# 第一部分 问题提出和研究思路

## 一、问题提出

### （一）调查背景

### （二）文献研究

### （三）研究思路与创新

## 二、研究思路框图

数据分析

问题提出

方案设计

市场调研

# 第二部分 基于弥勒景区评论数据的挖掘与分析

## 数据来源背景

随着国内经济的持续增长，人们的生活水平质量的提升，旅游市场不可替代的成为了助力经济增长的主力军。据不完全统计，近几年来，旅游业带来的消费是居民消费领域增长最快的。而弥勒凭借其发达的交通以及独特的景观，旅游人次逐年上升，拥有丰富的旅游数据，从《弥勒市文化旅游产业发展报告》来看，2016年以来，推进红河水乡、东风韵、太平湖、可邑小镇、甸溪河治理等一批文旅重点项目，累计完成文化旅游重点项目固定资产投资43.04亿元，年均增长27.15%;另外, “十三五”以来至2019年，全市接待国内外旅游者从2015年的434.34万人次增加至1205.61万人次，实现旅游收入从26.51亿元增加至142.54亿元，分别年均增长29.06%、52.28%， “十三五”期间，全市累计接待游客4506万人次，实现旅游收入442.74亿元，分别是“十二五”期间的2.85倍、5.08倍，占全州接待游客17.96%，旅游收入15.35%。接待游客年均增长23.35%，旅游收入年均增长38.45%.除此之外，从最近的旅游数据来看，2021年弥勒市第一季度旅游接待游客152.84万人次，同比增长74.32%，旅游收入16.64亿元，同比增长58.85%。弥勒市“五一”小长假共接待游客893599人次，旅游总收入60826.27万元，人数及收入同比分别增长61.82%、35.25 %。端午节假期间，共接待游客32.36万人次，旅游总收入2.23亿元，人数及收入同比分别增长51.16%、37.47%。

弥勒市的旅游业正处于蓬勃发展阶段，现如今大部分消费者的决策往往会收到评论的影响。市场分析显示：超过75%的消费者在购买商品前都会参考互联网用户发表的在线评论；超过90%的企业相信，在影响消费者购买决策的决定性因素中用户评论占据重要的位置。然而海量评论让消费者难以找到对自己有用的信息，而那些截然相反的评论会让消费者更难做出判断。因此，对在线评论进行情感分析，得到书面语言的观点、情绪、情感、评价、态度等消费者主观感受显得尤为重要。

大部分富有价值的信息往往杂乱的分布在海量的文本数据中。运用自然语言处理技术和机器学习的方法处理非结构化或半结构化的文本数据成为大数据时代的一个关注热点。本次文本评论数据的来源渠道主要是携程、马蜂窝、美团、等旅游 APP，通过对云南红酒庄、东风韵、太平湖森林小镇、万花谷玫瑰庄园温泉、甲玉泉·天池都温泉度假村、鑫甲玉泉温泉旅游度假村、锦屏山风景区、可邑小镇、湖泉温泉水世界、湖泉半山温泉十个个弥勒主流景点评论数据进行挖掘分析，充分利用这些信息，挖掘消费者心声，探究游客对弥勒主流景点总体印象， 为弥勒景点未来维持可持续发展提供改进方向指引。

根据上述数据来源背景及实际情况，本节通过python 语言编写好数据采集逻辑后，运行程序爬取主流弥勒景点评论数据，通过 NLP 分词方式进行词频统计，知网 Hownet 情感词典进行情感分析、构建语义网络，gephi 进行词共现分析， 实现对评论数据的隐藏信息的挖掘与分析，进而优化实地调查环节中的调查问卷，改进实地调查方案。

## 数据的搜集

### （一）有效数据的爬取

通过对各个平台汇总后，本次数据共搜集评论条27356条（其中马蜂窝126条，美团24675条，携程旅行2555条），通过程序算法，抓取无效评论11151条，有效率仅有59.07%(分析发现有效率低的原因是大部分游客敷衍评论以及平台为刷评分水军过多) 此次分析中，将非结构化的文本数据转化为可检索的结构化数据，用表格形式储存，结构示意图如下：

|  |  |  |
| --- | --- | --- |
| 用户名称 | 评论时间 | 评论内容 |
| 肯羊吃不饱 | 2021年10月5日 | 国庆人很多，有游泳池和温泉泡池，还是很好玩的。 |
| 黄国荣゜ | 2021年10月11日 | 不记得了，很久以前去的。感觉离城很远 |
| wxjwqx | 2021年8月23日 | 标间窗户低下就是温泉池，大晚上十点多还有孩子在嘻嘻哈哈，比较吵 |
| 匿名用户 | 2021年5月25日 | 非常好的一次体验，下次还要带上朋友一起去 |
| 叫你和头 | 2021年8月29日 | 只能说还可以，四个池子，其他的都是单项目收费 |

