doi:10.3772/j.issn.2095-915x.2017.03.008

基于 LDA 的游客网络评论主题分类：以故宫为例

1. 北京联合大学旅游信息化协同创新中心 北京 100101；
2. 北京联合大学北京市信息服务工程重点实验室 北京 100101；
3. 美国弗吉尼亚理工大学 布莱克斯堡 VA 24061 美国

黎嘤[1] 谢宗彦[2] 张公鹏[1] 郝志成[1] 向征[3]

摘 要 游客的网络评论由于能够真实反映对旅游服务的真实体验及感受，正在逐渐影响旅游者对旅游目的地的看法甚至旅游消费行为。如何将碎片化的旅游评论转化为对其他用户和旅游经营者有价值的且直观的信息，成为旅游信息挖掘的热点。本文提出了基于 LDA（Latent Dirichlet Allocation）主题发现模型的游客评论挖掘方法，以大众点评、携程及马蜂窝中关于故宫的用户在线评论为例，挖掘游客关于故宫的关注主题并分析其情感倾向。实验结果表明，故宫的游客网络评论主题包含入口服务、历史文化、体验感受以及遗址文物四个方面，游客对该四个主题的情感倾向均为正向；其中，大众点评和马蜂窝在体验感受方面的情感极性值较高。该方法对定位旅游目的地游客关注点具有实践意义。

关键词： LDA, 游客 , 网络评论 , 情感分析 , 故宫

中图分类号： G35

Topic Classification of Tourist Online Reviews Based on LDA: The Case of the Forbidden City, Beijing

1. Collaborative innovation center of tourism informatization, Beijing Union University, Beijing 100101, China;
2. Beijing Key Laboratory of Information Service Engineering, Beijing 100101, China;
3. Pamplin College of Business, Virginia Tech, Blacksburg VA 24061, USA

LI Nao[1] XIE ZongYan[2] ZHANG GongPeng[1] HAO ZhiCheng[1] XIANG Zheng[3]

基金项目： 本文受国家自然科学基金青年项目“基于Agent的景区游客游憩行为仿真建模研究”（41101111），北京联合大学学术（科研）创新团队资助项目“旅游大数据研究方法、关键技术与应用研究”（Rk100201509）的资助。

作者简介： 黎嘤（1975–），女，博士，副教授，研究方向：旅游大数据分析与挖掘，游客行为仿真，Email：lytlinao@buu.edu.cn；谢宗彦（1992–），女，硕士研究生，研究方向：文本分析与挖掘；张公鹏（1978–），男，博士研究生，研究方向：旅游大数据分析与挖掘；郝志成（1977–），男，硕士，工程师，研究方向：数据挖掘，计算机仿真；向征（1967–），男，博士，副教授，研究方向：信息技术和旅游体验，旅游大数据分析和旅游营销策略。

doi:10.3772/j.issn.2095-915x.2017.03.008

Abstract Tourist online reviews can reflect their real experiences and feelings on travel service, which gradually influence other tourists' opinions on tourist destinations and even consumption behaviors. How to convert the fragmented travel comment data into useful information for other users and tour operators becomes a hotspot of tourism information mining. This paper proposed an information mining method based on the LDA (Latent Dirichlet Allocation) topic discovery model, and carried on a topic classification and sentimental analysis of the online reviews of the Forbidden City in dianping.com, Ctrip.com and mafengwo.cn. The results showed that there were four topics of online reviews of the Forbidden City, these are entrance service, history and culture, experience and feeling, heritage and cultural relic, and the sentimental on these four topics is positive. Wherein, dianping.com and mafengwo.cn have high sentimental polarity values in terms of experience and feeling. This method has practice meaning for addressing what tourist pay attention to on destinations.

