东北大学秦皇岛分校

Northeastern University at Qinhuangdao

基于在线评论情感分析、AHP 和直觉模糊-TOPSIS 方法的共享单车品牌选择模型

Brand selection model of shared bicycle based on online comment Sentiment analysis, AHP and Intuitionistic Fuzzy Sets-TOPSIS

院　　别	管理学院
专业名称	信息管理与信息系统
团队成员	b1ub1u
指导教师	赵　　萌

2022 年 6 月

基于在线评论情感分析、AHP 和直觉模糊-TOPSIS 方法的共享单车品牌选择模型

摘要

随着社会的不断发展，经济、素质、科技等方面都有了巨大的提升，人们越来越注重绿色生活与可持续发展的理念。伴随着共享经济模式在社会各行业的广泛应用，共享单车这一“解决城市最后一公里”问题的绿色出行方式应运而生，并且发展成为了一个庞大的行业，具有众多单车品牌。本文旨在通过对在线评论进行抓取并进行情感分析，确定所要研究的单车品牌与评价指标，结合 AHP 方法进行各项指标综合权重的确定，最后通过直觉模糊-TOPSIS 综合评价方法进行结果的一致性检验，确定最终的排序序列。本文利用问卷结果对美团单车、青桔单车、哈啰单车三个单车品牌进行了比较分析，从而得到了可供用户进行选择的优先排序序列，为用户进行选择提供了合理参考依据，有利于用户得到最大满足，解决了在众多各有利弊的品牌与褒贬不一的用户使用评价中，如何帮助用户根据自身需求，选择最优的单车品牌，得到最大化满足这一问题。该研究对于用户进行选择以及单车品牌对于自身进行改进、创新，提高单车质量，调整价格策略和提升用户体验等方面具有重要参考价值。

关键词：在线评论；情感分析；AHP；直觉模糊集；选择犹豫性；TOPSIS 方法

Brand selection model of shared bicycle based on online comment Sentiment analysis, AHP and Intuitionistic Fuzzy Sets-TOPSIS

Abstract

With the continuous development of society, the economy, quality, science and technology have been greatly improved. People pay more and more attention to the concept of green life and sustainable development. With the wide application of the sharing economy model, sharing bicycles, a travel mode that "solves the problem of the last mile in the city", came into being, and has developed into a huge industry with many bicycle brands. The purpose of this paper is to determine the bicycle brand and evaluation index to be studied by capturing the online comments and conducting emotional analysis, and determine the comprehensive weight of each index in combination with AHP method. Finally, the consistency of the results is checked through the intuitive fuzzy TOPSIS comprehensive evaluation method to determine the final ranking sequence. In this paper, the questionnaire results are used to compare and analyze the three bicycle brands of meituan bicycle, Qingju bicycle and hello bicycle, so as to obtain the priority sequence for users to choose, which provides a reasonable reference for users to choose, and is conducive to users' maximum satisfaction. It solves how to help users according to their own needs in the use evaluation of many brands with advantages and disadvantages and users with mixed praise and criticism, Choose the best bicycle brand to maximize the satisfaction of this problem. The study has important reference value for users to choose and for bicycle brands to improve and innovate themselves, improve the quality of bicycles, adjust price strategies and improve user experience.

Keywords：Online comments；Sentiment analysis；AHP；Intuitionistic fuzzy set ; Choice hesitation；TOPSIS

目录

1 研究问题背景

随着社会经济的不断发展，人们对于美好生活的需要日益增长。为了解决城市“最后一公里”的问题，符合绿色共享理念的共享单车应运而生。共享单车不仅节约了用户等待出租车的时间成本，降低了服务的费用成本，还大大减少了碳排放，符合绿色出行理念。自 2006 年出现首家专业租车网站以来，中国在线出行服务业经历了“线下重资产+线上服务”向“互联网+共享经济/轻资产重服务”的转变，同时也实现了 PC 端向移动端使用场景的转变；作为在线出行业的主流服务，共享出行包含专车、快车等网约车服务，分时租赁服务，以及 2016 年火爆市场的共享单车服务，相比分时租赁，共享单车使用方便，取车还车灵活，使用性价比高，目前共享单车用户覆盖率增长迅速，远超分时租赁用户群体。

但是共享单车种类繁多，品牌琳琅满目，骑行后的体验评价褒贬不一，使得消费者眼花缭乱，但更多的人没有时间和精力去逐一比较从而挑选出最舒服的骑行品牌。伴随着网民数量的增长，越来越多的评价信息在网上出现。用户可以在微博、美团等公共平台上发布对于共享单车的使用体验，表达自己的情感态度。在商品和服务评论分析问题中，对文本评论进行筛选降重，抓取主要关键词，识别评论中的情感倾向性，对用户挑选商品，以及商家改进商品/服务具有一定的辅助作用。在大数据时代，通过情感分析实现有效的主题抓取，以便进行后续的其他分析，是一个有巨大意义的任务。

2 研究现状综述

互联网上关于商品以及消费的评论往往具有与其密切相关的主题，基于主题的情感分析方法，可以快速在众多纷杂的评论中，抽象出其共有的某个主题，快速挖掘出文本主题之间的相似性和与之对应的情感信息，从而对大量信息进行筛选，可以极大的提高情感分析的效率。并且可以优化传统情感分析方法中存在的只能判断整句评论的情感信息，无法大量缩减信息数量，以及无法对深层次语义对象进行情感分析的问题[1]。

AHP 方法是由美国运筹学家 Thomas L.Saaty 提出的一种用于解决复杂问题进行决策排序以及弥补主观定权方法缺陷的评价方法。AHP 以系统分层分析为主要思想，对评价对象总目标进行连续性分层分析，进行两两比较确定各项指标权重，最后进行加权求出综合权重，来实现对目标的排序[2]。

模糊综合评判法是多属性决策方法之一，近年来，直觉模糊集在不确定多属性决策中得到了广泛的应用[3]。直觉模糊集，考虑到了人们对事物评判所具有的一定犹豫性，是一种更符合实际情况的决策方法。TOPSIS 方法利用归一化的决策矩阵，找出理想方案与负理想方案，通过各方案到理想方案与负理想方案的距离，来评价对象的优劣[4]。对于直觉模糊集与 TOPSIS 方法的结合应用，是近年来的研究热点，有多篇基于 TOPSIS 原理探讨直觉模糊多属性决策方法的文献。

3 系统环境分析

3.1 政策环境

在社会经济达到现如今的发展水平之后，人们开始意识到可持续发展的重要性，绿色发展已经成为一个重要的理念和趋势。在全世界范围内，很多国家把发展绿色产业，突出绿色的理念和内涵作为自身发展道路中的重要一环。绿色发展是对可持续发展的继承,是可持续发展的一种实现，也是可持续发展中国化的理论创新，是中国特色社会主义应对全球生态环境恶化的客观现实做出的重大理论贡献,符合历史潮流的演进规律。绿色发展理念,多次在国家重要会议上被提起，并已经成为我国五大发展理念之一。中共十九大报告明确指出：加快建立绿色生产和消费的法律制度和政策导向，建立健全绿色低碳循环发展的经济体系[5]。可以见得，共享单车这一行业作为绿色发展理念和碳中和理念的重要践行，其发展的政策环境可以说是十分优越的。作为政策支持的发展行业，具有无比强大的活力性。

3.2 经济环境

随着科技水平的不断发展，互联网+、物联网等技术在社会生活的各个方面得到广泛应用，使得社会网络生态日益成熟。共享经济这一全新的经济概念，逐渐成为潮流,共享单车、共享汽车、共享医疗设备等共享行业如雨后春笋般涌现。目前，共享经济已经广泛且深入的渗透了社会的各类产业，为产业创新与转型升级提供了巨大的推动力。共享经济模式的本质，其实是实现资源的优化配置，使其拥有共享渠道的商业运营模式。共享模式极大地满足了绿色发展可持续的需求。共享模式实现了消费模式从“扔掉型”转变为“再利用型”；实现了物品从“占有空闲”转变为“共享使用”。通过社会存量资产的调整，实现了商品价值的最大程度利用。每辆共享单车平均可替代 20 辆普通自行车，这不仅大大降低了，

