
第三届“泰迪杯”

全国大学生数据挖掘竞赛

优
秀
作
品

作品名称：基于电商平台家电设备的消费者评论数据挖掘分析

荣获奖项：二等奖

作品单位：韩山师范学院

作品成员：何钢浩 高童强 韦振宇

指导教师：刘 波

基于电商平台家电设备的消费者需求及产品数据挖掘分析

摘要

本文通过对各电商平台的评论用八爪鱼软件进行爬虫式爬取净水器消费数据，运用情感分析，统计分析以及多元回归模型等多种方法从多个角度进行消费者的个性化需求分析，并分析各品牌的特点及现有的电商平台的情况，得出各品牌及电商运营优势及不足，最后从消费者的消费途径分析得出一些消费习惯，针对电商及品牌给出一些意见跟建议。

针对问题一，用网络爬虫工具爬取消费者对净水器的选择及评价，运用 R 语言进行关键词提取各产品主要属性，绘制词云，找出消费者关心的个性属性为价格，售后服务，质量，物流，品牌，利用数理统计对爬取的数据统计汇总得出各属性的高峰属性值，逐步分析，得出消费者的普遍的个性化需求中价格接受尺度为 2500 之间，净水级别 4 级以上的免费安装产品。

针对问题二、三、四，采用情感分析的方法计算各品牌的属性的情感分，绘制表格，得到各属性分值表，据此绘制属性条形图进行比较，可以看出，各品牌的特点显著不同，但总体而言，消费者普遍关心的产品属性，这里品牌优势高于同行其他品牌。又由于不同消费者的不同消费需求，销量也紧随其后，这些不同的属性差异形成了个品牌之间的各自特色。根据问题一得出的个性化需求结果，可以发现 3M 跟美的的销量优势与它的属性跟消费者的个性化需求吻合优势离不开，这也造成了该类产品的销量领先于其他产品，从这一方面看，各品牌的属性化差异是否与消费者的个性需求对接很大程度上影响了消费者的消费行为。

针对最后一个问题，基于多元回归模型分析消费者的购买行为，其拟合的效果较好，一般而言，消费者更多地根据产品的质量及售后服务选择品牌，从购买途径来看，消费者网上购物搜索的关键词以及购买时的关注点与各品牌间的差异化特色紧密关联，而消费者购买后的使用及评价又会在以后反过来影响消费者的个性化需求。

关键词：文本挖掘，情感分析，逐步回归，个性化需求，R 语言

Abstract

The comment on each business platform with Octopus software for crawler crawling water purifier consumption data, using sentiment analysis, clustering model and statistical analysis and other methods from multiple angles were personalized consumer demand analysis, and analyze the characteristics of the brand and the existing business platform, and come to the conclusion that the brand and business operational advantages and disadvantages, finally from the consumer's consumption path analysis draw some consumption habits, the business and brand gives some opinions and suggestions.

To solve a problem, with web crawler tool climbing consumers of water purifier selection and evaluation, using R language to extract keywords from the main product attributes, the word cloud rendering, find out the personality attributes of consumers concerned about the, by means of mathematical statistics and statistics to take up the summary that the attributes of the peak values, clustering model is established on the basis of the analysis, concluded that the consumer generally personalized needs.

Aiming at the problem of the second, the third and the fourth by sentiment analysis method for calculation of properties of various brands of emotion, drawing table, each attribute value table, thus rendering the properties bar charts are compared, as can be seen, significantly different from the characteristics of the brand, but overall, consumers are generally concerned about product attributes, here brand strengths higher than the counterparts of other brands. Because of the different consumer's different consumer demand, the sales volume is closely followed, these different attribute difference forms the respective characteristic of the brand. According to problems come to the demand for personalized results, it can be found with the sales advantage and its property with the consumer demand for personalized identical advantage inseparable. This is caused by the sales of the products leading to other products, from this, the brand attribute differences whether personality demand docking with consumers, to a large extent, affected the consumer behavior.

For one last question, analysis of consumer purchase behavior, from the perspective of the means to buy, purchase of focus and consumer online shopping search keywords and the brand differentiation features closely related, and consumer use and evaluation after purchase will in later come to affect the consumer's personalized demand.

Keywords: sentiment analysis, stepwise regression, clustering analysis, personalized needs, R language

目 录

1. 问题重述	1
1.1.问题背景	1
1.2. 问题提出	1
2. 问题分析	1
2.1. 概论	1
2.2. 模型建立分析	2
3. 模型假设	3
4. 模型的建立与求解	3
4.1. 数据预处理	3
4.2. 模型准备	4
4.2.1. 文本情感分析	4
4.2.2. 回归模型	6
4.3. 模型的建立与求解	6
4.3.1. 个性化需求分析	6
4.3.2. 商家产品优劣势分析	8
4.3.3. 消费者情况反馈机制	11
5. 模型评价	16
5.1. 模型优点	16
5.2. 模型局限与展望	16
6. 参考文献	17

一、问题重述

1.1 问题背景

随着互联网与移动互联网的到来及快速发展，越来越多的海量数据出现在我们的生活中，如何不被海量的数据淹没并从中挖掘到其潜在的价值以成为各行业的研究课题。

就拿电商的出现及强劲的发展势头来说，截止 2014 年 6 月，我国的网民规模达 6.32 亿，互联网普及率为 46.9%，2015 年中国网民的渗透率将接近 50%。2014 年天猫双十一的交易额达 571 亿，网上购物将成为人民生活的一部分。网民在电商平台上浏览和购物，产生了海量的数据，如何利用好这些碎片化、非结构化的数据，也直接影响到企业产品在电商平台上的发展，也是大数据在实际企业经营中的应用。对于用户在电商平台上留下的评论数据，运用文本分析方法，了解用户的需求、抱怨，购买原因以及产品的优点、缺点，对于改善家电设备产品及用户体验有着重要的意义。