Keywords: LDA, tourist, online review, sentimental analysis, the Forbidden City

1 引言

Web 2.0 时代，越来越多的游客通过网络评论分享其对旅游产品和服务的购买体验，表达自己的观点和感受。截止 2015 年底，美国 Yelp 产生了 95 000 000 条评论，2015 年第四季度平均每月就有 86，000,000 位移动用户访问该网站[1]；截止 2016 年底，马蜂窝（www.mafengwo.cn）产生了 210000000 条评论。游客网络评论作为一种不同于传统问卷、访谈等研究方法所获取的，反映游客对产品体验和评价的丰富数据源，被用于研究接待与旅游的许多问题[2]，如，游客满意度[3]，服务质量评价[4]，旅游目的地形象[5]等。

随着社交媒体分析技术的不断发展，相关研究持续增长，集成网页爬取、计算语言学、机器学习以及统计等技术，采集、分析和解释互联网大数据，基于商业目标追踪热门话题、普遍情绪以及对于产品的观点和信念[6]。其中，LDA（Latent Dirichlet Allocation）[7] 作为一种自动文本分类模型，能够很好地将不同文本通过隐含主题联系起来，挖掘包含大量冗余与不完备信息的网络评论的隐藏信息[8]，目前在网络评论分析中得到了广泛应用。例如，利用 LDA 主题模型来计算文本的相似度[9]，基于 LDA 进行微博当前热点话题的抽取[10]，基于 LDA 实现对文本流的有效主题挖掘[11]等。近期，LDA 开始被用于游客网络评论分析，基于 LDA 来确定酒店顾客所关心的客户服务关键维度[12]，利用 LDA 对酒店在线评论关键词进行降维[6]。

目前，基于 LDA 的游客网络评论主题分类研究主要针对接待业，如 Guo&Jia[12] 和 Xiang[6] 等的研究均面向酒店业，鲜有针对景区、目的地的相关研究。这是因为接待业尤其是酒店业提供的是标准化产品，而景区、目的地类别繁多，仅从资源角度的分类就有 8 个主类，31 个亚类，155 个基本类[13]，游客对不同资源类型的景区表现出不同的关注点，比如，游客对山岳型景区的关注点为自然风光，而对历史文化遗迹则关注历史文化要素，且即便是相同类型的景区，由于本身具有鲜明的特点，也会

对游客产生不同的关注点。因而，景区的游客评论所表达的关注点，无法标准化，为主题分析带来困难。

本研究尝试基于 LDA 针对景区进行网络评论主题划分，以补充目前以针对酒店为主的游客评论研究。由于景区类别繁杂，本研究以目前客流量较大的著名文化遗址类景区 —— 故宫作为案例进行分析，以期对其他类别景区，甚至目的地的游客评论研究带来启示。为了避免单一社交媒体数据源造成的内在特征与潜在偏差[14]，本研究从多信息源获取评论数据，考虑覆盖不同类型的游客评论数据，分别从携程、马蜂窝以及大众点评网三种不同类型的旅游在线服务网站采集故宫网络评论数据。通过数据预处理，基于 LDA 的主题分类，最终给出三大网站故宫游客评论的主题类别；基于主题类别，进一步采用情感分析，给出每种主题游客的情感倾向。

2 LDA主题模型

2.1 LDA的基本思想

LDA（Latent Dirichlet Allocation）[7] 是一个三层贝叶斯概率模型，包含词、主题、文档三层结构。LDA 是一个生成模型，对于某篇文章中的第 n 个词，首先从该文章的主题分布中采样一个主题，然后在这个主题对应的词分布中采样一个词。不断重复这个随机生成过程，直到篇文章全部完成上述过程。LDA 对文本信息的主题建模，如下图所示（图 1）。

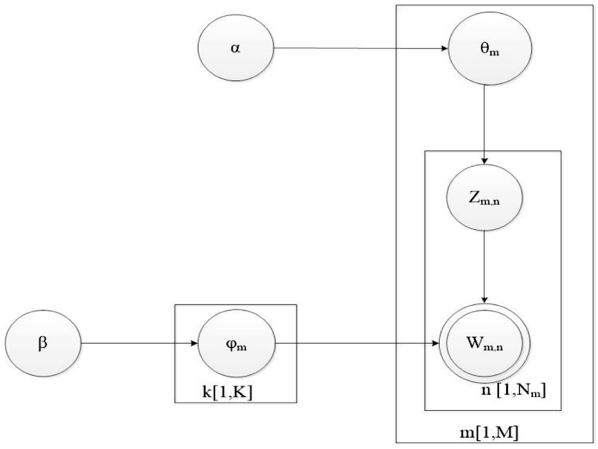

图1 LDA图模型

其中 K 为主题个数，M 为文档总数，N_m 是第 m 篇文档的单词总数。β 是每个主题下的词的多项分布的 Dirichlet 先验参数，α 是每个文档下主题的多项式分布 Dirichlet 先验参数，$Z_{m,n}$ 是第 m 篇文章中第 n 个词的主题，$W_{m,n}$ 是第 m 篇文档中的第 n 个词。

