

基于内容特征的评论效用排名预测 ——以豆瓣书评为例

聂 卉

(中山大学资讯管理学院, 广州 510275)

摘要: 本文基于双路径模型理论,以书评为对象,深入探究表征评论内容信息量、结构、语言、论据及情感的特征变量对评论感知效用的影响,据此构建评论效用预测模型,实现基于内容的评论排名预测。研究分两个层面:解释层面,运用基于随机森林的特征优选算法,探析影响评论感知效用的重要文本特征;预测层面,采用树回归模型预测评论效用,实现基于内容的评论排名推荐。研究结果表明,对于长篇幅书评,内容蕴含的信息量、组织结构以及内容中主客观论据的呈现对提升评论感知效用预测精度有重要影响力;基于优选的内容特征,评论效用预测模型的解释力达78%,误差小于0.001;对于投票得分较高的评论,基于内容的效用预测排名与投票排名基本保持一致。这些结论验证了依据评论内容能够比较准确地预测评论的感知效用的判断,揭示出评论的感知效用与评论内容的密切关系。这一结论为网站进行评论质量控制和有效利用提供了依据及可行方案。

关键词: 在线评论; 预测; 评论效用; 文本挖掘

引 言

移动互联时代,在网上抒发感悟、分享观点成为人们的生活日常。网络上由此涌现大量用户评论,对人的观念行为产生深远影响。但复杂的网络环境中,评论资讯泛滥,质量良莠不齐的问题严重。人们从中获得有价值资讯的认知成本越来越高。为减轻用户认知负担,具有评论功能的网站采用投票机制控制评论质量。网站依据基于投票机制获得的评论效用指标对评论进行排名,帮助人们快速定位有价值资讯。尽管如此,有研究指出,基于投票机制形成的排名存在偏差。新近发布的评论,因获得投票少,在网站呈现位置靠后。排名靠前的评论,获得关注多,但可能是虚高评价^[1]。网站评价机制的这种局限性,随着评论资讯的迅猛增长,日显突出。学术研究逐渐从聚焦于在线评论对消费者购买决策的影响机制,转变为探索引导在线评论的内容与特征以提升在线评论的效用价值,后者更具现实意义^[2]。

对这一问题的探讨,大量研究工作主要集中在评论元特征、评论人及外围的评论特征层面。但作为带给评论阅读者直接感知的评论“主体”,内容的挖掘更具价值。评价评论质量,预测评论是否有用,内容始终无法回避。随着研究推进,近年来,研究工作逐渐深入内容展开^[3-6,9]。

本文以实现基于内容的评论感知效用预测与排名为目标,强调评论内容。本研究认为,如果能够从内容层面明晰影响评论效用的重要因素,则能够直接构建有效模型辨识高质量评论,及时呈现有价值信息。目前,中文语料环境下,专门以内容为主题的评论效用研究相对较少。本研究拓展丰富了这一研究方向,实践意义则在于寻求有效控制评论质量的可行方案,本质上解决网络资讯“信息过载”与“知识匮乏”间的矛盾。

文献回顾

1、评论感知效用研究

评论感知效用是测度评论阅读者感知评论价值的指标。作为主观概念,感知效用与阅读者关联密切。商务领域,产品评论阅读者往往被当作潜在消费者,有用评论被界定为:能具体描述产品的特征或性能等信息,辅助潜在消费者做出适当购买决策的评论^[10]。“消费者认为怎样的在线评论更有用?”^[11]成为评论感知效用研究探讨的核心。这一研究以评论的感知效用或质量为目标变量,从不同视角探究感知效用的影响机理,分

收稿日期: 2017-12-04

基金项目: 国家社会科学基金项目(15BTQ067)。

作者简介: 聂卉,中山大学资讯管理学院副教授,硕士生导师,博士。

析作用于评论感知效用的各类型特征要素,进而构建预测模型,实现对评论资讯的分类、排序及推荐。

2、影响评论感知效用的特征要素

从模型解释或预测变量的视角考虑,基于不同研究目标,作用于评论感知效用的特征涉及评论元数据、评论人和评论内容三个层面。

(1) 元数据特征

元特征能够表征评论人与阅读者之间的互动属性^[1],以数值形式直接反馈阅读者对评论或评论对象的印象,是最常采用的评论特征,如评论评分、产品评分、评论投票量、评论发表时间等。以“评论评分”为例,有研究^[9,11-13]将其命名为“评论极性”,有的则称其为“评论效价”^[14]。本质上,“评论评分”反映阅读者对评论的认同。相关研究^[4,8,9,11-14]指出,阅读者对评论的认同显著影响评论的感知效用,而这一影响还会随产品类型不同而不同^[12]。

(2) 评论人特征

评论人特征涵盖评论人的方方面面,包括社交网络、经验知识、网站声誉、身份公开程度以及行为模式等。在研究模型中引入评论人特征是为了探究评论人的资质能力与其发文质量间的关系。Ghose 和 Ipeirotis^[15]以及 Huang 等^[16]强调了评论人的经验和信誉,其过往的发文质量被验证与评论的感知效用正相关,即口碑好的评论人发表的评论效用值更高;Willemsen 等^[8]则发现评论人的专业性对评论的感知效用有影响。殷国鹏^[11]以网络社区评论人的入度中心度表征其受关注程度,指出评论人的受关注度与评论效用呈倒 U 型关系。闵庆飞等^[17]的研究则验证了评论人内向社交网络质量对评论效用有显著影响力。

(3) 评论内容特征

作为评论主体,评论内容直接激发阅读者,阅读者从中获取信息,同时对信息价值产生判断。内容中蕴含的评论特征十分丰富,分语法、体裁和语义三大类^[1]。

语法特征主要基于词汇,如各种词性的统计量以及情态词、感叹词、比较级形容词等特殊词性在内容中的呈现。以预测为目标的研究常以语法类特征表征内容。Martin 等^[12]的研究引入了词性特征,发现尽管简单,词性却能显著提升高质量评论的识别精度。Krishnamoorthy^[18]利用心理学领域的 LCM(Linguistic Category Model)语言模型解析内容,探析蕴含心理特征的用词对评论感知效用的作用,结论是,反映心理状态和行为的词汇会影响阅读者的认知过程,进而对评论感知效用产生效应。

