西安交通大学

硕士学位论文

基于倾向性分析的股票资讯服务系统的设计与实现

学位申请人：马丹阳

指导教师：饶元　副教授

类别（领域）：工程硕士（软件工程）

2017 年 3 月

Design and Implementation of Stock Information Service System Based on Orientation Analysis

A thesis submitted to
Xi'an Jiaotong University
In partial fulfillment of the requirement
for the degree of
Master of Engineering

By
Danyang Ma
Supervisor: Associate Professor Yuan Rao
(Software Engineering)
March 2017

论文题目：基于倾向性分析的股票资讯服务系统的设计与实现
类别（领域）：工程硕士（软件工程）
学位申请人：马丹阳
指导教师：饶　元 副教授

摘　要

自证券市场建立以来，高收益和高风险并存的股票一直是众多投资者关注的对象。随着互联网络平台的快速发展与大数据时代的到来，传统的股票技术指标数据已经不能满足人们分析股票价格走势的需求。由于我国股票市场的弱有效性，投资者的倾向性分析及其对股票的影响越来越引起研究者的重视。

股票评论由于包含大量的专业词汇、特殊的词性特征以及复杂的篇章结构，使用传统倾向性分析方法的计算结果并不理想。针对股评的这些特点，本文提出了一种词典语义和 SVM 相结合的股评倾向性分析方法。模型首先使用半监督的方式构建股票专业词汇词典库，然后提出了一种新的词性情感特征提取方法，对带有情感极性的语料句按词法规则找出所有带有情感倾向的词性特征，并通过词性最大匹配算法依据情感识别准确率和占有率提取词性情感特征。最后将词典语义分析方法与 SVM 模型相结合，得到最终的倾向性识别结果。实验证明，本文提出的改进分析方法查准率达到 91.8%，查全率达到 86.2%，相较传统的词典方法或 SVM 方法有一定方面的提升。在对股评文本的倾向性进行分析之后，本文设计并实现了一个基于股票评价倾向性分析的股票资讯服务系统。该系统可以为用户提供最新的股票资讯，并对各大财经门户的股票评论倾向性进行计算，分析网友的投资意向，从而为用户的股票交易提供参考。系统同时提供模拟交易功能，为用户熟悉交易流程、练习交易策略提供了真实的环境。此外，用户可以在系统的股票论坛中交流交易经验，分享股票新闻。

本文将从系统的需求分析、建模设计、实现测试三个角度进行详细的描述，并对系统按照测试用例进行了测试，给出运行结果。测试结果表明，本系统能够满足用户对股票资讯查询和模拟交易的需求，在证券领域有重要的意义和价值。

关 键 词：股票评论；倾向性分析；词性情感特征
论文类型：应用研究

Title: Design and Implementation of Stock Information Service System Based on Orientation Analysis
Professional Fields: Software Engineering
Applicant: Danyang Ma
Supervisor: Associate Professor Yuan Rao

ABSTRACT

Since the establishment of the securities market, stock with high yield and high risk has been the concerning object of many investors. With the rapid development of Internet platform and the arrival of big data age, the traditional stock technical indicators data can not meet the needs of people to analyze the trend of stock prices. Due to the weak effectiveness of China's stock market, the investor's sentiment analysis and its impact on the stock is attracting increasing attention of researchers.

As stock comments contain large number of professional vocabulary, special part of speech features and complex chapter structure, the results of traditional orientation analysis method is not ideal. Aiming at these characteristics of stock comments, a method of stock orientation analysis combining dictionary semantics and SVM is presented in this thesis. The model first applies the semi-supervised way to construct the stock vocabulary dictionary, and puts forward a new method of POS emotion feature extraction which finds all the POS features that have emotional tendencies in the corpus sentence with emotional polarity by lexical rules, then extracts the POS features according to feature’s occupancy rate and the accuracy and of maximum the POS matching algorithm. Finally, the dictionary semantic analysis method and SVM model are combined to obtain the final sentiment recognition results. Experiments show that the improved method proposed in this thesis can effectively improve the precision and recall rate， which is 91.8% and 86.2%. After analysis of the sentiment of the comment texts, a stock information service system is designed and implemented in this thesis. The system provides users with the latest stock related news, and computes the stock comments orientation of major financial portals, thus analysing investment intention of users to offer a reference for the user's stock trading. In addition, users can exchange trading experience in the system's stock forum and share stock news.

In this thesis, the system is described from aspects of needs analysis, modeling design and implementation, and tested according to the test cases, with running results given. The test results show that the system can meet the needs of users for stock information query and simulation transactions, making certain importantace and value in the securities field.

KEY WORDS: Stock comment; Orientation analysis; Speech emotion feature
TYPE OF THESIS: Application Research

目录

Contents

1 绪论

1.1 选题背景和意义

以 Facebook、Twitter 为代表的各类社交网络成为互联网用户发表、传播、交流信息的重要平台。在我国，微博、贴吧等社交平台因其便携、快捷等特性在网民中迅速传播，尤其在移动互联网的时代浪潮下，大量的信息通过这类平台得到传播。在大数据时代，通过数据挖掘方法对此类平台交互数据传导的情感分析可从侧面反映出舆论倾向。

自证券市场建立以来，作为高风险与高收益并存的股票市场，一直受到众多投资者的关注。股票不仅能给投资者带来个人利益，也为国家的经济发展做出了巨大的贡献。近年来，我国经济飞速发展。股票市场作为国民经济的重要组成部分，是反映经济发展的风向标，股票市场也在互联网技术的支持下取得了很大进步，量化金融、高频交易等新兴概念与技术也不断出现在国内的证券市场。目前，得益于电子交易平台的不断完善和信息技术的发展、市场结构的不断完善和交易品种日趋多样化、计算机自动交易理念和技术的发展，高频交易已经逐渐成为发达国家证券交易的主要手段，而在中国，随着金融市场的不断创新和发展，这些条件也都已逐步具备。

随着中国经济的蓬勃发展，金融领域在国民经济中地位的逐步提高，吸引了更多的学者对金融领域内的微观结构进行理论研究。研究的重点包括，决策者的情绪化行为对他们的投资决定的影响；新闻舆论对股票市场价格的影响；公众面对突发时间的反应规律，以及各类反应如何影响证券市场等等。这些研究包括对投资者们的情感进行分析、对行为进行建模，并寻找数学模型之间的相互关系。这方面的研究有利于发现金融舆论对证券市场的具体影响，挖掘金融市场的内在规律，具有重要的现实意义。

目前，对评论的倾向性研究主要集中在网络产品、电子商城、银行服务、电影评论、旅游评价以及电子产品评测等领域，而专门针对股票评价的倾向性研究还处于较为起步的阶段，因此有很大的研究空间[1]。由于移动网络中信息的迅速传播以及我国股票市场的弱有效性[2]，股民再收到信息影响之后情绪波动会对股票市场有较大的影响，羊群效应引发的市场行为也会对证券市场形成波动[3]。因此，如何挖掘、分析并利用网民的股票评论，如何研究用户情感与行为与股票市场的关系，并在此基础上寻找合理的分析模型，与传统的指数方法结合共同完善对股票的分析与投资决策成为当下的热点问题[4]。这一技术的研究对完善金融学理论，指导证券投资有着重要的实际意义。

1.2 国内外研究现状分析

1.2.1 文本倾向性研究

目前，最为主流的文本倾向性研究方法是基于机器学习的方法和基于语义分析的

方法[5]。基于语义的倾向性算法的主要有两大研究方向。第一种是首先抽取待分析文本中的具有情感含义的词语或短语，然后逐一计算这些词语或短语的倾向性并量化，最后利用所有的倾向值得到文本的篇章级倾向性。该研究方向的代表性工作有 Turney 等人的 SO-PMI 算法[6]。张美娜等人通过对输入文本进行分析，并划分为全文章节、段落、复句、分句五个层次，用文本结构树来表示，给出了标记方法，并实现了文本篇章结构的自动标引[7]。另一个研究方向是预先建立语义模式库，然后将待分类文本中的主观性语句参照模式库做模式匹配，最后通过相关算法获得整个文档的倾向性[8]。该研究方向的代表性工作有 Riloff 等人的工作。乔春庚等人采用模式的方法对句子进行倾向性分析，再对所有句子的倾向性进行累加，得到文本的倾向性。

基于机器学习的算法是将文本倾向性分析的问题转化为传统的文本分类问题，基本思想是人工标注文本的倾向性，然后作为训练集，通过机器学习的方法构造分类器，利用分类器对文本的倾向性进行分类[9]。该研究方向的代表性工作是 Pang 等人的工作。他对比了三种传统的文本分类方法（分别是 Bayesian、SVM、最大熵）在文本的倾向性分类领域的效果，实验结果显示 SVM 分类器的分类结果最优。

以上两类方法应用于金融领域进行倾向性分析，但是效果不佳[10]，原因在于金融领域的文章，尤其是股评，与一般性的网络评论相比，在特点上有所不同。

在金融领域，国内外也进行了文本倾向性分析的相关研究。Tetlock 将基于规则和基于机器学习的情感分析方法相结合，但是他只是将商品评论等领域的倾向性分析方法应用于金融领域内，并没有针对金融的文本特点进行具体的分析[11]。而且，他将重点放在了研究文本的情感波动与股票价格之间的关系，因此在文本倾向性分析上的改进较少[12]。Li 和 Wan 利用语素法，集合 HowNet 对文本进行语义相似度的计算，从而得到金融领域内新词汇的情感特征，并进一步进行金融舆情的分析[13]。他们的方法能够在一定程度上解决金融领域中的新生词汇问题，但是他们的分析方法过于以来具体词汇的分析，缺少对文本在 COAE2011 上整体的把握[14]。

在 COAE2011 上，徐睿峰、王亚伟等使用支持向量机的机器学习方法，先对金融舆情文本进行粗粒度的划分，再对划分的结果赋予权重，综合不同权重的计算结果来对句子的情感倾向进行量化表示[15]。他们考虑的权重计算因子包括句子中情感词的数量、句子中的各类词性、全文的情感倾向等。他们提出的这种方法在会议评测中实现了最优效果，但是该方法没有针对金融领域文本中的句子结构，及情感词的位置进行分析，也忽视了情感词之间的相互修饰关系[16]。

1.2.2 文本倾向性分析在证券领域的应用

目前的证券交易模型主要分为四类[11]：证券投资分析，时间序列分析，非线性系统分析以及组合分析。证券投资分析法[12]和时间序列分析法[13]是比较传统的方法，近年来比较新兴的预测方法集中在非线性系统分析法和组合分析法。其中，支持向量机回归方法是目前最热门应用最多的证券分析模型之一[14]。本文也将在这一算法上进行研究。

目前，国外对文本情感分析用于金融领域的研究已经相对成熟，并且开始出现将Twitter 或 FaceBook 中的金融舆情进行量化、并作为股票分析的模型指标的分析方法[21]。Zhang 等人通过对 Twitter 上预测的道琼斯指数、纳斯达克指数等股票技术指标的预测数据，发现股票的大盘走势与 Twitter 上的情感变化存在关联，并且具体的与道琼斯指数、纳斯达克指数以及标普 500 指数负相关[22]。Bollen 等人利用 OpinionFinder 对 Twitter 中的文本进行情感分析，将分析结果分类为正、负两个维度，同时利用谷歌情感从 Calm，Alert，Sure，Vital Kind 和 Happy 这 6 个情感分类维度分析 Twitter 的文本内容，取得了较为良好的效果[23]。

国内对证券市场的研究分析多采用支持向量机(support vector machines，SVM)方法，因为这种机器学习方法可以适应证券价格非线性变化和随时间变化的两个特点。金桃等人提出了一种基于 SVM 的股票价格预测方法，这种方法结合时间序列算法，引入开盘价、最高价、最低价和收盘价四个技术指标，并将股票舆情作为 SVM 的输入向量之一，实验结果证明这一方法具备良好的泛化能力，具有一定的实用价值[24]。王超等人在支持向量机预测方法中，进一步引入了互联网中社交平台的舆情信息，分析他们的褒贬情况，并通过实验证明了金融市场的波动与社交网络中的时事新闻有紧密的联系。随着研究的不断深入，目前已经有多个分析股票市场投资者情绪的数据来源，包括“央视看盘”、好淡指数、以及许多其他看盘软件等。这些资料依据自身的算法提供了他们的股票相关舆情的倾向性分析，为股民的投资提供了参考。

由于金融市场十分复杂，导致股票的研究本身就成为一个非常困难的任务。股票价格具有随机波动的特点，这使得对于股票的研究变得更加困难。本文采用的研究方式与其他方法不同，不再直接分析股票的大盘趋势和金融舆情之间的关系，而是通过对某只股票的网络评价进行追踪，分析其评价的倾向性，将研究目标细化到某一只股票，继而从投资者情绪的角度进行分析。我们希望以此作为一次十分有价值的实证探索，能够给股市市场的实际研究领域提供新的思路和研究方法。

1.3 论文的主要研究内容

本文以各大门户网站的股评作为大数据分析的切入点，通过对突发事件以及影响个股消息、公告发布后，股评信息传递过程中的情感倾向传导，通过对股评文本数据进行情感计算，得出关于特定个股的情感倾向性。本文对网络股评数据进行数据挖掘，分析舆情情感状态所反应特征对投资者行为影响，并将股评的倾向性分析与股票模拟交易相结合，实现用户的舆情掌握与股票交易的统一。通过以上背景问题以及现状的分析，本文旨在深入研究股票数据情感分析的实际应用，结合证券交易数据的特点与规律，力求寻找出一个较好的结合方法，来实现一个有效、可靠的股票资讯服务系统，为股民进行股票的交易提供有效的参考数据，为相关领域的研究与发展提供具有重要创新意义的理论、方法与技术工具。

本文首先针对现有的文本倾向性分析方法在金融领域的不足进行了全面的分析，

随后提出一种基于倾向性词典和 SVM 相结合的股票评价倾向性分析算法。实验验证算法的有效性之后，通过这一算法，获取股票的网络评价，掌握网民对该股票的情绪观点。随后，本文将计算出的股评舆情值与股票本身的技术指标相结合，作为股票的资讯提供给用户作为股票交易的参考。

本文在上述理论分析与实证研究的基础上，设计并实现一个基于股评倾向性分析的股票资讯服务系统。该设计将严格遵守软件设计规范，进行系统的需求分析，分析系统用例，建立系统静态模型和动态模型。然后在需求分析的基础上完成系统的设计，包括系统的总体结构设计、类的设计、各模块功能的设计、数据库设计和系统界面设计等。最后在需求分析和系统设计的基础上完成系统的实现和测试。本文完成的基于股评倾向性分析的股票资讯服务系统可以为用户提供个股的详细指数与最新的股票新闻，同时为用户提供各大财经门户的股票评论倾向性分析，作为用户交易的参考。系统提供完全仿真的模拟交易功能，方便用户熟悉股票交易流程，演练股票交易策略。同时，系统将提供股票论坛的功能，提升系统的社交属性，方便用户之间交流。

1.4 论文的组织结构

本文一共分五章，各章节内容具体安排如下：

第一章绪论

论述本论文的选题背景以及意义，并调研分析了国内外对文本情感分析和情感分析运用于金融领域的研究现状，简单概括了论文的研究内容，并列出了论文的组织结构。

第二章相关理论以及技术

介绍了本论文涉及到的相关概念和技术，包含本文提出算法的理论支持，金融舆情的介绍，并对本文使用的 SVM 支持向量机做了介绍。

第三章股评倾向性分析研究

本章主要介绍了一种针对股票评价的文本倾向性分析方法。本章首先对研究的主要问题进行描述，指出现有文本倾向性分析方法的问题，然后描述了股评文本的挖掘过程，并就这些问题提出了一种基于倾向性词典和 SVM 相结合的股评文本倾向性分析算法，最后将本章提出的算法与常用 SVM 分析算法进行了实验对比。

第四章基于倾向性分析的股票资讯服务系统的需求分析

从本章开始，将设计并实现一个基于股评倾向性分析的股票资讯服务系统。本章将主要从需求描述和分析模型两个角度进行论述。前者立足全局，将整个系统的框架和设计进行全面的概括；后者则将系统功能细分，详细介绍了系统需要实现的各大功能的具体设计。本章主要使用 UML，从系统的功能模型，静态模型以及行为模型三个角度三个方面对需求设计进行规范化的描述。

第五章基于倾向性分析的股票资讯服务系统的设计

本章将本章将从系统构架设计、系统概要设计、系统详细设计和数据库设计四个

方面对系统设计进行讨论，为系统的实现打下基础。

第六章基于倾向性分析的股票资讯服务系统的实现与测试

本章将基于股票评价的股票资讯服务系统系统的实现环、实现的核心代码、功能截图以及测试过程进行详细的展示，以描述系统的实现与测试。

第七章总结与展望

总结全文，并以本文对股评倾向性及其在股票分析中的应用的研究为出发点，指出了股评分析在股票领域进一步研究和发展的方向。

2 相关理论及技术

本章介绍了本文涉及到的相关理论及技术，包含股票交易的影响因素、常见的舆情事件分类方法与倾向性分析方法，并对本文将涉及到文本情感倾向性分析技术进行了详细的介绍。

2.1 文本倾向性分析常用技术

网络文本的倾向性分析需要先对文本进行获取、分词，随后才是文本的特征处理与倾向性分析。文本的获取需要爬虫技术，分词则涉及到分词技术[25]。常见的倾向性分析技术则包括基于词典的方法、基于语义的方法、机器学习方法等。下面，本文将逐个对这些技术进行介绍。

2.1.1 网络爬虫

网络爬虫的工作原理是从一个或若干初始网页的链接开始进而得到一个链接队列。伴随着网页的抓取又不断从抓取到的网页里抽取新的链接放入到链接队列中，直到爬虫程序停止。其工作流程如下[26]：

1）首先选择需要抓取的网页，作为种子 URL；

2）将需要抓取页面的 URL 组成队列，即待抓取 URL 队列；

3)从上一步由 URL 组成的队列中依次取出待抓取的 URL，解析其中的 DNS 特征，将它们对应的网页下载下来，并按照事先设定好的结构存储进数据库中。已经抓取过的网页 URL 会被放入已抓取 URL 队列。

4）对已经分析过的 URL 队列，分析页面中包含的其它 URL 信息，如果出现没有抓取过的页面，则需要将新的 URL 放入待抓取 URL 队列中进入下一个循环。

一个通用的网络爬虫用的框架如图 2-1 所示[27]：

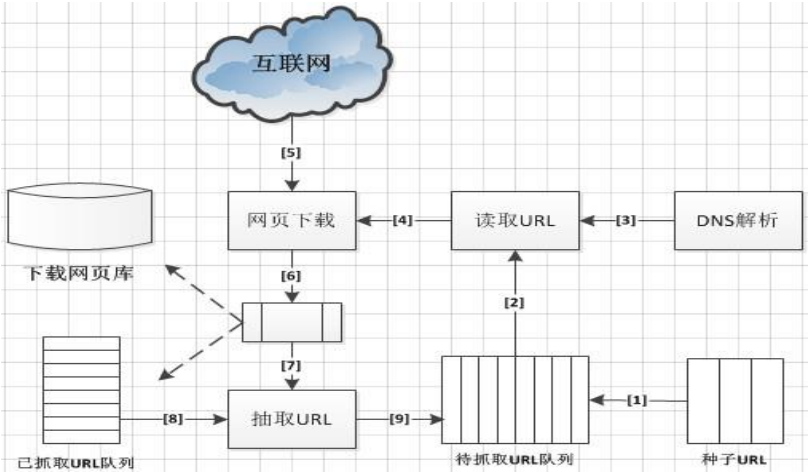


图 2-1　网络爬虫框架图

2.1.2 常用分词技术

中文分词(Chinese Word Segmentation)，是指将一整个篇幅的汉字分割为独立的汉字词汇，或单独的字。分词就是将字的连续序列按照某种规则重新整合成特定的汉字词汇的序列的过程。在英语中，由于英文可以利用空格作为词语词之间的分界，其与汉语是不同的。汉语的最小意义成分是汉字，而汉字所组成的词汇之间并没有分隔符，因此，如何为汉语进行分词是当下非常热门的研究课题。中文分析是中文文本分析与理解的基础与关键，是其他汉语信息处理的前提。目前，最为常见的分词方法主要有以下三类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法[28]。

1）基于字符串匹配的分词方法

基于字符串匹配的方法又叫做机械分词方法，它的基本思想是依据某种策略，将未分析的汉字串按照一定的策略，与一个实现准备好的汉字词典进行逐条匹配，如果再词典中找到了这个字符串，那么就认为匹配是成功的，也就是成功识别出了一个汉字的词汇。基于字符串匹配的分词方法根据扫描方向的不同分为正向匹配和逆向匹配两种方法；按照不同长度优先匹配的情况，则又可以分为最大（最长）匹配和最小（最短）匹配；常用的几种机械分词方法如下：

（1）正向最大匹配法（由左到右的方向）；

（2）逆向最大匹配法（由右到左的方向）；

（3）最少切分（使每一句中切出的词数最小）；

（4）双向最大匹配法（进行由左到右、由右到左两次扫描）

还可以将上述各种方法相互组合，例如，可以将正向最大匹配方法和逆向最大匹配方法结合起来构成双向匹配法。由于汉语单字成词的特点，正向最小匹配和逆向最小匹配一般很少使用。一般说来，逆向匹配的切分精度略高于正向匹配，遇到的歧义现象也较少。统计结果表明，单纯使用正向最大匹配的错误率为 1/169，单纯使用逆向最大匹配的错误率为 1/245。但这种精度还远远不能满足实际的需要。实际使用的分词系统，都是把机械分词作为一种初分手段，还需通过利用各种其它的语言信息来进一步提高切分的准确率。

一种方法是改进扫描方式，称为特征扫描或标志切分，优先在待分析字符串中识别和切分出一些带有明显特征的词，以这些词作为断点，可将原字符串分为较小的串再来进机械分词，从而减少匹配的错误率。另一种方法是将分词和词类标注结合起来，利用丰富的词类信息对分词决策提供帮助，并且在标注过程中又反过来对分词结果进行检验、调整，从而极大地提高切分的准确率。

2）理解法

这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判

断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

3）统计法

从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息。定义两个字的互现信息，计算两个汉字 X、Y 的相邻共现概率。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词。这种方法只需对语料中的字组频度进行统计，不需要切分词典，因而又叫做无词典分词法或统计取词方法。但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”等，并且对常用词的识别精度差，时空开销大。实际应用的统计分词系统都要使用一部基本的分词词典（常用词词典）进行串匹配分词，同时使用统计方法识别一些新的词，即将串频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

2.1.3 常用倾向性分析技术

目前有如下几种常用的倾向性分析技术[29]：

情感极性分类：文本根据才体现的主观情感可分为两种或三种类型，即正向的和负向的，或者正向、负向和中性的。具体到股票评价领域，往往使用看涨、看跌和看平表示这三种观点。根据使用的技术手段的不同，常见的倾向性分析方法可以分为如下三类：

1）基于词典的方法：基于词典的方法在现有的倾向性分析技术中是较为基础的一种。 基于词典的倾向性分析方法需要统计分析文本中正向和负向词汇的数量，并依据这些词汇的特征来进一步判断整个文本的情感。基于词典的倾向性分析方法在进行倾向性分析时，需要进行“情感词典”的构建。情感词典是一个用于进行情感词匹配的词库工具，它事先收录了已经标注好情感的词汇，并在进行倾向性分析时对文本中的词汇进行分词，然后逐一进行匹配。随后面对词汇的情感进行加权计算，就可以得到整个文本的情感状况。目前，常用的情感词典包括英文词典 WoredNet，以及中文词典 HowNet。

2）基于机器学习的方法：基于机器学习的方法是现在最为流行的倾向性分析方法之一。基于机器学习的方法往往需要一些已经标注好情感的文本作为分类算法的训练集，然后用这些训练集进行分类器的训练。当训练完成后，只需要将需要分类的文本使用分类器进一步的分析就可以了。目前常用的机器学习分类方法有支持向量机、朴素贝叶斯、最大熵、以及决策树等。其中，支持向量机（SVM）在实际的测试中往往具有更好的效果，在领域内的文本倾向性分析中常常能达到 80%以上 [30]。

3）基于语义的分析方法：这是一种基于语义分析的倾向性分类方法。基于语义分析的方法需要对已经被赋予标签的文本进行语句组成成分的标注，并在确定好文本语句中的词性之后，对谓语（往往是动词或名词）与句子中其他词汇的关系。目前，比较主流的方法是在已经确定好的句法中，利用及其学习，基于词性的特征向量来进行语义的分析。然后，研究表明，在进行语义标注时，效果非常依赖句法分析的性能，因此改进句法分析方法，提升词性标注的准确率也非常重要。

2.2 基于机器学习的文本倾向性分析方法

所谓基于机器学习的文本倾向性分析方法，是把文本的情感分析作为一种特殊的分类问题。机器学习需要进行的工作包括对文本表示模型进行建立、抽取文本特征、利用特征抽取完成并标注好的预料对分类器进行训练、并最终利用训练完成的分类器对需要进行倾向性分析的文本进行情感分类[31]。下面，将按步骤对基于机器学习的文本倾向性分析方法进行进行介绍。

2.2.1 文本表示模型

基于机器学习的文本倾向性分析算法并不能对文本进行直接的处理，而需要事先将自然语言文本进行转化，使之成为分类算法能够识别、处理的形式，也就是文本表示模型的转化。目前，常用的文本表示模型有布尔模型(Bool Model) 、向量空间模型(Vector Space Model)、概率模型(Probabilistic Model)等等。下面，将对常用的文本表示模型进行介绍。

（1）布尔模型。布尔模型是一种基于布尔代数和集合论的简单检索模型，基于特征项的严格匹配模型，可以看做一种向量模型的特例。布尔模型采用布尔表达式进行文本标识，并定义一个二值函数。当定义的函数表达式的值为1时，函数表达式为真，反之则为0。也就是说，布尔模型的表达形式是0或1的向量集合[32]。

（2）概率模型。概率模型主要基于概率排序理论，该模型会考察用户的查询与自然语言文本相关的概率，并对给定的文本通过概率来判断该文本是否与集合中的其它某种类型的文本相关联。二者之间的概率越大，则两者的相关性就越大。

（3）向量空间模型(VSM)。空间向量模型是目前使用最多的文本表示模型，并且在文本倾向性分类中效果良好。向量空间模型最早在上世纪七十年代由Salton 等人提出。他们认为，将文本内容的处理向向量运算进行转化，并引入特征权重这一数值，将提升计算机识别文本特征的效率。对于任意的自然语言文本，如果这个文本由一些特征项组成，那么每个特征项都表示向量空间的一个维度。文档D_i可以视为一组特征项$(t_1, t_2, .. t_n)$所组成。那么只需要对所有特征项进行特征权重的计算就可以计算出n维向量的特征量。在空间向量模型中，每个特征项都要计算其特征权重，这是为了能够使不同的文本之间有较高的区分度。目前，较为常见的文本特征加权方法有布尔权值，绝对词频、IFIDF权值等[33]。下面，将对每种方法进行简单的介绍：

布尔权值是一种较为简单的特征项权值计算方法。它的思路主要是在特征项t_k出

现在文档D_i中时，就认为权值为1，否则就为0。其公式表示如下：

$$W(t_k, d_i) = \begin{cases} 1 & tf(t_k, d_i) > 0 \\ 0 & tf(t_k, d_i) = 0 \end{cases} \quad (2-1)$$

