

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：特等奖

作品单位：华南师范大学

作品成员：周 涛 吴家舜 邵悦涵

指导教师：杨 坦

基于情感分析、语义网络和主题模型的评论文本分析

摘 要：随着网上购物在中国越来越流行，人们对于网上购物的需求变得越来越高，这给京东、淘宝等电商平台得到了很大的发展机遇，但是与此同时，这种需求也推动了更多的店商平台的崛起，引发了激烈的竞争。而在这种电商平台激烈竞争的大背景下，除了提高商品质量、压低商品价格外，了解更多消费者的心声对于店商平台来说也变得越来越有必要，其中非常重要的方式就是对消费者的文本评论数据进行内在信息的数据挖掘分析。而得到的这些信息，也会有利于对应商品的生产厂家自身竞争力的提升。本文将基于数据挖掘技术对京东三种品牌型号的热水器的评论数据进行内在信息的挖掘与分析。

在本次数据挖掘过程中，我们首先对获取到的评论数据利用 python 以及 ICTCLAS 工具进行数据预处理、分词以及停用词过滤操作，实现了对评论数据的优化，并提升了其可建模度。

接着，采用多种方法来进行数据挖掘模型的构建，为后面的评论分析构建分析的基础。为此我们先利用深度学习的方法，通过多种工具构建栈式自编码神经网络；其次，运用武汉大学的 ROSTCM6 系统为三种品牌型号热水器的好差评文本构建语义网络；再有，利用 LDA 主题模型的思想，结合统计学的角度实现评论主题模型的构建。

最后，运用构造出来的多种数据挖掘模型的结果，对这些评论数据进行多方面多角度的评论文本分析，以提取评论中隐藏的信息。栈式自编码神经网络被用以进行情感倾向性分析；语义网络重建了有价值高频词之间的关系，在共词矩阵以及评论定向筛选回查的帮助下，一定程度上得到了京东三种品牌型号热水器包括特有优点、抱怨点等信息；LDA 主题模型则滤取出了从统计学角度上的给予不同型号热水器好差评的消费者的关注点，以了解热水器消费者一般关注的对象。

关键词：评论数据；文本分析；信息提取；语义网络；LDA；栈式自编码

comments analysis based on sentiment analysis, semantic network and Latent Dirichlet Allocation

Abstract: With the prevalence of online shopping in China, consumer has paid more and more attention on online shopping, which at the same time, brings opportunities and challenges to E-business such as Jingdong and Taobao. With the background of challenges, studying what people virtually think based on data analyzing and mining plays an important role in improving the quality of the products and service. What's more, the study will strengthen the competitiveness of E-business. Therefore, in this thesis, some research are done on the products comments of three different brands based on data mining.

Firstly, in order to optimize the comment data and enhance the ability of our model, we pre-process the comment data and stoplists filtering by using python, and utilizing ICTCLAS to do word segmentations.

Secondly, in order to analyze the information of the comments, we choose various methods to establish the data mining model. First, deep learning is applied on the construction of Stacked AutoEncoder (SAE) neural network. Then, we utilize the ROSTCM6 system(building by Wuhan university) to build up semantic network on account of favorable and unfavorable comments. Finally, combined with statistical perspectives, we establish the Latent Dirichlet Allocation (LDA) model to study the information of favorable and unfavorable comments.

Finally, the above models are applied comprehensively to analyze the comments from different perspectives, which can discover the latent information in comments. And we do emotional tendency analysis through Stacked AutoEncoder (SAE) neural network model. furthermore, semantic network rebuilds the relationship between the valuable high-frequency words. With the help of co-word matrix and the comment directional filter checkback, we attain the strengths and weaknesses of three different brands-Midea, Haier, Wanhe. Additionally, combined with statistical perspectives, we apply LDA model to study the core concerns and their attitude from consumers on three different brands.

Key words: comment data, text analysis, Stacked AutoEncoder (SAE), semantic network, Latent Dirichlet Allocation (LDA), information extraction

目 录

1. 挖掘目标.....	1
2. 分析方法与过程.....	1
2.1. 总体流程	1
2.2. 具体步骤	1
2.3. 结果分析	18
3. 结论.....	27
4. 参考文献.....	27

1. 挖掘目标

本次建模针对京东电商平台海尔、美的、万和三种品牌型号的热热水器的消费者的文本评论数据，在对文本进行基本的机器预处理、中文分词、停用词过滤后，通过建立包括栈式自编码深度学习、语义网络与 LDA 主题模型等多种数据挖掘模型，实现对文本评论数据的倾向性判断以及所隐藏的信息的挖掘并分析，以期得到有价值的内在内容。

2. 分析方法与过程

2.1. 总体流程

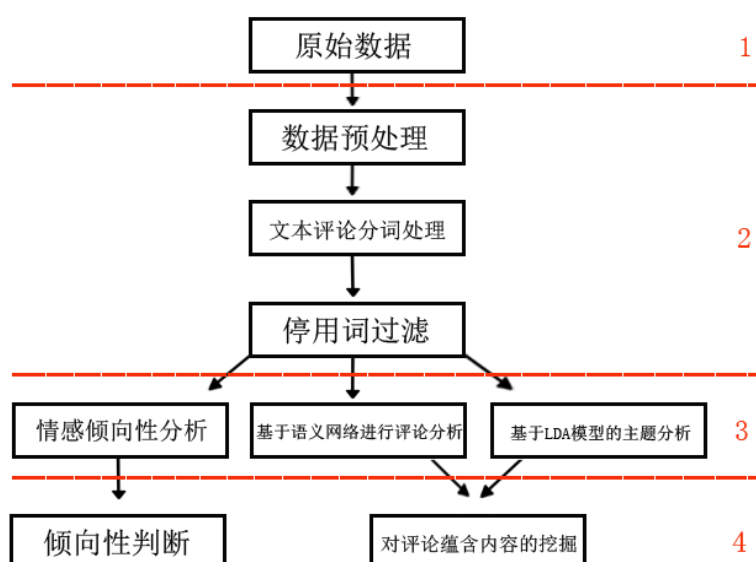


图 1 总体流程图

本论文的分析流程可大致分为以下四步：

第一步：获取分析用的原始数据（文本评论语料），部分数据自行爬取；

第二步：对获取的数据进行基本的处理操作，包括数据预处理、中文分词、停用词过滤等操作；

第三步：文本评论数据经过处理后，运用多种手段对评论数据进行多方面的分析；

第四步：从对应结果的分析中获取文本评论数据中有价值的内容。

2.2. 具体步骤

2.2.1 数据介绍

本文使用的实验数据为从京东得到的海尔 ES50H-Q1（ZE）热水器（50 升）、美的 F50-15A1 热水器（50 升）以及万和 DSCF50-T4A 热水器（50 升）的文本评论数据（前两者来自所给数据，可通过

筛选的方式得到，而最后一种则来自我们的自行爬取），即选取三个品牌的热水器，且每个品牌选取一个型号来研究。从总体上来说，京东作为国内最大型的电商平台之一，在该网站购买的顾客比较多，因此相关产品的评论也就会比较多，而且评论质量也较一些其它网站上的要好。

2.2.2 文本评论预处理

取到文本后，我们首先要进行文本评论数据的预处理。文本评论数据里面存在大量价值含量很低甚至没有价值含量的条目，如果将这些评论数据也引入进行分词、词频统计乃至情感分析等，则必然会对分析造成很大的影响，得到的结果的质量也必然是存在问题的。那么在利用到这些文本评论数据之前就必须要先进行文本预处理，把大量的这些无价值含量的评论去除。

我们运用 Python2.7 对这些文本评论数据的预处理主要由三个部分组成：文本去重、机械语料压缩以及短句删除。按照各自处理的特性，我们依照这个顺序进行文本评论数据的预处理。

2.2.2.1 文本去重

(1) 文本去重的基本解释及原因

文本去重，顾名思义，就是去除文本评论数据中重复的部分。无论获取到什么样的文本评论数据，首先要进行的预处理应当都是文本去重。文本去重的主要原因如下：

①一些电商平台往往为了避免一些客户长时间不发表评论，会设置一道程序，如果用户超过规定的时间仍然没有做出评论，系统会自动替客户做出评论，当然这种评论的结果大多都会是好评，比如国美。但是这类数据显然没有任何分析价值，而且这种评论是大量重复出现的，必须去除。

②同一个人可能会出现重复的评论，因为同一个人可能会购买多种热水器，然后在进行评论过程中可能为了省事，就在多个热水器中采用同样或相近的评论，这里当然可能不乏有价值的评论，但是即使有价值也只有第一条有作用。

③由语言的特点我们知道，在大多数情况下，不同人之间的有价值的评论都不会出现完全重复，如果出现了不同人评论之间的完全重复，这些评论一般都是毫无意义的，诸如“好好好好好”、“XX 牌热水器 XX 升”等等或者说就是直接复制粘贴上一人的评论，这种评论显然就只有最早评论出的才有意义（即只有第一条有作用）。而如果不是完全重复，而比较相近的，也存在一些无意义的评论。

(2) 常见文本去重算法概述及缺陷

在前人的研究下，有许多的文本去重算法，大多都是先通过计算文本之间的相似度，再以此为基础进行去重，包括编辑距离去重，Simhash 算法去重等等，但是大多存在一些缺陷。以编辑距离算法去重为例，编辑距离算法去重实际上就是先计算两条语料的编辑距离，然后进行阈值判断，如果编辑距离小于某个阈值则进行去除重复处理，这种方法针对类如：

“XX 牌热水器 XX 升 大品牌 高质量”

以及

“XX 牌热水器 XX 升 大品牌 高质量 用起来真的不错”

的接近重复而又无任何意义的评论文本的去除的效果是很好的，主要为了去除接近重复或完全重复

的评论数据，而并不要求完全重复，但是当这种方法测到都有意义，但是有相近的表达的时候就可能也会采取删除操作，这样就会造成错删问题，比如如下的例子：

“还没正式使用，不知道怎样，但安装的材料费确实有点高，380”

以及

“还没使用，不知道质量如何，但安装的材料费确实贵，380”。

这组语句的编辑距离只是比上一组大 2 而已，但是很明显这两句就是都有意义的，如果阈值设为 10（该组为 9），就会带来错删问题。可惜的是，这一类的评论数据组还是不少的，特别是差评的语料，许多顾客不会用太多的言语表达，直至中心，问题就来了。

（3）文本去重选用的方法及原因

既然这一类相对复杂的文本去重的算法容易去除有用的数据，那么我们就需要考虑一些相对简单的文本去重思路。由于相近的语料存在不少是有用的评论，去除这类语料显然不合适，那么为了存留更多的有用语料，我们就只能针对完全重复的语料下手。那么处理完全重复的语料直接采用最简单的比较删除法就好了，也就是两两对比，完全相同就去除的方法。

从上述的总结我们知道存在文本重复问题的条目归结到底只有 1 条语料甚至 0 条语料是有用的，但是透过观察评论我们知道存在重复但是起码有 1 条评论有用的语料，即 1.1 中情况②的语料很多，而我们运用比较删除法显然只能定为留 1 条或者是全去除，因此我们只能设为留 1 条，以确保尽可能存留有用的文本评论信息。

观察比较删除法实现后的结果，我们发现总体效果还是很不错的。

2.2.2.2 机械压缩去词

（1）机械压缩去词的思想

由于电商品台的文本评论数据质量参差不齐，没有意义的文本数据很多，因此透过文本去重就已经可以删除掉非常多的没有意义的评论文本。但是文本去重远远不够，经过文本去重后的评论仍然有很多评论需要处理掉，比如：

“非常好非常好非常好非常好非常好非常好非常好”

