基于 Lasso-LDA 的酒店用户偏好模型

赵志杰 刘 岩 张艳荣 周婉婷 孟令跃
(哈尔滨商业大学计算机与信息工程学院 黑龙江 哈尔滨 150028)
(哈尔滨商业大学黑龙江省电子商务与信息处理重点实验室 黑龙江 哈尔滨 150028)

摘 要 随着首个在线旅游数据生态共建倡议书的发布,在线评论数据更加真实、准确地表达顾客的客观感受,成为商家和消费者情报的重要来源。结合 LDA、TF-IDF 算法获取不同类型酒店客户评论特征权值,采用 AipNLP 获得情感倾向性估计值。利用 Lasso 算法进行特征筛选构建基于 Lasso-LDA 的用户偏好模型,将该模型应用于携程网上五种类型用户的偏好分析中。研究结果表明,与传统的多元线性回归及岭回归相比,该模型有更好的预测效果。

关键词 酒店用户偏好 LDA TF-IDF Lasso 情感分析

中图分类号 TP399 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.02.004

HOTEL USER PREFERENCE MODEL BASED ON LASSO-LDA

Zhao Zhijie Liu Yan Zhang Yanrong Zhou Wanting Meng Lingyue
(*School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, Heilongjiang, China*)
(*Heilongjiang Key Laboratory of Electronic Commerce and Information Processing, Harbin University of Commerce, Harbin 150028, Heilongjiang, China*)

Abstract With the publication of the first online tourism data ecology co-construction proposal, online commentary data more truthfully and accurately express customers' objective feelings, and become an important source of business and consumer information. This paper combined LDA and TF-IDF algorithms to obtain the feature weights of different types of hotel customer reviews, then used AipNLP to obtain the emotional orientation estimates. The Lasso algorithm was used to filter the features and construct a user preference model based on Lasso-LDA, which was applied to the preference analysis of 5 types of users on Ctrip. The results show that, compared with the traditional multiple linear regression and ridge regression, the hotel user preference model constructed in this paper has better prediction effect.

Keywords Hotel user preference LDA TF-IDF Lasso Sentiment analysis

0 引 言

互联网与新兴信息技术的快速发展使得人们不再只是信息的传递者同时也是信息的创造者。互联网海量数据的存在,使用户难以高效地获取自己感兴趣的数据,导致“信息过载”现象的存在。2018 年年末在线旅游数据生态与治理峰会上八大 OTA 平台联合发布行业数据治理倡议书《在线旅游行业内容和数据生态共建》。这一倡议书建议为消费者提供更真实可靠的旅游数据,帮助用户正确、高效地选择和决策。随着移动互联网基础设施的不断完善,互联网的普及率急速上升,多元化、专业化的酒店顾客需求开始觉醒。中国互联网络信息中心发布的《第 43 次中国互联网络发展状况统计报告》显示,截止到 2018 年 12 月,30.3% 的网民在网上预订酒店[1],这一举措给酒店业的建设提出挑战。由此可见,研究消费者的用户偏好对如今的酒店业而言意义非凡。

本文以 OTA 巨头“携程网”上的五种类型酒店顾客产生的酒店评论为基础数据,运用文本挖掘技术、情

收稿日期:2019-09-02。教育部人文社会科学研究项目(18YJAZH128)。**赵志杰**,教授,主研领域:图像处理,电子商务。**刘岩**,硕士生。**张艳荣**,副教授。**周婉婷**,硕士生。**孟令跃**,硕士生。

感分析技术和机器学习算法分别对五类用户评论数据进行分析处理,通过对文本数据进行特征聚类、权值计算、情感倾向性估计值计算、特征优选,构建基于 Lasso-LDA 的用户偏好模型。采用 LDA 主题模型聚类,总体得出顾客对于酒店的一系列偏好因素;运用 Lasso 回归进行特征筛选,基于每一类型的顾客剔除不重要的特征因素以达到特征优选,最后得到用户偏好模型。模型有助于顾客根据自己的需求精准地选择适合的酒店,而对于商家,也可以依此有针对性地打造出个性化服务和创新经营方式,提升市场竞争力。

1 相关研究

1.1 LDA 模型聚类

LDA 是最先由 Blei 等在 2003 年提出的包含文档-主题-词 3 层贝叶斯文档主题生成模型,LDA 是一种无监督的机器学习方法,用来识别隐藏在文档集或语料库中的主题信息。对于 LDA 在文本挖掘中的应用,文献[2]使用 LDA 对小红书中的评价文本数据进行主题建模,将聚类得出的高频词划分为 8 个主类目,构建结构方程模型研究小红书用户粘性形成的动态机制。文献[3]将 LDA 这种半监督方法与其他的半监督方法和监督分类方法对比,结果表明,在文本分类精度方面 LDA 方法远优于其他方法。同时,实验证明 LDA 方法可适用于标签文本缺失的情况下。

