Statistics and Application 统计学与应用, 2021, 10(6), 1053-1059
Published Online December 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.106112

Hans汉斯

基于LDA主题模型的重庆市5A级景区旅游评价研究

龚乃林，赵胜利*

重庆理工大学理学院，重庆

收稿日期：2021年11月25日；录用日期：2021年12月9日；发布日期：2021年12月27日

摘　要

随着“互联网 + 旅游”模式的普及，网络中旅游信息呈爆炸式增长，其中用户评论由于数据的非结构特点及多样性，难以直接使用。本文通过挖掘用户评论对景区进行旅游评价，有利于游客和景区负责人多方面了解景点形象。本文以重庆市5家著名“AAAAA”级景区为研究对象，爬取12,571条评论数据。首先，借助LDA主题模型挖掘评论文本中的主题信息，并构建综合评价体系。然后，以游客对主题的关注程度为依据，对评价指标进行相应的赋权处理，得到评价得分。最后，对5家景区进行得分排序，并根据结果进行分析。从提取的主题信息来看，游客关注的五个主题为：平台服务、景区管理、自然景观、夏冬体验、性价比；从高到低，综合得分排序结果为：金佛山、武隆喀斯特、巫山小三峡、云阳龙岗、四面山。

关键词

LDA，综合评价体系，综合得分排序

Research on Tourism Evaluation of 5A Scenic Spots in Chongqing Based on LDA Theme Model

Nailin Gong, Shengli Zhao*

School of Science, Chongqing University of Technology, Chongqing

Received: Nov. 25th, 2021; accepted: Dec. 9th, 2021; published: Dec. 27th, 2021

*通讯作者。

文章引用: 龚乃林, 赵胜利. 基于 LDA 主题模型的重庆市 5A 级景区旅游评价研究[J]. 统计学与应用, 2021, 10(6): 1053-1059. DOI: 10.12677/sa.2021.106112

Abstract

With the popularization of the "Internet + Tourism" model, the tourism information in the network has exploded. Among them, user reviews are difficult to be directly used due to the unstructured characteristics and diversity of data. This paper evaluates the tourism of scenic spots by mining user comments, which is conducive to tourists and scenic spot leaders to understand the image of scenic spots in many aspects. This paper takes five famous "AAAAA" scenic spots in Chongqing as the research object, and crawls 12,571 comment data. Firstly, LDA topic model is used to mine the topic information in the comment text and construct a comprehensive evaluation system. Then, based on the tourists' attention to the theme, the evaluation index is weighted accordingly, and the evaluation score is obtained. Finally, the five scenic spots are ranked and analyzed according to the results. From the extracted theme information, the five themes that tourists pay attention to are: platform service, scenic spot management, natural landscape, summer and winter experience, and cost performance. From high to low, the composite score ranking results are: Jinfo Mountain, Wulongkast, Wushan Little Three Gorges, Yunyang Longgang and Simian Mountain.

Keywords

LDA, Comprehensive Evaluation System, Comprehensive Score Ranking

1. 引言

重庆位于中国内陆西南部、长江上游地区，旅游资源丰富，有长江三峡、世界文化遗产大足石刻、世界自然遗产武隆喀斯特和南川金佛山等景观[1]。在 2021 年国庆假期，重庆 A 级旅游景区累计接待游客 1273 万人次，较 2020 年增长 31.2%。重庆全市 A 级旅游景区累计接待游客 1273 万人次，按可比口径分别较 2020 年、2019 年增长 31.2%、2.6% [2]。重庆市正逐渐成为全国最热门的旅游城市之一，因此研究重庆的热门景区，对来渝游客及重庆市相关旅游部门有一定程度上的参考意义。

在“互联网 + 旅游”背景下，旅游网络评论得到迅速发展，不仅为游客的旅行决策提供了全面和丰富的信息，也为旅游景区管理者提供了真实的反馈信息。与传统的问卷调查相比，网络评论由于在不受周边干扰环境中根据自身体验提供的信息，所以更具真实性和客观性[3]。本文采用网络平台中的用户评论数据，进行文本挖掘分析。