以及

“好呀好呀好呀好呀好呀好呀好呀好呀好呀好呀”。

这一类语料是存在连续重复的语料，也是最常见的较长的无意义语料。因为大多数给出无意义评论的人都只是为了获得一些额外奖励等，并不对评论真正抱有兴趣，而他们为了省事就很可能进行这样的评论。显然这一类语料并不显得就会重复，但是也是毫无意义的评论，是需要删除的。

可惜的是，计算机不可能自动识别出所有这种类型的语料，比如“非常好”可以有从 1 到无上限的有穷个的叠加，即使运用词典透过某些方式识别了这一类的文本评论数据，比如算出“非常好”比较多意味着可能是无意义评论，一位制造无意义评论的顾客还可以以任何一个词进行重复，还可以重复某词，但次数不一定多，而这种显然只需要保留第一个即可，若不处理，可能会影响情感倾向的判断，比如：

“15 分钟就出热水了，感觉还不错，但是安装费实在是太贵太贵太贵太贵”

与

“15 分钟就出热水了，感觉还不错，但是安装费实在是太贵太贵太贵”

是没有差别的，但是若不处理，就会出现差别。

因此，我们就需要对语料进行机械压缩去词处理，也就是说要去掉一些连续重复累赘的表达，比如把：

“哈哈哈哈哈”

缩成

“哈”

不过这样仍然会保留无意义的评论（比如上述的评论），但是这些评论在经过这步处理后，在最后一个预处理环节：短句删除环节就会被去除掉。当然，机械压缩去词法不能像分词那样去识别词语。

（2）机械压缩去词处理的语料结构

机械压缩去词实际上要处理的语料就是语料中有连续累赘重复的部分，从一般的评论偏好角度来讲，一般人制造无意义的连续重复只会在开头或者结尾进行，比如：

“为什么为什么为什么安装费这么贵，毫无道理！”

以及

“真的好好好好好好”

等等，而中间的连续重复虽然也有，但是非常少见（中间重复在输入上显得麻烦，无意义评论本就是为了随意了事），而且中间容易有成语的问题，比如

“安装师傅滔滔不绝的向我阐述这款热水器有多好”

这种语料显然在去掉一个“滔”字后肯定就会出现重复，因此我们只对开头以及结尾的连续重复进行机械压缩去词的处理。

（3）机械压缩去词处理过程的连续累赘重复的判断及压缩规则的阐述

连续累赘重复的判断可通过建立两个存放国际字符的列表来完成，先放第一个列表，再放第二个列表，一个个读取国际字符，并按照不同情况，将其放入带第一或第二个列表或触发压缩判断，若得出重复（及列表 1 与列表 2 有意义的部分完全一对一相同）则压缩去除，这样当然就要有相关的放置判断及压缩规则。在机械压缩去词处理的连续累赘重复的判断及压缩规则设定的时候，必然要考虑到词法结构的问题，综合文字表达特点，我们设定如下 7 条规则（说明：1、这里为了初始化列表而放入的空格不算输入了国际字符；2、由于批量的评论里头可能会存在某些评论无法识别，因此在进行这一步时我们需要结合运行进程人工删除一些无法识别语句）：

规则 1：如果读入的这个字符与第一个列表的第一个字符相同，而第二个列表没有任何放入的国际字符，则将这个字符放入第二个列表中。

解释：因为一般情况下同一个字再次出现时大多数都是意味着上一个词或是一个语段的结束以及下一个词或下一个语段的开始，举例如下：

真的很快加热完毕，真的马上就能用。




图2 机械压缩去词规则1的示例图

规则2：如果读入的这个字符与第一个列表的第一个字符相同，而第二个列表也有国际字符，则触发压缩判断，若得出重复，则进行压缩去除，清空第二个列表。

解释：判断连续重复最直接的方法，举例如下：

重复！

为什么为什么为什么安装费这么贵，毫无道理！

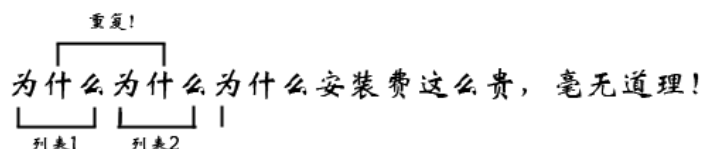


图3 机械压缩去词规则2的示例图

规则3：如果读入的这个字符与第一个列表的第一个字符相同，而第二个列表也有国际字符，则触发压缩判断，若得出不重复，则清空两个列表，把读入的这个字符放入第一个列表第一个位置。

解释：即判断得出两个词是不相同的，都应保留，举例如下：

不重复！

真的很好！真的很便宜！真的加热很快！

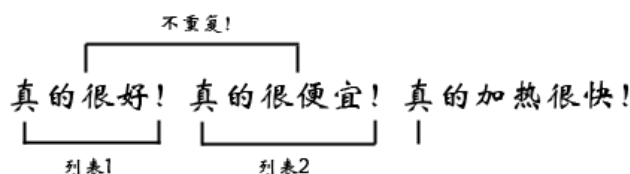


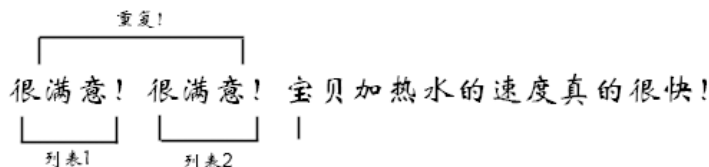
图4 机械压缩去词规则3的示例图

规则4：如果读入的这个字符与第一个列表的第一个字符不相同，触发压缩判断，如果得出重复，且列表所含国际字符数目大于等于2，则进行压缩去除，清空两个列表，把读入的这个字符放入第一个列表第一个位置。

解释：用以去除下图情况的重复，并避免类如“滔滔不绝”这种情况的‘滔’被删除，并可顺带压缩去除另一类连续重复，亦见下图示例：

重复！

很满意！很满意！宝贝加热水的速度真的很快！



顺带可以处理的语料：

重复！ 重复！

真的真的很好很好用！

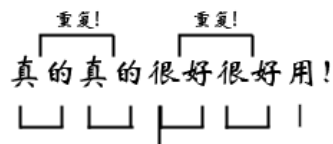


图5 机械压缩去词规则4的示例图

规则5：如果读入的这个字符与第一个列表的第一个字符不相同，触发压缩判断，若得出不重

复，且第二个列表没有放入国际字符，则继续在第一个列表放入国际字符。

解释：没出现重复字就不会有连续重复语料，第二个列表未启用则继续填入第一个列表，直至出现重复情况为止。

规则 6：如果读入的这个字符与第一个列表的第一个字符不相同，触发压缩判断，若得出不重复，且第二个列表已放入国际字符，则继续在第二个列表放入国际字符。

解释：类似规则 5，此处省略叙述。

规则 7：读完所有国际字符后，触发压缩判断，对第一个列表以及第二个列表有意义部分进行比较，若得出重复，则进行压缩去除。

解释：由于按照上述规则，在读完所有国际字符后不会再触发压缩判断条件，故为了避免下图实例连续重复情况，补充这一规则。

很好很好
□ □
列表1 列表2

图 6 机械压缩去词规则 7 的示例图

(4) 机械压缩去词处理操作流程

根据上述规则，便可以完成对开头连续重复的处理。类似的规则，亦可以对处理过的文本再进行一次结尾连续重复的机械压缩去词，算法思想是相近的，只是从尾部开始读词罢了。从结尾开始的处理结束后就得到了已压缩去词完成的精简语料。

2.2.2.3 短句删除

(1) 短句删除的原因及思想

完成机械压缩去词处理后，我们进行最后的预处理步骤：短句删除。我们知道，虽然精简的辞藻在很多时候是一种比较好的习惯，但是由语言的特点我们知道，从根本上说，字数越少所能够表达的意思是越少的，要想表达一些相关的意思就一定要有相应量的字数，过少的字数的评论必然是没有任何意义的评论，比如三个字，就只能表达诸如“很不错”、“质量差”等等。为此，我们就要删除掉过短的评论文本数据，以去除掉没有意义的评论，包括：

①原本就过短的评论文本，如“很不错”。

②经机械压缩去词处理后过短的评论文本，即原本为存在连续重复的且无意义的长文本，如“好好好好好好好好好好好好”。

(2) 保留的评论的字数下限的确定

显然，短句删除最重要的环节就是保留的评论的字数下限的确定，这个没有精确的标准，可以结合特定语料来确定，一般 6 到 10 个国际字符都是较为合理的下限，在此处我们设定下限为 7 个国际字符，即经过前两步预处理后得到的语料若小于等于 6 个国际字符，则将该语料删去。

2.2.3 文本评论分词

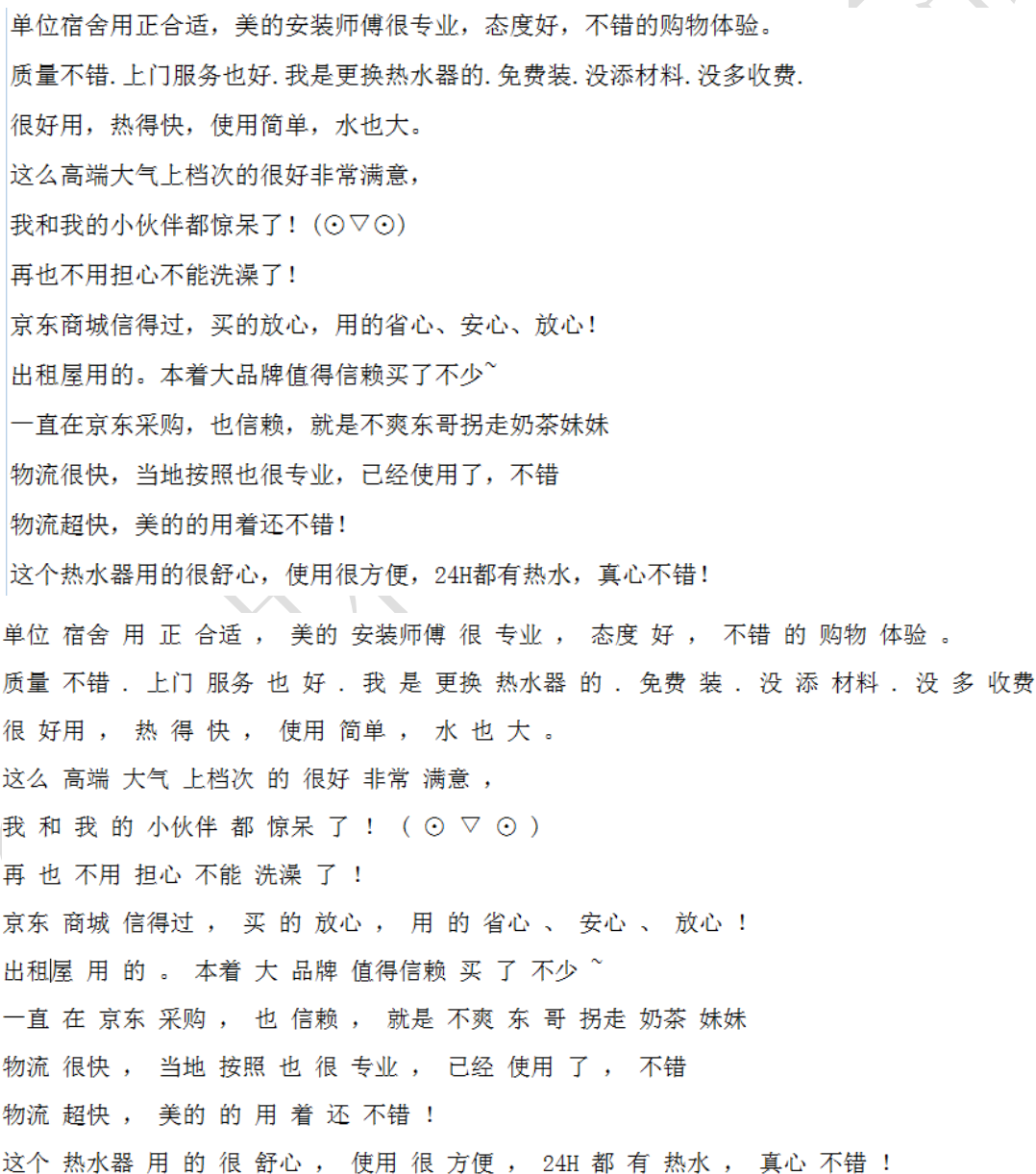
在中文中，只有字、句和段落能够通过明显的分界符进行简单的划界，而对于“词”和“词组”

