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：一等奖

作品单位：南京财经大学

作品成员：纪明明　李翰林　王攀

指导教师：李冠艺

基于电商平台热水器的消费者需求及产品数据挖掘分析

摘要：本文对三大电商平台、六大热水器品牌和五大热水器类型的评论数据通过数据清洗、数据集成和融合、数据变换、数据规约等方法进行了预处理；在此基础上，使用情感词典和语义规则进行极性累加，进行评论的情感分析；最后采取了消费者决策的 AHP-FCE（层次分析法与模糊综合评判）分析，结合参考百度指数及 F-IDF 评论词频得出的分层评判及模糊综合评判的数值化求解，最终得出对某一种类热水器品牌差异化评分，并得出及探究了各热水器品牌类型的用户购买原因和差异化卖点，实现数据挖掘后对数据的实际应用。

关键词：数据预处理、情感分析、层次分析、模糊评判

Data mining analysis based on water heaters consumer demand and product of e-commerce platforms

Abstract: This paper uses comments data from top three electric commercial platforms, top six water heater brands and top five water heater types to implement data preprocessing through data cleaning, data integration and fusion, data changing, and data Statute. This paper uses emotional dictionary and semantic rules to implement sentiment analysis by the method of Polar cumulative. And then, this paper uses Baidu index and frequency of F-IDF comments to achieve the numerical solution of layered evaluation and fuzzy synthetic evaluation, reaches the score of a certain type of water heater brand differences, explores the user purchased causes and differences of selling of different water heater brands and implements practical application of data mining.

Key words: Data Preprocessing, Sentiment Analysis, Analytic Hierarchy Process, Fuzzy Sets

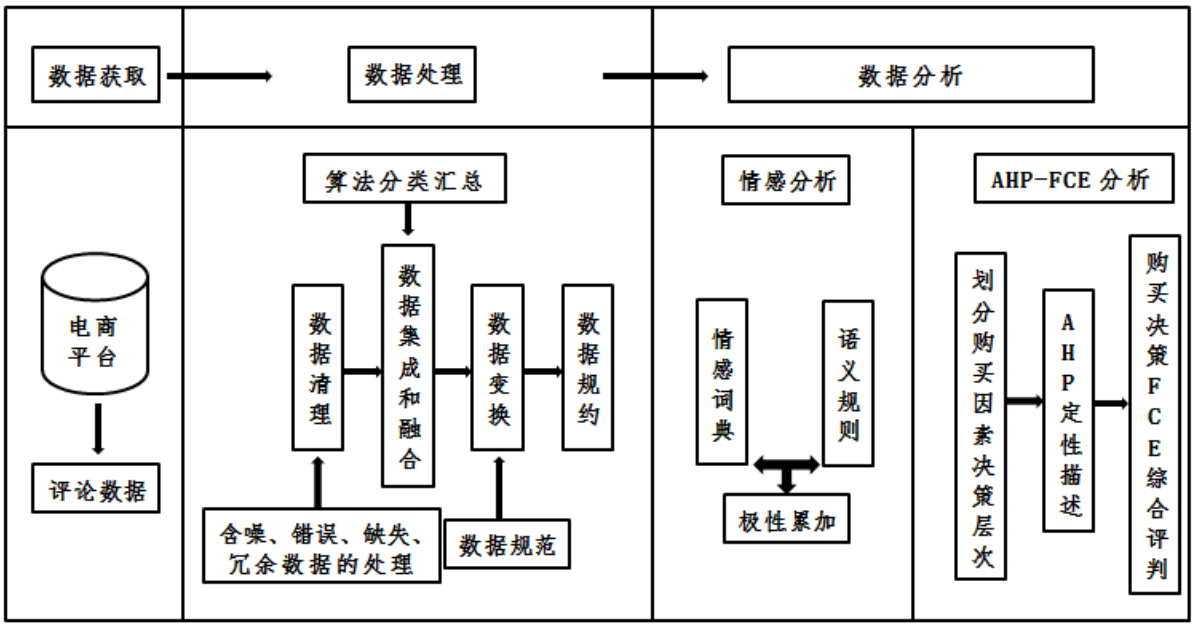
目录

一、 研究目标

本次数据挖掘通过火车头和八爪鱼两个软件实现，通过对三大电商平台、六大热水器品牌和五大热水器制热类型的热水器进行价格、型号、评论时间、评论数据的挖掘，获取到电商平台数据后，对数据进行处理，剔除造假数据和无意义数据。对处理过后的数据进行分析，运用情感分析方法分析评论数据，发掘用户情感倾向，进一步分析个热水器产品的优势和劣势、差异化买点和用户个性化需求。

二、 分析方法与过程

(一) 总体流程



数据获取：通过火车头和八爪鱼两个软件实现，通过对三大电商平台、六大热水器品牌和五大热水器制热类型的热水器进行价格、型号、评论时间、评论数据的挖掘。

数据处理：通过简单的对评论数据去重以后，对接下来的数据进行数据清理，以此对含噪、错误、确实、冗余的数据进行处理；在数据集成和融合的基础上，再对数据进行数据变换以此使数据规范化；最后对数据进行数据规约，并以可视化呈现。

数据分析：采用了情感分析和 AHP-FCE 分析法。情感分析主要通过情感词典和语义规则的方法进行分析，在此基础上进行极性累加；AHP-FCE 分析首先对购买的决策层级进行划分，再对 AHP 进行定性描述，最后对购买决策 FCE 进行综合评判。

