Operations Research and Fuzziology 运筹与模糊学, 2022, 12(3), 980-990
Published Online August 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.123104

Hans汉斯

游客目的地印象分析

成　盟

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2022年7月13日；录用日期：2022年8月9日；发布日期：2022年8月15日

摘　要

旅游目的地作为旅游活动开展的载体，近年来面临日趋增大的竞争压力。当前，中国各地区旅游品牌声誉度呈现发展不均衡的态势，怎样提升景区及酒店等旅游目的地美誉度，吸引优质游客、扩大品牌影响、提高竞争能力，成为各地区文旅主管部门和旅游相关企业关注的重点问题。游客满意度与目的地美誉度紧密相关，游客对旅游目的地的满意度越高，目的地美誉度就越大。本文通过分析景区及酒店等旅游目的地的游客互联网评价，在TF-IDF模型基础上，提出综合考虑词频与时间跨度的TF-ITH词汇热度计算模型，能够准确反映随时间变化的不同景区的游客评论热门词汇；引入预训练的Bert模型提取网评文本的观点，采用多元线性回归以MSE为评价指标来预测景区评分，为文本信息更加可观化提供一种新的方法；提出一种基于有效性的网络评论文本排序与筛选模型，能准确地剔除旅游目的地游客无效评论，从而为提升各地区旅游目的地的差异化竞争能力提供借鉴，进一步探索旅游目的地声誉塑造与维护的实现路径。

关键词

TF-ITH模型，Bert模型，评论观点提取，细粒度情感分析，评论有效性，差异化系数

Tourist Destination Impression Analysis

Meng Cheng

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Jul. 13th, 2022; accepted: Aug. 9th, 2022; published: Aug. 15th, 2022

Abstract

As a carrier of tourism activities, tourist destinations have faced increasing competitive pressure in recent years. At present, the reputation of tourism brands in various regions of China is developing unevenly. How to improve the reputation of tourist destinations such as scenic spots and hotels, attract high-quality tourists, expand brand influence and improve competitiveness has become a key issue of concern for cultural and tourism authorities and tourism-related enterprises

文章引用: 成盟. 游客目的地印象分析[J]. 运筹与模糊学, 2022, 12(3): 980-990.
DOI: 10.12677/orf.2022.123104

in various regions. Tourist satisfaction is closely related to the reputation of the destination. The higher the tourist satisfaction with the destination, the greater the reputation of the destination. By analyzing the tourist internet evaluation of scenic spots, hotels and other tourist destinations, this paper proposes a TF-ITH vocabulary heat calculation model based on the TF-IDF model, which comprehensively considers the word frequency and time span, and can accurately reflect the hot words of tourist comments in different scenic spots over time; pre-trained Bert model is introduced to extract the viewpoint of online evaluation text, and multiple linear regression is used to predict the score of scenic spots with MSE as the evaluation index, which provides a new method for more observable text information. This paper proposes a text sorting and screening model of online comments based on effectiveness, which can accurately eliminate invalid comments from tourists in tourist destinations, so as to provide a reference for improving the differentiated competitiveness of tourist destinations in various regions, and further explore the realization path of reputation shaping and maintenance of tourist destinations.

Keywords

TF-ITH Model, Bert Model, Comment Extraction, Fine-Grained Sentiment Analysis, Comment Effectiveness, Differentiation Coefficient

1. 引言

提升景区及酒店等旅游目的地美誉度是各地文旅主管部门和旅游相关企业非常重视和关注的工作，涉及到如何稳定客源、取得竞争优势、吸引游客到访消费等重要事项。游客满意度与目的地美誉度紧密相关，游客满意度越高，目的地美誉度就越大。

当前，中国各省份旅游品牌声誉呈现发展不均衡的态势，东南沿海经济发达省份品牌声誉指数排名靠前，西部及西北地区品牌声誉指数排名相对靠后。目的地整体品牌形象、游客满意度的高低直接关乎各地区旅游品牌的美誉度，各旅游城市相关部门应大力支持旅游业发展，积极探索旅游新模式，创新旅游文化宣传，进而提升城市旅游品牌美誉度，从而推动旅游业与经济社会各领域深度融合[1]。

近年来，随着网络技术的高速发展，在线旅游订票平台成为了游客们获取信息、发表观点、互相交流的新途径，游客们通过在线评论的方式，分享旅途体验，产生了大量真实有效的文本信息。吴宝清、吴晋峰、吴玉娟[2]等采用内容分析法和对应分析法，基于网络论坛的文本数据，研究了距离对西安旅游形象的影响；庄小丽、程仕菊等[3]爬取微博评论为研究样本，采用文本分析法、社会网络分析法，探索了游客对峨眉山风景区的旅游形象感知；李凤佼[4]运用 TF-IDF 算法和 LDA 主题提取模型，以百度旅游、携程网马蜂窝等多家在线旅游平台的网络点评数据为样本，探究了哈尔滨市冰雪旅游形象感知。