来说，它们的边界模糊，没有一个形式上的分界符。因此，进行中文文本挖掘时，首先应对文本分词，即将连续的字序列按照一定的规范重新组合成词序列的过程。

分词结果的准确性对后续文本挖掘算法有着不可忽视的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择。

本文采用由中国科学院计算机技术研究所研发的基于多层隐马尔科夫模型的分词系统 ICTCLAS，对 TXT 文档中的商品评论数据进行中文分词。ICTCLAS 系统提供分词、词性标注、未登录词识别，支持用户词典等功能。经过相关测试，此系统的分词精度高达 98.45%。为进一步进行词频统计，分词过程将词性标注作用去掉。

本文使用 ICTCLAS 提供的 API 接口进行 C++ 编码，实现了中文文本分词。部分结果示例如下：



单位宿舍用正合适，美的安装师傅很专业，态度好，不错的购物体验。

质量不错。上门服务也好。我是更换热水器的。免费装。没添材料。没多收费。

很好用，热得快，使用简单，水也大。

这么高端大气上档次的很好非常满意，

我和我的小伙伴都惊呆了！（☺▽☺）

再也不用担心不能洗澡了！

京东商城信得过，买的放心，用的省心、安心、放心！

出租屋用的。本着大品牌值得信赖买了不少~

一直在京东采购，也信赖，就是不爽东哥拐走奶茶妹妹

物流很快，当地按照也很专业，已经使用了，不错

物流超快，美的的用着还不错！

这个热水器用的很舒心，使用很方便，24h都有热水，真心不错！

单位 宿舍 用 正 合适 ， 美的 安装师傅 很 专业 ， 态度 好 ， 不错 的 购物 体验 。

质量 不错 。 上 门 服 务 也 好 。 我 是 更 换 热 水 器 的 。 免 费 装 。 没 添 材 料 。 没 多 收 费

很 好 用 ， 热 得 快 ， 使 用 简 单 ， 水 也 大 。

这 么 高 端 大 气 上 档 次 的 很 好 非 常 满 意 ，

我 和 我 的 小 伙 伴 都 惊 呆 了 ！（ ☺ ▽ ☺ ）

再 也 不 用 担 心 不 能 洗 澡 了 ！

京 东 商 城 信 得 过 ， 买 的 放 心 ， 用 的 省 心 、 安 心 、 放 心 ！

出 租 屋 用 的 。 本 着 大 品 牌 值 得 信 赖 买 了 不 少 ~

一 直 在 京 东 采 购 ， 也 信 赖 ， 就 是 不 爽 东 哥 拐 走 奶 茶 妹 妹

物 流 很 快 ， 当 地 按 照 也 很 专 业 ， 已 经 使 用 了 ， 不 错

物 流 超 快 ， 美 的 的 用 着 还 不 错 ！

这 个 热 水 器 用 的 很 舒 心 ， 使 用 很 方 便 ， 24h 都 有 热 水 ， 真 心 不 错 ！

图 7、8 部分商品评论数据及其分词效果

2.2.4 停用词过滤

经过中文分词这一步骤，将初始的文本处理成为词的集合，即 $d = (\omega_1; \omega_2; \dots; \omega_N)$ ，其中 N 为文本 d 中出现词语的个数。但是文本中含有对文本含义表达无意义的词语，应进行删除，以消除它们对文本挖掘工作的不良影响，此类词称为停用词。停用词的两个特征为：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。例如中文中的“的”、“了”、“地”、“啊”等，英文中的“is”、“are”、“the”、“that”等，在特征选取的过程中，停用词的介入可能会造成选出的特征几乎都是停用词，从而影响结果的分析。但是在停用词的去除中，应注意要保留其中的否定词，可以对停用词表进行人工筛选相结合的方式，对停用词进行处理。

文本采用基于停用词表的文本停用词过滤方式，将分词结果于停用词表中的词语进行匹配，若匹配成功，则进行删除处理。结果示例如下：

物流 超快 ， 美的 的 用 着 还 不错 ！

物流 超快 美的 用 还 不错 ！

图 9 停用词过滤效果前后对比

2.2.5 情感倾向性分析

为了得到一个商品的总体情感倾向，我们可以对该商品的评论集做情感倾向分析，以得到对商品的总体印象。传统的情感分析是基于情感词典的方法，对每条评论中的情感词做加权，来得到每条评论的情感倾向值，进而获得整个评论集的情感倾向。这种方法直观，容易理解；此外，运用机器学习的方法进行情感二极分类也是当前的流行方法：对每条评论抽取特征（tf，tf-idf 值等）构成特征向量，然后采用朴素贝叶斯法或者 SVM 进行分类。如今，这种方法的运用也是比较成熟。

本文抛弃这些传统的方法，大胆尝试采用新的方法：基于词向量和深度学习方法对评论集做情感倾向性分析。

2.2.5.1 训练生成词向量

我们首先训练以得到词向量，为了将文本情感分析（情感分类）转化为机器学习问题，首先就是需要将符号数学化。在 NLP 中，最常见的词表示方法就是 One-hot Representation：将一个词映射成一个很长的单位向量，向量的长度就是词表的大小，如：“学习”表示成[0 0 0 1 0 0 0 0 0 0 0 0 0 0 ...]，“复习”表示成[0 0 0 0 0 0 0 1 0 0 0 0 0 0 ...]；这样就完成了词语的数学化表示。

但是，这样就存在“词汇鸿沟”的问题：即使两个词之间存在明显的联系但是在向量表示法中却体现不出来，无法反映语义关联。然而，Distributed Representation 却是能反映出词语与词语之间的距离远近关系，而用 Distributed Representation 表示的向量专门称为词向量，如：“学习”可能被表示成[0.1,0.1,0.1,0.15,0.2.....]，“复习”可能被表示成[0.11,0.12,0.1,0.15,0.22.....]，这样，两个词义相近的词语被表示成词向量后，它们的距离也是较近的，词义关联不大的两个词的距离会较为远。

一般而言,不同的训练方法或语料库训练得到的词向量是不一样的,它们的维度常见为 50 和 100 维。

现今常用的词向量模型如下:

a) LSA 矩阵分解模型

采用线性代数中的奇异值分解方法,选取前几个比较大的奇异值所对应的特征向量将原矩阵映射到低维空间中,从而达到词矢量的目的。

b) PLSA 潜在语义分析概率模型

从概率学的角度重新审视了矩阵分解模型,并得到一个从统计,概率角度上推导出来的和 LSA 相当的词向量模型。

c) LDA 文档生成模型

按照文档生成的过程,使用贝叶斯估计统计学方法,将文档用多个主题来表示。LDA 不只解决了同义词的问题,还解决了一次多义的问题。目前训练 LDA 模型的方法有原始论文中的基于 EM 和差分贝叶斯方法以及后来出现的 Gibbs Samplings 采样算法。

d) Word2Vector 模型

word2vec 是 Mikolov 的新作,最近几年刚刚火起来的算法,通过神经网络机器学习算法来训练 N-gram 语言模型,并在训练过程中求出 word 所对应的 vector 的方法。

我们在这里使用最后一种词向量模型。

word2vec 采用神经网络语言模型 NNLM 和 N-gram 语言模型,每个词都可以表示成一个实数向量。模型如下图所示:

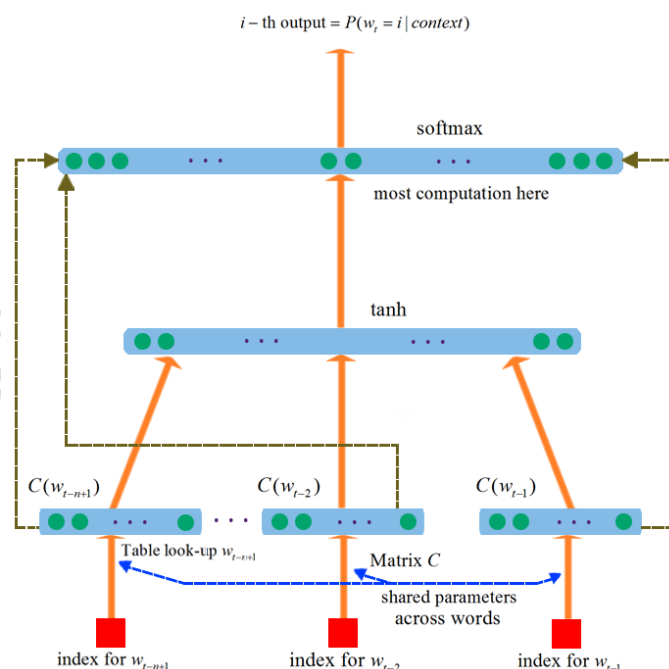


图 10 word2vec 模型展示图

图中最下方的 $w_{t-n+1}, \dots, w_{t-2}, w_{t-1}$ 就是前 $n-1$ 个词。现在需要根据这已知的 $n-1$ 个词预测下一个词 w_t 。 $C(w)$ 表示词 w 所对应的词向量, 存在矩阵 C (一个 $|V| \times m$ 的矩阵) 中。其中 $|V|$ 表

示词表的大小（语料中的总词数）， m 表示词向量的维度。 w 到 $C(w)$ 的转化就是从矩阵中取出一行。

网络的第一层（输入层）是将 $C(w_{t-n+1}), \dots, C(w_{t-2}), C(w_{t-1})$ 这 $n-1$ 个向量首尾相接拼起来，形成一个 $(n-1)m$ 维的向量，记为 x 。

网络的第二层（隐藏层）就如同普通的神经网络，使用 \tanh 作为激活函数。

网络的第三层（输出层）一共有 $|V|$ 个节点，每个节点 y_i 表示下一个词为 i 的未归一化 \log 概率。最后使用 softmax 激活函数将输出值 y 归一化成概率。最终， y 的计算公式为：

$$y = b + Wx + U \tanh(d + Hx)$$

其中 U 是隐藏层到输出层的参数，整个模型的多数计算集中在 U 和隐藏层的矩阵乘法中。

2.2.5.2 评论集子集的人工标注与映射

利用词向量构建的结果，我们进行评论集子集的人工标注，正面评论标为 1，负面评论标记为 2。（或者采用 python 的 NLP 包 `snownlp` 的 `sentiment` 功能做简单的机器标注，减少人为工作量），然后将每条评论映射为一个向量，将分词后评论中的所有词语对应的词向量相加做平均，使得一条评论对应一个向量。

2.2.5.3 训练栈式自编码网络

自编码网络是由原始的 BP 神经网络演化而来。在原始的 BP 神经网络中我们从特征空间输入到神经网络中，并用类别标签与输出空间来衡量误差，用最优化理论不断求得极小值，从而得到一个与类别标签相近的输出。但是在编码网络并不是如此，我们并不用类别标签来衡量与输出空间的误差，而是用从特征空间的输入来衡量与输出空间的误差。其结构如图所示：

