专题 海南资源环境遥感产品数据集

基于社交媒体的海南旅游景区评价数据集

林振宇[1,2,3]，解吉波[1,2*]，覃佐淼[1,2,3]，杨腾飞[2]，赵静[2]

1. 海南省地球观测重点实验室，海南三亚 572000
2. 中国科学院遥感与数字地球研究所，数字地球重点实验室，北京 100094
3. 河南理工大学，河南焦作 454000

摘要：本文从社交媒体中采集并处理了2012–2018年海南所有4A及5A级景区的评论数据构建了海南旅游景区评价数据集。本数据集旨在用于对海南旅游景区的质量评估、景区的容量管理、景区传播效果评价、景区网络舆情监测预警、景区网络口碑管理、景区形象管理、景区个性化推荐等研究。同时，结合多源化数据，本数据集可为研究海南省旅游发展提供数据支持。

关键词：海南旅游；社交媒体；景区评价；个性化推荐

数据库（集）基本信息简介

数据库（集）名称	基于社交媒体的海南旅游景区评价数据集
数据作者	林振宇，解吉波，覃佐淼，杨腾飞，赵静
数据通信作者	解吉波（xiejb@radi.ac.cn）
数据时间范围	2012年1月至2018年10月
地理区域	地理范围包括海南岛（北纬18°10′–20°10′，东经108.37°–111.03°）。
数据量	58.8 MB
数据格式	*.rar, *.sql, *.xlsx
数据服务系统网址	http://www.sciencedb.cn/dataSet/handle/714
基金项目	海南省重大科技计划项目（ZDKJ2016021）
数据库（集）组成	数据集由1个压缩包组成，主要包括5个文件夹，数据量约125 MB，压缩后数据量约58.8 MB。5个文件夹分别为美团、同程、途牛、携程、样例数据，每个文件夹下由各旅游网站的景区评论数据组成，以两种数据形式存放（*.sql, *.xlsx）。

引　言

旅游是海南省的经济支柱产业之一，对其他相关产业的发展有着较强的带动作用。研究和提高海南各景区的服务质量，满足游客多元化的旅游需求，对进一步促进海南旅游产业的发展至关重要。

随着旅游互联网的快速发展，大量和旅游景区相关的用户评论信息为旅游业的发展研究提供了有力数据支持。更多的潜在游客，会在出行前根据这些评论信息制定旅游路线，协助旅游决策[1–3]。通常，这些数据信息多以文本、图片的形式

ISSN 2096-2223
CN 11-6035/N

文献 DOI:
10.11922/csdata.2018.0094.zh
数据 DOI：
10.11922/sciencedb.714
文献分类：地球科学

收稿日期：2018-12-17
开放同评：2019-02-20
录用日期：2019-04-28
发表日期：2019-05-22

* 论文通信作者
解吉波：xiejb@radi.ac.cn

出现在各大社交媒体平台上。这些信息通常表达了游客对于相关景区的意见、建议和满意度，从而为景区质量和服务的进一步提升提供有效参考。目前，国内外已有不少学者对景区的社交媒体信息展开相关研究，并从不同方面探讨它们的应用。如文献[4]以众包的形式收集秦皇岛高校大学生对当地旅游景区的评论信息，并结合这些数据提供者的个人信息开展用户画像的旅游情境化推荐服务研究；文献[5]利用多模态的景点信息（文本、地理标记图片以及视频生成景点的信息摘要），根据用户的查询为用户个性化地推荐景点；文献[6–8]基于签到记录数据来进行旅游路线的推荐等。然而目前，可用的基于社交媒体的开放旅游景区评论数据集并不多，这严重制约了旅游信息挖掘的研究。为此，本文以海南岛为研究对象，从主流旅游网站（包括美团网、途牛网、同程网以及携程网等）上收集和处理了 2012–2018 年间所有 4A 和 5A 级旅游景区的评论数据构建了数据集。