给定一篇文档集合，$W_{m,n}$ 是可以观察到的已知变量，α 和 β 是根据经验给定的先验参数，其他的变量 $Z_{m,n}$、θ_m、φ 都是位置的隐含变量，需要根据观察到的变量来学习估计。根据 LDA 的图模型，所有变量的联合分布：

$$p(W_m, Z_m, \theta_m, \varphi \mid \alpha, \beta) = \prod_{n=1}^{N_m} p(W_{m,n} \mid \varphi_{Z_{m,n}}) \cdot p(Z_{m,n} \mid \theta_m) \cdot p(\theta_m \mid \alpha) \cdot p(\varphi \mid \beta)$$

（公式 1）

每个文档中出现主题 k 的概率乘以主题 k 下出现此 w 的概率，然后枚举所有主题求和得到整个文档集的似然函数为：

$$p(W \mid \Theta, \Phi) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(W_{m,n} \mid \theta_m, \varphi)$$（公式 2）

其中 W 为整个文档集的词，Θ 为整个文档集的主题分布，Φ 为整个文档集的词分布。

doi:10.3772/j.issn.2095-915x.2017.03.008

2.2 Gibbs采样

LDA 模型的构建过程中需要对相应的参数进行估计，比较常用的方法有期望传播算法、变分贝叶斯估计和 Gibbs Sampling 等。基于 Gibbs 采样的参数推理方法容易理解且实现简单，能够非常有效的从大规模文本集中抽取主题。Gibbs Sampling 算法的运行方式是每次选取概率向量中的一个维度，给定其他维度的变量值采样当前维度值。不断迭代直到收敛输出待估计的参数。根据图模型，可得到一篇文本的概率值为：

$$p(W_m, Z_m \mid \alpha, \beta) = p(W_m \mid Z_m, \beta) p(Z_m \mid \alpha)$$

（公式 3）

其中 W_m 为文档中的词，β 是每个主题下的词的多项分布的 Dirichlet 先验参数，α 是每个文档下主题的多项式分布 Dirichlet 先验参数，Z_m 是文档的主题分布。

Gibbs Sampling 算法避开了实际待估计的参数，转而对每个单词的主题进行采样，一旦每个单词的主题确定下来，参数就可以在统计词频后计算出来。

3 数据采集与预处理

3.1 数据采集

本研究分别从携程、马蜂窝以及大众点评网三种不同类型的旅游在线服务网站采集故宫网络评论。携程属于目前国内规模最大的在线旅行社，马蜂窝属于旅游社交网站，大众点评属于综合性社区网站。采集时间为 2017 年 2 月，携程网显示故宫评论 38810 条，马蜂窝 10525 条，大众点评网 7922 条。为了减少噪音数据对主题分类的干扰，需要对采集到的评论进行去重、去除无意义数据（广告，无意义的超短文本）等数据清洗。最终，获取到关于故宫的有效游客评论共 19526 条，其中马蜂窝 10515 条，大众点评 7922 条，携程网 1019 条。

3.2 数据预处理

数据预处理对于主题分类的质量至关重要。本研究进行三步数据预处理。首先进行分词和词性标注。采用中科院计算所张华平、刘群研制的 ICTCLAS 分词词性标注一体化系统 [15]。ICTCLAS 使用层叠隐马尔可夫模型，基于 N 最短路径的切分排歧策略和基于角色标注的未登录词识别策略，将汉语分词、切分排歧、未登录词识别、词性标注等词法分析任务融合到一个相对统一的理论模型中，能够取得良好的分词效果（http://ictclas.nlpir.org/）。分词后共得到 1066474 个词汇。