随着 LDA 模型的不断完善,LDA 被广泛应用于各行各业的文本分析。文献[4]采用 LDA 模型对汽车保险欺诈索赔中的文字信息进行文本分析,结合深度神经网络对数据进行训练。实验结果表明,结合深度神经网络和 LDA 的框架适用于判断汽车保险欺诈问题。文献[5]描述一个使用电子请愿数据训练和验证 LDA 的框架,通过严格的训练和评估,87% 的 LDA 生成的主题对法官了解请愿者的主要诉求有参考意义,发现 LDA 主题可以比通过手动内容分析提取的主题更具一些优势。LDA 能够反映文本中表达的多个主题,提取人类编码器未突出显示的新主题,并且不易受人类偏见的影响。

1.2 Lasso 特征优选

Lasso 是由 Robert Tibshirani 于 1996 年首次提出的一种基于压缩估计的特征选择方法并且应用于各个行业领域。文献[6]将 Lasso 框架应用于虚拟金融上,把返回的 21 个潜在因素优化替换为 8 个因素,找出影响强度最重要的两种变量。文献[7]将 Lasso 应用于船舶业中,用以预测不同海况和天气下船舶的燃油消耗,得到大量的特征变量,应用 Lasso 实现特征选择,提出一种新的预测模型。文献[8]应用 Lasso 研究与金融因素、市场驱动指标和宏观经济预测因素相关的市场隐含信用评级的决定因素,记录了实质性的预测能力,将 Lasso 选择的模型与基准有序概率模型进行比较,发现 Lasso 选择的模型具有卓越的预测能力,在全部样本预测中都优于基准有序概率模型。文献[9]将 Lasso 应用在医药行业上,提出一种新的药物-靶标相互作用预测方法,使用 Lasso 减少提取的特征信息维度,然后使用合成少数过采样技术(SMOTE)方法处理不平衡数据。最后,将处理后的特征向量输入随机森林(RF)分类器以预测药物-目标相互作用。文献[10]提出一种自适应特征提取算法,预先生成各种大气条件下的光谱特征,然后利用 Lasso 算法进行快速特征优选,选择出最优目标-背景组合重构背景光谱,最后提取目标特征。文献[11]将 Lasso 应用于金融领域,不同于以往常规的变量选择,提出针对时间序列的改进自适应 Lasso 方法,提高对未来的预测能力。

1.3 用户偏好

新兴信息技术推动着消费结构从生存型消费向享受型、发展型消费转变,消费者不再被动地接受来自商家提供的服务,而是通过自身的参与和网络生成内容主动地发表自己的偏好。文献[12]提出一种从一组评论中提取评论贡献者偏好的方法。提取的偏好用于酒店推荐,使得贡献者给出的具有类似于用户偏好的评估值被赋予更大的权重,用此方法可以推荐符合用户偏好的酒店。文献[13]提出用于从评论文本中学习和表示用户的偏好知识,利用所获得的表示来支持评级预测的一种混合方法,并用此方法对亚马逊产品数据集进行实验,揭示用户偏好知识表现的能力以及对评论预测的影响。文献[14]利用用户的评分与评论数据,提出一种基于贝叶斯网络的用户偏好建模方法。利用隐变量确定模型的初始结构约束和初始参数约束,使用亚马逊电影评价数据集作为测试数据,对用户偏好模型进行验证。文献[15]针对高维、稀疏的评分数据提出一种基于深度信念网络和贝叶斯网络的用户偏好建模方法,分别利用深度信念网络和贝叶斯网络对评分数据进行分类以及描述相关属性间的不确定性,最后使用 MovieLens 和大众点评数据对模型进行验证。

CNNIC 报告显示,截止到 2019 年 6 月我国在线旅行预订用户占网民整体的 48.9%。随着中国经济发展加速,“人均 GDP 1 万美金俱乐部”成员呈指数上

升,越来越多的新人口进入旅游消费市场,使得酒店预订需求进一步增长。Trustdata 移动大数据监测平台于 2019 年 8 月 29 日发布的《2019 上半年中国在线酒店预订行业发展分析报告》显示,主流在线酒店预订平台用户粘性均超 20%,其中携程表现最优达 24.3%。因此,本文基于携程网平台进行调研,将本文所得情感倾向性估计值与之相比,发现存在评论与分值具有偏差的问题。本文利用 AipNLP 计算情感倾向性估值对存在偏差的数据进行剔除,以便得到实验所需的真实数据,本文构建的模型进一步提升酒店的管理经营模式。携程有着自有的评价指标,分别是环境、设施、服务和卫生四个方面,但分析大量的评论数据后,发现评论的文本与携程自有的用户偏好特征不能完全地进行匹配,评论文本本身包含更多和更详细的信息。为了获得更加客观和细致化的用户偏好特征,本文在评价指标的获取中使用 LDA 模型进行用户偏好特征聚类,为使获取的特征更理想,使用 Lasso 算法剔除掉聚类中不重要的特征,得以分辨出五种不同类型的顾客所关注的特征指标的不同,使得不同类型的顾客个性化偏好存在差异。例如,假设用户重视交通的便利程度,则对于这类顾客而言个性化偏好为交通方面,使用 Lasso 特征优选尽可能地剔除与偏好特征不一致的特征,从而使商家有效地对不同类型的顾客提供不同的酒店服务。