1 数据采集和处理方法

本数据集的生产流程包括数据采集与清洗，数据管理和数据分类。数据制作流程如图 1 所示。

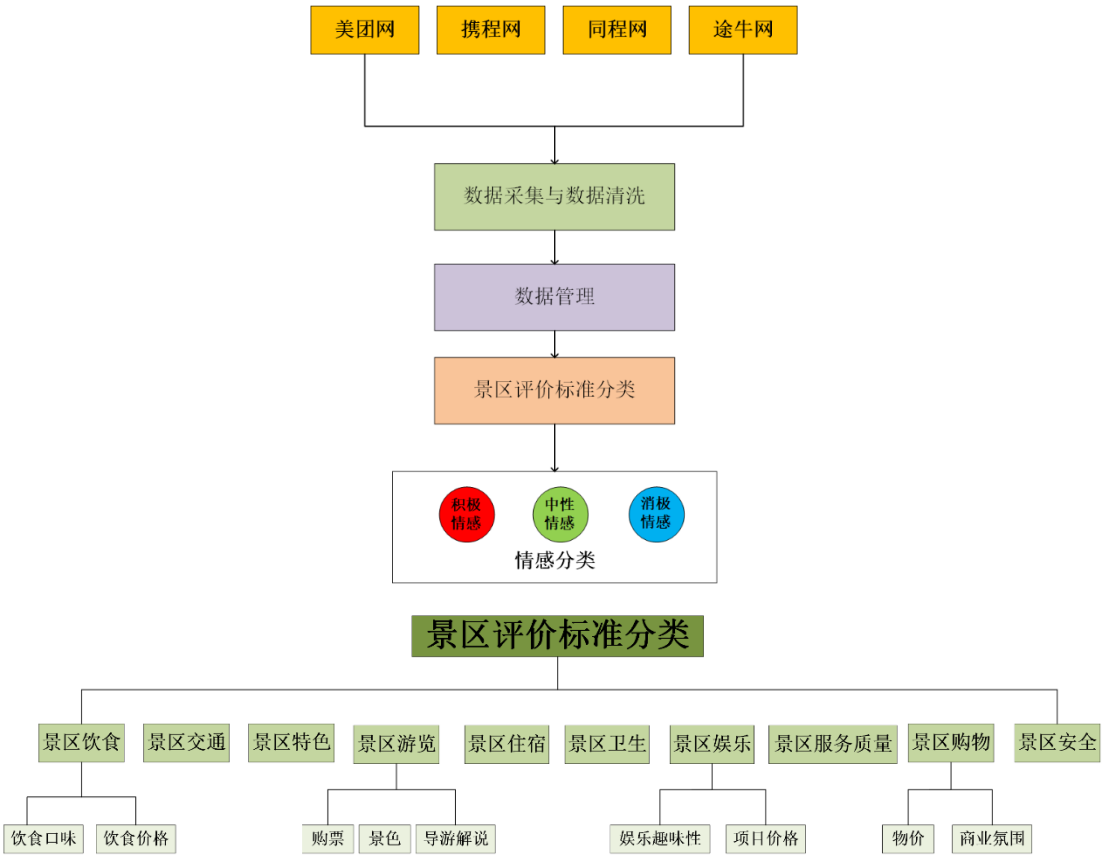

图 1　数据集制作流程图

1.1 景区评论数据的采集与清洗

该部分数据以海南岛 4A 和 5A 级景区为研究对象，将 4A 和 5A 级景区分为 4 种类型，分别为自然景区、历史人文景区、民俗风情景区、休闲度假景区。获取了 2012–2018 共 7 年的评论数据，这些数据主要来源于包括美团、携程、同程以及途牛在内的 4 个旅游网站。原始数据形式为 HTML，本文通过 Java 编程语言对其进行了解析和清洗，最终形成了 283 072 条结构化文本数据。其中，数

