基于文本挖掘的西部城市旅游满意度研究

——以携程西安和成都为例

李 爽，张 政，刘娅娅

（西安财经大学 统计学院，西安 710100）

摘要：利用携程网西安和成都的 5 万条旅游评论数据，构建旅游满意度指标体系，以评估不同城市旅游消费者的满意度。采用基于词云图、社会网络语义图以及 LDA 模型的特征分析来提取评价指标，构建适用于旅游评论的情感词典，并据此为评论情感赋值，结合层次分析法确定指标权重，分别计算满意度。结果显示：成都市的综合满意度高于西安市；游客对于成都旅游服务中的导游情况和行程安排的满意度高于西安；而西安携程旅游服务中地区游览特色更好，景点体验更佳。

关键词：文本挖掘；情感分析；满意度；LDA 模型；旅游

中图分类号：F063.2； **文献标志码**：A **文章编号**：1671－1807(2022)04－0326－08

随着信息技术的发展和居民生活水平的提高，特别是近年来移动终端设备的普及，全国网民数量进一步增多，形成了旅游市场巨大的潜在消费群体。第 48 次《中国互联网络发展状况统计报告》显示，中国 2021 年网民数量为 10.11 亿，互联网普及率达到 71.6%[1]。据前瞻产业研究院数据，截至 2019 年，全国在线旅游市场交易规模已突破万亿元，用户规模突破 4 亿人，其中携程月活跃用户规模居榜首[2]。

成都和西安都位于西部地区，均是传统旅游城市。2018 年，西安市旅游总收入达到 2 013.2 亿元，同比增长超过 50%；成都市旅游总收入达到 3 712.6亿元，同比增长 22.4%。作为西部旅游强市，对两市旅游业的分析研究具有重要意义。翟若琳[3]采用 DEA 模型对于西安城市旅游效率进行了对比分析；欧启均等[4]以西安市为例研究了城市旅游客流的网络信息驱动特征变化问题；徐茜[5]以成都为例研究了文创与旅游融合的路径问题。鲜有文献对西安、成都两市的旅游满意度进行比较研究。

用户满意度指消费者对产品购买后的使用感知效果与期望进行对比后所产生的情绪状态。以往学者已对用户满意度做了很多工作[6]，但通过对文献的梳理发现，研究者通常采用问卷调查[7-10]来收集数据，但问卷调查耗费人力物力较多，且样本数据有限[11]，局限性较大。事实上，在线评论中蕴含着用户的情感倾向，若基于线上评论探索满意度，则可补充传统问卷调查获取数据的不足。近年来，诸多学者在多个领域运用了文本挖掘的方法研究用户满意度。蔡珺哲[12]采用文本挖掘方法研究中国当前社交媒体的用户情感问题；张琰等[13]基于酒店评论研究了不同档次酒店顾客满意度的因素对比；黎晶[14]基于决策树方法对移动互联网服务进行满意度研究；赵杨等[15]利用 CNN-SVM 方法最终构建起评论满意度得分；郭立秀[16]运用了特征词匹配方法，精细化研究了生鲜电商满意度问题；王媛等[17]基于文本挖掘技术收集博客中的文本数据，据此研究了游客对古镇旅游形象的感知问题；刘阳[18]通过文本挖掘研究了哪些因素影响了在线旅游产品的销量因素；范宁[19]依托在线旅游网站评论，研究了消费者对于民宿满意度特征；刘婷[20]使用 IPA 分析法对提高三亚旅游购物满意水平进行研究。

关于文本的情感分析，Cambria 等指出情感分析通过判断评论的情感极性和强度[21]，在一定程度上能够反映旅客的满意度。情感分析方法一般为

收稿日期：2021-11-05

基金项目：国家社会科学基金(17BTJ022)。

作者简介：李爽(1980—)，男，河南唐河人，西安财经大学统计学院，教授，硕士研究生导师，研究方向为应用数理统计；张政(1994—)，男，陕西西安人，西安财经大学统计学院，硕士研究生，研究方向为数量经济建模；刘娅娅(1995—)，女，四川达州人，西安财经大学统计学院，硕士研究生，研究方向为文本挖掘。