据观研天下行业分析：近年来我国家电设备销量增长迅速，以电热水器为例，2011 年电热水器市场销量比 2010 年增长 2.29%，销售额增长 5.23%；2013 年热水器零售量达到 2842 万台，零售额达到 459 亿元，2014 年热水器整体规模向上，但增速较 2013 年有所回落，零售量达到 2985 万台，零售额达到 504 亿元。

1.2 问题提出

根据网络爬取到的各类数据，从以下几个方面对数据进行分析挖掘：

- (1) 分析用户对于热水器/净水器产品的个性化需求；
- (2) 分析现有电商热水器/净水器的产品劣势（用户抱怨点）及产品优势（用户赞点）；
- (3) 分析各品牌的产品间的差异，进行差异化卖点提炼；
- (4) 分析用户购买的原因；
- (5) 对用户的购买行为进行分析挖掘（搜索关键字、购买时关注点、购买步骤、使用、评价）

二、问题分析

2.1 概 论

随着人民生活水平的提高，许多行业提供了比以往任何时代更多的人文关怀，信息时代在带来物质成果的同时，也将逐步关注到了人的需要和个性的充分满足。根据消费者的身份、个人爱好、经济能力的方面选择适合自己的商品，追求消费的个性化现象的出现标志着个性化需求时代的到来。

各商家要做的就是众多的评论数据里运用分析方法，充分把握消费者个性化需求，并结合自身产品的特点定制尽可能符合消费者需求属性的产品。

而消费者从购买时的关注点到使用产品、得出心得过程中途径的选择反过来又将影响到自身需求的逐渐转变以及商家产品定制的不断优化。

以上三个方面相互影响，共同作用于消费者的购买欲望，最终影响消费者的购买行为。为此本文将围绕这三点展开讨论，建立消费者购买欲分析模型（汇总总程序得出所有程序结果分析），分析当下产品的销售现象及潜在的需求前景。

2.2 模型建立分析

本文研究思路如下图 1—1，模型的建立主要包括以下三个步骤：

首先，对消费者进行个性化需求分析。包括以下三步：第一步，产品的属性拟定，指从该产品难以计数的产品属性中根据属性关注程度高低识别出高关注度的属性。第二步，分析属性效用值，通过统计分析对消费者需求属性进行量化，第三步，划分消费者类别，进行聚类分析，获得基于不同属性重要度的消费者类别。

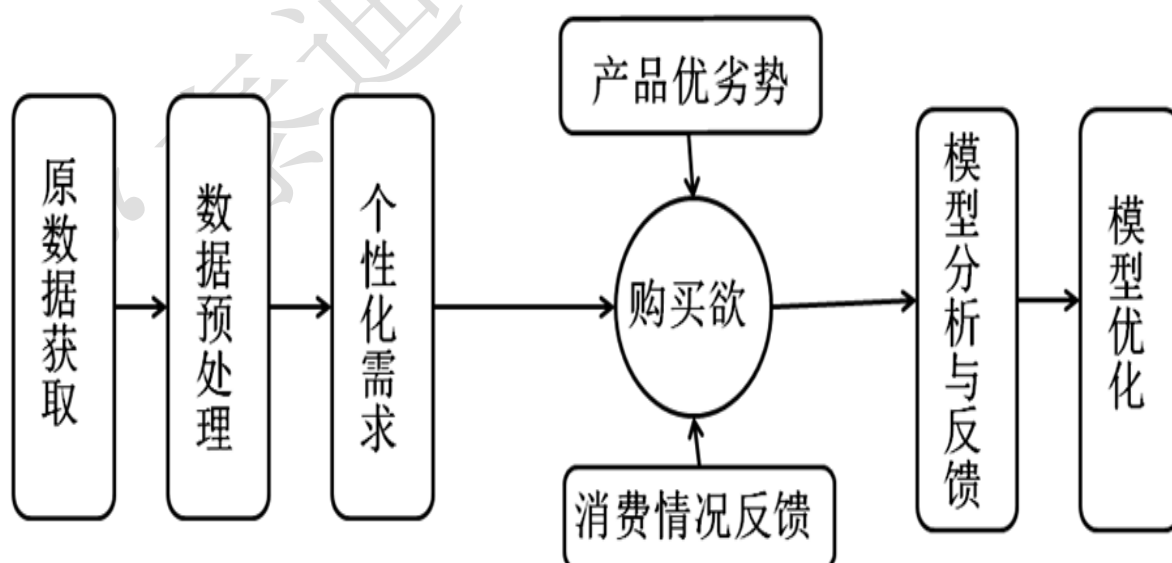
其次，对消费者购买商品的品牌进行情感分析，通过导入情感词典以及赋予的情感权值计算出评语和每个属性的情感分，统计品牌间的属性差异，判断品牌的优劣势，进行差异化卖点分析。

最后，对消费者的购买行为的途径进行分析，得出这个途径中某一行为所表现出的个性需求及产品属性的选择。

综合考虑以上因素对消费者购买欲的影响，建立多元回归模型分析消费者需求的差异性，得出隐藏在数据背后的市场规律。

本文的研究思路如下所示：

图 2.1 本文研究思路



三、模型假设

- 1、假设采集提供的评论具备代表性，能够反应广大消费者的消费分布情况；
- 2、假设分析产品主属性时没列举到的其他属性对消费者的影响水平均相同；
- 3、假设数据的处理有其合理性，能保留普遍消费者的情况；

四、模型的建立与求解

4.1 数据预处理

对于网络数据而言，其一个最鲜明的特点便是数据的随意性大，参差不齐、概念不清、数量级不同等诸多差异给后续的数据分析和挖掘带来诸多麻烦甚至错误。故对原始数据进行按如下步骤处理以提高程序效率及模型的准确性。

