

武汉大学学报(信息科学版)

Geomatics and Information Science of Wuhan University

ISSN 1671-8860, CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目: 利用网络游记分析不同类型游客的旅游流网络特征差异——以云南省为例
作者: 李秋萍, 陈宇, 栾学晨
收稿日期: 2021-09-01
网络首发日期: 2021-09-16
引用格式: 李秋萍, 陈宇, 栾学晨. 利用网络游记分析不同类型游客的旅游流网络特征差异——以云南省为例. 武汉大学学报(信息科学版).
<https://kns.cnki.net/kcms/detail/42.1676.TN.20210915.1006.005.html>



网络首发: 在编辑部工作流程中, 稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定, 且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式(包括网络呈现版式)排版后的稿件, 可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定; 学术研究成果具有创新性、科学性和先进性, 符合编辑部对刊文的录用要求, 不存在学术不端行为及其他侵权行为; 稿件内容应基本符合国家有关书刊编辑、出版的技术标准, 正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性, 录用定稿一经发布, 不得修改论文题目、作者、机构名称和学术内容, 只可基于编辑规范进行少量文字的修改。

出版确认: 纸质期刊编辑部通过与《中国学术期刊(光盘版)》电子杂志社有限公司签约, 在《中国学术期刊(网络版)》出版传播平台上创办与纸质期刊内容一致的网络版, 以单篇或整期出版形式, 在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊(网络版)》是国家新闻出版广电总局批准的网络连续型出版物(ISSN 2096-4188, CN 11-6037/Z), 所以签约期刊的网络版上网络首发论文视为正式出版。

引用格式：

李秋萍,陈宇,栾学晨. 利用网络游记分析不同类型游客的旅游流网络特征差异——以云南省为例[J].武汉大学学报·信息科学版, 2021, DOI: 10.13203/j.whugis20210045 (LI Qiuping, CHEN Yu, LUAN Xuechen, . Tourism Flow Network Structures of Different Types of tourists Using Online Travel Notes: A Case study of Yunnan Province [J].Geomatics and Information Science of Wuhan University,2021, DOI: 10.13203/j.whugis20210045)

利用网络游记分析不同类型游客的旅游流网络特征
差异——以云南省为例

李秋萍^{1,4} 陈宇² 栾学晨^{3,4}

1 中山大学地理科学与规划学院, 广东 广州, 510275

2 武汉大学遥感信息工程学院, 湖北 武汉 430079

3 广东方纬科技有限公司, 广东 广州 510275

4 广东省公共安全与灾害工程技术研究中心 广东 广州, 510275

摘 要：旅游流网络分析对理解游客的目的地选择以及目的地在旅游流网络中承担的角色有重要意义。以往有关旅游流网络结构特征的研究大多将所有游客当作一个整体，较少关注不同类型游客的旅游流网络特征差异。本文以云南省为例，基于网络游记数据挖掘游客的多维度偏好，并以此对游客聚类，进而划分出不同类型的游客群体。针对各类游客游记中的旅游目的地序列建立旅游流网络，并从多个角度分析各类游客旅游流网络的结构特征和各目的地节点的角色特征。结果表明：不同类别游客的旅游流网络在整体结构上各有特点，反映出旅游目的地不同的空间交互模式和网络中心化程度。此外，部分旅游目的地在不同类别旅游流网络中承担截然相反的角色。本文的分析有助于优化旅游流网络中各节点的协作机制、辅助旅游目的地制定差异化的旅游产品。

关键词：旅游流网络；文本挖掘；网络分析；游客偏好；游记数据

中图分类号：P208

文献标识码：A

旅游流是指游客在旅游空间场内的迁移^[1]，反映游客从客源地向目的地流动的人员数量和流动模式^[2]，也代表着旅游目的地与客源地以及目的地内不同旅游节点间的空间交互。对旅游流构成的空间交互网络的分析有助于从宏观上把握游客在旅游目的地网络内集聚、扩散的空间特征，对辅助旅

游景区优化旅游产品、合理配置旅游资源具有积极作用。

随着游客网络游记数据、微博签到数据、手机定位数据等获取越来越容易，基于位置大数据的旅游流网络分析成为旅游流研究的热点^[3-8]。目前大部分相关研究是通过社会网络分析方法来探究旅游流网络的各类结构特征，如网络密度^[9]、核心-边缘特征^[10]、

收稿日期：2021-09-01

项目资助：国家自然科学基金（41971345）；广东省自然科学基金（2020A1515010695）

第一作者：李秋萍，博士，副教授，主要从事交通地理的理论与方法研究。liqp3@mail.sysu.edu.cn。

通讯作者：栾学晨，博士。luanxuechen@fundway.net。

旅游节点的中心性^{[3]、[6]}等,反映旅游流的空间分布模式及不同旅游节点在网络中的重要性。例如,网络密度代表了旅游网络中所有节点的联系程度。中心性指标高的旅游节点往往是旅游流网络中的核心节点,承担着集聚与扩散旅游流的作用。Lueng 等 (2012) 利用 500 位北京入境游客的网络游记数据对比分析了 2008 年北京奥运会前后旅游流网络结构的变化,发现入境游客流网络逐渐从单中心转变为多中心的结构^[11]。闫闪闪等 (2019) 利用社交媒体数据构建洛阳市区各景点间的旅游流网络,探究了洛阳市区旅游流网络中的核心-边缘特点以及旅游流节点明显分层聚集的态势^[12]。Mou 等 (2020) 根据网络游记数据发现青岛市的旅游流空间模式结构松散,核心节点分布不均衡,形成内部竞争激烈的网络结构模式^[6]。上述研究重点关注整个旅游流网络的整体结构特征。然而,由于需求的多样性(如交通方式、景点选择等),不同游客的旅游流网络呈现出相异的特点。例如,与一般游客相比,自驾游客的旅游流网络呈现出更加分散的空间分布形态^[13]。由于旅游景点选择上的偏好,不同客源地游客的旅游流网络也具有各自的形态和结构特点^[7]。对这些差异的分析和探讨能够帮助旅游目的地更好细分市场、进行精细化的旅游管理和运营。然而文献[7]和文献[13]目前都是从单一维度对游客进行划分,只能反映游客某一方面的需求偏好。此外,对各旅游节点在不同类型游客旅游流网络中的重要性和角色差异也缺乏多角度的量化分析。

鉴于此,本文以云南省为例,基于在线旅游网站中的真实旅游路线和游记文本数据,挖掘游客的多维度偏好并利用 K-prototype 算法对游客进行聚类,进而划分出不同类型的游客群体。针对各类游客游记中记录的游玩景点序列建立旅游流网络,量化分析各类游客旅游流网络的网络整体特

