ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.16, No.35, December . 2020
E-mail: lyq@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于Python和Echarts的商品评价文本的可视化设计

穆翠霞

(中华女子学院 计算机系,北京 100101)

摘要:为了用户更直观、全面、高效地了解商品评价信息,以京东为例,设计和实现了商品评价文本可视化。采用八爪鱼采集器进行了数据采集,结合Python和jieba进行了分词和词频统计,Echarts实现了词云图、旭日图和主题河流图等文本可视化形式,帮助用户多角度多形式地了解商品评价情况。

关键词:文本可视化;商品评价;Echarts;jieba

中图分类号:TP311.1 文献标识码:A

文章编号:1009-3044(2020)35-0011-04

开放科学(资源服务)标识码(OSID):

Visualization Design of Commodity Evaluation Text Based on Python and Echarts

MU Cui-xia

(China Women's University, Beijing 100101, China)

Abstract: In order to understand the commodity evaluation information more intuitively, comprehensively and efficiently, taking JD as an example, this paper designs and realizes the visualization of commodity evaluation text. Using Octopus collector for data collection, combined with Python and Jieba for word segmentation and word frequency statistics, realizing text visualization forms such as word cloud, sunburst and theme River, which helps users understand commodity evaluation from different angles.

Key words: text data visualization; commodity evaluation; Echarts; jieba

ChannelAdvisor通过调查发现[1],90%的消费者在购买商品前会浏览在线评论,而且他们中的83%消费者认为最终购买决策会受到在线评论影响。在线评论作为一种口碑形式,通常没有明显的商业目的,更容易获得消费者信赖。Jupiter Research调查数据[2]显示超过90%的大企业相信,在影响消费者是否购买的决定性因素中网民意见是至关重要的。以京东为例,商品评价通常包含好评度、评价标签、评价条数、好评中评差评各自条数、各条评价详情(用户、时间、星级、文本等)等,如图1所示。消费者可以通过好评度、评价标签获得对商品的初步总体印象,通过好评、中评、差评条数情况进一步了解用户对商品的反馈倾向,还可以查看评价详情了解具体评价内容。但是,在查看评价详情时,虽然可以按照默认系统推荐排序,也可以选择按照时间排序,但是评价条数成千上万,不可能依次全部浏览。为了让用户更全面、高效、直观地了解商品评价情况,从而为用户的购买决策提供支持,本文研究商品评价文本的可视化设计和实现。文本可视化的一般流程包括数据采集、数据清洗与预处理、文本分词与统计、数据可视化设计与实现。下文将以京东商城某型号的投影仪评价数据为例,结合具体实现工具、方法和过程,阐述商品评价文本的可视化设计与实现。

图1 京东商品评价

1 数据采集与预处理[3]

八爪鱼采集器是一款免费的网页数据采集软件,使用简单,功能强大,还可以根据软件内置模板进行数据采集。配置八爪鱼采集参数,采集了京东某款投影仪的商品评价原始数据,如图2所示,包括用户账号、级别、评价星级、评价内容、评价日期、评价关键词、评价类型等信息。

会员	级别	评价星级	评价内容	时间	好评度	评价关键词	评论
n***9	PLUS会员	star5	好	2020-08-11 22:33	96%	声音清亮(1282)/简单方便(514) 好评	
l***1		star5	音质音效：非常的好	2020-08-11 21:42	96%	声音清亮(1282)/简单方便(514) 好评	
大***-		star5	很清晰，不刺眼，非常适合孩子看。	2020-08-11 15:51	96%	声音清亮(1282)/简单方便(514) 好评	
jd_ru_jun		star5	外形外观：漂亮，投影亮度：满意	2020-08-11 15:06	96%	声音清亮(1282)/简单方便(514) 好评	
j***6		star5	外形外观：不错投影亮度：非常好	2020-08-11 14:29	96%	声音清亮(1282)/简单方便(514) 好评	
id_13775	PLUS会员	star5	小巧美观，使用方便	2020-08-11 14:21	96%	声音清亮(1282)/简单方便(514) 好评	

图2 采集数据部分

收稿日期:2020-08-28
基金项目:中华女子学院2014年度科研规划课题阶段成果(项目编号:KG2014-0402);中华女子学院2020年度本科教学改革创新项目:基于混合式教学的数据可视化课程建设
作者简介:穆翠霞(1978—),女,山东滨州人,讲师,硕士,研究方向为信息系统、数据挖掘等。

