基于 LDA 主题模型和偏序集的在线商品评论研究

崔 宁[1] 赵宗良[2] 吴瑞雪[2]

(1.辽宁工程技术大学公共管理与法学院 辽宁阜新 123000)

(2.辽宁工程技术大学工商管理学院 辽宁葫芦岛 125000)

摘 要 [目的/意义] 从大量在线商品评论中得到有价值的信息,为商家掌握顾客消费需求以及消费者选择商品提供依据。[方法/过程] 以 LDA 主题模型和偏序集理论为基础,构建在线商品评论分析模型,并以天猫商城空调的在线评论文本为例进行实证分析。首先,利用八爪鱼采集器爬取商品评论文本;其次,利用 LDA 主题模型从在线商品评论文本中提取出影响顾客消费的因素;最后,基于偏序集理论对调查问卷收集的数据进行商品排序。[结果/结论] 通过实证分析证明该模型可行有效,不仅能帮助商家掌握顾客消费需求和潜在倾向,也可以使消费者迅速筛选出心仪的商品。

关键词: 在线评论;LDA 主题模型;偏序集

中图分类号: C934 **文献标识码:** A **doi:** 10.3969/j.issn.1005-8095.2021.12.010

Research on Online Commodity Review Based on LDA Topic Model and Partial Order Set

Cui Ning[1] Zhao Zongliang[2] Wu Ruixue[2]

(1. School of Public Administration and Law, Liaoning Technical University, Fuxin Liaoning 123000)

(2. School of Business Administration, Liaoning Technical University, Huludao Liaoning 125000)

Abstract: [Purpose/significance] The valuable information obtained from a large number of online commodity reviews can provide a basis for merchants to grasp the consumer's consumption demands and choice of commodities. [Method/process] Based on LDA topic model and partial ordered set theory, the paper constructs an online commodity review analysis model, and takes the online review text of Tmall air conditioner as an example for empirical analysis. The paper uses octopus collector to crawl the product review text, adopts LDA topic model to extract the factors influencing customer consumption from online product review text, and ranks the commodities by the data collected from the questionnaire based on partial ordered set theory. [Result/conclusion] The empirical analysis proves that the model is feasible and effective, which can not only help merchants to grasp the consumer's consumption demands and potential tendency, but also enable consumers to quickly screen out the desired products.

Keywords: online review; LDA topic model; partial order set

0 引言

中国互联网络信息中心(CNNIC)发布的第 47 次《中国互联网络发展状况统计报告》显示,截至 2020 年 12 月,中国网购注册用户数量已达 7.82 亿户,占网民整体数量的 79.1%,而且网购平台商品零售额高达 11.76 万亿元[1]。在线评论(Online Reviews)是指网购用户购买某件商品后,根据真实的购物体验和感受并以文本的形式对商品的评价[2],用户的在线商品评论是商品口碑的载体,评论文本是开放式结构,可以帮助顾客迅速有效的筛选出符合个人需求的商品[3]。在线商品评论文本不仅可以使商家掌握消费者需求以提高销量,而且可以使潜在消费者得到有价值的信息[4-5]。因此,越来越多的学者开始研究在线商品评论文本。例如:Susan M. Mudambi 等[6]利用亚马逊平台中顾客对商品的评论,证明评论深度、产品类型会对消费者的购买行为产生相应的影响;刘灵芝等[7]利用天猫旗舰店中水禽熟食的在线评论文本,研究证明在线评论文本中的偏好差异性对商品销量产生负向效果;梁霞等[8]考虑在线评论文本的评论效用和情感倾向,利用

收稿日期: 2021-06-18

作者简介: 崔宁(1980—),女,博士,副教授,硕士生导师,研究方向为工业工程、决策理论与方法;赵宗良(1994—),男,2019 级硕士研究生,通讯作者,研究方向为决策理论与方法;吴瑞雪(1996—),女,2019 级硕士研究生,研究方向为决策理论与方法。

