ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 5,Sep. 2020,pp. 965－977
DOI:10. 16337/j. 1004-9037. 2020. 05. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于LDA主题模型和直觉模糊TOPSIS的农产品在线评论情感分析

王珠美[1,2]，胡彦蓉[1,2]，刘洪久[1,2]

(1. 浙江农林大学信息工程学院,杭州，311300;2. 浙江省林业智能监测与信息技术研究重点实验室,杭州，311300)

摘　要: 提出了基于LDA主题模型和直觉模糊TOPSIS的农产品在线评论情感分析方法。该方法使用情感词典对在线评论进行情感倾向分析,并计算农产品的积极情感值;运用LDA主题模型计算各个属性的权重,结合直觉模糊TOPSIS方法计算农产品的综合评价值;采用SPSS统计分析软件进行有效性检验。结果表明,综合评价值与月销售量、积极情感值呈显著的正相关性,说明该方法具有合理性,为挖掘农产品在线评论中的情感信息提供一种新的思路。

关键词: 在线评论;LDA主题模型;直觉模糊;TOPSIS

中图分类号: TP391　　**文献标志码:** A

Emotional Analysis of Agricultural Product Online Reviews Based on LDA Thematic Model and Intuitionistic Fuzzy TOPSIS

WANG Zhumei[1,2], HU Yanrong[1,2], LIU Hongjiu[1,2]

(1. School of Information Engineering, Zhejiang A & F University, Hangzhou, 311300, China; 2. Key Laboratory of Forestry Intelligent Monitoring and Information Technology Research of Zhejiang Province, Hangzhou, 311300, China)

Abstract: An emotional analysis method based on LDA thematic model and intuitionistic fuzzy TOPSIS for agricultural product online reviews was proposed. The method uses the affective dictionary to analyze the emotional tendency of online comments and calculate the positive emotional value of agricultural products. The LDA thematic model is used to calculate the weight of each attribute, and the comprehensive evaluation value of agricultural products is calculated with the intuitive fuzzy TOPSIS method. SPSS statistical analysis software was used to verify the validity. The results show that the comprehensive evaluation value has a significant positive correlation with monthly sales volume and positive emotional value, which indicates that the method is reasonable and provides a new idea for mining emotional information in the online evaluation of agricultural products.

Key words: online reviews; LDA thematic model; intuitionistic fuzzy; TOPSIS

基金项目: 浙江省哲学社会科学规划基金(17NDJC262YB,19NDJC240YB)资助项目;教育部人文社会科学研究规划基金(18YJA630030)资助项目;浙江省自然科学基金(LY17G020025,LY18G010005)资助项目;杭州市软科学(20190834M27)资助项目。

收稿日期: 2020-04-15;**修订日期:** 2020-08-11

引 言

随着互联网与信息技术的迅猛发展,我国的网络购物正在急速发展。根据第45次《中国互联网络发展状况统计报告》数据统计,截至2020年3月,我国网络购物用户数量达到7.10亿个,互联网普及率达64.5%,即超过一半的中国公民都在通过网络来购物。但由于网络购物的虚拟性和产品的不可触摸性,商品的在线信息成为消费者评判商品的重要依据。根据《2015年中国网络购物市场研究报告》数据统计,消费者在网上购物时,商品口碑、价格、商家的信誉成为消费者评判商品的主要考虑因素,其中网络口碑的百分比最大,达到77.5%。在线评论作为口碑的主要载体,成为消费者获取信息的主要来源,也是商家了解消费者需求、产品需求改进、促进商品销量的主要渠道。因此,越来越多的学者开始研究评论中包含的隐藏信息,通过挖掘评论的情感信息进一步分析评论中的有效信息。

情感分析又称情感极性分析,它是对文本进行表达出的情绪积极、消极以及不确定的判断。在现阶段,情感分析主要有通过构建情感词典进行分类的方法,也有机器学习方法。通过构建情感词典的方法主要是通过情感词典对文本进行词语分析,计算情感值,然后通过判断情感值确定文本表达的情感倾向。在基于情感词典进行分类的方法方面,Baccianella等[1]提出一种通过构建情感词典来挖掘情感特征进行情感判断的方法。郭顺利等[2]将用户情感倾向细致划分,通过构建中文图书评论的情感词集,同时结合改进的SO-PMI算法和同义词词林,提出一种判别词语情感类别的方法。也有很多学者对于特定领域构建情感词典。陈柯宇等[3]提出一种结合扩展的情感词典以及word2vec工具的情感倾向分析方法。蒋盛益等[4]通过改进的Hevner情感模型,利用HowNet中语义相似度计算的思想,构建音乐领域的中文情感词典。通过机器学习分析文本情感倾向的主要思想是将文本情感分析转化为一个分类问题,然后利用算法进行训练得到一个模型,最后通过这个模型进行文本情感判断。在机器学习方法方面,Singh等[5]运用相同的数据对机器学习方法和基于语义信息的方法进行情感分类实验,实验表明了基于机器学习方法的有效性。赵刚等[6]对餐厅评论情感分析时,通过比较几种经典的机器学习算法,包含了Ada Boosting、Bayes Network、Decision Tree、C4.5分类树、Naïve Bayes分类器以及Ripper等算法,实现了适合于发掘隐含属性、展现商品间关联性和判断客户情感倾向的网上商品评论情感分析模型。然而在机器学习中,文本大多都是通过词袋模型来表示,这样易造成文本中包含的语义信息和情感信息等问题不能很精确地描述出来,而新兴的深度学习方法恰好能够弥补这些缺点。通过神经网络模型,能够计算得到文本中词语的分布式向量,可以用低维且连续的形式来表达词,能够较好地应用到其他深度神经网络模型,利用多层网络的学习,可以更加具体地表达文本特征,提高了模型的准确性和工作效率。近年来,许多学者将卷积神经网络[7](Convolutional neural network, CNN)、长短时记忆网络[8](Long-term memory network,LSTM)、双向长短时记忆网络[9](Bidirectional long-term memory network, BLSTM)等深度学习模型运用到产品在线评论情感分析中去并取得了较好的成果。