本文具体研究流程如下：1) 文本评论采集：利用 Python 网络爬虫技术在携程网、去哪儿网两大平台上抓取游客用户评论信息。2) 中文预处理：首先，删除重复评论与字符数小于 5 的评论；其次，借助 Jieba 库进行“精确匹配模式”下的分词，对评论语料进行预处理；然后，利用 TF-IDF 剔除部分与旅游目的地无关的词汇。3) 主题评价体系构造：首先，基于困惑度指标，综合选择最优主题个数；其次，根据 LDA 模型中高频主题特征词的内部关系，给主题赋予现实意义的主题名称；最后，根据主题分布构建主题评价体系。4) 多主题得分排序：首先，基于游客用户关注度确定主题权重；然后，计算多主题的得分及综合得分；最后进行排序和解释。

本文提出将 LDA 主题模型和基于游客关注度的赋权相结合的一种全新的方法构建旅游景区的评价体系，具有一定的现实指导意义。主要创新：1) 把非结构化的文本数据转换为结构化的主题分布数据，便于后续的数据分析。2) 基于游客关注度确立指标权重，能较好反映游客对主题的关注程度，从而得出较为客观的评价。3) 基于主题分布构建主题评价体系，具有承接的作用，使得对文本评论的主题挖掘工作更进一步。

2. 数据来源与研究方法

2.1. 数据来源

对网上的各大旅游平台受众群体和受欢迎程度进行综合对比，本文选取携程网、去哪儿网作为数据来源。先借助网站自带的综合分析和热度排名对景区进行初步筛选，再选取景区类型相同的景区，便于同类型比较。最终通过 Python 网络爬虫技术获取重庆市 5 家以山为核心要素的自然风景类 5A 级景区的用户评价，共计 12,571 条评论数据，以此来推断重庆市 5A 级景区的特点和综合口碑。为了保证评论的时效性以及更好的体现景区近年来的发展程度和趋势走向，本研究选取的采集对象为 2018~2021 年的用户评价。下表 1 为选取的五家重庆市 5A 级风景区。

Table 1. Five 5A tourist attractions in Chongqing
表 1. 五家重庆市 5A 级旅游景区

景区名称	景区简称	质量等级	评定年月
重庆巫山小三峡－小小三峡	巫山小三峡	5A	2007 年 5 月
重庆武隆喀斯特旅游区	武隆喀斯特	5A	2011 年 6 月
重庆南川金佛山景区	金佛山	5A	2013 年 9 月
重庆江津四面山景区	四面山	5A	2015 年 10 月
重庆云阳龙缸景区	云阳龙缸	5A	2017 年 2 月

注：表 1 数据来源为百度百科词条[1]。

2.2. 研究方法

2.2.1. LDA 主题模型概括

LDA 模型(Latent Dirichlet Allocation)由 Blei [4]等提出，核心思想是文档生成包括两个步骤：主题生成和词汇生成，其中文档服从主题分布、主题服从词汇分布。LDA 模型就是文档生成的逆过程，是一个三层的贝叶斯结构，该模型的出现完成了主题模型在贝叶斯层面的拓展并取得广泛的应用[5]。

2.2.2. 主题数选择

困惑度(Perplexity)是目前自然语言处理中最常见的评价指标[6]。困惑度得分越低，说明模型的效果相对越好。本研究根据困惑度取值来确定模型的主题数。

3. 分析过程

3.1. LDA 主题模型

用户评论经中文预处理后，有效评论共计 11,342 条评论。根据 LDA 主题模型，对处理后的用户评论进行主题划分。使用 2.2 章节的困惑度指标，确定主题个数。一般情况下，困惑度会随着主题数的增

多而逐渐减少，困惑度越小，主题的泛化能力越强。若某点为最优主题数，那么该点的困惑度与前一个点的困惑度差值较大，与后一个点的困惑度差值较小，称为“肘形”[5]。图 1 为不同主题数的困惑度函数训练的输出结果。