PROMETHEE Ⅱ和随机占有准则方法对备用商品进行排序选择。

在线商品评论文本是一种非结构化数据,使用传统的研究方法分析数据具有一定的局限性,因此,随着计算机语言技术的蓬勃发展,出现了许多处理非结构数据的方法,如关键字提取、主题模型、情感分析等。隐狄利克雷分布(Latent Dirichlet Allocation,LDA)是一种将文本内容以概率分布的形式形成某种主题的无监督学习算法,主要应用于文本主题识别、文本分类等文本挖掘领域方面[9]。目前,部分学者利用 LDA 主题模型从在线评论文本中挖掘出用户对产品或服务某些方面的影响因素。例如:Yue Guo 等[10]利用 LDA 主题模型对酒店的在线评论文本提取出影响顾客体验的影响因素,并认为这些因素对酒店的星级评定至关重要;Y. J.Jung 等[11]利用公司官网上雇员的在线评论文本,基于 LDA 主题模型挖掘出影响雇员工作满意度的因素,如假期、组织文化、工作时间等;Sharan Srinivas 等[12]应用 LDA 模型分析某大学的学生在线评论文本,得到影响学生对学校感观的因素,如学术水平、学校环境等因素。

偏序集决策理论是一种处理多指标对象的评价、选择问题的蕴含权重的多准则决策方法[13],具有稳健性,有效避免了决策过程的不确定性。目前,部分学者将偏序集评价方法应于多个领域。例如:赖文哲等[14]基于偏序集理论对长江及嘉陵江干流段 9 个监测点水质进行水源质量分类,并提出相应的建议;陈亮等[15]基于偏序集评价理论获得样本银行财务绩效排名,并分析出样本银行财务绩效的层级及银行之间的竞争格局;毛志勇等[16]利用偏序集理论构建高校图书馆服务质量评价模型,为高校图书馆服务质量评价提供了新思路。

目前,商家掌握线上消费者需求通常利用问卷法、访谈法、文献法等研究方法,这些研究方法本身具有较强的主观性和片面性,容易受到研究人员自身知识的限制,而且,浪费大量的时间和人力物力仅能得到少量的有效数据,导致研究结论具有局限性和不准确性。另外,偏序集决策理论具有较强的稳健性,不需要各影响因素的具体权重,有效避免了方案决策过程的不确定性。因此,本文构建了基于 LDA 主题模型和偏序集的在线商品评论分析模型,首先,利用 LDA 主题模型从在线商品评论文本中提取出影响顾客消费的因素;其次,基于偏序集理论对调查问卷收集的数据进行商品排序;最后,以商品空调为例进行实证分析,证明模型可行有效,既可以帮助商家掌握顾客消费需求,又可以帮助消费者迅速筛选出心仪的商品。

1 基于 LDA 主题模型和偏序集的在线商品评论分析模型

1.1 问题描述及模型构建

如今,中国是全球最大的线上零售市场,2020 年受疫情的影响,大量的商家涌进线上零售市场,导致线上市场的竞争日趋激烈,掌握顾客的消费需求成为商家提高商品销量和扩大市场占有率的重要渠道。另外,如果顾客打算购买某种商品时,在网购平台搜索栏输入关键字后,虽然符合要求的范围缩小,但是商品的品牌和种类还是比较多,此时需要顾客有一定的筛选能力,但是因为诸多因素的影响,顾客无法直接从海量在线评论中得到有效的信息,导致顾客无法从众多商品中选出符合自己要求的商品。

本文首先利用大量的在线商品评论文本,基于 LDA 主题模型提取出评论文本的主题,并且根据词频确定各主题的权重大小;其次,把提取出的主题作为影响顾客选择商品的因素形成调查问卷;最后,运用偏序集理论对备选商品进行排序,基于 LDA 主题模型和偏序集的在线商品评论分析模型如图 1。

1.2 LDA 主题模型

LDA 主题模型是一种包含“主题—文档—词”三层贝叶斯结构模型,具体训练过程如下[17]:

(1)第 d 条评论包含的特征词数量 N_d 服从泊松分布,即 $N_d \sim Poinsson(\xi)$。

(2)第 d 条评论生成主题分布,即 $\theta_d \sim Dirichlet(\alpha)$;$\theta_d$ 表示第 d 条评论中的主题概率分布,其中 $d \in [1,D]$,D 表示在线评论总数,α 是每条评论下主题的多项分布的 *Dirichlet* 先验参数。

(3)第 k 个主题生成特征词分布,即 $\varphi_k \sim Dirichlet(\beta)$;$\varphi_k$ 表示第 k 个主题下特征词概率分布,其中 $k \in [1,K]$,K 表示主题总数,β 是该主题下特征词的多项分布的 *Dirichlet* 先验参数。

(4)根据主题概率分布 θ_d 生成特征词 $W_{d,n}$ 的主题,即 $Z_{d,n} \sim Multinomial(\theta)$,其中,$W_{d,n}$ 表示第 d 条评论中的第 n 个词,$Z_{d,n}$ 表示第 d 条评论中第 n 个词的主题,其中 $n \in [1,N_d]$;根据特征词概率分布 $\varphi_{Z_{d,n}}$ 生成主题下的特征词,则 $W_{d,n} \sim Multinomial(\varphi_{Z_{d,n}})$。

综上所述,基于 LDA 的在线商品评论文本主题提取模型如图 2 所示。

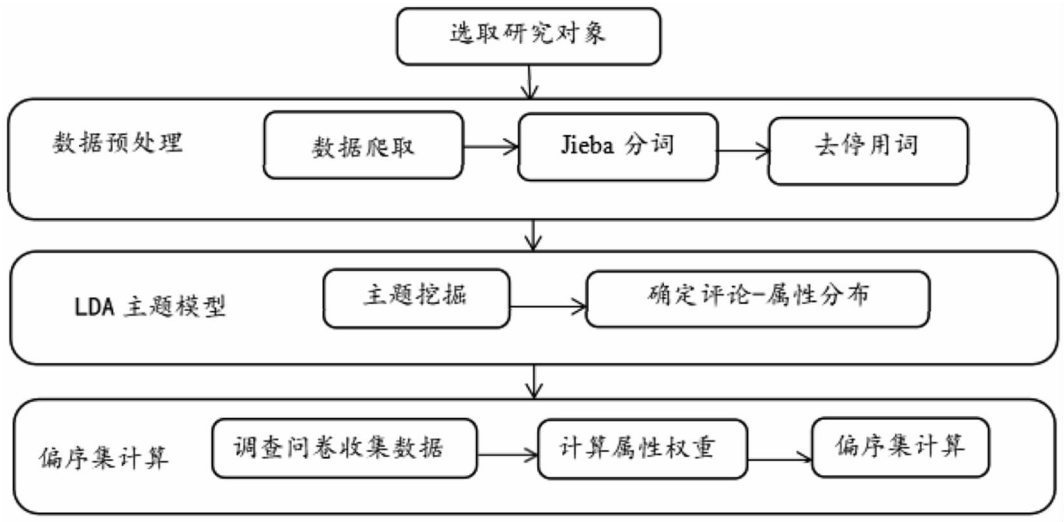

图 1　基于 LDA 主题模型和偏序集的在线商品评论分析模型

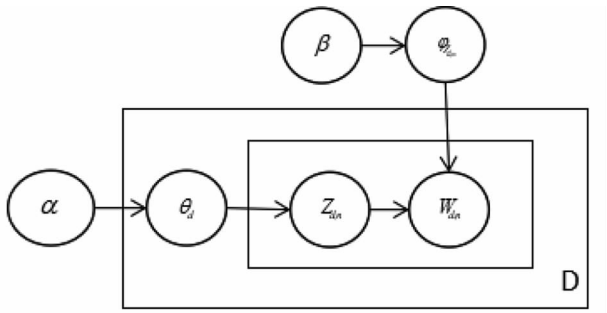

图 2　在线商品评论文本主题提取模型

在 LDA 主题模型中,隐变量的分布概率很难通过计算得到准确数值,所以需要使用一些抽样方法近似估计未知参数,基于现有的研究成果,本文采用吉布斯抽样算法来实现参数估计,从而实现文本中主题抽取。

