分类号_____　　密级　公开
U D C _____　　编号　10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目 基于文本和网络搜索信息的游客流量预测研究
——以海南为例

研 究 生 姓 名：曹静如

指导教师姓名、职称：孙景云、教授

学 科、专 业 名 称：统计学、应用统计

研 究 方 向：大数据分析

提 交 日 期：2023 年 5 月 30 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：翦静如 签字日期：2025.5.30

导师签名：孙景云 签字日期：2025.5.30

导师(校外)签名：张卫东 签字日期：2025.5.30

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：翦静如 签字日期：2025.5.30

导师签名：孙景云 签字日期：2025.5.30

导师(校外)签名：张卫东 签字日期：2025.5.30

Research on tourist flow forecasting based on text and web search information: A Case Study of Hainan

Candidate : Cao Jing Ru

Supervisor : Sun Jing Yun

摘 要

随着 5G 时代的到来，越来越多的游客利用互联网提前了解旅游目的地，进而制定旅游计划。基于这些互联网数据信息能够很大程度上提高对旅游人数的预测准确率，不仅可以动态监测游客行为，同时又克服了传统旅游数据的时滞问题。因此，基于互联网数据对游客流量进行预测是十分有必要的。

本文基于已有文献的研究，以海南省月度游客流量为例进行旅游预测的实证研究，主要研究工作如下：（1）针对网络搜索信息存在多噪声、非线性、高波动等特点，在关键词的选择和指数合成过程中存在诸多困难。本文提出一种新的搜索关键词选择及指数合成技术——R/S-TDC-EMD-KPCA 方法，首先利用重标极差法（R/S）和时差相关法（TDC）选择具有预测能力的关键词，并对选择的关键词搜索量分别进行经验模态分解（EMD）降噪，最后利用核主成分(KPCA)方法合成网络综合搜索指数。通过对比验证所提取的网络综合搜索指数在游客流量预测中的有效性。（2）互联网数据信息代表了游客的不同行为特征，可全面反映游客的关注点、兴趣和情感倾向。本文提出一种基于百度指数、微博文本等互联网数据融合的旅游预测新方法。首先，基于 R/S-TDC-KPCA 方法将百度指数合成网络综合搜索指数。其次，从中国主流社交平台新浪微博中提取与最优关键词有关的文本数据信息，对提取出的文本信息实施数据清洗，并采用基于正负情感简单相加和基于正负情感非对称的方法来构建情感指数。最后，将网络综合搜索指数、情感指数以及历史游客流量作为输入变量，构建 SARIMAX 模型进行实证预测研究。

实证结果表明，与其他传统预测模型相比，基于 R/S-TDC-EMD-KPCA 方法的网络综合搜索指数结合 BP 神经网络在海南旅游预测中具有较低的平均绝对百分比误差(MAPE)和归一化均方根误差(NRMSE)，其中 MAPE 从 10.44%下降到 7.11%，NRMSE 从 14.66 下降到 9.81。因此，提出的 R/S-TDC-EMD-KPCA 方法能高质量的提取和合成网络搜索信息，进而可有效用于游客流量的辅助预测。其次，研究发现将网络综合搜索指数与微博情感指数同时作为预测因子时，可以有效提高预测精度，其水平预测精度均低于其他基准模型，MAE 下降到 15.23，

MAPE 下降到 2.62%，RMSE 下降到 21.77，RMSPE 下降到 3.47%。另外，本文采用了两种不同的方法编制情感指数，分别是基于正负情感简单相加的情感指数和基于正负情感非对称的情感指数，研究发现不同的情感指数编制方法会对预测结果有一定的影响。基于不同人类心理行为构建正负情感非对称情形下的情感指数相比于正负情感的简单加总更能反映游客情感倾向，可以获得更好的预测效果。因此，基于文本和网络搜索信息的游客流量预测是有效的，这为旅游需求的精准预测提供了新的途径。

关键词： 旅游预测 KPCA 百度指数 SnowNLP 情感指数

Abstract

With the advent of the 5G era, more and more tourists use the Internet to understand tourist destinations in advance and then make travel plans. Based on these Internet data information, the accuracy of forecasting the number of tourists can be greatly improved, which can not only dynamically monitor the behavior of tourists, but also overcome the time lag problem of traditional tourism data. Therefore, it is necessary to predict tourist traffic based on Internet data.

Based on the research of existing literature, this paper takes the monthly tourist flow of Hainan Province as an example to carry out the empirical research of tourism forecasting, and the main research work is as follows: (1) In view of the characteristics of multiple noise, nonlinearity and high fluctuation of network search information, there are many difficulties in the selection of keywords and index synthesis. This paper proposes a new search keyword selection and exponential synthesis technology, R/S-TDC-EMD-KPCA method, which first uses the rescale range method (R/S) and time difference correlation method (TDC) to select keywords with predictive ability, performs empirical mode decomposition (EMD) noise reduction on the search volume of selected keywords, and finally synthesizes the network comprehensive search index by nuclear principal component (KPCA) method. The effectiveness of the extracted

web comprehensive search index in tourist flow prediction is verified by comparison. (2) Internet data information represents the different behavioral characteristics of tourists, which can fully reflect tourists' concerns, interests and emotional tendencies. This paper proposes a new method of tourism forecasting based on the integration of Internet data such as Baidu index and Weibo text. Firstly, based on the R/S-TDC-KPCA method, the Baidu index is synthesized into a comprehensive online search index. Secondly, the text data information related to the optimal keywords is extracted from the mainstream Chinese social platform Sina Weibo, the extracted text information is cleansed, and the emotion index is constructed by using the methods of simple addition of positive and negative emotions and asymmetry based on positive and negative emotions. Finally, the comprehensive search index, sentiment index and historical tourist flow are used as input variables, and the SARIMAX model is constructed for empirical prediction research.

The empirical results show that compared with other traditional prediction models, the network comprehensive search index based on R/S-TDC-EMD-KPCA method combined with BP neural network has a lower mean absolute percentage error (MAPE) and normalized root mean square error (NRMSE) in Hainan tourism forecasting, in which MAPE decreases from 10.44% to 7.11%, and NRMSE decreases from 14.66 to 9.81. Therefore, the proposed R/S-TDC-EMD-KPCA method can extract and

synthesize web search information with high quality, and then can be effectively used for auxiliary prediction of tourist flow. Secondly, it is found that when the network comprehensive search index and Weibo sentiment index are used as predictors at the same time, the prediction accuracy can be effectively improved, and the horizontal prediction accuracy is lower than that of other benchmark models, MAE drops to 15.23, MAPE drops to 2.62%, RMSE drops to 21.77, and RMSPE drops to 3.47%. In addition, this paper adopts two different methods to compile the sentiment index, namely the emotion index based on the simple addition of positive and negative emotions and the emotion index based on the asymmetric emotion of positive and negative emotions, and it is found that different emotion index compilation methods will have a certain impact on the prediction results. The emotional index under the asymmetric situation of positive and negative emotions based on different human psychological behaviors can better reflect the emotional tendency of tourists and obtain better prediction effect than the simple sum of positive and negative emotions. Therefore, the prediction of tourist flow based on text and web search information is effective, which provides a new way to accurately predict tourism demand.

Keywords： Tourism forecast; KPCA; Baidu index; SnowNLP; Sentiment index

目 录

1 引言

1.1 研究背景及意义

1.1.1 研究背景

旅游业是一个产业规模不断扩大、从业人员不断增加、运行制度不断创新、运行环境持续改善的劳动密集型产业。旅游业成为重要的服务业，不仅拉动了经济的增长，还拓宽了就业的渠道。由于我国社会经济的不断发展，人民生活的不断提高，对旅游的需求量也日益增多，这促使旅游观光成为国家经济持续增长的主要支撑，其重要性不可低估。

自 1978 年以来，中国政府开始重视并大力推动旅游业的发展，从而促进了中国旅游业的发展。随着社会经济发展的进步，人们物质水平的提升和消费方式的变化，旅游的门槛开始降低，境外旅游、国外旅游和国内旅游逐渐成为旅游发展的三大方式。随着我国居民消费水平的不断提升，居民对旅游活动的意向和支付能力不断增强，再加上便捷的公共交通设施，可能产生的旅游需求不断转化为现实有效的旅游需求。随着中国旅游市场需求的不断上升，中国旅游市场发展迅速。根据国家文化和旅游部的统计结果显示，从 2012 年到 2019 年，国内游客数量呈现出稳步增长的趋势。2020 年，受新冠疫情影响，游客总数减少至 28.79 亿人次，同比减少 52.1%。但从国内市场来看，游客量和收入继续稳步回升，2021 年增至 32.46 亿人次，同比增长 12.8%。用户出行意愿正在逐步恢复。2023 年疫情放宽，生活状况逐步恢复，旅游人数逐渐回升至疫情前的水平，旅游业将迎来全面复苏，甚至进入一个更加高级的发展阶段。另外新冠疫情也促使游客的旅游方式发生了变化，其中“互联网+旅游”变得愈加重要。

这些旅游方式催生了各类互联网数据，如游客搜索、评论、拍照的数据信息在互联网上被记录存储，这些数据信息能够反映游客对旅游目的地的关注程度和兴趣范围，同时还具有数据量大、时效性强等特征。因此，在分析旅游行业时，互联网数据的作用是不可忽视的。当前，旅客流量预测研究中所采用的互联网数据主要包括搜索引擎数据和社交媒体数据。但是，目前所获得的互联网数据往往呈现多而杂的现象，仅有很少一部分数据是具有预测价值的。在实际应用中，在

提取这些数据信息时要避免冗余信息的干扰，尽可能提取有效且有高价值的互联网数据。

1.1.2 研究意义

游客流量的准确预测对旅游业和相关行业至关重要。通过对未来客流的准确预测，可提高政府管理部门的预警能力。根据预测结果，政府可及时采取有效措施，合理配置资源，优化旅游景点的软硬件服务设施，防止旅游景点超载等混乱局面的产生。对旅游服务类企业而言，准确的预测有助于制定合理的营销策略，优化旅游线路等。因此，预测游客流量能对未来的规划和决策提供重要的参考信息，从而具有重要的现实意义。

基于传统数据的旅游预测，主要依赖于政府发布的结构化统计数据，此类数据公开发布的频次较低，通常以月或季度为发布周期，且存在发布时间滞后的情况，导致预测的精度不高。而随着 5G 时代的到来，越来越多的游客开始利用互联网提前了解旅游目的地的风土人情和旅游攻略，进而制定旅游计划。互联网大数据为优化旅游需求预测效率提供了有效途径。

网络搜索引擎为游客提供了在线查询渠道，游客在网络搜索引擎中留下的与旅游相关的搜索信息可以作为旅游预测模型的辅助信息源。但网络搜索信息数据依赖于关键词选取，信息庞杂，搜集和有效提取存在各种困难。因此，筛选出有效的网络搜索信息数据并合成网络综合搜索指数对于旅游预测研究是至关重要的。基于此，本研究提出了一种新的网络搜索信息的提取合成，并用于海南省游客流量预测，为网络搜索数据合成方法提供了一种新的思路。

随着互联网的不断发展，网络搜索引擎已经不再是人们获取信息的唯一通道，越来越多的游客开始转向社交平台获取信息，制定旅游计划。现有的大多数研究都是基于网络搜索引擎的搜索索引进行预测，没有考虑搜索数量背后的情感。在现实生活中，搜索量的增加并不一定反映人们前往旅游目的地的积极倾向。网络发布的负面报道也会致使在线搜索量的增加。因此，仅根据搜索指数预测游客流量并不全面。对于游客来说，可以在论坛、博客、微博等公共社交平台上发布旅游体验，表达自己的情感态度；对于旅游媒体来说，也可以通过这些公共社交平台为游客提供丰富的旅游资源。这些情感信息的传递在一定程度上体现了游客对旅游目的地的感知行为，从而对旅游预测产生重大影响。

互联网数据信息代表了游客的不同行为特征，可全面反映游客的关注点、兴趣和情感倾向。将来自不同平台的互联网数据进行有效融合，可以使旅游部门更好的了解游客行为，进而优化资源配置，合理进行旅游管理决策。本研究将网络搜索数据与社交平台数据结合起来，通过对海量数据的处理及分析，构建多个基准模型，选择预测效果最佳的模型，对游客流量进行预测，从而为未来的旅游规划和决策提供重要的参考信息。

1.2 文献综述

1.2.1 基于历史游客流量的预测

旅游需求预测最初主要采用时间序列模型和计量经济学模型。其中 ARIMA 常被用来作为基准模型来衡量预测效果，是常用的时间序列模型[1]；在计量经济学模型中向量自回归模型(VAR)[2]、自回归分布滞后模型(ADLM)[3]等也常用于旅游需求预测，均取得了良好的预测效果。但时间序列和计量经济学模型存在非线性拟合能力差的缺点，为解决这一问题，基于机器学习的人工智能技术开始受到学者的广泛关注，由于人工智能技术在解决非线性问题具有较强的处理能力，因此其精度往往优于时间序列和计量经济模型。如 Law[4]通过引入 BP 神经网络大大提升了游客需求预测的准确性。实证表明，与其他基准模式相比，该模型具有更优越的预期性能。贾鹏等[5]基于 BP 神经网络构建了邮轮旅游需求预测模型，并通过美国邮轮市场数据验证了模型的有效性。最后将其应用于我国邮轮旅游需求预测，取得了良好的效果。但有一些时间序列存在明显的季节性周期变化，为解决这一问题，可以使用 SARIMA 模型来表示。SARIMA 模型不但可以处理非平稳时间序列，而且在模型中考虑了季节信息。Tendai 等[6]通过 SARIMA 模型预测国际游客人数的季节和趋势性变化。Abu 等[7]采用 SARIMA 模型和指数平滑模型预测彭亨国家公园的游客人数，结果表明，指数平滑模型的预测性能不如 SARIMA 模型。此外，考虑到外生因素的影响，延伸出了 SARIMAX 模型，考虑外部因素对预测信息的作用。譬如，Arunraj 等[8]通过构建一个具有外部变量的

SARIMAX 模型预测零售商店中易腐食品的日销量。Park 等[9]利用 SARIMAX 模型预测来自中国大陆和美国的香港旅游人数。这些研究为旅游需求预测提供了有价值的参考。

1.2.2 基于网络搜索指数的预测

网络搜索引擎中特定关键词的搜索量反映了用户对相关问题的关注程度和相关事务的需求倾向。利用网络搜索信息数据可以在一定程度上反映游客在一段时期内的旅游倾向，并作为解释变量对游客流量进行预测，提高旅游预测的准确性。关于网络搜索信息数据在预测领域的研究最早由 Ginsberg 等[10]用于美国流感的预测。之后，大量学者开始对于网络搜索信息数据进行了深入探讨和广泛应用。譬如，张瑞等[11]以上海商品零售价格指数为研究对象，结合 RPI 相关的网络搜索数据进行预测研究。实证表明，互联网搜索数据的引入有助于提高零售价格指数的预测性能。Clark[12]通过谷歌趋势数据构建预测模型来预测美国国家公园游客数量，研究发现，谷歌趋势模型优于自回归模型，显著提高了预测精度。黄先开[13]以北京故宫为研究对象，搜集百度搜索数据和景区游客数据进行预测研究，实证表明，百度搜索信息能够提高预测精度。因此利用网络搜索信息捕捉游客对旅游目的地的关注度，能够提高旅游需求的预测性能。

网络搜索数据已然成为提高预测精度的有效数据来源。但网络搜索信息数据依赖于关键词选取，信息庞杂，搜集和有效提取存在各种困难。Xie 等[14]认为将网络搜索信息纳入预测模型时，过少的关键词变量可能会丢失大量有用信息，而过多的关键词变量可能引起多重共线性或过拟合问题。Li 等[15]提出尽管网络搜索信息数据能够反映游客的关注度信息且具有数据收集成本低的优势，但网络搜索信息数据的质量问题，譬如噪声、不相关数据可能会影响预测效果。因此，筛选出有效的网络搜索信息数据并合成网络综合搜索指数对于旅游预测研究是至关重要的。

如何对网络搜索关键词进行有效的筛选，已有大量学者进行研究。譬如，文献[16][18]利用时差相关法(TDC)筛选关键词：即计算每个关键词与游客流量提前

0~N 期的皮尔逊相关系数，设置阈值 R，最终选取皮尔逊相关系数大于 R 的关键词。陆利军和廖小靖[19]提出一种关键词优化选择方法——关键词集中度，选择具有稳定特征的关键字。Peng 等[20]将赫斯特指数(HE)和 TDC 法相结合，提出了 HE-TDC 关键词筛选新方法，通过对九寨沟游客流量数据的实证分析，发现该方法能够较好地筛选出具有预测能力的关键词，进而提高预测效果。Yao 等[21]通过 R/S 和时差相关分析(TDR)构建了一个综合网络搜索指数集 S，发现综合网络搜索指数集 S 中的关键词不但与九寨沟旅游人数的时间序列高度相关，并且与其历史游客人数序列呈现出一致的变动趋势。