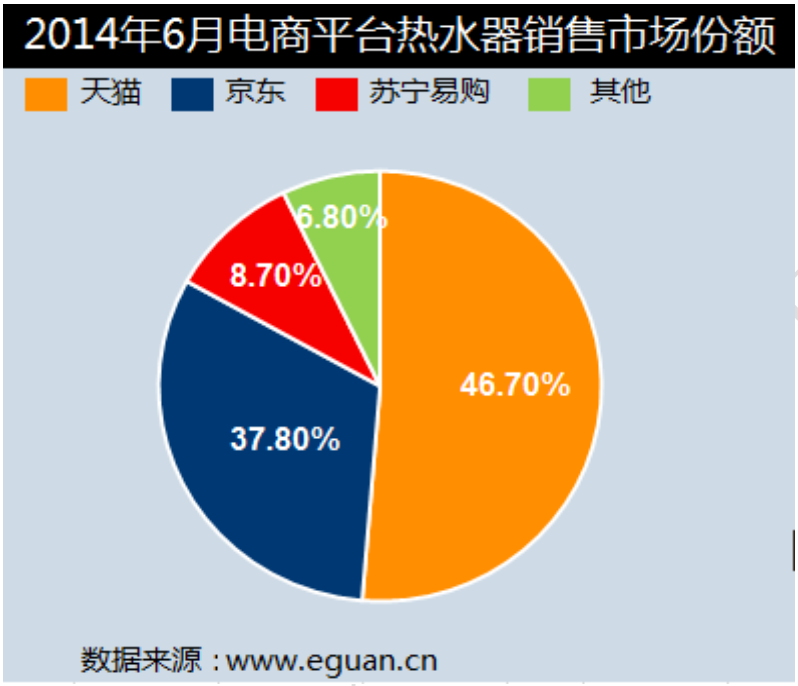
(二) 具体步骤

1. 数据获取

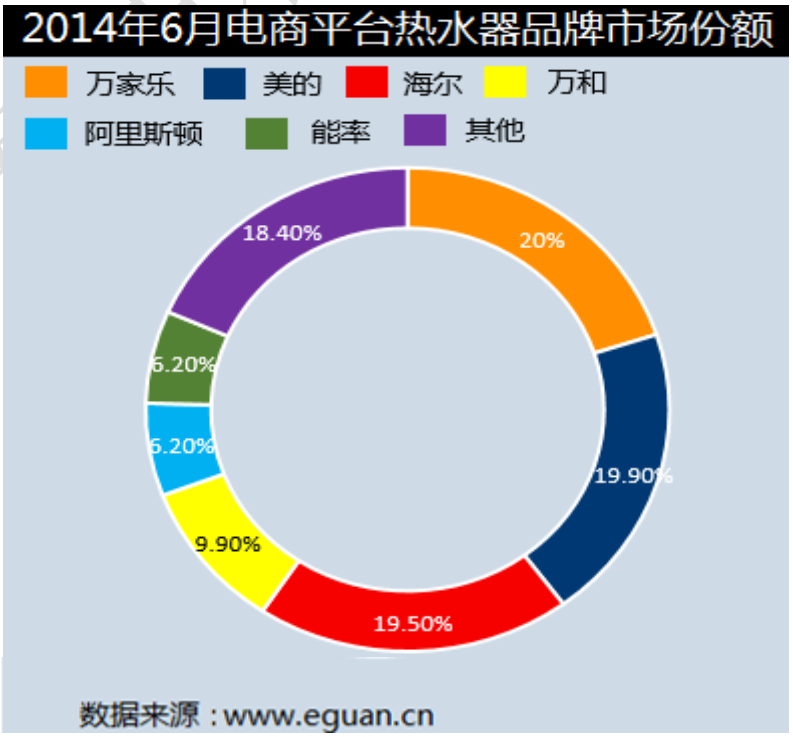
根据中国网商行情系统数据，截至 2014 年 6 月，我国电商平台热水

器销售市场份额情况如下图所示，天猫商城、京东商城和苏宁易购位列电商平台销售市场份额前三，其中天猫商城占整个市场份额的46.7%，京东商城占37.8%，苏宁易购占8.7%，其余6.8%是其他电商平台（国美、易迅等）所占市场份额。

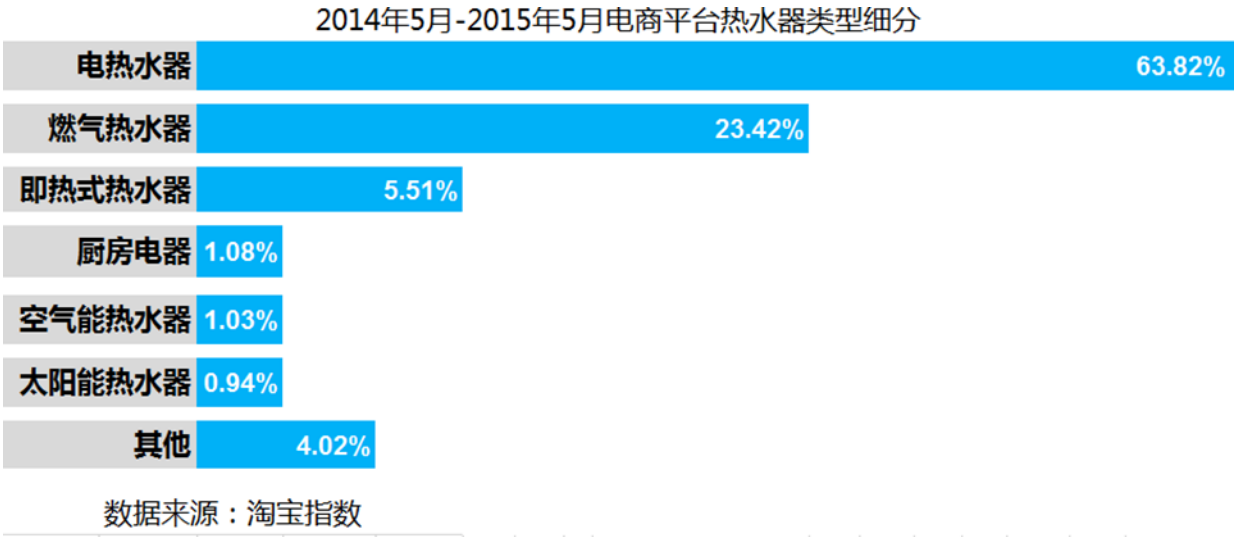
由于销售市场份额前三的电商平台所占市场份额之和已达到所有电商平台销售市场份额的90%以上，所以我们的数据主要从天猫商城、京东商城和苏宁易购采集，并且采集的数据具有足够的代表性。