1.3　偏序集理论

岳立柱等[13] 利用决策问题中指标权重次序 $\omega_{11}>\omega_{12}>\cdots\cdots>\omega_{1n}$,用矩阵形式处理蕴含权重信息的多样本多指标的决策问题:

$$D=(d_{ij})_{m\times n}=X\cdot E=\begin{bmatrix} x_{11} & x_{11}+x_{12} & \cdots & x_{11}+x_{12}+\cdots+x_{1n} \\ x_{21} & x_{21}+x_{22} & \cdots & x_{21}+x_{22}+\cdots+x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m1}+x_{m2} & \cdots & x_{m1}+x_{m2}+\cdots x_{mn} \end{bmatrix}.$$

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,n} \\ x_{21} & x_{22} & \cdots & x_{2,n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{m3} \end{bmatrix}.$$

$$E=\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

假如 D 中第 j 行系数≤第 i 行系数,说明 i 方案优于 j 方案,由此得出全部方案的层级关系。对矩阵 D 进行累加变换后,对每一行进行系数比较得到比较关系矩阵; 即存在偏序集$(A,\leqslant)$,对于 $\forall a_i,a_j\in A$,若 $a_j\leqslant a_i$,记 $r_{ij}=1$; 若 $a_i\leqslant a_j$ 或者 a_i 和 a_j 不可比,记 $r_{ij}=0$,称 R 为$(A,\leqslant)$ 的比较关系矩阵。由比较关系矩阵得到 *Hasse* 矩阵,范懿[13] 给出了比较关系矩阵 R 和 *Hasse* 矩阵 H_R 之间的转换公式:

$$H_R=(R-I)-(R-I)*(R-I).$$

式中,I 为主对角元素构成的单位矩阵,* 为布尔运算。

Hasse 图是一种研究偏序集理论的重要工具,Hasse 图清晰表示了方案间的传递性和逻辑性,Hasse 图可以由 Hasse 矩阵得出。所以,偏序集理论的计算步骤: 首先,确定样本的多个指标; 其次,对指标数据进行预处理,并确定指标权重次序; 最后,计算得出 Hasse 矩阵,并绘制 Hasse 图,进行结果分析。

2　实证分析

2.1　数据获取与预处理

根据奥维云网(AVC) 的数据统计,2020 年中国空调市场线上零售量为 5134 万台,零售额为 1545 亿元,空调行业线上零售量占比首次超过线下。AVC 发布的《2020 年中国空调市场年度报告》显示,2020 年线上销售额最多的空调品牌是美的和格力。中国空调市场零售额前十名的空调品牌如表 1。

表 1 中国空调市场零售额前十名的品牌

品牌	零售额份额/%
美的	34.3
格力	29.0
海尔	10.7
奥克斯	10.2
TCL	3.6
海信	2.3
科龙	1.8
华凌	1.6
米家	1.3
松下	0.7

本文的数据来源于天猫商城的商品评论文本，搜索栏输入空调，按商品销售量从高到低进行排序，分别选取美的、格力、海尔和奥克斯销量第一的商品作为分析对象，各品牌均选择官方旗舰店，保证数据的公平性和有效性，并利用八爪鱼采集器爬取商品评论文本。在线评论文本具有较强的自由性，所以消费者有时对于商品的评论是比较随意的，而且有些消费者没有发表评论，导致获取的数据文本中存在许多需要剔除的无意义评论，例如“此用户没有填写评论”“999”“嗯嗯”等，经过剔除无效评论后共得到 4940 条数据文本，将这些数据文本作为本文的数据来源。然后，对原始数据文本进行预处理，利用 Python 语言中的 Jieba 库进行分词处理，并使用中文停用词表对停用词进行删除，得到特征词数据集。