综上所述,目前国内对酒店用户偏好模型的构建还有待完善,大多数学者只是从酒店本身总体的经营情况进行建模,得出的一系列特征指标是针对酒店总体性的,并没有从酒店客户群体进行考虑,未细分顾客群体,盲目地将总体的偏好强加于各类顾客上。因此,本文基于这一问题,首先使用 LDA 主题模型将所得到的数据进行总体聚类,得出一系列特征因素;然后针对每种类型客户的 TF-IDF 权值计算每种类型客户的个性化偏好属性值;最后通过对比三类回归方法,利用更为精准的 Lasso 特征优选得到每种类型客户的优选特征,构建基于 Lasso-LDA 的用户偏好模型,为酒店管理者随时追踪顾客认知和服务质量提供客观、真实、有效的信息,从而能快速有效地为不同的用户群体提供其满意的个性化服务,而不再局限于现有酒店行业一成不变的服务,为酒店提升行业竞争力。

2　模型设计

本文主要运用 LDA 模型对用户偏好特征聚类,基于 TF-IDF 对用户偏好权值进行计算,结合情感倾向性分析方法对酒店用户评论进行统计分析,确定用户偏好程度,最后运用 Lasso 算法对用户偏好特征进行筛选,构建出基于 Lasso-LDA 的用户偏好模型。该模型按照信息处理的先后顺序分为三个部分:数据的采集及预处理,基于 LDA 的用户特征偏好的确定,基于 Lasso-LDA 的用户偏好模型的构建。本文的研究框架如图 1 所示。

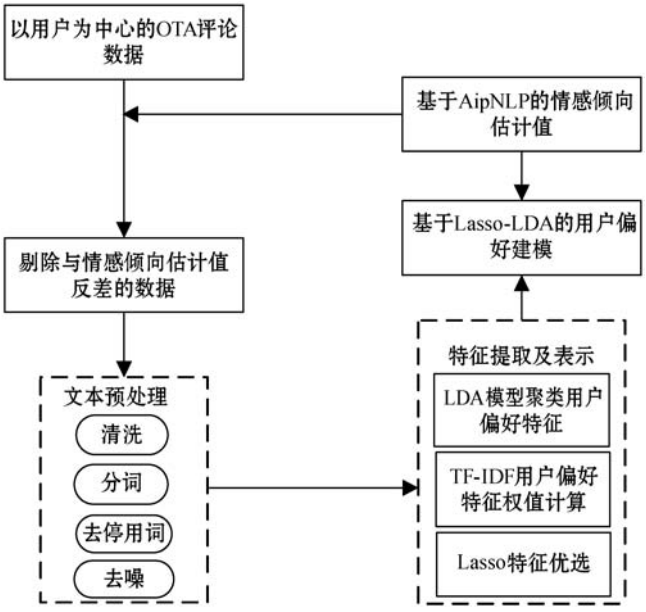

图 1　基于 Lasso-LDA 用户偏好模型研究基本框架

2.1　数据来源及预处理

(1) 数据来源。携程财报公布数据显示,截至 2018 年 12 月 31 日,携程全年住宿预订收入为 116 亿元人民币,同比增长 21%,全年旅游度假业务营业收入为 38 亿元,同比增长 27%,整体行业发展呈上升趋势,行业优势明显。因此,本文主要以携程网上的酒店评论数据为数据源,借助网络信息采集工具“八爪鱼采集器”对数据进行采集,并将采集到的每一条记录内容通过八爪鱼采集器以 Excel 表格形式导出。

(2) 基于 AipNLP 的反差评论数据剔除。由于在所收集的数据中会存在评论数据与评分数据不一致的数据,因此采用情感倾向性分析方法对这类数据进行排除,确保数据的有效性。本文采用百度自然语言处理平台进行情感倾向性估值计算,该平台可自动对包含主观信息的文本进行情感倾向性判断,为口碑分析、话题监控和舆情分析等应用提供基础技术支持。同时,该平台基于深度学习训练,在相对长的句子上仍能确保较高的效果,可得到整体精度很高的情感倾向性分析结果。此外,该平台垂直类效果优,在酒店、汽车等多个垂直类上情感倾向性分析可达到 95% 以上的准确率,并且已应用于实际电商产品销售分析中。在

测试过程中本文应用情感倾向分析接口对包含主观观点信息的文本进行情感倾向性类别(积极、消极和中性)的判断,例如用户评论:“前台的服务意识没有达到星级标准,体验超差! 直接给安排的吸烟区房间,这季节根本不满房,离店时又说没提早和她说开发票,服务和体验超差!”经过 AipNLP 处理之后,可得到表 1 所示的结果,其中:positive 代表积极类别的概率;negative 代表消极类别的概率;confidence 代表分类的置信度;sentiment 代表情感倾向性分类结果。在测试过程中主要应用 post 方式进行调用,JSON 作为返回格式。由于携程平台上的酒店用户评分采用 5 分制原则,为了便于对比,本文根据 5 × ′positive′将得出的情感倾向性估值与酒店评分进行对比,将评论数据与评分数据不一致的数据剔除。通过分析示例用户评价内容可知该评论为差评,而用户给出的星级评分为 5 分,这明显高于情感倾向性估值 0.03 分,为无效数据,需剔除。在实验数据处理中将采集到的每条评论数据运用 AipNLP 进行上述处理,将反差数据排除,由于 AipNLP 计算出的情感倾向性估值较携程平台上用户星级评分值更加客观和具体,因此,将得到的情感倾向性估值数据进行保存,方便后续建模使用。