资源的闲置率，同时也减少了因购车后边际成本降低而诱增的无效出行。可以说，共享模式以接近免费的方式分享绿色能源和一系列基本商品和服务，是最具生态效益的模式，也是切实可行的可持续发展模式。

3.3 社会发展情况

党的十八大以来，我国经济社会发展和生态文明建设取得了具有里程碑意义的重大成就。十年来，我国扎实推进绿色发展，使生态环境状况实现了历史性的转折。植树造林占全球人工造林的 1/4 左右，单位 GDP 二氧化碳排放量累计下降了大约 34%。绿色理念越来越深入人心。这也是扎实推进共享发展的十年。我们历史性地解决了困扰中华民族几千年的绝对贫困问题，近 1 亿农村贫困人口全部脱贫，为世界减贫事业作出了巨大贡献。我们建成了世界上规模最大的教育体系、社会保障体系和医疗卫生体系，基本养老保险参保人数达 10.3 亿，基本医疗保险的参保人数增加到 13.6 亿，人均预期寿命由 75.4 岁提高到了 77.9 岁。人民生活的质量和社会的共享水平取得了历史性进步、全方位跃升。教育的普及使得社会整体素质得到了提高，使得“共享”成为了一种新可能。保障体系的建立健全使得人们的追求不再是温饱，而是吃的好、穿的好，人们对于美好生活的需要不断提高[5]。共享单车不仅贴合绿色共享的理念，而且能很好的解决城市最后一公里问题，可以在满足人们出行需要的同时，符合社会绿色发展可持续的理念。因此，被社会广泛认可。

3.4 科技发展水平

在大数据与信息化时代，互联网技术越来越普及，物联网技术也得到了更加深远的发展，物联网技术实现了物与物、物与人的泛在连接，实现对物品和过程的智能化感知、识别和管理。可以实时采集任何需要监控、连接、互动的物体或过程，共享单车正是物联网系统的产物。共享单车停放在路边，通过 GPS 定位模块，定期将定位信息告知给设备商的云服务器，当用户想要使用时，可以通过手机 APP，访问云服务器的数据，查看周边的单车停放位置信息。当用户来到单车旁边，扫描单车二维码，就可以通过 APP 获取单车 ID，发送开锁信息给云服务器，云服务器发送开锁信息给单车。在这之后，单车通过 GPRS 通讯模块收到解锁命令，就会由主控模块控制车锁进行解锁。用户也会收到解锁成功的消息，并进入计费状态。共享单车，对网络的要求并不是大数据量，所以早期采用短信的

方式实现单车与云服务器之间的信息传递。但随着 5G 技术以及正在试行的 6G 技术的落实与普及，单车与云服务器之间传输信息的速度将会越来越快，从而可以缩短开锁时间，使用户有更好的体验。

4 目标分析

本文主要研究大众对于不同共享单车品牌的不同体验因素的在线评论以及问卷填写情况，所体现出的偏好性以及由问卷结果得来的客观数据，旨在建立一个基于在线评论情感分析，结合 AHP 方法、模糊综合评价法和 TOPSIS 方法的共享单车选择推荐模型，实现根据用户真实体验以及在线评论，对不同的共享单车品牌进行优先级排序，从而供用户参考，选择最佳方案的目的。并且可以据此，进一步向共享单车行业发展提出改进建议，使其更加满足大众消费心理，从而增大共享单车使用用户及使用频率，更好的践行绿色发展可持续的理念。

5 系统结构及其影响要素分析

5.1 用户骑行评论的作用

在“互联网+”时代下，网络评论因满足用户在生活中分享欲而迅速走红，与此同时，用户体验与消费者对服务的情感态度对消费者对服务品牌的影响也越来越重要，因此，通过对于各种品牌单车的在线评论进行数据分析，可以更好的了解消费者暴露出来的选择偏好，更好的关注到消费者的的痛点，由此也可以帮助消费者做出更加符合自身需要的决策，同时也可以为共享单车的品牌商提供一定的可行的改进意见，促进共享单车行业的健康发展，推动共享经济更好地融入人们日常生活。

5.2 影响用户选择的因素

根据问卷结果以及用户在线评论抓取分析，确定了本次研究主要评价因素。

5.2.1 停靠点数目与范围

停靠点数与范围是影响用户选择的首要因素，只有单车可存放地点范围包含用户目的地或者在其附近时，用户才会选择骑行。毕竟选择共享单车的目的也是解决“最后一公里”，同时也要方便存取停靠，即“好借好还”。

5.2.2 价格

商品价格一定要匹配商品的作用大小，价格也就成为决策的重要因素，即使服务的可预料性极佳，倘若寄托于过高的价格，也无法让用户为此买单。性价比

是当今用户十分重视的消费因素。立足于前往目的地需求，倘若价格如果高，大多数客户就会选择出租车等服务，导致失去客户。共享单车的价格追求也就成了对于服务性价比的追求。

5.2.3 开锁速度

开锁速度对用户是否使用起着潜移默化的影响，开锁是骑行使用的第一步，侧面上决定着用户对于此次服务的初印象，虽然影响弱于停靠点数目与范围和价格，但是良好的初印象对于用户的感情倾向趋于积极方向。决定选择之后，骑行前开锁速度快慢、是否有广告弹出等影响开锁到使用快慢的因素都是影响用户决策的的因素。

5.2.4 骑行体验

骑行体验是用户在骑行过程中的感受、对于共享单车实体自身的评判，包括共享单车的完整性、座椅的舒适程度、携带物品放置的位置、电量充足、骑行速度等。其中，共享单车的完整性主要是共享单车是否已被破坏，影响总体的美观性，应满足部分用户的审美需要；座椅舒适度以及是否方便随身物品的放置追求用户使用的舒适度和便捷性；电量充足可以避免中途换车的糟糕体验，避免使用增加麻烦；骑行速度体现着用户对于时间利用的需求，提高用户的时间利用率也会增加用户的好感。因此，整体的骑行体验对于用户的选择也发挥着重要影响。

6 备选方案

根据前期发放调查问卷，进行结果回收，统计了 174 份有效结果，以此为根据，选定了，哈啰单车、美团单车、青桔单车为本次研究的对象。利用 Python 爬取微博中关于三种单车品牌的在线评论，并利用情感分析技术，提取评论主题，汇总主题，建成可视化词云，展现用户对于单车使用体验的倾向性，而后利用 AHP、直觉模糊-TOPSIS 方法进行相关分析，以得出三种单车品牌的优先排序序列，为用户进行选择以及单车品牌进行改进提供合理依据。

本次研究根据在线评论分析结果与问卷调查结果，选取“停靠点数目及范围”、“价格”、“开锁速度”、“骑行体验”4 项因素为评价指标。

7 建立总体模型

7.1 总体流程

首先通过发布问卷调查获取消费者在选择共享单车时的影响因素以及影响

权重。之后通过 Python 爬取三个主流品牌在微博上的在线评论数据并进行数据预处理，通过特征提取和主题模型将数据进行分类，获取不同因素下的在线文本评论。而后通过情感分析获得用户在微博平台上在线评论的情感数值，根据两种方法获得权重。最后，根据 AHP 方法和直觉模糊-TOPSIS 方法得到共享单车的一般选择和侧重不同影响因素下的选择。