体裁特征通常用来刻画评论内容的组织结构和语言风格,涉及段落数、语句数、平均句长、可读性等指标。如 Chen 和 Tseng^[7]依据设定窗口宽度下的词性统计量获得了两个与词汇复杂度有关的指标,用来表征语言风格。郝媛媛等^[19]用简单的平均句长刻画评论可读性。Ghose 和 Ipeirotis^[15]则采用标准的易读指数和拼写错误率反映可读性,并验证了可读性与评论感知效用显著相关。这种相关性解释为内容表达对阅读者认知负荷的影响,进而对评论感知效用产生作用。

较之语法和体裁,语义特征反映内容实质,强调信息呈现与情感表达。信息量、评论客观性、专业性与信息呈现有关。众多研究^[7,11-13,15]以评论长度表征信息量。Mudambi 和 Schuff^[12]指出信息量能反映内容深度,提升阅读者信息诊断力,帮助阅读者消除认知上的不确定性,通过建立回归模型,Mudambi 和 Schuff 验证了信息量对评论感知效用的显著影响力。Nikolaos 等^[13]重点分析评论效用、评论极性与内容深度三者的交互作用,指出内容深度同时还影响用户对评论的评价(评论极性)。Willemsen 等^[8]从客观的角度提出“论据”这一概念,并根据内容中蕴含的产品特征词提炼出论据密度和多样性指标,指出“论据”是评论感知效用的重要影响因素。Chua 和 Banerjee^[3]关注内容的可理解性以及专业性,他们用信息量和易读指标表征可理解性,用特征词统计量以及名词密度刻画内容的专业性,得出有产品细节描述的评论更有用(对于搜索类产品)的结论。王智生等^[9]同样以评论长度和产品特征在内容中的占比描述内容中蕴含的信息,发现产品特征提及度对评论感知效用有显著影响力。由此可见,阅读者感知评论的价值与评论内容密切相关,充分、客观、专业、细致的内容提升阅读者的信息诊断力,从而使评论获得更多认同。

情感是语义特征强调的另一重要元素,涉及面宽泛,从主客观性、褒贬观点,到细粒度情绪。Martin 和 Pu^[6]从内容中提炼情感特征,整合词性及可读性,预测评论感知效用,结论是基于情感要素的模型显著优于基于语言和结构特征的模型。Ghose 和 Ipeirotis^[15]、郝媛媛等^[19]对主客观性进行了深入研究,发现主客观均衡的评论被读者感知更有帮助。Willemsen 等^[8]采用半自动方法标识观点句以获得情感值,发现源自观点句的情感值显著影响评论感知效用,但作用方向与评价产品类型有关。Salehan 和 Kim^[20]则进一步指出,对于感知效用,情感值只对褒贬兼具的中性评论才有积极作用。对于细粒度情感的效应,Felbermayr 和 Nanopoulos^[21]

利用情感词典构建了细粒度情感向量,验证了细粒度情感在预测评论感知效用时发挥的重要作用。而 Yin^[5]的研究则发现内容中的“愤怒”情绪对阅读者感知评论是否有用无影响,但“焦虑”却会产生显著作用力。

综上,信息与情感是构成评论语义特征的核心要素,直接作用于评论感知效用,对语义特征的深度挖掘,可使研究深入评论“主体”探究影响评论感知效用的机理。

3、文献评述

整体看来,对评论感知效用的研究,视角不同,测度变量定义、特征选取及建模方法各异。

探究评论感知效用影响机理的研究(以文献[11-13]为代表)立足于理论或基于经验,多从帮助阅读者消除认知不确定性,辅助潜在消费者决策的视角来解释引起评论感知效用变化的动因。研究模型以线性回归为主,解释变量少,相互独立,主要源自元特征、评论人,内容分析较浅(评论深度是表征内容质量的主要指标)。对内容分析的简化会使研究结论存有一定局限。

预测评论感知效用或辨别高质量评论为目标的研究(以文献[7,15,18]为代表)更关注评论内容,内容特征被充分挖掘并引入模型,包括语法特征^[12,17]、语义特征^[3,5,8,9,15,18]以及体裁特征^[7,15,18]。语义特征获取富有挑战,自动方法要借助文本挖掘和自然语言处理技术^[7,15],或采用实验方案^[5],或者人工标注手段^[2,8]。由于引入变量多且关系复杂,基于机器学习的支持向量机(SVM)、随机森林(RanF)、支持向量回归(SVR)是较常采用的建模方案,这类研究多以预测精度为评测标准,对影响评论感知效用特征的分析相对粗略,理论依据较弱,解释性不强。尽管如此,内容特征的丰富性及内容特征在预测评论感知效用时发挥的效用使内容越来越受到重视。最新研究^[2,3,5,6,18,20]更强调语义特征的效应,对内容深度挖掘的要求也越来越高。因而,理想的研究方案应该将机器学习与实证分析相结合,克服计量模型对引入变量的限制,同时增强模型的解释力。

在中文领域,主流研究仍是因素分析,特征选择多元化,但基于元数据、评论人及浅层内容的研究居多,深入语义层面的探讨相对少。现有研究中,郝媛媛等^[19]从内容中提取观点句,计算句子的主观度均值及混杂度来解释情感要素对评论感知效用的作用;汪旭晖等^[2]采用人工标注方法从内容中提炼反映评论解释性的指标来解释内容的感知效用;王智生等^[9]抽取高频特征词生成测度内容质量的指标进行评论效用影响因素分析。可以看出,内容分析也逐渐受到关注。自然语言与文本挖掘技术的成熟为基于内容的评论感知有用性研究提供了有力支持。中文领域,内容层面的研究有更多值得探索的空间。

理论基础与模型

1、双路径模型与信息效用

双路径模型是描述信息特征影响个体行为的代表性理论,该理论从中枢路径和边缘路径的视角解释信息对信息接收者态度及行为产生影响的机理。中枢路径上,接受者经过对信息的精密思考形成认知判断;边缘路径上,则依据简单线索进行决策。显然,中枢路径上信息接收者将付出更多认知努力,由此产生的态度与行为相对稳定和持久。基于双路径模型理论,在信息系统研究领域,Sussman 和 Siegal^[22]提出了信息采纳模型。该模型被应用于探讨管理建议被采纳的状况,发现信息质量和信源可信度会显著影响接收者的采纳意愿,接收者对信息的感知效用则对接纳意愿起调节作用。依据信息采纳模型,Davis^[23]进一步验证了信息质量和信源可信度与信息的感知效用正相关。