采集京东某款投影仪原始数据共729条(受限于采集软件和京东平台,采集的并非全部评价数据,重点仅在研究可视化设计与实现),去除无效评价记录,包括重复记录(同一用户的相同评价),评价文本与星评不一致的记录,用户未进行文本评价的记录。京东默认4星和5星为好评,2星和3星为中评,1星为差评。评价文本与星评不一致的记录,比如评价文本中出现差评而星评为4星,评价文本为一般而星评为4星或5星。最后保留评价记录共706条,将好评358条、中评150条和差评198条保存为3个不同的txt文件,后面用于设计词云图和旭日图等。另外按照时间顺序将好评、差评数据各自分别保存为5个不同txt文件,后面用来设计主题河流图。

2 文本分词及词频统计[4-5]

下面将利用Python和jieba结巴中文分词实现商品评价文本的分词和词频统计。jieba是一款优秀的Python第三方中文分词库,jieba支持三种分词模式:精确模式、全模式和搜索引擎模式,其中精确模式将语句最精确的切分,只输出最大概率组合,不存在冗余数据,适合做文本分析。

2.1 自定义词典和自定义停用词表

针对投影仪商品评价这一特定文本的分析需求,通过jieba分词效果测试,把部分特定词添加到自定义词典中,比如"侧投""不刺眼""还原度""自动对焦"等。

为了保证可视化效果,去掉一些无效词的干扰,还可以自定义停用词表,这些词对于表达商品评价信息没有实际意义,比如"部分""联系""整体""应该""最后"等。另外,有的用户进行商品评价采用了模板,类似"外观外形:……投影亮度:……",考虑到这些评价分类词汇包括:"外形外观""投影亮度""投影效果""音质音效""操作难易""其他特色"等,不是对商品的实质评价或描述,因此将分类词汇删除,减少对实质评价词汇信息的提取和可视化表达的影响。

2.2 分词及词频统计

Python采用jieba分词工具,调用上文的自定义词典和自定义停用词表,实现分词及词频统计,并按照Echarts可视化实现的格式要求写入文件。实现关键代码如下图3,结合可视化具体需求,对整理好的不同评价文本进行分词和词频统计处理。

```
import jieba
with open('xiaomi1.txt', 'r',encoding='UTF-8') as f:
    xiaomi=f.read()  #读取评价文本文件
jieba.load_userdict("camDict.txt")  #添加自定义词典
seg_list = jieba.lcut(xiaomi, cut_all=False) #调用jieba分词
tf = {} #词频统计
for seg in seg_list:
        if seg in tf:
            tf[seg] +=1
        else:
            tf[seg] = 1

ci = list(tf.keys())
with open('stopword.txt', 'r', encoding='UTF-8') as ft:
    stopword=ft.read()
for seg in ci:  #停用词判断, 词频处理
    if tf[seg]<5 or len(seg)<2 or seg in stopword :
        tf.pop(seg)

ci, num, data = list(tf.keys()), list(tf.values()),[]
for i in range(len(tf)):
    data.append((num[i],ci[i]))
data.sort()      #词频排序
data.reverse()
```

图3　分词及词频关键代码

2.3 评价词标注

以好评文本分词结果为例,将按照用户体验、产品性能、外观外形、其他评价、物流客服、性价高低等6个评价属性类别对分词进行标注,然后按照不同属性类别内部进行词频排序,选择排名前10评价词及词频用于分属性词云图和旭日图可视化设计,如下图4所示。

用户体验541		产品性能526		外观外形181		其他评价98		物流客服43		性价高低16	
不错	108	效果	64	小巧	45	小米	39	物流	12	性价比	16
可以	78	清晰	58	外观	21	京东	19	售后	7		
简单	59	白天	29	美观	14	孩子	13	发货	6		
方便	52	投影仪	28	漂亮	13	很快	8	服务	6		
操作	41	投影	25	外形	10	家里	7	速度	6		
喜欢	32	晚上	23	大方	8	品牌	7	速度快	6		
满意	26	亮度	19	简约	8	家庭	5				
使用	18	色彩	19	设计	7						
容易	15	投屏	19	包装	6						
购买	13	手机	15	简洁	6						

