文章编号：2095-2163(2020)10-0170-04 中图分类号：TP393.0 文献标志码：A

基于网络爬虫的用户评论数据分析

瞿 娟，郁舒兰
（南京林业大学 家居与工业设计学院，南京 210037）

摘 要：互联网的繁荣改变了人们的消费方式，网络购物愈来愈多，用户评论也随之增多。用户评论中不仅仅包含着产品的反馈信息，还蕴藏着产品改进的需求。因此，利用网络爬虫和文本分析技术研究电商平台的商品评价中的用户需求具有很重要的现实意义。本文以某购物网站中的一款折叠桌为例，对采集到的评论数据进行多角度分析，为家具企业掌握用户的消费习惯和行为特征，为产品的迭代以及更新提供一个有据可依的参照，最终制定更精准的营销策略。

关键词：用户评论；网络爬虫；文本分析；家具

Analysis of user comment data based on Web crawler

QU Juan，YU Shulan
（College of Furniture and Industrial Design，Nanjing Forestry University，Nanjing 210037，China）

【Abstract】 The Internet boom has changed the way people consume，with more and more online shopping and more user reviews. User reviews not only contain product feedback information，but also contain the need for product improvement. Therefore，it is of great practical significance to use web crawler and text analysis technology to study user demand in commodity evaluation of ecommerce platform. Taking a folding table in a shopping website as an example，this paper analyzes the collected review data from multiple perspectives，so as to provide furniture companies with a reliable reference for the user´s consumption habits and behavior characteristics，and for the iteration and update of products，and finally develop more accurate marketing strategies.

【Key words】 User comments；Web crawler；Text analysis；Furniture

0 引 言

各种各样的网购平台促使人们的消费方式从线下转到线上，用户评论数据呈爆发式增长。用户会在购物平台上发表大量有关产品、服务、物流等个人体验的评价。用户评论是消费者了解产品真实情况的重要途径。目前，个性定制、个性服务已成为主流的趋势，如何通过消费者的评论来挖掘产品的发展趋势，将成为厂商盈利、扩大市场份额的重要手段。能不能迎合消费者的需求，引领产品的发展方向，将关系到一个企业的生死存亡。家具电商在新的商业模式下也面临着艰难的挑战。本文通过大量分析用户在网购平台上的文本数据，有效挖掘出有价值的信息，帮助家具行业在战略、营销或技术上寻找相应的变革机会和发展对策。

本文选取某购物平台中的某款家具产品的用户评论数据，使用现有的网络爬虫和文本分析技术进行分析。通过对评论数据的语义挖掘分析出用户对家具产品的关注侧重点，了解用户对已购家具产品的态度和意见，进而帮助未购买用户全方位了解已购买用户对家具产品的评价，同时也帮助家具企业更好的掌握家具用户的消费习惯和行为特征，把握自身产品的后续优化方向，并制定更加精准的营销策略。以某购物平台的一款折叠桌作为实例，对如何设计爬虫程序获取信息，及对获取的信息快速分析进行了深入探讨与研究。

1 爬虫的设计

1.1 程序需求及分析

网络爬虫程序的开发成功取决于程序是否能够实现用户定制功能，达到预期设计目的。本次研究即以某购物平台的一款折叠桌为例，通过爬虫对当前此款折叠桌的商品评论详情做出科学分析，而受技术、数据库以及服务器的限制，该购物网站只能显示前 100 页内容，故而针对此情况只能从天猫网站中获取该款折叠桌的前 100 页的商品评论内容和评论日期，在程序设计中，具备了较强的针对性。

1.2 爬虫程序设计

(1)爬虫程序设计思路。首先，需要获得所有该款折叠桌网页的源码；其次，在网页源码中寻找出

基金项目：2019 年研究生案例库项目(163104056)。

作者简介：瞿 娟(1994-)，女，硕士研究生，主要研究方向：交互设计、文本挖掘、可视化。
通讯作者：郁舒兰 Email：372468296@qq.com

收稿日期：2020-05-07

与需求相匹配的信息，此时就需要连接爬虫系统和数据库，将每次成功匹配到的信息均存入数据库中，直至所有网页检索完毕。在数据爬取的过程中，针对天猫网站的高度反爬，还要引入适当的反扒策略，以此保证数据爬取的连续性[1]。爬虫程序的流程如图 1 所示。