第二步为去停用词。本研究采用一个较为通用的停用词表（http://www.cnblogs.com/ibook360/archive/2011/11/23/2260397.html），共有 2825 个停用词。用该停用词表作为初始停用词表，根据多次主题分析结果，对初始停用词表进行扩展，增加主题分类实验中出现的对于主题分类无意义的高频词，如：故宫、北京等。

第三步为语义去重与合并。利用 Hownet 将相同含义的词或短语合并，即先对预处理得到的数据集中所包含的特征词项进行语义分析，通过词项相似度的计算，删除、合并语义相似的词项。

经过以上三步数据预处理，分词后的 26% 词汇被去除，用于主题分类的噪音被大规模减少。

4 数据分析

4.1 主题分类

基于 LDA 分别对三个网站的评论词汇进行主题分类，其中，采用 MCMC Gibbs 算法进行参数的估计，具体设置为 topic 的初始值 K=10，α=0.5，β=0.1，Gibbs 采样的迭代次数为 100 次，实验过程中主题数 K 的取值依次为 10、8、6，4，利用不同的主题数进行多次试验，获取最优的主题数 K 为 4。实验所得主题分类及其对应的主题词如表 1 所示。

表1　故宫游客评论的主题分类

	主题 1：入口服务	主题 2：历史文化	主题 3：体验感受	主题 4 遗迹文物
大众点评	讲解 0.03469	建筑 0.03935	景点 0.01111	皇帝 0.01825
	门票 0.02698	中国 0.03223	不错 0.01089	宫殿 0.01730
	时间 0.02328	历史 0.03206	历史 0.00983	御花园 0.01588
	导游 0.02298	宫殿 0.02274	皇家 0.00748	太和 0.01517
	午门 0.01053	古代 0.01850	大气 0.00748	珍宝 0.00925
	身份证 0.0203	文化 0.01799	气派 0.00663	建筑 0.00902
	排队 0.01335	紫禁城 0.01392	震撼 0.0064	大殿 0.00878
	网上 0.01276	明清 0.0132	景色 0.00812	坤宁 0.00499
携程网	门票 0.03637	中国 0.0383	喜欢 0.01804	珍宝 0.02251
	讲解 0.01936	建筑 0.03315	皇帝 0.01771	值得 0.01791
	排队 0.01713	历史 0.02680	历史 0.00982	喜欢 0.01280
	身份证 0.01380	宫殿 0.02101	宏伟 0.00875	钟表 0.01717
	网上 0.01253	文化 0.01981	印象 0.00752	历史 0.01218
	便宜 0.01215	古代 0.01891	壮观 0.00715	宫殿 0.00862
	时间 0.01065	紫禁城 0.01664	宫殿 0.00575	大殿 0.00673
	安检 0.00881	明清 0.00862	气派 0.00571	景点 0.01164
马蜂窝	讲解 0.03469	历史 0.05495	不错 0.01706	建筑 0.03753
	门票 0.02815	建筑 0.02969	喜欢 0.01563	宫殿 0.03180
	时间 0.02493	中国 0.02657	值得 0.01078	皇帝 0.01945
	导游 0.01957	值得 0.02457	游客 0.00992	文物 0.015734
	午门 0.01226	文化 0.01682	景点 0.00921	保存 0.01359
	排队 0.01144	皇家 0.01344	时间 0.00664	太和 0.01159
	网上 0.00971	古代 0.01232	壮观 0.00635	规模 0.01069
	身份证 0.00741	宏伟 0.01163	推荐 0.00614	大殿 0.01056

