

基于 LDA 模型及情感分析的电影 主题公园评论文本研究

林 轶, 曹清芳

(广西大学 工商管理学院, 南宁 530000)

摘要:以中国电影主题公园为例,收集多个旅游网站游客评论数据,采用 LDA(latent dirichlet allocation)主题模型方法、共现网络分析方法及情感分析方法对游客评论文本进行研究。结果表明,国内电影主题公园游客评论可以分为旅游服务主题、旅游景观主题、整体感知主题以及活动体验主题 4 个主题。结合共现网络分析及情感分析结果识别当前景区现状及痛点,为文旅企业经营者提供较为准确的游客偏好信息,也为企业后续管理及决策提供相关理论及数据支撑。

关键词:电影主题公园;游客评论;LDA 主题模型;情感分析

中图分类号:F592.6 **文献标志码:**A **文章编号:**1671-1807(2023)22-0093-06

当前我国旅游业强势回暖,旅游业是各地方经济增长的产业选项。国内诸多影视 IP 成为文旅消费市场的新主题,《“十四五”中国电影发展规划》中明确提出要拓展和延伸电影产业链条,稳步提高电影综合收入。影视型主题公园是电影产业链条的重要一环,也是影视旅游的重要载体。主题公园“舞台化”体验的核心价值就是为消费者创造美好体验,留下难忘的回忆,使消费者愿意再次、多次光临主题乐园。电影主题公园的优势在于可以用电影场景的思维去构建每一处景观和体验,整个电影乐园有专业演员的戏剧表演游走在整个园区,这也是电影主题公园的独特之处^[1]。但目前我国本土电影主题公园存在多种问题,比如影视 IP 资源未能充分实现文化资源的转换及文化空间零散化^[2]、缺乏针对性的市场推广及文化建设的疏忽。这些问题最终会导致游客无法获得身临其境的游玩体验。

智能手机的普及使得智能手机在客户体验中的作用越来越重要,智能手机超越空间和时间的能力可以让用户在旅行过程中获得非常个性化的服务^[3],游客在旅游中获得的旅行体验受到科技的影响,互联网平台上的游客分享经验会对后来的游客决策产生一定的影响,旅游目的地有巨大的机会通过网络评论文本挖掘背后大量有价值的信息。这些评论能展现游客的关注点和情感态度。已有研

究证实利用网络文本进行游客情感分析的可行性^[4]。例如,白健和洪小娟^[5]利用文本挖掘技术多维展现了网民的关注焦点;孙宝生等^[6]利用网络文本挖掘技术构建了游客满意度评价指标体系。因此,本文基于互联网平台的游客评论数据,运用文本主题建模、共现网络分析及情感分析方法,调查和分析中国电影主题公园旅游发展的客观情况,助推中国文旅产业发展。

1 相关文献研究梳理

电影主题公园属于主题公园,目前关于主题公园的研究已有相当数量。例如, Park 等^[7]结合社交媒体分析理论、地理空间分析理论及影响环形模型理论,对迪士尼乐园游客在社交媒体表达的情感进行了研究,并根据研究结果开发了一条唤起愉快情绪的旅游路线。有不少学者基于网络点评文本对主题公园形象感知进行定量研究。例如,程雨等^[8]采用内容分析法和共现网络分析法相结合的方式,对华谊兄弟电影小镇(郑州)进行了形象感知分析。

借助前沿技术,运用文本挖掘方法能使国内外不同领域学者对用户评论进行更深入的研究,隐含狄利克雷分布(latent dirichlet allocation, LDA)主题模型方法由于能准确提取大数据的研究主题而得以运用在不同研究领域。例如,王涛^[9]结合 LDA 主题模型方法和语义网络方法对电商平台酒店的

收稿日期:2023-08-07

作者简介:林轶(1976—),女,广西桂林人,广西大学工商管理学院,教授,博士,研究方向为少数民族经济、旅游经济与管理;曹清芳(1996—),女,湖南常德人,广西大学工商管理学院,硕士研究生,研究方向为民族文化旅游、旅游经济与管理。