根据特征词数据集绘制出特征词分布情况，如图 3 所示。

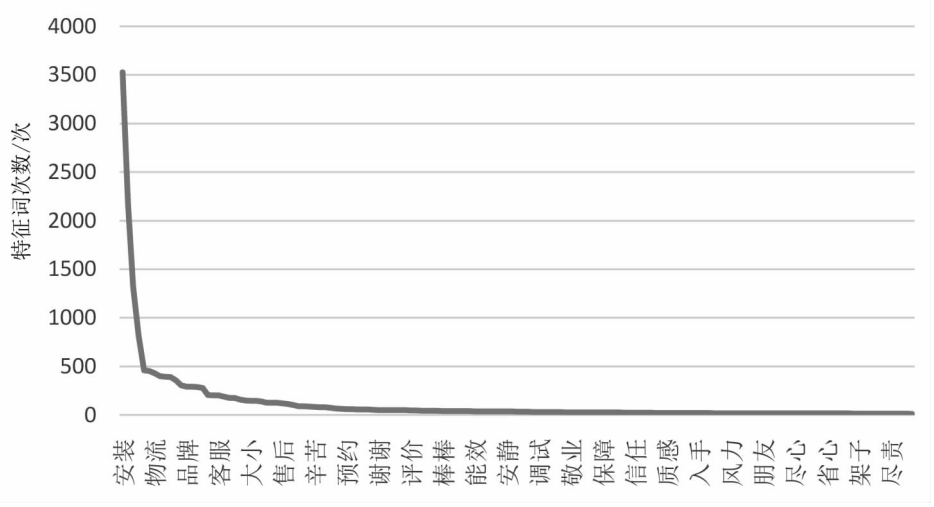

图 3 空调在线评论特征词分布

在线空调评论文本中的特征词分布有典型的“长尾效应”，由于消费者常常依据自己对空调的真实体验和偏好对商品进行评价，导致商品的在线评论文本中使用的词语比较随意，所以，在线商品评论文本是一种明显的非结构化数据。

2.2 基于 LDA 主题模型的影响因素的提取

利用 Python 语言中的 gensim 库对在线评论特征词数据集进行 LDA 主题建模，提取出的主题作为影响顾客消费的因素。首先，依据相关的数据计算，把在线空调评论文本中主题总数设置成 K = 5，并得到 5 个主题下排名前 10 的特征词及其相对应的权重大小; 为了说明 LDA 主题建模效果，展示前两个主题特征词分布情况，如表 2 所示。其次，由各主题下特征词间的逻辑关系确定各主题的名字，表 2 中的主题 1 的名称为产品质量，主要来自于关键词外观、制冷、静音、制热等关键词确定的; 主题 2 的名称为送货速度，主要来自于关键词送货、速度、发货等关键词确定的。同理可以得到其余主题名称: 品牌效应、服务态度和售后安装。

表 2 LDA 主题模型提取结果

主题 1: 产品质量	相对权重	主题 2: 送货速度	相对权重
专业	0.134 918 84	服务	0.240 639 02
效果	0.105 417 74	快	0.106 383 58
外观	0.050 030 02	送货	0.090 797 29
制冷	0.048 459 16	及时	0.084 989 29
静音	0.041 432 02	态度	0.062 900 74
服务到位	0.040 067 45	安装	0.061 454 95
热	0.038 142 05	速度	0.058 566 07
材质	0.037 753 37	发货	0.038 341 83
已经	0.026 658 9	收到	0.029 43 45
好看	0.025 114 11	装	0.023 137 82