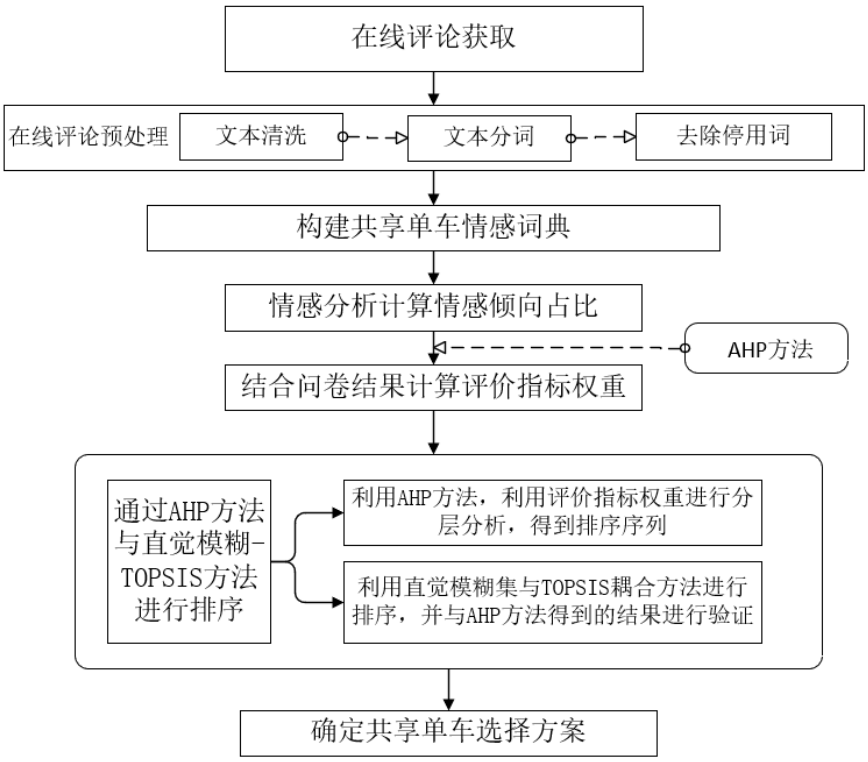

图 1　总体流程

7.2 具体步骤

7.2.1 数据收集

对于共享单车的评论在大众点评平台上评论较少且分布在少数几个城市，而微博平台在各个共享单车品牌均设有话题，话题下评论可以反映用户的情感，故选用微博平台的在线评论作为数据爬取对象。通过 python 编写爬虫程序爬取了微博下三个共享单车品牌在线文本评论。

7.2.2 数据预处理

对于爬取的三个品牌的在线文本评论，分别进行数据预处理，过程如下：

7.2.2.1 文本清洗

微博数据相较于其他平台在线评论，包含了很多无用信息，比如@其它用户、

//对其它用户评论转发、包含了很多微博表情、在评论前后含有话题标识符##以及一些URL网址链接。通过python利用正则表达式进行清洗文本去除无用字符。

7.2.2.2 文本分词

利用python，分词使用 jieba 分词的精确模式，对于一些未登录词，jieba库采用了基于汉字成词能力的隐马尔可夫模型，使用了维特比算法，能更好的处理微博文本中的一小部分网络词汇。

jieba.cut(line, cut_all=False, HMM=True)

7.2.2.3 去除停用词

对“哈工大停用词库”、“四川大学机器学习智能实验室停用词库”、“百度停用词表”等各种停用词表进行加总去重并去除其中的英文词，整理出比较全面的中文停用词表。并利用 python 进行去除停用词。

至此得到了三个品牌的清洗分词去除停用词的数据，进行词云绘制，得出美团单车词云图如图 7.2 所示，可以看出微博中用户在线评论的主要构成要素。

图 2　美团单车词云图

同时词云结果也比较符合早期问卷调查的结果，停车、车少、贵、开锁慢、定位等问题是用户较为不满意的因素，同时也是问卷所展现出用户所关注的因素。

7.2.3 TF-IDF 特征处理

TF-IDF（term frequency－inverse document frequency）是一种用于信息

检索与文本挖掘的常用加权技术。TF-IDF 可以作为一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。在信息筛选时，有的词出现的频率很高，但是作用却没那么大；有的词出现的频率很低，却表达了重要的信息，TF-IDF 的提出正是为了解决这一问题[6]。其主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

其中 TF 是词频，通常会被归一化(一般是词频除以文章总词数)，以防止它偏向长的文件。

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

即

$$TF_w = \frac{\text{在某一类词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

其中$n_{i,j}$是该词在文件d_j中出现的次数，分母则是文件d_j中所有词汇出现的次数总和。

IDF 是逆向文件频率，是指包含某词条的文档在文档集中的分布情况。包含某词条的文档越少，IDF 越大[7]，词T_i的 IDF 计算公式为。

$$\mathrm{idf_i} = \log \frac{|D|}{1+\left|\left\{ j : t_i \in d_j \right\}\right|} \tag{2}$$

即

$$IDF = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条 } w \text{ 的文档数} + 1}\right)$$

其中，|D|是语料库中的文件总数。$\left|\left\{ j : t_i \in d_j \right\}\right|$表示包含词语$t_i$的文件数目（即$n_{i,j} \neq 0$的文件数目）。如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用$1+|\{d \in D : t \in d\}|$

而 TF-IDF 是 TF 与 IDF 相乘得到。在某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

$$TF-IDF=TF\times IDF$$

在本例中，使用 Python 的 Sklearn 实现 TF-IDF 算法进行，将在线文本评论中的词语转换为词频矩阵，方便接下来对文本评论在不同因素下进行划分。

7.2.4 LDA 对数据进行划分

在 TF-IDF 进行特征处理得出关键词后，需要对现有的在线评论进行划分，得到每组关键词下的在线评论。LDA 模型是从词的角度进行话题构建，在话题发现领域使用较多。它可以将非结构化的文本数据转化成高维稀疏矩阵，利用概率分布的形式给出文档中的潜在主题，一定程度上起到了降维的作用。对于 LDA 选取的主题数目，根据微博话题数据的模糊性，采取 LDA 模型提出者 Blei 使用的一致性来选择模型的主题数，选取不同主题数目下一致性最大的模型作为最优的模型。

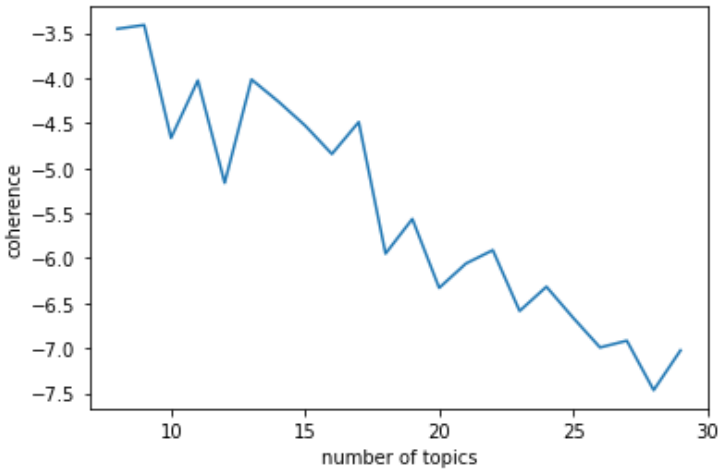

图 3　一致性-话题数选择

以美团单车文本评论为例可以看出主题数目在 7、8 附近时一致性最大，如图 7.3 所示，故选取 7 为主题数目。每个主题下初步选取 5 个关键词，但是效果不够理想，之后决定选取 12 个关键词，在关键词中寻找我们的调查因素，然后手动对主题进行划分，最终可以得到 4 个调查因素下的在线文本评论。

7.2.5 情感分析

情感分析主要利用了通过 GooSeeKer 工具进行情感分析，但初步的情感分析发现平台的情感词典与研究内容的情感词语相差较大，故重新根据现有评论构建新的正面词、负面词、程度词的情感词典并导入至工具中，得到情感分析的结果。