实验表明，三个不同类型网站针对故宫的游客评论主题类别是相同的，即，三个网站游客评论所隐含的游客关注主题均为入口服务、历史文化、体验感受以及遗迹文物四类。其中，历史文化与遗迹文物均关注建筑与宫殿，但前者侧重历史、年代（明清、古代）等文化内容，

doi:10.3772/j.issn.2095-915x.2017.03.008

后者侧重具体遗迹和文物对象，如珍宝、钟表、太和殿等。每类主题包含的若干主题词带有后验概率，表示该主题词属于特定主题的概率（为节省空间，每个主题仅列出来 8 个主题词）。如表 1 所示，故宫的游客评论主题类别及其主题词，具有景区独特性。首先，故宫作为明清两朝皇宫，拥有独特藏品与世界上规模最大、保存最完整的宫殿建筑群，文化历史与遗址文物受到游客关注；其次，故宫作为全人类的珍贵文化遗产，游客的体验和感受与以上特性相关；最后，入口服务尽管可能也会在其他景区受到关注，但其主题词所包含网上（门票预约）、午门（南入口）以及安检等，体现了故宫的特有服务与管理措施受到了游客的关注。

尽管主题类别相同，但每个主题在每个网站的评论中所占的比重不同（图 2）。如图 2 所示（横坐标表示主题，纵坐标表示每个网站属于该主题的评论数占总评论数的百分比），总体上，各个主题在三个网站中的比例没有出现较大差距。主题 1 即入口服务，在三个网站所占的比例大致相同。这表明，三个类型网站的故宫评论都对入口服务表现出了相同程度的关注。其他三个主题在三个网站中的比例有略微差别。比较而言，主题 2 即历史文化，在大众点评中占比较大，在携程网占比较小，似乎说明大众点评网的游客比携程的游客更加关注历史文化；主题 3 即体验感受，在马蜂窝中占比较大，在携程和大众点评中占比大致相同，这种情况似乎印证了社交媒体网站游客更加注重景区带给自身的体验与感受；主题 4 即遗址文物，在携程中占比最大，其次为大众点评。

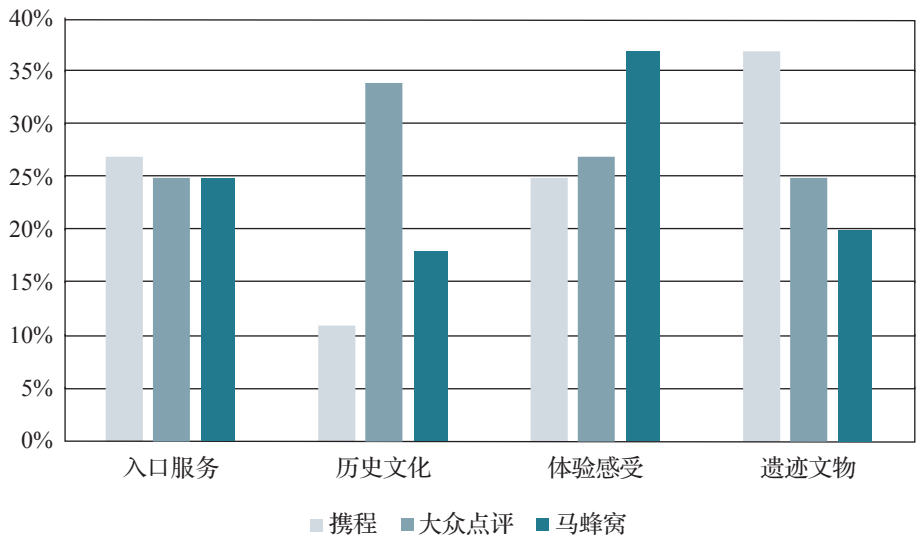

图2 大众点评、携程和马蜂窝故宫评论的主题分布

4.2 情感分析

由于旅游领域没有情感词典，为了计算每一条评论的情感值，我们首先建立了适应旅游领域以及网络评论的旅游情感词典。然后，利用情感极性计算规则，我们将大众点评网、携程网和马蜂窝中得到的故宫的游客评论逐一计算其情感极值（大于 0 表示正向情感，小于 0 表示负向情感；值越大表示极性越强），