两种,一种是情感词典法。Zhang 等[22]定义了相关领域的情感词典,并按照词语倾向性建立了情感倾向分析模型;丁蔚[23]将情感词典法与机器学习方法相结合对评论进行分类,提高了预测的准确性和泛用性;崔志刚[24]将情感分析方法用于分析电商用户的喜好;刘楠[25]构建了多元词特征分析法对微博短文本进行分类,并进行情感倾向的研究。另一种是机器学习方法。Li 等[26]对评论的情感倾向进行聚类分析,最终统计了正面、中性、负面评论的占比;徐军等[27]利用朴素贝叶斯和最大熵两种方法研究情感倾向分类,并比较两者间最高准确率;徐琳宏[28]建立文本倾向识别机制,并使用 SVM 模型进行文本感情倾向分析。张英[29]基于深度神经网络研究了微博短文本分类和情感预测问题。鲜有文献构建适用于旅游领域的情感词典,因此现有研究对旅游评论的情感分析存在偏差。

基于上述讨论,利用携程网西安和成都的旅游在线评论数据,通过词云图、社会网络语义图、LDA 主题模型挖掘旅游满意度影响因素,构建了城市旅游满意度评价体系。进一步构建适用于研究的情感词典,计算指标情感值,结合层次分析法确定指标权重,据此计算游客的满意度。所得的指标体系不仅可用于西安、成都两地,也可用于其他城市的旅游满意度研究,具有一定的通用性。

1　基于文本挖掘的游客满意度指标体系构建

1.1　数据来源

爬取 2016 年 1 月 1 日至 2019 年 3 月 1 日来自携程旅游网西安和成都页面中的跟团游项目的在线评论数据。西安市实际爬取 28 268 条数据,经过预处理得到 18 452 条有效数据。成都市实际爬取 22 450 条数据,经过预处理得到 14 568 条有效数据。

1.2　游客满意度影响因素挖掘

1.2.1　基于词云图和语义图的影响因素初探

由于采集到的数据量较大,过滤掉部分与研究无关的词语,绘制游客满意度影响因素词云图,如图 1 所示。图字体的颜色和大小反映了词频高低和重要程度。综合来看,两市在线旅游评论反映出的主题有很多相似之处,占两市词云图中显示面积最大的词语均是“行程”和“安排”,这两词基本可视为同义词,由于行程路线安排会反映在旅行的价格上,而这些因素又很大程度上决定了游玩的综合体验,因此游客的关注度最高。词云图中也显示出了“峨眉山”“大雁塔”等景点名称,表明游客对这些知名景点产生了深刻印象,游客可能正是为了此景点而来的。此外,从词云图中也可以看出“导游”“讲解”也是游客关注的重点。最后,成都的词云图中提及了旅游线路购物方面的问题,这点是西安词云图不曾涉及之处。

词云图只能大致看出游客们关注哪些影响因素,但是很难发现之间的关联,因此可利用网络语义图,进一步分析旅游评论的特征。采用 ROSTCM 6 软件分别对两市评论进行语义网络分析,得出的部分结果如图 2、图 3 所示。

(a) 成都　　(b) 西安

图 1　两市满意度影响因素的词云图对比

从两城网络语义图总体来看,两城的影响因素较为相似,“行程”“安排”“讲解”“导游”“服务”这几个词作为中心节点连接了其他节点。进一步分析可知,以“导游”为节点的密切相连的评价词有“详细”“认真”“满意”等词,说明游客对于导游总体上的评价是相对正面的。另外,导游的“态度”也是游客关注的一个要点。这些可以归纳为游客对于导游的服务非常敏感。以“行程”“安排”为节点的,与