表 1 反差数据用例

酒店评分	positive	negative	confidence	sentiment	情感倾向性估值
5.0	0.006	0.994	0.987	0	0.03

(3) 数据预处理。数据预处理是为了保证数据的有效性,是数据处理过程和分析过程中不可缺少的关键步骤。在本文数据预处理过程中主要对数据进行清洗、分词、去停用词及去噪处理。为了保证模型构建的准确度,采用中科院谭松波教授整理的酒店评论数据集作为本文模型构建时数据处理的数据集。该数据集共 10 000 篇评论,将其 80% 的评论作为训练集,20% 的评论作为测试集。在对所收集到的数据进行分析测试时发现,需要清洗掉的数据主要包括:① 同一个用户进行多次评论,且评论内容相同,此时必须对重复数据进行删除,否则会对所测试的真实的正负面评论产生“虚高”影响;② 有些用户评论为无效评论,比如评论内容全部为标点符号或表情符号,这些数据需全部删除。接下来针对清洗后的评论语句,在处理过程中运用 jieba 分词工具进行分词处理,同时加载哈工大的停用词表,停用词表会根据本文的需要剔除一些词汇。最后利用过滤函数过滤如日期、英文等噪声数据,将经过预处理后的数据保存进行后续处理。

2.2 基于 LDA 的用户特征偏好确定

本文采用 LDA(隐含狄利克雷分布)主题模型聚类方法面向处理过的数据,聚类一定量的因素来确定用户对酒店服务的特征偏好。LDA 是判断两个文档的关联程度使用的方法,主要查看两个文档中出现相同单词的个数,一个文档表示一些主题所构成的概率分布,一个主题代表一些单词所构成的概率分布。同时,词袋方法被应用于 LDA 中,该方法使每篇文档被看作一个词频向量,并将文本信息转化为易于建模的数字信息。由于词袋方法不考虑两个词之间的顺序,因此问题的复杂性也就被简单化。LDA 概率图模型如图 2 所示。

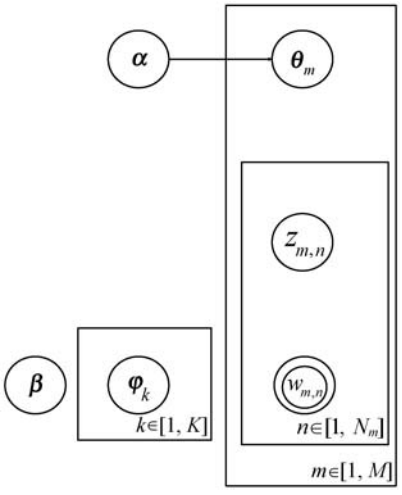

图 2 LDA 的概率图模型结构

图 2 中,m 表示文章序号;k 表示主题个数;n 表示词袋长度;N_m 表示第 m 篇文章中单词的总数;$\boldsymbol{\alpha}$ 表示每篇文章的主题分布的先验分布狄利克雷(Dirichlet)分布的参数(也被称为超参数,简称 Dir);$\boldsymbol{\beta}$ 表示每个主题的词分布的先验分布 Dirichlet 分布的参数,是一个 V 维向量,V 代表词汇表里的所有词的个数;$\boldsymbol{\theta}_m$ 是一个 K 维列向量,表示第 m 篇文章的主题分布;$\boldsymbol{\theta}_m \sim Dir(\boldsymbol{\alpha})$ 表示本文所需参数;$\boldsymbol{\varphi}_k$ 是一个 V 维向量,表示第 k 个主题的词分布;$\boldsymbol{\varphi}_k \sim Dir(\boldsymbol{\beta})$ 也为本文所需参数;$z_{m,n}$ 表示第 m 篇文章第 n 个词被赋予的主题;$w_{m,n}$ 表示第 m 篇文章第 n 个词。主题分布表示为:

$$p(\boldsymbol{z}\mid\boldsymbol{\alpha}) = \prod_{m=1}^{M} p(\boldsymbol{z}_m\mid\boldsymbol{\alpha}) = \prod_{m=1}^{M}\frac{\Delta(\boldsymbol{n}_m+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \tag{1}$$

词分布表示为:

$$p(\boldsymbol{w},\boldsymbol{z}\mid\boldsymbol{\alpha},\boldsymbol{\beta}) = p(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{\beta})p(\boldsymbol{z}\mid\boldsymbol{\alpha}) = \prod_{k=1}^{K}\frac{\Delta(\boldsymbol{n}_k+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}\prod_{m=1}^{M}\frac{\Delta(\boldsymbol{n}_m+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \tag{2}$$

根据式(1)-式(2)结合代码可知 LDA 工作流程为:① 将预处理后的数据集、关键词数量、主题数量三

个参数传入主题模型函数中,并使用 gensim 接口,将文本转为向量化表示,构建词空间,使用 BOW 模型向量化,根据 TF-IDF 算法对每个词进行加权计算,得到加权后的向量表示;② 选择加载的模型 LDA,得到数据集的主题-词分布;③ 对词分布和文档分布的相似度进行计算,将相似度最高的词作为关键词,再对输入文本与每个词的主题分布进行相似度计算;④ 取相似度最高的前 8 个词作为用户特征偏好影响因素。LDA 实验结果如表 2 所示。