其中，$tf(t_k, d_i)$用来表示文档d_i中特征项t_k出现的次数，而$W(t_k, d_i)$则表示文档d_i中特征项t_k的权重。

布尔权值在实际的使用中具有一些不足，因此人们使用绝对词频进行改进[34]。布尔权值仅使用0或1判断特征项是否出现，却无法对特征项的重要性进行判断。此时，采用文档c_i中特征项出现的次数作为处理对象特征的权值，也就是当特征项c_i在目标文本中出现的越多，说明在文档c_i中t_k越重要。由于特征项出现的次数与文档篇幅的大小有段，因此目标文档较长时，它的特征项出现的机会较多，特征项的频率就高，也更为重要。反之，文本较短时，特征项就会出现较少。因此，为保证公式的合理性，对$W(t_k, d_i)$的表示做出如下改进：

$$W(t_k, d_i) = \frac{tf(t_k, d_i)}{\sum_{t_k} tf(t_k, d_i)} \quad (2-2)$$

TFIDF权值在上述结论的基础上，进一步考虑整个文本中特征项的分布情况，将特征项t_k在测试文本中出现的频率与整个文本中包含这一特征向的文本数量进行加权，就可以得到。TFIDF的计算公式如下：

$$W(t_k, d_i) = tf(t_k, d_i) \times \log(N / df(t_k, d_i)) \quad (2-3)$$

2.2.2 文本特征提取

在对文本进行建模之后，提取相应的特征项就可以构成初始特征集合。如果每个特征词代表一个特征维度，那么目标文本的特征维度会非常高，因此，需要对文本的特征进行降维。此外，由于自然语言文本的特征本身比较稀疏，如果把所有的特征都考虑在内，进行特征向量的分类，那么计算会非常复杂，效率低且效果差。由上可见，如果想提升文本情感分析的正确率，提高算法的计算效率，就必须对文本的特征进行降维，并去掉出现频率很低的特征。在对分类器进行训练时，我们应该保留关联度高、出现频率高的特征，而对分类效果影响不大的特征进行过滤。通常情况下，我们使用常用的特征提取和特征选择方法来对建模后的文本进行降维，并以此简化算法的复杂度，提升分类器的分类效果[35]。

文本特征提取的方法很多，针对不同的应用领域往往有各自的优化方法。但是它们的基本思路往往较为统一，就是构建一个从复杂的维度向简单的维度进行转换的方法，并通过特征向量维度的降低实现算法复杂度的降低，提升系统的计算速度与分类器分类的准确性。目前，最常见的特征提取方法包括主成分分析、非负矩阵分解以及潜在语义索引三种。

特征的选择方法，是指按照一定的标准，从原始的特征集合中挑选出最具有区分度，对分类效果影响最大的特征项。这需要我们能够减少对文本分类没有帮助的特征，减少无用特征对分类算法的干扰，从而进一步降低向量空间的维度。特征选择的方法

已经被广泛的应用在各个领域，降低了文本分析的计算复杂度，提升了特征对文本语义描述的相关性。目前，常见的特征选择方法包括互信息(Mutual Information)、χ^2统计量(CHI)、信息增益(Information Gain)、以及倒排文档频率(Inverse Document Frequency)和文档频率(Term Frequency)等等。在不同的领域下，他们往往有着不同的应用。下面，本文将对这几种办法进行简单的介绍[36]。

互信息(Mutual Information，MI)通常用来表示不同变量间的相关程度[37]。互信息通过计算类别与特征项之间的相关度来对特征进行选择，并进一步对类别和特征项之间的相关性进行量化表示。一般情况下，互信息越大，类别和它对应的特征项之间的相关性就越高，反之则相关性越小。在使用互信息时，需要预先设定好一个阈值，那么低于该阈值的特征项就需要从文本的特征向量空间中移除，从而降低了文本特征向量的维度。下面将就互信息的计算方法进行简单的介绍：假设A属于文本类型c，且同时属于特征项t的文档频率，B是不属于文本类型c但是属于t的文本频率，C表示属于文本类别c但是不属于特征项t的文档频率，N是所有预料的总和，那么则有t和文本类型c的互信息计算公式如下所示：

$$\mathrm{MI}(\mathrm{t},\mathrm{c}) \approx \log\frac{A\times N}{(A+C)(A+B)} \tag{2-4}$$

互信息计算方法在进行特征选择时充分考虑了频率较低的词汇和稀有词汇的信息，以及他们对文本分类效果的影响，但是这样会导致有些低频词汇的互信息数值比常见词汇的互信息数值还要高，这显然是不合理的。这种现象会导致特征选择函数过滤掉了出现频率较高的普通词汇，而选择互信息高的低频词汇。这就使得特征抽取不具有全局代表性，导致特征选择效果不佳。

信息增益(Information Gain, IG)是一种基于熵的计算的特征选择方法。信息增益是指文本中的特征项在文本中的信息熵之差。信息增益常常用来进行特征中包含的类别信息的度量。文本集合中的一个特征项是否出现会导致整个文本的信息量发生变化，而这个变化中出现的差值就是这个特征项为整个文本造成的信息增益。对于特征项$\mathrm{t_k}$和文本的类别$\mathrm{c_i}$，二者之间的信息增益的计算公式如下所示：

$$\begin{aligned}\mathrm{IG}(t_k) = &-\sum_{i=1}^{m}P(c_i)\log P(c_i) + P(t_k)\sum_{i=1}^{m}P(c_i|t)\log P(c_i|t)\\ &+\ P(\bar{t_k})\sum_{i=1}^{m}P(c_i|\bar{\mathrm{t}})\log P(c_i|\bar{\mathrm{t}})\end{aligned} \tag{2-5}$$

CHI 统计(Chi-Square Statistic, CHI)是目前较为常用的一种特征选择算法。该方法主要度量特征项$\mathrm{t_k}$与文档类别$\mathrm{c_i}$之间的关联程度，并假设$\mathrm{t_k}$和$\mathrm{c_i}$之间与一阶自由度的χ^2分布相符合。如果某个类别下的特定特征的统计值越高，那么这个类别与这个特征值的相关性就越大，这个特征项所包含的信息也就越丰富。如果使用N表示所有文本的总数，A表示属于文本类型$\mathrm{c_i}$而且包含特定特征项$\mathrm{t_k}$的文本的数量，B表示不属于文本类型$\mathrm{c_i}$但是包含特定特征项$\mathrm{t_k}$的文本数量，C表示属于文本类型$\mathrm{c_i}$类却不包含特定特征项

$t_k t_k$的文本数量，D 是既不属于文本类型c_i也不包含特定特征项t_k的文档数量。那么t_k对于c_i的CHI 值的计算公式如下所示：

$$\chi^2(t_k, c_i) = \frac{(A \times D - C \times B)^2}{(A + C)(B + D)(A + B)(C + D)} \quad (2-6)$$

由上，特征项t_k在文本集合中的计算公式如下所示：

$$\chi_{avg}{}^2(t_k) = \sum_i^m P(c_i)\chi^2(t_k, c_i) \quad (2-7)$$

2.2.3 分类模型

基于机器学习的分类模型包括监督式、半监督式和无监督式三种。其中常用的分类方法包括K近邻、神经网络、支持向量机SVM等等。下面，对这几种常见的分类方法进行简单的介绍。

1）K近邻

该方法是利用了各模式类的分布特征，即直接利用各类的概率密度函数、后验概率等，或隐含地利用上述概念进行分类识别。按照判别准则来划分统计分类方法，包括最小误判概率准则和最小损失判决规则等。K 近邻的基本思想如下图 2-2 所示。

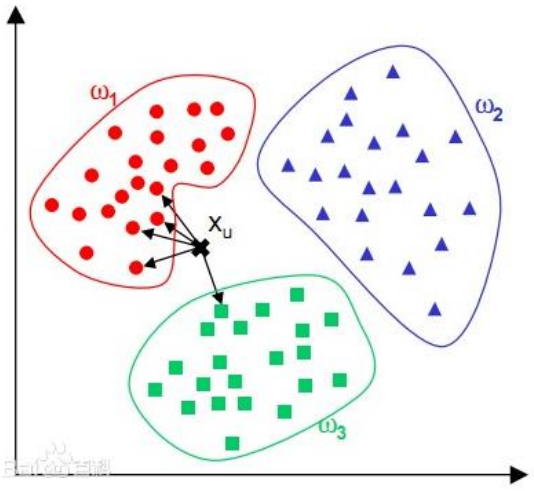


图 2-2 KNN 算法思想示意图

2）神经网络

神经网络算法模拟生物中的神经网络的功能和构成，是由大量的简单基本元件——神经元相互联接组成的自适应系统。它可以充分利用状态信息，对来自于不同状态的信息逐一进行训练而获得某种映射关系。而且网络可以连续学习，如果环境发生改变，这种映射关系还可以自适应地调整。人工神经网络的学习需要遵循特定的准则，其特点有[18]：（1）神经网络具有一定的学习和适应能力，经过训练就可以开发出不同的功能；（2）泛化性：如果样本没有事先的训练，神经网络模型仍然可以进行很好的预测；（3）非线性映射能力：神经网络模型中，用户不需要对系统有完全的掌握，也无需了解其中的技术细节，只需要对输入和输出有规范的定义，就可以进行训练；（4）神经

网络模型是高度并行的。目前使用最多的神经网络是 BP 神经网络。

BP 神经网络是一种人工神经网络，引文全称是 Back Propagation Network，即反馈神经网络。顾名思义，BP 神经网络通过对期望结果的方向传播，实现对神经网络输入层、隐含层和输出层的训练，是一种多层的前馈网络[19]。BP 神经网络是一种有监督的神经网络模型，可以通过定义好的输入模型和输出模型，自动寻找输入与输出之间的映射关系。BP 神经网络的训练无需对中间过程有详细的了解，而是通过反馈计算自动寻找银蛇关系。BP 神经网络在反向传播时，需要以来各个网络层之间的权值和阈值，这些值会在网络初始化时进行随机的分配[20]。BP 神经网络的示意图如下图 2-3 所示：

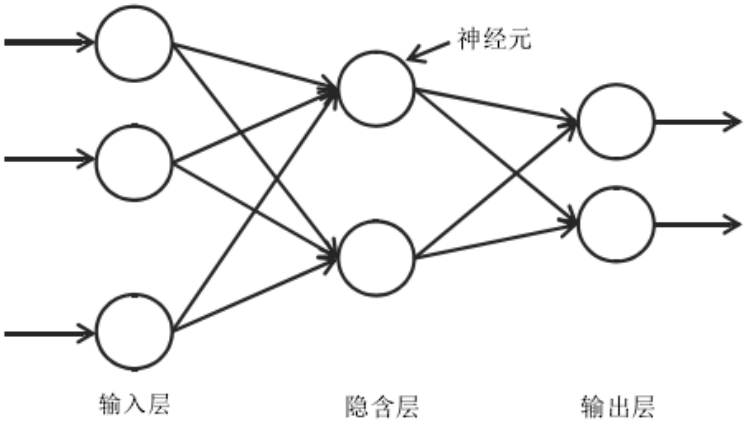


图 2-3　BP 神经网络结构图

3）支持向量机

支持向量机SVM是Vapnik等学者基于统计学理论提出的一种线性分类设计，可以非常准确的解决样本规模较少时的分类问题[38]。

支持向量机在面对两种类别的线性可分问题时，会直接进行最优超平面的构建。最优超平面的构建需要在一定条件的约束下，对二次规划问题进行求解。而这个过程就是最优超平面的构建过程。其最优分类函数如下公式所示：

$$f(x) = Sgn[\sum_{i=1}^{l} a_i y_i k(x_i, x) + b] \qquad (2-8)$$

支持向量机SVM的决策平面允许在一定的区域中进行平行移动。这种移动将不会造成支持向量机分类的误差。因此，在构建SVM分类器时，往往会使用多个决策平面，这些平面的margin值都不一样。而需找最佳的分类方法，就是找到具有最大的margin值的区间，并以这个最优超平面最为SVM分类器的决策平面。假设样本集$(x_i，y_i)$，$i = 1,2,\dots,n, x_i \in R^n, y_i \in \{1,-1\}$，线性判别函数的形式为g（x）=w*x+b，那么分类超平面的表达公式为w*x=0，将函数进行归一后，需要所有的样本在两个类别中都能够满足$|g(x)| \geq 1$，使得距离分类超平面最近的样本的$|g(x)| \geq 1$，则分类间距就是$2/\|w\|$。所以，如果要SVM分类其的分类效果最佳，就要保证$\|w\|$能够最大。所以要满足$\|w\|$能

够最大，从而使得分类效果最佳，有下述公式：

$$y_i[(w \cdot x_i) + b] - 1 \geq 0，\ i = 1,2,\ldots,n \qquad (2-9)$$

当计算公式满足上式，并且||w||的值最小时，SVM分类器的分类超平面就是最优超平面。支持向量机的原理图如下图2-4所示：

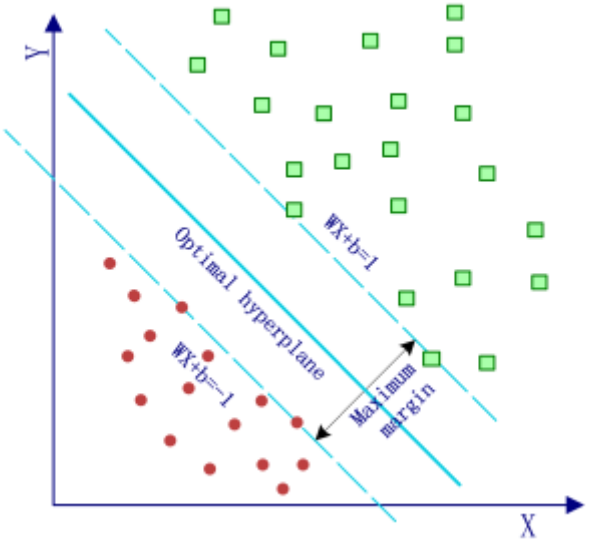


图2-4 支持向量机原理图

2.4 本章小结

本章主要介绍了本文所涉及的相关理论和技术。首先对股票交易的影响因素进行了介绍，包括市场有效性理论和行为金融学的相关概念。然后对文本倾向性分析的常见流程和分析技术进行了简单的阐述。最后，简要介绍了基于机器学习的文本倾向性分类技术。这些技术的调研为算法的改进和优化打下了基础，对于论文算法的研究，以及系统的设计和实现有着十分重要的意义。

3 一种基于词典语义和 SVM 结合的股评倾向性分析算法

上一章主要介绍了本文涉及到的基本概念与关键技术。从本章开始，将介绍一种针对股票评价的文本倾向性分析方法。本章首先对研究的主要问题进行描述，指出现有文本倾向性分析方法的问题，然后描述了股评文本的挖掘过程，并就这些问题提出了一种基于倾向性词典和 SVM 相结合的股评文本倾向性分析算法，最后将本章提出的算法与常用 SVM 分析算法进行了实验对比。

3.1 问题描述

3.1.1 股票评价的特点

目前主流的文本情感分析算法在对股票评论进行分析时效果往往不够理想，这主要是由于金融领域内的文本比起普通的网络文本有着显著的特征差异。这种差异主要体现在以下几个方面：

（1）术语的特定性。金融领域内的文章，尤其是股票评论，往往会包含大量的特定术语，例如对股市状态的描述“牛市”、“熊市”；描述股民投资意向的“补仓”、“减仓”；描述股民观点的“看多”、“看空”等。对专业术语缺乏相关的补充显然会导致情感分析结果不够准确。

（2）倾向性特征词中包含大量的动词。股票评价中，股民往往会使用动词描述自己对特定股票的看法，利用描述自己可能的买入卖出行为来表达对股市走势的预期，或者，运用“上行”、“高走”、“反弹”、“井喷”等词具有积极的含义，即具有看多的倾向，而“下跌”、“下挫”、“低迷”、“走低”、“减仓”等词通常具有看空的倾向。由于无法象对一般性网络评论一样计算情感词的倾向性并量化，因此传统的倾向性分析算法在股票领域的应用效果并不理想。本文所设计的算法通过分析股票评价文本词性、篇章结构的特点，对传统的基于词典的倾向性分析方法进行改进，从而比较准确地判断股票评价倾向性。

（3）内容形式多样，部分篇幅较长。除了网民针对特定股票发表的短评之外，还有股评专家或证券机构发布的大量长篇幅的股评。这些的目的是帮助或者引导投资者，一篇股评文章要想表达出对股市的预测，必然要进行相关说明和论证来证明其预测的准确性，否则投资者将会对其产生不信任感。一篇股评中大部分的篇幅是现状描述，这部分对于观点分类来说往往是噪声数据，如很多股评用大量篇幅描述股市的现状是“下跌”，但是在预测的时候却可能说是“看多”，因此，使用基于模式的单一的方法对股票评价进行分析无法取得较为精确的结果。此外，在对股票评价文本进行倾向性分析时，必须保证能够对股评的客观性陈述和主观性的观点加以区分，才能保证股票评价倾向性分析的准确性。为了能够实现这一目标，在进行文本倾向性分析时需要充分分析股评文本的篇章结构，找准中心观点。

3.1.2 方法概述

基于文本的情感分析涉及到自然语言处理、文本挖掘、数据挖掘、机器学习等多个领域。如图 3-1 所示，一个基本和典型的情感分析问题的解决，包括下面几个步骤:

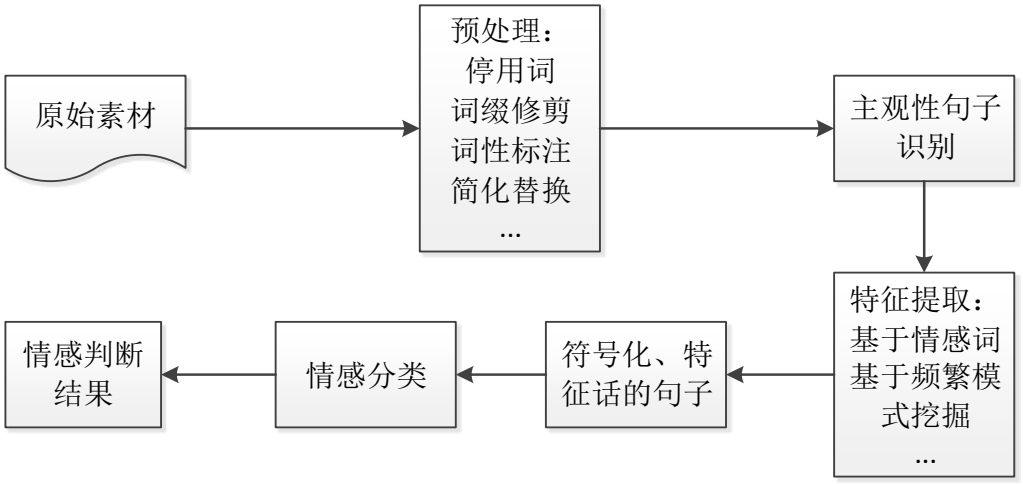


图 3-1 文本情感分析过程图

针对股票评论的词汇和结构特点，相关研究做出了以下几种改进方式：

基于模式的混合股评倾向性分析方法：莫倩使用了一种结合方法，将基于篇章结构的股票评价倾向性分析算法与基于模式的方法进行了结合，从而实现了一种混合的分析方法[39]。通过实验，莫倩提出的这种混合方法能够在各个方面对股票评价的准确性进行提升，包括查准率和查全率。但是，由于没有对金融专业术语进行扩充，也缺少针对性的词性特征分析，这种方法还存在较为明显的不足。

基于词典扩充的股评倾向性分析方法[40]：该方法首先针对股票评论中的专业词汇进行情感词典扩充，然后结合对多种特征选择算法的实现与分析，提出了一种改进的卡方统计法，该算法不仅可以保留卡方统计方法对具有区分度特征识别的优点，而且能够避免低频特征项对系统整体的干扰。实验证明该方法在股评文本的分类中能够较大地提升查准率，但由于缺乏对篇章结构的分析，该方法在查全率上提升较低。

本文结合目前针对股票评论的倾向性分析方法，提出了一种词典语义和 SVM 相结合的股评文本倾向性分析方法。该方法首先使用半监督的方式构建股票专业词汇词典库，然后对传统的词性特征提取方法进行了改进，对预料文本中的所有语句，按照本文所设计的规则进行分析，找出他们的情感极性，并分析所有带有情感极性的词汇的词性特征。随后，通过本文设计的词性匹配算法，根据倾向性分析的准确率，与词汇在文本中的出现比重提取词性的情感特征。最后将词典语义分析方法与 SVM 模型相结合，得到最终的倾向性识别结果。实验证明，本文提出的改进分析方法能够有效提高针对股评的倾向性分析查准率和查全率，尤其在反向查准率上进步显著。下面，将对本算法进行详细的描述。

3.2 算法描述

词典语义和 SVM 结合的分类算法模型主要是先对待分析文本语料进行一次基于情感词典语义的分析，然后就可以得到所有需要分析的文本中的一部分情感倾向性分析的结果。然后，本文设计了一个情感阈值，所有分析结果中情感极性大于该阈值的就被认为是成功分类。将所有成功分类的结果作为 SVM 分类器的训练集，训练完成后在对未能成功运用词典方法分类的文本进行分类。从算法的总体结构上说，本文设计的倾向性分析算法分为文本预处理、基于词典的文本语义分析以及基于 SVM 的倾向性第二部分是使用自己整理的股评情感词库和基于词性情感特征的方法通过特殊情感加权计算，得到部分文本情感分析结果。第三部分是通过使用支持向量机的分类方法对文本情感极性进行分类。分类模型图如下图 3-2 所示：

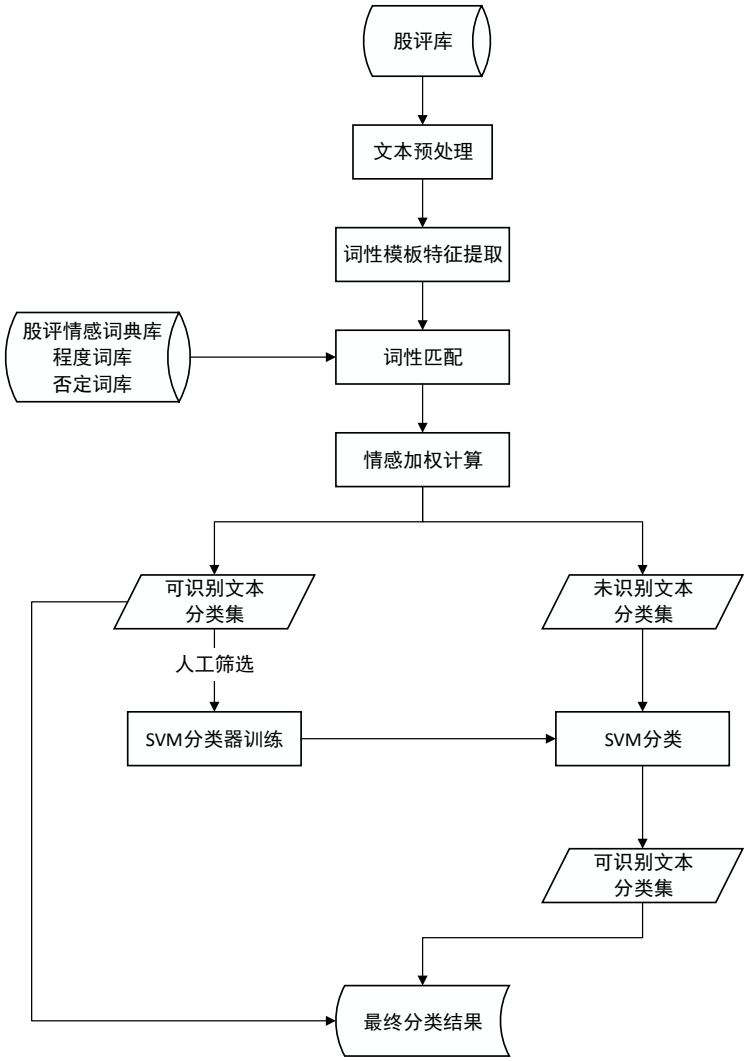


图 3-2 词典语义和 SVM 结合的分类模型图

3.2.1 网络股评倾向性词典库的构建

本节将介绍本文使用的网络股评倾向性词典库构建方法。本文引入从网上查找的种子倾向词，并利用词汇或短语的统计特征，采用半监督的方法构建股评倾向词词典。针对股评倾向词的特点，带倾向性的倾向词多为动词和形容词，故对用 ICTCLAS 完成

分词及词性标注的数据集上，提取形容词和动词，作为倾向性词典候选词，进行下一步处理。本文采用的网络股评倾向性词典库构建方法如下图 3-3 所示：

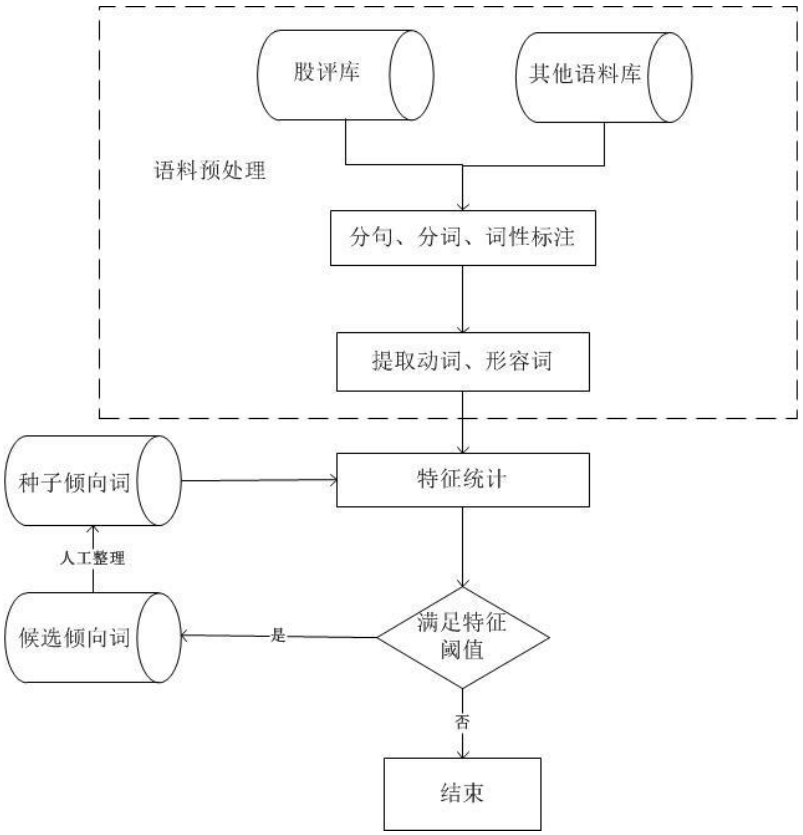


图 3-3 词库构建过程示意图

词典库构建的具体步骤包括语料预处理、特征统计、特征阈值计算等。下面对词典构建过程进行介绍。

1）语料库的处理

利用爬虫从 2015 年 5 月到 2016 年 4 月从多家证券网站采集下来的数据，本文随机选取 5000 篇股评文本作为训练文本，采用如下步骤进行预处理：

（1） 对原始语料进行分句、分词、词性标注，其中使用的是 ICTCLAS 分词程序并在其上做了改进，增加了股评领域的词典。

（2） 提取形容词和动词。

（3） 利用预先整理的转折词词表和并列词词表，识别转折复句和并列复句。

2）特征统计

本文主要利用以下四个词汇特征项进行特征的统计

（1） 词性：除了形容词和动词，对其他词性的词汇暂时不做统计

（2） 词频（TF）：词频指的是某个特定词汇在一个文本中的出现频率。其计算公式如下：

$$TF_w = \log\left(\frac{n_w}{n}\right) \qquad (3-1)$$

其中，n 表示本文中的句子总数，n_w表示语料中出现 w 的句子数。

（3） 反文档频率（IDF）：IDF 是一个词语普遍性和重要性的重要度量。IDF 越小说明该词越普通。其计算公式如下所示：

$$IDF_w = \log\left(\frac{T_w}{T}\right) \quad (3-2)$$

其中T_w表示出现词语 w 的训练文本的数量，T 表示总的训练文本数。

（4） 与种子倾向词汇共同出现次数：利用语料库中的并列复句和转折复句，计算词 w 和种子倾向词库中任意词 s 共同出现的次数，即为 C(w，s)。其中，种子倾向词库中有三类倾向词，分别是看多倾向词、看平倾向词和看空倾向词，即为S_p, S_u, S_n，那么共出现次数可以表示为 C（w，S_p）， C（w，S_u），C（w，S_n）。则有公式：

C(w，s)=Max{ C（w，S_p）， C（w，S_u），C（w，S_n）} (3-3)

