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：东南大学

作品成员：杨情　吴俊康

指导教师：

基于电商平台家电设备的消费者评论数据挖掘分析

摘　要：

在传统的市场中，销售人员与消费者是面对面的。在整个销售的过程中，有经验的销售人员可以敏锐的捕捉到消费者对商品的需求及对各产品的比较，从而总结出出消费者的消费模式。而在电商环境下，交易并没有人与人的互动，有的只是消费者在电商平台下留下的消费痕迹。这些痕迹描述了消费者和他们的需求的联系。这就需要我们运用数据挖掘的手段去探究这些足迹与消费者需要和愿望之间的关系。究竟如何实现这两者之间的模式挖掘正是本文的目的。

本文的研究主要基于消费评论数据，首先运用 ICTCLAS 分词工具对中文文本进行分词；为实现计算机对中文词语的理解，我们运用 word2vec 将中文词语转化为对应的词向量。最后对词向量进行分类和聚类，其中小部分数据基于支持向量机的好评差评分类正确率高达 95%，较好的实现了对评论数据的挖掘。

研究发现用户购买净水器的理由是净水器能够改善用户的水质，保障用户的身体健康。同时，由于购买净水器而不需要再购买瓶装水为用户节省了时间和金钱。安装的方便性是所有用户购买净水器时关注的焦点。服务，净水质量，以及产品的方便性也都是优先关注点。同时我们也发现不同购物平台上消费者的个性化需求，例如国美网站上的用户关心产品是否是正品；而苏宁上的用户则更为关注净水后的水质，而淘宝用户对卖家的发货及物流速度有比较高的期望。同时本文也对各大品牌进行了对比并提出了其主要卖点和需要加强的方面，3M 产品在用户关注的各方面都有较好的表现，用户较多，品牌效应较强，且能提供规范的服务；沁园在产品外观上具有明显优势且有专门的服务专员，但产品安装上需要用户提取预定，同时净水过程中产生的废水较多。特浩恩在四个品牌中不具备较强的竞争力，特别是快递服务有待加强。道尔顿价格公道，服务热忱，回头客较多，但受众面小，需要加强宣传力度。

关键词： 中文分词，词向量，支持向量机，K-均值聚类

The data-mining analysis based on consumers' feedback on household appliances in e-commerce platform

Abstract：

In the traditional market, sales staff and consumers are face to face. In the process of the whole sales, sales staff in experience can be keen to capture consumer's demand for goods, make comparison of various products, and figure out their consumption pattern. Instead, electricity trading lead to no human interactionbut onlyexpense calendar of consumption in the e-commerce platform , which makes important connection between consumers and their demand. It is, therefore, worthy to exploring the tracks and the relationship between the consumer's demands by data-mining methods, and the objective of this article was to realize the data mining and analysis between these two patterns.

In this article, segmentation tools (ICTCLAS) was primarily applied for Chinese text segmentation based on the consumers' feedback data. Then word2vec software was employed to convert Chinese words to the corresponding term vectors for Chinese word identification by computer. Finally, the vector classification and clustering was implemented for data mining, and results showed that the classification accuracy of segmental reviewer datas based on the segmented vector terms was as high as 95%.

We found that the reason consumers buying water purification products is that water purifiers could improve the user's water quality and assure their health. Besides, the possession of a water purifier could help them save lots of money and time because of no bought of bottled water. The convenience of installation was the most focus of attention when coming to a water purifier, and different products demand was shown in different shopping platform, such as the careness of product quality in Gome, the attention to water quality after purfication in Suning, and the expectation of quicker shipping and logistics speed in Taobao. Meanwhile, the merits and demerits of different brands of water purifier were also compared. 3M has good performance in many aspects, and it has many users, strong brand effect, and standardized services; Qinyuan has obvious advantage in the product appearance and special service commissioner, but it requires the user to extract scheduled on product installation and produces wastewater at the same time. Tehaoen in four brands does not have strong competitiveness, especially the express service needs to be strengthened. Dalton owns fair prices, service enthusiasm, so it attracts more repeat customers, but it needs to improve brand popularity

Key words： Chinese word segmentation, word vector, support vector machine (SVM), Kmeans clustering

目 录

1. 挖掘目标

电子商务平台作为一种新兴的商务交易模式，以其成本低廉、快捷、不受时空限制等优点而受到广大企业和消费者的青睐。用户在电商平台上购买产品时通常会留下大量的信息数据，如搜索关键字、购买时关注点、购买步骤、使用、评价。如果能从用户使用过程中留下的信息，敏锐的捕捉信号，从而能更好的为消费者提供服务，抓住消费者的心，保持消费者的忠诚度，必然也能在激烈的市场竞争中脱颖而出。

由于基于产品评论的数据挖掘可以帮助用户了解某一产品在大众心目中的口碑。收集数据并建立有效的情感倾向的中文客户评论情感文本平衡语料库；一方面总结并提出基于支持向量机的中文情感文本分类模型；将文本主题分类的关键技术应用于文本情感倾向分类，构建中文情感文本分类模型，将支持向量机方法应用于中文情感文本分类。另一方面运用现有的工具实现文本词间紧密关系，文本聚类分析。从而提取消费者对产品评论的有效信息。并将建模的模型运用到实际数据分析中，对用户为何购买净水器，及各净水器产品之间的优劣势进行挖掘；为用户和电商之间的双赢提供数据保证。

2. 总体流程

由于文本数据的非结构化特征，同时要处理的情感文本数据量巨大，普通的数据处理工具并不能实现对文本数据的处理；本文运用R语言及 MATLAB 平台编程实现对文本数据的处理。本文的总体架构及思路如下：

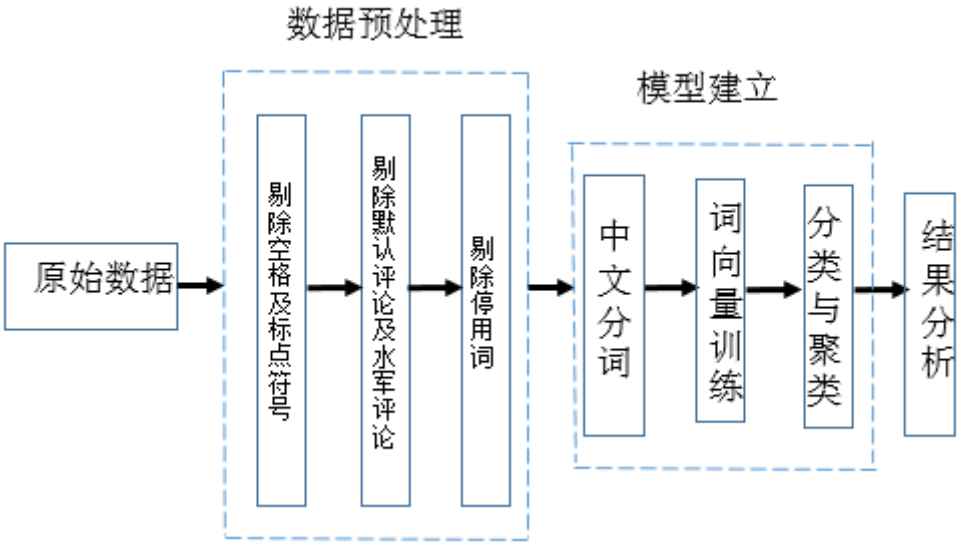


图 1 总流程图

本用例主要包括如下步骤：

步骤一：模型准备：对原始数据进行预处理，去除数据中的噪声，以免对后续数据的处理，及数据

处理的质量产生不良影响，为建模提高有效数据。

步骤二：模型建立：对处理后的模型进行特征提取，特征分类等操作，建立评论数据的情感模型。

步骤三：模型应用：将建立的模型，应用到实际数据处理中，实现对评论数据的挖掘，分析出大数据中的有效信息。

3. 数据预处理

3.1. 剔除空格及标点符号

考虑到中文文本中含有的符号的特殊性， 在剔除空格及标点符号时需要考虑以下几个方面的因素:

(1)半角格式输入的数字0～9， 英文字符(大小写)；

(2)全角格式输入的数字0～9， 英文字符(大小写)；

(3)完整的常用数学符号集；

(4)可能的其他语种的语言符号。

3.2. 剔除默认评论及水军评论

- 删除评论中出现的的默认评论，如：评价方未及时做出评价，系统默认好评!
- 认为连续重复出现一字未改的评论为水军评论，如天猫数据中出现的：

质量非常好，与卖家描述的完全一致，非常满意，真的很喜欢，完全超出期望值，发货速度非常快，包装非常仔细、严实，物流公司服务态度很好，运送速度很快，很满意的一次购物

.质量非常好，与卖家描述的完全一致，非常满意，真的很喜欢，完全超出期望值，发货速度非常快，包装非常仔细、严实，物流公司服务态度很好，运送速度很快，很满意的一次购物

质量非常好，与卖家描述的完全一致，非常满意，真的很喜欢，完全超出期望值，发货速度非常快，包装非常仔细、严实，物流公司服务态度很好，运送速度很快，很满意的一次购物

质量非常好，与卖家描述的完全一致，非常满意，真的很喜欢，完全超出期望值，发货速度非常快，包装非常仔细、严实，物流公司服务态度很好，运送速度很快，很满意的一次购物.

3.3. 剔除停用词

目前，在文本信息处理过程中，一般可以选择字、词或词组作为文本的特征项，但普遍认为选取词作为特征项要优于字和词组。因此，在基于向量空间模型的分类系统中，要将文本表示为向量空间中的一个向量，首先将文本分成词，由这些词作为向量元素来表示文本。但是中文分词器切分出来的所有词条中含有大量的单个独立字， 并且经过研究发现这些单个独立字不仅所携带的文本信息量较少， 而且还对其他实词起到一定的抑制作用， 降低了分类系统的处理效率和准确度， 因此，文本预处理过程有必要将所有的单个独立字过滤。此外，含有数学符号或者英文字符的中文词条对

于正确分类也没有很大的贡献，其携带的信息量也可以忽略。最后，中文文本自动分类系统需要过滤掉纯英文词条，如中文文本中的英文摘要等，以保证代表中文文本的特征为中文词。上述的预处理方法，不仅大大降低了初始文本词条向量的维数，而且有效提高了文本特征向量的中文词纯度。

为了找出噪声词，需要一些标准用于估计词的有效性。最基本和直接的计算标准是基于词频，高频词通常与高噪声值具有相关性。词频标准导致了英文中像the，to 等稳定高频词作为噪声词的合理结果。 在各种文本中出现的频率可以作为一个噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应被看作是噪声词。Luh n认为，当对文本容量不加限制时，词在各个文本中出现的平均概率具有最大的区分力。Ha rman认为，利用一些基于单词出现的文本频率，对于区分词具有较高的精度。

1. 文档频数(DF)

DF是一种简单的评估函数，其值为训练集合中包含此单词的文本数。 DF评估函数的理论假设是当一个词在大量文本中出现时，这个词通常被认为是噪声词。

2. 词频(TF)

T F同样是一种简单的评估函数，其值为训练集合中此单词发生的词频数。 TF评估函数的理论假设是当一个词在大量出现时， 通常被认为是噪声词。

3. 熵计算(entropy calculation， EC)

