生态游客需求与偏好的文本挖掘研究

肖伟

南京旅游职业学院，江苏 南京 211100

摘　　要： 近年来，生态旅游在全球范围内呈现出迅猛的增长势头，年增长率高达25% ~ 30%，显示出其已经成为世界性旅游的潮流，也为游客提供了更多元化的旅游选择。本研究旨在通过文本挖掘技术，深入分析生态游游客的需求与偏好，旨在为生态旅游业提供定制化服务策略。选取携程旅游网站内丰富的生态游景点信息作为数据源，运用BeautifulSoup4爬虫框架获取杭州景区游客的评论、星级评价以及旅游时间等原始数据。采用Pandas和NumPy进行数据预处理和分析，结合中文分词工具Jieba，对游客评论进行深度挖掘，通过百度情感分析大模型NLP分析从而揭示游客对生态旅游的真实态度与情感倾向。研究结果表明，游客对生态旅游的喜爱程度日益增加，对景区服务质量的关注也在不断提升，这有助于理解现代游客在生态游中的行为模式，为旅游业提供定制化服务策略，同时推动生态旅游的可持续发展。

关 键 词： 生态旅游；大数据技术；需求分析；服务策略

Research on Text Mining of Eco-Tourists 'Needs and Preferences

Xiao Wei

Nanjing institute of tourism and hospitality, Nanjing, Jiangsu 211100

Abstract: In recent years, eco-tourism has seen a rapid growth worldwide, with an annual increase of 25%~30%, marking it as a global tourism trend and offering diverse travel options. This study uses text mining to analyze the preferences of eco-tourists, aiming to provide tailored services for the eco-tourism industry. We sourced data from Ctrip's website, employing BeautifulSoup4 to extract comments, ratings, and visit times from Hangzhou attractions. Using Pandas and NumPy for analysis, along with Jieba for Chinese segmentation, we delved into tourist feedback. The Baidu NLP model revealed tourists' sentiments towards eco-tourism. Results indicate growing affection for eco-tourism and rising concern for service quality, guiding customized strategies for sustainable tourism development.

Keywords: eco-tourism; big data technology; demand analysis; service strategy

引言

随生活品质提升，生态旅游受广泛关注，年增长率达25% ~ 30%。但快速发展带来诸多问题，如服务质量不一、生态保护与体验矛盾、游客需求多样性等，影响可持续发展。因此，研究生态旅游、找出问题和解决方案尤为重要。信息化、数字化时代下，大数据技术为解决问题提供新思路。通过分析大规模数据，深入了解游客需求和行为，为景区提供科学服务策略[1]。本研究基于大数据，分析游客行为和偏好，揭示真实态度和情感倾向，为生态旅游规划、管理、市场营销提供数据支持和决策参考，并为理论研究和实践应用提供理性视角和方法。

一、文献综述

生态旅游概念由墨西哥专家谢贝洛斯·拉斯喀瑞在1983年提出，并在1986年的国际环境会议上得到确认，此后受到广泛关注并快速发展。自20世纪90年代起，全球生态旅游稳步增长，成为旅游业的热点。随着科技进步，大数据和文本挖掘技术在生态旅游中得到应用，如分析社交媒体和在线预订数据，以揭示游客偏好、趋势和客流量，为景区管理提供决策支持[2]。这些技术的应用有助于理解游客需求，促进生态旅游业的可持续发展。

二、研究方法

本研究使用BeautifulSoup4爬虫框架从携程旅游网站抓取杭州景区的游客评论、评分及旅游时间等数据。这些数据反映了游客的真实态度和需求，是文本挖掘的理想数据源。在采集过程中，我们用Python编写的爬虫程序访问并解析了多时段的网页数据，确保了数据集的多样性。通过Pandas和NumPy对原始数据进行预处理，包括清洗、缺失值处理和去重[3]。然后，利用Jieba中文分词工具对评论进行分词，为后续分析打下基础。情感分析

基金项目：南京旅游职业学院校级课题(2023KYC004)；基金项目：江苏省高校“青蓝工程”优秀青年骨干教师资助项目；2024年度全国高等职业院校信息技术课程教学改革研究项目；江苏省社科应用研究精品工程课题(23SYC-072)。

采用百度的 NLP 大模型来识别经过预处理的游客评论中的情感倾向，包括正面、负面和中性情绪。通过汇总分析评论的情感标签，本研究可以定量和定性地了解游客对杭州生态旅游景点的满意度及其变化趋势，并为景区提供改进服务的具体方向[4]。

三、数据分析结果

（一）描述性统计分析

为了更有效地制定市场策略与资源分配，本研究对游客的来源地进行了深入分析，了解游客的地理分布，分析结果显示，大多数游客来自江苏省和广东省显示出较强的地域集中性。此外，来自北京地区也呈现出稳定的游客增长趋势，成为潜在的重点市场[5]。

