分类号：TP311　　密　　级：公开
UDC：　　单位代码：10076

工程硕士学位论文

基于情感分析的商品评价系统设计与实现

作　者　姓　名：崔新宇
指　导　教　师：贾东立 副教授
企　业　导　师：任自贤 高级工程师
申请学位级别：工程硕士
学　科　专　业：计算机技术
所　在　单　位：信电学院
授予学位单位：河北工程大学

A Dissertation Submitted to
Hebei University of Engineering
For the Degree of Master of Engineering

Design and Implementation of Commodity Evaluation System Based on Emotional Analysis

Candidate : Cui Xinyu
Supervisor : Associate Prof. Jia Dongli
Pluralistic Supervisor : Senior Engineer Ren Zixian
Academic Degree Applied for : Master of Engineering
Specialty : Computer Technology
College/Department : College of Information and Electronic Engineering

Hebei University of Engineering

June, 2020

独创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的研究成果，也不包含为获得**河北工程大学**或其他教育机构的学位或证书而使用过的材料。对本文的研究做出重要贡献的个人和集体，均已在论文中作了明确的说明并表示了谢意。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：崔新宇　　签字日期：2020 年　6 月 6 日

学位论文版权使用授权书

本学位论文作者完全了解 **河北工程大学** 有关保留、使用学位论文的规定。特授权 **河北工程大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和电子文档。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：崔新宇　　签字日期：2020 年　6 月　6 日

导师签名：贾东立　　签字日期：2020 年　6 月　6 日

摘 要

网络购物在方便人们生活的同时，也存在着一些问题，由于网络购物不能像线下购物一样能够接触到商品，而且商品的所有信息都是由商家所给出的，这就造成了信息的不对等，容易导致用户买到了假货、残次品、或与自己期望不符的商品，造成一定的损失。虽然用户可以通过查看商品评论来获取有价值的信息，但随着商品评论的不断的累积，这种信息获取方式的效率越来越低。针对商品评论累积过多的问题，重点是从海量的商品评论中提取出有价值的信息，本文通过对现有情感分析方法进行研究，设计并实现了基于情感词典和基于 SVM 的商品评价系统，利用情感分析方法对商品评论进行分析处理，向用户提供客观且全面的购物参考信息。

本文主要工作体现在以下方面：

（1）数据的采集。本文通过对电商平台页面进行分析，设计了基于 Scrapy 框架的网络爬虫，该网络爬虫能够从电商平台上爬取相应商品的各项信息。本文利用该网络爬虫从电商平台累积收集了数十万条数据。

（2）情感分析方法的研究。为了从商品评论中挖掘出有价值的信息，本文对现有情感分析方法进行研究，采用了基于情感词典和基于 SVM 这两种情感分析方法对商品评论进行处理。在情感词典方法的研究中，本文以通用情感词典为基础利用 Word2vec 模型通过计算词语相似度的方法进行扩展，构建了专用情感词典，并设计了相应的评分规则。在 SVM 方法的研究中，本文首先采用词典与用户评分相结合的方法构建训练集，之后使用 Word2vec 模型生成特征词的词向量，并且为了在词向量中更好的表现出词语的重要程度，本文采用了词向量加权的方式对生成的词向量加以改进。

（3）商品评价系统的设计与实现。本文使用 Flask+vue 框架，采用前后端分离的开发方式，以情感分析方法为核心完成了商品评价系统的设计与实现。用户可以通过与 UI 界面进行交互来获取商品的情感分析结果、商品特征分析结果等各项信息，从而知晓该商品的整体评价情况以及商品特征的优缺点。为用户提供客观且全面的购物参考信息。

关键词：商品评价；情感分析；领域词典；SVM

Abstract

While online shopping is convenient for people's life, there are also some problems. Because online shopping can't touch the goods as offline shopping, and all the information of the goods is given by the merchants, this causes the information inequality, which easily leads to the users to buy fake goods, defective goods, or goods that don't meet their expectations, causing certain losses. Although users can get valuable information by viewing product reviews, with the continuous accumulation of product reviews, the efficiency of this way of information acquisition is getting lower and lower. Aiming at the problem of excessive accumulation of commodity reviews, the key point is to extract valuable information from the mass of commodity reviews. This paper designs and implements the commodity evaluation system based on emotion dictionary and SVM through the research of existing emotion analysis methods, analyzes and processes commodity reviews by using emotion analysis methods, and provides users with objective and comprehensive shopping reference information.

The main work of this paper is as follows:

(1) Data collection. Based on the analysis of e-commerce platform page, this paper designs a crawler based on scrapy framework, which can crawl the information of corresponding products from e-commerce platform. In this paper, hundreds of thousands of data are collected from e-commerce platform by using the crawler.

(2) The research of emotion analysis method. In order to extract valuable information from commodity reviews, this paper studies the existing affective analysis methods, and adopts two affective analysis methods based on affective dictionary and SVM to deal with commodity reviews. In the research of emotion dictionary method, based on the general emotion dictionary, this paper uses word2vec model to expand by calculating word similarity, constructs a special emotion dictionary, and designs the corresponding scoring rules. In the research of SVM method, this paper first uses the combination of dictionary and user score to build the training set, then uses word2vec model to generate the word vector of feature words, and in order to better show the importance of words in the word vector, this paper uses the way of word vector weighting to improve the generated word vector.

(3) The design and implementation of commodity evaluation system. In this paper,

using the framework of flask + Vue, adopting the development mode of front-end and back-end separation, the design and implementation of commodity evaluation system is completed with the emotional analysis method as the core. Users can get the results of emotional analysis, product characteristics analysis and other information through interaction with UI interface, so as to know the overall evaluation of the product and the advantages and disadvantages of the product characteristics. Provide users with objective and comprehensive shopping reference information.

Keywords: Commodity evaluation; emotion analysis; domain dictionary; SVM

目 录

第 1 章 绪论

1.1 研究背景及意义

在这个因特网爆炸式发展的时代里，互联网已经融入了人们的生活中。据 CNNIC 发表的第 44 次《中国互联网络发展状况统计报告》统计显示，截至 2019 年 6 月，我国网民规模已达 8.54 亿，手机网民规模达 8.47 亿。庞大的网民规模为我国电子商务的发展提供了良好的环境，目前，我国的电子商务总体发展水平已走在世界前列，网络购物用户的规模已达到 6.39 亿，网民使用率高达 74.8%。2018 年全国网上零售额达 90065 亿元，已经连续六年稳居世界第一。电子商务这个随着互联网发展而诞生的新兴产业，以其便捷、高效等特点迅速为用户所接受，网络购物逐渐成为了主流。随着网络购物的人数越来越多，如何在网络购物中购买到合适的商品也逐渐成为了人们关注的重点。

相较于传统的购物模式，网络购物不能像线下购物一样通过近距离接触商品来获取第一手的商品信息。在网络购物中，用户了解一件商品的途径大致有三种，商品的图片、商品的参数以及其他用户留下的商品评论。商品的图片虽然能够为用户提供直观的商品信息，但这些图片往往是经过商家美化后的宣传图片，可能会与商品的实际情况有所差别。商品的参数信息一般包含了商品的性能、技术指标、使用规则等关键信息。这些信息由于其具有一定的专业性，就会导致一些用户出现看不懂、理解错误等问题。而商品的评论中往往包含着用户的意见反馈、使用体验等重要的信息。所以查看评论就成了用户获取商品信息的重要途径，也正是由于商品评论能够为用户提供丰富且真实的商品信息，使得商品评论成为用户在进行购物前重要的参考信息来源。因此用户在进行网上购物时，可以通过充分查看商品的评论，来获取相应的商品信息，抵消不能直接接触商品的劣势，避免在购物中出现错误的决策，使消费购物更趋于合理化。

然而，随着时间的推移，商品的评论数量不断累积，一些热门商品的评论条数就多达数十万乃至数百万条。虽然评论的数量越多其所包含的商品信息就越详细，但用户无法短时间内阅读完全部评论来获取足够详细的商品信息，这就面临着信息过载问题。虽然各大电商平台推出了一些措施，来改善该问题，如根据用户为商品的打分情况将评论归纳分类为好评、中评和差评三类，用户可以根据自己的兴趣去查看相应的评论。虽然该类方法能在一等程度上缩减了用户在购物中的时间成本，但仍不能让用户在短时间内获取到直观且详细的商品信息。为了解

决由于商品评论过多造成的问题，重点是如何从海量的评论中挖掘出有价值的信息，这时就需要引入情感分析的方法来对商品的评价进行处理。

情感分析（Sentiment Analysis）又称意见挖掘或观点挖掘[1]。其目的是通过对带有情感色彩的文本数据进行分析，挖掘出该文本中蕴含的情绪、观点等有价值的信息。随着 web2.0 的兴起，人们越来越热衷于在网络上发表各样的言论，所以情感分技术也被广泛应用于对网络文本的分析中，挖掘其中有价值的信息。目前，根据该技术应用方向的不同，大致可分为 2 大类。第一类主要是对新闻评论文本的分析。政府通过新闻评论文本的分析，获取网络舆情的走向，及时了解群众对社会性事件的看法和观点，合理引导舆论导向，监督和约束网络行为，避免由于监管不力而导致出现群体性事件，进而引发难以估量的损失[2]。第二类主要是对商品评论文本进行分析。对于用户而言，通过分析评论信息，用户可以获取到商品的客观评估信息，为用户进行购物决策时提供客观且全面的参考信息，避免出现电商欺骗等问题。对于商家和厂家而言，通过情感分析技术，分析用户的意见反馈，获取其中的情感倾向，并对负面观点进行及时的改进，提升自己的行业竞争力。因此针对商品评论不断累加的问题，本文通过对情感分析方法的研究，设计并实现了基于情感分析的商品评价系统，以一种较为直观的方式为用户展现出商品的情感分析结果，为用户提供重要的参考消息。

1.2 国内外研究现状

1.2.1 文本情感分析

文本情感分析是自然语言处理（NLP）领域的热门研究方向之一，吸引了大量中外学者对其展开研究，目前已有不少学者在中英文的相关领域中取得了不错的成果[3,4,5]。通过对现有文献的研究，可以发现对于情感分析方法的研究大多集中在情感词典和机器学习这两类上。

基于机器学习的情感分析方法是利用训练集数据对机器学习中的分类模型进行训练，得到一个情感分类器，之后利用训练好的模型来对测试文本语料进行情感倾向的分类。常用的分类模型有支持向量机（SVM）、朴素贝叶斯法（NB）、最大熵模型（MaxEnt)）等[6]。

Pang 等人是最早将机器学习的方法引入到对电影的评论的情感分析中，分别采用支持向量机、朴素贝叶斯法和最大熵这三种分类器来对文本进行情感分类，通过实验数据表明，SVM 在文本情感分类中的取得了不错的效果[7]。Boiy 等人针对多语种多领域的 web 文本展开研究，根据不同的特征表示来对多种分类模型进行组合，结果证明朴素贝叶斯法适用于英文，其准确率高达 83%[8]。Kang 针

对使用朴素贝叶斯进行文本分类时会出现正面分类和负面分类的准确率相差过大问题，提出了一种改进的 NB 算法，并使用餐馆评价文本进行测试，结果表明，该算法能够有效缩小正负分类间准确率的差距，并且其精度高于 SVM 和传统贝叶斯[9]。Liu 通过优化、无标记数据选择和自适应特征扩展三个步骤将普通支持向量机转换成为一个自适应的分类模型，并通过实验证明，该模型的准确度比传统的分类模型提升不少[10]。Dasgupta 提出了一种半监督的机器学习模型，该模型不需要大量人工标注的训练集就能取得不错的分类效果[11]。Troussas 等人将朴素贝叶斯应用于在线文本情感分析中，并通过实验验证了该方法的可行性[12]。Yih 等人针对现有向量空间模型中正反义词的词向量相近的问题，将 LSA 模型引入到词向量的训练中，有效增加了正反义词词向量的距离[13]。Jain 等人为了挖掘 Twitter 上的公共意见信息，将机器学习的方法应用其中，取得了较好成果[14]。Rajput 等人将机器学习方法对股市评论者意见进行情感分析，并验证了该方法的有效性[15]。

中文领域中李婷婷等人提出一种基于 SVM 和 CRF 的情感分析方法，并通过研究不同特征组合情况下 SVM 和 CRF 的表现情况，得出了这两种模型的最佳特征组合方式，有效提高了这两种模型得准确率[16]。朱梦通过对朴素贝叶斯分类方法的研究，提出了一种改进型的 K-Bayes 算法，该算法是将特征词的分布情况与朴素贝叶斯相结合，进而突出了特征词的权重，提高了情感分析方法的表现效果[17]。刘勇等人针对传统随机森林算法在文本分类中出现的问题，提出了随机森林算法的优化方法，通过对算法的投票机制和超参数进行优化，提高了随机森林算法在文本分类中的性能 [18]。陈强等人使用句法分析模型对文本语料进行处理，将处理好的语料作为训练集用于训练模型，实验表明，该方法训练出来的分类模型的准确率有所提升[19]。

张林等人为了提高机器学习在短文本中的表现，提出了一种短评论特征共现的特征筛选方法，并通过实验验证了该方法的有效性[20]。陈叶旺等人针对网络中存在的非结构化形式的文本数据，提出了一种改进的朴素贝叶斯分类方法，实验证明该方法不仅易于建立，而且其分类的准确率也得到提升[21]。李琼等人针对传统支持向量机存在的训练和测试时间过长的问题，提出了一种改进型的支持向量机多类分类方法，降低了训练和测试过程中的时间消耗，且在一定程度上提高了多类文本分类的识别准确率[22]。

经过大量学者的研究，基于机器学习的情感分析方法已经得到了很大的改进。但该方法仍有诸多不足之处。一是在对分类模型进行训练时仍需要大量已标注的语料文本作为训练集对其进行训练，并且其分类的效果取决于训练集的质量。对于语料集的标注需要拥有专业知识的人员进行人工标注，既费时又费力，二是机

器学习的在迁移到其他领域时表现效果较差，在原领域内往往能够取得不错的分类效果，但当对其他领域的文本进行分类时就不能达到理想的效果。

基于情感词典的方法主要依赖于词典和规则，通过情感词典来识别文本中的情感词，之后结合评分规则来分析该文本的情感倾向。基于情感词典方法的优点是不需要耗费大量人工成本去构建高质量的训练集，其分类的准确率依赖于情感词典的完备程度，情感词典的覆盖程度越高，情感分析的结果也就越精确。

英文领域内，较为出名的通用情感词典是由 Esuli 等人在 WordNet 语义词典的基础上构建的 SentiWordNet 通用情感词典，该通用情感词典以其较高的词语收录率在情感分析中被广泛应用[23]。

Fengs 等人针对文本中存在的隐晦词，使用 PageRank 和 HITS 来对其进行挖掘，并通过实验证明其挖掘出来的隐晦词对情感分析极具价值[24]。Nakagawa 等人针对主观性句子中可能包含着逆转其他词情感极性的词汇的问题，使用了句法依存分析结合条件随机场的方式对语句进行分析，根据实验表明，该方法的准确率要比仅使用词典分析时更高[25]。Kcuc 等人将情感分析技术应用于 Twitter 的文本研究中，并提出了机器学习和情感词典相结合的情感分类器，并通过实验验证了该方法的有效性[26]。Zagibalov 等人针对点互信息方法难以获得大规模的情感词信息的问题，引入迭代机制来提高情感分析的准确度[27]。Mudinas 提出了 psenti 情感分析系统，通过将词典方法和机器学习方法进行融合，提升了情感极性分类和情感强度检测方面的准确率[28]。Whitelaw 等人提出了一种细粒度的情感分析方法，通过情感词典与词袋模型的结合，提高了情感分析的准确率 [29]。Gyamfi 等人利用 Wordnet 情感词典对构建的基础情感词典进行扩展，提高了情感词典的完备程度[30]。

在中文领域中，使用比较广泛的通用词典为台湾大学的 NTUSD 简体中文情感极性词典、知网（HowNet）情感分析用词语集、大连理工大学情感词汇本体库等。

黄仁等人使用 word2vec 通过计算语义相似度的方法扩展情感词典，实验结果显示，经该方法扩充的情感词典在文本分类中取得了不错的效果[31]。王志涛等人针对微博文本的特性，提出了一种基于情感词典和规则集的情感计算方式，并利用采集的微博数据对该方法进行验证，证明了该方法的有效性[32]。原多多利用近义词和规则的扩充方法对情感词典进行扩展，并通过对比试验证明了该词典扩展方法的有效性[33]。刘亚桥等人提出了一种基于改进 word2vec 的情感词典扩展方法，并通过实验证明了该方法在摄影领域文本中取得了不错的效果[34]。周莉等人针对突发事件领域内通用情感词典表现不佳的问题，采用机器学习和人工构建相结合的方法对通用情感词典进行扩展，取得了不错的效果[35]。杨奎等人通过计

算词汇相似度的方法对 HowNet 情感词典进行扩充，并通过研究情感词对文本情感影的响程度设计了情感得分策略，提高了情感分类的准确率[36]。蒋翠清等人利用 Word2vec 与 PMI（点互信息）相结合的情感词典扩展方法构建了社交媒体领域的情感词典，并通过实验证明了该方法的可行性[37]。杨小平等人利用 Word2vec 对知网情感词典、大连理工大学情感词典等通用词典进行筛选，构建出了 SentiRuc 词典，并在通用领域数据集上取得了不错的实验结果[38]。

虽然众多学者对基于情感词典的方法提出许多改进策略，但仍有不足之处。其一是情感词典的构建仍需要较高的成本，由于中文语言的复杂性，构建一个通用情感词典往往需要语言学的专家花费数年的时间才能完成。二是通用情感词典的覆盖度问题。任何情感词典都不能做到收录所有的词语，而且网络上每天都会产生新的词汇，因此在对情感词典的研究时，结合相关领域对通用情感词典进行扩展，才能保证情感词典在进行情感分析时达到良好的效果。

1.2.2 商品评价情感分析

随着网络购物产业的不断发展，越来越多的学者对商品评论展开分析，将情感分析方法引入到商品评论的分析中。目前可根据使用方法的不同将对商品评价的研究工作分为两类分别为是使用基于机器学习的方法以及基于情感词典的方法。

在使用基于机器学习的商品评价分析方面。李明等人提出了一种基于 SVM（支持向量机）和 PMI（点互信息）相结合的商品评论分析方法，对商品评论进行细粒度分析，取得了不错的效果[39]。叶锦等人通过对商品评论中配图的研究，发现这些配图往往包含着用户最直观的情感信息，因此以图片和文字的角度对商品评论进行研究，构建了包含有图片和评论文本的数据集，并利用一种深度塔克融合的方法对数据集中的文本特征和图像特征进行融合，在测试中验证了该方法的有效性[40]。