综上，以在线评论等文本信息为样本数据，运用文本挖掘技术来研究旅游形象感知，为旅游管理领域提供了一种新的视角。本文通过分析景区及酒店等旅游目的地的游客互联网评价，提出一种改进的 TF-ITH 词汇热度计算模型，能够准确反映随时间变化的不同景区的游客评论热门词汇；引入预训练的 Bert 模型提取网评文本的观点，采用多元线性回归来预测景区评分；提出一种基于有效性的网络评论文本排序与筛选模型，能准确地剔除旅游目的地游客的无效评论。为提高游客满意度，最终提升目的地美誉度，提供一种新的参考方法。

2. 数据预处理

对数据所作预处理如图 1，数据预处理顺序：删除完全重复网评文本(所给数据的全部指标均重复)→英文网评文本译为中文→繁体网评文本转为简体→删除无中文网评文本(经过翻译与繁化简处理后仍无中文信息的文本，一般为全符号文本)→网评文本错字纠正。

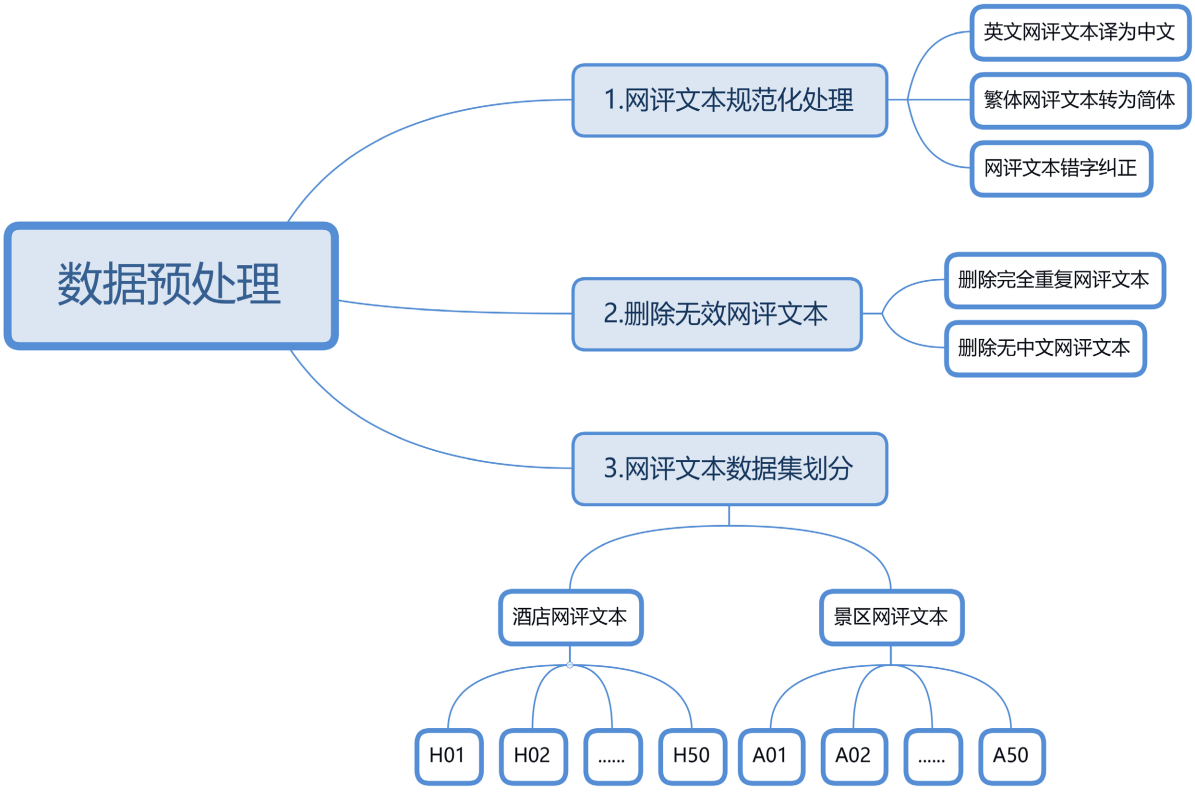

Figure 1. Flow chart of data preprocessing
图 1. 数据预处理流程图

3. 景区及酒店印象分析

3.1. 网评文本数据二次处理

对经过数据预处理的景区及酒店网评文本进行二次数据处理，包括数据清洗、分词及去重，从而保证通过建模求解的不同景区及酒店游客评论的热门词汇真实可靠合理。

首先，本文将通过数据预处理的景区及酒店网评文本中所包含的标点符号等无价值信息删除，使网评文本只保留相关文字信息，保证热门词汇的合理性。其次，采用 jieba 分词的精确模式进行分词，把网评文本精确切分，不存在冗余词汇。最后，对同一条网评文本中出现的相同词汇只保留一次，保证词汇在每条网评文本中词频相同。并通过加载 Hanlp 自然语言处理类库中的预训练词性标注模型对网评文本进行词性标注，去除掉每条网评文本中标记为语气词、动词、时间词等词。