并按照主题分类求情感极性平均值，得到三个网站中故宫游客评论各主题的情感分类均值（图 3）。

如图 3 所示（横坐标表示网站，纵坐标表示每个主题在各网站中的情感极性均值），三个网站中游客对故宫入口服务、历史文化、体验感受以及遗迹文物四个主题的评论均表现出正向情感。如果我们假设情感极性值与用户满意度一致，大于 0 表示满意，小于 0 表示不满意，且值越高满意度越高，那么可以认为这三个网站的游客对故宫的满意度为正向。总体上，三个网站评论对故宫入口服务、历史文化和遗迹文物三个主题的情感极性均值相差不大。而对体验感受这一主题的情感极性均值则存在显著差异，大众点评网和马蜂窝中的游客评论对于故宫的体验感受表现出相比携程更为强烈的正向情感。这表明大众点评网和马蜂窝的游客在故宫的体验感受上满意度比携程的游客要高。然而，如果考虑三个网站的用户可能存在不同特性或偏好，该差异也可以从另外一方面解释，即大众点评网和马蜂窝的游客更愿意使用情感较为强烈的词汇来表达体验与感受。

前文提到，主题 2 即历史文化在大众点评中占比较大，在携程网占比较小；而两个网站评论对于主题 2 的情感均值却相差极小。这说明评论数量与情感值相关性较小，评论数量多并不意味着情感值高。此外，我们还计算了总评论的情感极性均值，其中，携程网的总评论的情感均值为 0.44，大众点评为 1.01，马蜂窝为 1.14。这表明，总体比较而言，大众点评和马蜂窝的用户比携程网的的用户对于故宫有相对较强的正向情感体验。

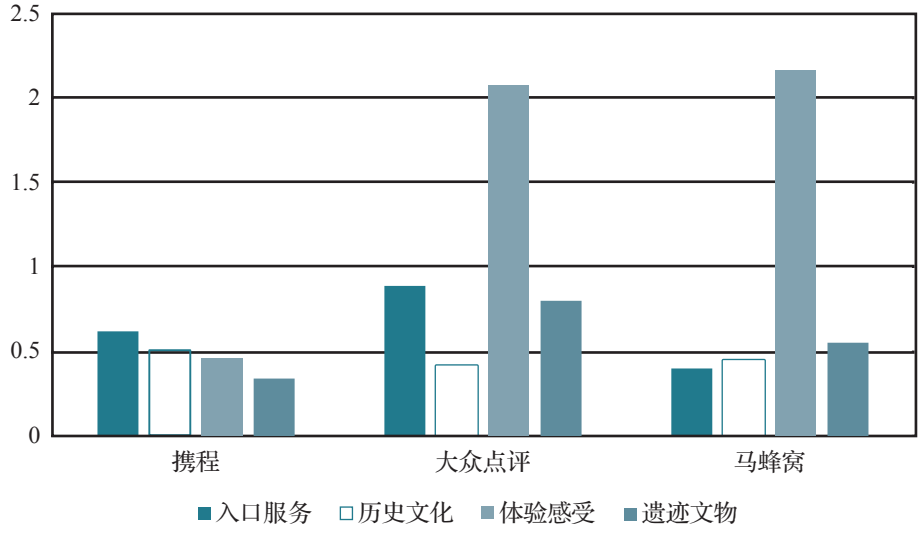

图3 大众点评、携程和马蜂窝故宫评论的各主题的平均情感值

5 结论

本文提出了一种用游客网络评论获取游客关注主题的方法。本文对于 LDA 主题模型、情感分析以及以网络评论为代表的用户生成内容的获取技术的使用，使得本文能够挖掘出传统旅游研究方法难以获得的非预定义的游客关注维度、内容以及游客在不同维度的情感倾向。

doi:10.3772/j.issn.2095-915x.2017.03.008

传统旅游研究在进行游客关注点研究时，常采用预定义理论模型与结构的方法，而即便是使用焦点小组并配合编码式文本分析这种非预定义模型的方法，也很难脱离人为的预定义模式，如人工或半人工的编码与分类。本文为旅游研究贡献了一种方法上的新启示。