为了解决传统机器学习在进行分析时仍需大量标注的商品评价语料问题，一些学者将深度学习引入到商品评价的分析中。於雯等人将 Dropout 算法融入到 LSTM 模型中，在对商品评论的情感分类中取得了 99%以上的准确率[41]。刘智鹏将 CNN 模型和 RNN 模型相结合，提出了 Shunt-C&RNN 文本评价情感分类模型，该模型通过制定分流规则，实现了用 CNN 模型处理评论中的短语句，用 RNN 模型处理长语句，发挥了 CNN 模型和 RNN 模型各自的优势[42]。张佳悦在进行商品评论的情感分析时使用了 Word2vec 结合 LSTM 的方法，该方法是使用 Word2vec 对文本进行向量化，然后输入到已经训练好的 LSTM 模型的嵌入层，对 LSTM 模型再次进行训练，经过该方法训练的模型准确率提升至 89%[43]。

在使用基于情感词典的商品评价分析方面。冯仓龙等人从商品细粒度方面入手对商品评论展开研究，在商品特征与情感词的抽取过程中，通过将句法关系信息、情感要素信息以及聚类代码信息融入到随机场模型中，提高了对目标特征提取的准确率[44]。王名扬等人利用词语的卡方值与累积覆盖度相结合的方法对基础情感词典进行扩展，提高了基础情感词典的完备程度[45]。常丹等人为了提升情感词典的完备程度，采用词频共现的方法对情感词典进行扩充，并通过实验验证了该方法的有效性[46]。

在商品的评价的研究领域中，众多学者针对特定领域的商品的评价来对情感分析的方法进行了有针对性的改进，取得了不错效果，也为本文提供了思路。本文采用机器学习中的 SVM 方法和情感词典方法来对商品的评论信息进行情感分析，为用户提供更为直观和准确的情感分析结果。

1.3 本文主要研究内容

本文以商品评论作为研究重点，分别使用基于 SVM 的情感分析方法和基于情感词典的情感分析方法对商品评论进行分析，分析出隐藏在评论中的用户情感倾向。论文的研究内容主要是评论数据的采集、基于 SVM 和基于情感词典这两种情感分析方法的研究、商品评价系统的设计与实现。

（1）商品的各项参数信息和商品的评论信息的获取。本文通过对电商网站的页面进行分析，设计并实现了基于 Scrapy 框架的网络爬虫，并利用该网络爬虫收集了大量数据，为商品评价系统提供了数据基础。

（2）对基于情感词典和基于 SVM 的情感分析方法进行研究。本文通过对目前情感分析方法的研究，选定基于情感词典和基于 SVM 的方法作为本文的情感分析方法。在基于情感词典的方法研究中，本文以知网情感词典为基础，利用 Word2vec 模型通过计算词语相似度的方法对基础情感词典进行扩充，构建了专用情感词典，增加了情感词典的完备程度。此外，为了提高情感词典的准确率，本文从评论文本的语法规则方面入手，制定了情感词典方法的评分规则，在一定程度上提升了情感词典方法的准确率。

在基于 SVM 的情感分析方法研究中，本文采用情感词典与用户评分相结合的方法来构造 SVM 的训练集数据。在特征选择方法中，本文使用了 TF-IDF 的特征选择方法对商品评论文本进行处理，筛选出了对商品评论贡献较大的特征词。在文本向量化处理过程中，本文使用了 Word2vec 模型来对特征词进行向量化处理。此外，为了能够使词向量矩阵能够更好的表示出词语的重要程度，本文利用词语的 TF-IDF 值来对词向量进行加权处理。

（3）完成商品评价系统的设计与实现。本文使用 Flask+vue 框架，采用前后

端分离的开发方式，完成了商品评价系统的开发工作，为用户提供客观且全面的商品整体评价情况和商品特征分析情况。

1.4 论文的组织结构

本文一共分为五个章节，每章节的主要内容如下：

第一章为绪论部分，主要介绍了本文的研究背景和意义，国内外的研究现状，主要研究内容以及论文的组织结构。

第二章为相关技术部分，主要介绍了本文使用的相关算法和技术，包括网络爬虫的技术、文本预处理技术、情感分析方法以及特征选择和 word2vec 模型。

第三章为数据的采集部分，主要介绍了基于 Scrapy 框架的网络爬虫的设计与实现。

第四章为情感分析方法的研究，主要介绍了基于情感词典的情感分析方法与基于 SVM 方法的研究与实现。

第五章为系统的设计与实现，主要对商品评价系统进行设计与实现。用户可通过与 UI 界面的交互来获取自己感兴趣的信息。

第 2 章 相关理论和方法介绍

2.1 网络爬虫

评论数据的获取是整个评价系统的基础，如果仅通过人工的方式对评论数据进行收集，不仅效率低下又会浪费大量的时间和精力，因此本文利用网络爬虫技术从电商网站上大量收集所需的商品评论数据，下面将简单的介绍网络爬虫的相关知识。

网络爬虫，又称网络蜘蛛[47]，它是一段按照特定规则自动提取特定网页内容的程序，主要用于收集网络上的各种资源。网络爬虫的一般流程为：

（1）给出初始 URL 地址，放入 URL 等待队列中；

（2）从队列中取出 URL 地址，进行读取，DNS 解析，下载对应 URL 的网页并储存到数据库中；

（3）放入新的 URL 地址，重复之前的操作，直至爬取完所有的 URL 地址，爬虫停止进行网页爬取；

（4）将存入到数据库的网页数据进行分析、过滤、清洗等操作。目前常用的网络爬虫的框架有许多，如 Scrapy、Beatiful Soup、selenium 等。网络爬虫的框架图如图 2-1 所示。

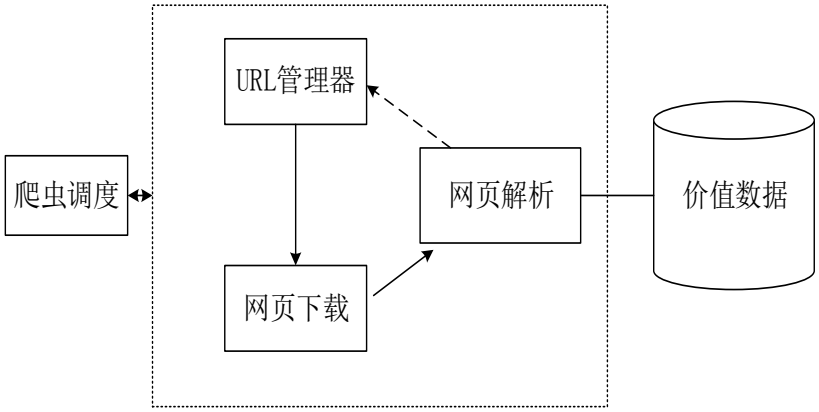

图 2-1 爬虫框架图

Fig.2-1 Crawler Framework

2.2 文本预处理

在进行情感分析之前，需要先对文本进行预处理操作，将文本中半结构化和非结构化的信息转换为结构化信息，以便于计算机进行识别和处理。文本预处理

的一般流程包括了中文分词、词性标注以及停用词处理。

2.2.1 中文分词

在自然语言处理中，词是最小的、且具有意义的语言成分。中文分词的目的就是将中文语句按照既定的逻辑规则拆分成符合语言实际且具有意义的词语，它是自然语言处理中研究重点之一。在文本预处理步骤中最重要的就是对评论文本进行分词处理，分词的质量会直接影响到之后情感分析的结果。目前常用的分词方法有三类，分别为基于词典的方法、基于理解的方法以及基于统计的方法[48]。以下将对这三种分词方法进行介绍。

（1）基于词典的分词方法

该方法是按照一定的策略将待处理的中文语句与词典进行匹配，当发现与词典中登录的词语一致时，匹配成功，进行切割。该方法使用广泛、易于实现、耗时短、切分速度快。但这种方法对词典的依赖程度较高，对于新词的发现能力很差，当出现词典中未登录的词语时，就会降低分词的质量。此外这种分词方法对于歧义的识别能力也表现不佳，不能良好的应对歧义问题。目前分词系统一般将该方法用于对文本的初分，之后利用其它手段提升分词的质量。

（2）基于理解的分词方法

该方法又称人工智能分词，其方法是将句法分析和语义分析融入到分词中，利用句法和语义信息来进行对句子的分割，以期达到最符合语句原意的分词结果。该方法能够有效的解决未登录词的问题和歧义问题，同时也是 3 种分词方法中分词效果最佳的一种方法。但由于中文本身复杂的句法和语言知识，难以将复杂的中文语句组织成计算机直接可读取的形式，因此基于理解的分词方法目前仍处于探索阶段。

（3）基于统计的分词方法

该方法是通过对相邻共现汉字的组合频率进行统计，一般认为共现频率较高的相邻汉字成词的概率也较高。该方法常用的分词模型有 N—Gram 元分词模型、最大熵分词模型、最大概率分词模型等。该方法不需要依靠词典，在新词的识别和消除歧义方面表现不错。但该方法需要大量的训练语料文本用于训练模型，而且该方法有时抽取出一些共生频率高，但并不是词语的组合。

由于这三种方法各有优点和缺点，因此分词系统通常将这些方法综合起来进行分析，提升分词的准确率。目前国内较为完善的分词系统有哈工大云 LTP 平台、NLPIR 汉语分词系统、庖丁解牛分词器、结巴（jieba）中文分词等。

（4）结巴分词工具介绍

结巴（jieba）分词工具是一款使用广泛的中文分词工具，该工具是由国内的

研发人员针对中文领域的分词处理而开发的，该分词工具为了保证分词的精度，以人民日报等权威语料为基础，整理得出了 dirct.txt 词典，该词典包含有 2W 多条词汇。结巴分词工具的分词模式有 3 种：

①全模式：该模式是通过对文本进行分析，扫描出句子中所有能够成词的词语加以切分。在该模式下的分词速度较快，但是不能解决词语的歧义问题。如句子“送的全面膜不错”在全模式下就会被切分成“送/的/全面/面膜/不错”。

②精确模式：该模式是通过分析句子中成词概率最高的词语来对句子进行精确切分，该方法能在一定程度上解决词语的歧义问题。

③搜索引擎模式：该模式是以精确模式为基础，将精确模式切分下的长词再次进行切分，提高分词的召回率。该模式常用于搜索引擎上，用以降低搜索的难度。

2.2.2 词性标注

在完成分词之后就可以进行词性标注操作，词性标注就是通过判定每个词的语法范畴，为每一个词标注上词性，从而确定该词是属于动词、名词、形容词或其他词性的过程，部分中文词性对照表如表 2-1 所示。词性标注也是自然语言处理领域的基础步骤。目前，主流的词性标注方法有两种：基于规则的词性标注方法和基于统计的词性标注方法[49]。

表 2-1 部分中文词性对照

Table 2-1 part of Chinese part of speech comparison

符号	词性	符号	词性
a	形容词	d	副词
ad	副形词	c	连词
m	数词	e	叹词
n	名词	ng	名词素
nr	人名	ns	地名
r	代词	v	动词
w	标点符号	y	语气词

（1）基于规则的词性标注方法

该方法是利用人工制定的规则来进行词性标注，首先使用词典对语料进行标注，给出所有的可能词性，之后结合上下文信息，使用规则集进行筛选，最后得到该词的最终词性。这种方法对于规则库的要求较高，需要专业人员消耗大量时间和精力进行规则的制定，这就导致了这用方法实用水平较低。

（2）基于统计的词性标注方法

该方法通过分析输入的语料，给出语料词性的所有可能性，通过概率统计的方式对语料标注合理的词性。该方法在新词识别和消除歧义方面表现不错，是自

然语言处理领域中使用最为广泛的词性标注方式。

2.2.3 停用词

停用词通常是指在在文章中出现频率很高，但是在文章中没有实际意义，仅起到结构作用的词语，这类词大多为介词、副词等。在情感分析过程中，为了降低文本的维度，提升文本处理效率，一般会将停用词进行排除。停用词去除的方法一般以现有停用词表为基础，结合自己的任务需求来进行扩展，之后对经过分词后的数据对照停用词表进行匹配，从而将停用词排除。常用的停用词表有哈尔滨工业大学停用词表、四川大学停用词表以及百度停用词表等。

2.3 情感分析

情感分析可根据分析文本的粒度不同分为词语级、句子级和篇章级，其中词语级的粒度最小，由于商品评论多是一些句子，因此对于商品评论的情感分析应从句子级出发。目前对于句子级的情感分析方法可分为 2 类，一是基于机器学习的方法，二是基于词典的方法。

机器学习的方法就是利用训练集对分类模型进行训练，从而得到一个能够自动对文本的情感倾向进行分类的分类器。基于机器学习的情感分析方法可分成 3 类：有监督的机器学习、半监督的机器学习以及无监督的机器学习。有监督的分类学习是利用大量人工标注的语料作为训练集用于训练模型，对于该方法的研究已经取得了不少成果，情感分类的效果较为理想。但由于有监督的机器学习需要大量高质量人工标注的样本数据才能取得较好的分类结果，而高质量的样本数据又需要拥有专业知识的人去进行标记，既费时又费事。半监督的分类学习是一种较为折中的方法，该方法使用较少的标记语料结合大量未标记语料为训练集去训练分类模型，该方法既降低了对训练集质量的要求，又避免出现因没有足够标记样本带来的分类效果的降低。该方法对于提升机器学习在情感分析中的性能方面有很大的使用价值。无监督分类学习方法是使用无标记的语料数据作为训练集进行学习，进而去分析待处理的文本数据。

目前，较为经典的机器学习分类模型有最大熵、支持向量机、朴素贝叶斯等，这些分类模型大都需要高质量的人工标注的样本数据。

基于机器学习进行情感分析一般流程可分为 3 阶段，第一阶段为先对获取到的训练文本进行预处理，之后对训练文本进行特征选取和提取，然后使用训练集对分类模型进行训练，得到情感分类分类器。在得到训练完成的分类器后就可进入第二阶段，第二阶段对测试数据进行处理后，使用训练好的分类模型对其进行

情感分析，对分类模型进行性能评估。第三阶段就可利用得到的情感分析模型对待处理的文本进行情感分析。

基于词典和句法分析的情感分析方法注重词典的构建和规则的制定。该方法通过拆解文本、进行句法分析以及情感值计算的方式完成对文本的情感分析，该方法是根据文本的情感得分情况来判别文本的情感倾向。该方法分类的准确性依赖于规则的制定和情感词典的完备程度，因此目前对于该方法的研究大多集中在对于词典的扩充以及规则的制定方面，如吴潇等人以知网情感词典为基础，通过计算词语的 TF-IDF 的方式对基础情感词典进行扩充，构建出领域情感词典，并通过实验的方式证明了该方法建立起的情感词典在情感分析中的表现优于普通情感词典[50]。

由于情感词典的构建成本较高，往往需要具备相关知识的专家进行人工标注，既费时又费事，因此大多数的学者在研究基于情感词典的情感分析时，通常都是以通用情感词典为基础，通过扩充的方式构建出相应的情感词典。目前国内较为通用的情感词典有台湾大学的 NTUSD 简体中文情感极性词典、知网（HowNet）情感分析用词语集、大连理工大学情感词汇本体库。下面将对这些通用情感词典进行介绍：

（1）知网（HowNet）情感词典

知网情感词典是由董振东等人花费数年时间构建的情感分析用语集，该词典共包括 17877 个词汇，其中不仅包含了中英文正面和负面的情感词，也对包含了评价词、程度词、主张词等 12 个部分的词汇。

（2）大连理工大学情感词汇本体库（DUTIR）

该词典是由大连理工的徐琳宏等人构造的，共计整理标记了 27466 个情感词语，不仅包含了大量积极和消极情感词，并对每个词语的情感强度和极性进行了标记，并对情感词进行了进一步的划分，共分成 7 大类 21 个小类。

（3）台湾大学的 NTUSD 简体中文情感极性词典

该词典是由台湾大学自然语言处理实验室整理发布的，该词典分为简体中文和繁体中文两个版本，共包含了 2810 个正面词汇和 8325 个负面词汇。表 2-2 展示了通用情感词典包含情感词的数量。

表 2-2 通用情感词典情感词数量

Table 2-2 Number of Emotional Words in General Emotional Dictionary

情感词典	正面词语个数	反面词语个数	情感词总数
HowNet	4320	4528	8848
DUTIR	10783	11229	22012
NTUSD	2810	8276	11086

2.4 特征选择

目前常用的特征选择方法如下所示。

（1）信息增益（IG）

信息增益是通过计算文本中包含特征项t_i的信息熵与文本中不包含特征项t_i的条件熵之间的差值来衡量该特征所携带的信息量，信息增益越大时，该特征对于文本来说也就越重要。信息增益的表达式如公式（2-1）所示。

$$IG(t_i)=-\sum_{j=1}^{M}p(C_j)\times\log p(C_j)+p(t_i)\times\sum_{j=1}^{M}p(C_j\mid t_i)\times\log p(C_j\mid t_i)+ p(\bar{t_i})\times\sum_{j=1}^{M}p(C_j\mid \bar{t_i})\times\log p(C_j\mid \bar{t_i}) \tag{2-1}$$

式中，$p(C_j)$——C_j类别的文本在文本数据集中出现的概率

$p(t_i)$——特征项t_i在文本数据集中出现的概率

$p(C_j\mid t_i)$——数据集中包含特征项t_i且属于C_j类别的概率

$p(\bar{t_i})$——数据集中不含有特征项t_i的概率

$p(C_j\mid \bar{t_i})$——数据集中不包含特征项t_i但属于C_j类别的概率

M——文档的类别数

（2）TF-IDF

词频逆文档频率是一种常用的特征选择方法，该方法是以数学统计学为思想来对文本进行特征选择的，词频逆文档频率由两部分组成，词频（TF）与逆文档频率（IDF），词频和逆文档频率的计算公式分别为公式（2-2）和公式（2-3）。

$$tf_i=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2-2}$$

式中，tf_i——词语t_i的词频（TF）

$n_{i,j}$——词语t_i在文档d_j出现的次数

$\sum_k n_{k,j}$——文档d_j的词语总数

词频代表着一个单词对与该文档的重要程度，当一个词语多次出现在一个文档中时，可以认定该词对该文档很重要。

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|+1} \tag{2-3}$$

式中，idf_i——词语t_i的逆文档频率

$|D|$——文档的总数

$|\{j:t_i\in d_j\}|$——包含词语t_i的文档总数

逆文档词频代表着词语t_i的类别区分能力，IDF 越大，该词语的类别区分能力越强。考虑到词语不存在文档中会导致分母为 0 的情况，本文在其分母上加上 1。最终 TF-IDF 的表达式为公式（2-4）。

$$tidf_{ij} = tf_i \times ifd_i \tag{2-4}$$

式中，tf_i——词语t_i的词频

ifd_i——词语t_i的逆文档频率

为了进一步提升 TF-IDF 的准确率，本文采用了目前常用的一种改进方式，其改进型表达式如公式（2-5）所示。

$$tidf_{ij} = \frac{tf_i \times ifd_i}{\sqrt{\sum_{p=1}^{k}(tf_i \times ifd_i)^2}} \tag{2-5}$$