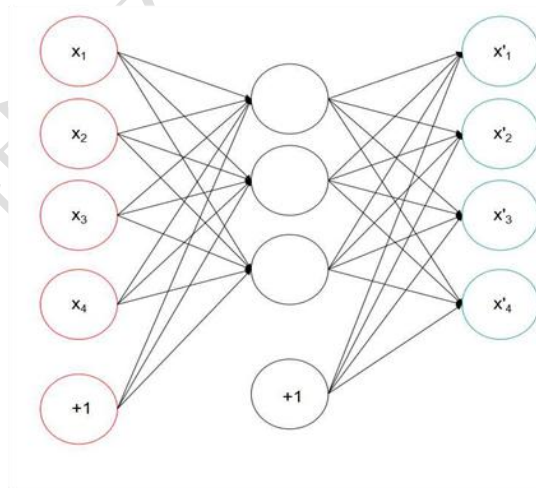


图 11 自编码网络结构示意图

我们把特征空间的向量 (x_1, x_2, x_3, x_4) 作为输入，把经过神经网络训练后的向量 (x'_1, x'_2, x'_3, x'_4) 与输入向量 (x_1, x_2, x_3, x_4) 来衡量误差，最终得到了一个能从原始数据中自主学习特征的一个特征提取的神经网络。从代数角度而言，亦即从一个线性相关的向量中，寻找出了一组低维的基，而这组基线性组合之后又能还原成原始数据。自编码网络正是寻找了一组这样的基。

神经网络的出现，时来已久，但是因为局部极值，梯度弥散，数据获取等等问题而构建不出深层的神经网络，直到 2007 年深度学习的提出，才让神经网络的相关算法得到质的改变。而栈式自编码就属于深度学习理论中一种能够得到优秀深层神经网络的方法。

栈式自编码神经网络是一个由多层稀疏自编码器组成的网络。它的思想是利用逐层贪婪训练的方法，把原来多层的神经网络剖分成一个个小的自编码网络，每次只训练一个自编码器，然后将前一层自编码的输出作为其后一层自编码器的输入，最后连接一个分类器，可以是 SVM，SoftMax 等等。上述步骤是为了得到一个好的初始化深度神经网络的权重，当我们连接好一个分类器后，还可以用 BP 神经网络的思想，反向传播微调我们神经元的权重，以期得到一个分类准确率更好的栈式自编码神经网络。

完成评论映射后，我们将标注的评论划分为训练集和测试集，在 MATLAB 下，利用标注好的训练集（标注值和向量）训练栈式自编码网络（SAE），对原始向量做深度学习提取特征，并后接 Softmax 分类器做分类，并用测试集测试训练好的模型的正确率。

2.2.5.4 情感分析

当 SAE 模型训练好后，我们便可以对整个评论集进行情感倾向分析。

2.2.6 基于语义网络的评论分析

本文综合使用语义网络以及 LDA 主题分析的方法对评论进行进一步的分析，包括各产品独有优势、各产品抱怨点以及顾客购买原因等，并结合以上分析对品牌产品的改进提出建议，这两种方法各有特点，各自能够解决相应的一些问题，比如语义网络方法通过语言关系构建有利于滤取产品的独有优势，而 LDA 主题分析方法通过调节评论集中各个特征词在潜在主题上的概率分布情况，抽取得到某种产品的热门关注点及其评论词。

首先我们进行基于语义网络的方法，这一部分我们主要通过由三种品牌型号的好、差评文本数据生成的语义网络图，结合共词矩阵以及评论定向筛选回查来完成对评论的分析。

2.2.6.1 语义网络的概念、结构与构建本质

语义网络是由 R.F.Simon 提出的用于理解自然语言并获取认知的概念，是一种语言的概念及关系的表达。语义网络实际上就是一幅有向网络图，举例下图所示：

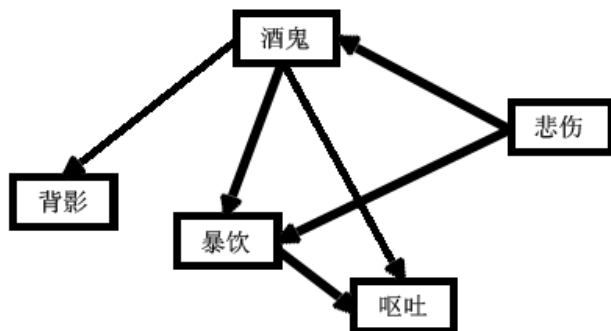


图 12 语义网络举例示意图

节点中的物体可以使各种用文字所表达的事物，而节点之间的有向弧则被用以表达节点之间的语言意义上的关系，其中的弧的方向是语言关系的因果指向，比如 A 指向 B 就意味着 A 与 B 有语言关系牵连且 A 与 B 分别是语义复杂关系的主动方与从动方。当然，这种用语言意义上的关系往往是复杂的。以上图为例，由于是一名酒鬼，那么他或她就经常会在特定情况之下（诸如朋友聚会、婚宴等）暴饮；一个人因受到各种挫折而显得的悲伤，长期的悲伤无法释怀，只能通过借酒浇愁就可能会成为酒鬼。这些里面就都是些复杂的关系。

虽然每一个语义网络结构里头事物（节点）之间的关系是复杂的，但是语义网络的每一道弧的形成从本质上看就是由于这种语义关系的存在。不同的用词语表达的特定事物之间就是因为存在千丝万缕的联系，才会形成一个个的语义网络。

2.2.6.2 基于语义网络进行评论分析的优势

从前面的论述当中我们知道，要想对中文的热水器评论进行合理的分析，我们必须采取的一项措施是分词，因为计算机不可能像我们人一样去识别每一个整句的语义，不能直接识别语句的整体结构思想，但是分词又会使得语句的整体结构变得凌乱，从而对分词后的语句直接进行诸如产品差异等复杂的分析变得不合实际，所以我们必须要采取方法尽可能将这种原已凌乱关系重新整合起来，使得复杂的分析重新变为可能。那么建立起事物之间（这里分出的每一个词代表一项事物）的语义网络关系就能够使得原已凌乱的关系得以整合，特别是那些可以连成通顺语料的词语的关系（即连接“因果”关系）的重新整合，而这种关系的成功重建能够清晰的还原语料中所反映出来的许多内容，特别是单独的词语无法清晰表达相应的情况的时候，比如：

“安装”与“方便”分开的时候，任何一方都不能清晰表达相关的情况，单独一个“安装”可以表达很多的东西，可以是“安装很容易”，也可以是“有师傅上门帮忙安装”，还可以是“安装要收手续费”等等；而单独一个“方便”也可以表达很多的东西，可以是“使用十分方便”，也可以是“商品签收方便快捷”，还可以是“交款方式方便简易”等等，但是如果“安装”和“方便”通过语义网络方式连接起来，如下图所示，就可以清晰的反映出是相关热水器产品在安装的时候比较便利。再比如“热水”与“不足”也是这样的情况，此处就不再赘述。



图 13 “安装”和“方便”的语义网络连接示意图

当这种语义网络建立起来后，我们就可以借助它进行各种各样的特定的分析，特别是判断特定产品优点、抽取各品牌的顾客关注点等上都具有一定的优势。以判断特定产品优点为例，如果某种产品相对于其他产品具有某种特定的优势，那么由该种商品的正面评论形成的语义网络上就会生成与其他产品正面评论形成的语义网络不一样的且蕴含着这种优势的关系连接，透过可视化，我们就能够从中抽取出来。

2.2.6.3 基于语义网络进行评论分析的前期步骤与解释

进行语义网络分析实际上所需要的前期步骤实际上就是在二分类文本情感分析的基础上增添，

语义网络的分析之所以要以二分类文本情感分析的结果为基础的原因在于评论是正面的以及评论是负面的大多都会具有不同的语意结构，且对于同一商品而言，正面以及负面的评论必然从根本上说关注的点是不完全一样的，信息也是不完全一样的，毕竟正面以及负面评论之间是存在逻辑冲突的。而这种正面负面评论的分割需要用到情感分析的技术。具体前期步骤如下：

①数据预处理，分词以及对停用词的过滤；

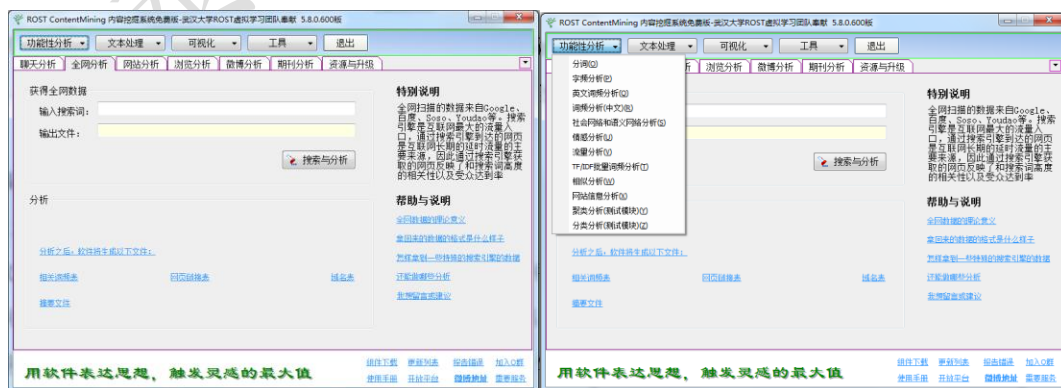
②进行情感倾向性分析，并借助此将评论数据分割成正面（好评）、负面（差评）、中性（中评）三大组；

③抽取正面（好评）、负面（差评）两组，以进行语义网络的构建与分析。

第一步我们可以直接按照原有的流程来进行，第三步的抽取只需要在第二步分成的三组结果中抽取即可，我们不对中性评论进行分析是因为中性评论往往携带着比较复杂的信息，难以对细节进行倾向性提取。

而第二步的情感倾向性分析并将评论数据分类可以在原有的情感分析工作基础上做出修改来完成，但是在此处我们使用ROSTCM6来完成该项操作。ROST系统是由武汉大学开发的一款免费反剽窃系统（ROSTCM6全称为ROST Content Mining System (Version 6.0)），可用以检测论文抄袭的现象；而同时ROST系统又是一款大型的免费用以社会计算的软件，可以用以实现多种类型的分析，包括情感倾向性分析以及后面我们将要进行语义网络的构建等。之所以我们使用ROSTCM6来完成情感分析是因为ROSTCM6软件的情感倾向性分析使用的是基于优化的情感词典的方法，其准确率目前来讲会比基于词向量以及基于神经网络的情感分析方法的正确率会高，而我们前述用于情感倾向性分析的方法是基于词向量以及基于神经网络的情感倾向性分析方法。另外，受限与现今中文分词技术的缺陷以及评论本身的特性，能够透过中文评论所挖掘出来的内容还是偏少的，因此对情感倾向性分析的正确率要求就要更高。当我们需要以此为基础进一步分析的时候，我们就需要利用基于情感词典的方法。第二步的具体流程如下：

单击“功能性分析”，再点击“情感分析”，然后将待分析的文件地址输入“待分析文件路径”对应框内，点击“分析”选项就得到了情感倾向性分析的结果，三种情感倾向被放入三个不同的txt文件内。



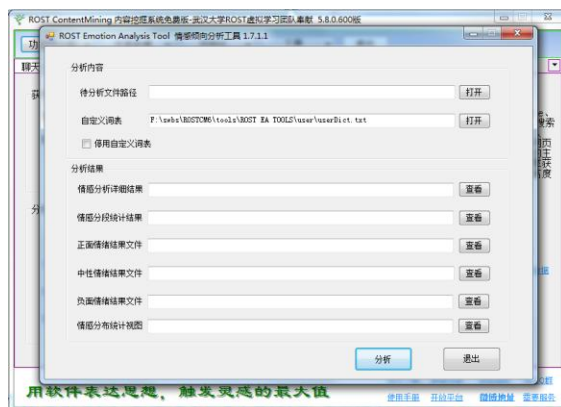


图14、15、16 ROSTCM6实现情感倾向性分析的步骤示意图

这三步完成后，我们便可以开始语义网络分析。

2.2.6.4 基于语义网络进行评论分析的实现过程

要进行语义网络分析，首先我们要分别对两大组重新进行分词处理，并提取出高频词（为了实现更好的分词效果，我们引入在分词词典中引入更多的词汇）。因为只有高频词之间的语义联系才是真正有意义的，个性化词语间关系不具代表性。然后在此基础上过滤掉显著的无意义的成分，减少分析干扰。最后再抽取行特征，处理完后便可进行两组的语义网络的构建。

