基于社交媒体的旅游时空行为数据库的建立 | | | |
| 英文论文题目 | | Establishment of Tourism Time and Space Behavior Database Based on Social Media | | | |
| 导师评语  （包含对论文的性质、难度、分量、综合训练等是否符合培养目标的目的等评价） | | 基于POI签到微博提取用户的旅游行为，具有很高的准确度（92%），但是召回率低（29%）。其原因是该方法只利用了签到微博而忽略了非签到地理微博。本论文中，作者以景区POI签到微博作为训练集，使用密度制图方法作为学习模型训练得到分类器进行旅游语义标注，实现了从坐标点到景区语义；并在此基础上，提出了基于景区访问规则的旅游行为提取方法，建立了旅游行为数据库；实验评估结果显示这种数据驱动方法在具有较高准确度的基础上，大大提高了召回率；最后，基于旅游行为数据库作者实现了两个具体应用用以评估旅游行为数据库的有用性，包括（1）建立景区网络并基于网络中心度得到了景区等级结构，（2）基于社区发现算法计算城市旅游空间。  旅游已成为国民经济中的支柱产业，基于大数据的定量旅游研究成为研究热点。而从社交媒体中识别旅游行为建立旅游行为数据库是实现旅游定量研究的关键一环，因此本论文研究很高的现实意义。  本文的成果是在综述了国内外相关文献基础上独立完成的，说明作者具有一定的理论基础和实际科研能力，符合本专业培养的目标要求。  本论文文字较为流畅，研究逻辑清晰，符合本科毕业论文的要求。    导师签名： | | | |

（此表格供院系参考用，各院系可根据实际情况制定和使用原有本单位的论文评阅表）

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

近年来，随着智能手机的普及和通讯网络的进步，人们在社交网络上的活动越来越多。在社交媒体上，人们热衷于分享自己的生活，感想，经历，其中也就包含着旅游活动。无时无刻，社交媒体上产生着大量的数据，信息繁杂有待挖掘，可以说社交媒体为研究游客的旅游时空行为提供一种新的数据源。

在本文的研究中，我们旨在基于社交媒体数据，提取用户的旅游时空行为，建立数据库以供后续研究。在本文中，首先定义了旅游时空行为模式，为研究提供了理论框架，然后提出和讨论了旅游时空行为的提取方法，并实际应用抓取了社交媒体数据，建立行为数据库，最后利用行为数据研究目的地景区网络。

本文研究目的地城市以苏州市为例，社交媒体以新浪微博为例，编写程序抓取苏州相关的地理微博，并建立苏州旅游时空行为数据库。本文对旅游语义标注方法和旅游行为识别方法进行了实验评估，在模型参数K取0.17时，旅游语义标注方法F值最大，为85.96%，在模型参数k取0.17时，旅游行为识别方法F值最大，为81.35%。实验证实了本文提出的旅游时空行为提取方法的可行性，并给出了参数适当的选择范围 。

本文给出了一种普适性的从社交媒体数据提取旅游时空行为并建立数据库的技术路线，实验结果表现良好。旅游语义标注方法可以扩大可用数据的范围，使研究人员挖掘更多的信息。在后续研究中，可以应用至其他城市，分析全国范围内的旅游时空行为。

关键词：旅游,社交媒体,大数据,行为分析

Establishment of Tourism Time and Space Behavior Database Based on Social Media

Qi Sun (Geographic Information Science)

Directed by Yong Gao

ABSTRACT

In recent years, with the popularity of smart phones and communications network progress, people in the social network activities have been more and more. In social media, people are keen to share their lives, feelings, experiences, which also contains tourism activities. There is a lot of data on social media, the information is complicated to be excavated, it can be said that social media for the study of tourists travel time and space behaviors to provide a new data source.

In this study, we aim to extract the user's travel time and space behavior based on social media data and establish a database for subsequent research. In this paper, we first define the pattern of tourism time and space behavior, provide a theoretical framework for the study, and then propose and discuss the method of tourism time and space behavior extraction, and the actual application of social media data, the establishment of behavior database, and finally the use of behavioral data Destination area network.

In this paper, we take the city of Suzhou the destination city, and Sina microblogging as the social media. We prepare the procedures to capture the geography microblogging of Suzhou and build the Suzhou tourism time and space behavior database. In this paper, the tourism semantic annotation algorithm and the tourism behavior recognition algorithm are evaluated experimentally. When the model parameter K is 0.17, the tourism semantic annotation algorithm has the largest F value of 85.96%. When the model parameter k is 0.17, the tourism behavior recognition algorithm has the largest F value of 81.35%. The experiment proves the feasibility of the proposed method of tourism time and space behavior extraction, and gives the appropriate range of parameters.

This paper presents a general technical route to extract the spatial and temporal behavior of tourism from social media data and establish a database. The experimental results are in good condition. Tourism semantic annotation algorithm can expand the scope of available data, so that researchers dig more information. In the follow-up study, it can be applied to other cities, and can be used to analyze the nationwide tourism time and space behavior.

Keywords: Tourism, Social media, Big data, Behavior analysis

目录

[第一章 概述 1](#_Toc485045435)

[1.1 研究背景 1](#_Toc485045436)

[1.2 研究现状 1](#_Toc485045437)

[1.3.研究目的与内容 4](#_Toc485045438)

[1.4 技术框架 6](#_Toc485045439)

[第二章 旅游时空行为模型和旅游行为数据库设计 7](#_Toc485045440)

[2.1 旅游时空行为模式 7](#_Toc485045441)

[2.2 数据库设计 9](#_Toc485045442)

[第三章 方法 13](#_Toc485045443)

[3.1 常居地的提取 13](#_Toc485045444)

[3.2 旅游语义标记 13](#_Toc485045445)

[3.3旅游行为识别 18](#_Toc485045446)

[3.3.1 大尺度旅游行为 18](#_Toc485045447)

[3.3.2 中尺度旅游行为 20](#_Toc485045448)

[3.4 景区网络的建立与景区等级识别 21](#_Toc485045449)

[3.5 旅游区提取 23](#_Toc485045450)

[第四章 数据 25](#_Toc485045451)

[4.1 研究城市 25](#_Toc485045452)

[4.2 数据采集 25](#_Toc485045453)

[4.2.1 微博数据结构 26](#_Toc485045454)

[4.2.2 数据采集 29](#_Toc485045455)

[4.2.3 社交媒体数据抓取系统的设计与开发 30](#_Toc485045456)

[第五章 结果及评估 33](#_Toc485045457)

[5.1 常居地提取 33](#_Toc485045458)

[5.2 旅游语义标记 33](#_Toc485045459)

[5.2.1 旅游语义标记结果 33](#_Toc485045460)

[5.2.2 实验评估 37](#_Toc485045461)

[5.2.3 结果分析 40](#_Toc485045462)

[5.3旅游行为 41](#_Toc485045463)

[5.3.1 旅游行为识别结果 41](#_Toc485045464)

[5.3.2精度评价 42](#_Toc485045465)

[5.3.3 结果分析 43](#_Toc485045466)

[5.4 应用 44](#_Toc485045467)

[5.4.1 景区中心度及景区等级 44](#_Toc485045468)

[5.4.2 旅游区提取 45](#_Toc485045469)

[第六章 结论与展望 49](#_Toc485045470)

[6.1 结论 49](#_Toc485045471)

[6.2 展望 49](#_Toc485045472)

[6.2.1 旅游语义标注方法的改进 49](#_Toc485045473)

[6.2.2 旅游行为识别方法的改进 50](#_Toc485045474)

[6.2.3 小尺度旅游行为提取 50](#_Toc485045475)

[6.2.4 扩展旅游要素 51](#_Toc485045476)

[6.2.5 与信令数据结合 51](#_Toc485045477)

[参考文献 53](#_Toc485045478)

[致谢 55](#_Toc485045479)

[北京大学学位论文原创性声明和使用授权说明 57](#_Toc485045480)

第一章 概述

1.1 研究背景

传统旅游数据收集包括问卷调查，访谈等形式，然而这些方法不仅消耗大量人力财力，还十分耗时，数据收集时间与数据分析时间往往间隔很长，这种数据收集方法正在被逐步地淘汰。

随着时代发展，几乎所有人都会在互联网上留下自己的足迹，在新地理学(Neogeography)领域中，每个人都有着双重身份，既是数据的使用者也是数据的生产者。社交媒体是一个巨大而不断增长的地理位置信息的来源，具有巨大的潜力但仍然没有被充分利用(Chareyron et al., 2014)。每时每刻，人们在社交媒体上发布着自己的位置信息，由此产生海量的地理信息。正如大数据的一般特点，用户们不断生产的社交媒体数据数量大，种类多，获取速度快，但其的准确性和有效性还有待提高。我们需要从中提取感兴趣的信息，加之以利用。

大数据时代的到来，影响着各行各业，也包括这旅游方面。人们在旅游过程中，往往会将一些旅游信息发布在社交媒体上，我们可以从全部社交媒体数据中筛选出其中与旅游相关的数据，加以规则使之结构化后，建立旅游行为数据库，用以分析目的地形象或游客画像等。

1.2 研究现状

随着大数据时代的到来，数据源趋于多样化，使得研究者们可以通过不同数据源来研究旅游时空行为。主要使用的数据包括GPS数据，出租车行驶轨迹数据，手机终端定位数据，社交媒体数据等，其中社交媒体数据又可以细分为带地理标签的照片分享平台(如Flickr)和带地理位置的微博类平台(如twitter，微博)。

GPS数据是指志愿者穿戴GPS接收机设备进行旅游时空行为，记录下旅游者的时空轨迹，Modsching等人使用基于位置的跟踪数据来分析城市游客的运动，探究了GPS数据在研究旅游行为的可行性(Modsching et al.,2008)。李渊等人则将GPS数据与问卷相结合起来，研究鼓浪屿景区旅游者空间行为(李渊等,2016)。但是这种数据依赖穿戴者的意愿，加之成本的限制导致数据量较小，无法大规模推广使用。

出租车是人们出行不可或缺的一种交通方式。出租车数据是记录装有全球定位系统的出租车的行使轨迹。LiuYu等人使用上海出租车数据计算距离和方向分布，并构建了一个整合地理异质性和距离衰变效应的模型用以研究人类移动模式，其蒙特卡罗模拟结果与观察到的模式密切相关(Liu et al., 2012)。但是从中得到的人类移动模式被限制在城市路网上，而且无法从中提取游客的乘车行为，分析较困难。

手机同样是现代人常携带的电子设备，由基站信息可以获取手机用户准确的定位，近年来，智能机的普及更是提高了其定位精度。手机数据大致有两种被动移动定位和蓝牙轨迹。Kuusik等人利用手机被动移动定位(PMP)来研究目的地营销的重复访问，区分不同忠诚度的用户(Kuusik et al., 2011)。Versichele等人通过为期15天的蓝牙跟踪方法，在旅游景点访问中关联规则挖掘游客旅游行为，说明了旅游管理实践中蓝牙跟踪方法的潜力(Versichele et al., 2014)。手机数据覆盖范围广，数据量大，连续性较好(如信令数据)，但是涉及个人隐私，运营商不会提供大范围的数据，应用受限。

社交媒体，即在线社交网络，是一种在信息网络上由社会个体集合及个体之间的联系构成的社会性结构(方滨兴等, 2014)。社交媒体大致分为四类(1)即时消息类应用，如MSN，微信；(2)在线社交类应用，如人人网，Facebook；(3)微博类应用，如新浪微博，Twitter；(4)共享空间应用，如博客，BBS等(Huijboom N. et al., 2011)。而关于旅游行为的研究，主要集中在后两类。

Flickr是雅虎旗下的一个图片共享网站，基于其开放的数据，国内外学者进行了多项研究。Steven Van Canneyt等人从Flickr上的大量地理参考照片开始，使用Mean Shift聚类来确定城市内的兴趣点，然后估计随机用户在给定时间上下文中访问给定地点的概率，将该系统与基准系统进行比较，其中基准系统的唯一标准是该地方的总体受欢迎程度，效果可嘉(Steven Van Canneyt et al., 2011)。YAN-TAO ZHENG等人利用大量在线照片，利用马尔可夫链模型来研究不同的与景点相关的旅游运动模式之间的旅游流量，通过在旅游路线上进行序列聚类来分析旅行路线的拓扑特征。最后，通过在旅游路线上进行序列聚类来分析旅行路线的拓扑特征(YAN-TAO ZHENG et al.,2012)。Kohya Okuyama 和Keiji Yanai通过分析大量的地理标记照片，提取了许多Web用户的流行旅游地点和旅游路径，并提出了使用提取的旅行信息的旅行路线推荐系统(Kohya Okuyama et al., 2013)。Rob Feick等人通过Flickr API采集了2001-2012年加拿大温哥华市地理标记的照片，建立了照片地理数据库，并发现标签空间语义与空间聚合之间存在显着的相互作用(Rob Feick et al., 2014)。Luke Bermingham和Ickjai Lee提出从Flickr照片提取时空元数据的概念，将这些照片的集合结合在一起得到空间-时间实体移动轨迹，描述个人的运动的轨迹；同时提出了新兴的时空轨迹感兴趣区域挖掘和序列模式挖掘框架，得到更高质量的序列模式 ( Luke Bermingham et al., 2014)。Quan Vu等人引入了基于密度聚类(P-DBSCAN)和马尔可夫链的这两种数据挖掘技术来去除地理标记照片中噪声，并挖掘旅行行为模式(Quan Vu et al., 2015)。

