

基于在线评论的客户偏好趋势挖掘

沈超^{1,2,3}, 王安宁^{1,3}, 陆效农^{1,3}, 彭张林^{1,3}, 张强^{1,3}

(1. 合肥工业大学管理学院, 安徽 合肥 230009;

2. 安徽工程大学经济与管理学院, 安徽 芜湖 241000;

3. 过程优化与智能决策教育部重点实验室, 安徽 合肥 230009)

摘要: 为了有效地从在线评论数据中获取客户的需求偏好, 提出了一种客户偏好趋势挖掘方法. 该方法采用时间序列模型预测下一阶段产品属性重要性, 利用决策树模型分析客户偏好的变化趋势, 将产品属性分为关键属性和非关键属性. 并进一步, 根据 Mann-Kendall 趋势检验将非关键属性分为过时属性、增值属性和稳定属性. 此外, 以汽车产品为案例, 验证了该方法在产品设计与开发过程中起到的重要作用. 研究结果可以为企业的汽车产品开发提供决策支持, 从而使产品最大化地满足客户的需求.

关键词: 在线评论; 偏好趋势; 产品属性; 时间序列分析; 决策树模型

中图分类号: TP273

文献标识码: A

文章编号: 1000-5781(2021)03-0289-13

doi: 10.13383/j.cnki.jse.2021.03.001

Customer preference trend mining based on online reviews

Shen Chao^{1,2,3}, Wang Anning^{1,3}, Lu Xiaonong^{1,3}, Peng Zhanglin^{1,3}, Zhang Qiang^{1,3}

(1. School of Management, Hefei University of Technology, Hefei 230009, China;

2. School of Economics and Management, Anhui Polytechnic University, Wuhu 241000, China;

3. Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009, China)

Abstract: In order to effectively acquire customer demand preferences from online reviews data, this paper proposes a customer preference trend mining method. The time series model is used to predict the importance of product attributes of the next stage, and decision tree model is built to analyze the changing trend of customer preferences, by dividing product attributes into key attributes and non-critical attributes. Further, according to the Mann-Kendall trend test, non-critical attributes are classified into obsolete attributes, value-added attributes, and stable attributes. In addition, the automotive product is taken as an example to verify the important role of this method in product design and development. The research results can provide decision support for the development of automobile products, so that products can best meet customer needs.

Key words: online reviews; preference trends; product attributes; time series analysis; decision tree model

1 引言

随着科学技术的发展和物质生活的丰富, 消费者的需求变化越来越快. 企业能否及时把握市场需求变化是决定产品成功的关键. 产品设计者通常使用一些需求模型来预测客户偏好, 以确定下一代产品应该包含

收稿日期: 2018-10-22; 修订日期: 2019-09-30.

基金项目: 国家自然科学基金资助项目(71690235; 72071060; 71901086); 安徽省科技重大专项项目(18030901028); 安徽省自然科学基金项目(2008085QG336).

和排除的产品属性. 对于高度波动的市场, 传统的需求建模技术运用单次获取的调研数据生成预测模型, 往往无法反映产品趋势的演变过程.

在线评论作为一种新的口碑形式, 包含了消费者对产品的全方面评价, 有助于企业了解客户的偏好. 同时, 在线评论的收集成本低、更新速度快, 对制造企业及时准确地获取客户偏好变化趋势具有重要价值.

客户的需求往往是动态的, 获取客户需求及其变化趋势是一项复杂的工作. 为了及时、准确地从在线评论中获取客户需求的变化, 本文提出了一种客户偏好趋势挖掘方法. 首先, 利用信息增益方法测量产品属性对客户满意的重要性, 并基于时间序列分析预测下一时间节点产品属性的重要性, 实现产品属性重要性的测量与预测. 然后, 基于产品属性重要性的测量, 采用决策树模型识别出随时间变化的关键产品属性. 最后, 结合产品属性重要性的预测结果, 运用 Mann-Kendall 统计趋势识别技术, 依据产品属性重要性的变化趋势将非关键产品属性分成三类: 过时属性、增值属性和稳定属性. 通过一个汽车产品的示例验证了本文所提出模型的有效性. 研究表明, 本文提出的产品属性分类有助于指导产品架构, 支持在下一代产品设计中何时包含或排除某些产品功能的决策; 所建立的客户偏好趋势挖掘模型在 $k+1$ 阶段样本分类准确性预测上要优于单纯的决策树模型, 在 $k+2$ 阶段的预测效果更加明显. 本文的研究结果将对新产品开发过程中产品设计者制定产品开发策略起到辅助决策作用.

2 相关研究

2.1 决策树分类模型

在产品设计中, 经常会应用到一些需求建模和偏好测量的方法, 例如质量功能展开、联合分析、离散选择模型和监督机器学习模型. 本节只讨论决策树分类模型, 它在产品设计领域的应用比较普遍^[1].

决策树代表的是对象属性与对象值之间的一种映射关系. 产品设计领域常运用 C4.5 等决策树分类算法分析大规模消费者数据用于产品概念生成. 决策树作为一种预测模型, 算法的基础是经典信息熵理论. 在产品设计术语中, 信息熵可以表示为在数据集 S 中, 区分一个类别和另一个类别的不确定性, 可以表示为

$$\text{Entropy}(S) = - \sum_{r=1}^k p(c_r) \log_2 p(c_r), \quad (1)$$

其中 $p(c_r)$ 表示数据集 S 中的类变量 c_r 的概率, k 表示类变量值的个数.

为了确定最大能力的属性, 减少选择集的不确定性, 根据属性变量的取值划分为 n 个子数据集. 给定一个特定的属性 a , 信息熵是该属性的每个唯一值的信息熵的总和为

$$\text{Entropy}_a(S) = \sum_{j=1}^n \frac{|S_j|}{|S|} \text{Entropy}(S_j), \quad (2)$$

其中 S_j 表示训练数据 S 的子集, 包含属性的互斥结果值, $|\cdot|$ 表示集合所包含的元素个数. 例如, 属性有三个互斥结果值(高, 中和低), 则训练集 S 将被划分为三个数据子集, S_1 将包含属性值为低的所有数据实例.

ID3 决策树分类算法使用信息增益作为属性选择的度量, 实质上是属性提供的类变量不确定性的减少量. 属性 a 的 $\text{Entropy}_a(S)$ 越低, 其增益 $\text{gain}(a)$ 越高, 这两者的关系可表示为

$$\text{gain}(a) = \text{Entropy}(S) - \text{Entropy}_a(S). \quad (3)$$

C4.5 算法是继 ID3 算法后的决策树分类算法, 使用增益率进行度量, 克服了包含更多互斥结果属性带来的偏差. 增益率可以定义为^[2]

$$\text{GainRatio}(a) = -\text{gain}(a) \left(\sum_{j=1}^n \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|} \right)^{-1}. \quad (4)$$

决策树分类模型克服了离散选择模型对高维数据进行分析时通常会出现的多重共线问题, 更加适合大

规模数据分析. 此外, 由于决策树分类模型和离散选择模型都是处理静态数据, 单纯的决策树分类模型限制了分析客户偏好变化趋势的能力. 因此, 需要对决策树模型进行优化, 从而提高模型分析偏好趋势的有效性.