2.3 基于偏序集理论的产品排序

2.3.1 问卷设计与数据收集

根据 LDA 主题模型所提取出的主题，作为顾客购买空调时的影响因素，制作网络调查问卷，本文调查问卷借助问卷星平台，通过微信邀请用户填写问

卷,在全国范围内随机抽样调查近两年在天猫旗舰店购买空调的消费者,共收到 336 份问卷,有效问卷 328 份,有效率为 97.6%。调查问卷的选项采用 Likert 5 级量表法进行设计,以“非常不满意—非常满意”(或“非常不影响—非常影响”) 的答案选项形式从 1 分到 5 分依次赋分。先对原始问卷进行预处理,把答案选项进行赋分; 再依次求出各品牌各指标的平均分,调查问卷处理后的各品牌相应指标的评分,如表 3 所示。

表 3　调查问卷处理后的评分

品牌	送货效率	售后安装	服务态度	外观性能	品牌效应
格力	4.0625	4.0536	4.0892	4.0982	4.1696
美的	4.1111	3.9136	4.1481	4.0988	4.0247
海尔	4.0556	3.9445	4.1111	4.1111	3.8889
奥克斯	4.1	3.6	4	3.6	3.8
TCL	3.875	3.75	4	3.875	4.125

2.3.2　信度检验

信度是为保证调查问卷收集处理后的数据具有一定的可靠性,其中克朗巴哈系数的划分标准: 系数小于 0.6,通常认为数据信度太低,不具有可靠性; 系数在 0.7~0.8 时,说明数据具有良好的信度,具有一定的可靠性; 系数在 0.8~0.9 时,表示数据具有非常好的信度,具有较好的可靠性。本文采用 SPSS23.0 软件对处理后的数据进行信度检验,得到总体量表的 Cronbach's Alpha 值为 0.775,标准化后的 Cronbach's Alpha 值为 0.798,说明调查问卷得到的数据具有良好的信度,具有稳定性和可靠性,证明本次研究是有效的。

2.3.3　基于偏序集理论排序

通过 LDA 主题模型提取出影响顾客购买空调的因素,并得到每个因素下的特征词的分布情况,用特征词出现的次数来计算各影响因素的权重,即某一因素权重由该因素下所有特征词出现的次数和所有因素下特征词出现的次数之和的比重计算而得到; 计算后顾客购买空调影响因素的权重大小如表 4 所示。由表 4 可知影响因素的权重次序: 送货效率>售后安装>服务态度>产品质量>品牌效应。

$$w_i = \frac{n_i}{\sum_{i=1}^{m} n_i}(i = 1,2,\cdots,m).$$

表 4　顾客购买空调影响因素的权重

影响因素	权重
产品质量	0.118
送货效率	0.314
服务态度	0.21
售后安装	0.245
品牌效应	0.113

按照偏序集理论的计算步骤,先计算比较关系矩阵,再计算 Hasse 矩阵,最后得到空调品牌层级分布的 Hasse 图,如图 4 所示。由图 4 可知: 第一层的空调品牌是格力和美的,说明顾客在选择空调品牌时,会首选美的和格力; 第二层是的空调品牌是海尔和奥克斯; 第三层是 TCL。与奥维云网(AVC) 发布的《2020 年中国空调市场年度报告》中 2020 年线上销售额空调品牌相比,本方法最后得出的结果与其结论完全一致,说明本文研究方法具有有效性和可靠性。