3）特征阈值计算

对候选词进行选择评价。针对词 w，确定其词性，计算其词频（TF），反文档频率（IDF），与种子词的共现次数这四个词汇统计特征项。经过测试，当词 w 的词性为形容词或动词，并且 TF、IDF、与种子倾向词的共现次数这三个参数的取值满足如下条件：

a. TF≥0.00035

b. IDF < -0.77

c. 与种子倾向词的共现次数≥3

则将词 w 放入候选倾向词词库，然后通过人工整理和极性确认，将筛选出来的词放入到种子倾向词库中。循环执行直到没有新的候选倾向词出现为止。这样倾向词典就构建成功了。

为了进行股评文本的情感分析，在完成专业词汇的整理与情感词库的建立之外，也需要一个程度词库和一个否定词库，以便在进行倾向性分析时对预料文本中的句子进行情感的加权计算。其中，程度词库主要收录程度副词，这是一种表示词汇程度或范围的修饰词，是倾向性极性偏离程度的表示。当程度副词对情感词进行修饰时，这个词汇组合的情感值将有这两个词共同决定。否定副词是一种表示否定含义的修饰词，当它对情感词进行修饰时表示的是相反的情感含义。

3.2.2 情感特征提取

1）词性情感特征提取方法

基于倾向性词典的分析方法对文本进行情感分析，除了需要一个非常完备的倾向性词典库外，还需要一个准确的倾向性特征提取方法，才能保证文本倾向性分析的准确性。本文针对股评文本中情感词词性的特殊性，选取基于词性特征的提取方法对倾向性特征进行提取，需建立词性特征模板。在股评文本中，影响句子情感倾向的词汇有动词（v）、名词（n）和形容词（a），以及虽然不具有情感倾向但是会对情感词产生影响的修饰词，例如副词（d）或否定词（f）。为了得到句子的准确情感特征，急需要

从文本中提取所有表达情感语义的情感特征块。

通过对现有词性提取文献的分析以及对采集的实验文本中包含情感特征句子的词性规则进行分析，可以将股评文本中的具有情感倾向性的表达方式分为七个类型，24种词性特征，如下表所示：

表 3-1 词性特征分析表

类型	词性特征
Senti-word	a、na、va、vn
Adverb + Senti-word	da、dva、nda、ndv
Neg-word + Senti-word	fa、nfa、fvn、nfv
Adverb + Neg-word + Senti-word	ndfa、dfv、dfa
Neg-word + Adverb + Senti-word	fda、ddv、nfda、nfdv
Adverb + Adverb + Senti-word	dda、ddv、ndda
Neg-word + Neg-word + Senti-word	ffv、nffa

上表中对词性特征的描述可以看出，词性的特征具有包含性，也就是一些特征可能会被另一些特征所包含。然而在进行情感倾向性分析时，一个情感语义单元应当只能够被唯一的特征所提取，因此，本文设计了一个情感特征匹配算法如表3-2所示：

表 3-2 词性特征最大匹配算法伪代码

算法：词性特征最大匹配算法
输入：测试语料集 Corpus，词性特征集 POSFeature
输出：匹配成功的测试语料词性特征集 FeatureSET
步骤：
1. for 1:cNum
2. for POSFeature Length 4:1
3. if Corpus Unit Match Items in POSFeature
4. Add Unit to FeatureSET
5. Delete matched POSFeature
6. return FeatureSET

从伪代码中可以看出，本文设计的词性特征匹配方法主要依靠词性特征的长度进行匹配。当算法匹配到相应的特征时，就从文本中进行抽取。这种方法可以保证情感特征能够最大尺度进行匹配，又能保证不会重复的进行特征提取，同时也能保证所有的情感特征形式得到匹配。本文使用股评库中随机选取的 1000 篇股评作为语料，进行情感特征匹配测试，由于对所有 24 种词性特征进行匹配会导致算法复杂度较高，因此选取了词性模板所占百分比超过 1.5%的所有词性特征进行匹配，对其中占比较高的词性情感特征提取分析如表 3-3，这样所选取的词性情感特征占总词性特征的 97.1%，仍然能够保证较高的情感词性特征的匹配完整度，进而保证情感极性分析的准确度。

表 3-3 部分词性特征提取分析表

词性特征	占比	准确率
da	21.6%	93.4%
va	20.3%	82.5%
na	16.1%	86 %
a	12.7%	92%
vn	8.5%	22.7%
nda	7.4%	96%
dda	3.1%	90.6%
nfa	2.2%	94.7%
dfa	1.9%	86%
dvn	1.7%	30.1%
ndda	1.6%	90.2%

特别需要注意的是，由于股票评价文本的特殊性，动词与其他词语组合的情感倾向性需要进行特别的分析。仅仅通过上述算法进行分析，则vn、dvn两种特征的准确率普遍较低。通过分析词性特征可以发现每一个词性情感特征都会有一个核心词，核心词一般都会有一定的情感极性。所以可以针对不同的词性情感特征项，找到对应的核心词。为此，还需要对这两类词性情感特征添加判断，判断其核心词是否带有情感极性，从而提高词性情感特征匹配的准确度。改进的词性情感特征最大匹配算法代码如下：

表 3-4 改进的词性特征最大匹配算法伪代码

算法：改进词性特征最大匹配算法
输入：测试语料集 Corpus，词性特征集 POSFeature
输出：匹配成功的测试语料词性特征集 FeatureSET
步骤：
1. for 1: cNum
2. for POSFeature Length 4:1
3. if Corpus Unit Match POSFeature Items of vn or dvn
4. if Core Word of Corpus Unit Contains Sentiment
5. Add Unit to FeatureSET
6. Delete matched POSFeature
7. else
8. Add Unit to FeatureSET
9. Delete matched POSFeature
10. return FeatureSET

通过改进的词性情感特征最大匹配算法再次进行测试，可以得到下表 3-5：

表 3-5　改进后的词性情感特征提取分析表

词性特征	改进前准确率	改进后准确率
vn	22.7%	89.3%
dvn	30.1%	94.7%

将该进后的词性匹配算法与改进钱的最大匹配词性特征算法进行比较可以发现，改进后的词性特征最大匹配算法在对相同预料文本进行分析时，特征识别的准确率得到了较大的提升，而此时只要对整理出的情感词典、程度词库、以及否定词库进行进一步的分析就可以得到样本预料的完整倾向性。

2）情感词性特征计算

通过对情感词性特征准确率和占有率的评估分析，总共提取了十一种词性情感特征，每一种词性情感特征都带有情感极性。本文也设计了一系列情感计算公式，用于计算倾向性语义单元的情感值。情感计算公式如下表 3-6 所示：

表 3-6　词性特征计算公式表

情感词性特征	情感计算公式
a	E=e(d)*e(a)
da	E=e(d)*e(a)
va	E= e(a)
na	E= e(a)
vn	E=e(v)|e(n)
nda	E=e(d)*e(a)
dda	E=e(d) *e(d) *e(a)
nfa	E= - e(a)
dfa	E= - e(d)*e(a)
dvn	E= e(d)*(e(v)|e(n))
ndda	E=e(d)*e(d)*e(a)

3.2.3 情感语义加权

基于倾向性词典的文本倾向性分析，最后一步就是对提取的倾向特征进行计算。每条需要分析的文本预料都由多个句子组成，而每个句子又有多个情感词组成，所以在对整个文本的倾向性进行计算时，只需要对它们进行加权就可以得到股评文本的倾向性。本文使用的加权计算公式如公式 3-4 所示。

其中 pol 表示句子 t 的情感倾向性。 E（t_i）表示句子中的情感特征t_i的情感极性值。对于 E（t_i）的情感值计算公式已经在上文表中给出。P(t)表示整个文本中的正向情感数值的大小，N(t)是指预料文本中的负向情感数值的大小。Pos 表示整个文本中，每个句子的正向情感数值的大小，neg 是指预料文本中每个句子的负向情感数值的大小，mid 则表示文本预料中的句子包含的正向情感和负向情感值是相等的，也就是句子的倾向性是中性。在股评文本中，也就是看平。随后，将之前分析过的情感词性特征带入

的公式中，就可以得到整个测试文本的倾向性。

$$\mathrm{pol}(t)=\begin{cases} pos: & P(t)-N(t) & P(t)>N(t) \\ neg: & N(t)-P(t) & N(t)<P(t) \\ mid: & 0 & P(t)=N(t) \end{cases}$$

$$P(t)=\sum_{i=0}^{m}E(t_i)>0 \qquad (3-4)$$

$$N(t)=-\sum_{i=0}^{m}E(t_i)<0$$

3.2.4 基于词典语义和 SVM 结合的股评倾向性分析算法流程

采用改进的情感词典分析算法可以较好的对专业术语以及词性的特点进行分析，但是还不能够解决专业股评中，篇幅较长，情感词极多且较为复杂的问题。在这种情况下，本文将情感词典分析方法与机器学习方法结合起来，提出一种基于词典语义和 SVM 结合的股评倾向性分析算法，以提升文本情感分析的分类效果。

本文所设计的基于词典语义和SVM相结合的股票评价倾向性分析算法在设计上需要首先使用基于词典的情感分析方法进行预料的分析。由于并非所有文本都能通过基于词典的方法进行准确分析，因此该方法只能得到部分高准确率的文本情感。随后，通过本文设计的情感精度评价公式，可以进一步进行分析。计算公式如下：

$$\mathrm{pol}(t)=\begin{cases} pos: & P(t)-N(t) & P(t)-N(t)>0.5\cup N(t)=0 \\ neg: & N(t)-P(t) & N(t)-P(t)>0.5\cup P(t)=0 \end{cases}$$

$$P(t)=\sum_{i=0}^{m}E(t_i)>0 \qquad (3-5)$$

$$N(t)=-\sum_{i=0}^{m}E(t_i)<0$$

由算法公式可知，只要当分析句子中两者的倾向性取值相差超过阈值0.5或者句子中只含有一种倾向性时可以计算出句子的倾向性，而其他的情况都视为不能识别句子的倾向性。这种方法的目标是能够利用阈值的计算提高倾向性计算的准确率和召回率。首先，通过计算出的阈值提升基于词典语义的识别方法的准确率，这样就可以减少对SVM进行训练时人工标注的成本。这种方式由于在股票评价领域进行了专门的优化，因此进一步优化了SVM计算的结果。算法的流程图如下图3-4所示。

从图中可以看出，本文设计的基于词典语义和SVM相结合的股票评价分析算法，是一种综合了词典语义情感分析方法和SVM机器学习情感分析方法的改进算法。本文提出的股评倾向性分析方法基于股票评价专业词汇多、词性复杂、篇章结构不易分析的特点，利用扩充后的股票情感词库，提出了一种改进的词性特征最大匹配算法，将股评文本的词性特征进行分析、抽取，从而充分体现了股票评价文本的词性特点。随后，通过阈值的设置，优化了SVM的训练过程，减少了人力成本。

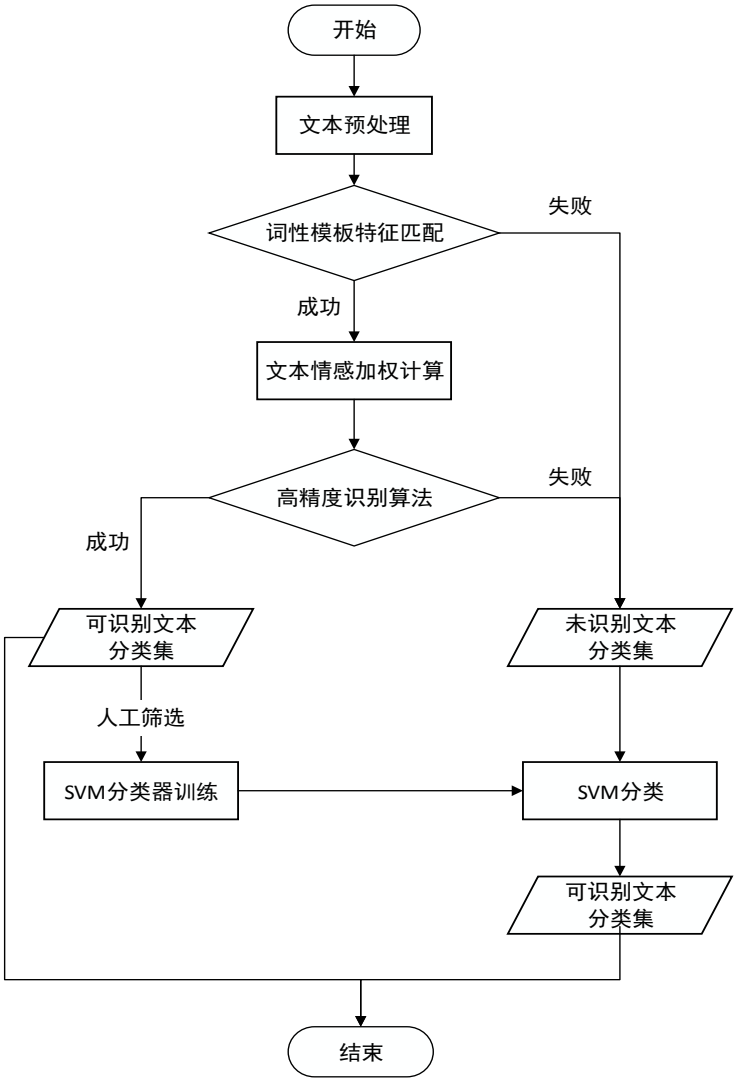


图 3-4　词典语义和 SVM 结合的算法流程图

3.3 实验结果与分析

由于目前针对网络股票评论的语料还没有一个公开标注的数据集可以直接使用，因此本文选用了来自东方财富网、和讯网、新浪财经等热门股票评论网站的股评文章，作为本次实验的语料。为了保证实验的数据集一致，本文将主要就提出的改进算法与改进前的算法进行对比。

3.4.1 实验性能指标

实验结果在进行评价时，指标的对比应当全面、客观、公正，才能保证实验结果的分析能够正确反映算法的价值。通常情况下，进行文本倾向性分析的算法实验需要考虑的指标包括准确率（查全率）、召回率（查准率）、以及 F 值三种。下面，用公式说明这三个评价指标的的计算方法。其中 A 代表算法能够准确分类的文本数量，B 代表算法错误分类到某个类别的文本数量，C 代表分类算法没有将文本进行准确分类的数量。

准确率，也叫查全率，通常记为 P。准确率用于判断算法在分类时的准确程度。准确率的计算公式如下所示：

$$\mathrm{P}=\frac{A}{A+B}\times 100\% \qquad (3-6)$$

召回率，也叫查准率，通常记为 R。召回率用于判断算法在分类时的完备程度，其计算公式如下所示：

$$\mathrm{R}=\frac{A}{A+C}\times 100\% \qquad (3-7)$$

F值记为F，是为了寻求查准率和召回率之间的平衡而引入的指标。这是因为在实际的算法应用中，当查全率高时，召回率会低一些，反之当召回率高的时候，查全率会低一些。因此，引入了F值。其计算公式如下：

$$\mathrm{F}=\frac{2PR}{P+R} \qquad (3-8)$$

3.4.2 实验设计与结果分析

本章的实验主要针对两个方面，一个是改进后的词典分析算法在词性特征分析上是否有所改进；另一个则是基于倾向性词典和 SVM 相结合的股评文本倾向性分析算法与传统 SVM 算法，以及传统词典分析算法的比较。

1）词性特征的情感分析实验

在此次算法比较中，PP 表示正向的正确率，RP 表示正向的召回率，FP 表示正向的 F 值，PN 表示反向的准确率，RN 表示反向的召回率，FN 表示反向的 F 值，F 表示正向和反向的平均 F 值。在词性特征的提取方法上，最为常见的方法词性模板提取特征方法和词性特征组合方法。下面的实验即是分别使用这两种方法对 3.2 中股评库的 1000 篇股评进行词性特征识别，并进行上述参数的比较的结果。其实验结果数据如下表 3-7 所示：

表 3-7 基于不同词性特征的情感分析实验结果数据

特征方法	PP(%)	RP(%)	FP(%)	PN(%)	RN(%)	FN(%)	F(%)
词性模板	90.1	49.6	64	92.7	36.2	52.1	58.1
特征组合	91.4	66.3	76.9	91.5	49.7	64.4	70.7
本文方法	92.3	71.8	80.8	93.2	56.9	70.7	75.8

通过表 3-7 的数据可以发现，基于词性模板的方法在对股票评价文本进行分析时，反向识别的查准和查全率都非常低，导致 F 值也不够理想，总的情感极性识别结果较差。而基于词性组合的分析方法比词性模板方法有着较大提升，尤其是反向的查准率和查全率都进步明显。本文的算法最为显著的优势在反向的查全率上，由于采用了针对股评文本的分析方法，因此在股票评价领域提升较大。

2）基于倾向性词典和 SVM 相结合的股评文本倾向性分析算法实验

为了研究基于倾向性词典和 SVM 相结合的股评文本倾向性分析算法在股票评价

领域内是否能够对最终的分析结果有所提升，本文首先选取股评库中 3000 篇股评进行基于词典的倾向性分析，并对分类成功的文本进行人工标注，选择其中分类准确的 2140 篇作为 SVM 的训练集。随后选取了 160 篇股票专家评论作为样本进行测试。其中看多股票 95 篇，看平股票 32 篇，看空股票 33 篇，其实验结果如下表 3-8 所示：

表 3-8 股票评论的倾向性分析对比实验结果数据

分析方法	查准率	查全率	F 值
基于词典的分析方法	89.3%	61.6%	72.8%
传统 SVM 分析方法	81.2%	85.6%	83.3%
本文方法	91.8%	86.2%	88.9%

将实验数据用柱状图表示如下图 3-5 所示：

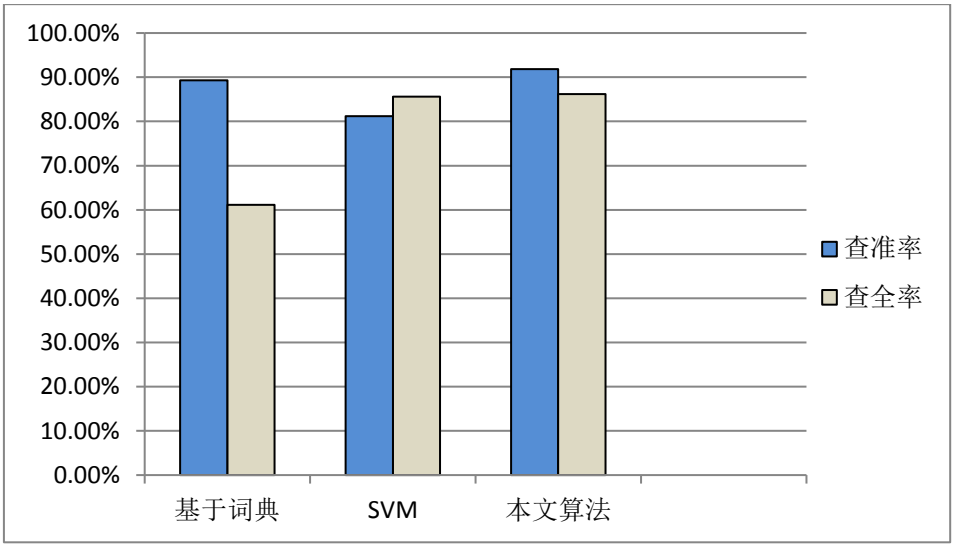


图 3-5 股票评论的倾向性分析对比实验结果数据

从柱状图中可以发现，本文改进的算法在针对股评文本的查准和查全率上都有着显著的提升。尤其在引入了针对股评的术语词典和词性分析改进之后，查全率得到了很大的提升。

随后，本文选取了 3000 条股评，此次的股评包含了大量网友评论，而不再是长篇幅的专家股评。其中看多、看平、看空的评论各 1000 条，并一次作为测试样本进行了股票评论的倾向性分析。最终的分析结果如下表 3-9 所示：

表 3-9 股票评论的倾向性分析实验结果数据

	看多	看平	看空	查准率
看多	830	27	45	92.1%
看平	65	703	18	89.4%
看空	43	51	811	90.6%
查全率	83%	70.3%	81.1%	

从实验数据中可以看出，针对一般性的股评本算法的查准率仍保持了较高的水平，

但在看平股评的查全率上仍然存在不足，这是由于网民的发言中往往带有很多口语化、网络化的表达内容，这仍需后续的工作中进一步的研究。

3.4 本章小结

将介绍一种针对股票评价的文本倾向性分析方法。本章首先对研究的主要问题进行描述，指出现有文本倾向性分析方法在股票评价领域存在的问题，然后分析了股评文本的特征，并就这些特征提出了一种基于倾向性词典和 SVM 相结合的股评文本倾向性分析算法，最后将本章提出的算法与常用 SVM 分析算法进行了实验对比。本章为后续进一步的研究打好了基础。

4 基于倾向性分析的股票资讯服务系统需求分析

上一章主要介绍了本系统涉及到的基本概念与关键技术。从本章开始，将对该系统的框架结构、建模设计、实现测试进行详细介绍。在本章，主要从需求描述和分析模型两个角度进行论述。前者立足全局，将整个系统的框架和设计进行全面的概括；后者则将系统功能细分，详细介绍了系统需要实现的各大功能的具体设计。本章主要使用 UML，从系统的功能模型，静态模型以及行为模型三个角度三个方面对需求设计进行规范化的描述。

4.1 基于倾向性分析的股票资讯服务系统需求描述

在第二章中我们已经介绍过，股票市场的交易情况会受到股民情绪的影响。因此，分析网络股评的倾向性，掌握股民的购买意向，对指导个人投资有重要的参考意义。然而，大数据时代，信息爆发性增长，股民在“信息过载”的浪潮中很难方便快捷地获取自己最感兴趣的资讯，而现有的财经网站和股票门户上，也很少提供股票评论倾向性分析的服务。

股票资讯服务系统基于实时的股票数据，能够为用户提供真实的股票操作环境，熟悉股票交易流程，演练股票买卖策略，已经是最为常用的股票工具之一。然而，目前的股票资讯服务系统与财经信息之间往往是分散的，用户无法在进行模拟交易的同时参考最新的股票资讯。如上文所说，一个准确的股票评论倾向性分析可以为股民的决策提供有力的参考，而模拟交易系统则为用户提供了实战演练的平台。本文秉承个性化、实用性的原则，集合股票评论文本的倾向性分析，实现一个简单易操作、可以实时获取股票数据、结果可视化，并同时能为用户提供个性化的股票资讯与股评分析的股票模拟交易平台。

本文的目标是建立一个基于倾向性分析的股票资讯服务系统。用户在这里可以就自己感兴趣的个股进行股票舆情的查询，了解最新的相关资讯。同时，本系统提供一定的个性化功能，用户可以查看自己关注的财经网站的相关信息，包括头条新闻或个股资讯。此外，系统会就主要门户网站与股吧的股评进行倾向性分析，因此，用户可以很方便的了解到自己感兴趣的股评意向。作为股票资讯服务系统，用户可以在该系统中进行股票的买卖、复盘等操作，实现对真实股票交易的模拟。本系统也提供股票论坛的功能，用户可以在论坛中发表自己的股评，也可以就交易心得、最新的消息等等进行讨论。下面，本文将从系统的功能模型、静态模型和行为模型三个角度对系统进行需求分析。

4.2 基于倾向性分析的股票资讯服务系统的分析模型

本文设计的基于倾向性分析的股票资讯服务系统将主要包括三种用户，即游客、会员以及管理员用户。下面的系统需求分析中，将先以用户角色的角度进行粗粒度的功能

划分，再按主要功能，结合用例图进行详细的功能模型论述。

4.2.1 基于倾向性分析的股票资讯服务系统的功能模型

1）系统总体需求

本文的总体需求分析将从用户角色的角度进行描述。本文主要分为游客、会员和管理员三种角色。其中，游客可以浏览并查询系统中的股票资讯，不仅是股票的相关新闻、股评等，也包括个股的技术数据等内容。会员需要登陆系统，并且除了股票资讯外，还可以进行股票模拟交易操作，也可以使用股票论坛的相关功能。会员也可以对个人信息进行管理，包括密码、用户名等。将系统功能按照角色进行划分，可以得到如下图 4-1 的系统游客和会员角色用例图：