在线文本进行分析,提出了主题聚类与语义网络模型(LDA topic clustering and semantic network model, LTC-SNM)文本挖掘法;冯立杰等^[10]采用了LDA-SNA主题共现网络识别、多维空间专利地图及创新机会识别相结合的方法,应用于钛白粉浆料制备技术研发;曹荣等^[11]同样采用SNA和LDA相结合的方法,利用大数据分析对我国慈善政策的演变路径进行深度挖掘;黄昌莉等^[12]采用LDA主题模型和扎根理论相结合的方法,探讨了旅游城市形象感知内容和影响因素之间的关系。LDA主题模型在旅游研究领域运用较少,现有文献中缺乏对游客的评价探索,更多是对于城市形象感知的整体评价展开研究。因此,本文将运用LDA主题模型挖掘游客评论主题,并引入社会网络分析方法和情感分析方法,探究电影主题公园游客评论背后所蕴含的情感倾向和关注点,从而完善景区管理机制,提高游客体验价值。

2 研究设计

2.1 理论模型及研究思路

LDA主题模型是一种无监督机器学习技术,广泛应用于文本特征挖掘、社交媒体、文本聚类及图像处理等领域,当前大量的非结构化社交媒体数据使得无监督的主题建模已经成为文本挖掘领域的一项重要任务^[13]。LDA可对数据进行主题聚类或文本分析,能依据文本中词与词的共现概率挖掘文本中隐含的主题信息^[14],是从大量文档中寻找主题的一种方法。

对电影主题公园历年游客评论进行文本预处理,然后进行LDA主题建模,运用困惑度指标确定游客评论的最优主题数,使用社会网络分析(social network analysis, SNA)方法构建主题-主题和主题-特征词的社会网络, SNA为分析主题词重要程度与关联词语提供支持,词语的共现程度从侧面反映关联关系。然后利用gephi软件将主题社会网络结果进行可视化展示,最后运用Python的SnowNLP模块量化游客对于电影主题公园的整体及各主题的情感倾向。

2.2 数据来源及处理

数据来自国内大型旅游网站(如携程、去哪儿、途牛、美团、马蜂窝、同程等)的游客评论,可以较为全面地反映游客对电影主题公园的体验情况。为保证数据的时效性,选取近五年新开业的电影主题公园,搜索“电影”“影视”“主题公园”等关键字样,综合采用后羿采集器及八爪鱼爬虫软件对海口长影环球100奇幻乐园、北京环球影城、郑州、苏州及

长沙华谊兄弟电影小镇的游客评论进行爬取,采集截止到2023年6月的所有游客评论数据,数据集字段包括用户ID、评论内容、评论时间,删除重复评论、无效评论后,最终获得有效评论32 998条。

将收集到的评论数据进行预处理,评论中包含大量数字、英文、标点符号或无意义中文词汇等与主题内容不相关的字词或符号,运用Python的jieba分词功能对原始语料进行切分,运用哈工大停用词表将与主题无关的标点符号、数字、url、中文词汇剔除,设置自定义词库保留与主题相关的专用名词。数据处理的结果会直接影响模型拟合的效果,因此数据清洗步骤十分关键。

2.3 主题模型及共现网络分析

2.3.1 LDA主题模型分析

在进行LDA主题模型分析前,需要确定不同主题数目下的困惑度,以找到一个最优主题数目,保证LDA模型训练结果的合理性,困惑度(perplexity)可用于评价LDA主题模型指标,困惑度越低说明模型越好。借助Python的gensim库对预处理过后的评论文本进行LDA模型训练,设置模型先验参数 $\alpha=0.1, \beta=0.01$ 。根据困惑度变化图可以计算得出,主题数目为4时出现了明显的拐点,因此将主题数目确定为4。图1是不同主题数目下的困惑度变化情况,结果表明,随着K(主题数目)的增大,困惑度逐渐减小并且当K约为4的时候,存在一个显著的拐点,故拐点4即为K的最佳值,因此将LDA主题的最优主题数选定为4。

表1为主题-关键词矩阵,显示了4个主题以及主题前10个关键词。结合这些词汇并根据特征词的语义关系进一步确认了主题描述名称。同时参考以往的旅游管理相关文献为主题进行编码归纳,归纳为“活动体验”“整体感知”“旅游服务”“旅游景观”4个主题。

