◆计算科学◆

基于UGC数据的旅游数据挖掘研究进展*

俸亚特[1,2],徐正丽[3**],文益民[1,4]

(1.桂林旅游学院,广西文化和旅游智慧技术重点实验室,广西桂林 541006;2.桂林旅游学院旅游数据学院,广西桂林 541006;3.桂林电子科技大学商学院,广西桂林 541004;4.桂林电子科技大学计算机与信息安全学院,广西桂林 541004)

摘要:随着互联网和社交媒体的蓬勃发展,用户生成内容(User Generated Content,UGC)数据逐渐成为旅游大数据的重要组成部分。UGC数据能够体现游客旅游行为,数据类型丰富、真实性强、噪声大。本文回顾了过去几年旅游UGC数据研究的发展,分别从文本、照片、多模态3种数据类型角度进行综述,总结了近年来旅游UGC数据挖掘研究取得的成果,并探讨了未来研究的方向。

关键词:UGC;多模态;数据挖掘;旅游大数据

中图分类号:TP181 **文献标识码**:A **文章编号**:1005-9164(2024)01-0087-13

DOI:10.13656/j.cnki.gxkx.20240417.009

互联网的飞速发展把世界带入了自媒体时代,人们比过去任何时候都更加乐意在社交媒体上分享自己在旅游时的所见、所闻和所感。用户生成内容(User Generated Content,UGC)数据指由用户生成的高度原创数据,具有客观性、真实性、及时性、丰富性及噪声大等特点,能够全方位表现游客的各类旅游行为。旅游景区、旅游服务提供商以及旅游行政管理部门可以利用UGC数据洞察游客的动机,深入了解游客的偏好和需求从而改善其服务以吸引更多游客。因此,UGC数据逐渐成为研究旅游的新兴数据源[1-9]。目前,基于UGC数据的研究在旅游大数据研究中占比很高[10],Marine-roig等[1]基于关键词的方法分析了来自加泰罗尼亚地区的游客游记;Cheung等[2]分析了不同形式的UGC数据对游客行为的影响;Zhang等[3]讨论了UGC数据在目的地营销中的重要性以及对目的地营销组织的影响。这些以UGC数据为驱动的数据挖掘研究可为旅游景区、旅游行政管理部门以及旅游服务提供商提供数据支持以及决策建议。

游客往往通过文字以及镜头记录旅游体验,因此所形成的UGC数据大体包括文字、照片以及视频3种基本类型,不同类型的数据组合起来可构成多模态数据,这些非结构化数据有着信息量大、多样性高的

收稿日期:2023-10-21 修回日期:2024-01-03

*国家自然科学基金项目(62366011),广西重点研发计划项目(桂科AB21220023),广西图像图形与智能处理重点实验室项目(GIIP2306),广西高校中青年教师科研基础能力提升项目(2023KY0850)和桂林市重点研发计划项目(20220115-1)资助。

【第一作者简介】

俸亚特(1996—),男,硕士,主要从事数据挖掘和人工智能等研究。

【**通信作者简介】

徐正丽(1982—),女,副教授,主要从事大数据及信息化管理等研究,E-mail:xu_zhengli@gsgsg.uum.edu.my。

【引用本文】

俸亚特,徐正丽,文益民.基于UGC数据的旅游数据挖掘研究进展[J].广西科学,2024,31(1):87-99.

FENG Y T,XU Z L,WEN Y M. Research Progress in Tourism Data Mining Based on UGC Data [J]. Guangxi Sciences,2024,31(1):87-99.

特点。传统的人工分析方法不仅效率低,而且主观性强;先进的人工智能(AI)技术既可以从照片中自动检测目标[11],又可以从文章中提炼主题[12],这为挖掘旅游 UGC 数据中的潜在知识奠定了技术基础。因此,学者们深入研究了旅游 UGC 数据挖掘,如 Liang 等[13]以网站上的游客评论为数据源,运用词频分析、语义网络分析和情感分析等技术,分析了游客对无锡的文化形象感知;Yim 等[14]用深度学习算法识别照片中包含的对象及其类别,再使用回归模型分析照片的内容、用户对照片的主观理解与用户对照片的点赞、评论和浏览之间的关系。

到目前为止,学术界已经发表了一些关于旅游大数据研究的综述性论文。Li 等[10]根据数据来源的差异将旅游大数据分为 UGC 数据、设备数据和事务型数据 3 大类,并对这些不同类型的旅游数据从研究重点、数据特征、分析技术、主要挑战和未来方向 5 个角度进行了全面综述。文益民等[15]从推荐内容、推荐方法和利用到的数据 3 个层面总结了利用旅游数据进行个性化旅游推荐的研究工作,并简述了旅游推荐的实际应用和面临的挑战。Li 等[16]对应用于旅游大数据分析的文本挖掘技术进行了详细综述,总结并讨论了基于自然语言处理的主题提取、文本分类、情感分析和文本聚类等技术在旅游市场统计、旅游目的地形象分析、市场需求等方面的应用。Samara 等[17]介绍了大数据和人工智能技术在旅游中的应用,探索了新兴技术对旅游业各方面的影响,论证了大数据和人工智能可以在预测、产出、提高、提供 4 个方面为旅游业创造价值。Lyu 等[18]通过分析旅游大数据的研究现状,从伦理学层面对旅游大数据研究的未来发展趋势、哲学意义、方法论以及实际意义进行了讨论。尽管以上综述论文说明了旅游 UGC 数据所蕴藏的巨大价值,但学界目前还缺乏对旅游 UGC 数据挖掘方法及技术创新的介绍。考虑到针对视频数据的工作还较少,本文试图从机器学习的视角对旅游 UGC 数据挖掘中针对文本、照片类型数据的算法进行综述。另外,由于多模态数据分析的兴起,本文还介绍了对旅游 UGC 多模态数据的挖掘。最后,本文还对旅游 UGC 数据挖掘研究的未来趋势进行了展望。

1 旅游 UGC 文本数据挖掘

旅游 UGC 文本数据是游客对自己的旅游过程进行描述或评价而生成的文本,包括在线评论、游记、攻略等。文本通常能比较清楚地描述游客在旅游过程中的所见、所闻及所感。由于社交媒体平台以及在线旅行社(Online Travel Agent,OTA)的快速发展,游客乐于在不同网络平台(如微信、QQ、小红书、马蜂窝和携程等)上发表游记、攻略和评论。

游客做出旅游决策前往往会受到其他游客的旅游评论、博客或游记等文本数据的影响[19]。游记、攻略和评论数据可以为潜在游客提供旅游建议,帮助他们凝练旅游动机和兴趣、选择更合心意的旅游目的地、优化旅游路线等。这些信息也能帮助旅游服务提供商提取游客画像,从而了解游客对旅游产品的满意度,进而改进旅游服务。旅游行政管理部门则可以通过目的地评论数据了解游客的需求和满意度、掌握目的地形象的投射以及旅游服务提供商的服务质量,以加强对本地区旅游行业的管理,从而提高游客满意度。