其次，本文针对旅游景区这一非标准化旅游服务设施开展网络评论主题研究，是对目前以酒店等标准化旅游产品与服务为主的相关研究的有益补充。非标准化旅游产品不仅复杂多样，且每一景区都有其独特特性。这种独特性一方面使得针对景区的游客网络评论难以像酒店客户评论那样统一处理，另一方面使得领域词、情感词等方面的处理复杂性加剧。从工程意义上讲，本文扩展了目前旅游研究中基于LDA的相关研究能够处理的复杂程度，使得基于LDA的主题分类能够在更宽泛的旅游范畴（而不仅仅是酒店）处理更为多样的游客网络评论。

本文也为旅游实践者提供了有价值的研究发现。旅游景区能够意识到基于游客网络评论可以分析出游客关心的主题维度和游客对于这些维度的情感倾向。这一方面会让旅游景区认可互联网用户生成内容的价值；另一方面，也更为重要的是，旅游景区能够基于这些维度及其情感倾向，有针对性的改进和提升相关管理与服务。以故宫为例，研究所得出的入口服务、历史文化、体验感受以及遗址文物四个游客特别关注的主题，尽管均表现出正向情感但大部分均值不高，暗示故宫管理者应在包含了排队、安检、导游等方面的入口服务进行持续的改进和提升，在历史文化、遗址文物等方面进行更为深入的能够提升游客体验与感受的价值挖掘。

本文存在一些局限。首先，本文仅选取了以携程为代表的的在线旅行社、以马蜂窝为代表的在线旅游社交网站以及以大众点评为代表的综合社区三种类型的在线旅游服务网站进行数据采集，研究结果受限于这些网站的类别、用户的特性与偏好，样本规模需要进一步扩大。其次，本文以故宫为例，故宫仅代表文化遗址类景区，而景区的类别复杂多样且独特性较强，如何处理同类型多景区，甚至多类型景区，是下一步值得关注的问题。

参考文献

[1] Lee S H, Ro H. The Impact of Online Reviews on Attitude Changes: The Differential Effects of Review Attributes and Consumer Knowledge[J]. International Journal of Hospitality Management, 2016, 56: 1-9.

[2] Schuckert M, Liu X W, Law R. Hospitality and Tourism Online Reviews: Recent Trends and Future Directions[J]. Journal of Travel & Tourism Marketing, 2015, 32(5): 608-621.

[3] Li H Y, Qiang Y, Law R. Determinants of Customer Satisfaction in the Hotel Industry: An Application of Online Review Analysis[J]. Asia Pacific Journal of Tourism Research, 2013, 18(7): 1-19.

[4] Bogicevic V, Yang W, Bilgihan A, et al. Airport Service Quality Drivers of Passenger Satisfaction[J]. Tourism Review, 2013, 68(4): 3-18(16).

[5] Garay L, Cànoves G. Barcelona: New Stakeholders and New Images in Social Media Reviews[J]. Tourism Analysis, 2016, 21(3).

[6] Xiang Z, Du Q, Ma Y, et al. A Comparative Analysis of Major online Review Platforms: Implications for Social Media Analytics in Hospitality and Tourism[J].

Tourism Management, 2017, 58: 51-65.

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[8] 阮光册 . 基于 LDA 的网络评论主题发现研究 [J]. 情报杂志 , 2014(3): 161-164.

[9] 王振振 , 何明 , 杜永萍 . 基于 LDA 主题模型的文本相似度计算 [C]. 全国智能信息处理学术会议 . 2013: 229-232.

[10] Liu G, Xu X, Zhu Y, et al. An Improved Latent Dirichlet Allocation Model for Hot Topic Extraction[C]// IEEE Fourth International Conference on Big Data and Cloud Computing. IEEE, 2014: 470-476.

[11] Wang Y, Agichtein E, Benzi M. TM-LDA: Efficient online Modeling of Latent Topic Transitions in Social Media[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012: 123-131.

[12] 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.

[13] GB/T 18972-2003, 旅游资源分类、调查与评价 [S]. 北京 : 中国质检出版社 , 2003.

[14] Ruths D, Pfeffer J. Social Media for Large Studies of Behavior[J]. Science, 2014, 346(6213): 1063-1064.

[15] 刘群 , 张华平 , 俞鸿魁 , 等 . 基于层叠隐马模型的汉语词法分析 [J]. 计算机研究与发展 , 2004, 41(8): 1421-1429.