1、首先对原始数据用 Excle 中的 VBA 进行分类汇总，实现对不同商家中同一品牌的汇总（程序见附件 VBA.txt）。

2、数据清洗。遗漏值处理、噪音数据处理，不一致数据处理。对于这种情况，可对其进行如下方法处理：忽略该数据、人工填写、使用属性平均值填充。如本数据中的评论文本就对其中的诸多特殊符号用 R 语言清洗。

```
guomei_data$评论<-gsub(pattern="[a-zA-Z\\.\0-9\\$\\^\\&]+", "", guomei_data$评论)
```

3、数据集成和替换。即把多个原数据中的数据整合存放到一个数据存储，或替换掉原始数据的空值，重复值，本文在此基础上引入同义词林，替换原则如下表（详见附件 dictionary.csv）。本文对这种情况通过 R 语言编写去空去重函数对其进行替换处理处理水军（程序见附录文件）。

```
guomei_data_1<-deduplicated(guomei_data$评论)#去重
```

```
guomei_data<-guomei_data[!guomei_data_1,]
```

```
或：guomei_data = unique(guomei_data,fromLast=TRUE)
```

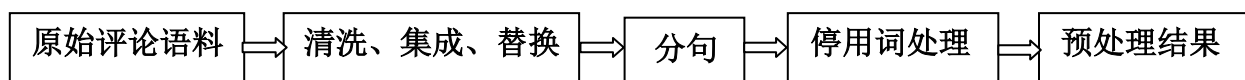
售后词林	物流词林	价格词林	质量词林	品牌词林
售后服务	物流	物美价廉	质量	牌子
送货上门	神速	实惠	品质	名牌
安装	快递	费用	性能	Midea
... ..				
维修	送货	经济	水质	品牌
保修	寄	价格	过滤	正品
服务	快速	大洋	效果	3M

4、根据分析的需求性，用 Excle 对评论数据进行分句处理，并导入停用词表，过滤评论停用词。分句，每条评论可能有多个句子组成，每一句谈到的内容或者产品特征均不相同，用以每条评论为单位来进行产品特征评论语句来分类防止产生混淆。

例：“净水器送货速度很快，净水效果也很好。”

对每条评论语句进行分句处理，这样上面例句就将分为“净水器送货速度很快”和“快递员服务态度好”两个句子，分别可以很好的表示“物流”和“质量”这两个特征。

整个流程如下所示：



综上，对原始数据进行预处理，得到较为规则、便于分析数据评论（部分处理结果见附录）。

4.2 模型准备

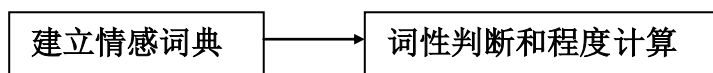
4.2.1 文本情感分析

文本情感分析是自然语言处理研究领域中的一个非常重要的课题，亦可称作意见挖掘，这是对人们的观点，情绪，评论，态度以及针对产品、服务、组织、个体、时间、主题等实体的情感倾向作出有效的挖掘和分析，然后进一步对挖掘出来的信息归纳和推理的一类计数。它代表着一个庞大的问题空间，这也意味着有许多不同的研究内容。目前，文本情感分析研究中的主要内容有：文本情感信息分类，文本情感信息抽取以及文本情感分析技术的应用等。主客观分类和褒贬情感倾向判别、强度分类是文本情感信息研究中非常重要的几个研究课题，而信息抽取方向中的主要研究任务包括观点句中的几类相关要素的抽取，例如：观点持有者的抽取，评价此或者情感词的抽取以及评价搭配的抽取等，在情感分析技术应用的研究方面则主要是基于情感信息分类以及信息抽取为基础的。

文本预处理

情感极性判断与程度计算

利用情感词典进行情感极性和极性程度的计算大致步骤如下图：



建立情感词典：首先我们建立情感极性词性，及正面情感词典和负面情感词典，其中包含情感极性的词，例如“好”、“漂亮”，“差”、“烂”等词。

在得到了情感极性词典之后只能对语句进行情感极性的判断，要计算极性程度还需另外一些词典，其中最重要的是副词词典。

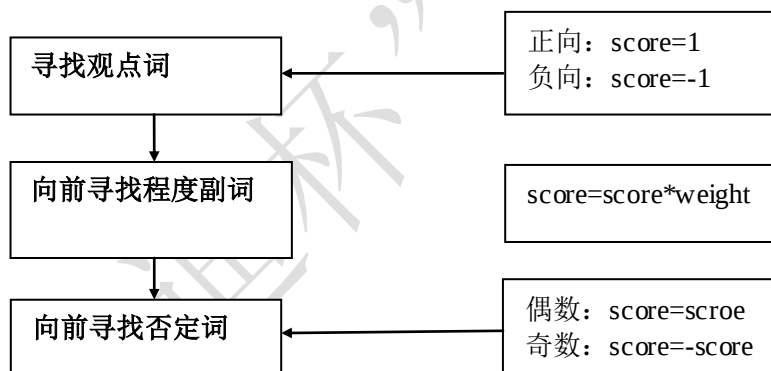
副词词典包含的主要是否定副词和程度副词，否定副词可以在一定程度上提高情感词典对语句极性判断的准确率。

例：“质量不好”，在使用了否定副词词典后结果就为负向；另外通过程度副词比如“较”，“极”，“稍微”、“有点”，并给不同程度的副词赋予不同的权值，可以在情感极性基础上计算极性程度。

完整的情感词典与权值：

词典名称	词典内容	例子	权值
Posidive	正向情感词	好、漂亮	1
Negative	负向情感词	差、烂	-1
Inversedict	否定副词	不、没	-1
mostdict	最级别程度词	绝对、之极	4
Verydict	很级别程度词	分外、很是	3
Moredict	较级别程度词	较为、越发	2
isbdict	稍级别程度词	略为、稍微	0.5
insuffientdion	欠级别程度词	不怎么、半点	0.25