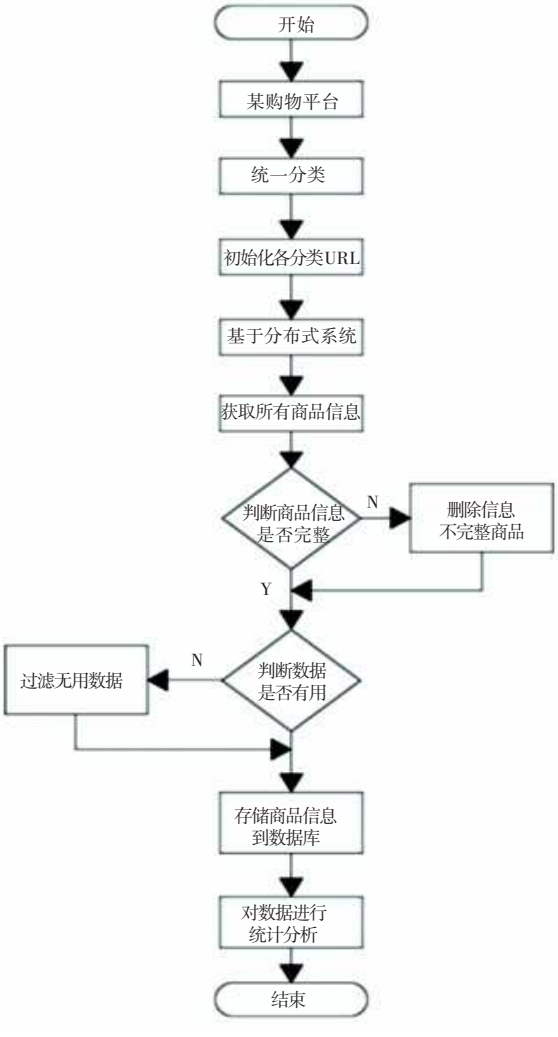

图 1　爬虫程序的流程框架

Fig. 1　The process framework of the crawler

（2）网页抓取。网页抓取是爬虫程序中最重要的部分，由于同一个 IP 在短时间内的多次爬取，会被网站屏蔽，因此采用代理 IP 技术去访问，还需要加入 User Agent 将自己伪装成代理服务器。通过构造代理 IP，每次随机选择访问 IP 与用户代理的搭配，将自己伪装成来自不同 IP 的用户访问，大大降低了被反爬虫的概率。

（3）网页源码分析。在提取好第一层 URL 的源码后，分析当前文本，寻找用户需要的关键信息，根据用户的需求，还需要了解每一类工作的名称与对应网页链接，通过对 Elements 的寻找，发现每一个商品评论都位于 < div class = "tm-rate-fulltxt" > 标签中，每一个评论日期都在 < div class = "tm-rate-date" > 中。将所有的商品评价存入 rateContent 列表，将所有的评价日期存入与 rateContent 列表对应的 rateDate 列表。

（4）信息获取。使用 requests 库实现当前网页解析，同样也可以运用代理 IP 加上用户代理池随机选择与搭配的方法以便能够更加流畅地爬取信息。网页解析 JSON 格式数据，将获取到的页面数据转换为字典类型。

（5）MongoDB 数据库的联合使用。某购物网站上的这款折叠桌的用户评论的信息相对来说是一个比较大的数据，MongoDB 数据库开源，易操作、并且速度、可靠性以及适应性，因此选择 MongoDB 数据库对爬取的数据进行存储。使用 MongoDB 8.0，并通过 pymongo 库去对数据库进行操作，在程序开端，利用 API 建立数据库的链接。

提取网页分析信息主要包括评价日期、评价内容和颜色分类，将这些数据导入所创建的数据库的表中，为下一步的用户评价分析奠定基础。本文利用数据库可视化工具 STudio 3T 展示部分爬取数据如图 2 所示。

2　数据分析

2.1　数据处理

利用网络爬虫程序从某购物网站上爬取了 2019 年 10 月 18 日 11:43 至 2019 年 12 月 13 日 14:35 的所有用户评论（共有 9003 条）数据，采集的内容包括用户评论的发布时间、评论内容、颜色分类等，研究与分析折叠桌的数据研，对用户、家具行业、产品设计师可起到一个初步指导的作用。

