情报学报 ISSN 1000-0135
第34卷第10期1088-1097,2015年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 34 No. 10, 1088-1097 October 2015

doi:10.3772/j.issn.1000-0135.2015.010.009

基于在线评论的用户需求挖掘模型研究[1)]

涂海丽[1,2] 唐晓波[1] 谢 力[1]

(1.武汉大学信息管理学院,武汉 430072;2.东华理工大学经济与管理学院,抚州 344000)

摘要 用户需求挖掘是产品/服务质量提升的重要前提,在线评论真实反映了用户对产品/服务的满意与否。本文针对在线评论数据构建了一个用户需求挖掘模型。该模型首先获取关于某产品/服务的评论数据,经预处理后提取评论文本的主观句;结合构建的领域本体和依存句法分析确定该产品主题属性和相应的主观评论,按产品/服务主题属性对评论内容进行正负向分类;并运用LDA模型对用户评论进行聚类分析,展示用户重点关注主题属性的评价向量及其情感;同时运用KANO模型对分类结果进行KANO转换与评价,对评价结果进行分析,得出用户关于该产品/服务各主题属性特征需求满足情况;在此基础上提出该产品/服务改进的方向。本文以庐山旅游为例,验证了模型的可行性。

关键词 在线评论情感分类 LDA聚类需求挖掘 KANO模型

Research on User Needs Mining Model Based on Online Reviews

Tu Haili[1,2], Tang Xiaobo[1] and Xie Li[1]

(1. *School of Information Management, Wuhan University, Wuhan* 430072;
2. *Faculty of Economics and Management, East China Institute of Technology, Fuzhou* 344000)

Abstract User needs mining is an important prerequisite for the product/service quality improvement, online reviews truly reflect whether users aresatisfied with the product/service. This paper built a user needs mining model aiming at online reviews. Firstly, this paper gets the reviews data about some product/service, after pretreatment, subjective sentences from the reviews text are extracted. With domain ontology customized and dependency syntax analysis, this paper determines the product attributes and its corresponding subjective sentences and classify them to positive and negative ones by product/service themes characteristics. Then, using LDA model, this paper clusters consumer reviews to show evaluation vector and emotion of the theme property that users focus on. At the same time, this paper carries out KANO conversion and KANO evaluation with the classification results using KANO model. Lastly, this paper analyzes the evaluation results, and gets users' need-satisfaction fettle about each theme characteristic of the product/service, based on this, some improvement advice on the product/service is given. Taking Lushan tourist as an example, this paper verifies the feasibility of the model.

Keywords online reviews sentiment classification, LDA clustering needs mining, KANO model

收稿日期:2015年4月2日

作者简介:**涂海丽**,女,1979年生,博士研究生,讲师,主要研究方向:数据挖掘与知识服务,E-mail:69417380@qq.com。**唐晓波**,男,1962年生,教授,博士生导师,博士,主要研究方向:知识组织与情报研究。**谢力**,男,1991年生,硕士研究生,主要研究方向:知识组织与情报研究。

1) 基金项目:本文系国家自然科学基金资助项目"社会化媒体集成检索与语义分析方法研究"(项目编号:71273194)和抚州市社会科学规划项目"基于情感分析的抚州旅游推荐系统构建"(项目编号:14sk035)的研究成果之一。

1 引 言

Web2.0 给网民创造了畅所欲言的机会，网民可以在微博、在线社区、博客、QQ 空间等平台上发表随想、评论事物、晒当前生活。社交网站上每天产生着大量的用户生成内容(UGC)，其中，用户关于产品/服务的评论是极具研究价值的数据，它是用户关于产品/服务的真实看法，通过这些评论内容挖掘用户需求知识，将是商家急需的商业情报，商家可以从中了解用户对其所提供商品/服务的态度，获取用户最新需求，从而改进产品质量或提升服务水平，更好地为用户服务。目前关于用户需求挖掘方法的研究主要是基于数据挖掘的技术和基于心理学建模，前者主要考虑了用户的交易数据，运用分类、聚类、关联规则等数据挖掘技术，重在发现显性需求背后的隐性关联[1~3]；而后者运用用户心理分析模型，借助访谈或问卷调查，旨在发现用户心理深层次的隐性需求[4,5]。这两种方法都有其优势和弊端，优势在于：数据挖掘方法基于客观交易数据，真实可靠，技术先进可行，结论可信；心理学建模触及用户心理深层，从源头出发分析问题，能够挖掘用户需求产生的本质。但弊端也显而易见：前者的交易数据涉及用户隐私，一般研究者难以获取；后者的访谈或调查数据的真实性有待商榷。本文尝试规避以上两种方法存在的弊端，采用用户在线评论数据，综合运用数据挖掘和心理学建模方法来挖掘用户需求。通过数据挖掘技术，对主题评论文本进行情感倾向分类，运用心理学建模对分类文本进行用户需求类型转化，从而分析用户需求。