征和网络中各景点的起终点特征、中转点特征和中间人角色特征差异,从而为进一步理解旅游流网络的形成机制、辅助旅游景区精细化管理和制定差异化的旅游产品提供理论指导。

1 研究区与数据

1.1 研究区概况

云南省位于中国西南方,省域总面积 39.41 万平方千米,拥有丰富的自然资源和人文景观,如:丽江古城、云南石林等许多世界遗产,还有丰富的特色文化资源,如:滇剧、沧源崖画等^[14]。根据云南省统计局发布的 2018 年国民经济和社会发展统计公报^[15],该年云南省共接待旅客约 6.88 亿人次,旅游业收入占该地区生产总值的 48.6%。

1.2 数据来源与预处理

许多游客通过在线旅游网络平台分享其在旅途过程中的游记和拍摄的照片。其中游记文本数据可以反映游客的兴趣内容、出行特点等,因此可以用于分析游客的旅行偏好特征。本文爬取了携程 (www.ctrip.com)、去哪儿 (www.qunar.com)、马蜂窝 (www.mafengwo.cn) 三大旅游网站云南省的旅游信息数据,包括旅游目的地信息、2019 年全年的游记数据和对应的作者信息。此外,还爬取了这些作者发布的历史游记文本。对三个网站获取的旅游目的地信息数据进行清洗,合并重复目的地信息,相互补充缺省的字段,最终得到 2255 条目的地数据。整合后的目的地信息主要包括旅游目的地的名称、经度、纬度、所属行政区、推荐游玩时间(小时)以及开放时间。

对三个网站获取的游记文本进行整合,共得到 44418 条游记数据以及 15267 条有效游记路径。处理后的游记数据主要包括游客的旅行日期、旅行天数、旅行同伴类型、花

费、路线及游记正文内容。

2 游客特征提取及聚类方法

2.1 基于网络游记的游客兴趣偏好提取

本文利用词袋模型 (bag of words model)^[16]从游客历史发表游记中提取各游客的词袋向量,进而提取其兴趣偏好。首先对游客的游记文本进行分词,获得该文本的词向量。例如,一条游记文本是“我在大理拍了很多照片”,则该文本的词袋为{“我”,“在”,“大理”,“拍”,“了”,“很多”,“照片”}。接着,利用文本挖掘中常用的 TF-IDF (Term frequency-Inverse Document Frequency)算法^[17]评估游客游记文本中各单词的重要性,并以此构建带权重的词袋向量。

在 TF-IDF 算法中,词的重要性与它在文档中出现的次数成正比,与它在语料库中出现的频率成反比。游客 j 的游记词袋中词 i 的重要性用 $Sum_{TF-IDF}(j, i)$ 表示,其计算过程如下:

$$Sum_{TF-IDF}(j, i) = \sum_k^{N_j} TF - IDF_j(i, k) \quad (1)$$

$$TF - IDF_j(i, k) = TF_j(i, k) \times IDF_j(i) \quad (2)$$

$$TF_j(i, k) = \frac{word_freq_j(i, k)}{word_count_j(k)} \quad (3)$$

$$IDF_j(i) = \log \frac{N_j}{d_count_j(i)} \quad (4)$$

TF(Term frequency) 为词频, $word_freq_j(i, k)$ 表示词 i 在游客 j 第 k 篇游记中出现频次, $word_count_j(k)$ 表示游客 j 文档 k 中的总词数。IDF(Inverse Document Frequency) 为逆文档频率, N_j 表示游客 j 历史游记的总数, $d_count_j(i)$ 表示游客 j 历史游记中包含词 i 的游记数量。 $TF - IDF_j(i, k)$ 为词频和逆文档频率的乘积, $Sum_{TF-IDF}(j, i)$ 为词 i 在游客 j 所有游记文档中 TF-IDF 值的和,其值越大,表示游客在该特征上的兴趣偏好越强。

2.2 基于 K-prototype 聚类算法的游客聚类

由于游记中的词袋向量具有高维稀疏的特点,本文利用 t-SNE(t-Distributed Stochastic Neighbor Embedding)^[18]方法对 TF-IDF 方法生成的高维稀疏词袋向量进行降维,以降维后的词向量分布作为游客的兴趣特征。此外,游客还有一些实际的旅行出行特征,包括:旅伴类型、游玩天数、旅行月份及旅行花费。各聚类特征的维度和取值如表 1 所示。由于这些聚类特征包括离散型特征和连续型特征,本文利用适合于对混合特征进行聚类的 K-prototype 算法^[19]进行游客聚类,并利用轮廓系数 (silhouette coefficient)^[20-21]指标来确定最佳聚类个数。轮廓系数越接近 1 表明聚类效果越好。

表 1 游客聚类特征

Tab. 1 Features for Tourist Clustering

特征	维度	取值
游客兴趣	2	词向量分布
旅伴类型	1	夫妻、和父母、和朋友、 家庭、亲子、情侣、一 个人
游玩天数	1	实数域
旅行月份	1	1 月~12 月
旅行花费	1	实数域

3 游客目的地网络特征分析方法

将游记文本数据集中的旅游路线数据拆分为前驱目的地-后继目的地等若干条短路径,即表示游客从某一旅游目的地向另一目的地转移。以目的地为网络节点,目的地之间转移的路线方向、频次作为节点的有向、加权连接边,据此构建旅游目的地有向网络 $G(N, A, W)$, N 为节点(即旅游目的地)集合, A 为边集合, W 为边的权重。

利用社会网络分析方法^[22-23]中的网络密度、出(入)度中心势和中介中心势等指标分析各类游客目的地网络的结构特征。此

外,还进一步在节点层面分析网络中各个节点的角色特征,包括:起、终点特征、中转点特征、中间人角色特征。

3.1 目的地网络的结构特征

1) 网络密度

网络密度可用于度量网络边的密集程度和网络结构的均衡性,一个具有 N 个节点和 L 条边的网络,其网络密度 D 可计算为:

$$D = \frac{2L}{N(N-1)} \quad (5)$$

2) 度中心势

度中心势可分为出度中心势 $C_{RD,out}$ (或外向程度中心势) 和入度中心势 $C_{RD,in}$ (或内向程度中心势), 可用于度量网络结构的均衡性。 $C_{RD,out}$ 、 $C_{RD,in}$ 可分别表达为:

$$C_{RD,out} = \frac{\sum_{i=1}^N (C_{Dmax,out} - C_{D,out}(i))}{N-2} \quad (6)$$

$$C_{RD,in} = \frac{\sum_{i=1}^N (C_{Dmax,in} - C_{D,in}(i))}{N-2} \quad (7)$$

其中 $C_{Dmax,out(in)}$ 分别为网络中最大的相对出、入度值, $C_{D,out}(i)$ 为节点 i 的出度, $C_{D,in}(i)$ 为节点 i 的入度。