近年来,诸多研究采用双路径模型对网络用户的信息接受行为和信息感知效用展开探索。目标不同,研究各有侧重。有些研究强调中枢路径上信息深度加工产生的效应,重点分析信息质量;有些研究则对边缘路径上影响用户信息行为的要素感兴趣,关注可信度。本研究的直接目标是实现基于内容的评论效用预测与排名,因而研究要点是中枢路径上反映信息质量的内容要素。

2、面向评论内容的特征要素

研究基于双路径理论模型。在边缘路径上,通过选择研究对象(第三方评论网站的书评),对信源可信度进行一定控制。在中枢路径上,借鉴文献[7]提出的评论质量评测框架,从评论内容中提炼出 5 个指标,共 14 个特征,对评论质量进行描述。5 个指标分别为信息量(informativeness)、信息结构(structure)、语言特质(linguistic)、论据(argument)及情感表达(subjectivity)。

(1) 信息量

信息量反映信息丰富度和内容深度。内容丰富意味着有充分的文字量来描述事物或阐述观点。“长”文本才能帮助阅读者消除认知上的不确定性,提升其信息诊断力,从而感知信息有用。但认知负荷理论同时指出,纳入信息超过人的记忆容量,将引发认知资源分配不足,产生认知超载^[11]。可见,内容充分有助于提升信

息诊断力,但超出认知负荷,则适得其反。可见,信息量对评论感知效用的影响是非线性的。

(2) 信息结构

结构是信息的组织形式。Li 和 Zhan^[24]认为信息的可理解性主要指信息结构是否易于阅读、易被理解与信息是否被接受或推荐正相关。在网络环境下,面对海量资讯,人们的阅读模式倾向于浏览或基于兴趣的“跳跃式”阅读,信息组织形式将直接影响读者的阅读体验,进而对读者感知信息效用产生作用。本研究依据前人研究结论^[7,19]和经验,从评论内容中提炼出针对段落和语句的4个基本特征来表征信息结构。

(3) 语言特质

文本信息的可读性与其语言特质密切相关。文献[7]指出,能够轻松阅读的高质量评论不仅清晰直观地描述了产品特征,抒发了用户观点,其语言风格更趋向于词法单纯,句式简单。根据双路径理论模型,这种趋向反映出阅读者进行信息加工时的一种潜意识行为,以尽量少的认知成本来获取有用信息。显然,能够减轻认知负荷的语言往往易得到更多阅读量,提升感知效用,语言风格与信息的感知效用有关联。本研究中,测度语言风格的特征指标计算借鉴了文献[7]的方法。用文档词频反映用词特点,高频词对应常用词,低频词对应生僻词,平均文档词频可表征语言风格是简单朴素还是别具特色。词性统计量则从词汇多样性的角度描述语言。该指标能够反映用词变化,均值高表明词性丰富,用词复杂,阅读难度增大。

(4) 论据完备性

作为蕴含观点的主观性文本,论据特征为其特有。依据经验,对于评论,感知信息内容是否有用的关键是论述是否得当,论据是否有力。有说服力的评论应兼具主客观两方面,客观层面描述评价对象,主观层面呈现观点。评论被感知是否有用体现在其内容是否有充分的观点陈述和支持观点的客观依据。评论内容的论据完备性和适切度也会增强评论的“说服力”,感知有“说服力”的评论易获得有用投票。据此,本研究从评论内容中提取了一组语法特征测度论据完备性。主观层面取观点词和情态词的统计量,反映观点陈述;客观层面主要取数量词,作为客观论据的代理指标。

(5) 情感表达

对于主观性文本,评论的情感表达同样对其感知效用产生作用。Ghose 和 Ipeirotis^[15]提出两个测度评论情感的特征指标,即基于语句的评论主观度均值和主观度标准差。主观度均值刻画情感值,情感值越高,意味着评论内容包含更多或更强烈的观点或情绪;主观度标准差则反映情感的波动。Ghose 和 Ipeirotis^[15]检验了这两个情感特征对评论感知效用的影响力,主观度均值越高,评论感知效用越低;但主观度标准差越大,即评论主客观表述适当,有述有评,则会对评论感知效用产生积极影响。Ghose 和 Ipeirotis^[15]的研究针对商品评论,本研究面向第三方网站书评,评论内容中的情感因素是否具有同样的作用?值得探索。为此,借鉴 Ghose 和 Ipeirotis^[15]提出的情感特征,本研究从内容中提炼出评论的主观度的均值和标准差,作为情感因素引入模型。

3、预测模型与方法

本研究以预测评论感知效用为核心目标。特征源自内容,由于内容特征多且彼此信息交叠,本文没有选择对自变量有独立性要求的线性回归模型,而是采用了基于机器学习的非线性模型。在以预测为目标的类似研究^[7,15,18]中,支持向量机 SVM、随机森林 RanF 和支持向量回归 SVR 是较常采用的建模方法。SVM 建立在计算学习理论结构风险最小化原则之上,通过在高维空间中寻找超平面实现类分割,其优点是能够处理线性不可分情况,采用不同的内积核函数可形成不同算法,SVM 在文本分类中有出色表现。RanF 属于树模型,是一种利用多个树学习器进行分类和预测的组合算法,分类准确率高,能够评估分类问题中特征的重要性,稳健性强。Ghose 和 Ipeirotis^[15]的研究比较了两种算法在评论有用性判别中的表现,RanF 性能优于 SVM。借鉴 Ghose 和 Ipeirotis^[15]的方案,本研究同时采用了 SVM 和 RanF,但特征重要性分析选用的是基于 RanF 的 Boruta 特征优选算法,该算法全局寻优,有更好的稳定性。评论效用预测实验中则采用随机森林回归,以确保分析的一致性。

