[文章编号] 1003－4684(2020)01-0110-04

基于 R 语言的商品评论情感可视化分析

陈俊宇，郑 列

（湖北工业大学理学院，湖北 武汉 430068）

[摘 要] 通过自然语言处理技术，实现对商品评论数据的可视化流程分析，利用八爪鱼采集器对京东商品评论数据进行采集，对文本数据进行去重、分词、去停词等预处理，再结合 TF-IDF 算法来提取文本数据特征词，利用 R 软件建立 LDA 主题模型并提取主题，使用 LDAvis 可视化工具对主题模型进行交互式可视化分析，并结合词云图将评论文本数据以更直观的方式展现出来，从而挖掘消费者重点关注的评论词语，结合主题模型和词云图两种可视化方法将用户评论情感数据通过丰富的图形进行内容展示，可以使情感分析的结果更准确、更全面反映产品和客户需求，为消费者的购买和商家的改进提供依据。

[关键词] 可视化分析；TF-IDF 算法；词云图；LDAvis

[中图分类号] F724.6 [文献标识码] A

在线评论来自消费者使用产品后的自身感受，能够反映卖家产品质量和服务的好坏。通过情感分析，对评论者的褒贬态度、意见进行判断或评估，从而了解用户对商品的情感态度，评论情感可视化是将评论文本抽象量化的过程，将用户评论情感数据通过丰富的图形或图像进行内容展示，可以使情感分析的结果直观化，便于被有效接纳和应用，进而帮助商家洞察文本数据中隐含的产品信息和顾客需求。本文通过自然语言处理技术，实现对文本数据的可视化流程分析，对评论数据进行高频词汇和主题词的提取，挖掘和分析文本数据所包含的隐含信息，通过对用户的评论进行文本挖掘，能够从大量网络评论中提取反映评论褒贬极性的特质词语，避免消费者所需信息被大量的评论噪音掩盖，从而为消费者的购买决策和企业的营销策略提供支持[1-2]。

1 设计框架与思路

评论情感可视化分析可归纳为数据采集、数据预处理、特征词提取、主题模型建立和情感可视化五个步骤，通过将量化的数据转换为直观感受图形以便大众感知。本文研究的文本对象是京东商城华为荣耀系列 magic2 的评论数据，目的在于通过对购买者的文本评论数据的信息挖掘，获取此类文本数据隐含的消费者关注的评论观点。先对数据进行情感分类，再对数据进行包括文本去重、中文分词，去停用词等预处理；再对预处理后的数据提取特征词，对评论数据进行词频分析并制作词云图，并结合 LDA 主题模型，运用 LDAvis 可视化工具对评论数据作可视化分析，流程见图 1。

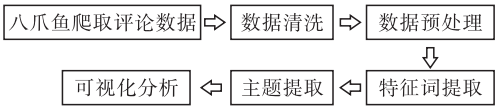

图 1 评论数据可视化分析过程

2 在线评论商品属性的特征抽取

2.1 数据预处理

2.1.1 分词 在对文本数据分析前，需要对文本做分词处理，分词的过程是将连续的词句按照一定的规范重新排列组合，分割成单个词序列的过程。文本分词对后续文本挖掘有着十分重要的影响。本文采用的文本分词方法是基于 R 软件，调用 jiebaR 安装包来实现。jiebaR 包拥有自己的系统词典，且词汇量相当丰富，在此次文本分析中够用[3]。

2.1.2 去停用词 停用词是指某一行业领域目前不再使用的词条。删除这类词，既可以基于现有的停用词词典，也可以根据需要手动建立词典。另外，文本中使用频率不高的非停用词往往对文本特征的表示没有价值，故对这类词也可以进行筛选，即可根据词的长度或出现频率高低进行过滤处理。

[收稿日期] 2019－09－26

[基金项目] 教育部人文社会科学研究规划基金项目(17YJA790098)