我们亦利用软件 ROSTCM6 来完成这一部分及语义网络构建的操作。打开 ROSTCM6 软件，单击“功能性分析”选项，再点击“社会网络与语义网络分析”，我们便得到社会网络与语义网络分析的界面，如下所示：

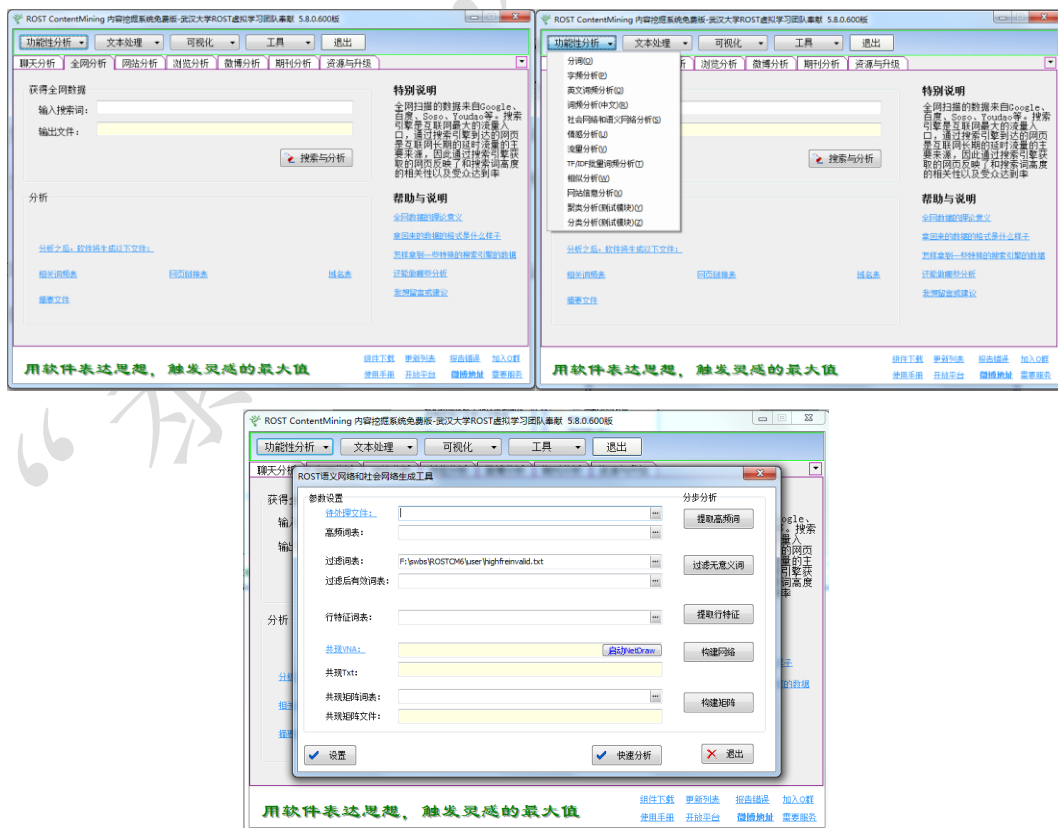


图 17、18、19 ROSTCM6 实现语义网络构建的步骤示意图

我们先将分好的好评差评两个文本文档当中的好评文档的地址输入“待处理文件”对应框内，并单击“提取高频词”、“过滤无意义词”以及“提取行特征”，这样我们便完成了对应的操作，系统还会自动生成对应处理后的文件。在此之后，我们依次单击“构建网络”与“启动 NetDraw”，然后我们就得到了好评文档的语义网络图（其生成的语义网络图可能不便观察，我们可以移动 NetDraw 生成的语义网络结果中的节点以增强该网络的可读性），为了方便分析，我们再单击“构建矩阵”，形成被挑选出的节点词的矩阵词表，该操作会生成一个 xls 文件。完成好评文档的语义网络图的构建后再对差评文档进行同样的操作，我们也将得到相应的语义网络图。三种牌子三种型号对应就会有总共六个好评文档及差评文档，对应就会生成六个语义网络图，并以此为基础，结合共词矩阵（可在语义网络生成后再点击“构建矩阵”形成）与评论定向筛选回查，我们便可进行相关评论分析。

2.2.7 基于 LDA 模型的主题分析

基于语义网络的评论分析进行初步数据感知后，我们从统计学习的角度，对主题的特征词出现频率进行量化表示。本文运用 LDA 主题模型，用以挖掘三种牌子评论中更多的信息。

主题模型在机器学习和自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型。直观上来说，传统判断两个文档相似性的方法是通过查看两个文档共同出现的单词的多少，如 TF、TF-IDF 等，这种方法没有考虑到文字背后的语义关联，可能在两个文档共同出现的单词很少甚至没有，但两个文档是相似的，因此在判断文档相似性时，应进行语义挖掘，而语义挖掘的有效工具即为主题模型。

如果一篇文档有多个主题，则一些特定的可代表不同主题的词语会反复的出现，此时，运用主题模型，能够发现文本中使用词语的规律，并且把规律相似的文本联系到一起，以寻求非结构化的文本集中的有用信息。比方说，对于热水器的商品评论，代表热水器特征的词语如“安装”、“出水量”、“服务”等会频繁地出现在评论中，运用主题模型，将与热水器代表性特征相关的情感描述性词语，同相应的特征词语联系起来，从而深入了解热水器评价的聚焦点及用户对于某一特征的情感倾向。LDA 模型作为其中一种主题模型，属于无监督的生成式主题概率模型。

2.2.7.1 LDA 主题模型介绍

潜在狄利克雷分配 (Latent Dirichlet Allocation, LDA) 是由 Blei 等人在 2003 年提出的生成式主题模型。生成模型，即认为每一篇文档的每一个词都是通过“一定的概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语”。LDA 模型也被称为三层贝叶斯概率模型，包含文档 (d)、主题 (z)、词 (w) 三层结构，能够有效对文本进行建模，和传统的空间向量模型 (VSM) 相比，增加了概率的信息。通过 LDA 主题模型，能够挖掘数据集中的潜在主题，进而分析数据集的集中关注点及其相关特征词。

LDA 模型采用词袋模型 (Bag Of Words, BOW) 将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。

定义词表大小为 V ，一个 V 维向量 $(1, 0, 0, \dots, 0, 0)$ 表示一个词。由 N 个词构成的评论记为

$\mathbf{d} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)$ 。假设某一商品的评论集 D 由 M 篇评论构成，记为 $D = (\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_M)$ 。 M 篇评论分布着 K 个主题，记为 $z_i (i = 1, 2, \dots, K)$ 。记 α 和 β 为狄利克雷函数的先验参数， θ 为主题在文档中的多项分布的参数，其服从超参数为 α 的 Dirichlet 先验分布， ϕ 为词在主题中的多项分布的参数，其服从超参数 β 的 Dirichlet 先验分布。LDA 模型图示见下图：

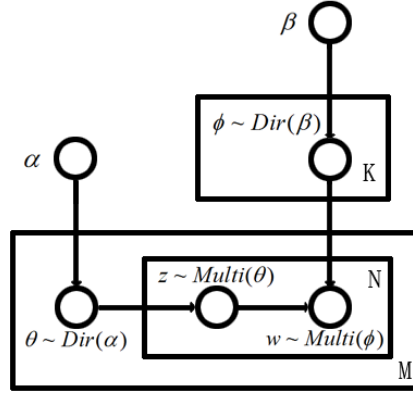


图 20 LDA 模型结构示意图

LDA 模型假定每篇评论由各个主题按一定比例随机混合而成，混合比例服从多项分布，记为：

$$Z | \theta = \text{Multinomial}(\theta)$$

而每个主题由词汇表中的各个词语按一定比例混合而成，混合比例也服从多项分布，记为：

$$W | Z, \phi = \text{Multinomial}(\phi)$$

在评论 d_j 条件下生成词 w_i 的概率表示为：

$$P(w_i | d_j) = \sum_{s=1}^K P(w_i | z = s) \times P(z = s | d_j)$$

其中， $P(w_i | z = s)$ 表示词 w_i 属于第 s 个主题的概率， $P(z = s | d_j)$ 表示第 s 个主题在评论 d_j 中的概率。

2.2.7.2 LDA 主题模型估计

LDA 模型对参数 θ 、 ϕ 的近似估计通常使用马尔科夫链蒙特卡洛 (Markov Chain Monte Carlo, MCMC) [1] 算法中的一个特例 Gibbs 抽样。利用 Gibbs 抽样对 LDA 模型进行参数估计，依据下式：

$$P(z_i = s | Z_{-i}, W) \propto (n_{s,-i} + \beta_i) / \left(\sum_{i=1}^V n_{s,-i} + \beta_i \right) \times (n_{s,-j} + \alpha_s)$$

其中， $z_i = s$ 表示词 w_i 属于第 s 个主题的概率， Z_{-i} 表示其他所有词的概率， $n_{s,-i}$ 表示不包含当前词 w_i 的被分配到当前主题 z_s 下的个数， $n_{-j,s}$ 表示不包含当前文档 d_j 的被分配到当前主题 z_s 下的个数。

通过对上式的推导，可以推导得到词 w_i 在主题 z_s 中的分布的参数估计 $\phi_{s,i}$ ，主题 z_s 在评论 d_j 中的多

项分布的参数估计 $\theta_{j,s}$:

$$\phi_{s,i} = (n_{s,i} + \beta_i) / \left(\sum_{i=1}^V n_{s,i} + \beta_i \right)$$

$$\theta_{j,s} = (n_{j,s} + \alpha_s) / \left(\sum_{s=1}^K n_{j,s} + \alpha_s \right)$$

其中, $n_{s,i}$ 表示词 w_i 在主题 z_s 中出现的次数, $n_{j,s}$ 表示文档 d_j 中包含主题 z_s 的个数。

LDA 主题模型在文本聚类、主题挖掘、相似度计算等方面都有广泛的应用, 相对于其他主题模型, 其引入了狄利克雷先验知识, 因此, 模型的泛化能力较强, 不易出现过拟合现象。其次, 它是一种无监督的模式, 只需要提供训练文档, 它就可以自动训练出各种概率, 无需任何人工标注过程, 节省大量人力及时间。再者, LDA 主题模型可以解决多种指代问题。例如: 在热水器的评论中, 根据分词的一般规则, 经过分词的语句会将“费用”一词单独分割出来, 而“费用”是指安装费用, 还是热水器费用等其他情况, 如果简单地进行词频统计及情感分析, 是无法识别的, 从而也无法准确了解用户反映的情况。运用 LDA 主题模型, 可以求得词汇在主题中的概率分布, 进而判断“费用”一词属于哪个主题, 并求得属于这一主题的概率和同一主题下的其他特征词, 从而解决多种指代问题。

2.2.7.3 运用 LDA 模型进行主题分析的实现过程

在本文商品评论关注点的研究中, 即对评论中的潜在主题进行挖掘, 评论中的特征词是模型中的可观测变量。一般来说, 每条评论中都存在一个中心思想, 即主题。如果某个潜在主题同时是多则评论中的主题, 则这一潜在主题很可能是整个评论语料集的热门关注点。在这个潜在主题上越高频的特征词将越可能成为热门关注点中的评论词。