据清洗操作包括全半角字符的转化、繁简体文字的转化、去除同一用户的多次评论以及文本去重等。同时，为方便读者使用，该部分数据以 sql 和 xlsx 两种格式存储。如下表 1–4 显示了数据的基本结构信息，如图 2 展示了旅游景区在海南岛的分布情况。

表 1　海南岛自然景区名称及评论数据量

序号	名称	等级	美团	携程	同程	途牛	地址
1	海南分界洲岛旅游区	5A	8350	2771	966	285	陵水县
2	三亚大小洞天旅游区	5A	6321	2835	3232	651	三亚市
3	七仙岭温泉国家森林公园	4A	1756	626	133	15	保亭
4	海南热带野生动植物园	4A	11 410	1389	851	75	海口市
5	中国雷琼海口火山群世界地质公园	4A	0	1283	1007	278	海口市
6	南湾猴岛生态旅游区	4A	5130	1977	729	530	陵水县
7	天涯海角游览区	4A	30 510	2966	3146	3509	三亚市
8	亚龙湾热带天堂森林旅游区	4A	20 000	2978	5789	4017	三亚市
9	东山岭文化旅游区	4A	740	560	196	16	万宁市
10	兴隆热带植物园	4A	1773	2011	375	63	万宁市
11	三亚水稻公园	4A	0	98	66	10	三亚市
12	鹿回头风景区	4A	23 100	2846	4827	3341	三亚市

表 2　海南岛历史人文景区名称及评论数据量

序号	名称	等级	美团	携程	同程	途牛	地址
1	三亚南山文化旅游区	5A	26 633	2944	2839	2028	三亚
2	海南文笔峰盘古文化旅游区	4A	4016	250	162	17	定安县
3	博鳌亚洲论坛永久会址景区	4A	1725	1125	341	28	琼海市

表 3　海南岛民俗风情景区名称及评论数据量

序号	名称	等级	美团	携程	同程	途牛	地址
1	槟榔谷黎苗文化旅游区	5A	5095	2620	1139	260	保亭县
2	海南呀诺达雨林文化旅游区	5A	9693	2876	2362	1270	保亭县

表 4　海南岛休闲度假景区名称及评论数据量

序号	名称	等级	美团	携程	同程	途牛	地址
1	三亚蜈支洲岛度假中心	5A	22 340	2978	4168	1261	三亚市
2	海口观澜湖旅游度假区	4A	46	334	103	30	海口市
3	海口假日海滩旅游区	4A	0	1217	0	0	海口市
4	三亚大东海旅游区	4A	305	2793	0	0	三亚市
5	三亚西岛海洋文化旅游区	4A	8290	2824	1937	0	三亚市
6	亚龙湾爱立方滨海乐园	4A	785	194	168	9	三亚市