其次，根据品牌划分，我国电商平台的热水器品牌市场份额中，万家乐、美的、海尔、万和、阿里斯顿和能率六个品牌位列市场份额前六，总计占电商平台市场份额的81.6%，所以在数据采集时，主要采集天猫、京东、苏宁易购三大电商平台中，该六个品牌的热水器的数据。



根据淘宝指数数据，在 2014 年 5 月至 2015 年 5 月期间，各种类型的热水器的销售市场份额占比如上图所示，电热水器、燃气热水器即热式热水器位列前三，紧随其后的分别是空气能热水器和太阳能热水器，以上五种热水器类型占到了整个市场的 94.27%。因此，在采集数据时，我们选择了以上五种类型的热水器进行评论采集，使得采集的数据具有充分的代表性。



2. 数据处理

数据处理是数据分析过程中最花费时间、最乏味的,但也是最重要的一步.该步骤处理得当，可以有效地提升数据质量，减轻下一步工作量，并作为数据精准分析的基础.本文数据处理的过程主要分为以下几步：

1) 数据粗处理

在充分采集三个电商平台相关数据的基础上，获取了海量而驳杂的数据。出于评论内容真实有效的考虑，在整个数据集中，本文选取了三个无效属性进行并集删除操作（A∪B∪C），分别为：

A. 评论项为空
B. 评论不含中文
C. 不含关键词的评论

作为预处理之前的粗处理，京东、天猫和苏宁三个平台的平均去除率在 7.3%左右。

2) 数据预处理

a)数据清洗

与资讯，微博不同，商品评论文本的噪声更少，除了粗处理去除的无效数据，主要就在于商家恶意刷的重复评论。在对文本的进一步观察和相应网站的考证基础上，再次发现了大量短时间内不断重复的虚假评论。基于层次分析法的模型，采用凝聚层次聚类的算法，对整个数据集中的五个相关联的属性进行交集删除操作（A∩B∩C∩D∩E）,分别为：

A. 评论时间（不同评论相隔 60s 以内）

B. 评论内容（100%相似度）
C. 相同平台
D. 相同品牌
E. 相同型号

三家电商平台热水器的清洗结果见表 1：

电商平台	京东	天猫	苏宁
平均去重率	36.4%	39.7%	90.0%以上

b) 数据集成和融合

本文的集成合并多家电商平台中采集到的多个热水器品类数据，存放到一个一致的数据存储中。本文的融合仅限于数据层的数据融合,即把数据融合的思想引入到数据预处理的过程中,加入数据的智能化合成,产生比单一信息源更准确、更完全、更可靠的数据进行估计和判断。按照模式集成和对象匹配的原则，通过对数值属性的相关系数 Rab（见公式 1）的判定检测冗余，按照统一的构造集成融合，进而提升其后挖掘过程的准确度和速度。

$$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{(n-1)\sigma_a \sigma_b}$$

$$\sigma_a = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}}$$

$$\sigma_b = \sqrt{\frac{\sum_{i=1}^{n}(b_i - \bar{b})^2}{n-1}}$$

其中,n是数据集样本个数，a_i和b_i分别是元组i中a和b的值，

$\bar{a}$和$\bar{b}$分别是a和b的均值；

σ_a和σ_b分别是a和b的标准差。

c)数据变换

本文进行数据变换的目的在于将多维数据压缩成较少维数的数据,消除不同平台各型号热水器数据在时间、空间、属性及精度等特征表现方面的差异。这类方法虽然对原始数据都有一定的损害,但其结果往往具有更大的实用性，主要步骤如下：

数据平滑去噪,使连续数据离散化,增加粒度数据聚集对数据进行汇总；数据概化减少数据复杂度,用 excel 中的分类汇总使各数据落入指定条目区域属性构造。

d) 数据规约

本文用数据规约技术得到数据集的规约表示，主要通过数据立方体聚集、属性子集的分类选择得到更易于处理的文本数据，且不损伤原始数据的完整性。

三、 结果分析

1. 评论情感分析

已采集的数据中，评论所体现的复杂信息含有很多隐藏价值，本文在此从情感分析的角度对其进行挖掘。与传统较长的文本（新闻、博客、微博等）不同，商品的评论信息文本简短，字数一般在 10-30 字左右，语句随意，一条评论语句中除了文字信息，还穿插着标点符号商品评论的这些特征对文本的情感分析会产生一定影响，通常一条评论的情感与它所含符号的情感也是相符的。例如~表达的一般是正面的情感。

本文采用基于情感词典和基于语义规则的两种方法，对处理后的评论进行情感分析。

1) 基于情感词典的方法

词典资源是基于情感词典方法的前提，本文使用台湾大学 ntusd（简体中文 2810 正面词语 + 8276 负面性词语)和 HowNet 评价词词典(9,193 个中文评价词语/短语）作为词库，然而词库有褒贬分类，但是没有标注情感极性强度。本文把褒义词语的情感极性值设为 0.7，贬义词语情感极性值设为-0.7，作为评判基准，采用极性累加的算法进行估算。

基于情感词典的方法首先对每条评论进行分词、词性标注等预处理，然后依据情感词典判断每条商品评论中出现的所有情感词以及强度，并采用极性累加的方法来计算每条评论的情感极性，算法如下：

$$\text{Polarity(T)} = \sum_{i=1}^{n} Polarity(w_i)$$

其中：w_i为一条微博中所含的情感词；

$Polarity(w_i)$为一个情感词的情感极性；

Polarity(T)为一条评论的情感极性；

若结果大于零，评论为褒义倾向，
若结果小于零，表明结果为贬义倾向，否则为中性。

2) 基于语义规则的方法

考虑到基于情感词典的方法有着明显的缺点：对独立的词语来进行分析的，也就是把词语从句子中孤立出来，忽略词语的前后联系。简单举例，海尔的热水器不好，只提取情感词难以得出正确的结论。孤立地分析情感词，并不能完全正确地反映评论信息的情感倾向，必须将语句的联系考虑

进来，才能够提高分析的准确度。因此，在词语情感计算的基础上，本文同时也着眼于语句中能够改变词语情感倾向或者情感强度的修饰副词等。将会改变词语极性强度的修饰副词分为两类，第一类是否定词，它会改变极性倾向，比如上个例子的“不”就是反义；第二类是程度词，它会改变极性强度，如“比较”、“非常”等。