通过分析不同时间段的游览时间显示，到2022年生态游的游客有了明显的上升趋势，游览时间在夏季和秋季较长，尤其是在暑假和国庆假期达到峰值。相比之下，春冬两季的游览时间相对较短，特别是在非假日时段[6]。

在基于词频－逆文档频率（TF-IDF）算法提取的关键词中，选取前500个生成词云图，如图3所示。在词云图中，所选取的500个关键词，出现的频次越高，字体就越大。由此发现杭州生态景区相关讨论的核心内容主要集中在地点和评论：“西湖”“景色”“雷峰塔”“值得”“杭州”等出现频率较高，表明杭州作为著名的生态旅游地区，是公众讨论的热点之一，特别是关注其作为观赏地及其价值[7]。

> 图1 关键字云图

（二）评论数据情感分析

1. 评论分析

（1）情感倾向分析

在情感倾向分析中，本研究首先对原始评论数据进行去重处理，以保留有价值的信息。通过 pandas 库的 drop_duplicates 方法，从10000条评论中移除了118条完全重复的数据，剩余99882条独特评论。为保证数据质量，将评分转化为正面和负面标签，并将非实质性的数字和字母以及与生态旅游景点评价无关的高频词如“杭州”“乌镇”等剔除，以清洗数据[8]。同时研究采用基于词典的匹配方法，利用知网发布的“情感分析用词语集（beta版）”进行情感分析。该词典包含正面和负面的情感词汇，经过优化后，加入了“好评”“差评”等网络购物相关的词汇，以适应研究场景。每个词汇根据情感倾向被赋予权重，正面词汇为1，负面词汇为 -1。分词后，使用 merge 函数将分词结果与情感词典进行匹配，识别出评论的情感倾向。为直观展示分析结果，研究进一步使用 wordcloud 生成正负面评论的词云图。这种方法不仅提高了情感分析的准确性，也使得结果更为直观，有助于理解游客对生态旅游景点的真实感受和偏好[9]。

> 图2 正面评论词云图

> 图3 负面评论词云图

（2）情感倾向分析预测

为了进一步查看情感分析效果，假定用户在评论时选了好评的标签，而写了差评内容的情况，比较原评论的评论类型与情感分析得出的评论类型，绘制情感倾向分析混淆矩阵：

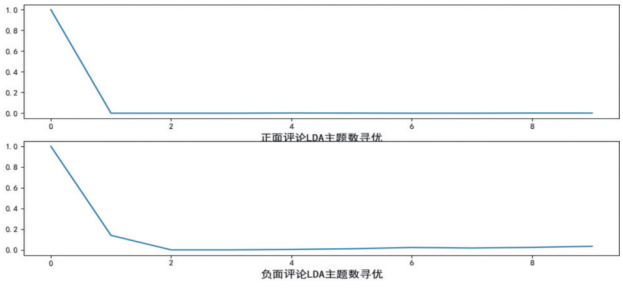
表1 情感倾向分析混淆表

		neg	pos	All
实际	neg	228	141	369
	pos	29	591	620
	All	257	732	989

通过比较原评论的评论类型与情感分析得出的评论类型，基于词表的情感分析的准确率达到了74.01%，证明通过词表的情感分析去判断某文本的情感程度是有效的[10]。

2. 主题分析

潜在狄利克雷分配，即 LDA 模型，用于识别大规模文档集或语料库中的潜在隐藏的主题信息。基于相似度的自适应最优 LDA 模型选择方法，确定主题数并进行主题分析。实验证明该方法可以在不需要人工调试主题数目的情况下，用相对少的迭代找到最



> 图4 寻找最优主题图

优的主题结构[11]。使用 LDA 主题模型，找出不同主题数下的主题词，每个模型各取出若干个主题词，合并成一个集合。生成任何两个主题间的词频向量，计算两个向量的余弦相似度，值越大表示越相似；计算各个主题数的平均余弦相似度，寻找最优主题。

对于正面评论数据，当主题数为2时，主题间的平均余弦相似度就达到了最低。因此，对正面评论数据做 LDA，可以选择主题数为2；对于负面评论数据，当主题数为2时，主题间的平均余弦相似度也达到了最低。因此，对负面评论数据做 LDA，也可以选择主题数为2。