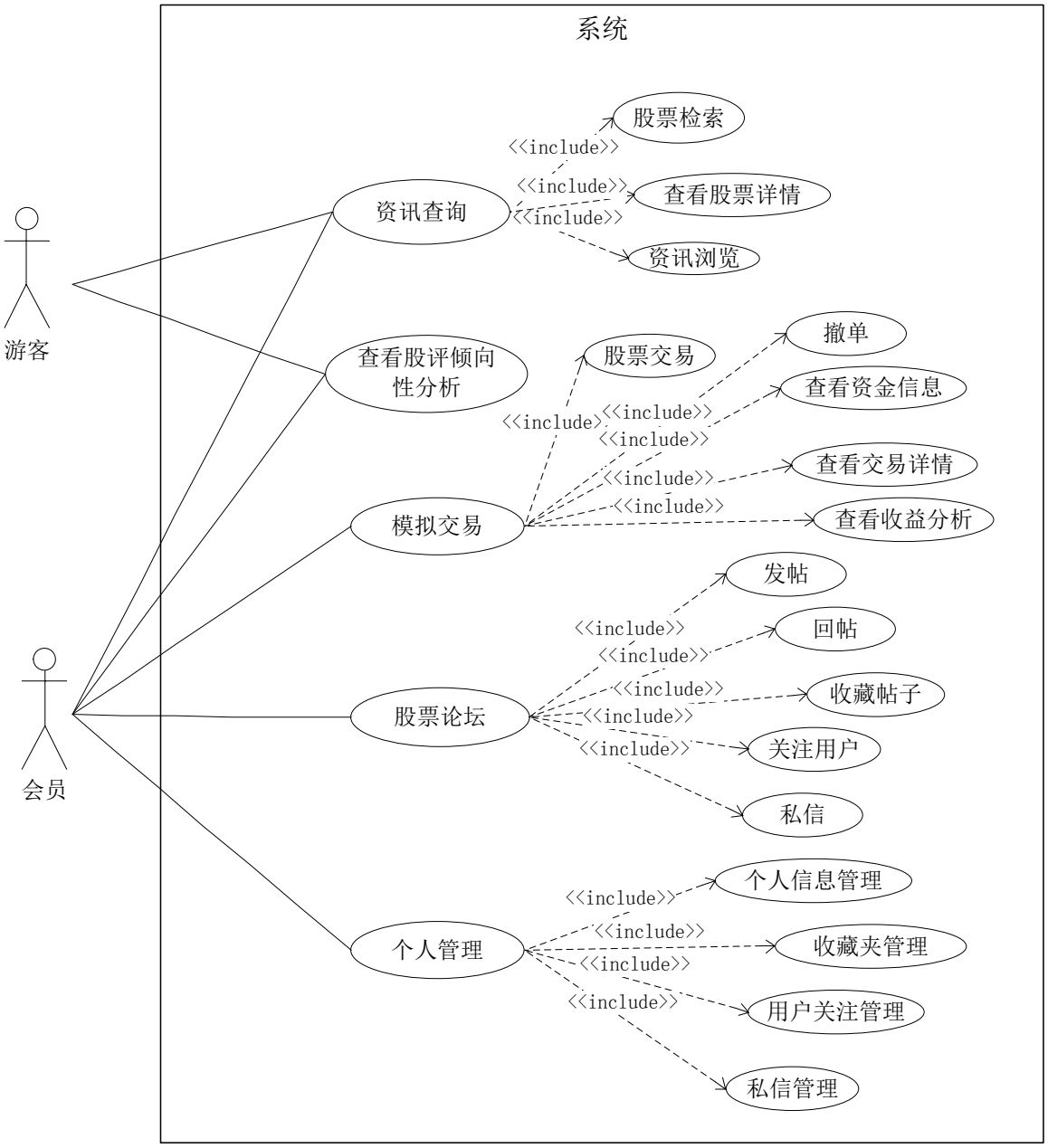


图 4-1 游客与会员角色用例图

管理员是系统的管理者，可以对用户的信息、股票的交易数据、资讯数据、股票论坛的相关数据等进行管理。管理员角色的用例图如下图 4-2 所示：

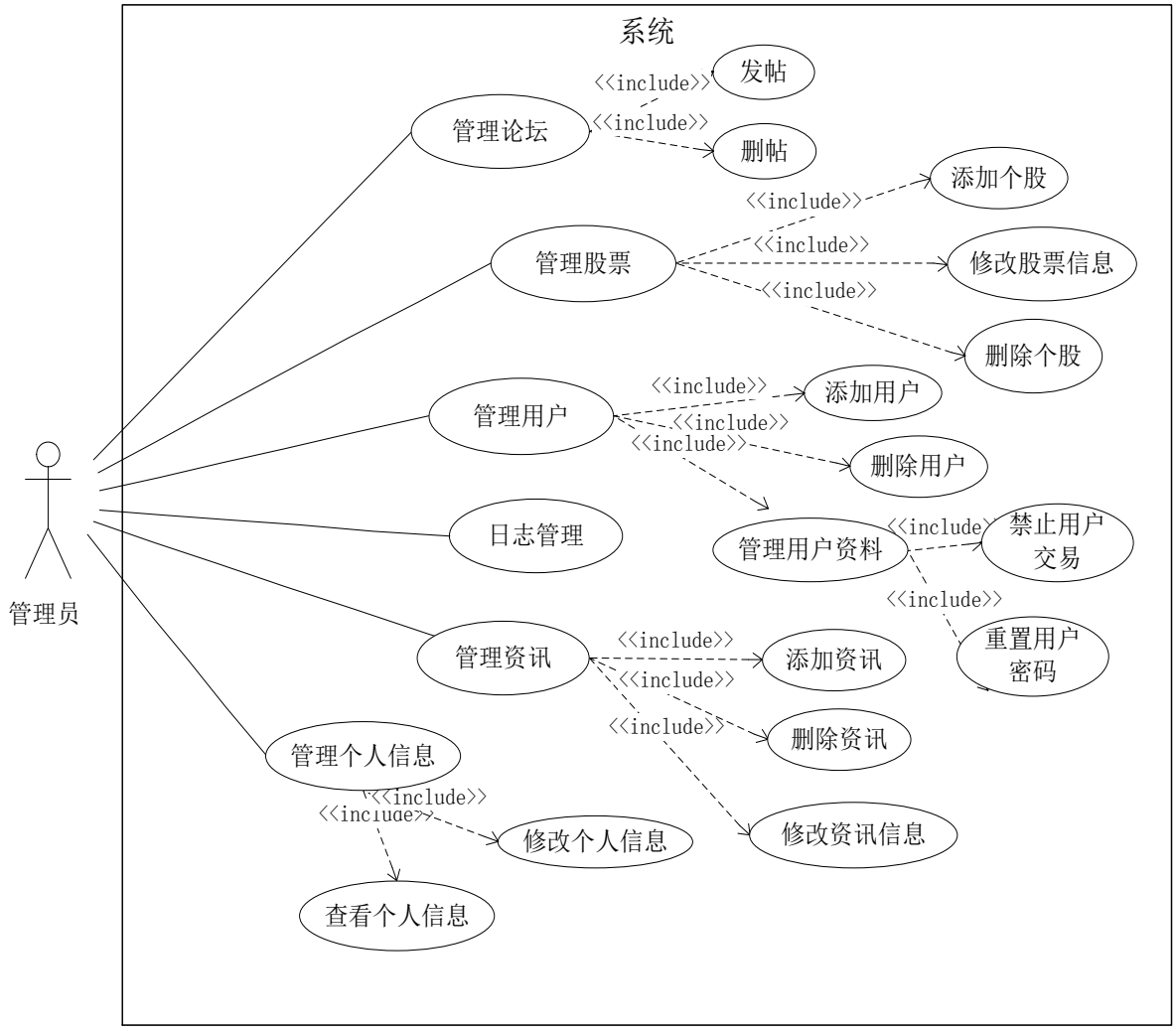


图 4-2 管理员角色用例图

在对系统进行总体用例的分析之后，下面将就系统的具体功能需求，利用用例描述进行详细的分析。

2）资讯查询

本系统的资讯查询功能能够让用户方便的浏览感兴趣的股票新闻与股票评论，同时也能为用户提供个性化的股票评论倾向性分析。

具体地说，资讯查询功能可以分为股票检索、股票详情查看、资讯列表浏览、定制信息查看和股评分析查看 5 个用例。其中，游客和会员都可以使用的功能是股票检索、股票详情和资讯列表浏览。股票检索为用户提供股票的查找功能，用户可以按股票编号、股票名称等关键词进行股票的信息检索。股票详情是用户查看股票详细信息的功能，详情内容包括股票的发行方、股票类型、上市时间、历史价格走势等。资讯列表浏览是指，用户可以查看系统资讯首页的新闻等信息。用户也可以通过检索股票来查找相关新闻。

下面以资讯查询功能中的查看股票资讯为例，对用例进行描述。查看股票资讯的用

例描述如下表 4-1 所示：

表 4-1 资讯查询用例描述

表项	说明
用例名称	查看股票资讯
来源	需求
主要参与者	会员用户
描述	该用例描述会员浏览股票资讯的过程
前置条件	该用户已经进入系统并登陆
典型事件过程	1.用户进入系统资讯首页 2.点击新闻标题 3.系统显示股票资讯详情
后置条件	系统保存用户浏览记录
结论	当用户完成股票资讯查询所有步骤，该用例结束

3）查看股评倾向性分析

用户可以在本系统中查看各大主流财经网站的股评倾向性分析，作为股票交易的参考。用户需要首先在搜索框中输入股票编号或股票名词，在检索出需要查询的股票之后，进入股票详情页面。系统会对股票详情，包括各种技术指标和股票评论的倾向性分析结果。系统会对各大门户网站的股评倾向性进行分析，并统计总的倾向性分析结果，显示在股票详情页面上。用户也可以只选择其中某个网站，查看更加详细的倾向性分析情况。查看股评倾向性分析的用例描述如下表 4-2 所示：

表 4-2 查看股评倾向性分析用例描述

表项	说明
用例名称	查看股评倾向性分析
来源	需求
主要参与者	会员用户
描述	该用例描述会员用户查看股评倾向性分析的过程
前置条件	该用户已经进入系统并登陆
典型事件过程	1.用户在搜索框中输入股票编号或名称 2.点击搜索按钮 3.系统进行股票信息的检索 4.系统显示检索到的股票 5.用户查看股票详情 6.系统显示股评倾向性分析
后置条件	系统显示股评倾向性分析
结论	当用户完成股评倾向性分析查看的所有步骤，该用例结束

4）模拟交易

模拟交易部分是只有会员才可以使用的功能。具体可以分为资金信息显示、股票交易、交易详情、撤单和收益分析 5 个功能。

下面以模拟交易中的买进股票为例，买进的用例描述如下表 4-3 所示：

表 4-3 模拟交易用例描述

表项	说明
用例名称	股票交易
来源	需求
主要参与者	会员用户
描述	该用例描述会员用户进行模拟交易的过程
前置条件	该用户已经进入系统并登陆
典型事件过程	1.用户在搜索框中输入股票编号 2.点击搜索按钮 3.系统进行股票的检索 4.系统显示检索到的股票 5.用户点价买入股票 6.用户选择买入股票的数量 7.系统计算需要金额并显示 8.用户确认下单 9.系统提示买入成功，并显示单号
后置条件	系统添加交易记录
结论	当用户完成购买股票的所有步骤，该用例结束

5）股票论坛

股票论坛是只允许会员使用的系统功能。用户在登陆系统后，点击股票论坛板块，即可进入相应的系统功能页面。股票的论坛功能包括发帖、回帖、关注其他用户、收藏帖子以及向其他用户发送私信等功能。

下面以股票论坛中的发帖功能为例，对用例进行描述。股票论坛用例描述如下表 4-4 所示：

表 4-4 股票论坛用例描述

表项	说明
用例名称	股票论坛
来源	需求
主要参与者	会员用户
描述	该用例描述会员用户在股票论坛中发帖的过程
前置条件	该用户已经进入系统并登陆

表 4-4 （续）

表项	说明
典型事件过程	2.系统进入发帖界面 3.用户填写帖子内容，点击发送按钮 4.系统显示已发送的帖子
后置条件	系统中增加一条发帖记录
结论	当用户完成发帖的所有步骤，该用例结束

6）个人管理

会员在登陆系统后,可以进行个人管理的相关操作。用户可以进行个人信息的修改，也可以对收藏的股票、帖子、关注的用户等相关内容进行管理操作。除了和其他用户之间的私信操作，当用户对系统信息的准确性、系统检索的结果、或系统收录的信息的真实性有所怀疑时，都可以通过私信想管理员进行反馈。

下面以系统中的个人管理为例，对用例进行描述。个人管理的用例描述如下表 4-5 所示：

表 4-5 个人信息管理用例描述

表项	说明
用例名称	个人信息管理
来源	需求
主要参与者	普通用户
描述	该用例描述普通用户进行个人信息管理的过程
前置条件	该用户已经进入系统并登陆
典型时间过程	1.用户点击“用户管理”按钮，进入用户管理界面 2.用户点击“基本信息”按钮，进行个人信息修改 3.用户对需要调整的信息进行修改或补充 4.点击“保存”，保存修改后的个人信息
后置条件	系统中对用户最新的个人信息进行记录
结论	当用户完成个人信息管理的所有步骤，该用例结束

7）管理员管理

系统的管理员管理功能只有管理员用户才可以使用，包括系统论坛的管理、股票信息的管理、用户信息的管理等。其中，股票论坛管理是指管理员可以对论坛中的帖子进行添加、删除、查询、修改等操作；股票信息管理是指对股票及其相关资讯的添加、删除、查询、修改等操作；用户管理是指对系统会员用户信息的添加、删除、查询、修改等操作。

下面以管理员管理中的股票管理为例，对用例进行描述。股票管理的用例描述如下表 4-6 所示：

表 4-6 管理员管理用例描述

表项	说明
用例名称	管理员管理
来源	需求
主要参与者	管理员用户
描述	该用例描述管理员进行后台股票管理的过程
前置条件	该用户已经进入系统并登陆
典型事件过程	1.用户点击“股票”按钮，进入用户管理界面 2.用户点击“基本信息”按钮，进行股票信息修改 3.用户对需要调整的信息进行修改或补充 4.点击“保存”，保存修改后的股票信息
后置条件	系统中对用户最新的股票信息进行记录
结论	当用户完成股票信息管理的所有步骤，该用例结束

4.2.2 基于倾向性分析的股票资讯服务系统的静态模型

本文使用静态模型对系统进行分析的主要目标，是能够获得基于倾向性分析的股票资讯服务系统的所有核心类，以及类之间的基本关系。系统的核心类图如下图 4-3 所示：

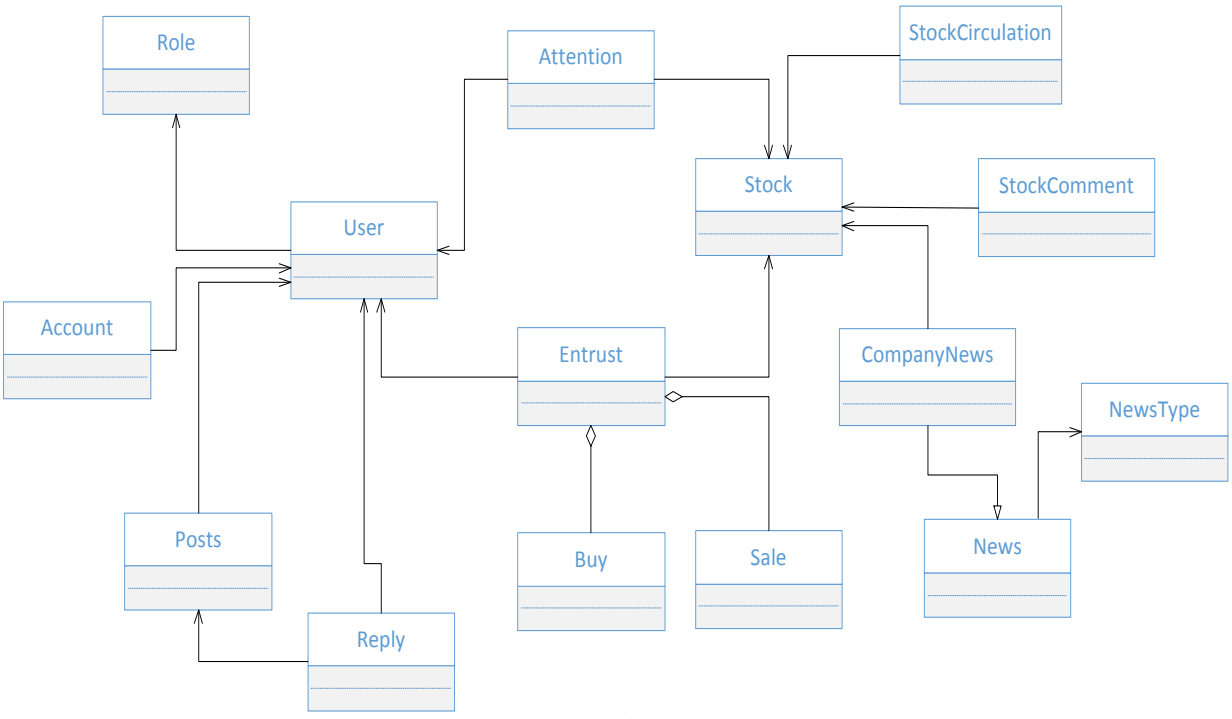


图 4-3 系统核心类图

系统的核心类包括 User 类、News 类、Stock 类、Account 类、Role 类、Posts 类、Reply 类、Attention 类、Entrust 类、Buy 类、Sale 类、StockCirculation 类以及 StockComment 类等。其中，User 类表示用户类， Stock 表示股票类，这两个类是系统中关键的部分。Buy 类和 Sale 类分别表示买进和卖出类，是股票交易相关的类。Role 类表示用户角色

类，用于区分不同的用户类型。Ensrust 类是委托类，在股票交易中有着重要的作用。News 类表示系统的资讯类，是系统首页资讯查询功能的相关类。StockComment 类是股评类，负责股评分析相关的类。在对系统的静态模型进行说明后，下面将对基于倾向性分析的股票资讯服务系统的行为模型进行详细的描述。

4.2.3 基于倾向性分析的股票资讯服务系统的行为模型

本文将使用活动图来对基于倾向性分析的股票资讯服务系统的行为模型进行说明。活动图用来描述业务用例内部事件流中不同活动之间的动作序列，实现不同活动之间的控制，特别适合于工作流和并发的处理行为，同时可以按照需要来定义活动中的对象和对应的状态、角色和属性的改变。下面，将按照需求分析中的功能划分，对系统对象之间的交互过程进行详细的说明。

1）资讯查询

系统的资讯查询功能主要涉及游客、会员用户和系统。下面以检索股票资讯为例，说明用户对股票资讯进行查询的过程：用户进入系统后，选中搜索框，并在检索框中输入需要检索的关键词。关键词可以是统一的股票编号、股票名称，也可以是股东信息等。在输入检索关键词后，用户需要选择检索的信息类型。信息类型分为两类，一类是直接对股票进行检索，用户点击检索结果后就可以查看股票的技术指标等详细信息；另一类是该股票的相关新闻。系统资讯查询的活动图如下图 4-4 所示：

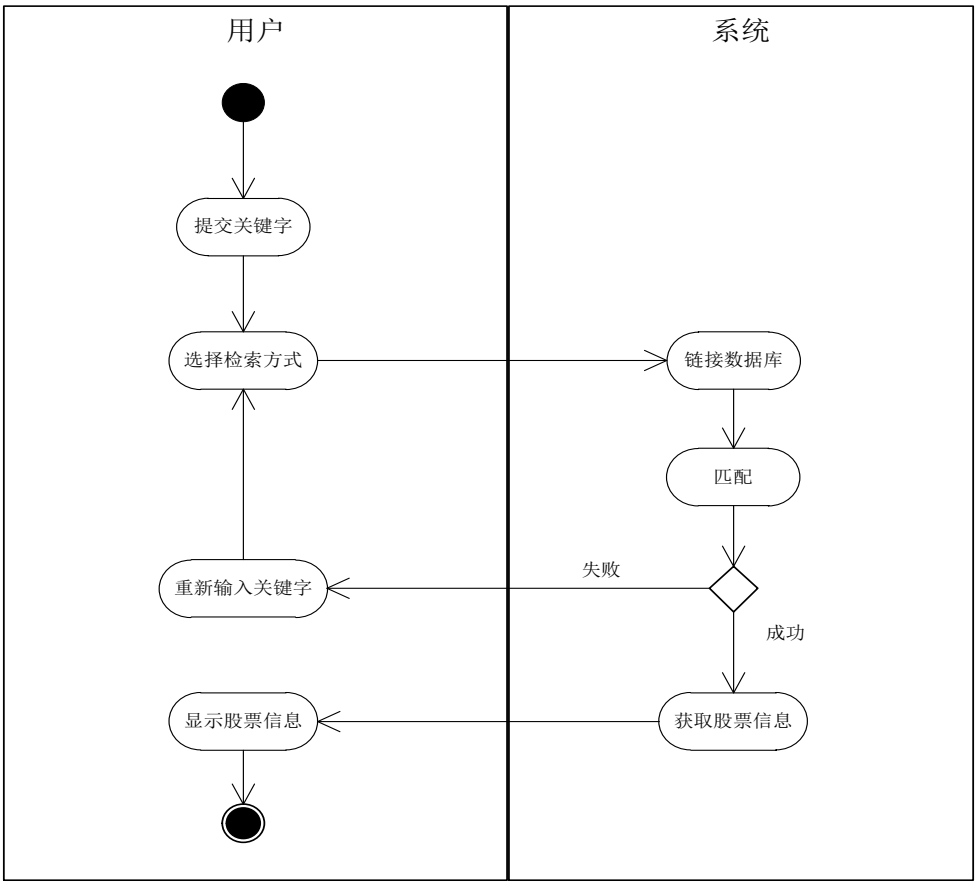


图 4-4　资讯查询活动图

2）查看股评倾向性分析

用户可以在本系统中查看各大主流财经网站的股评倾向性分析，作为股票交易的参考。用户需要首先在搜索框中输入股票编号或股票名词，在检索出需要查询的股票之后，进入股票详情页面。系统会对各大门户网站的股评倾向性进行分析，并统计总的倾向性分析结果，显示在股票详情页面上。用户也可以只选择其中某个网站，查看更加详细的倾向性分析情况。查看股评倾向性分析的活动图如下图 4-5 所示：

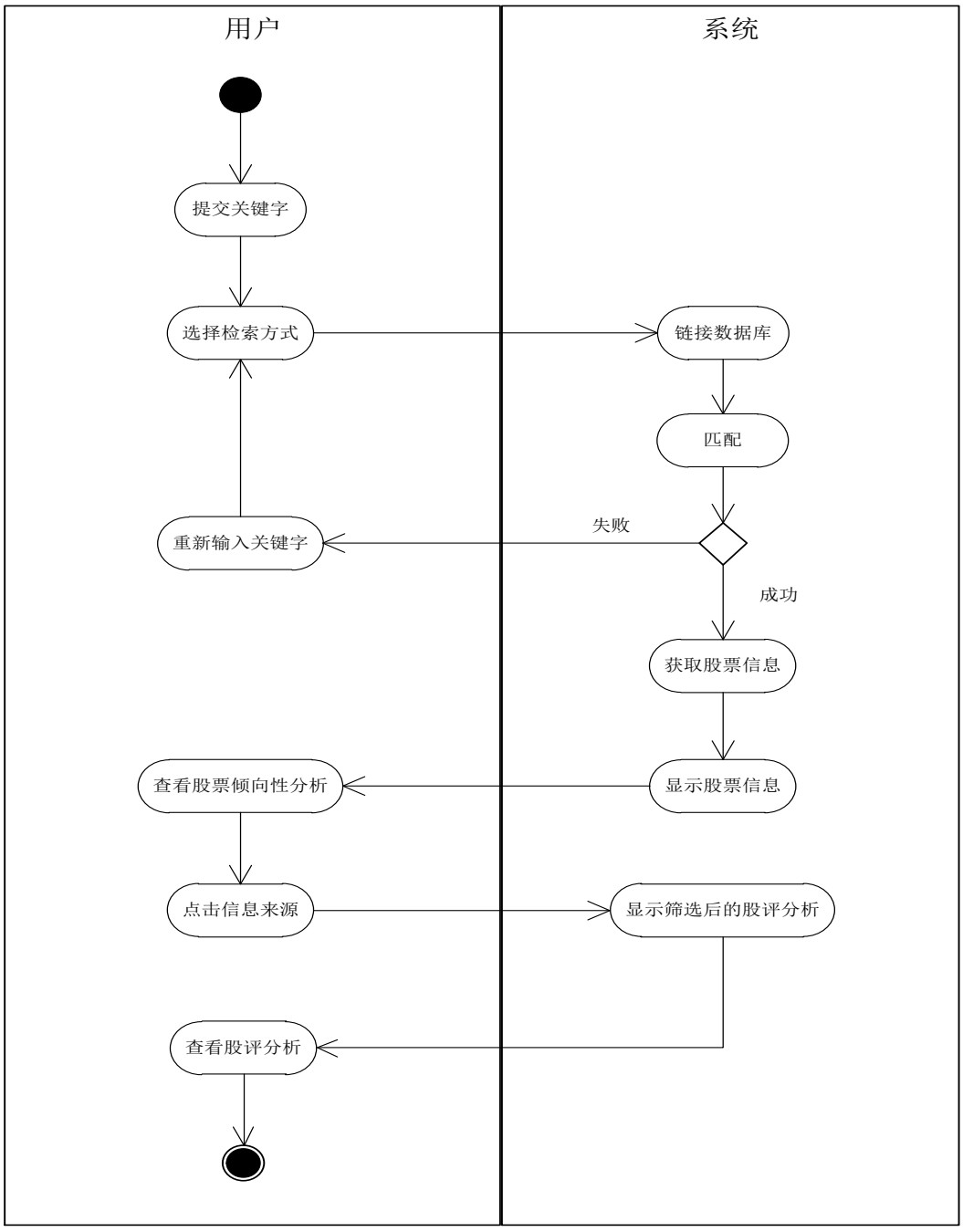


图 4-5　查看股评倾向性分析活动图

3）模拟交易

模拟交易是只有会员才能使用的系统功能。以买进股票为例，用户与系统之间的交

互包括：用户登陆会员账号，系统验证之后，用户进入系统模拟交易板块。用户首先选择股票市场然后点击交易按钮。买入股票的活动图如下图 4-6 所示：

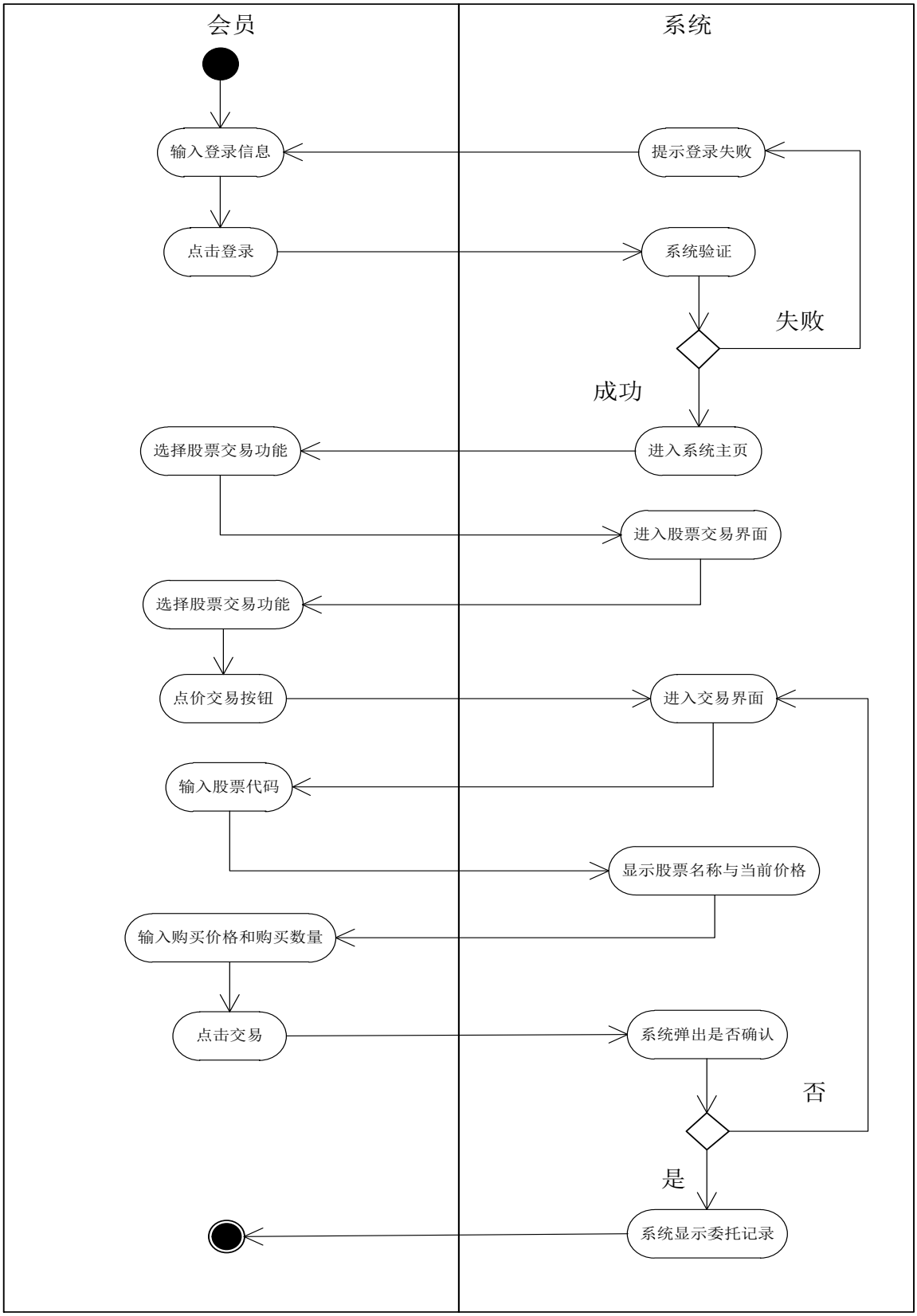


图 4-6　模拟交易活动图

4）股票论坛

股票论坛功能的使用需要会员先登陆系统并验证身份。用户可以在股票论坛中发帖、回帖、关注用户、收藏帖子或发送私信。用户发帖的活动图如下图 4-7 所示：

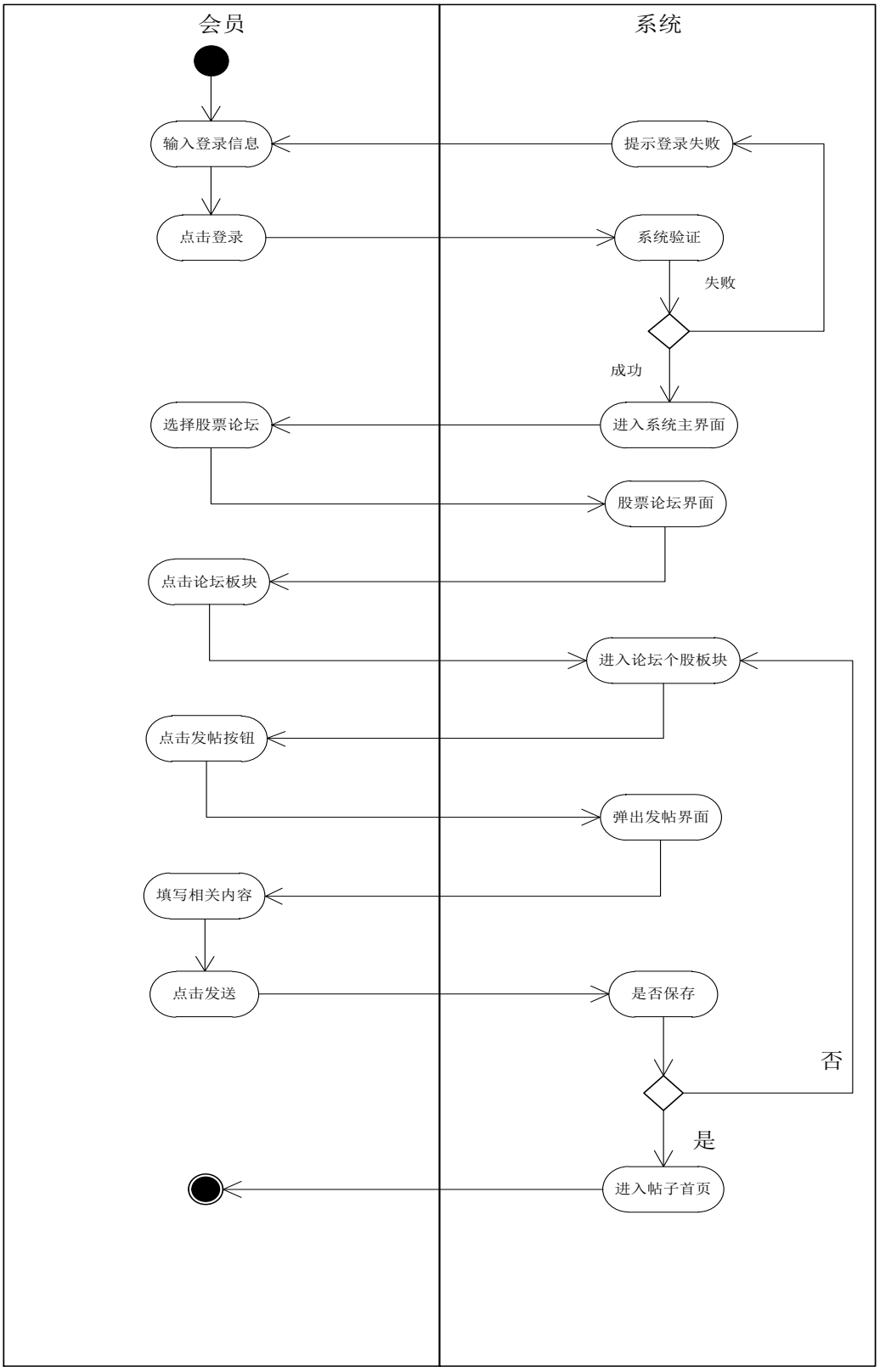


图 4-7　发帖活动图

5）个人管理

个人管理的本质是用户在登陆自己的账号后，对自身的信息，例如用户名、密码、收藏夹、歌单、私信等相关信息，进行增加、删除、修改、查找的操作。当用户正确登陆自己的账号之后，便可以进行相应的操作。现在以用户修改个人信息为例，介绍系统中用户进行个人信息管理的活动图，如图 4-8 所示：