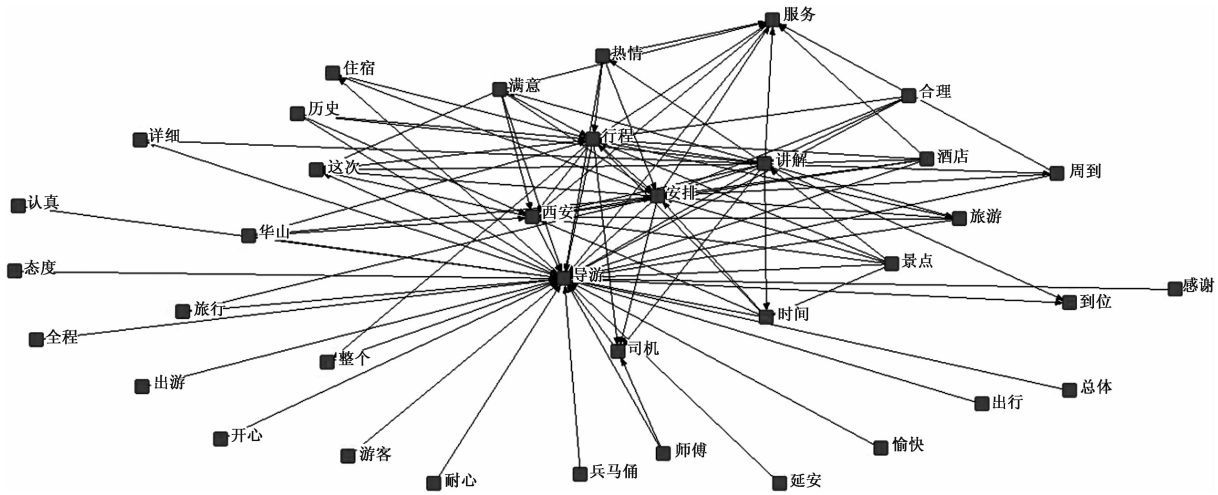

图 2　西安旅游影响因素网络语义图

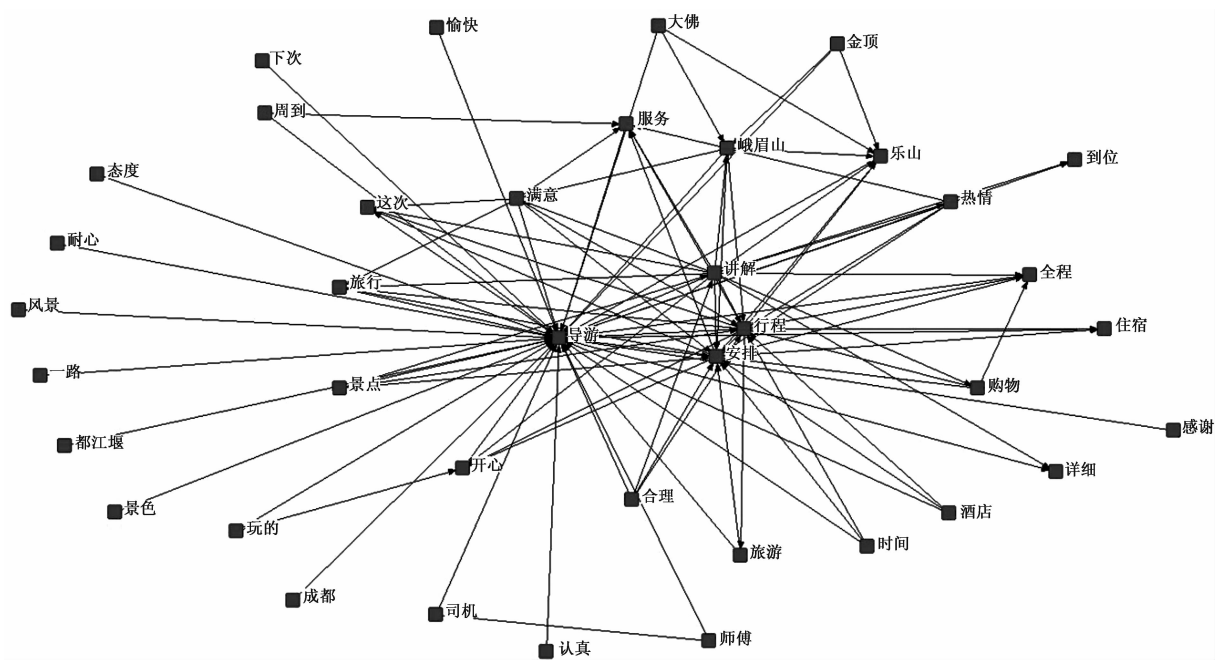

图 3　成都旅游影响因素网络语义图

周边的“住宿”“时间”等形成语义关系，行程一般是提前在旅游网站上预定好的线路行程，而实际旅游时由于当地情况变化，临时安排也成为决定游客体验的一个重要因素。此外，这几个节点也都指向了“司机”这个节点，可见司机情况也影响了游客在旅途中的感受。在现实情况中，跟团游的游客每天可能会有较长的时间在车辆中，司机也会相当程度上影响游客的体验。以“讲解”为节点的，与“景点”“热情”形成了语义关系，说明游客的旅游体验中，在景点能否受到好的服务也是很重要的一点。