熵计算是一个基于单词出现的平均信息量，对词的有效性进行计算：

$$W(w)=1+\frac{1}{\ln(n)}\sum_{i=1}^{n}P_i(w)\ln[P_i(w)]$$

式中 $P_i(w)$ 为单词 w 在文本 i 中出现的概率。 每个词的熵被计算出来以后，其结果列表按熵的升序排。

4. 联合熵(union entropy ， UE)

基于词在句子中出现的频率与包含该词的句子频率的联合熵分别计算词条在语料库中各个句子内发生的概率，以及包含该词条的句子在文本中发生的概率 P_j ，计算它们的熵，并依据它们的联合熵选取停用词：

$$W(w_i)=H(w_i)+H(s\mid w_i),$$

$$H(w_i)=-\sum_{j=1}^{n}P_j(w_i)\lg P_j(w_i),$$

$$H(s\mid w_i)=-\sum_{l=1}^{n}P_l(s\mid w_i)\lg P_l(s\mid w_i)$$

$$P_j(w_i)=\frac{f_j(w_i)}{\sum_{j=1}^{n}f_j(w_i)}$$

$$P_l(s\mid w_i)=\frac{f_l(s\mid w_i)}{\sum_{j=1}^{m}f_l(s\mid w_i)}$$

式中： $f_j(w_i)$ 单词 w 在句子 j 中出现的频率； n 为句子数. $f_l(s|w_i)$ 为包含 w_i 的句子在文本 l 中出现的频率； m 为文本数

采用联合熵作为噪声词的提取方法的理论依据是，当一个词在句子中出现的平均信息量和包含该词的句子的平均信息量较大时，表示该词较为普通.研究表明，应用该方法可以有效避免语料选取不均衡造成的停用词选取错误.

本文选用DF方法筛选出如下停用词：我，有，的，了，是，所有的英文词。将筛选出的停用词从文本中剔除，以确保分词的正确性。

4. 中文分词

4.1. 中文分词方法概述

中文分词就是将中文连续的字序列按照一定的规则重新组合成词序列的过程，是中文信息处理的基础。自然语言处理的过程中词语往往表述为具有独立语法、语义的最小的自然语言构成单位，要进行自然语言的处理，必然需要选择一个单位来对语言进行分割处理研究。在西文中，词与词之间有明显的标记，由空格隔开，词语不需要进行其它处理；而在汉语、日语等东方语言中，由于词与词之间没有明显的切分标记，要对文本进行研究则需要将文本进行切分，即分词。此外，还需要对词语进行词性标记等。

中文文本的自动分词技术是智能检索、机器翻译、文献标引、自然语言理解与处理的基础，也是中文文本分类的一个关键的环节，对中文进行处理首先要对文本进行分词。中文分词的理论与方法决定了中文分词系统的性能与效率。同比于西文文本分类相比较，中文文本分类不同于西文文本分类的一个重要环节就是文本数据的预处理，对中文文本进行分词处理时文本预处理的首要环节。目前，中文分词技术的发展阶段从最初的基于词典的方法，发展到了基于统计语言模型的分词方法并且已经取得了较好的研究效果。 国内主流的对分词系统所做的研究中，正在研究的或者采用的方法主要有三种：基于词典规则的机械分词方法、基于人工智能的分词方法以及基于语料库的统计分词方法。

1. 基于词典规则的机械分词方法

它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配。若在词典中找到某个字符串，则匹配成功(识别出一个词)。该方法有三个要素，即分词词典、文本扫描顺序和匹配原则。文本的扫描顺序有正向扫描、逆向扫描和双向扫描。匹配原则主要有最大匹配、最小匹配、逐词匹配和最佳匹配。

优点：简单，易于实现。

缺点：1)匹配速度慢；2)存在交集型和组合型歧义切分问题；3)词本身没有一个标准的定义，没有统一标准的词集；4)不同词典产生的歧义也不同；5)缺乏自学习的智能性

2. 基于人工智能的分词方法

其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统和总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。目前基于理解的分词方法主要有专家系统分词法和神经网络分词法等。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

3. 基于语料库的统计分词方法

该方法的主要思想：词是稳定的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻出现的概率或频率能较好反映成词的可信度。可以对训练文本中相邻出现的各个字的组合的频度进行统计，计算它们之间的互现信息。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可以认为此字组可能构成了一个词。该方法又称为无字典分词。

该方法所应用的主要的统计模型有：N元文法模型、隐Markov模型和最大熵模型等。在实际应用中一般是将其与基于词典的分词方法结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

综上所述，本文选用基于语料库的统计分词方法进行分词，利用ICTCLAS分词工具实现。

4.2. 基于 ICTCLAS 的分词

4.2.1 分词原理

1. 基于层次隐马模型的汉语词法分析

针对汉语词法分析各个层面的处理对象及问题特点，我们引HHMM 统一建模。该模型包含原子切分、普通命名实体识别、嵌套的复杂命名实体识别、基于类的隐马切分、词类标注共五个层面的隐马模型，如图 3-1 所示。其中，N-最短路径粗切分可以快速产生N个最好的粗切分结果，粗切分结果集能覆盖尽可能多的歧义。在整个词法分析架构中，二元切分词图是个关键的中间数据结构，它将未登录词识别、排歧、分词等过程有机的进行了融合，在分词模型中会详细地介绍。

原子切分是词法分析的预处理过程，主要任务是将原始字符串切分为分词原子序列。分词原子指的是分词的最小处理单元，在分词过程中，可以组合成词，但内部不能做进一步拆分。分词原子包括单个汉字、标点以及由单字节、字符、数字等组成的非汉字串。如“2002.9，ICTCLAS 的自由源码开始发布”对应的分词原子序列为“2002.9/，/ICTCLAS/的/自/由/源/码/开/始/发/布/”。在这层HMM中，终结符是书面语中所有的字符，状态集合为分词原子，模型的训练和求解都比较简单。过程本质一样，在这里不重复论述。

2. 基于类的隐马尔科夫分词算法

首先，我们可以把所有的词按照图 3-2 分类，其中，核心词典中已有的每个词对应的类就是该

词本身。这样假定核心词典中收入的词数为|Dict|，则我们定义的词类总数有：|Dict|+6。

给定一个分词原子序列S，S的某个可能的分词结果记为$W=(w_1,\cdots,w_n)$，W 对应的类别序列记为$C=(c_{1,}\cdots,c_n)$，同时，我们取概率最大的分词结果$W^{\#}$作为最终的分词结果。则：

$$W^{\#}=\arg\max_{W} P(W)$$

利用贝叶斯公式进行展开，得到：

$$W^{\#}=\arg\max_{W} P(W\mid C)P(C)$$

将词类看作状态，词语作为观测值，利用一阶 HMM 展开得：

$$W^{\#}=\arg\max_{W}\prod_{i=1}^{n} p(w_i\mid c_i)p(c_i\mid c_{i-1})$$

其中c_0为句子的开始标记 BEG，c_n为句子的结束标记 END，下同 为计算方便，常用负对数来运算，则：

$$W^{\#}=\arg\max_{W}\sum_{i=1}^{n}[-\ln p(w_i\mid c_i)-\ln\, p(c_i\mid c_{i-1})]$$

根据图 3-2 中类c_i的定义，如果w_i在核心词典收录，可以得到$c_i=w_i$，因此$p(w_i\mid c_i)=1$。在分词过程中，我们只需要考虑未登录词的 $p(w_i\mid c_i)$。最终所求的分词结果就是从初始节点 S 到结束节点 E 的最短路径，这是个典型的最短路径问题，可以采取贪心算法，如 Dijkstra 算法快速求解。

3. N-最短路径的切分排歧策略

其基本思想是在初始阶段保留切分概率 P(W)最大的 N 个结果，作为分词结果的候选集合。在未登录词识别、词性标注等词法分析之后，再通过最终的评价函数，计算出真正最优结果。实际上，N-最短路径方法是最少切分方法和全切分的泛化和综合。一方面避免了最少切分方法大量舍弃正确结果的可能，另一方面又大大解决了全切分搜索空间过大，运行效率差的弊端。

4.2.2 分词流程图

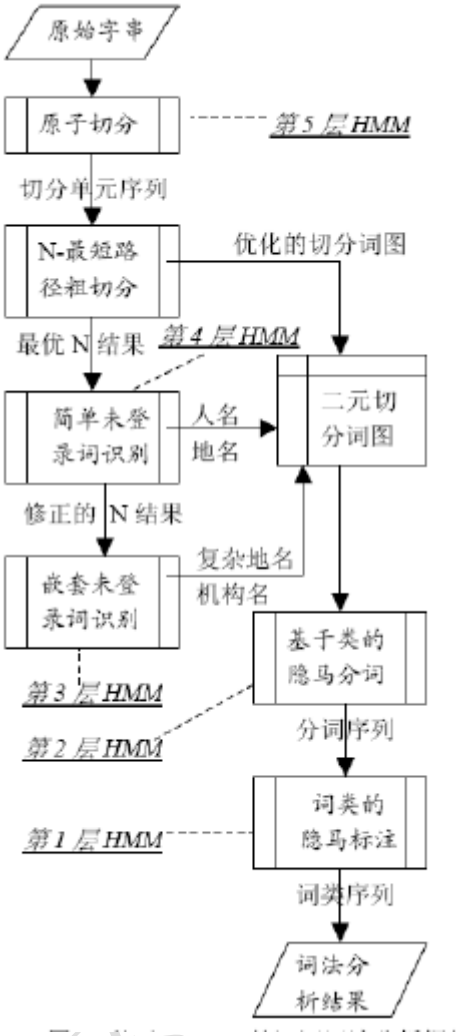


图 2 分词流程图

4.3. 分词结果示例

分词前：

东西看着不错

可以看到里面的脏东西

前置到了净水器还没到

安装好了不错的东西

今天安装好了，等橱柜装好了再买个终端净水

收到了，很不错等安装

这个价格挺值的，比以前买的那个还要好

性价比挺高的

很满意的东西

怎么安装呢，不会啊

性价比不错

性价比挺高的，不错的产品

很满意的东西，好用

这是多久清洗一次啊？

东西很不错，看着不错

分词后：

东西 看着 不错

可以 看到 里面 脏 东西

前 置 到 了 净水 器 还 没 到

安装 好 了 不错 的 东西

今天 安装 好 了 等 橱柜 装 好 了 再 买 个 终端 净水

收到 了 很 不错 等 安装

这个 价格 挺 值 的 比 以前 买 的 那个 还要 好

性 价 比 挺 高 的

很 满意 的 东西

怎么 安装 呢 不 会

性 价 比 不错

性 价 比 挺 高 的 不错 的 产品

很 满意 的 东西 好 用

这 是 多久 清洗 一次

东西 很 不错 看着 不错

4.4. 本章小结

经过分词处理之后，原始的文本会被处理成一系列由空格隔开的单个词。接下来可以使用词频统计或神经网络算法构建语言模型，在构建语言模型的过程中得到词向量，从而实现对文档的进一步处理。