(1) 甄别“有用”评论的分类预测模型

根据评论的有用性投票占比,占比得分高的评论被认为“有用”。甄别“有用”评论属分类任务。因变量为 0/1 变量,分别表示评论有用(1)和无用(0)。自变量对应 14 个内容特征,分属 5 个指标。采用逐层建模策略,建立基于内容的有用评论分类器。分别用 SVM 和 RanF 进行实验。考虑到特征变量间复杂的非线性关系,SVM 采用径向基核函数 RBF(radial basis kernel function)。RanF 模型主要涉及决策树数量 m 以及分枝时需考虑的特征数量 k ,较小的 k 更适合复杂关系建模。两个模型通过预实验进行调参。

(2) 评论效用值预测模型

效用预测模型选择随机森林回归。自变量为优选特征,因变量为评论感知效用 ($helpfulness \in [0, 1]$)。RanF 以回归树的形式实现效用预测。若模型包含 m 棵回归树,依据公式 (1) 计算评论感知效用:

$$\hat{f}_{review}(helpfulness) = \frac{1}{m} \sum_{t=1}^m \hat{f}_{review}^t(helpfulness) \quad (1)$$

RanF 回归的结果同样与 k 的选择有关,合适的 k 可减少偏差,得到更稳定的结果。模型性能评价参考均方误 MSE (mean square of error) 和拟合优度 R^2 。 MSE 越小,预测值与实际值(有用性投票占比)的偏差越小,预测越精准。 R^2 越大,模型拟合程度越高,预测能力越强。

(3) Boruta 特征选择算法

在变量多且关系复杂的情况下,特征选择是构建优化模型的重要步骤。本研究选用 Boruta 算法实现特征优选,同时对特征重要性进行探测,以获得最优预测模型。

Boruta 算法基于 RanF,有 4 个关键步骤:①所有特征随机进入模型;②采用 RanF 建立分类器,评估每个特征的重要度,重要度为剔除某特征后,模型预测精度的平均降幅,降幅越大意味着特征越重要;③评估是一个迭代过程,每次迭代,剔除一个特征,检测该特征剔除后预测精度的降幅,即获得该特征重要度,若重要度高于其他特征,特征被模型保留;④所有特征得以确认(保留或拒绝)或运行达到设定的迭代次数上限,算法终止。Boruta 遵循对所有特征进行检测的原则,处理每个迭代过程中表现不佳的特征,从而最大限度地减少了模型误差,最终生成一个最小化的全局最优特征子集。

研究设计与问题描述

R 为评论数据集。评论 r 被表述为特征向量,对应信息量、结构、语言特质、论据以及情感 5 个维度,共 14 个特征变量。评论同时具有“感知效用”($helpfulness \in [0, 1]$);并通过设定区分有用性的阈值 ε ,被赋予类标 ($helpful.or.not = 1/0$)。研究框架如图 1 所示。

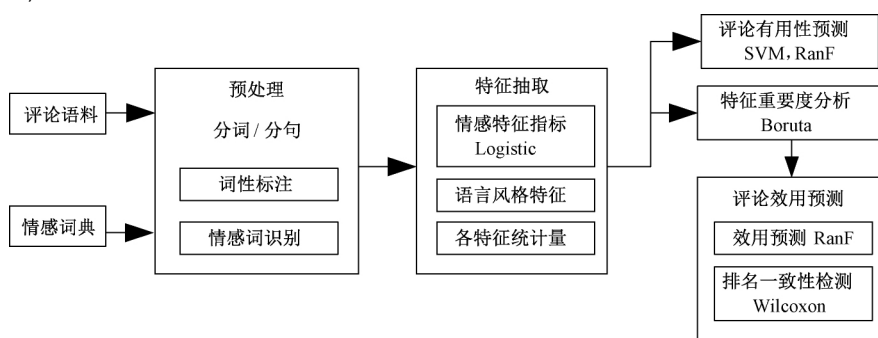


图 1 研究框架

对评论效用的预测本质上是对评论内容质量进行评价。本研究基于内容分析,回答以下三个问题:

- (1) 基于内容特征是否能准确预测评论的感知效用?
- (2) 以预测为目标,哪些内容特征是影响评论的感知效用的重要因素?
- (3) 基于内容的预测效用排名是否能够达到基于有用性投票排名一致的效果?

基于评论内容特征的评论效用预测实验

1、数据与数据采集

研究以书评对象。书评杜撰成分少,文字叙述中有鲜明的语言特色。实验语料源自豆瓣读书频道 (douban.com),豆瓣为第三方评论网站,可信度普遍高于电商网站。本研究根据知名购书网站的图书排名,选择了 14 本文学类畅销书,在 2016 年 10 月至 11 月期间,编写爬虫程序抓取长篇幅书评,共 34928 条。剔除有用性投票少(<3)或总投票量少(<6)的评论,研究用语料共 4647 条评论。显然,大量评论会因投票量过少而被忽视。

2、数据预处理和特征变量

评论语料经过预处理,从内容中提炼特征变量,用 python 编程实现。如表 1 所示。

初始评论语料首先经过分句、分词和词性标注,继而统计词量、句量等指标。情感词借助中文情感词汇本体库^①识别并进行统计。语言风格特征借鉴文献[7]的计算方法。 F_6 为评论中词项文档频率的均值,反映评论用词是简单还是独特。 F_7 为移动取词窗口内词性种类的均值,反映评论用词的变化。情感特征 F_{13} 和 F_{14} 运用了基于机器学习的情感分析法。从评论语料中随机抽取30%的评论语料,分句过滤后,经由5名研究生标注语句的主客观属性($\kappa > 0.7$),生成实验数据。采用词袋模型和Logistic算法构造语句主观度预测模型。运用该模型预测评论句归属主观句的概率,即语句主观度。求评论内容中所有语句的主观度均值和主观度标准差,获得 $F_{13}(\text{avg.subjectivity})$ 和 $F_{14}(\text{dev.subjectivity})$ 两个情感指标(见公式(2))。

$$\begin{aligned} \text{avg.subjectivity}(r) &= \frac{\sum_{k=1}^n \text{sub}(s_k)}{n} \\ \text{dev.subjectivity}(r) &= \frac{\sqrt{\sum_{k=1}^n (\text{sub}(s_k) - \text{avg.subjectivity}(r))^2}}{n} \end{aligned} \quad (2)$$