同时，简短的评论文本有其自身的特征，本文只考虑最高频出现的消息文本中的符号，如“!”、“~”等符号，其他对评论文本的情感极性没有什么影响的，不予以考虑。

3) 基于 PMI-IR 算法与搜索引擎结合进行分类

使用 PMI-IR 算法，以情感词语为中心,通过搜索引擎返回的结果来计算文本中的情感要索和背景情感词之间的点互信息值,从而对文本进行情感分类，方便下一步情感词汇的整合。

a)情感词汇

情感词是判断电热水器评论文本是否具有情感倾向的一个重要特征。根据人们留言习惯和大量语料分析得知,人们在商品评论中反馈大多是通过情感词的形式实现的，情感词的褒贬也通常代表这句子的褒贬。

通常情况下商品评论文本中都是比较简单的句子,情感词的倾向很多时候决定了商品评论的情感倾向,情感词的数量和情感强度对每条商品评论的情感倾向有较大的影响，因此仍然采用极性累加的方法，即通过情感词极性累加的公式 2 来计算每条商品评论的情感极性。

b) 细分程度副词语态

情感词典分析用词表中提供了程度级别词语，本文以此为基础，参考商品评论评论中高频出现的词汇添加人工整理的程度副词表，把程度副词分为三个级别。第一级的程度词对所修饰的情感词的情感强度大大加强，例如“极”、“最”；第二级的程度词对所修饰的情感词的情感强度是加强作用，如“很”、“非常”。第三级的程度词对所修饰的情感词的情感强度是削弱作用，如“有些”、“稍微”。三个级别程度词对所修饰情感词的情感强度扩大倍数分别设置为第一级 2 倍、第二级 1.5 倍、第三级 0.5 倍。

倘若句子中情感词语前面有程度词修饰，那么被修饰的情感词语的情感强度必然发生改变，进而会影响到这个句子的情感强度。一个程度副词后面可以有多个情感词，同样一个情感词也可以被多个程度副词所修饰。本文处理程度副词的方法是把情感强度加到其后修饰的第一个情感词上，情感强度

对情感词 w_i 的影响因子 σ 定义为：

$$\sigma(w_i) = \prod_{k=1}^{m} Deg(d_k)$$

其中：$\sigma(w_i)$为程度副词的情感强度扩大倍数。

c)增加否定词影响

本文选取“不想、不会、不要、没有”等 30 个常见否定词作为否定副词表，并将其极性强度设置为-1。

例如“没有配套的上门安装服务，虽然热水器很好”，在情感词前面加上否定词“没有”，整个句子的情感极性就会发生改变。本文处理否定词的方法是将否定加到其后的第一个情感词上，当一个情感词前面出现不只一个否定词时，根据否定词出现的次数来判断情感词的极性。出现奇数次则情感词的极性逆转，否则情感词的极性不发生改变。所以，否定词对情感词 w_i 的影响因子 ϵ 定义为：

$$\epsilon(w_i) = (-1)^n$$

d) 增加符号的影响

很多买家在发布评论时喜欢加上一些表符号，比如“~”表示褒扬，“…”表示无语或不满等，本文将常用表情符号分为正向和负向两类。一般情况下，如果一条评论信息包含表情符号，将正向和负向符号转化为上文程度副词中的第三级词汇，再加以进行计算。

4) 情感极性计算

综合考虑上述几个特征，使用公式 3 对评论信息的情感倾向值 Polarity(T) 进行计算：

$$\text{Polarity(T)}=\sum_{i=1}^{n} Polarity(w_i) * \sigma(w_i) * \epsilon(w_i)$$

若 Polarity(T) 计算结果大于零，表明评论信息为褒义倾向，若结果小于零，表明评论信息为贬义倾向，否则为中性。

5) 主客观判断

因为本文立足于无监督学习的方法，将采取人工检验的结果进行二次检验。本文将处理的不同子属性进行整理，分散给学校不同专业背景的成绩正常的学生进行人工主客观判断，采用分级制度，用 0（不确定）、1（基本确定）、2（确定）、3（非常确定）对结果进行标注，来达到主客观判断的目的。

以预处理的一条子属性结果为例，对京东-能率-燃气热水器进行分析。首先进行分词，词性标注等预处理，然后分别采用上文介绍的基于情感词典的方法和基于语义规则的方法分析处理评论，最后分别得到正面、负面和中性的评论数目，并且通过上文的主客观程度判断方法，计算 2 种方法的准确率（P）。实验结果如表 2 所示。

方法 类别	情感词典			增加语义规则		
	自动识别数	正确数	准确率	自动识别数	确数	准确率
能率 **GQ-1150FE**	189	107	56.4%	189	128	67.8%

能率 GQ-1350FE	3050	1818	59.6%	3050	2086	68.4%
能率 GQ-1650FE	2519	1484	58.9%	2519	1683	66.8%
能率 GQ-1070FE	2315	1356	58.6%	2315	1525	65.9%
能率 GQ-1180 AFE	1699	980	57.7%	1699	1147	67.5%
能率 GQ-1680 AFE-A	129	79	61.3%	129	90	70.0%
能率 GQ-1380C AFE	91	46	50.2%	91	59	65.1%

由表 2 可以看出，结合基于语义规则分析的方法相对于基于情感词典的方法在准确度方面有了明显的提升，极性累加的算法对于简短的热水器商品评论的情感的自动判断已经达到了一定程度的准确率，可以对用户在选择商品所查看评论的大体情感反馈中起到辅助决策的作用，具有实际商业应用的价值。

2. 消费者决策的 AHP-FCE 分析

1) 方法介绍

AHP(Analytical Hierachy Process,应用层次分析法) 是匹兹堡大学 T. L. Saaty 教授在 20 世纪 70 年代初期提出对定性问题进行定量分析的一种渐变灵活的多准则决策方案，其特点在于把复杂问题中的各种因素通过划分为相互联系的有序层次，使之条理化，根据对有一定客观现实的主观两两比较，把专家意见和分析者的客观判断结果直接有效的结合起来，而后利用数学方法计算每一层元素相对重要性次序的权值，最终通过所有层次间的总排序计算所有元素的相对权重并进行排序从而分析消费者决策。