3) 中介中心势

中介中心势可以反映网络中核心节点的主导地位以及判断网络是否具有核心-边缘结构特征。中介中心势越大,表明网络中的核心节点在网络中处于主导地位,较多节点在网络中需要通过核心节点产生连接,该网络具备明显的核心-边缘结构特征。中介中心势 C_{RB} 可表达为:

$$C_{RB} = \frac{\sum_{i=1}^N (C_{Bmax} - C_B(i))}{N-1} \quad (8)$$

$$C_B(i) = \sum_{j,k}^N \frac{g_{jk}(i)}{g_{jk}} \quad (j \neq k \neq i) \quad (9)$$

其中 C_{Bmax} 分别为网络中最大的相对中介中心性指数, $C_B(i)$ 为节点 i 的中介中心性, $g_{jk}(i)$ 为节点 j 和节点 k 之间经过节点 i 的最短路径数, g_{jk} 为节点 j 和节点 k 之间的最短路径总数。

3.2 网络节点的角色特征

1) 节点的起、终点特征

利用目的地的出、入度差值与入度之比来表征目的地的起终点特征。若目的地起终点特征较大,则其可能常常作为旅行路线的起点,反之则更有可能作为终点出现在旅行路线中。节点 i 起终点特征值 $F(i)$ 计算如下:

$$F(i) = \frac{C_{D,out}(i) - C_{D,in}(i)}{C_{D,in}(i)} \quad (10)$$

2) 节点的中转特征

节点的中转能力可以由中介中心性指标反映。中介中心性越大,节点的作为中转点的功能就越明显。在不同网络中,相同节点 i 的中转能力大小可采用 i 的相对中介中心性指标来度量和比较。相对中介中心性 $RC_B(i)$ 表示为: 在一个网络中, 节点 i 的中介中心性 $C_B(i)$, 占与该网络具有相同节点数的全连接网络中节点 i 中介中心性的比例。

$$RC_B(i) = \frac{C_B(i)}{(N-1) \times (N-2)} \quad (11)$$

3) 节点的中间人角色特征

除了度量节点的中转能力,本文还利用中间人分析方法分析游客在各景点组团间转移的过程中所扮演的中间人角色。首先,将旅游目的地网络中的各节点进行组团划分。以游客在目的地之间转移的频次和目的地之间的距离作为聚类特征,利用近邻传播聚类算法(Affinity Propagation Clustering, AP 聚类)^[24]进行节点聚类,并构建景点组团。

一个节点的中间人角色可以分为五大类,如下图 1 所示。虚线圈内的节点为同一组团的节点,中间人在组团内、外联系中扮演的角色包括“协调员”、“顾问”、“守门人”、“代理人”、“联络人”五类。

“协调员”(图 1(a)): 中间人及其联系的节点均在同一组团,该中间人扮演“协调员”的角色,表明游客在同一个景点组团内游玩。

“代理人”(图 1(b)): 若中间人节点与其前驱节点在同一个组团,而其后继节点在

其他组团，该中间节点为“代理人”，这表明游客在前一个景点组团连续游玩至少2个景点后转移到另一个景点组团中。

“守门人”（图 1(c)）：若中间人及其后继节点在同一景点组团，而其前驱节点在其他组团，该中间人扮演“守门人”角色。在旅游流网络中，表明游客转移到另一个景点组团中，且下一旅游目的地仍然在该景点子群中。

“顾问”（图 1(d)）：依靠中间人联系的两个节点在同一组团，但不与该中间人在同一组团，那么该中间人为“顾问”，表示游客从一个景点组团到中间人景点游玩，之后又回到之前的景点组团。

“联络人”（图 1(e)）：若三个节点均位于不同组团，那么中间节点为“联络人”，表明游客在不同景点组团内短暂游玩并转移。

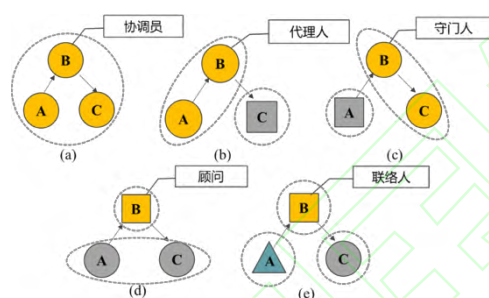


图 1 中间人角色示意图

Fig.1 Example of Five Intermediary Roles

4 结果分析

4.1 游客聚类结果

基于 2.2 节中的 K-prototype 算法对游客进行聚类。该算法中主要的参数为聚类个数 k 。由于 K-prototype 算法具有随机性，本文对每个输入的聚类参数 k 进行 20 次聚类。在每个 k 值下会得到 20 个轮廓系数值。图 2 显示了聚类个数 k 取不同值时的轮廓系数分布。从图中可以看出，当 $k=5$ 时，轮廓系数值接近最大，且分布集中。因此，本文将游客最佳聚类个数设置为 5。

对聚类结果进行 ANOSIM (Analysis of similarities) [25] 组间差异显著性检验。

ANOSIM 是一种用于分析高维度数据组间相似性的非参数检验方法，组间与组内比较的差异程度用统计量 R 表示。若 $R>0$ 说明组间差异大于组内差异，显著性 P 值越低表明差异检验结果越显著。经过 999 次置换检验，聚类后各类游客的差异显著性分析得到 $R=0.4375$, $P=0.001$ ，表明不同类别游客之间存在显著差异。

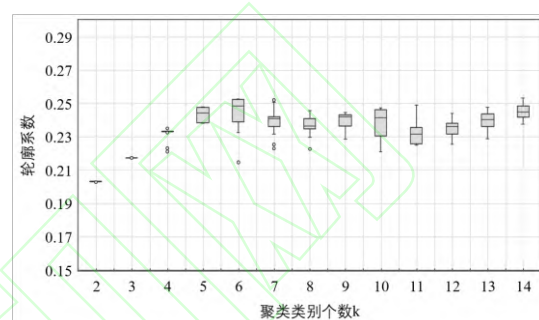


图 2 不同聚类类别个数 k 下聚类结果轮廓系数

Fig.2 The Values of Silhouette Coefficient Under Different Number of Clustering Categories (K)