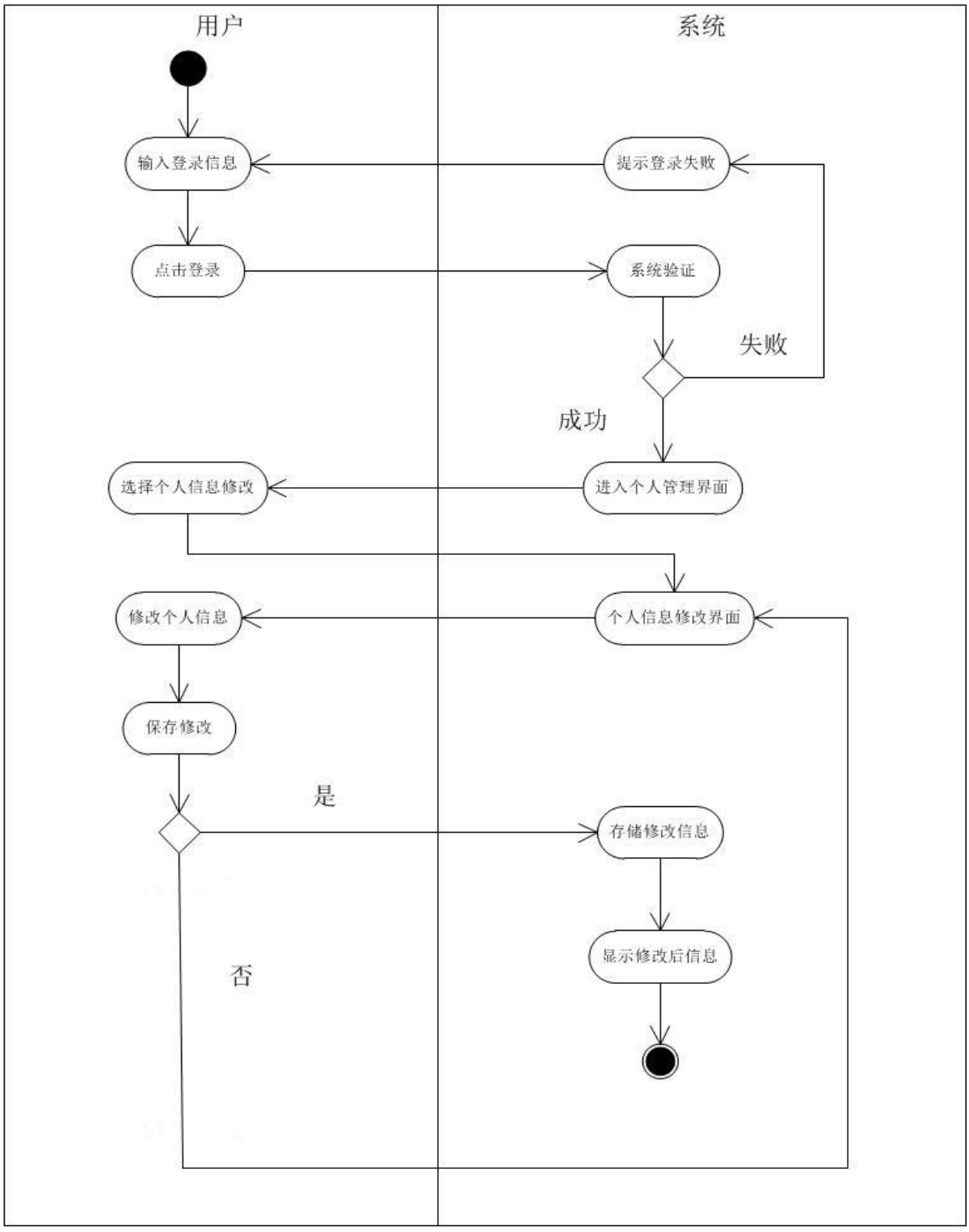


图 4-8　个人信息管理动图

6）管理员管理

管理员用户有权限对股票资讯、会员用户、个股数据、股票论坛的信息进行管理。其中，管理员管理的核心功能是对以上信息的增加、删除、修改和查询操作。下文以添加股票资讯信息为例，活动图如图 4-9 所示：

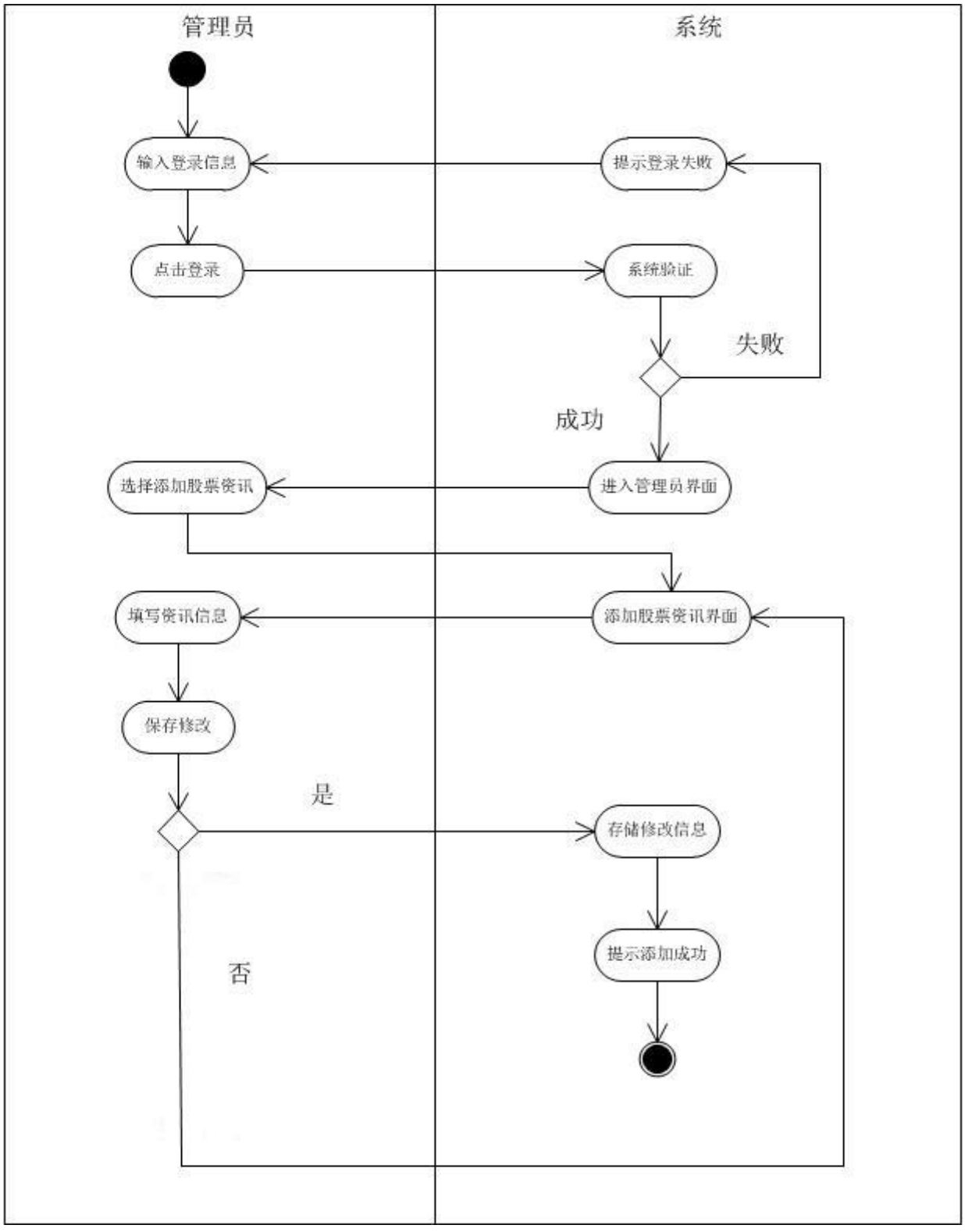


图 4-9　资讯添加活动图

4.3 基于倾向性分析的股票资讯服务系统的非功能性需求

系统的非功能性需求，是指产品为满足用户业务需求而具有的功能意外的其他特性。针对本文的基于倾向性的股票资讯服务系统，应当具有以下非功能性需求：

1）性能需求: 并发用户数小于 1000，在 1000 个并发用户进行股票检索时，业务处理响应时间在 5 秒以内；对股票进行交易时，不计入网络传输时间，买进和卖出的响应时间在 3 秒以内。当使用股票论坛功能时，回帖、发帖的响应时间应该在 5 秒以内。

2）可靠性：可靠性是指，在规定的一段时间内与条件下软件维持其性能和水平的属性。软件的可靠性包括成熟性、容错性和易恢复性。尽管本文设计的交易行为是模拟、仿真的，但仍然应该对故障情况有严格的要求。此外，一旦发生故障，应当能够迅速重建系统，恢复数据，以免影响用户体验。

3）易用性：易用性是与一组规定或者潜在的用户为使用其软件所需做的努力和对这样的使用所作的评价有关的一组属性，包括易理解性、易学习性和易操作性。本文所设计的基于倾向性分析的股票资讯服务系统，在模拟交易方面应当具备和真实交易完全一致的使用体验，在股票舆情查询上则应该简便易学，便于用户迅速习惯。

4）可移植性：可移植性是指与软件可从某一环境转移到另一环境的能力有关的一组属性。具体包括系统的适应性、易安装性、遵循性、以及可替换性。软件的移动化是当前的趋势，股票的交易操作也越来越多得由手机完成，因此开发基于移动网络的版本对本系统有重要的意义。

5）安全性：即与防止对程序技术局的非授权的故意或者意外访问的能力有关的软件属性。如用户权限、动态口令、数据库字段加密等。系统应该能够保障用户的个人信息安全。

4.4 本章小结

本章对基于倾向性分析的股票资讯服务系统的主要需求进行了分析和归纳，对系统需要实现的主要功能进行了清晰的表达和描述。本章首先立足全局，将整个系统的框架和设计进行了全面的概括；然后将系统功能进行了细分，详细介绍了系统需要实现的各大功能的具体设计。本章主要使用 UML，从系统的功能模型，静态模型以及行为模型三个角度三个方面对需求设计进行规范化的描述。

5 基于倾向性分析的股票资讯服务系统的设计

上章对基于倾向性分析的股票资讯服务系统的整体需求做了介绍，对于系统需要完成的主要功能点进行了阐述，对主要角色的功能通过用例图进行了展示，同时进行了系统静态建模和系统动态建模。本章将从系统构架设计、系统概要设计、系统详细设计和数据库设计四个方面对系统设计进行讨论，为系统的实现打下基础。

5.1 基于倾向性分析的股票资讯服务系统的架构设计

本系统打算采用基于 B/S（浏览器/服务器）模式的分层体系结构，表示层是返回给用户显示的页面，方便用户对系统的操作；数据层用来存放股票、用户信息等；业务层主要是事务处理，包括处理用户的请求、系统与用户的交互和实现系统的功能。

根据上一节对系统的需求分析，本文采用基于 SSH（Structs+Spring+Hibernate）框架来构建系统，可以实现搭建结构清晰、可复用性好、维护方便的 Web 应用程序。其中使用 Struts 作为系统的整体基础架构，负责 MVC 的分离，在 Struts 框架的模型部分，控制业务跳转；利用 Hibernate 框架对持久层提供支持；Spring 做管理，管理 Struts 和 Hibernate。本系统的体系结构从职责上分为四层，依次为表示层、业务层、数据持久层和域模型层，如图 5-1 所示。

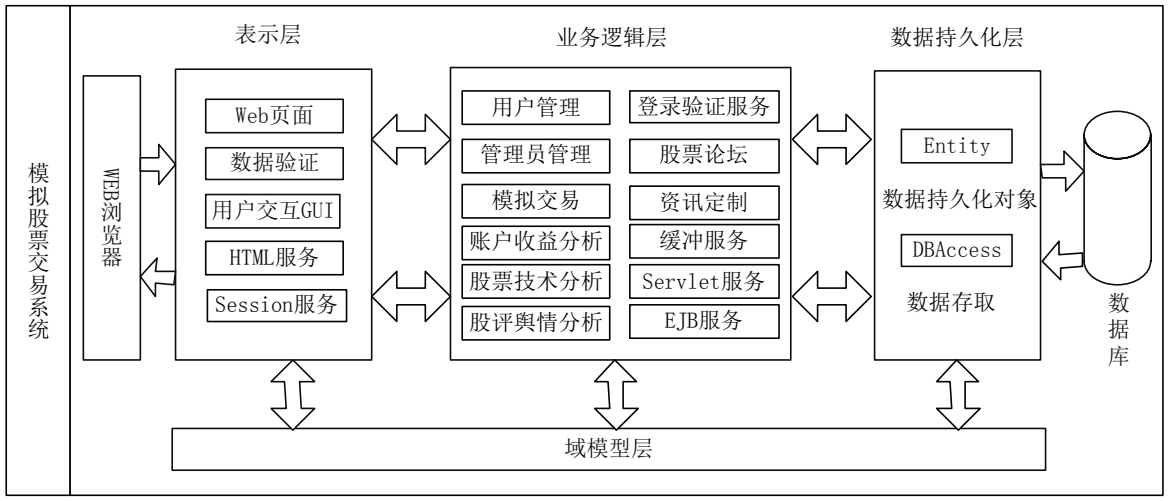


图 5-1 系统的体系结构图

从图 5-1 中可以看出，表示层使用了 Struts 框架构建了系统页面与业务逻辑的分离，使页面显示和业务逻辑实现低耦合，业务层由 Spring 来完成事务的处理功能，数据持久层采用了 Hibernate 框架实现的 DAO 类来实现 Java 类与数据库之间的转换和访问，使业务逻辑与数据持久化分离。下面对各层功能做详细的论述。

表现层：包含系统与用户交互 GUI（图形用户界面）和数据验证等。该层用于向系统用户提供图像交互界面，允许用户在显示页面中输入和编辑信息，同时该层还提供

数据验证功能，用于验证数据的有效性。该层主要包含的模块有系统各个功能模块的显示页面和相关请求处理页面。

业务层：包含业务规则的处理代码，即程序中与业务相关的专业算法、业务政策等。该层用于执行业务流程和指定数据的业务规则。业务层是面向业务应用，为表示层提供业务服务。主要包含的模块有资讯查询、模拟交易、股票论坛和股评舆情分析。还包括个人信息管理、用户管理、帖子管理、评论管理模块中需要进行处理的部分，需要与数据库之间交互处理。

数据持久层：包含数据处理代码和数据存储代码。该层主要包括数据存取服务，负责与数据库之间的通信。包括股票、资讯、股评、账户信息、委托订单、买入卖出订单、用户和管理员的个人信息，还有股票论坛的帖子和评论等存储在数据库中的信息。

5.2 基于倾向性分析的股票资讯服务系统的概要设计

5.2.1 功能模块设计

本文设计的基于倾向性分析的股票资讯服务系统的最终目的，是要实现用户方便查询股票舆情，实施模拟操作的需求。用户在该系统中可以获得最新、最准确的股票舆情信息，了解到各大门户与知名分析师的股票评论倾向性，掌握其他网友的股票购买意向，从而帮助自己决策。同时，用户可以在本系统中进行仿真的股票模拟交易，练习股票交易技术，实践股票交易策略。

从系统的功能角度，可以将本系统具体的分为 6 个功能模块，分别是资讯查询模块、模拟交易模块、股票论坛模块、个人管理模块、管理员管理模块和舆情分析模块。下面对系统的各个功能模块进行详细的介绍。系统功能模块图如下图 5-2。

资讯查询模块，是用户进行股票资讯浏览，查看股票评论倾向性分析的功能。普通游客用户可以对具体的个股进行检索，以检索自己感兴趣的股票或者相关的新闻信息。游客也可以浏览系统资讯首页的资讯列表，对系统默认显示的新闻资讯进行查看。用户登陆系统并认证为会员后，就可以使用系统资讯查询的个性化功能。会员可以选择查看的股票舆情的信息来源，也可以查看股票评论的倾向性分析，从而掌握其他股民的股票交易意向和专家的意见。

个人管理模块，是指会员用户对个人信息等相关信息进行管理的功能。系统使用者首先可以使用用户注册的功能，成为系统的注册用户，即会员。在注册成为会员后，用户就可以进行登陆操作，并使用会员的相关功能。用户在登陆后，可以进入个人信息管理界面，对账号信息，例如用户名、密码、头像等进行修改。用户在个人管理模块下，也可以进行收藏的管理，浏览、删除已经收藏的股票或帖子。用户同时可以对关注的其他用户进行删除。此外，在个人管理功能中，用户可以对私信进行查看、删除等操作。

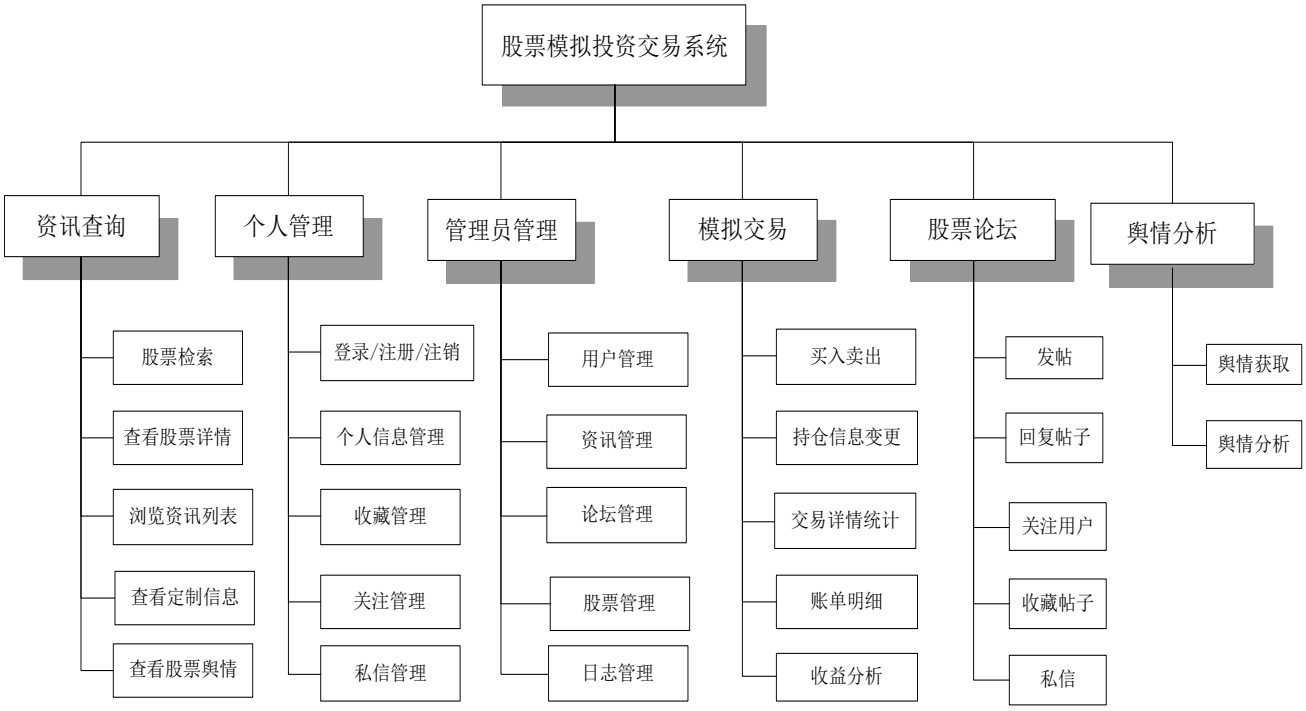


图 5-2 系统的功能模块图

管理员管理模块：管理员管理是针对系统管理者的，系统管理员不需要注册，他们的账号是在系统后台中直接分配的。当系统管理员管理系统的时候首先需要以管理员的身份登录，登录以后可以管理个人资料、可以对系统用户进行管理，包括系统用户的增加和删除等。系统管理员还可以对用户的评论和反馈进行管理，可以通过查看使用系统时产生的日志管理查看系统是否正常运行。当然，管理员也可以对股票的信息进行管理，并就论坛的内容进行添加、搜索、修改、删除的操作。当管理员退出系统时需要注销系统，保证系统的安全。

模拟交易模块，会员在登陆系统后，可以使用模拟交易功能。会员可以在模拟交易中使用真正股票交易的所有功能，包括股票的买进、卖出，对持仓信息的查询，历史交易查看，账单明细，盈利情况分析等。系统中的股票数据基于实时获取的真实股票数据，以便于为用户提供最为真实的交易体验。

股票论坛模块，会员在登陆系统后，可以使用股票论坛中的相关功能。用户可以在股票论坛中交流自己的交易经验，也可以发表对股票的看法等等。本模块的功能包括发帖、回帖、关注其他用户、收藏帖子以及向其他帖子发送私信。该模块为系统增加了社交属性，增强了用户黏性，为用户的交流、沟通提供了平台。

舆情分析模块。该模块主要为会员用户提供系统的舆情分析功能。系统在使用网络爬虫获取数据源之后，进行文本的预处理，例如分词等，并使用改进的针对股票评论的倾向性分析算法进行倾向性计算。

5.2.2 类的设计

系统的详细类图如下图 5-3 所示：

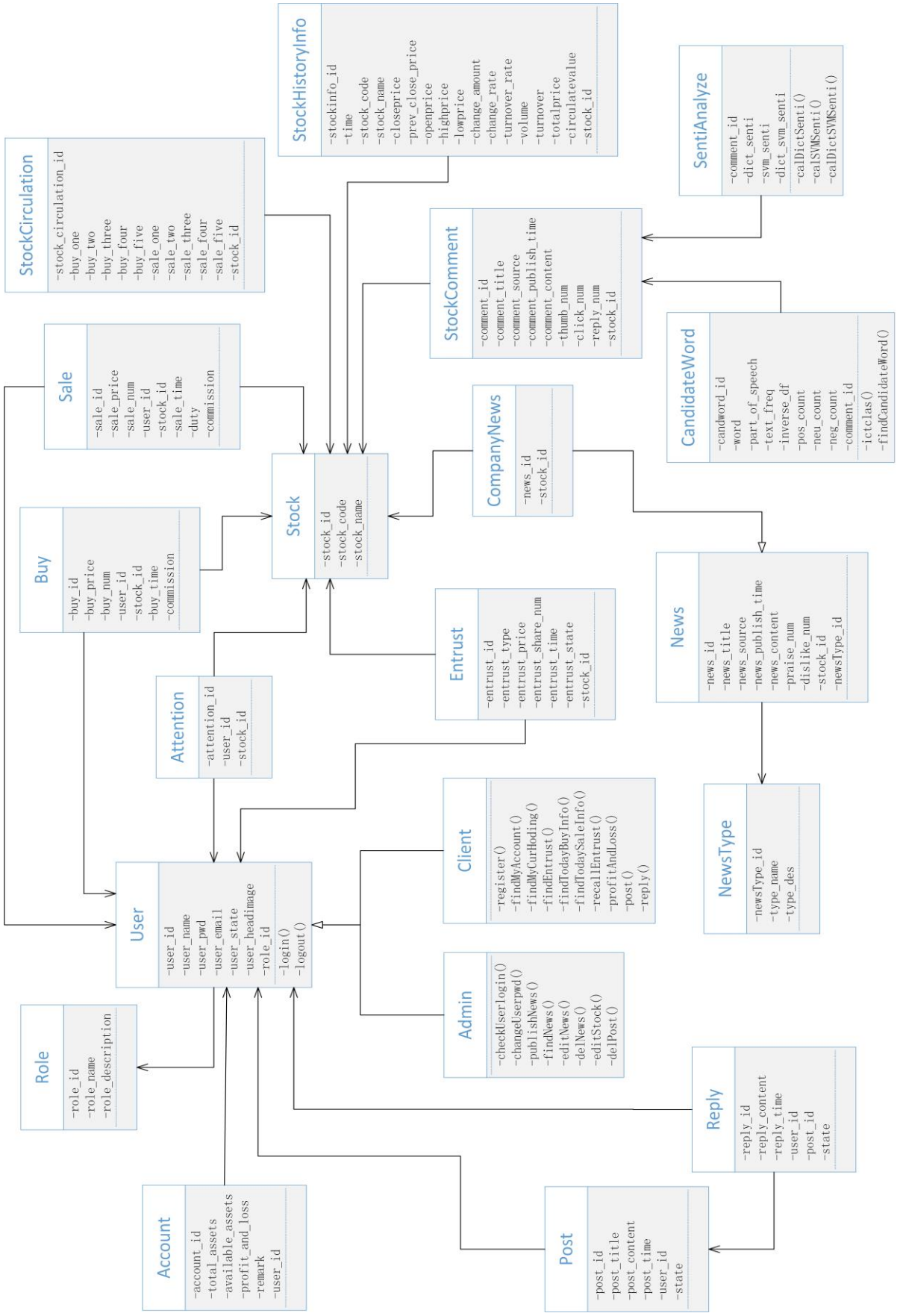

StockHistoryInfo
-stockinfo_id
-time
-stock_code
-stock_name
-closeprice
-prev_close_price
-openprice
-highprice
-lowprice
-change_amount
-change_rate
-turnover_rate
-volume
-turnover
-totalprice
-circulatevalue
-stock_id
StockCirculation
-stock_circulation_id
-buy_one
-buy_two
-buy_three
-buy_four
-buy_five
-sale_one
-sale_two
-sale_three
-sale_four
-sale_five
-stock_id
SentiAnalyze
-comment_id
-dict_senti
-svm_senti
-dict_svm_senti
-calDictSenti()
-calSVMSenti()
-calDictSVMSenti()
StockComment
-comment_id
-comment_title
-comment_source
-comment_publish_time
-comment_content
-thumb_num
-click_num
-reply_num
-stock_id
CandidateWord
-candword_id
-word
-part_of_speech
-text_freq
-inverse_df
-pos_count
-neu_count
-neg_count
-comment_id
-ictclas()
-findCandidateWord()
Sale
-sale_id
-sale_price
-sale_num
-user_id
-stock_id
-sale_time
-duty
-commission
CompanyNews
-news_id
-stock_id
Stock
-stock_id
-stock_code
-stock_name
Buy
-buy_id
-buy_price
-buy_num
-user_id
-stock_id
-buy_time
-commission
News
-news_id
-news_title
-news_source
-news_publish_time
-news_content
-praise_num
-dislike_num
-stock_id
-newsType_id
Entrust
-entrust_id
-entrust_type
-entrust_price
-entrust_share_num
-entrust_time
-entrust_state
-stock_id
Attention
-attention_id
-user_id
-stock_id
Client
-register()
-findMyAccount()
-findMyCurHoding()
-findEntrust()
-findTodayBuyInfo()
-findTodaySaleInfo()
-recallEntrust()
-profitAndLoss()
-post()
-reply()
NewsType
-newsType_id
-type_name
-type_des
User
-user_id
-user_name
-user_pwd
-user_email
-user_state
-user_headimage
-role_id
-login()
-logout()
Admin
-checkUserlogin()
-changeUserpwd()
-publishNews()
-findNews()
-editNews()
-delNews()
-editStock()
-delPost()
Role
-role_id
-role_name
-role_description
Reply
-reply_id
-reply_content
-reply_time
-user_id
-post_id
-state
Account
-account_id
-total_assets
-available_assets
-profit_and_loss
-remark
-user_id
Post
-post_id
-post_title
-post_content
-post_time
-user_id
-state