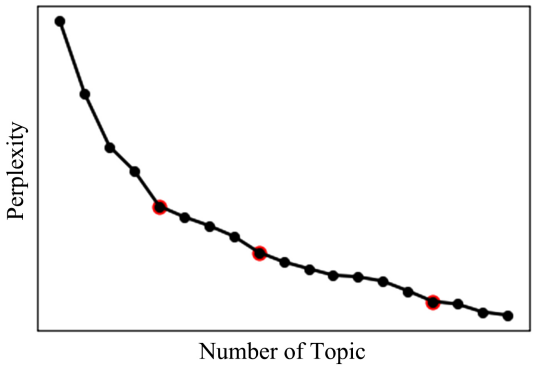

Figure 1. Number of topic-perplexity line graph
图 1. 主题数－困惑度折线图

在折线图的红色圆圈点处(主题数为 5、9、16)，出现了“肘形”，困惑度的递减趋势变缓。经比较后，取主题数为 5 为最优主题个数，利用 LDA 模型得到各用户评论所属主题的概率矩阵。设主题关联度为 0.2 (1/主题数)，将每一条用户评论依据 LDA 模型输出的主题概率分布，根据主题关联度进行归类。若用户评论的主题概率大于主题关联度，则将该条用户评论划入该主题。因此，一条评论可能属于多个主题。所属某主题的用户评论数占比(所属该主题的评论数/总评论数)代表了该主题在景区的热门程度。

Table 2. The distribution of five types of topic-feature words in passenger reviews
表 2. 旅客评论 5 类主题－特征词分布

主题	评论比例(%)	主题命名	特征词
主题 1	24.53%	平台服务	排队、网上、门票、身份证、订票、购票、取票
主题 2	33.15%	景区管理	工作人员、热情、导游、接待、现场、设施、观光车
主题 3	57.62%	自然景观	大自然、山上、云端、清新、奇观、震撼、刺激
主题 4	46.03%	夏冬体验	夏天、下雪、滑雪、乘凉、山上、凉快、冬天
主题 5	22.61%	性价比	便宜、性价比、值得一看、下次、不错、小贵、价格

根据表 2 的高概率特征词，可对每个主题命名。本文只选取与主题相关的 7 个的高概率特征词，发现主题内部的特征词高度相关。如主题 1 中高概率特征词“排队”、“网上”、“门票”、“身份证”、“订票”、“购票”、“取票”内部含义紧密相关，证明 LDA 模型在提取本文旅游用户评论数据的潜在主题方面是有效的。大部分研究者是通过主题含义来自行判定的，但是尽管根据相同的文本来对主题进行凝练也有可能每个人得出的结果会不一样[3]。面对这一问题，暂无统一、有效的办法，本文暂用人工的方式来凝练主题含义，并对主题命名。

根据主题 1 中的 7 个高频词，将主题 1 命名为“平台服务”；在主题 2 中高概率词“工作人员”、“热情”、“导游”、“接待”、“现场”、“设施”、“观光车”均内部语义相关，说明游客对景区管理的表现表示赞许，将主题 2 命名为“景区管理”；在主题 3 中的“大自然”、“山上”、“云端”、

“清新”、“奇观”、“震撼”、“刺激”，体现了游客在旅游时充分感受到大自然的魅力，因此将主题 3 命名为“自然景观”；在主题 4 中出现了较多的季节性的词汇，如“夏天”和“冬天”等高频词，可以将主题 4 命名为“夏冬体验”；在主题 5 中出现了“便宜”、“性价比”、“值得一看”、“下次”和“不错”等高频词，体现了游客对旅游景区性价比的重视程度，因此，将主题 5 命名为“性价比”。

在以山为主要特点的 5 家重庆市 5A 级景区游客评论中，涉及自然景观和夏冬体验的两个主题，评论所占比例高，表示游客对这两个话题热度较高；涉及性价比的评论比例较低，表示游客对此话题热度较低，原因可能是自然景观类景区收费普遍较低，游客对此没有太多关注。

3.2. 主题评价体系

根据 LDA 算法的运行结果，得到 11,342 条有效旅游评论文本的主题分布，如下表 3 所示。表 3 中第 1 行数据表示：金佛山景区第一条用户评论数据的主题分布，评论的主要内容为主题 1 的概率为 0.069，评论的主要内容为主题 3 的概率为 0.025，其余部分以此类推。