自然语言处理(Natural Language Processing,NLP)技术为旅游 UGC 文本数据挖掘提供了方法支持,旅游 UGC 文本数据挖掘的一般流程如图 1 所示。本章节分别针对旅游 UGC 文本的情感分析、主题抽取、文本分类及文本聚类 4 种技术进行述评。

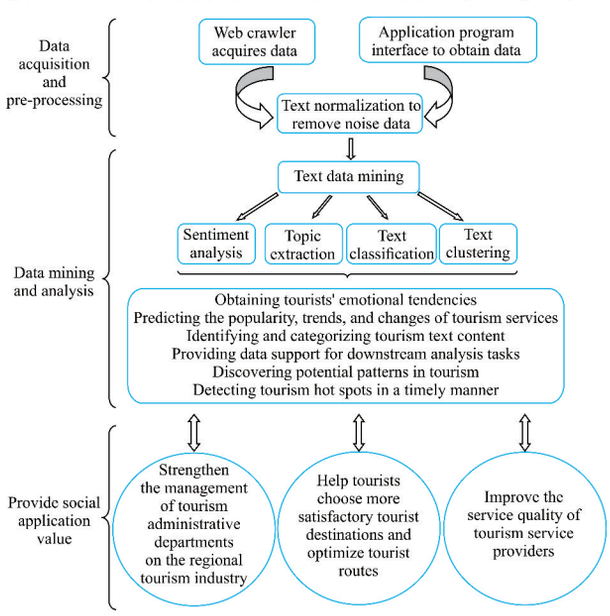

图 1 UGC 文本数据挖掘流程

Fig. 1 Flowchart of UGC text data mining

1.1 旅游 UGC 文本数据挖掘中的情感分析

情感分析是旅游 UGC 文本数据挖掘的主要任务之一,是指通过算法对旅游 UGC 文本表达的情感进行识别和分类,以确定作者对某个特定话题、产品等的态度是积极、消极还是中立。对用户在社交媒体、旅游平台以及其他在线渠道上发布的评论、评价

和反馈进行情感分析,可以深入了解用户对旅游体验的情感倾向,有助于旅游提供商深刻了解用户需求,提高服务水平,进而提升竞争力和旅游体验质量。情感分析方法可分为两类:第一类为基于情感词库的方法,即使用预置的情感词库对文本进行情感分析;第二类为基于机器学习的方法,即先构建一个机器学习模型,再利用训练数据优化模型参数,最后使用模型对未学习过的文本进行情感识别。

早期旅游UGC文本数据的情感分析大多通过手动构建情感词库的方法实现。刘逸等[20]首先利用网络爬虫爬取了来自百度旅游网、去哪儿网和携程网平台的120 731条评论数据,然后人工构建情感词库,最后利用人工设计的情感乘数与计算规则计算整个文本的情感得分,该研究利用世界旅游组织给出的总体游客满意度监测数据作为参照优化模型参数,具有机器学习的基本思想。但是随着时代的发展,涌现出大量由用户创建的"新词汇",这导致情感词库不再具有普适性。Li等[21]针对旅游评论中用户创建的"新词汇"多且难以权衡的问题,提出一种新词检测和语义校准相结合的方法以提高情感词库的质量,实验结果验证其准确率和$F1$值均高于其他情感分析算法。

由于上述方法依赖人工设定的规则和情感词库,当文本结构较复杂或包含未收录的情感词时,会存在一词多义和缺乏单词上下文语义信息等问题[22],因此不受词汇限制的机器学习方法逐渐进入旅游文本情感分析领域。刘文远等[23]针对细粒度文本分类中的类别不平衡问题,设计出一种双向循环卷积注意力网络,并利用批处理平衡技术以提高预测准确率,实现了旅游场景下细粒度文本情感分类。张旭辉等[22]通过构建融合注意力机制的卷积神经网络模型对旅游餐饮文本的情感取向进行分类。Li等[24]提出带有主题词增强嵌入和情感分类注意机制的BiGRULA循环神经网络模型,其在一个酒店评论数据集上的实验结果显示所提出的模型比其他算法能够更加准确地进行情感分类。Luo等[25]基于地质公园旅游发展的需求,爬取了120 532条关于地质公园的旅游评论,通过分词、词向量提取和主成分分析等数据预处理技术提取文本特征,再利用支持向量机进行情感分类,分析了游客对地质公园旅游的满意度。Li等[26]使用网络爬虫技术从TripAdvisor平台上爬取了约7 522家东京餐厅的143 131条原始评论,并从中选择2 335条评论作为基准数据交付专家进行人工标注,使用预训练过的BERT[27]词向量层将评论数据进行词向量化转换,利用长短期记忆(Long Short-Term Memory,LSTM)[28]网络进行语义编码,最后使用交互式叠加注意力(Attention-Over-Attention,AOA)网络[29]拼接两个特征表示用于情感分类。

以上研究尽管可以通过量化情感值来达到分析游客情感的目的,但人类情感表达具有丰富且复杂的形式,包括暗示、比喻、讽刺等,情感量化模型容易忽略情感表达的复杂性。另外,人的情感总是针对具体对象而产生,游客对一个景点内的不同景观或服务项目可能具有不同的情感体验。因此,细粒度的情感分析更具针对性和准确性,从而极大提高旅游UGC文本数据情感分析结果的可用性。

1.2 旅游UGC文本数据挖掘中的主题抽取

主题抽取是指自动识别和提取文本内容所属的主题。通过主题抽取,可以预测旅游服务的受欢迎程度及其趋势和变化。Hu等[30]采用结构主题模型(Structural Topic Modeling,STM)检测负面评价以及不同等级酒店的公众关注度的变化。Loureiro等[31]使用基于层次贝叶斯模型的主题挖掘方法从210篇文章中提取了包括"酒店""场所依恋""个体意识"等8个潜在主题,从而分析环保行为对旅游和酒店管理的影响。通过主题抽取,可以识别和归类例如游记、评论等相关旅游文本内容,有助于旅游从业者更好地了解用户兴趣,从而提供更符合用户需求的旅游服务。Guo等[32]为确定影响消费者满意度的潜在因素,从TripAdvisor网站爬取了包括16个国家25 670家酒店的266 544条在线评论,并使用隐狄利克雷分布(Latent Dirichlet Allocation,LDA)[33]模型进行主题抽取,得到"办理入住和退房手续""差评""度假村设施"等30个潜在的评论主题,这些主题反映了酒店评论中涉及的各个方面,可帮助酒店更好地了解顾客的需求。

使用文本主题抽取技术可以为旅游文本分类任务设置类别。Lee等[34]利用LDA模型对Flickr文本数据进行主题提取,得到"市场/美食街""文化遗产/历史""景观平台""文化/节日""公园/自然风景""宗教场所""购物/城市""海岸"和"文化村"9个主题类别,之后使用LSTM网络对Flickr文本数据进行分类,并将分类结果与位置信息结合起来,在地图上将游客分布可视化,最后通过划分训练、验证、测试数据集进行实验验证,结果证明该算法能够准确、快速地挖掘游客的喜好和旅游目的地,并提供有价值的旅