3.2. TF-ITH 词汇热度计算模型

由于题中给出的景区及酒店网评文本存在时间跨度，同一评论词汇在不同时间热度存在变化可能，本文在结合 TF-IDF 模型[5]和 Reddit 热点排行算法[6]基础上，综合考虑词频 TF (Term Frequency)和逆向

时间热度(Inverse Time Heat)，提出了改进后的 TF-ITH 词汇热度计算模型。

TF (Term Frequency)：网评文本中所涉及词汇的词频，词汇 w_j 在评论 r_i 中出现则标记为 1，否则标记为 0。则有式(1)：

$$TF_{ij} = \begin{cases} 1, & \text{if } w_j \subset r_i \\ 0, & \text{if } w_j \not\subset r_i \end{cases} \tag{1}$$

ITH (Inverse Time Heat)：逆向时间热度，评论发表日期距离当前时间越近，则评论所涉及词汇热度越高。具体逆向时间热度计算公式如式(2)：

$$ITH_i = \frac{1}{\log_2\left(\frac{(T-L)-(c_i-L)}{7}+2\right)} = \frac{1}{\log_2\left(\frac{T-c_i}{7}+2\right)} \tag{2}$$

WH (Word Heat)：词汇热度值，具体计算公式如式(3)：

$$WH_j = \sum_{i=1}^{n}\left(TF_{ij} \times ITH_i\right) \tag{3}$$

其中，TF_{ij} 表示词汇 w_j 在评论 r_i 中是否出现；ITH_i 为第 i 条评论的逆向时间热度；T 为本文固定的基准日期，2021-04-27；c_i 为第 i 条评论的发表日期；L 为最早的一条评论发表日期。同时，为防止出现日期间隔为 1 导致分母为 0 的现象，采用加 2 平滑；WH_j 表示评论中某一词汇的热度值，热度值越大，该词汇热度就越高。综上，第 j 个词汇的热度值可表示为所有涉及该词汇的网评文本的逆向时间热度之和。

3.3. 景区和酒店游客评论热门词分析

本文对每个景区游客的网络评论分别进行热门词汇提取，首先采用 TF-ITH 词汇热度计算模型对经过数据二次处理后的景区游客网评文本计算所涉及的词汇热度值，并按热度值从高到低对词汇进行排序，从而选出前 20 热门词汇。景区 A01 游客评论前 20 热门词如图 2(左)所示，可以发现景区 A01 游客评论前 20 热门词对景区 A01 特征的反映较为合理。

Figure 2. Scenic spot A01 (left) and hotel H01 (right) top 20 popular words in tourist comments
图 2. 景区 A01 (左)与酒店 H01 (右)游客评论前 20 热门词云图

4. 景区及酒店的综合评价

4.1. 评论观点提取模型简介

Bert 模型是一种基于 Transformer 架构的神经网络语言模型[7]，具有双向深度编码能力。传统的神经网络语言模型得到的词向量是单一的、固定的，不能代表词的多义词。预先训练好的语言模型很好地解决了这个问题，可以结合上下文来表示一个单词。本文采用了 Bert 预训练语言模型，该模型使用长期关注机制，可以准确地提取景区及酒店的服务、位置、设施、卫生、性价比等方面的信息，充分捕捉关系和词与词之间的关系[8]，在一个句子中有很强的模型泛化能力和鲁棒性。

Bert 利用了 Transformer 的 encoder 部分。Transformer 是一种注意力机制，可以学习文本中单词之间的上下文关系的。Bert 的目标是生成语言模型，所以只需要 encoder 机制。Transformer 的 encoder 是一次性读取整个文本序列，而不是从左到右或从右到左地按顺序读取，这个特征使得模型能够基于单词的两侧学习，相当于是一个双向的功能。如图 3 所示，在 Transformer 的 encoder 部分中，输入是一个 token 序列，先对其进行 embedding，称为向量，然后输入给神经网络，输出是大小为 H 的向量序列，每个向量对应着具有相同索引的 token。