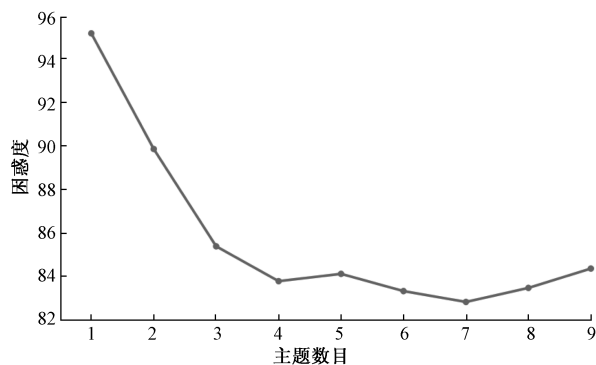


图1 不同主题数目下的困惑度

表 1 主题-关键词矩阵

主题 1 活动体验	概率 1	主题 2 整体感知	概率 2	主题 3 旅游服务	概率 3	主题 4 旅游景观	概率 4
哈利·波特	0.046	主题乐园	0.057	小朋友	0.064	适合拍照	0.084
霸天虎过山车	0.046	有意思	0.039	工作人员	0.062	小朋友	0.064
小朋友	0.030	意大利	0.034	很好玩	0.060	强烈推荐	0.053
过山车	0.017	环境优美	0.027	性价比	0.033	很漂亮	0.044
游乐园	0.015	强烈推荐	0.025	服务态度	0.020	意大利	0.041
变形金刚	0.017	基本上	0.024	万圣节	0.020	很适合	0.035
超级好玩	0.017	性价比	0.018	小姐姐	0.019	还可以	0.029
太棒了	0.013	挺好玩	0.016	华谊兄弟	0.018	好去处	0.024
值得推荐	0.011	游乐场	0.016	挺好玩	0.018	值得一看	0.019
太棒了	0.011	婚纱照	0.010	一家人	0.016	建筑风格	0.019

2.3.2 主题共现网络分析

通过对高频词汇的共现关系进行分析,可以进一步揭示各主题之间的关系。首先对每个主题下的特征词进行整理,对文本中不同主题下的特征词用相应的主题词进行同义词替换的方式,以此为基础构建主题与主题间的共现矩阵(表 2),对角线显示的数目代表主题词在游客评论文本中出现的频次,非对角线的数目是两两关键词在同一条评论中出现的次数^[15]。其次,为了揭示单一主题下特征词关联关系,根据特征词-特征词共现关系构建关键词共现矩阵(表 3)。

最后,利用 gephi 软件将电影主题公园的主题共现网络结果进行可视化展示,并通过共现网络考察主题与主题间的关联关系,以及单一主题下特征词的共现关系。共现网络分析的可视化结果如图 3 所示,关键词共现网络图中节点的大小由节点的度值决定,节点度值越大,节点的圆圈也越大,该主题的游客关注程度越高。同时各主题间连接线条的粗细度及颜色与对应节点主题共同出现的数量成正比,线条越粗颜色越深代表两个主题之间的共现程度越高。

由图 2 可知:①主题与主题的连接线条的粗细

度能反映对应节点主题共同出现的数量,因此从整体而言,旅游服务主题和旅游景观主题的共现程度较高,整体感知与旅游服务主题及旅游景观主题的共现程度次之,而活动体验主题与其他三个主题的共现程度较弱。说明游客对旅游服务和旅游景观两个主题尤为关注,而整体感知与这两者的主题联系较为密切。②从旅游服务主题与特征词的关系来看,“服务态度”“工作人员”“小朋友”这三个关键词的联系较为密切,说明游客在游玩中所感受的旅游服务更多是受到工作人员的影响,而小朋友的服务感知对于游客评判服务表现发挥了重要的作用。③从旅游景观主题与特征词的关系来看,主题旅游景观的关键词联系较为密切,其中“意大利”“意大利建筑”“适合拍照”“建筑风格”“小朋友”的联系较为密切。可以推测出电影主题乐园的建筑风格以

表 2 主题-主题的共现矩阵

共现次数	活动体验	整体感知	旅游服务	旅游景观
活动体验	5 650	1 275	1 981	560
整体感知	1 275	3 897	1 349	591
旅游服务	1 981	1 349	6 375	838
旅游景观	560	591	838	2 995