图 5-3 系统的详细类图

图中可以看到，系统中最为核心的类是 User 类，News 类，Stock 类以及 SentiAnalyze 类。User 类涉及 login()和 logout()方法，分别用于系统的登陆和注销。Client 类继承自 User 类，是系统的会员类，所涉及的方法有 findMyAccount()，findMyCurholding()，分别是查看那自己的个人账户以及查看当前持仓，findEntrust()和 recallEntrust()表示查看当日委托和撤单，profitAndLoss()用于查看用户当前盈利，post()和 reply()表示在股票论坛发帖和回帖。管理员类 Admin 涉及的方法包括 checkUserlogin()、changeUserpwd()、publishNews()、findNews()、editNews()、delNews()、editStock()、delPpst()等等，这些方法分别是用户管理、股票新闻管理、股票管理、论坛管理等涉及到的功能。

News 类是系统的资讯类，用于定制化展示系统从新浪财经、东方财富、和讯网等各大主流财经门户网站爬取的不同类别的资讯，主要的资讯类别有个股点评、股吧精华、大盘分析、门户头条、机构专栏等。

Stock 类则与股票的交易相关，与股票信息类 StockHistoryInfo、委托类 Entrust 以及股票流通类 StockCirculation 关联，支撑系统的模拟交易业务，根据股票当日流通信息完成客户的委托下单。

在情感分析类 SentiAnalyze 中，calDictSenti()、calSVMSenti()和 calDictSVMSenti()是倾向性分析所需要的方法，并记录各种不同方法计算得出的情感分类结果。它与 StockComment 类相关联，对个股相关的股票评论进行倾向性分析，最终对分类结果进行统计呈现给用户。

在对系统的类进行描述之后，将利用类中的方法，进一步对系统的功能进行详细设计。

5.3 基于倾向性分析的股票资讯服务系统的详细设计

本节对系统的详细设计的阐述将类的设计以及各功能模块的详细设计方面展开。在此前对于系统功能分析的基础上，本节将对系统功能进行进一步的详细设计，主要分析系统的 6 个功能模块，并给出各模块的顺序图。

通过前面对系统功能的分析，本节对系统功能的详细设计主要从资讯模块、舆情分析模块、股票模拟交易模块、论坛模块和个人管理模块等给予详细的阐述，并给出各功能模块的顺序图。

5.3.1 资讯查询

股票检索是系统资讯查询模块的子功能。用户进入系统主页，选择搜索方式为检索股票或是检索股票资讯，在搜索框中输入查询关键词，比如股票代码或名称，系统对输入的关键词进行匹配，数据库最终返回检索结果给用户。用户检索股票的时序图如图 5-4 所示。

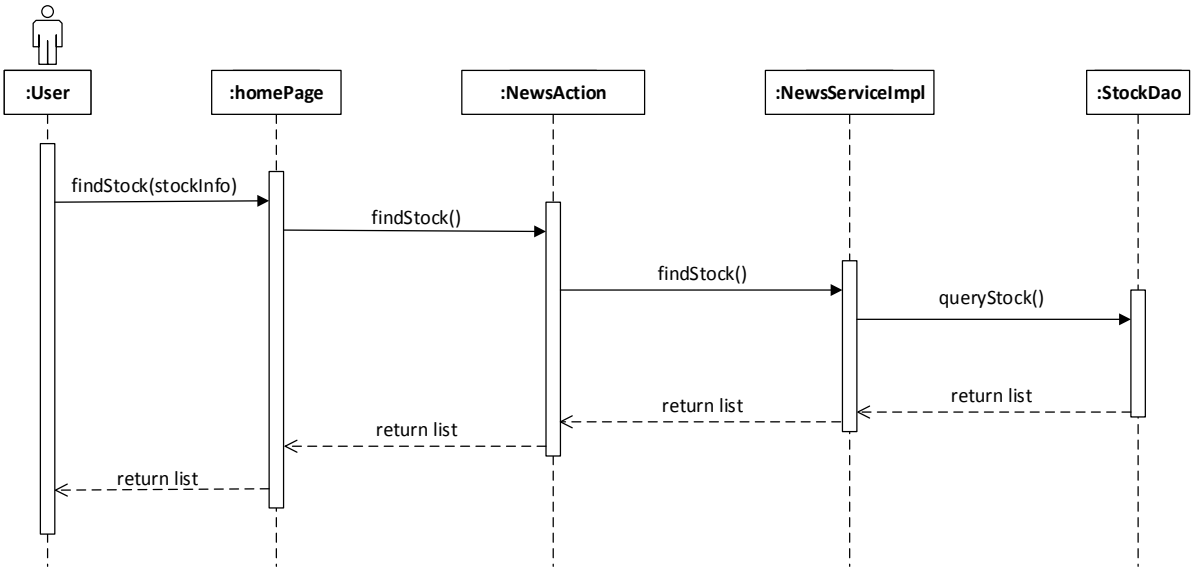


图 5-4 用户股票检索时序图

5.3.2 股评分析

股评分析是系统针对各大门户个股股吧内的股评文本进行倾向性分析的过程。系统首先对股评文本进行分词等预处理操作，随后进行词性情感特征匹配，之后根据本文改进的词典语义的算法进行情感加权计算，并结合 SVM 算法最终得出股评文本的情感分类，统计所有相关股评的情感分类，将结果返回给股票详情页面，用于展示个股的舆情数据。系统股评分析的时序图如图 5-5 所示。

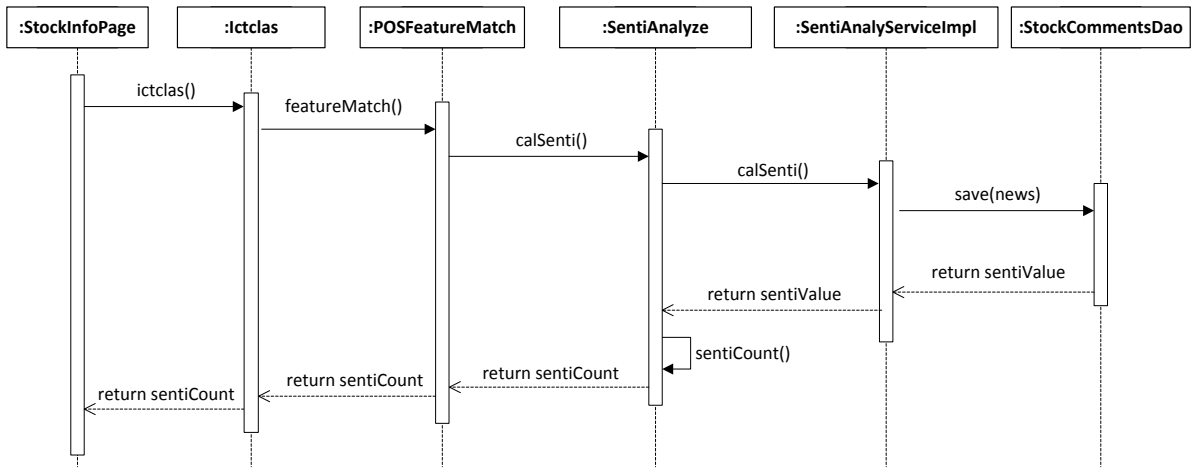


图 5-5 系统股评分析时序图

5.3.3 模拟交易

用户可以在系统中进行股票的模拟交易，即股票买卖操作。此处以股票买入操作为例进行详细分析。用户进入股票模拟交易首页，点击交易按钮，随即进入股票买入界面，在买入界面输入股票名称、买入价格和买入数量等信息，点击买入下单向系统

提出委托下单请求，系统返回当日委托列表给用户。系统在当日委托列表中选择需要委托下单的股票，根据股票流通信息表进行股价撮合等判断处理，若股价撮合成功，则生成买入委托订单，将信息存储入买入订单表中，并更新成交状态为成功，若撮合失败，则返回成交状态为失败。用户买入股票的时序图如图 5-6 所示。

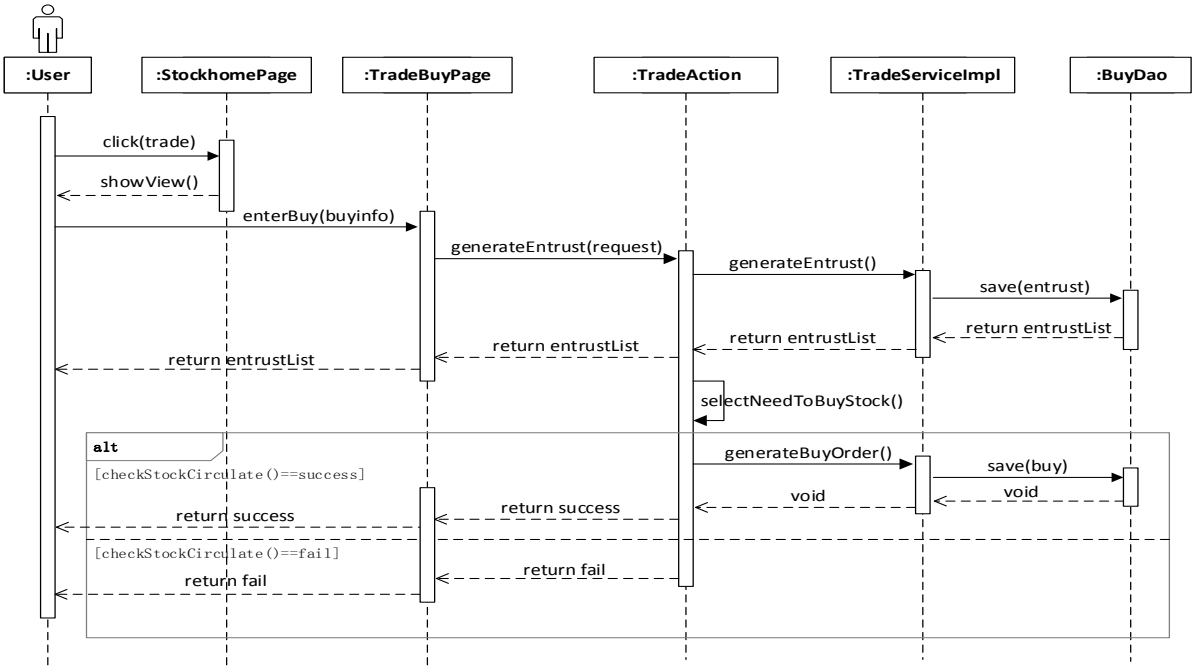


图 5-6 用户模拟买入股票时序图

5.3.4 股票论坛

股票论坛模块，用户可以进行发帖和回复。用户在论坛发帖的顺序图如图 5-7 所示。

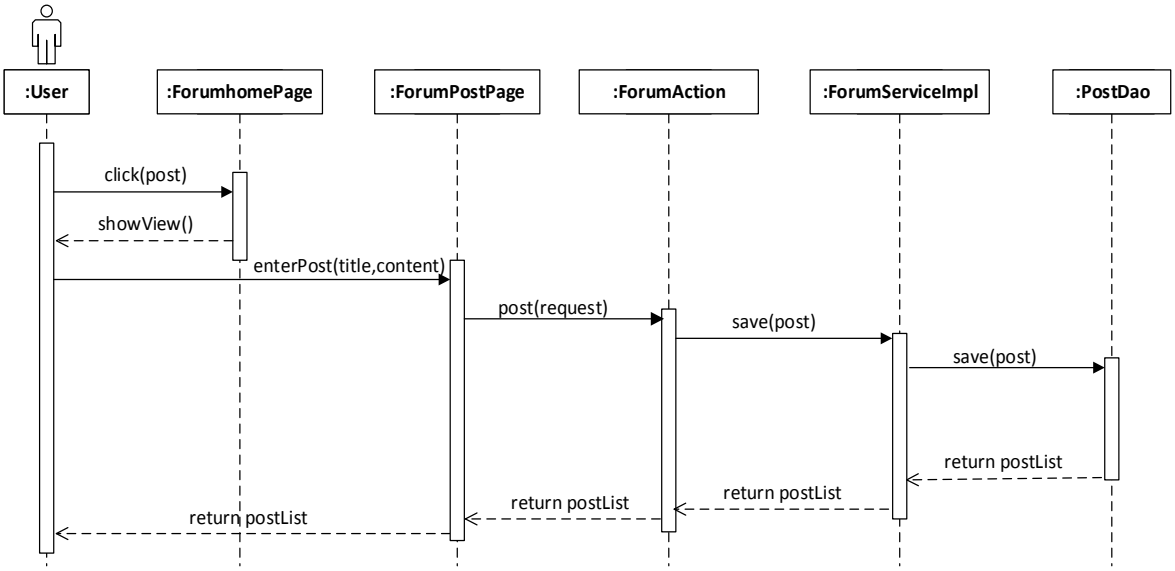


图 5-7 用户论坛发帖时序图

5.3.5 个人管理

在个人管理模块中，用户可以对自己的个人信息进行修改，包括登录密码、邮箱、头像等账号信息，以及收藏信息，关注用户等。个人信息修改的时序图如图 5-8 所示。

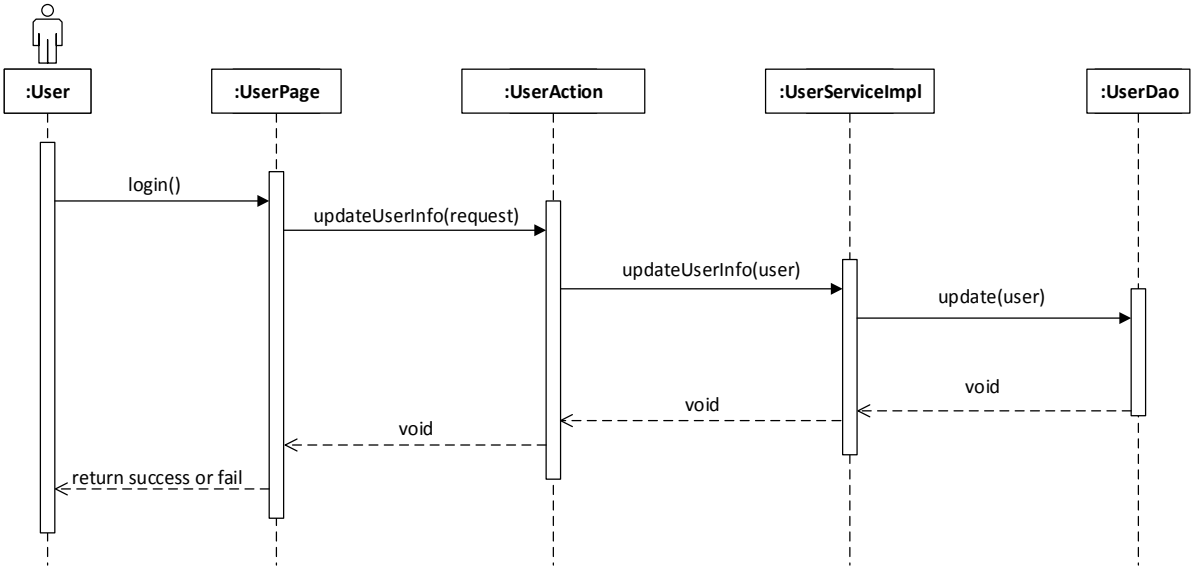


图 5-8 个人信息修改时序图

5.3.6 管理员管理

管理员在系统管理员管理模块能够实现对数据库中的用户信息、资讯信息、股票信息和论坛等的管理和维护。管理员在后台添加资讯的时序图如图 5-9 所示。

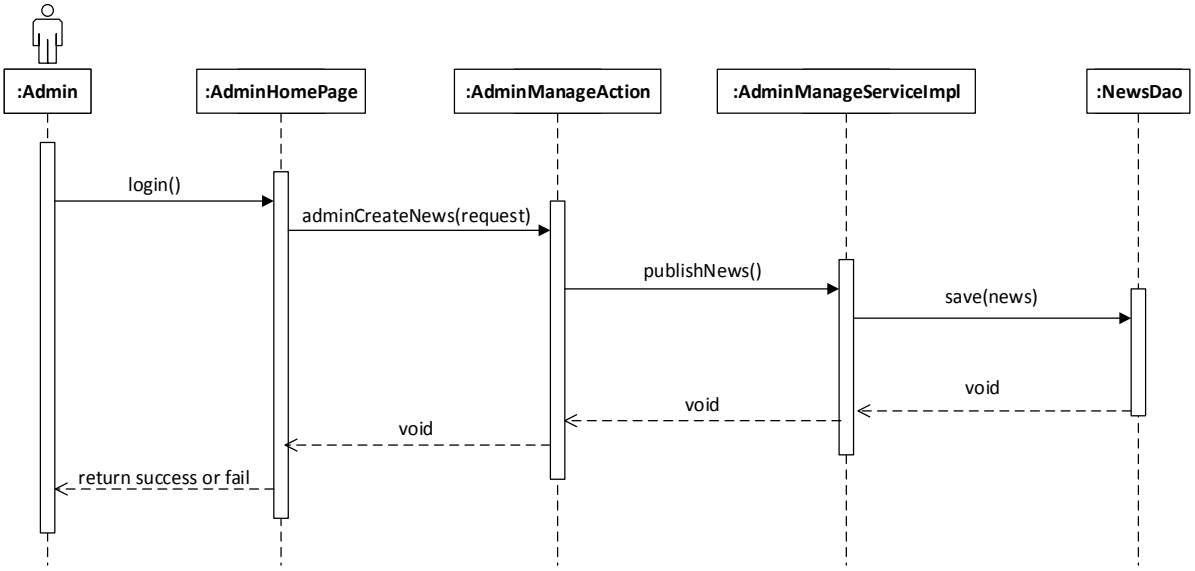


图 5-9 管理员添加资讯时序图

5.4 基于倾向性分析的股票资讯服务系统的数据库设计

本文使用 MySql 数据库作为基于倾向性分析的股票资讯服务系统的数据库。系统

的数据库设计如下图 5-10 所示:

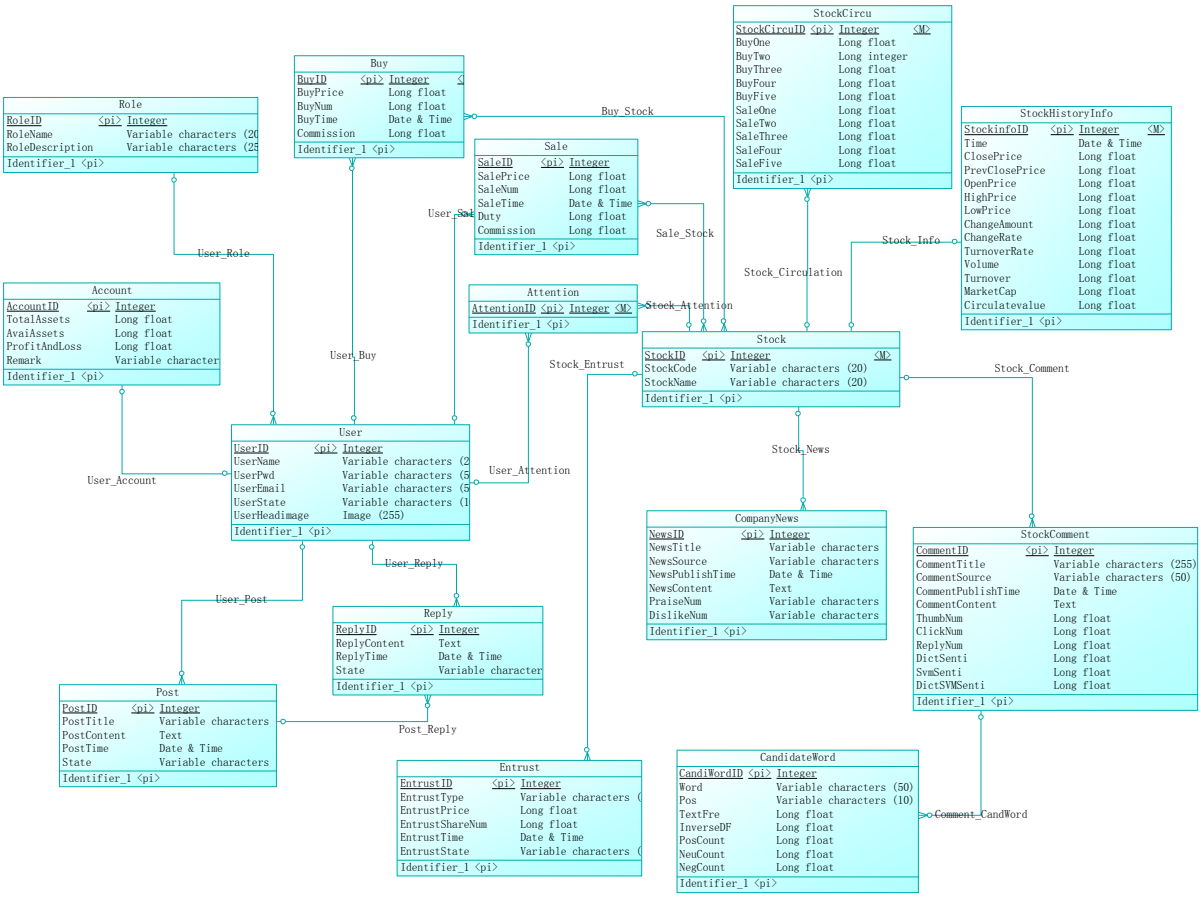


图 5-10 系统数据库的实体关系模型

数据库的业务围绕 User 表,News 表、Stock 表以及 StockComment 表进行设计。整个系统的业务，均由 User 表发起。User 表根据角色类型分为管理员和普通用户。News 表存储股票资讯，支持系统的资讯业务。Stock 表存储股票基本信息，方便和其他表单进行关联，StockHistoryInfo 表存储股票的基本技术指标信息。StockCircu 表存储股票流通信息，是进行模拟交易的关键表单，和 Stock 表是一对一的关系。Buy 表和 Sale 表分别是买入表单和卖出表单，Entrust 表是委托表单，Stock 表和这些表的关系均为一对多。Account 表是用户的账户表单，和 User 表是一对一的关系。StockComment 表支持系统的股评分析业务，Stock 表和 StockComment 表是一对多的关系。CandidateWord 表支持倾向性词库构建业务。

有了以上的分析，下面可以对数据库的各个表格，及表格中字段和类型进行具体的设计。由于篇幅的原因，本文就系统最重要的几张数据库表进行介绍，包括用户表、资讯表、账户表、股票历史信息表、委托表、股票评论表和候选词表，各个表的结构如下 。

1）用户表

用户信息表记录用户的基本信息，其中的 UserID 代表用户编号，也是用户信息表的主键，UserName 代表用户名，UserPwd 代表用户密码， RoleID 代表用户的角色编

号，用户角色一共有两种，分别为普通用户和管理员用户。用户信息表如表 5-1 所示。

表 5-1　用户信息表

字段名称	字段代码	数据类型
用户 ID	UserID	integer
角色 ID	RoleID	integer
用户名	UserName	varchar(20)
用户密码	UserPwd	varchar(50)
用户邮箱	UserEmail	varchar(50)
用户状态	UserState	varchar(10)
用户头像	UserHeadimage	long binary

2）资讯表

资讯表记录股票相关资讯，包括资讯标题、资讯类型、资讯来源、资讯内容等信息。咨询表如表 5-2 所示。

表 5-2　资讯表

字段名称	字段代码	数据类型
资讯 ID	NewsID	integer
资讯标题	NewsTitle	integer
资讯类型	NewsType	varchar(50)
资讯来源	NewsSource	varchar(10)
资讯发布时间	NewsPublishTime	double
资讯内容	NewsContent	double
点赞数	PraiseNum	double
倒彩数	DislikeNum	double

3）账户表

账户表记录用户的账户信息，AccountID 代表账户编号，TotalAssets 是用户的总资产，用户的可用资产是 AvaiAssets，收益为 ProfitAndLoss。账户表如表 5-3 所示。

表 5-3　账户表

字段名称	字段代码	数据类型
账户 ID	AccountID	integer
用户 ID	UserID	integer
总资产	TotalAssets	double
可用资产	AvaiAssets	double
收益	ProfitAndLoss	double
备注	Remark	varchar(255)

4）股票历史信息表

股票历史信息表记录股票的基本技术指标信息，包括收盘价，开盘价，最高价、最低价、涨跌额、成交额、换手率和流通市值等信息。股票历史信息表如表 5-4 所示。

表 5-4　股票历史信息表

字段名称	字段代码	数据类型
股票信息 ID	NewsID	integer
股票 ID	StockID	integer
时间	Time	timestamp
收盘价	ClosePrice	double
前收盘价	PrevClosePrice	double
开盘价	OpenPrice	double
最高价	HighPrice	double
最低价	LowPrice	double
涨跌额	ChangeAmount	double
涨跌幅	ChangeRate	double
换手率	TurnoverRate	double
成交量	Volume	double
成交额	Turnover	double
总市值	MarketCap	double
流通市值	Circulatevalue	double

5）委托表

委托表记录用户委托下单的信息，委托类型 EntrustType 有两种，分别是买入和卖出。委托表如表 5-5 所示。

表 5-5　委托表

字段名称	字段代码	数据类型
委托 ID	EntrustID	integer
股票 ID	StockID	integer
用户 ID	UserID	integer
委托类型	EntrustType	varchar(50)
委托价格	EntrustPrice	double
委托金额	EntrustShareNum	double
委托时间	EntrustTime	timestamp
委托状态	EntrustState	varchar(10)

6）股票评论表

股票评论表用来记录爬取的各大金融财经类网站的股票评论信息。其中的 DictSVMSenti 代表该条股评的情感值。股票评论表如表 5-6 所示。

表 5-6　股票评论表

字段名称	字段代码	数据类型
股评 ID	CommentID	integer
股票 ID	StockID	integer
股评标题	CommentTitle	varchar(255)
股评来源	CommentSource	varchar(50)
股评发布时间	CommentPublishTime	timestamp
股评内容	CommentContent	long varchar
词典情感值	DictSenti	double
SVM 情感值	SvmSenti	double
SC 情感值	DictSVMSenti	double

7）候选词表

候选词表用来记录构建倾向性词库所选出的候选词信息。候选词表如表 5-7 所示。

表 5-7　候选词表

字段名称	字段代码	数据类型
候选词 ID	CandiWordID	integer
股评 ID	CommentID	integer
词	Word	varchar(50)
词性	Pos	varchar(10)
词频	TextFre	double
逆文档频率	InverseDF	double
看涨共现数	PosCount	double
看平共现数	NeuCount	double
看跌共现数	NegCount	double