图 3 展示了 5 类游客簇在旅行特征和兴趣偏好上的差异。在雷达图中对游玩天数和旅行花费这两个连续型的旅行特征进行分段统计，例如游玩天数分为： ≤ 3 天、4-9 天、9-14 天、 ≥ 15 天。此外，由于经过 t-SNE 方法降维后的游客兴趣词袋向量依然较长，为了更直观的展示各类游客兴趣特征的差异，将每类游客兴趣词袋向量按照 TF-IDF 均值倒序排序。然后，取词袋向量中的前 300 个关键词加入核心关键词集合，并将核心关键词划分为“景区特点”、“路途交通”、“时间”、“消费”、“主观感受”和“住宿”六大主题。最后，用每个主题下各关键词的 TF-IDF 值总和来表示游客对该主题的关注度。

类别 1（图 3(a)）有 81% 的游客选择一个人或和朋友一起旅行，旅行天数主要为 4~9 天。有超过四分之一的游客在 3 月份出行，旅行花费小于 3000 元的游客超过 60%。可以看出该类别的游客在旅行搭档、花费、

以及日期方面比较敏感。此外，该类别中的游客游记中倾向于对景区特点进行描述和讨论，除此之外还喜欢表达对景点的主观感受。类别1中的游客数约占样本总数的26%。

类别2（图3(b)）的游客除了与朋友出行外，和父母或亲子出游的比例也较高(约占37%)，旅行天数较短（旅行天数 ≤ 3 天）的游客占比达到78%。大部分游客的旅行花费低于3000元，且这类游客的旅行日期往往在节假日（如五一劳动节、春节、寒暑假等），具有旅游日期敏感的特点。游记中讨论的内容集中在景点特点、主观感受和景点的消费。类别2中的游客数约占样本总数的23%。

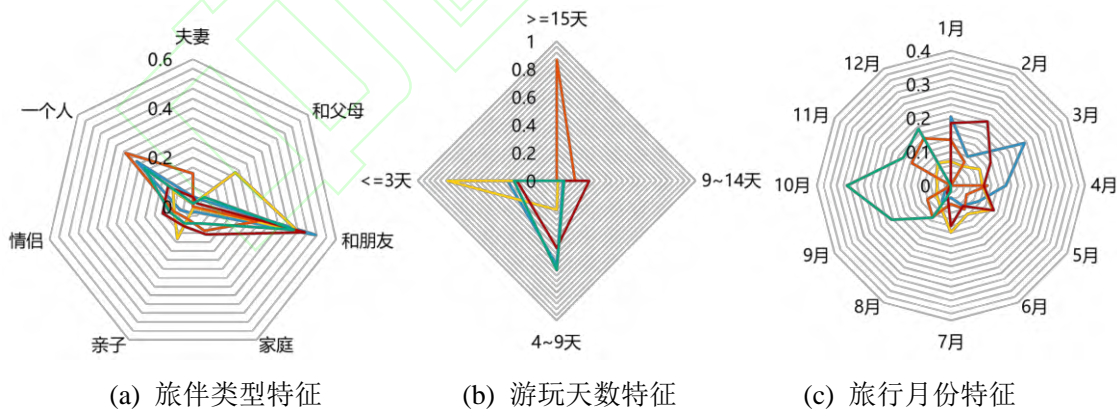
类别3（图3(c)）内约87%的游客旅行天数超过15天，出行日期往往集中在11月份~1月份以及8、9月份(分别占约43%，19%)，接近50%的游客旅行花费在3000-8000元，消费相对较高。游客游记更加关注路途交通、景点消费和住宿相关的内容。类别3中的游客数约占样本总数的4%。

类别4（图3(d)）中接近50%的游客旅

行天数大部分为4~9天。这类游客的旅行日期集中在8月之前。与类别3相似，游客中除了表达对景点特点的关注外，对路途交通也关注较多。类别4中的游客数约占样本总数的21%。

类别5（图3(e)）中旅行时间在4~9天的游客占比约64%，出行集中在8月之后(接近100%)，以10月为主（国庆假期）。与类别1相似，这一类游客的游记中对景点特点和个人主观感受讨论较多，倾向于选择更符合自身兴趣的景点作为旅行目的地。类别5中的游客数约占样本总数的26%。

总体来看，可以将游客类型分为旅行花费敏感型（类别1）、旅行天数敏感型（类别2）、长时间深度游型（类别3）、旅行月份敏感型（类别1、4、5）、景点特点敏感型（类别2、类别5）。