5. 词频统计

为了能精确的对实验文档进行表示，基于统计的方法需要对整篇文章进行分词，进而对每个单词的出现频率进行统计。词是最小的、能独立活动的、有意义的语言成分。计算机的所有语言知识都来自机器词典(给出词的各项信息)、句法规则(以词类的各种组合方式来描述词的聚合现象)以及有关词和句子的语义、语境知识库。

在中文文档中，从形式上看，词是组合稳定的字而得到的，因此，在上下文环境中，同时出现

相邻的字次数越多，就表示该相邻字组合越有可能构成一个词。因此相邻字的组合共现的频率或概率能够较好的反应组成词的可信度。这就是词频统计的基本原理，这种技术发展至今已经有许多不同的统计原理。

1. 互信息原理

对有序汉字串 AB 中汉字 AB 之间的互信息定义如公式：

$$\log_2 \frac{P(A,B)}{P(A)P(B)}$$

互信息体现了汉字之间结合关系的紧密程度，当紧密程度高于某一个阈值时，便可认为该字组合可能构成了一个词。其中，$P(A,B)$ 为汉字串 AB 联合出现的概率，$P(A)$ 为出现汉字串 A 的概率，$P(B)$ 为汉字串 B 出现的概率

2. N-Gram统计模型

N-Gram统计计算语言模型的思想是：一个单词的出现与其上下文环境中出现的单词序列密切相关，第n个词的出现只与前面n-1个词相关，而与其它任何词都不相关，设W1W2W3…Wn是长度为n的字串，则字串W的似然度用方程表示如公式：

$$P(W)=\prod_{i=1}^{n} P(W_i \mid W_{i-n+1}W_{i-n+2}\cdots W_{i-1})$$

不难看出，为了预测词Wn的出现概率，必须知道它前面所有词的出现概率。从计算上来看，这种方法太复杂了。如果任意一个词Wi的出现概率只同它前面的两个词有关，问题就可以得到极大的简化。这时的语言模型叫作三元模型(tri-gram) 如公式：

$$P(W)\approx P(W_1)P(W_2 \mid W_1)\prod_{i=3\ldots n} P(W_i \mid W_{i-2}W_{i-1}))$$

符号 $\prod_{i=3\ldots n} P(W_i \mid W_{i-2}W_{i-1})$ 示概率的连乘。一般来说，N元模型就是假设当前词的出现概率只同它前面的N-1个词有关。重要的是这些概率参数都是可以通过大规模语料库来计算的，比如三元概率如公式：

$$P(W_i \mid W_{i-2}W_{i-1})\approx \frac{count(W_{i-2}W_{i-1}W_i)}{count(W_{i-2}W_{i-1})}$$

式中count(…)表示一个特定词序列在整个语料库中出现的累计次数。

3. 基于匹配的词频统计方法

单关键词匹配算法在国内外都对其进行了深入的研究，主要有BF(Brute Force)算法、KMP (Knuth-Morris-Pratt)算法、BM(Boyer-Moore)算法等，都已基本上完善。BF算法直观、简单，但涉及多次回溯，算法效率低，时间复杂度为O(m*n)；KMP算法实现了无回溯匹配，避免了BF算法中频繁的回溯，文本串中的每个字符只匹配一次，普遍提高了模式匹配的工作效率，时间复杂度为O(n+m)；BM算法实现了跳跃式匹配，文本串中的部分字符不需要匹配，时间复杂度为O(m*n)，但其最优情况下的时间复杂度为O(n/m)。

本文对评论信息分别进行词频统计，计算出重复的关键词出现次数，然后直接按照由大到小进行词频统计排序。部分统计结果如下：

word freq

不错 40522

安装 33534

非常 18263

效果 17088

质量 15462

满意 14728

过滤 12311

服务 12230

方便 12194

净水 12161

水质 11437

为方便观察词频统计中的信息特征，本文将词频转换为词云图如下：词的大小代表该词在文本中出现的频率，词越大，表示出现的频率越高。

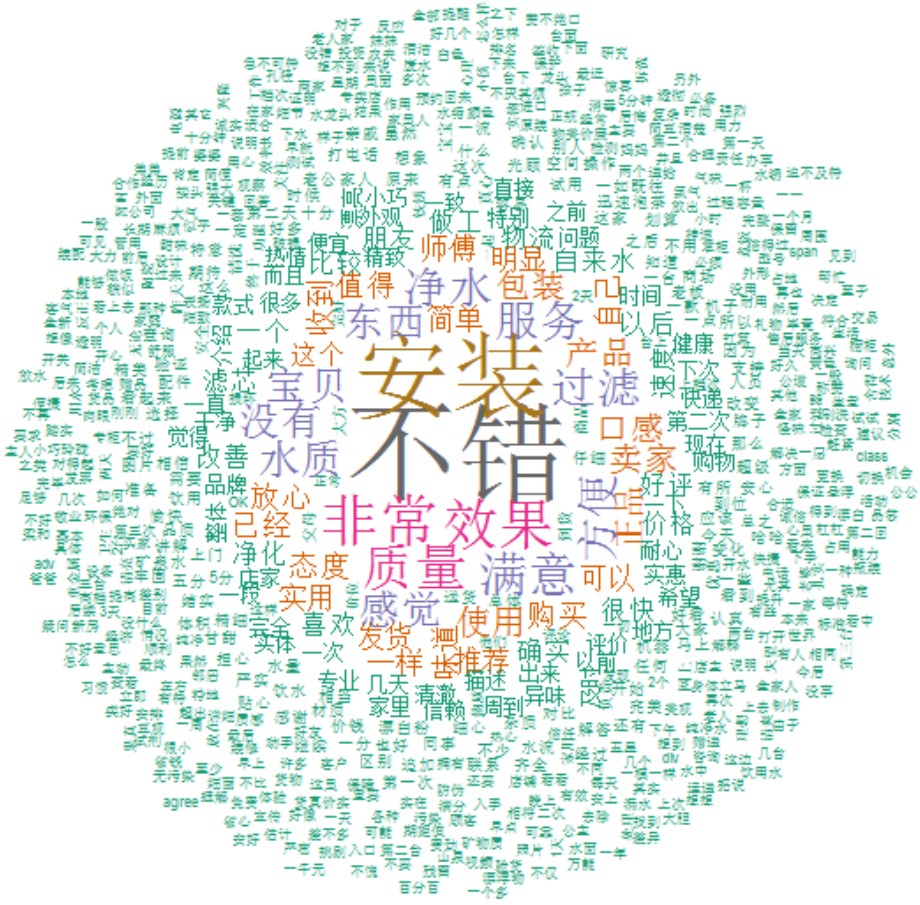


图3 词频图

6. 文档模型

文本分类器不能对情感文本直接进行分类，所以首先要将文本表示成分类器能够处理的表示方法，这个过程主要是建立文档模型。

6.1. 文档模型概述

1. 布尔模型

布尔模型(BM， Boolean Model)是第一个被提出的模型，布尔模型的关键是：一个文档被表示为特征项的集合，而特征项的权值只有 0 和 1。布尔模型的优点主要是模型简单，容易理解。但是布尔模型表征文本的能力不强，因为仅有 0，1 权值，而没有一定的权值大小，因此，特征项对文本语义的重要程度无法在布尔模型中被反映出来，并且其逻辑表达式过于严格。虽然 G Salton 等人在波尔模型的基础上增加了特征项的权重，推出了扩展布尔模型，有所改进，但由于种种限制仍未

能达到理想的效果。

2．向量空间模型

向量空间模型(VSM， Vector Space Model)是将基于这种模型的每篇文档都转化为高维向量空间中的一个向量，向量空间模型的关键是：把文档看成一组独立的 n 维特征向量，对每一个特征都赋予一个权值。VSM 模型的基本思想是将文本数据表示成一个向量空间，将文本中的词作为文本向量的维度特征，这种表示方法的提出时基于假设：

文本数据中词特征出现的先后次序无关，每一个词对应特征向量空间中的一个特征维度。其基本概念为： 文本数据表示成由 S 个特征项的向量，文本数据 T={t1，t2，…，ts}，其中 ts 为文本向量的第 S 维特征项，特征项 ts 通常会被赋予一定的权值，表示其在文本 T 中的重要程度，并将 ts 的权值表示为 WI。

其中，权值的计算主要根据两方面的原则：首先，特征词在某篇文档中出现的频率越高，它与文档主题的相关度越高；另外，特征词在整个文档数据集中出现的频率越高，其表征某篇特定文本特征的能力越小。

自从 Salton 等人将 VSM 模型提出，并且成功应用于著名的 Smart 系统后，VSM 模型被广泛的应用于文本分类、自动标引、信息检索等领域。VSM 模型的主要优点：文本表示方法上的巨大优势，将对文本的处理转化为了向量空间中的向量运算，大大降低了问题的复杂度，提高了文本处理的速度。不足表现在：VSM 模型的假设的提出是建立在文本向量中的特征项是相互独立的基础上，但是显然这一假设在自然语言文本中是不成立的。

在大多数情况下，文本分类基本都采用了 VSM 模型，近年来，VSM 模型在文本挖掘系统中显示出了普遍优良的性能，目前已成为最简便、高效的文本表示方法之一。

3．潜在语义索引

潜在语义索引(LSI， Latent Semantic Indexing)模型是从语义相关的角度，利用概念标引代替关键词标引，为文本选择标引词，可以看做是对向量空间模型的一种改进，其基本思想是利用文本中词之间的关系来构造词和文本间的关系。主要优点是：对向量空间模型进行了改进，客服了同一篇文档词多义性和统一性带来的问题，可以较大限度的保证文档特征信息量，同时能够大大缩减文档向量空间的维度。但是，LSI 模型增加了文本预处理的复杂度，在分类文本数据预处理的效度上得不到很好的效果。

4．概率模型

概率模型(Probabilistic Model)是由 Stephen Robertson 和 Spark Jones 等人提出的，该模型综合考虑了词频、文档频率和文档长度等因素，将文档和用户兴趣按照一定的概率关系融合，形成了著名的 OKAPI 公式。其主要的优点是使得关键词和文档之间的相关性得到更准确的描述。但是，由于需要事先确定关键词和文档之间的相关概率，对于大量的文本数据来说工作量极其巨大，不能保证分类数据预处理的高效性，因此 PM 模型并没有被广泛的应用。

以上介绍的四种常用特征表示模型分别从不同的角度，采用不同的方法处理特征权值、类别学

习和相似计算等问题。VSM 模型是目前公认最好的文档表示模型。

6.2. 基于向量空间模型的词向量

6.2.1 向量空间模型的词向量模型

向量空间模型的基本思想是将文本文档 id 看作由一组特征项) $(T_1,T_2,\cdots T_n)$ 构成，特征项通常是对文本类别有区分能力的字词或短语。对于每个特征项 T_i，可根据其在文档中的重要程度赋予一定的权重 W_i，这样每个文档就表示成由 n 个特征权重组成的向量 $(W_1,W_2,\cdots W_n)$。文档集可以看成是一个 n 维空间，每一个文档对应文档空间中的一个点，$(W_1,W_2,\cdots W_n)$ 便是其对应的坐标值。于是向量空间模型便把对文本的计算转换成了对向量空间中点的计算，极大的方便了计算机的分析处理