综合以上分析，得出旅游满意度的影响因素大致包括“行程”“安排”“司机”“服务”“导游讲解”等。

1.2.2　基于 LDA 主题模型的影响因素挖掘

上述两种方法对两市旅游评论的特征进行了初步研究，下面选用 LDA 主题模型进一步分析。在 LDA 主题模型中，每一句评论可视为一个文档，找出文档的主题，通过观察主题中的特征词，最终归纳出影响旅游评论的特征因素。LDA 的数学模型过程描述如下：

词是组成语料的基本要素，词库中的词汇量视为 V，那么可以将该词表示成一个 V 维向量，这样第 v 个词出现时，即为向量 w 的第 v 个分量 $w^v=1$，其他分量 $w^u=0$（$v\neq u$）。

文档是由 N 个词组成的序列，可为 $d=(w_1,$

$w_2,\cdots,w_N$),其中 w_n 是文档中的第 n 个词。

文档集是 M 个文档组成的集合,可为 $D=(d_1,d_2,\cdots,d_M)$,生成的过程为。

选择 $N\sim$ Possion(ξ) 和 $\theta\sim$ Dir(α)。

在文档中生成第 n 个词 w_n :

依据多项式分布 $z_n\sim$ Mutinomial(θ) 抽样所得的主题 z_n ;

根据概率 $p(w_n\mid z_n)$ 抽样得到具体的词 w_n 。

给定参数 α 和参数 β,LDA 生成文档 d,N 个主题 Z,N 个词语 w 的联合概率分布为

$$p(d,z,w\mid\alpha,\beta)=$$

$$p(d\mid\alpha)\prod_{n=1}^{N}p(z_n\mid d)p(w_n\mid z_n,\beta)\qquad(1)$$

通过期望最大化算法求最大似然式(2)从而估计 α 和 β 的参数值,进而确定模型

$$l(\alpha,\beta)=\sum_{i=1}^{M}\ln\ p(d_i\mid\alpha,\beta)\qquad(2)$$

LDA 主题建模的主要问题在于主题数的确定,可用式(3)困惑度来判断,其判别标准为在合理数量范围内选择困惑度小的主题数为最优,困惑度的计算公式为

$$\text{perplexity}(D)=\exp\left[-\frac{\sum\ln\ p(w)}{N}\right]\qquad(3)$$

式中:$p(w)$ 为测试集中出现的每一个词的概率;N 为测试集中出现的所有词的个数。

分别将预处理好的数据输入 LDA 主题模型,此处使用 Python 机器学习工具包 scikit-learn 进行训练,选择 GIbbs Sampling 估计模型的后验参数。关于主题数目,首先根据词云图和语义图可大致确定主题数目的范围为 3～8,经过人工测试并分析困惑度发现,当主题数设为 6 个时,模型的困惑度较低,特征词拥有较好的分布,主题的区分度比较合适,模型的涵盖度较高。

综合 LDA 主题模型来看,景点体验、行程安排、酒店住宿、导游服务、导游讲解、地区旅游特质、司机情况这 7 种特征基本涵盖了游客对西安和成都旅游方面的大部分特征,因此可以选这 7 个因素作为特征因素作为进一步分析满意度的依据。地区旅游特质指的是某地的旅游特色。

1.2.3　满意度评价体系构建

1.2.3.1　词性标注

词性标注是对句中词汇确定词性,词性成分主要包括名词、动词和形容词等,为下一步确定特征对评论精细分类做准备。采用 Python 中的 jieba 模块对两市评论进行分词标注。

1.2.3.2　特征情感词对的匹配

用分词和词性标注的结果,结合特征词和情感词对的抽取,总结特征词与情感词共同出现的词法模板,根据这些模板匹配评论数据,从而得到所有的关系对。匹配好的词对数量和示例见表 1。

表 1　西安-成都两市评论各特征词对数量

单位:对

词对搭配	西安	成都	示例
名词＋形容词	7 056	6 653	导游不错
名词＋副词＋形容词	3 091	2 985	车程比较干净
动词＋形容词	1 012	689	游览舒适
双副词	317	335	导游都非常好

匹配之后进行特征情感词对的抽取。选用谷歌开发的 Word2vec 模型对已经确定的特征词对进行扩充,将这些词输入 CBOW 模型进行训练,最终计算出各个词汇的相似度,使用 K-means 方法对词汇进行聚类。将指标“景点”“行程安排”“酒店住宿”“导游服务”“导游讲解”“西安旅游特征”“成都旅游特征”“司机情况”使用上述模型,分别得出 7 个特征词对应的相似度最高的前 10 个词汇,得出与这 7 个特征方面关系相近的词汇,见表 2。