8 系统评价方法

8.1 使用 AHP 方法进行选择

在调研开始我们就确定了决策目的和决策方案，在后续的数据分析中又通过 LDA 模型整理划分得到四个重要决策因素，基于此我们建立了多要素、多层次的评价系统。

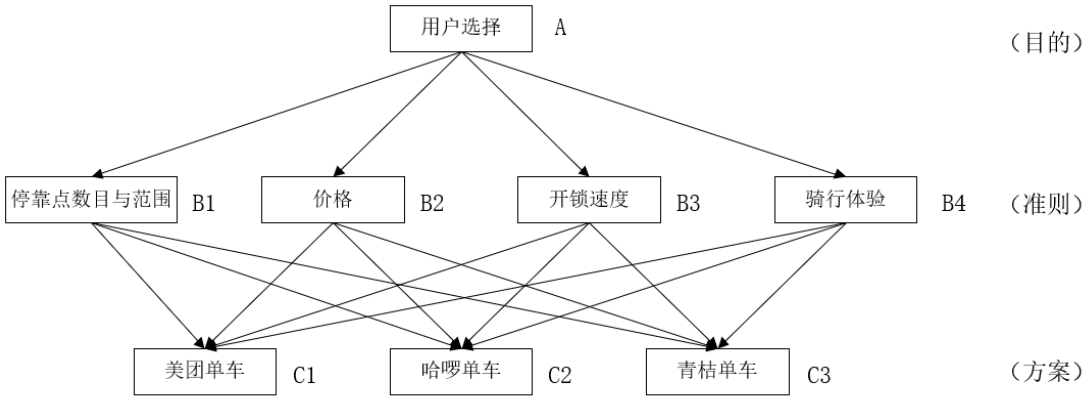

图 4　系统评价

基于此决策系统，主要分为以下四个步骤进行决策分析：

(1)对同一层次的各元素关于上一层次的某一准则的重要性进行两两比较。

(2)构造比较矩阵，进行一致性检验。

(3)由矩阵计算出比较元素相对于某准则的相对权重。

(4)整合各层要素矩阵，计算出相对总目标的合成权重，并进行一致性检验，一致性检验通过后得到方案排序。

根据情感分析得到的四个决策因素，我们发布了问卷，共得到 174 份有效问卷，28 份无效问卷。将参与问卷调研的人看做专家，让他们对各个要素的看重程度进行打分，得到如表 1 所示结果：

表 1　问卷打分结果

	0	1	2	3	4	5
停靠点数目与范围	5(2.91%)	6(3.49%)	11(6.4%)	28(16.28%)	45(26.16%)	77(44.77%)
价　格	7(4.07%)	15(8.72%)	20(11.63%)	37(21.51%)	41(23.84%)	52(30.23%)
开锁速度	11(6.4%)	10(5.81%)	20(11.63%)	44(25.58%)	42(24.42%)	45(26.16%)
骑行体验	6(3.49%)	4(2.33%)	11(6.4%)	28(16.28%)	53(30.81%)	70(40.7%)

经过加权求和以及数据差异度扩大，得到四个决策因素的重要程度如表 2 所示：

表 2　重要程度

决策因素	停靠点数目与范围（B1）	价格（B2）	开锁速度（B3）	骑行体验（B4）
重视程度	18.262	14.325	13.715	18.312

由上述数据可列出决策因素层针对决策目的层的比较矩阵，如表 3 所示：

表 3　判断矩阵

A	B1	B2	B3	B4
停靠点数目与范围（B1）	1.000	1.276	1.332	0.997
价格（B2）	0.784	1.000	1.045	0.783
开锁速度（B3）	0.751	0.957	1.000	0.8032
骑行体验（B4）	1.003	1.278	1.245	1.000

利用方根法进行 AHP 层次分析权重计算

$$w_i = \left(\prod_{j=1}^{m} A_{ij} \right)^{\frac{1}{n}} \tag{3}$$

$$W_i^0 = \frac{W_i}{\sum_{i=1}^{n} W_i^0} \tag{4}$$

$$CI = \frac{\lambda \max - \mathrm{N}}{N-1} \tag{5}$$

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

得到此判断矩阵的权重以及一致性检验，如表 4 所示：

表 4　A-B 判断矩阵结果

	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
停靠点数目与范围（B1）	1.141	0.283	4.001	0.0003	0.882	0.0003	通过
价格（B2）	0.895	0.222					
开锁速度(B3)	0.872	0.216					
骑行体验(B4)	1.124	0.279					

可得，停靠点数目与范围（B1）的权重得分为 0.283，，价格（B2）的权重得分为 0.222，开锁速度（B3）的权重得分 0.216，骑行体验（B4）的权重的份为 0.279，且决策因素的比较矩阵通过了一致性检验。

根据前期的数据整理分类以及情感分析，由于爬取的微博评论大多数为负面

评论，我们得出了不同方案对应不同决策因素的不满意度情况如表 5 所示：

表 5　不同决策因素不满意度情况

	停靠点数目与范围（B1）	价格（B2）	开锁速度（B3）	骑行体验（B4）
美团单车（C1）	-2.197	-1.426	-1.040	-2.301
哈啰单车（C2）	-3.212	-3.057	-1.809	-2.812
青桔单车（C3）	-4.556	-1.684	-1.300	-3.437

考虑到目前得到的是不满意度，但我们希望找到更令用户满意的骑行选择，所以在计算比较矩阵的时候将原本满意度之比改为，不满意度之比的倒数进行计算，以此获得用户偏向的结果，其中各个方案针对四个决策因素的比较矩阵和一致性检验结果如表 6-表 13 所示：

表 6　B1 与 C 决策矩阵

停靠点数目与范围（B1）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	1.462	2.075
哈啰单车（C2）	0.684	1.000	1.418
青桔单车（C3）	0.482	0.705	1.000

表 7　B1 与 C 一致性检验

停靠点数目与范围（B1）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.448	0.462	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	0.990	0.316					
青桔单车（C3）	0.698	0.223					

表 8　B2 与 C 决策矩阵

价格（B2）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	2.146	1.181
哈啰单车（C2）	0.466	1.000	0.551
青桔单车（C3）	0.847	1.816	1.000

表 9　B2 与 C 一致性检验

价格（B2）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.363	0.432	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	0.636	0.202					
青桔单车（C3）	1.154	0.366					

表 10　B3 与 C 决策矩阵

开锁速度（B3）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	1.739	1.250
哈啰单车（C2）	0.575	1.000	0.718
青桔单车（C3）	0.800	1.392	1.000

表 11　B3 与 C 一致性检验

开锁速度（B3）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.295	0.421	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	0.745	0.242					
青桔单车（C3）	1.037	0.337					

表 12　B4 与 C 决策矩阵

骑行体验（B4）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	1.222	1.494
哈啰单车（C2）	0.818	1.000	1.222
青桔单车（C3）	0.669	0.818	1.000

表 13　B4 与 C 一致性检验

骑行体验（B4）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.222	0.402	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	1.000	0.329					
青桔单车（C3）	0.818	0.269					

以上比较矩阵全部通过一致性检验，最后进行整体一致性检验得到排序结果，整体一致性检验的公式如下所示：

假设 p-1 层有 p_k 个因素，第 p 层的一致性指标为 CI_1^P，CI_2^P，$CI_3^P\cdots CI_{pk}^P$
第 p 层的随机一致性指标为 RI_1^P，RI_2^P，$RI_3^P\cdots RI_{pk}^P$

$$CI^P=\left(\mathrm{CI}_1^P,CI_2^P,CI_3^P\cdots\mathrm{CI}_{pk}^P\right)W^{(P-1)}$$

$$RI^P=\left(RI_1^P,RI_2^P,RI_3^P\cdots RI_{pk}^P\right)\mathrm{W}^{(P-1)}$$

$W^{(P-1)}$ 为 p-1 层对第一层的排序权向量，由此可得第 p 层对第一层的组合一致性比率为 $CR^P=CR^{(P-1)}+\frac{CI^P}{RI^P}$

表 14　总体一致性检验

	停靠点数目与范围（B1）	价　格（B2）	开锁速度（B3）	骑行体验（B4）	总权重	排序	一致性检验
	0.283	0.222	0.216	0.279			
美团单车（C1）	0.462	0.432	0.421	0.402	0.430	1	通过
哈啰单车（C2）	0.316	0.202	0.242	0.329	0.280	3	
青桔单车（C3）	0.223	0.366	0.337	0.269	0.292	2	