1. 统计语言模型

统计语言模型是用于刻画一个句子出现概率的模型。给定一个由 n 个词语按顺序组成的句子 $S=(w_1,w_2,\cdots w_n)$，则概率 $p(S)$ 即为统计语言模型。通过贝叶斯公式，可以将概率 $p(S)$ 进行分解，即 $p(S)=P(w_1)\bullet P(w_2\mid w_1)\bullet P(w_3\mid w_1^2)\cdots P(w_n\mid w_1^{n-1})$。所以，要计算一个句子出现的概率，只需要计算出在给定上下文的情况下，下一个词为某个词的概率即可，即 $p(w_i\mid context(w))$。当所有条件概率 $p(w_i\mid context(w))$ 都计算出来后，通过连乘即可计算出 $p(S)$。所以，统计语言模型的关键问题在于找到计算条件概率 $p(w_i\mid context(w))$ 的模型

2. 词向量

词向量具有良好的语义特性，是表示词语特征的常用方式。根据文献表述，词向量每一维的值代表一个具有一定的语义和语法上解释的特征。所以，可以将词向量的每一维称为一个词语特征。词向量具有多种形式，distributed representation 是其中一种。一个 distributed representation 是一个稠密、低维的实值向量。distributed representation 的每一维表示词语的一个潜在特征，该特征捕获了有用的句法和语义特性。可见，distributed representation 中的 distributed 一词体现了词向量这样一个特点：将词语的不同句法和语义特征分布到它的每一个维度去表示。

3. 神经网络语言模型

2003 年，Bengio 等提出一个基于 3 层神经网络的自然语言估计模型 NNLM（Neural Network Language Model），如图 4 所示。NNLM 可以计算某一个上下文的下一个词为 wt 的概率，即 p（wt=i丨context），词向量是其训练的副产物

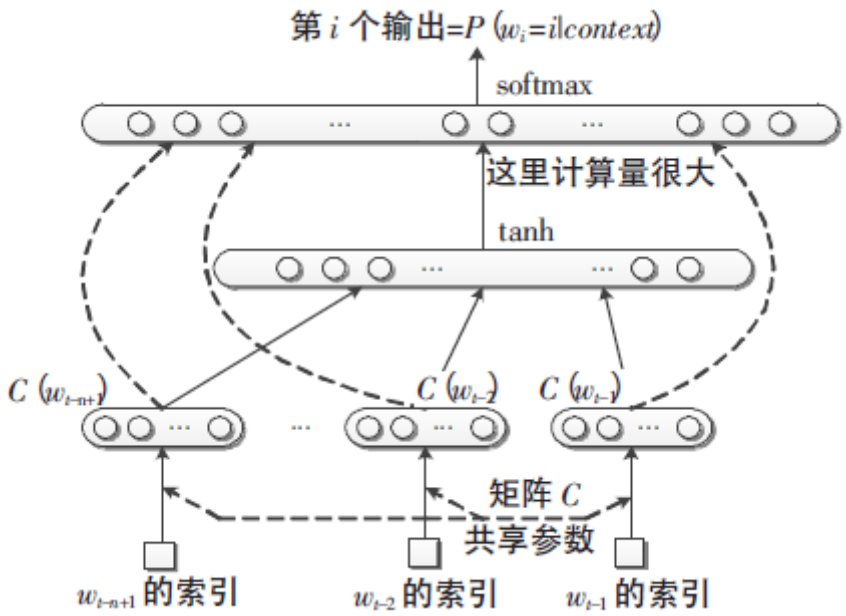


图 4　神经网络语言模型

NNLM 根据语料库 C 生成一个对应的词汇表 V。V 中的每一个词汇都对应着一个编号 i。为了确定神经网络的参数，需要通过语料库来构建训练样本并作为神经网络的输入。样本的构建过程为：对于 C 中的任意一个词 w_i，获取其上下文 $context(w_i)$（例如前 n-1 个词），从而得到一个元组 $(context(w_i), w_i)$。以该元组作为神经网络的输入进行训练。由图 4 可知，对于每一个词，NNLM 都将其映射成一个向量 $C(w_i)$，该向量即为词向量。NNLM 在输出层上的 softmax 归一化函数为 $p(w_i \mid w_{i-1},\cdots,w_{i-n+2},w_{i-n+1}) = \frac{e^{y_{w_i}}}{\sum_i e^{yi}}$，其中 y_i 是下标为 i 的词语对应的未归一化的 log 概率值。可见，当词汇表很大时，softmax 的操作将非常耗时。为了解决这个问题，Bengio 等在 NNLM 的输出层上进行层次化改进，大大提高了 NNLM 的计算效率。层次化操作的基本思想是对词语进行分类。例如，在包含 10000 个词的词汇表的情况下，对于每一个概率 $p(w_i \mid context(w))$，NNLM 的输出层要对 w_i 的每个可能取值进行一次归一化，每个取值总共需要计算 10000 次。对词汇表中的词进行分类后，比如分为 100 个类别，每个类别为 100 个词。则对于 w_i 的每个可能取值，只需要进行 200 次计算即可，加速比达到 50。可以对词汇表进行更多层次的划分来提高计算的速度。

4. Log__Linear 和 Log__Bilinear 模型

Log__Linear 模型可用于自然语言建模。同时，Log__Linear 模型也是 word2vec 训练模型的基础。Log__Linear 模型的组成部分包括：

（1） 一个可能的输入集合 X；

（2） 一个有限的标记集合 Y；

（3） 一个模型中的特征和参数数目的正整数 d；

（4） 一个将(x,y)映射到一个特征向量$f(x,y)$的函数$f:X\times Y\to R^d$

（5） 一个参数向量$v\in R^d$

Log__Linear 模型定义了一个条件概率，即对于任意的$x\in X, y\in Y$

$$p(y\mid x;v)=\frac{\exp(v\bullet f(x,y))}{\sum_{y'\in Y}\exp(v\bullet f(x,y'))}$$

其中$\exp(x)=e^x$

$f(x,y)$是一个代表(x,y)的特征向量，$f(x,y)$的每一维$f_k(x,y)$代表一个特征，每一个特征与v的一个维度v_k相关联。v中每一维的参数v_k需要通过训练得到。

在 Log__Linear 模型基础上，Hinton 提出了 Log__Bilinear 模型。Log__Linear 和 Log__Bilinear 模型的主要区别在于映射函数部分$f(x,y)$。在 Log__Bilinear 模型中，$f(x,y)$将输入(x,y)直接映射成一个 d 维特征向量，而 Log__Bilinear 模型直接使用输入y对应的向量。通过借鉴神经网络语言模型中的层次化思想，Hinton 提出了层次化的 Log__Bilinear 模型，这也是 Word2vec 所使用的模型。

6.2.2 word2vec 词向量训练

1.word2vec 的训练模型

Word2vec 是 Mikolov 等所提出模型的一个实现，可以用来快速有效训练词向量。Word2vec 包含了两种训练模型，分别是 CBOW 和 Skip_gram，如图所示：

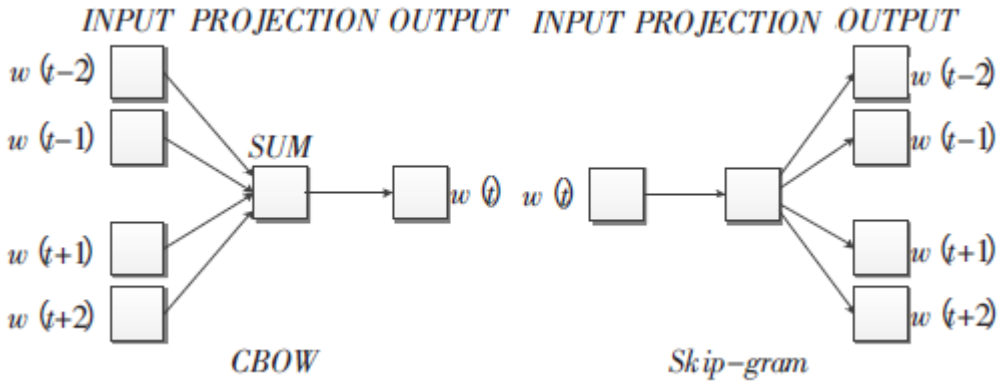


图 5 CBOW 和 Skip_gram

从图 5 可以看出，CBOW 和 Skip__gram 模型均包含输入层、投影层和输出层。其中，CBOW 模型通过上下文来预测当前词，Skip__gram 模型则通过当前词来预测其上下文。Word2vec 提供了两套优化方法来提高词向量的训练效率，分别是 HierachySoftmax 和 Negative Sampling。将训练模型和优化方法进行组合可得到 4 种训练词向量的框架，如表 1 所示。

表 1 word2vec 词向量训练框架

模型	CBOW	Skip__gram
Hierachy Softmax	CBOW+HS	Skip__gram+HS
Negative Sampling	CBOW+NS	Skip__gram+NS

- CBOW＋HS 和 Skip__gram＋HS

从 Word2vec 的源码中可以归纳出 CBOW＋HS 和 Skip__gram＋HS 模型的训练框架，如图 6 所示。

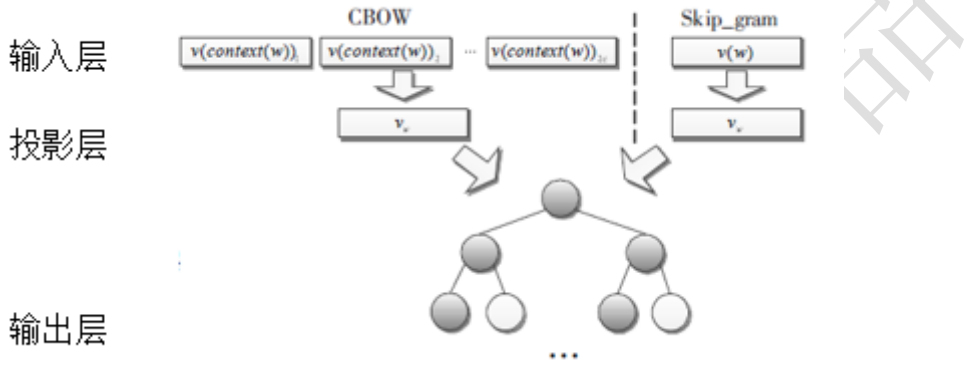


图 6 CBOW+HS 和 Skip__gram+HS 训练框架

根据文献，总结得到两种训练框架的共同点和区别，如表 2 所示

表 2 CBOW+HS 和 Skip__gram+HS 训练框架对比

对比项	CBOW+HS	Skip__gram+HS
训练样本	已知上下文，预测下一个词	已知当前词预测上下文
输入层	词语 w 前后各 c 个词对应的词向量	当前词 w 对应的词向量
投影层	输入层的 2c 个词向量的累积	当前词 w 对应的词向量
输出层	以语法在语料库中的词频作为权值构造的一颗二叉树。叶子节点对应词汇表中的所有词语。假设叶子节点为 N 个，则非叶子节点为 N-1 个。叶子节点和非叶子节点对应一个向量。其中叶子节点对应的向量即为词向量，而非叶子节点对应的向量是一个辅助向量。	
训练参数	词汇表中词语对应的词向量，输出层非叶子节点对应的辅助向量	