游建议。

目前针对旅游 UGC 文本数据进行主题提取大多基于 LDA 模型,LDA 模型作为一种传统的主题模型,高度依赖词袋模型,难以捕捉复杂的语义信息和词汇之间的语境关系,也难以处理长距离的语义依赖。相比之下,深度学习模型能够更好地学习文本的语义表达,更适用于捕获文本数据的长距离依赖,具有更优秀的性能。因此,将深度学习应用于主题抽取任务将进一步提高其效率和准确性。

1.3 旅游 UGC 文本数据挖掘中的文本分类

旅游 UGC 文本数据分类可为下游的综合分析任务提供基础数据支持。池云仙[35]提出一种基于粒计算的旅游文本分类与热度挖掘方法,以解决旅游文本挖掘过程中的文本信息粒化、粒计算模型构建、数据集构建、文本关联特征选择、数据粒自动分类和旅游热度计算等问题。王祥翔等[36]提出一种基于朴素贝叶斯分类器的文化旅游文本分类模型,该模型首先构建文化专题词库,利用向量空间模型将景点描述文本转换为向量,通过信息增益进行词汇特征选择以及使用词频-逆文档频率进行权重赋值,再构建分类器模型,实现旅游文本的自动分类,实验选取了 1 447 个景点描述文本,按照闽南文化、客家文化、红色文化和生态文化进行分类。马喆康等[37]提出一种集成词级卷积神经网络(WL-CNN)与句级双向长短期记忆(SL-Bi-LSTM)网络的旅游问句文本分类算法,该算法利用 WL-CNN 和 SL-Bi-LSTM 分别学习词序列子空间向量和句序列深层语义信息,通过多头注意力机制将两种深度学习模型进行集成以实现旅游问句文本的语法和语义信息互补。Chang 等[38]为探究高质量的酒店回复用户评论策略,收集 113 685 条酒店的用户评论和商家回复并将其转换为文档矩阵,通过搭建一个卷积神经网络模型对酒店评论中的主动和非主动回复进行分类,并在模型全连接层之前加入评论的情感得分、回复间隔以及评审员等级 3 类特征以提高分类准确率。

虽然旅游 UGC 文本数据挖掘中的文本分类目前已经取得了一定的成果,但是同样存在一定的缺陷。上述方法大都使用特定领域的数据集,这可能导致模型在其他领域或更大规模数据上的泛化性能不足。未来仍有诸多研究方向可推动其快速发展。例如,如何运用迁移学习从而在少量标注或无标注样本情况下实现旅游 UGC 文本数据挖掘中的文本分类任务;如何将与旅游领域相关的先验知识融入迁移学习模型中以提高模型抽象概括的能力,尤其是利用现有的大模型去构建更加高效的适应旅游行业垂直应用的分类模型等。

1.4 旅游 UGC 文本数据挖掘中的文本聚类

文本聚类是一种无监督算法,可以将文本数据聚集成若干个簇,并利用这些簇将文本数据划分到不同的簇中以发现不同簇的模式和特征,利用文本聚类技术可以从大规模旅游 UGC 文本数据中发现旅游行业的潜在规律。由于 UGC 文本数据往往是非结构化数据且无标签类别,因此文本聚类算法在旅游文本数据挖掘领域有着广泛的应用场景。

旅游 UGC 文本数据的聚类有助于更及时地发现旅游热点和突发事件。李娟等[39]通过构建高频词共现矩阵进行聚类分析以获取旅游热点,并利用社交网络地图中的度中心性测量西藏区内与区外之间的联系权重,获得旅游热点及景点之间的相互关系并用于旅游规划。丁晟春等[40]使用突发性主题表示文本,同时使用内聚分层聚类方法来检测旅游中的紧急事件。

利用文本聚类可以将游客自动划分为不同的人群,从而了解他们对旅游产品的不同评价和需求。为发现伊朗出境游游客的兴趣点和旅游目的地之间的关系,Sohrabi 等[41]利用社交媒体中的评论数据建立评论数据集,使用 X-means 聚类算法对游客评论进行聚类,将游客分为 4 个簇,从而得到不同类型的游客对旅游目的地的偏好程度。此外,旅游 UGC 文本数据的聚类也可以应用于旅游市场的细分[42,43],从而帮助企业分析游客需求、制定更合理的产品策略和拓展更精准的市场营销方案。

目前的旅游 UGC 文本数据的聚类一般依赖高频词共现矩阵或手动选择特征,这种表示方法可能无法充分捕捉文本的丰富语义和上下文信息,从而导致文本特征表达不够准确。未来可以结合深度学习技术,实现语义表示和特征提取的自动学习,提高聚类效果。

在旅游 UGC 文本数据挖掘中,情感分析能够解析用户的情感倾向,但在面对多种情感对象以及多种复杂情感时则存在细粒度挑战;主题抽取有助于辨识旅游焦点,但可能产生模糊主题或在捕捉上下文语义信息方面遇到困难;文本分类有利于数据整理,但要求大量标记数据和特定领域的数据集;文本聚类虽能发现潜在关联,挖掘旅游热点和突发事件,但却高度依赖特征选择和表示方法。因此,在实际应用中,为

更准确地分析和挖掘旅游 UGC 文本数据,需综合考虑所用技术的特点与局限性,从而给旅游业提供更优质的决策与支持。

2 旅游 UGC 照片数据挖掘

旅游照片不仅是感知目的地形象的重要媒介,还是传达目的地感知的第一手数据,既包含了对旅游目的地的客观描述,又反映了游客的内心世界[4],是了解游客旅游行为和旅游感受的主要信源之一。与文本数据相比,图像照片数据可以更快速地传达游客的旅游感知。如何通过旅游照片研究游客旅游行为的内在规律,挖掘照片的潜在价值,进而帮助旅游服务提供商改进营销策略、提高服务质量,逐渐成为研究旅游 UGC 照片数据的重点。UGC 照片数据挖掘的基本流程见图 2,可针对旅游照片中的元数据和视觉内容设计不同的数据挖掘方法。