表 2　主题分布相似度

主题名称	符号表示	主题分布相似度
总体感受	X1	0.9 182 690 503 970
设备设施	X2	1.3 549 815 420 700
餐饮	X3	0.5 526 764 027 320
位置	X4	0.5 655 912 957 002
交通	X5	0.7 492 234 331 770
价格	X6	0.6 558 748 838 880
服务	X7	1.3 570 225 016 100
卫生	X8	0.4 833 989 937 250

数据结果显示,酒店用户在总体感受、设备设施、餐饮、位置、交通、价格、服务和卫生八个方面的主题分布相似度测试数据位于测试结果的前八位,其中:主题分布相似度最高的是服务属性,设备设施属性位于第二。因此可知酒店用户通常会将入住酒店的服务作为首要关注点,其次为酒店提供的设备设施条件。毋庸置疑,好的服务水平和设备设施条件从感官上会直接带给用户舒适的入住体验。同时,总体感受、交通、价格、餐饮、位置、卫生这六个用户特征偏好也会得到很高的用户关注,因此,酒店管理人员应及时调整各方面的服务水平,确保酒店良好运营。

2.3　模型构建

(1) 基于 TF-IDF 的用户偏好权值计算。TF-IDF 是词频和反文档频率两个算法的综合应用,利用 TF-IDF 算法结合情感倾向性分析方法对评论文本数据特征进行赋值,并将情感倾向性估计值作为用户的偏好程度。一个文档里的词汇重要性计算式表示为:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

式中:$n_{i,j}$表示该词汇在文件 d_j 中出现的次数;$\sum_k n_{k,j}$ 表示所有词汇在文件 d_j 中出现的次数总和。逆向文件频率 IDF 计算式表示为:

$$idf_i = \log \frac{|D|}{1 + |\{j:t_i \in d_j\}|} \tag{4}$$

式中:$|D|$表示语料库中存在的文件总数。如果该词不在库中,则被除数为零,因此式(4)被除数由式子 $1 + |\{j:t_i \in d_j\}|$代替,最后得到 TF-IDF 值为:

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{5}$$

由式(5)可知,一个文件内的词频率乘以该词在整个文件集合中的文件频率,可得到 TF-IDF 值。一般来说,文本表示方式分为离散式和分布式两种,结合本文的数据情况,采用离散式文本表示方法中的 TF-IDF 算法对评论数据进行权重计算以得到特征属性表示值,具体执行过程为:① 获取总的文档数,记录每个词出现的文档数;② 按公式将其转换为 IDF 值,然后进行拉普拉斯平滑处理,使用该方法目的是将分母加 1,对于没有在字典中出现的词,将该词默认为只在其中一个文档中出现过,最后得到默认的 IDF 值;③ 按公式计算 TF-IDF 值,根据 TF-IDF 的排序,取排名前 *keyword_num* 个词作为关键词,在评论中每个因素如果有多个就进行 TF-IDF 值的求和运算,如果评论中未出现某影响因素,则赋值为 0。例如评论:“位置距离哈站只有几分钟的车程,打车起步价。刚开业三个月大堂豪华,室内干净高档完全不像这个价位的酒店,性价比极高,就是距离地铁站有点小远步行大概十几分钟,总之住宿体验很好”,实验结果如表 3 所示。

表 3　TF-IDF 实验结果

属性	特征值
X1	0.394
X2	0.303
X3	0
X4	0.185
X5	0.735
X6	0.307
X7	0.364
X8	0

(2) 基于 Lasso 的用户特征偏好筛选。本文主要利用 Lasso 回归,剔除相关性较小因素,得到 Lasso 预测模型,对用户特征偏好进行筛选。Lasso 是一种处理具有复共线性数据的有偏估计,它利用所构造的惩罚函数确定相对精炼的模型,利用这个模型压缩一些系数,同时设定某些系数为零,通过这个方法能够将子集收缩的优点保留下来。Lasso 回归又叫线性回归的 L1 正则化,它通过对最小二乘估计加入 L1 范数作为罚约束,使某些系数估计为 0,因此可以减少参数数量,

Lasso 回归预测模型目标函数表示为：

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=RSS+\lambda\sum_{j=1}^{p}|\beta_j| \tag{6}$$

式中：*RSS* 是实际值减去估计值的差的平方和；λ 是调优参数；p 为参数个数。根据式(6)可知，由于 Lasso 回归模型的目标函数包含惩罚项系数 λ，因此在计算模型回归系数前，需要得到最理想的 λ 值，λ 值的确定可以通过定性的可视化方法和定量的交叉验证方法。同时，Lasso 作为一种 λ 特征选择方法相比于岭回归，其在完成系数估计的同时就能够完成特征的选择，还能够降低过拟合，是近几年备受关注的特征选择工具，综合以上研究结果结合用户偏好相关理论研究，可得不同类型用户的偏好模型表示为：

$$user\_preferences=Intercept+\sum_{i=1}^{n}\omega_i s_i \tag{7}$$

式中：*user_preferences* 代表用户偏好；*Intercept* 代表截距项；s_i 代表用户偏好特征因素；ω_i 代表对应 s_i 的系数。

3 实 验

3.1 总体方案

本文利用八爪鱼数据采集器从携程网的酒店社区共采集 15 000 条用户评论数据作为数据源，在采集过程中主要以用户类型为独自出行、朋友出游、亲子旅行、情侣出游、商务出差的五类人士，对酒店进行的评论以及对应的酒店总评分和环境、设施、服务、卫生四个方面的评分为采集数据。采集后利用 AipNLP 剔除评论反差数据，对剩余有效数据再进行预处理，然后采用 LDA 主题聚类的方法提取用户特征偏好，并通过 TF-IDF 统计特征值对评论文本数据特征进行赋值，利用情感倾向性估计值作为用户的偏好程度，最后采用 Lasso 进行特征的筛选及预测。