[第一作者] 陈俊宇(1994－)，女，江西萍乡人，湖北工业大学硕士研究生，研究方向为大数据分析

[通信作者] 郑 列(1963－)，男，湖北英山人，湖北工业大学教授，研究方向为应用统计

2.2　文本特征提取

特征选择是特征降维的一种技术，目的在于从样本所有特征中筛选出具有区分性和代表性的特征，通过减少无关特征来提高模型的性能。特征选择一般要先构造目标评分函数，然后基于评分函数来筛选出高评分的特征。本文采用的算法是 TF-IDF 算法。其中，TF（Term Frequency）代表词频，IDF（Inverse Document Frequency）表示逆文档频率[4]。

如果一个词在文章中出现多次并且不是停用词，那么在这种情况下，它很可能就代表了文章的特性，也就是要提取的特征词。文本特征提取公式如下：

$$\mathrm{TF}\times\mathrm{IDF}(i,j)=tf_{ij}\times idf_i=\frac{n_{ij}}{\sum_k n_{kj}}\log p\left(\frac{|D|}{1+|D_i|}\right)$$

其中：

$$\mathrm{TF}=\frac{\text{某个词在文章中出现的次数}}{\text{文章总次数}}$$

$$\mathrm{IDF}=\log\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}$$

$|D|$表示语料库中的文档总数。

3　文本主题挖掘

商品评论主题挖掘是从大量评论中找到消费者关注的主题。LDA 主题模型是一种文档主题生成模型，包含词、主题和文档三层结构，通过训练语料，生成文档主题、主题词语概率矩阵，同时 LDA 主题模型也是一种非监督的学习方法，运用词袋模型，将每篇文档视作一个词频的向量，从而识别文档中隐藏的主题信息[5]。

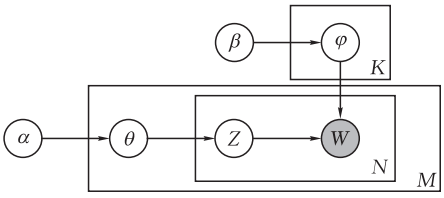

图 2　LDA 模型图

如图 2 所示，假设某个商品评论集由 M 篇评论文档构成，K 为主题的个数，N 为文档的单词总数，α 和 β 分别是文档主题分布和主题词语分布的参数，Z 是文档中生成词 W 的主题，整个过程就是利用“文档-主题”概率分布模型来选取某个主题，然后根据选到的主题，利用“主题-词语”概率分布模型来抽取该主题下的某个单词。不断重复上述步骤，最后形成文档。

4　实证分析

4.1　商品评论数据获取

商品评论中隐含的信息，必须通过获取商品的评论数据分析得到。网络爬虫技术作为一种自动爬取网页、获取网页内容的方式被广泛应用。八爪鱼数据采集器是一种分布式云计算平台，它可以在很短的时间内从不同的网站或网页轻松获取大量的标准化数据，并帮助任何需要从网页获取信息的客户实现自动化采集数据，从而降低获取信息的成本，提高效率[6]。本文利用八爪鱼软件采集了京东商城华为荣耀系列 magic2 手机的在线评论（表 1）。

表 1　商品部分评论数据

ID	评价星级	评价内容	时间
嘉＊＊＊号	star5	帮朋友买的一款手机，据说还挺好的，物流也快，好评好评，必须好评！	2018-11-06 13:47
提＊＊＊娘	star5	颜色大气，外观好看，华为的东西现在越做越好看了，物流很快，还有两样赠品，不错	2018-12-06 23:50
我＊＊＊t	star5	收到货第一时间，就迫不及待的打开了，很惊艳，麒麟 980 速度杠杠的，很流畅，外观很好看，滑盖也感觉牢固可靠，很赞	2019-3-03 08:41
j＊＊＊l	star5	拍照超级赞，就是指纹识别有时候迟钝，拍照整面双摄，拍照不用说运行速度也很好	2019-4-07 10:11
j＊＊＊f	star5	外观漂亮。运行流畅。价格优惠。值得拥有。京东商城服务点赞。送货及时	2019-5-07 09:53