而微博类应用，由于包含了更加丰富的信息，用户行为的多样化，使得研究起来困难些，但同样有许多学者进行研究。张子昂等人基于微博签到数据，从时段月份季节尺度分析旅游活动时间特征，采用Kernel核密度分析方法研究旅游活动空间特征(张子昂 等，2015)。谷文林等人抓取新浪微博，通过提取关键词的方式，对农家乐乡村旅游的时空分布规律进行研究，分析乡村旅游的热点所在(谷文林等, 2015)。Hawelka等人分析了有地理位置的Twitter消息，考虑了流动性，回转半径，目的地多样性以及流入和流出的平衡等特征，以发现全球人类流动模式。对Twitter移动网络的社区结构的分析，揭示了跟随世界区域划分的空间连贯的区域，证实Twitter可以作为量化全球移动模式的数据源(Hawelka et al., 2014)。

各个数据源的属性特点总结如下表：

表1.1 不同数据源属性对比表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据属性 | GPS数据 | 出租车数据 | 手机数据 | 社交媒体 | |
| 共享空间类 | 微博类 |
| 数据规模 | 小 | 大 | 大 | 大 | 大 |
| 定位精度 | 高 | 高 | 中等 | 低 | 低 |
| 公开性 | 保密 | 部分公开 | 保密 | 公开 | 公开 |
| 空间尺度 | 小(景区) | 中(城市) | 大(全球) | 大(全球) | 大(全球) |
| 轨迹完整性 | 完整 | 完整 | 部分完整 | 不完整 | 不完整 |
| 语义信息 | 无 | 无 | 无 | 丰富 | 极丰富 |

可以看出，通过社交媒体中微博类应用对旅游时空行为的相关研究较少，其中主要研究方面集中在旅游区热点，旅游路线推荐以及旅游移动模式上。基于社交媒体对用户旅游时空行为的研究还有所欠缺；在数据源方面，以往研究多采用微博的签到数据，而未能使用带有地理坐标的普通数据，本文提供一种技术路线用来从普通地理微博中快速建立旅游目的地用户的时空行为数据库。

1.3.研究目的与内容

本文的研究目的，是要从社交媒体数据中，提取旅游者的旅行时空行为，构建旅游时空行为数据库。

研究内容主要有：

(1)对旅游时空行为建模，设计旅游行为数据库。

给出旅游时空行为的定义及相关概念的定义，包括游客，出发地，出发时间，返回时间，逗留时间。确定旅游时空行为尺度等级，给出大尺度，中尺度，小尺度时空行为定义。根据旅游行为模型，设计对应的数据表结构，构建旅游行为数据库。

(2)旅游时空行为提取技术的研究。

按照制定的规则，提取社交媒体用户的常居地城市。结合空间密度聚类和DENCLUE算法思想，提出一种旅游语义标记方法，以实现从坐标到旅游场所语义的转换。基于规则的方法，提取社交媒体的旅游时空行为。

(3)抓取社交媒体数据，建立旅游行为数据库。

编写程序，基于社交媒体提供的API，抓取公开的社交媒体数据，建立地理社交媒体数据库。应用旅游时空行为提取技术，从数据库中，提取用户常居地，对有地理位置的社交媒体记录进行旅游语义标记，扩展原有数据库。从扩展后的社交媒体数据库，依据规则提取时空行为，并根据旅游时空行为模型所定义的模式，建立旅游行为数据库。

(4)基于行为数据库的应用

利用建成的行为数据库，提取旅游城市的景区网络，再由景区网络，应用社区发现的算法，提取景区模块结构。计算景区网络中心度，应用K-Means聚类算法，获得景区等级。

1.4 技术框架



第二章 旅游时空行为模型和旅游行为数据库设计

2.1 旅游时空行为模式

在世界旅游组织推荐的《旅游统计推荐标准》中，旅游被定义为“人们为了休闲、商务和其它目的，离开他们的惯常环境，去往他处并在那里逗留连续不超过一年的活动”(吴必虎等, 2010)。旅游行为则是游客搜集旅游产品相关信息进行决策、购买、消费和评价旅游产品全过程中所反映出来的心理过程和行为表现。旅游行为可以划分为四个主要行为过程：旅游动机行为、决策行为、时空行为和后续评价行为。其中，旅游时空行为是指，游客在常居地至目的地、目的地内部和返回常居地的旅游过程中的空间移动行为和时间分配行为。本文提出了一个多尺度的旅游时空行为形式化框架，如下表所示：

表2.1 旅游时空行为概念框架

|  |  |
| --- | --- |
| 组成要素 | 表示 |
| 游客 |  |
| 出发地 |  |
| 大尺度旅游时空行为 |  |
| 中尺度旅游时空行为 |  |
| 小尺度旅游时空行为 |  |
| 出发时间 |  |
| 返回时间 |  |
| 逗留时间 |  |

游客：进行该次旅游行为的游客。在传统旅游调查统计中，记录了游客的常居地、年龄、性别、学历、职业、出行目的、出现方式等信息。每个人都有着独一无二的信息，这些信息在我们分析用户行为中发挥了很大作用。名称，每个人的姓名，唯一标识人的属性，在社交媒体中，为尊重他人隐私，以互不相同的一串数字作为代替。

出发地：该次旅游出行的出发城市，通常是游客的常居城市。常居地，即用户经常居住的地点，在此处定义在城市尺度，这在区分客源地起到关键作用。

出发时间：在一次旅游中，游客离开常住地的时间。

返回时间：在一次旅游中，游客返回常住地的时间。

出游时间：一次旅游的时间，它等于出发时间和返回时间只差。

旅游者空间行为(Tourist Spatial Behavior)：研究广义上包含旅游者的动机行为，决策选择行为，旅游行为和体验行为4个过程，而狭义上的旅游者空间行为是指旅行行为的地域移动过程，旅游流是其核心研究内容(李渊 , 2016)。



图2.1 旅游者空间行为的概念内涵

大尺度时空行为(Large - scale space - time behavior)：在一次旅游中，游客从常居地出发，通过某种交通工具，依次到访目的地，最终返回常居地，形成一条时空序列。可以用一个三元有序对集合表示，其中是到访的第n个目的地，是到达第个目的地时间，是离开第个目的地的时间。在地域移动过程中，主要包含两个要素，交通和时间。

中尺度时空行为(Medium-scale space - time behavior)：在一次旅游中，游客在目的地内，依次到访过的旅游场所(主要包括景区、酒店、交通、饭店等) 所形成一条时空序列，用一个二元有序对集合表示，其中，是到访的第个旅游场所，是到访第个旅游场所的时间。

小尺度时空行为(Small - scale space - time behavior)：在一次旅游中，游客依次到访过的旅游有关位置(主要包括景区中的景点，地理坐标等) 形成的序列，用一个二元有序对集合表示，其中，是到访的第个位置，是到访第个位置的时间。

在本次实验中，我们只研究旅游者的大尺度和中尺度的时空行为，在旅游要素中，我们仅限定在时间上，其他的要素在以后的研究中再做讨论。

2.2 数据库设计

旅游行为数据库的设计，包含表的命名，字段的设计，表间的关系等等。旅游行为数据库至少包含5张表。

表2.2 用户社交媒体数据表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 含义 | 类型 | 备注 |
| Id | 唯一标识社交媒体记录的序号 | BigInt(20) | Prime Key |
| Create\_at | 创建时间 | Datetime |  |
| Text | 社交媒体记录文本 | VarChar(512) |  |
| Source | 社交媒体客户端类型 | VarChar(128) |  |
| Latitude | 纬度 | Double |  |
| Longitude | 经度 | Double |  |
| Userid | 发布社交媒体的用户序号 | BigInt(20) | Foreign Key |
| CityName | 社交媒体记录所在城市名称 | VarChar(255) |  |
| CityCode | 社交媒体记录所在城市序号 | VarChar(256) |  |
| ProvinceName | 社交媒体记录所在省份名称 | VarChar(256) |  |
| UserResidence | 用户常居地 | VarChar(256) |  |
| TravelBehaviorId | 用户行为序号 | Int(11) |  |
| MarkedPlace | 旅游语义标记场所 | VarChar(256) |  |
| PositionCity | 定位城市 | VarChar(256) |  |

表2.3 用户信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 含义 | 类型 | 备注 |
| Userid | 唯一标识用户的标识符 | Bigint(20) | Prime Key |
| Screen\_name | 昵称 | VarChar(128) |  |
| Create\_at | 账号注册时间 | Datetime |  |
| Gender | 性别 | VarChar(2) |  |
| Province | 常居地省份 | Int(11) |  |
| City | 常居地城市 | Int(11) |  |
| Weibo\_Count | 在常居地发社交媒体记录数目 | Bigint(20) |  |
| Active point | 常居地活跃点 | Point |  |

表2.4 行为信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 含义 | 类型 | 备注 |
| UserID | 用户序号 | Bigint(20) | Prime Key |
| TravelBehaviorId | 用户行为序号 | Int(11) | Prime Key |
| Source | 出发地 | Varchar(128) |  |
| Starting time | 开始时间 | Datetime |  |
| End time | 结束时间 | Datetime |  |
| Local behavior | 是否为本地行为 | BOOL |  |
| Behavior category | 行为性质 | Int(4) | -1：非旅游或未知；0：旅游；1：观光；2：休闲； |
| Travel mode | 出行模式 | Int(4) | -1：未知；0：环状；1：星状 |
| WeiboCount | 行为所包含社交媒体记录数目 | Int(11) |  |
| CitySequence | 用户所到城市序列 | VarChar(512) |  |
| PlaceSequence | 用户所到场所序列 | VarChar(512) |  |

表2.5 大尺度信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 含义 | 类型 | 备注 |
| Behavior Id | 行为ID | Bigint(20) |  |
| Large scale behavior ID | 大尺度行为ID | Int(11) | 记录城市次序，以0开始；其中0代表入口目的地 |
| City | 城市 | Varchar(128) |  |
| Starting time | 开始时间 | Datetime |  |
| End time | 结束时间 | Datetime |  |
| Transportation | 交通方式 | Int(4) | 到达城市的交通方式 |
| Local behavior | 本地行为 | Bool |  |
| Behavior category | 行为性质 | Int(4) | -1：非旅游或未知；0：旅游；1：观光；2：休闲； |

表2.6 中小尺度行为信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 含义 | 类型 | 备注 |
| Behavior Id | 行为ID | Bigint(20) |  |
| Large scale behavior ID | 大尺度行为ID | Int(11) | 记录城市次序，以0开始；其中0代表入口目的地 |
| Text | 社交媒体记录内容 | Varchar(512) | 社交媒体内容 |
| Location | 社交媒体记录发布位置坐标 | Point |  |
| Time | 社交媒体记录发布时间 | Datetime |  |
| Region | 位置对应的旅游场所 | Varchar(128) |  |
| Attraction | 位置对应的景点 | Varchar(128) | 如果场所景区，则坐标对应的景点 |
| Media ID | 该信息来自的社交媒体记录ID | Bigint(20) |  |

表2.7 目的地城市信息表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 含义 | 类型 | 备注 |
| TravelPlace | 旅游场所 | Varchar(256) |  |
| PlaceType | 场所类型 | Varchar(256) |  |
| TrainingSet | 样本集 | PointsArray |  |

第三章 方法

3.1 常居地的提取

常居地是限定在城市尺度上的，因为城市是具有明确的行政边界的，在社交媒体数据中，一个地理位置可以通过点到面的空间查询，得到点与面的空间关系， 以此确定地理位置所处的城市。

在微博的用户表中，存有用户的所在城市和省份，但由于这个信息是用户自己可以设置的，并不是十分准确，因此我们采用如下方法确定用户常居地：统计每个用户在各个城市所发微博数量，选取微博数量最多的城市，作为用户的常居地。

设一个用户，城市序列为，用户在各个城市所发微博数记为，则用户的常居地(，记为)应满足

若，则有。

若存在一个以上这样的，则选取用户在待选的常居地城市中，最早的一条微博所在城市视为用户U的常居地。

3.2 旅游语义标记

由于缺少旅游区，景区的边界数据，游客在发布了一个带地理坐标的社交媒体记录，如何界定这条社交媒体记录是属于哪个POI或场所就变得十分困难。在社交媒体记录数据中，签到社交媒体记录是关联了POI的，因此可以知道签到的用户去了哪些地点场所，但是更多的地理社交媒体记录没有关联POI，因此不能直接得到用户所取的地点，需要从坐标语义解释为文字语义。

由于所有的数据均为点的形式，我们考虑采用空间聚类算法来提取POI影响区域。考虑到景区场所的形状并不规则且多样化，我们需要找到一个能发现任意形状聚类形状的空间聚类算法，密度聚类DBSCAN是一个很好的选择。



图3.1 常居地提取流程图

社交媒体记录数据不同于一般数据，发社交媒体记录的位置遍布城市各处，且由于社交媒体记录定位机制的缘故，在同一位置(经度纬度完全相同)上可以进行多次定位，这就导致了数据点的重合问题，如果单纯的使用密度聚类，两个参数的选择就会变得很困难，最后分类结果，将使得所有的数据点划分为一类。因此不能进行简单的密度聚类。

同样社交媒体记录数据还不同于一般的实验数据只有待分类的数据点，社交媒体记录数据中的签到社交媒体记录是包含分类信息的，可以作为训练样本，以此根据训练样本得到训练区，来确定为分类数据点的类别。考虑到在样本数据中同样存在噪音和干扰点，先对样本数据进行处理。