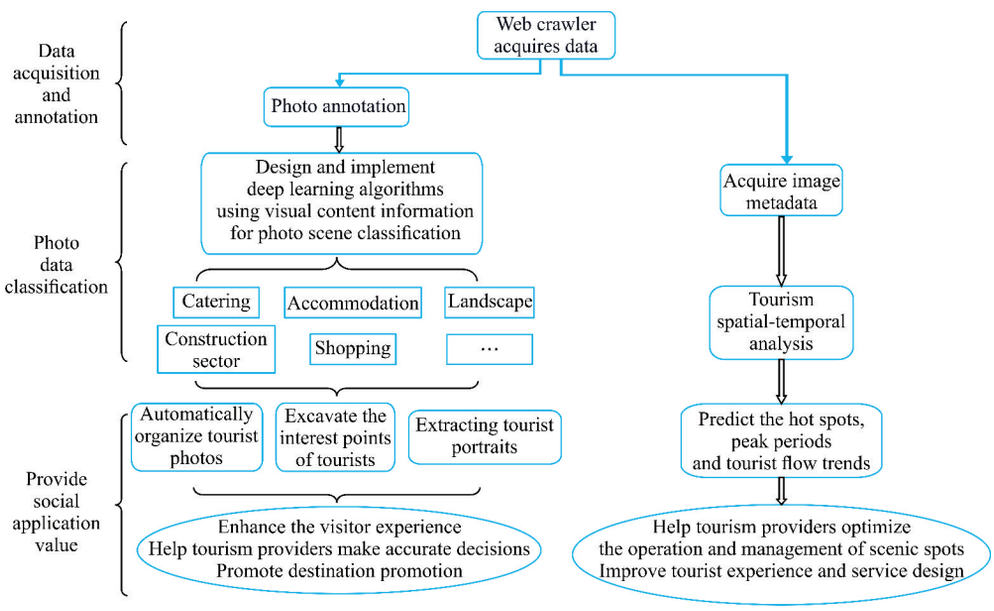

图 2 UGC 照片数据挖掘流程

Fig. 2 Flowchart of UGC photo data mining

元数据信息和视觉内容信息共同构成了一张照片的数据。元数据是指与照片相关的附加信息,包括拍摄时间、地理位置、设备信息、拍摄者信息等。元数据蕴含的如拍摄时间和地理位置等时空信息可揭示旅游业的季节性变化、热门目的地以及旅游群体偏好等规律。视觉内容信息是对游客在旅游过程中所见的直接反映,准确地析取旅游照片的内容,可以在很大程度上帮助旅游服务提供商或研究者掌握游客兴趣点、目的地形象、游客行为、旅游趋势等[44]。例如,根据旅游照片是否展示自然景观或城市景观,可以帮助研究人员了解游客对大自然和城市环境的偏好,从而更好地规划城市和自然资源的开发和保护。

2.1 旅游 UGC 照片数据挖掘的应用

2.1.1 时空分析

早期针对旅游照片挖掘的研究侧重于使用照片的地理标记信息进行空间特征分析[45,46]或时空分析[47]。时空分析通过对照片的时间和空间变化进行分析,可以直观地反映出游客活动规律和目的地选择偏好等信息。

Giglio 等[48]利用 Flickr 图片分享网站提供的地理照片来研究人类流动性和旅游景点的关系,收集、分析与 6 个意大利城市相关的 26 392 张图片的样本,并使用聚类分析自动挖掘兴趣点聚簇。Kisilevich 等[49]通过来自 Flickr 和 Panoramio 等网站的数百万张带地理标记的照片来分析游客的行为轨迹、景点热度以及游客的兴趣点。Li 等[50]以华山景区为例,利用社交平台照片中自带的时空标签分析游客的兴趣地区,进而划分华山的功能区域,并进一步揭示游客行为的地理特征。Ma 等[51]为探究特殊兴趣旅游与一般大众旅游之间的关系,对 2017 年美国 37 652 位日食爱好者在 Instagram 上分享的 41 747 张带地理标记的照片进行聚类分析,获得了游客日食观测位置的空间分布,并通过分析日食前后 3 个月内日食爱好者拍摄照片的空间和时间,确定了存在机会型和狂热型两种类型的游客,并发现两类游客的旅游活动频率因旅游距离、收入和受教育程度而异。

旅游照片中的元数据信息为旅游时空分析提供了新的数据和研究方向。旅游中游客地理位置不断发生变化,游客产生的行为数据本质是序列数据。如何利用循环神经网络来深度分析这些海量照片中的旅游位置数据是一个充满挑战的问题。

2.1.2 场景分类

深度学习在图像分析领域得到了广泛应用并取得了巨大成功[52-54]。2012 年,基于卷积神经网络(Convolutional Neural Network,CNN)[55]设计的 AlexNet[56]模型通过参数共享实现多层网络的连接,大大降低了模型参数的规模。使用深度学习进行旅游 UGC 照片数据挖掘,可以自动提取照片数据的特征,从而无须人工进行特征选择,简化了特征处理流程。目前,深度学习已经成为旅游 UGC 照片场景分类的主要方法。

Zhang 等[57]基于 ResNet 残差神经网络[58]分析了来自欧洲、北美和亚洲的游客在香港旅游时分享的 29 081 张照片,并将这些旅游照片分为 12 个场景类别,使用了卡方检验和频率分析来比较来自 3 个不同地区的游客对香港旅游形象的感知差异,其中,卡方检验用于确定不同地区游客对各个旅游形象类别的感知是否存在显著差异,而频率分析则用于分析这些感知在不同地区的分布情况。Ding 等[59]将旅游照片数据中的视觉内容与时空属性相结合,根据旅游领域专家的意见将景观照片设置为生物景观、动物景观、自然景观等 9 种类别,将收集的 43 234 张景观照片利用 Inception v3 深度学习模型[60]进行分类,再利用照片的时间戳、纬度、经度、海拔等属性信息,分析了在不同时间尺度、保护区和海拔高度下,游客对螺小山自然保护区的具体景观偏好,为研究游客的景观偏好提供了全新视角。Zhang 等[61]利用 ResNet101 残差神经网络作为主干,将 Flickr 平台上 35 356 张来自北京的旅游照片划分成 103 个场景,对来自不同国家和地区游客的行为和感知进行了统计分析。Yim 等[14]提出用深度学习算法识别照片中包含的对象及其类别,再使用回归模型分析照片的真实性、创造性和情感属性对游客参与度的影响,并将该方法用来分析 Trover 上的 416 164 张旅游照片,实验结果验证了该方法的合理性。

当前工作大多利用深度学习技术对旅游 UGC 照片进行自动分类。然而,旅游 UGC 照片数据挖掘中的深度学习技术不仅可用于图像分类,还可用于以下 3 个领域:(1)目标检测。通过目标检测技术,能够识别并定位旅游照片中的多种特定目标,例如著名景点、建筑物等,以分析游客感兴趣的景观。(2)图像描述。通过图像描述技术可以为旅游照片自动生成对照片中的景点、人物或其他相关信息的自然语言描述,帮助游客更轻松地表达旅游感受。(3)内容生成。通过深度学习中的生成式模型,可以进行旅游照片画质修复、景区照片合成等应用,或对已有景观照片进行再创作,增强游客在旅游过程中的互动体验。

2.2 旅游 UGC 照片数据挖掘中的预训练模型使用

由于特定场景中数据的缺乏,基于大量数据训练的预训练模型对提高深度神经网络的泛化能力具有非常重要的价值。Places365 数据集[62]是麻省理工大学团队制作的图像类数据集,拥有包括机场、阁楼、礼堂等 365 个场景类别共 180 万张照片。每个类别照片的验证集、测试集规模分别为 50 和 900,剩余照片为训练集。由于其多样性与权威性,Places365 数据集被广泛用于场景识别类任务的预训练和评估。Places365-CNN[62]是一个基于 Places365 的预训练模型集合,包括 AlexNet、ResNet 和 DenseNet161[63]等神经网络结构。Figueredo 等[64]提出一种能够基于照片检测游客旅游偏好的方法,该方法利用 Places365 数据集中与旅游相关的 25 个场景的照片训练一个基于卷积神经网络的特征提取模型,该模型提取到旅游照片的特征后,将得到的特征依次输入 25 个逻辑回归分类器,再输出照片所属的场景类别,之后再根据场景类别,使用模糊分类方法判断游客的偏好。Xiao 等[65]利用 Places365-CNN 模型将 531 629 张江西旅游照片分为山、湖、园等 365 个场景,之后应用 LDA 模型,根据预设的真实建筑和村庄、山、城市、水和乡村等 5 个主题的照片建立模型,将每张照片的主题按照概率分布形式给出,实现了对旅游照片的主题分类。Bui 等[66]针对旅游照片中存在的独特类别,将 Places365 数据集进行扩充,使用 ResNet18 卷积神经网络训练得到一个新模型以适应新的场景识别,研究共选取 491 705 张照片作为训练集,训练集总共包括 49 个类别,每类照片最少 5 000 张,最多 15 000 张,每个类别的验证集包含 100 个图像,每个类别的测试集包含 200 个图像。