首先, 为提高主题分析在不同情感倾向下热门关注点反映情况的精确度, 本文在语义网络的情感分类结果的基础上, 对不同情感倾向下的潜在主题分别进行挖掘分析, 从而得到不同情感倾向下用户对热水器不同方面的反映情况。例如, 选取差评中的一条评论“售后服务差极了, 不买他们的材料不给安装, 还谎称免费安装, 其实要收挺贵的安装费, 十分不合理。这也算了, 安装费之前说二百, 安好之后要四百, 更贵了, 更加不合理, 不管是安装师傅自己还是美的规定, 都是很差很差的体验, 我看其他人的了, 一样的安装, 比别人贵的安装费。而且安装师傅做事粗糙, 态度粗鲁。”在这条评论中, “安装费”和“安装师傅”在这条评论中出现频率较高, 可作为潜在主题。同时, 可以得到潜在主题上特征词的概率分布情况, 反映潜在主题“安装费”的特征词包括“贵”、“不合理”, 反映“安装师傅”的特征词包括“粗糙”、“粗鲁”。

接着, 分别统计整个评论语料库中正负情感倾向的主题分布情况, 对两种情感倾向下, 各个主题出现的次数从高到低进行排序, 根据分析需要, 选择排在前若干位的主题作为评论集中的热门关注点, 然后根据潜在主题上的特征词的概率分布情况, 得到所对应的热门关注点的评论词。

本文运用 Python2.7 软件编写 LDA 主题模型的算法, 并采用 Gibbs 抽样方法对 LDA 模型的参数

进行近似估计，由上文的模型介绍可知，模型中存在 3 个可变量需要确定最佳取值，分别是狄利克雷函数的先验参数 α 和 β 、主题个数 K 。本文中将狄利克雷函数的先验参数 α 和 β 设置为经验值，分别是 $\alpha = 50 / K$ ， $\beta = 0.1$ 。而主题个数 K 采用统计语言模型中常用的评价标准困惑度[6]来选取，即令 $K=50$ 。

2.3. 结果分析

有了以上的理论基础，现在我们进行对结果的分析。

2.3.1 情感倾向性分析结果

我们首先分析情感倾向性分析的结果，结果如下图所示：

表 1 深度学习法实现情感倾向性分析的结果表

品牌型号	训练迭代次数	测试正确率 1	测试正确率 2	测试平均正确率	
Midea_F50_15A1	100	77.687%	75.822%	76.75%	
	150	76.673%	77.363%	77.02%	
Haier_ES50H_Q1(ZE)	100	78.808%	79.526%	79.17%	
	150	77.613%	78.798%	78.21%	
Vanward_DSCF50_T4A	100	69.479%	66.612%	68.05%	
	150	70.929%	69.908%	70.42%	

上述结果都是在 1000 条训练样本的条件下训练出的模型，用评论集余下的评论作为测试集测试的正确率，其中迭代次数是控制模型训练终止的阈值，由于 SAE 的参数初值都是随机化的，所以对于每次条件下，测试 2 次正确率取平均值。经过多次调参总结，当训练迭代次数在 100 附近时模型的泛化能力达到最优，而且模型并非随着迭代次数的增多而渐优，因为迭代次数的增多仅仅意味着在训练集上的分类效果越佳。

在多个评论集的测试下，可以看出模型的自动学习分类能力尚可，正确率最高能达到 79.526%，模型在词向量的基础上，通过 SAE 特征提取，挖掘出更深层的语义信息，从而达到较好的分类能力，同时，SAE 模型不需要我们过多的关注特征的选择和构造问题（如 tf，tf-idf 特征等），只需将关注点聚集在如何训练出更好的词向量的问题上，这也是深度学习吸引人的优势之一。

相对地，本文受限于词向量的训练不足和训练样本的正负倾斜，模型的精度还有待提高，尤其是 Vanward 评论集由于规模受限导致精度降低的问题。但随着语料库的完备和词向量模型的发展，相信 SAE 模型的精度会得到进一步的提升。

2.3.2 语义网络的结果与分析

首先我们观察海尔品牌型号为 ES50H-Q1 (ZE) 的热水器的好评与差评的语义网络图，其好评与差评的语义网络图分别如下所示：

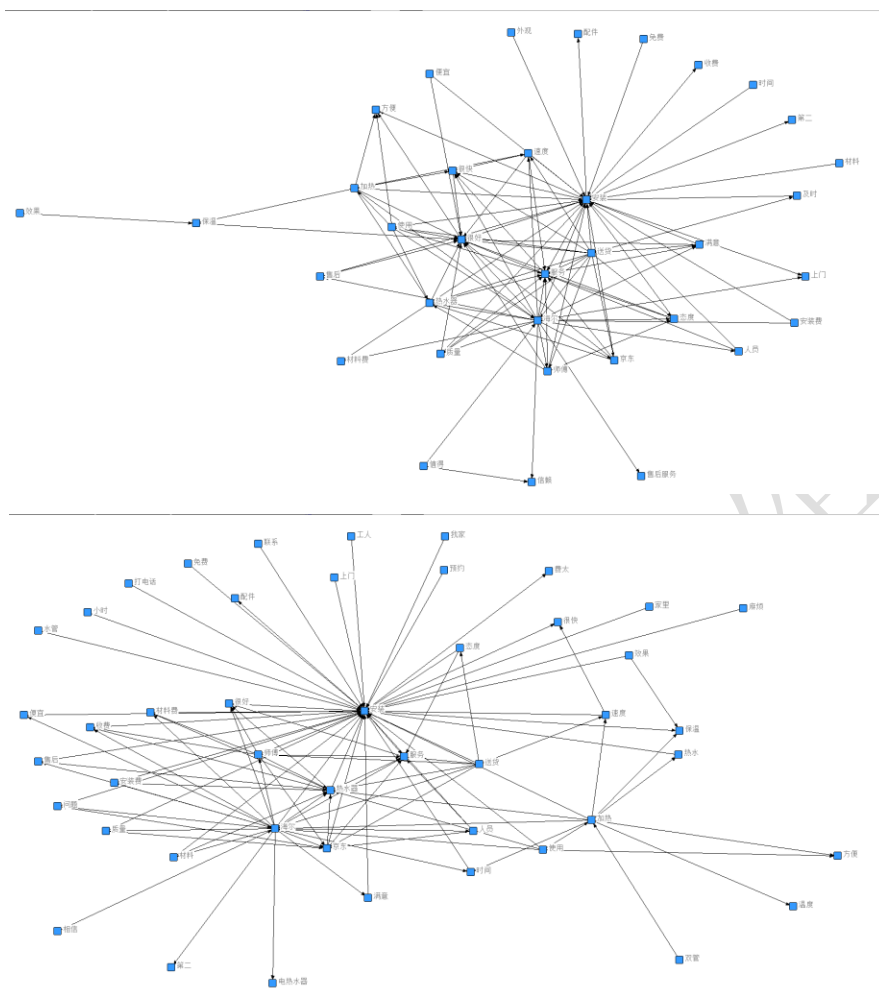


图 21、22 海尔品牌型号为 ES50H-Q1 (ZE) 的热水器的好评与差评的语义网络图

从生成的两个语义网络的结果并对比，我们可以得知给予海尔品牌型号为 ES50H-Q1 (ZE) 的热水器当中比较明显的赞点在于安装比较方便以及热水器的保温效果比较好，因为这两道弧只在好评的语义网络中有构建出来，即没有或没有多少类似“安装实际上不怎么样，只是使用起来比较方便而已”的差评语段（一条评论中的情感为差的那一部分语料段）。换句话说，安装方便以及热水器的保温效果比较好基本上都是在好评里面出现的，而在好评中出现上述的类似差评语段的几率比较低，即海尔该种型号的这种优势是确实存在的。另外，它们在共词矩阵中的值都大于等于 100，而共词矩阵特定元素代表着特定的两个高频词同时在评论中出现的相应的评论数值，该数值从人与人的表达以及所处环境的固有差异来看已经比较高，这也能很好证明这两者的优势确实存在的这一点。

而从差评生成的语义网络来看，排除一些类似“我家”和“安装”等有意义但是又无情感意义的连接以及类似“速度”和“很快”等属于表达好评的连接关系（即导致总体差评的源头不在此，虽然他们同时在语料中出现的可能性高，或者是实际上是类似上段所提的差评语段的结构的抽取），与好评生成的相比比较明显的差异在于多出了“双管”与“加热”以及“水管”与“安装”，但是由于从这他们的这种结构关系比较难以判断他们是否就是用户抱怨点等，为此我们就需要将语料导入

EXCEL 文件，透过查找关键字并大致浏览每条评论的方式来了解这里面是否真的存在有价值的信息。从我们的观察结果看，“双管”和“加热”实际上所反映的是该款热水器的特点，是双管加热的，且并不是负向情感表达的载体。但是“水管”与“安装”构成的结构就反映出来了问题，通过“水管”与“安装”的结构提示查找与水管有关的评论发现水管包括安装的各种问题是海尔该种型号热水器的一个用户抱怨点，这从筛选出来的一些例子里面就能够发现（这里列举 3 条）：

- ① “刚装上之后，晚上回来发现水管没接好，既然漏水。郁闷”，
- ② “好黑啊！收了我 130 元的安装费！太黑了！！我水管都已经在热水器底下了！！说是材料费！！”，
- ③ “调节加热温度的按钮不方便不科学，非常费劲。尤其是我近视眼在浴室洗澡不戴眼镜的话，只能靠手感和经验操作。家中的老人更难掌握。而且有一条泄水管在外面非常难看”。

至于海尔品牌的购买原因，从网络结构上看，就可能显得比较多，海尔送货及时，价格实惠等都可能成为原因。

现在转到美的品牌 F50-15A1 型号的热水器的好评与差评的语义网络图，如下所示：

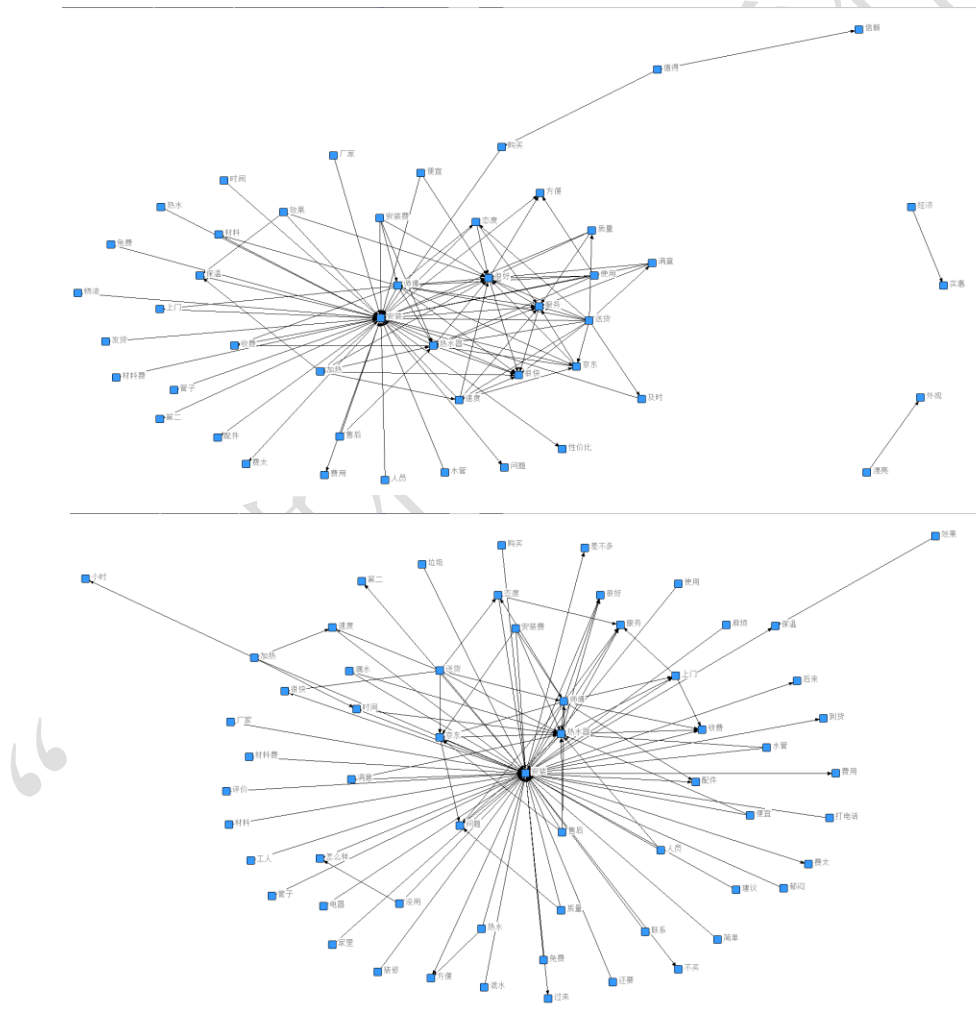


图 23、24 美的品牌型号为 F50-15A1 的热水器的好评与差评的语义网络图

按照上述类似分析海尔的型号的方法与思路，透过两个语义网络图的对比，我们可以得知热水器的漂亮外观成为了给予美的品牌 F50-15A1 型号的热水器好评的赞点（特别是它们作为孤立的连接