式中，k——词语t_i的在文档d_j的个数

（3）互信息（MI）

互信息是通过计算特征与类别之间的关联性来进行特征选择，当互信息越大时，该特征与该类别之间的联系也就越大，也就意味着该特征对于该类别也就越重要。互信息的表达式为公式（2-6）。

$$I(t,c) = \log \frac{p(t,c)}{p(t) \times p(c)} \tag{2-6}$$

式中，t——文本的特征项

c——文本的类别

$p(t)$——特征项t在文本数据集中出现的概率

$p(c)$——类别C的文本在文本数据集中出现的概率

$p(t,c)$——特征项t属于类别C的概率

$I(t,c)$——特征项t与类别C之间的互信息值

2.5 Word2vec 介绍

Word2vec 与 glove、ELMo 等模型是目前 NLP 领域内较为经典的几个语言模型方法，Word2vec 最初于 2013 年由谷歌公司发布的一个用于训练词向量的语言模型，该模型是基于神经网络模型实现的，它可以通过训练将文本中的词语转化为 N 维的向量来对词语进行表示，对比传统的 one-hot 词向量表示模型，该方法产生的词向量可以有效地避免了维度爆炸等问题。由于该模该方法能够快速准确的训练词向量，并且特别适用于处理大规模，乃至超大规模的语料文本，因此 Word2vec 经常被用于进行情感分析中进行词向量的训练工作。

Word2vec 中常用的模型有两个，分别为 Skip-Gram 模型和 CBOW 模型。这

两个模型的结构如图所示 2-2。从图中可以看出 Skip-Gram 模型和 CBOW 模型都为 3 层结构，分别为输入层（Input）、投影层（Projection）以及输出层（Output），其中 Skip-Gram 模型是利用当前词 w_t 去预测该词的上下文 w_{t-2}、w_{t-1}、w_{t+1}、w_{t+2}，而 CBOW 模型则与之相反，该模型是利用上下文去预测当前词 w_t。

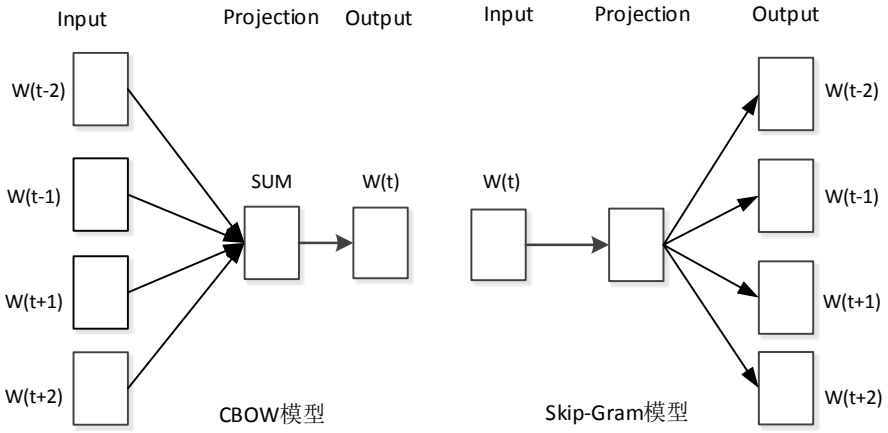

图 2-2 Word2vec 模型

Fig.2-2 Word2vec Model

2.6 本章小结

本章主要介绍实现商品评价系统所需要的相关算法和技术，首先介绍了网络爬虫的原理和流程，之后针对情感分析中的文本预处技术进行介绍，包括去停用词、词性标注以及中文分词。然后本文介绍了情感分析的方法，包括基于机器学习的方法和基于情感词典的方法，并对基于情感词典方法中的常用情感词典进行了介绍，最后本文介绍了特征选择的几种常用的方法和 Word2vec 模型。

第 3 章 基于 Scrapy 框架的数据采集

商品评价系统的目的是利用情感分析方法对商品评价进行分析和处理，而情感分析方法的研究又需要以大量的评论数据为基础，当获取的评论数据不足时就可能导致情感分析的效果不佳，影响整个评价系统的性能，所以数据的采集工作是商品评价系统重要的组成部分也是整个系统的基石所在。目前获取商品评价的方法有基于 API 的方法、基于数据集的方法以及基于网络爬虫的方法，其中基于 API 的方法虽然简单容易实现，而且大多数网站也开放了自己的 API 接口，但是该方法需要获取到官方的授权才能获取到相应的权限来进行对应的操作。

基于数据集的方法虽然能够将分享的数据集直接用于分析工作，但由于数据集的质量与其制作者有关，这就导致了数据集的质量参差不齐，可能影响到分析的效果，而且数据集不能及时更新，导致无法保证数据的时效性。

基于网络爬虫的方法是目前最为常用的方法，该方法可以根据使用者的需求来制定相应的爬虫来获取对应的数据，方便灵活。因此本系统使用基于 Scrapy 框架的爬虫来进行数据的爬取工作。

3.1 Scrapy 框架介绍

Scrapy 框架是一个基于 python 语言的爬虫框架，它能够对 web 资源实现高层次、快速的抓取，经常应用于各类网站的抓取工作，提取其中有价值的结构化数据，如信息处理、数据挖掘等工作。Scrapy 最吸引人的地方在于其采用了框架结构设计，使用者可以在根据自己的需求在该框架内通过开发其中的几个模块就可实现爬虫功能，用于抓取所需内容。此外 Scrapy 框架不仅能够通过抓取网页来获取数据，也可以通过访问 API 接口获取对应到的数据。Scrapy 框架以其简单、高效、易于扩展的特性被频繁应用于各个领域的数据采集中。

3.1.1 Scrapy 框架组件

Scrapy 框架如图 3-1 所示，它通常由引擎（Scrapy Engine）、调度器（Scheduler）、爬虫（Spider）、下载器（Downloader）、管道（Item Pipeline）、下载中间件（Downloader Middlewares）、Spider 中间件（Spider Middlewares）组成。其各个组件的功能如下所示。

（1）引擎：用于整个框架的调度，管理爬虫、管道、下载器与调度器之间

数据的传递。

（2）调度器：用于接收由引擎传递过来的请求，按照一定的规则进行储存，并在引擎需要时，将储存的请求发送给引擎。

（3）下载器：用于在引擎发送访问请求后将网站的返回的响应数据进行打包下载，并经由引擎交给爬虫进行解析。

（4）爬虫：用于对网站的响应数据进行解析，获取用户想得到的 item 类的数据，如果解析出 URL 则经由引擎存入调度器。

（5）管道：用于处理由爬虫解析出来的 item 数据，并对这些数据进行清理、验证等操作，如数据去重、数据储存等。

（6）下载中间件：用于自定义扩展下载功能的组件，介于引擎与下载器之间的组件。

（7）Spider 中间件：用于自定义扩展引擎与爬虫中间的通信的组件，介于爬虫和引擎之间的组件。

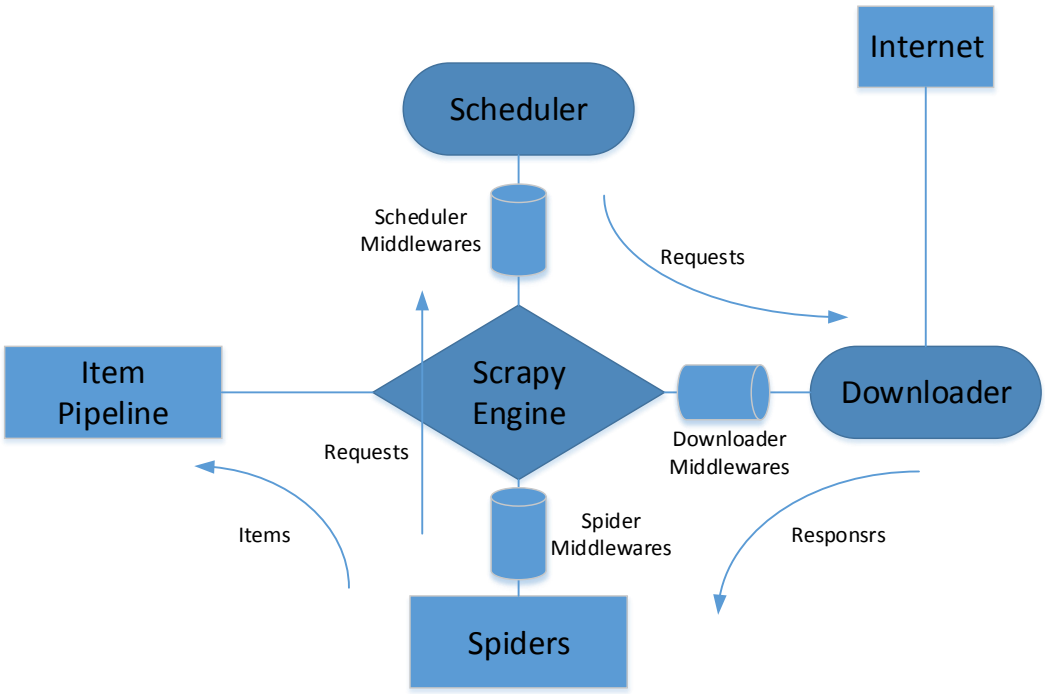

图 3-1 Scrapy 框架

Fig.3-1 Scrapy frame

3.1.2 Scrapy 工作流程

Scrapy 的框架运行大致流程如下所示。

（1）爬虫将设置好的初始 Request 发送给引擎；

（2）引擎将 Request 交给调度器，进行入队操作；

（3）调度器将队列中的 Request 经由引擎交给下载器；

（4）下载器向目标网站发送 Request，并将的网站的 Response 下载，之后经由引擎交给爬虫；

（5）爬虫对响应数据进行分析，解析出其中的 Item 或 Request，然后发送到引擎；

（6）引擎根据返回的数据，将 Item 发送给管道，将 Request 发送给调度器进行入队；

（7）如果调度器的队列中无请求则结束，有请求则重复之前的操作。

3.2 数据源的选择

目前，针对商品评价的研究，网络上暂时没有提供合适的数据集，主要是通过对各大电商网站进行数据的收集工作。目前主流的电商网站有京东商城、淘宝网、苏宁易购、当当网、卓越亚马逊等网站。考虑到本文主要研究的目标是针对电子商品的评论进行研究，并且为了保证数据来源的可靠性，本文选取了京东和淘宝作为本文的数据源。

3.3 爬取策略设计

在上一小节中，确定了本文的数据源为京东、淘宝等电商平台，在本节中将制定具体的评论爬取策略。本文需要获取的数据是商品的各项信息（商品的 ID、价格、品牌、产品参数等）以及对应的评论信息。因此可以将爬取过程分为两部分，商品信息爬取和商品评论爬取。对于京东商城的商品信息爬取策略，可以从商品的品牌入手，通过对京东某一类商品的品牌分类页面进行分析，获取该类商品的所有品牌的 URL，之后利用各个品牌的 URL 进入到该品牌下所有商品页面，然后获取到该品牌下的所有商品的 URL，然后通过这些 URL 进入到商品的详细页面，之后获取到该商品的各项信息。商品信息的爬取策略如图 3-2 所示。

对于淘宝商城的商品信息爬取策略则可以从搜索页面入手，通过对搜索页面进行分析来获取商品详细页面的 URL 和对应的商品的 ID 信息，然后根据商品详细页面的 URL 来进入到该商品的详细页面，之后通过分析页面获取到该商品的各项信息。

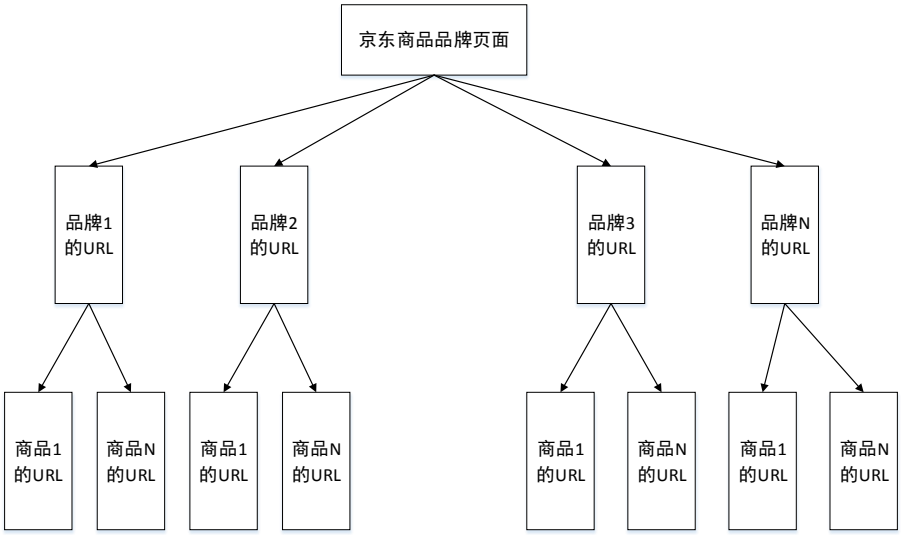

图 3-2 京东商城商品信息爬取策略

Fig.3-2 Crawling strategy of commodity information

由于京东商城和淘宝商城中商品的详细页是动态加载的，无法从页面中直接解析出商品的所有评论，需要对页面进行分析获取用于保存商品评论的 URL，然后根据商品的 ID 来构造每个商品对应的评论信息的 URL，之后利用爬虫进行商品评论的爬取，商品评论的爬取流程如图 3-3 所示。

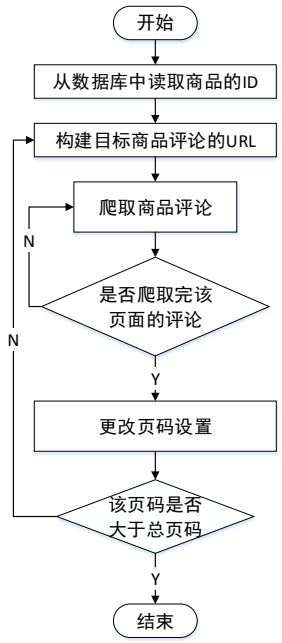

图 3-3 商品评论爬取流程

Fig.3-3 Product review crawling process

3.4 数据获取及分析

在前几个小节中本文确定了所需数据的来源网站，并制定了商品信息和商品

评论的爬取策略，本小节将通过对电商网站的网页数据结构的解析来获取所需的数据。根据爬取制定策略，本小节将分为两部分进行介绍。

3.4.1 商品信息的收集

（1）京东商城商品信息采集

本文主要针对电子商品进行分析，因此以笔记本电脑为例进行数据的收集工作介绍。首先以京东笔记本电脑的品牌分类页面为起始进行分析，该页面的 URL 为 https://list.jd.com/list.html?cat=670,671,672。该页面包含了京东所有笔记本电脑的品牌商的名称，之后利用浏览器的开发者模式对该页面的 HTML 源码进行分析获取所需的品牌名称及其对应的 URL，如图 3-4 所示，发现京东用于存储品牌的标题及其 URL 信息存放在 div 标签下的 u1 id="brandsArea"中，其中 URL 在存放<a>标签中的 href 属性中，品牌标题放在<a>标签中的 title 属性中。之后根据分析结果制定获取所有品牌标题和 URL 的语法，本文使用 xpth 语法对 HTML 页面进行匹配选取所需数据，xpath 是 XML 路径语言，由于 HTML 与 XML 同样是树状结构，也常被用于匹配选取 HTML 中的特定数据。

图 3-4 京东品牌源码截图

Fig.3-4 screenshot of JD brand source code

完成对品牌的分析收集后，就可以对各个品牌的笔记本电脑列表页进行分析并制定相应的爬取规则。在商品的列表需要收集的数据是商品的详细页的 URL 与商品的图片 URL 以及商品对应的 ID。通过对各品牌下笔记本电脑列表页的 HTML 结构如图 3-5 所示分析发现，用于存储各型号笔记本详细页 URL 以及图片 URL 的位置在 div 标签下的 class=“p-img”中，其中<a>标签中的 href 是存储笔记本电脑详细页的 URL，src 是用来存储图片的 URL。之后通过制定相应的 xpath 语法对笔记本详细页面 URL 以及图片 URL 进行收集。由于详细页面 URL 中包含了商品的 ID 信息，如 https://item.jd.com/1000051714 61.html 该链接的数字部分即为该商品的 ID，所以本文通过正则匹配的方法对详细页面的 URL 进行处理，获取商品的 ID。

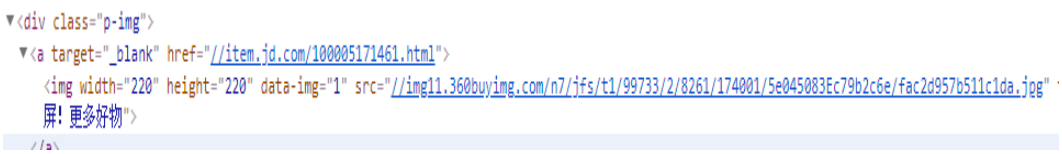

图 3-5 京东商品信息源码截图

Fig.3-5 screenshot of Jingdong commodity information source code

最后是对笔记本的详细页面进行分析获取到该笔记本的标题、参数以及价格。经过对详细页的 HTML 的结构分析发现，笔记本的标题存放在 head 标签中的 title 中，参数则存放在 div 标签下的 class=“PTable”中，根据分析结果制定相应 xpath 语法获取笔记本的标题以及参数，对于笔记本的价格则是使用商品的 ID 使用京东免费查询单个商品价格的接口来获取相应的价格。

（2）淘宝商城商品信息采集

淘宝商城的商品信息采集的策略与京东略显不同，本文以淘宝网的搜索页面为起始页面进行分析，淘宝网的搜索页面的 URL 为 https://s.taobao.com。本文以笔记本电脑为例进介绍。设置搜索关键词为笔记本电脑并进行搜索。在进入搜索结果页面后，通过对图 3-6 的分析可以发现商品详细页面的 URL 存储在 div 标签下的 class="row row-2 title"中，其中<a>标签中的 href 是用来存储商品详细页面的 URL，<data-nid>是用来储存 ID 信息。对于图片的 URL 则储存在 div 标签下的 class="pic"中，其中<a>标签中的 src 是用来储存商品图片的 URL。

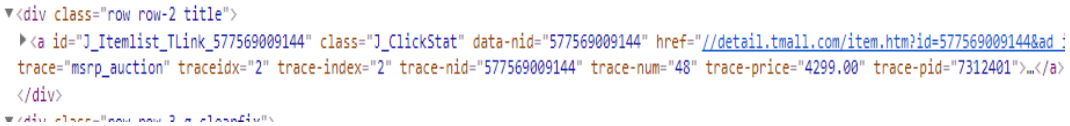

图 3-6 淘宝商品信息源码截图

Fig.3-6 screenshot of Taobao commodity information source code

获取到商品详细页面的 URL 后就可以对商品的详细页面进行分析，来获取该商品的标题、参数等信息。对详细页面进行分析后，可以发现商品的标题是存放在 div 标签下的 class="tb-detail-hd"中，价格是存放在 div 标签下的 class="tm-promo-price"中，参数信息存放在 table 标签中的 class="tm-tableAttr"中。之后根据分析的结果通过制定相应的 xpath 语法来获取笔记本的相关信息。

3.4.2 商品评论的收集

在对商品信息的收集中，本文通过分析网页的 HTML 结构来获取相应的数据，但由于包含有商品评论的网页是动态加载的，如果仍使用之前的分析方法则会导致仅能获取少量的商品评论，因此需要通过分析网页的 network 属性来获取到商品评论的真实 URL。本文以某一型号的笔记本电脑为例进行介绍，在登录页面后进入开发者模式，选定该页面的 network 属性，查找其中名称包含有 comment 的文件，通过预览可以得知在 content 中包含有商品评论，因此该文件即为储存有京东商品评论的文件，而用于储存商品评论的 URL 可以在文件的 headers 中找的，如图 3-7 所示。