Kim 等[67]从 TripAdvisor 上采集游客发布的照片,再使用由 Places365 数据集预训练得到的 VGG-16[68]网络,对采集到的旅游照片分别用 512 维特征向量表示,然后使用 t 分布随机邻域嵌入法(t-SNE,t-Distributed Stochastic Neighbor Embedding)将其

简化为二维向量,最后使用 HDBSCAN 算法[69]来识别嵌入空间中的簇并将其作为一个类别,从而实现照片分类。Feng 等[70]针对多标签旅游照片数据的大规模检索,使用基于 Places365 数据集预训练的 VGG-16 网络学习图像特征,并通过改变模型的最后一层,使输入照片图像获取具有区分性的二进制哈希码,从而实现旅游图像的检索。Kang 等[71]自 Flickr 社交平台爬取了 168 216 张首尔的照片,首先使用 ImageNet 数据集对 Inception-v3 网络进行预训练,同时对网络权重进行初始化,然后在人工标记的包括 30 000 张 Flickr 社交平台照片的数据集进行权重微调,以使网络能够适应包含 75 个场景以及 13 个类别的新任务,实验结果表明了模型的优越性。

DeepSentiBank[72]是一种在 ILSVRC2012[73]数据集上进行训练得到的卷积神经网络模型,可以从图像中自动提取形容词名词对(Adjective Noun Pairs,ANPs)。作为一个视觉情感分类模型,DeepSentiBank 可对图像中所蕴含的情感类别进行统计,因此被旅游研究者当作视觉分析工具使用[74-76]。Deng 等[76]发现超过 50%的照片包含游客的面部信息,基于此提出一种利用人脸信息及照片内容识别游客的旅游模式和偏好的技术,该技术利用 DeepSentiBank 提取照片内容和情感关键词,以分析游客的情感,再根据年龄和性别对游客进行分组,发现分组游客在拍摄的兴趣点和背景方面表现出不同偏好。Huang 等[77]利用 DeepSentiBank 提取旅游照片的 ANPs,以实现对旅游目的地地域特征和认知形象的客观表达。

尽管预训练模型的使用提高了模型对旅游场景的理解能力,但是预训练模型很容易发生数据集偏移(Dataset shift)或领域偏移(Domain shift)现象。这主要是由于旅游照片与普通图像相比具有鲜明的旅游特色,包括更多的美景、美食、人物等,且普遍带有强烈的情感。自监督模型的运用可以在一定程度上缓解这一现象。利用图像中的自监督训练技巧,如图像补全和修复、图像颜色化等方式,可在没有人工标注的情况下学习到足够泛化的特征和知识,从而提高模型对旅游场景的理解能力。

3 旅游 UGC 多模态数据挖掘

人们可以用眼睛看到物体、用耳朵听到声音、用手感觉纹理、用鼻子闻到气味,对一个缤纷世界的体会往往需要通过多种感官共同完成[78]。现在,5G 网络的覆盖为游客采用多种形式表达自己的旅游感受提供了技术支撑,游客也更愿意通过文字加照片或视频的形式直观地分享自己的旅游体验,表达自己的情感与观点,图文并茂的旅游游记或评论更加生动。因此,UGC 数据通常以多模态形式呈现,这有助于保证信息的完整性。UGC 多模态数据挖掘的基本流程如图 3 所示,相比单模态数据,多模态数据在进行数据挖掘时要考虑模态对齐和模态融合的问题。

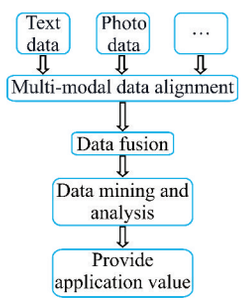

图 3 UGC 多模态数据挖掘流程

Fig. 3 Flowchart of UGC multimodal data mining

Ma 等[79]首次将多模态机器学习技术引入旅游 UGC 数据挖掘领域,利用深度神经网络来分析游客提供的照片对在线酒店评论感知的影响程度,首先将每个用户评论中的文本和图像进行预处理,对文本数据使用如分词(Tokenization)以及词干提取(Stemming)等标准自然语言处理技术来提取文本序列的特征,对图像数据则使用预先训练好的 ResNet 模型来提取相应的图像特征,然后利用 LSTM 模块融合整个文本和图像的特征,实验结果表明深度学习模型在多模态场景下比决策树、支持向量机、逻辑回归这 3 种模型的性能更佳。邓宁等[80]利用 ResNet101 模型和 COCO 数据集进行训练,选择卷积神经网络作为图片编码器,提取视频图像特征,并将提取结果传入 LSTM 网络处理再输出照片的文本描述。从生成的文本描述中进一步分析用户生成视频同目的地营销组织发布视频之间的联系,以分析旅游目的地形象的差异。

在旅游评论中存在讽刺性评论。反讽作为一种隐性而间接的情感表达方式,在社交评论中被广泛使用,正确识别用户的反讽情感对旅游服务提供商具有重要意义[81]。张继东等[81]基于多模态在线旅游评论数据,分别对文本、表情符号和图片使用 BiLSTM[82]、emoji2vec 和 FCNN 模型抽取特征,再将提取到的不同模态数据的特征进行特征融合,并在此基础上进行反讽识别,与单模态识别模型相比,多模态融合算法具有更高的准确率。刘洋等[83]构建了多模态的深度学习模型,利用 BERT、BiLSTM、ResNet

对文本和图像进行编码,通过图神经网络提取文本与图片中的交互信息,利用注意力机制强化多模态特征,最后进行反讽识别。

旅游评论数据的真实性使得潜在游客更喜欢在旅游网站上查看旅游评论,然后再选择目的地计划出游,因此利用旅游评论中的多模态数据设计旅游推荐系统有着较大的经济价值。Shao 等[84]为了挖掘在线旅游网站中多模态数据的潜在语义,提出一个综合考虑游客照片、评论以及情感 3 种类型数据的多模态主题模型(SMTM),用于发现用户喜好和目的地感知之间的关系。在特征提取方面,SMTM 分别利用 SIFT-Bow[85]、Part-of-Speech 软件[86]以及 SentiWordNet 算法[87]提取照片、评论文本以及文本情感的特征,并基于此构建了旅游推荐框架。