4.2　数据清洗

通过八爪鱼采集器内置的京东评论数据采集规则，共爬取了 magic2 手机从上市至今的购买者评论数据共计 3687 条，针对反爬虫机制出现的重复爬取和噪声数据，需要对文本数据作去重删除处理，将剩余的 2850 条评论数据，最终存储到 txt 文本中，将其作为实验的样本数据。对这些评论数据进行手动分类，类别包括情感倾向为正面、负面、中性的评论以及噪声评论，最终得到如表 2 所示的样本集统计。

表 2　京东手机评论数据样本集

	正面评论	负面评论	中性评论	噪声评论
数量/条	1861	785	173	31
比例/%	65.3	27.5	6.1	1.1

4.3 词频统计

本文采用的文本分词方法是基于R软件，调用jiebaR安装包来实现。jiebaR包拥有自己的系统词典，且词汇量相当丰富，在此次文本分析中够用。在对文本分词之后，全文共分成了36905个词语，但其实这些词中包括了语气助词、副词、介词、连接词等，这些词语没有太大的分析意义，但出现的频率却很高，比如“得、呢、了、还、于是、那么”等。为了避免后期统计词频时增加许多的噪音，所以一般都会将这些词进行过滤处理。本文采用的是哈工大停用词，在筛出了停用词后剩余27277个词。经过分词和去停用词处理后，提取词频如表3所示。

表3 词频统计结果(前10)

关键词	词频	关键词	词频
手机	818	满意	132
指纹	276	很快	131
速度	268	充电	126
屏幕	158	拍照	125
华为	147	流畅	124

4.4 特征词提取

文本处理中一个非常重要的环节是特征词提取，然后由IDF来算出每个词的权重，词语出现的频率越高则IDF值越大。得到"词频"(TF)和"逆文档频率"(IDF)以后，数值相乘即得到这个词的TF-IDF值。一个词对文章的重要性与该词的TF-IDF值大小成正比关系。最后只需要选取TF-IDF值排在最前面的几个词，即为文章的特征词。根据算法，将手机评论的特征词提取出来，大致分为6类(表4)。

表4 属性特征词

属性	特征词
性能	不错、快、运行、正品、续航、拍照
外观	屏幕、惊艳、外观、颜值
体验	指纹、流畅、手感、满意、喜欢、好评
快递	物流、发货、服务态度、快递
价格	价格、赠品
其他	手机、京东、国产

4.5 可视化分析

4.5.1 词云图 对情感分类后的评论数据，在完成分词等一系列预处理操作后，按照词频降序排列，画出排在前100热词的词云图，其中正面评论词云图见图3，负面评论词见图4。

图3 正面评论词云图

图4 负面评论词云图

正面评论词云图中词频较高的词语是手机、华为、不错、喜欢、流畅等，从这些评论中可以看出该款手机的外观、运行速度都给用户带来了很好的体验；负面高频词语主要是退货、售后、降价、问题、划痕等，从这些评论中可以看出手机的硬件、细节及卖家售后等方面并没有让顾客满意，主要原因在于该款手机上市时期正值双十一狂欢节，活动期间商铺推出的优惠活动不一样，导致前后价格波动较大，另外负面情绪还集中在京东售后服务方面，大多数买家对京东非自营卖家的服务态度表示失望，并且在商品退换过程中出现纠纷等问题。

4.5.2 LDA主题可视化 利用LDAvis作为可视化工具对主题模型进行交互式可视化分析，结果为可以交互的html页面，左边面板代表主题气泡，当选定一个主题气泡时，右边面板就变成与选定主题最相关的30个术语，红色横条代表该术语在选定主题中出现的频次，而浅蓝色横条代表该术语在语料料库中出现的频次[7]。

某个词语主题的相关性，由右边面板上方的参数λ来调节，以确定最相关的30个术语是出现频率最高的，还是该主题最独特的。当λ接近1时，在该主题下频繁出现的词跟主题正相关，所以可以通过调节λ的大小来改变词语与主题的相关性。本文选择参数λ为0.8，点击主题1后的可视化结果如图5所示。该主题与手机性能相关，代表主题的主要关键词有“喜欢、清晰、像素、还行、效果、指纹、软件”等。当点击右边面板的术语时，左边面板代表主题的气泡也会随之发生变化，每个气泡的位置不变，但面积变成由该术语在这些主题上的分布比例决定，图6为点击“像素”术语后的可视化结果，可以看出该术语主要出现在主题1中，在主题18中也占少许。