2 相关研究

用户需求挖掘/获取的研究多见于产品设计、软件开发、数字图书馆服务、信息推荐等方面，但早期的研究大多是基于问卷调查、访谈、网络日志、用户使用行为等数据，较少涉及在线评论数据。在线评论是用户关于某产品/服务的看法、用户体验后的感受或意见，评论的情感倾向有褒义/正面、贬义/负面和中性之分，目前关于在线评论的研究主要集中在主观句识别、产品特征提取、情感倾向判断、评论挖掘系统构建等方面，运用自然语言处理、机器学习、统计学等方法对评论意见进行挖掘建模与自动分类[6~9]。我们认为，在线评论真实反映了用户关于产品/服务的用后感受，中肯地给出了关于产品/服务的褒贬意见，这些意见中蕴藏着用户对于产品/服务满意与否的态度，而且数据量大，内容丰富，是用户需求获取重要的数据源，获取这些数据将提高用户需求数据的完整性。而且，现有的评论挖掘方法为基于在线评论的用户需求挖掘提供了很好的技术支持和方法上的借鉴，使得自动化用户需求获取成为可能。近年来，基于在线评论的用户需求挖掘研究越来越受到学者们的关注。姜巍等提出一种基于复杂网络的评论有用性分析方法，利用评论间的语义关联，从宏观的角度分析评论对于用户需求识别的有用程度，结果发现评论中出现的高频主题特征及其主流意见是准确描述用户需求的评论[10]。Gebauer 等为了发现移动技术的用户需求，运用结构方程模型对在线评论的内容进行分析，发现四个因素显著影响用户需求，分别是是功能性，便携性，性能和可用性，文章还讨论了网上用户评论对用户需求的评估方法的适用性[11]。徐芳平从网络上获取某一特定产品的评论，利用基于关联规则的 Apriori 算法挖掘出产品属性，提取各属性的评价词，对评价词进行模糊化表示，利用模糊的方法计算出该产品各属性的评价值，最后对计算的结果进行分析，得到产品的改进点，从而辅助产品再设计方案的制定[12]。那日萨和钟佳丰利用语义情感计算技术，对用户评论中的显式和隐式属性进行模糊化表示，并与所构建的产品推荐模糊规则结合，实现了基于在线评论的个性化产品模糊智能推荐系统平台的开发[13]。李敏等为了探索用户签到及相关行为的规律及背后动机，更好地了解用户的需求，利用 GooSeeker 抓取国内典型的 LBSN(基于位置的社会网络)嘀咕网的用户数据，使用分类工具 SVMCLS 将用户对签到地麦当劳的评论划分为不同的倾向级别，得到用户对麦当劳的主观情感倾向性，结果表明，用户倾向于在签到地做出正面的评论[14]。周朴雄和陈涛通过对标签间的相似性进行计算、聚类，以聚类的标签群作为纽带，形成信息资源的语义链条，进而挖掘出用户的需求信息，最后以标签云的方式展示给用户，完成信息推荐过程[15]。何炎祥等提出一种针对用户生成内容和用户关注信息的用户兴趣发掘方法。首先通过启发式初始化的概率潜在语义分析模型训练得到贴近兴趣类别的话题模型，然后抽取可靠的话题并以此构建分类器，对用户的分享数据进行分类，最后根据用户的分享数据分类结果来识别用户的兴趣类别[16]。

从以上的相关研究成果可以看出，目前本领域的研究侧重于运用分类聚类等数据挖掘技术、文本语义分析技术、机器学习方法和统计学方法对获取的在线评论数据进行主题属性识别、情感极性分析和主题-情感识别等研究，在此基础上获取用户需求信息或兴趣。但以上研究并没有明确说明在线评论正负倾向与用户需求之间的关系，也没有从心理角度分析评论信息所代表的用户需求类别。本文认为，并不是所有的评论都反映了用户对于产品/服务的需求，而只有那些具有褒贬情感的评论才反映用户对于产品/服务的满意或不满意的态度。从心理学角度来讲，这种满意与否的态度正反映了用户的期望，也是评论受众决定是否购买该产品/服务的重要依据。一般认为，用户满意度与产品/服务质量特性满足状况之间是正相关的线性关系，即产品/服务质量越好，用户越满意，反之亦然。而日本著名质量管理专家NoriakiKano则认为用户满意度与产品/服务质量特性存在非线性关系，及产品/服务质量越好，用户不一定满意，甚至更不满意。并于1984年建立了关于产品质量认知的心理学模型，表达了质量特性满足状况与用户满意程度的双维度认知关系，该理论得到质量管理和市场需求分析研究领域的高度认同。该模型将产品质量特性分为魅力质量、期望质量、基本质量、无差异质量和反向质量五类[17]。其中，魅力质量是让用户意料之外的、感到惊喜的质量特性，它将大大提高用户满意，但没有用户也不会感觉不满；期望质量是指该质量特性满足，用户满意，否则不满意；基本质量是不需要顾客表达出来的理所当然的期望，它的满足对顾客满意度贡献不大；无差异质量是指用户不关注的质量特性，它的满足与否不会引起用户满意或不满，用户在评论中也不会表达出来；反向质量是指该质量特性充足时，用户反而不满，不充足时反而满意。从这里可以看出，在线评论反映的主要是用户关于产品/服务的魅力质量、期望质量和反向质量特性。段黎明和黄欢认为产品的设计与生产是以用户需求为目的的，因而产品质量特性满足状况可以理解为用户需求满足状况，并认为用户需求分析主要需关注三类需求，即基本需求、期望型需求和魅力型需求三类[18]。综合以上研究结论，基于在线评论所要挖掘的用户需求是期望需求和魅力需求。孟庆良等认为KANO模型中魅力质量和期望质量的判断高度个性化，是用户的隐性知识，通过KANO模型的三个工具（KANO调查表、KANO评价表、KANO结果表），能够挖掘上述五类质量因素，从而将用户隐性知识显性化[19]。如果将用户需求看成用户知识的一种，即需求知识，那么，魅力型需求和期望型需求是用户的隐性需求。我们要做的就是从在线评论中挖掘这些隐性需求。孙寄凌等进一步指出KANO模型中魅力型需求和期望型需求是消费者期望获得满足的需求，这部分需求如果被及时满足将帮助运营者建立竞争优势[20]。因此，基于在线评论的用户需求挖掘重在发现用户关于产品/服务魅力型需求和期望型需求，在此基础上提出该产品/服务优化改进的方向。