得到总体一致性检验，如表 14 所示，可知对共享单车的选择最好的方案为美团单车，在各个因素下总得分最高为 0.430，是明显好于其他两个品牌的共享单车的，而哈啰单车和青桔单车其实相关差不大，更看重停靠点数目与范围或者骑行体验的，可以考虑哈啰单车，更看重价格优惠和开锁速度的可以选择青桔单车。

考虑到我们是以负情感度为初始条件进行分析，可能会存在误差，因此采用新的计算方式确定情感分析后的满意度，进行二次计算，满意度如表 8.1.15 所示：

表 15　满意度情况

	停靠点数目与范围（B1）	价格（B2）	开锁速度（B3）	骑行体验(B4)
美团单车（C1）	0.537	0.732	0.640	0.536
哈啰单车（C2）	0.461	0.512	0.564	0.495
青桔单车（C3）	0.396	0.595	0.617	0.496

将满意度两两比较，得到比较矩阵并进行一致性检验如表 16-表 24 所示：

表 16　B1 与 C 决策矩阵

停靠点数目与范围（B1）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	1.166	1.357
哈啰单车（C2）	0.858	1.000	1.318
青桔单车（C3）	0.737	0.859	1.000

表 17　B1 与 C 一致性检验

停靠点数目与范围（B1）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.166	0.385	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	1.000	0.331					
青桔单车（C3）	0.859	0.284					

表 18　B2 与 C 决策矩阵

价格（B2）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	1.429	1.231
哈啰单车（C2）	0.700	1.000	0.861
青桔单车（C3）	0.813	1.161	1.000

表 19　B2 与 C 一致性检验

价格（B2）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.207	0.398	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	0.845	0.279					
青桔单车（C3）	0.981	0.323					

表 20　B3 与 C 决策矩阵

开锁速度（B3）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	1.134	1.038
哈啰单车（C2）	0.882	1.000	0.915
青桔单车（C3）	0.964	1.093	1.000

表 21　B3 与 C 一致性检验

开锁速度（B3）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.056	0.352	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	0.931	0.310					
青桔单车（C3）	1.017	0.339					

表 22　B4 与 C 决策矩阵

骑行体验（B4）	美团单车（C1）	哈啰单车（C2）	青桔单车（C3）
美团单车（C1）	1.000	1.083	1.081
哈啰单车（C2）	0.924	1.000	0.998
青桔单车（C3）	0.925	1.002	1.000

表 23　B4 与 C 一致性检验

骑行体验（B4）	特征向量	权重值	最大特征根	CI	RI	CR	一致性检验
美团单车（C1）	1.054	0.351	3.000	3.000	0.520	0.000	通过
哈啰单车（C2）	0.973	0.325					
青桔单车（C3）	0.975	0.324					

各层次一致性检验通过后，进行整体一致性检验。

表 24　总体一致性检验

	停靠点数目与范围（B1）	价格（B2）	开锁速度（B3）	骑行体验（B4）	总权重	排序	一致性检验
	0.283	0.222	0.216	0.279			
美团单车（C1）	0.385	0.398	0.352	0.351	0.367	1	通过
哈啰单车（C2）	0.331	0.279	0.310	0.325	0.306	3	
青桔单车（C3）	0.284	0.323	0.339	0.324	0.313	2	

由表 24 结果可知，虽然采用了两种不同的情感满意度确定方法，从而引起的最终方案得分有细微差别，但是整体排序并无区别。

对共享单车的选择最好的方案为美团单车，在各个因素下总得分最高为0.367，是好于其他两个品牌的共享单车的，而哈啰单车和青桔单车之间评分仍然差距不大，可以根据个人选择去决定。

综合以上两种方法计算得出的最终排序和各个因素下方案权重比较可以得出以下结论，如表 25 所示：

表 25　结论

决策因素	不同因素方案排序		
停靠点数目与范围（B1）	1、美团	2、哈啰	3、青桔
价格（B2）	1、美团	2、青桔	3、哈啰
开锁速度（B3）	1、美团	2、青桔	3、哈啰
骑行体验（B4）	1、美团	2、哈啰	3、青桔
总计	**1、美团**	**2、青桔**	**3、哈啰**

8.2 使用直觉模糊-TOPSIS 方法进行检验

8.2.1 直觉模糊数及其距离

定义 1 设 C 为论域 E 上的一个直觉模糊集，则 C 定义如下[8]：

$$C=\left\{\left[\left\langle x,\mu_c(x),v_c(x)\right\rangle\right]\middle|\ x\in E\right\} \tag{7}$$

其中$\mu_c(x)$为E中元素x属于 C 的隶属度，$v_c(x)$为 E 中元素x属于 C 的非隶属度，二者均属于[0,1]区间，且满足条件$0\le\mu_c(x)+v_c(x)\le1$，$x\in E$ 则可称

$$\pi_c(x)=1-\mu_c(x)-v_c(x) \tag{8}$$

为 E 中元素 x 属于 C 的犹豫度.

一个直觉模糊集 C 的隶属度$\mu_c(x)$、非隶属度$v_c(x)$、犹豫度$\pi_c(x)$分别表示对象属于其好、差、一般这三种证据的程度.

定义 2 E 中元素 x 属于 C 的隶属度与非隶属度所组成的有序对$\langle\mu_c(x),v_c(x)\rangle$称为直觉模糊数，简记为[8]

$$\alpha=\mu_\alpha,v_\alpha \tag{9}$$

α的犹豫度记为

$$\pi_\alpha = 1-\mu_\alpha-\nu_\alpha \tag{10}$$

其中，$\mu_\alpha \to [0,1], \nu_\alpha \to [0,1], \pi_\alpha \to [0,1]$，且满足

$$0 \le \mu_\alpha + \nu_\alpha \le 1, \quad \mu_\alpha + \nu_\alpha + \pi_\alpha = 1$$

定义 3 设$\alpha_1 = \langle \mu_{\alpha 1}, \upsilon_{\alpha 1} \rangle$，$\alpha_2 = \langle \mu_{\alpha 2}, \upsilon_{\alpha 2} \rangle$为直觉模糊数，$\alpha_1$和$\alpha_2$之间的距离定义为[9]

$$d(\alpha_1,\alpha_2) = \sqrt{\frac{1}{3}\left[\left(\mu_{\alpha 1}-\mu_{\alpha 2}\right)^2 + \pi_{\alpha 1}-\pi_{\alpha 2}\right)^2 + \nu_{\alpha 1}-\nu_{\alpha 2}\Big)^2\Big]} \tag{11}$$

8.2.2 直觉模糊信息多属性优化决策模型的建立

步骤 1 建立直觉模糊决策矩阵[10]

对于m个方案、n个属性，属性值均可以使用直觉模糊数形式给出的决策问题，决策者可通过“好”、“一般”、“差”3 粒度语言表示，经汇总给出方案的属性值为$\alpha_{ij} = \langle \mu_{\alpha ij}, \pi_{\alpha ij}, \upsilon_{\alpha ij} \rangle$，于是有直觉模糊决策矩阵

$$A = \begin{pmatrix} \langle \mu_{\alpha 11}, \pi_{\alpha 11}, \upsilon_{\alpha 11} \rangle & \cdots & \langle \mu_{\alpha 1n}, \pi_{\alpha 1n}, \upsilon_{\alpha 1n} \rangle \\ \vdots & \ddots & \vdots \\ \langle \mu_{\alpha m1}, \pi_{\alpha m1}, \upsilon_{\alpha m1} \rangle & \cdots & \langle \mu_{\alpha mn}, \pi_{\alpha mn}, \upsilon_{\alpha mn} \rangle \end{pmatrix} \tag{12}$$

步骤 2 确定直觉模糊“理想最优方案”和“理想最差方案”[10]

