基于数据挖掘技术的商品评价分析

阎宇航

（北京工商大学人工智能学院，北京　102488）

摘　要：数据挖掘一般是指从大量数据中自动搜索出具有特殊关系的信息的过程。通过挖掘顾客购买商品的评价数据，企业可以获得更多的信息，从而制定合理的销售策略，制造出成功的产品。首先，在剔除无用或错误的数据后，使用TF-IDF算法和余弦相似度算法对文本信息数据进行量化，并将所有文本评论转换为数字形式。其次，结合星级评分、评论评分和助益评分数据，建立线性回归模型，得出三者之间的关系。最后，使用权重公式和星级评分、评论评分和助益评分的数据，计算各产品在不同时间段的综合得分，用于定性分析。

关键词：数据挖掘；权重分配；线性回归；TF-IDF算法

在亚马逊创建的在线商城中，亚马逊为顾客提供了对其购买的商品进行评级和评价的机会，购买者可以使用1到5分的评分来表达对产品的满意程度，这些将作为其他客户购买产品的参考信息。阳光公司计划根据历史评分和评论在网上商城推出和销售三款新产品，以制定其在线销售策略，并确定潜在的重要设计功能，从而增强产品的有效性。

一、数据预处理

第一步：设置阈值K，如果级次概率大于阈值，则为单据中的关键字。第二步：使用TF-IDF算法中的IDF方法定义关键字的权重。第三步：领域词典的构建考虑了文本上下文的语义相关性。第四步：使用Word2Vec构建以下语义相关的关键字：组织语料库，将所有相关评论数据组成一个大的文本数据集，并记录为语料库；设置阈值K，将大于阈值K的所有关键词组成一个集合：

$$K_{corpus}=\{w_1,w_2,\cdots,w_n\}$$

使用Word2Vec的跳词模型对文本上下文进行索引，找出与K中义相关的词。最后，形成一组语义相关的词。

这些语义相关的词汇集，与先前的域M一一对应，以形成领域词典。

分别计算这些词的权值，形成向量空间模型，用余弦距离计算向量之间的相似度：

$$y=4.0281-0.0319x \qquad y=4.0281-0.0319.$$

对文本信息数据集进行预处理后，根据统计语言模型计算数据集的词频。将文本集中的所有词映射到k维向量，并根据余弦距离判断词之间的语义相似度。将评论分成句子，统计每个句子中包含的领域词典的词数。当相关领域词的语义相关词汇集中的一个或几个词出现在句子中时，将这些词的出现频率统一计算为相关领域词的出现频率。通过对评论句子的切分和基于领域词典的关键词频度统计，对评论进行量化，并计算文本数据集中每个字段的权重，权重值越高，该属性对注释就越重要。根据词汇量的不同，采用TF-IDF算法和余弦相似度算法将文本信息量化为数字形式，数值范围为1～3(1对应1分，3对应5分)。

二、模型建立

最小二乘法通过最小化误差平方和来寻找数据的最佳函数匹配。利用最小二乘法可以方便地获得未知数据，并使获得的数据与实际数据之间的误差平方和最小。使用最小二乘法进行曲线拟合，y=a+bx是线性回归方程。

三、模型求解

吹风机：对星级、评论、助益评级数据进行拟合，得到星级评分与帮助度评分之间的关系，可以表示为：y=4.028−0.0319x(置信区间在0.65以内)。其中，y为星级，x为帮助度评分。星级与评论之间的关系可以表示为：y=1.8302+0.9873x(置信区间在0.61以内)，其中y为星级，x为评论的星级。

微波炉：将评论和助益得分数据进行拟合，帮助度评分和评论星级之间的关系可以表示为：y=0.2658±0.0633x(置信区间在0.44以内)。其中，y为帮助度评分，x为评论的星级。

如果购买者不同，他们提供的评论的参考价值也不同。如果对没有购买的产品有评论，那么参考就没有意义了。因此，使用TF-IDF权重公式来计算每个案例的权重。在计算权重后，根据综合评价方法，将星级评分、量化评论等级和权重一起计算，最终得到产品的综合得分。（如图1、2所示）

图1　2011—2015年吹风机综合得分

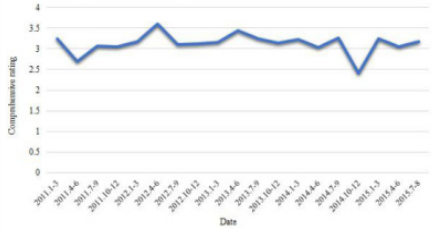

图2　2011—2015年微波炉综合得分

四、结语

吹风机的综合得分普遍持平，且略有下降趋势，说明顾客对吹风机的满意度较低。微波炉的综合得分虽然有一定波动，但相对稳定。假设公司有一批新产品的售后数据，上述模型可以先对文本进行量化，然后利用权重公式对星级和评论进行计算和分析，从而获得有用的信息。

参考文献：

[1]Liang Hao.The Impact of Iime Lapse in the Promotion Countdown on Consumer[D].Nanjing:Nanjing University,2017.

[2]Zhang Zhan.Research on the Impact of Reverse Negative Additional Comment Intervals on Consumers'Purchase Intention[D].Wuxi:Jiangnan University,2019.

作者简介：阎宇航(1999—　)，男，汉族，山西太原人。主要研究方向：人工智能。