旅游评论数据的有用性判断对过滤无效评论至关重要,Li[88]基于旅游产品在线评论中的多模态数据(文字和图片评论)构建评论有用性分类模型,文章利用 BERT 进行文本词嵌入得到文本向量,利用 SqueezeNet[89]进行图像识别,得到排名前 5 的标签词和它们对应的词向量。处理后的文本和图像向量具有相同的维度,因此使用向量相加的方法进行特征融合。最后,通过实验比较了 BiLSTM、TextCNN[90]、支持向量机和逻辑回归等算法在单模态和多模态数据下的准确率,结果表明与仅包含文本或图片的单一模态评论相比,结合文本和图片的多模态评论能更好地预测在线评论的有用性。

表 1 分析了不同旅游 UGC 多模态数据挖掘方法中的模态特征融合方式与具体使用的模型。每种挖掘方法都有各自的优缺点,应用中应根据数据特点、任务需求以及模型设计选择合适的特征融合方式。

表 1　旅游 UGC 多模态数据挖掘方法技术对比

Table 1　Comparison of tourism UGC multimodal data mining methods and techniques

数据类型 Data type	模型 Model	特征融合方法 Feature fusion method	任务 Task	文献 Reference
Text,photo	LSTM,ResNet152	Concatenation fusion	Binary classification	[79]
Text,video	ResNet101,LSTM	No feature fusion	Generative tasks	[80]
Text,emoticon,photo	BiLSTM,emoji2vec,VGG-16	Addition fusion	Binary classification	[81]
Text,photo	BERT,BiLSTM,ResNet50	Graph model/fully connected layer	Binary classification	[83]
Photo,text,emotion	SIFT-Bow,Part-of-Speech,SentiWordNet	Not mentioned	Recommendation system	[84]
Text,photo	BERT,SqueezeNet	Addition fusion	Binary classification	[88]

旅游 UGC 多模态数据挖掘是一个复杂而具有潜力的新研究领域。尽管已经在过去几年中取得了一些进展,但仍面临不少的瓶颈和局限。多模态数据通常以不同的格式存在,存在语义鸿沟的问题。现有旅游数据挖掘研究对多模态数据的使用往往通过简单的模态对齐方式进行,无法充分捕捉和保留每个模态中的重要信息,导致信息的部分丢失。现有旅游 UGC 多模态数据挖掘的应用多以判别式模型为主导,缺乏生成式模型的使用。通过使用具有强模态对齐性的数据来训练生成模型,可以利用游客拍摄的照片自动生成与游客情感相符的游记文本,或者使用游客提供的游记描述来增强、重构或生成照片图像,进一步提升游客的旅游体验。旅游 UGC 多模态数据挖掘的发展离不开大数据的支持,若能构建一个可信、公正、广泛使用的多模态旅游 UGC 数据集,必将有力地推动旅游 UGC 多模态数据挖掘的发展。未来,随着深度学习、自然语言处理和计算机视觉等技术的进一步发展,旅游 UCG 多模态数据挖掘将会有更多的创新和研究视角。

4　展望

本文从数据挖掘算法角度,以旅游 UGC 数据中的文本、照片及多模态数据这 3 类数据为线索,综述了国内外旅游 UGC 数据挖掘的相关工作,阐明了这 3 类数据的潜在挖掘价值。未来针对旅游 UGC 数据挖掘的研究工作包括:(1)高额的数据标注成本使得当前的旅游 UGC 数据挖掘研究更倾向于利用小模型解决单一方面的问题。这使得当前模型的针对性过强,缺乏通用性。因此,研究针对旅游行业的大模型,或者是利用现有的如 ChatGPT 等的大模型挖掘

旅游UGC数据价值是非常值得探索的方向。(2)目前,针对文本、旅游照片的研究一般是将文本和照片整体当作一个样本,而少有更细粒度层面的工作。因此,在进行文本情感分类时,进一步获取游客情感针对的对象;在分析旅游照片时,进一步分析照片中人、物之间相互关系等将是未来非常有趣的选题。(3)旅游UGC多模态数据中存在着天然的自监督信息,比如一张照片附近的文本与照片存在很强的语义相关。怎样利用好多模态数据中模态间的关系,进行自监督多模态数据挖掘也是未来旅游UGC数据挖掘的新方向。(4)多模态旅游UGC数据挖掘中还存在诸如模态对齐、用户对齐等困难,影响模型的准确率。(5)自动分析游客投诉中的图文及视频数据也是多模态旅游UGC数据挖掘中的一个热点问题,图文问答作为一种根据图像信息进行文本回答的技术也将备受旅游产业关注。

参考文献

[1] MARINE-ROIG E,CLAVE S A. A method for analysing large-scale UGC data for tourism:application to the case of Catalonia [C]//Information and Communication Technologies in Tourism 2015. Lugano:Springer,Cham,2015:3-17.

[2] CHEUNG M L,LEUNG W K,CHEAH J H,et al. Exploring the effectiveness of emotional and rational user-generated contents in digital tourism platforms [J]. Journal of Vacation Marketing,2022,28(2):152-170.

[3] ZHANG Y,GAO J,COLE S,et al. How the spread of user-generated contents (UGC) shapes international tourism distribution:using agent-based modeling to inform strategic UGC marketing [J]. Journal of Travel Research,2021,60(7):1469-1491.

[4] BURGESS S,SELLITTO C,COX C,et al. User-generated content (UGC) in tourism:benefits and concerns of online consumers [C]//European Conference on Information Systems(ECIS). Verona:AIS,2009:439.

[5] DOS SANTOS M L B. The "so-called" UGC:an updated definition of user-generated content in the age of social media [J]. Online Information Review,2022,46(1):95-113.

[6] LU Q L. Research on the impact of online travel platform UGC on user usage depth [J]. Frontiers in Economics and Management,2021,2(7):209-213.

[7] LI C L,CAO M Q,WEN X L,et al. MDIVis:visual analytics of multiple destination images on tourism user generated content [J]. Visual Informatics,2022,6(3):1-10.

[8] 刘逸,李广涵,李晓娟. 基于UGC评论和TSE模型的我国游客爱国情感研究[J]. 旅游导刊,2021,5(4):79-96.

[9] WANG Y,HUANG W D,YAO X K. Research on the evaluation of tourism destination image based on user generated content [C]//ICAIIS 2021:2021 2nd International Conference on Artificial Intelligence and Information Systems. Chongqing:ACM,2021:261.

[10] LI J,XU L,TANG L,et al. Big data in tourism research:a literature review [J]. Tourism Management,2018,68:301-323.

[11] GIRSHICK R. Fast R-CNN [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago:IEEE,2015:1440-1448.

[12] ZHANG Y,WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing. Taipei:Asian Federation of Natural Language Processing,2017:253-263.

[13] LIANG F,PAN Y,GU M L,et al. Cultural tourism resource perceptions:analyses based on tourists' online travel notes [J]. Sustainability,2021,13(2):519.

[14] YIM D,MALEFYT T,KHUNTIA J. Is a picture worth a thousand views? Measuring the effects of travel photos on user engagement using deep learning algorithms [J]. Electronic Markets,2021,31(3):619-637.

[15] 文益民,史一帆,蔡国永,等. 个性化旅游推荐研究综述[C]//2015中国旅游科学年会论文集. 北京:[出版者不详],2015:2-13.