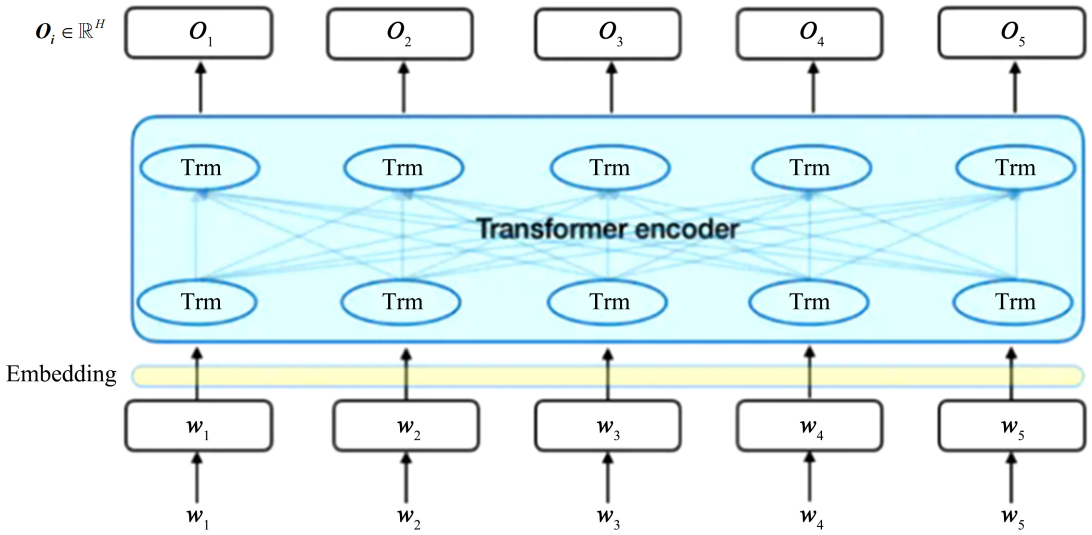

Figure 3. Schematic diagram of encoder part of transformer
图 3. Transformer 的 encoder 部分示意图

Transformer 中 encoder 部分的核心思想是计算一个句子中每个单词与句子中所有单词的相互关系，然后认为这些单词之间的相互关系在一定程度上反映了不同单词在句子中的相互关系和重要性程度。在此基础上，利用这些相互关系来调整每个单词的重要性，可以得到每个单词的新表达式。这种新的表示法不仅包含了词本身，而且还包含了其他词与词的关系，因此它是一种更全面的表示法，而不仅仅是一个词向量。与其他语言模型相比，Bert 预训练模型可以充分利用单词左右两边的信息，得到更好的分布式的单词表示。

4.2. 基于外部数据集的 Bert 模型预训练

4.2.1. 外部数据集引入与指标选取

本文通过直接对题中所给网评文本数据进行建模，发现建模得到的网评文本对旅游目的地五个方面

的评分效果较差，因此引入外部数据集。该数据集包括对目的地位置的交通便利性、位置距商业区的远近、位置的易被发现程度、服务等待时间、服务人员态度、服务的停车便利性、服务速度、性价比、环境装饰、环境噪音程度、环境空间、环境的卫生程度十二项指标的评价，−2 对应的评论标签为不相关，−1 对应的评论标签为差，0 对应的评论标签为中，1 对应的评论标签为好，x_1~x_{12} 对应上述十二项指标。

4.2.2. 数据增强

对本文所引入的外部数据集各项指标进行统计分析，发现各指标的不相关标签和评价为好的标签占比很高，该数据集属于不平衡数据集。以数据集的服务人员态度指标为例，不相关、差、中和好的数量分别为 33,937、6968、9954、33,141 条。因此，对外部数据集进行数据扩充和数据删除处理。

1) 数据扩充

首先对数据集中包含 0 和−1 类标签的网评文本按句号进行拆分，并统计网评文本拆分后的子文本个数。当子文本个数大于 3 时，将同一网评文本中的不同子文本打乱顺序，并重新进行排列组合[9]，每条网评文本组合 3 段，排列 6 次。假定网评文本 r_i 包括 x_1, x_2, x_3, x_4 四个子文本，即 $r_i = \{x_1, x_2, x_3, x_4\}$，该网评文本的子文本个数大于 3，则将该网评文本随机组合为 $\{x_1\}$，$\{x_2\}$，$\{x_3, x_4\}$ 三段新文本，并以不同顺序排列为新的文本，从而使原有的少数类网评文本新增 5 条。新文本如表 1 所示。

Table 1. Text set generated based on sub text permutation and combination
表 1. 基于子文本排列组合生成的文本集

网评文本属性	网评文本内容	网评文本属性	网评文本内容
原有文本	$\{x_1, x_2, x_3, x_4\}$	新生成文本	$\{x_3, x_4, x_2, x_1\}$
新生成文本	$\{x_2, x_1, x_3, x_4\}$	新生成文本	$\{x_1, x_3, x_4, x_2\}$
新生成文本	$\{x_3, x_4, x_1, x_2\}$	新生成文本	$\{x_2, x_3, x_4, x_1\}$