出现时，即可以说它们这种关系是更加紧密而非纯统计意义上的关系），而同时由这个赞点的性质可知它有时候也会成为用户购买这种热水器的原因，我们随机抽取了几条好评数据，从抽取结果来看，这个赞点与购买点确实存在（这里列举 3 条）：

① “外观挺漂亮，但因为是圆形有点占空间，安装要收五金费，收了 188 元，美的的安装工作很上心，装的挺好的，烧水很快，挺费电的。保温效果也不错。整体还不错。”，

② “外观漂亮，性价比高，大品牌值得信赖”，

③ “外观很漂亮。价格挺便宜点。”。

再到美的品牌抱怨点的挖掘，对比可看出美的品牌的水器存在着热水器漏水的问题，为此我们筛选相关的评论，可以得到更详细的信息已证实这个抱怨点的存在（这里列举 3 条）：

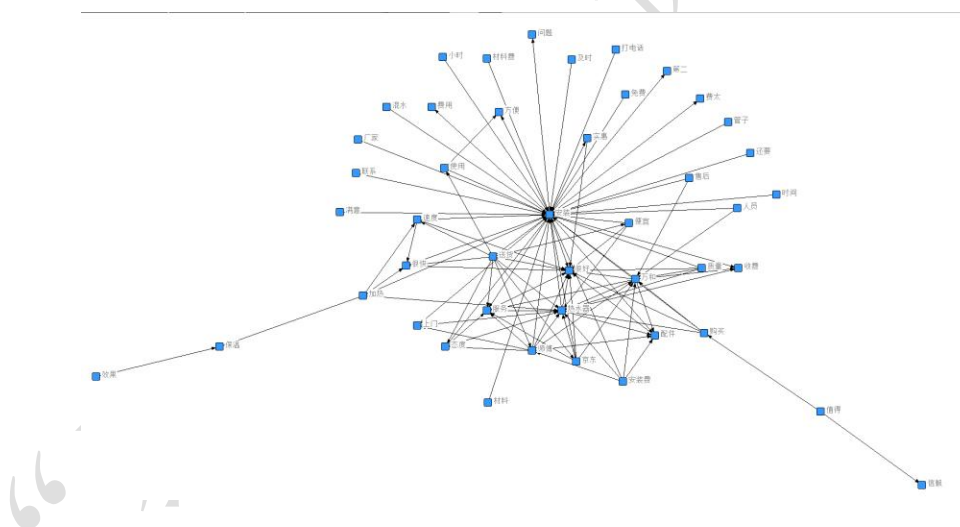
① “产品除了混水阀使用时有点漏水，别的问题不大……”，

② “买的价格*售后安装*元~美的的售后服务还不错 但是装完 我回家看怎么漏水？怎么京东送货人员 态度不好 急急忙忙的 我还没说完就挂我电话了，怎么袋里写家电下乡啊？什么意思”，

③ “上周买了两个 50 升的热水器，昨天我回家一看靠左边的热水管漏水了，郁闷 退货太麻烦了 给怎么办？ 我就想凑胡用 换条好点的管子，另一个还没用呢，网上不太靠谱啊”。

另外，和抽取的海尔品牌型号一样，水管的安装等问题也是抽取的美的品牌型号的问题，这里就不再赘述。

最后到万和 DSCF50-T4A 型号的，对应的两个语义网络图如下所示：



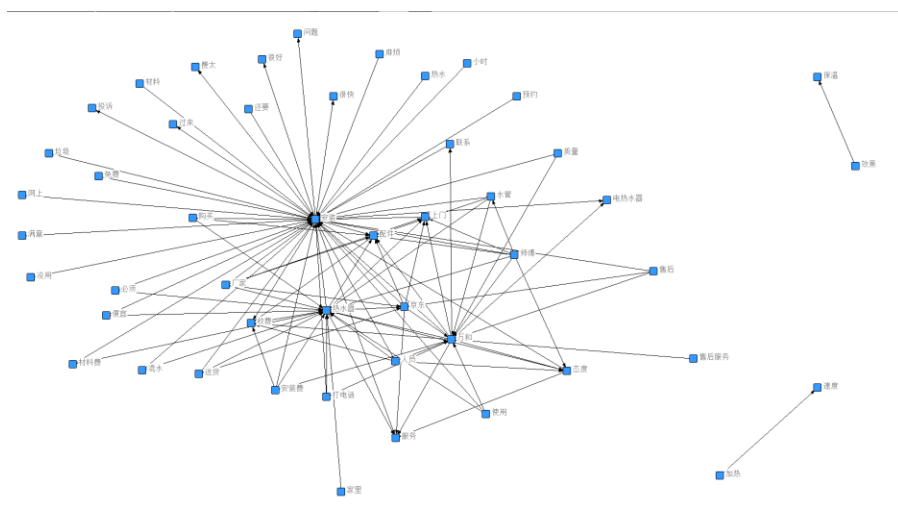


图 25、26 万和品牌型号为 DSCF50-T4A 的热水器的好评与差评的语义网络图

从语义网络结构上我们可以看出，万和该种型号的热水器除了可能类似价格比较便宜、使用比较方便一类的在电商品台上比较常见的评论（即比较常见的称赞之处）外，并没有十分突出的与其它热水器相比的优势。换句话说，从好评的语义网络图来看，并不能展现出该种热水器的产品优势所在，这也与该种热水器的好评度相对美的、海尔等一些品牌热水器低相吻合。显然，没有足够的优势就会缺乏相应的竞争力。

至于抱怨点，从差评语义网络图上看，最明显的一点，与通常的热水器加热速度为赞点相比，万和该种型号的热水器的加热速度就成为了不少消费者的抱怨之处，由于加热速度的表达方式较多，我们直接用“加热”来筛选，筛选后发现确实差评中这类的评论就有不少（这里列举 3 条），显示出了这方面的问题：

- ① “加热速度慢了些 毕竟单管”，
- ② “适合一两个人使用，加热速度一般”，
- ③ “加热慢点，水温可以。”。

另外，从列举的第一条反映加热速度慢的差评中我们也得知这种热水器是单管的，相比于抽取的海尔型号的双管，这是万和该种型号的一个缺陷之处。

综合上述六个语义网络图，我们可以十分清晰的发现，大多数评论都是围绕着“安装”，这个话题展开（这在共词矩阵当中也能清晰地反映出来），这也是电商平台评论制度的通病，一般情况下，顾客都会在购买不久后就会主动或被请求做出评论，而这个时间段顾客可能仅仅是完整经历了安装过程而已，实际使用时间并不长，要想使得评论能够更好的反映一些其它的好处或问题，相关的评论制度可能就要更改，比如延迟多日再评送更多积分等等，以吸引顾客给出更多更有价值的评论，真正体现出给予商品评论的意义。

2.3.3 LDA 模型构造结果与分析

本文选取正负情感倾向下各 3 个主题（共 6 个）进行分析，每个主题中提取 10 个的高频特征词进行挖掘分析。

本文首先对海尔品牌型号为 ES50H-Q1 (ZE) 热水器的好评和差评评论集分别进行分析。运用 LDA 模型，从运行结果中选取正负情感倾向下各 3 个主题的结果如下：

表 2 海尔好评下的 3 个潜在主题

主题 1		主题 2		主题 3	
w	p	w	p	w	p
不错	0.0337	不错	0.0882	不错	0.0610
买	0.0315	很好	0.0504	加热	0.0463
好	0.0302	好	0.0450	使用	0.0243
热水器	0.0271	品牌	0.0346	很好	0.0236
送货	0.0209	价格	0.0305	保温	0.0232
服务	0.0202	便宜	0.0206	好	0.0226
天	0.0176	性价比	0.0182	时间	0.0179
很好	0.0166	高	0.0177	挺	0.0166
京东	0.0166	产品	0.0176	方便	0.0159
花	0.0157	实惠	0.0133	热	0.0156

表 3 海尔差评下的 3 个潜在主题

主题 1		主题 2		主题 3	
w	p	w	p	w	p
安装	0.0715	安装	0.2208	加热	0.1941
买	0.0635	热水器	0.1016	热水	0.0702
热水器	0.0411	买	0.0951	速度	0.0596
次	0.0352	材料费	0.0887	慢	0.0586
货	0.0347	安装费	0.0548	水	0.0548
质量	0.0336	花	0.0500	点	0.0481
客服	0.0289	贵	0.0500	时间	0.0442
服务	0.0283	售后	0.0253	使用	0.0375
师傅	0.0267	价格	0.0246	天	0.0365
问题	0.0219	比较	0.02399	次	0.0358

根据海尔热水器好评的 3 个潜在主题的特征词提取，主题 1 中的高频特征词，即热门关注点主要是送货（0.0209）、服务（0.0202）、京东（0.0166）等，由此可以看出，主题 1 主要反映的是京东送货服务质量好；主题 2 中的高频特征词，即热门关注点主要是价格（0.0305）、便宜（0.0206）、性价比（0.0182）等，由此可以看出，主题 2 主要反映的是海尔品牌价格实惠、性价比高；主题 3 中的高频特征词，即热门关注点主要是加热（0.0463）、保温（0.0232）、好（0.0226）等，由此可以看出，主题 3 主要反映海尔热水器实用又方便，保温效果良好。

从海尔热水器差评的 3 个潜在主题中，我们可以看出，主题 1 中的高频特征词主要是安装（0.0715）、次（0.0352）、客服（0.0289）、师傅（0.0267）等，即主题 1 主要反映的是海尔热水器安装师傅、客服服务不周，热水器质量较次；主题 2 中的高频特征词主要是材料费（0.0887）、安装费（0.0548）、贵（0.0500）等，由此可以看出，主题 2 主要反映的是海尔热水器安装费、材料费贵；主题 3 中的高频特征词主要是加热（0.1941）、速度（0.0596）、慢（0.0586）等，由此可

以看出，主题 3 主要反映海尔热水器加热速度慢。

综合以上对主题及其中的高频特征词可以看出，海尔热水器的优势有以下几个方面：价格实惠、性价比高、热水器实用、使用起来方便、保温效果良好。

相对而言，用户对海尔热水器的抱怨点主要体现以下几个方面：海尔品牌的客服服务不周、安装费用贵及安装人员服务不周、热水器加热速度慢。

在好评的主题 2 中，“品牌”一词出现的频率较高。在 EXCEL 中将“品牌”定位回原来评论集中，发现许多用户购买海尔热水器的原因在于认准海尔是一个大品牌。例如，有评论提到“一直认准海尔这一大品牌，买它的东西放心。”，再如“海尔是个老品牌了，家里一直使用这一牌子的家电，物美价廉。”因此，用户的购买原因可以总结为以下几个方面：海尔大品牌值得信赖，海尔热水器价格实惠，性价比高。