表1 研究变量

变量类型	变量名	变量测量	变量操作
因变量/ 响应 变量(2)	helpfulness	评论有用性投票量与总投票量的比值,取值[0,1]	评论感知效用值,直接计算
	helpful.or.not	0/1变量, $\text{helpfulness} \geq \varepsilon$, 有用1; 否则没用0	通过预实验确定划分阈值 ε ,将评论的感知效用转换为0/1变量。本研究中设 $\varepsilon=0.8$
	F_1 : num.word	评论篇章的总词量	对预处理后的评论内容进行词量统计。连续变量
	F_2 : num.sen	评论篇章的总句量	
	F_3 : num.para	评论篇章的段落数	对预处理后的评论内容进行句量和段落数统计,直接计算句平均词量以及段落的平均词量。连续变量
	F_4 : avg.word.sen	句平均词量	
自变量/ 预测 变量(14)	F_5 : avg.word.para	段落平均词量	
	F_6 : avg.word.df	评论用词的平均文档频率	统计评论内容中的每个词项的文档频率并求均值;预先设定取词窗口宽度 W (缺省100),对窗口内词项的词性进行统计,滑动窗口遍历整篇评论,计算窗口内词性种类的均值。连续变量
	F_7 : avg.moving.type.ratio	基于移动窗口的词性统计量	
	F_8 : num.opinion	评论篇章的情感词统计量	
	F_9 : num.adv	评论篇章的副词统计量	通过词典匹配标注评论篇章中所有词项词性,获得统计量,并进行标准化处理。连续变量,取值范围[0,1]
	F_{10} : num.modal	评论篇章的情态词统计量	
	F_{11} : num.numeral	评论篇章的数词统计量	
	F_{12} : num.measure	评论篇章的量词统计量	
	F_{13} : avg.subjectivity	评论篇章的主观度均值	运用Logistic回归算法预测单句的主观度,利用公式(2)计算。连续变量
	F_{14} : dev.subjectivity	评论篇章的主观度标准差	

其中 $S_k(k=1, \dots, n)$ 为评论 r 中的语句, $\text{sub}(S_k)$ 为其主观度。 F_{13} 测度情感值, F_{14} 测度情感波动幅度。 F_{13} 得分高表明评论内容偏感性; F_{14} 得分高则表明评论中观点表达和理性陈述相互交错,即有明确观点,也有理性分析。

3、实验与结果

本研究进行了4个实验:实验I确定评论有用性划分阈值 ε ;实验II构建有用评论的分类预测模型,探测内容特征要素对评论效用的影响力;实验III明晰内容特征的重要度及彼此关联,对特征进行优选,构建基于最优特征集的评论效用预测模型;实验IV实现基于预测效用的评论质量排名并检测与投票排名的一致性。数据分析采用R语言。

(1) 实验I: 确定评论有用性划分阈值 ε

通过实验确定评论有用性划分阈值 ε 。 ε 取0.6~1.0的一组值,依据阈值 ε 将评论划分为“有用”和“无用”。引入所有特征构建分类模型,考察模型在 ε 下的表现(F 值),结果见表2。综合考虑模型对“有用”和“无用”评论的分类表现, $\varepsilon=0.8$ 时获最佳 F 值。因此设 $\varepsilon=0.8$,即当评论效用 $\text{helpfulness} \geq 0.8$ 的评论为有

① 中文情感词汇本体库由大连理工大学开发,涵盖27466个中文情感词。见: <http://www.datatang.com/datares/go.aspx?dataid=619774>。

用(1),否则无用(0)。

表 2 有用性划分阈值 ε 设定

ε	0.6	0.7	0.8	0.9
有用评论	0.939	0.886	0.800	0.513
无用评论	0.102	0.211	0.353	0.617
平均	0.520	0.549	0.577	0.565

(2) 实验 II: 分类预测模型 SVM 和 RanF

分类模型考虑了 SVM 和 RanF。先通过实验选取模型参数。SVM 模型主要涉及 $cost$ 和 γ , $cost$ 为调节参数,控制模型偏差与方差的均衡, γ 为径向基核函数的主要参数,影响模型性能。本文利用 R 语言 e1071 包内置的参数优选函数 tune 引入全体内容特征进行实验,采用 10 折交叉验证,以分类错误率 Error 为评测依据,结果见表 3。考虑问题规模不大,对于 RanF,预设决策树 $m=50$,若特征变量数为 p ,对于分枝节点的特征数量 k ,通常取 $[\sqrt{p}, p]$ 区间的若干值进行优选,以 F 和精确度 Accuracy 为评测依据,结果见表 4。表 3 显示, SVM 最佳参数 $cost=10$ 或 100 , $\gamma=0.5$ 。表 4 显示,对于 RanF 最佳 $k=14$ 。基于优选参数构建模型,将每个层面的内容特征依次叠加引入,观测特征项逐层引入后,分类模型的性能表现,分析各层面内容特征对预测评论感知有用性产生的影响。结果见表 5。

表 3 SVM 模型最优参数选择(SVM 采用径向核函数,全体特征)

Error	$\gamma=0.5$	$\gamma=1$	$\gamma=2$	$\gamma=3$	$\gamma=4$
$cost=0.1$	0.137	0.143	0.144	0.145	0.145
$cost=1$	0.084	0.122	0.133	0.134	0.134
$cost=10$	0.074	0.112	0.129	0.131	0.132
$cost=100$	0.074	0.112	0.129	0.131	0.132

表 4 RanF 模型的最优参数选择($m=50$ $p=14$,全体特征)

	$k=4$	$k=6$	$k=8$	$k=12$	$k=14$
Accuracy	0.698	0.698	0.701	0.706	0.709
F	0.802	0.802	0.804	0.807	0.809

由表 5 可见,初始模型只引入信息量,称之为基线。较之基线,随着各层面特征的引入,模型性能指标发生变化。两个模型均显示,内容特征的引入能有效提升模型的预测能力。对于 SVM, F 从 0.785 提升至 0.814;对于 RanF, F 则从 0.715 提升至 0.809。对比两个模型, F 指标反映 SVM 略好,但 RanF 的预测精度更高。另一方面,考察各层面特征的预测力,结构、论据、情感特征的引入均提升 F 。其中,结构特征提升幅度最大(RanF: $F+0.067$; SVM: $F+0.026$),其次是情感特征(RanF: $F+0.014$; SVM: $F+0.002$);论据特征则使 F 降低,暗示论据特征的预测力不强。但这一结果非全局特征优选获得的结论,后续实验运用特征选择算法考察所有特征组合,以获得全局最优预测模型。