在模型的训练过程中，梯度是训练参数更新的依据。为了获得梯度公式，需要构造出训练模型的目标函数。在 CBOW＋HS 框架的训练中，给定一个训练样本 $(context(w_i), w_i)$，则在输出层从根节点到词 w 的路径上，对于每一个非叶子节点均对应一个辅助向量 θ_j^w 和一次二分类且左右两边各对应一个哈夫曼编码 l_j^w。其中，j 表示从根节点到叶子节点 w 路径上非叶子节点的编号。

对于l_j^w，规定向左的分支对应 1，向右的分支对应 0。这样，通过若干二分类后，最终可到达叶子节点 w。可将非叶子节点向左的分支定义为负类，将向右的分支定义为正类。通过逻辑回归知识，可以得到一个二分类被分成正类的概率为$\delta(v_w^T\theta)=\frac{1}{1+e^{-v_w^T\theta}}$，被分成负类的概率为$1-\delta(v_w^T\theta)$。其中，$v_w^T$是$context(w_i)$中包含的词所对应词向量的累加和，而$\theta$为非叶子节点对应的辅助向量。沿着从根节点到叶子节点 w 的路径将每个非叶子节点的二分类概率相乘即为$p(w_i\mid context(w))$。

对于 Skip__gram＋HS，给定一个训练样本$(context(w_i),w_i)$，其中$context(w_i)$包含 2c 个词。可以将通过 w 预测$context(w_i)$的问题，即$p(w_i\mid context(w))$转换成 2c 个通过 w 预测下一个词为u的问题$p(u\mid w)$，其中$u\in contex(w)$。$p(u\mid w)$可以利用 CBOW＋HS 框架的思路来解决，即将 u 视为叶子节点。不同的是，在 CBOW＋HS 框架中，v_w^T是$context(w_i)$中包所有词所对应词向量的累加和，而在 Skip__gram＋HS 中指的是w对应的向量。Skip__gram＋HS 框架的目标函数为$p(contex(w)\mid w)=\prod_{u\in contex(w)}p(u\mid w)$。

- CBOW＋HS 和 Skip__gram＋HS

CBOW＋HS 和 Skip__gram＋HS 框架输出层的哈夫曼树构造过程相对复杂。所以，CBOW＋NS 和 Skip__gram＋NS 框架采用了一种替代的方法，即采用相对简单的负取样来提高词向量训练的速度。

对于 CBOW＋NS 框架，已知样本$(context(w_i),w_i)$，则w为一个正样本，而词汇表中的其他词为负样本。可以采用不同的负采样算法来选择负样本。在确定关于$context(w_i)$的一个非空的负样本集合 NEG（w）后，对于词汇表中的任何词w'，若$w'=w$，给其为 1 的标签，否则给其为 0 的标签。这样，正样本的标签为 1，负样本的标签为 0。这里的标签可类比成 CBOW＋HS 框架输出层哈夫曼树中非叶子节点的左右二分类编码。

对于 Skip__gram＋NS 框架，已知样本（$(context(w_i),w_i)$。则对于一个上下文词w'，当$w'=w$，给其为 1 的标签，否则给其为 0 的标签。这里的标签同样可类比成 CBOW＋HS 框架输出层哈夫曼树中非叶子节点的左右二分类编码。

6.2.3 Word2vec 的应用

从文献［10］中导出的 Word2vec 项目包含了若干个 C 语言原文件和 linux 脚本。其中，与训练词向量相关的文件主要包括 Word2vec.c、demo－word.sh 和 distance.c 3 个文件。Word2vec.c 文件包含了 Word2vec 各个模型的实现，demo－word.sh 包含了模型训练需要指定的参数列表，而 distance.c 文件可以计算不同词向量间的余弦值。在 linux 平台下，通过 demo－word.sh 脚本即可启动 Word2vec 进行词向量的训练。distance.c 文件以训练好的词向量文件作为输入，可以获取与某个

词语在语义上最相近的词语列表。

Word2vec 提供了许多超参数来调整训练过程。不同参数的选择对生成的词向量的质量以及训练的速度有影响。Word2vec 可调整的参数如表 3 所示

表 3 word2vec 的训练参数

参数	说明	参数	说明	参数	说明
train	输入文件的路径	output	输出文件位置	size	词向量的维数
window	窗口大小	hs	是否采用softmax体系	negative	负样本的数量
threads	开启的线程数量	min-count	词语出现的最小阈值	alpha	学习率初始值
binary	是否用模式保存数据	cbow	是否采用CBOW算法	classes	输出词类别

本文参数选择如下：

./word2vec 4_fenci.txt -output vector2fencis.bin -cbow 0 -size 200 -window 5 -negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 0

6.3. 向量模型及应用

将分好词的中文语料作为 Word2vec 的输入文件并指定合适的训练参数，即可进行中文词向量的训练。

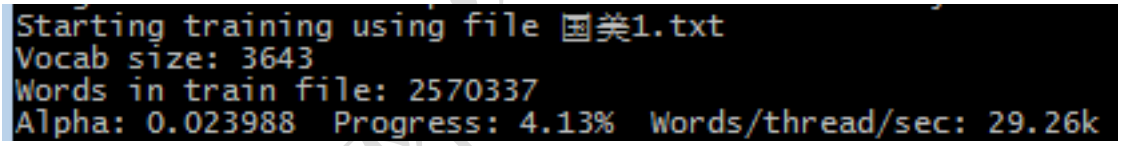


图 7 词向量训练过程

训练所得词向量：

```
</s> 0.002001 0.002210 -0.001915 -0.001639 0.000683 0.001511 0.000470 0.000106
-0.001802 0.001109 -0.002178 0.000625 -0.000376 -0.000479 -0.001658 -0.000941
0.001290 0.001513 0.001485 0.000799 0.000772 -0.001901 -0.002048 0.002485
0.001901 0.001545 -0.000302 0.002008 -0.000247 0.000367 -0.000075 -0.001492
0.000656 -0.000669 -0.001913 0.002377 0.002190 -0.000548 -0.000113 0.000255 -
0.001819 -0.002004 0.002277 0.000032 -0.001291 -0.001521 -0.001538 0.000848
0.000101 0.000666 -0.002107 -0.001904 -0.000065 0.000572 0.001275 -0.001585
0.002040 0.000463 0.000560 -0.000304 0.001493 -0.001144 -0.001049 0.001079 -
0.000377 0.000515 0.000902 -0.002044 -0.000992 0.001457 0.002116 0.001966 -
0.001523 -0.001054 -0.000455 0.001001 -0.001894 0.001499 0.001394 -0.000799 -
```

图 8 词向量结果

通过训练得到的词向量文件，我们可以计算词间关系距离。图展示了与输入词关系最紧密在语词。

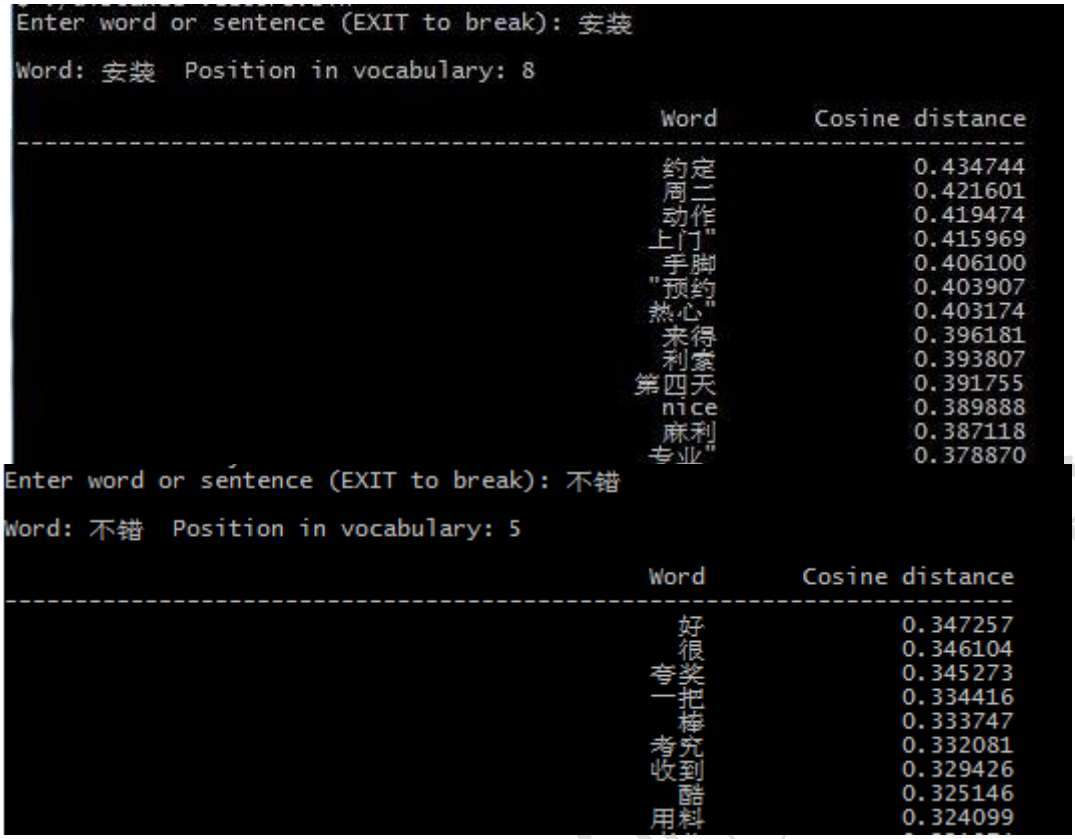


图 9 词间关系距离

7. 词向量聚类及分类

7. 1. 基于词向量的支持向量机文本情感分类

7.1.1 支持向量机（SVM）原理

SVM 的基本思想是：对于特征空间中两类点不能靠超平面分开的非线性问题，SVM 采用映照方法将其映照到更高维的空间，并求得最佳区分二类样本点的超平面方程，作为判别未知样本的判据。

线性可分情形：

SVM算法是从线性可分情况下的最优分类面（Optimal Hyperplane）提出的。所谓最优分类面就是要求分类面不但能将两类样本点无错误地分开，而且要使两类的分类空隙最大。d维空间中线性判别函数的一般形式为 $g(x)=w^Tx+b$，分类面方程是 $w^Tx+b=0$，我们将判别函数进行归一化，使两类所有样本都满足 $|g(x)|\geq 1$，此时离分类面最近的样本的 $|g(x)|=1$，而要求分类面对所有样本都能正确分类，就是要求它满足

$$y_i(w^Tx_i+b)-1\geq 0, i=1,2,\cdots,n \text{。}$$

式中使等号成立的那些样本叫做支持向量（Support Vectors）。两类样本的分类空隙（*Margin*）的间

隔大小：

$$Margin=2/\|w\|$$

因此，最优分类面问题可以表示成如下的约束优化问题，即在约束条件下，求函数

$$\phi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w^T w)$$

的最小值。为此，可以定义如下的Lagrange函数：

$$L(w,b,\alpha)=\frac{1}{2}w^T w-\sum_{i=1}^{n}\alpha_i[y_i(w^T x_i+b)-1]$$

其中，$a_i\geq 0$为Lagrange系数，我们的问题是对w和b求Lagrange函数的最小值。把上式分别对w、b、α_i求偏微分并令它们等于0，得：

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b}=0\Rightarrow \sum_{i=1}^{n}\alpha_i y_i=0$$

$$\frac{\partial L}{\partial \alpha_i}=0\Rightarrow \alpha_i[y_i(w^T x_i+b)-1]=0$$