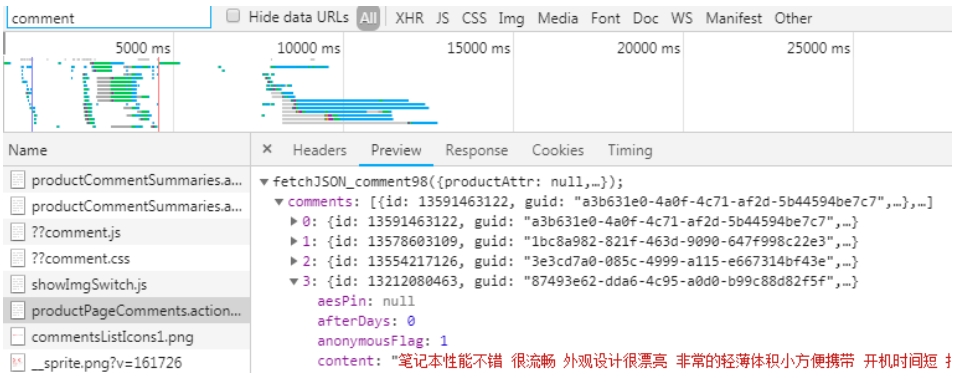

图 3-7 京东源码截图

Fig. 3-7 JD source code screenshot

通过同样的分析方法，可以发现淘宝用于存储商品评论的文件为包含有 list_detail 的文件，商品评论的 URL 同样可以在 headers 中找到，如图 3-8 所示。

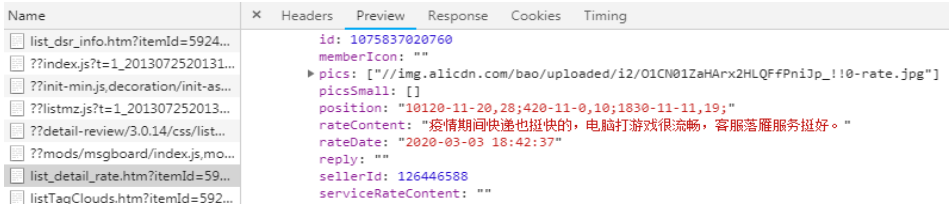

图 3-8 淘宝源码截图

Fig. 3-8 screenshot of Taobao source code

在对其他笔记本电脑进行同样的分析后，可以发现这些用于存储商品评论的 URL 的区别仅在于商品的 ID 不同，所以本文依此构造了商品评论的初始 URL，并通过更改 URL 中的商品 ID 来构造出相应商品评论的 URL。获取评论的 URL 后，就可分析电商平台返回的数据来制定相应的爬虫规则来获取评论内容、评分等信息。本文以京东商城的返回数据为例进行分析，如图 3-9 所示，可以发现返回的响应数据是 json 格式的数据，所以可以通过查找数据中的 key 来获取对应的数据，其中存放评论的 key 是 content，存放评论创建时间的是 creationTime，存放用户评分的是 score。因此通过遍历数据中的 content、createTime、score 来获取到相应的数据。

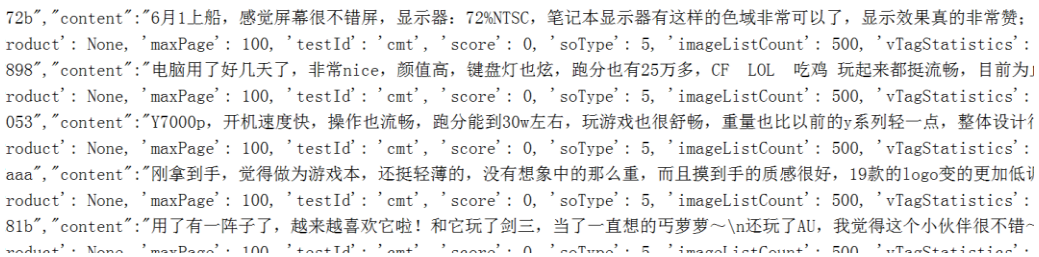

图 3-9 京东响应数据截图

Fig. 3-9 JD response data screenshot

3.5 反爬虫措施

由于网络爬虫在短时间内向爬取的网站发送大量请求，这就增加服务器的负担，影响其他正常用户的使用。所以各大网站会制定一些反爬措施，来限制网络爬虫的使用。另一方面各大网站为了保护自己的数据，防止被第三方平台滥用，也会在后台上嵌入反爬系统，来限制网络爬虫。本文进行数据采集的电商网站都拥有着完善的反爬机制，因此需要制定相应的对策来绕过反爬机制。

（1）使用代理 IP 池，由于 IP 是网络之间用于传输数据协议，当电商网站监测到某一 IP 访问异常时，就会将该 IP 纳入黑名单，拒绝该 IP 进行访问，所以本文为了保证爬虫正常工作，通过收集网络上免费提供的 IP，如站大爷（www.zdaye.com）等，在 setting.py 文件中建立代理 IP 池，当检测到使用的 IP 多次访问超时时，从代理 IP 池中抽取新的 IP 继续进行爬取。

（2）设置延时程序，由于网络爬虫在进行数据爬取时会在短时间内发送大量的访问请求，容易触发电商的反爬机制，所以本文在每次爬虫循环结束后，调用延时函数，进行延时 3 秒，用于模拟人为访问操作。

（3）使用随机 user-agent，user-agent 是用户代理，是服务器用于识别用户的浏览器信息的一组参数数据，本文使用第三方开源用户代理库 fake-useragent 结合 python 自带的 random 函数，实现随机更换用户代理，模拟不同的浏览器信息，从而降低触发反爬机制的概率。

3.6 数据的储存

通过网络爬虫，本文在电商平台上爬取了大量的商品信息以及对应的商品评论，其中各型手机共计 2370 条，各个手机的评论共计 135200 条，各型笔记本电脑共计 1340 条，各个笔记本电脑的评论共计 89650 条。在对采集到的数据进行去重、筛选等操作后，存入数据库中用于下一阶段分析中。本文使用的数据库是 mysql 数据库，该数据库是目前流行的一款轻量化的数据库。其中手机信息储存在 goods_phone 表中，手机评论存储在 phone_comment 表中，笔记本电脑信息储存在 goods_computer 表中，笔记本的评论存储在 computer_comment 表中。部分储存数据如图 3-10 所示。

comment_id	id	score	create_time	content
1000018203671132227684 6	100001820367	5	2018-03-19 00:2	京东购买方便快捷，送货上门很好
1000018203671132494576 3	100001820367	5	2018-03-19 21:2	非常客观的说，优点是轻薄，键盘手感非常好，性能商务办公绝对够用，玩游戏就算了
1000018203671133100965 8	100001820367	5	2018-03-21 21:4	对比了苹果和微软，经过专业IT朋友的推荐最后选择了这款X1，首先不得不佩服京东的
1000018203671133629972 1	100001820367	5	2018-03-23 19:0	物流很快，出乎意料，这次包装很好看，打开包装还挺香！电脑也很爽，超快！

图 3-10 部分电脑评价数据

Fig. 3-10 part of computer evaluation data

3.7 本章小结

本章针对数据采集工作进行了详细的阐述，首先介绍了本文使用到的基于 Scrapy 框架网络爬虫的相关知识及其工作流程，之后通过数据源的选择和数据爬取策略的设计，选定了京东和淘宝作为数据的来源。然后通过对京东和淘宝页面的分析制定了网络爬虫的爬取规则。最后针对电商平台的反爬机制，制定了相应的应对措施。

第 4 章 情感分析算法研究

情感分析又称意见挖掘，它是自然语言处理方向的一个热门的研究方向，同时也是本系统的核心所在。本章将对情感分析方法中的基于情感词典的方法和基于 SVM 的方法进行研究，设计出本文所需要的情感分析方案，用于处理本文在第三章中收集到的商品评论数据。

4.1 数据清洗

商品评论文本不同于其他规范化的文本数据，许多评论都存在着随意性、口语化严重等问题，而且由于电商平台存在着评论的奖励机制，当评论达到一定字数时就能获取到积分奖励，这就导致了一些用户在进行评论时会进行“恶意评论”。这些问题就导致了获取到的评论文本质量不一，一些低质量的评论文本会降低情感分析的精确度，所以为了保证文本的质量，需要对爬虫收集到的数据进行清洗。

本文通过对收集到商品评论进行分析后发现，这类文本在形式上可分为两类，第一类通常是以字符或数字的形式出现，如一些评论中会出现“111111111”或者“@#￥%&%”，这些评论一般都没有实际意义，因此本文利用正则匹配的方式对这些评论进行识别删除。

最值得注意的是第二类，这类评论可以概括为垃圾信息，通常是以文字的形式出现，大致表现为，一是有些评论为了凑字数，会重复使用几个文字，如“很方便，方便，方便，方便，方便”等，这种评论出现的次数不多，本文通过人工的方式进行处理将重复的词语进行删除。二是电商平台的默认好评，这种评论是由于买家未及时填写评价内容，电商平台就会默认评价，对于这种文本，本文通过设置模板来进行匹配删除。三是“刷单”导致的垃圾评论，这些评论会误导用户，也会降低情感分析的准确性，通常情况下用于刷单的评论一般会有一些固定的万能评语，如“不错的卖家，谢谢啦。我和同事都很喜欢，下次还会再来”或“一个字！！值！！”等固定句式。这些评论一般不涉及到商品的相关信息，适用于任何商品下的评论，因此本文通过收集这些通用评论模板，对出现的刷单评论进行匹配删除。

4.2 中文商品情感词典的构建

本文在 2.3 节中介绍了用于情感分析的方法，并介绍了情感分析中常用的基

于情感词典和机器学习方法的流程。在基于情感词典的分析方法中需要利用词典识别和提取文本中的情感词、程度副词等词语。所以情感词典的词语覆盖程度影响着该方法的准确率，因此需要构建一个相对完善的情感词典。由于情感词典通常是语言学的专家结合自身的知识耗费大量时间构建的，所以为了降低情感词典的构建难度，本文以通用情感词典为基础构建基础情感词典。

经过对国内主流情感词典进行分析对比后，本文选定了广泛使用的知网情感词典作为本文的基础情感词典。基础情感词典中的部分情感词如表 4-1 所示。除了情感词典，本文还构建了程度副词词典、否定词词典以及连词词典。

表 4-1 基础情感词典部分情感词

Table 4-1 part of emotion words in basic emotion dictionary

情感极性	情感值	情感词
正面	+1	表扬、称誉、有效、善、壮美、皮试、安全、优良
负面	-1	嘴笨、招风、弱小、令人厌恶、汗颜、名过其实

4.2.1 专用情感词典

虽然通用情感词典收录了大量的词语，但随着网络新词的不断涌现，用户在进行评论时往往会使用一些网络词汇，这就导致了通用情感词的完备程度不足以应对本文的情感分析工作，需要对其进行扩展，增加其情感词的收录量。对于情感词典的扩展大多采用情感倾向点互信息算法（SO-PMI）和 word2vec 这两种方法，其中 Word2vec 方法是基于词语的词向量来实现扩展的，有利于挖掘词语间的语义信息。因此本文采用 Word2vec 来对基础情感词典进行扩展。

情感词典的扩充过程可以分为两部分进行。第一部分，首先是收集商品评论中出现的情感词，然后与基础情感词典进行对比，剔除掉已存在于词典中的情感词，组成未收录情感词集合。第二部分，首先构建种子情感词典，之后通过计算未收录情感词与种子情感词典之间词语相似度的方式，完成未收录情感词的标注。情感词是指在一句话中能够表达用户主观情感倾向的词语，通常为动词、形容词，如“满意”、“高兴”“可以”等。

第一部分是对评论中的情感词进行收集。首先对商品评价进行预处理操作，调用结巴分词工具将商品评价文本，分割成带有词性的单独词汇，然后根据词性提取出商品评价中的动词、形容词等情感词，之后对提取出的情感词进行词频统计，剔除掉词频较低的情感词，本文设定的阈值为 10，即剔除掉词频不足 10 的情感词。完成低频情感词的剔除后，将余下的情感词与基础情感词典进行匹配，剔除掉已经出现在词典中的情感词，最后得到候选情感词集合。具体流程如图 4-1 所示。

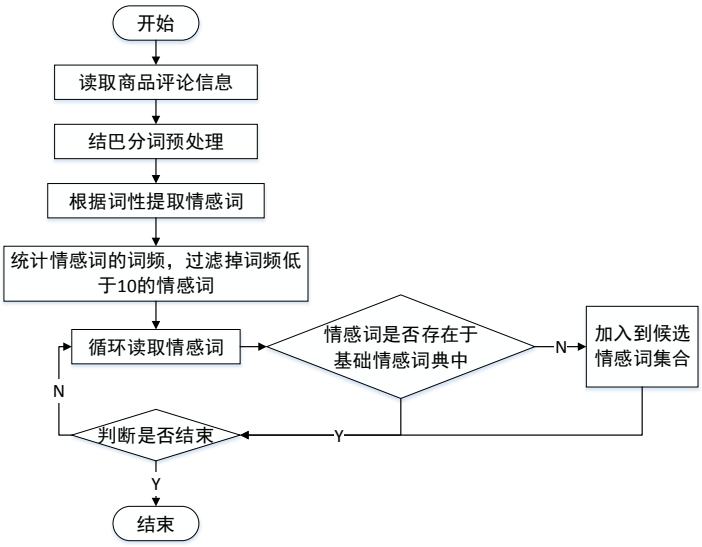

图 4-1 情感词收集流程图

Fig. 4-1 flow chart of emotional words collection

第二部分是对候选情感词集合进行情感值的标注。在第一部分中，本文收集到了大量的未标注的候选情感词，这些未标注的情感词是无法直接扩充进基础情感词典中，需要对其进行标注，考虑到人工标注的方法往往费事费力，本文采用 Word2vec 模型对其进行标注工作。具体的过程如图 4-2 所示。

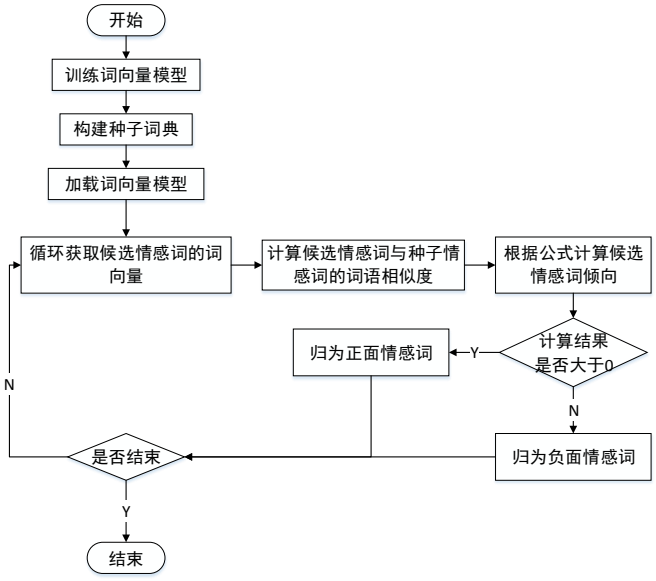

图 4-2 情感词标注流程图

Fig. 4-2 flow chart of emotional words annotation

利用 Word2vec 模型进行情感词标注的详细描述如下所示。

第一步，将收集到的十几万条商品评价文本输入 Word2vec 中进行训练，得到商品评论的词向量模型文件。本文在第二章中介绍过 Word2vec 常用的模型为 CBOW 和 Skip-Gram 这两个模型，其中 Skip-Gram 模型训练的词向量较 CBOW 模型训练的词向量的准确率要高 20%[51]。因此本文使用 Skip-Gram 模型进行词

向量训练。

训练词向量的具体的参数设置为：词向量维度为 200，窗口大小为 10，最低词频为 10，训练算法为 skip-gram，采样阈值为 1e-3，其余参数使用默认值。训练参数与其取值如表 4-2 所示。

表 4-2 训练参数及其说明

Table 4-2 training parameters and description

参数名称	参数值	释义
-train	sppj.txt	输入文本语料
-out_model	评论.model	输出训练完成的词向量模型
-size	200	词向量的维度
-min_count	10	最低词频
-window	10	上下窗口大小
-cbow	0	是否使用 cbow 模型（0 是不使用）
-sample	le-3	采样的阈值

第二步，构建种子情感词典，由于知网情感词典没有对情感词的强度进行划分，因此本文从大连理工大学情感词汇本体库中挑选种子情感词。在情感词汇本体中，情感词根据其情感强度化分为 5 档，其中情感词强度最高的为第 5 档，因此本文提取词汇本体中第 5 档的情感词，并统计这些情感词在文本语料中的词频，之后提取词频排名前 70%的情感词作为种子情感词，用于构建种子情感词典。

第三步，读取训练完成的词向量模型文件，将候选情感词与种子情感词输入，获取到其对应的词向量，之后计算候选情感词与种子情感词之间的相似度，最后根据公式（4-1）来计算该候选情感词的情感倾向，根据计算结果标注为正面或负面情感词。

$$Value(w)=\frac{1}{m}\sum_{i=1}^{m}sim(w,Pword_i)-\frac{1}{n}\sum_{j=1}^{n}sim(w,Nword_j) \qquad (4\text{-}1)$$

式中，$sim(word_1,word_2)$——$word_1$ 与 $word_2$ 之间的语义相似度

$Pword$——正面种子情感词

$Nword$——负面种子情感词

$value(w)$——候选情感词的得分结果，当 $value(w)>0$ 时，为正面情感词，否则为负面情感词。

计算词语相似度的方法通常是通过计算词向量间的余弦值，当余弦值越大时，这两个词语的相似度也就越高，越小时则表明这两个词语相似度也就越低，具体的表达式为公式（4-2）。

$$\cos\theta=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}} \qquad (4\text{-}2)$$

式中，n——向量的维度

x, y——两个词语的向量

第四步，将情感词集合添加进基础情感词典中，完成领域词典的建立。领域词典中部分情感词如表 4-3 所示。

表 4-3 领域情感词典部分情感词

Table 4-3 part of emotion words in the domain emotion dictionary

情感极性	情感值	情感词
正面	1	耐用、耐摔、奈斯、全五星、好评、一流、有品位、低调奢华、物超所值、顺手、时尚、帅炸了、蛮好的、很溜、还会光顾、大品牌、实用、惊艳、可以、流畅、棒棒哒、给力、还可以、正品、行货、小巧、稳定、新潮、没问题、蛮不错
负面	-1	无语、醉了、卡死、垃圾、丑、卡、慢、假货、差评、卡顿、难看、心疼、亏、会发热、有瑕疵、断流、烫手、丑死了、退货、掉漆、失望、后悔、一般般、就那样、呵呵、不值、水货、不值、不行、店大欺人、将就、渣、接触不良、厚