[16] LI Q,LI S B,ZHANG S,et al. A review of text corpus-based tourism big data mining [J]. Applied Sciences,2019,9(16):3300.

[17] SAMARA D,MAGNISALIS I,PERISTERAS V. Artificial intelligence and big data in tourism:a systematic literature review [J]. Journal of Hospitality and Tourism Technology,2020,11(2):343-367.

[18] LYU J Y,KHAN A,BIBI S,et al. Big data in action:an overview of big data studies in tourism and hospitality literature [J]. Journal of Hospitality and Tourism Management,2022,51:346-360.

[19] YE Q,LAW R,GU B,et al. The influence of user-generated content on traveler behavior:an empirical investigation on the effects of e-word-of-mouth to hotel online bookings [J]. Computers in Human Behavior,2011,27(2):634-639.

[20] 刘逸,保继刚,朱毅玲. 基于大数据的旅游目的地情感

评价方法探究[J]. 地理研究,2017,36(6):1091-1105.

[21] LI W,GUO K,SHI Y,et al. DWWP:domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain [J]. Knowledge-Based Systems,2018,146:203-214.

[22] 张旭辉,张郴,李雅南,等. 城市旅游餐饮体验的注意力机制模型建构:基于机器学习的网络文本深度挖掘[J]. 南京师大学报(自然科学版),2022,45(1):32-39.

[23] 刘文远,郭智存,郭丁丁. 旅游场景下的基于深度学习的文本方面级细粒度情感分类[J]. 高技术通讯,2022,32(1):22-32.

[24] LI Q,LI S B,HU J,et al. Tourism review sentiment classification using a bidirectional recurrent neural network with an attention mechanism and topic-enriched word vectors [J]. Sustainability,2018,10(9):3313.

[25] LUO Y Y,HE J J,MOU Y,et al. Exploring China's 5A global geoparks through online tourism reviews:a mining model based on machine learning approach [J]. Tourism Management Perspectives,2021,37:100769.

[26] LI N,YANG X Y,WONG I A,et al. Automating tourism online reviews:a neural network based aspect-oriented sentiment classification [J]. Journal of Hospitality and Tourism Technology,2023,14(1):1-20.

[27] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies (NAACL-HLT 2019). Minneapolis:ACL,2019:4171-4186.

[28] HOCHREITER S,SCHMIDHUBER J. Long short-term memory [J]. Neural Computation,1997,9(8):1735-1780.

[29] 周纯洁,黎嶷,杨晓宇. 基于交互式叠加注意力网络的实体属性情感分类[J]. 计算机应用与软件,2022,39(2):194-200.

[30] HU N,ZHANG T,GAO B J,et al. What do hotel customers complain about? Text analysis using structural topic model [J]. Tourism Management,2019,72:417-426.

[31] LOUREIRO S M C,GUERREIRO J,HAN H. Past,present,and future of pro-environmental behavior in tourism and hospitality:a text-mining approach [J]. Journal of Sustainable Tourism,2022,30(1):258-278.

[32] GUO Y,BARNES S J,JIA Q. Mining meaning from online ratings and reviews:tourist satisfaction analysis using latent Dirichlet allocation [J]. Tourism Management,2017,59:467-483.

[33] BLEI D M,NG A Y,JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research,2003,3:993-1022.

[34] LEE H,KANG Y. Mining tourists' destinations and preferences through LSTM-based text classification and spatial clustering using Flickr data [J]. Spatial Information Research,2021,29(6):825-839.

[35] 池云仙. 基于粒计算的旅游文本分类与热度挖掘方法[D]. 石家庄:河北师范大学,2020.

[36] 王祥翔,方荟,陈崇成. 基于朴素贝叶斯的文化旅游文本分类技术研究[J]. 福州大学学报(自然科学版),2018,46(5):644-649.

[37] 马喆康,迪力亚尔·帕尔哈提,早克热·卡德尔,等. 一种集成深度学习模型的旅游问句文本分类算法[J]. 计算机工程,2020,46(11):70-76.

[38] CHANG Y C,KU C H,CHEN C H. Using deep learning and visual analytics to explore hotel reviews and responses [J]. Tourism Management,2020,80:104129.

[39] 李娟,褚玉杰,赵振斌,等. 基于共现聚类分析的西藏入境旅游热点研究[J]. 旅游学刊,2015,30(3):35-43.

[40] 丁晟春,龚思兰,李红梅. 基于突发主题词和凝聚式层次聚类的微博突发事件检测研究[J]. 现代图书情报技术,2016(S1):12-20.

[41] SOHRABI B,VANANI I R,NASIRI N,et al. A predictive model of tourist destinations based on tourists' comments and interests using text analytics [J]. Tourism Management Perspectives,2020,35:100710.

[42] XU S H. Studies on the spatial subdivision and its influencing factors of domestic tourist markets in Wuhan [C]//International Conference on Intelligent Information Processing. Bucharest:ACM,2021:39-43.

[43] CELARDO L,LEZZI D F,VICHI M. Multi-mode partitioning for text clustering to reduce dimensionality and noises [C]//Proceedings of the 13th International Conference on Statistical Analysis of Textual Data. Nice:IRIS,2016:7-10.

[44] LI S,TAKAHASHI S,YAMADA K,et al. Analysis of SNS photo data taken by foreign tourists to Japan and a proposed adaptive tourism recommendation system [C]//2017 International Conference on Progress in Informatics and Computing (PIC). Nanjing:IEEE,2017:323-327.

[45] SONG X P,RICHARDS D R,TAN P. Using social media user attributes to understand human-environment interactions at urban parks [J]. Scientific Reports,2020,10(1):808.

[46] SONTER L J,WATSON K B,WOOD S A,et al. Spatial and temporal dynamics and value of nature-based recreation,estimated via social media[J]. PLoS One,2016,11(9):e0162372.

[47] SCHIRPKE U,MEISCH C,MARSONER T,et al. Revealing spatial and temporal patterns of outdoor recreation in the European Alps and their surroundings [J]. Ecosystem Services,2018,31(Pt C):336-350.

[48] GIGLIO S,BERTACCHINI F,BILOTTA E,et al. Machine learning and points of interest:typical tourist Italian cities [J]. Current Issues in Tourism,2020,23(13):1646-1658.

[49] KISILEVICH S,KRSTAJIC M,KEIM D A,et al. Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections [C]//2010 14th Internatial Conference Information Visualization. London:IEEE,2010:289-296.

[50] LI C X,GAO L F,LIU Y,Z et al. Cluster analysis of China's inbound tourism market:a new multi-attribute approach based on association rule mining of tourist preferences at scenic spots [J]. Asia Pacific Journal of Tourism Research,2021,26(6):654-667.

[51] MA S H,KIRILENKO A P,STEPCHENKOVA S. Special interest tourism is not so special after all:big data evidence from the 2017 great American solar eclipse [J]. Tourism Management,2020,77:104021.

[52] LIU W,ANGUELOV D,ERHAN D,et al. SSD:single shot multibox detector [C]//European Conference on Computer Vision (ECCV 2016). Amsterdam:Springer,2016:21-37.