表2 正面评价潜在主题表

Topic1	Topic2
西湖	值得
不错	景色
太	美
古镇	玩
真的	地方
夜景	推荐
酒店	门票
晚上	体验
时间	特别
想	江南水乡

表3 负面评价潜在主题表

Topic1	Topic2
差	票
走	门票
太	体验
地方	坐
点	太
进	排队
路	小时
船	真的
不值	工作人员
失望	游客

四、讨论

从来源地分析，江苏和广东游客占较大比例，显示这两个地区居民对杭州生态游兴趣高。北京地区虽增长但有差距。季节性分析显示，夏秋季是游览高峰期，尤其在暑假和国庆假期，而春冬季节相对较短，尤其在非假日时段。这可能与季节气候有关，夏秋季气候适宜出行，春秋则可能因天气影响出行意愿。从游览时间趋势看，2022年生态游游客明显上升，说明疫情后期人们亲近自然、体验生态环境需求增加。夏季和秋季游览时间较长，符合学生暑假和国庆节旅游高峰。关键词云图显示，“西湖”“景色”“雷峰塔”等地标性词语出现频率较高，表明这些景点在杭州生态旅游中的重要性和吸引力。“值得”一词频繁出现，暗示游客对杭州生态游整体评价较正面[12]。

评论数据的处理显示，通过去重和清洗后，保留了近十万条有效评论。情感分析结果表明，基于词表的情感分析准确率达到了74.01%，这是一个相对较高的准确率，说明该方法能有效区分评论的情感倾向[13]。LDA 主题模型的应用进一步揭示了正负面评论的潜在主题，为理解游客的具体需求和不满因素提供了更深入的视角。通过 LDA 模型识别出游客关注的主要话题。例如，在正面评价中，“西湖”“景色”“值得”等词汇频繁出现，而在负面评价中，“差”“不值”“失望”等表达不满的词汇较为集中。这种信息有助于旅游管理者针对性地改善服务质量和游客体验[14]。

五、结论

研究显示，生态旅游的受欢迎程度在提升，游客更加关注景区服务质量，尤其是杭州西湖地区。高峰旅游时段集中在七八月份，游客追求性价比。情感分析揭示，游客主要关心价格、卫生、体验和服务等方面[15]。因此，发展生态旅游应考虑：一是利用本地资源制定独特旅游战略，避免盲目模仿；二是减少硬件建设，注重旅游的体验性和参与性。

参考文献

[1] Feature Selection Based Data Mining Method［J］. Chemical Research in Chinese Universities, 2021, 27(01): 87-93.

[2] Identifying Metabolite-Protein Biomarkers in In-patients with Unstable Angina［J］. Chemical Research in Chinese Universities, 2022, 79(01): 87-105.

[3] On the Coordinated Development Mechanism of Rural Tourism Development and Ecological Environment Protection in the International Cultural Tourism Demonstration Area in Southern Anhui Province［C］. Yingda Wang International Conference on Urban Engineering and Management Science . 2020,17（10）: 4-26.

[4] Community-Based Tourism Development and Its Effects on the Local Community: the Case of Penglipuran Village, Indonesia［D］. Briliyanti, Astri. 2021,17（18）: 24-28.

[5] Organically Linking Green Development and Ecological Environment Protection in Poyang Lake China Using a Social-Ecological System (SES) Framework［O］. Ji Feng, Zheng Zhao, Yali Wen, 2021,17（07）: 114-136.

[6] 邢露雨胡润鸿，汤陈松. Python 大数据挖掘安徽黟县全境民宿住客体验［J］. 电脑知识与技术，2021,17（09）: 154-158.

[7] 李文华. 基于 Python 的网络爬虫系统的设计与实现分析［J］. 内江科技，2021,42（02）58-59+26,.

[8] 时梨，蔡林. 基于 Python 语言构建神经网络识别手写数字的研究［J］. 电脑编程技巧与维护，2021,（02）: 117-118+130.

[9]Lan Man, Lin Aiwen, Jin Tian, et al. Quantitative analysis of knowledge maps of natural resources accounting and assessment research in China based on CiteSpace［J］. Resources Science, 2020,42(4):621-635.

[10] 李慧，聂寒玉，靳梦菲. 生态旅游景区产品创新对游客体验的影响研究——基于模糊集定性比较分析［J］. 生态经济，2020,36(12):112-117.

[11] 吴殿廷，郭来喜，刘锋，刘宏红，王彬. 世界旅游强国建设：国际经验与中国方略［J］. 中国生态旅游，2022, 12(4): 533-549.

[12] 卢畅，罗芬，王琛. “互联网+”生态旅游产品价值实现机制研究——以南山国家公园线路为例［J］. 中国生态旅游，2022, 12(2): 291-306.

[13] 张光生，林天飞，朱蓉. 俄罗斯国家公园建设与管理体系及其对中国的启示［J］. 中国生态旅游，2022, 12(2): 320-329.

[14] 李文华. 基于 Python 的网络爬虫系统的设计与实现分析［J］. 内江科技，2021,42（02）58-59+26.

[15] 时梨，蔡林. 基于 Python 语言构建神经网络识别手写数字的研究［J］. 电脑编程技巧与维护，2021,（02）: 117-118+130.