综上所述，在线评论中的正负向评论是用户根据自己的体验和认知自由生成，高度个性化，包含了用户关于产品/服务各主题特征好恶的正负向情感，也是关于产品/服务魅力型需求和期望型需求的宿主。在线评论数据中正向情感的主题特征反映了用户对产品/服务的满意态度和需求取向，用户满意的产品/服务的一些方面，会提高用户忠诚度，也会产生口碑效应，蕴含着用户的魅力需求和期望需求，是经营者需要保持或精益求精的方面；负向评论反映了用户对产品/服务的需求得不到满足，蕴含中用户的期望需求，对于用户不满意的方面，会负面影响潜在消费者的购买意愿，也是产品/服务需要改进的方向。基于此，本文构建一个基于在线评论的用户需求挖掘模型，该模型运用自然语言处理及情感分类方法对采集的在线评论数据进行情感极性分类，并运用LDA模型对用户评论进行聚类分析，展示用户重点关注主题属性的评价向量及其情感。同时运用KANO模型对分类集进行KANO转换与评价，得出用户对评论主题的满意情况，分析此满意结果所属的KANO质量特征类型，提出该主题优化改进的方向。本文以庐山旅游为例，验证该模型的可行性。

3 用户需求挖掘模型构建

本模型结合产品评论挖掘模型与用户需求分析的KANO模型，构建基于在线评论的用户需求挖掘模型。该模型按信息处理的先后顺序分为五个部分：数据采集与预处理、主题—情感分类、LDA主题聚类、KANO转换与评价、主题内容改进建议。该模型流程图如图1所示。

3.1 数据采集与预处理

3.1.1 数据采集

本文数据来源于网络社区中用户评论，利用网

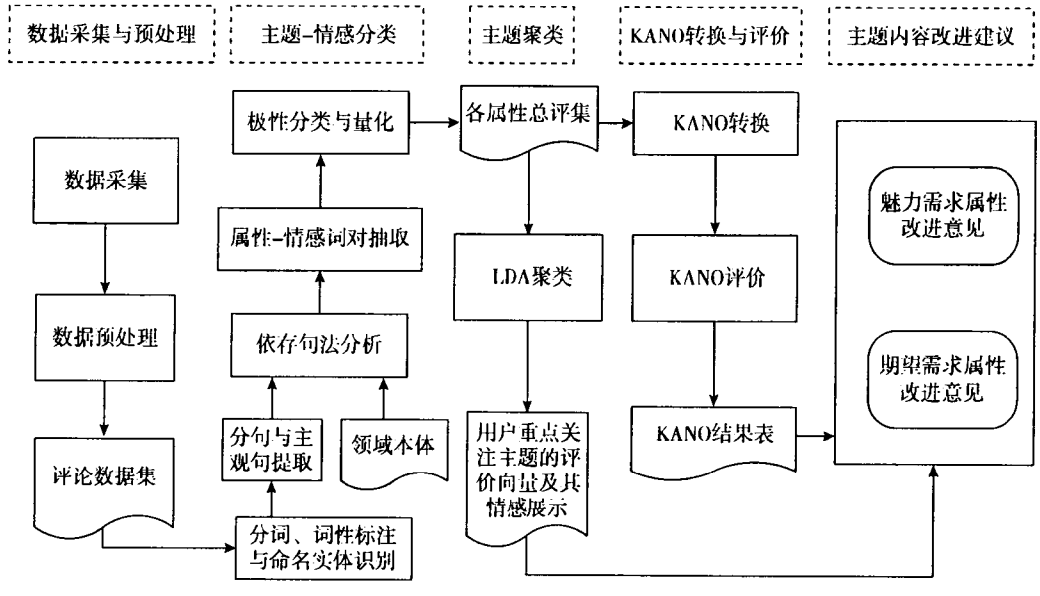

图1 基于在线评论的用户需求挖掘模型图

络信息采集工具获取。为了获取网络社区中用户评论数据，我们首先打开包含网络社区的网站，进入到在线社区模块，运用网站中提供的站内搜索功能，确定关键词后进行在线评论的搜索，然后运用网络信息采集工具"八爪鱼采集器"对搜索结果进行采集，采集的每一条记录内容包括用户名（或用户号）、评论内容、回复数（指回复该评论的用户数），采集到的所有数据导出到EXCEL文件保存。

3.1.2 数据预处理

为了确保采集到的数据对研究问题有价值，需要把一些无用的数据清理掉，以减少噪音数据干扰。通过对网络社区的数据进行分析发现，需要过滤掉的数据主要包括：①与主题无关的信息，有些广告信息或与检索关键词匹配但与主题无关的信息，如键入关键词"庐山旅游"，本来想搜索庐山旅游相关信息，搜索结果中却有"终于见到厦门的庐山真面目了"关于厦门旅游的信息，需要删除。②某一用户多次重复的评论，这种评论数据是用户为了赚取积分等满足自己某个目的而产生，对统计真实正负面评论会造成"虚高"干扰，因而，需将重复记录删除，只保留该用户重复记录中的一条记录作为该用户关于某主题的评论数据。经过预处理后的数据为与检索关键词匹配的评论集，将其保存以备下一步使用。