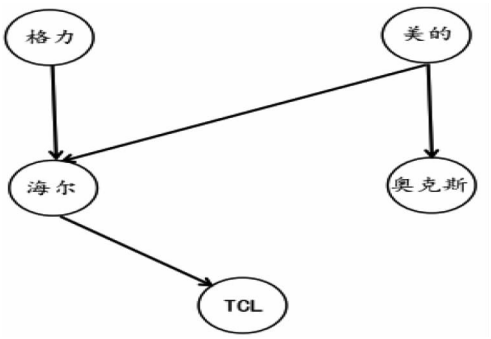

图 4　空调品牌层级分布的 Hasse 图

3　结语

本文提出了一种基于 LDA 主题模型和偏序集的在线商品评论研究方法。该方法利用 LDA 主题模型提取顾客购买商品的影响因素,具有一定的客观性,比文献调研、问卷调查、访谈等方法得到的影响因素更具有可靠性; 计算特征词出现次数确定影响因素的权重次序,避免了人为给定影响因素权重的主观性和局限性。实验结果表明: 顾客在线上购买空调时,更加关注送货效率,最终空调品牌排序结果是与权威机构发布的文件一致的,验证了本文研究的评价方法可行有效。在商品过程中,通过调查问卷的方式得到研究数据,存在一定的局限性,因此,利用客观的方式对网购平台商品进行选择是下一步研究的重点。

总而言之,本文提出的基于 LDA 主题模型和偏序集的在线评论分析方法具有合理性和实际应用价值,能够帮助网购平台商家掌握顾客的消费需求和潜在倾向,及时调整产品结构,从而占领更多的线上市场份额; 同时也为顾客在网络购物平台选择商品时提供参考建议,为分析商品信息提供一种新的思路。

参考文献

[1]　中国互联网络信息中心(CNNIC) . 第 47 次《中国互联网络发展状况统计报告》发布中国将建成全球最大数字

社会[J]. 网络传播,2021(2):68-75.

[2] 宋晓晴,孙习祥. 消费者在线评论采纳研究综述[J]. 现代情报,2015,35(1):164-169.

[3] YUBO C,JINGHONG X. Online consumer review: word of mouth as a new element of marketing communication mix[J]. Management science,2008,54(3):477-491.

[4] FILIERI R,HOFACKER C F,ALGUEZAUI S. What makes information in online consumer reviews diagnostic over time? The role of review relevancy,factuality,currency,source credibility and ranking score[J]. Computers in Human Behavior,2018,80:122-131.

[5] 陶晓波,张欣瑞,杨建坤,等. 在线评论、感知有用性与新产品扩散的关系研究[J]. 中国软科学,2017(7):162-171.

[6] SUSAN M M,DAVID S. Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon. com[J]. MIS Quarterly,2010,34(1):185-200.

[7] 刘灵芝,胡天娇,肖邦明. 熟食品消费的网络评论对线上销量的影响研究:以水禽熟食产品为例[J]. 中国农业大学学报,2018,23(5):208-217.

[8] 梁霞,姜艳萍,高梦. 基于在线评论的产品选择方法[J]. 东北大学学报(自然科学版),2017,38(1):143-147.

[9] TIRUNILLAI S,TELLIS G J. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation[J]. Journal of Marketing Research,2014,51(4):463-479.

[10] YUE G,STUART J B,QIONG J. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation[J]. Tourism Management,2017,59:467-483.

[11] JUNG Y J,SUH Y M. Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews[J]. Decision Support Systems,2019,123:113074.

[12] SHARAN S,SUCHITHRA R . Topic-based knowledge mining of online student reviews for strategic planning in universities[J]. Computers & Industrial Engineering,2019,128:974-984.

[13] 岳立柱,张志杰,闫艳. 蕴含权重的偏序集多准则决策法[J]. 运筹与管理,2018,27(2):26-31.

[14] 赖文哲,毛志勇,岳立柱,等. 基于熵权-偏序集的水质评价方法[J]. 长江科学院院报,2021,38(3):32-38

[15] 陈亮,刘欣慧,李春友. 基于偏序集理论的商业银行财务绩效评价[J]. 统计与决策,2019,35(20):178-181.

[16] 毛志勇,崔鹏杰. 基于偏序集的高校图书馆服务质量评价模型研究[J]. 情报探索,2020(6):20-25.

[17] 冯坤,杨强,常馨怡,等. 基于在线评论和随机占优准则的生鲜电商顾客满意度测评[J]. 中国管理科学,2021,29(2):205-216.

[18] 范懿. 一个有关哈斯图的解析方法[J]. 上海第二工业大学学报,2003,20(1):17-22.