4.2.2 程度副词词典

程度副词的使用会改变商品评论文本的情感强度，因此需要在情感分析时考虑程度副词会给评价文本带来多大的影响。当程度副词被用于修饰情感词时，情感词的情感倾向虽然没有发生变化，但情感词的情感强度会随着程度副词的不同会被加强或削弱。本文使用了知网情感词典中程度级别词语表作为程度副词词典，由于程度级别词语表中仅是将程度副词分成了超、最、很、较、稍、欠、这 6 类，并没有赋予权值，因此本文将这 6 类程度副词分别赋予了一定权值。部分标注结果如表 4-4 所示。

表 4-4 程度副词词典部分程度副词

Table 4-4 some degree adverbs in the dictionary of degree adverbs

类别	权值	程度副词
超	2.0	超、超级、过、过于、何止、不为过
最	1.7	极、极为、百分之百、充分、最、最为、无比
很	1.5	多、多么、分外、格外、好、很、很是、尤其
较	1.2	更加、更为、还要、较、较为、那么、那样
稍	0.8	略微、略加、稍微、稍稍、稍许、一点、有些
欠	0.5	半点、不大、相对、不怎么、没怎么、轻度

4.2.3 否定词词典

否定词可以用于翻转词语或者语句的情感极性，如：我喜欢这部手机，与“我不喜欢这部手机”这两个句话的表达的意思完全相反，句子的情感倾向发生了改

变，但如果出现“不可能不满意”这类语句时，其情感倾向没有发生改变，仍然是正向的。所以在进行情感分析过程中需要充分考虑到句中否定词出现的次数，一般的，在一句话中，语句情感极性发生改变的次数会与否定词出现的次数一致。当句子中否定词出现的次数为单数时，语句的情感极性发生改变，当用于修饰的否定词出现的次数为双数时，语句的情感极性不发生变化。部分否定词如表 4-5 所示。

表 4-5 否定词词典部分否定词

Table 4-5 some negative words in the dictionary of negative words

权值	否定词
-1	不、非、无、没、否、不是、勿、弗、毋、未、 别、休、無、不曾、没有、不然、不必、未必

4.2.4 连词词典

在一些评论中，用户为了使句子看起来更加通顺，通常会使用一些连词，有些连词的使用会影响到评论中的一些子句的关系，如“虽然这部手机很贵，但它的性能超强”。这句话中用户的表达重心在后一段话上，表达了对该手机的认可之情。常用的连词可分为并列连词、并折连词、递进连词和因果连词，其中转折连词是用于表示两句话之间的转折关系，其要表达的重心主要在后句上。递进连词是表达分句间的递进关系，在递进关系中，后一句的情感强度比前一句的情感强度更强。为了识别出文本中的连词，本文通过建立连词词典来进行匹配识别。部分连词如表 4-6 所示。

表 4-6 连词词典部分连词

Table 4-6 some conjunctions in the dictionary of conjunctions

序号	权值	连词
1	2	但是、偏偏、况且、何况、但、指
2	1.5	不但、也、其次、不仅、就是
3	1.2	不过、岂知、不过、不料、虽然

4.3 基于词典与规则的情感分析

4.3.1 词语组合规则

由于中文语法复杂，在使用基于词典方法进行情感分析时，仅仅依靠情感词典来对商品评论进行分析，会导致分析结果的准确率不高，因此需要结合中文的语法结构，制定一些评分规则，来提高基于情感词典方法的准确率。

（1）情感词

当文本中只出现情感词时，该文本的词语组合得分为情感词的权值，如“手

机不错”，其中情感词的权值为 1，则词语得分为 1，计算形式为词语组合得分=情感词。

（2）程度副词与情感词的组合

当程度副词与情感词同时出现时，可以根据程度副词在词典中的权值划分情况，将权值与情感词相乘得到词语组合的得分，如“很高兴”，其中情感词的权值为 1，程度副词的权值为 1.2，那么该词语组合的得分为 1.2。计算形式为：词语组合得分=程度副词*情感词。

（3）否定词和情感词的组合

当这种组合出现时，需要对否定词出现的频率 n 进行统计，通过取其权值的 n 次方与情感词的权值相乘得到该词的最终得分。计算形式为：词语组合得分=（-1）^{n}情感词。

（4）程度副词与否定词同时出现的组合

除了前两种组合外，还存在程度副词与否定词同时出现的组合情况，对于这种组合方式需要根据程度副词与否定词的位置来计算词语组合的得分，当程度副词出现在否定词前时，计算方式为程度副词乘以否定词乘以情感词，如“很不满意”的词语组合得分为 1.2*-1*1=-1.2，计算形式：词语组合得分=程度副词*否定词*情感词。

当否定词出现在程度副词前面时，其整体的情感表达结果就会与之前的组合产生改变，如“我不是很满意”它就与“我很不满意”的整体情感表达有所不同，前者表达的是正面情感，但情感强度要稍弱些。而后者表达了较为强烈的负面情感。对于出现否定词在程度副词前的组合形式，本文通过设定权值来进行计算，计算形式为：词语组合得分=0.5（程度副词 *情感词）。词语组合及其计算公式如表 4-7 所示。

表 4-7 词语组合及其计算公式

Table 4-7 word combination and calculation formula

序号	词语组合	计算公式
1	情感词	$E = SW$
2	程度副词+情感词	$E = A*SW$
3	否定词+情感词	$E = (-1)^{n} SW$
4	程度副词+否定词+情感词	$E = A(-1)^{n} SW$
5	否定词+程度副词+情感词	$E = 0.5(A*SW)$

注：表中 E 为词语组合得分，SW 为情感词权值，A 为程度副词权值，n 为否定词出现次数。

4.3.2 基于词典和规则的算法设计

基于词典和规则的情感分析流程如图 4-3 所示。

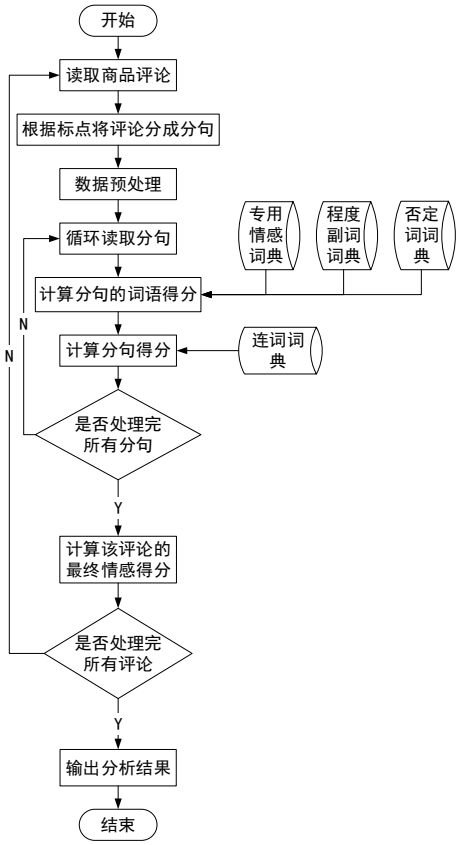

图 4-3 基于词典和规则的情感分析流程

Fig. 4-3 emotion analysis process based on dictionary and rules

在商品的评论中，由于一条评论中可能包含多个句子，因此本文以逗号为分割点，将评论分割为多个分句，调用结巴分词工具对文本进行预处理操作，然后利用构建好的专用情感词典、程度副词词典和否定词词典对评论文本中的情感词、程度副词和否定词进行抽取，计算分句中的词语组合得分，之后利用连词词典抽取连词，用以计算分句的得分情况，最后将分句得分汇总，计算出该条评论的最终情感得分。如果得分大于 0，则将该评论标记为正面评论，如果小于 0，则标记为负面评论。具体方法流程如表 4-8 所示。

表 4-8 基于词典和规则的情感分析方法流程

Table 4-8 emotion analysis method flow based on dictionary and rules

基于词典和规则的方法流程
输入：商品的评论文本数据 C
输出：基于词典和规则的分析结果
过程
第一步：调取数据库中的商品评论 C。
第二步：根据标点将 C 切割成数个分句。
第三步：调用结巴分词对每个分句进行预处理操作。
第四步：读取情感词典，对处理完成的分句进行情感词匹配：

IF 分句中存在情感词：

读取程度副词词典和否定词词典，提取并记录情感词的位置，以情感词为中心建立字典结构，并在情感词的位置设置检测窗口，在三个窗口内检测程度副词和否定词，将检测到的程度副词和否定词按照位置存入字典中，之后按照表 4-7 中的计算公式来计算分句的词语组合得分 E 。

IF 分句中不存在情感词：

读取下一条分句进行分析。

第五步：读取连词词典，对分句中的连词词典进行匹配：

IF 分句中存在分词：

按照公式（4-3）来计算分句的得分。

$$S = B * E \tag{4-3}$$

式中，S ——分句的情感得分

B ——连词的权值

E ——分句的词语组合得分

IF 分句中不存在分词：

分析下一条分句。

第六步：根据公式（4-4）计算整条评论的情感得分。

$$E(S) = \sum (S_i) \tag{4-4}$$

式中，$E(S)$ ——整体评论的最终情感得分

S_i ——第 i 条分句的情感得分

第七步：根据评论的最终情感得分来标注评论，如果得分大于 0，则标记为正面评论，如果小于 0，则标记为负面评论。

第八步：重复第一步，直至标记完成所有评论。

4.4 基于特征组合 SVM 的情感分析

在 4.2 小节中本文使用 Word2vec 模型在通用情感词典到的基础上构建了专用情感词典并通过制定评分规则完成了基于词典的情感分析方法的设计，在本小节中，本文将完成基于 SVM 情感分析方法的设计。

4.4.1 SVM 算法

SVM（support vector machine），支持向量机是一种经典的机器学习算法，最初是由 Cortes 和 Vapnik 等人研究发表。SVM 是一种有监督的二元分类算法，常用于文本分类和回归分析中，它的核心思想是在一个给定的空间中找到能够将数据按照其类别来进行划分的最优分类超平面，使该分类器在满足准确率的条件下，实现分类间隔的最大化，增强该分类模型的泛化能力。

支持向量机的线性分类如图 4-4，在该图中，中间的实线代表着最优分类超平面，超平面两边的红色圆点和绿色方形表示不同类别的数据，虚线上的红色圆点和绿色方形代表着距离超平面最近的数据点。

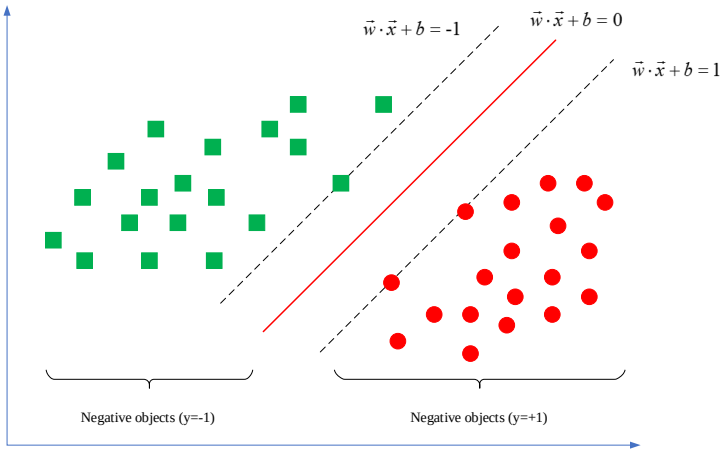

图 4-4 SVM 线性分类

Fig. 4-4 SVM linear classification

SVM 的分类过程就是在给定的训练集样本 $D=\{(x_1,y_1),(x_2,y_2),\cdots(x_i,y_i)\}$，通过训练找到一个能够将样本 D 中的不同类别数据进行分类的最优超平面，其中 x_i 为训练集第 i 个样本的特征，y_i 训练集第 i 个样本的结果标签，取值为{-1,+1}，当该样本为正例时 y 的值为+1，反之为-1。

SVM 在样本 D 中寻找最优分类超平面的表达式如公式（4-5）所示。

$$w^T x+b=0 \tag{4-5}$$

式中，w——法向量

b——截距

通过公式（4-5）可知，SVM 的分类超平面是由参数 w 与参数 b 共同决定的，即图 4-4 中的实线。而训练的过程就是寻找这两个参数使其满足公式（4-6）和公式（4-7）的约束条件下，使图 4-4 中实线两侧的虚线部分上的点都能成立。

$$w^T x_i+b\geq+1,\ \ y_i=+1 \tag{4-6}$$

$$w^T x_i+b\geq-1,\ \ y_i=-1 \tag{4-7}$$

其中，满足约束条件的距离平面最近的点成为“支持向量”，支持向量距离超平面的间隔值为 $2/\|w\|$ ，最优超平面即为间隔值最大时，可以等价为求 $\|w\|$ 最小时，最优分类超平面的公式（4-8），约束条件为公式（4-6）和公式（4-7）。

$$\min\frac{1}{2}\|w\|^2 \tag{4-8}$$

在使用支持向量机进行分类问题的研究时，如果遇到非线性分类问题时，线性分类器就不能完成分类工作，对于非线性分类问题，SVM 提供了核函数来进行解决，通过核函数将原始的数据映射到高维空间上，以此来解决原始数据空间中存在的线性不可分的情况。SVM 使用核函数将非线性的原始数据映射到高维

空间转化成线性可分的具体情况如图 4-5 所示。

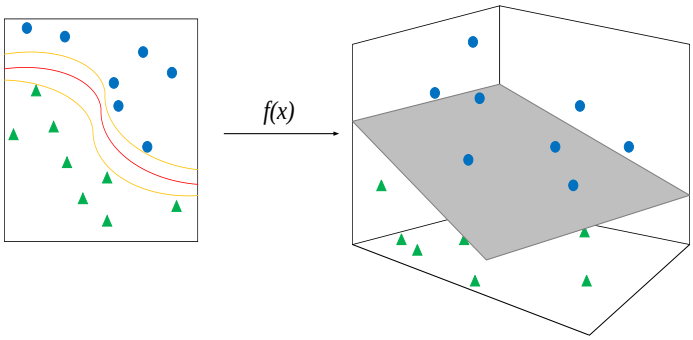

图 4-5 SVM 映射到高维空间

Fig. 4-5 mapping SVM to high dimensional space

支持向量机中常用的核函数有四种，分别为高斯核函数、多项式核函数、线性核函数、Sigmoid 核函数。

（1）高斯核函数

高斯核函数是特殊的径向基核函数（RBF）,它属于局部核函数，通过高斯核函数将样本数据映射到高维的空间中。高斯核函数的表达式如公式（4-9）所示。

$$f(x_i, x_j) = \exp(\frac{-\| x_i - x_j \|^2}{2\delta^2}),\ x_i x_j \in D \tag{4-9}$$

式中，δ——高斯核的带宽

高斯核函数可以根据使用者的需求通过调控δ来实现映射到不同维数，具有很高的灵活性。

（2）多项式核函数

该核函数是 SVM 的一种全局核函数，它适合处理正交归一化类型的数据。多项式核函数的表达式如公式（4-10）所示。

$$f(x_i, x_j) = (x_i^T x_j)^d \tag{4-10}$$

式中，d——多项式的次数

多项式核函数的参数d越大时，数据映射的维数就越高，但当d越大时，其计算的复杂度就越高，就容易造成“过拟合”情况的发生。

（3）线性核函数

线性核函数具有参数少，速度快的优点，可以与 SVM 相配合快速找到最优超平面。一般分类问题使用线性核函数就能取得不错效果。其表达式为公式（4-11）。

$$f(x_i, x_j) = x_i^T x_j \tag{4-11}$$

（4）Sigmoid 核函数

该核函数是 SVM 中较为广泛的核函数，其来源于神经网络，Sigmoid 核函

数的表达式如公式（4-12）所示。

$$f(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta),\ \beta > 0, \theta < 0,\ x_i x_j \in D \qquad (4\text{-}12)$$

Sigmoid 核函数能够借助神经网络对样本数据进行泛化，求出其最优解。

4.4.2 基于词典和用户评分的训练集构建

本文选用 SVM 作为第二种情感分析方法，该方法需要大量已经标记完成的文本作为训练集，对分类器进行训练，而且训练集的质量直接关系到了分类器的训练效果。由于本文是针对商品评论进行分析，因此为了保证 SVM 分类器的准确率，需要以本文收集到商品评论为基础构建训练集数据。在构建训练集方面，通过人工标注来构建训练集的方式既费时又费事，目前常用的训练集构建方法是利用情感词典与用户评分相结合的构建方法[52,53]。

训练集的构建可分两步进行，首先从评价中收集用户评分为 5 分和 1 分的评论，其中用户评分为 5 分的评论在电商平台中为好评，1 分评论为差评。之后利用本文设计的基于词典和规则的情感分析方法对 5 分和 1 分的评论进行分析。构建训练集的具体流程如表 4-9 所示。

表 4-9 基于词典和用户评分的训练集构建流程

Table 4-9 training set construction process based on dictionary and user score

基于词典和用户评分的训练集构建流程
输入：商品评价的用户评价文本 C
输出：SVM 的训练集
过程
第一步：调取数据库中的商品评论 C。
第二步：提取商品评价中评分为 5 分和 1 分的评论。
第三步：调用 jieba 分词工具对提取的评论进行预处理操作。
第四步：调用 4.3 章中构建的基于词典的情感分析方法对预处理后的数据进行处理。
第五步：IF 评论的情感分析得分>0 将评论归为正面评论。 ELSE 将评论归为负面评论。
第六步：遍历正面与负面评论，将正面评论中用户评分为 1 的评论与负面评论中用户评分为 5 的评论剔除。
第七步：按照情感分析的得分的绝对值对两类评论按照得分大小排序。
第八步：分别提取正面和负面评论中排序前 1000 的评论，组成正例与负例 1:1 的训练集，其中正例标记为+1，负例标记为-1。

4.4.3 特征选择

将评论文本数据进行预处理后，就可以得到了一个原始的特征空间，该特征

空间不仅较大而且较为复杂，其中包含了大量不同的特征，这些特征对于文本的分类贡献不一，如果不对特征空间进行处理，将特征空间中贡献较低的特征剔除，则会增加分类算法的计算复杂度，降低分类的效率，影响分类模型的准确率。目前常用的降维方法有特征选择和特征提取，其中特征选择是利用数学的统计学方法，通过计算每个特征的贡献值，将特征贡献较大的特征选择出来，从而降低特征空间的大小。目前常用的特征选择方法有词频逆文档频率（TF-IDF）、互信息（MI）、卡方统计（x^2）以及信息增益（IG）。各个特征选择方法的公式本文在 2.4 小节中已经说明。本文选择词频逆文档频率（TF-IDF）为特征选择的方法。特征选取的流程如图 4-6 所示。