Table 3. Theme distribution of tourism reviews
表 3. 旅游评论的主题分布

评论序号	景区简称	主题 1	主题 2	主题 3	主题 4	主题 5
1	金佛山	0.069	0.000	0.025	0.822	0.082
2	金佛山	0.000	0.000	0.148	0.735	0.115
3	金佛山	0.344	0.041	0.535	0.040	0.040
…	…	…	…	…	…	…
11,340	巫山小三峡	0.026	0.025	0.025	0.025	0.899
11,341	巫山小三峡	0.013	0.013	0.014	0.013	0.946
11,342	巫山小三峡	0.583	0.015	0.016	0.015	0.370

通过 LDA 模型得出旅游目的地的评价指标，以主题作为评价指标，其好处为：1) 指标蕴含实际生活意义，便于最终结果的解释；2) 每条评论相当于一个样本，有效的把非结构化数据转换为结构化数据，其对应的数值便于后续评价计算。对于评价指标的权重分配，本文基于游客关注度确立指标权重，即以评论数占比作为指标权重分配的依据，各所属主题的评论数占比除以所有所属主题的评论数占比之和，得到各所属主题指标的指标权重，如下表 4 所示。

Table 4. Evaluation index weight
表 4. 评价指标权重

主题	评论数占比(%)	指标名称	指标权重
主题 1	24.53%	平台服务	0.133
主题 2	33.15%	景区管理	0.180
主题 3	57.62%	自然景观	0.313
主题 4	46.03%	夏冬体验	0.250
主题 5	22.61%	性价比	0.123

3.3. 多主题得分排序

主题综合得分按照“加权平均”的方式计算。首先，计算各景区的对应评价主题的概率，其计算方式：计算每条评论对应主题的概率值的累加值，然后对其累加值进行求均值。然后，基于游客注意力机制，确定出指标权重，评价主题概率与对应的评价主题指标权重相乘得到主题得分。最后，对各景区的5 个主题得分求其均值作为该景区的主题综合得分，并进行景区的得分排序。因赋权相乘后，得分数值较小，为方便观察，所有得分都乘 10。

Table 5. Theme scores of five 5A tourism scenic spots in Chongqing
表 5. 重庆市 5 家 5A 级旅游风景区主题得分

景区	平台服务		景区管理		自然景观		夏冬体验		性价比		综合	
	得分	排序	得分	排序	得分	排序	得分	排序	得分	排序	得分	排序
金佛山	0.227	1	0.395	1	0.518	2	0.392	2	0.102	4	0.327	1
四面山	0.163	3	0.198	4	0.327	4	0.373	3	0.111	3	0.234	5
云阳龙岗	0.214	2	0.355	3	0.301	5	0.286	5	0.186	1	0.268	4
武隆喀斯特	0.132	4	0.187	5	0.576	1	0.351	4	0.179	2	0.285	2
巫山小三峡	0.064	5	0.366	2	0.455	3	0.407	1	0.097	5	0.278	3

由上表 5 可知，对于金佛山景区，其主题综合得分为 0.327，排序第 1。金佛山景区在“平台服务”、“景区管理”主题得分最高；金佛山景区在“性价比”主题得分较低。该景区在今后可以继续保持现有的优质服务，以及合理地调整景区门票价格，不断提高游客满意度。

对于四面山景区，其主题综合得分为 0.234，排序第 5。四面山景区在“景区管理”、“平台服务”主题得分较低。该景区在今后可以着力整顿景区管理，为游客提供更优质的服务和管理。

对于云阳龙岗景区，其主题综合得分为 0.268，排序第 4。云阳龙岗景区在“夏冬体验”、“自然景观”的主题值最低；云阳龙岗景区在“平台服务”、“性价比”的主题得分较高。该景区可以继续保持优势项，在科学合理的前提，继续挖掘景区特色景点，进而不断提高游客满意度。