表 5 内容的影响力分析

引入特征	RanF($m=50, k=14$)				SVM($cost=10, \gamma=0.5$)			
	Accuracy	Precision	Recall	F	Accuracy	Precision	Recall	F
信息量(F_1)	0.613	0.719	0.720	0.715	0.692	0.683	0.922	0.785
+结构($F_2 \sim F_5$)	0.676	0.730	0.844	0.782	0.696	0.708	0.953	0.811
+语言(F_6, F_7)	0.702	0.741	0.872	0.800	0.696	0.704	0.966	0.813
+论据($F_8 \sim F_{12}$)	0.691	0.729	0.877	0.795	0.687	0.692	0.984	0.812
+情感(F_{13}, F_{14})	0.709	0.737	0.900	0.809	0.688	0.690	0.994	0.814

(3) 实验 III: 特征重要度分析

本实验采用严格的特征优选算法 Boruta,以全局最优为目标,对特征重要度进行分析。算法实现借助 R 语言的 Boruta 包。因变量为 helpful.or.not,同时引入全体特征,经过 60 次迭代运算,算法依序确认 13 个重要特征。实验过程及结果如图 2、图 3 所示。

图 2 记录算法运行过程,横坐标表示迭代序次(classifier run),纵坐标为重要度(importance),线迹记录了

60 次迭代运算中特征变量重要度的变化轨迹。可以看到, 整个运行过程中, 被保留特征(灰色线迹)的重要度为 $[2.5, 15]$, 显著高于被拒绝特征(虚线线迹)的重要度 $[0, 5]$ 。依据 60 次迭代结果绘制的箱线图(图 3) 同样显示, 被保留特征(灰色)对分类性能的影响显著高于被拒绝特征(白色)。两个图中, 深色实线迹和深色箱线图对应影子特征, 是运算过程中算法随机引入的特征, 对最终结论不具参考价值。

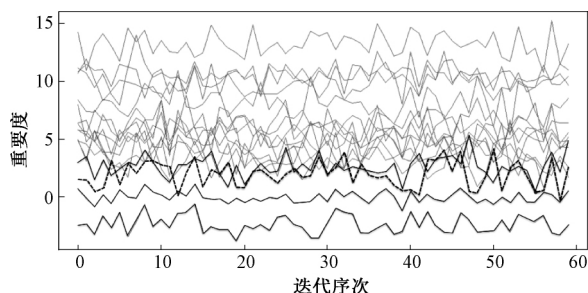


图2 Boruta 特征优选算法运行过程图

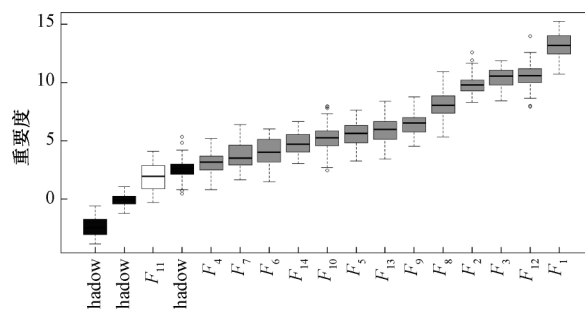


图3 文本特征重要度排序(基于 Boruta 特征选择算法)

进一步观察图 3 可发现, 14 个特征中 13 个被确认重要, 1 个被拒绝。重要特征则大致分三层: 首先是信息量 $F_1(\text{num.word})$ (重要度均值 13.195); 其次是论据特征 $F_{12}(\text{num.measure})$ 和 $F_8(\text{num.opinion})$; 最后是两个结构特征 $F_2(\text{num.sen})$ 和 $F_3(\text{num.para})$ 。这一结果验证了内容要素对评论感知效用的预测力。信息量始终是阅读者感知评论有用的最重要指标, 前人研究^[7, 11-13, 15]用信息量表征内容深度, 从信息诊断力的视角很好地解释了信息量与信息效用存在关联的必然性。论据特征的重要性则印证了观点陈述和客观依据对阅读者感知评论效用的作用。而表征结构的段落数和语句数对阅读体验的影响也会直接引发评论的感知效用的变化, 其作用也不容忽视。

(4) 实验 IV: 基于内容的评论效用排名预测

从 14 本书中随机选取 8 本书评论构建训练语料集, 剩余 6 本书评论为测试集。基于实验 III 获得的最优特征集, 采用 RanF 回归构建评论效用预测模型, 分别对测试集中 6 本书的评论效用进行预测并排名。将每本书评论的效用预测排名与投票排名进行 Wilcoxon 一致性检验。结果显示, 6 本书评论的 MSE 和 R^2 均值分别达到 0.004 和 0.772, 整体效用预测精度令人满意。但排名检验中, 只有《小王子》的预测排名与投票排名获得一致性 ($Sig. = 0.295 > 0.05$), 其他 5 本在统计意义上均显示有差异 ($Sig. = 0.0 < 0.05$)。考虑到有用评论的排名在实际应用中的意义更大, 进一步实验对有用评论子集进行排名预测。结果见表 6。

表6 基于内容的评论效用预测(有用评论)

编号	书名	评论量	MSE	R^2	$Sig.$	与投票排名是否一致
1	白夜行	108	0.0011	0.714	0.000	不一致
2	平凡的世界	73	0.0009	0.813	0.132	一致
3	三体	35	0.0012	0.742	0.280	一致
4	月亮与六个便士	70	0.0006	0.788	0.788	一致
5	追风筝的人	71	0.0009	0.745	0.009	不一致
6	小王子	110	0.0008	0.878	0.315	一致

注: 缺省显著度水平 0.05。

表 6 显示, 对于有用评论, 预测效果更好, MSE 均值小于 0.001, R^2 均值为 0.780。6 本书中, 4 本书的预测排名与投票排名一致, 只有《白夜行》($Sig. = 0.0 < 0.05$) 和《追风筝的人》($Sig. = 0.009 < 0.05$) 的检测结果为不一致。图 4 和图 5 展示两种情况的排名预测结果(采用 R 语言绘制)。