3.2 实验结果及分析

在筛选过程中针对用户类型为独自出行、朋友出游、亲子旅行、情侣出游、商务出差这五类人士在总体感受、设备设施、餐饮、位置、交通、价格、服务和卫生八个方面的数据利用 Lasso 回归与线性回归和岭回归做对比，以商务出差用户评论数据为例，将 80% 的数据作为训练集，20% 的数据作为测试集，采用 sklearn 子模块 linear_model 中的 Lasso 类及 Ridge 类对 Lasso 回归和岭回归中目标函数所包含的惩罚项系数进行计算，如图 3 和图 4 所示。

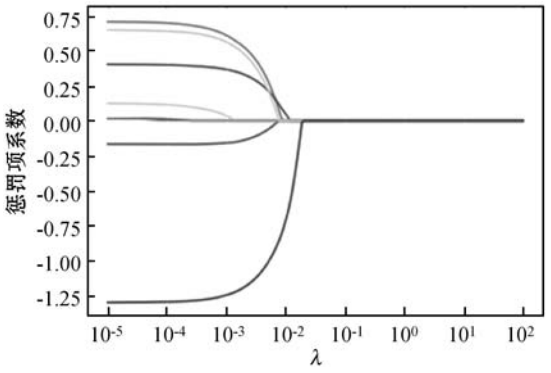

图 3 LASSO 回归结果图

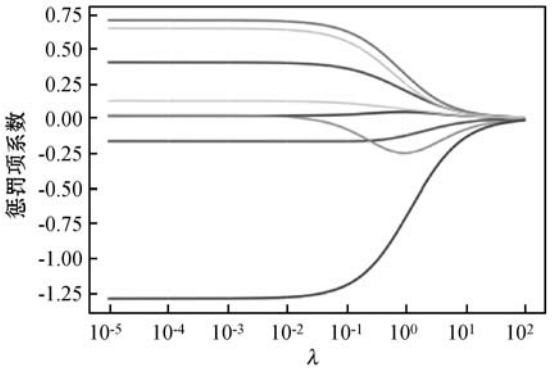

图 4 岭回归结果图

可以看出，初始迭代的 λ 值落在 $10^{-5}\sim10^{2}$ 之间，图中的每条曲线指代不同的变量。由于出现了喇叭形曲线，说明该变量存在多重共线性，图 3 中 λ 值落在 0.000 5 附近，图 4 中 λ 值落在 0.05 附近，此时绝大多数变量的回归系数趋于稳定，所以可以锁定合理的 λ 值范围。接下来分别采用 sklearn 子模块 linear_model 中的 LassoCV 类及 RidgeCV 类，采用 10 重交叉验证的方法分别得到 Lasso 回归与岭回归的最佳的 λ 值，$Lasso_\lambda=0.000\,8$，$Ridge_\lambda=0.074\,1$，与可视化方法确定的 λ 值范围基本一致。最后基于最佳的 λ 值分别得到 Lasso 和岭回归的模型回归系数，采用 statmdels 子模块 api 类对数据进行训练得到多元线性回归模型的系数。基于以上回归系数分别得到多元线性回归、岭回归及 Lasso 回归的表达式：

$$\begin{aligned}Y_1=&3.511\,3+0.017\,2X_1-0.166\,8X_2+\\&0.708\,5X_3+0.403\,5X_4+0.125\,0X_5+\\&0.021\,3X_6+0.649\,1X_7-1.299\,8X_8\end{aligned} \tag{8}$$

$$\begin{aligned}Y_2=&3.570\,6+0.023\,5X_1-0.168\,3X_2+\\&0.646\,7X_3+0.369\,6X_4+0.111\,6X_5-\\&0.057\,2X_6+0.577\,7X_7-1.226\,2X_8\end{aligned} \tag{9}$$

$$\begin{aligned}Y_3=&3.594\,9-0.161\,6X_2+0.629\,8X_3+0.373\,3X_4+\\&0.052\,4X_5+0.580\,4X_7-1.258\,2X_8\end{aligned} \tag{10}$$