此外，由于网络搜索信息数据存在较大噪声，往往导致数据呈现非线性和非平稳的特征。李晓炫等[22]采用 EMD 方法对数据进行噪声处理，利用去噪后的网络搜索数据预测游客流量，取得了良好的效果。陆利军[16]采用 EMD 去噪方法对游客流量和网络搜索信息数据进行去噪处理，并构建关于游客流量的预测模型。结果表明，EMD 去噪方法可以提高其预测精度。梁小珍等[23]通过 EMD 方法对关键字搜索量进行去噪，剔除了其中的高频噪声，并将其应用于中国民航旅客流量的综合预测，结果表明该方法具有良好的预测性能。

为了将网络搜索信息更方便地纳入预测模型，需要将重要关键词的搜索信息合成一个综合指数。文献中综合指数的合成方法主要有简单加合法[22]、相关系数加权法[24]、移位加合法[25]和移位加权合成法[26]。此外也有学者针对网络搜索信息数据存在的问题在方法上进一步改进，例如孙毅等[27]针对网络搜索信息具有共线性和评价指标权重的非客观性等问题，采用主成分分析的思想对网络搜索指数进行综合。Li 等[28]使用 PCA 方法去除冗余信息，结合改进的 BP 神经网络构建预测模型。然而，基于 PCA 的网络搜索指数只包含原始网络搜索数据中的线性信息，而忽略了非线性信息；为了有效地捕获非线性信息，Xie 等[14]提出了一种利用 KPCA 方法合成网络搜索指数的方法，并应用在旅游市场上，可以有效提升预测效果。

1.2.3 基于文本情感分析的预测

许多信息来源如社交媒体、新闻文章和在线评论，对旅游市场有很大的影响因此需要对其文本数据进行挖掘和分析，以提取有价值的信息辅助预测。而情感

分析成为了近年来文本挖掘中最受关注的研究课题，其主要分为两类：基于情感词典和基于机器学习的情感分析，对此学者们做出了大量的研究工作。基于情感词典的情感分析是根据经验对情感词进行总结、整理和归纳，构建相关情感词典，将预处理之后的海量信息与情感词典中收纳的词进行匹配，从而判断该文本的情感情极性[29]。Liu 和 Zhang[30]通过人工方法、基于词典的方法以及基于语料的方法构建出三种不同的情感词典。赵妍妍等[31]爬取海量新浪微博数据，利用情感词语种子来获取表情符种子对微博中的大规模语料使用简单的文本统计算法，构建出一个包含 10 万词语/词组的情感词典。基于机器学习的情感分析是对带有情感标签的数据进行网络训练形成一个情感分类器，再利用该情感分类器预测新句子的情感极性。用于分类的机器学习算法主要有人工神经网络（ANN）、朴素贝叶斯、支持向量机(SVM)等。Tao 等[32]利用人工神经网络（ANN）机器学习方法测量新浪微博文本数据的情感价值。实证表明，基于社交媒体数据的人工神经网络应用有助于丰富现有的游客情感研究方法。陈新元等[33]对微博中的表情图片及符号提取情感特征建立数据库，并使用 10 折的朴素贝叶斯分类器对微博文本进行分类。Wawre 等[34]在文本分类中分别利用支持向量机和朴素贝叶斯模型进行分类，结果表明朴素贝叶斯在训练样本数较多的情况下，该方法的分类准确率要高于其他模型。尚永敏等[35]用八爪鱼采集器对京东商城热销笔记本顾客评论进行采集，选用支持向量机、朴素贝叶斯以及 SnowNLP 分类方法进行情感分析，结果表明基于情感词典的 SnowNLP 库分类模型效果最好。

多项研究也证明情感分析能够提高预测效果。刘苗等[36]基于百度搜索采集消费者情感相关新闻文本，结合模型计算新闻文本的情感倾向，用于反映中国消费者的情况，同时从侧面反映了舆论对消费者的影响。虽然社交媒体可能会提供过度放大的文本数据，但旅游业中的媒体及其在促进旅游目的地方面的主导作用已经引起了相当大的关注。Colladon 等[37]利用在线旅游论坛进行文本分析，讨论从

在线社区研究中提取的变量的有用性，以便预测欧洲各国首都机场的国际旅客人数。实证研究表明，带有在线文本变量的模型比单变量模型和基于谷歌趋势数据的模型具有更好的预测性能。王建成等[38]提出了一种基于神经主题的对话情感分析模型。旅游研究人员需要利用情感分析来研究图像形成、旅游体验的表征、情感和行为意图。Brochado 等[39]以葡萄牙为例，构建了两种基于谷歌搜索数据的正面情感指数和负面情感指数验证是否影响股票市场。结果表明，构建的情感指数对市场回报和成交量的短期预测有积极作用。虽然社交媒体可能会提供过度放大的文本数据，但旅游业中的媒体及其在促进旅游目的地方面的主导作用已经引起了相当大的关注。Fan 等[40]将 Bass 模型与情感分析相结合，通过历史销售数据和在线评论数据来预测商品销售情况。结果发现，基于 Bass 模型的情感分析的加入对其预测有显著提升。因此，情感分析有助于改善预测效果。

此外如何构建情感指数使其更好的达到预测效果也是至关重要的。在相关的研究中，情感指数有不同的构建方法，如对情感指数简单相加，基于权重或正负情感的非对称性计算等。Checkley 等[41]探讨从微博网站提取的情感指数预测股市能力的研究中衍生出了两种情感度量指标，分别为看涨情感指数、积极和消极情感之间的一致性指数。Liang 等[42]在研究所构建的情感指数对上证综合指数实际波动率(RV)的预测能力中，采用对所定义的情感倾向值赋予权重合成综合看涨情感指数。Zhang 等[43]在预测三亚游客人数研究中采用简单相加的方法生成情感指数。陈晓红等[44]在股票预测研究中，基于 Mao[45]使用的体系，根据三种不同情感的发贴数量，定义情感指数。王晓丹等[46]通过对比正向和负向情感指数的对数形式，研究了互联网新闻媒体对我国股市的影响，以此来反映整体舆情的变化趋势。

1.2.4 基于多源数据融合的预测

来自单一来源或平台的数据可能会限制模型预测性能的稳定性和泛化性。解

决这个问题的一种方法是融合文本和网络搜索信息进行预测。Yang 等[47]应用包括传统经济数据和谷歌搜索数据等数据信息预测月度原油价格；研究表明，多源数据融合了传统经济数据和谷歌搜索数据的优势，在水平和方向度方面都获得了最佳的预测性能。Pan 和 Yang[48]结合搜索引擎查询、网站流量和每周天气信息等多源数据预测目的地每周酒店入住率，证明不同数据源的融合可以减少预测误差，提高预测精度。

随着社交媒体数据中文本信息的提取，文本中反映的情感也在一些文献中得到应用。Zhang 等[43]根据网络搜索数据和社交媒体数据构建搜索指数和情感指数预测中国三亚的游客量。研究发现，基于搜索指数和情感指数的预测模型具有较好的预测效果。陈晓红等[44]基于社交平台新浪微博和搜索引擎百度指数构建情感指数和搜索指数，考察了情感指数和搜索指数共同作用下对股价预测的影响，研究表明，当两种指数同时作用时，较大提升了股价的预测精度。

综上所述，目前所提出的旅游需求预测模型虽具有一定优点，但也存在以下几点问题：

首先，在大数据时代，利用网络搜索信息可以在一定程度上反映游客在一段时期内的旅游倾向，并作为解释变量对游客流量进行预测，能够提高旅游预测效果。但网络搜索数据的质量，包括关键词的覆盖范围和准确性、网络搜索指数合成的有效性，对网络搜索数据作为辅助变量进行旅游预测起到了关键性作用。因此，在利用网络搜索数据对游客流量进行预测时，应该重点提出更严格更规范的网络搜索关键词选择以及网络搜索指数合成方法，从而在较大程度上减少网络搜索数据的不相关信息和噪声的干扰。本文针对上述问题提出一种网络搜索信息提取合成新技术——R/S-TDC-EMD-KPCA 方法。该方法综合考虑了网络搜索信息数据存在噪声、复杂性和非线性的特点，将 R/S 和 TDC 方法相结合，选择具有预测能力的关键词，确保所选关键字的准确性；然后将这些具有预测能力的关键

词进行 EMD 降噪，提高网络搜索信息数据的质量；最后通过 KPCA 合成网络综合搜索指数，有效地捕获网络搜索信息数据的非线性信息，从而得到了更加精准和稳定的预测结果。

其次，社交平台数据尽管具有一定的优势，例如包含反映游客积极或消极情感的信息，但大部分研究主要集中在如何更好地对富含感情倾向的文本数据进行情感极性判断，或是利用特定公式构建情感指数，对于如何将情感指数有效融合在旅游预测模型中也是至关重要的。因此，在情感分析研究中，应充分考虑自然语言处理和情感分析的融合创新。本文针对这个问题采用 SnowNLP 方法对微博文本进行自然语言处理，并两种不同方法编制情感指数，分别是基于正负情感简单相加的情感指数和基于正负情感非对称的情感指数，将这两种情感指数一同作为辅助预测的变量输入到模型中进行旅游预测研究，从而获得更好的预测效果。

最后，在旅游预测中来自单一来源或平台的数据可能会限制模型预测性能的稳定性和泛化性。基于文本和网络搜索信息融合的旅游预测模型有助于旅游目的地更好地了解游客行为，并及时做出改善旅游需求的决定。本研究以百度指数为数据来源生成网络综合搜索指数，新浪微博为数据来源生成情感指数，基于基于文本和网络搜索信息融合构建 SARIMAX 预测模型，从而提高旅游预测精度。

1.3 研究内容及创新点

1.3.1 研究内容

在旅游预测于基于文本和网络搜索信息融合有助于旅游目的地更好地了解旅游行为，并及时做出改善旅游需求的决定。具体研究内容如下：

第一，由于网络搜索信息存在多噪声、非线性、高波动等特点，在关键词的选择和指数合成过程中存在诸多困难。因此，提出了一种新的网络搜索指数选择及合成技术——R/S-TDC-EMD-KPCA 方法，首先收集海南月度游客流量及相关关键词的网络搜索信息数据，利用 R/S-TDC-EMD-KPCA 方法提取网络搜索信息

合成网络综合搜索指数；然后检验网络综合搜索指数与海南游客流量之间的协整关系和格兰杰因果关系；最后将网络综合搜索指数和游客流量作为输入变量，分别利用 ARIMA、ARIMAX 和 BP 神经网络方法建立预测模型，针对引入基于不同合成技术的网络综合搜索指数的预测模型的预测效果进行评价。

其次，游客流量的时间序列具有周期性和高波动性的特点，来自单一来源或平台的数据可能会限制模型预测性能的稳定性和泛化性。由此，提出一种基于文本和网络搜索信息融合的旅游预测新方法。首先基于 R/S-TDC-KPCA 方法合成网络综合搜索指数；同时，从新浪微博中提取与最优关键词有关的文本数据，对提取的文本信息实施数据清洗，并采用基于正负情感简单相加和基于正负情感非对称的方法来构建情感指数。最后，基于网络综合搜索指数和情感指数建立 SARIMAX 预测模型对海南游客流量进行预测。

1.3.2 创新点

本文的创新与特色之处在于：

第一，提出一种网络搜索信息提取合成新技术——R/S-TDC-EMD-KPCA 方法。该方法综合考虑了网络搜索信息数据存在噪声、复杂性和非线性的特点，将 R/S 和 TDC 方法相结合，选择具有预测能力的关键词，确保所选关键字的准确性；然后将这些具有预测能力的关键词进行 EMD 降噪，提高网络搜索信息数据的质量；最后通过 KPCA 合成网络综合搜索指数，有效地捕获网络搜索信息数据的非线性信息，从而得到了更加精准和稳定的预测结果。

第二，采用两种不同方法编制情感指数，分别是基于正负情感简单相加的情感指数和基于正负情感非对称的情感指数，研究发现不同的情感指数编制方法会对预测结果有一定的影响，基于正负情感非对称的情感指数相比于正负情感的简单加总更能反映游客情绪，将基于不同人类心理行为的情感指数进行有效综合，可以获得更好的预测效果。

第三，以百度指数为数据来源生成网络综合搜索指数，新浪微博为数据来源生成情感指数构建预测模型，验证了基于文本和网络搜索信息融合的有效性。通过对 SARIMAX 模型与 SARIMA 模型的预测精度比较发现，无论添加网络综合搜索指数还是情感指数在预测海南省游客人数时都对提高预测准确性起到了积

极的作用。网络搜索量以及微博文本的内容都会反映游客对旅游目的地的关注度和正负情感倾向，并在决定旅游需求方面发挥作用。

1.4 研究结构安排

全文共分为 5 章进行阐述，研究结构如图 1.1 所示。

第 1 章：引言。本章在研究背景及意义的基础上，从四个角度出发对现有文献的研究内容进行综述，分别是：基于历史游客流量的预测；基于网络搜索指数的预测；基于情感分析的预测；基于多源数据融合的预测。阐释现有旅游预测模型中所存在的一些问题与不足，并提出本文构想。

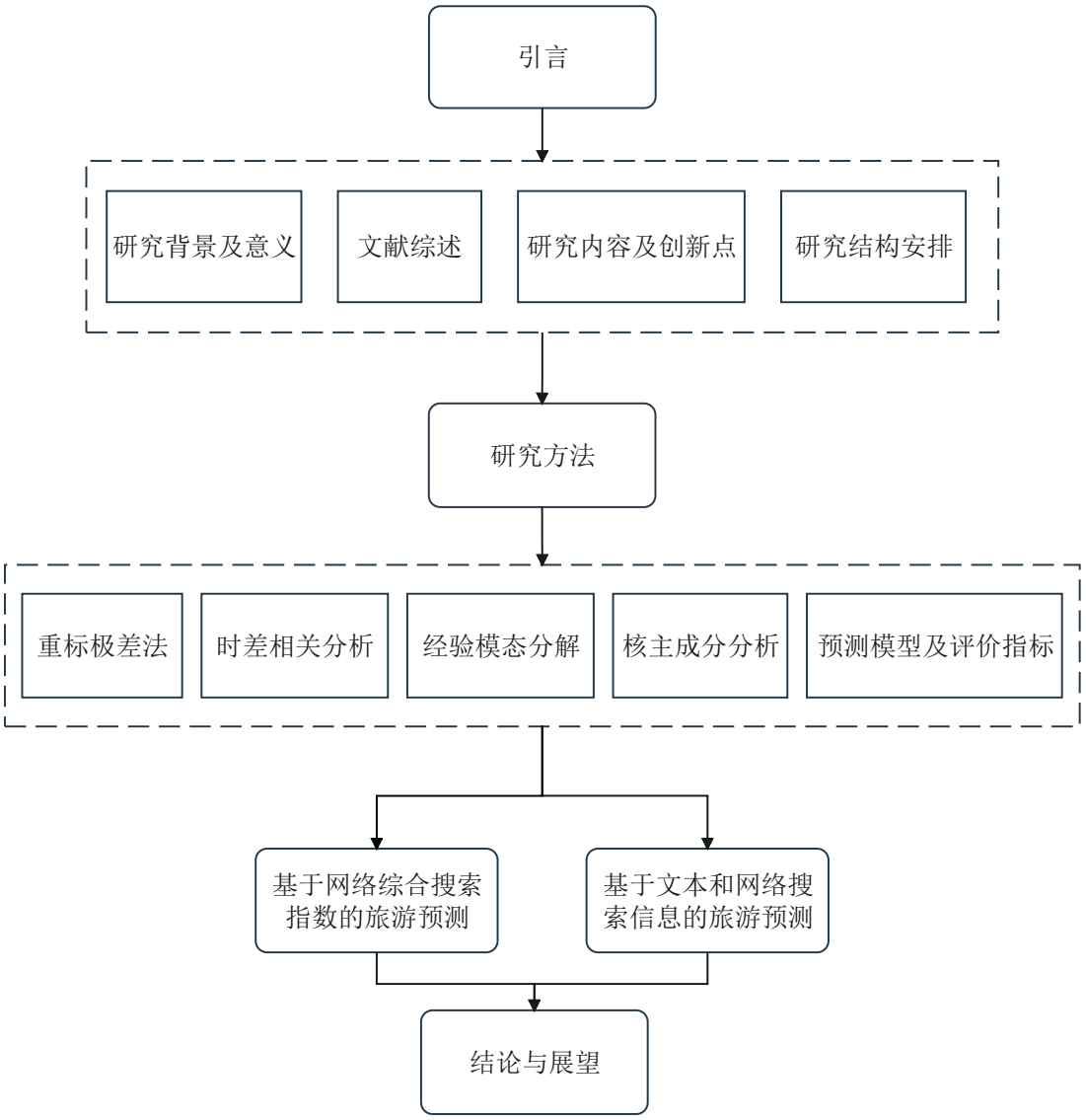


图 1.1 论文框架示意图

第 2 章：研究方法。本章对主要的研究方法重标极差法、时差相关分析法、经验模态分解、核主成分分析、SnowNLP 以及所用的预测模型和评价指标分别进行介绍。

第 3 章：基于网络综合搜索指数的游客流量预测。本章是为了充分验证所提出的网络搜索数据合成技术——R/S-TDC-EMD-KPCA 方法的有效性，以海南省的月度游客流量作为研究对象，将构建的网络综合搜索指数和游客流量作为输入变量，分别使用 ARIMA、ARIMAX 和 BP 神经网络方法构建预测模型，并对不同预测模型的预测效果展开评估。