序号	名称	等级	美团	携程	同程	途牛	地址
7	清水湾旅游区	4A	0	330	0	0	陵水

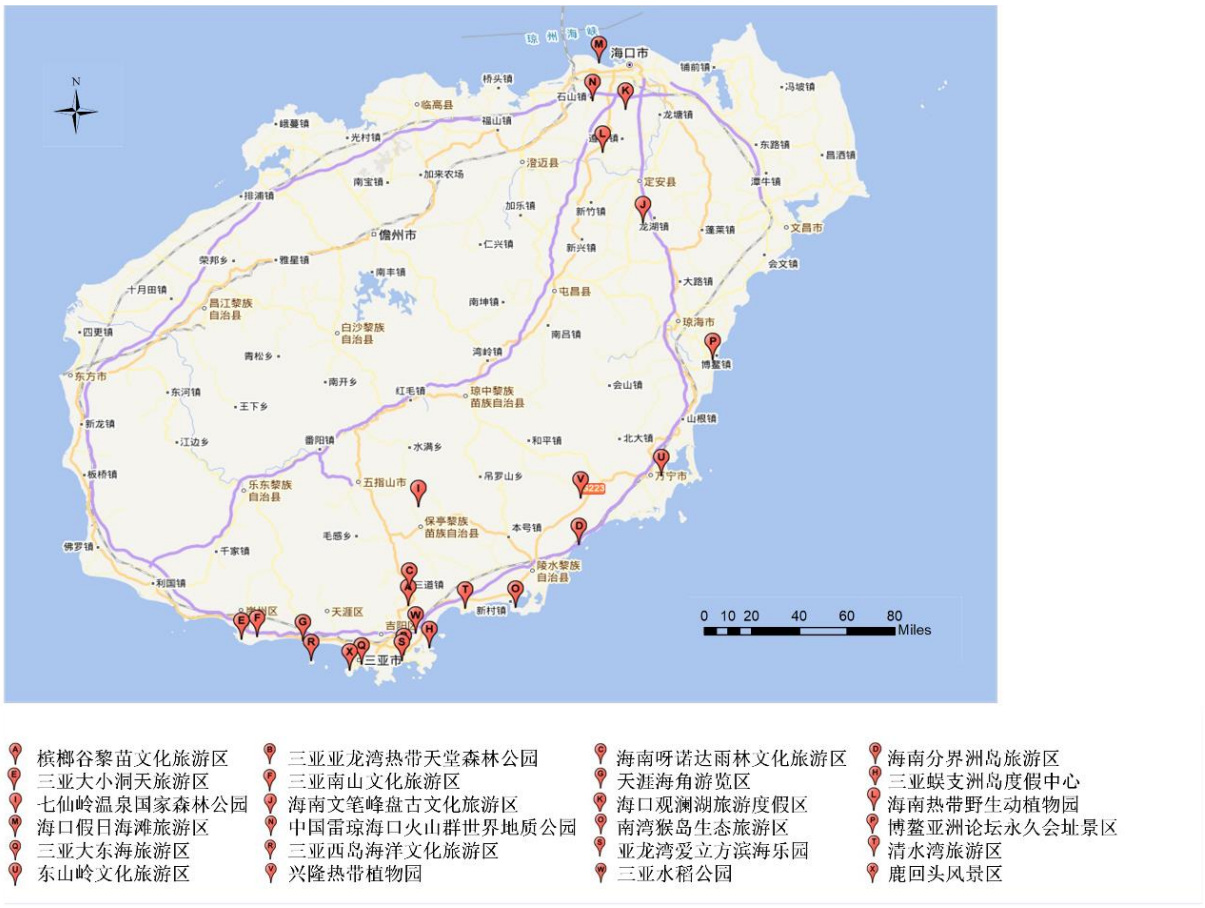

图 2　旅游景区在海南岛分布情况

1.2　数据分类

景区社交媒体评论信息蕴含着公众对于景区不同方面的评价，这对于发现和解决旅游景区存在的问题，提高游客满意度等具有重要的参考价值。为此，本数据集从多个主题对这些评论信息进行公众情感分类。

我们根据整个文本的情感倾向，将该文本分为积极情感、消极情感和中性情感 3 个类别[9]，用以从宏观上对该景区作出评价。从细粒度主题上分，我们则根据国家 A 级景区的评价指标，基于这些评价指标对该景区作出情感分类，旨在从多个主题方面刻画景区质量，以提供个性化服务需求。其中细粒度的主题指标包括景区饮食（饮食口味、饮食价格）、景区娱乐（娱乐趣味性、项目价格）、景区购物（物价、商业氛围）、景区游览（购票、景色、导游解说）、景区特色、景区卫生、景区交通、景区住宿、景区服务质量、景区安全 10 个方面。图 3 为根据国家 A 级景区评价指标的细粒度分类标准。