5.5 本章小结

本章根据上一章的需求分析和功能设计，对该系统需要使用到的主要模块进行了详细的分析和设计。在整体框架的指导下，分别对系统的资讯查询模块、模拟交易模块、股票论坛模块、个人管理模块、管理员管理模块和舆情分析模块进行了详细的刻画。借助 UML 工具，将系统的主要内容进行了设计，包括架构设计、概要设计、详细设计和数据库设计等。在整个设计的过程中，该系统遵循系统分析的要求，对系统的具体功能进行了完整的覆盖，并在舆情分析模块对关键的算法进行了改良，使得系统的股票评价倾向性分析更加优化，并且具有更好的可用性。本章为下一章的系统实现奠定了基础。

6 基于倾向性分析的股票资讯服务系统的实现与测试

6.1 系统的开发环境

1）系统的软件配置

操作系统：Windows864bit

数据库：MySQL5.6.16

JAVA 开发环境：MyEclipse8.5，JDK1.6.0_17

服务器：Tomcat6.0

2）硬件平台参数

设备：ThinkPadT420

CPU：酷睿 i52520M

内存：4GB

硬盘：500GB

6.2 系统主要功能的实现

6.2.1 资讯查询模块的实现

1）关键代码

资讯查询模块主要用于系统首页按照资讯分类展示不同类型的资讯，用户也可以根据信息来源筛选来自不同的网站的资讯。核心代码如下：

```
public String findNewsByTypeName(String stockCode) {
        // TODO Auto-generated method stub
        NewsType newsType = (NewsType)getCurrentSession().createCriteria(NewsType.class).add(Restrictions.eq("typeName", "门户头条"+stockCode)).setMaxResults(1).uniqueResult();
        return newsType.getNewsTypeId();
    }
    @Override
public List<News> findNewsByNewsTypeId(String newsTypeId) {
        List<News> newsList = (List<News>)getCurrentSession().createCriteria(News.class).add(Restrictions.eq("newsType",findById(NewsType.class, newsTypeId))).list();
        return newsList;
    }
    public String findNewsByTypeName1(String stockCode) {
        NewsType newsType = (NewsType)getCurrentSession().createCriteria(NewsType.class).add(Restrictions.eq("typeName", "个股点评"+stockCode)).setMaxResults(1).uniqueResult();
        return newsType.getNewsTypeId();
    }
```

```
@Override
public String findNewsByTypeName2(String stockCode) {
    // TODO Auto-generated method stub
    NewsType newsType = (NewsType)getCurrentSession().createCriteria(NewsType.class).add(Restrictions.eq("typeName", "大盘评述"+stockCode)).setMaxResults(1).uniqueResult();
    return newsType.getNewsTypeId();
    }
```

2）实现效果

图 6-1 资讯查询效果图

6.2.2 股评倾向性分析模块的实现

1）关键代码

本文设计的股评倾向性分析算法是一种词典语义结合 SVM 的方法。因此，在 JAVA 实现时，需要分别对两种算法进行实现。首先，是基于词典语义的文本倾向性分析的核心代码。下面对其中的几个关键步骤进行代码分析：

```
public void OpinionAnalyser() throws SQLException
{
    ConnDB conndb;
    PreparedStatement stmt = null;
//    PreparedStatement stmt = null;
    ResultSet rs = null;
    conndb = new ConnDB(SERVER, USER, PASSWORD, DATABASE);
    conndb.executeUpdate("SET NAMES 'utf8mb4'");
    //获取倾向性词表
    String strSQL = "select word,polar,weight from twordlist";
    try {
        stmt = conndb.getConnection().prepareStatement(strSQL);
```

```
                rs = stmt.executeQuery();
                } catch (SQLException e1) {
                    e1.printStackTrace();
                }
        }
        //句子倾向性得分
        public int sentenceScore(String sentence)
        {
            int opinionScore=0;
            //是否出现倾向词
            int opinionPosition=0;
            for(int i=0;i<words.size();i++)
            {
                //找到倾向性词表
                opinionPosition=sentence.indexOf(words.get(i).getWord());
            //      System.out.println(opinionPosition);
                if(opinionPosition!=-1)
                {
                    int flag=0;
                    for(int j=0;j<adjectives.size();j++)
                    {
                        StringBuffer wordPair=new StringBuffer();
                        wordPair.append(adjectives.get(j).getWord());
                        wordPair.append(words.get(i).getWord());
                        int pairPosition =0;
                        pairPosition=sentence.indexOf(wordPair.toString());
                                            if(pairPosition!=-1)
                        {
                        //      System.out.println("yeyeyeyey");
                            flag=1;
                            int
tmpScore=words.get(i).getWeight()*adjectives.get(j).getWeight()*words.get(i).getPolar()*adjectives.get(j
).getPolar();
                            if(tmpScore>0)
                                opinionScore +=tmpScore;
                            else
                                opinionScore +=tmpScore*NEG_WEIGHT;
                        }
                    }
                    //没出现修饰词只计算倾向词本身的权重
                    if(flag==0)
                    {
                    //      System.out.println(opinionPosition);
                    //      System.out.println("nnnnnnnnnnnn");
                        if(words.get(i).getPolar()==1)
                        {
                            opinionScore+=words.get(i).getWeight()*words.get(i).getPolar();
```

```
                    //              System.out.println(words.get(i).getWord());
                    //              System.out.println("wwwwwwwww");
                        }
                        else if(words.get(i).getPolar()==-1)
                        {
opinionScore+=words.get(i).getWeight()*words.get(i).getPolar()*NEG_WEIGHT;
                    //              System.out.println(words.get(i).getWord());
                        }
                    }
                }
            }
            //System.out.println("最后得分:"+opinionScore);
            return opinionScore;
    }
```

2）实现效果

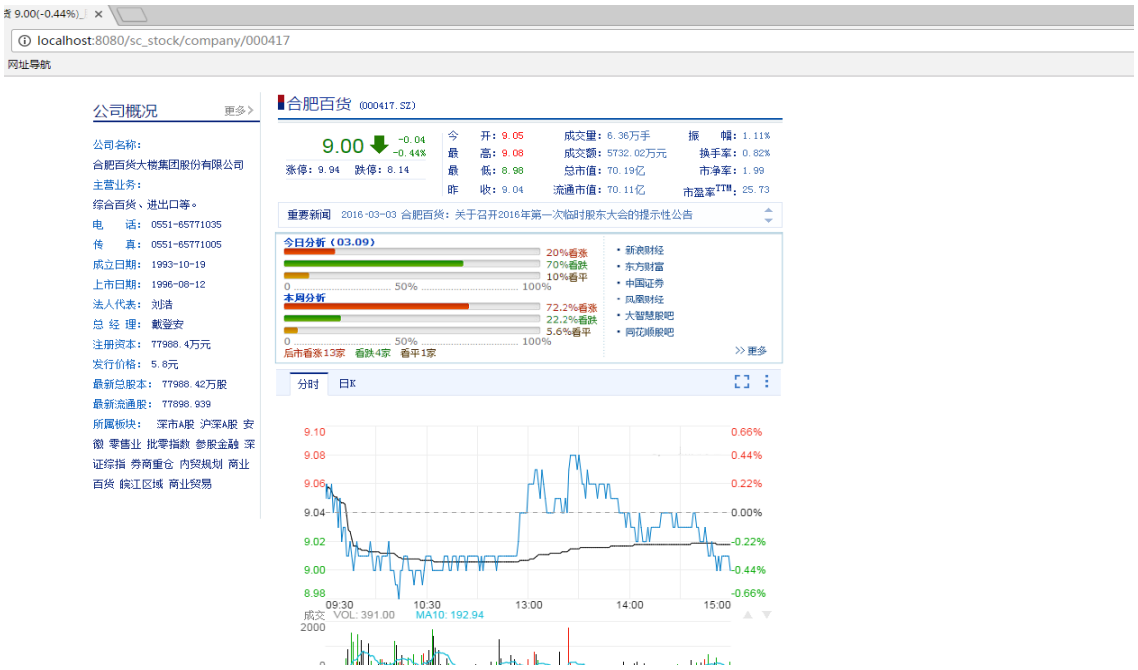


图 6-2 股评倾向性分析效果图

6.2.3 模拟交易模块的实现

1）关键代码

模拟交易是系统最主要的功能之一，其实现的关键在于对股票价格的撮合过程，核心代码如下：

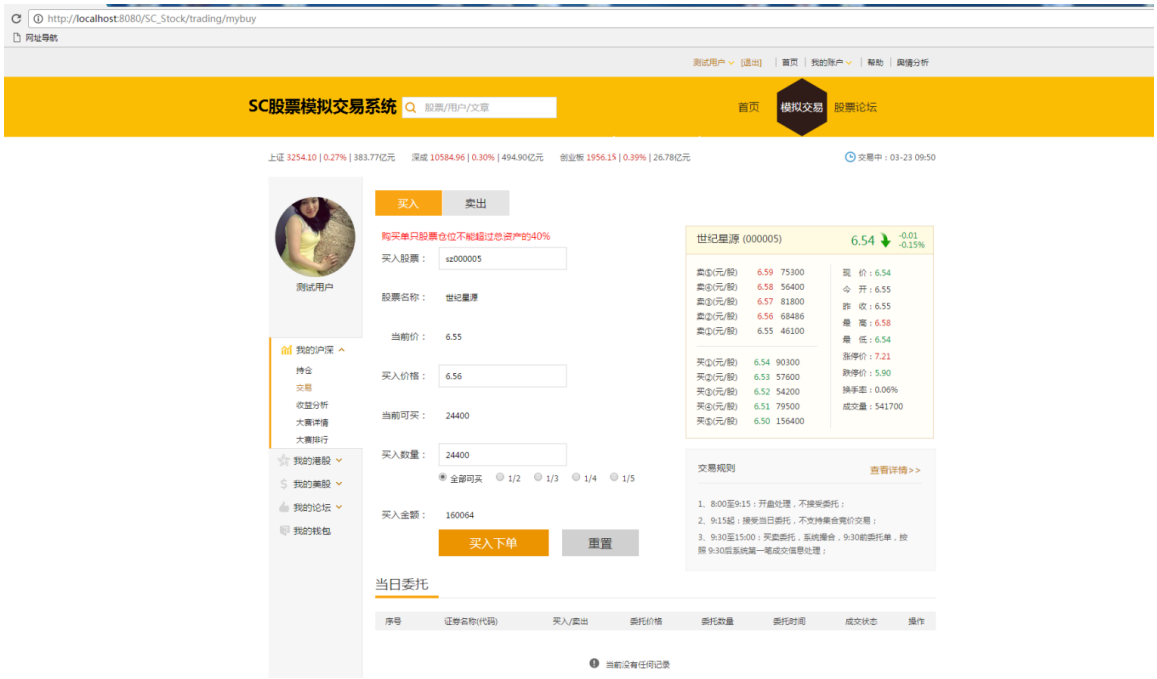
```
public void run() {
        System.out.println("=====开始价格撮合=====");
        List<Stock> stockList = myHoldShareImpl.findAllStock();
        for (Stock stock : stockList) {
            DetachedCriteria dCriteria = DetachedCriteria.forClass(Entrust.class)
                    .add(Restrictions.eq("stock", stock))
```

```
                    .add(Restrictions.eq("entrustState", "未成交"))
                    .add(Restrictions.eq("entrustType", "买入"))
                    .addOrder(Property.forName("entrustPrice").desc())
                    .addOrder(Property.forName("entrustTime").asc());
            //股票购买委托列表
            List<Entrust> entrustBuyList = service.queryAllOfCondition(Entrust.class, dCriteria);
            //合并相同委托价格的股票,内层 List 存相同的价格的股票(买家)
            List<List<Entrust>> buyCombineEntrustList = new ArrayList<List<Entrust>>();
            //处理最开始的一个
            if(entrustBuyList.size() > 0){
                List<Entrust> buyFirstEntrustList = new ArrayList<Entrust>();
                buyCombineEntrustList.add(buyFirstEntrustList);
            }

            //处理中间的委托
            for (int i = 0; i < entrustBuyList.size(); i++) {
                if((i != entrustBuyList.size() - 1)){
                    double d1 = entrustBuyList.get(i).getEntrustPrice() ;
                    double d2 = entrustBuyList.get(i + 1).getEntrustPrice();
                    //注意：下面那个写法必须用 equals，因为 get 到的是对象 不是基本类型
                /*  if(entrustList.get(i).getEntrustPrice()  !=  entrustList.get(i  +
1).getEntrustPrice()){*/
                    if(d1 != d2){
                        List<Entrust> newEntrustList = new ArrayList<Entrust>();
                        buyCombineEntrustList.add(newEntrustList);
                    }
                }
            }
  //…省略部分代码
  /*====查询出与成交价相关的数据信息并做修改====*/
  //例如账户卖家可用的资产、总资产(账号表)；购买表的可用资金;卖表的可用资金；股票流通信
息;股票历史信息
  /*=========价格撮合(为了避免单只股票一直被交易,则每次股票成交完一笔交易之后,则进行下
一只股票的交易)=======*/
    //按照时间顺序，由先竞价的人购买
    for (int i = 0; i < combineEntrustList.size(); i ++) {
    for (int j = 0;j < combineEntrustList.get(i).size(); j ++) {
    double salePrice = combineEntrustList.get(i).get(j).getEntrustPrice(); //卖方的价格
    long saleNum = combineEntrustList.get(i).get(j).getEntrustShareNum(); //卖方的数量
    Date saleTime = combineEntrustList.get(i).get(j).getEntrustTime(); //卖方挂单时间
    User saleUser = combineEntrustList.get(i).get(j).getUser(); //卖方用户
      for(int m = 0; m < buyCombineEntrustList.size(); m++){
        for (int r = 0; r < buyCombineEntrustList.get(m).size(); r++) {
        double buyPrice = buyCombineEntrustList.get(m).get(r).getEntrustPrice(); //买方的价格
```

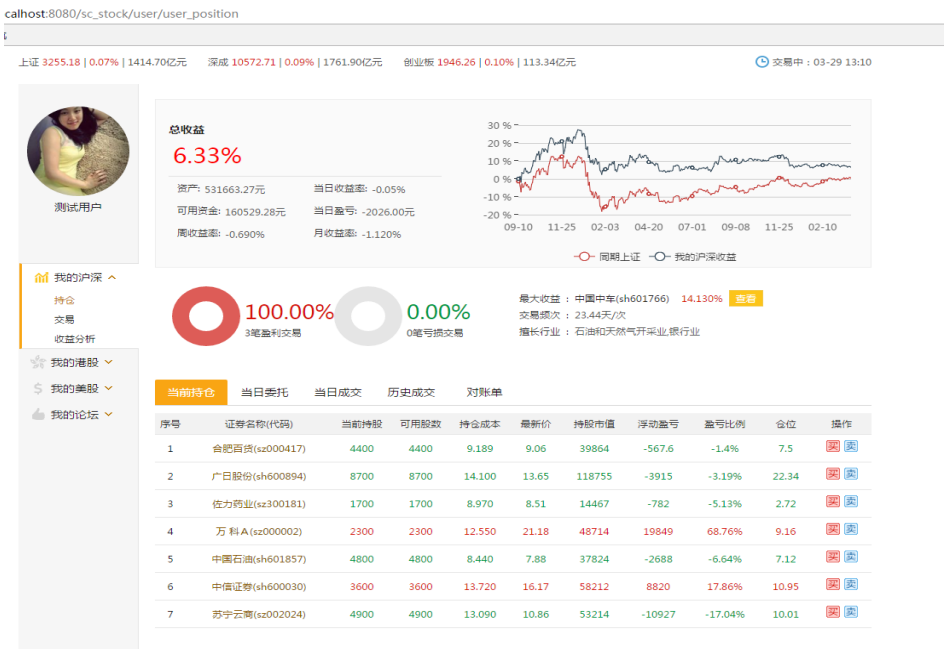
```
long buyNum = buyCombineEntrustList.get(m).get(r).getEntrustShareNum(); //买方的数量
Date buyTime = buyCombineEntrustList.get(m).get(r).getEntrustTime(); //买方的时间
User buyUser = buyCombineEntrustList.get(m).get(r).getUser();//买方用户
//当买方价格高于卖方价格时，则进行交易股数的判断
if(buyPrice >= salePrice){
//如果是卖家的低价先挂单，买家高价出价，那么买家以低价买入；如果买家的高价先挂单，卖家低价出价，那么卖家以高价卖出;
//同时挂单则以中间价为成交价
    if(saleTime.after(buyTime)){
    //买家先挂单，以买家的高价为成交价
    //当买方股数高于卖方股数时，则买方委托股数减少，并生成一笔交易
    //更新股票流通信息中的成交价
    transactionManageServiceImpl.updateBargainPrice(stock, buyPrice);

        if(buyNum > saleNum){
        //计算卖家应获得的钱数(扣除印花税和租金,各占 0.1%)
        double salerObtainMoney = service.round((saleNum * buyPrice) - ((saleNum * buyPrice) * 0.002));
        //生成买卖订单
        transactionManageServiceImpl.geneateSaleOrder(stock,
        combineEntrustList.get(i).get(j).getUser(), buyPrice, saleNum,salerObtainMoney);
        transactionManageServiceImpl.generateBuyOrder(stock,
        buyCombineEntrustList.get(m).get(r).getUser(), buyPrice, saleNum,0.0);
        //更新买入表数量
        transactionManageServiceImpl.updateBuyNumAndEntrustNum(stock, saleUser, saleNum);
        //更新今日股票历史信息
        try {
    transactionManageServiceImpl.updateTodayStockHistoryInfo(stock, buyPrice, saleNum);
            } catch (ParseException e) {
        // TODO Auto-generated catch block
        e.printStackTrace();
    }
    Entrust entrust = new Entrust();
    entrust.setEntrustId(UUIDGenerator.randomUUID());
    entrust.setEntrustPrice(buyPrice);
    entrust.setEntrustShareNum(saleNum);
    entrust.setEntrustState("已成交");
    entrust.setEntrustTime(buyCombineEntrustList.get(m).get(r).getEntrustTime());
    entrust.setEntrustType("买入");
    entrust.setStock(stock);
    entrust.setUser(buyCombineEntrustList.get(m).get(r).getUser());
    service.save(entrust);
    //更新买方委托的数量
    buyCombineEntrustList.get(m).get(r).setEntrustShareNum(buyNum - saleNum);
```

```
service.update(buyCombineEntrustList.get(m).get(r));
//卖方数量少，则只需更新卖方的状态
combineEntrustList.get(i).get(j).setEntrustState("已成交");
service.update(combineEntrustList.get(i).get(j));
//该次交易完成，跳出循环
break;
}}}
```

2）实现效果



（a） 用户模拟买入页面



（b）用户当前持仓页面

图 6-3 股票模拟交易效果图

6.2.5 个人管理模块的实现

1）关键代码

个人管理模块，用户可以修改自己的收藏，关注的股票、关注的用户和个人信息。会员在登陆系统后，可以进行个人管理的相关操作。除了和其他用户之间的私信操作，当用户对系统信息的准确性、系统检索的结果、或系统收录的信息的真实性有所怀疑时，都可以通过私信想管理员进行反馈。用户需要登录自己的账号才能使用该功能模块。用户登录的核心代码如下：

```
//处理用户登录校验
    @RequestMapping("checkUserName.htm")
    public void checkUserName(HttpServletRequest request,
            HttpServletResponse response) throws Exception {

        String username = request.getParameter("userName");
        Map<String, Object> map = new HashMap<String, Object>();
        String result = "";
        boolean b = loginService.queryByName(username);
        if (b == false) {
            result = "false";
        } else {
            result = "true";
        }
        map.put("result", result);
        // map.put("dateResult", dateResult);
        JSONObject json = JSONObject.fromObject(map);
        System.out.println("json       "+json);
        response.getWriter().print(json.toString());
    }
```

用户登录账号后，就可以使用个人管理模块的相关功能。以修改个人信息为例核心代码如下：

```
    @RequestMapping("userupdate.htm")//用户修改个人信息
    public ModelAndView updateUser(HttpServletRequest request,HttpServletResponse response) throws Exception{
        //用户的信息
        String userid=request.getSession().getAttribute("userId").toString();
        User user=(User) request.getSession().getAttribute("user");
        User user2=new User();
        user2=(User) uservice.getCurrentSession().createCriteria(User.class).add(Restrictions.eq("userId", userid)).uniqueResult();
        //获取修改的信息
        String username=request.getParameter("userprofile_name");
        String useremail=request.getParameter("email");
        if(!(username ==null|| username =="")){
        user2.setIntro(username);
```

```
user.setIntro(username);
}
if(!(useremail ==null|| useremail =="")){
user2.setCity(useremail);
user.setCity(useremail);
}
//保存信息
uservice.update(user2);
return new ModelAndView("redirect:gouser.htm");
}
```

2）实现效果

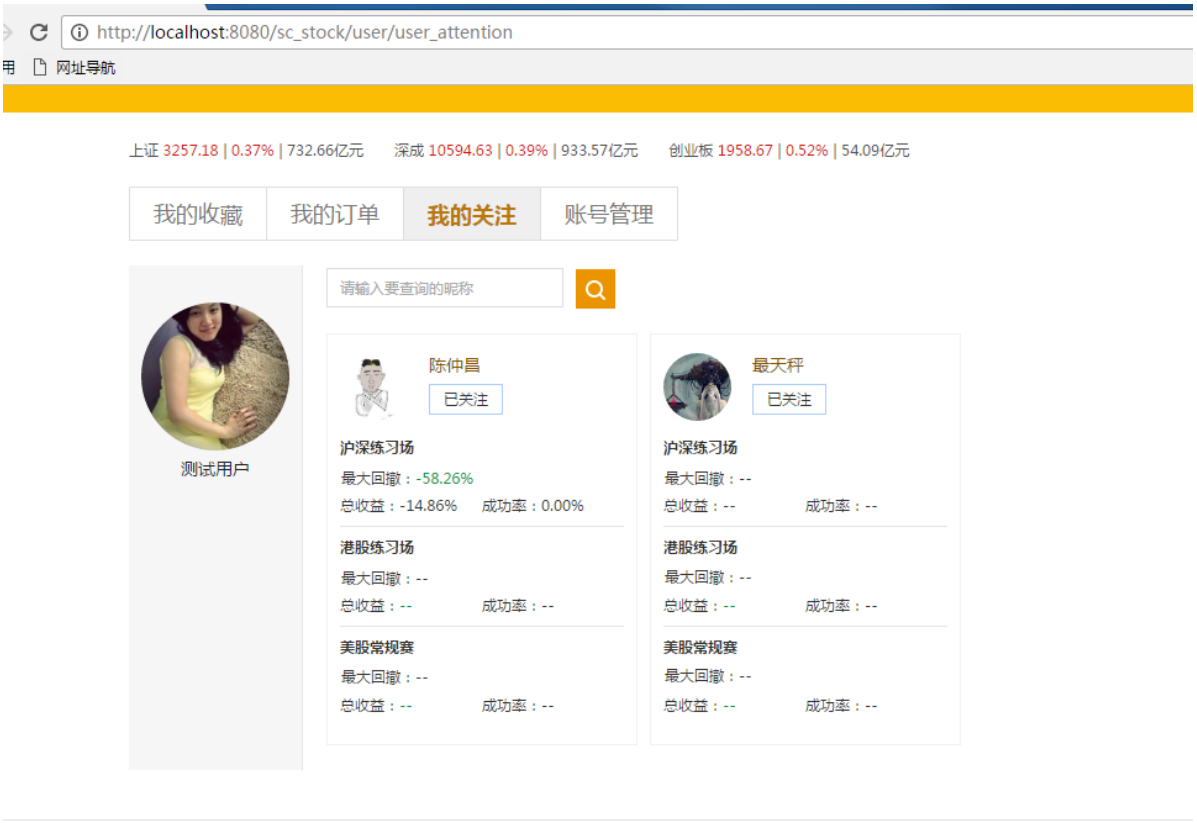


图 6-4 用户个人管理效果图

6.2.4 股票论坛模块的实现

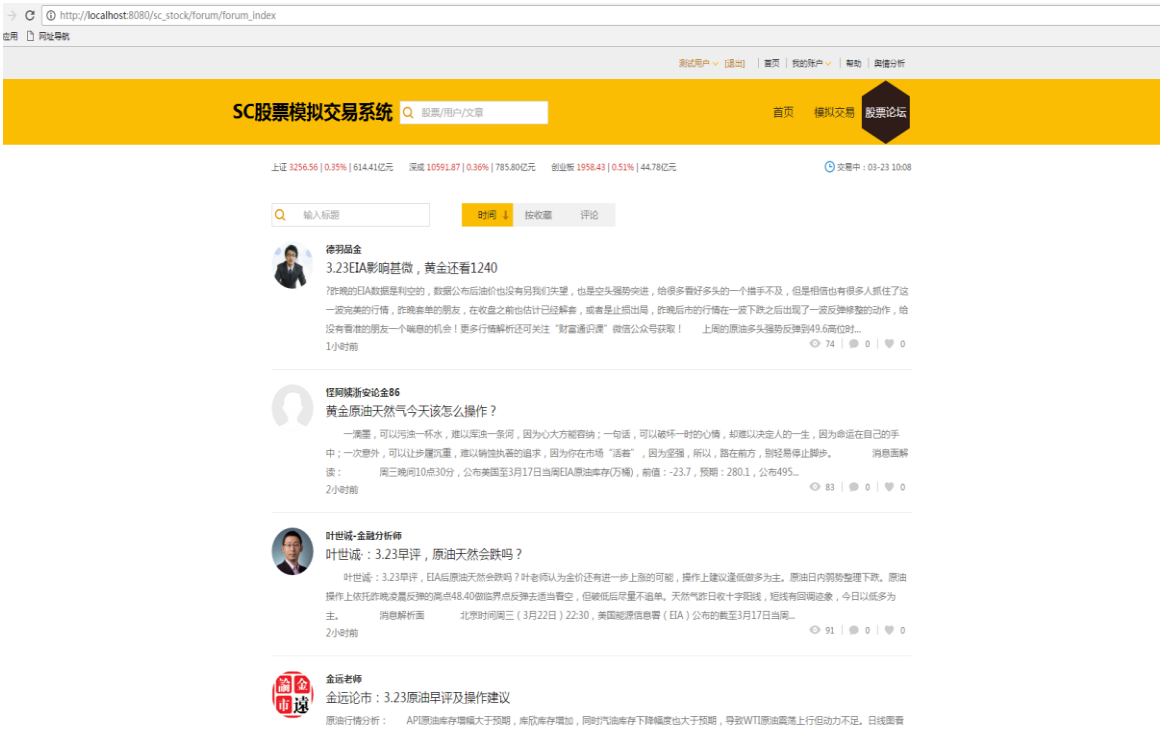
1）关键代码

股票论坛模块的主要功能是用户发帖和回复，核心代码如下：

```
//发帖
@RequestMapping("posting.htm")
public ModelAndView posting(HttpServletRequest request,HttpServletResponse response) throws
ParseException{
User user = (User)request.getSession().getAttribute("user");
String stockBarId = request.getParameter("stockBarId");
StockBar stockBar = barService.findById(StockBar.class, stockBarId);
String title = request.getParameter("title");
```

```
String content = request.getParameter("content");
Thread thread = new Thread();
   thread.setStockBar(stockBar);
   thread.setThreadContent(content);
   thread.setThreadId(UUIDGenerator.randomUUID());
   thread.setThreadName(title);
   thread.setThreadTime(barService.currentTime());
   thread.setUser(user);
   barService.save(thread);

DetachedCriteria detachedCriteria = DetachedCriteria.forClass(Thread.class);
detachedCriteria.add(Restrictions.eq("stockBar", stockBar));
detachedCriteria.addOrder(Order.desc("threadTime"));
List<Thread> threadList =    barService.queryAllOfCondition(Thread.class, detachedCriteria);
   /*ModelMap map = new ModelMap();
map.addAttribute("stockBarId", stockBarId);
//分页
int pageSize = 5;
List<Thread> threadList = barService.findProjectByState(pageSize,stockBar);
int totalPage = (int)threadList.size()/5+1;
PageHelper.forPage(totalPage,pageSize,map);*/
request.setAttribute("stockBarId", stockBarId);
request.setAttribute("tList", threadList);
return new ModelAndView("redirect:barPostList.htm?stockBarId="+stockBarId);
 }
```