第 4 章：基于文本和网络搜索信息的游客流量预测。本章考虑到来自单一历史信息或互联网平台的数据可能会限制模型的预测精度和泛化能力，以百度指数为数据来源生成网络综合搜索指数，新浪微博为数据来源生成情感指数构建 SARIMAX 预测模型。以中国海南省的月度游客流量为预测对象进行实证研究，最后基于不同评估方法对提出的方法的预测性能进行评估。

第 5 章：结论与展望。本章主要探讨论文的研究成果，并对其中的不足进行总结，最后对未来的研究方向进行展望。

2 研究方法

2.1 重标极差法

重标极差法（Rescaled Range Analysis, R/S）主要用于计算 Hurst 指数[49]，该指数最初是由 Hurst 提出，并以他的名字命名。该指数的发现是基于 Hurst 对尼罗河进行长期的水文观测，在此基础上提出用 R/S 法建立 Hurst 指数，用于考察时间序列的自相关特性。该指数的计算方法如下：

1. 将长度为 m 的时间序列划分为长度为 n 且不重叠的 a 个子序列，记为 $T_1,T_2,T_3,\cdots,T_a$，其中 $a=m/n$；则 $T_i=\{x_{ji}\}$，其中 $j=1,2,\cdots,n;i=1,2,\cdots,a$。$x_{ji}$ 表示序列 T_i 中的第 j 个观测值。

2. 分别计算每个子序列 T_i 的均值，记为 $\overline{x}_{T_i}$，

$$\overline{x}_{T_i}=\frac{1}{n}\sum_{j=1}^{n}x_{ji} \tag{2.1}$$

3. 分别计算每个子序列 T_i 的标准差，记为 S_{T_i}，

$$S_{T_i}=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\overline{x}_{T_i})^2} \tag{2.2}$$

4. 在每个子序列 T_i 中，逐一计算前 k 个数的相对该子序列 T_i 均值 $\overline{x}_{T_i}$ 的累积离差，其中 $k=1,2,\cdots,n$

$$y_{k,i}=\sum_{j=1}^{k}(x_{ji}-\overline{x}_{T_i}) \tag{2.3}$$

5. 计算每个子序列 T_i 的累积离差的极差，记为 R_{T_i}

$$R_{T_i}=max(y_{k,i})-min(y_{k,i}) \tag{2.4}$$

6. 对于每个子序列 T_i，计算其重标极差，记为 $\frac{R_{T_i}}{S_{T_i}}$.

7. 从而 Hurst 推出如下关系：

$$E\left[R_{T_i}/S_{T_i}\right]=c\cdot n^{H} \tag{2.5}$$

式中，R_{T_i}/S_{T_i} 是每个子序列 T_i 的重标极差变量，$E\left[R_{T_i}/S_{T_i}\right]$ 是 a 个子序列的平均重标极差，c 为常数，H 即为 Hurst 指数。

对(2.5)式两边同时取对数可得：

$$ln(E[R/S]) = H \cdot ln(n) + ln(c) \tag{2.6}$$

若以 $ln(E[R/S])$ 为因变量，$ln(n)$ 为自变量，根据 n 的不同取值，计算对应子序列的平均重标极差值，进而将它们作为数据样本进行最小二乘回归，从而可得 Hurst 指数 H。

Hurst 指数 H 反映了时间序列的自相关性，度量了序列的趋势强度。这一指数与趋势的关系如下：

1）当 $0.5 < H \le 1$ 时，说明序列具有长记忆性，未来增量与过去增量正相关；

2）当 $0 < H < 0.5$ 时，说明序列具有反持续性，未来增量与过去增量负相关；

3）当 $H = 0.5$ 时，说明序列的变化趋势接近于随机，即未来增量和过去增量没有关系。

2.2 时差相关分析

时差相关分析(Time Difference Correlation Analysis, TDC)是测量时间序列相关系数领先、同步或滞后关系的常用方法。r_l 表示时差为 l 的相关系数，计算公式如下：

$$r_l = \frac{\sum_{t=1}^{n}(x_{t-l} - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_{t-l} - \bar{x})^2 \sum_{t=1}^{n}(y_t - \bar{y})^2}} \tag{2.7}$$

本文中 y_t 表示每月游客流量，$\bar{y}$ 是平均每月游客流量，时间序列 x_t 表示 t 时期的关键词搜索量，$\bar{x}$ 是平均搜索量，l 是时差数。当 $l < 0$ 时表示超前，$l > 0$ 时表示滞后。

2.3 经验模态分解

经验模态分解(Empirical Mode Decomposition, EMD)是 Huang 提出的一种处

理非平稳信号的新方法[50]。根据经验模态分解，将原始信号划分为若干本征模函数（IMF）和一个残差项，IMF 分量是具有时变频率的振荡函数，能够反映非平稳信号的局部特征。

EMD 的基本步骤如下：

第一步，将需要分解的序列进行局部处理，得到局部极值点。第二步，连结相邻局部极小值点和局部极大值点，得到一组均值为零的内插线段。第三步，对每一条内插线段进行包络分解，得到上包络和下包络。第四步，将每个内插线段与其上包络和下包络的平均形态进行相减，得到该内插线段对应的本征模态函数，并将其分离出来。第五步，重复上述步骤，至所得的模态函数之和接近原信号，这些函数称为本征模态函数。最终 EMD 结果为一组本征模态函数和一个残差项。

其中，EMD 的使用存在一些限制条件：

(1) 在一段时间内，函数的局部极值点和过零时的数量应该保持一致或最多相差一个；

(2) 在任何时间点，局部最大值的包络(上包络)和局部最小值的包络(下包络)的平均值必须为零。

2.4 核主成分分析

PCA 是一种经典的特征提取方法，但该方法只能处理具有线性相关特性的变量。核主成分分析(Kernel Principal Component Analysis, KPCA)是一种新的非线性信息提取方法，基本思想是通过非线性映射将输入空间投影到高维特征空间，然后在高维特征空间中对映射数据做主成分分析[51]。为解决“维数灾难”问题，利用核函数来替代特征空间中的样本内积运算，进而实现对非线性数据的线性化处理。

给定训练数据集 $X=[x_1,x_2,\cdots,x_m]\in R^{n\times m}$，其中 $x_1,x_2,\cdots,x_m$ 为 n 维列向量，m 是样本个数（本文为关键词搜索量的统计月数），n 是变量的个数（本文为关键词个数）。非线性投影映射到高维特征空间 F（记作 d 维）上得到一个 $d\times m$ 的新矩阵：

$$\phi(X)=[\phi(x_1),\phi(x_2),\cdots,\phi(x_m)] \tag{2.8}$$

假设矩阵 $\phi(X)$ 已经进行中心化处理，即有

$$\sum_{i=1}^{m}\phi(x_i)=0 \tag{2.9}$$

特征空间的协方差矩阵为：

$$C=\frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\phi(x_i)^T \tag{2.10}$$

根据 PCA 理论，求解协方差矩阵 C 的特征值 λ 和特征向量 v：

由等式 $Cv=\lambda v$ 得到，存在 $\alpha_1,\alpha_2,\cdots,\alpha_m$ 使得

$$v=\sum_{i=1}^{m}\alpha_i\phi(x_i) \tag{2.11}$$

定义核矩阵 K:

$$K\left[(x_i,x_j)\right]_{m\times m}=\phi(x_i)\cdot\phi(x_j) \tag{2.12}$$

将式(2.10)、(2.11)和(2.12)代入 $Cv=\lambda v$ 化简得：

$$K\alpha=\lambda\alpha \tag{2.13}$$

通过式(2.13)可解出 $\alpha_1,\alpha_2,\cdots,\alpha_m$。对于任意样本 x，v^j 是第 j 个特征向量，F_j 是第 j 项核主成分，则高维映射 $\phi(x)$ 在特征空间上的投影公式为：

$$F_j=(v^j\cdot\phi(x))=\sum_{i=1}^{m}\alpha_i^jK(x_i,x) \tag{2.14}$$

2.5 SnowNLP

情感分析是自然语言处理（Nature Language Process）的一种，旨在识别文本中的情感信息，并对其进行加工、归纳和推理，以更好地理解文本内容，并将其应用于日常生活中，以提高工作的效率和质量。SnowNLP 是一个专门用于中文文本挖掘的 Python 库，其主要功能包括分词、词性标注、情感分析、关键词提取等。它运用的情感分类方法是朴素贝叶斯原理，通过计算出的先验概率与每个属性特征词的条件概率相乘，从而得到情感概率值；最后，将较大的情感概率值

的极性作为语句的情感倾向。输出的概率区间为[0,1] ，即积极情感倾向越强；若概率值越接近 1，则表示消极情感倾向越强。

朴素贝叶斯是一种基于贝叶斯决策理论的分类方法，假设存在一个二元分类问题，如c_1类与c_2类，样本具有 x 和 y 两个特征，则需要分别求解条件概率$P(c_1|x,y)$和$P(c_2|x,y)$。而$P(c_i|x,y)$（$i=1,2$）可以表达为：

$$P(c_i|x,y)=\frac{P(x,y|c_i)P(c_i)}{P(x,y)} \tag{2.15}$$

由于特征之间一般假设是独立的，所以$P(c_i|x,y)=P(x|c_i)P(y|c_i)$。

此后进行分类，其贝叶斯分类准则为：如果$P(c_1|x,y)>P(c_2|x,y)$，那么属于c_1类；如果$P(c_1|x,y)<P(c_2|x,y)$，那么属于c_2类。

2.6 预测模型及评价指标

2.6.1 季节性 naïve 模型

Seasonal Naïve 模型（SNaïve）是季节性数据的典型基准模型。它使用上一个时间周期中对应位置的观测值作为预测值。为了预测每月的旅游需求，一般模型是

$$\hat{y}_t = y_{t-12} \tag{2.16}$$

其中，$\hat{y}_t$是游客在 t 时刻的预测值，y_{t-12}是从 t 时刻滞后 12 月的游客到达实际值。

2.6.2 指数平滑模型

指数平滑模型（Exponential Smoothing）最早是在 1959 年美国经济学家 Brown 提出的，随后在不同学者的研究中不断发展完善。通常使用指数平滑模型的三个基本变体：简单的布朗指数平滑[52]；Holt 趋势修正的经验平滑[53]；以及 Holt-Winters 方法[54]。指数平滑模型是一种统计预测模型，它的原理是：生成的预测

值是过去观测值的加权平均值，当离观测值愈接近，则给予愈大的权重，这一模型能够快速准确地进行预测，在时间序列预测中具有普遍的适用性。它可以分为一次指数平滑模型、二次指数平滑模型和三次指数平滑模型。

一次指数平滑模型适合于无明显趋势变化的时间序列。公式如下：

$$S_t^{(1)} = \alpha Y_t + (1-\alpha) S_{t-1}^{(1)} \tag{2.17}$$

其中，α 为平滑系数，$S_t^{(1)}$ 为 t 时刻的一次指数平滑值，Y_t 为实际观测值。α 控制权重的下降速度，越接近 1，历史观测值的权重越大，因此平滑系数α 的选择在指数平滑模型很重要，在 R 中通常利用真实值和预测值的残差平方和来选择最合适的α 。

二次指数平滑模型适用于具有趋势特性的时间序列，由两部分组成：平滑值和趋势值。公式如下：

$$S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + b_{t-1}) \tag{2.18}$$

$$b_t = \beta(S_t - S_{t-1}) Y_t + (1-\beta) b_{t-1} \tag{2.19}$$

$$Y_{t+T} = S_t + b_t \cdot T \tag{2.20}$$

其中，α,β 为平滑系数，S_t 为 t 时刻的平滑值，b_t 为 t 时刻的趋势值，Y_{t+T} 为预测值，即为平滑值加上趋势值乘以预测超前期数。

三次指数平滑模型适用于具有季节周期特性的时间序列，其由三部分组成：平滑值、趋势值和季节分量，公式如下：

$$S_t = a\frac{Y_t}{I_{t-m}} + (1-a)(S_{t-1} + b_{t-1}) \tag{2.21}$$

$$b_t = \beta(S_t - S_{t-1}) + (1-\beta) b_{t-1} \tag{2.22}$$

$$I_t = \gamma\frac{Y_t}{S_t} + (1-\gamma) I_{t-m} \tag{2.23}$$

$$Y_{t+T} = (S_t + b_t \cdot T) I_{t-m+T} \tag{2.24}$$

其中，α,β,γ 为平滑系数，I_t 代表周期性，k 是周期的长度。

2.6.3 SARIMA 模型

自回归差分移动平均模型（ARIMA）是一种经典的时间序列模型。但有部分时间序列具有较强的季节性变化，通常使用季节性自回归差分移动平均模型（SARIMA）来表示，它是从 ARIMA 模型中衍生而来，集成了季节性、自回归和移动平均成分。通过使用季节性自回归阶数 P、季节性移动平均阶数 Q 和季节性差分阶数 D 建立 $SARIMA(p,d,q)(P,D,Q)$ 模型，公式如下：

$$\phi(B)\Phi_S(B)(1-B^S)^D(1-B)^d Y_t = \theta(B)\Theta_S(B)\varepsilon_t \tag{2.25}$$

其中 Y_t 是 t 月的旅游人数。$\phi(B)=(1-\sum_{i=1}^{p}\phi_i B^i)$ 是自回归算子，$\phi_1,\phi_2,\cdots\phi_p$ 为对应自回归算子参数；$\theta(B)=(1-\sum_{i=1}^{q}\theta_i B^i)\varepsilon_t$ 是移动平均算子，$\theta_1,\theta_2,\cdots\theta_p$ 为对应算子参数；$(1-B)^d$ 是 d 阶差分，$(1-B^S)^D$ 是季节性 D 阶差分；$\Phi_S(B)=1-\Phi_1 B^S-\Phi_2 B^{2S}-\cdots-\Phi_P B^{PS}$ 为 P 阶季节性自回归算子，$\Phi_1,\Phi_2,\cdots\Phi_P$ 为季节性自回归算子参数；$\Theta_S(B)=1-\Theta_1 B^S-\Theta_2 B^{2S}-\cdots-\Theta_Q B^{QS}$ 为 Q 阶季节性移动平均算子，$\Theta_1,\Theta_2,\cdots\Theta_Q$ 为季节移动算子参数。

当外生变量纳入模型时，SARIMAX 模型的具体形式为：

$$Y_t=\beta_0+\beta_1 X_{1,t}+\beta_2 X_{2,t}+\cdots+\beta_k X_{k,t}+\frac{\theta(B)\Theta_S(B)}{\phi(B)\Phi_S(B)}\varepsilon_t \tag{2.26}$$

其中，Y_t 是因变量的第 t 个观测值，$X_{1,t},X_{2,t},\cdots,X_{k,t}$ 为加入的外生变量，$\beta_1,\beta_2,\cdots,\beta_k$ 是回归系数。

2.6.4 BP 神经网络

BP 神经网络（Back-propagation Neural Network）是一种具有反向传播算法的人工神经网络，也是目前最流行的神经网络之一。BP 神经网络的结构是由多层神经元构成，分为输入层、隐藏层和输出层三部分，通过反向传播算法可以实现网络的训练和优化。它的优势在于，可以有效地处理复杂的机器学习问题，其中包括两个关键步骤：信号的前向传递和误差的后向传播，从而实现有效预测。