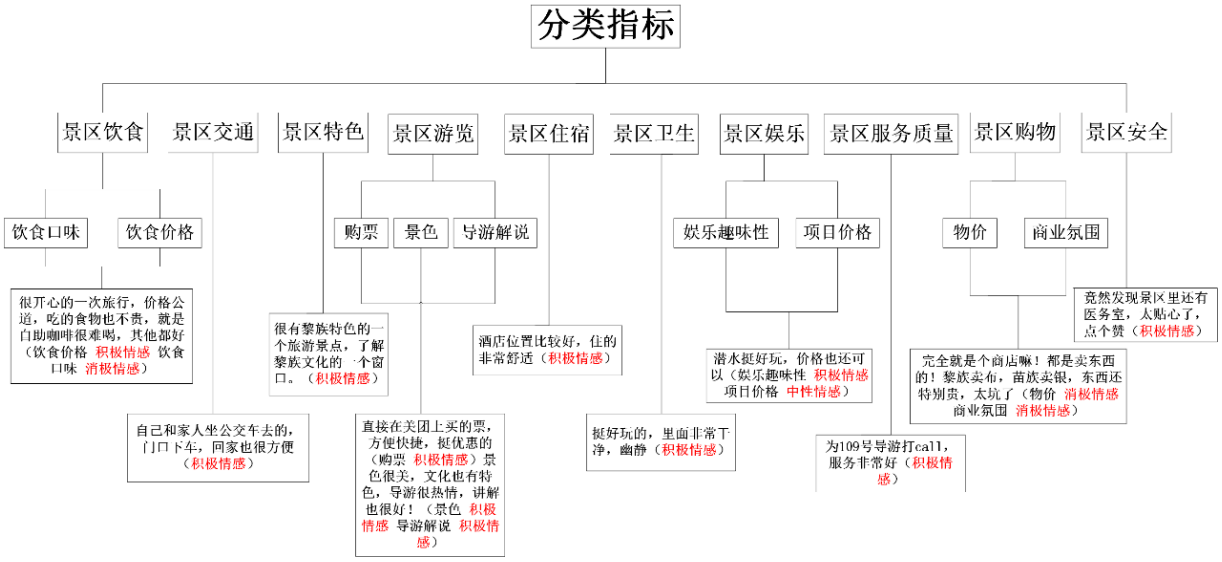

图 3　细粒度分类标准

2　数据样本描述

2.1　数据集信息

本数据集从美团、携程、同程和途牛 4 个旅游网站上收集并处理了海南岛所有 4A 和 5A 级景区的评论数据，数据的时间跨度为 7 年（2012–2018 年）。

整个数据集由 1 个压缩文件包组成，压缩文件包由 5 个文件夹组成。其中 4 个文件夹分别以上述 4 个旅游网站命名，每个文件夹下分别为 24 个景区在该旅游网站中的所有评论数据，数据储存格式包括 sql 和 xlsx，命名规则为“旅游网站+景区名称”；第 5 个文件夹以样例数据命名，该文件夹下为经过分类处理的数据，命名规则为“旅游网站+景区名称+积极/消极/中性+分类细则序号〔1 景区饮食（11 饮食口味，12 饮食价格），2 景区交通，3 景区特色，4 景区游览（41 购票，42 景色，43 导游解说），5 景区住宿，6 景区卫生，7 景区娱乐（71 娱乐趣味性，72 项目价格），8 景区服务质量，9 景区购物（91 物价，92 商业氛围），10 景区安全〕”。数据集详细信息如表 5。

表 5　海南岛旅游数据集信息

序号	文件夹名称	数据格式	文件数量和大小
1	美团	sql，xlsx	40；75.2 MB
2	携程	sql，xlsx	48；28.4 MB
3	同程	sql，xlsx	42；14.2 MB
4	途牛	sql，xlsx	40；7.00 MB
5	样例数据	sql，xlsx	80；642 KB