但目前的研究存在以下问题:(1)文本属性权重确定方式不精确。在情感分析方法中有多种属性权重计算方式,其中,词频-逆文本频率(Term frequency-inverse document frequency,TF-IDF)是一个被广泛应用数学统计模型,表示在文档中词语的重要程度,如余苗等[10]运用TF-IDF分类算法挖掘用户兴趣模型,从而实现了情报的按需分发,但该方法的推荐精度还需要进一步提高。(2)文本情感描述不明确。传统的情感分析方法是需要人工标注文本特征后,利用机器学习构建分类模型,判断文本的情感倾向,这样的处理方法对于文本的情感特征描述处理不够客观[11],没有办法准确地描述消费者的情感倾向。

因此,为解决信息的有效提取和分析在线评论与商家绩效之间的关系,本文提出了一种基于潜在狄利克雷分布(Latent Dirichlet allocation,LDA)的主题模型和直觉模糊TOPSIS的农产品在线评论情

感分析方法。该方法的主要特点在于:(1)根据属性出现的次数来确定各个属性的权重。Pang等[12]研究表明,使用词语的出现次数能够获得比词频-逆文本频率方法更好的实验结果。因此,本文将用属性出现的次数来确定各个属性的权重,避免了人为给定权重的不确定性。(2)利用LDA主题模型进行主题建模,通过计算混乱度来确定在线评论的最佳主题数。Chiru[13]通过对现有的主题建模算法在处理大量文档和对已识别潜在主题进行解析方面的比较,确定LDA主题模型具有最高性能。同时根据LDA模型相关参考文献,混乱度是测量LDA预测能力的标准方法[14]。通过混乱度计算在线评论的最佳主题数目,保证了文档的聚类效果。(3)采用直觉模糊数来反映消费者不同的情感。针对消费者情感的不确定性,直觉模糊理论可以反映评论中消费者表达的支持、犹豫和反对程度,全面地描述评论中的情感倾向,弥补了只考虑消费者情感极性的不足。

1 基于LDA主题模型和直觉模糊TOPSIS的农产品在线评论情感分析算法

1.1 问题描述及解决框架

随着科技的发展,人们对于网上购物的依赖越来越大。在生活中,假设消费者想要购买某种农产品,经过关键字搜索后缩小了条件符合农产品的范围,但搜索结果往往还是呈现了数目较多的农产品,这时候进一步的选购就需要消费者具有一定的筛选能力,由于诸多因素限制,消费者无法有效地得到需要的评论信息,在多种商品之间无法便捷轻松地做出购买决定[15]。本文从产品在线评论信息过载出发,设计基于LDA主题模型和直觉模糊TOPSIS的产品在线评论情感分析方法对关键字搜索后的商品进行分析,挖掘在线评论中的有效信息,为消费者挑选商品提供建议,其解决框架如图1所示。

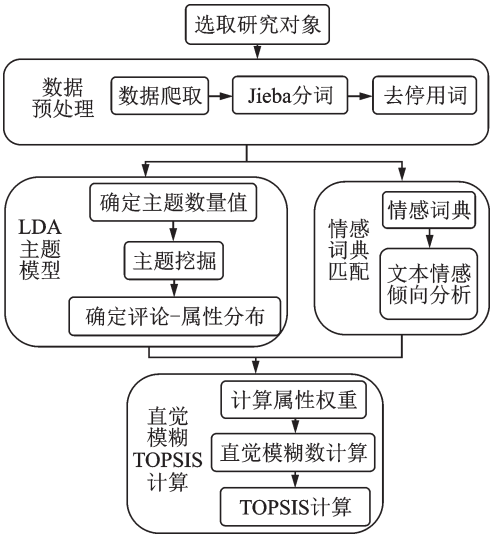

图1 农产品在线评论情感分析结构

Fig.1 Emotional analysis structure of online agricultural product reviews

1.2 LDA主题模型

统计主题模型近年来得到了学者的广泛应用,它能够在计算机没有完全了解文本结构的情况下,分析出易理解且相对平稳的语言结构,为数据集中的文本寻找一个相对简短的描述[16]。统计主题模型最早来源于隐含语义检索(Latent semantic indexing, LSI)[17],重大突破是Hofmann提出的PLSI(Probabilistic latent semantic indexing)模型,PLSI模型主要是通过概率模型来计算文档集中词产生的过程,但是PLSI对于文本的产生不能用概率来描述,只是简单地对部分文本进行拟合,得到指定文本的主题混合比例[16]。针对这些不足,Blei[18]于2003年提出的一种生成主题概率模型LDA,在PLSI的基础上,用一个服从Dirichlet分布的隐含随机变量表示文档的主题混合比例来模拟文档产生的过程,其模型结构更为完整清晰,采用概率去推断算法处理文本,可以将文本表示的维度大大降低,从而避免维度灾难,因此在文本分类、信息检索等领域取得了非常好的实践效果。

1.2.1 LDA主题模型

LDA模型即是3层贝叶斯概率模型,模型包含词—文档—主题3层结构,具体如图2所示,通常用来对大规模文档数据进行建模[19]。文档中某个主题的词汇构成存在一定的概率,且从主题中心选择了某个词语也可以用概率来分析。具体训练过程如下[20]:

(1)评论 m 包含的特征词数量 N_m 服从泊松分布，及 N_m ～泊松(ξ)。

(2)对于评论 m 生成主题分布，其中 $m\in\{1,2,\cdots,M\}$，即 θ_m～Dirichlet(α)，其中 M 表示数据集评论的总数量，θ_m 表示第 m 个评论的主题概率分布，α为每个评论下主题的多项分布的Dirichlet先验参数。