表× 从美团旅行平台爬取评论结构示意表

|  |  |  |
| --- | --- | --- |
| 用户名称 | 评论时间 | 评论内容 |
| 偏执老猫 | 2021-10-15 | 人工景点，建筑有点风格。大片的花海，有山有水，路过的话在里面散散步还是挺好的。 |
| 行走的兔子 | 2021-10-31 | 景点建筑有特色.建筑师的眼光独特.蚁工坊和东风韵两个景点都去了.很满意 |
| 川南一纤夫 | 2021-10-31 | 建筑还可以，以知青文化为支撑，内容很少，好在适合于年青情侣拍照，交通也还方便！ |
| 追风筝的人23333 | 2021-05-18 | 非常值得一去，门票40很值得，就是一定要做好防晒，不要像我一样被晒伤！ |
| 簋鱼 | 2021-09-27 | 风景很美 非常适合打卡 从弥勒高铁站、市区前往景区都非常的方便 工作人员的态度也很和善 介绍了很多注意事项 详细看图吧！ |

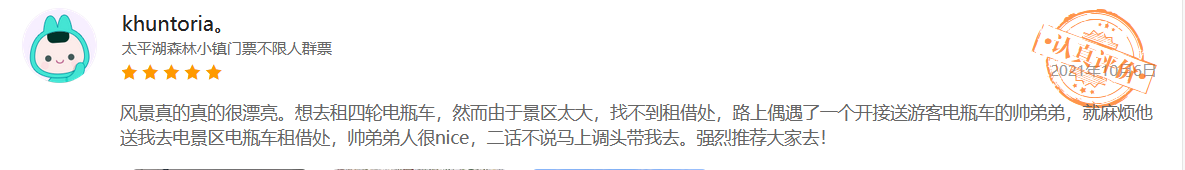
表× 从携程旅行平台爬取评论结构示意表

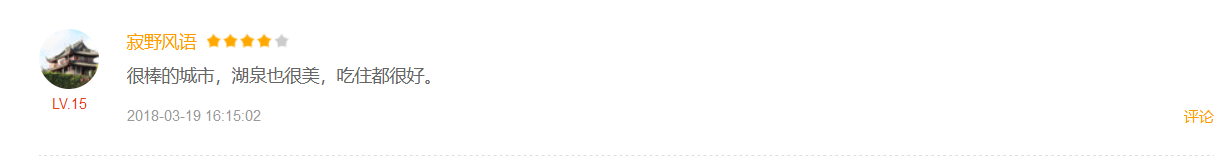
|  |  |  |
| --- | --- | --- |
| 用户名称 | 评论时间 | 评论内容 |
| 缘来是你 | 2019-08-10 17:15:14 | 好玩，好吃，风景优美，景色怡人。 |
| 沙鹰 | 2019-06-12 15:09:30 | 湖泉生态园的景色，服务，食品相当不错，已经来了好多次了。 |
| 平常心 | 2018/11/20 15:55 | 湖泉生态园，一个美美的免费景区，值得一去哦！ |
| 小兔 | 2018-08-31 10:40:10 | 不错，尤其是里面那个酒店非常好，适合度假 |
| 铁胆神侯云中鹤 | 2018-08-21 20:10:15 | 一般吧，坏境还是很不错的，晚上散散步还不错 |

表× 从马蜂窝旅行平台爬取评论结构示意表

分析发现携程、马蜂窝、美团三个平台评论有效率处于前三位，分别截取三大平台详细评论界面示意图如下：

图××× 携程旅行用户对东风韵景区评论结构示意图



图××× 美团用户对东风韵景区评论结构示意图

图××× 马蜂窝用户对湖泉生态园评论结构示意图

### （二）无效数据处理

对于旅游第三方平台的评论数据来说，由于存在雇佣水军采取作弊行为人为刷高销量好评的现象。未经处理的文本中通常包含大量重复性评论、无语义评论，例如数字、字母和网络语义的特殊字符，以及“该用户未填写评论”或“默认好评”等类似网站自动生成的评论文本以及字符长度小于2 的极短评论，这类评论内容传递的信息量少，且增加文本分析的复杂度，容易造成高稀疏性文本模型， 因此进行原始数据的清洗尤为重要。