图4　评价词标注示例

3 Echarts可视化设计与实现

ECharts是一个使用JavaScript实现的开源可视化库,提供直观、交互丰富且可高度个性化定制的数据可视化图表,适用于多种不同的可视化场景。本文设计了不同的文本可视化形式,包括词云图、旭日图和主题河流图,试图从多角度多方式地满足用户快速、全面、直观地了解商品评价的需求[4-6]。词云图通过字体大小、位置和颜色等表达不同关键词的重要程度。旭日图(Sunburst)由多层的环形图组成,既能像饼图一样表现局部和整体的占比,又能像矩形树图一样表现层级关系,本文中用来表达对于商品不同方面的评价情况。主题河流图主要用来表示事件或主题等在一段时间内的变化,本文用于表达随着时间推移的评价变化情况。下面将阐述商品评价文本的不同可视化设计与实现。

3.1 不同款商品的评价标签词云图

评价标签能直接通过八爪鱼采集器爬取,如图5所示,然后分别提取其中的标签词和数值,采用JavaScript和Echarts实现词云图,如图6和图7所示。这样可以直观、初步地对比不同产品,比如两个不同品牌的价位相当的投影仪评价标签情况。词云图实现的关键代码如图8所示。

图5　评价标签爬取结果

图6　品牌a好评印象

图7　品牌b好评印象

```
series: [{
    type: 'wordCloud',
    sizeRange: [10, 50],
    rotationRange: [-90, 90],
    rotationStep: 30,
    gridSize: 2,
    shape: 'circle', //五角星

    drawOutOfBound: true,
    textStyle: {
        normal: {
            color: function() {
                return 'rgb(' + [
                    Math.round(Math.random() * 160),
                    Math.round(Math.random() * 160),
                    Math.round(Math.random() * 160)
                ].join(',') + ')';
            }
        },
        emphasis: {
            color: 'red'
        }
    },
    data: xingjia1.sort(function(a, b) { //降序
        return b.value - a.value;
    })
}]
```

图8　词云图关键配置

3.2 好评、中评、差评词云图

根据商品评价的情感倾向，分别将好评、中评、差评文本以词云图形式展示，如图9、图10、图11所示，这样用户可以从视觉上直观、全面地了解不同评价类型的整体情况。比如差评中“客服”“京东”“退货”等较为突出，在好评中“不错”“清晰”“简单”“小巧”“方便”等较为突出。

图9　好评

图10　中评

图11　差评

3.3 好评与差评分属性旭日图[7]

根据评价属性分类后，可以设计和实现好评和差评文本的分属性旭日图。下图12为好评文本的旭日图及下钻效果，通过旭日图可以直观地看到，好评文本中不同的属性包括用户体验、产品性能、外观外形、物流客服、性价高低、其他评价等所占比例情况，还有不同属性中的各评价词占比情况。

比如好评文本的用户体验中“不错”“简单”“方便”，产品性能中的“效果”“清晰”“白天”，外观外形中的“小巧”“美观”等都占较大比例。单击某属性比如“用户体验”，通过旭日图的下钻效果可以进一步更清晰地查看“用户体验”的评价词比例分布情况。另外，通过对比好评和差评的旭日图，用户可以清晰地对比了解用户满意和不满意的方面主要集中在哪些方面，然后可以进一步查看评论详情来获取相关信息。下图13为差评的分属性旭日图，差评中“物流客服”明显占比增大，“外观外形”明显占比减小。

图12　好评旭日图

图13　差评旭日图

3.4 好评与差评分属性词云图

按照不同的属性分类分别设计词云图并进行对比，对好评和差评文本实现分属性词云图，结果如图14和图15，可以进一步直观对比不同属性的评价情况。比如差评中对“外观外形”的评价明显减少，性价高低的评价主要集中在“降价”“价格”“保价”等。

图14　好评分属性词云图

图15　差评分属性词云图

3.5 差评主题河流图[8]

随着时间的变化，评价也可能发生变化，设计主题河流图可以直观展示评价关键词的变化情况，比如差评变化情况如下图16所示，差评中“客服”“退货”“京东”等在3月增多明显。主题河流图的实现关键代码，如图17所示。