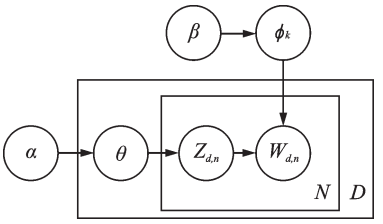

图2 LDA模型的生成过程

Fig.2 LDA model generation process

(3)对于主题 n 生成特征词分布，其中 $z\in\{1,2,\cdots,K\}$，ϕ_k～Dirichlet(β)，K 为总的主题数，β 为每个主题下的词多项分布的Dirichlet先验参数。

(4)评论 m 中的特征词 $w_{m,n}$($n\in\{1,2,\cdots,N_m\}$)的生成过程，N_m 为第 m 个主题包含的特征词①根据主题分布 θ_m 生成评论 $w_{m,n}$ 的特征词主题，即 $z_{m,n}$～Multinomial(θ_m)，$z_{m,n}$ 表示的是第 m 个评论的第 n 个词的主题。②根据词项分布 $\phi z_{m,n}$ 生成所选词主题词项，即 $w_{m,n}$～Multinomial($\phi z_{m,n}$)。

1.2.2 吉布斯抽样

LDA模型中变量的联合分布较难理解，对计算隐含变量概率分布难度很大，常见的抽样方法有接受-拒绝抽样、重要性抽样、吉布斯抽样。吉布斯抽样是应用于马尔科夫蒙特卡洛(MCCM)的一种算法，通常用来分析随机样本的多变量概率分布，由于其在混乱度和运行速度等方面优于接受-拒绝抽样和重要性抽样，且易于实现和推广应用，因此本文采用吉布斯抽样来实现对LDA主题模型进行主题抽取，主要的抽取过程如下：

(1)计算主题-特征词的概率分布

$$P(w,z|\alpha,\beta)=P(w|z,w)*P(z|\alpha) \tag{1}$$

(2)根据贝叶斯公式和Dirichlet先验分布，计算Dirichlet分布期望

$$\theta_{m,k}=\frac{n_{m,(k)}+\alpha_k}{\sum_{k=1}^{K}n_{m,(k)}+\alpha} \tag{2}$$

$$\varphi_{k,t}=\frac{n_{k,(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,(t)}+\beta} \tag{3}$$

式中：$\theta_{m,k}$ 表示数据 m 中主题 k 的概率，$\varphi_{k,t}$ 表示主题 k 中特征词 t 的概率，$n_{m,(k)}$ 表示评论 m 中主题 k 的特征词汇，$n_{k,(t)}$ 表示的是特征词 t 在主题 k 中出现的次数。

(3)通过吉布斯抽样得到概率分布

$$P(z_i=k|w,z_\varepsilon)=\frac{n_{m\varepsilon}^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_{m\varepsilon}^{(k)}+\alpha_k}*\frac{n_{k\varepsilon}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k\varepsilon}^{(t)}+\beta_t} \tag{4}$$

式中：$n_{m\varepsilon}^{(k)}$ 表示数据 m 中没有分配到主题 k 的特征词个数，$n_{k\varepsilon}^{(t)}$ 表示特征词没有分配给主题词 k 的次数。

对于文本数据集来说，LDA模型的主题挖掘过程就是通过文档主题概率分布 θ 和文档对应的主题向量 z，求出式(4)中的最大超参数 α 和 β 的值。在LDA主题模型中，所有文档以及文本的特征词都是可见变量，但是文本的主题是不可见变量，所以通过已有的数据和文本生成规则，LDA主题模型可以实现参数估计，分析出文本中不可见主题，有助于进一步分析文本内容[21]。

1.2.3 确定主题数

在文本预处理后获取文本评论，使用LDA主题模型对其建模，通过吉布斯抽样确定LDA模型参

数。虽然构建好了LDA模型,但文本的主题数无法由模型直接确定,而主题数对抽取主题分布影响较大。当主题数过大时,会产生很多不具明显分类语义信息的主题;当主题数量过少时,会产生比较粗粒度的主题,这样对分类影响也很大[22]。因此,如何科学地确定主题数量非常重要。本文采用混乱度(Perplexity)来确定最优主题数量值。

混乱度在对文档建模过程中特别有用,它关于测试文档概率单调递减,在代数上等价于所有词概率的几何平均值倒数。其实,混乱度可以理解为对于一篇文章d,所训练出来的模型对文档属于哪个主题有很多的不确定,混乱度就可以用来描述这个不确定的程度。混乱度越小,说明聚类的效果越好。计算公式为

$$\text{Perplexity}(D)=\exp\left\{-\frac{\sum_{d=1}^{M}\log P(w_d)}{\sum_{d=1}^{M}N_d}\right\} \tag{5}$$

式中:D为需要测试的文档集,w_d为文本d词汇序列,N_d为文档d的词汇数量,$P(w_d)$为文档中产生w_d的概率。

1.3 产品在线评论情感词典构建

情感分类主要是通过自动分析某种商品评论的文本内容,将其分为正面情感、负面情感和中性情感这3类。常用情感词语又称极性词、评价词,特指带有情感倾向性的词语。显然,情感词语在情感文本中处于举足轻重的地位,情感词语的抽取和极性判断在情感分析创建开始的时候就引起了极大的兴致[23]。

目前,常用的公共情感词典有知网(HowNet)发布的情感词典、台湾大学自然语言处理实验室提供的中文情感词典(National Taiwan University sentiment dictionary,NTUSD)以及清华大学提供的褒贬义词典。本文的情感词典构建如图3所示,具体步骤如下:

(1)选用爬取到的评论数据作为数据集,对原始评论数据进行结巴分词以及去停用词。

(2)将预处理后的评论数据进行筛选,按词性对数据进行筛选。

(3)按词性不同对HowNet、NTUSD和中文褒贬义词典进行筛选。

(4)因为中文语法的复杂性,除了基本情感词典外,还需要标点符号词典、连接词词典、短语词典等,本文根据知网情感词典整理出这3个词典。

(5)按词性的类别合并去重,并且人工对其进行打分,得到本文构建的情感词典,分别如下:副词词典、连接词词典、否定词词典、短语词典、消极词汇词典、积极词汇词典和标点符号词典。

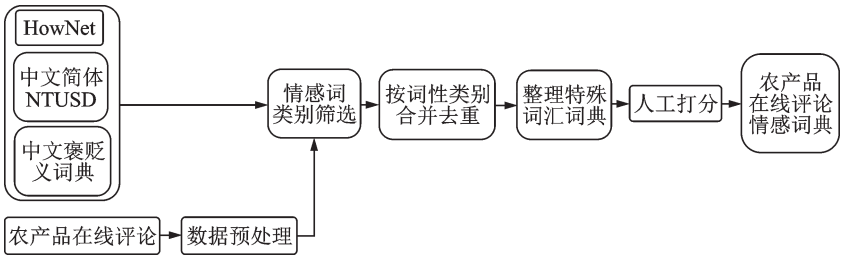

图3 农产品在线评论情感词典构建

Fig.3 Build an emotional dictionary for online reviews of agricultural products

1.4 直觉模糊TOPSIS模型

1.4.1 直觉模糊数的计算

直觉模糊集理论是处理模糊性和犹豫的有用工具,直觉模糊可以同时反映支持、犹豫和反对程度[24]。基于直觉模糊理论,关键字搜索之后的商品在线评论的情感分析可以通过直觉模糊数简单而完整地表示。

q_{ij}^{pos}表示商品A_i的特征j评论中积极情感评论数(k_{ij}^{pos})的占比,也称为积极评论占比,同理可计算得消极评论占比(q_{ij}^{neg})、中性评论占比(q_{ij}^{neu})。表达式为

$$q_{ij}^{\mathrm{pos}}=\frac{k_{ij}^{\mathrm{pos}}}{k_{ij}^{\mathrm{pos}}+k_{ij}^{\mathrm{neg}}+k_{ij}^{\mathrm{neu}}}\quad i=1,2,\cdots,n;j=1,2,\cdots,m \tag{6}$$

因此,根据直觉模糊数的解释,一个直觉模糊$Y_{ij}=[q_{ij}^{\mathrm{pos}},q_{ij}^{\mathrm{neg}}]$可被构造用于关键字搜索后商品$A_i$的特征$f_j$的性能。

1.4.2 TOPSIS模型

TOPSIS方法避免了数据的人为主观性,不需要目标函数,能够很好地刻画多个影响指标的综合影响力度。同时对于数据分布及样本量没有严格的要求,既适用于小样本数据,也适用于多评价单元、多指标的大样本数据,适用性较强。该方法基本思想如下:在确定各个属性指标权重的基础上,归一化原始数据矩阵,分别计算关键字搜索后商品与最优方案和最劣方案间的距离,获得各商品与最优方案的相对接近程度,作为评价商品优劣的依据。具体算法步骤如下:

(1)根据关键字搜索之后商品的整体模糊数构造矩阵决策矩阵$\boldsymbol{A}=(a_{ij})_{n*m}$,其中$a_{ij}=A_{ij}$,表示关键字搜索之后商品$A_i$的特征$f_j$的直觉模糊数,$n$为关键字搜索之后的商品个数,$m$为商品的特征数。

(2)为了消除不同属性之间的量纲效应,使每个属性特征都具有同等的表现力,首先对原始数据进行标准化处理。

$$b_{ij}=\frac{a_{ij}-\bar{a}_j}{s_j}\quad i=1,2,3,\cdots,n;j=1,2,\cdots,m \tag{7}$$

式中:$\bar{a}_j=\frac{1}{n}a_{ij}$,$s_j=\sqrt{\sum_{i=1}^{m}(a_{ij}-\bar{a}_j)}$。

(3)构成加权规范化矩阵

$$\boldsymbol{c}_w=(c_{ij})_{n*m}\quad i=1,2,\cdots,n;j=1,2,\cdots,m \tag{8}$$

通过LDA模型的构建,得到评论-属性的分布情况,统计评论的主题归属情况,用属性出现的次数来计算各个主题的权重$\boldsymbol{W}=(w_1,w_2,\cdots,w_m)^{\mathrm{T}}$。

$$w_j=\frac{n_j(d)}{\sum_{j=0}^{m}n_j(d)}\quad j=1,2,\cdots,m \tag{9}$$

式中:$n_j(d)$为第j个属性在商品评论中出现的次数,属性的权重由该属性出现的次数和所有属性出现的次数之和的比重计算而得到[25]。

$$c_{ij}^{w}=w_j*b_{ij}\quad i=1,2,\cdots,n;j=1,2,\cdots,m \tag{10}$$

(4)确定正理想解C^+和负理想解C^-。正理想解是每个属性评价值最好时的取值,负理想解是每个属性最差时的取值。设正理想解C^+的第j个属性值为c_j^+,负理想解C^-第j个属性值为c_j^-。

(5)计算各方案到正理想解C^+和负理想解C^-的距离。关键字搜索之后的商品A_i到正理想解的距离为S_i^+的计算公式如式(11)所示,同理可以求得S_i^-。

$$S_i^+ = \sqrt{\sum_{j=1}^{m}(c_{ij} - c_j^+)^2} \quad i = 1,2,\cdots,n \tag{11}$$

$$S_i^- = \sqrt{\sum_{j=1}^{m}(c_{ij} - c_j^-)^2} \quad i = 1,2,\cdots,n \tag{12}$$