极性程度计算：对含有特征词的语句进行一下极性程度计算，并以均值代表该产品特征的极性程度



根据情感分析原理，有 R 语言实现，程序见附录文件。

算法设计

第一步：读取评论数据，对评论用 Excle 进行分句。

第二步：查找对分句的情感词，记录积极还是消极，以及位置。

第三步：往情感词前查找程度词，找到就停止搜寻。为程度词设权值，乘以情感值。

第四步：往情感词前查找否定词，找完全部否定词，若数量为奇数，乘以-1，若为偶数，乘以 1。

第五步：计算完一条评论所有分句的情感值，用数组（list）记录起来。

第六步：计算并记录所有评论的情感值。

4.2.2 回归模型

回归分析是研究一个变量（被解释变量）关于另一个（些）变量（解释变量）的具体依赖关系的计算方法和理论。从一组样本数据出发，确定变量之间的数学关系式对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著。利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度。

这里：前一个变量被称为被解释变量或因变量后一（些）变量被称为解释变量或自变量或回归变量。由于变量间关系的随机性，回归分析关心的是根据解释变量的已知或给定值，考察被解释变量的总体均值，即当解释变量取某个确定值时，与之统计相关的被解释变量所有可能出现的对应值的平均值。

一元线性回归模型的显著性检验，从总体中随机抽取一个样本，根据样本的 n 对 X 与 Y 的资料导出的线性回归模型，由于受到抽样误差的影响，它所确定的变量之间的线性关系是否显著，以及按照这个模型用给定的自变量 X 值估因变量 Y 值是否有效，必须通过显著性检验才可作出结论，一元线性回归模型的显著性检验包括回归系数 b 的检验和模型整体的 F 检验。

多元回归模型，用来进行回归分析的数学模型(含相关假设)成为回归模型，只含有一个回归变量的回归模型称为一元回归模型否则称为多元回归模型。

4.3 模型建立与求解

4.3.1 个性化需求分析

通过查询百度指数了解净水器的主要属性以及消费者的问题反馈中，，查询 3M，美的和 AO 等网站，以及其他净水器相关的网站。初步确定净水器属性和水平，再通过百度指数和净水器客户的评价展开初步分析，修正后的产品属性及其水平如下：

属性	过滤精度	是否需要电源	过滤级别	自动冲洗功能	价格
水平	0.01 以下	是	五级以上	是	小于 1500
	0.01~0.2	否	五级	否	1500~2500
	0.2 以上		五级以下		大于 2500
属性	安装方式	颜色	滤芯寿命	质量	
水平	厨下式	白色	一年以上	好	
		蓝色	一年以内	一般	
		银色		差	

表 4.1 预调研净水器属性机器水平表

通过网络爬虫获取的产品属性值进行识别，所确定产品的属性应该是影响消费者喜好的突出属性他们对市场而言是重要的信息指标，一个典型的个性化需求分析应尽量选择消费者关心的突出的属性，具备代表这种产品的特色，但在一般情况下地选择不超过6-7个显著因素，过多过少的属性选择分析都不利于模型的精确性和预测能力，此外，确定一些关键信息属性后，还应该确定这些属性恰当的水平，属性与属性水平的个数将决定分析过程中要进行估计的参数的个数，也影响模型的复杂度，这就需要限制属性水平的个数。

此外，通过对数据提供的信息利用八爪鱼软件爬取各型号净水器属性并对其统计分析，得出如下图所示的型号信息表（详见附录文件）：

表 4.2 不同型号净水器属性比较

型号	价格	过滤级别	外包装尺寸	毛重	过滤精度	是否需要电源	自动冲洗功能	销量
沁园 QG-U-1004	649	5 级以上	445×175×493	7	0.01	否	不支持	10839
3M AP3-765S	1499	5 级以上	450*320*270	3.92		否	不支持	52306
3M BEV120	1036	5 级以上	450*320*270	3.94		否	不支持	2660
3M SDW8000T-CN 3M	6899	5 级以上	450*300*500	8	0.2	否	不支持	52633
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
AO AR400-B1	5238	5 级以上	610*430*320	14.5	0.0001	是	支持	25
AO SR50-C3	2688	5 级	475*435*580	16		否	不支持	1141
爱惠浦 EF-900P	1599	2 级	530*170*275	4.14	0.5	否	不支持	4998
美的 MG901-D	1198	4 级	396*264*521	5.9		是	不支持	960

本研究在对净水器各属性进行数据爬取的同时对原数据的评论文本做情感分析,分析消费者对产品属性的选择及属性水平的接受范围,并利用 R 语言分析消费者情感重心,绘制评论的文本词云,确定个性化分析的选择的属性个数及主要属性的属性水平

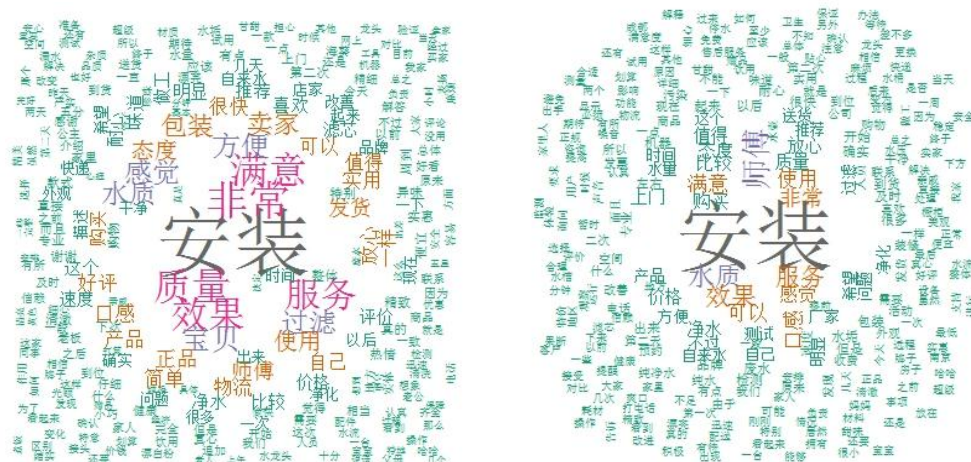


图 4.1 词云频率图

可以认为这些是产品应该具备的需求或功能，确定净水器的质量、服务、价格、物流以及品牌对消费者的影响深远，消费者非常看重这些属性，此外，根据数理统计分析，得到各主属性的属性水平，这一点也将在接下来的论文中继续说明。随着近年对客户需求的定制化，消费者的个性化需求是产品的必备属性，企业应当将可定制的需求重点集中于期望型需求，适度提高满足消费者的个性化需求。