以上三式加上原约束条件可以把原问题转化为如下凸二次规划的对偶问题：

$$\begin{cases}\max\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i^T x_j\right)\\ s.t\quad a_i\geq 0, i=1,\cdots,n\\ \sum_{i=1}^{n}a_i y_i=0\end{cases}$$

这是一个不等式约束下二次函数机制问题，存在唯一最优解。若α_i^*为最优解，则

$$w^*=\sum_{i=1}^{n}a_i^* y_i x_i$$

α_i^*不为零的样本即为支持向量，因此，最优分类面的权系数向量是支持向量的线性组合。

b^*可由约束条件$\alpha_i[y_i(w^T x_i+b)-1]=0$求解，由此求得的最优分类函数是 ：

$$f(x)=\mathrm{sgn}((w^*)^T x+b^*)=\mathrm{sgn}\sum_{i=1}^{n}a_i^* y_i x_i^* x+b^*)$$

sgn()为符号函数。

非线性可分情形：

当用一个超平面不能把两类点完全分开时（只有少数点被错分），可以引入松弛变量ξ_i（$\xi_i\geq$

0， $i=\overline{1,n}$），使超平面 $w^T x+b=0$ 满足：

$$y_i(w^T x_i+b)\geq 1-\xi_i$$

当0<ζ_i<1时样本点x_i仍旧被正确分类，而当ζ_i ≥1时样本点x_i被错分。为此，引入以下目标函数：

$$\psi(w,\xi)=\frac{1}{2}w^T w+C\sum_{i=1}^{n}\xi_i$$

其中C是一个正常数，称为惩罚因子，此时SVM可以通过二次规划（对偶规划）来实现：

$$\begin{cases}\max \sum_{i=1}^{n} a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i^T x_j\right)\\ s.t \quad 0\leq a_i\leq C, i=1,\cdots,n\\ \sum_{i=1}^{n} a_i y_i=0\end{cases}$$

7.1.2 SVM 参数选择

经分析对支持向量机有着重要影响的参数是:惩罚因子 C，核函数及其参数的选取。 惩罚因子 C 用于控制模型复杂度和逼近误差的折中，C 越大则对数据的拟合程度越高，学习机器的复杂度就越高，容易出现“过学习”的现象。而 C 取值过小，则对经验误差的惩罚小，学习机器的复杂度低，就会出现“欠学习”的现象。当 C 的取值大到一定程度时，SVM 模型的复杂度将超过空间复杂度的最大范围，那么当 C 继续增大时将几乎不在对 SVM 的性能产生影响。 SVM 的核函数包括线性核函数，RBF 核函数，多项式核函数，高斯核函数等，对于构建一个 SVM 模型来说首先需要做的就是选择核函数和核参数。根据 Vapnik 等人的研究表明，对于不同类型的核函数，SVM 模型所选择的支持向量的个数基本相同，但是其核函数参数和惩罚因子 C 的选择却对 SVM 模型的性能有着重要影响，如 RBF 核函数的参数 gamma 的取值就直接影响模型的分类精度，也就是说对于一个 RBF 核的 SVM 模型，要想提高其分类精度首先需要考虑的就是如何选择其核参数 gamma 和惩罚因子 C。

本文采用网格搜索的方法寻找最优参数。

网格搜索法

网格搜索法是目前比较常用的数据搜索方法之一，对于 RBF 核 SVM 模型来说，其算法流程如下所示：

首先，对于惩罚因子 C 和 RBF 核参数 gamma 分别确定其取值范围和搜索步长，从中得到 M 个 C 的值以及 N 个 gamma 的值。然后，根据 M 个 C 和 N 个 gamma 构建 M*N 组不同参数，使用每个参数构建 SVM 模型得到分类精度，以此确定最优的参数组 C 和 gamma。

最后，如何最佳分类精度依然没有达到要求，就可以根据分类精度曲线重新选择取值范围和搜索步长，进行细搜索直到满足要求为止。

网格搜索法的优点就是可以对多个参数同时进行搜索，参数之间相互联系相互制约的关系，使

其能够更好更快的得到最优解。而 M*N 组参数之间则是相互独立，这使得其可以进行并行搜索提高运算效率。其缺点就是当参数比较多时，如多项式核参数有 3 个，再加上惩罚因子 C，也就是说对于多项式核 SVM 模型来说，需要对 4 个参数同时进行选择，假设 4 个参数的取值个数分别为 M，N，P，Q，那么使用网格搜索法将计算 M*N*P*Q 组参数，计算量巨大。

交叉验证

由于本文手动设置的训练标签有限，训练样本数有限，为了能够更好的实现文本分类，本文采用 10 折交叉训练的方法提高分类正确率。

所谓交叉验证法就是指在训练 SVM 模型开始之前，对其训练数据进行一部分保留，然后利用这部分数据对训练后的模型进行评估。一般比较常用的是 K 折交叉验证法（K-foldcrossvalidation）。首先，将训练数据平均分成 K 组，然后取出其中一组进行保留，然后使用剩下的 K-1 组进行训练构建模型，最后用保留下的一组对训练出来的模型进行评估检测。将以上过程重复 K 次，保证每组数据都被保留测试过，然后根据 K 次评估检测得到的值来估计期望泛化误差，以此选择最优的参数。 交叉验证法是统计学习中的著名方法，被称为对泛化误差的无偏估计，它能够有效的防止过学习现象。它既具有一定的训练精度，又获得良好的泛化性能。

7.1.3 支持向量机分类

- 选择一部分好评与差评数据，训练处其相应的词向量，为降低训练难度，词向量的长度取为 1，训练部分结果如下

  满意 0.165834　　还 0.332710　　东西 0.130362

  好 0.008320　　不 -0.322917　　不错 0.248273

- 将评论数据匹配成向量 例：

  很 满意 东西 0.492280 0.165834 0.130362

  太 差 不 满意 -0.430087 0.018075 -0.322917 0.165834

- 给定训练数据标签 1 表示不满意 2 表示满意 例：

  很 满意 东西 0.492280 0.165834 0.130362　　2

  太 差 不 满意 -0.430087 0.018075 -0.322917 0.165834　　1

- 虽然训练样本和测试样本量较少，运用 5 折交叉训练，分类正确率为 95% 。说明支持向量机能很好的区分评论中的好评和差评数据，从而对评论中的信息进行挖掘。

7.2. 基于 K-均值聚类的文本分析

7.2.1 K-均值聚类原理

由于文本一开始并没有标签，而手动添加工作量太大，在有限的时间内，无法建立大量的训练样本，所以考虑对文本进行聚类。Kmeans 聚类算法算法解决的是将含有 n 个数据点(实体)的集合

$X=\{x_1,x_2,\cdots x_n\}$ 划分为 k 个类簇 Z_j 的问题， $j=1,2,\cdots,k$，算法首先随机选取 k 个数据点作为 k 个类簇的初始簇中心，集合中每个数据点被划分到与其距离最近的簇中心所在的类簇之中，形成了 k 个聚类的初始分布。对分配完的每一个类簇计算新的簇中心，然后继续进行数据分配过程，这样迭代若干次后，若簇中心不再发生变化，则说明数据对象全部分配到自己所在的类簇中，聚类准则函数收敛，否则继续进行迭代过程，直至收敛。这里的聚类准则函数一般采用聚类误差平方和准则函数。本算法的一个特点就是在每一次的迭代过程中都要对全体数据点的分配进行调整，然后重新计算簇中心，进入下一次的迭代过程，若在某一次迭代过程中，所有数据点的位置没有变化，相应的簇中心也没有变化，此时标志着聚类准则函数已经收敛，算法结束。

K-均值算法的步骤：

第一步：选 K 个初始聚类中心， $Z_1^1, Z_2^1, \wedge Z_k^1$，其中括号内的序号为寻找聚类中心的迭代运算的次序号。聚类中心的向量值可任意设定，例如可选开始的 *K* 个模式样本的向量值作为初始聚类中心。

第二步：逐个将需分类的模式样本*{x}*按最小距离准则分配给 K 个聚类中心中的某一个 $Z_j^{(1)}$。对所有的 *i*≠*j*，*j=1*，*2*，…，*K* ，如果 $Z_1^1, Z_2^1, \wedge Z_k^1$ 则 ，$X\in S_j^k$ 其中 *k* 为迭代运算的次序号，第一次迭代 *k=1*， S_j 表示第 *j* 个聚类，其聚类中心为 Z_j。

第三步：计算各个聚类中心的新的向量值 $Z_j^{(k+1)}$，*j=1*，*2*，…，*K*，求各聚类域中所包含样本的均值向量：

$$Z_j^{(k+1)} = \frac{1}{N_j} \sum_{X \in S_j^{(K)}} X$$

其中 N_j 为第 *j* 个聚类域 S_j 中所包含的样本个数。以均值向量作为新的聚类中心，可使如下聚类准则函数 *J* 最小：

$$J = \sum_{j=1}^{K} \sum_{X \in S_j^{(K)}} \left\| X - Z_j^{(k+1)} \right\|^2$$

在这一步中要分别计算 K 个聚类中的样本均值向量，所以称之为 K-均值算法。

第四步：若 $Z_j^{(k+1)} \neq Z_j^{(k+1)}$，*j=1*，*2*，…，*K*，则返回第二步，将模式样本逐个重新分类，重复迭代运算；若 $Z_j^{(k+1)} = Z_j^{(k+1)}$，*j=1*，*2*，…，*K*，则算法收敛，计算结束。

7.2.2 K-均值聚类效果

对本文的国美 3M 评论数据进行聚类，由于 k 均值聚类的 k 无法首先确定，因此我们从 k=2 开始尝试对数据进行聚类，K=5 时聚类效果相对较好，聚类部分结果如下图：

宝贝收到了，让朋友给装的，用机器过的水跟自来水不一样，感觉很好外观也很漂亮，值得购买 0
宝贝收到了，质量非常的满意。值得下次的购买. 0
宝贝收到了，质量非常好，和店家描述的一样，包装很仔细，过滤出来的水口感非常好，不错非常满意 0
宝贝收到了，质量很好，和卖家描述的一样，大品牌质量有保障，价格实惠是正品。赞一个 0
宝贝已经收到 0

安装清洗都方便，净水效果看得见。 1
安装上用了 1
安装师傅服务很好 1
安装师傅很给力的 1
安装时间算比较快的，我是当天送完货到家后就去预约 1
安装挺方便的，流出来的水质好了很多，没有氯水的味道，用着放心 1
安装也很顺利 1