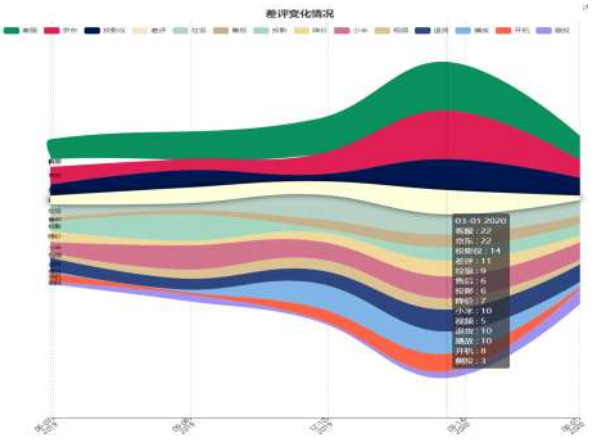

图16　差评主题河流图

```
    legend: {
        top: '6%',
        left: '2%',
        data:['客服', '京东', '投影仪', '差评', '垃圾', '售后', '投影',
        '降价', '小米', '视频', '退货', '播放', '开机', '侧投']
    },
    singleAxis: { //单轴
        type: 'time', // 轴类型
        top: 50,
        bottom: 50,
        left:'10%',
        maxInterval: 3600 * 24 * 90 * 1000, //ms 保证坐标轴分割刻度最大为90天。
        axisTick: {}, //轴刻度
        axisLabel: {
            rotate: 45 //标签旋转
        }, //轴刻度标签
        splitLine: { //网格线
            show: true,
            lineStyle: {
                type: 'dashed',
                opacity: 0.2
            }
        }
    },
    series: [{
        type: 'themeRiver',
        label:{ //主题河流中每个带状河流分支对应的文本标签
            show: true
        },
        emphasis: { //高亮格式
            itemStyle: {
                shadowBlur: 20,
                shadowColor: 'rgba(0, 0, 0, 0.8)'
            }
        },
        data: data
    }]
};
if(option && typeof option === "object") {
    myChart.setOption(option, true);
}
```

图17 主题河流图关键配置

4 结束语

对商品评价文本的可视化分析，可以帮助消费者更直观、全面、高效地了解商品情况，从而支持消费者的购买决策，同时也可以帮助商家更好地了解消费者的反馈和需求，进而改进商品和服务等，提升用户购物体验。本文基于Python和Echarts并结合jieba分词，对某款投影仪的评价文本设计了词云图、旭日图、主题河流图等多种可视化形式，让用户多角度更全面地了解商品，而且也适用于其他类型商品的评价文本可视化。但是，本文中数据爬取的完整性以及不同属性评价词的自动标注等有待后续深入研究。

参考文献：

[1] 宋苏娟,彭卫,王冲.基于手机评论数据探究在线评论有用性的影响因素[J].商场现代化,2020(11):1-4.

[2] 曹丽,郭恺强.基于在线评论的网络营销策略研究[J].轻纺工业与技术,2020,49(5):120-121.

[3] 陈俊宇,郑列.基于R语言的商品评论情感可视化分析[J].湖北工业大学学报,2020,35(1):110-113.

[4] 徐博龙.应用Jieba和Wordcloud库的词云设计与优化[J].福建电脑,2019,35(6):25-28.

[5] 李春芳,石民勇.数据可视化原理与实例[M].北京:中国传媒大学出版社,2018.

[6] 韩帅康,江涛,张顺.大数据评论采集分析系统的设计与实现[J].电脑知识与技术,2020,16(4):35-37.

[7] 易小群,李天瑞,陈超.面向评论文本数据的旭日图可视化[J].计算机科学,2019,46(10):14-18.

[8] 百度 Echarts[EB/OL].[2020-05-26].https://echarts.apache.org/zh/index.html.

【通联编辑：谢媛媛】

（上接第10页）

[46] 郭震,刘颖,于福华.FA优化BP神经网络的MEMS陀螺仪温度漂移补偿[J].微纳电子技术.2019,56(10):817-827.

[47] 毛君,郭浩,陈洪月.基于改进萤火虫算法神经网络的刮板输送机减速器故障诊断[J].机械强度,2019,41(3):544-550.

[48] 张明,张树群,雷兆宜.改进的萤火虫算法在神经网络中的应用[J].计算机工程与应用,2017,53(5):159-163.

[49] 吴华伟,张远进,叶从进.基于萤火虫神经网络的动力电池SOC估算[J].储能科学与技术,2019,8(3):575-579.

[50] 刘园园,贺兴时.基于自适应萤火虫算法的BP神经网络股价预测[J].渭南师范学院学报,2019,34(2):87-96.

[51] 彭新建,翁小雄.基于萤火虫算法优化BP神经网络的公交行程时间预测[J].广西师范大学学报(自然科学版),2017,35(1):28-36.

[52] 李敬明,倪志伟,朱旭辉,等.基于改进二进制萤火虫的BP神经网络并行集成学习算法[J].模式识别与人工智能, 2017, 30(2):171-182.

【通联编辑：梁书】