表 2　游客满意度指标和对应的相似词

指标	相似词
景点	时间;经典;游览;彰显;涵盖;调换;参观
行程安排	路线;线路;合理;规划;紧凑;安排;舒适;紧凑;舒适;感觉
酒店住宿	丰盛;酒店;干净;伙食;用餐;价格便宜;团餐;自助
导游服务	全程;负责;态度;专业;热情;耐心
导游讲解	解说;介绍;耐心;热情;认真;敬业;一路
西安地区旅游特征	风俗;文化;孤独;内涵;秦唐;气息;教育
成都地区旅游特征	峨眉;金顶;大佛;青城;修静;水利;佛教
司机情况	开车;责任心;技术;和善

在得出特征词相似词的集合后,下一步进行筛选特征情感词对操作,首先找出特定词性组合的词对,再进一步使用正则表达式方法找出含有特征词的词对集合,最后删去一些明显无意义和错误的词对,完成筛选。具体流程如图 4 所示。

抽取的词对示例:[旅游景点/n, 不错/a;景区/n];[最好/a 景点/n, 完全/a];[服务态度/n, 好/a 景点/n, 壮观/a];[整体/n, 不错/a 路线/n, 不错/a 导游/n, 不错/a]。

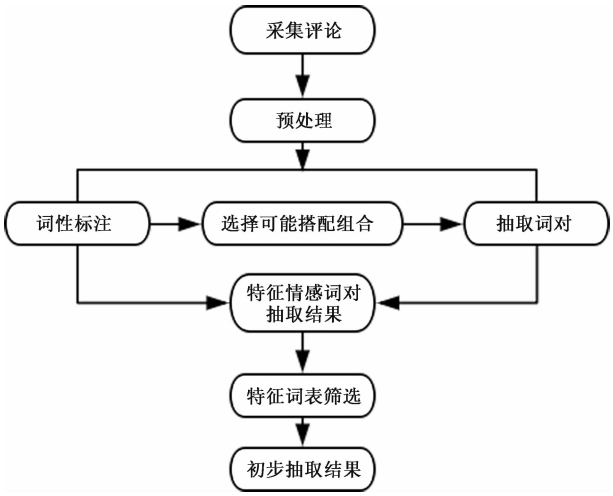

图 4　词对抽取流程图

2　游客满意度的度量

2.1　评论情感值的计算

2.1.1　构建情感词库

情感词典的含义是指表达人们日常语言中含有情感倾向词语所构成的词典，表示情感倾向的词语主要包括消极词和积极词。所使用的情感词典是知网情感词典和 Boson 情感词典。并将获取到的在线旅游评论中的网络词相关情感词也添加进去，如“太爽了”“炒鸡棒”“炫酷”等词。添加的方法仍然是将词缀为/a 的词用正则表达式的方法筛选出来，再人工筛选出其中能表示情感倾向的网络词语。

将上述相关网络词汇添加完毕后，就得到了可以使用的综合情感词汇表，此外还需要建立副词词表，因为“很”“非常”以及“有点”“稍微”“太”这些程度副词中蕴含了对于情感的暗中褒贬，隐藏了一定的情绪意义。当这些程度副词之前或之后跟随情感词时，这个词组表达的情感含义就将会和单独情感词表达的词义产生一定程度上的偏移。因此，在使用基于词典方法的情感分析时，需要将这些程度副词也加以挑选出并赋值。这里的副词词表一部分源于知网程度副词词表，一部分人工添加，这些程度副词分为 5 个等级，其中权重大小参考了已有的研究成果[28]，见表 3。

表 3　副词权重

程度	示例	权重
超级	极度；极端；绝	2
很，非常	实在；太；特别	1.5
较	较；较为；进一步	1.25
稍稍	略微；略加；挺	0.5
不足	不怎么；微；弱	0.25

否定词在性质上与程度副词的性质类似，但不同的是，否定词会直接改变情感词的原本情感指向。所使用的否定词词表是从互联网上获得的，再添加适合在线旅游评论的否定词。新添加的否定词的分数以－1 作为权值，共计 68 个。

2.1.2　短句情感值的计算

利用构建的情感词库对短句的特征词进行情感计算。通过确定特征值的位置，进一步确定特征词位置附近的相关词汇，进而综合计算出短句的情感得分。具体分以下几种情况，利用 Python 编程分别计算：