初步预处理的评论文本根据《哈工大停用词词库》和自建旅游类专有名词词典，使用jieba分词工具去除评论文本中停用词、识别旅游类专有词项，并逐条分词解析评论文本，形成由多个词项构成的字符串集合。因此，本节共对11151条之多的无效数据进行爬取并删除， 具体删除模式如下：

1、部分平台存在系统默认顾客评论行为，这些数据大量重复，不具有分析价值，予以删除，例如：

|  |  |  |
| --- | --- | --- |
| **Amandayoyotony** | 2021-08-09 21:05 | 该用户没有填写评论 |
| **AA万通庹庹庹** | 2021-10-13 00:17 | 该用户没有填写评论 |
| **BjergsenSun** | 2021-08-28 21:55 | 该用户没有填写评论 |

图×无效数据示例图1

2、部分用户为评论内容完全为英文字母、数字和标点符号的，视为随意发表的评论，予以删除， 例如：

|  |  |  |
| --- | --- | --- |
| **TFm340203386** | 2021-09-28 00:39 | 。。。。。。 |
| **能举铁的小燕子** | 2021-08-12 11:49 | ？？哈哈哈火狐后 |
| **小宇哥哥你好帅** | 2021-08-15 13:30 | 648ewhfyexdwqs |

图× 无效数据示例图2

3、部分平台雇佣水军进行随意评论，此类评论用户ID 相同，评论内容不具有分析价值，予以删除，例如：

|  |  |  |
| --- | --- | --- |
| **枫唁** | 2020-08-15 21:46 | 真不错 |
| **枫唁** | 2020-08-16 21:30 | 真不错 |
| **枫唁** | 2020-08-17 21:52 | 真不错 |

图× 无效数据示例图3

## 数据处理

### （一）文本数据特征抽取

在线评论是由自然语言构成的文档数据集合，每个文档由若干词项以一定语义逻辑组合而成。根据词项在文中出现的频率及其表达的特定主题，采用向量空间模型（Vector Space Model，VSM）对海量文本建立文本特征模型，使文本转化 为可量化表征的结构化数据进行特征挖掘。VSM基本原理是评论文档Di能够表示为Di=D (T1,Wi1;T2,Wi2;...;Tm,Wim）的文档集合，其中（T1,T2,...,Tm）为一个m维互异词项集合，（Wi1,Wi2, ...,Wim）为对应 m维词项在文档中的权重，即在文档Di中的重要程度， Wij一般定义为在文档Di（i=1,2,3,…,n）中词项 Tj （j=1,2,3,…,m）出现频率（Term Frequency, TF）的函数Tf（dij），本节中构建的文本特征模型见表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **文档** | **环境** | **温泉** | **设施** | **...** | **词项** |
| **文档 1（D1）** | w11 | w12 | w13 | ... | w1m |
| **文档 2（D2）** | w21 | w22 | w23 | ... | w2m |
| **文档 3（D3）** | w31 | w32 | w33 | ... | w3m |
| **文档 4（D4）** | w41 | w42 | w43 | ... | w4m |
| **︙** | ︙ | ︙ | ︙ | ︙ | ︙ |
| **文档n（Dn）** | wn1 | wn2 | wn3 | ... | wnm |