根据对京东平台上，海尔热水器的用户评价情况进行 LDA 主题模型分析，我们对海尔品牌提出以下建议：

1、在保持热水器使用方便、价格实惠等优点基础上，对热水器在的加热效率上进行改进，从整体上提升热水器的质量。

2、适度降低安装费用和材料费用，以此在大品牌的竞争中凸显优势。

接着，本文对美的品牌 F50-15A1 型号的热水器的好评和差评评论集分别进行分析。运用 LDA 模型，从运行结果中选取正负情感倾向下各 3 个主题的结果如下：

表 4 美的好评下的 3 个潜在主题

主题 1		主题 2		主题 3	
w	p	w	p	w	p
不错	0.0988	不错	0.1053	不错	0.0460
高	0.0969	很好	0.0646	时间	0.0392
性价比	0.0819	买	0.0446	加热	0.0381
加热	0.0593	好	0.0390	水	0.0357
很好	0.0564	东西	0.0318	好	0.0308
速度	0.0487	京东	0.0305	热水	0.0305
使用	0.0383	服务	0.0300	洗	0.0300
方便	0.0372	送货	0.0273	够	0.0279
外观	0.0345	质量	0.0257	保温	0.0261
满意	0.0333	给力	0.0236	烧	0.0258

表 5 美的差评下的 3 个潜在主题

主题 1		主题 2		主题 3	
w	p	w	p	w	p
热水器	0.1313	安装	0.3061	水	0.1525
安装	0.0978	上门	0.1047	阀	0.1017
买	0.0951	师傅	0.1007	热水器	0.0860
安装费	0.0911	服务	0.0948	装	0.0821
贵	0.0737	售后	0.0849	买	0.0782
装	0.0724	不好	0.0731	次	0.0704

花	0.0683	收费	0.0632	热水	0.0684
钱	0.0590	问题	0.0415	使用	0.0548
师傅	0.0536	使用	0.0395	问题	0.0548
东西	0.0536	阀	0.0375	售后	0.0528

根据海尔热水器好评的 3 个潜在主题的特征词提取，主题 1 中的高频特征词，即热门关注点主要是性价比（0.0819）、使用（0.0383）、外观（0.0345）等，由此可以看出，主题 1 主要反映的是美的热水器的性价比高、外观好看、使用方便；主题 2 中的高频特征词，即热门关注点主要是京东（0.0305）、送货（0.0273）、质量（0.0257）等，由此可以看出，主题 2 主要反映的是美的热水器质量较好、京东送货服务好；主题 3 中的高频特征词，即热门关注点主要是加热（0.0381）、热水（0.0305）、保温（0.0261）等，由此可以看出，主题 3 主要反映美的热水器加热速度快、保温效果好。

在差评的主题 3 中，“阀”这个字出现的频率较高，通过在 EXCEL 中将“阀”定位回原来评论集中，发现用户使用这个字构成的词语众多，例如：“混水阀”、“换水阀”、“加水阀”等，然而这些词表达的水器零件实际上是相同的。因此，在这一部分的模型运用中，对文本数据进行重新分词处理，即将“阀”一字单独分割出来，以真实反映用户对产品的抱怨关注点。同时，根据主题 1 的特征词可以看出，用户大多数反映的都是“阀”的质量问题。

因此，从美的热水器差评的 3 个潜在主题中，我们可以看出，主题 1 中的高频特征词主要是安装费（0.0911）、贵（0.0737）、钱（0.0590）等，即主题 1 主要反映的是美的热水器安装收费高；主题 2 中的高频特征词主要是安装（0.3061）、收费（0.0632）、问题（0.0415）等，主题 3 主要反映的是美的热水器安装人员乱收费；主题 3 中的高频特征词主要是阀（0.1017）、次（0.0704）、问题（0.0548）等，主题 3 主要反映美的热水器混水阀质量差。

综合以上对主题及其中的高频特征词可以看出，美的热水器的优势有以下几个方面：价格实惠、性价比高、外观好看、热水器实用、使用起来方便、加热速度快、保温效果良好。

相对而言，用户对海尔热水器的抱怨点主要体现以下几个方面：美的热水器安装的混水阀质量差、安装费用贵及安装人员乱收费。

因此，用户的购买原因可以总结为以下几个方面：美的大品牌值得信赖，美的热水器价格实惠，性价比高。

根据对京东平台上，美的热水器的用户评价情况进行 LDA 主题模型分析，我们对美的品牌提出以下建议：

1、在保持热水器使用方便、价格实惠等优点基础上，对热水器在的混水阀质量、进行改进，从整体上提升热水器的质量。

2、提升安装人员及客服人员的整体素质，提高服务质量。只能安装费用收取明文细则，并进行公开透明，减少安装过程的乱收费问题。适度降低安装费用和材料费用，以此在大品牌的竞争中凸显优势。

从以上对两个家电大品牌：美的和海尔的分析上来看，美的和海尔大品牌之间有许多相似性，例如，价格实惠、热水器使用方便、性价比高等优点，体现了大品牌大销量低价格的特征。而对于美的来说，热水器混水阀的质量问题是大多数用户抱怨的关注点，美的应提高热水器混水阀质量，以增加热水器的销量及客户满意度。对于海尔来说，客服服务态度问题是大多数用户抱怨的地方，海尔应注重对客服人员的培养，提高整体员工的素质，以增强客户的认可度。美的和海尔都存在安装费用乱收取的问题，由于两者均为大品牌，公司规模较大，员工人数较多，容易造成相关规定疏漏及人员管理较松问题。因此，美的和海尔应加强内部人员管理和培训，并制定相关的管理规定，以规范员工的办事行为，同时，应透明公开各项额外费用收取清单，协助用户及时发现乱收费问题，并避免问题发生。

差异化买点：美的品牌相对于海尔品牌，其明显优势在于热水器加热速度快、外观好看；而海尔品牌的明显优势在于安装工作人员基本按照安装收费标准收取，给客户留下的员工素质印象良好。

作为两个竞争激烈的大品牌，有许多共同优点，同时存在相同的弱点时，哪个品牌能够及时对问题采取有效的解决方法，在品牌竞争中形成明显的差异化优势，从长远的角度来看，这一品牌在长期竞争中更加具有竞争力。

最后，本文对万和 DSCF50-T4A 型号的热水器的的好评和差评评论集分别进行分析。运用 LDA 模型，从运行结果中选取正负情感倾向下各 3 个主题的结果如下：

表 6 万和好评下的 3 个潜在主题

主题 1		主题 2		主题 3	
w	p	w	p	w	p
安装	0.1067	价格	0.1904	不错	0.1018
买	0.0560	便宜	0.1674	好	0.0496
装	0.0291	不错	0.0923	加热	0.0451
热水器	0.0276	实惠	0.0923	使用	0.0424
配件	0.0274	很好	0.0750	安装	0.0369
安装费	0.0269	好	0.0726	挺	0.0326
好	0.0263	好用	0.0561	保温	0.3023
售后	0.0243	方便	0.0379	很好	0.0230
京东	0.0224	实用	0.0322	高	0.0289
送货	0.0216	东西	0.0305	速度	0.0279

表 7 万和差评下的 3 个潜在主题

主题 1		主题 2		主题 3	
w	p	w	p	w	p
安装	0.1708	安装	0.2346	买	0.2179
买	0.1636	太	0.1115	装	0.1809
热水器	0.1079	售后	0.1077	安装	0.1790
价格	0.1043	安装费	0.0961	热水器	0.1109
东西	0.0935	使用	0.0827	配件	0.0817
京东	0.0845	热水器	0.0750	水	0.0642
水	0.0773	收	0.0731	收	0.0506
配件	0.0683	京东	0.0635	安装费	0.0333

使用	0.0522	买	0.0615	使用	0.0331
售后	0.0449	水	0.0538	东西	0.0311

根据万和热水器好评的 3 个潜在主题的特征词提取，主题 1 中的高频特征词，即热门关注点主要是配件（0.0274）、安装费（0.0269）、京东（0.0224）等，由此可以看出，主题 1 主要反映的是万和热水器的安装费及配件收费便宜、京东送货情况良好；主题 2 中的高频词，即热门关注点主要是价格（0.1904）、便宜（0.1674）、好用（0.0561）等，由此可以看出，主题 2 主要反映的是万和热水器价格便宜、使用方便；主题 3 中的高频词，即热门关注点主要是加热（0.0381）、保温（0.0261）、速度（0.0279）等，由此可以看出，主题 3 主要反映万和热水器加热速度快、保温效果好。

从表 7 可以看出，对于万和热水器的差评评论集 3 个潜在主题都没有很明显的特征可供分析。结合数据实际情况进行分析，主要原因是万和热水器在京东商城的评论数据集总量本身较少。被归类为负情感倾向性的评论更少。而中文对于同一问题的表达又是各式各样的。因此，LDA 模型中词频统计这一步骤的效果不佳，继而造成后续的主题提取结果不明显。

相对于美的和海尔两个大品牌来说，万和热水器的优势主要在于安装收费便宜方面，同时万和热水器应积极提高热水器质量，以提高品牌知名度。

3. 结论

本文通过对处理过的京东三家品牌型号的热水器的文本评论数据利用栈式自编码神经网络等方法建立多种数据挖掘模型，得到了具有一定价值的结果，实现了对文本评论数据的情感倾向性分析以及一定程度上的对包括用户赞点、抱怨点、购买原因等在内的更细节的文本信息的挖掘与认识，而这些结果对于包括电商平台以及相关生产商家在内都具有一定性的指导意义，比如店商平台就可以将对文本评论数据挖掘得到的用户的抱怨点反馈到相关生产商家当中（比如海尔品牌的型号为 ES50H-Q1 (ZE) 的热水器的与水管相关的问题），促使其解决相关的问题，以重新赢得给予差评的消费者对于生产厂家的信任以及恢复在该电商平台购物的舒适度。

但是从我们的分析结果当中也可以看出总体来讲效果还不是特别的好，比如我们的情感倾向性分析结果就会与真实结果有一定程度上的出入，这里面既涉及到中文语言结构所必然导致的文本评论分析的缺陷的问题，也涉及到当今中文文本挖掘模型的不足以及评论数据本身所具有的问题，这也是我们在后期进一步的对中文文本数据的研究过程中可以继续深入探讨的地方。

4. 参考文献

[1]BERG B A .Markov Chain Monte Carlo Simulations and Their Statistical Analysis [M]. Singapore:

World Scientific. 2004

[2]朱少杰. 基于深度学习的文本情感分类研究[D].哈尔滨: 哈尔滨工业大学,2014

[3] Deep Learning in NLP (一) 词向量和语言模型[Z]. <http://licstar.net/archives/328>,2013-07-29

[4]崔志刚.基于电商网站商品评论数据的用户情感分析[D].北京: 北京交通大学, 2014 年 6 月

[5]大连理工大学信息检索研究室[Z]. <http://ir.dlut.edu.cn/NewsShow.aspx?ID=291>

[6]Cao Juan, Xia Tian, Li Jin Tao, A density method for adaptive LDA model selection[J]. Neurocomputing 2009(72):1775-1781

[7]张梦笑.基于 LDA 模型的观点聚类研究[D].山西: 山西大学, 2012 年 6 月

[8]董婧灵.基于 LDA 模型的文本聚类研究[D].武汉: 华中师范大学, 2012 年 5 月

[9]石晶, 范猛, 李万龙.基于 LDA 模型的主题分析[J].自动化学报, 2009, 35 (12), 1587-1592