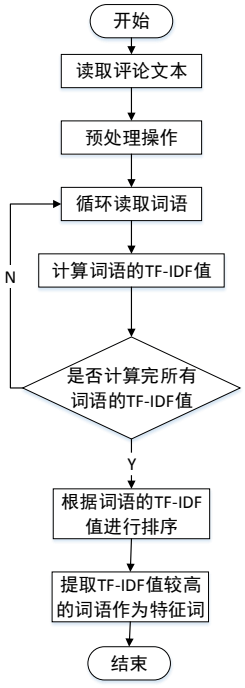

图 4-6 特征选择流程

Fig 4-6 feature selection process

在利用 TF-IDF 算法实现了商品评论的特征选择后，商品评论文本的特征空间维度得到了有效地降低了，也剔除掉了对文本贡献较低的干扰项。

4.4.4 文本向量加权表示

本文利用 TF-IDF 算法进行特征选择，筛选出评论中对句子贡献较大的词语作为特征词，但这样的特征词还不能直接输入 SVM 中进行情感分析，需要将特征词转化为计算机能够识别的词向量形式。目前常用的词向量表示方法有 One-Hot representation 和 Distributed representation。其中 One-Hot representation 的词向量表示方法是利用 0 和 1 来表示一个词语的词向量，如“西红柿”可以表示为[000010000…]“番茄”表示为[00100000…]。这种词向量的表示方法虽然能够简明

的表示一个词，但是随着词语的增多，维数也越来越大，这就容易导致“维数灾难”。Distributed representation 的表示方法是将词语投射到一个实体空间维度中，该空间维度通常是低维度（100 维或 200 维等），如“西红柿”表示为[0.645，-0.124,0.531，-0.358,0.347…]。这种词语的表示方法能够有效解决 One-Hot 中出现的“维度灾难”问题，而这种表示方法能够解决“词语鸿沟”现象，让相似或者关联性较强的词语在词向量空间中更为接近。

本文使用 Word2vec 训练产生的词向量模型是以 Distributed representation 表示形式为基础，利用神经网络训练而来的词向量，该词向量能够很好的反映出词语的语义特征。

Word2vec 虽然可以根据文中词语间的上下关系，构建一个能够反映词间关系的词向量矩阵，但该词向量不能很好地体现出词语的重要性，因此为了突显出词向量中词语的重要程度，改进 Word2vec 的词向量表达形式，可以通过对词向量进行加权的方式进行改进。本文采用的词向量加权的方法是利用词语的 TF-IDF 值与词向量相结合的方式[54,55]。词向量加权表示的步骤如下所示。

（1）利用 jieba 分词工具和停用词词典对商品评论文本进行预处理操作，预处理后的文本被分割为包含 n 个单独词汇的集合 D。

（2）使用 TF-IDF 算法对集合 D 进行特征选择，筛选出集合 D 中的特征词，并保留每个特征词的 TF-IDF 值 k。

（3）读取 4.2 小节中训练完成的词向量模型，将特征词输入获取每个特征词 K 维的词向量 $\vec{w}$。

（4）根据公式计算集合 D 的词向量 $\vec{V}(D)$。

$$\vec{V}(D)=\sum_{i=1}^{m}\vec{w}_i\times k_i \tag{4-13}$$

式中，m——特征词的个数

$\vec{w}_i$——特征词 i 的词向量

k_i——特征词 i 的 TF-IDF 值

4.4.5 基于 SVM 的情感分析设计

在 4.4.1 小节中本文介绍了 SVM 的计算公式和其提供的 4 种核函数，通过学者对 SVM 核函数应用场景的研究，本文选定更适合处理文本情感分析的线性核函数[56]。基于 SVM 的情感分析流程如图 4-7 所示。

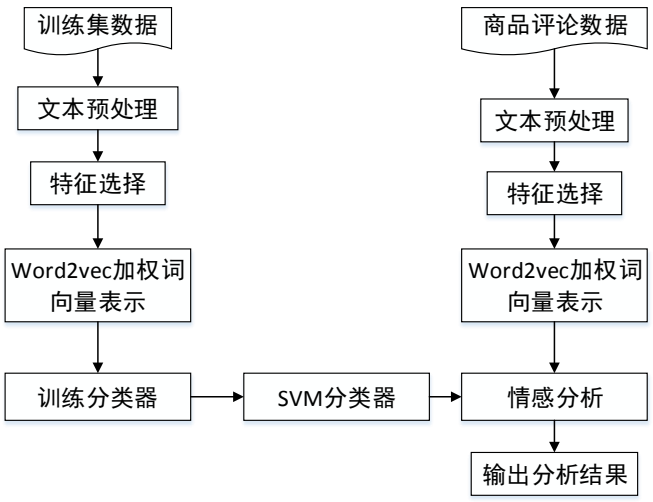

图 4-7 基于 SVM 情感分析流程

Fig. 4-7 emotion analysis process based on SVM

在使用 SVM 进行情感分析时首先需要对 SVM 分类器进行训练，之后利用训练完成的 SVM 分类器对经过特征选择、加权词向量表示的商品评论进行情感分析，具体方法流程如表 4-10 所示。

表 4-10 基于 SVM 的情感分析方法流程

Table 4-10 emotion analysis method flow based on SVM

基于 SVM 的情感分析流程
输入：商品评论文本 C
输出：带有情感倾向标记的商品评论
过程
第一步：调取数据库中的商品评论。
第二步：调用 jieba 分词工具读取自定义词典对商品评论文本进行预处理操作。
第三步：使用 TF-IDF 算法对预处理完成的评论进行特征选择，并保留特征词的 TF-IDF 值。
第四步：读取训练完成的词向量模型，获取特征词对应的词向量，之后根据公式（4-13）计算文本 C 的加权词向量。
第五步：调取 scikit-learn 中的 SVM，将文本的向量化数据加载进 SVM 中，进行情感分析，输出带有情感倾向标记的商品评论。

4.5 本章小结

本章主要对情感分析方法中的情感词典和 SVM 方法进行了研究，并制定了相应的方法流程。在情感词典方法的研究中，本文以知网情感词典为基础词典，通过计算词语相似度的方式，对基础情感词典进行了扩充，构建了专用情感词典，并制定了情感词典方法的评分规则。在 SVM 的研究中，为了构建高质量的训练集数据，本文采用了用户评分与情感词典相结合的训练集的构建方法，在词向量

表示过程中，本文为了更好的在词向量中表现出词语的重要程度，采用了词向量加权的方式。

第 5 章 系统实现与测试

通过前几章的研究，本文将在本章中设计并实现基于情感分析的商品评价系统，通过情感分析为用户从海量的数据中挖掘出有价值的信息，向用户提供客观且全面的购物参考信息。

5.1 系统需求分析

系统的需求分析是在进行系统开发前必不可少的一环。本文将对商品评价系统从功能性和非功能性这两方面进行需求分析。

5.1.1 功能性需求分析

用户在进行网络购物时一般是通过电商网站上的商品图片和产品参数信息来获取一件商品的直观信息，但这些信息一般是电商网站经过处理后的信息，不足以作为进行购物决策时的决定性信息，因此查看商品的评论信息就成为了用户补充商品信息的重要途径，但是随着商品评论的不断累积，就会导致信息过载问题，会增加用户购物的时间成本。因此本系统针对以上问题，从用户的角度出发进行需求分析。

（1）商品的查询

该功能需要系统为用户提供商品的查询功能，用户可以在搜索栏中搜索自己感兴趣的商品，此外考虑到用户在进行查询时可能会忘记自己所要查询商品名的全称问题，或者想要查看某一品牌的全部产品，因此需要在制定搜索策略时增加了模糊查询的功能，用户可以通过输入一些关键字来查询对应的商品，如输入“小米”就可以查询出小米品牌下所有的商品。

（2）商品整体评价信息的查看

该功能要求系统能够为用户提供商品整体评价信息的查看功能，用户可以通过查看该信息就可以了解到该商品的整体情感分析情况，并且用户还可以查看每类情感分析结果中所对应的商品评论。此外系统展示结果的方式应该尽量简洁直观。

（3）商品特征评价信息

除了商品的整体评价信息，用户也对商品的特征感兴趣。这就要求系统能够挖掘出隐藏在商品评论中的商品特征，并对商品特征进行相应的处理。此外在展

示商品特征评价信息时，应考虑到结果的直观性和简洁性。因此对于该结果的展示时应通过视图的方式。

（4）商品信息的查看

该功能需求要求商品评价系统不仅能够提商品的整体评价信息以及商品特征评价信息，也要求系统能够提供商品的详细信息，如该商品的图片、产品参数、商品的售价等，便于用户更加全面的掌握该商品的各项信息。

（5）商品对比功能

该功能需要系统能够为用户提供商品对比功能，用户可以自己输入需要对比的商品信息，通过商品的对比，用户可以直观的了解到两款商品的优缺点，为用户提供更好的购物参考信息。

5.1.2 非功能性需求

除了功能性分析，本系统也进行了非功能需求分析。非功能需求分析包括系统可用性、系统性能以及系统的可扩展性。

（1）可用性需求

这要求系统有较好的可操作性，用户可以像浏览淘宝、京东等网站时那样不需要复杂的操作就能够得到所需的信息。此外系统在设计界面时，应注重界面的简洁性，用户可以通过简单的操作就能获取到对应的商品的各项信息。

（2）性能需求

性能需求是对于系统运行速度的要求。这就要求提高系统的运行速度，在用户与系统进行交互时，避免系统的响应时间过长。

（3）系统可扩展性

系统可扩展性要求系统在进行开发过程中尽量保证系统的各个模块之间的独立性，以保证系统在添加新模块时，通过修改少量或不修改原始代码就可完成。

5.2 系统框架设计

根据之前的需求分析，本文将商品评价系统的整个业务应用划分为经典的三层结构，分别为表示层、业务逻辑层和数据访问层，这三层的软件设计结构有助于降低系统的耦合度，减少模块间的交互。商品评价系统的具体系统框架设计如图 5-1 所示。

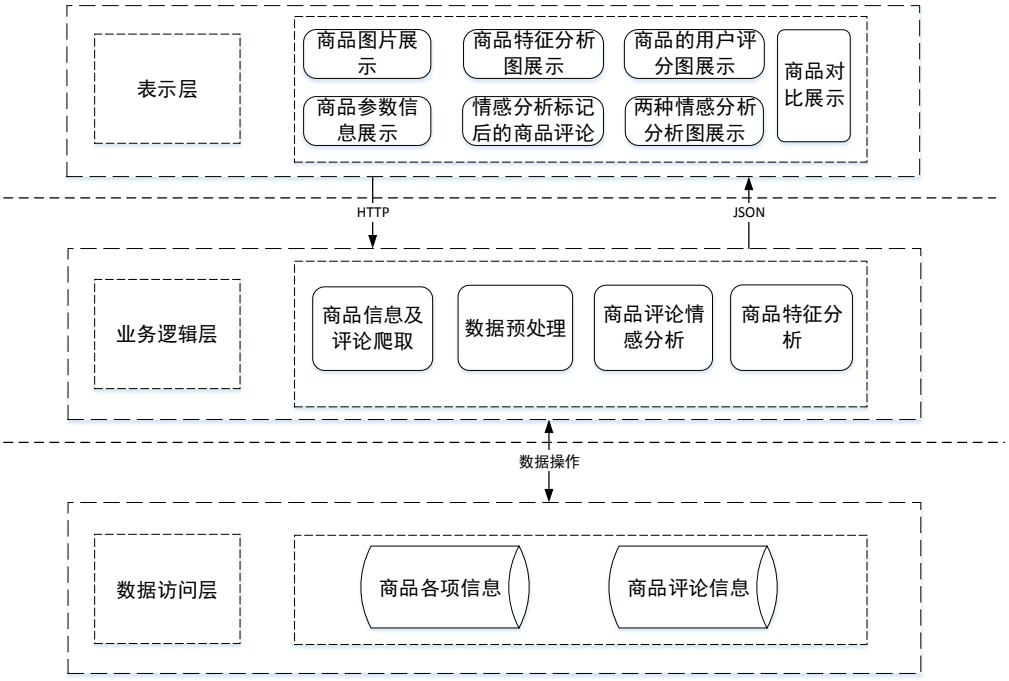

图 5-1 系统框架设计

Fig. 5-1 system framework design

系统的表示层即为系统的 UI 界面，用户可以通过浏览器查看商品评价系统的界面，并且可以通过浏览器与系统前端进行交互，从而查询自己感兴趣的信息，如商品的情感分析饼状图、商品特征分析饼状图、词云分析图、商品详细信息等。系统的表示层本文采用 node.js+vue 框架进行设计，并利用 Echars 实现数据的可视化展示。

业务逻辑层是商品评价系统的核心所在，它负责接收前端发送的 HTTP 请求，并向前端返回相应的 json 数据。业务逻辑层的主要功能是实现商品信息和评论数据的采集、文本数据的预处理、基于情感词典和基于 SVM 的情感分析以及商品特征分析等工作。系统的业务逻辑层主要采用了 python 面向 web 开发的 flask 框架进行研发。

数据访问层是实现商品评价系统的数据储存和数据访问功能，通过与业务逻辑层进行数据交互，实现商品评论信息、商品信息等数据的储存和访问。数据访问层采用的数据库是 MySQL，ORM 框架是 SQLAlchemy。

5.3 系统模块设计

根据之前的需求分析和框架设计，本文将商品评价系统的模块组成进行划分，共分为 4 个模块，分别为数据采集模块、数据预处理模块、情感分析模块和商品特征分析模块。系统的模块设计如图 5-2 所示。

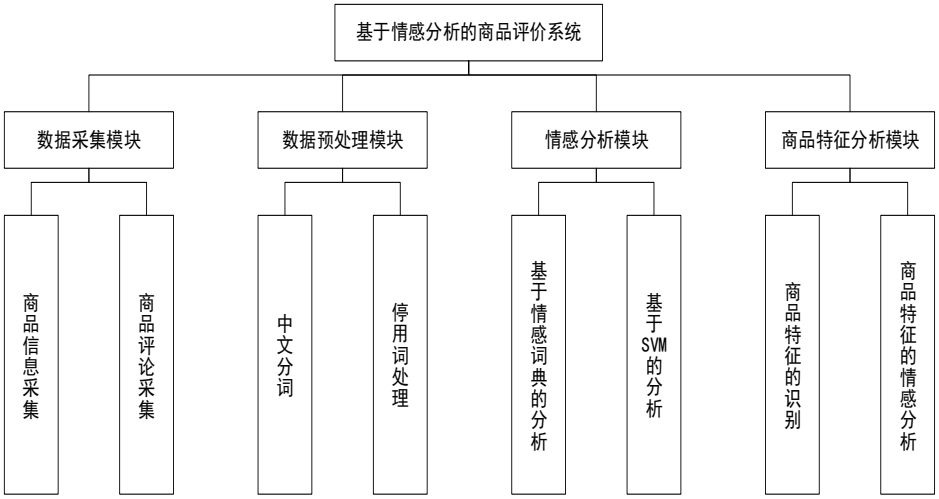

图 5-2 系统模块划分

Fig. 5-2 system module division

（1）数据采集模块。数据采集模块是通过网络爬虫对电商平台进行数据采集，需要采集的数据包括商品的信息（图片、售价、参数等）和商品的评论数据以及对应的用户评分。

（2）数据预处理模块。预处理模块是利用 jieba 分词和停用词词典对商品评论信息进行预处理操作。

（3）情感分析模块。情感分析模块是使用本文设计的基于情感词典和基于 SVM 这两种情感分析方法对商品的评论进行分析，得出评论的情感分析结果。

（4）商品特征分析模块。商品特征分析模块是通过本文构建的商品特征词典对包含有商品特征的评论进行识别分类，之后利用情感分析模块对分类后的商品评论进行分析，进而完成对商品特征的分析。

5.4 数据库结构设计

为了保证情感分析的准确性，本系统收集了大量的评论文本数据用于情感分析过程，数据量巨大且结构复杂，因此，设计好数据库表结构以及商品各项数据的储存方式也是很重要的。本系统的概念模型由以下组成。

（1）商品与商品的评论是一对多的关系，即一件商品能够对应着多条商品的评论，而一条商品评论只能对应着一件商品。

（2）商品与商品信息是一对一的关系，即一件商品对应着一个商品信息，而商品信息也只能对应着一件商品。

本文采用的数据库是 MySQL，该数据库具有轻量、免费开源等优点。数据库表结构的具体定义如下所示。

（1）用于存储手机商品的表为表 goods_phone，表中储存手机的各项信息，

包括手机的 ID、手机的品牌、手机的标题、手机的价格、手机的图片以及手机的参数信息。具体的如表 5-1 所示。

表 5-1 goods_phone 表

Table 5-1 goods_phone Table

Column	Type	Nullable	Describe
phone_id	bigint(20)	NO	手机的 ID
brand	varchar(30)	YES	手机的品牌
title	varchar(255)	YES	手机的标题
pirce	Int(11)	YES	手机的价格
Img_url	varchar(255)	YES	手机的图片
param	longtext	YES	手机的参数信息

（2）用于存储手机评论的表为表 phone_comment，表中储存手机的评论包括，手机评价的 ID、手机的 ID、用户评分、手机的评价。具体的表定义如表 5-2 所示。

表 5-2 phone_comment 表

Table 5-2 phone_comment Table

Column	Type	Nullable	Describe
comment_id	varchar(40)	NO	手机评价的 ID
phone_id	bigint(20)	NO	手机的 ID
score	bigint(20)	YES	用户评分
content	longtext	YES	手机的评价

（3）用于存储笔记本电脑的表为表 goods_computer，表中储存笔记本电脑的各项信息，包括笔记本电脑的 ID、笔记本电脑的品牌、笔记本电脑的标题、笔记本电脑的价格、笔记本电脑的图片以及笔记本电脑的参数信息。具体的如表 5-3 所示。

表 5-3 goods_computer 表

Table 5-3 goods_computer Table

Column	Type	Nullable	Describe
phone_id	bigint(20)	NO	笔记本电脑的 ID
brand	varchar(30)	YES	笔记本电脑的品牌
title	varchar(255)	YES	笔记本电脑的标题
pirce	Int(11)	YES	笔记本电脑的价格
Img_url	varchar(255)	YES	笔记本电脑的图片
param	longtext	YES	笔记本电脑的参数信息

（4）用于存储笔记本电脑评论的表为表 computer_comment，表中储存笔记本电脑的评论包括，笔记本电脑评价的 ID、笔记本电脑的 ID、用户评分、笔记本电脑的评价。具体的表定义如表 5-4 所示。

表 5-4 computer_comment 表
Table 5-4 computer_comment Table

Column	Type	Nullable	Describe
comment_id	varchar(40)	NO	笔记本电脑评价的 ID
phone_id	bigint(20)	NO	笔记本电脑的 ID
score	bigint(20)	YES	用户评分
content	longtext	YES	笔记本电脑的评价

5.5 系统实现

5.5.1 数据采集模块

数据采集工作是商品评价系统的基础，只有采集到的足够的数据才能为情感分析提供足够的样本数据，提高情感分析的表现效果，也能为用户提供丰富的商品信息。数据采集模块的流程如图 5-3 所示。

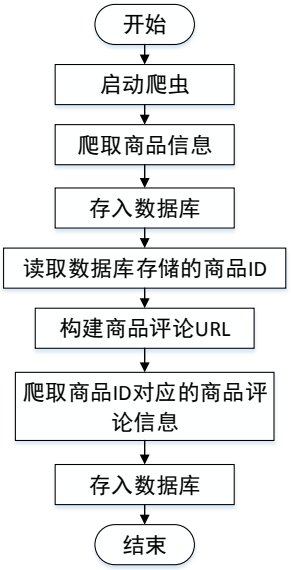

图 5-3 数据采集流程

Fig. 5-3 data collection process

数据的采集工作分为两步进行首先利用网络爬虫获取商品的各项信息，如商品的 ID、图片、参数等，第二步根据商品的 ID 构建存有该商品评论信息的 URL，之后利用网络爬虫采集该商品的评论数据。具体的数据采集过程本文在第三章中进行了详细的描述，在此就不过多赘述。