词云图能直观地将高频词汇通过颜色和大小展

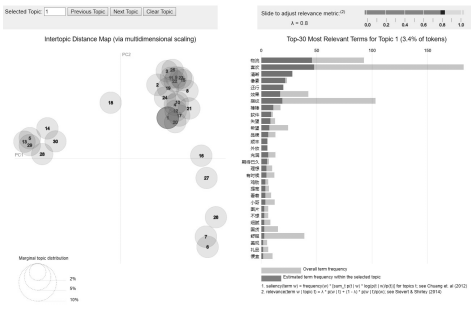

图 5　LDAvis 主题可视化

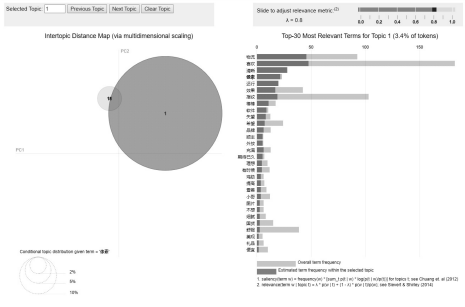

图 6　点击“像素”术语后的 LDAvis 主题可视化

示出来，同时 LDAvis 能够对主题模型进行交互式可视化分析，在高频词汇的基础上能够提取出词语与文本关联度高的主题词，将两种方法结合，能够将文本信息更直观、准确地呈现出来。

5　总结

对商品评论进行特征提取和可视化分析，能够帮助商家指导客户购买产品，而且能够让商家更好地发现用户的需求，进而改进产品，提升用户体验。基于 R 软件对手机评论进行可视化分析方法不仅适用于不同型号的手机评论分析，而且适用于不同种类的商品评论分析。结合主题模型和词云图两种可视化方法将用户评论情感数据通过丰富的图形进行内容展示，可以使情感分析的结果更准确，更能全方面了解产品和客户需求，此外本文的文本数据分类通过人工标注，这是本文的不足之处，后续可在这一模块深入研究。

[参 考 文 献]

[1] 刘 丹.基于客户评论的电商热水器数据分析与挖掘[J].理论观察，2017(11)：98-100.

[2] 刘徽.基于电商评价数据的情感分析[J].内江科技，2018(12)：104-105.

[3] 张梁均，云伟标，王路，等.R 数据分析与挖掘实战[M].北京：机械工业出版社，2015.

[4] 陈雪，朱名勋.淘宝网店铺评论数据高频词挖掘研究[J].金融经济，2018(1)：133-134.

[5] 刘少俊，方延风.基于主题模型的网络信息源可视化分析研究[J].图书情报导刊，2019(3)：32-38.

[6] 陈义.文本挖掘在网购用户评论中的应用研究[D].杭州：浙江工商大学，2018.

[7] 杨斯楠，徐健，叶萍萍.网络评论情感可视化技术方法及工具研究[J].数据分析与知识发现，2018(5)：77-86.

Visual Analysis of Commodity Comment Emotion Based on R Language

CHEN Junyu, ZHENG Lie

(*School of Sciences, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: The visualization process of commodity review data through natural language processing technology is realized. Jingdong commodity review data by using Octopus collector is collected. Preprocessing of the data such as deduplication, word segmentation and stop word on text data is performed, and TF-IDF algorithm is then combined to extract text data feature words. LDA topic model is built and topics are extracted by using R software. Topic models are interactively visualized by using LDAvis visualization tool, and word cloud diagram is combined to display comment text data in a more intuitive way to explore the commentary words that consumers focus on. The innovation of this paper is to combine the theme model and the word cloud map to display the user's commentary sentiment data through rich graphics, which can make the results of sentiment analysis more accurate. More comprehensive understanding of products and customer needs, providing an important basis for consumer purchases and business improvement.

Keywords: visual analysis; TF-IDF algorithm; word cloud map; LDAvis

[责任编校：张　众]