2) 数据删除

对数据集中包含 1 和−2 类标签的网评文本进行欠采样处理，从而使得包含−2 类标签的网评文本数量等于包含 0 和−1 类的网评文本数量之和；当包含 1 类标签的网评文本数量超过包含 0 或−1 类的网评文本数量二倍时，对包含 1 类标签的网评文本欠采样，使其数量为包含 0 和−1 类的网评文本数量之和。以数据集的服务人员态度指标为例，该指标经过数据增强后的四类标签数量分别为 33,860、21,070、24,136、33,141 条，可以发现，该指标各类标签数量基本达到均衡水平，数据质量符合模型训练标准。

4.3. 指标综合与评分模型选取

本文选择根据外部数据集预训练的 Bert 模型对景区与酒店网评文本进行观点提取，并将提取到的游客对旅游目的地十二项指标的观点综合为服务、位置、设施、卫生、性价比五个方面。其中服务方面包括十二项指标中的服务等待时间、服务人员态度、服务的停车便利性、服务速度四项指标；每一个指标又包含不相关、差、中、好四类标签，将四个指标的每一个标签都作为一个变量，则目的地的服务方面包括 16 个变量；每个变量的取值为该旅游目的地的全部网评文本中包含该变量对应标签的个数。旅游目的地五个方面的指标综合与变量个数具体情况如表 2 所示。

本文分别以景区及酒店的服务、位置、设施、卫生、性价比五个方面指标综合后所包括的标签取值作为输入变量，并分别以景区及酒店的服务评分、位置评分、设施评分、卫生评分、性价比评分为输出变量。其中服务方面包括 16 个自变量，位置方面包括 12 个自变量，设施方面包括 12 个自变量，卫生方

面包括 4 个自变量，性价比方面包括 4 个自变量。样本为 50 个景区和 50 个酒店在五个方面的评分及其对应的自变量取值。

分别对旅游目的地的五个方面构建模型进行拟合，以酒店的位置评分为例，本文分别选择梯度提升树模型、决策树模型、线性回归模型、随机森林等模型，并将 MSE 作为模型评价标准，采用五折交叉验证，对酒店的位置评分及其对应的自变量取值进行拟合。各个模型对酒店的位置评分拟合的 MSE 结果如表 3 所示。

Table 2. Tourism destination index synthesis and number of variables
表 2. 旅游目的地指标综合与变量个数

评价方面	指标综合	变量个数
服务	服务等待时间、服务人员态度、服务的停车便利性、服务速度	16
位置	位置的交通便利性、位置距商业区的远近、位置的易被发现程度	12
设施	环境装饰、环境噪音程度、环境空间	12
卫生	环境的卫生程度	4
性价比	性价比	4

Table 3. MSE results of hotel location score fitting of each model
表 3. 各模型对酒店位置评分拟合的 MSE 结果

模型选择	MSE	模型选择	MSE
GradientBoosting Regressor	0.01280	AdaBoost Regressor	0.01256
Decision Tree Regressor	0.02340	Bagging Regressor	0.01239
Linear Regression	0.00975	ExtraTree Regressor	0.02120
RandomForest Regressor	0.01218		

可以发现，线性回归模型对旅游目的地位置评分的拟合效果最好，MSE 相对较小。其原因一方面由于目的地位置的三个指标存在一定的相关关系，线性模型进行拟合损失较小；另一方面由于酒店位置评分的样本量为 50，选择机器学习模型进行拟合从而会存在一定的过拟合现象。因此，本文选择采用多元线性回归模型对旅游目的地的五个方面分别进行拟合。

4.4. 模型求解与评估

选择多元线性回归模型分别对景区和酒店服务评分、位置评分、设施评分、卫生评分、性价比评分进行拟合，以五个方面的指标综合后所包括的标签取值作为输入变量，并分别以服务评分、位置评分、设施评分、卫生评分、性价比评分为输出变量。随机抽取 40 个样本作为训练集，10 个样本作为验证集，拟合得到五个对应值为 y_1，y_2，y_3，y_4，y_5 并带入式 4 计算评价总分。

$$y = 0.3y_1 + 0.1y_2 + 0.15y_3 + 0.3y_4 + 0.15y_5 \tag{4}$$

景区预测总分和真实总分如表 4 所示，拟合效果如图 4 所示，可以发现各个回归模型的在验证集上的 MSE 都比较小，模型拟合效果较好。

Table 4. Predicted scores and real scores of some scenic spots
表 4. 部分景区预测评分与真实评分

名称	服务评分	位置评分	设施评分	卫生评分	性价比评分	预测总分	真实总分
A01	3.940785	4.813318	4.511846	4.749893	4.591018	4.453965	4.405
A02	4.084596	4.906879	4.783981	4.777942	4.495786	4.541415	4.495
A03	3.912413	5.011783	4.141553	4.563313	4.336011	4.315531	4.090
A04	4.138271	4.800222	4.716152	4.733425	4.624401	4.542614	4.345
…	…	…	…	…	…	…	…
A50	4.059710	4.738021	4.497342	4.810836	4.615191	4.501846	4.555