利用上述回归模型，分别在测试集上进行预测后，采用均方根误差 RMSE 对模型的预测效果进行衡量，三种回归的 RMSE 值如表 4 所示。

表 4　Lasso 回归与线性回归及岭回归比较数据

出游类型	回归类型	截距值	X1	X2	X3	X4	X5	X6	X7	X8	RMSE
商务出差	岭回归	3.570 6	0.023 5	−0.168 3	0.646 7	0.369 6	0.111 6	−0.057 2	0.577 7	−1.226 2	1.361 8
	多元线性	3.511 3	0.017 2	−0.166 8	0.708 5	0.403 5	0.125 0	0.021 3	0.649 1	−1.299 8	1.461 5
	Lasso 回归	3.594 9	0	−0.161 6	0.629 8	0.373 3	0.052 4	0	0.580 4	−1.258 2	1.161 3
独自出行	岭回归	4.052 5	0.000 4	−0.455 3	0.382 6	0.135 0	0.326 9	−0.426 0	1.404 9	−2.892 8	1.127 2
	多元线性	4.111 5	0.000 7	−0.546 2	0.496 8	0.015 7	0.523 3	−0.471 1	1.756 3	−3.462 8	1.213 3
	Lasso 回归	4.141 9	0	−0.522 6	0.450 2	0.025 4	0.449 0	−0.425 6	1.655 6	−3.433 0	0.990 6
朋友出游	岭回归	4.557 4	−0.071 0	0.042 4	0.249 5	0.007 1	−0.000 3	−2.192 7	0.383 6	−1.332 9	1.614 7
	多元线性	4.513 2	−0.076 6	0.066 1	0.382 7	−0.015 5	0.001 2	−2.844 2	0.575 8	−1.651 3	1.518 3
	Lasso 回归	4.563 2	−0.035 9	0.032 3	0.292 5	0	0	−2.606 1	0.426 8	−1.568 1	1.314 6
亲子旅行	岭回归	2.579 3	0.796 3	0.094 4	0.390 8	0.478 7	0.527 6	0.155 3	−0.418 0	−2.231 7	1.953 9
	多元线性	1.575 3	1.652 9	0.278 8	0.785 1	0.954 4	0.668 0	0.848 2	−0.546 6	−3.115 9	1.772 1
	Lasso 回归	2.453 9	0.964 4	0	0.464 8	0.587 9	0.510 7	0	0	−2.453 8	1.636 8
情侣出游	岭回归	2.563 2	0.397 3	0.576 9	0.260 2	0.593 9	0.574 7	−1.629 9	0.610 9	−1.303 7	1.447 3
	多元线性	1.847 4	0.502 5	0.909 4	0.415 3	0.852 1	0.838 5	−1.934 6	1.217 2	−2.007 4	1.369 8
	Lasso 回归	1.938 2	0.486 5	0.874 8	0.392 2	0.829 7	0.788 1	−1.921 7	1.128 1	−1.917 9	1.253 1

从商务出差类型用户的三种回归所对应的 RMSE 值中可知使用 Lasso 回归进行测试所得到的 RMSE 值最小,这表明使用 Lasso 回归确定的特征值更接近实际特征值。对比式(8)、式(9)和式(10)发现在 X1 和 X6 两个特征中,岭回归和线性回归测试结果虽然很小,但还有其测试值,不能贸然对该特征偏好进行删除。然而在 Lasso 回归测试结果中,发现其值为零,这就更加直观地反映出总体感受和价格对于商务出差用户来讲属于相关性较小特征因素,因此根据式(10)可知在计算用户特征偏好中 X1 和 X6 两个特征因素不加以考虑。同理,对用户类型为独自出行、朋友出游、亲子旅行、情侣出游的用户进行计算分析可知 X1 为独自出行用户的相关性较小特征偏好,X4 和 X5 为朋友出游用户的相关性较小特征偏好,X2、X6 和 X7 为亲子旅行用户的相关性较小特征偏好。

在对比剩余四类出行用户的三种回归方法中的 RMSE 值后发现四组数据中运用 Lasso 回归方法进行剔除相关性较小特征值所产生的数据离散程度比岭回归及线性回归方法产生的离散程度都要小,这进一步表明使用 Lasso 回归方法进行测试产生的数据结果更接近真实情况。

分析实验数据可知,用户类型为独自出行、朋友出游、亲子旅行、情侣出游和商务出差这五类人士的特征偏好主要表现在总体感受、设备设施、餐饮、位置、交通、价格、服务和卫生这八个方面,其中:用户类型为独自出行和朋友出游以及情侣出游的用户在服务和饮食两个特征方面表现出极高的兴趣;用户类型为亲子旅行的用户最为关注的是酒店位置及入住的总体感受;商务出差的用户比较关注饮食及酒店服务。同时通过对五种类型用户在八个特征方面运用 Lasso 回归和岭回归以及线性回归的方法进行测试,可知运用 Lasso 回归方法对特征偏好进行过滤所产生的 RMSE(均方根误差)值相对较小,因此本实验应用 Lasso 方法进行特征偏好筛选是符合实验要求的。

本文根据实验结果及分析对酒店提出几点建议:酒店作为服务行业,不单单要注重客户的总体感受、餐饮服务、酒店卫生,对酒店内的设备设施进行定期检查,制定合理的住宿价格,良好的服务态度也是至关重要的。针对本文研究成果,酒店管理人员可针对不同类型的用户提供不同的服务标准。面向独自出行及情侣出游类型的顾客,酒店需提供优质的入住环境。面向朋友出游类型顾客,由于除位置和交通两类特征偏好以外其余六种均为用户关注的特征偏好,因此酒店人员可在定期检查设备设施、及时满足顾客要求、制定合理价格等方面进行优化。面向亲子旅行类型客户需提供新鲜营养的餐饮服务,同时酒店可规划出足够的