5.5.2 数据预处理模块

数据预处理模块主要用于对收集到的商品评论数据进行分词、停用词处理等操作。在进行预处理后，评论数据就可以用于情感分析工作。数据预处理的

流程如图 5-4 所示。

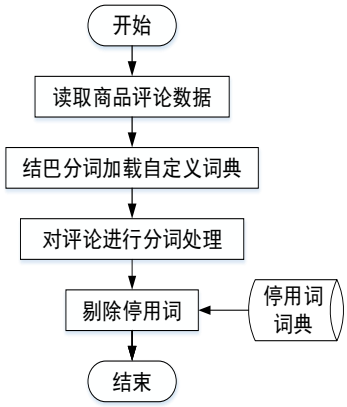

图 5-4 预处理流程

Fig. 5-4 pretreatment process

（1）中文分词

中文分词是文本预处理中最重要的一步操作，其分词的质量会影响到之后的情感分析的结果，所以需要选择一款较为成熟的分词工具来对商品评论数据进行处理。由于本系统是基于 Python 语言进行开发的，所以对比现有的 Python 分词工具，分析各个分词工具的优缺点，本系统采用结巴分词工具对商品评价文本进行分词操作。

由于网络新词的不断出现，这就导致结巴分词在处理一些未收录的新词时会出现词语歧义问题。为了提高结巴分词的准确率，本文通过构建自定义词典的方式，来提高结巴分词的准确率。未收录词语的来源，本文选择搜狗拼音提供的词库，该词库基本涵盖了所有热门词汇，所以可以作为结巴分词自定义词典的词语来源。结巴分词的结果如表 5-5 所示。

表 5-5 结巴分词结果

Table 5-5 result of word segmentation

商品评论	分词结果（未添加自定义词典）	分词结果（添加自定义词典）
钢化膜不错	钢化/膜/不错	钢化膜/不错
不错，五星好评	不错/,/五星/好评	不错/,/五星好评
玩吃鸡无压力	玩/吃/鸡/无/压力	玩/吃鸡/无压力

（2）停用词

在完成商品评论的分词处理后，就可以利用停用词表来对评论进行匹配，剔除停用词。停用词一般指的是评论中出现的代词、地点副词、称谓词、语气词等对本文的研究基本没有帮助的词语，这些词语的存在会增加数据库的负担，不利于数据的检索。因此本文以哈尔滨工业大学停用词表为基础结合评论中出现的一些无意义符号、字母等词汇构建了停用词表，并通过词典匹配的方式对经过处理后的数据进行匹配，剔除掉停用词，提高数据的质量，减少数据库的存储压力。

停用词操作结果如表 5-6 所示。

表 5-6 停用词操作结果

Table 5-6 operation results of stop words

停用词处理前	停用词处理后
收到了，老爸很喜欢	收到，很喜欢
京东物流就是快	物流快
周一收到的，手机不错	收到，手机不错

5.5.3 情感分析模块

情感分析模块是本系统的核心模块，该模块的作用是利用基于情感词典和基于 SVM 这两种情感分析方法对商品评论进行处理，为用户提供商品的情感分析结果。具体的情感分析流程如图 5-5 所示。

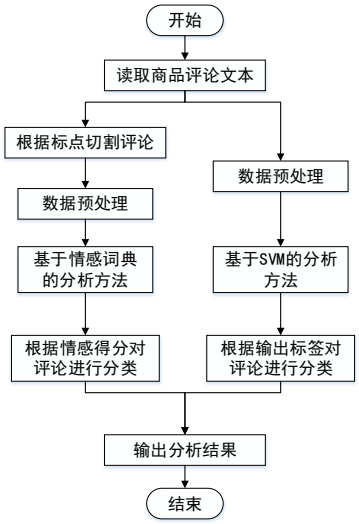

图 5-5 情感分析流程

Fig. 5-5 emotion analysis process

为了向用户提供更加全面且准确的情感分析结果，本文向用户提供了基于情感词典和基于 SVM 这两种情感分析方法，用户可以对照两种方法的分析结果来获取该商品评论的情感分析情况。在这两种情感分析方法中，情感词典的方法是根据评论的情感得分来对商品评论进行分类，如果评论的情感得分大于 0，则该评论是正面评论，反之为负面评论。SVM 的方法是利用分类模型对评论进行二分处理，根据输出的结果标签来对评论进行分类，其中结果标签为 1 是正面评论、0 为负面评论。这两种情感分析方法的具体实现过程本文在第 4 章中进行了详细说明，在此本文就不再赘述。

5.5.4 商品特征分析模块

对于商品评价系统，用户想了解的不仅仅是商品评论的情感分析，也想了解

到商品特征的分析情况。商品特征分析模块是以情感分析模块为基础，利用商品特征词典匹配识别商品评论中的商品特征，之后利用情感分析模块对包含有商品特征的商品评论进行分析，如果该评论是负面评论，则该商品特征也是负面评论，最后统计每类特征的正负评论数量，生成对应的正负比例的饼状图。由于商品评论中可能包含有多个特征词，因此本文以标点符号为分割点，将商品评论分割为数个分句，以此来确保每个分句中最多包含 1 个商品特征词。具体流程如图 5-6 所示。

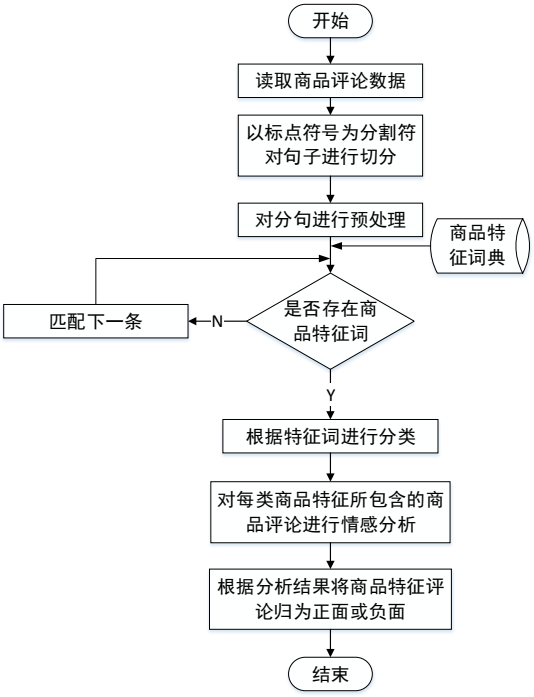

图 5-6 商品特征分析流程

Fig. 5-6 analysis process of commodity characteristics

对于商品特征词典的构建，其详细步骤如下。

第一步，首先参考京东商城上商品参数信息中提供的商品特征，之后通过对商品评论进行词频分析，筛选出用户最感兴趣的商品特征方向，其中手机类商品本文选定了服务、电池、屏幕、性能、相机、外观这 6 类特征，笔记本电脑选定了 CPU、电池、屏幕、性能、显卡、外观这 6 类。

第二步，调取 Word2vec，读取本文在第 4 章中训练完成的词向量文件。

第三步，调用 Word2vec 中的 most_similar 方法，将本文选定的特征词作为目标词汇，计算得出与目标词汇最相似的词语，提取相似词语中的名词。通过这种方式将商品评论中商品特征词的同义词提取出来，构建商品特征词典。商品特征词典如表 5-7 和表 5-8 所示

表 5-7 手机特征词典部分特征词

Table 5-7 some feature words in mobile phone feature dictionary

商品特征类别	特征词
服务	服务、服务态度、态度、服务质量、责任心、售后服务
电池	电池、电量、电池容量、电池电量、电力、待机、耗电量
屏幕	屏幕、画面、画质、触摸屏、屏显、显示屏、图像、分辨率
性能	性能、处理器、配置、开机、反应速度、硬件、流畅度、内存
相机	相机、照片、色彩、夜拍、美颜、闪光灯、摄影机、摄像头
外观	外观、外形、样子、外表、款式、样式、颜色、机身、机器

表 5-8 笔记本电脑特征词典部分特征词

Table 5-8 some characteristic words of notebook computer characteristic dictionary

商品特征类别	特征词
CPU	CPU、处理器、AMD、酷睿、i7、六核、inter、主频
电池	电池、续航、待机、待机时间、电量、容量、耗电量
屏幕	屏幕、显示、画面、图像、显示屏、画质、电脑屏幕
显卡	显卡、GPU、独显、集显、英伟达、显存、核显、核心
性能	性能、功能、效果、表现、流畅度、配置、内存
外观	外观、款式、样子、外表、样式、颜色、金属外壳

5.6 系统界面展示

根据之前的系统设计与实现，本文完成了商品评论系统的开发。在本节中本文将对商品评价系统的界面进行介绍和展示。

（1）搜索界面

在该界面上用户可以根据自己的兴趣对商品进行搜索，该搜索功能支持模糊查询，用户可以通过搜索关键词来检索商品，系统会根据关键词的相似度来为用户返回查询的商品。搜索界面如图 5-7 所示。

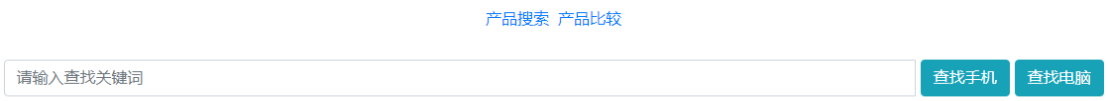

图 5-7 搜索界面

Fig. 5-7 Search interface

（2）查询结果展示界面

该界面是用来显示用户的搜索结果。系统会根据用户的输入，返回多个搜索结果，用户可以翻阅搜索结果，来查找自己感兴趣的商品。搜索结果界面如图 5-8 所示。

图 5-8 搜索结果展示界面

Fig. 5-8 Search results display interface

（3）商品信息展示界面

该界面展示了商品的一些基本信息，如商品的图片、品牌、产品名称等信息。商品信息展示界面如图 5-9 所示。

图 5-9 商品信息展示界面

Fig. 5-9 Commodity information display interface

（4）商品特征分析界面

该界面是用来展示该商品的商品特征分析结果饼状图。商品特征分析饼状图是根据商品特征模块的分析结果利用 echars 进行绘制展示，用户可以通过查看每类商品特征所对应的好评数和差评数，来了解该商品的优缺点。商品特征分析界面如图 5-10 所示。

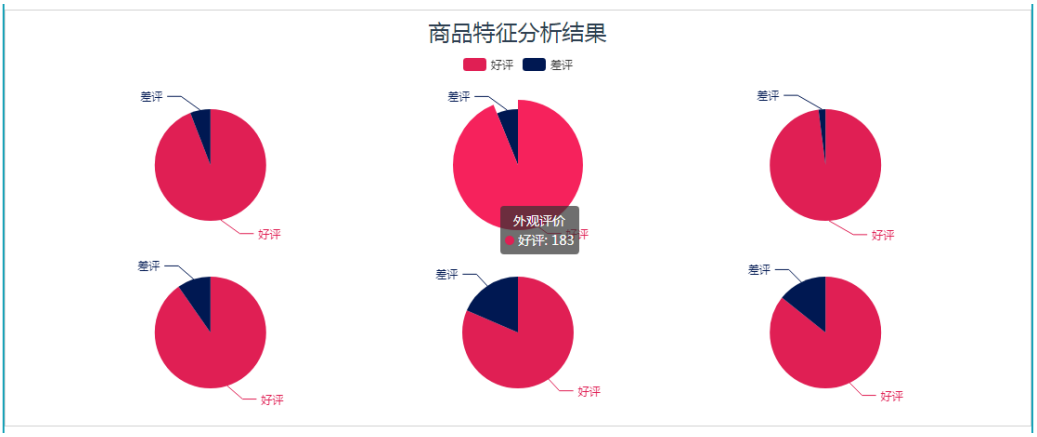

图 5-10 商品特征分析界面

Fig. 5-10 Commodity characteristic analysis interface

（5）情感分析界面

该界面是用来展示商品情感分析结果。其中用户评分饼状图是通过分析用户对该商品的评分情况而生成的，其中评分 5 分和 4 分为用户给出的好评、1 分和 2 分为差评。SVM 情感分析结果饼状图和情感词典分析结果饼状图是根据情感分析模块中的 SVM 方法和情感词典方法对商品评论进行分析后输出的结果生成的，用户可以通过查看饼状图来了解该商品好评和差评的数量和比例，获取该商品的整体评价状况，从而对该商品的整体情况进行评估。情感分析界面如图 5-11 所示。

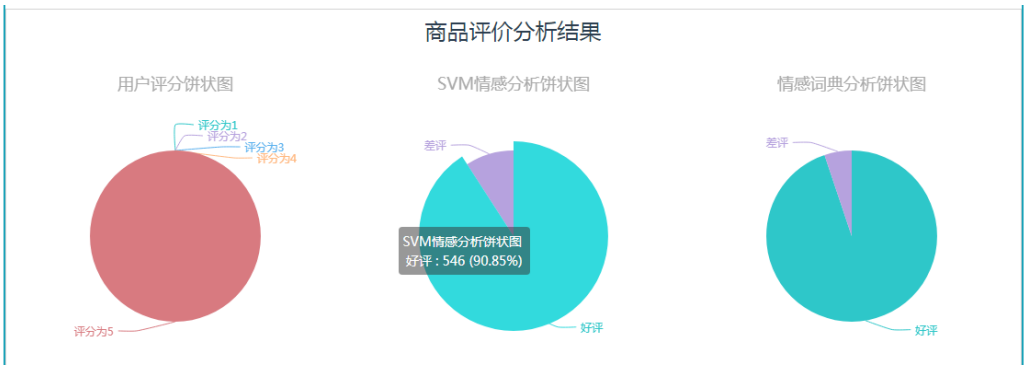

图 5-11 情感分析结果展示界面

Fig. 5-11 Emotional analysis result display interface

（6）词云分析界面

词云分析界面是用来展示该商品的词云分析结果。词云分析结果是根据该商品评论的分词结果，进行统计分析，筛选出商品评论中出现的高频特征词，出现频率越高的特征词其字体也就越大。用户可以通过查看该图来了解商品评论中出现的高频关键词。词云分析界面如图 5-12 所示。

图 5-12 词云分析界面

Fig. 5-12 Word cloud analysis interface

（7）商品评论展示界面

该界面是用来展示经过情感分析方法分类后的商品评论，默认展示的是使用 SVM 进行分析后的分类结果。用户可以分别查看该商品对应的正面评价和负面评价。商品评论展示界面如图 5-13 和 5-14 所示。

图 5-13 积极评论展示界面

Fig. 5-13 Positive comment display interface

图 5-14 消极评论展示界面

Fig. 5-14 Negative comment display interface

（8）商品参数信息界面

该界面是用来展示该商品的详细参数信息，用户可以通过点击按钮来查看该商品的详细参数信息。商品参数信息界面如图 5-15 所示。

图 5-15 商品参数信息展示界面

Fig. 5-15 Commodity parameter information display interface

（9）商品对比界面

该界面是用于向用户提供商品对比功能，用户可以通过输入商品的名称来进行对比，商品的对比结果包括了商品图片、情感分析结果、商品参数信息等。商品对比界面如图 5-16 所示。

图 5-16 商品对比界面

Fig. 5-16 Commodity comparison page

5.7 系统测试

5.7.1 系统整体测试

为了验证系统的整体功能，本文选定了笔记本电脑中的销量不错的戴尔灵越 5370 进行整体测试，并针对该型笔记本电脑进行分析。

从图 5-17 中可以看出该笔记本电脑的整体评价是好评居多，说明该商品的口碑不错，但在 SVM 和情感词典这两种分析方法的结果中，好评率要低于用户评分，这是由于部分用户虽然在评论中表达了对该笔记本电脑的不满之情，但在评分时出于一些原因并没有给出差评，这就导致了在电商平台中，一些商品的用户评分好评率要高于该商品的实际好评率。

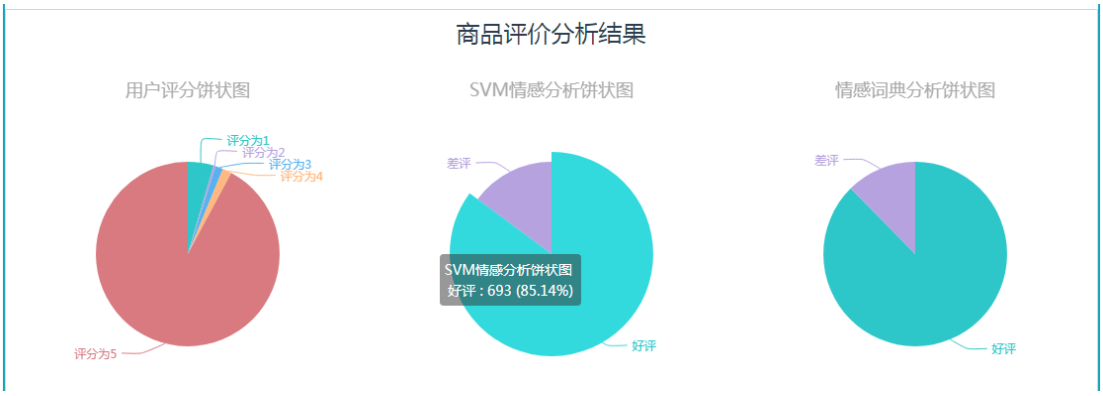

图 5-17 商品评价分析结果

Fig. 5-17 analysis results of commodity evaluation

此外从图 5-18 可以看出该型笔记本电脑的 6 个特征中，用户对电脑的性能最为满意，这也就说明了该商品的优点在于性能方面，但该商品的屏幕方面表现不佳。用户可以根据该商品的优缺点结合商品的整体口碑进行购物决策。