2.2 时间序列模型

为了克服静态需求模型的挑战, 在效用理论、数据挖掘和机器学习研究中, 时间序列建模技术应运而生. 现有文献针对量化客户偏好演变, 构建了时间序列效用函数模型. Mela 等^[3]调查了营销活动对消费者选择行为的短期、中期和长期影响, 并在多项 Logit 模型中选择一阶导数信息来量化客户偏好的时间敏感性. Jedidi 等^[4]提出了一个异方差、可变参数联合概率的回归模型, 研究了营销领域中促销和广告之间的权衡. Seetharaman^[5]提出了一种效用理论的品牌选择模型, 该模型考虑了四种不同的状态依赖性来源, 包括消费者选择和营销变量的滞后效应. Lachaab 等^[6]在选择模型中结合参数驱动的偏好动力学, 提出了贝叶斯状态空间框架, 建立了时间离散选择模型.

上述模型试图模拟不断变化的客户偏好, 但是它们主要关注模型参数的变化, 而不是属性-类变量关系的潜在演化. 此外, 这些时间序列模型不能为设计者提供下一代产品设计属性相关性的度量. 由于所提出的基于时间序列效用是面向营销领域的, 它们更关注于客户偏好的经济影响, 比如品牌偏好的演变, 广告影响等. 但是, 设计工程师面临的挑战是确定最佳属性组合, 以便产品架构设计能够满足不断变化的客户偏好.

除此之外, 上述时间序列挖掘算法没有表明哪些属性可能随时间推移表现出较弱预测能力. 因此, 必须建立出一种趋势预测模型使设计工程师能够及时有效地了解不断变化的客户偏好.

2.3 在线评论研究

在线评论作为一种新的口碑形式, 引起了信息系统、市场营销以及产品设计等领域研究者的极大关注^[7,8]. 对于产品设计研究而言, 在线评论是消费者主动发表的客户声音, 而非被动的问答, 对企业准确地获取客户需求和偏好具有十分重要的价值.

从在线评论中获取客户偏好, 首要任务是从评论文本中提取出重要的客户声音信息, 主要包括客户讨论的产品属性及其情感态度^[9,10]. 在线评论中提及的产品属性通常为名词及名词短语, 因此频繁出现的名词及名词短语通常被认为是潜在的产品属性^[11,12]. 基于这种考虑, 学者们通常会采用 POS 词性标注方法从评论文本中提取出产品属性. 对于产品属性的情感极性, 目前文献主要是采用情感词典方法和机器学习方法. 情感词典方法主要利用大量标注样本训练生成情感词典, 情感词典中包含了正面情感和负面情感的词语, 通过情感词典识别语句中带有情感极性的词语, 并以此判定产品属性的情感态度.

许多学者结合从评论文本中提取的客户声音和消费者的打分、购买等数据, 利用联合分析、离散选择模型等偏好模型, 分析产品属性对客户满意、产品价格或产品销量的影响机制^[13,14]. 刘晓君等^[15]运用复杂网络方法和隐含狄利克雷分配模型, 探讨了在线评论的相互关系及整体演化发展. 本文主要是采用决策树模型分析客户偏好的时序变化, 将产品属性分为关键属性和非关键属性. 对于非关键属性, 根据重要性变化趋势分为三类: 过时属性、增值属性和稳定属性. 这种产品属性分类, 有助于指导制造企业的产品架构设计, 以及辅助产品资金投入方案决策.

2.4 客户偏好趋势分析

客户偏好反映了客户习惯于消费某种产品的心理行为, 提升企业产品和服务价值的关键所在就是不断满足客户的需求偏好, 只有满足了不同客户的需求, 才能实现产品和服务的价值最大化. 客户偏好的研究还是企业进行市场细分的基础, 企业通过不同客户群的偏好从而合理的选择营销策略和销售管理方案.

目前, 相关文献主要基于营销理论和市场细分过程, 采用传统的统计调查和实证分析等方法分析客户偏好. 例如, Das 等^[16]通过多元统计技术对反馈调查获得的信息进行处理, 以获得顾客对鞋业零售市场中的特征偏好. Mihelis 等^[17]对民营银行客户满意度进行调查, 并基于多准则分析和偏好分解模型挖掘具有独特偏好和期望的客户集群. Kontot 等^[18]针对影响客户对伊斯兰银行存款偏好的因素进行面对面访谈, 并制定针对性的调查问卷, 从而发现影响顾客决策偏好的决定因素. Cao 等^[19]开发了偏好分类法对产品属性进行分

类, 实现将客户偏好映射到产品领域本体, 从而驱动高级概念生成, 并且通过有关手机产品的实证研究证明了客户偏好驱动的概念生成方法的有效性。

随着计算机技术的发展, 基于 Web 日志挖掘、网页数据挖掘逐渐成为获取客户偏好的主要方法. Zhou 等^[20]基于用户生成的在线产品评论的情绪分析, 来预测客户偏好的趋势, 并通过平板电脑的应用案例验证了所提方法的有效性. Zhang 等^[21]提出了一种基于在线产品购买数据的客户偏好预测方法, 并以玩具车的特征表征和顾客偏好预测为例说明了所提出的方法的有效性. Milovanović 等^[22]建立了一种基于社交网络分析(SNA)的客户偏好识别方法, 获取的客户偏好用于开发推荐系统以及营销策略的定制. Ye 等^[23]从时间和空间上对旅客的住宿偏好进行分析, 并运用酒店的在线评论挖掘来自不同地区旅客的季节性需求偏好. 此外, 研究者还通过建立智能算法系统实现客户偏好的收集. 例如, Li 等^[24]基于AP-KNN 算法开发了一个软件原型系统, 实现了客户反馈收集、客户偏好推理以及动机分析. Ma 等^[25]提出了连续偏好趋势挖掘(CPTM)算法以从交易数据中捕获客户购买模式的隐藏趋势。

从前期对客户需求偏好研究的文献来看, 客户偏好研究的数据获取经历了从传统调查到网络数据的转变, 相对应的研究方法也从实证研究发展到网络数据挖掘. 在当前研究的基础上, 本文基于在线评论数据, 运用决策树模型和趋势检验模型对产品属性进行分类, 分析客户偏好变化趋势。