[53] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified,real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas:IEEE,2016:779-788.

[54] REDMON J,FARHADI A. YOLO9000:better,faster,stronger [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu:IEEE,2017:7263-7271.

[55] FUKUSHIMA K,MIYAKE S. Neocognitron:a self-organizing neural network model for a mechanism of visual pattern recognition [C]//Competition and Cooperation in Neural Nets. Berlin:Springer,1982:267-285.

[56] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM,2017,60(6):84-90.

[57] ZHANG K,CHEN D Z,LI C L. How are tourists different? Reading geo-tagged photos through a deep learning model [J]. Journal of Quality Assurance in Hospitality & Tourism,2020,21(2):234-243.

[58] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas:IEEE,2016:770-778.

[59] DING Y,BAI Z,XIA H,et al. Tourists' landscape preferences of Luoxiao Mountain national forest trail based on deep learning [J]. Wireless Communications and Mobile Computing,2022,2022:4662818.

[60] SZEGEDY C,VANHOUCKE V,IOFFE S,et al. Rethinking the inception architecture for computer vision [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas:IEEE,2016:2818-2826.

[61] ZHANG K,CHEN Y,LI C. Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model:the case of Beijing [J]. Tourism Management,2019,75:595-608.

[62] ZHOU B L,LAPEDRIZA A,KHOSLA A,et al. Places:a 10 million image database for scene recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(6):1452-1464.

[63] HUANG G,LIU Z,VAN DER MAATEN L,et al. Densely connected convolutional networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu:IEEE,2017:2261-2269.

[64] FIGUEREDO M,CACHO N,THOME A,et al. Using social media photos to identify tourism preferences in smart tourism destination [C]//2017 IEEE International Conference on Big Data (Big Data). Boston:IEEE,2017:4068-4073.

[65] XIAO X,FANG C Y,LIN H. Characterizing tourism destination image using photos' visual content [J]. ISPRS International Journal of Geo-Information,2020,9(12):730.

[66] BUI V,ALAEI A R,VU H Q,et al. Revisiting tourism destination image:a holistic measurement framework using big data [J]. Journal of Travel Research,2022,61(6):1287-1307.

[67] KIM J,KANG Y. Automatic classification of photos by tourist attractions using deep learning model and image feature vector clustering [J]. ISPRS International Journal of Geo-Information,2022,11(4):245.

[68] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10)[2023-12-25]http://arxiv.org/abs/1409.1556.pdf

[69] MCINNES L,HEALY J,ASTELS S. HDBSCAN:hierarchical density based clustering [J]. The Journal of Open Source Software,2017,2(11):205.

[70] FENG J,SUN W. Improved deep hashing with scalable interblock for tourist image retrieval [J]. Scientific Programming,2021,2021:9937061.

[71] KANG Y,CHO N,YOON J,et al. Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos [J]. ISPRS International Journal of Geo-Information,2021,10(3):137.

[72] CHEN T,BORTH D,DARRELL T,et al. DeepSentiBank:visual sentiment concept classification with deep convolutional neural networks [EB/OL]. (2014-10-30)[2023-12-25]:http://arxiv.org/abs/1410.8586.pdf.

[73] RUSSAKOVSKY O,DENG J,SU H,et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision,2015,115(3):211-252.

[74] HE Z,DENG N,LI X,et al. How to "read" a destination from images? Machine learning and network methods for DMOs' image projection and photo evaluation [J]. Journal of Travel Research,2022,61(3):597-619.

[75] DENG N,LIU J Y,DAI Y,et al. Different cultures,different photos:a comparison of Shanghai's pictorial destination image between East and West [J]. Tourism Management Perspectives,2019,30:182-192.

[76] DENG N,LIU J Y. Where did you take those photos? Tourists' preference clustering based on facial and background recognition [J]. Journal of Destination Marketing & Management,2021,21:100632.

[77] HUANG X Y,HAN Y J,MENG Q L,et al. Do the DMO and the tourists deliver the similar image? Research on representation of the health destination image based on UGC and the theory of discourse power:a case study of Bama,China [J]. Sustainability,2022,14(2):953.

[78] BALTRUŠAITIS T,AHUJA C,MORENCY L P. Multimodal machine learning:a survey and taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,41(2):423-443.

[79] MA Y F,XIANG Z,DU Q Z,et al. Effects of user-provided photos on hotel review helpfulness:an analytical approach with deep leaning [J]. International Journal of Hospitality Management,2018,71:120-131.

[80] 邓宁,蘧浪浪. 基于视频机器分析的目的地形象差异对比:以北京 YouTube 视频为例[J]. 旅游学刊,2022,37(8):70-85.

[81] 张继东,蒋丽萍. 基于多模态深度学习的旅游评论反讽识别研究[J]. 情报理论与实践,2022,45(7):158-164.

[82] HUANG Z H,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. (2015-08-09)[2023-12-25]:http://arxiv.org/abs/1508.01991.pdf.

[83] 刘洋,马莉莉,张雯,等. 基于跨模态深度学习的旅游评论反讽识别[J]. 数据分析与知识发现,2022,6(12):23-31.

[84] SHAO X,TANG G,BAO B K. Personalized travel recommendation based on sentiment-aware multimodal topic model [J]. IEEE Access,2019,7:113043-113052.

[85] LOWE D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision,2004,60(2):91-110.

[86] Stanford NLP group. Software [EB/OL]. [2023-12-25]. http://nlp.stanford.edu/software/index.shtml.

[87] BACCIANELLA S,ESULI A,SEBASTIANI F. SENTIWORDNET 3.0:an enhanced lexical resource for sentiment analysis and opinion mining [C]//Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC). Valletta:ELRA,2010:2200-2204.

[88] LI M. Research on extraction of useful tourism online reviews based on multimodal feature fusion [J]. ACM Transactions on Asian and Low-resource Language Information Processing,2021,20(5):82.

[89] MAI S J,HU H F,XING S L. Modality to modality translation:an adversarial representation learning and graph fusion network for multimodal fusion [C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York:AAAI,2020,34(1):164-172.

[90] KIM Y. Convolutional neural networks for sentence classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha:ACL,2014:1746-1751.

Research Progress in Tourism Data Mining Based on UGC Data

FENG Yate[1,2], XU Zhengli[3**], WEN Yimin[1,4]

(1. Guangxi Key Laboratory of Culture and Tourism Smart Technology, Guilin Tourism University, Guilin, Guangxi, 541006, China; 2. School of Tourism Data, Guilin Tourism University, Guilin, Guangxi, 541006, China; 3. School of Business, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China; 4. School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China)

Abstract: With the vigorous development of the Internet and social media, User Generated Content (UGC) data has gradually become an important part of tourism big data. As an important data reflecting tourist behavior, UGC data has the characteristics of rich types, strong data authenticity, and large data noise. This article reviews the development of tourism UGC data research in the past few years, and summarizes the achievements of tourism UGC data mining research in recent years from the perspectives of text, photos, and multimodal data types, and discusses the direction of future research.

Key words: UGC; multimodality; data mining; tourism big data

责任编辑:陆　雁,陈少凡