下面做一些定义，数据点包含三个字段，纬度(,简记)，经度(，简记),重数(，简记)，经纬度标识该点的地理位置，重数表征在此位置重合的点数。设置重数，可以使减少重复计算。一个数据点可能关联着若干条社交媒体记录，因此数据点中不保存社交媒体记录的文本信息。

场所(Place)：表示一定的地理范围，表征了若干相近地物的集合体。

景区：若一个场所的所关联的地物都是与旅游相关的，那么就称它为景区。

景点：表示一个较小的地理范围，包含于景区内。

训练集()，我们需要首先确定一个场所所关联的社交媒体记录，并以这些记录作为训练集。我们要解决的问题就是是属于哪个的。

聚集类，简记为，是训练集在密度聚类后产生的若干个类别。每个类别包含若干的数据点。

在此我们提出一种密度聚类的改进方法。对于每个POI点，选取签到微博的数据点，作为训练样本，先对训练样本进行密度聚类，得到若干分类，依据数据点的重数与和其他数据点的空间关系，得到每个数据点的影响函数。对于每个实验数据点，计算每个类别对其的影响值，选取一定的阈值，若有类别影响值大于这一阈值，则选择影响值最大的类别作为该点的类别，若均小于阈值，则将该数据点标记为未分类。

(1)对训练样本进行密度聚类以去除噪声点，对于重数大于，可视为一类。

(2)经过密度聚类后，会得到个类，,若，则说明参数设置不合理，调整参数，再次聚类，或者训练样本本身存在错误，舍弃。若，则得到唯一的一个类。若,得到不止一个类，此处有两种做法，一是将这K类包含的所有的数据点都归为一类，在之后的计算中，将它们按一个整体来计算。二是保留这些类别，研究各类之间的空间关系。

(3)确定每个样本点p的影响函数,在此处，我们使用改进的高斯函数即

其中，表示实验点到样本点的欧氏距离，代表的重数，表示距离阈值，R表示地球半径，代表点的纬度值，表示点的经度值。

由于每个景区的空间尺度不尽相同，因此的取值不能一概而论，需要找到一个可以精确或近似衡量景区尺度的量。我们计算各个点距离其质心的平均距离

其中是所有样本点的几何中心，即的经纬度是所有样本点经纬度的算数平均数。

由此我们计算有公式如下

其中为一比例系数，用以调节距离阈值的大小。

每个实验点属于类别的密度函数为，其密度函数为 类中所有数据点影响函数之和，即

其中 是属于类 的样本点，为中包含的样本点数目。

每个实验点属于当前的密度函数,包含两部分，与每个类别的空间关系，以及每个类别之间的空间关系，那么密度函数可以如下表示，

其中为各类别之间的空间关系与所决定的密度函数。

若我们取密度聚类后的第二种做法，且设，即暂不研究各类别之间的空间关系则有

(4) 计算实验点在每个下的密度函数值，对超过阈值的区域，将实验点归为密度函数值最大的那个区域。



图3.2 旅游语义标注流程图

因此从旅游语义标记方法完整流程如下：

(1)选定一个要计算的区域

(2)确定该区域关联的POI集合

(3)从社交媒体数据库中提取这些上签到的微博，构成样本集

(4)对样本集进行去除噪声处理(DBSCAN)

(5)选定比例系数，阈值，构建区域影响模型。

(6)重复上述步骤，直到计算了所有区域，得到区域集合

(7)对于给定的坐标点P,计算其在各个区域的隶属度(影响值)，则有应满足且，对于

3.3旅游行为识别

采集到的社交媒体数据是不具有行为特征的，或者说不能明显体现用户活动行为的，需要对社交媒体数据进行分割。在本文中对行为划分，是按照一定的规则提取社交媒体用户的活动行为。

3.3.1 大尺度旅游行为

考察大尺度时空行为。我们制定如下规则，同一用户在连续若干天(N天)内发了数条(n条)社交媒体记录，且两条相邻的社交媒体记录的间隔时间不超过一个时间阈值，则将这n条社交媒体记录视作同一行为中发生的。

起止时间，一个行为的开始时间选取相关社交媒体记录中最早的时间，终止时间选取相关社交媒体记录中最晚的时间。

我们将重点放在社交媒体记录的大尺度定位上，即城市。由常居地提取方法我们可以知道一个用户的常居地，因此我们可以计算用户的一条行为中是否离开了常居地，去过哪些其他城市。因此制定如下规则：

用户的一条行为中，如果到达过一处及以上的非常居地城市，则这条行为是大尺度意义上的旅游行为。

具体方法如下：

(1)选取全部社交媒体记录，并按用户和时间的先后排序，添加字段，用以记录中尺度行为的编号。添加字段，用以标记旅游行为

(2)取第一条社交媒体记录，存下记录,用户与发布时间

(3)取下一条社交媒体记录，如果没有到末尾，存下记录，用户与发布时间，转至(4)；否则进行(8)

(4)如果，那么，转至(7)；否则进行下一步。

(5)如果，那么，转至(7) ；否则进行下一步。

(6)令

(7)将 值赋予 ，将值赋予，转至(3)

(8)对每个，，统计其包含的社交媒体记录中所到达的非常居地城市数目，如果，那么设



图3.3 大尺度旅游行为识别流程图

3.3.2 中尺度旅游行为

考察中尺度时空行为。我们制定如下规则，同一用户在连续若干天(N天)内发了数条(n条)社交媒体记录，且两条相邻的社交媒体记录的间隔时间不超过一个时间阈值，则将这n条社交媒体记录视作同一行为中发生的。

起止时间，一个行为的开始时间选取相关社交媒体记录中最早的时间，终止时间选取相关社交媒体记录中最晚的时间。

此处，我们将重点放在社交媒体记录的中尺度定位上，因此我们主要考察在目的地城市内定位的社交媒体记录。由常居地提取方法我们可以判断一个用户是否为外地用户。由旅游场所标记方法我们可以判断一个地理坐标(地理位置)是否位于景区内，因此我们可以计算用户的一条行为中是否去过景区，去过那些景区。因此制定如下规则：

外地用户的一条行为中，如果到达过一处及以上的景区，则这条行为是旅游行为。

具体方法如下：

(1)选取定位在研究城市内社交媒体记录，并按用户和时间的先后排序，添加字段，用以记录中尺度行为的编号。添加字段，用以标记旅游行为

(2)取第一条社交媒体记录，存下记录,用户与发布时间

(3)取下一条社交媒体记录，如果没有到末尾，存下记录，用户与发布时间，转至(4)；否则进行(8)

(4)如果，那么，转至(7)；否则进行下一步。

(5)如果，那么，转至(7) ；否则进行下一步。

(6)令

(7)将 值赋予 ，将值赋予，转至(3)

(8)对每个，，统计其包含的社交媒体记录中所到达的景区数目，如果并且不是研究城市，那么设



图3.4 中尺度旅游行为识别流程图

3.4 景区网络的建立与景区等级识别

设在3.2方法中定义了SN个区景区，可以建立景区与景区之间的关系矩阵。由3.3.2方法得到的用户旅游行为(中尺度时空行为)，设在一条旅游行为中出现了k个景区,则这k个景区中任意两个景区联系加一。我们以景区为节点，以两个景区同时出现在一次旅游行为移动序列中的次数作为边的权重，那么全部用户的全部旅游行为构成了一个无向有权图，为对称矩阵。

绝对程度中心度定义如下：

其中代表该景点与网络中其他景点相连接的个数。

相对程度中心度定义如下：

其中代表景区网络中所选择的景区个数。

计算景区的相对程度中心度，再对中心度进行聚类，得到景区K个中心度的等级。

具体算法：

(1)构建景区网络，阶方阵M

(2)对每个，，生成到达景区序列

(3)对于且，那么

(4)重复步骤(2)(3)，直至所有数据计算完成，景区网络构建完毕。

(5)利用软件，计算各个景区的中心度

(6)应用聚类算法，对进行聚类，得到K个等级。



图3.5 景区网络中心度及等级划分流程图

3.5 旅游区提取

基于旅游行为计算目的地城市旅游空间的基本流程如图3.5所示（详细内容见邵虎论文），主要包括以下步骤：

(1)使用规则网格（例如500m\*500m）对目的地城市进行空间划分；

(2)基于游客在目的地城市中的旅游时空行为构建以城市空间网格为结点的旅游行为网络G，用行为关联矩阵R(G)表示定义为：

(3)基于社区发现算法（例如算法）对城市网格进行社区划分，得到目的地城市的旅游空间。



图3.6 旅游空间提取流程

第四章 数据

4.1 研究城市

苏州，位于江苏省东南部，长江三角洲重要的中心城市之一，是国家历史文化名城和风景旅游城市，是中国首批24座国家历史文化名城之一。截止2015年1月，苏州市管辖5个市辖区：姑苏区（苏州国家历史文化名城保护区）、虎丘区、吴中区、相城区、吴江区；代管4个县级市：常熟市、张家港市、昆山市、太仓市。2016年末苏州市常住人口1062.57万人。市内交通包括苏南硕放国际机场、 苏州火车站、 苏州火车北站、苏州北广场汽车客运站、苏州汽车北站等。

苏州市以其独特的园林景观被誉为“中国园林之城”，素有“人间天堂”、“东方威尼斯”、“东方水城”的美誉。 苏州园林是中国南方园林文化的代表作。苏州市的著名景点包括太湖、 金鸡湖、 拙政园、 留园、 狮子林、 沧浪亭、 观前街、 虎丘、 寒山寺等。苏州古城遗存的古迹密度排名第三，苏州古城14.2平方公里。苏州古城和苏州园林为世界文化遗产和世界非物质文化遗产"双遗产"集于一身。

我们选择苏州市作为研究对象的主要原因是苏州市既是著名的旅游城市也是经济高度发达的大都市，因此外地人来苏州的目的即多又复杂，主要包括旅游、商务、学习、工作、探亲等，或者是以上目的的组合。

4.2 数据采集

新浪微博是中国的“twitter”， 是一个由新浪网推出，提供微型博客服务类的社交网站。用户可以通过智能手机、桌面终端等各种客户端组建个人社区，以 140 字左右的文字发布信息（2016年11月，新浪微博取消了发布器的 140 字限制，用户可以发布多达 2000 字的长微博），并实现即时分享。2009 年 8 月， 中国门户网站新浪推出“新浪微博”内测版，成为门户网站中第一家提供微博服务的网站，微博正式进入中文上网主流人群视野。截至2017年3月31日，新浪微博月活跃用户达3.4亿，已超过Twitter成为全球用户规模最大的独立社交媒体公司。

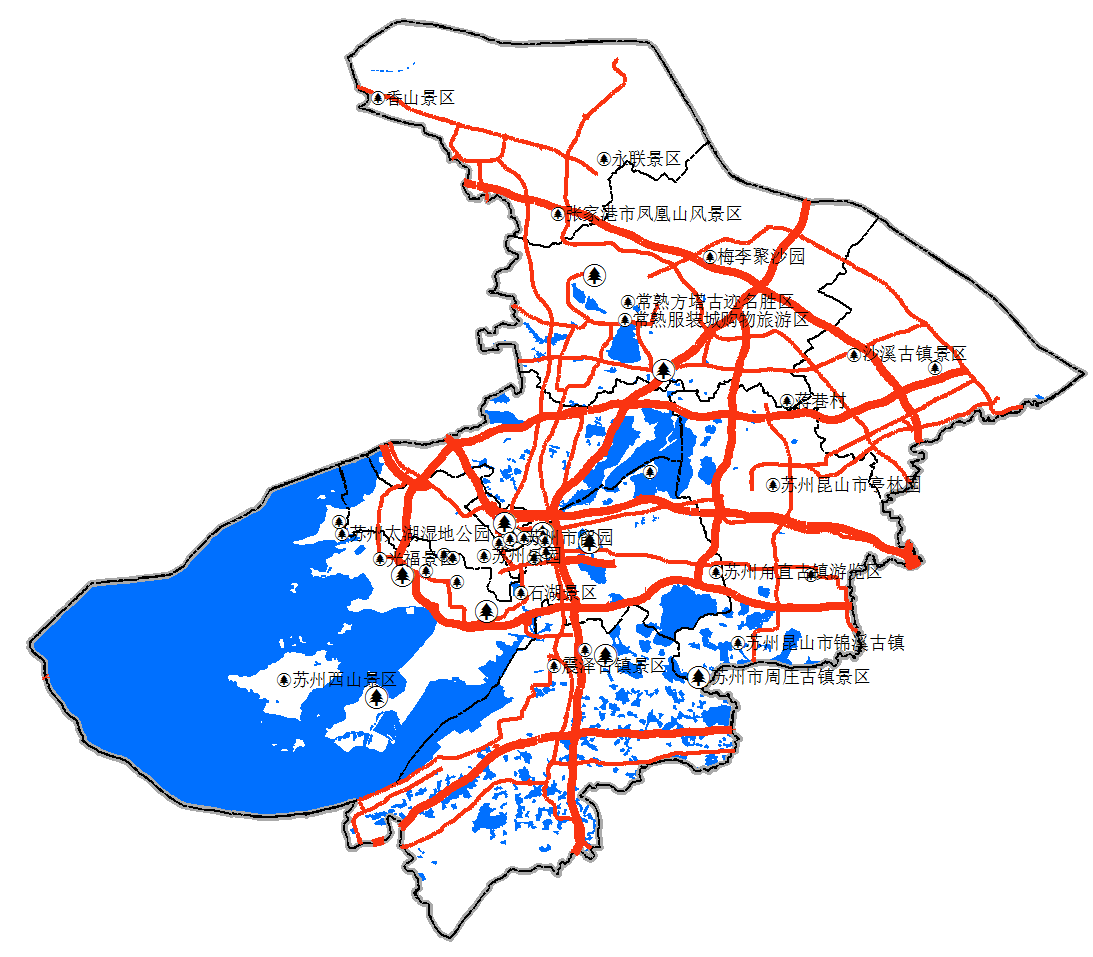


图4.1 苏州市行政区划、交通和旅游景区

4.2.1 微博数据结构

在本文研究中，我们通过新浪微博API获取了以下三类社交媒体数据用来提取游客旅游时空行为，建立旅游时空行为数据库。

**兴趣点信息**

通过新浪微博API，我们可以获取一个地理范围内的 POI 数据。每个 POI 数据都以一条记录的形式存在，并用Json格式传输。属性字段包括 POI编号（POIID）、名称（Title）、 经度（Longitude）、纬度（Latitude）、 类别（Category）、街道地址（Address）、在此 POI 签到的微博数量（Checkin\_num）和在此 POI 签到的用户数量（Checkin\_user\_num）。通过这些 POI 数据，我们可以获取目的地城市的旅游场所信息，主要包括景区、餐馆、宾馆等。

