电商产品评论数据情感分析

1）分析某一品牌热水器的用户情感倾向。

2）从评论文本中挖掘出该品牌热水器的优点与不足。

3）提炼不同品牌热水器的卖点。

分析方法与过程

1）利用爬虫工具——八爪鱼采集器，对京东商城进行热水器评论的数据采集。

2）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。

3）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。

4）从对应结果的分析中获取文本评论数据中有价值的内容。

评论数据采集

使用八爪鱼采集器。

评论预处理

- **文本去重**

1）原因

2）概述与缺陷

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，大都存在一些缺陷。

3）文本去重选用方法及原因

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，就只能对完全重复的语料下手。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。

从上述的总结我们可以知道，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，但是透过观察评论知道存在重复但是起码有 1 条评论有用的语料，而运用比较删除法显然只能定为留 1 条或者是全去除，因此只能设为留 1 条，以确保尽可能存留有用的文本评论信息。

- **机械压缩去词**

1）思想

2）语料结构

3）判断与规则

- **短句删除**

文本评论分词

- **情感倾向性模型**

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 0 0 …]，“复习”表示成 [0 0 0 0 0 0 0 0 1 0 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 维和 100 维。

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。

图 15-8 最下方的 $w_{t-n+1}, \cdots, w_{t-2}, w_{t-1}$ 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 $C(w_{t-n+1}), \cdots, C(w_{t-2}), C(w_{t-1})$ 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，直接使用 $d + Hx$ 计算得到。d 是一个偏置项。在此之后，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。矩阵 W（一个 $|V| \times (n-1)m$ 的矩阵），这个矩阵包含了从输入层到输出层的直连边。

（2）评论集子集的人工标注与映射

利用词向量构建的结果，再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 功能进行简单的机器标注，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。

（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，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 15-9 所示。

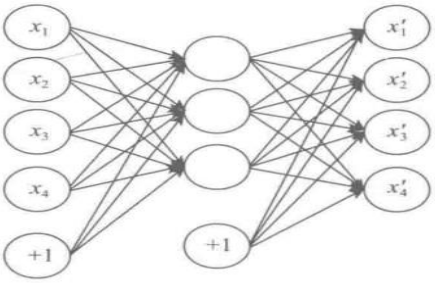

图 15-9　自编码网络结构示意图

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，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。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，将标注的评论划分为训练集和测试集，在 Python 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），对原始向量进行深度学习提取特征，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. 基于语义网络的评论分析

本节使用语义网络分析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。

这一部分主要通过对 3 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。

（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，举例如图 15-10 所示。

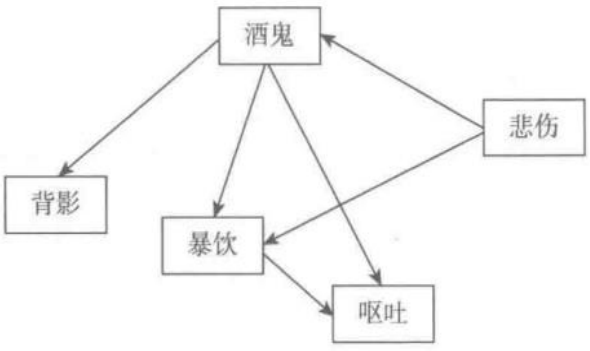

图 15-10　语义网络举例示意图

节点中的物体可以是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。例如，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。这些都是些复杂的关系。

虽然每一个语义网络结构中事物（节点）之间的关系是复杂的，但是从本质上看，语义网络的每一道弧的形成就是由于这种语义关系的存在。不同的用词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）基于语义网络进行评论分析的优势

从前面的论述中我们可知道，要想对中文的热水器评论进行合理的分析，必须要采取的一项措施就是分词，因为计算机不可能像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，因此对分词后的语句直接进行诸如产品差异等复杂的分析就不合实际，所以必须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。建立起事物之间（这里分出的每一个词料代表一项事物）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地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，例如，“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；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，但是如果“安装”和“方便”通过语义网络方式连接起来，如图 15-11 所示，就可以清晰地反映出是相关热水器产品在安装的时候比较便利。再如“热水”与“不足”也是这样的情况，此处就不再赘述。

图 15-11 “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，就可以借助它进行各种各样的特定的分析，特别是在判断特定产品优点、抽取各品牌的顾客关注点等方面具有一定的优势。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，通过可视化，就能够从中抽取出来。

（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就是在二分类文本情感分析的基础上进行增添，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，在于正面的以及负面的评论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关注的点是不完全一样的，信息也是不完全一样的，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，并将评论数据分割成正面（好评）、负面（差评）、中性（中评）3 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，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）），可用以检测论文是否抄袭；同时 ROST 系统又是一款大型的免费用于社会计算的软件，可以用来实现多种类型的分析，包括情感倾向性分析以及后面将要进行语义网络的构建等。之所以使用 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，目前来讲，其准确率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高，而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，受限于现今中文分词技术的缺陷以及评论本身的特性，能够通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更高。当需要以此为基础进一步分析的时候，就需要利用基于情感词典的方法。第二步的具体流程如下。

（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，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。

3. 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，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 3 种品牌评论中更多的信息。

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。

（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[24]。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 3 层贝叶斯概率模型，包含文档（d）、主题（z）和词（w）3 层结构，能够有效地对文本进行建模，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 L，一个 L 维向量 (1, 0, 0,…, 0, 0) 表示一个词。由 N 个词构成的评论记为 $d = (w_1, w_2, \cdots, w_N)$。假设某一商品的评论集 D 由 M 篇评论构成，记为 $D = (d_1, d_2, \cdots, d_M)$。$M$ 篇评论分布着 K 个主题，记为 $z_i (i = 1, 2, \cdots, K)$。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 模型图示如图 15-14 所示。

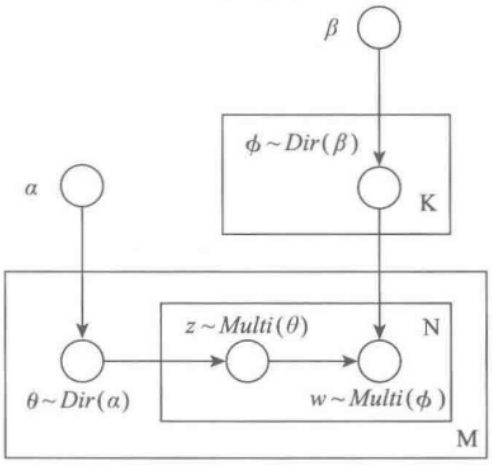

图 15-14　LDA 模型结构示意图

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：

$$Z|\theta = Multionomial(\theta)$$

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：

$$W|Z,\phi = Multinomial(\phi)$$

在评论 d_j 条件下生成词 w_i 的概率表示为：

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

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）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）[25] 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：

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

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，$n_{s,-j}$ 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 $\theta_{j,s}$，如下：

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

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

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、主题挖掘和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。

（3）运用 LDA 模型进行主题分析的实现过程

本文在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，每则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越可能成为热门关注点中的评论词。

（4）LDA 模型的实现

虽然 LDA 可以直接对文本做主题分析，但是文本的正面评价和负面评价混淆在一起，并且由于分词粒度的影响（否定词或程度词等），可能在一个主题下生成一些令人迷惑的词语。因此，将文本分为正面评价和负面评价两个文本，再分别进行 LDA 主题分析是一个比较好的主意。