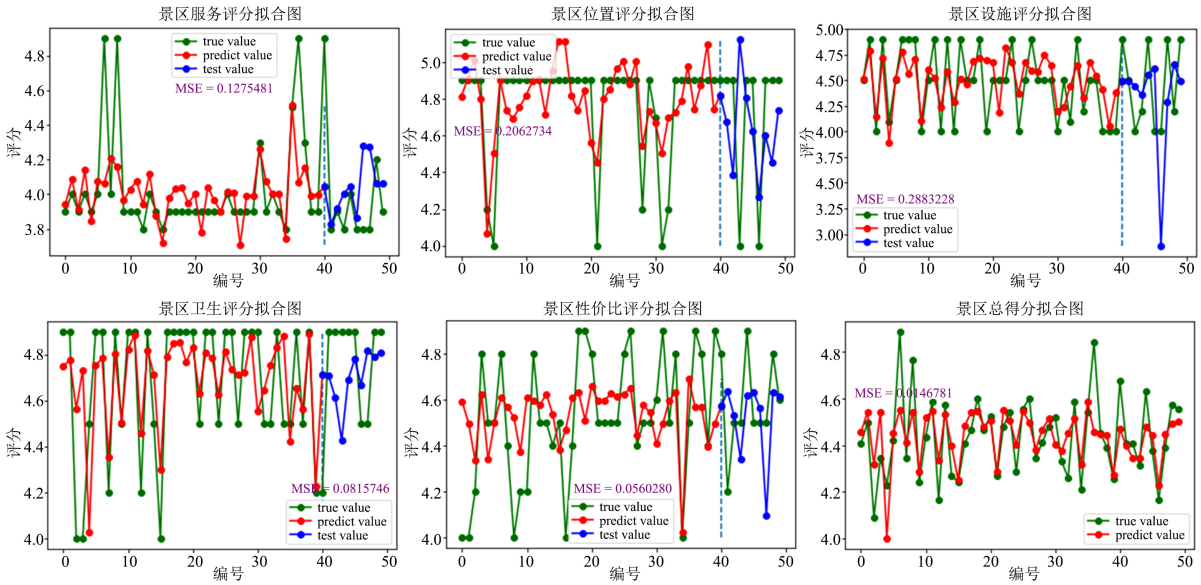

Figure 4. Fitting effect of scenic spot score linear regression model
图 4. 景区评分线性回归模型拟合效果

酒店预测总分和真实总分如表 5 所示，拟合效果如图 5 所示，可以发现各个回归模型的在验证集上的 MSE 都比较小，模型拟合效果较好。

Table 5. Forecast score and real score of some hotels
表 5. 部分酒店预测评分与真实评分

名称	服务评分	位置评分	设施评分	卫生评分	性价比评分	预测总分	真实总分
H01	4.827781	4.788751	4.766205	4.786522	4.090875	4.795744	4.6
H02	4.782249	4.730546	4.828300	4.857745	3.989731	4.800231	4.8
H03	4.760813	4.910539	4.754557	4.774841	4.055080	4.788583	4.8
H04	4.626933	4.665678	4.606252	4.718910	4.028817	4.654668	4.7
…	…	…	…	…	…	…	…
H50	4.660390	4.825611	4.618387	4.632776	4.046690	4.672825	4.8