FCE(Fuzzy Comprehensive Evaluation,模糊综合评判) 20 世纪 80 年代初，我国模糊数学领域的汪培庄教授提出了综合评判模型，并通过广大实际工作者的不断的补充发展，衍生出的适用于各种领域的评判方法。模糊综合评判的过程可简述为:决策者将价目标看成是由多重因素组成的因素集 U,再设定这些因素所能选取的评审等级，组成评语的评判集合 V,分别求出各单一因素对各个评审等级的模糊矩阵，然后根据各个因素在评价中的权重分配，通过模糊矩阵合成，求出评价的定量值。

但是这两种方法各有利弊：AHP 中的能够准确的对决策定性，但其决策过程过程需要经过大量数据比对来最终通过概率确定权重；而 FCE 中虽然有很好的定量评价但是无法很好地对决策定性。因此我们通过对二者的结合来寻求更加完善的问题解决方案。

AHP-FCE 模型需要经历以下三个步骤：

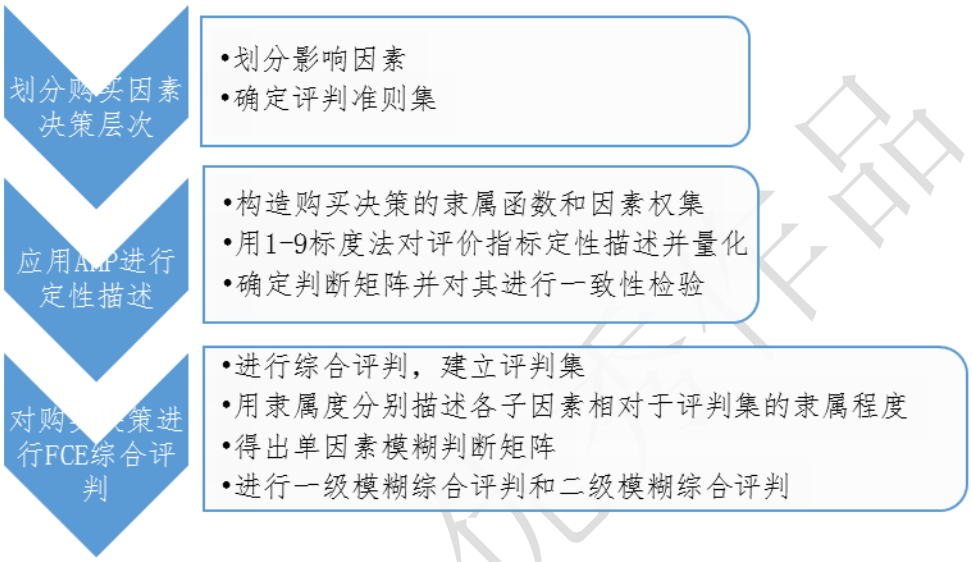
① 划分因素层

② 应用 AHP 构造消费者心理的隶属函数和因素权集合
③对所求结果进行综合评判
本次所建立模型需要经过两次模糊评判。

2) 模型建立

i. 模型概述

使用 AHP 与 FCE 对电商平台上热水器的购买决策分析。



ii. 层次划分及指标权重的建立

本次数据挖掘处理所涉及到的变量有：电商平台，热水器功能，热水器品牌，为了说明其中

我们从存在竞争关系的品类中，选取定元和不定元，得出我们需要分析的问题为

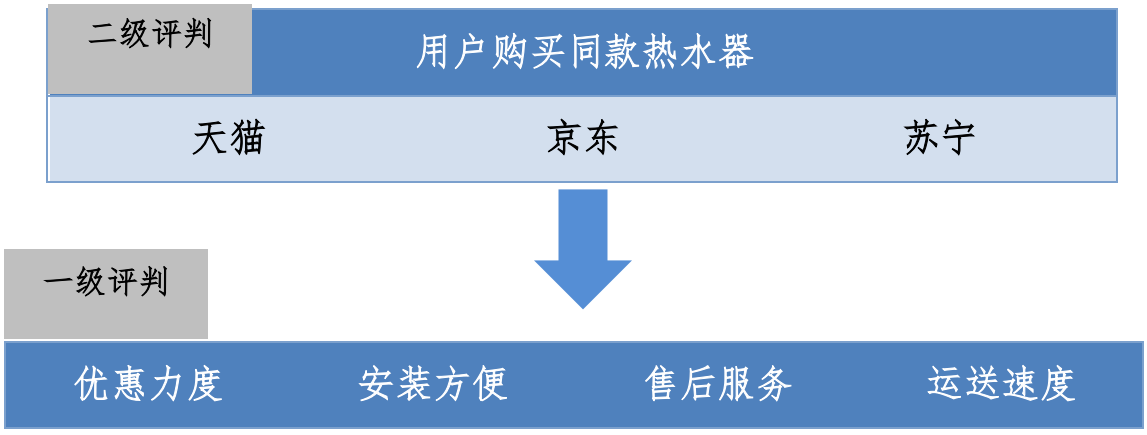
① 同一品牌同一型号在不同电商平台的决策
② 相同功能，不同品牌的用户决策
③相同品牌，不同功能的用户决策

根据不同的定元选取，需要建立不同的层次划分：

首先以体系 a 入手

用户电商平台热水器购买指标体系的建立体系假设

a：以用户购买同品牌同一产品分析



指标因素权重的确定：
本次选取本次与处理后的评论数据，并从评论数据中提取关键字的方法来确认各个评判指标中两两比较生成的对比程度，构造判断矩阵

$$B = n \times n$$

矩阵的元素b_{ij}代表b_i与b_j相比的重要程度，一般采用 Saaty 提出的 1-9 标度方法：

表 1 判断矩阵标度及其含义

标度	含义
1	Bi 和 Bj 同等重要
3	Bi 比 Bj 稍微重要
5	Bi 比 Bj 明显重要
7	Bi 比 Bj 强烈重要
9	Bi 比 Bj 极端重要
倒数	Bji=1/Bij
2，4，6，8	重要程度介于上述奇数之间