(6)计算每个商品与正理想解的相对贴近度(综合评价值)。商品$A_i(i=1,2,\cdots,n)$与正理想解C^+的相对贴近度定义为

$$C_i = \frac{S_i^-}{S_i^+ + S_i^-} \quad i = 1,2,\cdots,n \tag{13}$$

显然,$C_i \in [0,1]$,且C_i越大,则商品A_i越优。

(7)确定商品的优劣排序。综合评价值表示各种商品与正理想解、负理想解的距离进行比较,靠正理想解越近、离负理想解越远的备选方案的综合评价值就越大。可以按照综合评价值从大到小的商品优劣排序,确定其最优商品。

2 实　　验

2.1 数据源说明

本文选取天猫商城作为分析数据的来源,关键词设置为西湖龙井,按商品销售量从高到低进行排序,选取排名前200的商品作为分析对象,通过八爪鱼软件爬取商品评论数据。天猫商城是一个评论自由性较强的平台,消费者评论商品信息比较随意,因此获取的数据中存在很多需要剔除的垃圾评论,例如“哈哈哈哈哈哈”“666”等,经过去除垃圾评论之后一共得到110 824条评论数据,将这些在线评论作为本文实验的数据内容。

然后,对评论进行数据预处理。具体过程为:用Python中的Jieba分词软件包对评论数据进行分词处理;收集四川大学机器智能实验室停用词库、哈工大停用词库、百度停用词列表以及中英文停用词表,合并去重后作为本文实验的停用词表,经过Python编程对商品评论去除停用词。

最后,筛选评论中的词汇,根据情感词性进行打分,构成情感词典,手动检查词典的正确性,并根据商品的特性对情感词典进行补充。

2.2 基于LDA主题模型的农产品在线评论情感分析

2.2.1 最优主题数目的确定

使用主题模型建模的过程中,主题数量的最优值采用混乱度来确定,采用Gibbs抽样,抽样迭代参数值设为3 000。通过设置不同的主题数量对混乱度指标进行分析,获取最小混乱度的最优主题数目,具体结果如图4所示。从图4可以看出,当主题数目设置为20时,训练得到的LDA主题模型的混乱度最低,之后混乱度逐渐增长。因此,本文最优的主题数目为20。

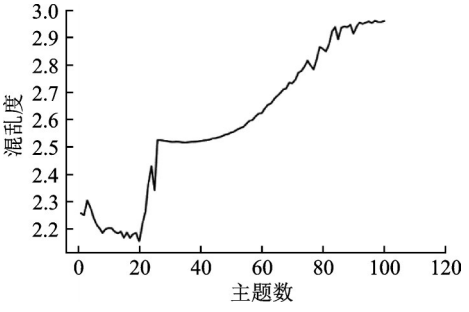

图4 LDA主题模型混乱度随主题数值变化趋势

Fig.4 Disorder degree of LDA topic model changes with the trend of topic value

2.2.2 基于LDA模型的主题挖掘

基于Python语言的机器学习包gensim对评论数据进行LDA主题建模,本文得到20个主题及其分布情况。为了展示建模效果,这里只展示其4个主题,每个主题的前10个词汇的分布情况,如表1所示。

LDA主题挖掘可以按照语义划分,得到语义相关词表达的若干个隐含主题。例如,Topic 0的词汇集合描述了主题“茶香”,Topic 1的词汇集合描述了主题“性价比”,Topic 2的词汇集合描述了主题“划算”,Topic 3词汇集合描述了主题“价位”,同理可得其余16个主题的挖掘结果描述的具体主题,如“服务、分量、促销、外包装、优惠、正宗、信赖、茶叶外观、满意、被推荐、品牌、颜色、图片、评论、上档次、完整”,详细见表2。

表1 主题挖掘结果

Table 1 Topic mining results

Topic 0	Topic 1	Topic 2	Topic 3
清香, 0.041 895 96	还行, 0.089 276 97	很棒, 0.052 666 03	贵, 0.038 477 29
实惠, 0.034 959 85	性价比, 0.086 251 84	正品, 0.052 208 64	小, 0.037 099 38
香味, 0.029 025 74	高, 0.073 557 6	划算, 0.046 935 70	特别, 0.029 615 87
价格, 0.027 893 35	不好, 0.043 211 70	特级, 0.037 047 20	评论, 0.023 515 70
高大, 0.025 620 58	完美, 0.024 473 71	同事, 0.036 446 31	价位, 0.022 900 92
口感, 0.021 823 65	甘甜, 0.022 699 45	合适, 0.033 374 35	产品, 0.020 632 92
泡, 0.021 742 13	回味, 0.020 295 62	很大, 0.032 616 57	算, 0.018 460 59
香气, 0.019 689 39	很足, 0.017 902 08	图片, 0.030 972 64	一盒, 0.015 907 92
不错, 0.019 333 44	没什么, 0.017 868 19	选择, 0.025 884 61	分, 0.015 710 90
龙井茶, 0.018 343 86	真不错, 0.016 141 05	双, 0.022 893 89	惊喜, 0.015 193 36

表2 主题权重

Table 2 Theme weight

主题	0	1	2	3	4	5	6	7	8	9
控掘结果	茶香	性价比	划算	价格	服务	分量	促销	外包装	优惠	正宗
权重	0.051	0.083	0.037	0.041 8	0.130	0.025	0.071	0.038	0.051	0.061
主题	10	11	12	13	14	15	16	17	18	19
挖掘结果	信赖	茶叶外观	满意	被推荐	品牌	颜色	图片	评论	上档次	完整
权重	0.054	0.025	0.046	0	0.052	0.081	0.027	0.034	0.064	0.029