2.2　分类样本描述

样本集以途牛网上分界洲岛旅游区的评论数据为基础，将这些原始数据进行多个主题的情感分类，从而得到表 6 所示的结果。

表 6　示例样本情况

分类标准细则	分类标准细则	总数量（条）	时间	情感分类数量（条）
景区饮食	饮食口味	4	2012.9–2018.10	积极情感 3 中性情感 0 消极情感 1
	饮食价格	6	2012.9–2018.10	积极情感 2 中性情感 3 消极情感 1
景区交通		4	2012.9–2018.10	积极情感 3 中性情感 0 消极情感 1
景区特色		20	2012.9–2018.10	积极情感 17 中性情感 1 消极情感 2
景区游览	购票	71	2012.9–2018.10	积极情感 53 中性情感 5 消极情感 13
	景色	49	2012.9–2018.10	积极情感 43 中性情感 1 消极情感 5
	导游解说	0	2012.9–2018.10	积极情感 0 中性情感 0 消极情感 0
景区住宿		2	2012.9–2018.10	积极情感 1 中性情感 0 消极情感 1
景区卫生		7	2012.9–2018.10	积极情感 5 中性情感 0 消极情感 2
景区娱乐	娱乐趣味性	27	2012.9–2018.10	积极情感 10 中性情感 3 消极情感 14
	项目价格	9	2012.9–2018.10	积极情感 1 中性情感 1 消极情感 7
景区服务质量		18	2012.9–2018.10	积极情 9 中性感情 2 消极情感 7
景区购物	物价	0	2012.9–2018.10	积极情感 0 中性情感 0 消极情感 0
	商业氛围	2	2012.9–2018.10	积极情感 1 中性情感 0 消极情感 1
景区安全		0	2012.9–2018.10	积极情感 0 中性情感 0 消极情感 0

3　数据质量控制和评估

评论海南景区旅游质量的社交媒体平台有很多。为保障数据的丰富性，我们通过比较选出了具有代表性的 4 个旅游网站，以确保最大程度地获取相关信息。数据收集完成后，我们人工检查了数据的有效性并删除了不完整的及与海南旅游景区无关的评论数据。在分类样例中，本文所用的细粒度主题则是根据国家 A 级景区的评价指标来拟定。对于分类样例中的文本情感类别标签，我们安排了 2 个同事进行人工判读，并对结果进行复议和讨论，以确保最终分类的正确性。

4　数据使用方法和建议

本数据集包含海南岛 4A 级以上所有景区 2012–2018 年以来 283 072 条社交媒体评论数据。研究人员可通过互联网文本情感分析算法抽取公众对景区不同主题特征的态度信息，同时结合时间维度从公众观测的角度探究景区质量的变化特征，为景区网络口碑、形象管理等提供数据参考。通过互联网主题聚类算法，如 LDA（Latent Dirichlet Allocation）、K-means 聚类算法、或者简单的词频计算等语义挖掘算法从各景区海量评论信息中抽取公众关注热点，以服务于旅游景区的个性化推荐、景区发展规划等。官方发布的诸如旅游景区统计年鉴等数据，可以与本数据集作为相互验证和补充的数据，将会在景区容量管理、景区传播效果评价、景区形象管理、景区热度分析、景区质量评价分析、景区发展趋势等研究上发挥重要作用。

数据作者分工职责

林振宇（1997—），女，河南省周口市人，硕士生，研究方向为 3S 技术理论与应用。主要承担工作：数据收集与处理，论文撰写。

解吉波（1977—），男，山东省青岛市人，博士，副研究员，研究方向为地理空间数据基础设施、遥感、地理计算。主要承担工作：数据集结构设计与技术指导。

覃佐淼（1994—），男，湖南省常德市人，硕士生，研究方向为空间数据挖掘。主要承担工作：数据收集与处理，论文撰写。

杨腾飞（1988—），男，河南省洛阳市人，博士生，研究方向为自然语言处理、灾害信息挖掘。主要承担工作：数据处理，技术指导，论文修改。

赵静（1988—），女，江苏省镇江市人，博士生，研究方向为信号与信息处理、全球变化（碳排放、气候和灾害）数据挖掘和分析。主要承担工作：数据收集与检查。

参考文献

[1] FANG B, YE Q, KUCUKUSTA D, et al. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics[J]. Tourism Management, 2016, 52: 498-506.