表 3 特征词-特征词共现矩阵(主题 1)

共现次数	哈利·波特	霸天虎过山车	小朋友	过山车	游乐园	值得推荐	强烈推荐	变形金刚	太棒了	超级好玩
哈利·波特	1 089	266	109	256	24	14	56	392	9	12
霸天虎过山车	266	538	29	89	12	5	18	229	3	4
小朋友	109	29	1 283	70	27	59	66	42	4	8
过山车	256	89	70	1 235	53	13	35	162	7	3
游乐园	24	12	27	53	893	2	13	11	0	0
值得推荐	14	5	59	13	2	825	10	14	8	4
强烈推荐	56	18	66	35	13	10	1 122	22	7	10
变形金刚	392	229	42	162	11	14	22	529	9	4
太棒了	9	3	4	7	0	8	7	9	117	1
超级好玩	12	4	8	3	0	4	10	4	1	148

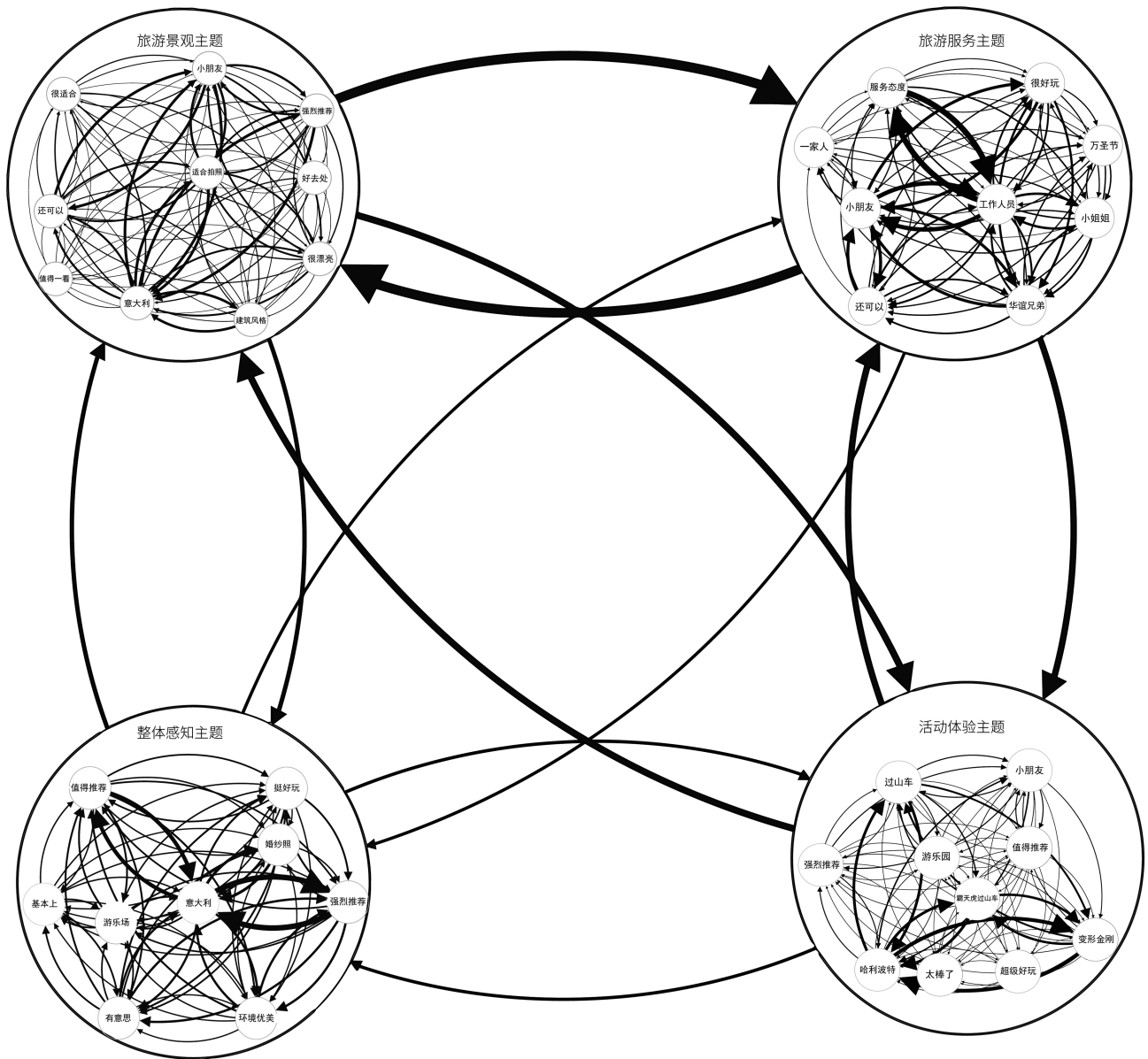


图 2 社会网络可视化图

及独特的意大利风格建筑比较吸引游客,因此游客愿意驻足停留拍照。④从整体感知主题与特征词的关系来看,“值得推荐”“强烈推荐”“意大利”“挺好玩”这几个关键词的联系较为紧密。可以推测出游客对于电影主题乐园的整体评价还是较高的,都有推荐的意愿,而“意大利”这一关键词的出现说明电影主题乐园建筑风格的特色对于游客整体感知评价能产生一定的影响。⑤从活动体验主题与特征词的关系来看,“哈利波特”“变形金刚”“霸天虎过山车”“过山车”这几个关键词联系较为密切。可以推测出游客更倾向这几类游乐项目,其中“过山车”出现的频率较高,说明游客对于这一游乐项目较为偏爱。

2.4 评论情感分析