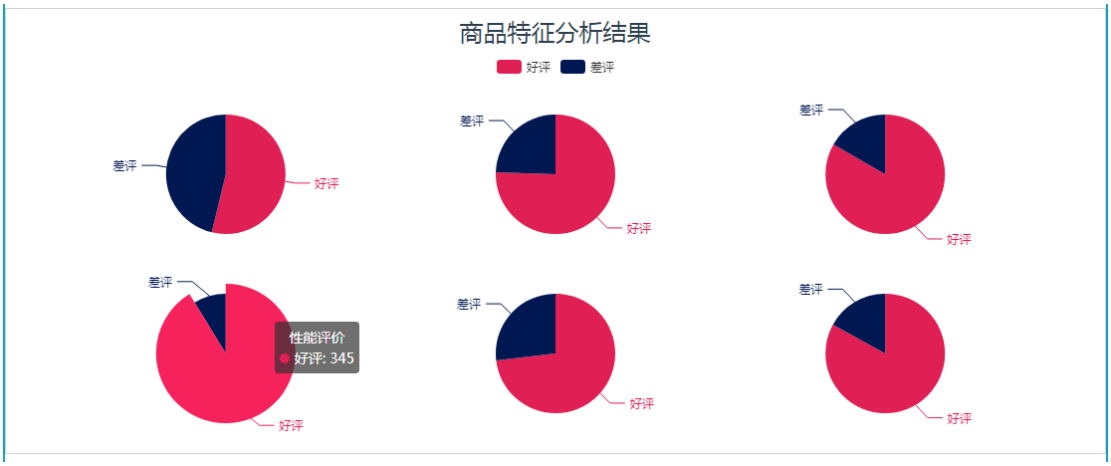

图 5-18 商品特征分析结果

Fig. 5-18 analysis results of commodity characteristics

5.7.2 功能测试

为了对系统的功能进行测试，本文通过黑盒测试的方法对系统测试，通过对系统进行合理的操作，来验证系统是否能够达到预期的效果。具体测试如表 5-9 所示。

表 5-9 功能测试

Table 5-9 function test

编号	测试操作	期望结果	测试结果
1	在商品搜索栏中输入已收录的商品名称进行搜索	系统展示商品的搜索结果	符合预期结果
2	在商品搜索栏中输入未收录的商品名称进行搜索	系统提示暂无该商品信息	符合预期结果
3	查看情感分析结果饼状图和商品特征分析饼状图	饼状图根据选定的区域展示该区域的占比以及该区域所包含的评论数量	符合预期结果
4	查看词云分析图	词云图可以根据选择的词语显示该词语的词频	符合预期结果
5	查看商品评论信息	系统可以根据用户点击的按钮展示出对应的商品评论	符合预期结果
6	查看商品的详细信息	系统可以在点击商品详细按钮后展示出商品的详细信息	符合预期结果
7	在商品对比栏中输入已收录的商品名称进行商品对比	系统根据用户输入的为用户展示相应的对比结果	符合预期结果
8	在商品对比栏中输入未收录的商品名称进行商品对比	系统提示暂无该商品信息	符合预期结果

根据以上的黑盒测试验证了本系统的所有功能运能达到预期效果。

5.7.3 分词实验测试

由于本文采用了基于情感词典和基于 SVM 的情感分析方法，这两种情感分析方法的准确率在一定程度上受分词质量的影响，因此为了提高分词的质量，本文通过添加自定义词典的方式对结巴分词工具进行改进，为了测试改进之后分词的效果，本文分别从手机和笔记本电脑的评论中随机抽取 800 条评论数据，并将这 800 条评论分为 4 组进行分词测试，为了保证测试的效果，本文采用了未使用自定义词典的结巴分词作为对照组进行测试，测试指标采用准确率，准确率的表达式为公式（5-1），测试结果如表 5-10 所示。

$$准确率=\frac{正确分词数}{测试分词总数} \tag{5-1}$$

表 5-10 分词测试

Table 5-10 word segmentation test

实验类别	编号	测试总数	正确数	准确率
实验组（添加自定义词典）	1	200	191	95%
	2	200	188	94%
	3	200	194	97%
	4	200	182	91%
对比组（没有添加自定义词典）	1	200	162	81%
	2	200	158	79%
	3	200	155	77%
	4	200	167	83%

从该表中可以看出经过引入自定义词典后分词的准确率维持在 90%以上，而没有添加自定义词典的对比组其分词的准确率明显低于实验组，证明了本文对结巴分词工具的改进是有效的，足以满意系统对分词的需求。

5.7.4 情感分析实验测试

（1）商品评论情感分析测试。情感分析作为本系统的核心所在，情感分析的准确率决定了本系统分析结果的质量，为了进一步测试本文设计的基于情感词典和基于 SVM 方法的表现情况，本文利用设计完成的网络爬虫从京东商城上爬取手机和笔记本电脑商品的商品评论信息各 1000 条，共计 2000 条评论，其中两类评论的正面评价和负面评论都各为 500 条。将这些评论通过分词、停用词等预处理后，分别利用知网情感词典、本文使用的专用情感词典以及基于 SVM 的分析方法这三种情感分析方法进行处理。本文使用的评估指标为精确率（Precision）、召回率（Recall）以及 F1 值，具体的表达式为公式（5-2）、公式（5-3）以及公式（5-4）。测试结果如表 5-11 和表 5-12 所示。

$$Precision=\frac{TP}{TP+FP} \tag{5-2}$$

式中，TP——将正面评论归为正面评论的数量

FP——将负面评论归为正面评论的数量

$$Recall=\frac{TP}{TP+FN} \tag{5-3}$$

式中，TP——将正面评论归为正面评论的数量

FN ——将正面评论归为负面评论的数量

$$Recall=\frac{2*P*R}{P+R} \tag{5-4}$$

式中，P ——精确率

R ——召回率

表 5-11 情感词典方法测试

Table 5-11 Emotion dictionary method test

分析方法	评论种类	精确率	召回率	F1
知网情感词典	手机	66.31%	66.24%	66.27%
	笔记本电脑	67.51%	68.42%	67.96%
专用情感词典	手机	80.12%	78.93%	79.52%
	笔记本电脑	81.34%	80.44%	80.88%

从表 5-11 中可以看出，本文利用 Word2vec 构建的专用情感词典在精确率、召回率以及 F1 值方面全面优于知网情感词典，这证明了本文采用的词典扩展方法是有效的，在一定程度上提高了情感词典的完备程度，提高了情感词典分析方法的准确率。

表 5-12 SVM 方法测试

Table 5-12 SVM method test

分析方法	评论种类	精确率	召回率	F1
TF-IDF+Word2vec+SVM	手机	82.12%	83.76%	82.93%
	笔记本电脑	83.53%	81.44%	82.47%
TF-IDF+Word2vec 加权	手机	85.41%	86.73%	86.06%
+SVM	笔记本电脑	86.21%	84.53%	85.36%

从表 5-12 中可以看出，本文在基于 SVM 方法中采用的 Word2vec 加权词向量表示方法能够提高 SVM 方法在情感分析中的表现情况，而且与表 5-11 对比可以看出，SVM 方法的表现情况要优于情感词典的表现情况，这是因为情感词典方法除了依赖词典的完备程度，也依赖于评分规则的制定，由于中文语法复杂，这就导致情感词典在面对复杂句式时表现不佳。

整体来看，本文设计的基于情感词典的分析方法和基于 SVM 的分析方法都达到不错的效果，能够满足用户对分析结果的性能要求，证明了本文所设计的系统的有效性。

（2）特征分析实验测试

为了测试商品特征分析的效果，本文从情感分析测试中收集到的评论中，挑选出包含有商品特征的评论 1000 条，其中手机和笔记本电脑各 500 条，利用人工标注的方法将商品评论中的商品特征进行标注。具体的测试结果如表 5-13 所示。

表 5-13 特征分析测试

Table 5-13 characteristic analysis test

分析方法	评论种类	准确率
专用情感词典	手机	77.23%
	笔记本电脑	76.45%
SVM	手机	80.12%
	笔记本电脑	79.55%

从上表中可以看出在进行商品特征分析时，SVM 方法与情感词典的分析方法都能够取得不错的效果。

5.8 本章小结

本章主要对商品评论系统进行设计与实现，之后通过测试验证了系统功能的有效性，并对本文设计的情感分析方法和商品特征分析方法进行了测试，均取得了不错的效果。

结论

随着互联网的不断普及，网络购物这个新兴的产业迅速发展，网络购物也成为了人们在进行购物时的首选之一。网络购物与传统的实体店购物最大的区别就是，用户无法直接接触到实际商品，虽然在各大电商平台上也为用户提供了商品的一些资料如商品的图片、商品的性能参数等，但这些信息往往是商家经过处理后的，可能会与实际不相符。所以查看商品的评论信息就成了用户获取商品信息的重要途径，但随着商品评价的不断累积，用户用于查看商品信息的时间不断增加。

针对商品评论不断累积的问题，重点是从海量的商品评论中提取出有价值的信息，本文通过对情感分析方法的研究，设计并实现了基于情感词典和基于 SVM 的商品评价系统，利用情感分析方法为用户提供直观且全面的购物参考信息。

本文主要工作体现在以下方面：

（1）数据的采集。本文通过对电商平台进行分析，利用 Scrapy 爬虫框架设计并实现了用于数据采集的网络爬虫，并利用设计的网络爬虫获取了大量的数据，为商品评价系统提供了庞大的数据基础。

（2）基于词典的情感分析方法的构建。本文为了保证情感词典的完备程度，以知网情感词典为基础利用 Word2vec 模型，通过计算词语相似度的方法构建了专用情感词典，提高了情感词典的完备程度。此外本文还通过分析评论文本的语法规则，制定了情感分析方法的评分规则。

（3）基于 SVM 的情感分析方法的构建。本文在对基于 SVM 方法的研究中，首先采用词典与用户评分相结合的方法构建了 SVM 的训练集数据，之后本文利用词向量加权的改进方法对 Word2vec 模型生成的词向量进行了改进，使词向量能够表现出词语的重要程度。

（4）商品评价系统的设计与实现。本文使用 Flask+vue 框架设计并实现了面向 web 的商品评价系统，用户可以通过浏览器与系统的 UI 界面进行交互，来了解自己感兴趣的商品，并通过查看系统展示的数据，如商品的情感分析结果、商品的特征分析结果等信息来获取重要的购物参考信息，提升用户的购物体验度。

基于情感分析的系统能够为用户提供各种可视化的商品信息，实现了为用户提供购物参考的功能，但是该系统仍有待完善之处。

首先是专用词典的领域迁移性较差，本文建立起的词典是针对商品领域的，当用于其他领域的分析时，效果不能令人满意，未来可以利用不同领域的情感词

来对情感词典进行扩充，来提高情感词典在不同领域的表现情况。

其次，系统为用户提供的服务还有可扩展的空间，未来可在系统上增加商品推荐功能，通过分析用户的搜索、浏览等记录，为用户推荐符合用户兴趣的高品质商品。

参考文献

[1] 洪巍,李敏.文本情感分析方法研究综述[J].计算机工程与科学,2019,41(04):750-757.

[2] Kramer A D I, Guillory J E, Hancock J T. Experimental evidence of massive-scale emotional contagion through social networks[J]. Proceedings of the National Academy of Sciences, 2014, 111(24): 8788-8790.

[3] Kim S M, Hovy E. Crystal: Analyzing predictive opinions on the web[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007: 1056-1064.

[4] Hu M, Liu B. Mining opinion features in customer reviews[C]//AAAI. 2004, 4(4): 755-760.

[5] 于游,付钰,吴晓平.中文文本分类方法综述[J].网络与信息安全学报,2019,5(05):1-8.

[6] 李继东,王移芝.基于扩展词典与语义规则的中文微博情感分析[J].计算机与现代化,2018(02):89-95.

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[8] Boiy E, Moens M F. A machine learning approach to sentiment analysis in multilingual Web texts[J]. Information retrieval, 2009, 12(5): 526-558.

[9] Kang H, Yoo S J, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews[J]. Expert Systems with Applications, 2012, 39(5): 6000-6010.

[10] Liu S, Li F, Li F, et al. Adaptive co-training SVM for sentiment classification on tweets[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013: 2079-2088.

[11] Dasgupta S, Ng V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification[C]// International Joint Conference on Acl. DBLP, 2009:701-709.

[12] Troussas C, Virvou M, Espinosa K J, et al. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning[C]//IISA 2013. IEEE, 2013: 1-6.

[13] Yih W, Zweig G, Platt J C. Polarity inducing latent semantic analysis[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 1212-1222.

[14] Jain A P, Dandannavar P. Application of machine learning techniques to sentiment analysis[C]//2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE, 2016: 628-632.

[15] Rajput V S, Dubey S M. Stock market sentiment analysis based on machine learning[C]//2016

2nd International Conference on Next Generation Computing Technologies (NGCT). IEEE, 2016: 506-510.

[16] 李婷婷,姬东鸿.基于 SVM 和 CRF 多特征组合的微博情感分析[J].计算机应用研究,2015,32(04):978-981.

[17] 朱梦. 基于机器学习的中文文本分类算法的研究与实现[D].北京:北京邮电大学,2019.

[18] 刘勇,兴艳云.基于改进随机森林算法的文本分类研究与应用[J].计算机系统应用,2019,28(05):220-225..

[19] 陈强,何炎祥,刘续乐,孙松涛,彭敏,李飞.基于句法分析的跨语言情感分析[J].北京大学学报(自然科学版),2014,50(01):55-60.

[20] 黄发良,冯时,王大玲,于戈.基于多特征融合的微博主题情感挖掘[J].计算机学报,2017,40(04):872-888.

[21] 张林,钱冠群,樊卫国,华琨,张莉.轻型评论的情感分析研究[J].软件学报,2014,25(12):2790-2807.

[22] 李琼,陈利.一种改进的支持向量机文本分类方法[J].计算机技术与发展,2015,25(05):78-82.

[23] Esuli A, Sebastiani F. Sentiwordnet: A publicly available lexical resource for opinion mining[C]//LREC. 2006, 6: 417-422.

[24] Feng S, Bose R, Choi Y. Learning general connotation of words using graph-based algorithms[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1092-1103.

[25] Nakagawa T, Inui K, Kurohashi S. Dependency tree-based sentiment classification using CRFs with hidden variables[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 786-794.

[26] Khuc V N, Shivade C, Ramnath R, et al. Towards building large-scale distributed systems for twitter sentiment analysis[C]//Proceedings of the 27th annual ACM symposium on applied computing. ACM, 2012: 459-464.

[27] Zagibalov, Taras, Carroll, et al. Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text[C]// International Conference. 2008. 15(1):1073-1080.

[28] Mudinas A, Zhang D, Levene M. Combining lexicon and learning based approaches for concept-level sentiment analysis[C]//Proceedings of the first international workshop on issues of sentiment discovery and opinion mining. ACM, 2012: 734-740.

[29] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis[C]//Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005: 625-631.

[30] Gyamfi Y, Wiebe J, Mihalcea R, et al. Integrating knowledge for subjectivity sense labeling[C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for

Computational Linguistics, 2009: 10-18.

[31] 黄仁,张卫.基于 word2vec 的互联网商品评论情感倾向研究[J].计算机科学,2016,43(S1):387-389.

[32] 王志涛,於志文,郭斌,路新江.基于词典和规则集的中文微博情感分析[J].计算机工程与应用,2015,51(08):218-225.

[33] 原多多. 产品评论文本的情感分析方法研究[D].甘肃:兰州财经大学,2019.

[34] 刘亚桥,陆向艳,邓凯凯,阮开栋,刘峻.摄影领域评论情感词典构建方法[J].计算机工程与设计,2019,40(10):3037-3042.

[35] 周莉,杨小俪.面向突发事件应急管理的情感词典构建——以“暴雨洪涝”灾害为例[J].武汉理工大学学报(社会科学版),2019,32(04):8-14..

[36] 杨奎,段琼瑾.基于情感词典方法的情感倾向性分析[J].计算机时代,2017(03):10-13.

[37] 蒋翠清,郭轶博,刘尧.基于中文社交媒体文本的领域情感词典构建方法研究[J].数据分析与知识发现,2019,3(02):98-107.

[38] 杨小平,张中夏,王良,张永俊,马奇凤,吴佳楠,张悦.基于 Word2Vec 的情感词典自动构建与优化[J].计算机科学,2017,44(01):42-47+74.

[39] 李明,胡吉霞,侯琳娜,严峻.商品评论情感倾向性分析[J].计算机应用,2019,39(S2):15-19.

[40] 叶锦,彭小江,乔宇,邢昊.基于深度学习的女装图片分类探索[J].集成技术,2019,8(02):1-10.

[41] 於雯,周武能.基于 LSTM 的商品评论情感分析[J].计算机系统应用,2018,27(08):159-163.

[42] 刘智鹏,何中市,何伟东,张航.基于深度学习的商品评价情感分析与研究[J].计算机与数字工程,2018,46(05):921-927.

[43] 张佳悦. 商品评论情感分析技术研究[D].北京:北京交通大学,2018.

[44] 冯仓龙,白宇,蔡东风.面向商品评价的情感要素抽取[J].沈阳航空航天大学学报,2016,33(06):71-76.

[45] 王名扬,吴欢,贾晓婷.结合 word2vec 与扩充情感词典的微博多元情感分类研究[J].东北师大学报(自然科学版),2019,51(01):55-62.

[46] 常丹,王玉珍.基于词典的商品评论情感分析[J].邵阳学院学报(自然科学版),2018,15(05):27-32.

[47] 马明阳,郭明亮,魏留强.网络爬虫的专利技术综述[J].科技视界,2018(22):12-13.

[48] 冯俐.中文分词技术综述[J].现代计算机(专业版),2018(34):17-20.

[49] 梁喜涛,顾磊.中文分词与词性标注研究[J].计算机技术与发展,2015,25(02):175-180.

[50] 吴潇,王磊.基于购物领域词典扩建的评论情感研究[J].计算机技术与发展,2017,27(07):194-199.

[51] 熊富林,邓怡豪,唐晓晟.Word2vec 的核心架构及其应用[J].南京师范大学学报(工程技术版),2015,15(01):43-48.

[52] 李长江. 基于酒店中文评论情感倾向分析[D].广东:华南理工大学,2016.

[53] 吴杰胜,陆奎,王诗兵.基于多部情感词典与 SVM 的电影评论情感分析[J].阜阳师范学院学

报(自然科学版), 2019,36(02):68-72.

[54] 唐明,朱磊,邹显春.基于 Word2Vec 的一种文档向量表示[J].计算机科学,2016,43(06):214-217+269.

[55] 王彦婕. 基于情感分析的汽车推荐系统的设计与实现[D].山西:山西大学,2018.

[56] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3):75-102.

攻读硕士期间发表的文章及参加项目

[1]贾东立,崔新宇,申飞.基于高频情感词扩充情感词典的商品评价系统[J].电脑知识与技术,2019,15(16):242-244.

致谢

时光匆匆而逝，转眼间研究生的学习生活已接近尾声，回首自己的研究生生涯，经历了许多也收获了许多。

首先我要感谢我的导师贾东立老师，贾老师学识渊博，平易近人，学术作风严谨。在我的学习中，贾老师对我悉心教导，从确立研究课题，到开题答辩，再到完成设计和毕业论文的撰写，贾老师一直在我身边进行耐心指导，让我得以顺利完成研究生的学业。在科研态度上，贾老师以严谨的科研态度一直激励我在科研的路上前进。最后感谢贾老师您付出的一切。

同样需要感谢的是我的同学郝光兆、申飞、高鹏、崔益豪等人，你们在我学习与生活中，给予了我无私的帮助，感谢你们的付出，也同样感谢你们陪伴我度过了一个有意义的研究生生涯。

最后我要感谢的是我的父母，是他们在我的背后默默付出，从不要求任何回报。每当遇到挫折，心情低落时，都是父母在为我打气，让我重新振作起来。感谢父母为我做出的一切。

最后我要由衷的感谢在百忙之中评阅此稿并给出修改意见的各位专家和老师，感谢你们为本文提出的宝贵意见和建议，谢谢。

作者简介

基本情况：

姓名：崔新宇

性别：男

出生年月：1994 年 2 月

民族：汉族

籍贯：河北省唐山市

个人简历：2013.9—2017.6 河北工程大学 电子信息工程

2017.9—2020.6 河北工程大学 计算机技术

研究生学业情况：共修了 22 个课程学分，平均绩点（GPA）为 3.09