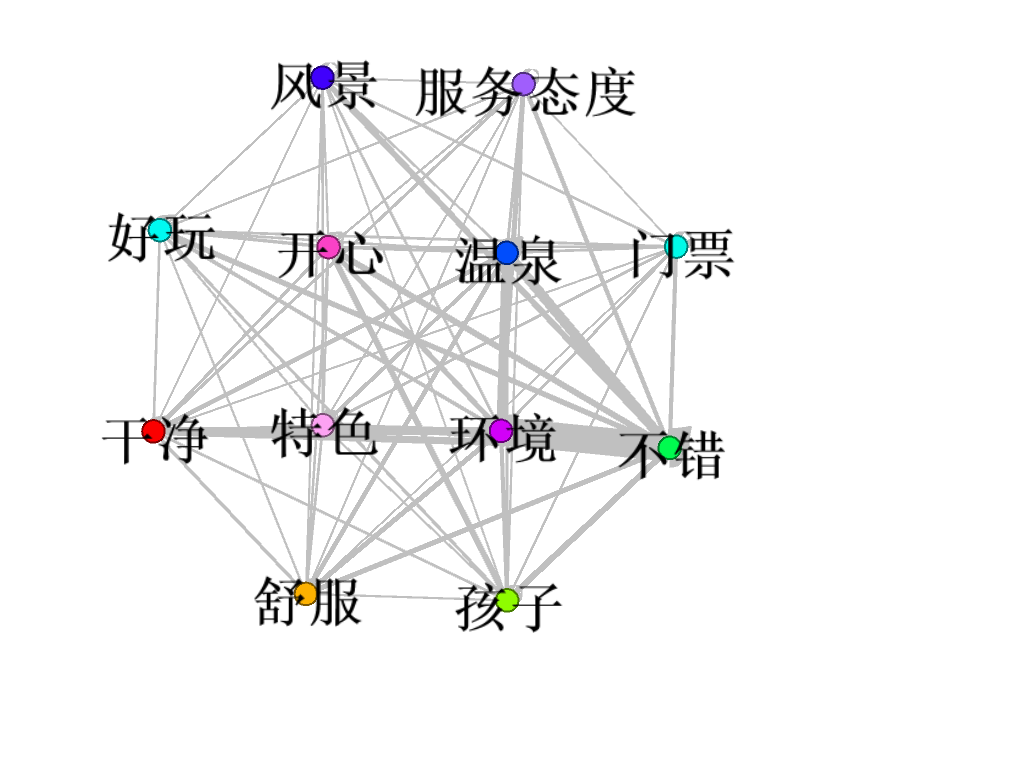
表XXXX

VSM 以词项在文档中权重系数Wij 构建的m×n 维文档-词项矩阵（Document Term Matrix， DTM） 集合了评论文本中的所有词项，导致DTM 通常具有高稀疏性和数据冗余性。因此，需采用词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）算法对DTM 提取文本特征。 TF-IDF 是Salton 等提出的单词权重统计方法。其中，TF 为词频，指某一给定词项在该文件中出现的次数；IDF 为逆文档频率，指含有该词项的文档数在总文档中所占比例取逆后的对数值，表征该词项区分文档的能力。当词项在一篇文档中出现频率越高，同时在其他文档中出现次数越少，表明该词项对该篇文档的区分能力越强，其权重 TF-IDF 则越大。TF-ID 计算公式为：



式中：为词项在文档中出现频率； 𝒅𝒇𝐢为出现 词项的文档数； N 为总文档数。

### （二）词共现分析

为了理清关键词之间的关系, 利用gephi共现分析数据中 12 个高频关键词进行共词分析, 生成 12\*12 的共现矩阵, 生成词共现网络图。相关矩阵中的数值(绝对值)越大,表示对应关键词的关联程度就越高，词共现网络图各个关键词连线越短，说明相应关键词联系越紧密。 

图XXXXX

### （三）词频统计及分词

为了较全面地反映游客对弥勒景区的总体印象，本节选取中高频关键词作为分析对象,这样既能避免词语数量太少导致的结果模糊，又能保证核心关键词的完整，不会产生词量过多的负担。下表格是部分累积词频分析情况，R\_S是高频词顺序，W\_N是词名，W\_F是词频，P\_S是词性。

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 标签名 | 词频 | 词性 |
| 1 | 环境 | 3286 | 名词 |
| 2 | 温泉 | 2035 | 名词 |
| 3 | 小孩 | 1642 | 名词 |
| 4 | 设施 | 1476 | 名词 |
| 5 | 弥勒 | 1446 | 名词 |
| 6 | 感觉 | 1437 | 名词 |
| 7 | 地方 | 1399 | 名词 |
| 8 | 项目 | 1077 | 名词 |
| 9 | 门票 | 978 | 名词 |
| 10 | 价格 | 977 | 名词 |
| 11 | 景区 | 951 | 名词 |
| 12 | 时候 | 878 | 名词 |
| 13 | 池子 | 827 | 名词 |
| 14 | 特色 | 725 | 名词 |
| 15 | 朋友 | 676 | 名词 |
| 16 | 东西 | 618 | 名词 |
| 17 | 风景 | 607 | 名词 |
| 18 | 建议 | 590 | 名词 |
| 19 | 服务态度 | 576 | 名词 |
| 20 | 时间 | 538 | 名词 |