BP 神经网络的基本步骤如下：1.初始化网络权值和偏置：对于一个给定的问题，我们需要首先设计一个神经网络的结构，并对网络的权值和偏置进行初始化。2.前向传播：将输入样本通过网络层和激活函数进行计算，得到输出值，并将输出值与实际值进行比较，得到网络的误差。3.反向传播误差：将误差反向传递回网络，计算偏差和权值的梯度，并根据学习率和梯度下降算法更新网络参数。4.重复迭代直到满足停止条件：对于所有训练样本，反复执行 2 和 3 步，直到网络的误差满足预设的停止条件。5.测试网络性能：使用测试数据集对网络进行测试，计算预测精度和性能指标，如准确率、召回率、F1 值等。需要注意的是，在训练过程中还需要进行一些调整，如决定网络的层数和神经元个数、选择激活函数、设置学习率和停止条件等。

BP 神经网络的学习过程是通过一定量的训练数据来训练网络得到最优权重，从而实现网络对输入数据的正确分类或预测。反向传播算法是 BP 神经网络实现学习的核心，通过计算误差反向传递调整权值和偏置，从而实现网络的优化。

2.6.5 评价指标

为了评估不同模型的水平预测精度，本文通过以下公式计算平均绝对误差（Mean Absolute Error，MAE）、平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）、均方根误差（Root Mean Square Error，RMSE）、归一化均方根误差（Normalized Root Mean Square Error，NRMSE）均方根百分比误差（Root Mean Square Logarithmic Error，RMSPE），具体如表 2.1 所示。其中，y_i 是实际的月度游客流量，$\hat{y}_i$ 是游客流量的预测值，$\bar{y}$ 为游客流量的平均值。

表 2.1　评价指标

评价准则	定义	公式
平均绝对误差	预测值与样本真实值之间距离的平均值，其范围为[0，+∞)。	$MAE=\frac{\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert}{n}$
平均绝对百分比误差	其范围[0，+∞)，理论上，MAPE 的值越小，说明预测模型拟合效果越好，具有更好的精确度。	$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert$
均方根误差	在均方误差的基础上做平方根，衡量观测值与真实值之间的偏差，误差越大，该值越大。	$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
归一化均方根误差	在均方根误差的基础上进行归一化(除以均值)	$NRMSE=\frac{1}{\bar{y}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
均方根百分比误差	衡量模型预测值与序列观察值的差异程度。它与使用的单位无关，因此可用于比较不同单位的系列。	$RMSPE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i-\hat{y}_i}{y_i})^2}$

为了比较模型 1 相对于模型 2 在水平预测精度上的改进程度，利用 MAPE 定义相对改进指标 RI 值，公式如下：

$$RI_{Model_2}^{Model_1}=\frac{MAPE(Model2)-MAPE(Model1)}{MAPE(Model2)} \tag{2.27}$$

3 基于网络综合搜索指数的游客流量预测

3.1 预测框架

本文采用 R/S-TDC-EMD-KPCA 方法提取网络搜索信息，结合海南游客流量数据，分别以 ARIMA 模型、ARIMAX 模型和 BP 神经网络作为基准模型，构建预测模型，如图 3.1 所示。

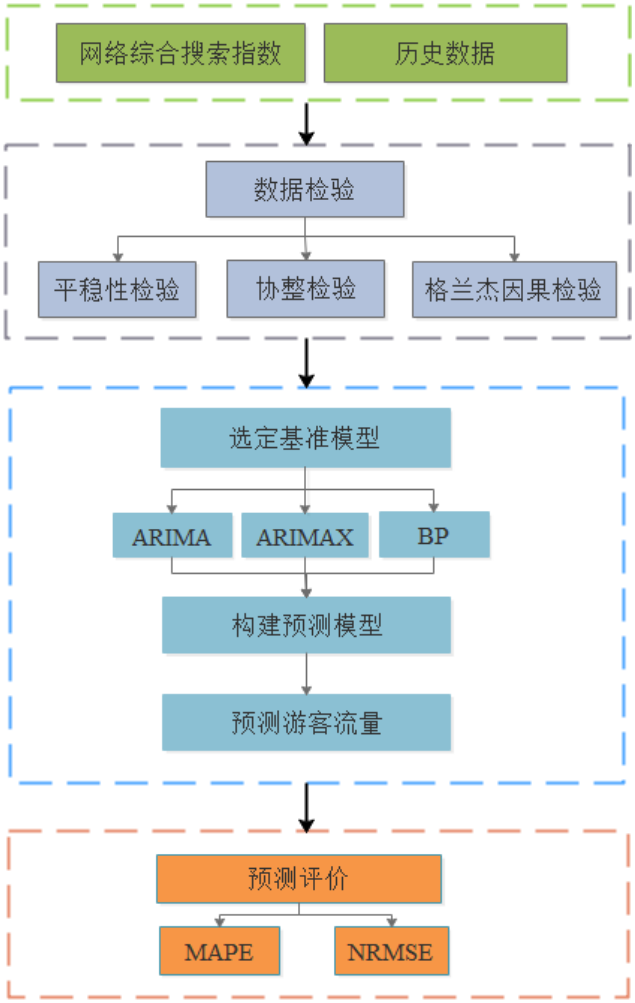


图 3.1　预测框架

3.2 数据采集

本文选取2011年1月至2019年12月的海南国内过夜游客人数的月度数据；网络综合搜索指数是通过关键词搜索量数据合成的。在中国，百度引擎占据了大部分市场份额，所以网络搜索关键字的所有数据均来自百度指数（http://index.baidu.com）。我们将2011年1月到2017年12月的84个数据作为训练样本，2018年1月至2019年12月的24个数据作为测试样本。

3.3 合成网络综合搜索指数

搜索引擎中特定关键词的搜索量反映了用户对相关问题的关注程度和相关事务的需求倾向[51]。利用网络搜索信息数据可以在一定程度上反映游客在一段时

期内的旅游倾向，并作为解释变量对游客流量进行预测，提高旅游预测的准确性。综合以往文献[55]的研究，结合旅游游客流量的数据特征，本文利用相关关键词的百度搜索信息合成网络综合搜索指数的具体步骤如下：

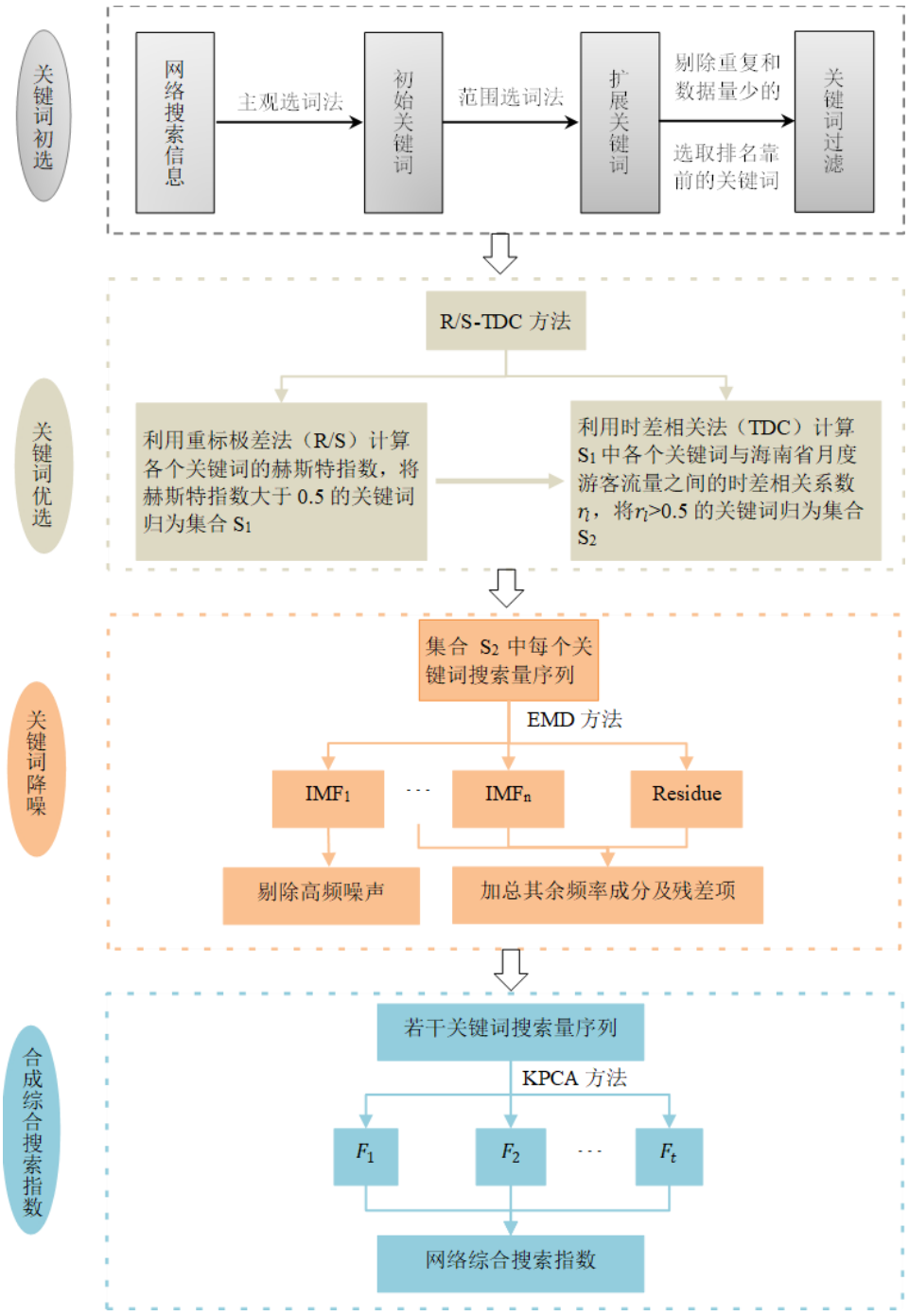


图 3.2　网络搜索指数合成框架

3.3.1 搜索关键词的初选

本文采用主观选词法，首先选取“海南天气”、“海南旅游”、“海南机票”、“海南住宿”、“海南美食”、“海南免税店”、“海南景点”等 7 个与海南旅游决策相关的词汇作为核心关键词，并搜索关键词对应的百度指数信息；并借助于百度指数的需求图谱能力，对这些关键字加以扩展到 63 个，如表 3.1 所示。

表 3.1 扩展的初始关键词

类别	序号	关键词	类别	序号	关键词
旅游	1	海南旅游	交通	32	海南机票
	2	海南旅游攻略		33	海南航空
	3	三亚旅游		34	海南高铁
	4	海口旅游		35	海南高铁最新时刻表
	5	海南地图		36	海南高铁票价
	6	海南旅游地图		37	海南汽车站
	7	海南旅游路线		38	海南机场
	8	海南旅游注意事项		39	海南航空机票查询
	9	海南旅游时间		40	海南航空官网
	10	海南旅行社		41	海南航空电话
	11	海南旅游局	餐饮	42	海南美食
	12	海南旅游图片		43	海南饭店
	13	海南自驾游攻略		44	海南特色美食
	14	海南旅游资讯		45	海南特色菜
	15	海南旅游价格		46	海南特色小吃
	16	海南旅游攻略自由行	住宿	47	海南住宿
	17	海口天气		48	海南酒店
	18	三亚天气		49	海南酒店预订
	19	海南天气		50	三亚酒店
	20	海南温度		51	海南酒店排名
购物	21	海南购物		52	海南酒店价格
	22	海南免税店	景点	53	海南景点
	23	海南特产		54	海南三亚
	24	海南特产休闲食品		55	海南旅游景点
	25	海南特产批发		56	三亚旅游景点
	26	海南购物中心		57	海南三亚旅游景点
	27	海南购物广场		58	海南岛
	28	海南特色水果		59	天涯海角
	29	海南特产水果		60	亚龙湾
	30	海南水果		61	槟榔谷
	31	海南椰子		62	蜈支洲岛
				63	海棠湾

在删除重复的关键词和数据不完整的关键词之后，最终确定 39 个有效的初始关键字，如表 3.2 所示。通过百度指数获取这些关键词的月度搜索量，作为我们选择关键字的基础数据集。

表 3.2　有效的初始关键字

类别	序号	关键词	类别	序号	关键词
旅游	1	三亚旅游	景点	21	蜈支洲岛
	2	海口旅游		22	槟榔谷
	3	海南旅游	交通	23	海南机票
	4	海南旅游攻略		24	海南航空
	5	海南地图		25	海南高铁
	6	海南旅游地图		26	海南航空官网
	7	海南旅游注意事项		27	海南航空电话
	8	海南旅游局	住宿	28	三亚酒店
	9	海南旅游图片		29	海南酒店
	10	海南旅游价格	餐饮	30	海南美食
	11	海口天气		31	海南特色小吃
	12	三亚天气	购物	32	海南免税店
	13	海南天气		33	海南特产
	14	海南温度		34	海南特色水果
	15	海南旅游景点		35	海南特产水果
	16	三亚旅游景点		36	海南水果
景点	17	海南岛		37	海南椰子
	18	天涯海角		38	海南景点
	19	亚龙湾		39	海南三亚
	20	海棠湾			

3.3.2 搜索关键词的优选

通过重标极差法（R/S）计算 Hurst 值，建立关键词集 C_1。计算 39 个候选关键词中每个关键词的H_i，保留满足$0.5 < H_i < 1$的关键词，筛选得到 23 个关键词，对应的 Hurst 值如表 3.3 所示，并将这些关键词构成的集合记 C_1。

表 3.3　关键词集 C_1 及其 Hurst 值

关键词	Hurst 值	关键词	Hurst 值
海口旅游	0.6804	海南岛	0.6091
海南旅游	0.7972	天涯海角	0.7399
海南旅游攻略	0.6451	亚龙湾	0.6679
海南地图	0.6014	海棠湾	0.8627
海南旅游地图	0.5343	蜈支洲岛	0.8865
海南旅游局	0.7116	槟榔谷	0.6131
海南旅游图片	0.6785	海南航空	0.9679
海南天气	0.5863	海南航空电话	0.9416
海南旅游景点	0.5377	三亚酒店	0.5816
三亚旅游景点	0.9224	海南酒店	0.7676
海南美食	0.6498	海南椰子	0.7308
海南免税店	0.6068		

然后，将筛选得到的每个关键词i通过 TDC 计算出滞后三期内的最大相关系数$r^{(i)}$，建立关键词集 C_2。在 23 个候选关键词中，有 9 个关键词满足$r^{(i)} \geq 0.5$，包括“海南旅游”、“海南地图”、“海南旅游局”、“海南旅游图片”、“海南航空电话”、“三亚酒店”、“海南美食”、“海南免税店”、“海南椰子”，如表 3.4 所示。

表 3.4　关键词集 C_2 及其时差相关系数

关键词	滞后期	最大相关系数
海南旅游	3	0.73
海南地图	1	0.54
海南旅游局	3	0.74
海南旅游图片	3	0.55
海南航空电话	3	0.81
三亚酒店	3	0.68
海南美食	1	0.67
海南免税店	1	0.58
海南椰子	2	0.60

综上所述，关键词集 C_2 中的关键词数据与海南月度游客流量具有较高的相关性，同时也呈现出与游客流量相同的变化趋势。

3.3.3 关键词搜索量降噪

由于关键词数据存在噪声，需要进一步降低噪声对有效网络搜索信息数据的干扰，将最终确定的关键词集 C_2 的每项关键词搜索量数据进行 EMD 分解。分解后分别得若干条 IMF 函数序列和一条残差，分解结果如图 3.3 所示。参考以往文献[16][22][23]的做法，最高频率的 IMF 函数可视为网络综合搜索指数的短期剧烈波动，将其作为噪声予以剔除。