通过 Jieba 对读取到的文本数据分词处理，利用现有的停用词词典对评价内容进行清理，去除对句子理解无意义的词，此时可对处理过后的文本数据进行分析。通过 TF-IDF 算法提取关键词，再采用共现分析技术进一步挖掘这些主题词的联系，获取共词矩阵。

2.2　数据分析结果

随着家具行业网购的人数每年不断上升，各个家具品牌店都纷纷建立属于自己的线上销售模式，用户评论数据也大大增加，在这种激烈的行业竞争环境下，如何通过研究激增的用户评论数据分析用户的关注点和行为特征从而改进产品、服务质量即已成为研究的热点与焦点。

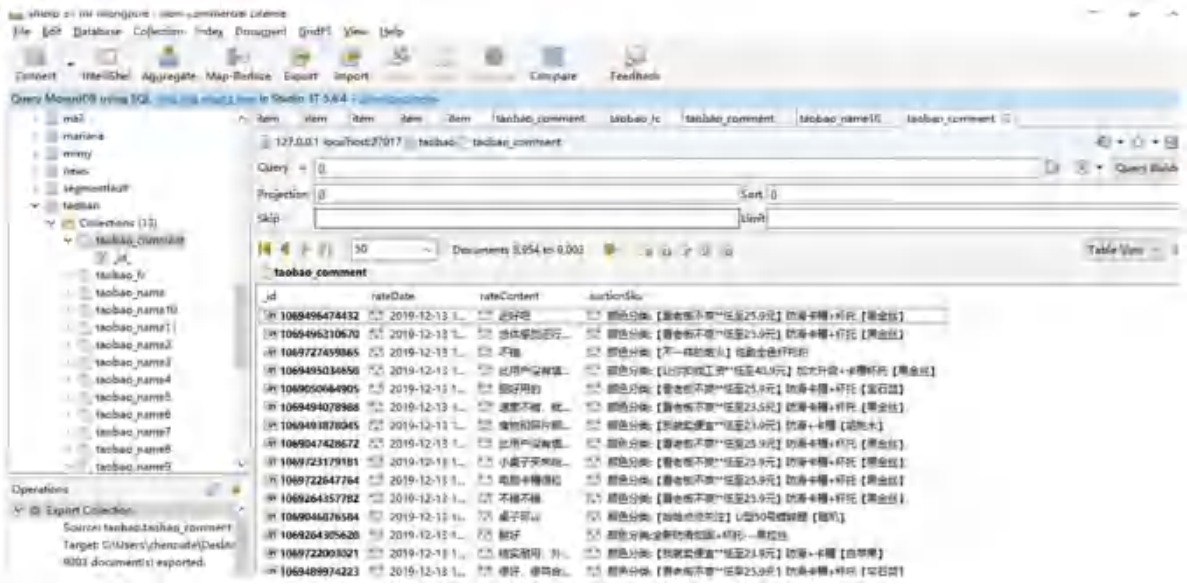

图 2 折叠桌的部分用户评论爬取数据截取

Fig. 2 Partial interception of user comment data from folding table

研究可得,大部分的购买人群比较关注折叠桌的质量、快递、稳定性、包装、外观、颜色、功能等,见表 1。因此后续折叠桌在改良或迭代更新时需要考虑上述方面的问题。至于“床上”一词则表明用户经常在床上使用该款折叠桌。“结实”一词则说明用户普遍认为该款折叠桌较结实。

表 1 天猫折叠桌的网购评论中的高频次关键词(前 20)

Tab. 1 Tmall folding table online comments in the high frequency keywords(The top 20)

序号	词语	词性	词频	频率/%
1	不错	形容词	2156	4.39
2	桌子	名词	1504	3.06
3	质量	名词	1257	2.56
4	喜欢	动词	955	1.94
5	结实	名词	847	1.72
6	床上	名词	703	1.43
7	满意	动词	675	1.37
8	好评	动词	547	1.11
9	快递	动词	532	1.08
10	收到	动词	479	0.98
11	东西	地名	461	0.94
12	好看	动词	440	0.90
13	稳定性	名词	429	0.87
14	产品	名词	412	0.84
15	包装	动词	399	0.81
16	真的	副词	372	0.76
17	外观	名词	366	0.75
18	颜色	名词	365	0.74
19	功能	名词	364	0.74
20	特别	副词	359	0.73