表4.1 新浪微博POI信息表示例

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| POIID | 名称 | 经度 | 纬度 | 类型 | 地址 | 签到数量 | 签到用户量 |
| B2094757D0  6EAAFB4893 | 苏州火车站 | 120.6108 | 31.32995 | 火车站 | 苏州市平江区车站路27号,京沪铁路,沪宁高铁,苏州站 | 23392 | 19539 |
| B2094650D1  6AABFF4098 | 苏州拙政园 | 120.6304 | 31.32469 | 国家级景点 | 平江区东北街178号 | 15903 | 13618 |
| B2094757D0  68A2F4469D | 周庄 | 120.846 | 31.11595 | 其他 | 昆山市西南隅(与吴江、吴县、上海青浦县交界) | 18032 | 12452 |
| B2094757D0  6EAAF44398 | 印象城购物中心(苏州店) | 120.6573 | 31.31779 | 商场 | 工业园区现代大道1699号(近苏嘉杭高速) | 15756 | 9918 |
| B2094757D0  6FA3FB459D | 同里古镇 | 120.7207 | 31.15749 | 其他 | 近郊吴江市同里镇 | 14040 | 8794 |

**用户信息**

本文的研究重点是游客的旅游时空行为，因此用户是必不可少的，新浪微博的API提供了单独的用户信息接口，同时在签到微博中也包含有用户信息。用户信息主要包括用户编号（Userid）、微博昵称（Screen\_name） 性别（Gender）、用户签名（Description）、省份代码（Province）、 城市代码（City）、 用户注册地位置（Location）、 好友数量(Friends\_count)、粉丝数量（Followers\_count）、 微博数量（Statuses\_count）、微博注册时间（Created\_at）等等。为保证社交媒体登的隐私安全，用户信息中并不涉及用户邮箱、手机、真实姓名、 生日等敏感信息。

表4.2 新浪微博用户信息表示例

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | 用户名 | 性别 | 用户签名 | 省份 | 城市 | 注册地 | 好友数 | 粉丝数 | 微博  数 | 注册时间 |
| 2738491671 | 冰蓝宇希 | f | 你发现了我，却已迟了 | 32 | 5 | 江苏 苏州 | 21 | 23 | 82 | Mon May 28 10:25:25 +0800 2012 |
| 2415213640 | HD-老兄弟 | m |  | 35 | 5 | 福建 泉州 | 132 | 166 | 558 | Wed Sep 21 22:58:21 +0800 2011 |
| 2668701633 | Miss徐小花 | f | 小花 | 21 | 1000 | 辽宁 | 85 | 391 | 2205 | Sat Mar 17 16:03:35 +0800 2012 |
| 1992451307 | 怀士堂 | m | 生命因经历而精彩 | 44 | 19 | 广东 东莞 | 296 | 76 | 384 | Sat Feb 26 14:17:20 +0800 2011 |
| 2896953953 | 魏忠- | m | 生活就像新闻联播，不是换台就能逃避得了的。——思想交流与探讨 | 35 | 7 | 福建 南平 | 38 | 22 | 618 | Mon Jul 23 01:51:35 +0800 2012 |

**地理微博**

用户微博包含了大量可挖掘的信息，因此用户微博是我们抓取数据的重点，特别地，我们研究带有地理坐标的微博。地理微博信息主要包含有微博 id（ID）、 发布用户的ID（Userid）、微博内容（Text）、 经度（Longitude）、纬度（Latitude）、发布时间（Created\_at）、 转发数量（Reposts\_count）、 评论数量（Comments\_count）。更进一步，如果用户在发布微博时位于某一POI附近，并且用户手动选择了这一个POI，则用户进行了一次签到，这样的微博称之为签到微博。

本文研究目标是从社交媒体数据中提取游客旅游行为，建立旅游时空行为数据库，所以我们只获取用户的地理微博和签到微博。

表4.3 地理微博示例

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 微博 id | 用户 id | 正文 | 经度 | 纬度 | 发布时间 | 转发数 | 评论数 | 签到 | 签到POI名称 |
| 3637754946762780 | 2785690262 | 阿布酸奶吧，小资情调额。 我在这里:http://t.cn/z8GvYI4 | 120.5934 | 31.21997 | Sat Oct 26 20:50:26 +0800 2013 | 0 | 0 | B2094650D169A2FA409E | 苏州工业职业技术学院 |
| 3533301149292510 | 1708293064 | 亲，我们房间好大的，双套居然是两张大床哦，你速来！@吉拉小姐 我在这里:http://t.cn/zle4R3u | 120.7077 | 31.45428 | Fri Jan 11 15:08:22 +0800 2013 | 3 | 22 | B2094757D16BA5F5409D | 阳澄万丽酒店 |
| 3572759118568130 | 1994306605 | 带娃去的好处！[偷笑] 我在这里:http://t.cn/zlmIz4M | 120.6 | 31.3098 | Tue Apr 30 12:20:15 +0800 2013 | 0 | 2 | B2094750D46EABF5449E | 海底捞火锅(广济南路店) |
| 3441616318929600 | 2028885851 | 新区出发全程15分钟到观前，我在这里:#苏州观前街# http://t.cn/z0spw6V | 120.6213 | 31.30798 | Thu May 03 15:05:35 +0800 2012 | 0 | 0 | B2094757D068A6FA4492 | 苏州观前街 |
| 3494805239404340 | 1714301067 | hey~裤衩~ 我在#现代大道# http://t.cn/zOgN6Vx | 120.6786 | 31.32331 | Thu Sep 27 09:39:22 +0800 2012 | 0 | 5 | B2094757D06BA5FA439A | 现代大道 |
| 3623237630070470 | 1237794980 | 酒吧里传出来的歌是《别再寂寞的拥抱》… 我在这里:http://t.cn/z8ApNUH | 120.598 | 31.32116 | Mon Sep 16 19:23:49 +0800 2013 | 0 | 4 | B2094757D06FA2FF459C | 山塘老街 |
| 3518157538637960 | 3002317225 | 在等车去无锡，明天一早回家 我在#苏州火车站二楼候车区# http://t.cn/zjq2OwV | 120.615 | 31.32941 | Fri Nov 30 20:13:05 +0800 2012 | 0 | 4 | B2094757D16DAAFA499D | 苏州火车站二楼候车区 |

4.2.2 数据采集

为了进行本文研究，我们通过新浪微博API抓取苏州市的社交媒体数据，抓取的基本思路如下（邵虎，2014）：

(1)选取一个目的地城市（本文是苏州市），通过新浪微博API(https://api. weibo.com/2/place/nearby /pois.json)，抓取该市的所有的 POI 数据。

(2)通过抓取到的 POI ，利用接口(https://api.weibo.com/2/place/poi\_timeline.json）抓取近期所有在该 POI 签到的签到微博和对应的用户信息。

(3)根据抓取到的用户信息表， 利用接口(<https://api.weibo.com/2/place/user> \_timeline .json)抓取用户过去所有的地理微博。 有相当一部分微博具有地理坐标但是并没有和任何一个 POI 关联，而本文正是要解决这个问题，如何实现从坐标语义到旅游场所语义的转换。

通过以上策略，我们抓取了苏州市2013年全年的社交媒体数据（如下表所示）作为提取游客旅游行为，建立旅游时空行为数据库的数据基础。

表4.4 苏州社交媒体数据

|  |  |
| --- | --- |
| 类型 | 数目 |
| POI | 171,045 |
| 签到微博 | 2,710,510 |
| 用户 | 739,331 |
| 地理微博 | 48,409,276 |
| 苏州市内地理微博 | 5,399,161 |

4.2.3 社交媒体数据抓取系统的设计与开发

根据上述数据采集的思想，我们编写了微博数据抓取系统。采用C#编程语言，目标框架.Net Framework 4，Windows窗体应用程序，采用多线程技术，可以较快较好地分步抓取POI信息，签到微博信息和地理微博信息，并实时掌控抓取进度。



图4.2 微博授权界面

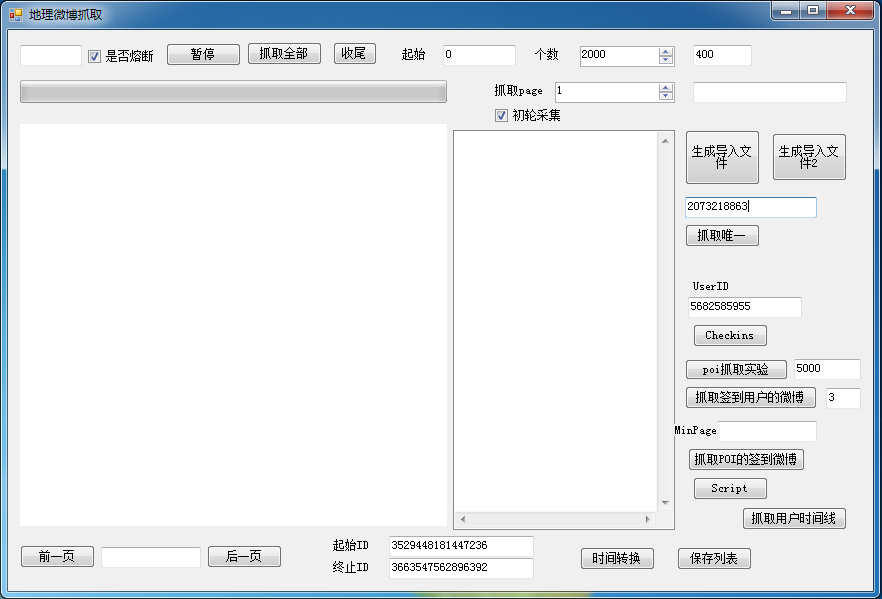


图4.3 程序主界面

第五章 结果及评估

5.1 常居地提取

应用3.1的方法提取用户的常居地，与微博信息中的注册地比对，有如下统计表：

表5.1 常居地提取方法与微博注册地比对表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 常居地提取的城市内累积发布微博条数 | 符合条件用户数 | 符合条件用户数占总用户数比例 | 常居地与注册地相同的用户数 | 常居地与注册地相同的用户占符合条件用户比 |
| >0 | 739331 | 100.00% | 510620 | 69.07% |
| >10 | 461160 | 62.38% | 359188 | 77.89% |
| >25 | 294935 | 39.89% | 239719 | 81.28% |
| >40 | 210392 | 28.46% | 174717 | 83.04% |
| >50 | 173090 | 23.41% | 145182 | 83.88% |

可以看出采用常居地提取方法所得结果与用户在微博信息中的注册地还是存在明显出入的，尤其是在用户所发的微博数较少的情况下，这种矛盾愈发突出。在用户微博较少的情况下(<10)，用户注册地可信度较高；而当微博数较多时(>50)，由提取方法得到的常居地城市更符合常理。为保证研究的一致性，本文所采用的用户常居地均采用3.1方法所提取到的。

5.2 旅游语义标记

5.2.1 旅游语义标记结果

首先，我们选择苏州市5A&4A景区以及主要旅游风情商业街作为旅游语义标注目标集，如表5.1第1列所示；其次，我们使用每个景区的签到微博作为景区样本集，如表5.1第2列和图5.1所示；再次，应用3.2中的方法，先对各个景区的样本集进行密度聚类，取值为500米，取值设为6，去出一些噪声，生成若干聚类。在上述的模型中，取阈值取值为6，取比例系数k取0.17，对苏州市内的微博进行旅游语义标注的计算；最后，旅游语义标注结果如表5.1第3列所示和图5.2所示。