4.3.2 商家产品优劣势分析

由于一种产品往往包括多个属性，每一个属性有包含多个水平值，因此如果漫无目的的设计概念化产品将导致数量的爆炸式增长，由此，分析个品牌及商家的产品优势也有利于从中受到启发，更好地为消费者服务。

如下图所示，根据型号表统计的产品价钱及销量也能从中发现统计分析的结果来看，价格对销售量的影响显著，随着价格的提高，销量明显下降，但随着价格的下滑，销售量在增加一段趋势之后区域平稳，总的来说，价格与需求的关系近似于正态分布。

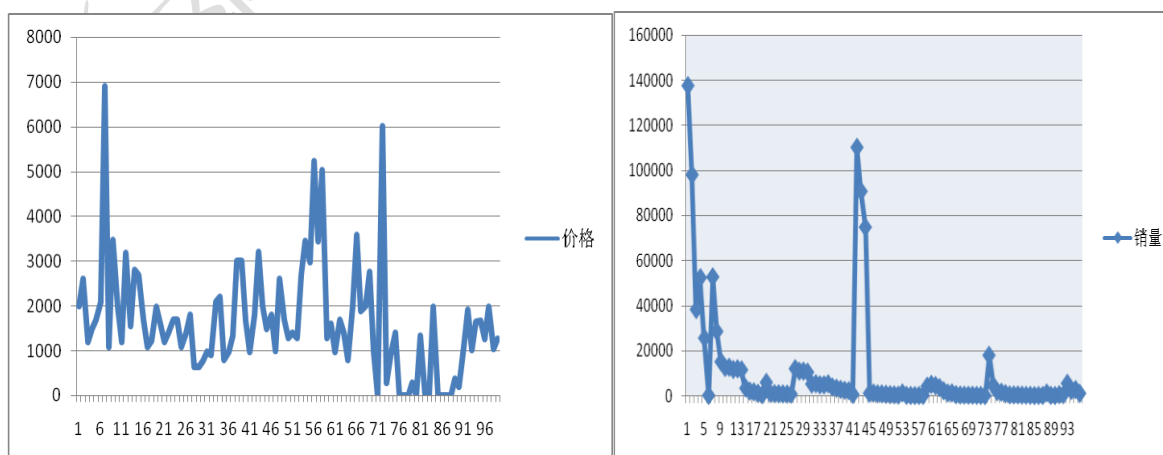


图 4.2 不同型号间价格与销量对比图

接下来，进行特征值提取，根据词云中初步获取消费者对净水器产品的关注点，得到如下表所示的个性属性表，统计评论数据中消费者的关注点导入同义词林理念（详见附录文件），进行同义词合并：由于中文对某一特征可以用多种方式表达，例如“价钱”“价格”“价位”表达的都是一个意思，在网络评论中，消费者往往都有自己的语言习惯，不会使用同一的表达方式，从而导致某些表达方式因为出现频率过高被选出来而其他的则不能被选出来，为了解决这个问题，用 R 语言提取属性高频词建立同义词林，将表达同一个意思的不同表达词汇统一用一个词来表示。统计出同义词林下各品牌属性的特征词频，如下表

	沁园	爱惠浦	安吉尔	道尔顿	海尔	佳尼特	AO	美的	3M	安之星
质量	51380	12905	6280	16246	2338	462	2889	3599	15932	2338
品牌	20201	3896	2033	4115	796	143	908	1209	4070	796
物流	44824	7898	8116	17368	3344	358	2461	5419	15759	3344
售后服务	80167	14797	7350	15675	3201	660	4479	6640	18244	3201
价格	35588	9408	3045	6949	1172	264	1515	2336	10559	1172

表 4.3 各品牌的特征属性的词频

由上表的统计结果，求出各品牌特征属性的比重表，如下

	安之星 a	爱惠浦 b	安吉尔 c	道尔顿 d	海尔 e	汇总
质量	0.2105263	0.3333333	0.2619048	0.4324324	0.2105263	1.4487232
品牌	0.0526316	0.0208333	0.0000000	0.0540541	0.0526316	0.1801505
物流	0.3157895	0.1041667	0.1428571	0.1081081	0.3157895	0.9867109
售后服务	0.2631579	0.2500000	0.4523810	0.1081081	0.2631579	1.3368048
价格	0.1578947	0.2916667	0.1428571	0.2972973	0.1578947	1.0476106
	佳尼特 f	AOg	美的 h	沁园 i	3M j	汇总
质量	0.3714286	0.3666667	0.3478261	0.3380282	0.2979298	1.7218793
品牌	0.0000000	0.0333333	0.0000000	0.1408451	0.1503150	0.3244934
物流	0.1428571	0.0666667	0.0869565	0.0000000	0.1836184	0.4800987
售后服务	0.2857143	0.2000000	0.4347826	0.3098592	0.1638164	1.3941724
价格	0.2000000	0.3333333	0.1304348	0.2112676	0.2043204	1.0793562