分析高频次关键词的词频统计，可以清楚知道该领域中的研究热点。然而，仅仅依据关键词的出现频次排列，并不能理清这些高频关键词之间的联系，因此采用共词分析的方法来进一步挖掘这些主题词之间的联系，见表 2。研究可得，“质量”、“物流”与折叠桌有较密切的联系，为购买人群比较关注的方面。

表 2　天猫折叠桌的网购评论中的共词矩阵

Tab. 2　Common word matrix in online comments for Tmall folding table

	质量	物流	材质	包装	客服	颜色	功能	服务态度	性价比	工艺水平	不错	好评	桌子	结实	满意	好看	实用
质量	1157	119	47	114	56	83	52	78	37	48	374	68	279	108	136	81	50
物流	119	312	8	68	29	9	7	58	12	4	77	25	98	33	73	15	8
材质	47	8	329	18	7	29	177	1	15	162	101	5	56	214	26	59	38
包装	114	68	18	383	21	29	18	50	15	17	72	29	142	49	81	22	7
客服	56	29	7	21	260	19	10	27	16	11	46	26	114	21	25	18	8
颜色	83	9	29	29	19	335	21	8	18	18	80	19	93	51	45	99	19
功能	52	7	177	18	10	21	356	2	15	171	98	6	58	337	30	55	33
服务态度	78	58	1	50	27	8	2	135	7	3	32	9	40	5	58	5	2
性价比	37	12	15	15	16	18	15	7	129	11	35	9	34	35	19	12	13
工艺水平	48	4	162	17	11	18	171	3	11	278	85	5	39	186	14	45	26
不错	374	77	101	72	46	80	98	32	35	85	1765	68	312	170	88	84	64
好评	68	25	5	29	26	19	6	9	9	5	68	342	72	24	18	17	8
桌子	279	98	56	142	114	93	58	40	34	39	312	72	1482	142	108	129	78
结实	108	33	214	49	21	51	337	5	35	186	170	24	142	649	48	95	51
满意	136	73	26	81	25	45	30	58	19	14	88	18	108	48	573	31	25
好看	81	15	59	22	18	99	55	5	12	45	84	17	129	95	31	405	41
实用	50	8	38	7	8	19	33	2	13	26	64	8	78	51	25	41	313

3　结束语

本文设计了一个基于某购物网站中某款折叠桌的用户评论的网络爬虫数据采集程序，该程序能够登录网站获取页面信息，分析页面中的 URL 链接，同时对筛选构造后的 URL 链接再一次进行数据筛选，将用户获取到的数据存储到数据库，在此基础上将对数据进行深层次的挖掘，即运用一系列的文本数据分析手段，获得关于折叠桌的用户评价中潜藏的一系列重要信息。用户对该折叠桌的关注点主要集中在质量、稳定性、外观、颜色、功能、材质等产品特征上，除此以外还有对快递、包装、物流等服务上也存在较大的关注度。而用户关注的这些产品特征大部分都是折叠桌的产品卖点。因此折叠桌在下一轮的产品迭代研发中，需要加强自身产品的特色，在质量上严格把关，与此同时还需要重新设计包装，并加强工人在包装时的监督管理，确保线上的用户群体在实际收到产品时不会出现破损、污渍等问题。此外，折叠桌的用户群体对于价格并不看重，因此后续可以向较高品质上发展。进一步分析用户对该款折叠桌的潜在需求为：该折叠桌的品质需要进一步提升，能更加结实；对于价格上有较高的追求，不能接受便宜又廉价的产品，可向高品质方向发展；功能上无需多样化，能满足折叠桌的基本功能即可，折叠起来的过程流畅、不卡顿；优化其外观，即从颜色等角度出发；包装上要更仔细，不能在运输途中产生破损、污渍；放置在床上或其他地方时要具备较强的稳定性；在物流服务上，能加快发货速度；在客服服务上，买家在发现货物出现问题时，客服要第一时间处理，安抚购买者的情绪，维护品牌形象，做好售后服务，不可言语激烈，发生不可调和的矛盾等。

参考文献

[1] 杨德清.在线产品社区中的用户需求分析研究[D]. 天津大学，2017.