3.2 主题—情感分类

这一步的目的是得到检索主题各属性的正负总评集。基本思路是：将上一步预处理后的评论集的每一条记录进行分词和词性标注；并按中文语法规则进行分句，将带有情感词或评价词的句子作为主观句提取；构建评论主题的领域本体；将提取的主观句中的名词逐一与领域本体中概念属性进行匹配，并将相匹配的句子进行依存句法分析（手动补充隐含了概念属性，只有情感词，但与本体概念属性有指代关系的句子，即隐含主题属性的句子，视为符合依存关系），抽取符合依存关系句子的属性－情感词对，分析对情感词进行修饰的否定词和程度副词，依据相关研究中的情感程度定级标准标注其情感极性值，保存该属性及极性值，依次累加，最终获得每个属性的情感累计值。主要步骤包括分词与词性标注、分句与主观句提取、领域本体构建、依存句法分析、属性－情感分类与量化。具体方法如下：

3.2.1 分词与词性标注

分词与词性标注是在线评论情感分析的基础，目前有很多分词工具，而分词的正确率是选择分词工具首要考虑的因素，中科院ICTCLAS3.0分词系统分词正确率达到98.45%，是用户公认的分词系统，具有分词、词性标注、命名实体识别和新词识别等功能，本文运用该系统对预处理后的评论集进行分词、词性标注和命名实体识别，我们主要关注的是名词、形容词、动词和副词，因为这些词对中文情感分析起关键作用。

3.2.2 分句与主观句提取

中文一般以句号、感叹号、问号为一句完整的话

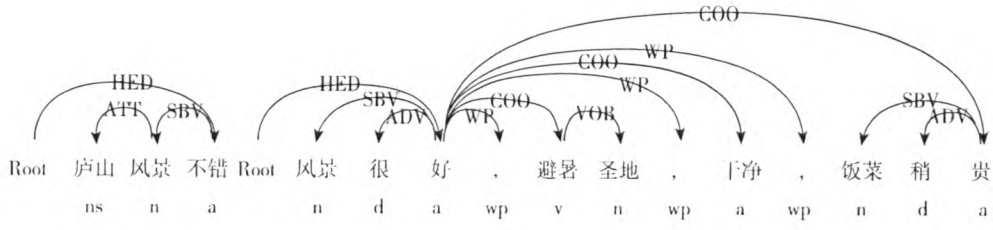

图2 规则一示例图

的断句符号，我们也以这三种符号为分句依据。由于在线评论语言表达的随意性，可能存在一个用户评论内容的末尾没有标点符号或不是断句符号的现象，我们将人工添加上句号，以使该用户的评论内容完整。将词性标注后的单句逐一与 HowNet 词典进行比对，如果不包含情感词语或评价词语，表示该句可能只是叙述或客观描述句，没有发表意见，不是主观句，不进行情感分类。否则，表示该句为主观句，我们将这些主观句进行保存，每个主观句保存为一条记录，每条记录除了保存一个主观句之外，还需保存这条记录原用户名/号、主题属性词（初始值为空）、该评论回复数（每个句子的回复数与该句所在的原用户名/号的回复数相同），另外，每条记录有一个自动编号字段。

3.2.3 领域本体构建

一般来讲，领域本体的构建是个复杂的工程，需要领域专家和本体构建专家的通力合作才能够完成，本文将借鉴实验中涉及领域的本体相关研究成果（旅游本体），构建本实验需要的本体库，简化本体构建过程，具体内容见实验部分。领域本体构建好后，本体概念属性就可以表达出来，将上一步提取的主观句中的名词逐一与领域本体中概念属性进行匹配，这里存在三种情况：①主观句中的名词至少有一个与领域本体属性匹配，保存该条记录，对该句做依存句法分析。②主观句中没有显式名词，只有情感词或评价词，将手动分析该句评价对象，如果与领域本体属性匹配，添加该属性到该条记录的主题属性词字段值中，并保存该条记录，对该句做依存句法分析；否则，将不做任何操作。③主观句中的名词与领域本体属性全部不匹配，说明该句评论与主题无关，不做任何操作。

3.2.4 依存句法分析

本文使用哈工大社会计算与信息检索研究中心研制的在线语言技术平台（Language Technology Platform，LTP）对上一步中需做句法分析的主观句进行依存句法分析[21]，得到每一句子的句法依赖关系图，根据句中词与词之间在语义层面上的修饰关系，可以识别和抽取出句中的属性－情感词对。抽取的规则如下：

规则一：匹配领域本体属性集，识别出评价集中的命名实体以及命名实体所依赖的语法元素。若依赖关系为 ATT（定中关系）、ADV（状中结构）或 SBV（主谓关系），而且其依赖的语法元素词性为形容词、其他名词修饰词、副词、动词或者习惯用语，则认为该命名实体为评价对象，其依赖的语法元素对应的词为情感评价词，如图4所示。

规则二：对于隐式主题属性的句子，将3.2.3中添加的主题属性词字段值作为该主题属性，将隐式主题属性显性化，然后按规则一抽取属性－情感词对。

3.2.5 属性－情感分类与量化

按上一步的规则抽取属性－情感词对，这里要特别处理否定词和程度副词对情感极性的影响。①否定词的处理，如果有否定词与情感词存在依存关系，这里可能有三种情况：如果是否定词修饰的是另一个否定词，则为双重否定，抽取的属性－情感词对的极性不变；如果否定词修饰的是情感词，则将抽取的属性－情感词对的极性进行反转；如果否定词修饰的是包含程度副词的情感词，则调整属性－情感词对的极性程度。②程度副词的处理，考察程度副词与情感词是否有依存关系，如果有，调整属性－情感词对的极性程度；否则，属性－情感词对的极性不变。