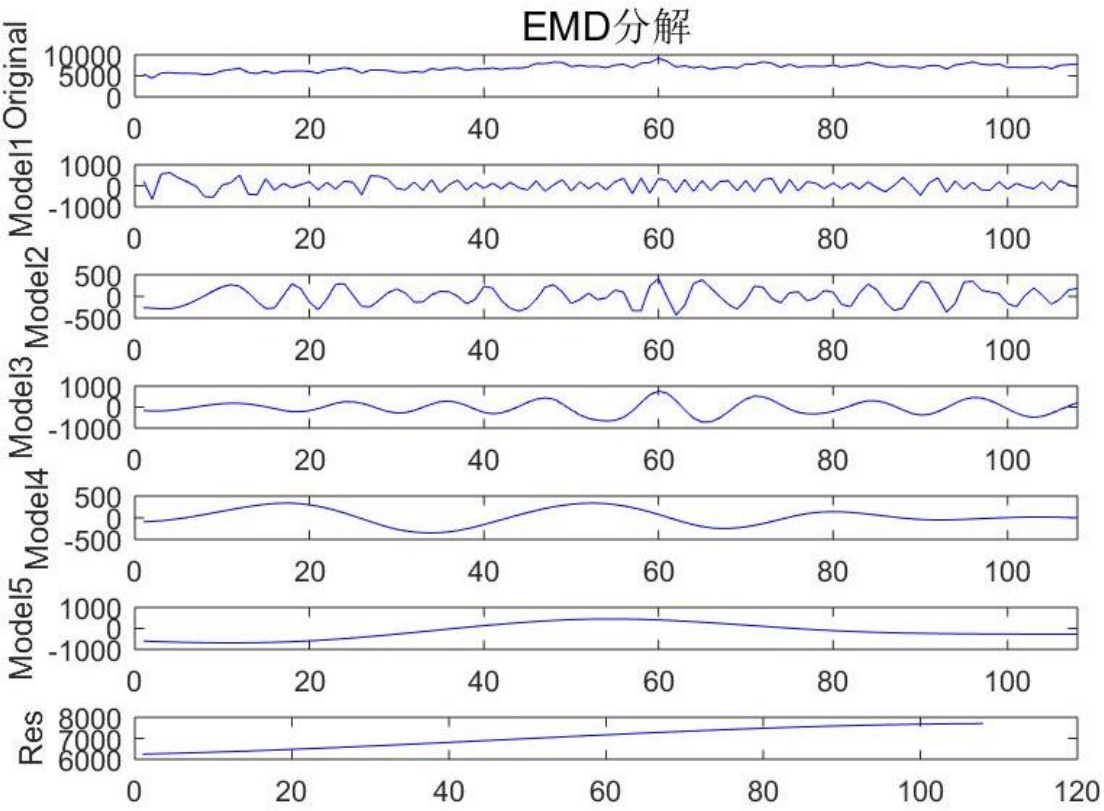
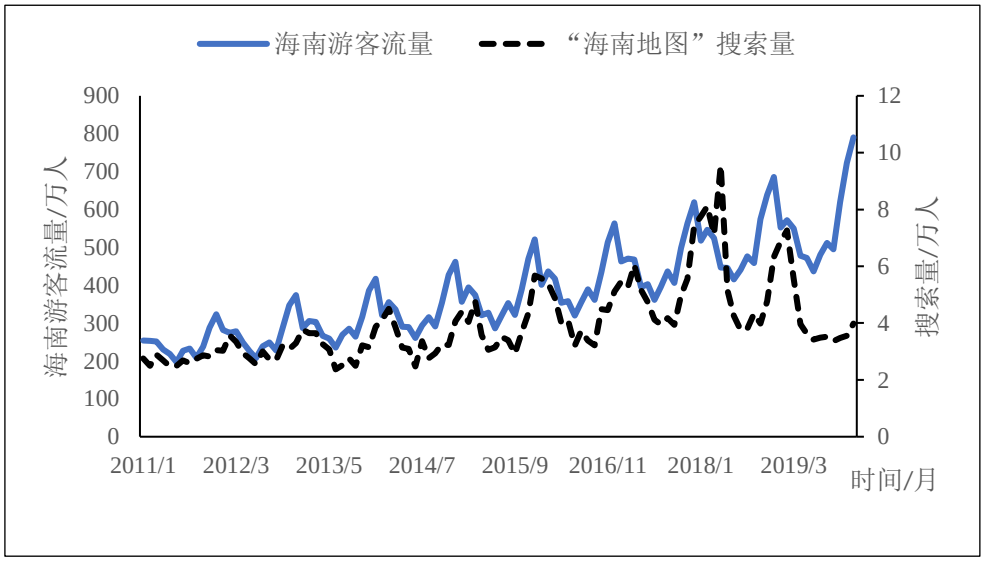
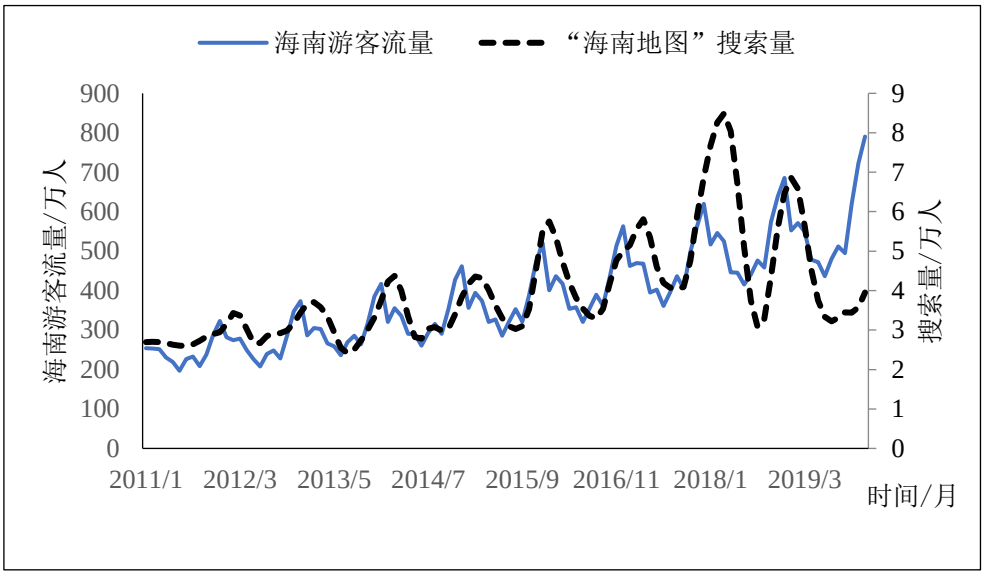


图 3.3 EMD 分解结果

以关键词“海南地图”搜索量为例，图 3.4 显示出降噪前和降噪后与海南游客流量的变化图，可以看出，在进行降噪处理之前，搜索量存在较多的尖峰，而降噪之后的搜索量和海南游客流量的变化波动更接近，其波峰、波谷以及周期性呈现大致相同的趋势，因此起到了减少噪声对网络搜索信息数据干扰的作用。



(a) 降噪前



(b) 降噪后

图 3.4　降噪前后“海南地图”搜索量与海南游客流量的变化图

3.3.4 KPCA 合成网络综合搜索指数

为了将网络搜索信息更方便地纳入预测模型，需要将重要关键词的搜索信息合成一个综合指数。考虑到网络搜索数据存在共线性问题，以及合成指数法存在赋权不客观的问题[27]，本文选择 Xie 等[14]提出的基于 KPCA 合成网络综合搜索指数的方法，能够有效捕捉原始网络搜索数据中的非线性信息，从而可以很好地预测旅游需求。具体内容如下：

首先对关键词集 C_2 数据进行标准化处理；采用归一化数据的径向基核函数

计算核矩阵，对其中心化得到中心化核矩阵，并求解其特征值和特征向量；然后对特征值进行降序排序，计算各核主成分的贡献率和累积贡献率，其计算结果见表 3.5。

表 3.5　核主成分结果

核主成分	特征值	贡献率%	累计贡献率%
1	4.383305	48.70	48.70
2	2.576448	28.63	77.33
3	0.491941	5.47	82.80
4	0.438654	4.87	87.67
5	0.342344	3.80	91.47
6	0.278614	3.10	94.57
7	0.228279	2.54	97.11
8	0.150341	1.67	98.78
9	0.110075	1.22	100.00

由表 3.5 可知，前 4 个核主成分 F_1、F_2、F_3、F_4 的累计贡献率已经超过了 85%，因此，将这 4 个核主成分替代原有的 9 个关键词数据，并以各个核主成分所相应的特征数值占所提取核主成分总特征值之和的百分比进行加权[27]，最后合成网络综合搜索指数 I 如下：

$$I = 0.5555F_1 + 0.3265F_2 + 0.0623F_3 + 0.0556F_4$$

同时为了减少异常值的影响，本文将海南旅游客流量和上述综合的网络综合搜索指数均转换为对数形式（lnT和lnI），其中 T 代表海南旅游客流量，I 代表网络综合搜索指数。图 3.5 是海南游客流量与网络综合搜索指数的序列图，可以看出两个序列谷峰和变化趋势具有一致性，说明游客流量与网络综合搜索指数之间存在相关性。

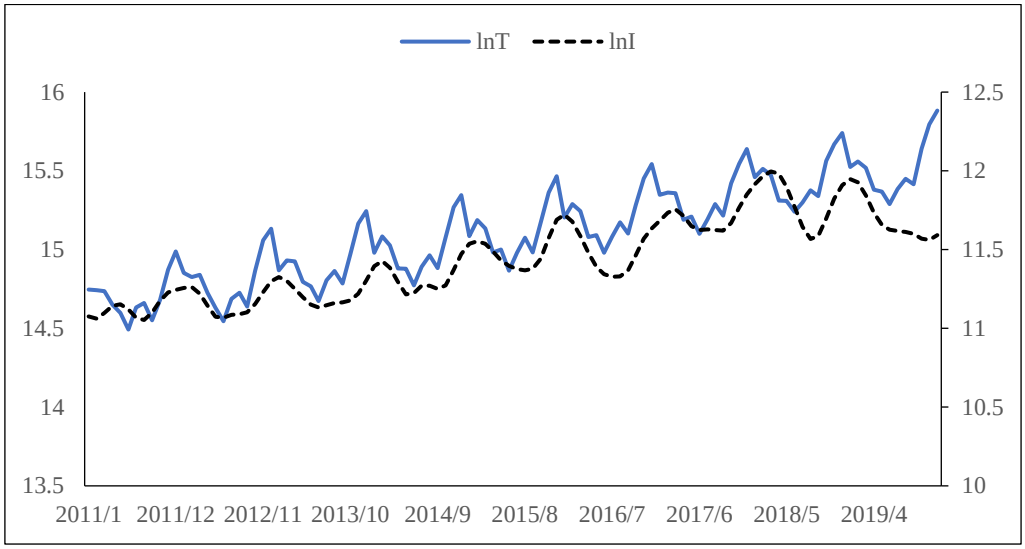


图 3.5　海南游客流量与网络综合搜索指数变化图

3.4 协整检验和格兰杰因果关系检验

通过检验海南游客流量与网络综合搜索指数之间的协整关系和格兰杰因果关系，探究两者之间的长期均衡关系，判断网络综合搜索指数是否具有可靠的预测能力。

由于海南游客流量与网络综合搜索指数均为时间序列，需要对其进行平稳性检验，以确保符合协整检验的前提条件——即所有序列须是同阶单整[56]。平稳性检验采用 ADF 检验，结果见表 3.6。

表 3.6　ADF 检验结果

变量	t 统计量	p 值	检验结果
lnT	-2.57	0.29	非平稳
lnI	-2.46	0.35	非平稳
$D(lnT)$	-5.44	0.00^{***}	平稳
$D(lnI)$	-6.25	0.00^{***}	平稳

注：***表示在 1%显著性水平下拒绝原假设。

结果表明，原序列lnT和lnI均为非平稳序列，在一阶差分下所有变量序列均为平稳序列，即为 1 阶单整，符合协整检验的前提条件。

然后进行协整检验，检验网络综合搜索指数与海南游客流量之间是否存在长期协整关系，Johansen 协整检验结果见表 3.7。

表 3.7　Johansen 协整检验结果

原假设	特征值	迹统计量	0.05 临界值	p 值
没有协整*	0.210344	28.08810	15.49471	0.0004
最多有一个协整关系	0.028413	3.055359	3.841465	0.0805

注：*表示在 0.05 显著性水平下拒绝原假设。

由表 3.7 可知，lnT和lnI是协整的。因此，说明网络综合搜索指数与海南游客流量之间存在长期的协整关系。下面进行格兰杰因果关系检验来验证网络综合搜索指数是否是海南游客流量的预测因子。格兰杰因果关系检验结果见表 3.8。

表 3.8　格兰杰因果关系检验结果

原假设	F 值	P 值	结论
lnI不是lnT的格兰杰原因	9.1268	0.0000	拒绝
lnT不是lnI的格兰杰原因	4.1798	0.0079	拒绝

结果表明，网络综合搜索指数与海南游客流量之间存在格兰杰因果关系，即表示网络综合搜索指数对海南游客流量具有预测能力。

3.5 预测模型构建及结果分析

本文分别用传统时序方法 ARIMA、ARIMAX，以及以 BP 神经网络为代表的机器学习方法来构建预测模型；其中本文提出的预测模型记为模型⑤——输入变量为原始序列和 R/S-TDC-EMD-KPCA 方法提取合成的网络综合搜索指数。

为验证加入网络搜索信息对预测性能的提高，选取输入变量为单一原始序列的预测模型①；为验证 R/S 与 TDC 结合的有效性，选取输入变量为原始序列结合 TDC-EMD-KPCA 方法提取合成的网络综合搜索指数的预测模型②；为验证对网络搜索信息数据进行降噪的必要性，选取输入变量为原始序列结合 R/S-TDC-KPCA 方法提取合成的网络综合搜索指数的预测模型③；为验证 KPCA 合成网络综合搜索指数的有效性，选取输入变量为原始序列结合 R/S-TDC-EMD-PCA 方法提取合成的网络综合搜索指数的预测模型④。

本研究中 ARIMA 和 ARIMAX 模型的最优阶数是通过赤池信息量准则（AIC）评判确定的；BP 神经网络模型的输入以及隐藏神经元的数目是通过试错测试来

确定的，以最小化样本预测误差。表 3.9 显示了五种模型的预测性能，并给出了 MAPE 与 NRMSE 的比较结果。

表 3.9　模型预测性能评价

输入信息	模型	样本外		样本内	
		MAPE(%)	NRMSE	MAPE(%)	NRMSE
①时间序列	ARIMA	10.62	14.87	6.01	8.00
	BP	10.44	14.66	5.35	6.60
②时间序列+TDC-EMD-KPCA 的百度指数	ARIMAX	10.32	14.42	4.44	5.44
	BP	8.90	13.42	3.72	5.40
③时间序列+R/S-TDC-KPCA 的百度指数	ARIMAX	8.89	12.72	4.45	5.47
	BP	7.83	11.04	3.58	4.88
④时间序列+R/S-TDC-EMD-PCA 百度指数	ARIMAX	7.70	12.06	3.21	4.10
	BP	7.66	9.73	2.42	3.77
⑤时间序列+R/S-TDC-EMD-KPCA 百度指数	ARIMAX	7.42	11.98	3.17	4.13
	BP	7.11	9.81	2.15	2.68

由表 3.9 可以得出如下分析结果：

（1）加入网络搜索信息的模型预测精度整体要好于单一历史数据。与没有网络搜索信息（模型①）的模型相比，加入网络搜索信息（模型②~⑤）的 ARIMAX 和 BP 模型的 MAPE 在样本内平均降低 2.1925%、2.3825%，在样本外平均降低 2.0375%、2.565%，NRMAE 在样本内平均降低 3.215、2.4175，在样本外平均降低 2.075、3.66。

（2）BP 神经网络构建的模型要好于传统时序模型。在样本内，五种情形的 BP 模型分别比 ARIMA 模型和 ARIMAX 模型的 MAPE 小 0.66%、0.72%、0.87%、0.79%和 1.02%，NRMSE 小 1.4、0.04、0.59、0.33 和 1.45，在样本外，相应的 MAPE 小 0.18%、1.42%、1.06%、0.04%和 0.31%，NRMSE 小 0.21、1、1.68、2.33 和 2.17。

（3）在从网络搜索信息的具体不同构建方法上来进行比较：

加入 R/S 方法的网络综合搜索指数的 ARIMAX 和 BP 预测模型（模型⑤）优于未加入 R/S 方法的预测模型（模型②），其中 MAPE 平均下降 2.085%、1.68%，NRMSE 平均下降 1.875、3.165，说明了 R/S 与 TDC 结合筛选关键词构建网络综合搜索指数能够进一步提高预测精度；

经过降噪处理的网络搜索信息数据的预测模型(模型⑤) 显著优于未经降噪

处理的模型（模型③），其中 MAPE 的平均降幅达 1.375%和 1.075%，NRMSE 的平均降幅达 1.04 和 1.715，证实了在预测前有必要对网络搜索信息数据进行降噪处理；

与基于 PCA 合成网络综合搜索指数（模型④）的 ARIMAX 和 BP 模型相比，基于 KPCA 合成网络综合搜索指数模型(模型⑤)的 MAPE 平均下降 0.16%、0.41%，NRMSE 平均下降 0.025、0.505，说明在网络综合搜索指数合成中考虑非线性信息对预测是有价值的。

综上所述，在这些预测模型中，结合历史数据和 R/S-TDC-EMD-KPCA 百度指数的 BP 组合预测模型在预测中表现出色，其 MAPE 值和 NRMSE 值都较低，显著优于其他模型。因此，利用 R/S-TDC-EMD-KPCA 方法提取网络搜索信息，并将其作为辅助信息而建立的 BP 模型具有更优的预测效果。