设I^+为越大越优的属性集合，I^-为越小越优的属性集合，I为越接近均值越优的属性集合，则直觉模糊“理想最优方案”为

$$A^+ = \left\{ \max_{1\le i\le m}\left(\mu_{ij}\right) | i \in I^+, \min_{1\le i\le m}\left(\mu_{ij}\right) | i \in I^-, mean_{1\le i\le m}\left(\mu_{ij}\right) | i \in I \right\}$$
$$i = 1,2,3\cdots m;\ j = 1,2,3\cdots n \tag{13}$$

“理想最差方案”为

$$A^- = \left\{ \max_{1\le i\le m}\left(\mu_{ij}\right) | i \in I^-, \min_{1\le i\le m}\left(\mu_{ij}\right) | i \in I^+, mean_{1\le i\le m}\left(\mu_{ij}\right) | i \in I \right\}$$
$$i = 1,2,3\cdots m;\ j = 1,2,3\cdots n \tag{14}$$

步骤 3 计算各方案到直觉模糊“理想最优方案”和“理想最差方案”的距离

设各方案A_{ij}到“理想最优方案”和“理想最差方案”的距离分别为$d(\alpha_{ij}, \alpha_j^+)$，$d(\alpha_{ij}, \alpha_j^-)$，则可得[10]

$$\mathrm{d}(\alpha_{ij},\alpha_j^+)=\sqrt{\frac{1}{3}\left[(\mu_{ij}-\mu_j^+)^2+\pi_{ij}-\pi_j^+)^2+\upsilon_{ij}-\upsilon_j^+)^2\right]} \tag{15}$$

$$\mathrm{d}(\alpha_{ij},\alpha_j^-)=\sqrt{\frac{1}{3}\left[(\mu_{ij}-\mu_j^-)^2+\pi_{ij}-\pi_j^-)^2+\upsilon_{ij}-\upsilon_j^-)^2\right]} \tag{16}$$

步骤 4 建立直觉模糊-TOPSIS 优化决策模型

设第k个方案贴近“理想最优方案”和“理想最差方案”的加权距离分别为$\overline{d}^+{}_k$、$\overline{d}^-{}_k$，根据各方案评价指标权重，可得

$$\overline{d}^+{}_k=\sum_{j=1}^{n}\left(\omega_j\times\mathrm{d}(\alpha_{kj},\alpha_j^+)\right) \tag{17}$$

$$\overline{d}^-{}_k=\sum_{j=1}^{n}\left(\omega_j\times\mathrm{d}(\alpha_{kj},\alpha_j^-)\right) \tag{18}$$

根据 TOPSIS 方法则可计算第个方案贴合度u_k为

$$u_k=\frac{\overline{d}^-{}_k}{\overline{d}^+{}_k+\overline{d}^-{}_k} \tag{19}$$

u_k的值越大，表示越贴近“理想最优方案，则可按u_k的值进行从大到小的顺序排列，得到方案优先序列。

8.2.3 使用直觉模糊-TOPSIS 方法进行选择

为了获取用户对于共享单车不同评价指标，体验感的数据，采取了发放问卷的方式。针对部分用户可能对事物评价表现出一定程度的犹豫，即除了有认为“好”和“差”之外，存在拿不准的情况，因此，我们采用了对各个指标进行最为简单的“好”、“一般”、“差”3 粒度语言打分，对于倾向性不太确定的，则按“一般”考虑，每项只选择一项。因为追求的使用户体验得到最大化满足，因此 4 项指标均为越大越好的效益型指标。

不同共享单车品牌使用体验的调查问卷

因素 品牌	停靠点数目与范围			骑行价格			开锁速度			骑行体验		
	好	一般	差	好	一般	差	好	一般	差	好	一般	差
哈啰单车												
青桔单车												
美团单车												

图 5　调查问卷

8.2.3.1 构造直觉模糊决策矩阵

本次调查，共收取有效问卷 60 份，按照问卷结果，将总票数除以 60 得到直觉模糊决策矩阵

$$A=\begin{pmatrix}\langle 0.4833,0.4833,0.0333\rangle & \langle 0.3667,0.5000,0.1333\rangle & \langle 0.4333,0.4833,0.0833\rangle & \langle 0.4333,0.5167,0.0500\rangle \\ \langle 0.4500,0.4833,0.0667\rangle & \langle 0.4333,0.4833,0.0833\rangle & \langle 0.4333,0.5167,0.0500\rangle & \langle 0.4167,0.5333,0.0500\rangle \\ \langle 0.5167,0.3500,0.1333\rangle & \langle 0.4667,0.5167,0.0167\rangle & \langle 0.4333,0.5167,0.0500\rangle & \langle 0.4333,0.4833,0.0833\rangle \end{pmatrix}$$

8.2.3.2 确定直觉模糊“理想最优方案”和“理想最差方案”

$$A^{+}=\left(\langle 0.5167,0.4389,0.0333\rangle \quad \langle 0.4667,0.5000,0.0167\rangle \quad \langle 0.4333,0.5056,0.0500\rangle \quad \langle 0.4333,0.5056,0.0500\rangle\right)$$

$$A^{-}=\left(\langle 0.4500,0.4389,0.1333\rangle \quad \langle 0.3667,0.5000,0.1333\rangle \quad \langle 0.4333,0.5056,0.0833\rangle \quad \langle 0.4167,0.5056,0.0833\rangle\right)$$

8.2.3.3 确定评价指标权重

根据上述 AHP 方法计算权重的结果得

$$W=\left(\langle 0.2830,0.2830,0.2830\rangle \quad \langle 0.2220,0.2220,0.2220\rangle \quad \langle 0.2162,0.2162,0.2162\rangle \quad \langle 0.2788,0.2788,0.2788\rangle\right)$$

8.2.3.4 计算直觉模糊贴合度并排序

由式(19)得直觉模糊贴合度u，$u_1=0.4381$，$u_2=0.5386$，$u_3=0.5951$，按直觉模糊贴合度u大小排序为$u_3\succ u_2\succ u_1$，即美团单车>青桔单车>哈啰单车。

9 评价结果分析

根据两种方法得到的结果，可以发现两种方法产生的排序序列均为<美团单车，青桔单车，哈啰单车>，且使用直觉模糊-TOPSIS 方法，可以较明显的区分出青桔单车与哈啰单车的排序情况。可以说明，本次研究，得到了较为准确且符合消费者倾向的结果，可供用户及单车品牌进行参考。

10 政策建议与研究意义

10.1 政策建议

10.1.1 对用户的建议

根据以上结果分析，用户对于共享单车的选择以美团单车为优。本次的研究结论可以给用户使用共享单车一定的指导，例如：在美团单车可以使用的情况下，用户可以选择美团单车以获得最佳骑行体验，而对于仅有哈啰单车和青桔单车时，性价比以及时间利用效率优先的用户可以优先选择青桔单车，而对于重视骑行体验以及方便存取的用户可优先选择哈啰单车。

同时，本次研究过程也为用户选择其他服务提供借鉴。本次数据爬取评论大多为相同的套话，去除行业广告的套话之后，好评率大打折扣，这也提醒消费者在选取服务时，不要被好评蒙蔽双眼，要学会适当规避行业套话，理智选择。用户也应该给予正确评论，为其他用户提供力所能及的帮助，利用评价功能为用户提供可靠的选择决策依据。

10.1.2 对共享单车品牌的建议

对于品牌运营，要即时改善目前的不足，如哈啰单车应该提升价格优惠力度，缩短开锁时间，青桔单车应该重视停靠地点与范围的扩充，遍布范围更广，优化单车性能，提供更好地骑行体验。同时，各品牌也应该积极调查市场，寻找新的优化方向，加以改进。

品牌的改善优化不仅提升用户体验，提高用户好感，也是增强品牌竞争力。在客户至上时代，提升用户体验，进行合理有效引导，避免浮夸宣传引起反感，是品牌壮大的基础。

10.2 研究意义

10.2.1 理论意义

本文面向不同的消费者群体，对在线评论背景下消费者在共享单车选择中的期望和偏好进行分析，从而为消费者决策提供更好的依据。本研究基于在线评论，结合不同类型消费者偏好对共享单车选择方法的研究进行补充，丰富了面向不同消费者群体的基于在线评论的共享单车推荐方法。