按属性将抽取的属性－情感词对分类，为了了解所有用户对同一属性的整体情感倾向，我们将情感词、带程度副词的情感词的极性进行量化，参考 HowNet 字典中给出的褒贬情感词的强烈程度，将不同程度分为4个等级，见表1。

表1 情感极性量化标准

程度副词	极性值
太、非常、极其、很、最……	2
较、稍……	1.5
还、欠、勉强……	0.5
没有程度副词的正向情感词	1
没有程度副词的负向情感词	-1

表1中，有程度副词修饰的情感词，其总极性值=程度副词的极性值*情感词的极性值。如"庐山很美"表达的情感值为2*1=2；"门票稍贵"表达的情感值为1.5*(-1)=-1.5。带否定词的情感极性计算按否定词处理办法的实际情况来定。另外，评论的回复数表达了其他用户对该评论的关注与认同，将一定程度调整该评论的情感值，我们对有回复的评论添加一个加权值，该值的计算方法为：加权值=该评论的回复数/总回复数，那么该条评论最终的情感极性值=计算的总极性值*(1+加权值)。

3.3 主题聚类

为了展示在线评论主要关注的主题内容及其情感倾向，我们利用LDA(潜在狄利克雷分配)模型对提取的属性-情感词对进行主题聚类。LDA是一种提取文档中潜在主题的方法，该方法的基本思想是：单个文档可以表示为隐含主题的概率分布，而隐含主题可以表示为词的概率分布[22]。LDA建模的目的是通过建立两个以主题数目为维数的多项式向量但参数不同的狄利克雷分布，计算文档-主题的概率分布和主题-词的概率分布。由于已有文献对LDA原理和计算公式的介绍很多，本文不再赘述。本文的基本方法是：首先运用Java编程将提取的属性-情感词对按记录个数分成若干个文本文档，即一条记录为一个文档，然后通过计算困惑度确定最优主题数，根据经验确定LDA的其他初始参数，通过主题抽取得到主题-词的概率分布。

3.4 KANO转换与评价

属性-情感分类量化后的结果是按属性分类的各属性情感总评集，计算各总评值的平均值，由此可以看出不同属性特征的情感极性值分布，如果该值大于0，表示总体来讲，用户对该属性具有正的情感倾向，该值越大，表示用户越喜欢这个属性；如果该值恰好为0，表示用户对该属性褒贬参半；如果该值小于0，表示用户对该属性具有负的情感倾向，用户对该属性特征不满意，没有达到用户的预期或期望，说明该属性特征有待改进。

此前在相关研究中分析得出，在线评论反映的是用户的魅力需求和期望需求。魅力需求是用户意料之外、感到惊喜的需求，美国营销管理研究大师菲利普·科特勒[23]认为，顾客之所以会对产品/服务特性高度满意或欣喜，是因为其可感知效果超过预期。赋之在线评论中表现为用户对于产品/服务的高度评价，如"太美了"、"非常好"、"特别舒服"等。表1的情感极性量化标准中，带有高强度程度副词的正向极性值为2，与没有程度副词的正向情感词1相乘，加上回复评论加权值(该值介于0~1之间)后，结果必然大于或等于2，因此，对于总评集中情感极性均值大于或等于2的属性，归为魅力质量属性。

期望需求表现为产品质量感知与用户预期之间正相关的线性关系。如果可感知效果低于预期，顾客就会不满意；如"酒店服务不是很满意"、"门票太贵"、"性价比低"等评论。如果可感知效果与期望(比较)匹配，顾客就(比较)满意；如"酒店服务还好"、"门票实惠"、"性价比高"等。因此，在线评论反映的期望需求，既包含情感均值小于2的正向评价，也包含评价值小于0的负向评价，不管是正向还是负向的期望需求，都反映了用户对改进产品/服务的期望，也是一种期望需求。

3.5 主题内容改进建议

以上结果得出了产品/服务的魅力质量属性和期望质量属性分类以及属性特征的情感倾向聚类，该结果反映了研究对象的哪些属性特征是令用户惊喜的(情感极性均值大于或等于2的属性)，哪些是用户比较满意的(情感极性均值小于2但大于0的属性)，哪些是有待改进的(情感极性均值小于0的属性)，以及哪些属性-情感是用户关注比较多的。针对这样的结果，分别提出魅力质量属性和期望质量属性的改进意见：对于魅力质量属性，根据魅力质量原理，产品/服务质量特性是会动态变化的，也就是说，随着竞争产品/服务此消彼长的影响，以及用户对产品/服务的适应和要求的提高，一种质量特性会向着离用户满意要求越来越远的质量特性演变[24]。因此，要保持用户对产品/服务的持续使用，产品/服务的提供者要做的是延长这一演变的时间，也就是持续保持对魅力质量的创造(Attractive Quality Creation)，特别是对于用户关注比较多的属

性，而实现这个目标的途径就是观察用户的日常作为，不断发现并满足用户的隐性需求[17]。对于期望质量属性，是用户期望从产品/服务中获得的体验满足，这些属性做的好，用户感到满意，如果做的不好，用户觉得需求没有获得满足，这部分需求如果被及时满足将帮助产品/服务在同类产品中建立竞争优势，脱颖而出。因此，对这部分属性要尽量做到改进不足，但求更好，提供比竞争对手更好的产品/服务。

4 实验环境与结果分析

4.1 实验环境

本文的实验环境为 Inter(R) Core(TM) 2.4GHz 的 CPU，4G 的内存，500 硬盘的 PC 机。操作系统为 Win7，实验工具为网络数据获取工具八爪鱼 4.1.5，中国科学院分词软件 ICTCLAS，哈工大在线语言平台，统计工具 Excel，Java 集成工具 Eclipse 3.7.3 和 JDK 1.6.0。