1）短句中不含情感词。识别出短句有特征词但没有情感词，则该短句的情感值为 0，可以将其视为中性句子。

2）短句中只包含情感词。即特征词与情感词组合，计算方法是依据情感词词表的相应权值，依次进行计算即可得出句子的情感值。

3）短句中包含了程度副词但不含否定词。计算方法为依据情感词表和副词词表的相应权重，综合计算情感值。

4）短句中含有否定词。共两种情况，一是特征词＋否定词＋情感词，二是特征词＋否定词＋副词＋情感词。其计算公式为

$$F = f(x)\prod_{i=1}^{n} A_i \prod_{j=1}^{m} g(y_j) \tag{4}$$

式中：n 代表副词个数；m 代表否定词个数；g 代表否定词权值的选定，否定词为奇数时，为－1，为偶数则为 1，再根据构建的情感词表、副词词表相应的权重，计算短句情感值。

2.1.3　西安-成都两市各指标评论情感分布

根据评论情感值的正负情况可以将评论分为正面情感和负面情感，两市各特征的情感分布情况如图 5 所示。

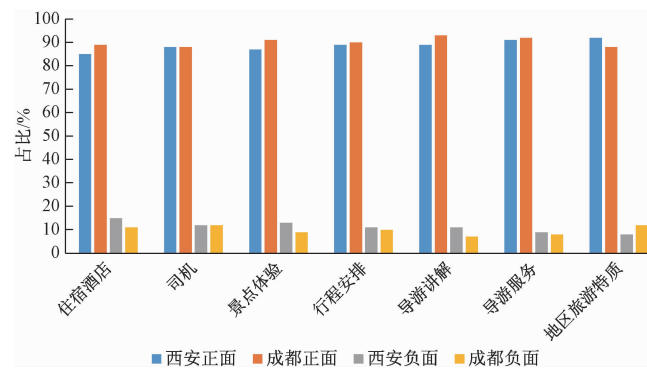

图 5　携程西安-成都旅游评论情感极性对比

由图5可以看出,两市所有特征的正面情感占比均在85%以上,在地区旅游特质方面,西安正面情感比例高于成都4个百分点;在司机服务方面,两市持平,占比为88%,相对于其他特征来说评价不算好;在其他5个方面,成都正面情感占比均高于西安。西安市的7个特征中,地区旅游特质的正面评论占比最高,而游客有关住宿酒店的正面情感占比最低;成都市的导游讲解与服务特征的正面情感占比最高,地区旅游特质和司机服务特征正面情感占比最低。

2.2　游客满意度指标体系构建与计算

2.2.1　确定指标权重

采用层次分析法计算各个评价指标的权重。层次分析法的判断矩阵构造并未采用传统的专家打分法,而是结合西安及成都客观数据,依据两指标的比例差,将其范围与标度对应起来。具体占比见表4。

表4　范围标度比例差

要素层	占比/%	指标层	占比/%
导游类 B_2	41.76	导游讲解 C_5	15.84
		导游素质 C_6	25.92
旅途安排类 B_1	69.25	住宿酒店 C_1	9.95
		司机 C_2	10.3
		景点体验 C_3	12.85
		行程安排 C_4	36.07
旅游特色类 B_3	8.77	地区旅游特质 C_7	8.77

根据表4得出的各指标之间的比例,构造判断矩阵并计算各指标的权重向量。判断矩阵采用Saaty的1～9级标度法,该法对比时采用相对尺度,以尽可能减少性质不同因素相互比较的困难,提高准确度。表4中计算出来的最大比例差为0.604 8,共有9级标度,因此以0.06为单位划分标度。

在使用层次分析法计算权重的过程中,需要对判断矩阵进行层次排序以及检验一致性。一致性检验是通过计算一致性指数CI与平均随机一致性指标RI的比值CR来检验,CI的公式为

$$\mathrm{CI}=\frac{\lambda_{\max}-n}{n-1} \tag{5}$$

$$\mathrm{CR}=\frac{\mathrm{CI}}{\mathrm{RI}} \tag{6}$$

式中:$\lambda_{\max}$为判断矩阵的最大特征根;RI则可以根据矩阵的阶数查表得出。当CR<0.1时,判断矩阵的一致性是可以接受的,若CR>0.1,则需要对矩阵进行修正。