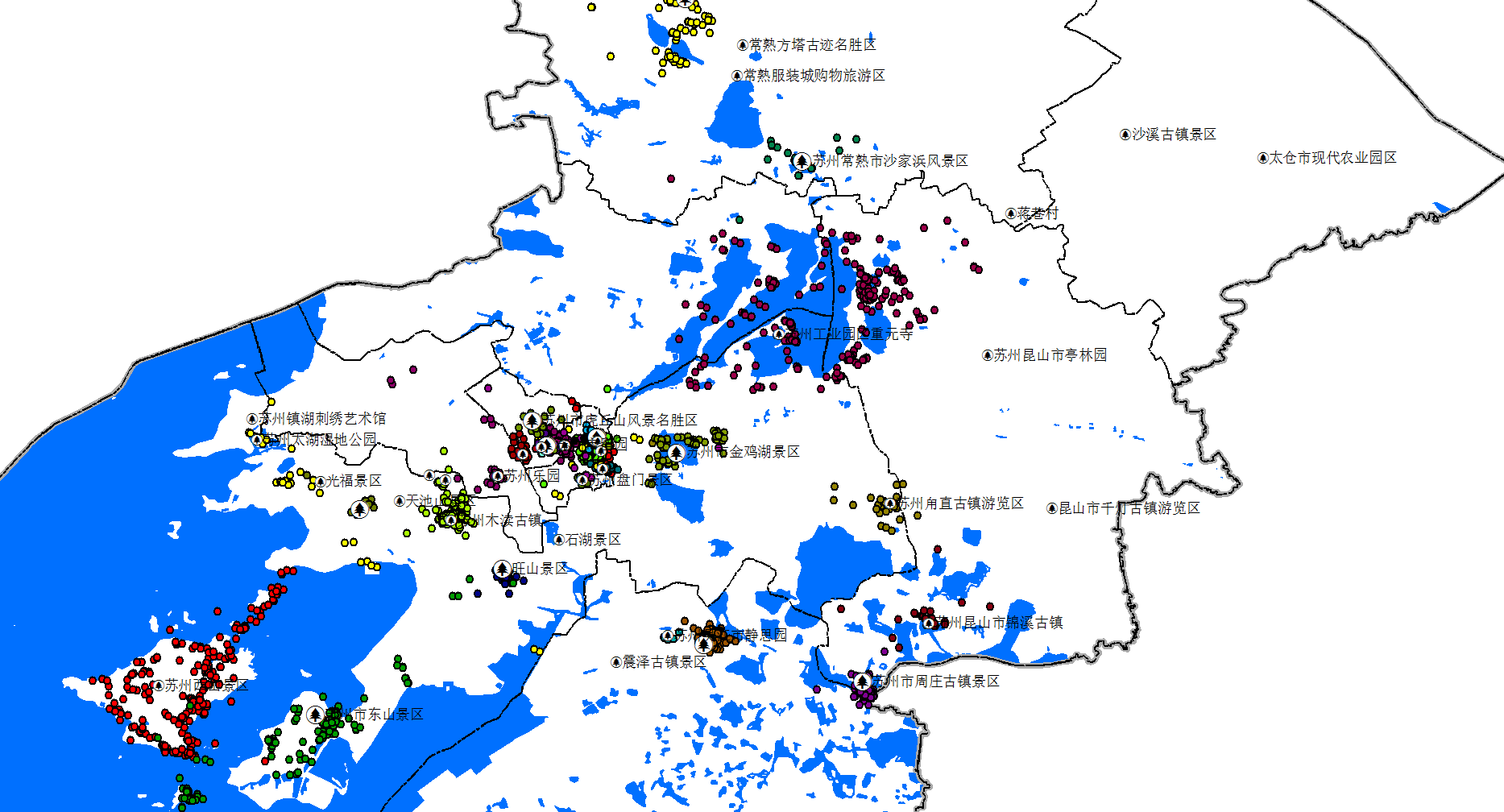


图5.1 苏州（主城区）景区样本集

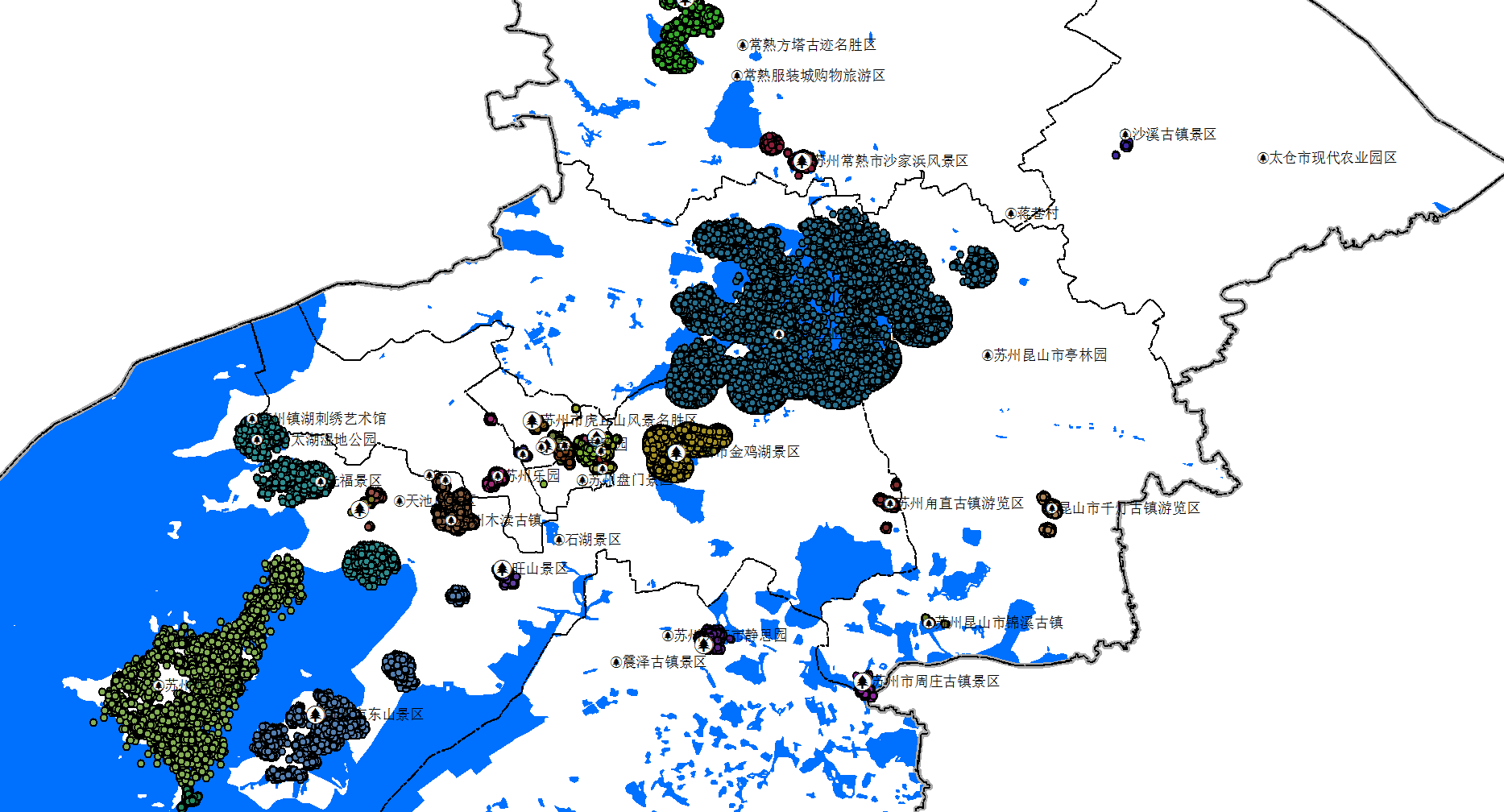


图5.2 苏州（主城区）旅游语义标注结果

表5.2 旅游语义标记统计数据(k=0.17)

|  |  |  |  |
| --- | --- | --- | --- |
| 景区名称 | 景区等级 | 样本集数目 | 语义标注总数（k=0.17） |
| 金鸡湖 | 5A | 28134 | 235178 |
| 观前街 | 商业街 | 37118 | 187337 |
| 阳澄湖 | 4A | 20477 | 82413 |
| 平江路 | 4A | 22370 | 75455 |
| 石路 | 商业街 | 4704 | 47352 |
| 木渎 | 4A | 10137 | 35302 |
| 山塘 | 4A | 9880 | 35769 |
| 周庄 | 5A | 16781 | 36437 |
| 同里 | 5A | 14518 | 31197 |
| 西山 | 4A | 8815 | 26312 |
| 苏州乐园 | 4A | 5653 | 22789 |
| 十全街 | 商业街 | 1658 | 14118 |
| 拙政园 | 5A | 13715 | 19594 |
| 东山 | 5A | 5343 | 16880 |
| 凤凰街 | 商业街 | 1712 | 13542 |
| 尚湖虞山 | 5A | 4497 | 13072 |
| 虎丘山 | 5A | 6626 | 10418 |
| 光福 | 4A | 3841 | 10406 |
| 寒山寺 | 4A | 8836 | 12102 |
| 狮子林 | 4A | 4969 | 8155 |
| 苏州博物馆 |  | 5710 | 9655 |
| 甪直 | 4A | 3581 | 7323 |
| 旺山 | 5A | 2051 | 7868 |
| 沙家浜 | 5A | 3167 | 5750 |
| 留园 | 5A | 2980 | 5115 |
| 千灯 | 4A | 2720 | 5285 |
| 杨舍 | 商业街 | 1429 | 4600 |
| 锦溪 | 4A | 2568 | 4145 |
| 十梓街 | 商业街 | 459 | 1975 |
| 沙溪 | 4A | 1468 | 2984 |
| 穹窿山 | 5A | 1806 | 2653 |
| 西园 | 4A | 897 | 1801 |
| 三山岛 | 5A | 1162 | 2147 |
| 藏书镇 | 4A | 795 | 1710 |
| 报恩寺塔 |  | 192 | 1193 |
| 网师园 | 4A | 629 | 1311 |
| 盘门 | 4A | 741 | 1165 |
| 静思园 | 4A | 508 | 851 |
| 香山 | 4A | 507 | 770 |
| 震泽 | 4A | 373 | 605 |
| 总计 |  | 263527 | 1002734 |

对于POI数据，进一步分类，手动标记识别POI所属景区，得到包括西山、阳澄湖、观前街、金鸡湖、周庄、平江路等40个景区分类。对每个景区包含的POI，选取在这些POI上签到的签到微博，作为这些景区的样本集。

可以看出，经过自动语义标注，数据量增加了2.81倍，一定程度上扩展了可用数据，为建立完整的旅游时空行为提供了数据支持。

更进一步，为研究模型中关键参数K的取值对标注结果的影响，本文又在对k的不同取值进行反复实验，得到如下表格：

表5.3 在K的不同取值下的旅游语义标注计算结果

|  |  |  |
| --- | --- | --- |
| 比例系数K的取值 | 有定位结果的微博记录数 | 占总微博数量的百分比 |
| 0.25 | 1226737 | 22.72% |
| 0.24 | 1196842 | 22.17% |
| 0.22 | 1140596 | 21.13% |
| 0.20 | 1086078 | 20.12% |
| 0.18 | 1031827 | 19.11% |
| 0.17 | 1002734 | 18.57% |
| 0.16 | 974757 | 18.05% |
| 0.15 | 943622 | 17.48% |
| 0.14 | 914625 | 16.94% |
| 0.13 | 882180 | 16.34% |
| 0.12 | 847079 | 15.69% |
| 0.11 | 816522 | 15.12% |
| 0.10 | 787161 | 14.58% |

5.2.2 实验评估

为了评估旅游语义标注效果，我们使用了信息检索领域中的三个基本指标准确率（Precision Rate）、召回率（Recall Rate）和F值。

准确率（P）衡量了语义标注的准确程度，它的定义如下：

召回率（R）衡量了语义标注的完整程度，它的定义如下：

F值是准确率和召回率的综合指标，它的定义如下：

建立测试集如下，对苏州市内的全部5399161条微博进行随机抽取，取得3029条记录进行人工判读，依据高德底图和文本内容进行推断，共计得到511条定位到景区内的微博。

根据不同比例系数K计算的结果，计算准确率和召回率如下表：

表5.4 在K的不同取值下旅游语义标注方法的各个评价指标值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 比例系数K的取值 | 自动分类正确 | 自动分类非空 | 准确率 | 召回率 | F值 |
| 0.25 | 484 | 673 | 71.92% | 94.72% | 81.76 % |
| 0.24 | 481 | 658 | 73.10% | 94.13% | 82.29 % |
| 0.22 | 476 | 627 | 75.92% | 93.15% | 83.66 % |
| 0.20 | 471 | 593 | 79.43% | 92.17% | 85.33 % |
| 0.18 | 462 | 565 | 81.77% | 90.41% | 85.87 % |
| 0.17 | 456 | 550 | 82.91% | 89.24% | 85.96 % |
| 0.16 | 450 | 544 | 82.72% | 88.06% | 85.31 % |
| 0.15 | 443 | 526 | 84.22% | 86.69% | 85.44 % |
| 0.14 | 437 | 508 | 86.02% | 85.52% | 85.77 % |
| 0.13 | 430 | 495 | 86.87% | 84.15% | 85.49 % |
| 0.12 | 415 | 474 | 87.55% | 81.21% | 84.26 % |
| 0.11 | 403 | 452 | 89.16% | 78.86% | 83.70 % |
| 0.10 | 393 | 433 | 90.76% | 76.91% | 83.26 % |