10.2.2 现实意义

近年来，随着国民经济不断发展，互联网技术不断普及，共享经济也脱颖而出，共享单车便是其一。因为对骑行服务的着重点不同，消费者对于所选择单车类型的决策得不到保障。故本文主要基于在线评论，以不同消费者群体为对象，优化消费者对单车的选择决策，已解决不同首要需求下的单车抉择问题。将为优化共享单车推荐系统提供实用性意见，在解决骑行者“最后一公里”问题的同时，给用户提供更佳骑行体验的选择方案，对提高消费者满意度和共享单车改进有重要的现实意义。

11 结论

本文基于在线评论的主题情感分析、AHP 方法以及直觉模糊-TOPSIS 综合评

价方法，对美团单车、青桔单车、哈啰单车三个单车品牌进行了比较分析，从而得到了可供用户进行选择的优先排序序列，为用户进行选择与单车品牌进行改进升级提供了合理参考依据，有利于用户得到最大满足，有利于单车行业的良性竞争及改革创新。

基于问卷结果，本文选取了使用人数最多的三种单车品牌，即美团单车、青桔单车、哈啰单车，进行分析。通过 python 编写爬虫程序爬取了微博下三个共享单车品牌在线文本评论，并进行了筛选，剔除了表情、符号等无用数据，得到了三个品牌的清洗分词去除停用词的数据，得到了主题词词云，根据词云结果确定了本次研究的评价指标。此后，利用问卷结果，对评价指标进行层次分析（AHP），得到了各项指标的综合权重，并通过情感分析计算满意程度与不满意程度，对三种品牌进行排序，得到了优先排序序列。最后，利用直觉模糊-TOPSIS 综合评价方法，进行再次验证，发现两种方法的优先排序序列具有一致性，从而得到了最终的选择序列，即美团单车优于青桔单车，青桔单车优于哈啰单车。

本文先是基于在线评论确定了研究的评价指标，而后根据 AHP 方法、情感分析确定各评价指标的权重以及排序序列，最后使用直觉模糊-TOPSIS 综合评价方法进行结果验证，检查一致性。本文综合考虑了单车用户的实际消费心理，考虑到用户对于单车评价的犹豫性，通过多属性决策分析。获得较为准确且符合消费者倾向的结果。

但本次研究也具有一定的局限性，目前只考虑了三种主流品牌、四种评价因素，并且因为时间问题，收集的问卷结果较少，数据可能在普遍性方面有所欠缺。纳入更多评价因素，对更多品牌，乃至整个共享单车行业进行分析，将作为后续研究的方向。

参考文献

[1]朱晓霞,宋嘉欣,张晓缇.基于主题挖掘技术的文本情感分析综述[J].情报理论与实践,2019,42(11):156-163.

[2]秦吉,张翼鹏.现代统计信息分析技术在安全工程方面的应用——层次分析法原理[J].工业安全与防尘,1999(05):44-48.

[3] Chen S M, Tan J M. Handling multi-criteria fuzzy decision-making problems based on vague set theory [J]. Fuzzy Sets and Systems1994, 67(2):163-172.

[4]王坚强. 几类信息不完全确定的多准则决策方法研究[D].中南大学,2005.

[5]决胜全面建成小康社会 夺取新时代中国特色社会主义伟大胜利[N]. 人民日报,2017-10-19(002).

[6]黄勃,陈欢,方志军,王明胜,刘文竹.基于微博的 COVID-19 热点话题分析[J].武汉大学学报(理学版),2020,66(05):425-432.

[7]敖长林,李凤佼,许荔珊,孙宝生.基于网络文本挖掘的冰雪旅游形象感知研究——以哈尔滨市为例[J].数学的实践与认识,2020,50(01):44-54.

[8] Atanassov K. Intuitionistic fuzzy sets [J]. Fuzzy Sets and Systems, 1986, 20(1):87-96

[9]马钰,张定海,王万雄.一种基于信息熵属性重要度的直觉模糊 TOPSIS 方法[J]. 兰州大学学报(自然科学版),2020,56(05):677-680+689.

[10]谭秋月,孙平安,吴迪,吴永波.直觉模糊信息多属性优化决策模型及其应用[J].数学的实践与认识,2022,52(03):132-139.

附录

附录 1

settings.py

```
# -*- coding: utf-8 -*-

BOT_NAME = 'weibo'
SPIDER_MODULES = ['weibo.spiders']
NEWSPIDER_MODULE = 'weibo.spiders'
COOKIES_ENABLED = False
TELNETCONSOLE_ENABLED = False
LOG_LEVEL = 'ERROR'
# 访问完一个页面再访问下一个时需要等待的时间，默认为 10 秒
DOWNLOAD_DELAY = 10
DEFAULT_REQUEST_HEADERS = {
    'Accept':
    'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
    'Accept-Language': 'zh-CN,zh;q=0.9,en;q=0.8,en-US;q=0.7',
    'cookie':                    '_T_WM=5cc9d8ad35d319f4893f95458d17e929;
SUB=_2A25Ppa3rDeRhGeFN61sS9CrLwziIHXVtaTOjrDV6PUJbktAKLWzgkW1N
QG2tqgTw7At2bA2eYMS9gL-Jeh3Ge6h4;
SUBP=0033WrSXqPxfM725Ws9jqgMF55529P9D9WFb7lIBhIjrPJdTdfS_9yR_5NH
D95QNe054e0BXS0nXWs4Dqcj_i--fiK.Ri-isi--ciKnRiK.pi--Ri-zciKnfi--NiKLWi-
88i--NiKLWiKnX; SSOLoginState=1654775227'
}
ITEM_PIPELINES = {
    'weibo.pipelines.DuplicatesPipeline': 300,
    'weibo.pipelines.CsvPipeline': 301,
    # 'weibo.pipelines.MysqlPipeline': 302,
    # 'weibo.pipelines.MongoPipeline': 303,
    # 'weibo.pipelines.MyImagesPipeline': 304,
    # 'weibo.pipelines.MyVideoPipeline': 305
}
# 要搜索的关键词列表，可写多个，值可以是由关键词或话题组成的列表，也可
以是包含关键词的 txt 文件路径，
# 如'keyword_list.txt'，txt 文件中每个关键词占一行
KEYWORD_LIST = ['# 青 桔 单 车 #']    # 或 者    KEYWORD_LIST =
'keyword_list.txt'
```

```
# 要搜索的微博类型，0 代表搜索全部微博，1 代表搜索全部原创微博，2 代表热门微博，3 代表关注人微博，4 代表认证用户微博，5 代表媒体微博，6 代表观点微博
WEIBO_TYPE = 0
# 筛选结果微博中必需包含的内容，0 代表不筛选，获取全部微博，1 代表搜索包含图片的微博，2 代表包含视频的微博，3 代表包含音乐的微博，4 代表包含短链接的微博
CONTAIN_TYPE = 0
# 筛选微博的发布地区，精确到省或直辖市，值不应包含“省”或“市”等字，如想筛选北京市的微博请用“北京”而不是“北京市”，想要筛选安徽省的微博请用“安徽”而不是“安徽省”，可以写多个地区，
# 具体支持的地名见 region.py 文件，注意只支持省或直辖市的名字，省下面的市名及直辖市下面的区县名不支持，不筛选请用“全部”
REGION = ['全部']
# 搜索的起始日期，为 yyyy-mm-dd 形式，搜索结果包含该日期
START_DATE = '2019-03-01'
# 搜索的终止日期，为 yyyy-mm-dd 形式，搜索结果包含该日期
END_DATE = '2022-06-09'
# 进一步细分搜索的阈值，若结果页数大于等于该值，则认为结果没有完全展示，细分搜索条件重新搜索以获取更多微博。数值越大速度越快，也越有可能漏掉微博；数值越小速度越慢，获取的微博就越多。
# 建议数值大小设置在 40 到 50 之间。
FURTHER_THRESHOLD = 46
# 图片文件存储路径
IMAGES_STORE = './'
# 视频文件存储路径
FILES_STORE = './'
# 配置 MongoDB 数据库
# MONGO_URI = 'localhost'
# 配置 MySQL 数据库，以下为默认配置，可以根据实际情况更改，程序会自动生成一个名为 weibo 的数据库，如果想换其它名字请更改 MYSQL_DATABASE 值
# MYSQL_HOST = 'localhost'
# MYSQL_PORT = 3306
# MYSQL_USER = 'root'
# MYSQL_PASSWORD = '123456'
# MYSQL_DATABASE = 'weibo'
```