对于武隆喀斯特景区，其主题综合得分为 0.285，排序第 2。武隆喀斯特景区在“自然景观”主题得分最高；武隆喀斯特景区在“景区管理”主题得分最低。这体现该景区在自然景观方面存在明显的优势，在今后可以不断完善其景区内景点，并且加强景区管理，可以通过与相关企业合作来实现景区的更好管理，从而不断提高提高游客满意度。

对于巫山小三峡景区，其主题综合得分为 0.278，排序第 3。巫山小三峡景区在“夏冬体验”主题得分最高；巫山小三峡景区在“平台服务”、“性价比”主题得分最低。该景区在夏天时旅游热度更高，景区可以做好夏季旅游旺季的充分准备，并且加强与网络平台的合作，丰富景区内的游览项目，通过薄利多销的方式提高景区的收入，进而使游客更加满意。

4. 结论与改进

4.1. 结论

本文以重庆市 5 家以山为核心要素的 5A 级自然旅游景区为研究对象，利用 Python 网络爬虫技术，获取了用户评论文本，然后基于 LDA 主题模型，进行文本挖掘。主要结论为：1) 从提取的主题信息来

看，游客关注的五个主题为：平台服务、景区管理、自然景观、夏冬体验、性价比。其中自然景观主题，游客热度最高；性价比主题，游客热度最低。2) 综合评价得分排序结果为：金佛山 > 武隆喀斯特 > 巫山小三峡 > 云阳龙岗 > 四面山。此外，在 5 个主题得分排序上，平台服务得分第 1 是金佛山、景区管理得分第 1 是金佛山、自然景观得分第 1 是武隆喀斯特、夏冬体验得分第 1 是巫山小三峡、性价比得分第 1 是云阳龙岗。3) 针对各景区可能存在的问题，提出各自的改进意见：金佛山景区可以继续保持现有的优质服务，以及合理地调整景区门票价格；四面山景区可以重点关注景区管理，为游客提供更优质的服务和管理；云阳龙岗景区可以在保留自身特色景点的前提下，科学合理地继续开发景区特色景点，进而提升自然景观方面的竞争力；武隆喀斯特景区可以继续加强景区管理，通过与相关企业合作来实现景区的更好管理；巫山小三峡景区可以做好夏季旅游旺季的充分准备，加强与网络平台的合作，从而丰富景区内的游览项目，通过薄利多销的方式提高景区的收入。

4.2. 改进

本文对游客用户评论进行文本挖掘分析，为游客外出旅行提供了良好的参考，同时对重庆市相关旅游部门具有一定的参考意义。但本文仅以重庆市 5 家 5A 级旅游景区为例，其评论数据量不够丰富，所得结论与实际情况具有一定程度的偏差，并不能完全反映重庆市的旅游现状，因此分析的结论还不够全面。后续改进：1) 扩大研究范围，后续相关研究将增加研究对象数量，使相关研究更具代表性；2) 数据来源多元化，后续相关研究需要增加评论数量，便于更好挖掘评论文本中蕴含的主题信息。

参考文献

[1] 百度百科. 重庆景点[EB/OL]. https://baike.baidu.com/item/重庆景点, 2021-10-25.

[2] 新浪财经. 较 2020 年增长 31.2%! 2021 年国庆假期重庆 A 级旅游景区累计接待游客 1273 万人次[EB/OL]. https://finance.sina.com.cn/jjxw/2021-10-07/doc-iktzqtyu0133479.shtml, 2021-10-25.

[3] 周文亮. 基于 LDA 改进的 AHP 旅游目的地评价研究——以江西省 5A 级景区为例[D]: [硕士学位论文]. 南昌: 江西财经大学, 2021.

[4] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2012) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[5] 赵蓉英, 戴祎璠, 王旭. 基于LDA模型与ATM模型的学者影响力评价研究——以我国核物理学科为例[J]. 情报科学, 2019, 37(6): 3-9.

[6] Azzopardi, L., Girolami, M. and Risjbergen, K.V. (2003) Investigating the Relationship between Language Model Perplexity and IR Precision- Recall Measures. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, 28 July-1 August 2003, 369-370. https://doi.org/10.1145/860435.860505