表5.5 在总体F值最大(k=0.17)时各景区评价指标值

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 景区名称 | 自动分类正确个数 | 人工分类个数 | 自动分类个数 | 召回率(%) | 准确率(%) | F值(%) |
| 金鸡湖 | 91 | 92 | 110 | 98.91 | 82.73 | 90.10 |
| 观前街 | 106 | 113 | 117 | 93.81 | 90.60 | 92.17 |
| 阳澄湖 | 33 | 39 | 43 | 84.62 | 76.74 | 80.49 |
| 平江路 | 31 | 35 | 39 | 88.57 | 79.49 | 83.78 |
| 石路 | 21 | 22 | 29 | 95.45 | 72.41 | 82.35 |
| 木渎 | 18 | 19 | 35 | 94.74 | 51.43 | 66.67 |
| 山塘 | 7 | 7 | 16 | 100.00 | 43.75 | 60.87 |
| 周庄 | 21 | 22 | 21 | 95.45 | 100.00 | 97.67 |
| 同里 | 22 | 23 | 22 | 95.65 | 100.00 | 97.78 |
| 西山 | 10 | 12 | 10 | 83.33 | 100.00 | 90.91 |
| 苏州乐园 | 13 | 13 | 15 | 100.00 | 86.67 | 92.86 |
| 十全街 | 6 | 11 | 8 | 54.55 | 75.00 | 63.16 |
| 拙政园 | 13 | 14 | 13 | 92.86 | 100.00 | 96.30 |
| 东山 | 6 | 8 | 6 | 75.00 | 100.00 | 85.71 |
| 凤凰街 | 7 | 7 | 9 | 100.00 | 77.78 | 87.50 |
| 尚湖虞山 | 4 | 6 | 5 | 66.67 | 80.00 | 72.73 |
| 虎丘山 | 4 | 4 | 5 | 100.00 | 80.00 | 88.89 |
| 光福 | 1 | 1 | 2 | 100.00 | 50.00 | 66.67 |
| 寒山寺 | 3 | 3 | 5 | 100.00 | 60.00 | 75.00 |
| 狮子林 | 6 | 7 | 6 | 85.71 | 100.00 | 92.31 |
| 苏州博物馆 | 6 | 6 | 6 | 100.00 | 100.00 | 100.00 |
| 甪直 | 5 | 8 | 5 | 62.50 | 100.00 | 76.92 |
| 旺山 | 3 | 3 | 3 | 100.00 | 100.00 | 100.00 |
| 沙家浜 | 4 | 4 | 4 | 100.00 | 100.00 | 100.00 |
| 留园 | 1 | 1 | 1 | 100.00 | 100.00 | 100.00 |
| 千灯 | 2 | 2 | 3 | 100.00 | 66.67 | 80.00 |
| 杨舍 | 2 | 3 | 2 | 66.67 | 100.00 | 80.00 |
| 锦溪 | 2 | 3 | 2 | 66.67 | 100.00 | 80.00 |
| 十梓街 | 0 | 5 |  | 0.00 | N/a | N/a |
| 沙溪 | 2 | 3 | 2 | 66.67 | 100.00 | 80.00 |
| 穹窿山 | 2 | 3 | 2 | 66.67 | 100.00 | 80.00 |
| 西园 | 1 | 2 | 1 | 50.00 | 100.00 | 66.67 |
| 三山岛 | 0 |  |  | N/a | N/a | N/a |
| 藏书镇 | 0 |  |  | N/a | N/a | N/a |
| 报恩寺塔 | 1 | 1 | 1 | 100.00 | 100.00 | 100.00 |
| 网师园 | 0 |  |  | N/a | N/a | N/a |
| 盘门 | 1 | 4 | 1 | 25.00 | 100.00 | 40.00 |
| 静思园 | 1 | 3 | 1 | 33.33 | 100.00 | 50.00 |
| 香山 | 0 | 1 |  | 0.00 | N/a | N/a |
| 震泽 | 0 | 1 |  | 0.00 | N/a | N/a |
| 总计 | 456 | 511 | 550 | 89.24 | 82.91 | 85.96 |

对各个景区按开放性进行分类，考察开放景区和封闭景区在旅游语义标记结果上的差异。其中开放景区包括金鸡湖，观前街，阳澄湖，平江路，石路，山塘，十全街，凤凰街，杨舍，十梓街；封闭景区包括木渎，周庄，同里，西山，苏州乐园，拙政园，东山，尚湖虞山，虎丘山，光福，寒山寺，狮子林，苏州博物馆，甪直，旺山，沙家浜，留园，千灯，锦溪，沙溪，穹窿山，西园，三山岛，藏书镇，报恩寺塔，网师园，盘门，静思园，香山，震泽。结果如下表：

表5.6 景区的开放性对旅游语义标记结果的影响(k=0.17)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 景区类型 | 自动标记正确个数 | 人工标记个数 | 自动标记个数 | 召回率 | 准确率 | F值 |
| 开放景区 | 304 | 334 | 373 | 91.02 % | 81.50 % | 86.00 % |
| 封闭景区 | 152 | 177 | 177 | 85.88 % | 85.88 % | 85.88 % |
| 总计 | 456 | 511 | 550 | 89.24 % | 82.91 % | 85.96 % |

5.2.3 结果分析

从旅游语义标记方法得到有类别微博的数量可以看出，随着模型中比例系数k的减小，微博数量也有所减小，这是由于k是与距离阈值σ直接相关的。而σ所表征的是一个数据点的影响半径，σ值越小，则区域的地理范围越小，因此被判断为区域内部的微博数量就有所减少。

从精度上来看，随着k值的减小，旅游语义标记方法的准确度有所提高，召回率下降。这是由于k值的下降，区域地理范围缩小，缩小过程中筛出了大部分错误微博和少量正确微博。其中方法的F值稳定在84%~86%之间，效果良好。当k =0.17时，F值取得最大值，为85.96 %。

产生错误的主要原因，模型计算出的区域范围未能与实际景区地理范围完全重合，主要由以下方面：

(1)模型偏差。由于计算一个点的影响范围是取欧氏距离，影响值随距离增大递减，因此表征在地图上是圆形的，在数十个点叠加起来后，其边界会不同程度地表现为圆弧形状，导致景区区域包含了部分周边无类别的数据点，本应数据景区的数据点被判读为无类别或其他。

(2)样本数据存在噪声。在构建方法模型前通过DBSCAN，去除了部分噪声，但是仍有部分噪声难以去除，噪声聚集在一起，甚至被分为一类，最后算出来一个本不该存在的区域。这种情况存在，但只占极少部分，可以通过人工筛选去除。

(3)样本数据空间分布不均。用户在发微博定位的时候，往往会选择热门区域，这就导致样本聚集在景区区域的一部分内，未能遍布景区各部，因此根据样本计算的结果会使得样本稀疏区域是空白。解决方法是调整阈值τ，但τ调至过低会导致噪声不能有效去除，需要值得注意。

(4)样本数据数量分布不均。一些热门景区，人们的签到数据很多，而与之临近的景区可能数据较少，从公式上可以看出，数据多的区域计算结果范围更大。尽管在公式中已经通过取自然对数削弱这种影响，但不可避免有“大景区”侵蚀“小景区”的现象。可以考虑从距离阈值σ上适当补偿，此部分日后再做研究。

(5)人工判读的误差。对于封闭景区来讲，有明显的边界(围墙，街道)，能够较好地识别景区区域。对于开放景区来讲，无明显边界，在人工判读上就会出现偏差，这部分错误是无法估计的，但从结果来看模型表现良好。

从表5.5中可以看出，封闭景区准确率和召回率相当，开放景区召回率高于平均，平均高于准确率。封闭景区的错误主要来源于模型的偏差，开放景区的错误则可能来源于人工判读的误差。

5.3旅游行为

5.3.1 旅游行为识别结果

根据3.3.2方法，取阈值T=3天，对苏州中尺度旅游时空行为进行提取，得到共计2073723条用户行为，同样，k的不同取值也影响着旅游行为识别的结果，具体数据如下表：

表5.7 在K的不同取值下的旅游行为识别计算结果

|  |  |  |
| --- | --- | --- |
| 比例系数K的取值 | 判定为旅游行为的行为数量 | 占总行为数量的百分比 |
| 0.25 | 653444 | 31.51 % |
| 0.24 | 641155 | 30.92 % |
| 0.22 | 619069 | 29.85 % |
| 0.20 | 596567 | 28.77 % |
| 0.18 | 574693 | 27.71 % |
| 0.17 | 562185 | 27.11 % |
| 0.16 | 550431 | 26.54 % |
| 0.15 | 536739 | 25.88 % |
| 0.14 | 524651 | 25.30 % |
| 0.13 | 509750 | 24.58 % |
| 0.12 | 493949 | 23.82 % |
| 0.11 | 481023 | 23.20 % |
| 0.10 | 467830 | 22.56 % |

5.3.2精度评价

同样，为了评估旅游行为识别的效果，我们也使用了信息检索领域中的三个基本指标准确率（Precision Rate）、召回率（Recall Rate）和F值。

准确率（P）衡量了语义标注的准确程度，它的定义如下：

召回率（R）衡量了语义标注的完整程度，它的定义如下：

F值是准确率和召回率的综合指标，它的定义如下：

建立测试集如下，对苏州全部微博进行行为分割，得到共计2073723条用户行为，随机选取2814条用户行为进行人工判读，依据高德底图和文本内容进行推断，共计得到1155条行为是旅游行为。

根据不同比例系数K计算的结果，计算准确率和召回率如下表：

表5.8 在K的不同取值下旅游行为识别方法的各个评价指标值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 比例系数K的取值 | 自动标注正确 | 自动标注旅游 | 准确率 | 召回率 | F值 |
| 0.25 | 1018 | 1379 | 73.82% | 88.14% | 80.35 % |
| 0.24 | 1017 | 1375 | 73.96% | 88.05% | 80.40 % |
| 0.22 | 1007 | 1351 | 74.54% | 87.19% | 80.37 % |
| 0.20 | 996 | 1321 | 75.40% | 86.23% | 80.45 % |
| 0.18 | 990 | 1291 | 76.68% | 85.71% | 80.95 % |
| 0.17 | 986 | 1269 | 77.70% | 85.37% | 81.35 % |
| 0.16 | 977 | 1251 | 78.10% | 84.59% | 81.21 % |
| 0.15 | 966 | 1233 | 78.35% | 83.64% | 80.90 % |
| 0.14 | 962 | 1221 | 78.79% | 83.29% | 80.98 % |
| 0.13 | 955 | 1207 | 79.12% | 82.68% | 80.86 % |
| 0.12 | 942 | 1184 | 79.56% | 81.56% | 80.55 % |
| 0.11 | 931 | 1167 | 79.78% | 80.61% | 80.19 % |
| 0.10 | 922 | 1148 | 80.31% | 79.83% | 80.07 % |

5.3.3 结果分析

随着k的减小，位于景区内的微博数的减小，同一用户的同一行为中所经过的景区数目不变或减少，若景区数目减小至0，则断定该行为不是旅游行为，因此旅游行为数量也随k的减小而减小。

随着k值的减小，旅游行为的准确率有所上升，召回率有所下降，这是由于k值的下降，区域地理范围缩小，缩小过程中筛出了大部分错误微博和少量正确微博。方法的F值稳定在80%~81%，效果同样良好，在k=0.17时，F值取得最大，为81.35%。

可以看出行为识别的结果较语义标注结果差一些，原因可能有以下几个方面：

(1)所用数据的准确率：场所的语义标注并非完全准确，存在约15%的错误率，这一部分错误数据会导致包含这些场所的行为被判断为旅游行为，因此就导致了行为识别的错误。

(2)识别规则的局限：外地人在旅游过程中，可能没有在景区附近发布微博，或者在两个景区间移动的过程中发布了一条微博，这样两种情况均不会被识别，就造成了行为的遗漏。

尽管如此，从结果上来讲，旅游行为识别还是可以找出大部分的旅游行为，方法结果可信度较高。

5.4 应用

5.4.1 景区中心度及景区等级

根据实验结果的精度评价，选取比例系数k值为0.17，此时判断景区的准确率为82.91%，召回率为89.24%，旅游行为准确率77.70%，召回率85.37%。

对苏州旅游行为数据进行统计分析，可以建立目的地的景区网络。在本文的研究中，我们选择用户中的两部分群体，一是苏州人和上海人，二是非江浙沪三地的外地游客；我们旨在分析不同旅行距离对景区结构的影响。我们分别统计两个景区同时出现在一次旅游行为移动序列中的次数，由此得到苏州40个景区的40阶方阵，据此进行后续研究。

我们采用Ucinet软件对上述景区网络计算中心度，对40个景区划分为3个等级，划分结果如下表：

表5.9 不同客源地下景区等级划分结果

|  |  |  |
| --- | --- | --- |
| 景区等级 | 苏州与上海游客 | 非江浙沪游客 |
| 一级景区 | 观前街，金鸡湖 | 观前街，平江路 |
| 二级景区 | 平江路，石路，阳澄湖，山塘，木渎 | 拙政园，金鸡湖，山塘，周庄，同里，苏州博物馆，狮子林，虎丘山，寒山寺，石路 |
| 三级景区 | 凤凰街，十全街，苏州乐园，西山，光福，拙政园，东山，狮子林，旺山，虎丘山，寒山寺，苏州博物馆，十梓街，留园，同里，甪直，网师园，周庄，穹窿山，尚湖虞山，报恩寺塔，西园，沙家浜，藏书镇，三山岛，千灯，锦溪，盘门，沙溪，杨舍，静思园，震泽，香山 | 留园，十全街，阳澄湖，木渎，苏州乐园，凤凰街，西山，甪直，光福，报恩寺塔，锦溪，网师园，东山，盘门，西园，千灯，十梓街，沙家浜，旺山，静思园，尚湖虞山，穹窿山，三山岛，藏书镇，杨舍，沙溪，香山，震泽 |