附录 2

clean.py

```
import re
def clean(text):
    text = re.sub(r"(回复)?(//)?\s*@\S*?\s*(:| |$)", " ", text)  # 去除正文中的@和回复/转发中的用户名
    text = re.sub(r"\[\S+\]", "", text)          # 去除表情符号
    text = re.sub(r"#\S+#", "", text)          # 保留话题内容
    URL_REGEX = re.compile(
        r'(?i)\b((?:https?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()<>]+\)))*\))+(?:\(([^\s()<>]+|(\([^\s()<>]+\)))*\)|[^\s`!()\[\]{};:\'".,<>?«»“”‘’]))',
        re.IGNORECASE)
    text = re.sub(URL_REGEX, "", text)           # 去除网址
    text = text.replace("转发微博", "")           # 去除无意义的词语
    text = text.replace("电单车", "")
    text = text.replace("单车", "")
    text = text.replace("美团", "")
    text = text.replace("哈啰", "")
    text = text.replace("青桔", "")
    text = text.replace("共享", "")
    text = re.sub(r"\s+", " ", text) # 合并正文中过多的空格
    return text.strip()

fout = open('青桔-清洗.txt', 'a', encoding='utf-8')
f = open("青桔.txt",encoding='utf-8')
while True:
    line = f.readline()
    if line:
        print (clean(line))
        fout.writelines(clean(line) + '\n')
    else:
        break
f.close()
fout.close()
```

cut and stop.py

```
import jieba


def stopwordslist(filepath):  # 定义函数创建停用词列表
    stopword = [line.strip() for line in open(filepath, 'r', encoding='utf-8').readlines()]
# 以行的形式读取停用词表，同时转换为列表
    return stopword


def cutsentences(sentences):  # 定义函数实现分词
    # print('原句子为：' + sentences)
    cutsentence = jieba.lcut(sentences.strip())  # 精确模式
    # print('\n' + '分词后：' + "/ ".join(cutsentence))
    stopwords = stopwordslist(filepath)  # 这里加载停用词的路径
    lastsentences = ''
    for word in cutsentence:  # for 循环遍历分词后的每个词语
        if word not in stopwords:  # 判断分词后的词语是否在停用词表内
            if word != '\t':
                lastsentences += word
                lastsentences += "/ "
    return(lastsentences)


filepath = 'C:/Users/13391/Desktop/系统工程/stop_words.txt'
stopwordslist(filepath)

fout = open('青桔-清洗分词去停用词.txt', 'a', encoding='utf-8')
f = open("青桔-清洗.txt",encoding='utf-8')
while True:
    line = f.readline()
    if line:
        fout.writelines(cutsentences(line) + '\n')
    else:
        break
f.close()
fout.close()
```

附录 3

美团 topic.ipynb

```
# -*- coding: utf-8 -*-
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import time
import warnings
warnings.filterwarnings('ignore')
import tomotopy as tp
import re
import jieba
from tqdm.notebook import tqdm
import jieba.posseg as pseg
from gensim import corpora, models

import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s',
level=logging.INFO)
#文档
data=pd.read_csv(r'data/美团单车.csv')
data.head()
data.isnull().sum()
#  只保留中文
data['微博正文']=data['微博正文'].astype('str')
data['content'] = data['微博正文'].str.replace("[^\u4e00-\u9fa5]", "")
#  移除常用词以及分词
stoplist = [i.strip() for i in open(r'data/stopwords.txt',encoding='utf-8').readlines()]
def segment(text):
    words = jieba.cut(text)
    words = [w for w in words if w not in stoplist if len(w)>1]
    return words
```

```
data['words'] = data['content'].apply(segment)
print("二次删除停用词和分词成功!!!  ")
data.head(1)
#查询主题个数比较合适
def find_k(docs, min_k, max_k, min_df):
    #min_df  词语最少出现在 2 个文档中
    import matplotlib.pyplot as plt
    scores = []
    for k in range(min_k, max_k):
        #seed 随机种子，保证运行的结果一样
        mdl = tp.LDAModel(min_df=min_df, k=k,seed=555)
        for words in docs:
            if words:
                mdl.add_doc(words)
        mdl.train(20)
        coh = tp.coherence.Coherence(mdl)
        scores.append(coh.get_score())

    #x = list(range(min_k, max_k - 1))   #  区间最右侧的值。注意：不能大于 max_k
    #print(x)
    #print()
    plt.plot(range(min_k, max_k), scores)
    plt.xlabel("number of topics")
    plt.ylabel("coherence")
    plt.show()

find_k(docs=data['words'], min_k=8, max_k=30, min_df=2)
from sklearn.feature_extraction.text import TfidfVectorizer,CountVectorizer
#  将文本中的词语转换为词频矩阵
cleanchap = [" ".join(w) for w in data['words']]

Tfidfvectorizer =TfidfVectorizer(max_df=0.4,
                                 min_df=2,
                                 max_features=2000,stop_words=stoplist)  # 创 建
```

```
词袋数据结构
data_vectorized = Tfidfvectorizer.fit_transform(cleanchap)
pd.DataFrame(data_vectorized.toarray(),columns=Tfidfvectorizer.get_feature_names(
)).to_csv('data/美团单车_Tf-idf.csv',encoding='utf_8_sig')
# 设定 LDA 模型
from sklearn.decomposition import LatentDirichletAllocation
n_topics=8
ldamodel = LatentDirichletAllocation(n_components = n_topics)
ldamodel.fit(data_vectorized)
# 主题词打印函数
def print_top_words(model, feature_names, n_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print("Topic #%d:" % topic_idx)
        print(" ".join([feature_names[i] for i in topic.argsort()[:-n_top_words - 1:-1]]))
    print()
n_top_words = 12
tf_feature_names = Tfidfvectorizer.get_feature_names()
print_top_words(ldamodel, tf_feature_names, n_top_words)
import joblib
#保存 Model(注:save 文件夹要预先建立，否则会报错)
joblib.dump(ldamodel, r'data/美团 model_1.pkl')
import numpy as np

# 构建文档-词频矩阵
lda_output = ldamodel.transform(data_vectorized)
# 列名
topicnames = ["Topic" + str(i)
                  for i in range(ldamodel.n_components)]
# 行索引名
#docnames = ["Doc" + str(i) for i in range(len(data.words))]

# 转化为 pd.DataFrame
df_document_topic = pd.DataFrame(np.round(lda_output, 4),
                                 columns=topicnames,
```

```
                                    index=data['bid'])

# Get dominant topic for each document
dominant_topic = np.argmax(df_document_topic.values, axis=1)
df_document_topic['dominant_topic'] = dominant_topic
df_document_topic.head()
result_data=pd.merge(data,df_document_topic,on='bid')
#  主题分布图
import seaborn as sns
result_type_counts = result_data['dominant_topic'].value_counts()
fig,axes = plt.subplots(1,2,figsize=(16,8),dpi=100)
axes[0].pie(result_type_counts.values,autopct="%.2f%%",labels=result_type_counts.i
ndex)
sns.barplot(result_type_counts.index,result_type_counts.values)
plt.savefig(r'data/美团单车_主题分布图.png')
plt.tight_layout()
df_topic_distribution                                                    =
df_document_topic['dominant_topic'].value_counts().reset_index(name="Num
Documents")
df_topic_distribution.columns = ['Topic Num', 'Num Documents']
df_topic_distribution
result_data.to_csv('data/美团单车_data.csv',encoding='utf_8_sig')
```