4.2 结果分析

首先进行数据采集，我们以“庐山旅游”为关键词，通过网络信息采集工具“八爪鱼采集器”从同城旅游网、携程网、穷驴网的旅游社区上抓取了 9949 条相关评论进行实验，预处理后得到 9751 条。通过中科院分词工具 ICTCLAS 对这些评论进行分词和词性标注，并按中文的句号、感叹号、问号进行分句，对一条评论结束而没断句的加上句号作为断句符号。将词性标注后的单句逐一与 HowNet 词典进行比对，将包含情感词或评论词的句子抽取，并进行保存，获得 10 327 个主观句。根据旅游体验的六大要素：吃、住、行、游、购、娱，参考已有旅游本体构建框架[25,26]和旅游网站中用户评论关注点，构建庐山旅游本体，如图 3 所示。

旅游评论中，主要涉及游客关于旅游目的地的游览、住宿、购物、交通、饮食、娱乐等六大主题的相关属性评价，我们将保存的关于庐山旅游的主观句逐一与旅游本体中各主题的相关属性进行比对，对关于相关主题或属性的评论句子进行依存句法分析，获得属性 - 情感词对 9662 对，部分属性 - 情感词对列表如图 4 所示。

运用 LDA 模型对属性 - 情感词对进行主题聚类，构建特征词文档和词频文档，设主题数目 T 为 5，每个初始值最多迭代 200 次，进行反复迭代，得到的主题 - 词的概率如图 5 所示。

从图 5 的聚类结果中可知，庐山评论主题 - 情感词对，存在 5 个潜在主题，图 4 中只展示了每个主题的部分高频关键词和隶属于该主题的概率信息。可以看出，Topic1 是关于景色和景区服务方面的主题，反映的是庐山景色秀美、壮观，服务周到；Topic2 是娱乐场所温泉、收费和交通方面的主题，主要反映的是温泉好玩、交通还行、收费贵；Topic3 是购票方便性和票价方面的主题，反映的是购票方便、实惠；topic4 是关于具体景点、服务、娱乐的主题，反映了秀峰、瀑布的气势，娱乐的愉悦性；topic5 是关于具体娱乐设施、食宿、服务的主题，反映了设施齐全、服务热情。将这些主题结果与属性 - 情感词对的描述对照，可以看出，根据 LDA 主题模型得到的主题，能够反映出用户关于庐山的主要关注点及其情感倾向。