图 4 和图 5 中, 横坐标为基于投票得分的排名序列(rank), 纵坐标为感知效用(helpfulness)。黑色点迹记录评论感知效用的变化轨迹, 浅色折线记录效用预测得分的变化轨迹, 深色实线则是预测线迹的拟合曲线。以基于投票的评论感知效用轨迹为参照(黑色点迹), 浅色折线的波动幅度以及深色实线与黑色点迹的贴近状态能够图示化效用排名预测的准确性。例如, 预测排名与投票排名验证一致的两本书《小王子》和《月亮与六个便士》(图 4), 预测拟合曲线的走势与参照线十分贴近。而不一致的两本书《白夜行》和《追风筝的人》(图 5), 预测分变化轨迹的波动幅度较大, 排名 30 之后, 拟合曲线偏离参照线较远, 说明 30 名后的预测存在偏差。但整体看来, 该模型对有用评论的排名预测能够与投票排名基本保持一致。也就是说, 依据评论内容就能够比较准确地预测评论的感知效用, 达到和用户投票基本一致的效果, 基于评论内容对评论进行排名推

荐具有合理性。

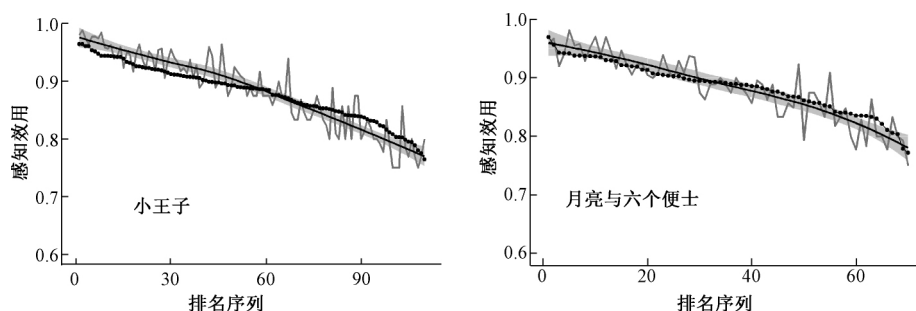


图 4 有用评论的预测排名序列 (有用评论 排名一致)

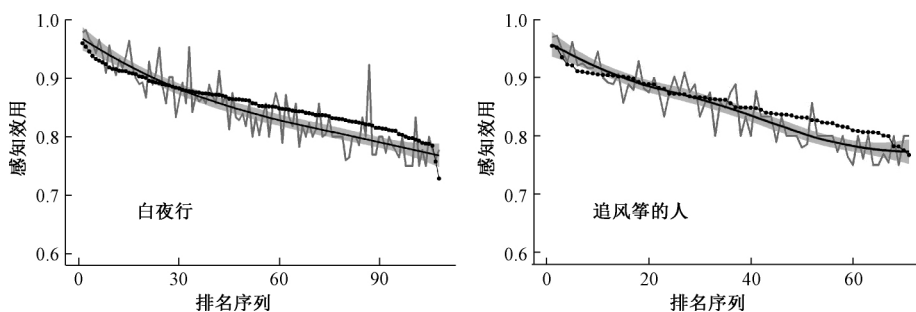


图 5 有用评论的预测排名序列 (有用评论 排名不一致)

研究结论与讨论

本研究以双路径模型为理论依据,探究评论内容对评论感知效用的影响机理,并运用特征工程分析内容特征的重要度,构建预测评论感知效用的回归模型,实现基于“内容”的评论排名推荐。研究结合实证分析,通过三个主体实验对提出的核心问题进行了解答。

1、研究结论

实验 II 验证了内容对评论感知效用有较强的预测力。评论内容与评论被感知有用性关联密切,评论的感知效用能够折射出阅读者对评论内容的认同。

实验 III 从信息诊断力及认知负荷理论的视角解释了评论信息量、组织结构、语言特质、论据表达及情感影响评论感知有用性的机理。结果揭示,信息量是预测信息价值的最重要指标,内容丰富的评论能够提升阅读者信息诊断力,评论被感知有价值。论据反映评论作为口碑的“说服”效应,兼具客观陈述和主观表达的评论易获认同,加强内容中论据的呈现能够有效提升评论感知效用。结构特征表征内容的组织形式,段落数和语句数对预测评论感知效用有重要意义。对于长篇幅书评,良好的组织结构显得尤为重要,层次划分有欠缺的评论影响读者阅读体验,降低读者对评论的感知效用。这些源自理论与经验的假设在本研究获得了支持。但可读性和情感指标被验证与评论感知有用性的关系较弱。分析原因,有其合理性,对文学类评论,读者对语言风格的偏好趋向多元化,相对于语言形式,实质性内容更受关注。情感特征在本研究中作用不突出的原因在于,论据与情感要素存在信息交叠,从全局角度进行特征优选,论据特征表现出更强的预测能力。这一研究揭示内容层面,影响评论感知效用的复杂性。

实验 IV 检验基于内容的预测排名与基于投票排名的一致性。结论是,基于内容的预测排名在有用评论集上准确率较高,低质评论效用预测的不稳定性会导致预测排名与投票排名的一致性。这进一步表明,感知有用的评论,其内容质量通常比较稳定,依据内容评估其效用具有合理性。这使得网站能够从内容层面评估评论的价值,为其进行信息质量控制和有效利用提供依据和可行方案。

2、理论贡献

目前,内容对在线评论感知效用的影响引起研究者关注,但对“内容”的挖掘尚不充分。本研究借助文本挖掘技术深入挖掘评论内容中的特征要素,重点考察“内容”对评论感知效用的影响力。研究基于双路径理论模型,依据信息质量评价框架,从内容中提炼评论的信息量、结构、语言、论据及情感指标,将影响因素分析问题转化为特征优选问题,用分类策略探析影响评论感知效用的内容要素。实验结果揭示出信息量、论据和结构的重要作用,并从信息诊断理论及认知负荷理论的视角进行了合理解释。本研究结合实证分析和文本挖掘技术,是对评论感知效用影响因素研究从内容到方法的补充与完善。另一方面,本研究特别注意了内容特