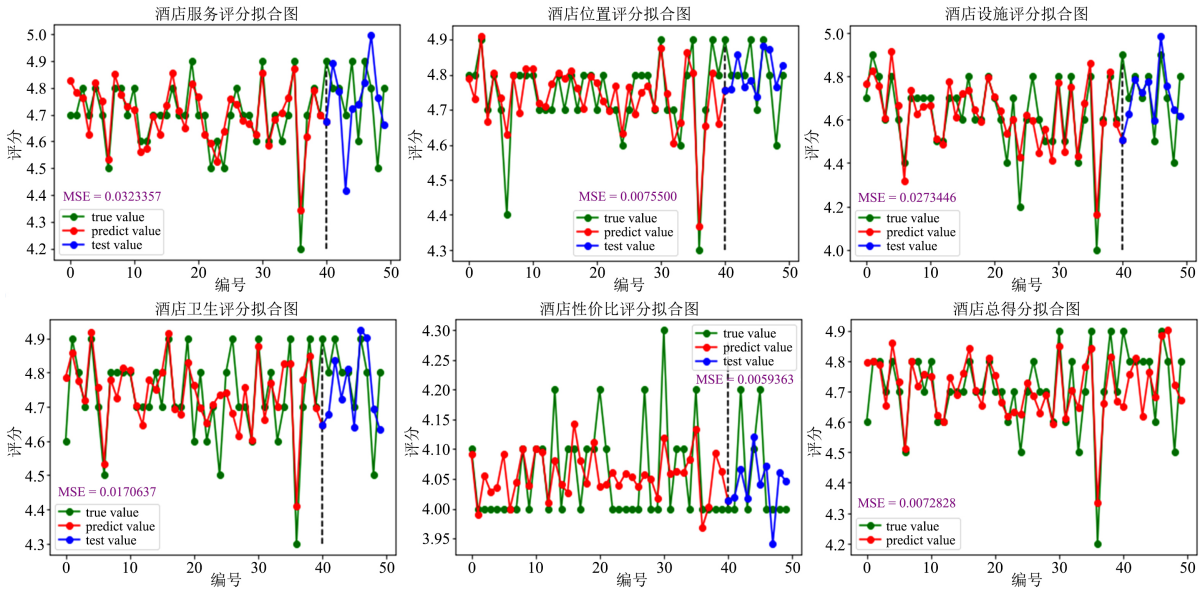

Figure 5. Fitting effect of hotel scoring linear regression model
图 5. 酒店评分线性回归模型拟合效果

5. 网评文本的有效性分析

5.1. 基于有效性的网络评论文本排序与筛选模型构建

针对景区及酒店的游客网络评论常常出现内容不相关、简单复制修改和无有效内容等现象[2]，本文通过提出一种基于有效性的网络评论文本排序与筛选模型，建模流程如下：

步骤一采用词性标注工具对评论进行标注；

步骤二对待排序评论集中的名词出现次数进行统计，并按词频从高到低提取出评论数乘以 1%之前的高频名词构建评论目标的特征集；

步骤三依次对待排序评论集中的每一条评论进行处理，得到每条评论中涉及的特征数；

步骤四对每一条网评文本中涉及的特征权重赋值为 2，并将该条网评文本中除涉及特征之外的所有名词权重赋值为 1；

步骤五依次将待排序评论集中的每一条评论的所有权重求和，并按照权重之和将评论从高到低进行排序；

步骤六将每条评论权重之和作为网评文本的有效性评分，筛选出有效和无效网评文本。

5.2. 网评文本词性标注与评论目标特征集构建

首先，本文首先采用 Hanlp 分词器对文档进行分词[10]，在 Hanlp 自然语言处理类库中封装好的 Hanlp 类中共有五种分词器，分别为维特比分词器、双数组 trie 树分词器、条件随机场分词器、感知机分词器、N 最短路分词器，本文选择默认的维特比分词器来对网评文本进行分词，并对不同网评文本分词后相似度超过 90%的词汇进行去重，只保留发表时间最早的网评文本中的词汇。

然后，本文采用 Hanlp 经过中文预训练的 fastText 词性标注模型，对经过分词处理后的网评分本进行词性标注，如“酒店，很，适合，家庭，出行”被标记为“n, d, v, n, v”。

最后，分别统计出不同景区及酒店的网评文本涉及的所有名词词频，在单个评论目标下，按词频从高到低提取出该目标的评论数乘以 1%之前的高频名词，并将这些高频名词作为构建特征集的评论目标特征。

5.3. 网评文本有效性分析

网评文本有效性评分统计

对景区及酒店的每一条游客网评文本中涉及的特征权重赋值为 2，并将该条网评文本中除涉及特征之外的所有名词权重赋值为 1，以此得到一组关于该条网评文本的权重值。将该组权重值求和，作为这条网评文本的有效性评分，从而可以得出不同的景区及酒店的网评文本的有效性平均评分，并筛选出单个景区或酒店的无效网评文本。部分网评文本有效性评分如表 6 所示。

以每个表格中的第一条网评文本为例进行分析，有效性评分较高的第一条网评文本“A01 欢乐世界 1、是个大型的游乐场，比较有名的是垂直过山车，惊险刺激。二、A01 水上乐园三、A01 野生动物世界 1、目前国内最大的原生态动物园，这里可以看到精彩的动物表演秀，比较特别的比如白虎表演。很适合带孩子来玩，可以在‘丛林发现’了解动物习性，还可以在‘儿童天地’玩游艺项目……”，可以发现，该条评论分层次分景点介绍了景区特色内容，帮助游客更加详细了解具体景区的可游玩性，加深游客对景区印象，并且给游客提供了游玩、交通、餐饮等方面的建议。

有效性评分一般的第一条网评文本“非常震撼史诗极的表演花得值外地来的朋友如果不住在旁边的话最好早点过去普通票是随便坐的占座要趁早在最前面可以和演员们互动接到小娃娃的机会也大些”，可以发现，该条评论并没有全方位的对景区进行介绍，而是选择景区的单独特色内容进行介绍，同样也对游客提供了借鉴与参考，对游客有一定的帮助性。