东西很不错，看着水干净多了，味也没了 2
东西很不错，终于喝到放心水了 2
东西很好，安装的师傅说这个产品做工很好 2
东西很好， 2
东西很好用哦，能做到这样已经是很完美的了，很实用，卖家服务态度也热情，产品值得推荐！2
东西很好正经品牌和商店的一样，发货很快，包装很好，卖家的服务更好，买东西就得找这样的卖家，没有后不 2
东西很好噢，售后也很给力

包装的很严实 3
包装盒很不错 3
包装很好，快递也给力，卖家很耐心地给我解释各种问题~ 3
包装很仔细，安装师傅也不错 3
包装看上去很好。正品行货 3
宝贝包装的很好，很结实，卖家还送了很多东西挺不错的！质量也很好，家人都很满意！用着也很放心，不错！ 3
宝贝包装的很仔细呀，店家人很好呀。安装也很快，感觉净水效果很好，好评 3
宝贝不错，净化好的水口感甘甜，和买的矿泉水一样好喝，质量不错，包装也很好，以后不用担心喝水水质的问 3
宝贝不错，质量很好，净水效果很满意，可以直接饮用，很实用。卖家发货快，物流给力！ 3

宝贝的质量很好。 3

宝贝非常好，包装的很精细，很实用，质量真的很不错，价格很划算，一定给赞 3

物流很给力 4

物流很快，东西很不错的，水喝起来和纯净水没什么区别。 4

物流很慢，三天才到。 4

物流速度很给力 4

物流速度很快，安装的也很快，态度很好，很满意 4

物流也给力 4

物流也很快~~~ 4

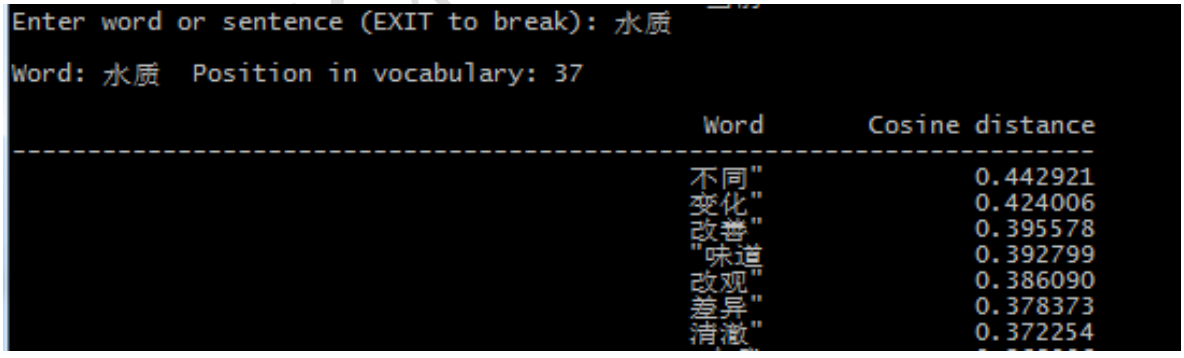
物有所值 4

8. 结果分析

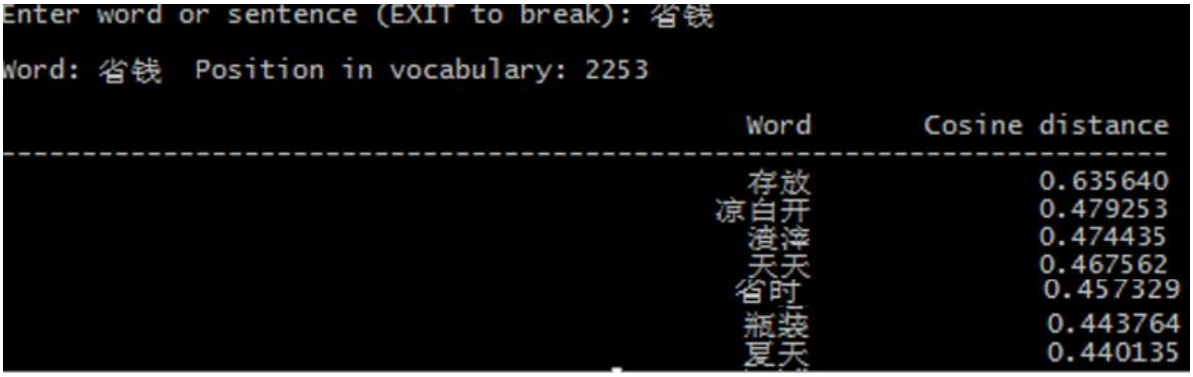
8.1. 用户为何购买净水器产品？

由经验可知及上述分析我们粗略可知，用户为使其饮用水的质量得到保障，所以需要通过净水器来实现其需求。一般分析用户为何购买某种产品通常从产品本身的作用，产品带来的经济效益，产品的作用对用户产生的影响三个方面来分析。因此我们设定的三个指标为净水作用——水质；经济效益——省钱；对用户产生的影响——健康。为此我们通过计算指标和相关的词向量的距离来量化分析究竟用户为何要购买净水器产品。

水质：



省钱：



健康：

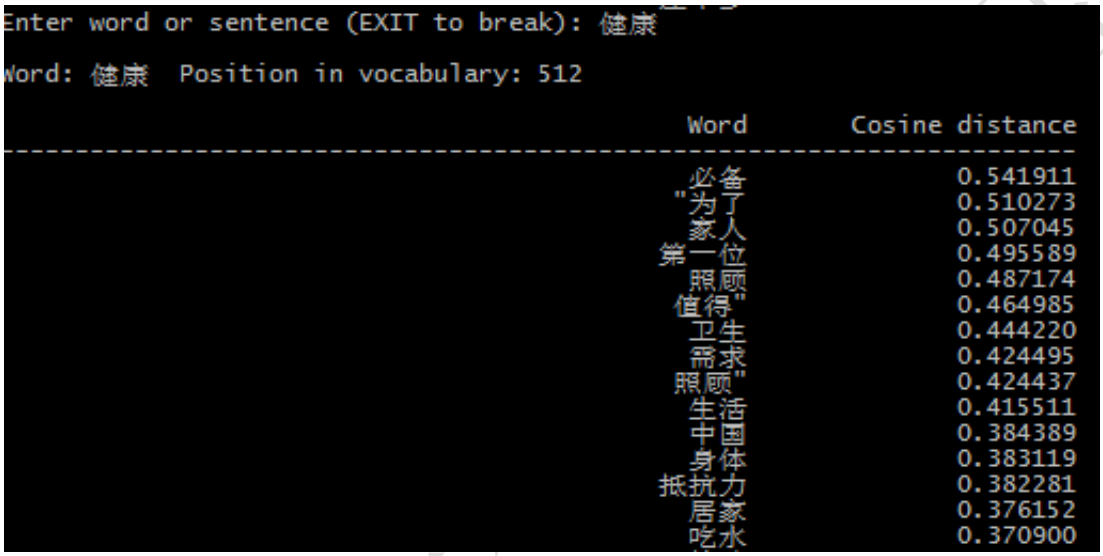


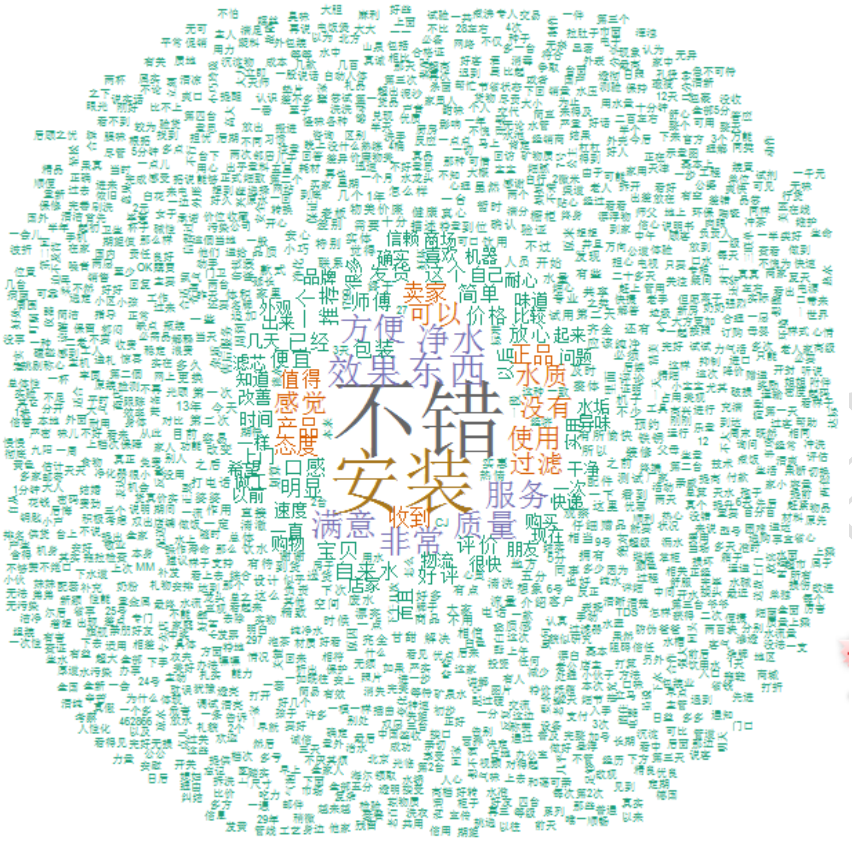
图 10 用户购买净水器的原因

根据上图我们可以看到因为净水器能够改善用户的水质，保障用户的身体健康，特别是中国这个水污染严重的大环境下，购买净水器保证用水卫生，提高孩子的抵抗力，守护家人的健康成了大趋势。同时，由于购买净水器而不需要再购买瓶装水为用户节省了时间和金钱。以上就是大部分用户购买净水器的理由。

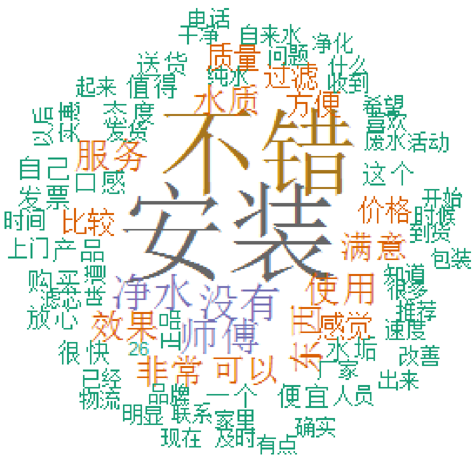
8.2. 用户对净水器产品的个性化需求分析

由经验可知，不同购物平台的顾客的消费习惯可能不一样，为了更好的进行个性化分析，我们将分不同的运营商进行产品个性化分析。由于天猫和淘宝，京东和苏宁在购物平台上具有相似性，为节省计算时间，本文主要分布分析国美，淘宝，苏宁，易迅四个平台上的用户个性化需求。