2）实现效果



（a）股票论坛首页

（b）股票论坛帖子详情页面

图 6-5 股票论坛效果图

6.2.6 管理员管理模块的实现

1）关键代码

管理员管理模块包括管理员对资讯、股票、论坛、用户和个人信息的管理。

```
public class AdminManageController {
    @Autowired
    private AdminManageService adminManage;
    //跳转至资讯管理界面
    /*@RequestMapping("newsList.htm")
    public ModelAndView goNewsList(){
        System.out.print("资讯管理界面");
        return    new ModelAndView("/admin/newsList");
    }*/
    //跳转至添加新闻页面
    @RequestMapping("goAddNewsPage.htm")
    public ModelAndView goAddNewsPage(HttpServletRequest request){
        List<NewsType> NewsTypeList = adminManage.findAllNewsType();
        request.setAttribute("NewsTypeList", NewsTypeList);
        System.out.print("添加新闻界面");
```

```
        return    new ModelAndView("/admin/newsAdd");
    }
    //添加新闻
    @RequestMapping("addNews.htm")
    public void adminCreateNews(HttpServletRequest request,
            HttpServletResponse response) throws ServletRequestBindingException, IOException,
ParseException{
        User user = (User) request.getSession().getAttribute("user");
        if (user != null) {

            /*
             * if(announcementTopic != null){ announcementTopic = new
             * String(announcementTopic.getBytes("iso-8859-1"),"utf-8"); }
             */
            String newsTypeId = ServletRequestUtils.getStringParameter(request,
                    "newsTypeId");

            String newsTitle = ServletRequestUtils.getStringParameter(
                    request, "newsTitle");
            String newsContent = ServletRequestUtils.getStringParameter(
                    request, "newsContent");
            String source = ServletRequestUtils.getStringParameter(
                    request, "source");
            String data = "";
            adminManage.publishNews(newsTypeId, newsTitle, newsContent, source);
            response.setCharacterEncoding("utf-8");
            response.setContentType("text/html; charset=utf-8");
            PrintWriter pw = response.getWriter();
            pw.print(data);
            pw.flush();
            pw.close();
        }
    }
    //添加股票
        @RequestMapping("addStock.htm")
        public void adminCreateStock(HttpServletRequest request,
                HttpServletResponse response) throws ServletRequestBindingException,
IOException{
            User user = (User) request.getSession().getAttribute("user");
            if (user != null) {
                String stockCode = ServletRequestUtils.getStringParameter(request,
                        "stockCode");
                String stockName = ServletRequestUtils.getStringParameter(
                        request, "stockName");
                long stockCirculation = Long.parseLong( ServletRequestUtils.getStringParameter(
                        request, "stockCirculation"));
                String plate = ServletRequestUtils.getStringParameter(
```

```
                request, "plate");
        String data = "";
        adminManage.createStock(stockCode, stockName, plate, stockCirculation);

        response.setCharacterEncoding("utf-8");
        response.setContentType("text/html; charset=utf-8");
        PrintWriter pw = response.getWriter();
        pw.print(data);
        pw.flush();
        pw.close();
    }
}
```

2）实现效果

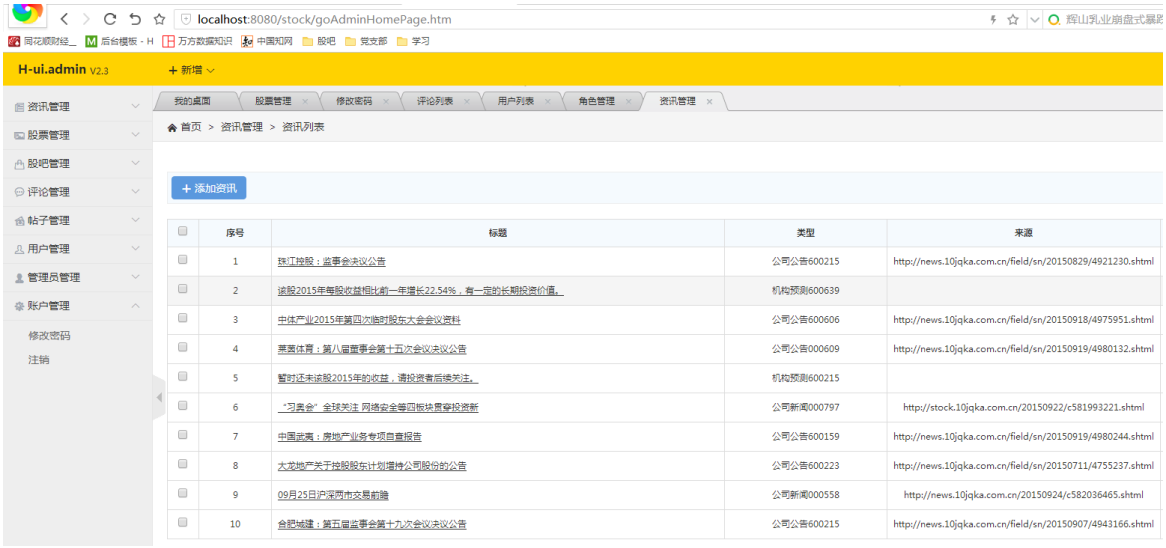


图 6-6 管理员管理效果图

6.3 系统的测试

6.3.1 系统测试环境

软件测试是软件开发过程中的一个重要组成部分，是贯穿整个软件开发生命周期、对软件产品进行验证和确认的活动过程，软件测试的目的是保证本文设计的系统能够正常运行，并且在测试的过程中能够即时发现问题，解决问题。本文的测试环境如下：操作系统为 Windows7，Java 开发环境是 jdk1.6.0_17，以 MyEclipse 为集成开发环境，以 Tomcat6.0 作为 web 服务器，以 MySQLServer5.0 作为数据库。

6.3.2 系统功能测试

本项目测试完成了对前面需求分析和系统实现描述的各功能模块，对各个功能是否与需求分析一致进行测试，测试用例表和功能测试结果如下表所示：

表 6-1 资讯查询功能测试用例表

测试模块	测试项目	执行步骤	测试结果	是否通过
资讯查询	资讯浏览	进入系统，浏览首页信息	系统显示首页资讯	是
	资讯检索	1.在搜索框中输入关键词，选择检索类型，点击检索按钮； 2.系统返回检索结果；	显示结果列表	是
	股评分析	1. 在搜索框中输入关键词，选择检索类型，点击检索按钮； 2. 系统返回检索结果。 3.点击查看股拼倾向性分析。	系统显示股评倾向性分析	是
	资讯定制	1.用户登陆账号后，进入资讯首页 2.选择资讯首页中的信息来源标签。	系统显示过滤后的资讯	是

个人管理模块的测试用例如下表 6-2 所示：

表 6-2 个人管理功能测试用例表

测试模块	测试项目	执行步骤	测试结果	是否通过
个人管理	系统初始化	检查系统在正常情况下启动是否正常	显示系统首页	是
	用户注册	1.点击页面右上角的注册，显示注册页面； 2.在注册页面填写用户名、登录密码等必填信息，点击确定； 3.弹出注册成功提示窗口，点击确定。	显示登录页面	是
	正常登录	1.点击页面右上角的登录图标，显示登录页面； 2.在登录页面填写用户名、登录密码信息和选择用户类型，点击确定。	按照用户的类型显示系统登录后的页面	是
	错误登录	1.点击页面右上角的登录图标，显示登录页面； 2.在登录页面填写错误的用户名、登录密码信息，或者填写与用户名不匹配的用户类型，点击确定。	弹出提醒用户名或密码错误的提示窗口	是
	修改个人信息	1.点击页面上方的个人信息； 2.在显示的个人信息页面选择修改按钮，填写需要修改的内容，点击确定。	个人信息页面显示修改以后的信息	是

表 6-2（续）

测试模块	测试项目	执行步骤	测试结果	是否通过
个人管理	查看个人信息	点击页面上方的个人信息。	显示个人信息页面	是
	删除收藏帖子	1.在会员登陆条件下点击页面上方的收藏夹管理； 2.在帖子管理页面点击查看帖子信息； 3.选择要删除的帖子，在其后面的删除按钮上点击确定。	删除所选帖子，显示剩余帖子的信息	是
	删除关注用户	1.在会员登陆条件下点击页面上方的歌单管理； 2.在歌单页面点击查看歌单信息； 3.选择要删除的歌单，在其后面的删除按钮上点击确定。	删除所选歌单，显示剩余歌单的信息	是
	注销	在会员登陆的条件下，点击页面右上角的注销按钮。	退出用户登陆状态，返回系统非登录状态的首页信息	是

模拟交易模块的测试用例如下表 6-3 所示：

表 6-3 模拟交易功能测试用例表

测试模块	测试项目	执行步骤	测试结果	是否通过
模拟交易	买进股票	1.进入个人交易界面； 2.输入需要买入的股票编号，系统显示股票信息； 3.选择要买入的股票数量，点击确认交易。	系统显示交易成功，交易记录增加	是
	卖出股票	1. 进入个人交易界面； 2. 输入需要卖出的股票编号，系统显示股票信息； 3. 选择要卖出的股票数量，点击确认交易。	系统显示交易成功，交易记录增加	是
	查看持仓	1.进入个人交易界面，选择查看持仓功能； 2.系统显示个人持仓情况。	系统显示个人持仓信息	是
	盈收分析	1. 进入个人交易界面，选择查看营收分析； 2.系统显示个人盈收情况	系统显示个人盈收信息	是

股票论坛是只允许会员使用的系统功能。用户在登陆系统后，点击股票论坛板块，即可进入相应的系统功能页面。股票的论坛功能包括发帖、回帖、关注其他用户、收藏帖子以及向其他用户发送私信等功能。股票论坛模块的测试用例如下表 6-4 所示：

表 6-4 股票论坛功能测试用例表

测试模块	测试项目	执行步骤	测试结果	是否通过
股票论坛	发帖	1.点击发帖按钮； 2.填写帖子内容，包括标题和正文； 3.点击发送按钮。	显示已发送的帖子	是
	回帖	1.进入他人的帖子，点击回帖按钮； 2. 填写帖子内容； 3. 点击发送按钮	显示已发送的帖子	是
	收藏帖子	在用户登录模式下，用户选择某一帖子下方的收藏按钮，将该帖子添加到个人的收藏夹	收藏列表显示收藏帖子	是
	关注用户	1.点击帖子页面中的用户头像； 2.在他人的主页中点击关注按钮。	系统提示已关注该用户	是
	发送私信	1.点击私信按钮； 2.填写私信内容，包括收件人ID和正文。 3. 点击私信发送	收件人收到私信	是

管理员管理模块的测试用例如下表 6-5 所示：

表 6-5 管理员管理功能测试用例表

测试模块	测试项目	执行步骤	测试结果	是否通过
管理员管理	查看用户信息	1.在管理员登陆条件下点击页面上方的用户管理； 2.在用户管理页面点击查看用户信息。	显示所有系统用户的信息	是
	删除用户	1.在管理员登陆条件下点击页面上方的用户管理； 2.在用户管理页面点击查看用户信息； 3.选择要删除的用户，在其后面的删除按钮上点击确定。	删除所选用户，显示剩余用户的信息	是
	查看帖子信息	1.在管理员登陆条件下点击页面上方的论坛管理； 2.在帖子管理页面点击查看帖子信息。	显示所有系统帖子的信息	是

表 6-5（续）

测试模块	测试项目	执行步骤	测试结果	是否通过
管理员管理	删除帖子	1.在管理员登陆条件下点击页面上方的帖子管理； 2.在帖子管理页面点击查看用户信息； 3.选择要删除的帖子，在其后面的删除按钮上点击确定。	删除所选帖子，显示剩余帖子的信息	
	修改帖子信息	1.在管理员登陆条件下，点击页面上方的帖子管理； 2显示帖子列表信息，选择要修改的帖子，填写修改信息，点击确定。	显示修改后的帖子信息	是
	查看资讯信息	1.在管理员登陆条件下点击页面上方的论坛管理； 2.在资讯管理页面点击查看资讯信息。	显示所有系统资讯的信息	
	删除资讯	1.在管理员登陆条件下点击页面上方的资讯管理； 2.在资讯管理页面点击查看用户信息； 3.选择要删除的资讯，在其后面的删除按钮上点击确定。	删除所选资讯，显示剩余资讯的信息	
	修改资讯信息	1.在管理员登陆条件下，点击页面上方的资讯管理； 2显示资讯列表信息，选择要修改的资讯，填写修改信息，点击确定。	显示修改后的资讯信息	
	管理日志	1.在管理员登陆条件下，点击页面上方的日志管理； 2.显示日志列表，并可以对每条日志做查看和删除操作。	显示管理之后的日志列表	是

6.3.3 系统性能测试

性能测试主要对响应时间、事务处理速率和其他与时间相关的需求进行评测和评估。性能评测的目标是核实在正常的预期工作量和预期的最繁重工作量情况下性能需求是否都已满足。由需求分析得到的性能需求指标如下：

并发用户数小于 1000，在 1000 个并发用户进行股票检索时，业务处理响应时间在 5 秒以内；对股票进行交易时，不计入网络传输时间，买进和卖出的响应时间在 3 秒以内。当使用股票论坛功能时，回帖、发帖的响应时间应该在 5 秒以内。

系统实现后，在搭建的 LoadRunner 测试环境中，让系统的服务器一直处于开启状

态，让系统持续运行了 3 天，没有发生意外中断的情况，可以有效工作，客观地证明了系统满足稳定性指标。性能测试的结果如下表 6-6 所示（秒：s）。

表 6-6　性能测试结果表

并发用户数	事务平均响应时间			事务最大响应时间		
	股票检索	股票交易	股票论坛	股票检索	股票交易	股票论坛
50	1.0s	0.6s	0.9s	1.4s	0.8s	1.1s
100	1.3s	0.7s	1.0s	1.6s	1.0s	1.2s
200	1.6s	0.9s	1.3s	1.8s	1.1s	1.6s
500	2.2	1.2s	1.5s	2.6s	1.5s	1.9s
1000	3.7s	1.8s	2.6s	4.3s	2.4s	4.1s

从测试结果来分析，在 LoadRunner 测试的过程中，发现 100 个以内用户同时使用系统时，系统的处理速度变化不大；200 个以内用户同时使用系统时，系统能够快速准确地完成业务的处理；在 200 至 500 个用户同时使用系统时，系统的处理时间加长，并且有处理失败的情况，但系统仍能正常运行。当并发数达到 1000 时，发帖的错误率有一定上升，并出现检索失败的情况，但系统人能够正常运行。测试说明该系统达到了支持 1000 个用户同时操作的目标，保证了系统在需求范围内正常工作，满足了业务处理能力指标的要求。

6.4 本章小结

本章将基于股票评价的股票资讯服务系统系统的实现、实现的核心代码、功能截图以及测试过程进行详细的展示，以描述系统的实现与测试。在代码与功能展示上，以模块为单位，具体描述了系统股评分析模块、资讯查询模块、模拟交易模块、用户管理模块和管理员管理模块的和核心代码与系统界面。随后，本章对系统的功能进行了用例测试，保证了系统的实用性。

7 结论与展望

7.1 总结

自证券市场建立以来，作为高风险与高收益并存的股票市场，一直受到众多投资者的关注。股票不仅能给投资者带来个人利益，也为国家的经济发展做出了巨大的贡献。近年来，随着信息科技的发展，股票市场也在互联网技术的支持下取得了很大进步，越来越多的学者将倾向性分析技术与股票结合起来，从而带来了金融舆情在证券领域的研究热潮。本文真实在这一背景下，针对股票评论的特点，提出了一种改进的文本倾向性分析方法，并设计实现了一个基于股评倾向性的股票资讯服务系统。本文的具体工作如下：

1）针对股票评论专业词汇量大、词性特征显著、以及篇章结构复杂的特点，本文提出了一种词典语义和 SVM 相结合的股评倾向性分析方法。模型首先使用半监督的方式构建股票专业词汇词典库，然后提出了一种新的词性情感特征提取方法，对带有情感极性的语料句按词法规则找出所有带有情感倾向的词性特征，并通过词性最大匹配算法依据情感识别准确率和占有率提取词性情感特征。最后将词典语义分析方法与 SVM 模型相结合，得到最终的倾向性识别结果。本文通过实验证明，这种改进的文本倾向性分析方法能够有效提高针对股评的倾向性分析查准率和查全率，尤其在反向查准率上进步显著。

2）在对股评文本的倾向性进行分析之后，本文设计并实现了一个基于股票评价倾向性分析的股票资讯服务系统。该系统可以为用户提供最新的股票资讯，并对各大财经门户的股票评论倾向性进行计算，分析网友的投资意向，从而为用户的股票交易提供参考。系统的模拟交易功能为用户熟悉交易流程、练习交易策略提供了真实的环境。此外，用户可以在系统的股票论坛中交流交易经验，分享股票新闻。本文从系统的需求分析、建模设计、实现测试三个角度对系统的构建过程进行了详细的描述，并对系统按照测试用例进行了测试，给出运行结果。测试结果表明，本系统能够满足用户对股票资讯查询和模拟交易的需求，在证券领域有重要的意义和价值。

7.2 展望

本文针对股票评论的特点，提出了一种针对股评的倾向性分析算法，并设计、实现了一个基于股评的股票资讯服务系统。然而，由于自身的知识和能力的限制，本文在算法设计和系统实现上仍然存在一些不足，具体如下：

1）股评倾向性分析算法中，我们主要选用了基于词典的方法和基于支持向量机的情感分类方法作为原型进行改进，缺乏和其他方法的对比，算法实验中也缺少多种算法的分析结果对比，从而导致本文提出的算法模型可能并不是问题的最优解，在其他

分类模型下可能有更大的改进空间。

2）在对股评的分析上，本文选取的数据来源，尤其是股票评论往往是以天为单位进行获取和分析。而事实上在量化交易、高频交易逐渐成熟的今天，以天为单位进行的金融舆情计算已经不能满足用户的需求，获取更多的金融舆情信息，以更高的速率进行计算，才能够为用户提供能加有价值的信息。

3）在金融舆情对证券交易的影响研究上，本文所设计的系统只能提供股票评论的倾向性分析，对股票的分析角度过于单一，没有深入探讨金融舆情、股评倾向性和用户购买行为之间的关系，从而不能从更高的角度为用户的交易提供指导。之主要是由于作者缺乏相关的金融学专业知识，因此在有限的时间内无法进行深入的研究。

4）从系统的设计与实现看，系统在测试阶段没有进行压力测试，因此无法确定在大流量访问下的性能是否稳定。

针对以上的问题，作者将在后续的工作中从算法和系统两个角度做出改进。算法上，将会考虑更多文本倾向性分类模型，对多种分类方法进行对比验证。同时，将对系统进行更多测试，以便于及时发现问题，完善本文的工作。

致　谢

三年的研究生生活很快就要结束了，在我读研和撰写论文的过程中，得到了很多人的帮助，在此表示衷心的感谢，并致以我深深的敬意。

我最需要感谢的是我的论文指导老师——饶元老师。从科研课题的确立到指导科研阶段，再到确定初期的整个论文的思路和方向，直至论文的精心审阅，饶老师都认真负责地指导我、教育我，不仅使我在技术上有了明显的进步，而且思想上提升了高度。尤其是在我撰写论文的过程中，饶老师对论文严格要求，多次指导，监督论文进度并提出很多有用的意见，使我受益匪浅。饶老师治学态度严谨，工作一丝不苟，对学术精益求精，这都将成为我以后工作和学习的榜样。感谢饶老师对我的悉心指导和热情鼓励，向饶老师致以我深深的谢意。

我还要感谢软件学院的各位领导和老师，是他们给予了我很多指导和帮助，为我们创造了和谐的学习氛围。谢谢各位老师们。

感谢我的父母及家人，感谢他们为我默默地奉献，支持我，关心我，在我遇到困难的时候给我信心和力量，这是我一生都难以回报的。感谢我的同学和朋友们，他们对我无微不至的关心和支持，给了我前进的动力。

感谢给予参考和引用权的资料、图片、文献、研究成就的作者，是他们的作品指导我在这一领域有了更大的研究和进步。

最后感谢参加论文评审和答辩的各位老师们，谢谢老师对本文的批评指正。

三年的研究生时光快要结束了，现在回想起来不免对研究生学习和生活有几分留恋。从三年前我入学到论文撰写的过程，我得到了很多人的帮助，在这里我表示由衷的励。家人的支持和包容是我人生奋斗和前进的动力。感谢和真挚的敬意。

我最需要感谢的是我论文的指导老师——饶元老师。我很庆幸在我研究生的时光里能遇到饶老师这样的良师。这三年来，饶老师不仅在学习上给了我很大的鼓励和帮助，同时在生活上也让我受益匪浅。他一丝不苟的科研态度、对学术饱满的热情以及忘我的工作精神都给我留下了深刻的印象。在我写论文期间，尽管饶老师远在美国，但仍然从论文的选题、文献调研到算法研究和最终的写作上，给予了我最耐心和深刻的讲解与指导。在我写论文阶段，饶老师也多次提出宝贵意见，帮助和指引我顺利的完成论文，他勤恳的工作作风以及严谨的科研态度对我影响至深，使我受用终身。在此，对饶老师的耐心指导和帮助表示深深的感谢。

感谢在学习上帮助和指引我的软件学院的各位老师和领导。同时还要感谢软件学

参考文献

[1]刘苇. 分析师评价对股票预测价值的差异性研究[D].大连理工大学,2016.

[2]张兵, 李晓明. 中国股票市场的渐进有效性研究[J]. 经济研究, 2003(1):54-61.

[3]孙培源, 施东晖. 基于 CAPM 的中国股市羊群行为研究——兼与宋军、吴冲锋先生商榷[J]. 经济研究, 2002(2):64-70.沈云霞.

[4]基于股吧舆情的投资者情绪与股票收益研究[D]. 天津工业大学, 2015.

[5]Deniz Aldoğan,Yusuf Yaslan. A Comparison Study On Active Learning Integrated Ensemble Approaches In Sentiment Analysis[J]. Computers and Electrical Engineering,2016,:.

[6]Yang A M, Lin J H, Zhou Y M, et al. Research on Building a Chinese Sentiment Lexicon Based on SO-PMI[J]. Applied Mechanics & Materials, 2012, 263-266:1688-1693..

[7]张美娜, 迟呈英, 战学刚,等. 基于篇章结构的文本自动标引算法[J]. 计算机应用与软件, 2008, 25(9):122-124.

[8]Chun Liao,Chong Feng,Sen Yang,Heyan Huang. Topic-Related Chinese Message Sentiment Analysis[J]. Neurocomputing,2016,:.

[9]蒋润. 基于语义的文本倾向性分析[D].华东理工大学,2014.

[10]黄进. 金融领域中基于 UGC 的情感分析[D].华东理工大学,2014.

[11]袁晨, 傅强. “T+1”交易制度下非线性证券价格动态模型及实证[J]. 管理科学学报, 2011, 14(3):83-96.

[12]吴云勇, 范树杰. 证券投资分析方法研究[J]. 中国市场, 2012(27):76-77.

[13]彭佳星, 肖基毅. 基于分型转折点的证券时间序列分段表示法[J]. 商, 2016(31):195-196.

[14]潘立先, 朱玉峰. SVM 在证券选股中的应用[J]. 商情, 2014(16):34-34.

[15]Nicolas Pröllochs,Stefan Feuerriegel,Dirk Neumann. Negation scope detection in sentiment analysis: Decision support for news-driven trading[J]. Decision Support Systems,2016,:.

[16]祁斌, 黄明, 陈卓思. 机构投资者与市场有效性[J]. 金融研究, 2006(3):76-84.

[17]Wei G, Zhang W, Zhou L. Stock trends prediction combining the public opinion analysis[C]// International Conference on Logistics, Informatics and Service Sciences. IEEE, 2015.

[18]石研. 中国财经媒体传播失灵现象研究[D].武汉大学,2010.

[19] Howell D. Moving Public Opinion--And Stock Prices[J]. Chain Store Age, 2006.

[20] 徐琳. 网络舆情对股价波动影响的实证研究[D].西南财经大学,2013.

[21]Ahuja R, Rastogi H, Choudhuri A, et al. Stock market forecast using sentiment analysis[C]// International Conference on Computing for Sustainable Global Development. IEEE, 2015:1008-1010.

[22]Zhang D, Qi J. Research on the Stock Price Shock Effects of the Internet Public Opinion of Enterprise's Emergency Crisis Incident Based on Microblog[J]. Journal of Intelligence, 2015.

[23]张超. 文本倾向性分析在舆情监控系统中的应用研究[D].北京邮电大学,2008.

[24]Farhan Hassan Khan,Usman Qamar,Saba Bashir. SWIMS: Semi-Supervised Subjective Feature Weighting and Intelligent Model Selection for Sentiment Analysis[J]. Knowledge-Based Systems,2016,:.

[25]Doaa Mohey El-Din Mohamed Hussein. A Survey on Sentiment Analysis Challenges[J]. Journal of King Saud University - Engineering Sciences,2016,:.

[26]Deepa Anand,Deepan Naorem. Semi-supervised Aspect Based Sentiment Analysis for Movies Using

Review Filtering[J]. Procedia Computer Science,2016,84:.

[27]Orestes Appel,Francisco Chiclana,Jenny Carter,Hamido Fujita. A Hybrid Approach to the Sentiment Analysis Problem at the Sentence Level[J]. Knowledge-Based Systems,2016,:.

[28]龙树全, 赵正文, 唐华. 中文分词算法概述[J]. 电脑知识与技术, 2009, 5(4):2605-2607.

[29]Andrea Ceron,Luigi Curini,Stefano Maria Iacus. iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content[J]. Information Sciences,2016,:.

[30]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,(04):978-981.

[31]Rui Neves-Silva,Marta Gamito,Paulo Pina,Ana Rita Campos. Modelling Influence and Reach in Sentiment Analysis[J]. Procedia CIRP,2016,47:.

[32]Matthijs Meire,Michel Ballings,Dirk Van den Poel. The added value of auxiliary data in sentiment analysis of Facebook posts[J]. Decision Support Systems,2016,:.

[33]Ioannis Korkontzelos,Azadeh Nikfarjam,Matthew Shardlow,Abeed Sarker,Sophia Ananiadou,Graciela H. Gonzalez. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts[J]. Journal of Biomedical Informatics,2016,:.

[34]杨伟杰,马博渊,刘雯. 基于意见目标句抽取的中文股评情感分析方法[J]. 计算机仿真,2014,(03):431-436.

[35]R. Piryani,D. Madhavi,V.K. Singh. Analytical Mapping of Opinion Mining and Sentiment Analysis Research during 2000 – 2015[J]. Information Processing and Management,2016,:.

[36]Tao Chen,Ruifeng Xu,Yulan He,Xuan Wang. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN[J]. Expert Systems With Applications,2016,:.

[37]丘桥云. 结合文本倾向性分析的股评可信度计算研究[D].哈尔滨工业大学,2014.

[38]Charalampos Karyotis,Faiyaz Doctor,Rahat Iqbal,Anne James,Victor Chang. A fuzzy computational model of emotion for cloud based sentiment analysis[J]. Information Sciences,2017,:.

[39]莫倩,张渝杰,胡航丽,张华平. 一种混合的股评观点倾向性分析方法[J]. 计算机工程与应用,2011,(19):222-225.

[40]张斌斌. 网络股评的倾向性分析[D].中央民族大学,2015.

攻读硕士期间发表的学术论文

1 Deng K, Wang K, Ma D. A New Solution of Distributed Disaster Recovery Based on Raptor Code[M]// Proceedings of the 2015 International Conference on Applied Mechanics, Mechatronics and Intelligent Systems (AMMIS2015). 2015:825-832.

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．交回学校授予的学位证书；

2．学校可在相关媒体上对作者本人的行为进行通报；

3．本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。

4．本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）：日期：年月日

学位论文独创性声明（2）

本人声明：研究生所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．学校可在相关媒体上对本人的失察行为进行通报；

2．本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。

3．本人接受学校按照有关规定做出的任何处理。

指导教师（签名）：日期：年月日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）：日期：年月日

指导教师（签名）：日期：年月日