有效性评分较低的第一条网评文本“还不错，就是品种不是很多”，可以发现，该条评论在字数和内容介绍方面和有效性较高的评论相比均比较少，并没有指出评论内容的描述对象与景区特色，对游客的参考借鉴意义并不是很大。因此将其归类为有效性较低的网评文本较为合理。

Table 6. Some online review texts and their effectiveness scores
表 6. 部分网评文本及其有效性评分

评论内容	得分	评论内容	得分	评论内容	得分
A01 景区一、A01 欢乐世……	38	非常震撼史诗极的表演……	5	还不错，就是品种不是很多	0
去过很多地方的 A01 景区……	34	超级好的动物园管理严格……	5	非常震撼，值得一看	0
我预订了 A01 旅游渡假区……	30	坐了垂直过山车，很刺激……	5	值得去，里面有很多树……	0

6. 总结

本文在 TF-IDF 模型基础上，提出综合考虑词频与时间跨度的 TF-ITH 词汇热度计算模型，该模型采用词频乘以逆向时间热度，来解决存在时间跨度的旅游目的地网评文本词汇的热度值计算问题。该模型简洁、易于理解，能够准确反映随时间变化的不同景区及酒店的游客评论热门词汇，切合游客评价实际情况，对于探索游客的目的地印象较为有效。

本文引入包括 12 项指标的外部数据集训练用于得出游客评价的 Bert 模型，并将训练好的模型用于题中所给网评文本的评价观点提取，针对每一项指标分别进行模型训练，从而得出游客网评文本在 12 项指标上的标签值。将游客对旅游目的地的十二项指标评价归纳为服务、位置、设施、卫生、性价比五个方面，并根据得到的游客对旅游目的地的评价变量拟合线性模型，并以 MSE 对线性模型拟合效果进行评估。该模型较为简便、准确率高，能够较为准确地反映景区及酒店的评分与评论之间的关系。

本文针对旅游目的地网络评论常常出现内容不相关、简单复制修改和无有效内容等现象，本文提出一种基于有效性的网络评论文本排序与筛选模型，该模型可以剔除旅游目的地游客无效评论，便于从游客评论中获取有价值信息，以此对景区及酒店网络评论进行有效性分析。在构建模型过程中，首先构建基于高频名词的目标特征集，然后依次对每一条待排序网评文本进行筛选，得到每条网评文本涉及的特征数，对网评文本中涉及的特征和剩余名词分别赋予不同的权重，并将依次将每条网评文本中所有的名词权重求和，从而求出每条网评文本的有效性评分。

更进一步地，我们还可以利用热词和有效的评论来分析每个酒店和景点的优劣势，为酒店和景区的整改提供科学依据。

参考文献

[1] 芦珊, 梁燕子. 探索旅游文化宣传新模式提升旅游品牌美誉度[EB/OL]. http://yuqing.people.com.cn/n1/2017/0421/c394872-29227476.html, 2017-04-21.

[2] 吴宝清, 吴晋峰, 吴玉娟. 基于网络文本的 TDI 地域差异研究——以西安的国内旅游形象为例[J]. 浙江大学学报: 理学版, 2015, 42(4): 474-482.

[3] 庄小丽, 程仕菊, 常雪萍. 基于文本挖掘的峨眉山风景区旅游形象感知[J]. 国土资源科技管理, 2020, 37(1): 106-117.

[4] 李凤佼. 基于网络文本挖掘的冰雪旅游形象感知研究[D]: [硕士学位论文]. 哈尔滨: 东北农业大学, 2019.

[5] Gu, Y.W., Wang, Y.R., Huan, J., Sun, Y.Q. and Xu, S.K. (2020) An Improved TFIDF Algorithm Based on Dual Parallel Adaptive Computing Model. *International Journal of Embedded Systems*, **13**, 18-27.

[6] Oschina. Reddit 的排名算法原理[EB/OL]. http://www.oschina.net/translate/how-reddit-ranking-algorithms-work, 2013-08-07.

[7] Lee, K., Filannino, M. and Uzuner, Ö. (2019) An Empirical Test of GRUs and Deep Contextualized Word Representations on De-Identification. *Studies in Health Technology and Informatics*, **264**, 218-222.

[8] 黄梅根, 刘佳乐, 刘川. 基于 Bert 的中文多关系抽取方法研究[J/OL]. 计算机工程与应用: 1-9. http://kns.cnki.net/kcms/detail/11.2127.TP.20210426.1358.004.html, 2021-05-04.

[9] 管箫笛, 李凡, 贺丽君. 基于数据扩充与迁移学习的真实失真图像质量评估算法[J]. 中国科技论文, 2021, 16(3): 241-246+270.

[10] 邱德钧, 冯霞. 谓词逻辑视角下 Hanlp 中文分词中对歧义的处理[J]. 科学经济社会, 2020, 38(1): 33-38.