首先我们来看四个平台数据词频统计的结果：


不错
安装
效果
东西
净水
方便
满意
非常
质量
服务
过滤
使用
没有
水质
可以
卖家
正品
收到
值得
感觉
态度

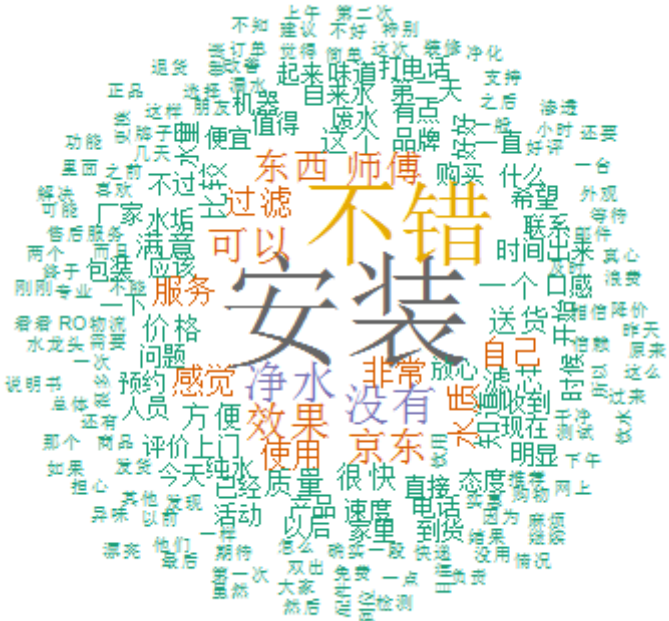
国美


不错
安装
净水
没有
师傅
效果
服务
使用
满意
价格
水质
质量
过滤
方便
比较
非常
可以
东西
感觉

苏宁


不错
安装
效果
净水
满意
质量
非常
卖家
服务
收到
宝贝
好评
包装
发货
方便
很快
东西
态度

淘宝



易迅

图 11 个性化需求图

通过这四张图我们发现，几乎所有的购物平台，用户提到最多的就是净水器的安装问题，这说明安装的方便性是所有用户都关注的焦点。同时我们也发现四张图中服务，净水质量，以及产品的方便性都是优先关注点。同时我们也发现四大购物平台上也存在着差异性，例如国美网站上的用户就关心产品是否是正品，因此电商平台因着重强调产品的正规性；而苏宁上的用户则更为关注净水后的水质，因此可通过视频演示等方式使消费者确信水质有保障；而淘宝用户对卖家的发货及物流速度有比较高的期望，因此电商平台应关注其物流质量。

8.3. 各品牌产品优劣势及卖点分析

本文首先取 3M 和沁园两个数据量较大的品牌探究分析，取每个数据的前 50000 条数据，剔除其中用户评价不明确的评论，对数据进行统计学分析得下表。

表 4 3M 和沁园部分评论数据统计表

3M	满意	很好	不错	还可以	一般	麻烦	不满意
数目	5330	3876	22840	1071	781	38	3
占有效评论的比例	15.70%	11.42%	67.30%	3.16%	2.30%	0.11%	0.01%

沁园	满意	很好	不错	还可以	一般	麻烦	不满意
数目	8587	18837	19012	480	3	0	2
占有效评论的比例	18.30%	40.15%	40.52%	1.02%	0.01%	0.00%	0.00%

由于很好，不错，这些表达本身就是抽象概念，无法表现出产品本身的特性，因此通过此方式实现产品的优劣分析不太合适，但是统计的过程中我们发现，要比较产品的优劣性，从用户的关注度出发，定为一下几个方面：1.用户总评，即用户对产品的总体体验和感觉

2.产品的外形

3.产品的安装

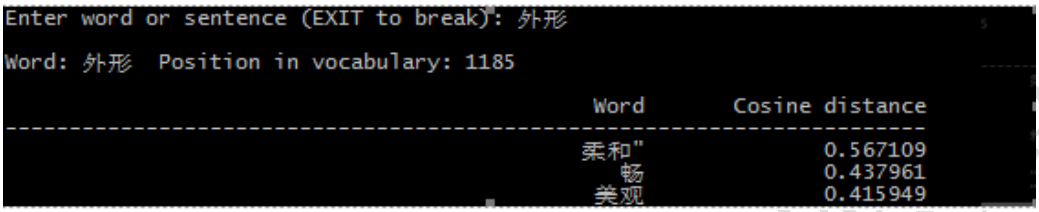
4.产品客服及快递的服务

5.产品的价格

6.对水质的改善

7.发货及快递的速度。

本文通过计算以上关键词在各产品中的相关性得下图：



3M 产品外形计算结果

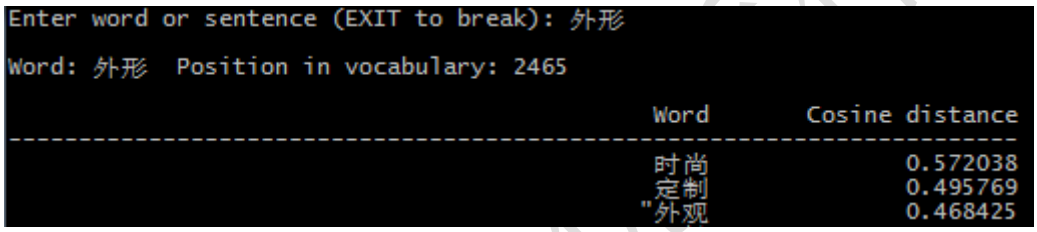


图 12 沁园外形计算结果

由上图我们可知沁园产品的外形上比 3M 更具优势，更时尚性，且其产品还可以进行量身定制，这对于重视产品外形美观的用户来说无疑是巨大的吸引力。

运用上述方法得 3M，沁园，特浩恩，道尔顿四个品牌的产品优劣比较如下表：

表 5 各品牌产品优劣势分析

品牌	总评	外形	安装	客服及快递服务	价格	水质的改善	发货及快递速度	缺点
3M	快捷 经济 品质	柔和 流畅 美观	上门 规范	热情 温馨 有责任心	实惠 有折扣活动	改善 立竿见影	快	
沁园	习惯性好评	时尚 定制	需要预定	有专员	价格一般	明显改变	很快，送货上门	废水有点多
特浩恩	品质稳定	尺寸刚好	需要发票	快递服务差	便宜 公道	改善	发货速度	
道尔顿	必须好评	柔和 尺寸刚好	热心 熟练	周到 热心 诚实	实惠 便宜	有变化	快递尽职尽责	

由上表可以发现 3M 产品在用户关注的各方面都有较好的表现，用户较多，品牌效应较强，且能提供规范的服务；沁园在产品外观上具有明显优势且有专门的服务专员，但产品安装上需要用户提取预定，同时净水过程中产生的废水较多。特浩恩在四个品牌中不具备较强的竞争力，特别是快递服务有待加强。道尔顿价格公道，服务热忱，回头客较多，但受众面小，需要加强宣传力度。

9. 结论

- 用户购买净水器的理由：净水器能够改善用户的水质，保障用户的身体健康。同时，由于购买净水器而不需要再购买瓶装水为用户节省了时间和金钱。
- 不同电商平台的客户需求存在差异性，需要区别对待。国美电商平台因着重强调产品的正规性；而苏宁上可通过视频演示等方式使消费者确信水质有保障；淘宝电商平台需重点关注其物流质量
- 不同品牌的定位不同，客户群不同，关注点也不同。如沁园因突出其产品外形优势，弱化产品安装的劣势。特浩恩的物流亟需提高。

10. 参考文献

[1]周练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济,2015,02:145-148.

[2]罗杰,王庆林,李原. 基于 word2vec 与语义相似度的领域词语聚类[A]. 中国自动化学会控制理论专业委员会、中国系统工程学会.第三十三届中国控制会议论文集（A 卷）[C].中国自动化学会控制理论专业委员会、中国系统工程学会:,2014:5.

[3]刘志明,刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与用,2012,01:1-4.

[4]郑文超,徐鹏. 利用 word2vec 对中文词进行聚类的研究[J]. 软件,2013,12:160-162.

[5]周钦强,孙炳达,王义. 文本自动分类系统文本预处理方法的研究[J]. 计算机应用研究,2005,02:85-86.

[6]顾益军,樊孝忠,王建华,汪涛,黄维金. 中文停用词表的自动选取[J]. 北京理工大学学报,2005,04:337-340.

[7]唐晓丽,白宇,张桂平,蔡东风. 一种面向聚类的文本建模方法[J]. 山西大学学报(自然科学版),2014,04:595-600.

[8]曹卫峰. 中文分词关键技术研究[D].南京理工大学,2009.

[9]朱鹤祥. Web 日志挖掘中数据预处理算法的研究[D].大连交通大学,2010.

[10]郭茂. 基于类中心向量的文本分类模型研究与实现[D].大连理工大学,2010.

[11]刘秀芹. 基于 Web 挖掘的 B2C 网站需求分析与应用研究[D].天津大学,2009.

[12]苏先宇. 基于潜在语义索引的文本分类研究与实现[D].哈尔滨工业大学,2008.

[13]马扬. 基于语义的中文文本自动分类系统的研究与实现[D].重庆大学,2009.

[14]陶敏. 基于支持向量机的中文客户评论情感文本分类研究[D].武汉纺织大学,2011.

[15]俞鸿魁. 基于层次隐马尔可夫模型的汉语词法分析和命名实体识别技术[D].北京化工大学,2004.

[16]王萌. 基于概念向量空间模型的中文自动文摘研究[D].华中师范大学,2005.

[17]炎士涛. 基于词频统计的文本分类模型研究[D].上海师范大学,2007.

[18]李丽. 基于本体的网页文本分类的研究[D].北京交通大学,2008.

[19]张海燕. 基于分词的中文文本自动分类研究与实现[D].湖南大学,2002.

[20]贾玉生. 基于 Hadoop 的分布式文本分类研究[D].北京工业大学,2013.

[21]孔照昆. 中文文本层次分类方法研究及应用[D].扬州大学,2013.

[22]杨鹏. 面向领域自然语言的文本自动分类及其在产品设计中的应用[D].西安电子科技大学,2007.

[23]于瑞萍. 中文文本分类相关算法的研究与实现[D].西北大学,2007.

[24]王倩. 中文文本分类技术的研究[D].北京化工大学,2007.

[25]刘伍颖. 面向垃圾信息过滤的主动多域学习文本分类方法研究[D].国防科学技术大学,2011.

[26]张京楣. 基于统计方法的文本风格分析研究[D].山东大学,2012.

[27]朱朝勇. 基于本体的知识库分类研究[D].中国科学技术大学,2013.

[28]谢晋. 基于词跨度的中文文本关键词提取及在文本分类中的应用[D].浙江工业大学,2011.

[29]李书明. 数字化学习中知识组织模型及应用研究[D].华中师范大学,2011.

[30]王霜霜. 中文农业网页多分类方法研究[D].新疆农业大学,2012.

[31]蒋伟. 基于 Hadoop 的电商商品文本分类研究与实现[D].武汉理工大学,2014.

[32]赵俊杰. 基于文本挖掘技术的论文抄袭判定研究[D].合肥工业大学,2009.

[33]杨武,宋静静,唐继强. 中文微博情感分析中主客观句分类方法[J]. 重庆理工大学学报(自然科学),2013,01:51-56.