运用 python 的 SnowNLP 库对游客评论文本进行情感分析,SnowNLP 是专门用于处理中文文本的类库,通常是利用贝叶斯机器学习方法训练出来的模型,可用于核心情感的分析处理^[16]。利用 SnowNLP 库将评论数据情感极性划分为正向、中性、负向并标注为 1、0、-1,文本情感分析结果如表 4 和图 3 所示。

表 4 各主题情感得分占比

主题类别	总体占比/%	正向	负向	中性
整体感知主题	19.21	0.823	0.109	0.068
旅游景观主题	30.03	0.835	0.090	0.075
旅游服务主题	31.17	0.711	0.178	0.112
活动体验主题	19.59	0.450	0.349	0.200

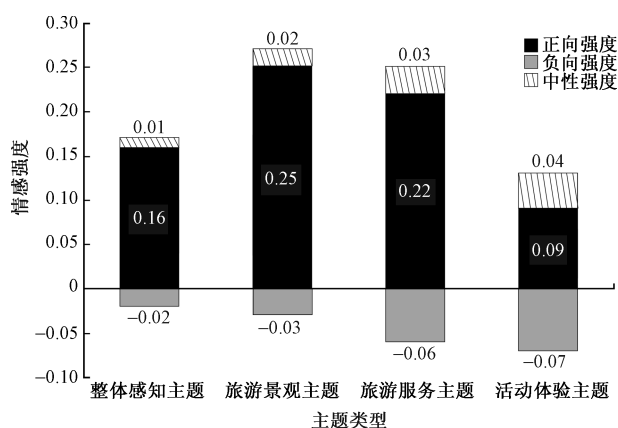


图3 主题情感强度结果

整体评论情感可以分为正向、中性及负向三个方面,且占比分别为70%、12%和18%。情感倾向数值越接近1表示情感表现越积极,根据本文情感分析结果显示整体评价靠近数值1,说明游客对于电影主题公园的游玩感受呈现较为积极的态度。从各主题情感分析结果显示,除了活动体验主题部分接近数值1,并表现出45%的正向情感,其他三个主题整体情感倾向接近数值1,说明游客对三个主题的评论态度为积极,其中各主题情感得分占比表现情况如表4所示。同时,本文通过情感极性判断评论文本得分,在计算出各个主题的情感得分之后,依据各评论数占总体的比例来利用公式加权平均得出情感强度,如图3所示。

由此可知,游客对于电影主题公园的总体满意度良好,虽然大部分游客对电影主题乐园游玩感受持正面评价,但是各主题之间仍存在差异。通过对正、负面评论的内容做进一步分析:旅游景观主题的正向强度为0.25,位于4个主题正向情感强度之首;活动体验主题的负向情感强度为-0.07,位于4个主题负向情感强度之首;且活动体验主题中性情感强度为0.04,同样位于4个主题中性情感强度之首。