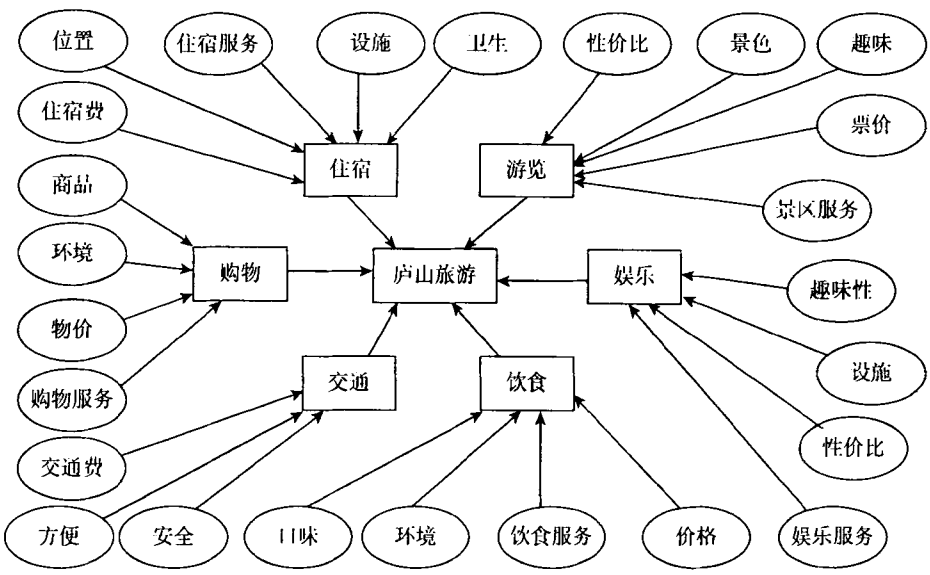

图 3 庐山旅游本体

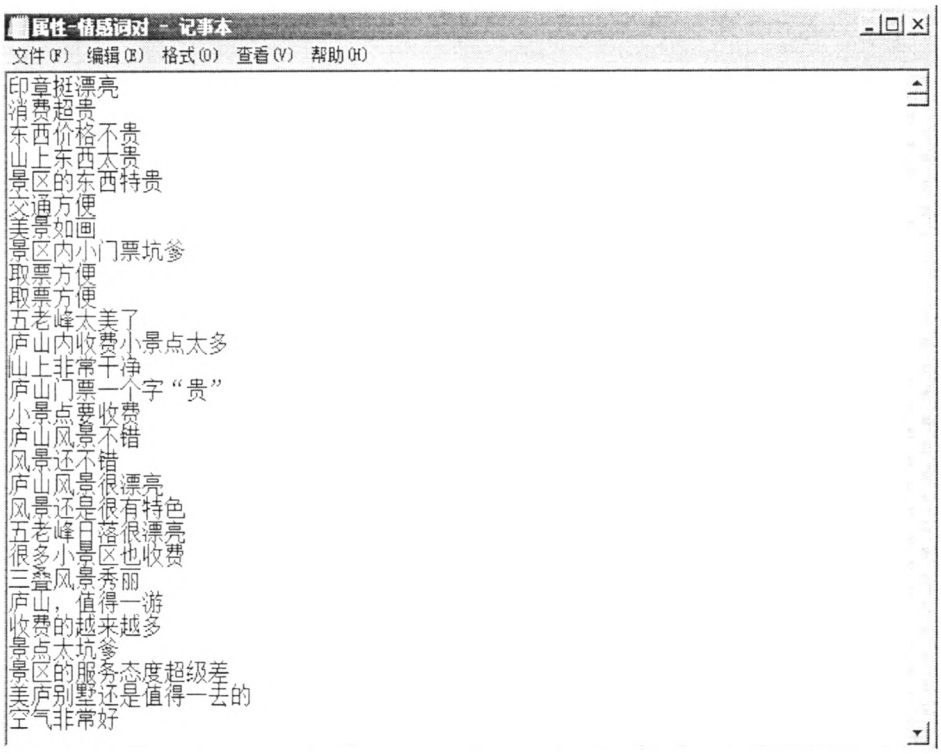

图 4　庐山旅游评论属性 - 情感词对

topic1		topic2		topic3		topic4		topic5	
好	0.537	景点	0.125	方便	0.160	干净	0.030	三叠	0.030
秀美	0.345	庐山	0.089	票	0.097	服务态度	0.021	池子	0.023
景色	0.233	温泉	0.056	实惠	0.062	瀑布	0.017	水	0.021
服务	0.110	设施	0.033	门票	0.047	订票	0.009	票	0.018
环境	0.089	贵	0.030	价格	0.030	秀峰	0.007	性价比	0.018
壮观	0.059	交通	0.025	舒服	0.029	气势	0.005	吃	0.017
整体	0.035	好玩	0.016	便捷	0.024	开心	0.005	很多	0.022
周到	0.033	还行	0.015	购票	0.015	避暑胜地	0.005	缆车	0.010
地方	0.023	收费	0.012	挺好	0.014	电影	0.005	热情	0.009
瀑布	0.012	漂亮	0.011	一般	0.008	酒店	0.005	住宿	0.009
…	…	…	…	…	…	…	…	…	…

图 5　LDA 主题聚类结果

对属性 - 情感词对进行分类和量化，得到按主题和属性分类的庐山旅游评论的总体情感均值，如表 2 所示。从表 2 中可以看出，庐山上的景色、商品、娱乐服务、娱乐设施、娱乐的趣味性、娱乐性价比、住宿位置、卫生等 8 个属性是魅力质量属性，是用户感觉很惊喜或非常满意的质量属性，由 LDA 聚类的结果看出，其中的景色、娱乐服务、娱乐趣味性和娱乐设施是用户重点关注的，特别是一些具体景点，如三叠泉、秀峰、瀑布是用户评论交多的景点；其他属性是期望质量属性，其中游览的趣味、游览的性价比、景点服务、购物服务、购物环境、饮食口味、餐饮服务、餐饮环境、住宿服务、住宿设施、交通的方便性及安全等 12 个属性是用户比较满意的，特别是游览的趣味性、服务是用户评论的较多的主题。值得注意的是，用户对门票价格、购物的物价、饮食价格、住宿费和交通费不满意，普遍认为这些价格太高，而其中的门票贵是用户关注最多的。

表2 庐山旅游评论分类均值

主题	属性	平均值	主题	属性	平均值
游览	趣味	1.93	饮食	口味	1.79
	性价比	1.55		价格	-0.3
	景色	2.27		餐饮服务	1.57
	票价	-0.58		环境	1.45
	景点服务	0.9	住宿	住宿服务	1.98
购物	商品	2.08		设施	1.96
	物价	-0.07		位置	2.08
	购物服务	1.97		卫生	2.28
	环境	1.68		住宿费	-0.08
娱乐	娱乐服务	2.10	交通	方便	0.66
	设施	2.05		交通费	-0.52
	趣味性	2.19		安全	0.4
	性价比	2.02			

针对魅力质量属性，庐山旅游管理部门要不断了解旅游者需求的变化，及时发现用户的需求动向，以提供优质的服务。庐山三叠泉、石门涧、秀峰、美庐等景区优美的风景和富有历史沉淀的景观给久经喧杂、污浊闹市的人们带来哪怕短暂的清静和神怡，也令游客赞不绝口，流连忘返，应该努力保护好这些自然景观和人文景观，避免在这些地方过度开采和开展商业化活动。庐山西海温泉度假村、庐山恋电影院是游客认为不管是在服务、设施还是趣味性、性价比方面都非常满意的娱乐场所，给人或休闲舒适或受教育启迪的美好体验，需要保持不断创新娱乐项目，延续口碑效应。住宿的位置和卫生是游客旅游体验最在意的方面之一，庐山上的这项服务成为魅力质量属性，无疑为庐山这个避暑胜地锦上添花了，这也可以为庐山赢得旅游竞争优势添砖加瓦。对于期望质量属性，其中12个用户比较满意，说明还有提升的空间，要使其向魅力质量属性转化或进一步提高用户满意度，重在提高庐山管理区的管理水平和庐山上经营者的服务意识，5个用户不满意的都是价格方面，突出表现在景区门票重复收费、缆车收费过高、吃住购较贵，这是庐山上相关经营者特别要警惕的地方，这种负面评价会放大用户的不信任感，从而失去很多潜在客户，因此，让价格回归理性才是出路。