[2] SCHUCKERT M, LIU X, LAW R. Hospitality and tourism online reviews: Recent trends and future directions[J]. Journal of Travel & Tourism Marketing, 2015, 32(5): 608-621.

[3] ZHU F, ZHANG X. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics[J]. Journal of marketing, 2010, 74(2): 133-148.

[4] 刘海鸥, 孙晶晶, 苏妍嫄, 等. 基于用户画像的旅游情境化推荐服务研究[J]. 情报理论与实践, 2018, 41(10): 87-92.

[5] WU X, LI J, ZHANG Y, et al. Personalized multimedia web summarizer for tourist[C]. Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 1025-1026.

[6] HSIEH H P, LI C T. Composing traveling paths from location-based services[C]. Sixth International AAAI Conference on Weblogs and Social Media , Toronto, Canada, 2012: 618-619.

[7] LIAN D, XIE X. Learning location naming from user check-in histories[C]. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2011: 112-121.

[8] ZHENG Y, XIE X. Learning travel recommendations from user-generated GPS traces[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(1): 2.

[9] 陆林, 朱申莲, 刘曼曼. 杭州城市旅游品牌的演化机理及优化[J]. 地理研究, 2013, 32(3):556-569.

论文引用格式

林振宇, 解吉波, 覃佐淼, 等. 基于社交媒体的海南旅游景区评价数据集[J/OL]. 中国科学数据, 2019, 4(2). (2019-04-17). DOI: 10.11922/csdata.2018.0094.zh.

数据引用格式

林振宇, 解吉波, 覃佐淼, 等. 基于社交媒体的海南旅游景区评价数据集[DB/OL]. Science Data Bank, 2018. (2018-12-17). DOI: 10.11922/sciencedb.714.

Evaluation data set for Hainan tourism scenic spots based on social media

Lin Zhenyu[1,2,3], Xie Jibo[1,2*], Qin Zuomiao[1,2,3], Yang Tengfei[2], Zhao Jing[2]

1. Key Laboratory of Earth Observation, Hainan Province, Sanya 572000, P.R. China;

2. Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, P.R. China;

3. Henan Polytechnic University, Henan Province, Jiaozuo 454000, P.R. China

*Email: xiejb@radi.ac.cn

Abstract: This paper collected and processed the review data of all 4A and 5A scenic spots in Hainan from 2012 to 2018 to construct the Hainan tourism scenic spot evaluation data set. This dataset is intended to be used for the quality assessment of Hainan tourist attractions, the capacity management of scenic spots, the evaluation of scenic spot communication effects, the monitoring and early warning of scenic spot network, the management of scenic spot network reputation, the management of scenic spot image, and the personalized recommendation of scenic spots. At the same time, combined with multi-source data, this data set can provide data support for the study of tourism development in Hainan Province.

Keywords: Hainan tourism; social media; scenic evaluation; personalized recommendation

Dataset Profile

Title	Evaluation data set for Hainan tourism scenic spots based on social media
Data corresponding author	Xie Jibo (xiejb@radi.ac.cn)
Data authors	Lin Zhenyu, Xie Jibo, Qin Zuomiao, Yang Tengfei, Zhao Jing
Time range	January 2012 – October 2018
Geographical scope	18°10′N–20°10′N, 108°37′E–111°03′E
Data volume	58.8MB
Data format	*.rar, *.sql, *.xlsx
Data service system	<http://www.sciencedb.cn/dataSet/handle/714>
Sources of funding	Major Science and Technology Program of Hainan Province (ZDKJ2016021)
Dataset composition	The dataset consists of 1 compressed package, which mainly includes 5 folders, the data volume of about 125MB, and the compressed data volume of about 58.8MB. The five folders are Meituan, Tongcheng, Tuniu, Ctrip, and sample data. Each folder is composed of scenic review data of each travel website and stored in two forms of data (*.sql, *.xlsx).