2.3 基于直觉模糊TOPSIS的农产品在线评论情感综合评价值计算

2.3.1 属性权重的确定

根据LDA主题模型得到的评论数据集中评论-主题概率,根据公式(9)得到20个主题的权重,从表2中可以看出主题4(服务)的权重最大,权重为0.130,可以看出消费者在挑选茶叶时最关注的是商家的服务;主题13(被推荐)的权重最小,权重为0,可以看出消费者在挑选茶叶时受别人推荐的影响最小。同时可以分别计算200种商品各自的评论-主题-权重分布,分析每种商品的具体情况,为调整商品特征结构提供参考信息。

2.3.2 直觉模糊决策矩阵

根据式(6)计算可得200个农产品的直觉模糊数组成的TOPSIS决策矩阵。这里只展示销售量前6名的商品的前10个主题决策矩阵,如表3所示。从表3中可以看出,各个商品-主题-情感倾向分布,例如,商品1中主题0(茶香)的直觉模糊矩阵[0.828,0.046],其中0.828表示的是商品1评论中属于主题0

(茶香)的积极评论占比,0.046表示的是商品1评论中属于主题0(茶香)的消极评论占比。由此可见,商品1主题0中的积极评论数量要远远多于消极评论数量,商品1的茶香这一商品特质符合了绝大部分购买此商品的消费者需求(如果有需要,笔者可以提供全部的数据)。

表3 直觉模糊矩阵

Table 3 Intuitionistic fuzzy matrix

主题	直觉模糊矩阵					
	商品1	商品2	商品3	商品4	商品5	商品6
0	[0.828,0.046]	[0.814,0.070]	[0.775,0.038]	[0.767,0.041]	[0.772,0.114]	[0.781,0.109]
1	[0.915,0.880]	[0.880,0.028]	[0.730,0.108]	[0.865,0.032]	[0.891,0.020]	[0.912,0.029]
2	[0.744,0.049]	[0.775,0.075]	[0.450,0.225]	[0.659,0.146]	[0.701,0.052]	[0.720,0.120]
3	[0.791,0.044]	[0.800,0.063]	[0.634,0.037]	[0.782,0.010]	[0.769,0.092]	[0.725,0.072]
4	[0.782,0.034]	[0.744,0.064]	[0.631,0.123]	[0.733,0.067]	[0.732,0.065]	[0.670,0.115]
5	[0.755,0.041]	[0.814,0.034]	[0.695,0.051]	[0.636,0.036]	[0.661,0.054]	[0.609,0.043]
6	[0.800,0.096]	[0.898,0.036]	[0.773,0.068]	[0.870,0.031]	[0.788,0.061]	[0.827,0.072]
7	[0.500,0.037]	[0.474,0.035]	[0.219,0.031]	[0.400,0.073]	[0.283,0.019]	[0.183,0.100]
8	[0.867,0.020]	[0.876,0.010]	[0.846,0.038]	[0.895,0.000]	[0.905,0.036]	[0.913,0.061]
9	[0.915,0.014]	[0.763,0.007]	[0.649,0.070]	[0.784,0.000]	[0.790,0.025]	[0.842,0.069]

2.3.3 加权规范矩阵

根据式(7)将农产品的整体模糊数构造决策矩阵进行标准化处理,结合特征权重,计算加权规范矩阵,部分商品的加权规范矩阵如表4所示。

表4 加权规范矩阵

Table 4 Weighted gauge matrix

主题	规范加权矩阵					
	商品1	商品2	商品3	商品4	商品5	商品6
0	[0.025,−0.010]	[0.022,−0.003]	[0.014,−0.013]	[0.013,−0.012]	[0.014,0.011]	[0.016,0.010]
1	[0.070,0.448]	[0.059,−0.030]	[0.009,0.015]	[0.054,−0.027]	[0.062,−0.034]	[0.069,−0.029]
2	[0.012,−0.004]	[0.016,0.003]	[−0.023,0.046]	[0.002,0.024]	[0.007,−0.003]	[0.009,0.016]
3	[0.019,−0.007]	[0.020,−0.001]	[−0.002,−0.009]	[0.018,−0.017]	[0.016,0.008]	[0.010,0.002]
4	[−0.019,−0.051]	[−0.025,−0.027]	[−0.044,0.022]	[−0.027,−0.024]	[−0.027,−0.025]	[−0.038,0.015]
5	[0.011,−0.003]	[0.015,−0.004]	[0.007,−0.001]	[0.004,−0.004]	[0.005,0.000]	[0.002,−0.002]
6	[0.006,0.010]	[0.028,−0.021]	[0.000,−0.004]	[0.022,−0.024]	[0.004,−0.008]	[0.012,−0.002]
7	[−0.017,−0.010]	[−0.019,−0.010]	[−0.047,−0.011]	[−0.027,−0.002]	[−0.040,−0.014]	[−0.050,0.005]
8	[0.023,−0.016]	[0.025,−0.021]	[0.020,−0.010]	[0.028,−0.024]	[0.029,−0.011]	[0.031,−0.001]
9	[0.020,−0.028]	[0.014,−0.030]	[−0.007,−0.004]	[0.018,−0.034]	[0.019,−0.023]	[0.029,−0.004]

2.3.4 基于TOPSIS的商品综合评价值

根据式(11)、(12)和(13),本文计算每种农产品在线评论情感倾向的正、负理解,以及每种农产品在线评论的情感综合评价值,本文选取了部分农产品的综合评价值,绘制了在线评论情感综合评价值

表,具体见表5所示。从表5中可以看出,200种商品的综合评价值最大的是第88种商品,综合评价值为0.614;综合评价值最小的商品有多个,综合评价值为0。

表5 在线评论情感综合评价值

Table 5 Online comments on the value ofcomprehensive emotional assessment

商品	综合评价值	正理想解	负理想解
1	0.042	1.768	0.499
2	0.048	1.768	0.090
3	0.085	1.811	0.168
4	0.053	1.778	0.094
5	0.063	1.781	0.119
6	0.071	1.792	0.140
⋮	⋮	⋮	⋮
87	0.102	1.779	0.201
88	0.614	1.082	0.700
89	0.092	1.762	0.178
⋮	⋮	⋮	⋮
198	0.142	1.778	0.142
199	0.116	1.807	0.237
200	0.228	1.815	0.535