进一步研究发现游客对于电影主题乐园的活动体验这一主题感受相对其他主题而言不太满意,具体负面情绪体现在游客太多排队时间长、游乐项目太少、工作人员的专业程度不够或者服务人员态度不好、疫情期间没有提前告知停业、部分表演节目不够满意等方面。比如部分游客表示“根本没人管,排队排了俩小时,最后全是插队的,好多项目都没有玩到”,但游客对部分景区的活动体验评价较高,有重游意愿。比如有游客认为“北京环球影城优速通可以节约排队时长,推荐哈利·波特禁忌之

旅,水世界的表演也很有意思,很多项目完成了二刷”。相比于表演节目,机动游玩项目会让游客产生更高的兴趣^[17]。由此可以看出,活动体验项目的特色及主题的吸引力会对游客情感产生一定的正向影响,但服务水平和景区管理的不佳会对游客产生不好的体验感受。

3 研究结论与建议

3.1 结论

本文采用LDA主题模型方法、共现网络分析方法及情感分析方法对电影主题公园游客评论主题及游客情感进行分析。从LDA主题模型研究结果可以发现,国内电影主题公园游客关注点主要集中在游客游玩的体验感受(活动体验)、游客对于整个景区的整体评价(整体感知)、游客感受到的服务(旅游服务)及景点景观(旅游景观)四个方面。

进一步对各主题及关键词进行共现分析发现:游客评论中旅游服务主题、旅游景观主题、整体感知这三者的关联性较高,其中旅游服务主题与旅游景观主题同时出现在游客评论中的频次最高。这反映了游客对于电影主题公园游玩的关注点主要集中在服务、景观这两方面,相比较而言活动体验主题与其他三者关联程度较弱,往往是独立出现。这说明部分游客对于活动体验主题的印象不够深刻,电影类主题公园活动的多样性和娱乐性十分重要,因此景区需要对活动的设计和文化价值引导方面进行加强。

从具体关键词的共现关系可以看出,旅游服务主题中“工作人员”与“服务态度”关键词会同时出现,侧面反映工作人员在旅游服务这一环节占据的重要作用,这往往是影响游客游玩感受的重要环节。旅游景观主题中“拍照”“建筑风格”“小朋友”“强烈推荐”这几个关键词的出现从侧面反映游客的拍照意愿和景区建筑风格的关联程度。

运用情感分析方法对游客评论进行深入研究表明,“活动体验”主题目前是电影主题公园的不足方面,主要问题集中在景区管理、排队时间过长、活动设计、人员服务等方面,为保证后期景区对游客的持续吸引力,景区需要得以重视并加以调整。针对上述问题,景区可以采取提高服务质量、对工作人员进行专业培训、丰富产品种类并完善结构体系、建立管理系统实现科学分流从而减少游客排队时长等系列手段进行改善。

3.2 建议

依托数字化手段实现信息化管理。完善景区

数字化基础建设,构建完善的信息处理体系,创新游客需求数字化分析。发挥社交网络连接作用,通过抖音、微博、微信公众号等媒体平台进行展示传播。充分利用数字技术对游客进行行为画像,从而实现旅游个性化服务和精准营销。重视信息化人才培育,实现科技信息知识普及以保证景区信息化管理运营。

创新项目开发提升游客旅游体验。由于游乐设施项目是电影主题公园的主要产品之一,因此创新思维能为具有独特个性的旅游产品的诞生提供新的要素。景区运用创新思维开发创新性旅游产品,综合景区内旅游吸引物、旅游基础性设施、相关服务并不断提炼景区卖点,开发深层次体验产品从而打造出全新的旅游吸引物。电影主题公园的客源主要是回头客,景区可在经营过程中对现有项目不断实施改造,对部分项目设施进行淘汰并引进新项目从而吸引消费者。

提高服务水平促进景区质量提升。提升景区旅游质量首先需加强工作人员的旅游服务意识,游客在整个旅游消费过程中体验到的旅游服务主要由景区工作人员来完成,游客感受到的服务质量需由景区全体工作人员和有关部门共同合作完成。其次,景区可通过员工培训教育使工作人员充分意识服务质量对于企业经营及效率的重要性,服务知识技能的教育可使工作人员掌握基本的服务技能,提高业务及服务水平。