3.6 本章小结

本章是为了充分验证所提出的网络搜索信息合成方法的有效性，以海南省的月度游客流量作为研究对象，首先收集海南月度游客流量及相关关键词的网络搜索信息数据，利用上述提到的 R/S-TDC-EMD-KPCA 方法提取网络搜索信息合成网络综合搜索指数；然后检验网络综合搜索指数与海南游客流量之间的协整关系和格兰杰因果关系；最后将网络综合搜索指数和游客流量作为输入变量，分别利用 ARIMA、ARIMAX 和 BP 神经网络方法建立预测模型，对不同模型的预测效果进行评价。结果表明，融合了网络综合搜索指数的模型在预测精度方面均优于其他基准模型，这说明本文提出的 R/S-TDC-EMD-KPCA 方法能高质量的提取和合成网络搜索信息，进而可有效地应用于游客流量的辅助预测。

4 基于文本和网络搜索信息的游客流量预测

4.1 预测框架

根据旅游客流时间序列的周期性和高波动性特点，本文不同于单一时间序列的预测建模方法，有效利用互联网百度搜索信息和微博文本信息，提出一种基于文本和网络搜索信息融合的旅游预测新方法，具体建模过程主要由如下四个步骤完成，图 4.1 为预测的流程框架图。

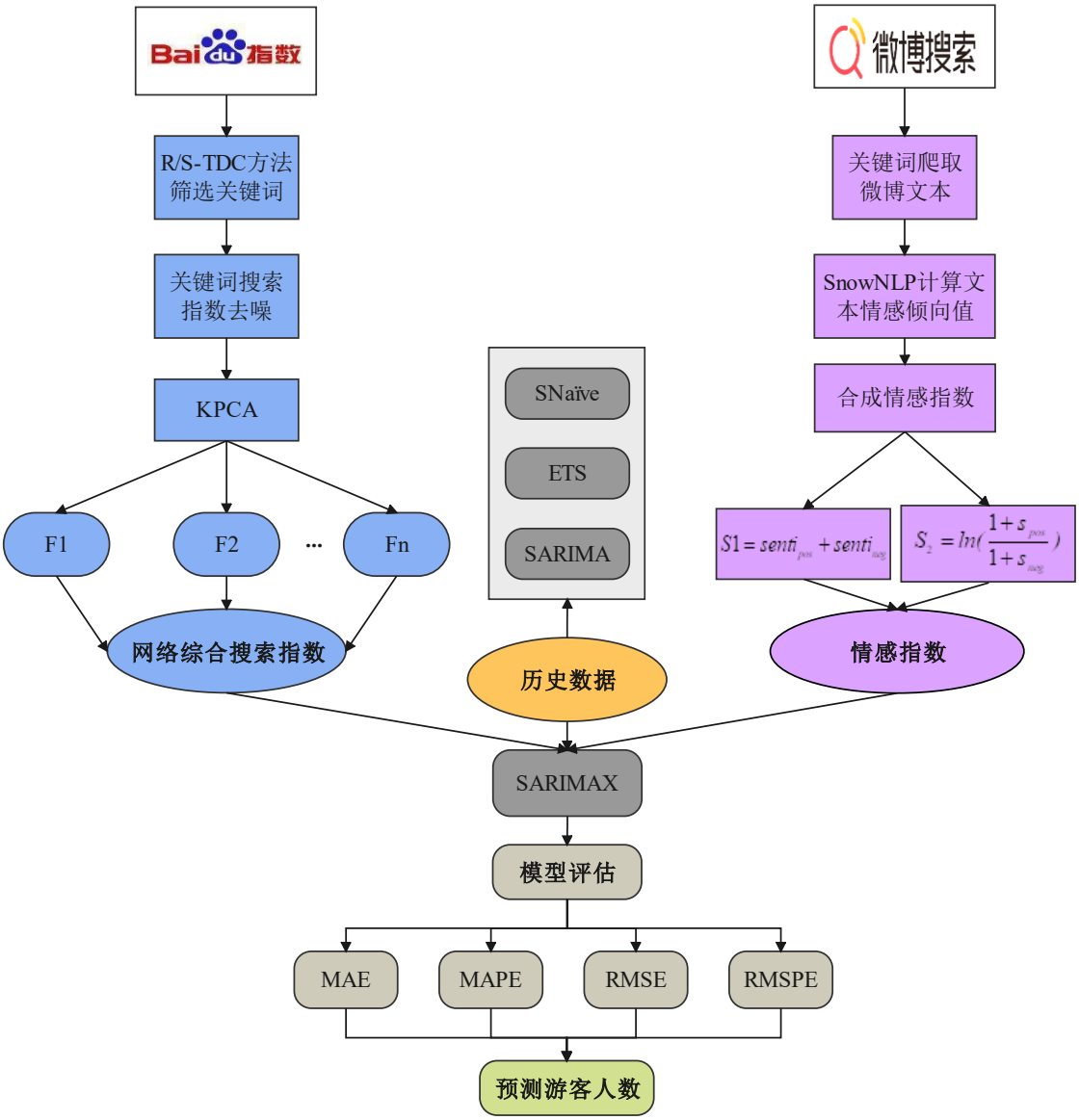


图 4.1　基于文本和网络搜索信息融合的旅游预测流程图

步骤一：构建网络综合搜索指数。初选关键词并获取网络搜索信息，并根据R/S-TDC 方法优选关键词，对优选后的关键词进行异常值处理，利用 KPCA 方法得到网络综合搜索指数。

步骤二：构建情感指数。以优选得到的关键词集作为搜索对象，从中国占主导地位的社交平台——新浪微博中获得文本数据，采用 SnowNLP 方法计算情感倾向值，采用基于正负情感简单相加和基于正负情感非对称的方法构建情感指数。

步骤三：建立预测模型。基于网络综合搜索指数、情感指数以及历史游客流量作为输入变量，构建 SARIMAX 模型预测短期的游客流量。

步骤四：评估模型。基于不同评估指标对提出的模型的预测性能进行评估。

4.2 数据采集

本研究选取的海南国内过夜游客人数的月度数据和网络搜索关键词数据与第 3 章数据相同；由于新浪微博作为社交平台，自上线以来覆盖用户人群逐年增加，截至 2021 年底，微博的月活跃用户数为 5.73 亿，成为国内使用最为广泛的社交类 APP，因此文本数据从新浪微博中获取。我们将 2011 年 1 月到 2017 年 12 月的 84 个数据作为训练样本，2018 年 1 月至 2019 年 12 月的 24 个数据作为测试样本。

4.3 网络综合搜索指数构建

本文利用相关关键词的百度搜索信息合成网络综合搜索指数的具体过程如下：其中，关键词初选和关键词优选同第 3 章方法相同，在此不再复述。

4.3.1 关键词搜索量异常值处理

由于网络搜索信息数据存在较大噪声，往往导致数据有异常值，在预测前不加剔除会对预测结果带来较大偏差。本文利用箱线图对关键词集C_2中的关键词数据进行异常值检测，结果如图 4.2 所示；剔除异常值后利用 K 近邻算法通过数据之间的相似性对缺失值进行填充，从而得到最终的关键词搜索量序列。

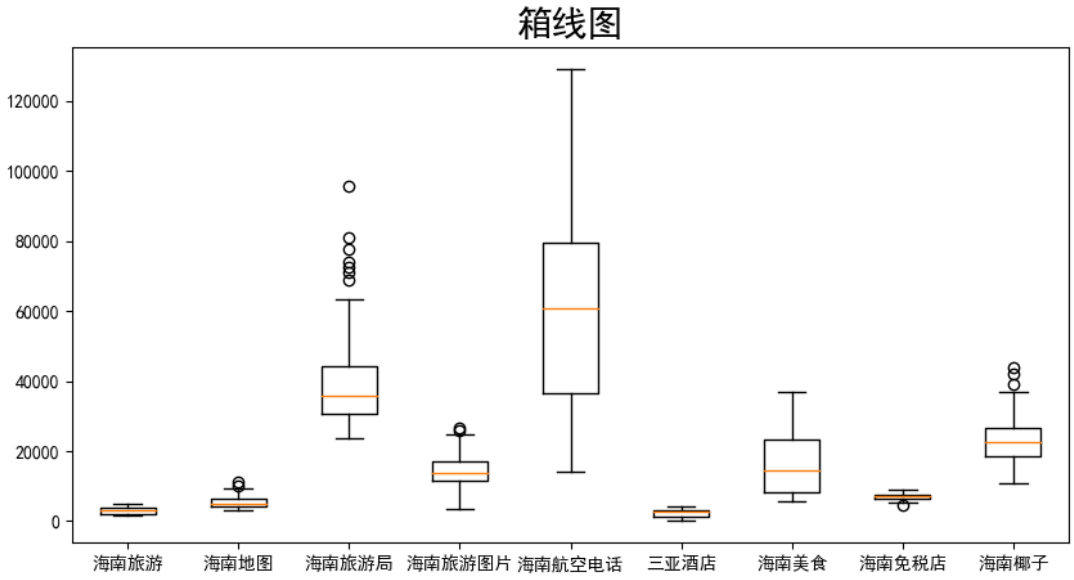


图 4.2　9 个关键词搜索量的箱线图

4.3.2 合成网络综合搜索指数

首先对关键词集 C_2 数据进行标准化处理；利用标准化数据通过径向基核函数计算核矩阵，对核矩阵进行中心化处理得到中心化核矩阵，并求解其特征值和特征向量；然后将特征值按照从大到小的顺序进行排序，计算每个核主成分的贡献率和累计贡献率，其计算结果见表 4.1。

表 4.1　核主成分结果

核主成分	特征值	贡献率%	累计贡献率%
1	0.4846	48.46	48.46
2	0.2707	27.07	75.52
3	0.0609	6.09	81.61
4	0.0504	5.04	86.65
5	0.0404	4.04	90.69
6	0.0326	3.26	93.95
7	0.0251	2.51	96.46
8	0.0186	1.86	98.32
9	0.0168	1.68	100.00

由表 4.1 可知，前 4 个核主成分 F_1,F_2,F_3 和 F_4 的累计贡献率已经超过了 85%，因此用特征值累计贡献率达到 85%的前 4 项核主成分代替原来的 9 个关键词数据；然后，在主成分分析中的加权合成方法，以每个核主成分所对应的特征值占所提取核主成分总的特征值之和的比例作为权重，最后合成网络综合搜索指数 F 如下：

$$F = 0.5591F_1 + 0.3124F_2 + 0.0702F_3 + 0.623F_4$$

图 4.3 是海南游客流量与网络综合搜索指数的序列图，可以看出两个序列谷峰和变化趋势具有一致性，说明游客流量与网络综合搜索指数之间存在相关性。

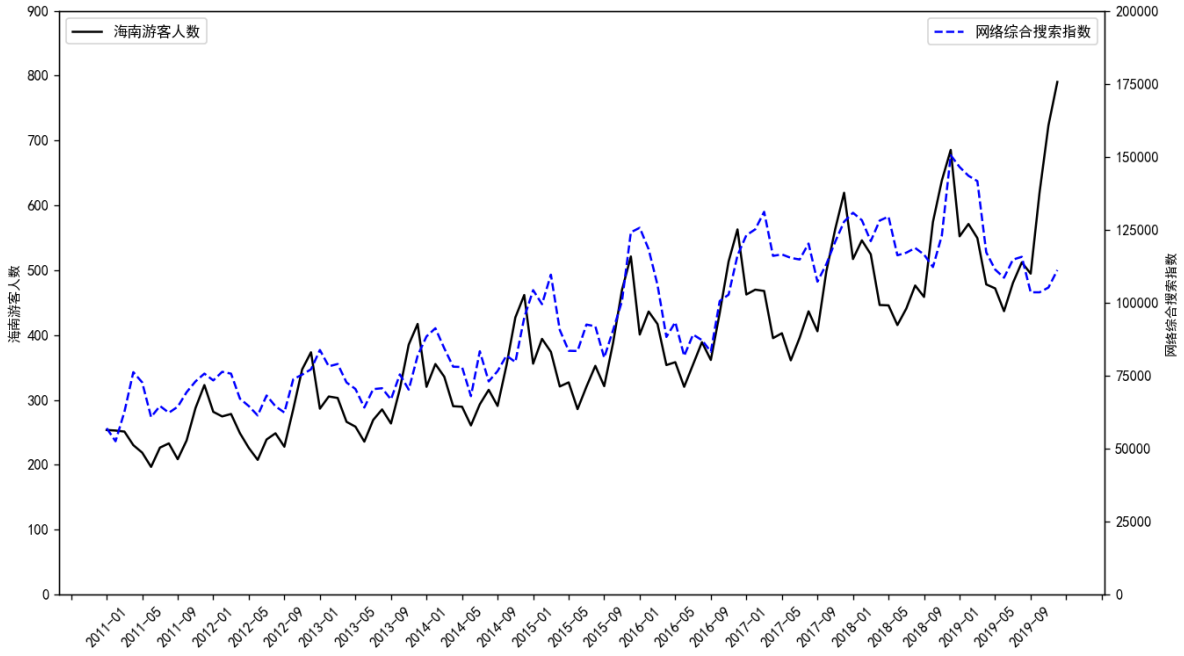


图 4.3　网络综合搜索指数 F 与游客人数 T 的关系

下面进行格兰杰因果关系检验来验证网络综合搜索指数是否是海南游客流量的预测因子。格兰杰因果关系检验结果见表 4.2。

表 4.2　格兰杰因果关系检验结果

原假设	F 值	p 值	结论
F 不是 T 的格兰杰原因	3.94872	0.0052	拒绝
T 不是 F 的格兰杰原因	9.66719	0.0000	拒绝

结果表明，网络综合搜索指数与海南游客流量之间存在格兰杰因果关系，即表示网络综合搜索指数对海南游客流量具有预测能力。

4.4 情感指数构建

4.4.1 数据采集及预处理

基于上述优选的关键词集C_2在新浪微博上进行关键词搜索，提取 2011 年 1 月至 2019 年 12 月期间的微博文本。而且不同在线用户可能发表重复的微博内容，一般来说这是没有研究意义的，因此，为了保留更有用的语料，对重复的文本数据进行剔除，以确保数据的有效性。对这些文本进行清理、删除重复后，最终保留了 34638 条微博文本。（下表 4.3 展示了部分微博文本数据）

在中文文本中，词语相互之间的界限往往不太明确，因此科学合理地分词显得尤为重要。本文采用 Python 中的中文分词第三方库 Jieba，利用中文词库，确定汉字之间的关联概率，汉字间关联概率大的组成词组，从而形成分词结果。

表 4.3 部分微博文本数据

time	content
2011/9/23	旅游地产黄金季之走遍海南，寻找海南旅游地产十大潜力区域论坛圆满结束。中国旅游地产服务集团副总经理孙天华进行了精彩总结：大家努力把盈滨半岛区域在短期内打造成海南真正的品牌区域，使得区位价值得到更大挖掘。
2011/9/24	热烘烘的海南岛旅游地图出炉了，兴奋的孩子有木有？祖国各地的兄弟姐妹们，来支持一下国际旅游岛建设，十一来海南玩吧！海南人民最热情啦！
2011/9/25	中原人爱旅游，会旅游，又地处内陆，海南旅游地产独特的滨海气质，得天独厚的自然资源，在河南乃至中西部区域的市场不可限量。
2011/10/2	//:以检查之名，行吃喝之实!要想真正为海南旅游做点事，不是住五星酒店，而是多想想普通民众，他们是旅游大军，然所遭遇的零负团'乱收费'强购物'行程缩水等等问题，不能病入膏荒却束手无策，甚至视而不见//:回复:是啊又走了！我们随旅游委检查全省旅游市场

4.4.2 情感倾向计算

利用 SnowNLP 对每条微博文本计算情感倾向，情感倾向值的区间在[0,1]之间，并判断每条微博文本的情感极性。在本文研究中，设定 0.4 以下的微博文本为负面情感，0.4 到 0.8 之间的微博文本标记为中立情感，0.8 分以上的微博文本标记为正面情感，同时把正面情感标记为 1，负面情感标记为-1，中立情感标记为 0，最后每条微博文本都被标记了不同的情感极性。

另外，由于本文用的海南历史游客流量是月度数据，而微博文本是日度数据，需要将日度微博文本数据整合为月度数据，表 4.4 是通过 SnowNLP 计算的日度微博文本数据的情感倾向值以及情感极性的部分举例，表 4.5 是整合为月度微博文本数据的情感极性数量的部分展示。