3 偏好趋势挖掘模型

图 1 介绍了偏好趋势挖掘模型的总体流程, 主要包括产品属性重要性测量和预测, 关键产品属性识别以及非关键产品属性分类. 通过这 3 个步骤, 可以预测产品每个属性的趋势模式是单调增加, 单调减少或是季节性趋势; 选择相关性(可以用信息增益度量)最高的属性作为关键属性, 运用决策树模型将客户对产品的打分进行分类; 而对于非关键属性, 运用 Mann-Kendall 方法将其分为过时属性、增值属性和稳定属性三个类别. 这一过程获得的结果可以为制造企业在开发产品时应该加入何种功能提供参考。

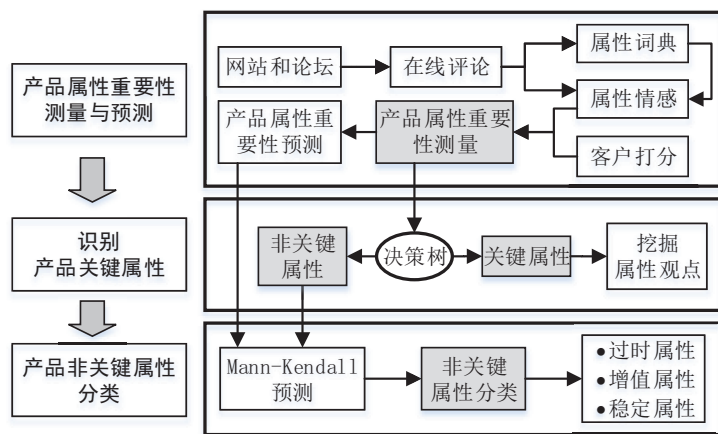


图 1 偏好趋势挖掘模型框架

Fig. 1 The framework of preference trend mining model

3.1 产品属性重要性测量与预测

在线评论数据通常包含评论文本和客户对产品的打分. 根据收集的大规模在线评论, 本文首先采用 POS 词性标记方法提取潜在产品属性词语^[26], 然后通过人工剔除其中常见的非属性词语, 并对剩余潜在属性词语进行同义词合并, 生成产品属性词典. 依据生成的属性词典, 可以识别每条评论提及的产品属性. 若属性出现在满意评论中, 则属性情感为正面; 若属性出现在不满意评论中, 则属性情感为负面^[27,28].

此外, 本文将客户对产品的打分作为类变量, 视为客户满意程度. 为方便计算, 可以将客户对产品的打分

划分为高, 中, 低三个类别. 结合属性情感(正面, 负面和未出现)以及客户对产品的打分(高, 中, 低), 可以运用信息增益方法计算每个产品属性对于客户满意度的影响大小. 因此, 产品属性 a 的信息熵为

$$\text{Entropy}_a(S) = \frac{|S^+|}{|S|} \text{Entropy}(S^+) + \frac{|S^-|}{|S|} \text{Entropy}(S^-) + \frac{|S^0|}{|S|} \text{Entropy}(S^0), \quad (5)$$

其中 S^+ , S^- 和 S^0 分别表示产品属性 a 为正面、负面以及未出现的评论.

结合式(3)可以计算每个产品属性的信息增益. 将处理好的数据集划分为 n 节时间段, 分别计算每节时间段产品属性的重要性, 并采用 Holt-Winters 模型, 从而实现产品属性重要性的测量和预测.

以属性 1 和属性 2 为例, 属性 1 和属性 2 的正负面评论数量及其信息增益如图 2 所示. 可以看出属性 1 负面评论数量较多, 且呈现增长趋势. 而属性 2 正面评论呈现增长趋势, 且远多于负面评论. 同时, 属性 1 重要性在后面几个时间阶段呈现缓慢下降趋势, 而属性 2 重要性却呈现不断上升趋势.

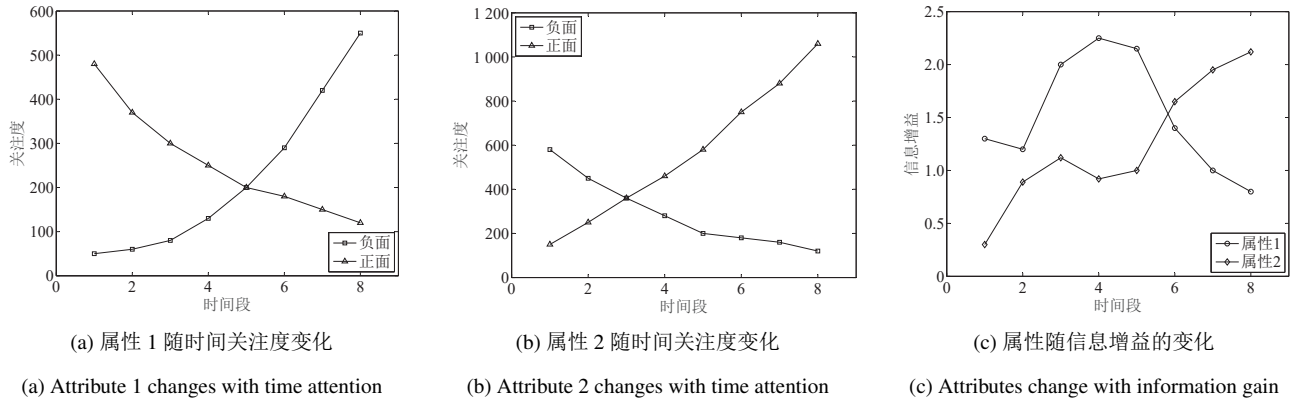


图 2 产品属性情感和信息增益变化趋势

Fig. 2 The trends of product attribute sentiment and information gain

从图 2(c)还可以发现, 在时间段 6 之后, 属性 2 的重要性开始超过属性 1. 在时间段 6 之前生成的产品属性重要性测量中, 将无法发现这一新兴客户偏好趋势. 为了克服这项挑战, 本文采用 Holt-Winters 指数平滑模型预测下一阶段的客户偏好, 运用加权平均技术, 同时考虑时间序列中数据趋势和季节性成分. Holt-Winters 是一种非参数模型, 可用于将具有线性趋势、季节变动和随机变动的时间序列进行分解, 并结合指数平滑法对属性重要性进行第 k 步预测, 有助于在市场空间中对新兴的偏好趋势进行判断. 一般把时间序列的系统性部分分为水平、趋势和季节性三种成分^[29]. 针对这三种成分, 分别对长期趋势、趋势增量和季节变动做出估计, k 步提前预测模型为

$$y_t(k) = L_t + kT_t + I_{t-s+k}, \quad (6)$$

式(6)中, 水平成分 L_t 可表示为

$$L_t = \alpha(y_t - I_{t-s}) + (1 - \alpha)(L_{t-1} + T_{t-1}), \quad (7)$$

趋势成分 T_t 可表示为

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}, \quad (8)$$

季节性成分 I_t 可表示为

$$I_t = \delta(y_t - L_t) + (1 - \delta)I_{t-s}, \quad (9)$$

其中 y_t 表示近期时间段 t 时刻的数据点, $y_t(k)$ 表示超过 y_t 的第 k 节时间段的预测值, 有 $y_t(k) = y_{t+k}$, s

表示季节性频率. 平滑参数 α , γ 和 δ 都在 $[0, 1]$ 范围内, 并通过最小化前一个时间段步长的误差平方和来估计^[30].