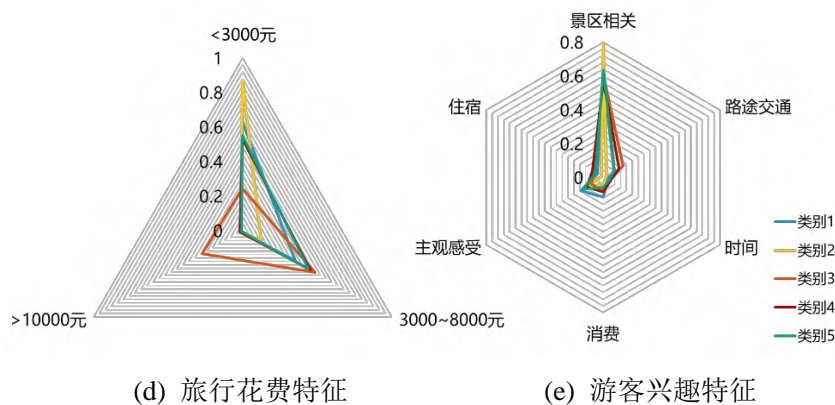


图 3 各类别游客的旅行偏好

Fig.3 Travel Preferences of Different Tourist Clusters

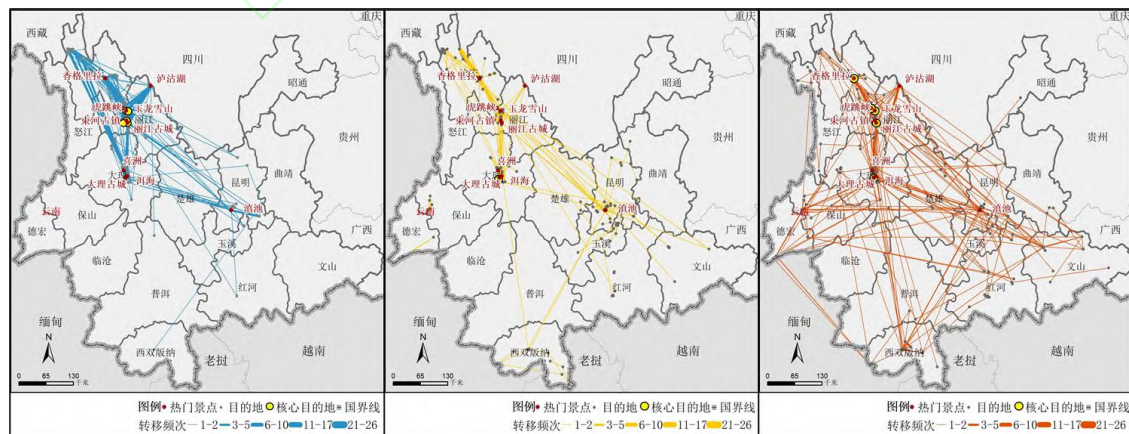
4.2 各类游客旅游流子网络特征分析

基于不同类别游客的旅游路线分别构建旅游流子网络，并对各个子网络的结构特征以及子网络中各节点的起、终点特征、中转能力及中间人角色特征进行分析。

4.2.1 子网络结构特征

不同类型游客的旅游流网络分别如下图所示 4 所示。可以看出，类别 1 游客（旅游花费敏感型）超过 50% 的旅行路线位于丽江市，旅游路线的跨度较小。类别 2 游客（旅行天数、景点特点敏感型）由于主要是节假日的短期旅行，旅行线路也以短线为主，因此路线的空间分布较为集中，大量的旅游流集中在大理和丽江两个城市之间。类别 1 和类别 2 游客的旅游流网络呈现出以丽江为核心的“单核线网状”的空间分布形态。研究区东南

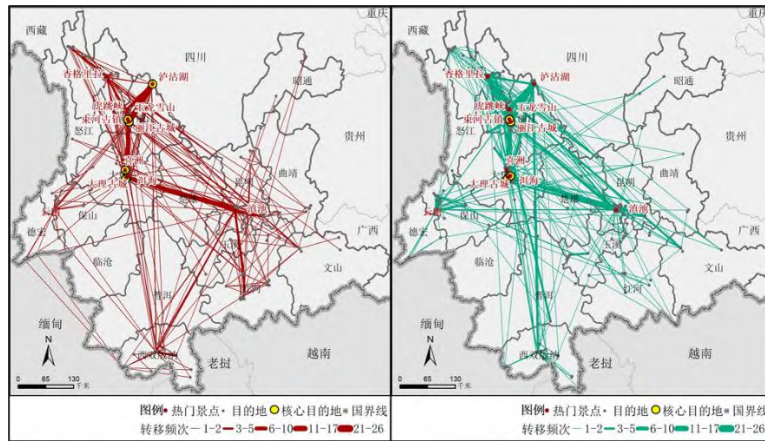
部的一系列景点在未来对类型 1 和类型 2 的游客有巨大的挖掘潜力。旅游和交通相关部门可以通过加大宣传、规划交通线路来引导这两种类型的游客（人数占到总样本的近 50%）的旅游流向，促进东南部地区各旅游节点的进一步发展。类别 3（长时间深度游型）旅行目的地更多样，旅行路线的空间上分布范围较类别 1、2 要大得多。由于有更加充足的游玩时间，这类游客也会选择距离跨度更大的旅游目的地，如西双版纳。类别 4（旅行月份敏感型）的旅游路线跨度相对较大，在大理古城和滇池间形成了很强的旅游流交互。类别 5（旅行月份、景点特点敏感型）路线中的旅游目的地在 5 类游客子网络中最丰富，而且游客长距离旅行也相对较多。类别 3，4，5 游客的旅游流网络呈现出多核心特点。



(a)类别 1

(b)类别 2

(c)类别 3



(d)类别 4

(e)类别 5

图 4 各类别游客路线空间分布

Fig.4 The Spatial Distribution of Travel Paths of Five Clusters of Tourists

各类别游客的子网络的结构特征如表 2 所示。在 5 个子网络中，类别 1 游客子网络的网络密度最高，因此节点联接程度最高，结构相对最为紧凑。类别 1、3、4 子网络的出、入度中心势、中介中心势相对较高，说明这些类别游客的旅游目的地集中在一些

核心景点，中心化程度较高，子网络的核心-边缘特征明显。类别 2、5 子网络的出、入度中心势小，且中介中心势较低，表明类别 2、5 子网络中各目的地对核心节点的依赖较低。

表 2 各类别子网络整体结构特征

Tab.2 Tourism Network Structure of Five Clusters of Tourists

子网络	节点数	路径数	网络密度	入度中心势	出度中心势	中介中心势
类别 1	171	997	0.0686	1.81%	2.37%	20.69%
类别 2	238	555	0.0197	0.70%	0.92%	17.44%
类别 3	232	447	0.0167	1.53%	1.53%	22.96%
类别 4	285	989	0.0244	1.47%	1.54%	22.81%
类别 5	336	1358	0.0241	0.78%	0.96%	15.00%

4.2.2 子网络中节点的起、终点特征

图比较了 5 个子网络在研究区域内 9 个热门景点（丽江古城、洱海、大理古城、香格里拉、泸沽湖、玉龙雪山、束河古镇、滇池、喜洲、虎跳峡）的起、终点特征。整体上，5 个类别的游客都倾向于将滇池、香格里拉等目的地作为旅行线路的起点，而虎跳峡在 5 个类别子网络中该特征值均为负，说明虎跳峡通常作为旅行线路的终点。类别 1 子网络中丽江古城、大理古城的起、终点特征值相比在其他子网络中大，表明类别 1 的

游客相比其他类型游客更倾向于选择它们作为旅行线路的起点。滇池的起、终点特征值在类别 2 中最大，该类别的游客常选择滇池作为旅行起点。部分节点在不同类别游客的旅游流网络中承担了截然相反的角色。比如类别 3 子网络中泸沽湖、喜洲的起、终点特征值为正，而在类别 2 和类别 5 子网络中该值为负，表明类别 3 的游客更倾向于选择这两个景点作为旅行线路的起点，类别 2 和类别 5 的游客则更倾向于将其作为旅行线路的终点。