表 4.4 日度微博文本数据的情感倾向值

time	content	情感倾向值	情感极性
2018/3/9	海南旅游遭遇“回家难”，返程机票近 2 万一张！航企算不算“坐地起价”？|唔哩头条	0.094506523	-1
2018/3/11	我下午拍的视频，海南都海~~真的很漂亮，今天在这真出了几张片，这次海南没有白来，自己给自己点个赞。	0.9836997	1
2018/3/15	#雄仔日评#为什么躺在沙发上叫懒惰，躺在沙滩上叫渡假？这跟学什么专业有前景一毛钱关系都没有！#海南旅游与美食#	0.0045943	-1
2018/3/16	娘亲去海南旅游后的第四天，我突然明白了一件事。她是我的主心骨。	0.5827892	0
2018/3/19	这个春天，我最想到海南旅游。	0.894973862	1
2018/3/21	远离喧嚣！到海南这些美丽的小镇，来一场说走就走的风情之旅！	0.993498483	1

表 4.5　月度微博文本数据的情感极性数量

time	积极文本数量	消极文本数量	中性文本数量
2019/1	181	186	31
2019/2	110	71	29
2019/3	430	174	58
2019/4	138	58	25
2019/5	147	57	28
2019/6	106	23	14
2019/7	113	27	23
2019/8	71	14	12
2019/9	38	22	19
2019/10	100	39	30
2019/11	176	42	61
2019/12	153	46	40

4.4.3 生成情感指数

代表整体情感趋势的情感指数可以通过上节计算的正向情感、负向情感以及中立情感指数来衡量。参考 Zhang 等[43]和 Liu 等[57]的文献，本文通过两种计算方法生成两个情感指数，第一种情感指数用 S_1 表示，第二种情感指数用 S_2 表示。计算公式如下：

$$\begin{cases} S_1 = senti_{pos} + senti_{neg} \\ senti_{pos} = \sum_{i=1}^{d} m_{t,i} \\ senti_{neg} = \sum_{j=1}^{d} (-1) \times n_{t,j} \end{cases} \tag{4.1}$$

其中 $senti_{pos}$ 为第 t 月的正面情感倾向值，数值为正值，$m_{t,i}$ 为第 t 月第 i 天的正面情感微博词条数量，$senti_{neg}$ 为第 t 月的负面情感倾向值，数值为负值，$n_{t,j}$ 为第 t 月第 j 天的负面情感微博词条数量，d 是第 t 月对应的天数。

$$\begin{cases} S_2 = ln(\frac{1+s_{pos}}{1+s_{neg}}) \\ s_k = \sum_{i=1}^{d} n_{t,i} (k = pos, neg) \end{cases} \tag{4.2}$$

其中，s_k 为第 t 天的新闻标题中情感类型 k 的倾向值，$n_{t,i}$ 为第 t 月第 i 天情

感类型为 k 的微博词条数量，同理 d 表示情感类型 k 中第 t 月对应的天数。情感指数为正，表示社交平台对旅游市场的变化趋势呈正向预期，其大小代表了正向预期的强烈程度，情感指数越大则表示正向预期越强烈。反之，情感指数为负，则表示对旅游市场的变化趋势预期呈负向预期，情感指数越小代表负向预期越强烈。计算的两种情感指数分别与海南历史游客人数的趋势图如图 4.4 和图 4.5 所示。

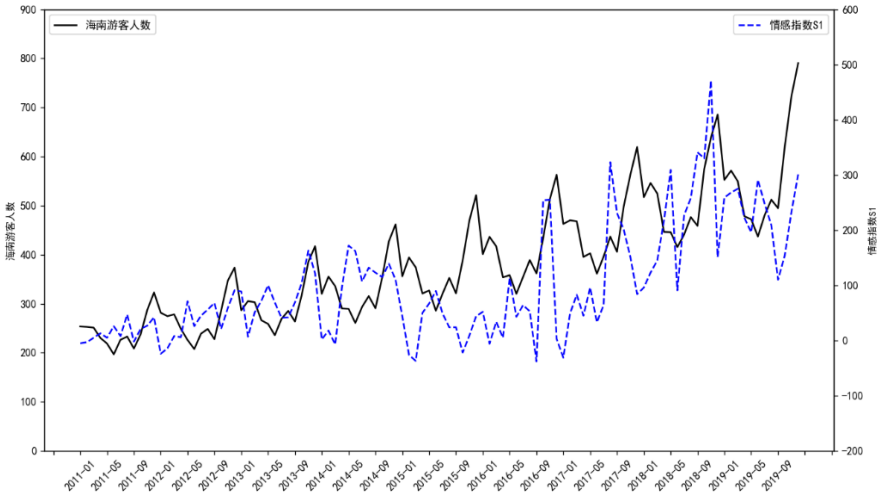


图 4.4　情感指数 S_1 与海南游客人数趋势对比图

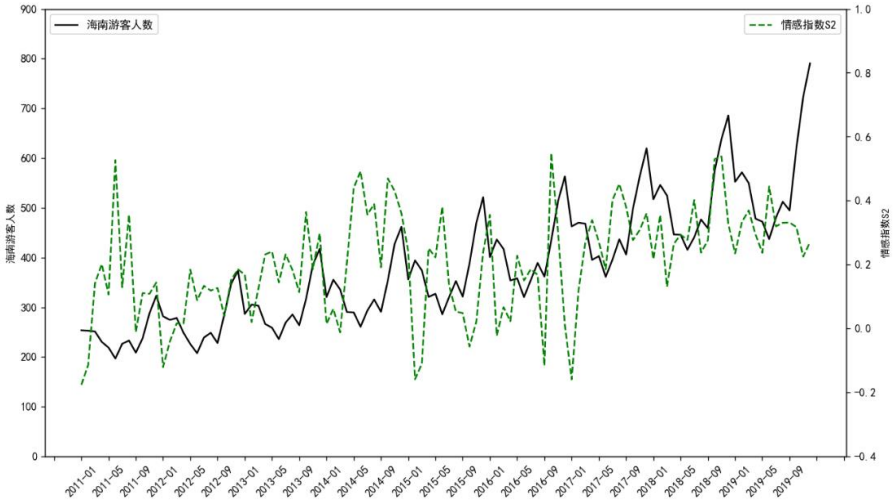


图 4.5　情感指数 S_2 与海南游客人数趋势对比图

由这两个趋势图可以看出，上述所构建的情感指数和旅游人数 T 的谷峰和变化趋势具有一致性，说明这两种方法构建的情感指数与海南游客流量之间存在相关性。后续将情感指数S_1和S_2共同作为输入变量对游客流量进行预测研究。

4.5 建立预测模型

常用的传统预测模型有 Naïve 模型，指数平滑模型（ETS）以及 ARIMA 模型，考虑到旅游数据具有季节性变动，本文采用季节性 naïve（SNaïve）模型和季节性 ARIMA（SARIMA）模型作为基准预测模型。

以下是本文所构建的模型：基于历史游客人数 T 构建 Snaïve 模型、ETS 模型和 SARIMA 模型作为传统基准模型；为了评估百度搜索信息的有效性，基于网络综合搜索指数和历史游客人数 T 构建 SARIMAX1 模型；为了评估情感指数的有效性，基于情感指数S_1和历史游客人数 T 构建 SARIMAX2 模型、基于情感指数S_2和历史游客人数 T 构建 SARIMAX3 模型，基于情感指数S_1、情感指数S_2和历史游客人数 T 构建 SARIMAX4 模型；为了评估融合文本和网络搜索信息后模型的有效性，基于网络综合搜索指数、情感指数S_1和历史游客人数 T 构建 SARIMAX5 模型、基于网络综合搜索指数、情感指数S_2和历史游客人数 T 构建 SARIMAX6 模型，以及本文提出的最佳预测模型——基于网络综合搜索指数、情感指数S_1、情感指数S_2和历史游客人数 T 构建 SARIMAX7 模型。

本研究中 SARIMA 和 SARIMAX 模型的最优阶数值*（p,d,q,P,D,Q）*以及周期 s 是通过 R 中 auto.arima()函数自动根据赤池信息量准则（AIC）最小原则评判确定的，并用 forcast()函数预测未来值。本文提出的 SARIMAX7 模型最终确定的最优阶数，即为$SARIMAX(1,0,2)(0,2,1)_{12}$模型。

从图 4.6 中的预测结果来看，游客人数显示出一定的波动性，并出现了峰值。基于网络综合搜索指数、情感指数S_1、情感指数S_2和历史游客人数构建的 SARIMAX7 模型可以很好地捕获其波动性，因此本文中所构建的预测模型比基准模型能够获得更好的预测结果。

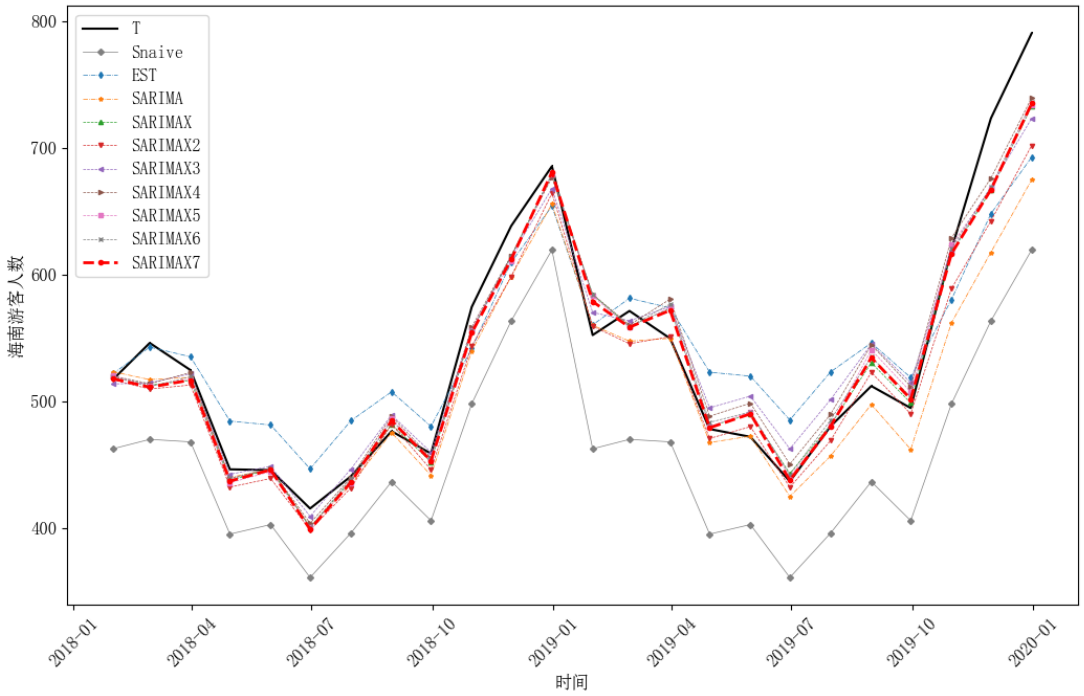


图 4.6　预测结果对比图

4.6 预测结果评价

表 4.6 显示了 10 种模型的预测性能评价结果，表 4.7 展示了本文所提出的最优预测模型 SARIMAX7 与其他 9 种模型的各种预测指标相比，各预测指标的相对改进值。

表 4.6　模型预测性能评价

	模型	MAE	MAPE(%)	RMSE	RMSPE(%)
	Snaive	78.83	14.41	85.19	14.91
历史	EST	33.72	6.30	39.69	7.14
	SARIMA	24.83	4.09	38.61	5.69
历史+百度	SARIMAX1	21.13	3.54	31.46	4.72
历史+S_1	SARIMAX2	20.29	3.40	30.09	4.49
历史+S_2	SARIMAX3	18.84	3.34	24.82	4.12
历史+S_1+S_2	SARIMAX4	17.16	3.05	22.09	3.72
历史+百度+S_1	SARIMAX5	16.46	2.87	22.60	3.65
历史+百度+S_2	SARIMAX6	15.67	2.72	21.86	3.52
历史+百度+S_1+S_2	**SARIMAX7**	**15.23**	**2.62**	**21.77**	**3.47**

通过对表 4.6 的分析，可以得到如下结果：

（1）从单一历史序列构建的预测模型来看，SARIMA 模型的预测效果最好，其 MAE、MAPE、RMSE 和 RMSPE 分别为 24.83、4.09%、38.61 和 5.69%；

（2）引入网络搜索综合指数的 SARIMA1 模型要比单一历史信息的 SARIMA 模型具有更高的预测精度，其 MAE、MAPE、RMSE、RMSPE 均有下降，其中 MAPE 下降到了 3.54%，RMSE 下降到了 31.46。这可能源于互联网的普及，网络综合搜索指数确实可以反映游客的对旅游目的地的关注度以及旅游意愿，因此可以作为旅游需求预测的有效预测因子。

（3）引入微博情感指数的模型（SARIMA2、SARIMA3 和 SARIMA4 模型）与单一 SARIMA 模型相比，在水平精度上有显著提升，尤其是将两种情感指数均作为预测因子时，SARIMA4 模型的预测精度进一步得到提升，其 MAE、MAPE、RMSE 和 RMSPE 分别下降为 17.16、3.05%、22.09 和 3.72%，这说明两种情感指数的综合信息比单一情感指数对游客人数的预测作用更显著，且比引入网络综合搜索指数的模型相比，预测精度的提升作用更强，反映出两种情感指数的结合能够更好的体现出游客的感知行为，对提升预测精度具有更大的贡献。

（4）进一步将网络综合搜索指数和情感指数融合后的 SARIMAX 模型总体比仅引入百度指数或情感指数的模型预测效果好。将网络综合搜索指数和历史数据分别与情感指数 S_1、情感指数 S_2 结合，其中与情感指数 S_2 相结合的预测模型表现较好，MAE、MAPE、RMSE、RMSPE 分别下降到 15.67、2.72%、21.86 和 3.52%。

（5）最后，本研究提出的基于文本和网络搜索信息融合的预测模型表现出最好的预测效果。实证研究发现，将历史数据、网络综合搜索指数、情感指数 S_1 和感指数 S_2 均作为输入变量对游客流量进行预测，其水平预测误差均低于其他基准模型，MAE 下降到 15.23，MAPE 下降到 2.62%，RMSE 下降到 21.77，RMSPE 下降到 3.47%。这是因为基于简单加总的情感指数 S_1 虽然能够反映微博文本的整体情感趋势，但其数值容易受微博文本数量的影响，而基于公式计算的情感指数仅与正向倾向值和负向倾向值有关，不受总体微博文本数量的影响。因此，添加两种不同方法计算的情感指数可以反映游客情感信息的不同方面，将网络综合搜索指数与情感指数相结合起来预测游客人数，可以使预测精度更高。

表 4.7　本文最优模型相比基准模型的相对改进值 RI

最优模型	基准模型	RI(%)
SARIMAX7	Snaive	81.79
	EST	58.36
	SARIMA	35.90
	SARIMAX1	25.96
	SARIMAX2	22.75
	SARIMAX3	21.40
	SARIMAX4	13.83
	SARIMAX5	8.41
	SARIMAX6	3.63

通过相对改进值 RI 表 4.7 可以进一步直观看出本文提出的 SARIMAX7 模型相比其他基准模型在预测指标 MAPE 上的提升效果。首先，SARIMAX7 与单一传统时序模型（Snaive、EST 以及 SARIMA）相比均有显著提升，其 MAPE 的相对改进值分别 81.79%、58.36%和 35.90%；其次，SARIMAX7 模型与仅加网络搜索综合指数的 SARIMAX1 模型和仅情感指数的 SARIMAX2、SARIMAX3 和 SARIMAX4 模型相比均有所提升，其 MAPE 的相对改进值分别为 25.96%、22.75%、21.40%和 13.83%。最后，SARIMAX7 模型相较于网络综合搜索指数与单一情感指数 S_1、单一情感指数 S_2 结合的预测效果均有所提升，其 MAPE 的相对改进值范围在 10%之内，这也进一步说明基于文本和网络搜索信息融合的游客流量预测模型表现出最好的预测效果。

4.7 本章小结

本章提出一种基于文本和网络搜索信息的旅游预测新方法，以中国海南省的月度客流量为预测对象进行实证分析验证其有效性。首先基于重标极差法（R/S）和时差相关法（TDC）选择具有预测能力的百度搜索关键词，并对所选关键词搜索量进行异常值处理，然后利用核主成分分析(KPCA)方法合成网络综合搜索指数。其次，从中国主流社交平台新浪微博中提取与最优关键词有关的文本数据信息，并采用两种不同的方法来构建情感指数。最后，基于网络综合搜索指数和情感指数构建 SARIMAX 预测模型。结果表明，将网络综合搜索指数与情感指数同时作为预测因子时，可以有效提高预测精度。因此，本文所提出的百度搜索信息与微博文本信息的提取方法是有效的，这为旅游需求的精准预测提供了新的途径。