3.2 关键产品属性识别

关键产品属性识别是需求预测模型的主要挑战之一, 本文采用决策树分类模型帮助制造企业理解关键产品属性. 为便于理解, 本文提供以下示例用于理解关键产品属性识别过程. 假定汽车产品的属性集合为{外观, 内饰, 动力, 油耗, 操控, 天窗}, 所有产品属性有正面、负面和未出现三个属性值. 客户对产品的打分被定义为依赖于多个属性情感的类变量, 取值集合为{高, 中, 低}.

对于某一时间段的关键产品属性识别, 图3给出了决策树模型的直观表示. 决策树模型一般包括收集和准备数据、选取划分算法、构造决策树、测试和使用算法等步骤. 通常, 决策树模型依据信息增益或 Gini 指数等算法进行数据集划分. 本文采用信息增益作为划分数据集的依据. 首先, 依据式(3)和式(5)计算每个产品属性对应的信息增益; 然后, 挑选信息增益最大的产品属性划分数据集. 通过决策树模型, 验证了四个关键产品属性“动力”、“内饰”、“油耗”和“外观”, 同时了解到这四个关键产品属性的分类能力是依次减弱的.

为了给设计工程师提供细粒度的客户需求, 本文采用点互信息(PMI)来挖掘每个时间段的属性观点^[31], 例如“外观-优雅”、“油耗-较高”等. PMI 可以用于衡量两个变量之间的相关性,

$$\text{PMI}(a, o) = \log_2 \frac{p(a, o)}{p(a)p(o)}, \quad (10)$$

其中 $p(a, o)$ 表示产品属性与属性观点 o 共同出现的概率, $p(a)$ 表示产品属性 a 出现的概率, $p(o)$ 表示观点 o 出现的概率.

根据 PMI 值的大小, 可以从评论文本中识别出每个产品属性的客户观点. 关键产品属性及其观点随着时间的推移也在不断变迁. 其演变趋势有助于制造企业加深对客户需求和偏好的了解. 为了便于理解, 本节给出一个示例, 如图4所示. 可以发现, 在 T_2 时刻, “配置”成为关键属性, 并且其受关注程度到 T_3 时刻逐步增大, 此外, 随时间变化, 关键属性“外观”的关注度在逐渐增大. 对于属性观点, 消费者越来越厌恶大众风格的“内饰”, 越来越偏好环保的“外观”.

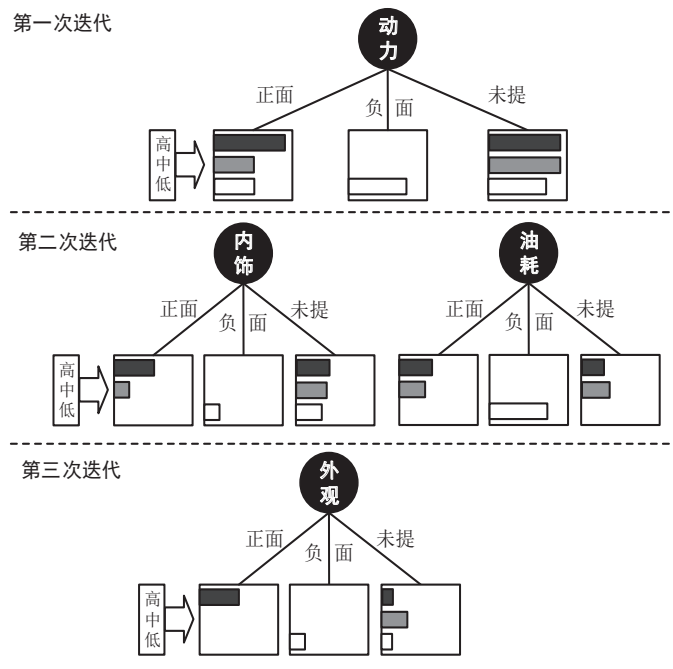


图3 决策树分类模型示例

Fig. 3 The example of decision tree classification model

此外, 可以发现产品属性“操控”和“天窗”并不在决策树模型分类规则中, 本文将其称为非关键产品属性. 对于非关键产品属性, 存在两点疑问: 1) 是否存在一些产品属性没有在决策树分类规则中, 但是在产品设计中设计工程师和制造企业十分重要? 2) 非关键产品属性在未来的产品设计中是否会转变成关键产品属性? 这些疑问提示设计工程师不能忽略非关键产品属性, 一旦忽略非关键产品属性将无法完整地获取客户需求, 从而容易导致产品开发的失败. 对于非关键产品属性, 将在 3.3 节予以详细的探讨.