可以看出两者结果有明显的区别，尤其是第一二级的景区。苏州上海用户在出行上较多地选择休闲游，集中在商业街，如观前街、金鸡湖、阳澄湖、石路、木渎；而远距离的游客多以观光为主，多去往5A、4A级景区，如平江路、拙政园、狮子林、周庄，同里、寒山寺等。

结果表明，应用旅游行为数据库可以有效地反映目的地城市的景区网络结构，在不同客源地中，结果差别较大。

5.4.2 旅游区提取

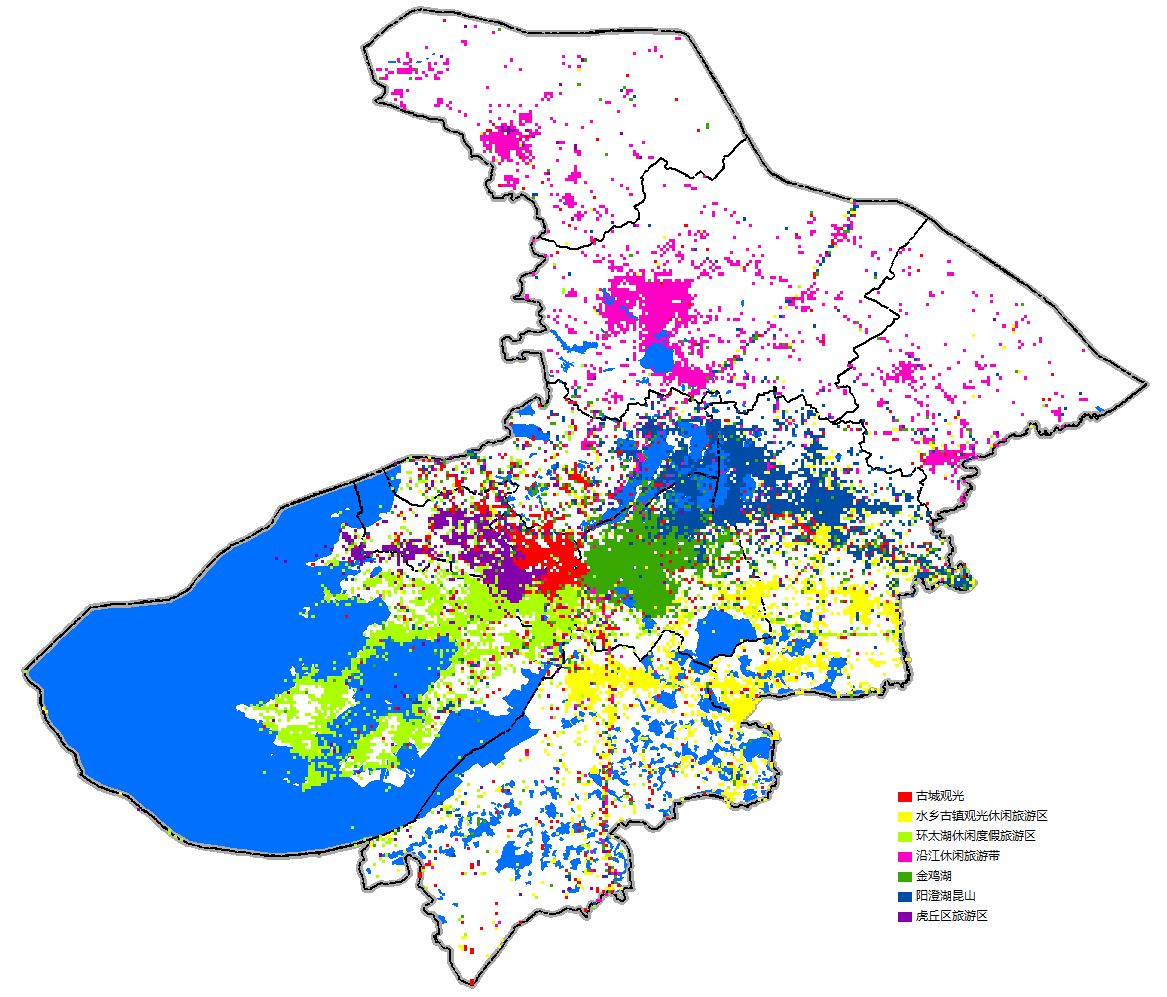
首先，我们采用500m\*500m网格对苏州市进行空间划分，得到了51263个网格。其次，为了对比，我们分别使用了旅游行为和非旅游行为构建了苏州行为联系矩阵。最后，基于算法对苏州市网格进行了划分得到了两个城市空间，如图所示。

图5.3 基于旅游行为计算的城市行为空间

图5.3是基于旅游行为计算的城市行为空间图，图中分为7个旅游区。图5.5是苏州旅游空间规划图，比较一下我们会发现两者是大致相同的，唯有规划图中的中心城区旅游发展级核被分为三个更小的旅游区（虎丘山，古城区，金鸡湖景区）。我们通过数据证明苏州的旅游规划是很有成效的。

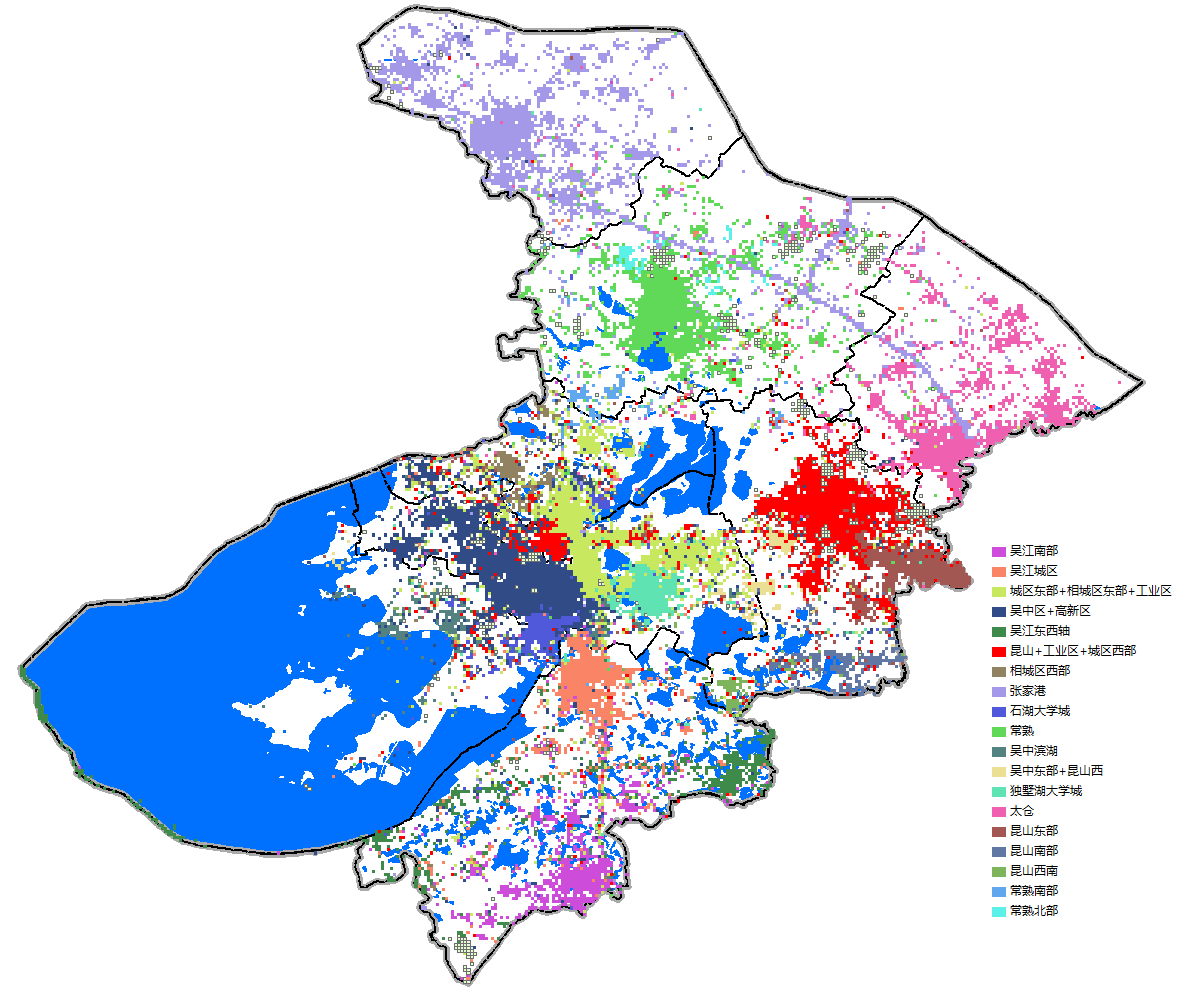


图5.4 基于非旅游行为计算的城市行为空间

图5.4是基于非旅游行为计算的城市行为空间，其中我们选取了外地用户的非旅游行为进行计算。可以看出图中被分为19个较小的区域，对于非旅游行为来说，一般活动范围小，移动距离短。如工作上班，日常生活，聚餐娱乐，往往选择地点较近。从图中所分区域也可以看出，基本符合行政规划范围，同时和旅游行为空间有较大差别，进一步证实了旅游行为与非旅游行为有较大差异。

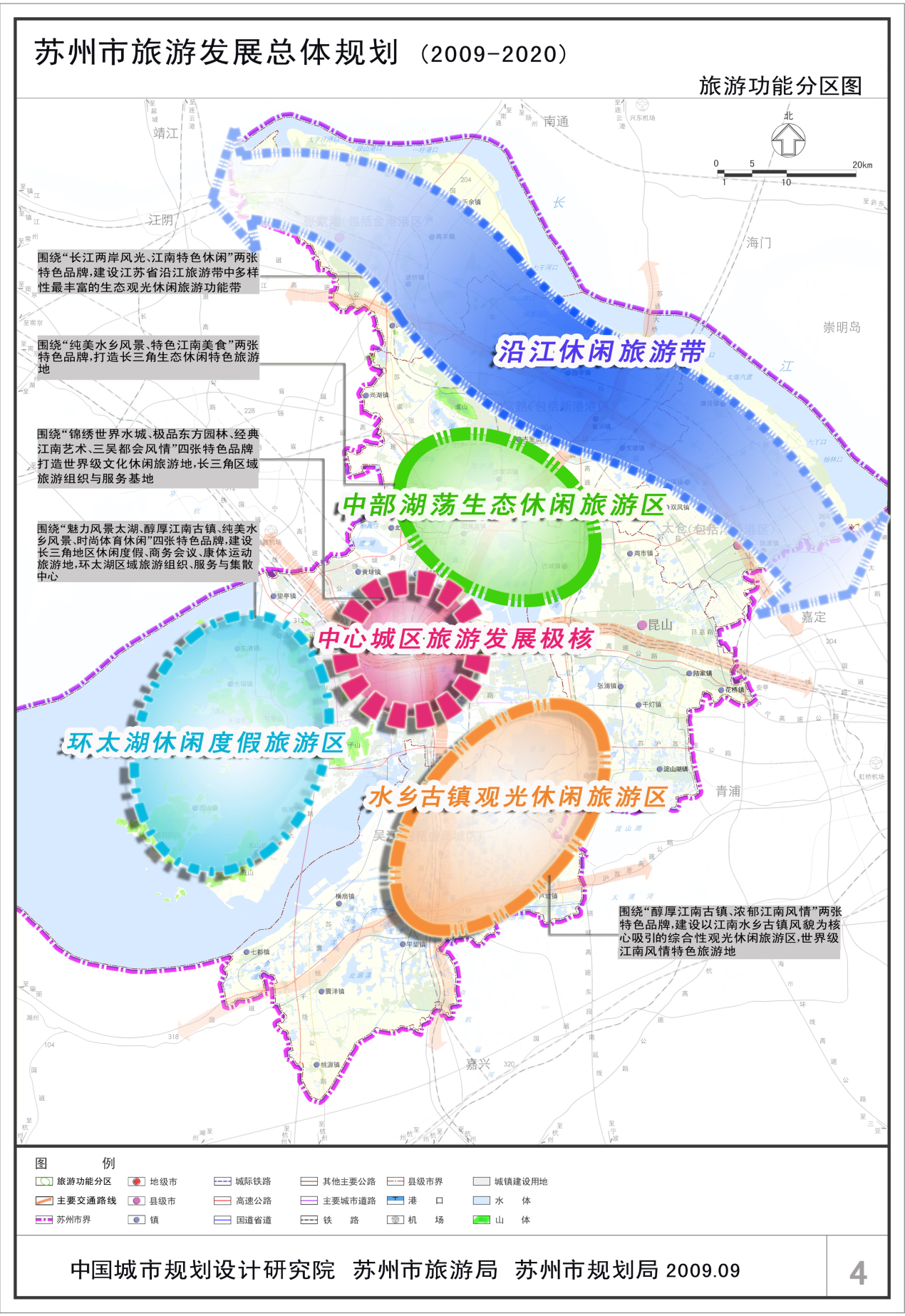


图5.5 苏州旅游空间规划

第六章 结论与展望

6.1 结论

本文主要研究了如何基于社交媒体建立旅游时空行为数据库。本文首先对旅游时空行为模式进行了定义，给出了行为数据库结构；然后对一些关键技术进行了研究，包括用户常居地的提取，旅游语义标注方法，旅游行为识别方法；最后，本文以苏州市为例，实际应用该技术路线，并研究苏州市旅游景区网络与格网行为空间。根据上述研究过程，本文所得研究成果如下：