停车区域等。面向商务出差类型的顾客,酒店可为其提供安静的办公区域、舒适的入住房间等。综上,酒店管理人员可为不同类型的顾客制定不同的服务方案,有助于提高酒店的服务标准。

4 结 语

酒店在线评论反映了用户对入住酒店的真实感受,如何分析用户评论并从中挖掘用户对酒店的需求是现如今酒店竞争情报研究领域的热点问题,对酒店经营领域具有重要的商业价值。本文根据酒店用户评论的直接性和客观性,将 TF-IDF 算法、LDA 聚类算法、情感分析技术、Lasso 特征优选方法结合起来,构建基于 Lasso-LDA 的用户偏好模型。通过该模型能够客观地对不同类型用户对入住酒店的影响因素进行量化打分,确定用户特征偏好,弥补酒店经营者和酒店住户之间信息交流的延迟性。实验结果表明:针对酒店用户可应用该方法对各酒店评论进行不同维度的情感倾向分析,并以此分析该酒店各项服务标准是否满足自己的需求,最终做出合理决策。面向酒店经营人员,能够及时准确地反馈用户特征偏好程度,帮助其准确地调整酒店经营模式及设备设施建设。本文主要是利用酒店预订系统中高星级酒店的用户评价数据进行建模,使得应用该研究模型分析出的用户特征偏好更适用于高星级酒店的调查。在后续调查研究中会结合市场中低星级酒店用户评价进行改进,为不同需求的用户提供合理的住宿条件,合理分配酒店流动资源。

参 考 文 献

[1] 中国互联网络信息中心. 第 43 次中国互联网络发展状况统计报告[R]. (2019 - 2 - 28),2019.

[2] 曹增栋,罗迪维,杨炳新,等. 基于文本分析和 SEM 模型的小红书用户粘性研究[J]. 电子商务,2019(10):60 - 61.

[3] Pavlinek M, Podgorelec V. Text classification method based on self-training and LDA topic models[J]. Expert Systems with Applications,2017, 80: 83 - 93.

[4] Wang Y B, Xu W. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud[J]. Decision Support Systems,2018,105:87 - 95.

[5] Hagen L. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? [J]. Information Processing & Management,2018,54(6): 1292 - 1307.

[6] Panagiotidis T, Stengos T, Vravosinos O. On the determinants of bitcoin returns: A LASSO approach[J]. Finance Research Letters,2018, 27: 235 - 240.

[7] Wang S Z, Ji B X, Zhao J S, et al. Predicting ship fuel consumption based on LASSO regression[J]. Transportation Research Part D: Transport and Environment,2018, 65: 817 - 824.

[8] Sermpinis G, Tsoukas S, Zhang P. Modelling market implied ratings using LASSO variable selection techniques[J]. Journal of Empirical Finance,2018, 48: 19 - 35.

[9] Shi H, Liu S M, Chen J Q, et al. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure[J]. Genomics,2019, 111(6): 1839 - 1852.

[10] 崔方晓,李大成,吴军,等. 基于 Lasso 方法的污染气体自适应探测算法[J]. 光学学报,2019,39(5):406 - 414.

[11] 王国长,梁焙婷,王金枝. 改进的自适应 Lasso 方法在股票市场中的应用[J]. 数理统计与管理,2019,38(4): 750 - 760.

[12] Takuma K, Yamamoto J, Kamei S, et al. A hotel recommendation system based on reviews: What do you attach importance to? [C]//2016 Fourth International Symposium on Computing and Networking, 2017.

[13] Chambua J, Niu Z D, Zhu Y F. User preferences prediction approach based on embedded deep summaries[J]. Expert Systems with Applications,2019,132: 87 - 98.

[14] 雷震,阚伊戎,孙正宝,等. 评价数据中的用户偏好建模:一种基于隐变量模型的方法[J]. 云南大学学报(自然科学版),2019,41(4):669 - 677.

[15] 潘良辰,吴鑫然,岳昆. 基于深度信念网和隐变量模型的用户偏好建模[J]. 计算机工程,2020,46(5):54 - 62.

(上接第 18 页)

[13] 杨慧炯,韩燕丽. 基于 B/S/S 架构的主-辅存储模式异构数据库集成系统设计[J]. 计算机工程与设计,2008,29(11):2987 - 2988,2991.

[14] 吴程熙. 混合型分布式存储系统中的存储策略研究与实现[D]. 南京:东南大学,2017.

[15] 李忠权,冷小鹏,梁军. 基于 SOA 的地质灾害实时监测预警平台设计[J]. 成都理工大学学报(自然科学版),2018,45(5):606 - 614.

[16] 彭文启. 河湖健康评估指标、标准与方法研究[J]. 中国水利水电科学研究院学报,2018,16(5):394 - 404,416.

[17] 魏明生,童敏明,訾斌,等. 基于粒子群-拟牛顿混合算法的管道机器人定位[J]. 仪器仪表学报,2012,33(11): 2594 - 2600.

[18] 徐静,张鹏,严华. 基于开源软件的山洪灾害监测预警系统设计[J]. 水电能源科学,2014,32(2):171 - 174.