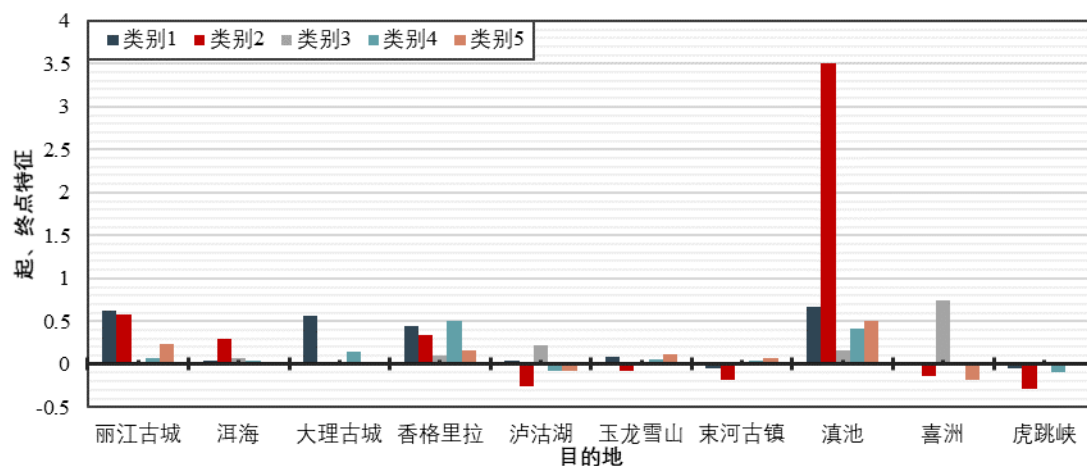


图 5 9 个热门景点在 5 类游客旅游流子网络中的起、终点特征值

Fig.5 The Beginning and Terminal Features of Nine Popular Attractions in the Tourism Networks of Five Clusters of Tourists

除了起、终点特征之外，5 类游客子网络中 9 个热门景点的中转能力也各有特点（图 6）。总体上，丽江古城和洱海这两个节点在各类游客子网络中的中介中心性都相对较高，说明 5 类游客都倾向于将这两个节点作为旅行线路的中转点。除去丽江古城和洱海，类别 3 子网络中介中心性较高的是大理古城、泸沽湖和滇池，表明这些节点在类别 3 的游客旅行路线安排中常作为中转点，

且中转能力相当。值得一提的是滇池，其在类别 3 的旅游流网络中的中介中心性值远高于其余 4 个类别（尤其是类别 1 和类别 2），说明在滇池中转的游客可能相当多属于类别 3，即长时间深度游型。结合起、终点特征的分析，可以发现滇池在不同类别游客中的角色转换。对于类别 2 游客，滇池是旅行线路的起点；对于类别 3 游客，滇池是旅行线路的中转点。

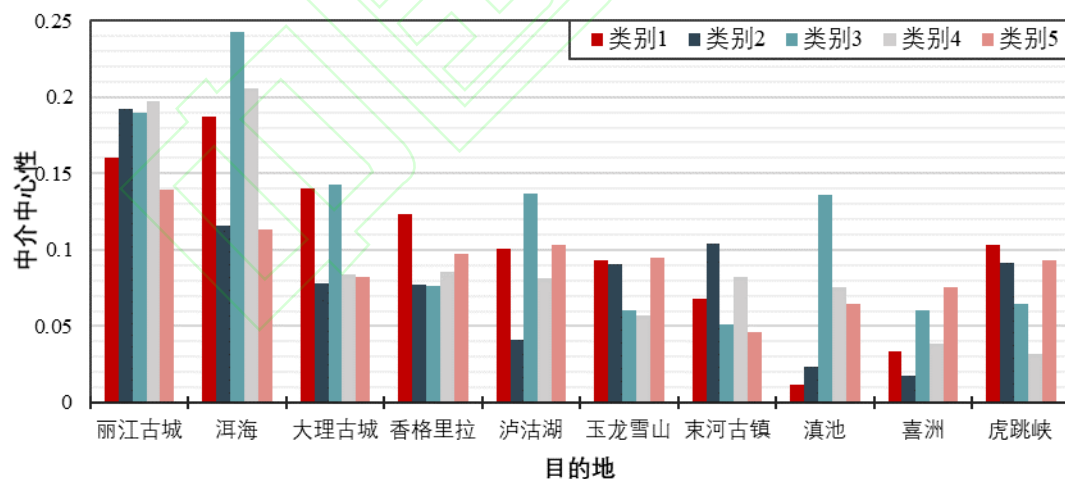


图 6 9 个热门景点在 5 类游客旅游流子网络中介中心性值

Fig.6 Betweenness Centrality Values of Nine Popular Attractions in the Tourism Networks of Five Clusters of Tourists

4.2.3 子网络中节点的中间人特征

以游客在目的地之间转移的频次和目的地之间的距离作为聚类特征，利用 AP 聚类算法为每类游客的旅游流子网络构建节点组团。AP 聚类算法不需要在外部输入聚

类个数，而是通过反复迭代，直到聚类中心在一定程度上不再更新时取得最佳节点组团效果。在本文的实验中当迭代次数达到 20 时，各类游客旅游流网络中的节点组团结构均已稳定。下文中旅游节点的中间人角色均

是基于聚类后的稳定节点组团结构和节点间的实际联系进行分析。

以具有高中转能力的丽江古城为例，图显示了丽江古城在每类子网络中百分比最高的中间人角色及其比例。丽江古城在类别 1、3、4 子网络中扮演最主要的角色都是“联络人”，其中类别 4 子网络中“联络人”角色的比例高达 66.1%，说明丽江古城在这些类别中主要作为不同节点组团间联系的纽带。在类别 2 中百分比最高的角色是“代理人”，说明类别 2（旅行天数敏感型）的游客常将丽江古城作为该节点组团最后一个游玩的目的地，之后转到下一个节点组团游玩。在

类别 5 中百分比最高的角色是“守门人”，表明类别 5 这类中短旅程的游客主要选择丽江古城作为该节点组团首个旅游目的地，并在丽江古城周边深度游。节点的中间人角色在不同类别游客的旅游流网络中存在差异，这主要是因为组团内部、组团间的旅游节点转移成本（路程、时间等）差别较大。此外，旅游节点自身特色上的差异也吸引了不同出行特征和兴趣偏好的游客。这些中间人角色的差异有助于优化旅游网络各节点的协作机制，为游客提供更好的景区设施和交通服务。