参考文献

- [1] 陆嘉宁. 特许权内容消费与沉浸式体验耦合策略——以迪斯尼乐园和环球影城为例[J]. 当代电影, 2021(2): 126-132.
- [2] 张卓. 文化传统与影视 IP 的创新融合——中国本土主题公园的发展路径[J]. 人民论坛, 2021(35): 110-113.
- [3] KIROVA V, THANH T. Smartphone use during the leisure theme park visit experience: the role of contextual factors [J]. Information & Management, 2019, 56(5): 742-753.
- [4] 阎友兵, 郭亮宏. 基于网络文本的红色旅游游客情感特征研究——以韶山风景名胜区为例[J]. 湘潭大学学报(哲学社会科学版), 2020, 44(3): 131-136.
- [5] 白健, 洪小娟. 基于弹幕的网络舆情文本挖掘与情感分析[J]. 软件工程, 2022, 25(11): 44-48.
- [6] 孙宝生, 敖长林, 王菁霞, 等. 基于网络文本挖掘的生态旅游满意度评价研究[J]. 运筹与管理, 2022, 31(12): 165-172.
- [7] PARK S B, KIM J, LEE Y K, et al. Visualizing theme park visitors' emotions using social media analytics and geospatial analytics[J]. Tourism Management, 2020, 80: 104127.
- [8] 程雨, 宋航, 程丽. 基于网络文本分析的影视型主题公园形象感知研究——以郑州建业·华谊兄弟电影小镇为例[J]. 市场论坛, 2021(1): 46-52.
- [9] 王涛, 李明. 基于 LDA 模型与语义网络对评论文本挖掘研究[J]. 重庆工商大学学报(自然科学版), 2019, 36(4): 9-16.
- [10] 冯立杰, 王文豪, 王金凤, 等. 基于 LDA-SNA 和多维空间专利地图的核心技术主题及创新机会识别研究[J]. 科技管理研究, 2023, 43(5): 165-174.
- [11] 曹蓉, 刘彦芝, 王铮. 中国慈善政策合作网络与主题热点演化研究——基于 SNA 和 LDA 的大数据分析[J]. 社会保障研究, 2023(1): 41-52.
- [12] 黄昌莉, 黄燕玲, 于海涛, 等. 基于 LDA 和扎根理论的桂林旅游地感知形象研究[J]. 地域研究与开发, 2022, 41(5): 91-97.
- [13] 陈嘉钰, 李艳. 基于 LDA 主题模型的社交媒体倦怠研究——以微信为例[J]. 情报科学, 2019, 37(12): 78-86.
- [14] 马誉蓁, 侯小妮, 段红梅, 等. 基于 LDA 主题模型的网络问答社区妊娠合并糖尿病孕妇健康信息需求分析[J]. 护理学杂志, 2023, 38(7): 86-89.
- [15] 周婷玮. 基于共现网络与情感分析的多平台消费者评论主题比较研究[J]. 知识管理论坛, 2023, 8(2): 79-91.
- [16] 邵俊利, 林育彬, 郭伟锋. 森林城市的旅游意象及情感认知研究[J]. 林业经济问题, 2022, 42(5): 540-549.
- [17] 鲍珊珊, 凌洁. 顾客满意度的主题公园游客在线评论情感分析[J]. 现代商业, 2020(25): 14-15.

Research on Movie Theme Park Criticism Based on LDA Model and Sentiment Analysis

LIN Yi, CAO Qingfang

(School of Business Administration, Guangxi University, Nanning 530000, China)

Abstract: Taking China's movie theme parks as an example, the visitor review data of multiple tourism websites were collected, and the LDA theme model method, co-occurrence network analysis method and sentiment analysis method was used to study the tourist review text. The results show that the domestic movie theme park visitor reviews can be divided into four themes: tourism service theme, tourism landscape theme, overall perception theme and activity experience theme. Combined with the results of co-occurrence network analysis and sentiment analysis, the current situation and pain points of the prospect area are identified, which provides more accurate tourist preference information for cultural tourism enterprise operators, and also provides relevant theoretical and data support for the subsequent management and decision-making of enterprises.

Keywords: movie theme parks; visitor reviews; LDA topic model; sentiment analysis