表 4.4 各品牌特征属性比重表

并根据比重表绘制条形比重图像，可以直观看出消费者对各品牌的主要关注点，如下图：

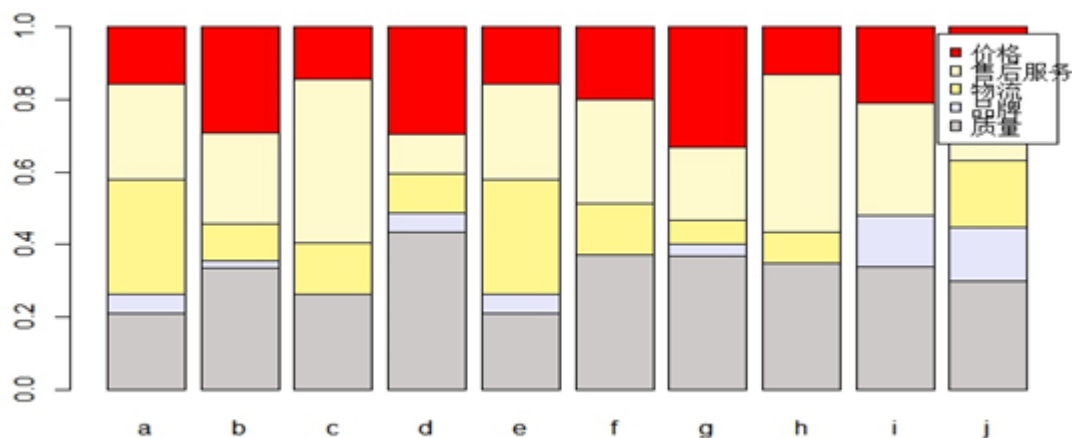


图 4.3 各品牌的属性占比

此外，根据消费者网上评论数据利用 R 语言进行属性间的差异分析，从纵向分析得到消费者对各品牌净水器的各属性需求，如道尔顿品牌中，消费者更关注于产品的质量及价格，可以看出，该品牌与其他品牌的差异性。另一方面，消费者对各品牌的需求不尽相同，每个品牌均有自身的优势，如安之星的物流相对其他品牌有其更鲜明的特色，各种品牌间的差异也给了消费者充足的选择空间，满足消费者不同的需求。



图 4.4 品牌间个性属性占比图

从横向分析，由上表可得到个属性的占比，如下图，回到消费者的个性化需求分析，可以看出，在净水器的行业中，大多数消费者对质量及售后的服务更加重视，有其自身的需求，对产品的要求也随着时间的改变而改变，我们可以借由此模型，统计总体的属性需求，这在商业化竞争激烈的当下尤为重要。

致此，我们分析了消费之的个性需求及其属性水平，各品牌之间的属性对比及特点，各商家也可以从中获取潜在市场价值，下一步，我们将建立起多元回归模型综合考虑以上几种因素叠加影响的效果，为此次的研究进行量化分析，提高模型准确性。

4.3.3 消费情况反馈机制

4.3.3.1 多元回归模型

根据消费者对属性的综合评分，由上面的情感分析得如下所示的表格，可以看到，对消费者的评论进行量化分析，从多种角度看出消费者的消费途径，得到更准确的效果。其部分表如下所示（详见附录）。

表 4.5 消费者评论情感分析部分表

4727	1	质量很好，安装方便，效果不错。
4728	0	还没安装，看上去是正品没有问题
4729	3	送货速度比较快，包装完整，安装也很快。
4730	0	到货很快，打售后安装的也很及时，安装师傅很细心，不过目前还喝不出水有什么特别的哎
4731	0	做工不错，就是刚要放分钟的水才没有塑料味。
4732	1	感觉还不错，感觉还不错
4733	2	水没有以前那么涩了，还是有点用的
633	-1	质量很好和卖家描述的一样
634	0	实在是不敢恭维。 安装有问题没有售后
635	0	净水效果很好，有需要还会再来的
636	-1	货收到了，宝贝很好，试和过了，净水之后比之前清了好多，口感也完全不一样了。非常满意的一次购物。
637	0	东西还好，比实体店便宜。售后也不错，我用了一下，水质还行。
638	1	比以前没装要好些吧只是水流很小。
639	0	用着挺好的，我老公自己安装的。和实体店里面看的是一样的，价格便宜好多
640	3	安装上以后水没有一点杂味 很好喝 我都喝凉的 泡茶也和以前不是一个味了

由于应不是所有的解释变量即产品特征都对销量有显著的影响，所以我们需要去掉不显著的解释变量。因此本研究使用 R 语言对该模型采用逐步回归方法，此方法可以选取对被解释变量显著影响的解释变量，剔除不显著的变量，并且可以在一定程度上解决解释变量之间的多重共线性问题。

根据 R 语言导入评论数据计算情感分并计算品牌各属性的属性评分：

质量评分	品牌评分	售后评分	物流评分	价格评分
0	3	-1	3	0
0	1	0	1	4
1	4	3	4	1
0	1	0	1	0
0	0	3	0	3
1	0	0	0	0
1	0	1	0	0
0	1	1	1	1
0	4	0	4	1
1	0	1	0	0
3	4	0	4	0
0	0	0	0	-4
1	1	4	1	0
3	2	1	2	1
3	0	3	0	0
0	1	3	1	0
0	0	0	0	1
1	1	1	1	0
0	1	0	1	1
3	0	0	0	0