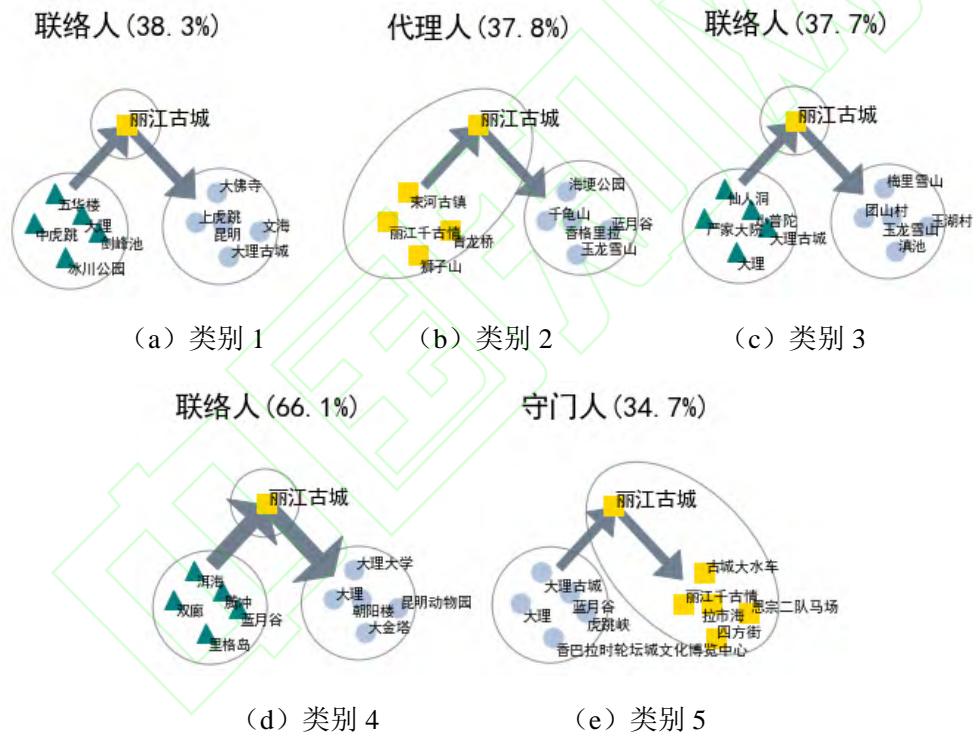


图 7 丽江古城在类别 1~5 旅游流网络中的主要中间人角色

Fig.7 The Broker Roles of Lijiang in Tourism Networks of Five Clusters of Tourists

5 结论

本文以云南省为例，基于在线旅游网站中游客上传的大量游记文本数据，挖掘游客多维度的出行特征和兴趣偏好，并以此对游客进行聚类。基于聚类结果分析了各类游客

的旅游流网络在整体结构和节点角色特征上的差异。结果显示，不同类别游客的旅游流网络在整体结构上各有特点，反映出旅游目的地不同的空间交互模式和网络中心化程度。旅行花费及旅行天数敏感型游客的旅游目的地主要集中在少数热门景点及周边，旅游流网络呈现“单核心线网状结构”；其他类型游客的旅行目的地更加多样，旅行线路

空间跨度更大,旅游流网络呈现明显的多核心结构。此外,部分目的地在不同类别旅游流网络中承担着截然相反的角色,如泸沽湖、喜洲、滇池。本研究的结果有助于辅助旅游景点明晰客流特征、优化旅游网络各节点的协作机制。未来的研究将进一步探讨旅游流网络特征差异的影响因素并将结果应用到个性化的旅游线路推荐中。

参 考 文 献

- [1]. 章锦河, 张捷, 李娜, 等. 中国国内旅游流空间场效应分析. 地理研究, 2005(02): 293-303.[Zhang Jinhe, Zhang Jie, Li Na, et al. An analysis on spatial field effect of domestic tourist flows in China. GEOGRAPHICAL RESEARCH, 2005, 24(2): 293-303.]
- [2]. 马耀峰, 李天顺. 中国入境旅游研究[M]. 北京: 科学出版社, 1999: 12-24. [Ma Yaofeng, Li Tianshun. Research on Inbound Tourist Flow in China [M]. Beijing: Science Press, 1999: 12-24.]
- [3]. 杨兴柱, 顾朝林, 王群. 南京市旅游流网络结构构建. 地理学报, 2007, 62(6): 609-620.[Yang Xingzhu, Gu Chaolin, Wang Qun. Urban Tourism Flow Network Structure Construction in Nanjing. Acta Geographica Sinica, 2007, 62(6): 609-620.]
- [4]. SAINAGHI R, BAGGIO R. Complexity traits and dynamics of tourism destinations[J]. Tourism Management, 2017,63: 368-382.
- [5]. Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research : A literature review. Tourism Management, 68, 301-323.
- [6]. Mou, N., Zheng, Y., Makkonen, T., Yang, T., Tang, J. J., & Song, Y. (2020). Tourists' digital footprint: The spatial patterns of tourist flows in Qingdao, China. Tourism Management, 81, 104151. doi: 10.1016/j.tourman.2020.104151.
- [7]. Xu, Y., Li, J., Belyi, A., & Park, S. (2021). Characterizing destination networks through mobility traces of international tourists — A case study using a nationwide mobile positioning dataset. Tourism Management, 82. <https://doi.org/10.1016/j.tourman.2020.104195>
- [8]. Xu, Y., Xue, J., Park, S., & Yue, Y. (2021). Towards a multidimensional view of tourist mobility patterns in cities: A mobile phone data perspective. Computers, Environment and Urban Systems, 86(January). <https://doi.org/10.1016/j.compenvurbsys.2020.101593>
- [9]. 闫闪闪, 靳诚. 基于多源数据的市域旅游流空间网络结构特征——以洛阳市为例. Spatial Network Structure of the City-Level Tourist Flow Based on Multisource Data: A Case Study of Luoyang[J]. 经济地理, 2019, 039(008):231-240.
- [10]. 季学峰, 吴国清. 基于旅游数字足迹的上海市旅游流网络结构分析[J]. 城市学刊, 2019.
- [11]. Leung, X., WangF, WuB., et al. A Social Network Analysis of Overseas Tourist Movement Patterns in Beijing: the Impact of the Olympic Games[J]. International Journal of Tourism Research, 2012, 14(5):469-484.
- [12]. 闫闪闪, 靳诚. 洛阳城区旅游流空间网络结构特征. 地理科学, 2019, 39(10): 1602-1611.[Yan Shanshan, Jin Cheng. Characteristics of Spatial Network Structure of Tourist Flow in Urban Area of Luoyang. SCIENTIA GEOGRAPHICA SINICA, 2019, 39(10): 1602-1611.]
- [13]. 罗秋菊, 梁思贤. 基于数字足迹的自驾车旅游客流时空特征研究——以云南省为例. 旅游学刊. 2016, 31(12)
- [14]. 云南省人民政府. 云南概况[EB/OL]. (2020-04-25)[4.25]. <http://www.yn.gov.cn/yn/gk/>. [People's Government of Yunnan Province. The General Situation of Yunnan Province. [EB/OL]. (2020-04-25)[4.25]. <http://www.yn.gov.cn/yn/gk/>]
- [15]. 云南省统计局. 云南省2018年国民经济和社会发展统计公报 [EB/OL]. (2019-06-14) [4.25]. http://www.yn.gov.cn/zwgk/zfxgk/jsshtj/201906/t20190627_179363.html. [Yunnan Provincial Statistical Bureau. Statistical Bulletin of the National Economic and Social Development of Yunnan Province 2018. [EB/OL]. (2019-06-14)]
- [16]. Zhang, Y., Jin, R., Zhou, ZH. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics. 2010, 1, 43-52.
- [17]. Salton G, buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988,24(5): 513-523.
- [18]. Laurens M, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research, 2008,9(2605): 2579-2605.
- [19]. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.