表XXXX部分词频统计分析表

不同于在英文中，每个单词之间都有空格来隔开。在中文里，只有字、句和 段落能够通过明显的分界符进行快速的划界，对于“词”和“词组”来说，其边界模 糊没有一个形式上的分界符。而词是最小的，能够独立运用的有意义的语言单位，文本分词的工作也就是将连续的字序列按照一定的规范重新组合成词序列。本文的分词工作是使用集搜客平台分析后得出部分词云图，如下：



图XXXX 部分词云图

上图中，词频越高，词的字号就越大。观察上图我们发现，弥勒景区最大的特色就是环境优美，从高频词可以看出，“温泉”是弥勒景区的特色，“服务”好坏和卫生环境的“干净”与否是游客关注的重点，其中也有部分体验、感官等词汇。进行词云图的分析可以便于得出弥勒景点吸引游客的关键点。由于机器不能直接识别语句的整体结构思想，在经过上一步的分词工作后，为了挖掘出各个词之间的联系和语义结构，得出单独的词语无法清晰表达的内容，基于分词后的文本，抽取其中的高频词和行特征词，得到一个共现矩阵，以此建立语义网络图。

## 四、基于情感词典的情感倾向性分析

情感倾向分析的通用方法都是对大规模语料集进行统计分析、预先对有代表性的词语采用人工标注方法选为基准词，然后对候选词作语义相似度计算来获取新的情感词，从而扩展情感词典的覆盖面。本文基于已有资源，使用中国知网发布的“情感分析用词语集”，作为判断文本情感倾向、情感极性的参照，进行情感倾向性分析的方法如下述公式：



上式中表示文本的情感极性值，表示文本中包含的正面情感词个数，表示文本中包含的负面情感词个数。对经过预处理的文本评论进行判别分析后，结论如下：

对评论数据分析指出，对于弥勒景点的游客认可度，52%左右的游客对弥勒市有好感，负向情绪占11%左右，另外还有37%的游客对弥勒景区持中性情绪。所以，如果弥勒市想要有更好的发展，了解游客的偏好和对景区相关设施的满意度是必须要做的事情

# 第三部分 调查方案设计

## 一、调查目的

## 二、调查内容

## 三、调查对象

## 四、调查方式

## 五、抽样设计

### （一）目标总体

### （二）抽样方法

### （三）样本量的分配

### （四）调查样本实施方法及过程

### （五）调查实施

#### 1.正式调查

#### 2.调查数据量控制

##### （1）调查前准备

##### （2）调查中的质量控制

##### （3）调查后处理

##### （4）数据录入阶段的质量控制

#### 3.数据预处理

##### （1）数据录入

##### （2）数据审核

##### （3）缺失值处理

### （六）数据分析方法简介

#### 1.描述性统计方法

#### 2.决策树

#### 3.Spearman秩相关检验和Kendall's τ检验

#### 4.结构方程模型

#### 5.文本挖掘

# 第三部分 调查数据分析

## 一、问卷效度和信度检验

## 二、“智慧旅游”总体认识描述分析

### （一）游客基本特征

#### 1、性别分布

#### 2、年龄分布

#### 3、地区分布

#### 4、职业分布

#### 5、月消费分布

### （二）景点认知渠道及宣传效果

### （三）游客对弥勒景区“智慧旅游”的体验情况

#### 1、各方面的满意程度

#### 2、“游云南”使用情况

#### 3、景区直播观看率

#### 4、景区预约及景区开放查询功能体验满意程度

#### 5、智慧停车场使用满意程度

#### 6、智慧厕所使用满意程度

#### 7、弥勒景区5G网络及无线网络覆盖满意程度

#### 8、一码通功能使用满意程度

#### 9、扫码识景功能使用满意程度

### （四）弥勒景区智慧旅游期望情况

#### 1、景区文化

#### 2、基础设施与配备

## 三、弥勒景区游客特征模型分析

### （一）二元选择模型的确立

### （二）二元选择模型的建立

### （三）对网红景点游客群体特征的研究

### （四）模型结果分析

#### 1、性别

#### 2、年龄

#### 3、收入

## 四、智慧旅游满意度影响因素分析

### （一）排序选择模型的建立

### （二）模型估计和分析

#### 1、模型估计

#### 2、平行回归假设检验

### （三）模型结果分析

#### 1、消费水平

#### 2、旅游频率

#### 3、旅游目的地偏好

#### 4、性别

#### 5、年龄

# 第四部分 结论与建议

## 一、结论

## 二、建议

# 参考文献

# 附录