为了更直观地观测每种农产品在线评论情感综合评价值情况,本文绘制了200种农产品在线评论情感综合评价值折线图,具体如图5所示。从图5中可以看出,200种农产品的综合评价值呈现无规律的波动。对200种农产品在线评论情感综合评价指数计算可得综合评价指数平均值为0.097,200种农产品中有76种农产品的综合评价指数超过了平均值,销售量前50的农产品中只有9种农产品的综合评价指数超过了平均值,由此可见,农产品的销售量并不是影响综合评价指数的主要因素。

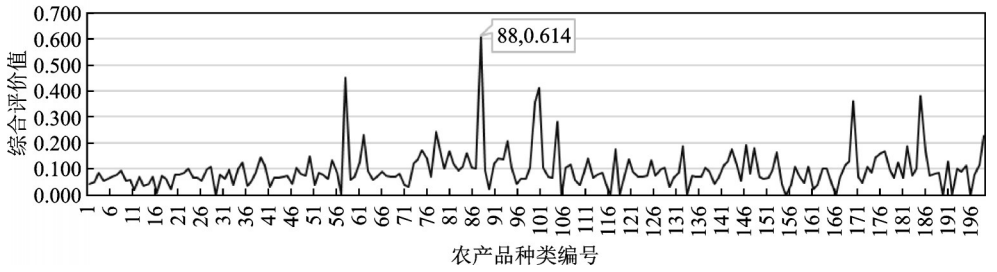

图5 农产品在线评论情感综合评价指数趋势

Fig.5 Agricultural products online review Sentiment comprehensive evaluation index trend

2.3.5 有效性分析

为验证基于LDA主题模型和直觉模糊TOPSIS的农产品在线评论情感分析方法的有效性,本文采用综合评价值与其他变量的相关性来验证,具体的指标包括综合评价值、月销量、积极情感值,其中积极情感值是指某农产品积极情感倾向的产品评论在该农产品全部文本评论中出现的比例,积极情感值

越大,情感倾向越强。变量分析具体结果如表6所示。从表6可以看出,在0.001水平上,综合评价值与店铺销量、积极情感值呈现显著的正相关性,说明本文的综合评价值具有合理性,评价方法是有效的。

表6 变量相关分析结果

Table 6 Results of variable correlation analysis

变量	月销售	积极情感值	综合评价值
月销售量	1.000		
	0.000		
积极情感值	0.135	1.000	
	0.005	0.000	
综合评价值	0.197	0.033	1.000
	0.000	0.493	0.000

注:表中每个变量的上方数据为相关系数大小,下方为sig值(<0.05为显著性意义)。

3 结束语

本文提出了一种根据在线评论对商品进行排序的方法。该方法通过计算属性出现的次数计算权重,避免人为给定权重的主观性和不确定性;充分考虑到评论的聚类效果,利用混乱度来确定最佳主题数目。除此之外,本文还考虑了消费者对不同商品的多种情感,利用直觉模糊数全面反映消费者的情感倾向,更符合消费者的实际购买情况。实验结果表明,本文提出的方法得到的综合评价值与月销售量、积极情感值呈显著的正相关性,这说明了该分析方法具有合理性,评价方法是有效的。在实验过程中发现,通过情感词典的方法来判断农产品在线评论的情感倾向,这一方法十分依赖人工构造的情感词典,存在一定的主观性。所以,客观评价在线评论的情感倾向成为下一步工作的重点。

总的来说,本文结合LDA主题模型和直觉模糊TOPSIS理论,提出了一种农产品在线评论情感分析方法。本文提出的情感分析方法具有合理性和实际应用价值,可以帮助商家了解消费者的购物需求,及时调整产品结构,同时也为消费者挑选商品提供参考建议,为当今分析商品信息提供了一种新的思路。

参考文献:

[1] BACCIANELLA S, ESULI A, SEBASTIANI F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of the International Conference on Language Resources and Evaluation. Valletta, Malta:[s.n.], 2010.

[2] 郭顺利, 张向先. 面向中文图书评论的情感词典构建方法研究[J]. 现代图书情报技术, 2016(2): 67-74.
GUO Shunli, ZHANG Xiangxian. Building sentiment analysis dictionary for chinese book reviews [J]. Modern Book Information Technology, 2016(2): 67-74.

[3] 陈柯宇, 何中市. 基于情感词典的酒店评论情感分类研究[J]. 现代计算机(专业版), 2017(6): 3-6.
CHEN Keyu, HE Zhongshi.Sentiment classification of hotel reviews based on sentiment dictionary[J]. Modern Computer, 2017 (6): 3-6.

[4] 蒋盛益, 阳垚, 廖静欣. 中文音乐情感词典构建及情感分类方法研究[J]. 计算机工程与应用, 2014, 50(24): 118-121.
JIANG Shengyi, YANG Yao, LIAO Jingxin. Research of building Chinese musical emotional lexicon and emotional classification[J].Computer Engineering and Applications, 2014, 50(24): 118-121.

[5] SINGH V K, PIRYANI R, UDDIN A, et al. Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches[C]//Proceedings of the 5th International on KST. [S.l.]: [s.n.], 2013.

[6] 赵刚, 徐赞. 基于机器学习的商品评论情感分析模型研究[J]. 信息安全研究, 2017, 3(2): 166-170.
ZHAO Gan, Xu ZAN. Research on the sentiment analysis model of product reviews based on machine learning[J]. Journal of Information Security Research, 2017, 3(2): 166-170.