- Data Mining and Knowledge Discovery, 1998,2(3): 283-304.
- [20]. Calinski T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods, 1974,3(1): 1-27.
- [21]. Rousseeuw J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987,20: 53-65.
- [22]. Freeman C. Centrality in social networks conceptual clarification. Social Networks, 1978,1(3): 215-239.
- [23]. 胡定利, 李锐, 孟瑶, 吴华意. 新闻地名共现视角下的中国城市网络[J]. 武汉大学学报信息科学版, 2020, 45(2): 281-288.[HU Dingli, LI Rui, MENG Yao, WU Huayi. China's Urban Network from the Perspective of Toponym Co-occurrences in the News[J]. Geomatics and Information Science of Wuhan University, 2020, 45(2): 281-288.]
- [24]. 刘军.整体网分析讲义:UCINET软件实用指南. 上海: 上海人民出版社, 2009. [Liu Jun. Lectures on Whole Network Approach: A Practical Guide to UCINET. Shanghai: Shanghai People's Publishing House, 2009]
- [25]. Oksanen J, Kindt R, Legendre P, et al. Vegan: community ecology package. version 1.11.14, 2008, <http://vegan.r-forge.r-project.org>.

Tourism Flow Network Structures of Different Types of Tourists Using Online Travel Notes: A Case Study of Yunnan Province

LI Qiuping^{1,4} CHEN Yu² LUAN Xuechen^{3,4}

1. School of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China

2. School of Remote Sensing and Information Engineering, Wuhan 430079, China

3. Guangdong Fundway Science and Technology Corporation Limited, Guangzhou 510275, China

4. Guangdong Provincial Engineering Research Center for Public Security and Disaster, Guangzhou 510275, China

Abstract: Objectives: The structure of the tourism flow network is of great significance for understanding the choices of tourists and the role of attractions in the network. The previous studies mainly focused on the structure of the tourism flow network of all tourists. However, the analysis on the disparities of tourism flow network for different tourists is still lack in the thorough research. Therefore, we analyzed the tourism flow networks constructed by different types of tourism routes from online travel notes. **Methods:** Based on the online travel notes, the text mining and social network analysis methods are used to construct and analyze tourism flow networks. First, we use text mining to extract the multi-dimensional preferences of tourists, and cluster tourists into different groups. Second, the destination sequences of different tourist groups are used to construct various tourist flow networks. Finally, the structural characteristics of these tourism flow networks and the role of each destination node are analyzed from multiple perspectives. **Results:** The experiment takes Yunnan Province as the case study area, and the tourists travelled in Yunnan in 2019 are clustered into five groups, then five travel flow networks are constructed. The results show that: (1) The tourism flow network structures of five clusters of tourists are distinct, demonstrating the disparities of spatial interaction patterns among travel destinations and different degrees of network centralization; (2) The travel destinations of cost sensitive and time sensitive tourists are primarily a few popular attractions and some attractions around them. The networks of these two types of tourists show a single-core structure. As for other

types of tourists, their travel destinations are more diverse and their travel routes have a larger spatial span. The networks of these types of tourists present a typical multi-core structure. (3) Some travel destinations like Lugu Lake, Xizhou and Dian Lake take opposite roles in the tourism flow networks of different clusters of tourists. **Conclusions:** Our research is helpful for tourism management department to clarify the characteristics of tourism flows and optimize the cooperation mechanism of travel destinations in the tourism network. In the future work, we will focus on exploring the influencing factors of different tourism flow network characteristics, and applying the results to personalized tourism route recommendations.

Keywords: tourism flow network; text mining; network analysis; tourist preference; online travel notes

First Author: LI Qiuping, Ph D, associate professor, specializes in the theories and methods of transport geography. E-mail: liqp3@mail.sysu.edu.cn.

Corresponding author: LUAN Xuechen, Ph D. E-mail: luanxuechen@fundway.net.

Foundation Support: The National Natural Science Foundation of China (41971345); the Natural Science Foundation of Guangdong Province (2020A1515010695).

网络首发:

标题: 利用网络游记分析不同类型游客的旅游流网络特征差异

作者: 李秋萍,陈宇,栾学晨

收稿日期: 2021-09-01

DOI: 10.13203/j.whugis20210045

引用格式:

李秋萍,陈宇,栾学晨. 利用网络游记分析不同类型游客的旅游流网络特征差异——以云南省为例[J].武汉大学学报·信息科学版, 2021, DOI: 10.13203/j.whugis20210045 (LI Qiuping, CHEN Yu, LUAN Xuechen, . Tourism Flow Network Structures of Different Types of tourists Using Online Travel Notes: A Case study of Yunnan Province [J].Geomatics and Information Science of Wuhan University,2021, DOI: 10.13203/j.whugis20210045)

网络首发文章内容和格式与正式出版会有细微差别, 请以正式出版文件为准!

您感兴趣的其他相关论文:

街景影像下的临街建筑风格映射及地图生成方法

徐虹, 王禄斌, 方志祥, 等.

武汉大学学报·信息科学版, 2021, 46(5): 659-671

<http://ch.whu.edu.cn/cn/article/doi/10.13203/j.whugis20200445>

九寨沟核心景区多源遥感数据地质灾害解译初探

董秀军, 许强, 余金星, 等.

武汉大学学报·信息科学版, 2020, 45(3): 432-441

<http://ch.whu.edu.cn/cn/article/doi/10.13203/j.whugis20190076>