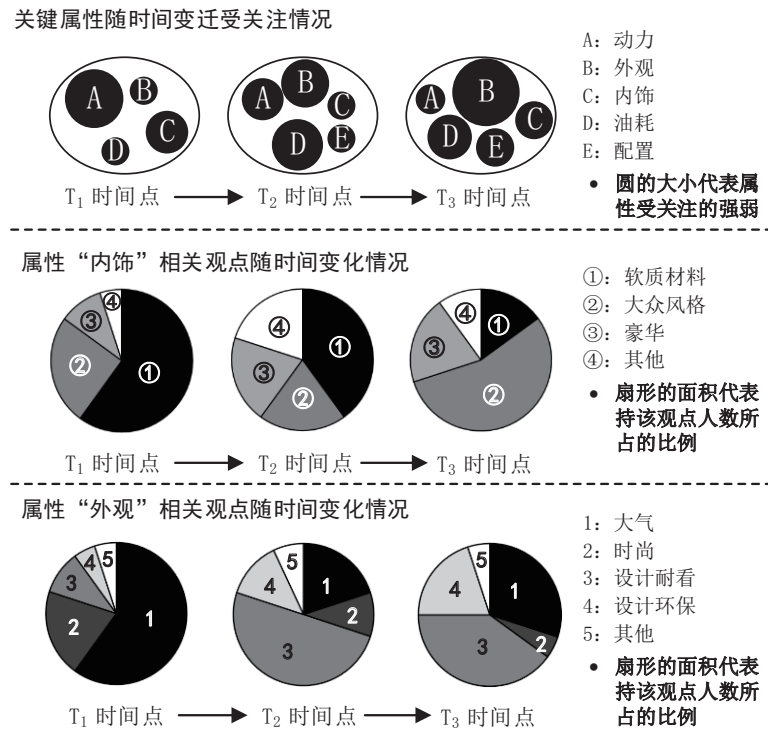


图4 关键产品属性及其观点变迁的示例

Fig. 4 An example of key product attributes and their views change

3.3 非关键产品属性分类

产品属性对消费者购买决策具有至关重要的影响, 关系到制造企业新产品开发策略是否能满足市场的需求. 在新产品开发策略过程中, 非关键产品属性不会显著影响客户满意程度. 但是, 为了避免忽略一些重要产品属性, 对于非关键产品属性, 采用 Mann-Kendall 趋势检测, 观察每个产品属性重要性是否表现出单调减少或单调增加趋势, 也就是产品属性的未来趋势是失去满足客户的能力还是获取不断满足客户的能力, 从而将产品属性确定为过时属性或增值属性. 对于检测过程中产品属性重要性没有表现出明显单调减少或者单调增加趋势, 比如, 表现出在某一范围内周期性波动趋势, 或者是平稳趋势, 将这类产品属性统一归为稳定属性. 这样, 所有的产品属性根据其重要性随时间变化, 都可以分成确定类别, 为产品开发工程师提供一定的参考.

1) 过时属性. 在给定的时间段 t_1 到 t_n 过程中, 如果在迭代过程中认为属性 a 是非关键的, 并且随着时间推移不会系统地影响客户满意度, 则属性 a 被定义为过时属性. 通过 Mann-Kendall 趋势检测, 如果属性 a 的重要性表现出单调减少趋势, 表明属性 a 随时间推移不断失去满足客户偏好的能力. 如果产品属性在给定的时间序列结束时属于该分类, 则在下一代产品设计中忽略该产品属性.

2) 增值属性. 给定的时间段 t_1 到 t_n 过程中, 如果在迭代过程中认为属性 a 是非关键的, 并且随着时间推移会系统地影响客户满意度, 则属性 a 被定义为增值属性. 通过 Mann-Kendall 趋势检测, 如果属性 a 的重要性表现出单调增加趋势, 表明属性 a 随时间推移始终获得满足客户偏好的能力. 如果产品属性在给定时

间序列结束时属于此分类, 尽管表征出一定的非关键性, 也应该视为对产品设计至关重要的产品属性。

3) 稳定属性. 给定的时间段 t_1 到 t_n 过程中, 如果在迭代过程中认为属性 a 是非关键的, 并且随着时间推移无法确定与客户满意度的可辨别关系, 则属性 a 被定义为稳定属性. 通过 Mann-Kendall 趋势检测, 稳定属性的重要性没有单调增加或减少的趋势. 在产品设计中表现出这种情况的产品属性可能是消费者尚未完全意识到的新颖属性. 这类产品属性不应该被忽略, 经过消费者的口碑传播最终可能成为增值属性。

本文使用非参数 Mann-Kendall 方法来检测迭代过程中被认为是非关键产品属性的重要性变化趋势. Mann-Kendall 检验^[32]方法不要求被分析样本遵从一定分布, 同时也不受异常值的干扰. 由于计算简便, 众多学者将其应用于分析关键要素的时间序列变化趋势, Mann-Kendall 测试的统计量^[33]为

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_j - x_i), \quad (11)$$

其中 n 表示时间序列数据点的总数, x_j 代表前一时刻数据点, x_i 代表当前数据点。

$$\text{sgn}(\theta) = \begin{cases} 1, & \theta > 0 \\ 0, & \theta = 0 \\ -1, & \theta < 0. \end{cases} \quad (12)$$

按下列方式标准化统计量 S , 即

$$Z = \begin{cases} (S - 1) (\text{Var}(S))^{-\frac{1}{2}}, & S > 0 \\ 0, & S = 0 \\ (S + 1) (\text{Var}(S))^{-\frac{1}{2}}, & S < 0. \end{cases} \quad (13)$$

统计量 Z 服从标准正态分布. 如果 p -值小于显著性水平 ($\alpha = 0.05$), 表明存在变化趋势, Z 为负值, 则为增值属性, Z 为正值, 则为过时属性. 如果 p -值大于显著性水平, 则为稳定属性. 企业决策者希望了解特定产品的整体市场需求, 以便可以做出潜在产品的开发决策, 基于时间序列趋势, 利用 Holt-Winters 方法可以估计产品属性在下一阶段的重要性水平, 并结合决策树分类模型和 Mann-Kendall 检验, 可以为企业的产品设计与开发提供辅助决策。

4 案例研究

4.1 数据收集与预处理

汽车作为重要的消费产品, 在汽车论坛上有着大量的客户评论. 因此, 本文以汽车产品为研究对象, 在易车网、太平洋汽车网上收集了 12 款国产汽车 SUV 品牌的所有客户评论, 数据内容包括满意部分、不满意部分、客户对产品的打分和评论时间. 考虑到产品的更新周期以及收集的数据集的大小, 即汽车公司的汽车产品一年更新 1 次到 2 次, 收集的评论数据集要能保证实验结果中的趋势变化较为明显, 本实验采用 3 年的在线评论数据, 共计有效评论 16 388 条, 从 2014 年 7 月到 2017 年 6 月, 分为 12 个阶段, 每个阶段代表 1 个季度(见图 7 和图 9)。

基于现有产品属性识别方法, 本案例研究识别出 24 个汽车产品属性. 根据产品属性在满意评论和不满评论中出现的频次, 得到产品属性正负面情感的分布, 如图 5 所示. 可以看出“外观”、“配置”和“价格”等产品属性的正面情感远多于负面情感; 相反, “用料”、“隔音”和“后备箱”等产品属性的负面情感远多于正面情感; 此外, “油耗”、“动力”和“中控”等产品属性的正负面情感相差不大. 同时, 可以得到客户打分的分布, 如图 6 所示. 由于 5 分和 4 分的评论远多于其他打分, 为确保类变量分布的均衡, 本案例研究将 5 分设定为高分, 4 分设定为中等分, 3 分及以下打分设定为低分。

4.2 结果与讨论

本文提出的基于在线评论的客户偏好趋势挖掘方法, 首先利用 Holt-Winters 方法预测下一阶段产品属性重要性, 然后结合决策树分类模型和 Mann-Kendall 检验分析客户偏好变化趋势. 因此, 首先展示产品属性重要性随时间变化趋势, 并检验 Holt-Winters 方法预测的效果. 图 7 展示了“动力”、“内饰”、“外观”、“油耗”和“导航”属性在 12 个阶段的变化趋势以及第 13 个阶段的预测值. 可以看出“动力”和“内饰”呈现增长趋势, “导航”呈现降低趋势, “油耗”呈现周期性变化, “外观”基本保持稳定趋势. 此外, 可以观察到“动力”在第 12 阶段开始高于“外观”, “内饰”在第 13 阶段开始高于“外观”, 与 Holt-Winters 模型的预测相吻合.