征的复杂关联,因此选用 RanF 树模型进行分析,实验结果验证了方法的先进性和可行性,为评论感知效用预测模型的构建提供了实践思路。

3、实践启示

基于投票的排名能直接反映读者对评论的感知有用性,但基于投票的排名存在“偏见”,许多高质量评论往往因时间效应无法及时上榜而被埋没。本研究验证了评论内容对评论被感知有用的重要影响力,据此提出评价预测模型,使具备潜在价值的高质量评论资讯有机会呈现,从而能最大程度地利用和发挥信息的价值。其应用价值在于,能帮助互联网企业更好地构建评论质量评测体系,为企业监控与有效利用在线网络资讯提供理论借鉴和实践建议。同时也为评论人撰写评论,获得读者认同与投票提供参考建议,帮助评论人贡献高质量评论资讯,促进在线评论社区的良性发展。

4、局限与后续研究思考

本研究也存在一些局限,有待进一步深入。如,研究以“评论内容”为分析对象,采用文本挖掘技术提炼的特征主要基于词汇,后续研究可在内容主题层面做深层次探讨;另一方面,对边缘路径上的可信度加强控制,分析对象可扩展到电商平台,明确信源真实性的条件下进行质量评价与预测。

参考文献:

- [1] 杨铭,祁巍,闫相斌,等. 在线商品评论的效用分析研究[J]. 管理科学学报,2012,15(5):65-75
- [2] 汪旭晖,聂可昱,陈荣. “解释行为”还是“解释反应”? 怎样的在线评论更有用[J]. 南开管理评论,2017,20(4):27-37
- [3] Chua Y. K. A., Banerjee S. Helpfulness of User-generated Reviews as a Function of Review Sentiment, Product Type and Information Quality[J]. Computers in Human Behavior, 2016, 54: 547-554
- [4] Weathers D., Swain D. S., Grover V. Can Online Product Reviews Be More Helpful? Examining Characteristics of Information Content by Product Type[J]. Decision Support Systems, 2015, 79: 12-23
- [5] Yin D. Anxious or Anger? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews[J]. MIS Quarterly, 2014, 38(2):539-560
- [6] Martin L., Pu P. Prediction of Helpful Reviews Using Emotions Extraction[C]. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014
- [7] Chen C. C., Tseng Y-D. Quality Evaluation of Product Reviews Using an Information Quality Framework[J]. Decision Support Systems, 2011, 50(4):755-768
- [8] Willemssen M. L., Neijens C. P., Bronner F. “Highly Recommended!” The Content Characteristics and Perceived Usefulness of Online Consumer Reviews[J]. Journal of Computer-mediated Communication, 2011, 17(1):19-38
- [9] 王智生,李慧颖,孙悦. 在线评论有用性投票的影响因素研究——基于商品类型的调节作用[J]. 管理评论,2016,28(7):143-153
- [10] 林煜明,王晓玲,朱涛,等. 用户评论的质量检测与控制研究综述[J]. 软件学报,2014,25(3):506-527
- [11] 殷国鹏. 消费者认为怎样的在线评论更有用?——社会性因素的影响效应[J]. 管理世界,2012,12(12):115-124
- [12] Mudambi S. M., Schuff D. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com[J]. MIS Quarterly, 2010, 34(1):185-200
- [13] Nikolaos K., García-Barriocanal E., Sánchez-Alonso S. Evaluating Content Quality and Helpfulness of Online Product Reviews: The Interplay of Review Helpfulness vs. Review Content[J]. Electronic Commerce Research and Applications, 2012, 11(3):205-217
- [14] 闫强,孟跃. 在线评论的感知有用性影响因素——基于在线影评的实证研究[C]. 中国管理科学学术年会,2013
- [15] Ghose A., Ipeirotis P. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(10):1498-1512
- [16] Huang H. A., Chen K., Yen C. D., et al. A Study of Factors that Contribute to Online Review Helpfulness[J]. Computers in Human Behavior, 2015, 48: 17-27
- [17] 闵庆飞,覃亮,张克亮. 影响在线评论有用性的因素研究[J]. 管理评论,2017,29(10):95-107
- [18] Krishnamoorthy S. Linguistic Features for Review Helpfulness Prediction[J]. Expert Systems with Applications, 2015, 42(7):3751-3759
- [19] 郝媛媛,叶强,李一军. 基于影评数据的在线评论有用性影响因素研究[J]. 管理科学学报,2010,13(18):78-88
- [20] Salehan M., Kim J. D. Predicting the Performance of Online Consumer Reviews: A Sentiment Mining Approach to Big Data Analytics[J]. Decision Support System, 2016, 81: 30-40
- [21] Felbermayr A., Nanopoulos A. The Role of Emotions for the Perceived Usefulness in Online Customer Reviews[J]. Journal of In-

- teractive Marketing , 2016 ,36: 60-76
- [22] Sussman S. M. , Siegal W. S. Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption [J]. Information System Research , 2003 ,14(1) : 47-65
- [23] Davis F. D. Perceived Usefulness , Perceived Easy of Use , and User Acceptance of Information Technology [J]. MIS Quarterly , 1989 ,13(3) : 319-340
- [24] Li J. , Zhan L. J. Online Persuasion: How the Written Word Drives WOM-Evidence from Consumer-Generated Product Reviews [J]. Journal of Advertising Research , 2011 ,51(1) : 239-257

Content-specific Ranking Prediction for Online Reviews
——Case of Douban Book Reviews

Nie Hui

(School of Information Management , Sun Yat-Sen University , Guangzhou 510275)

Abstract: In this paper , under the theory of dual-route model , the impact on book review helpfulness exerted by five aspects of its content , namely informativeness , structure , linguistic style , argument and subjectivity , is investigated. Thus , the significant content features can be confirmed and used for review helpfulness modeling. Two models are involved in the study. The interpretation one , built by employing a feature selection algorithm , is used for identifying the content features impacting on review helpfulness significantly; while the tree-based regression model is used for predicting review helpfulness and rank. For interpretation model , the research result indicates that informativeness , structure and argument related features are much more significantly related with review helpfulness. As for prediction model based the optimal features , its R^2 achieves 78% and the error index MSE is less than 0.001. Specifically , the predictive rank is basically in line with vote based ranking for reviews with higher score helpfulness. Overall , all results indicate the helpfulness of a review can be predicted quite accurately according to its content only , which means the study contributes to find out feasible solutions for the review quality control and effective utilization.

Key words: online review , prediction , review helpfulness , text mining