5 小 结

在线评论反映了用户关于产品/服务的真实感受，怎样准确分析用户评论并从中挖掘用户需求是竞争情报研究领域的热点，具有重要的研究意义和商业价值。本文根据在线评论语言表达的随意性和关于某一主题评论内容的分散性等特点，将文本挖掘、情感分析、领域本体构建、LDA聚类和KANO评价等方法结合起来，构建了一个基于在线评论的用户需求挖掘模型，该模型提出参照构建的领域本体和提取的主观句，进行依存句法分析，从而抽取属性-情感词对，并提出了用户关于某属性的情感量化计算方法，展示了用户重点关注主题评价向量及其情感，实现了在线评论情感倾向用户需求挖掘的KANO转化，得出用户需求类型，最后针对不同需求类型提出产品/服务改进的建议。从实验结果来看，本文提出的模型是可行的。但是，本文尚存在一些缺憾和不足，本文中隐式评论对象的识别，是通过人为添加的方式来完成的，应在自动识别和准确率上做进一步改进的研究；对于属性-情感词对的抽取方法也应做进一步改进，尽量获取更多的匹配数[27]。另外，由于用户需求受个人、情景等因素的影响，具有多样、动态变化的特征，如何跟踪用户需求的动态变化，是下一步需要研究方向。

参 考 文 献

[1] 胡浩，祁国宁，方水良，等. 基于产品服务数据的客户需求挖掘[J]. 浙江大学学报(工学版)，2009，43(3)：540-545.

[2] 赵军，王晓. 基于数据挖掘的第三方物流中心库存需求预测模型[J]. 物流技术，2014，33(2)：148-150，170.

[3] 刘斌,朱明,王景华,等.基于可拓数据挖掘的用户需求获取研究[J].合肥工业大学学报(自然科学版),2011,34(12):1823-1826.

[4] 孟文,韩玉启,何林,等.基于模糊 Kano 模型的顾客服务需求分类方法[J].技术经济,2014,33(6):54-58.

[5] 张魁,郭钢,陈宓,等.顾客潜在需求心理隐喻引出技术研究[J].市场研究,2006(4):41-45.

[6] Vinodhini G, Chandrasekaran R M. Sentiment analysis and opinion mining: a survey[J]. International Journal of Advanced Research in Computer Science and Software Engineering,2012, 2(6):282-292.

[7] Li C P, Guo L H, Lin N. Value Mining of Product Reviews Based on Sentiment Analysis [C]//Applied Mechanics and Materials, 2015, 1 (713-715): 2528-2531.

[8] 郗亚辉. 产品评论特征及观点抽取研究[J]. 情报学报, 2014, 33(3):326-336.

[9] 史伟,王洪伟,何绍义,等.基于微博的产品评论挖掘:情感分析的方法[J].情报学报,2014(12):149-171.

[10] 姜巍,张莉,戴翼,等. 面向用户需求获取的在线评论有用性分析[J].计算机学报,2013,01:119-131.

[11] Gebauer J,Tang Y,Baimai C. User requirements of mobile technology: Results from a content analysis of user reviews [J]. Information Systems and e-Business Management, 2008,6(4):361-384.

[12] 徐芳平. 基于在线评论的产品再设计需求研究[D]. 大连理工大学, 2012.

[13] 那日萨, 钟佳丰. 基于消费者在线评论的模糊智能产品推荐系统[J]. 系统工程, 2013(11):116-120.

[14] 李敏,王晓聪,张军,等.基于位置的社交网络用户签到及相关行为研究[J].计算机科学,2013,40(10):72-76.

[15] 周朴雄,陈涛. 虚拟社区中基于相似标签聚类的语义信息推荐[J].情报理论与实践,2013,10:100-104.

[16] 何炎祥, 刘续乐, 陈强,等. 社交网络用户兴趣挖掘研究[J]. 小型微型计算机系统, 2014 (11): 2385-2389.

[17] Kano N,Seraku N,Takahashi F,et al. Attractive Quality and Must-Be Quality [J]. Journal of the Japanese Society for Quality Control, 1984, 14(2):147-156.

[18] 段黎明,黄欢. QFD 和 Kano 模型的集成方法及应用[J]. 重庆大学学报,2008,05:515-519.

[19] 孟庆良,邹农基,陈晓君,等. 基于 KANO 模型的客户隐性知识的显性化方法及应用[J]. 管理评论,2009,12:86-93.

[20] 孙宵凌,赵宇翔,朱庆华. 在线商品评论系统功能需求的 Kano 模型分析——以我国主要购物网站为例[J].现代图书情报技术,2013,06:76-84.

[21] Che Wanxiang,Li Zhenghua, Liu Ting. LTP: A Chinese Language Technology Platform[C]//Proceedings of the Coling 2010:Demonstrations , Beijing, China. 2010,08: 13-16.

[22] Blei D M,Ng A Y,Jordan M I. Latent DirichletAllocation [J]. Journalof Machine Learning Research, 2003,3(4-5):993-1022.

[23] 菲利普·科特勒,凯文·莱恩·凯勒. 营销管理:第14版[M]. 王永贵,陈荣,何佳讯等,译. 上海:格致出版社,2012:124-135.

[24] 魏丽坤. Kano 模型和服务质量差距模型的比较研究[J].世界标准化与质量管理,2006,09:10-13.

[25] 冯欣, 王成良. 本体在旅游信息系统中的应用研究[J]. 计算机与现代化,2010,03:128-132.

[26] Marrese-Taylor E, Velásquez J D, Bravo-Marquez F. A novel deterministic approach for aspect-based opinion mining in tourism products reviews[J]. Expert Systems with Applications, 2014, 41(17): 7764-7775.

[27] 唐晓波,王洪艳.微博产品评论挖掘模型研究[J].情报杂志,2013, 32(2):107-111.

(责任编辑 魏瑞斌)