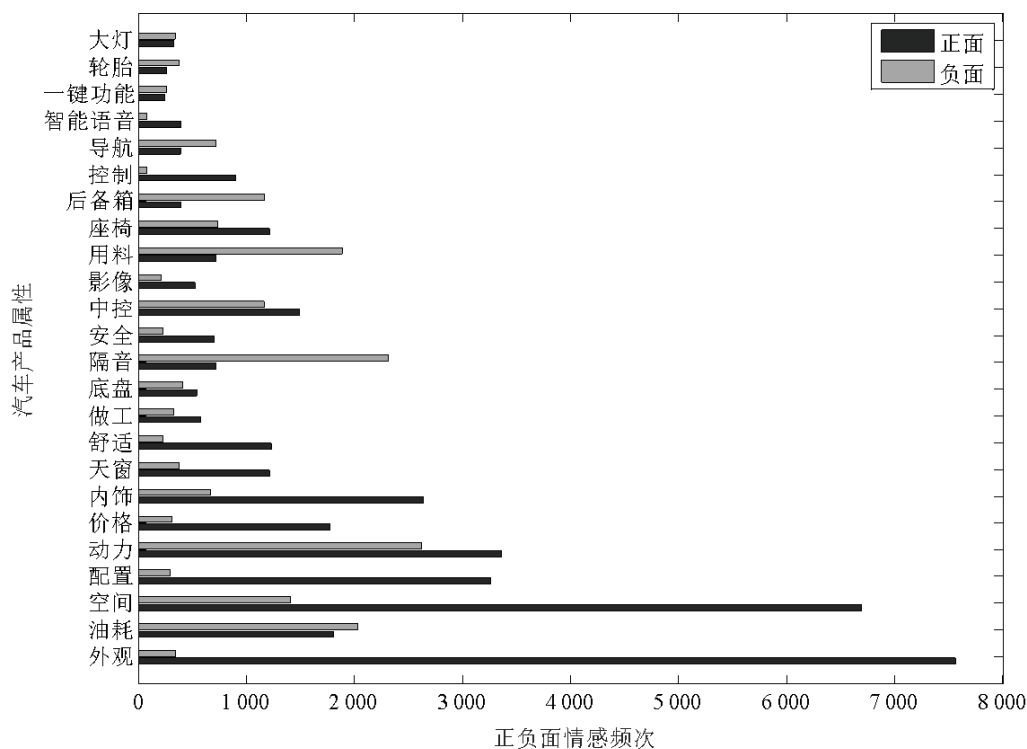


图 5 产品属性的正负面情感分布

Fig. 5 Positive and negative emotional distribution of product attributes

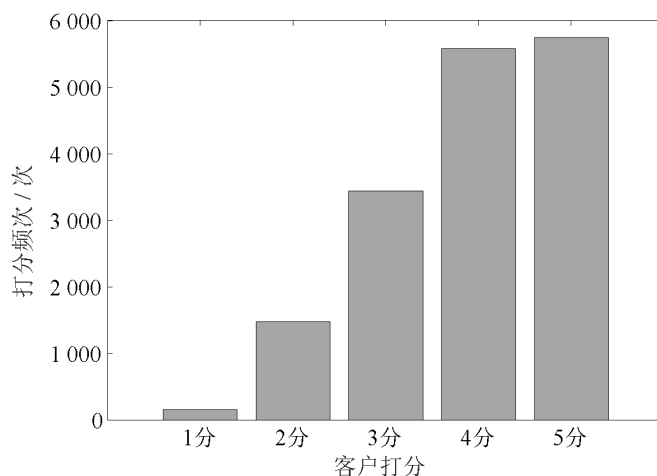


图 6 客户打分的分布

Fig. 6 The distribution of customers' score

为了展示决策树分类模型和 Mann-Kendall 检验的效果, 本节利用最新数据集(第 13 时间阶段)生成的

预测结果,如图 8 所示. 可以看出关键产品属性依次为“动力”、“内饰”、“外观”、“价格”、“油耗”、“空间”、“用料”和“配置”. 对于不在决策树分类规则中的非关键产品属性, 设计工程师并不能直接忽略. 非关键产品属性仅仅意味着信息增益不够高, 不代表对客户满意的影响可以忽略. 本文将这些非关键产品属性分为过时属性、增值属性或稳定属性, Mann Kendall 趋势检验结果如表 1 所示, 结果表明“智能语言”、“中控”、“隔音”、“做工”、“天窗”、“影像”和“后备箱”表征为增值属性; “大灯”、“轮胎”、“一键功能”、“座椅”、“安全”、“底盘”、“舒适”和“控制”表征为稳定属性; “导航”表征为过时属性.

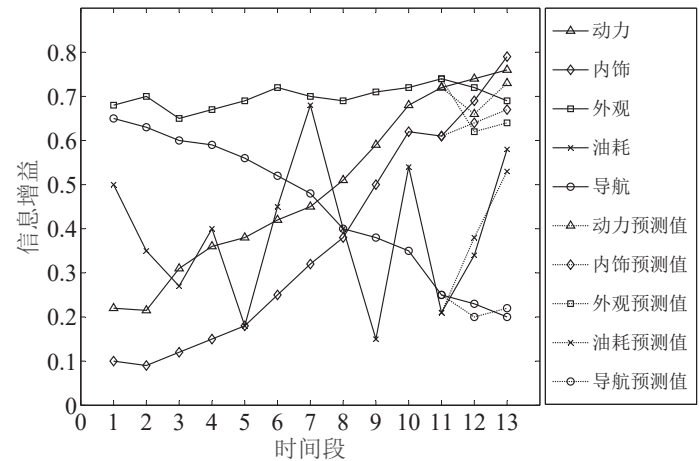


图 7 不同时间段各属性信息增益变化

Fig. 7 The change of information gain of each attribute in different time periods