确定权值
计算各指标的权值

1） 计算判断矩阵每一行元素的乘积P_i

$$P_i = \prod_{j=1}^{n} b_{ij} \ (i = 1,2,3 \ldots n)$$

2） 计算P_i的 n 次方根D_{ni}

$$D_{ni} = \sqrt[n]{P_i}$$

3） 权重计算，向量归一化

$$w_i = \frac{D_{ni}}{\sum_{i=1}^{n} D_{ni}}$$

4） 一致性检验
偏差的一致性指标

$$CI = \frac{\lambda_{max} - n}{n - 1}$$

$$\lambda_{max} = \sum_{i=1}^{n} \frac{(B\bar{w})_i}{nw_i}$$

$$\mathrm{B\bar{w}} = \begin{bmatrix} (Bw)_1 \\ (Bw)_2 \\ \cdots \\ (Bw)_n \end{bmatrix} = \begin{bmatrix} b_{11} & b_{11} & \cdots & b_{11} \\ b_{11} & b_{11} & \cdots & b_{11} \\ \cdots & \cdots & \cdots & \cdots \\ b_{11} & b_{11} & \cdots & b_{11} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{bmatrix}$$

$$\mathrm{CR} = \mathrm{CI}/\mathrm{RI}$$

RI 为平均随机一致性指标

矩阵阶数 n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52

当 CR 小于 0.1，矩阵一致性被认可

5) 分析评判矩阵

对于可能分析到的

①对**天猫用户**购买热水器判断矩阵的因素权值计算

百度指数

	B_1	B_2	B_3	B_4	W
B_1	1.000	0.308	0.334	0.216	0.084
B_2	3.247	1.000	1.086	0.701	0.271
B_3	2.994	0.921	1.000	0.645	0.250
B_4	4.630	1.550	1.550	1.000	0.400

$CI = 0.0652\ RI = 0.89\ CR = 0.0465$

F-IDF 评论权重

	B_1	B_2	B_3	B_4	W
B_1	1.000	2.272	0.612	0.738	0.247
B_2	0.440	1.000	0.269	0.325	0.109
B_3	1.634	3.717	1.000	1.206	0.404

B_4	1. 355	0. 829	0. 829	1. 000	0. 241

$CI = 0.0750\ RI = 0.89\ \ CR = 0.0831$

②对**京东用户**购买热水器判断矩阵的因素权值计算

百度指数

	B_1	B_2	B_3	B_4	W
B_1	1. 000	0. 635	0. 666	0. 196	0. 107
B_2	1. 575	1. 000	1. 081	0. 308	0. 170
B_3	1. 502	0. 925	1. 000	0. 285	0. 159
B_4	5. 102	3. 509	3. 509	1. 000	0. 564

$CI = 0.0412\ RI = 0.89\ \ CR = 0.0715$

F-IDF 评论权重

	B_1	B_2	B_3	B_4	W
B_1	1. 000	2. 873	0. 718	0. 958	0. 287
B_2	0. 348	1. 000	0. 250	0. 333	0. 100
B_3	1. 393	4. 000	1. 000	1. 320	0. 399
B_4	1. 044	0. 758	0. 758	1. 000	0. 213

$CI = 0.0414\ RI = 0.89\ \ CR = 0.0701$

③对**苏宁用户**购买热水器判断矩阵的因素权值计算

百度指数

	B_1	B_2	B_3	B_4	W
B_1	1. 000	0. 322	0. 349	0. 163	0. 076
B_2	3. 106	1. 000	1. 085	0. 506	0. 235
B_3	2. 865	0. 922	1. 000	0. 466	0. 216
B_4	6. 135	2. 146	2. 146	1. 000	0. 473

$CI = 0.0371\ RI = 0.89\ CR = 0.0572$

F-IDF 评论权重

	B_1	B_2	B_3	B_4	W
B_1	1. 000	2. 66	0. 526	0. 763	0. 241
B_2	0. 376	1. 000	0. 198	0. 287	0. 091
B_3	1. 901	5. 051	1. 000	1. 451	0. 458
B_4	1. 311	0. 689	0. 689	1. 000	0. 211

$CI = 0.0471\ RI = 0.89\ CR = 0.0871$

④对三种电商平台所提供的判断矩阵进行因素权值计算

百度指数

	B_1	B_2	B_3	W
B_1	1. 000	2. 517	0. 227	0. 220
B_2	0. 397	1. 000	0. 090	0. 110
B_3	4. 405	11. 111	1. 000	0. 670

$CI = 0.0652\ RI = 0.89\ CR = 0.0465$

F-IDF 评论权重

	B_1	B_2	B_3	W
B_1	1. 000	1. 397	0. 144	0. 165
B_2	0. 716	1. 000	0. 103	0. 128
B_3	6. 944	9. 709	1. 000	0. 706

$CI = 0.0346\ RI = 0.89\ CR = 0.0678$

6) 综合指数数据得出综合权值向量

天猫用户

$$A_1 = (0.165，0.190，0.327，0.320)$$

京东用户

$$A_2 = (0.197，0.135，0.280，0.388)$$

苏宁用户

$$A_3 = (0.158，0.163，0.337，0.342)$$

对三种电商平台重要程度的二级综合评判

$$\tilde{A} = (0.165，0.128，0.706)$$

iii. 热水器购买决策的模糊综合评判

1) 应用最大隶属原则综合考虑相关因素，进行等级和类别评判。

① 建立评判因素集 U

相应的评判因素集为 $U = (u_1, u_2, \ldots, u_n)$

取上一节的五个一级评判指标作为一级评判因素集，即：

$$U_1 = (优惠，安装，售后，配送)$$

取三种电商平台作为二级评价因素集，即：

$$U_2 = (天猫，京东，苏宁)$$

② 评语集 V 及因素在 V 上的隶属度 a_i

将评判一级因素集合的评语分为五级，及评论集

$$V_1 = \{好，较好，一般，较差，差\}$$

将二级因素集合的评语同样分为五级

$$V_2 = \{非常重视，重视，考虑，偶尔考虑，不考虑\}$$

根据 Satty 提出 1-9 标度方法，相应的等级矩阵值

$$C = \{9,7,5,3,1\}$$

③ 进行多层次模糊综合评判

U 中个元素 u,即各个评价项目包含不同的子因素，其影响权重不同，将其表现为 U 上的一个模糊子集 A，U 中元素 u 对 A 的隶属度，有

$$A = \{A(u_i)\} \text{ 且 } A(u_i) \geq 0, \sum_{i=1\to n} A(u_i) = 1$$

对评判矩阵 R 做加权模型

$$B = A * R, b_j = \sum_{i=1}^{n} (a_i \cdot r_{ij})(j = 1 \ldots m)$$

求多级评判的$\widetilde{B}$时，加权评判模型 B 构成次级评判矩阵$\widetilde{R}$并对次级$\widetilde{B}$ 归一化得$\underline{B}$