根据计算可得,要素层CR=0.086,旅途安排类CR=0.002,导游类特色CR=0,皆符合一致性检验要求。整合上述各因素权重,计算综合权重,结果见表5。

表5　指标体系综合权重

要素层	权重	指标层	权重	综合权重
旅途安排类 B_1	0.729	住宿酒店 C_1	0.094	0.069
		司机 C_2	0.094	0.069
		景点体验 C_3	0.148	0.109
		行程安排 C_4	0.664	0.484
导游类 B_2	0.217	导游服务 C_5	0.667	0.144
		导游讲解 C_6	0.334	0.072
旅游特色 B_3	0.054	旅游特色 C_7	1	0.054

可以看出在指标评价体系中,旅途安排类因素在整体评价中相对重要,其次是导游类因素,而旅游特色因素虽受到游客关注但占比不高,在指标层的因素中,行程安排和导游服务最受游客关注。

2.2.2　西安-成都两市旅游满意度的计算

依据已经得出的权重和特征词情感值,游客的满意度计算公式为

$$C=\sum W_i\sum W_jX_j \tag{7}$$

式中:C代表游客满意度;W_i代表要素层的权重;W_j代表指标层的权重;X_j代表各个特征词的综合情感值。

根据式(7),计算得到各个指标下的满意度,最终得到两市的综合满意度见表6。西安的综合满意度为11.3,成都为13.207,表明总体上来说游客对成都的旅游满意度高于西安1.907。西安在景点体验和旅游特色方面的满意度高于成都,主要是其拥有悠久的历史文化古迹,旅游特色满意度高;导游服务、导游讲解、形成安排和司机服务方面,成都的满意度高于西安。特别的,在行程安排方面成都的满意度高于西安1.175,表明成都在行程安排方面比西安做得好。

表6　两城市综合满意度

城市	住宿酒店	司机	景点体验	行程安排	导游服务	导游讲解	旅游特色	综合
西安	0.698	0.723	1.520	5.247	1.637	0.459	1.016	11.300
成都	0.909	0.983	1.483	6.422	1.942	0.663	0.805	13.207

3 结论与讨论

在线评论数据反映了游客的游览体验和情感，既可以为潜在消费者提供旅游借鉴，也可以为城市改进旅游服务提供信息反馈。本文结合词云图、网络语义图和 LDA 主题模型挖掘游客满意度影响因素，并根据同义词聚合方法，提取 7 项特征作为评价指标。进一步，基于情感分析和层次分析法计算出两市各旅游评价特征的权重和得分，构建满意度指标评价体系。结果显示，满意度影响因素中最重要的是行程安排和导游，成都市综合旅游满意度略高于西安市。

筛选出的 7 项评价指标不仅适于西安成都两市，也基本涵盖了国内旅游体验的各个方面，具有一定的普遍性。从基于文本挖掘的旅游满意度评价体系构建过程与传统基于问卷调查方式对比来看，后者一般是事前已经设计好评价指标体系，然后结合评价指标设置问卷问题，而前者则需要首先挖掘数据确定特征指标，并结合情感分析方法，构建评价体系。总的来说，基于在线评论的旅游满意度分析方法样本的代表性较为充足，可以节省大量人力、物力成本。当然这并不意味问卷调查方法已不重要，由于在线评论没有显示评价者年龄、职业等信息，这正好可结合问卷调查等方法来弥补。因此，探索多源数据的融合技术，是下一步值得研究的问题。

参考文献

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[EB/OL]. [2021-08-27]. https://finance. sina. com. cn/chanjing/cyxw/2021-08-27/doc-ikqcfncc5270431. shtml.

[2] 前瞻产业研究院. 2020 中国在线旅游行业市场分析[EB/OL]. [2020-08-29]. https://bg. qianzhan. com/trends/detail/506/200828-e61bb246. html.

[3] 翟若琳. 基于 DEA 模型的西安城市旅游效率分析：与成都市对比[J]. 中外企业家，2019(4)：64-65.

[4] 欧启均，李振亭，周晓丽. 城市旅游客流的网络信息驱动特征变化研究：以陕西西安为例[J]. 生产力研究，2018(7)：72-77.

[5] 徐茜. 文化创意产业与旅游产业融合的机理和路径探析：以成都为例[J]. 中华文化论坛，2015(6)：118-122.

[6] 王红兰，李洪涛. 旅游服务质量评价研究综述与展望[J]. 旅游纵览(下半月)，2016(12)：51-52.