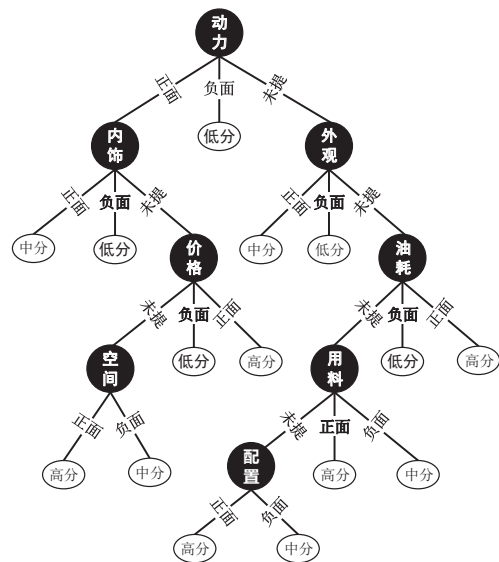


图 8 第 13 时间阶段生成的决策树

Fig. 8 Decision tree generated in the 13th time period

表 1 非关键属性趋势检验结果

Table 1 The trend test result of Non-critical attribute			
属性	统计量 Z	p -值	类别
智能语言	-0.036	0.032	增值属性
中控	-0.047	0.028	增值属性
隔音	-0.052	0.018	增值属性
做工	-0.063	0.046	增值属性
天窗	-0.065	0.035	增值属性

续表 1

Table 1 Continues

属性	统计量 Z	p-值	类别
影像	-0.067	0.028	增值属性
后备箱	-0.046	0.019	增值属性
大灯	—	0.135	稳定属性
轮胎	—	0.158	稳定属性
一键功能	—	0.156	稳定属性
座椅	—	0.089	稳定属性
安全	—	0.096	稳定属性
底盘	—	0.157	稳定属性
舒适	—	0.074	稳定属性
控制	—	0.134	稳定属性
导航	—	0.038	过时属性

4.3 模型验证

由于产品开发需要一定的周期, 因此客户偏好模型必须具备能够预测未来一段时间客户需求的能力. 本节将对决策树模型和本文提出的客户偏好趋势模型. 单纯的决策树模型主要是根据最新数据集生成的分类规则来判别未来阶段的客户满意度. 本文的客户偏好模型可以利用时间序列数据, 先剔除非关键属性中的过时属性和稳定属性, 然后根据最近数据集生成的决策树分类规则判别未来阶段的客户满意度. 举例来说, 当前是第 10 阶段, 单纯的决策树模型根据第 10 阶段的数据生成分类规则, 然后用于测试第 11 或 12 阶段样本分类的准确率. 客户偏好趋势模型根据第 1 阶段至第 9 阶段, 判别产品属性的类型, 然后选择关键属性和增值属性, 生成第 10 阶段的决策树分类规则, 并用于测试第 11 阶段或 12 阶段样本分类的准确率.

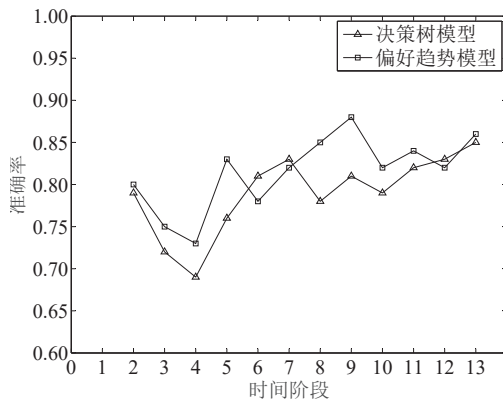
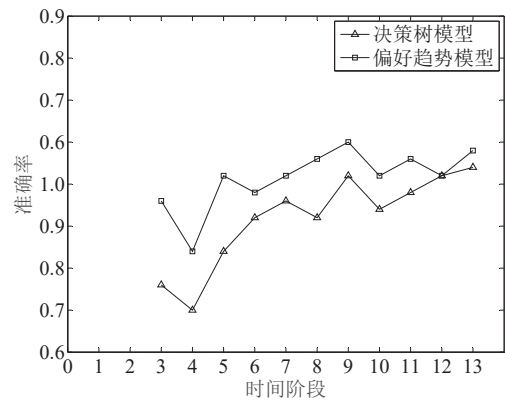
(a) $k+1$ 时间段预测准确率(a) Predicting accuracy in time period $k+1$ (b) $k+2$ 时间段预测准确率(b) Predicting accuracy in time period $k+2$

图9 偏好趋势模型和决策树模型预测准确率对比

Fig. 9 Comparison of predictive accuracy between preference trend model and decision tree model

本文研究选择预测 $k+1$ 阶段和 $k+2$ 阶段的样本分类准确率, 结果如图 9 所示. 可以发现客户偏好趋势模型明显优于单纯的决策树模型. 此外, 在 $k+2$ 阶段客户偏好趋势模型的优越性更加明显. 比较结果表明本文提出的客户偏好趋势挖掘模型能够较好地预测未来一段时间的客户需求.

5 结束语

本文主要的贡献是利用在线评论数据生成客户偏好趋势. 首先, 使用信息增益测量产品属性的重要性,

并运用时间序列模型预测产品属性重要性的变化趋势;然后,运用决策树分类模型识别随时间推移的关键产品属性;最后,运用 Mann-Kendall 统计趋势识别技术将非关键产品属性分为过时属性、增值属性和稳定属性. 本文研究可以有助于制造企业洞察市场变化,并且指导产品架构,从而使生产的产品最大化地满足市场客户的需求和偏好. 由于在线评论数据样本存在着自我选择偏差性,今后可以结合传统调研数据与在线评论数据的各自特点,从而使获得的客户偏好更加真实. 另一方面,在线评论提及的产品属性主要是客户的主观感知属性,而对一些企业设定的产品结构细节属性缺乏讨论,所以,在今后的研究工作中,可以构建客户主观属性与产品客观属性的映射关系,从而有效地利用在线评论中的客户主观信息,更加细粒度地支持企业的产品架构设计.