[7] 王盛玉, 曾碧卿, 胡翩翩. 基于卷积神经网络参数优化的中文情感分析[J]. 计算机工程, 2017, 43(8): 200-207.
WANG Shengyu, ZENG Biqing, HU Pianpian. Chinese sentiment analysis based on parameter optimization of convolutional neural network[J]. Computer Engineering, 2017, 43(8): 200-207.

[8] 颜端武, 杨雄飞, 李铁军. 基于产品特征树和LSTM模型的产品评论情感分析[J]. 情报理论与实践, 2019, 42(12): 134-138.
YAN Duanwu, YANG Xiongfei, LI Tiejun. Sentiment analysis of product reviews based on product feature tree and LSTM model[J]. Information Studies: Theory & Application, 2019, 42(12): 134-138.

[9] 李洋, 董红斌. 基于卷积神经网络和BLSTM网络特征融合的文本情感分析[J]. 计算机应用, 2018, 38(11): 3075-3080.
LI Yang, DONG Hongbin. Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network[J]. Journal of Computer Applications, 2018, 38(11): 3075-3080.

[10] 余苗, 杨瑞娟, 程伟, 等. 基于TF-IDF分类算法的雷达情报分发技术[J]. 计算机工程与设计, 2012, 33(5): 1822-1826.
YU Miao, YANG Ru ijuan, CHENG Wei, et al. Research on intelligence distribution based on TF-IDF classifier[J]. Computer Engineering And Design, 2012, 33(5): 1822-1826.

[11] 魏志远, 岳振军. 基于直觉模糊集的情感分析研究方法[J]. 通信技术, 2017, 50(12): 2692-2697.
WEI Zhiyuan, YUE Zhenjun. Sentiment analysis based on intuitionistic fuzzy sets[J]. Communications Technology, 2017, 50 (12): 2692-2697.

[12] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques[C]// Proceedings of Empirical Methods in Natural Language Processing.[S.l.]: [s.n.], 2002: 79-86.

[13] CHIRU R. Comparison between LSA-LDA-lexical chains[D]. [S.l.]: WEBIST, 2014.

[14] WANG B, LIU S, DING K, et al. Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: A case study in LTE technology[J]. Scientometrics, 2014, 101(1): 685-704.

[15] 林杰, 王梦娇, 张振宇. 基于在线评论的直觉模糊TOPSIS商品购买决策方法[J]. 上海管理科学, 2019, 41(1): 26-30.
LIN Jie, WANG Mengjiao, ZHANG Zhenyu. A decision-making method of product based on online review using intuitionistic fuzzy TOPSIS method[J]. Shanghai Management Science, 2019, 41(1): 26-30.

[16] 曹娟, 张勇东, 李锦涛, 等. 一种基于密度的自适应最优LDA模型选择方法[J]. 计算机学报, 2008(10): 1780-1787.
CAO Juan, ZHANG Yongdong, LI Jingtao, et al. A method of adaptively selecting best LDA model based on density[J]. Chinese Journal of Computers, 2008(10): 1780-1787.

[17] DEERWESTER S. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990(41): 391-407.

[18] BLEI N J. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(3): 993-1002.

[19] 彭雨龙. 基于VSM和LDA模型相结合的新闻文本分类研究[J]. 山东工业技术, 2016, 212(6): 202-203.
PENG Yulong. Research on text classification of news based on VSM and LDA model[J]. ShanDong Industrial Technology, 2016, 212(6): 202-203.

[20] 祖弦. LDA主题模型研究综述[J]. 合肥师范学院学报, 2015, 33(6): 55-58.
ZU Xian. Review of the latent dirichlet allocation topic model[J]. Journal of Hefei Normal University, 2015, 33(6): 55-58.

[21] 李莉, 林雨蓝, 姚瑞波. 基于LDA模型的交互式文本主题挖掘研究——以客服聊天记录为例[J]. 情报科学, 2018, 36(10): 64-70.

LI Li, LING Yulan, YAO Ruibo. Interactive text theme mining based on LDA model—Take customer service chat as an example[J]. Information Science, 2018,36(10): 64-70.

[22] 关鹏, 王曰芬. 科技情报分析中LDA主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016(9): 42-50.
GUAN Peng, WANG Yuefen. Identifying optimal topic numbers from sci-tech information with LDA model[J]. Data Analysis and Knowledge Discovery, 2016(9): 42-50.

[23] 刘挺, 赵妍妍, 秦兵. 文本情感分析[J]. 软件学报, 2010,21(8): 1834-1848.
ZHAO Yanyan, LIU Ting, QIN Bing. Sentiment analysis[J].Journal of Software, 2010,21(8): 1834-1848.

[24] 俞锦涛. 基于犹豫直觉模糊相关系数的多属性决策[J]. 重庆大学学报, 2019,34(3): 1-7.
YU Jintao, WANG Jinshan, WANG Peng, et al. Multi-attribute decision making based on correlation coefficient of hesitant-intuitionistic fuzzy set[J]. Journal of Chongqing University of Technology, 2019,34(3): 1-7.

[25] 由丽萍. 基于情感分析和VIKOR多属性决策法的电子商务顾客满意感测度[J]. 情报科学, 2015,34(10): 1098-1110.
YOU Liping. Measurement of e-commerce customer satisfaction based on sentiment analysis and VIKOR[J]. Journal of the China Society For Scientific And Technical Information, 2015,34(10): 1098-1110.

作者简介:

王珠美(1993-),女,硕士研究生,研究方向:决策管理研究,E-mail: judewzm@163.com。

胡彦蓉(1970-),**通信作者**,女,副教授,研究方向:预测与评价研究,E-mail: rosehyr2004@aliyun.com。

刘洪久(1969-),男,教授,研究方向:决策管理研究,E-mial:164291073@qq.com。

(编辑:夏道家)