5 结论与展望

5.1 结论

本文主要从两方面进行对旅游预测的展开研究。一方面，提出了一种从网络搜索关键词筛选，搜索信息去噪，以及网络综合搜索指数合成的 R/S-TDC-EMD-KPCA 方法；另一方面，以百度指数为数据来源生成网络综合搜索指数，新浪微博为数据来源生成情感指数，基于文本和网络搜索信息构建 SARIMAX 预测模型。主要研究结论如下：

（1）以海南省游客流量为例，通过建立原始游客流量数据结合网络搜索信息数据的预测模型，检验了 R/S-TDC-EMD-KPCA 方法的有效性。第一，预测游客流量时，在模型中加入网络搜索信息数据作为辅助输入信息能够有效地提高预测性能。R/S-TDC 方法筛选的关键词具有较好的预测能力，可以有效提升预测精度。第二，因为网络搜索信息数据中含有噪声，所以在进行预测前，需要对它进行 EMD 去噪，以提高模型的预测能力。第三，由于网络搜索信息数据中不仅包含线性信息，还包含非线性信息，因此为了从网络搜索信息数据中有效提取非线性信息，使用 KPCA 方法合成网络综合搜索指数；通过实证也验证了基于 KPCA 的网络综合搜索指数的有效性。第四，由于游客流量数据具有非线性和复杂性特征，实证表明 BP 模型的精度高于 ARIMA 模型，说明机器学习模型可以比传统的线性模型实现更高的预测精度。

（2）以海南省的月度游客流量为预测对象，采用本文提出的基于文本和网络搜索信息的游客流量预测模型，得到如下结论：首先，通过对 SARIMAX 模型与 SARIMA 模型的预测精度比较发现，无论添加网络综合搜索指数还是情感指数在预测海南省游客人数时都对提高预测准确性起到了积极的作用。这一结果表明，网络搜索以及微博文本的内容都会反映游客对旅游目的地的关注度和正负情感倾向，并在决定旅游需求方面发挥作用。其次，在 SARIMAX 模型中加入网络综合搜索指数和情感指数比单一数据源作为辅助输入信息能够更有效地提高预测性能。值得说明的是，本文提出的研究方法具有新颖性，不同于现有方法，采用了两种不同的方法编制情感指数，分别是基于正负情感简单相加的情感指数和基于正负情感非对称的情感指数，研究发现不同的情感指数编制方法会对预测结

果有一定的影响，基于不同人类心理行为构建正负情感非对称情形下的情感指数相比于正负情感的简单加总更能反映游客情感倾向，可以获得更好的预测效果。为旅游预测提供了更好的选择，丰富了数据驱动旅游预测方法的研究。

5.2 展望

尽管本文提出的模型获得了不错的预测表现，但仍存在一定的局限性。

第一，研究方法只应用在海南省旅游市场上，仍然需要利用更多目的地的旅游数据来检验本文提出的基于文本和网络搜索信息融合预测模型的稳健性。

第二，本文只涉及了非结构化数据中的其中一种类型——文本数据，对于图片、语音以及视频等数据类型未考虑到预测模型中。这些非结构数据蕴含着丰富的有效信息，可为了解旅游行为、旅游管理和旅游市场提供新的视角。例如，大多数景区都配备了监控系统，可以生成大量的视频数据，或是旅游博主分享的旅游视频，再或是网络社交平台上游客的互动等都是能够反映游客行为的数据信息，直接影响了游客的旅游决策和行为。在未来的研究中，可以将这些影响因素作为辅助预测变量对旅游目的地进行深入研究，从而有助于为相应的旅游管理服务。

第三，未考虑其他社会环境因素对旅游预测的影响。新冠肺炎对旅游业造成极大影响，尤其是在爆发初期，2020 年旅游业发展出现了剧烈的下滑，并且到 2022 年的旅游数据均因疫情防控的影响，其波动性较大。在研究时间和工作能力的限制下，未对 2020 年至 2022 年新冠疫情期间的旅游市场进行深入研究。而能够剔除干扰事件在将来预测影响的干预分析模型适用于这种情况[58]。在未来，有考虑通过干预分析模型结合本文提出的融合文本和网络搜索信息的预测模型对游客流量进行预测，对本文缺陷之处予以弥补。虽然在研究框架上没有考虑疫情影响，但是对疫情放开之后的旅游预测以及未来的旅游规划提供了重要的参考信息。

参考文献

[1] 康俊锋,郭星宇,方雷.基于百度指数时空分布的旅游趋势预测研究——以上海市为例[J].西南师范大学学报(自然科学版), 2020,45(10):72-81.

[2] 任婕.基于向量自回归模型的旅游热门景点预测方法研究[J].现代电子技术, 2020,43(03):158-161.

[3] 姜国华.上海入境旅游需求建模分析与预测——以亚洲市场为例[J].旅游研究, 2016,8(05):68-74+85.

[4] Law R. Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting[J]. Tourism Management, 2000,21(4):331-340.

[5] 贾鹏,刘瑞菊,孙瑞萍,杨忠振.基于 BP 神经网络的邮轮旅游需求预测[J].科研管理, 2013,34(06):77-83.

[6] Tendai M, Chikobvu D . Modelling international tourist arrivals and volatility to the Victoria Falls Rainforest, Zimbabwe: Application of the GARCH family of models[J]. African Journal of Hospitality Tourism and Leisure, 2017,6(4):1-16.

[7] Abu N, Syahidah W N, Afif M M, et al. SARIMA and Exponential Smoothing model for forecasting ecotourism demand: A case study in National Park Kuala Tahan, Pahang[C]//Journal of Physics: Conference Series. IOP Publishing, 2021,1988(1):012118.

[8] Arunraj N S , Ahrens D , Fernandes M . Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry[J]. International Journal of Operations Research and Information Systems, 2016,7(2):1-21.

[9] Park E, Park J, Hu M. Tourism demand forecasting with online news data mining[J]. Annals of Tourism Research, 2021,90:103273.

[10] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data [J]. Nature, 2008,457(7232):1012-1014.

[11] 张瑞,刘立新,唐晓彬,张斌儒.大数据背景下基于网络搜索数据商品零售价格指数预测研究[J].统计与信息论坛, 2020,35(11):49-56.

[12] Clark M, Wilkins E J, Dagan D T, et al. Bringing forecasting into the future: Using Google to predict visitation in U.S. national parks[J]. Journal of Environmental Management, 2019,243:88-94.

[13] 黄先开,张丽峰,丁于思.百度指数与旅游景区游客量的关系及预测研究——以北京故宫为例[J].旅游学刊, 2013,28(11):93-100.

[14] Xie G, Li X, Qian Y T, Wang S Y, et al. Forecasting tourism demand with KPCA-based web search indexes[J]. Tourism Economics, 2021,27(4):721-743.

[15] Li X, Law R, Xie G, Wang S Y. Review of tourism forecasting research with internet data[J]. Tourism Management, 2021,83(3):104245.

[16] 陆利军.基于网络搜索指数和 EMD-ARIMA-BP 组合模型的游客流量预测——以张家界为例[J].吉首大学学报(社会科学版), 2019,40(01):138-150.

[17] 王兰梅,陈崇成,叶晓燕,潘淼鑫.网络搜索数据和 GWO-SVR 模型的旅游短期客流量预测[J].福州大学学报(自然科学版), 2019,47(05):598-603.

[18] 任乐,崔东佳.基于网络搜索数据的国内旅游客流量预测研究——以北京市国内旅游客流量为例[J].经济问题探索, 2014(04):67-73.

[19] 陆利军,廖小靖.选择域视角下的旅游搜索指数构建及其预测效果分析——以四姑娘山为例[J].中南林业科技大学学报(社会科学版), 2021,15(02):100-110.

[20] Peng G,Liu Y,Wang J,et al.Analysis of the prediction capability of web search data based on the HE-TDC method prediction of the volume of daily tourism visitors[J]. Journal of Systems Science and Systems Engineering, 2017,26(2):1-20.

[21] Yao L, Ma R, Wang H. Baidu index-based forecast of daily tourist arrivals through rescaled range analysis, support vector regression, and autoregressive integrated moving average[J]. Alexandria Engineering Journal, 2021,60(1):365-372.

[22] 李晓炫,吕本富,曾鹏志,刘金烜.基于网络搜索和 CLSI-EMD-BP 的旅游客流量预测研究[J].系统工程理论与实践,2017,37(01):106-118.

[23] 梁小珍,张晴,杨明歌.面向网络搜索数据的航空客运需求两阶段分解集成预测模型[J].管理评论,2021,33(05):236-245.

[24] 魏瑾瑞,崔浩萌.基于网络搜索数据的区域旅游指数及其微观动态:以西安为例[J].系统科学与数学, 2018,38(02):177-194.

[25] 胡倩倩.基于百度指数的海南旅游量预测[D].长春:吉林大学,2019.

[26] 张玲玲,张笑,崔怡雯.基于聚类方法的百度搜索指数关键词优化及客流量预测研究[J].管理评论, 2018,30(08):126-137.

[27] 孙毅,戴维,董纪昌,吕本富.基于主成分分析的网络搜索数据合成方法研究[J].数学的实践与认识, 2014,44(21):121-128.

[28] Li S, Chen T, Wang L, et al. Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index[J]. Tourism Management, 2018,68:116-126.

[29] 洪巍,李敏.文本情感分析方法研究综述[J].计算机工程与科学,2019,41(04):750-757.

[30] Liu B, Zhang L. A survey of opinion mining and sentiment analysis[M]//Mining text data. Springer, Boston, MA, 2012:415-463.

[31] 赵妍妍,秦兵,石秋慧,刘挺.大规模情感词典的构建及其在情感分类中的应用[J].中文信息学报, 2017,31(02):187-193.

[32] Tao Y, Zhang F, Shi C, et al. Social media data-based sentiment analysis of tourists' air quality perceptions[J]. Sustainability, 2019,11(18):5070.

[33] 陈新元,谢晟祎,陈庆强,张力.结合图片语义规则和机器学习的情感分类方法[J].计算机应用与软件,2021,38(07):173-181.

[34] Wawre S V, Deshmukh S N. Sentiment Classification Using Machine Learning Techniques[J].International Journal of Science and Research, 2016,5(4):819-821.

[35] 尚永敏,赵榆琴.基于机器学习的在线评论情感分析与实现[J].大理大学学报, 2021,6(12):80-86.

[36] 刘苗,李蔚,朱述政,喻燕君,刘扬,纪宏.基于互联网文本情感分析的消费情感指数构建[J].统计与信息论坛,2018,33(08):31-38.

[37] Colladon A F, Guardabascio B, Innarella R. Using social network and semantic analysis to analyze online travel forums and forecast tourism demand[J]. Decision Support Systems, 2019,123:113075.

[38] 王建成,徐扬,刘启元,吴良庆,李寿山.基于神经主题模型的对话情感分析[J].中文信息学报, 2020,34(01):106-112.

[39] Brochado A. Google search based sentiment indexes[J]. IIMB Management Review, 2020,32(3):325-335.

[40] Fan Z P, Che Y J, Chen Z Y. Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis[J]. Journal of business research, 2017,74:90-100.

[41] Checkley M S, Higón D A, Alles H. The hasty wisdom of the mob: How market sentiment predicts stock market behavior[J]. Expert Systems with applications, 2017,77:256-263.

[42] Liang C, Tang L, Li Y, et al. Which sentiment index is more informative to forecast stock market volatility? Evidence from China[J]. International Review of Financial Analysis, 2020,71:101552.

[43] Zhang C, Liu H, Chen Z, et al. Tourism Forecast Based on Web Search Data and Sentiment Analysis of Social Network[C]//The 2nd International Conference on Computing and Data Science. 2021:1-6.

[44] 陈晓红,彭宛露,田美玉.基于投资者情绪的股票价格及成交量预测研究[J].系统科学与数学, 2016,36(12):2294-2306.

[45] Mao H, Counts S, Bollen J. Predicting financial markets: Comparing survey, news, twitter and search engine data[J]. arXiv preprint arXiv: 1112.1051,2011.

[46] 王晓丹,尚维,汪寿阳.互联网新闻媒体报道对我国股市的影响分析[J].系统工程理论与实践, 2019,39(12):3038-3047.

[47] Yang Y, J Guo, Sun S. Forecasting crude oil price with a new hybrid approach and multi-source data[J]. Engineering Applications of Artificial Intelligence, 2021,101:104217.

[48] Pan B, Yang Y. Forecasting Destination Weekly Hotel Occupancy with Big Data[J]. Journal of Travel Research, 2016,56(7):957-970.

[49] Hurst H E. Long-term storage capacity of reservoirs[J]. Transactions of the American society of civil engineers, 1951,116(1):770-799.

[50] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings Mathematical Physical & Engineering Sciences, 1998,454(1971):903-995.

[51] 任武军,李新.基于互联网大数据的旅游需求分析——以北京怀柔为例[J].系统工程理论与实践, 2018,38(02):437-443.

[52] R.G.Brown.Statistical forecasting for inventory controlM].McGraw/Hill, 1959.

[53] Holt C C. Forecasting seasonals and trends by exponentially weighted moving averages[J]. International journal of forecasting, 2004,20(1):5-10.

[54] Winters P R. Forecasting sales by exponentially weighted moving averages[J]. Management science, 1960,6(3):324-342.

[55] Liu Y, Chen Y, Wu S, et al. Composite leading search index: a preprocessing method of internet search data for stock trends prediction[J]. Annals of Operations Research, 2015,234:77-94.

[56] 陈海燕.共同因子结构下非平稳面板数据检验的一致性研究[J].数理统计与管理, 2019,38(03):460-472.

[57] Liu Q, Lee W S, Huang M, et al. Synergy between stock prices and investor sentiment in social media[J]. Borsa Istanbul review, 2022,23(1):76-92.

[58] 刘晓孟,马晓燕,周爱民.新冠疫情对全球股指的冲击影响研究[J].金融论坛,2021,26(10):58-69.

攻读硕士学位期间承担的科研任务及主要成果

一、研究成果

曹静如,孙景云.基于 R/S-TDC-EMD-KPCA 方法的网络搜索信息提取及游客流量预测——以海南省为例[J].哈尔滨师范大学自然科学学报,2022,38(05):46-57.

致谢

行文至此，回顾学生期间的时光，总以为来日方长，却不知白驹过隙。从懵懵懂懂的小孩儿独自从河南到甘肃开启自己的大学生活，本科和研究生的时光占据我人生七年时间在兰州这个城市度过。对于这个城市，是我除家乡之外最熟悉的地方。很多人对甘肃很陌生，当真正了解这座城市后，确是发现它远比外界想象的要精彩很多。很高兴在这段经历中结识很多老师和朋友。

自己的研究生学术生涯始于 2020 年的疫情，终于 2023 年的疫情放开。这三年我们一道经历了很多，国家保护了我们三年，何其有幸，生于华夏，感谢强大的祖国，背负我们行进了三年。三年的疫情即将成为历史，愿未来的祖国海清河晏，山河无恙！

有师如斯，庆幸之至。感谢我的导师。从研一到研三，每周的讨论班从不缺席，从最开始的推荐书籍帮我们打实基础，慢慢教我们查阅英文文献，再后来在遇到问题时的耐心教导。从选题到开题再到定稿，非常感谢导师字句斟酌地帮我审阅论文，耐心批注和指导，给予我很大帮助，让我顺利完成毕业论文。在未来的日子里，愿老师事业顺利，教泽绵延。

父母爱子，为之深远。感谢我的父母还有姥姥姥爷。从出生到现在给予我物质以及精神的支持，成为我在成长道路上坚实无比的后盾，成为我人生启蒙的第一位老师。自此之后，唯有继续努力才能成为他们的骄傲，成为他们的依靠。愿我的家人们平安喜乐，身体健康。

愿岁并谢，与友长兮。感谢我的挚友们。蔓云、鸣宇、荟荟我们“313 业务群”的组合，晨曦、林芸我们“学术垃圾互助群”的结合，很高兴在研究生期间结识你们，一路走来幸得有你们相互鼓励和相互陪伴，给予欢笑，在枯燥无味的研究生生活中添加一些趣味。我会记着曾经上课匆忙而奔跑的我们，记着一起吃饭一起学习的时光，记着承载我们生活点滴的照片和视频，记得疫情期间各种吐槽的我们，记着在路上因搞笑话题而捧腹大笑的我们，记得每次熬夜搞论文的我们。愿我们在以后的日子里每一个都成为更棒的自己。希望在未来，我们可以有时间聚，聊聊那时候的我们。

最后，这里的谢意就留给自己吧。很感谢那个走得很慢但一直前行的自己，在求学路上有过迷茫又不断求索的自己。回望过去，原来自己也走了很长的路。我的学术之路也许在此结束，但自己的生活也在此进入新篇章。

段家滩 496 号的故事并不是结束，只是开始另一段新的旅程。