参考文献:

- [1] Tucker C S, Hoyle C, Kim H M, et al. A comparative study of data-intensive demand modeling techniques in relation to product design and development // ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. San Diego: American Society of Mechanical Engineers, 2009: 371–383.
- [2] Quinlan J R. Induction of decision trees. *Machine Learning*, 1986, 1(1): 81–106.
- [3] Mela C F, Gupta S, Lehmann D R. The long-term impact of promotion and advertising on consumer brand choice. *Journal of Marketing Research*, 1997, 34(2): 248–261.
- [4] Jedidi K, Mela C F, Gupta S. Managing advertising and promotion for long-run profitability. *Marketing Science*, 1999, 18(1): 1–22.
- [5] Seetharaman P B. Modeling multiple sources of state dependence in random utility models: A distributed lag approach. *Marketing Science*, 2004, 23(2): 263–271.
- [6] Lachaab M, Ansari A, Jedidi K, et al. Modeling preference evolution in discrete choice models: A Bayesian state-space approach. *Quantitative Marketing and Economics*, 2006, 4(1): 57–81.
- [7] 刘 洋, 廖貅武, 刘 莹. 在线评论对应用软件及平台定价策略的影响. *系统工程学报*, 2014, 29(4): 560–570.
Liu Y, Liao X W, Liu Y. The impact of online review on software and platform's pricing strategies. *Journal of Systems Engineering*, 2014, 29(4): 560–570. (in Chinese)
- [8] 毕建武, 刘 洋, 樊治平. 依据在线评论的商品排序方法. *系统工程学报*, 2018, 33(3): 422–432.
Bi J W, Liu Y, Fan Z P. Method for ranking products through online reviews. *Journal of Systems Engineering*, 2018, 33(3): 422–432. (in Chinese)
- [9] Lee A J T, Yang F C, Chen C H, et al. Mining perceptual maps from consumer reviews. *Decision Support Systems*, 2016, 82: 12–25.
- [10] 李良强, 袁 华, 叶 开, 等. 基于在线评论词向量表征的产品属性提取. *系统工程学报*, 2018, 33(5): 687–697.
Li L Q, Yuan H, Ye K, et al. Extraction product features from online reviews based on word-vector-representation. *Journal of Systems Engineering*, 2018, 33(5): 687–697. (in Chinese)
- [11] Hu M Q, Liu B. Mining and summarizing customer reviews // Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168–177.
- [12] Ni X, Sai Y H, Wang A N, et al. Intelligent discovery of notable product features by mining large scale online reviews // IEEE Second International Conference on Data Science in Cyberspace. Shenzhen: IEEE, 2017: 644–649.
- [13] Archak N, Ghose A, Ipeirotis P G, et al. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 2011, 57(8): 1485–1509.
- [14] Decker R, Trusov M. Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 2010, 27(4): 293–307.
- [15] 刘晓君, 那日萨, 崔雪莲. 基于隐含狄利克雷分配模型的消费者在线评论复杂网络构建及其应用. *系统工程学报*, 2017, 32(3): 305–312.
Liu X J, Na R S, Cui X L. Construction and application of the complex network about consumer online reviews based on latent Dirichlet allocation model. *Journal of Systems Engineering*, 2017, 32(3): 305–312. (in Chinese)
- [16] Das P, Mukherjee S. Modeling of customer preferences on product features and comparing the competitors' performances. *Quality Engineering*, 2007, 20(1): 53–62.
- [17] Mihelis G, Grigoroudis E, Siskos Y, et al. Customer satisfaction measurement in the private bank sector. *European Journal of Operational Research*, 2001, 130(2): 347–360.

- [18] Kontot K, Hamali J, Abdullah F. Determining factors of customers' preferences: A case of deposit products in islamic banking. *Procedia: Social and Behavioral Sciences*, 2016, 224: 167–175.
- [19] Cao D X, Li Z J, Ramani K. Ontology-based customer preference modeling for concept generation. *Advanced Engineering Informatics*, 2011, 25(2): 162–176.
- [20] Zhou F, Jiao J R, Yang X J, et al. Augmenting feature model through customer preference mining by hybrid sentiment analysis. *Expert Systems with Applications*, 2017, 89: 306–317.
- [21] Zhang J, Simeone A, Gu P, et al. Product features characterization and customers' preferences prediction based on purchasing data. *CIRP Annals*, 2018, 67(1): 149–152.
- [22] Milovanović S, Bogdanović Z, Labus A, et al. An approach to identify user preferences based on social network analysis. *Future Generation Computer Systems*, 2019, 93: 121–129.
- [23] Ye B H, Luo J M, Vu H Q. Spatial and temporal analysis of accommodation preference based on online reviews. *Journal of Destination Marketing & Management*, 2018, 9: 288–299.
- [24] Li X, Shi D M, Charatrakul V, et al. Advanced P-tree based K-nearest neighbors for customer preference reasoning analysis. *Journal of Intelligent Manufacturing*, 2009, 20(5): 569–579.
- [25] Ma J, Kim H M. Continuous preference trend mining for optimal product design with multiple profit cycles. *Journal of Mechanical Design*, 2014, 136(6): 1–14.
- [26] 李实, 叶强, 李一军, 等. 中文网络客户评论的产品特征挖掘方法研究. *管理科学学报*, 2009, 12(2): 142–152.
Li S, Ye Q, Li Y J, et al. Mining features of products from Chinese customer online reviews. *Journal of Management Sciences in China*, 2009, 12(2): 142–152. (in Chinese)
- [27] Xiao S S, Wei C P, Dong M. Crowd intelligence: Analyzing online product reviews for preference measurement. *Information & Management*, 2016, 53(2): 169–182.
- [28] Qi J Y, Zhang Z P, Jeon S, et al. Mining customer requirements from online reviews: A product improvement perspective. *Information & Management*, 2016, 53(8): 951–963.
- [29] Brockwell P J, Davis R A. *Introduction to Time Series and Forecasting*. New York: Springer, 2002.
- [30] Grubb H, Mason A. Long lead-time forecasting of UK air passengers by Holt–Winters methods with damped trend. *International Journal of Forecasting*, 2001, 17(1): 71–82.
- [31] 王伟, 王洪伟. 特征观点对购买意愿的影响: 在线评论的情感分析方法. *系统工程理论与实践*, 2016, 36(1): 63–76.
Wang W, Wang H W. The influence of aspect-based opinions on user's purchase intention using sentiment analysis of online reviews. *Systems Engineering: Theory & Practice*, 2016, 36(1): 63–76. (in Chinese)
- [32] Kendall M, Gibbons J D. *Rank Correlation Methods*. New York: Oxford University Press, 1990.
- [33] Yue S, Pilon P, Cavadias G. Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, 2002, 259(1): 254–271.

作者简介:

沈超(1990—), 男, 安徽濉溪人, 博士生, 研究方向: 产品需求开发, Email: shenchao@mail.hfut.edu.cn;

王安宁(1991—), 男, 安徽无为, 博士生, 研究方向: 产品开发与设计, Email: waning@mail.hfut.edu.cn;

陆效农(1988—), 男, 安徽合肥人, 博士, 讲师, 研究方向: 智能制造与产品开发, Email: xnl@hfut.edu.cn;

彭张林(1984—), 男, 安徽潜山人, 博士, 副教授, 研究方向: 产品众包服务设计, Email: pengzhanglin@hfut.edu.cn;

张强(1984—), 男, 四川绵阳人, 博士, 副教授, 研究方向: 产品设计与产品创新, Email: qiang_zhang@hfut.edu.cn.