④ 综合评价系数

$$W = B \cdot C^T$$

2) 样本提取及数据采集

从预处理后的某一个热水器评论集（见附件）分层提取有效评论集合。

3) 进行一级评判

选取一种热水器（以能率燃气热水器为例）在三种电商平台的评论结果

该热水器在天猫的对其各因素的评价集U_1频数如下表

因素集 U	V_1(好)	V_2(较好)	V_3(一般)	V_4(较差)	V_5(差)
U_1(优惠)	3534	4507	1007	904	105
U_2(安装)	1755	2877	4632	1247	207
U_3(售后)	2043	1756	4672	2561	507
U_4(配送)	1454	2716	2704	1720	104

令$r_{ij} = \frac{c_{IJ}}{S}(i = 1,2,3,4,5)$,且 S 为每行评论总数(行归一化处理),得出因素集U_1单因素集评判矩阵

$$R_{11} = \begin{pmatrix} 0.351 & 0.448 & 0.100 & 0.090 & 0.010 \\ 0.164 & 0.268 & 0.432 & 0.116 & 0.019 \\ 0.187 & 0.161 & 0.372 & 0.234 & 0.046 \\ 0.136 & 0.254 & 0.346 & 0.254 & 0.010 \end{pmatrix}$$

则对第一种热水器U_1的第一级综合评判结果为：

$$B_{11} = A * R$$

$$=(0.165,\ 0.190,\ 0.327,\ 0.320)\begin{pmatrix}0.351 & 0.448 & 0.100 & 0.090 & 0.010\\ 0.164 & 0.268 & 0.432 & 0.116 & 0.019\\ 0.187 & 0.161 & 0.372 & 0.234 & 0.046\\ 0.136 & 0.254 & 0.346 & 0.254 & 0.010\end{pmatrix}$$

$$=\ (0.1937,0.2587,0.3312,0.1949,0.0237)$$

4) 进行二级评判

将上述计算结果 B 作为二级评判时的评价矩阵$\widetilde{R}$

第一种热水器的综合评判结果

$$\widetilde{B_1}=\widetilde{A_1}*\widetilde{R_1}=\widetilde{A_1}*\begin{bmatrix}A_{11} & R_{11}\\ A_{12} & R_{12}\\ A_{13} & R_{13}\\ A_{14} & R_{14}\end{bmatrix}$$

$$=(0.165,\ 0.128,\ 0.706)*\begin{pmatrix}0.351 & 0.448 & 0.100 & 0.090 & 0.010\\ 0.164 & 0.268 & 0.432 & 0.116 & 0.019\\ 0.187 & 0.161 & 0.372 & 0.234 & 0.046\end{pmatrix}$$

将综合评判结果归一化得

$$\underline{B_1}=(0.2108,0.2216,0.3346,0.1950,0.0369)$$

综合价值系数

$$W_1=\underline{B_1}\cdot C^T=(0.2108,0.2216,0.3346,0.1950,0.0369)\cdot[9\ 7\ 5\ 3\ 1]^T=5.7438$$

5) 同理，对不同产品运用综合评判得出

万和：

$$W_2=\underline{B_2}\cdot C^T=(0.3008,0.4212,0.0446,0.1731,0.0603)\cdot[9\ 7\ 5\ 3\ 1]^T=6.4582$$

万家乐：

$$W_3=\underline{B_3}\cdot C^T=(0.2973,0.1475,0.3176,0.1531,0.0845)\cdot[9\ 7\ 5\ 3\ 1]^T=5.8400$$

史密斯：

$$W_4=\underline{B_4}\cdot C^T=(0.1054,0.1720,0.1756,0.4101,0.1369)\cdot[9\ 7\ 5\ 3\ 1]^T=4.3978$$