[7] 刘福承，刘爱利，刘敏. 游客满意度的内涵、测评及形成机理：国外相关研究综述[J]. 地域研究与开发，2017，36(5)：97-103.

[8] 袁静静，密艳秋，张静文，等. 消费者对烟台葡萄酒文化酒庄旅游满意度及参与行为研究[J/OL]. 酿酒科技：1-6[2019-09-12]. https://doi. org/10. 13746/j. njkj. 20191150.

[9] 夏瑞洁，谷亮亮. 服务质量对游客满意度影响研究：基于张家界国家森林公园的实地调查[J]. 中国市场，2019(27)：125-127.

[10] 杜羽焱. 基于 IPA 分析法的南京汤山温泉旅游满意度分析[J]. 管理观察，2019(18)：86-89.

[11] 张补宏，周旋，广新菊. 国内外旅游在线评论研究综述[J]. 地理与地理信息科学，2017，33(5)：119-126.

[12] 蔡珺哲. 基于文本挖掘的中国社交平台用户情感分析[D]. 上海：上海师范大学，2018.

[13] 张琰，俞越，潘华丽. 基于网络文本分析的不同档次酒店顾客满意度影响因素对比研究[J]. 旅游论坛，2017，10(3)：45-57.

[14] 黎晶. 基于决策树的移动互联网满意度评价研究[D]. 成都：西南交通大学，2018.

[15] 赵杨，李齐齐，陈雨涵，等. 基于在线评论情感分析的海淘 APP 用户满意度研究[J]. 数据分析与知识发现，2018，2(11)：19-27.

[16] 郭立秀. 基于文本挖掘的生鲜电商顾客满意度研究[D]. 成都：西南交通大学，2018.

[17] 王媛，许鑫，冯学钢，等. 基于文本挖掘的古镇旅游形象感知研究：以朱家角为例[J]. 旅游科学，2013，27(5)：86-95.

[18] 刘阳. 基于文本挖掘的在线旅游产品销量影响因素分析[D]. 北京：首都经济贸易大学，2018.

[19] 范宁. 基于文本挖掘在民宿满意度中的研究[D]. 桂林：广西师范大学，2019.

[20] 刘婷. 三亚旅游购物满意度研究[D]. 海口：海南大学，2014.

[21] CAMBRIA E，SCHULLER B，XIA Y，et al. New avenues in opinion mining and sentiment analysis[J]. IEEE Intelligent Systems，2013(2)：15-21.

[22] ZHANG L，LIU B. Identifying noun product features that imply opinions[C]//Meeting of the Association for Computational Linguistics：Human Language Technologies：Short Papers. Association for Computational Linguistics，2011：575-580.

[23] 丁蔚. 基于词典和机器学习组合的情感分析[D]. 西安：西安邮电大学，2017.

[24] 崔志刚. 基于电商网站商品评论数据的用户情感分析[D]. 北京：北京交通大学，2014.

[25] 刘楠. 面向微博短文本的情感分析研究[D]. 武汉：武汉大学，2013.

[26] LI H，YE Q，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)：784-802.

[27] 徐军，丁宇新，王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报，2007(6)：95-100.

[28] 徐琳宏. 基于语义理解的文本倾向性识别机制[C]//第

三届学生计算语言学研讨会论文集.北京:中国中文信息学会,2006:5.

[29] 张英. 基于深度神经网络的微博短文本情感分析研究[D].郑州:中原工学院,2017.

Research on Tourism Satisfaction of Western Cities Based on Text Mining:

Taking Ctrip Xi'an and Chengdu as examples

LI Shuang, ZHANG Zheng, LIU Yaya

((School of Statistics,Xi'an University of Finance and Economics,Xi'an 710100,China))

Abstract: Taking the travel commentary information of Xi'an and Chengdu on Ctrip as example, the influencing factors of tourism satisfaction are explored by using word cloud map, social network semantic graph and LDA topic model. Together with the emotion dictionary method to calculate index sentiment value and the analytic hierarchy process to determine index weight, the urban tourism satisfaction evaluation system is constructed based on online text comments. The results show that Chengdu's satisfaction is higher than that of Xi'an. Tourists are more satisfied with the tour guides and itinerary of Chengdu tourism service than Xi'an, while Xi'an has better regional tour features and scenic experience than that of Chengdu.

Keywords: text mining; sentiment analysis; satisfaction; LDA model; tourism