(1)定义旅游时空行为模式，给出行为数据库结构。本文为后续研究提供了一个良好的研究框架，有助于明确研究目标；研究者可以快速而有效地从社交媒体数据中提取旅游时空行为，并构建数据库。

(2)提出旅游语义标注方法，有效扩展社交媒体可用数据量。根据苏州市的实验结果，本文提出的语义标注方法得到了不错的结果，在F值稳定在85%左右，可以进行推广使用。语义标注方法可以补全用户时空行为的轨迹，填补了非签到数据不能直接使用的空缺，更为准确地研究目的地旅游空间。

(3)旅游行为识别方法细化了用户行为，为旅游研究提供了更高质量更高数量的数据。在以往的一些研究中，研究者没有提取用户的旅游行为，而是将用户的全部行为一并研究，导致其结果与旅游认知偏差较多；而旅游者多是以GPS数据为数据源，样本量较小，不能全面地反映游客的时空行为。从苏州市行为空间上来看，旅游行为识别可有效地反应旅游区域规划的内容，效果良好。

6.2 展望

6.2.1 旅游语义标注方法的改进

在方法的理论模型上，还存在着不足，如聚集类之间的空间关系的分析，如何反映各个聚集类之间的差异，如何体现各类的相对空间位置和距离，对两类中间的区域如何进行判定和标注；距离阈值的确定，如何更为精准地反映各类内的空间关系，如何确定类内数据点的影响范围，使得小的聚集类不至于消失而大的聚集类不至于覆盖其他类。

在本文中，只对模型中的关键参数K进行了讨论，模型中的，，以及均会影响最后的标注结果。的选取同样重要，高斯函数随距离非线性递减，故标注结果也并非随着线性变化。本文研究有限，的取值还待后续研究。

6.2.2 旅游行为识别方法的改进

本文是基于规则对旅游行为进行识别的，不可避免有所缺陷。在后续的研究中，可以结合社交媒体的文本，对用户发布社交媒体的意图进行分类，以剔除因错误标注到过景区而被识别的旅游行为。

不仅如此，后续研究还可以区分旅游类别(观光，度假)。就观光游来说，旅行距离较长，在短时间内较多地游览各个景区，而度假，旅行距离较短，往往在一个景区内进行小范围的活动。在对旅游行为进行提取后，计算旅行距离，停留时间等指标，可以定量判断旅游行为类别。

6.2.3 小尺度旅游行为提取

在进行密度聚类的时候，发现一些景区会产生多余一类的结果，以阳澄湖，金鸡湖，东山，西山最为显著，均在10类以上。考虑这些景区的空间尺度较大，因此景区中景点间的距离也较大。在密度聚类中选取单一固定得搜索半径时，两部分的数据点不能连接到一次，最后聚集为两类。在空间尺度较大的景区中，景点和景点间的距离也相对较大，因此更可能聚集的类代表了一个或若干个小景点的聚集。

考虑在密度聚类时，调整搜索半径，则很可能得到多个聚集类，借助文本内容确定聚集类所代表的真实语义，以此可以进行小尺度旅游行为的提取。

6.2.4 扩展旅游要素

本文所研究的旅游要素，仅仅限定在游玩这单一的要素上。在之后的研究，可以不再局限于景区这一要素，分析游客的吃，玩，住，行等等要素，即在哪吃，去哪玩，在哪住，交通工具。进一步分析游客的游玩体验，而体验可以从文本入手。

可以提取跟“吃”相关的POI，如餐馆，酒店等；提取跟“住”相关的POI，如宾馆，酒店，青旅等；提取跟“行”相关的POI，如火车站，机场，港口等等；若用户在上述POI签到，或按本文方法定位到上述区域，可以视为用户在此处进行了上述行为。

6.2.5 与信令数据结合

手机信令数据可以与微博数据结合，前者仅限时空维，数据量大，后者较散，不够连续，但是有文本信息。两者相结合，信令数据可以弥补微博数据的不连续性，补全用户两条微博间的活动记录。

需要解决的技术问题是，如何将信令数据与微博用户一一对应起来，一种想法如下：可以通过对用户轨迹进行比对，用有时间和定位的微博与信令数据比较，加之用户常居地等进行辅助比较。

在得到了信令与微博结合后的数据，对无文本信息的数据点可用本文方法进行定位，转化为文字语义，用以随后对用户旅游行为的研究。

参考文献

Chareyron G, Da-Rugna J, Raimbault T. Big Data: a new challenge for tourism[C]// IEEE International Conference on Big Data. IEEE, 2014:5-7.

Feick R, Robertson C. A multi-scale approach to exploring urban places in geotagged photographs[J]. Computers Environment & Urban Systems, 2015, 53:96-109.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. Geo-located Twitter as proxy for global mobility patterns[J]. Cartography and Geographic Information Science, 2014.41(3), 260-271.

Huijboom N, Broek T V D, Frissen V. Key areas in the public sector impact of social computing[J]. 2011.

Kuusik A, Tiru M, Ahas R, et al. Innovation in destination marketing: The use of passive mobile positioning for the segmentation of repeat visitors in Estonia[J]. Baltic Journal of Management, 2011, 6(3):378-399.

Liu Y, Kang C, Gao S, et al. Understanding intra-urban trip patterns from taxi trajectory data[J]. Journal of Geographical Systems, 2012, 14(4):463-483.

Luke B., Ickjai L. Spatio-temporal Sequential Pattern Mining for Tourism Sciences[J]. Procedia Computer Science. Volume 29, 2014, Pages 379–389

Modsching M, Kramer R, Hagen K T, et al. Using location-based tracking data to analyze the movements of city tourists.[J]. Information Technology & Tourism, 2008, 10(1):31-42.

Okuyama K, Yanai K. A Travel Planning System Based on Travel Trajectories Extracted from a Large Number of Geotagged Photos on the Web[M]// The Era of Interactive Media. Springer New York, 2013:657-670.

S. Van Canneyt, et al. Time-dependent recommendation of tourist attractions using Flickr[C]. In Proc. of the Belgian/Netherlands Artificial Intelligence Conference (BNAIC), 2011.

Versichele M, Groote L D, Bouuaert M C, et al. Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: a case study of Ghent, Belgium.[J]. Tourism Management, 2014, 44(13):67-81.

Vu H Q, Li G, Law R, et al. Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos[J]. Tourism Management, 2015, 46:222-232.

Zheng Y T, Zha Z J, Chua T S. Mining Travel Patterns from Geotagged Photos[J]. Acm Transactions on Intelligent Systems & Technology, 2012, 3(3):1-18.

方滨兴, 许进, 李建华. 在线社交网络分析[M]. 电子工业出版社, 2014.

谷文林, 任敏. 大数据时代乡村旅游数据挖掘与分析[J]. 江西农业学报, 2015(8):143-146.

李渊. 基于GPS的景区旅游者空间行为分析——以鼓浪屿为例[M].科学出版社,2016.

邵虎. 基于微博数据的游客时空行为特征提取与分析[M].

吴必虎. 旅游规划原理[M]. 中国旅游出版社, 2010.

张子昂, 黄震方, 靳诚,等. 基于微博签到数据的景区旅游活动时空行为特征研究——以南京钟山风景名胜区为例[J]. 地理与地理信息科学, 2015, 31(4):121-126.

致谢

转眼之间，我在北京大学已经度过四载春秋。在这四年中，我在北京大学中结识了更多的老师同学朋友，也见识了更为广阔的世界。从大一的浮躁，到大二的懵懂，再到大三的醒悟，最后到大四的沉稳，这段时光里我也在不断成长。谨以此文，感谢曾经帮助我，支持我，关系我的老师同学朋友以及亲人们，感谢大学期间遇到的所有人，正因有你们我的本科时光才如此丰富多彩。

首先，我要感谢我的导师张毅老师。张老师不仅学识渊博，治学严谨，而且心胸宽广，平易近人。从本科生科研到毕业设计的过程中，我常常有很多疑问和困惑，张老师都一一且耐心地为我解答。无论从学术方向，研究方法，生活态度上，张老师都给予了我细致入微的教诲与指导。每每与张老师探讨学术研究时，总是感到亲切轻松的。张老师从为人为师上都令我受益匪浅。

感谢实验室的邬伦老师，刘瑜老师，高勇老师，田原老师等人，感谢各位老师在平日里对我的指导与帮助。感谢实验室的师兄师姐，在学术和生活上对我的关心和照顾。感谢同组张瑞洁师姐，王雯夫师兄，陈子豪师兄在研究上给予的帮助，使我更快地进步。

感谢13级GIS班的各位同学，在四年中我们一同上课，一同写大作业，一同整理复习资料，互帮互助，相亲相爱如一家人，共同努力共同进步。感谢我的室友刘政，从棒垒、数双、到学生会乃至科研等等，可以说是在大学期间交集最多的朋友，无论从生活上还是学术上都对我有很大的帮助。感谢我的室友李嘉政，让我感受到了一个学霸的魄力，激励我努力奋斗地前行。感谢室友刘宝剑用他不懈的刻苦练习促使我早出晚归地自习提升自我。感谢地空学院和其他兄弟院校的同学，大学时光因他们而精彩。

感谢我的父母对我多年来的养育之恩，即便远隔千里，仍然对我情绪的波动了若指掌，在我苦闷的时候，及时地开导我，安慰我，给我无微不至的关怀。常常叮嘱我吃早饭多运动多锻炼，教给我做人的道理和为人处世的方式。感谢你们对我的关爱和付出，愿你们永远健康快乐。

感谢我的女友秦琳，大二那年的男生节，你我相识的那一刻我永远不会忘。你与我一路走来，在困难时相互扶持，在快乐时分享喜悦，感谢你容忍我不定期的坏脾气，持续性的拖延症，给予我鼓励和信心。因为有你我愿加倍努力，完善自我，为我们的未来而奋斗。相信我，我们会有猫的。

感谢我的基友们，尤其是刘笑宇和芦熙文。每当感到生活不称意的时候，总是相约畅谈人生感悟和理想，或是在混沌要塞抵抗堕落天使，或是时空枢纽迎战黑心船长，或是在星际间运筹帷幄，或是在伊里奥斯拯救世界，亦或是简简单单搓几局麻将，顿觉豁然开朗。

论文至此，意味着我也即将结束我的本科生生涯，愿我能在今后的生活中，不忘初心，奋勇前行。

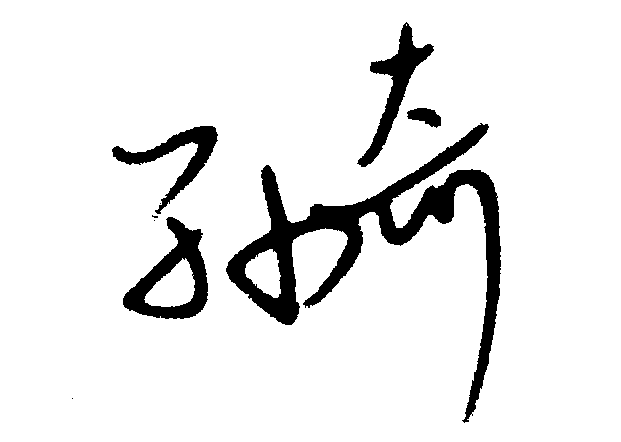
孙奇

2017年6月于北京大学

北京大学学位论文原创性声明和使用授权说明

**原创性声明**

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。



论文作者签名： 日期： 2017 年 6 月 3 日

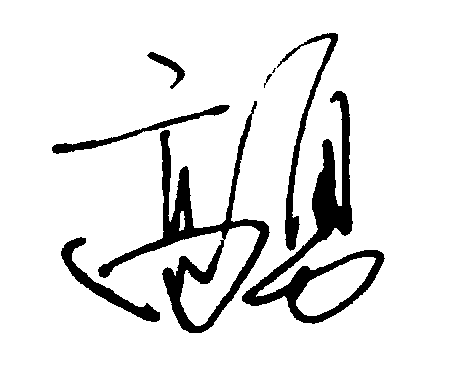
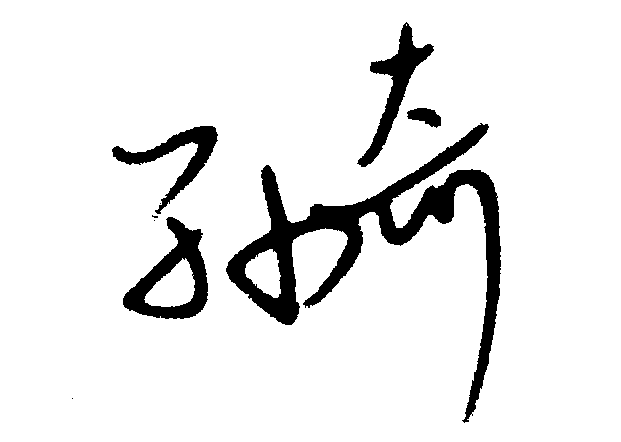
**学位论文使用授权说明**

(必须装订在提交学校图书馆的印刷本)

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
* 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
* 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

(保密论文在解密后遵守此规定)



论文作者签名： 导师签名：

日期： 2017年 6 月 3 日