4.6 各品牌属性得分表

根据上表分析结果，用 R 语言编程求出每个属性的平均得分如下表

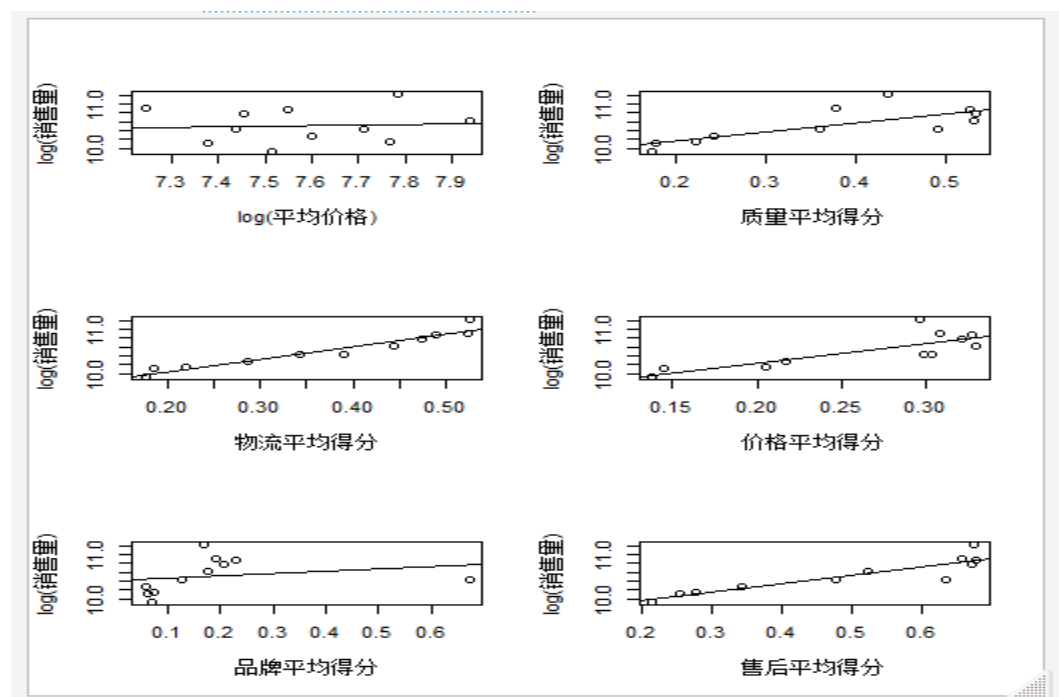
品牌	平均价格	物流平均得分	品牌平均得分	售后平均得分	质量平均得分	价格平均得分	销售量
3M	2000	0.286624	0.057324	0.343948	0.242204	0.216561	28475
A0	2800	0.444542	0.176016	0.522903	0.532681	0.329388	40814
爱惠浦	2365	0.219286	0.072769	0.275839	0.22203	0.205544	25513
安吉尔	1400	0.52378	0.190244	0.656707	0.378049	0.307927	53690
安之星	1729	0.473708	0.205761	0.672154	0.534979	0.32053	48723
道尔顿	2400	0.525424	0.168362	0.674576	0.436158	0.296045	74712
海尔	1899	0.488737	0.227228	0.678746	0.526934	0.32713	52633
美的	2235	0.343361	0.673829	0.633567	0.492649	0.302873	33522
沁园	1600	0.184483	0.063188	0.255398	0.176651	0.144581	24555
特洁恩	1830	0.176796	0.06814	0.21547	0.173112	0.138582	20211
其他	1700	0.389344	0.126138	0.477231	0.3602	0.298383	34214

表 4.7 品牌属性平均分结果

假设回归方程的各个属性，如下

Y	Ln(销售量)	X4	售后平均得分
X1	ln(价格)	X5	质量平均得分
X2	价格平均得分	X6	品牌平均得分
X3	物流平均得分		

根据品牌属性平均分结果绘制如下图形：



由散点图进一步进行显著性分析，分析其效果，各值如下

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0081	1.0100	5.949	0.00401 **
log(平均价格)	0.4814	0.1321	3.643	0.02190 *
质量平均得分	-0.8867	0.5466	-1.622	0.18010
物流平均得分	1.0201	1.1951	0.854	0.44143
品牌平均得分	-0.7470	0.3593	-2.079	0.10616
售后平均得分	2.9077	0.9299	3.127	0.03529 *
价格平均得分	-1.8276	1.2259	-1.491	0.21027

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06293 on 4 degrees of freedom

Multiple R-squared: 0.9901, Adjusted R-squared: 0.9752

F-statistic: 66.59 on 6 and 4 DF, p-value: 0.0005817

> |

根据显著分析：平均价格和售后得分对销售量影响较大，从而我进行逐步回归，剔除不显著变量，进行对模型分析，如物流平均得分，进而得到多元的回归模型，以下是剔除后的结果：

```

              Df Sum of Sq    RSS    AIC
<none>                 0.01873 -58.132
- 价格平均得分      1    0.00598 0.02471 -57.084
- 质量平均得分      1    0.01979 0.03852 -52.200
- log(平均价格)     1    0.07069 0.08942 -42.936
- 品牌平均得分      1    0.18608 0.20481 -33.819
- 售后平均得分      1    0.45120 0.46993 -24.684
summary(lm.test.star)

```

剔除后的显著性分析如下

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.6699	0.9035	6.276	0.001509	**
log(平均价格)	0.5216	0.1201	4.344	0.007400	**
质量平均得分	-1.0943	0.4761	-2.299	0.069899	.
品牌平均得分	-1.0258	0.1455	-7.048	0.000888	***
售后平均得分	3.6460	0.3322	10.976	0.000109	***
价格平均得分	-1.1598	0.9178	-1.264	0.262086	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0612 on 5 degrees of freedom
 Multiple R-squared: 0.9883, Adjusted R-squared: 0.9766
 F-statistic: 84.35 on 5 and 5 DF, p-value: 7.972e-05

> 可以看出, R-Squared=0.9766:模型拟合程度效果良好, 并且 p 值<0.1:最后筛选出来的解释变量对被解释变量显著影响。

如下是剔除后的效果图像显示:

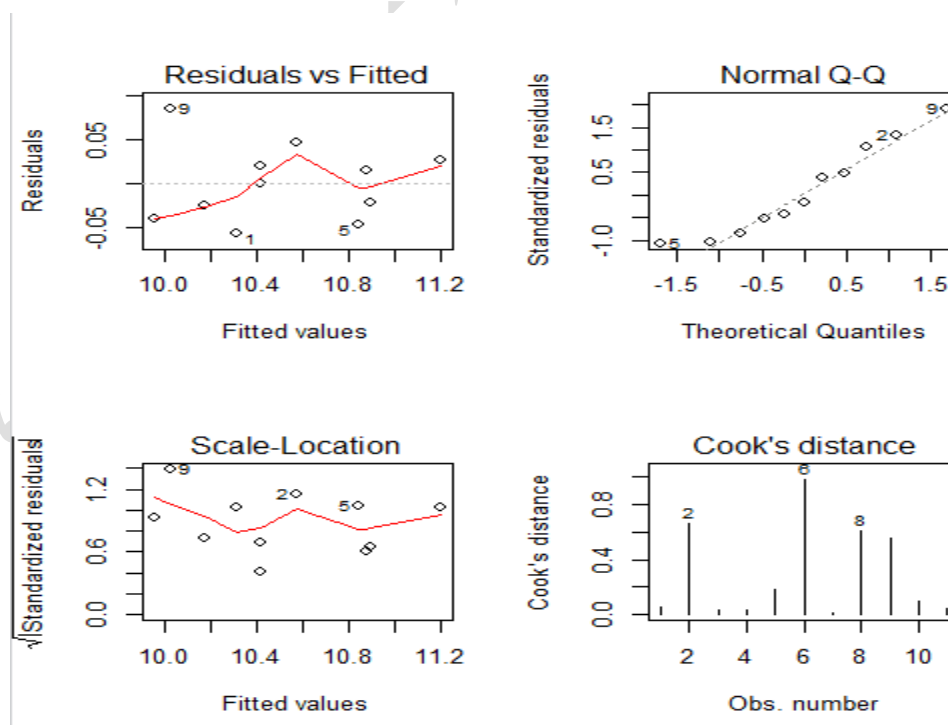


图 4.5 不同因素对销量的影响效果

4.3.3.2 多元回归模型分析

由此我们通过逐步回归模型的确立通过对以上主属性的综合考虑得到逐步回归方程为如下：

$$\ln(y) = 5.6699 + 0.5216 \ln(x_1) - 1.0943(x_2) - 1.0258(x_4) - 3.6460(x_5) - 1.1598(x_6)$$

可以看出质量和品牌分数为负数——质量跟售后服务越好，即销量越高。

其模型拟合的结果与现实情况的趋势较为满意。

综上，我们可以看到消费者与商家、品牌间的关系甚为微妙，消费者的个性化需求影响着商家品牌的方向，而商家品牌的服务有为消费者的个性化需求提供了可能，通过消费者购买欲分析模型的分析结果，我们可以从中分析出一些价值规律及现象，为消费者和商家提供更快捷准确的反馈机制模型。

五、模型评价

5.1 模型优点

本文对消费者的评论数据进行文本挖掘，对模型中涉及到的众多影响因素进行了量化分析，对以消费者个性化需求为基础的消费者购买欲分析模型进行分析，了解到当前产品的销售现象，使得各品牌及商家能够及时对产品需求量预估做一定程度的调整，而且还掌握了以消费者消费途径为特点的需求前景分析，得到了较为精细的分析结果。另外，模型建立过程中考虑的因素也相对全面，所采用的一些研究方法如聚类分析，数理统计等均具备较强的可移植性，可信度比较高。

5.2 模型局限与展望

尽管本文研究是基于规范的理论分析和实证方法对众多消费者的海量评论进行了系统有效地研究，获得了较丰富的分析结果，但仍然存在很多局限和不足，在今后的研究中有待深入和改进。

本文的局限性主要有以下几点：

第一，本文的原始数据爬取的时间周期还比较短，未能体现出产品销量随一些季节性因素的变化规律；

第二，由于网络及市场的发展速度迅猛，产品更新换代的节奏不断加快，我们的模型也局限于近年的销量分析及对短期的需求前景作预测。

第三，该模型的影响因素较多，在综合考虑各因素的情况下不能对其进行全面的分析统计，造成与实际情况有一定的差异。

在以后的研究中，一方面是突破以上研究的局限性，另一方面是扩展研究的内容和方法。

对于此次模型，还可以就分析的文本内容与评分进行比较，确定两者的吻合程度，进而弥补情感计算所带的偏差，解释现有消费者购买欲模型中尚未验证得到解释的因

素。另外，进一步探索消费者消费途径如购买步骤的影响，由于本文在描绘这方面的数据相对还是比较少的，且每个消费者的购买习惯千差万别，受性别、年龄等影响显著，因此，可以通过一些网站公布的数据如百度指数等对网民群体作出的统计进一步探索不同消费者的购买习惯，对其进行有效性的解读，这是下一步要探索的内容，结合本模型或着手这方面的信息挖掘将变得非常有意义。

六、 参考文献

- [1] 谢金星，基于SVM的中文微博情感分析的研究[D]. 2012;
- [2] Liu B.Sentiment Analysis and Opinoon Mining[J].Synthesis Lectures on Human Language Technolotics,2012;
- [3] 徐冰，基于统计学系的文本情感分析关键技术研究[D].2011
- [4] 赵威，冰箱产品个性化需求分析方法和应用研究，河北工业大学，20050101;
- [5] 冷羽，数据预处理技术浅谈[DB/OL]，<http://blog.sina.com.cn/u/1984071801>;
- [6] 张丽，在线评论的客户参与动机与评论有效性研究[D].2011;
- [7] 乐国安，情感分析技术及应用[D].2015;
- [8] 徐晓松，军的发现和算法研究，2015年4月;
- [9] 基于评论挖掘技术的在线商品评论对销量的影响——以亚马逊手机销量为例，2015;