$$W_2>W_3>W_1>W_4$$

从评价结果得知，运用 AHP-FCE 对不同品牌的燃气热水器的评判结果为：

万和 > 万家乐 > 能率 > 史密斯

对比 2014 年燃气热水器市场份额，基本可以得出本次评判结果符合大众消费心理。

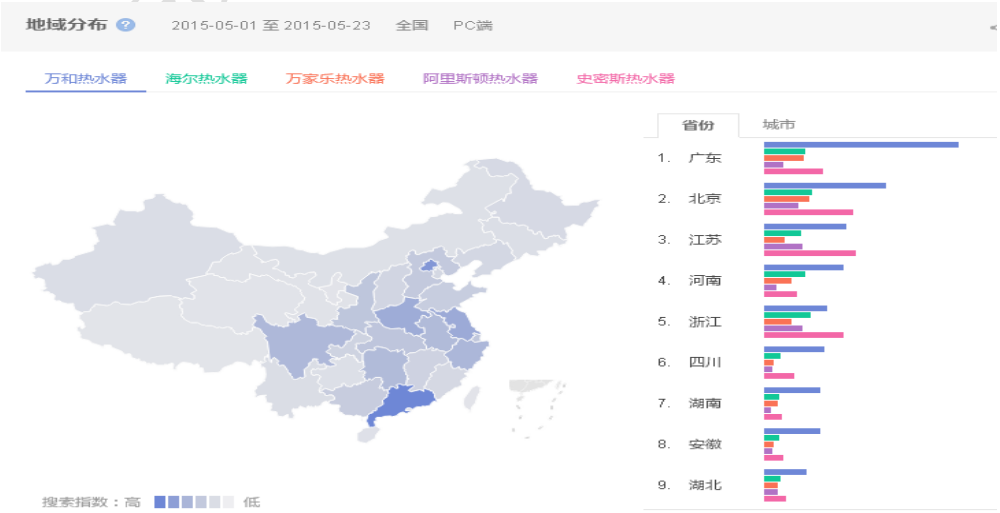
iv. 基于百度指数的用户需求倾向性分析

结合百度指数中的热点新闻分析：以空气能为代表的更为高能效、清洁、安全的热水器将获得更多的政策性支持，并迎来新一轮的发展。

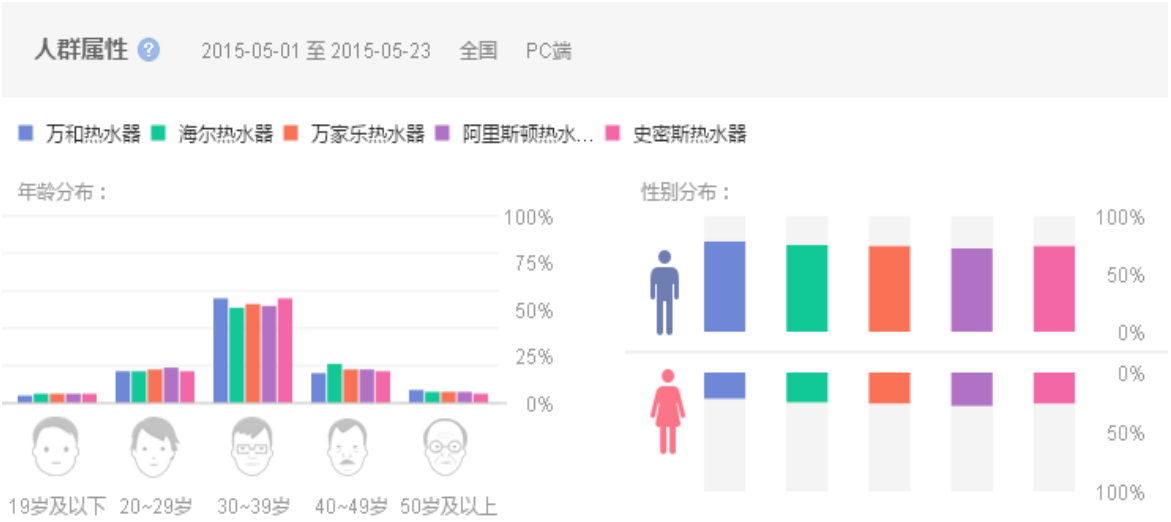
对一年的数据进行分析，以四季为区间：从全国的范围来看，人们在春夏两季的需求比较平淡，在秋冬两季的需求量开始猛增，商家可以针对消费需求衡量营销的主要时间和金钱投入比例，以达到最大的回报率

对月度的数据进行分析，按周为区间划分：分析发现，一般情况下的峰值都为周六或者周日，结合人们的购买习惯，商家可以把营销的重点细分，更多精力放在周末

结合百度知道的关键词提问分析：商家可以此针对自己的商品找出需要改进的问题，不断从消费者的角度出发，生产更符合市场需求的产品。



从地域和品牌的角度来分析：不同品牌可以结合自身情况，从地域的角度来调整自己的推广策略，从而提升广告投放的准度，同时市场份额在某些地域比较低的商家也可以学习该地区市场份额最高的竞争对手的模式，进一步开发市场



从人群属性的角度来分析：从目标年龄，性别的属性来划分，在不同人群集聚地做更为精准的推广

四、 结论

本文是团队三人两周以来对于电商平台热水器消费者需求的数据挖掘分析这一命题，从认知到应用方法解决全程的总结。

文章可分为三个部分，也正好体现团队合作的三个阶段：

1. 从数据挖掘的实现方法入手，结合运用八爪鱼、火车头等数据挖掘工具，通过直接获取和官方数据二次挖掘得出原始数据；之后进行的数据粗、预处理，筛选有效数据并以此评判三个电商平台的差异化，对分析后期处理数据的提供了良好条件。
2. 情感分析阶段，本文参考极限累加方法，基于情感词典和语义规则按照情感词的成分进行数值转化，对褒贬、中性词提取分析，并把情感强度转化为分析每条评论的情感数值。通过统计学方法对情感分析结果的检验，基本满足置信区间。
3. 在对用户需求分析方面，我们采取建立评价体系的数学建模思想，结合层次分析法和模糊综合评判的优点，对热水器选购构建一级和二级评价指标，结合参考百度指数及 F-IDF 评论词频得出的分层评判及模糊综合评判的数值化求解，最终得出对某一种类热水器品牌差异化评分，并得出及探究了各热水器品牌类型的用户购买原因和差异化卖点，实现数据挖掘后对数据的实际应用。

五、 参考文献

[1]. Applying a bilingual model to mine e-commerce satisfaction sentiment. Journal of Management Analytics, 2014. 1(4): p. 285 - 300.
[2]. 张紫琼, 叶强与李一军, 互联网商品评论情感分析研究综述. 管理科学学报, 2010. 13(6): 第 84-96 页.
[3]. 郑安怡, 用于文本情感分析的特征加权改进算法. 计算机工程与应用, 2015.
[4]. 赵文清, 侯小可与沙海虹, 语义规则在微博热点话题情感分析中的应用. 智能系统学报, 2014(1): 第 121-125 页.
[5]. 谢丽星, 周明与孙茂松, 基于层次结构的多策略中文微博情感分析和特征抽取. 中文信息学报, 2012(01): 第 73-83 页.
[6]. 徐健, 基于网络用户情感分析的预测方法研究. 中国图书馆学报, 2013(03): 第 96-107 页.
[7]. 张四维.AHP 在优化高校教学活动管理研究中的应用[J].太原工业大学学报, 1994,4:106-113.
[8]. 周泽义, 樊耀波, 王敏健. 视频污染综合评价的模糊教学方法[J].环境科学, 2000,21(3):22-26