清 华 大 学

综 合 论 文 训 练

题目：基于公众关注度的股票市场的分析与预测

系　　别：计算机科学与技术系

专　　业：计算机科学与技术

姓　　名：李雨田

指导教师：余宏亮副教授

2015 年 6 月 29 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签　名：_____ 导师签名：_____ 日　期：_____

中文摘要

本文通过公众关注度信息对股票市场进行分析预测。本文首先基于异步事件循环机制实现了公众关注度数据采集及预处理系统，能够高效抓取网络公众关注度数据，经过预处理序列化成通用的数据结构。通过抓取东方财富网股吧讨论区数据，然后利用整合数据及相关数学模型搭建基于 Pandas 和 StatsModels 的分析预测平台，分析论帖点击量与股票成交量和价格的关系。例如使用格兰杰因果关系检验发现内在相关关系，使用向量自回归模型进行回归拟合。本文还通过引入基于神经网络模型的情感极性分析提取讨论帖点击量中的情感因素，得到公众关注度信息中的各情感成分。使用以上数据进一步得到基于公众关注度的股票市场的预测模型。

实验结果表明，公众关注度对股票市场成交量与价格有很强的格兰杰因果关系，基于该关系可以构建出回归预测模型。结合公众关注度情感信息及其各情感分量，可以优化构建出相对精确的股票市场成交量与价格的预测模型。

关键词：股票市场；成交量；价格；预测；公众关注度

ABSTRACT

This paper utilizes public attention to analyze the stock market. Based on event driven architecture, this paper implements a data collection and preprocessing system, capable of highly efficient scraping of public attention data on the Internet. After preprocessing, the data is serialized into general data structure. Guba posts and the stock market are collected and aggregated, combined with mathematical models. I constructed an analysis and prediction platform based on Pandas and StatsModels, which is able to analyze the relationship between click count and the stock market. For example, Granger causality test is used to discover correlation relationship, and vector auto regression model is used for regression analysis. This paper also takes advantage of sentiment analysis tool, which is based on convolutional neural networks, to extract sentiment components from data, obtaining different sentiment components in public attention. All data is used to construct a predction model for the stock market.

As shown by the result, public attention holds a strong Granger causality towards trading volume and prices. It is possible to construct a regression prediction model based on this relationship. Combined with sentimental data, a rather accurate prediction model could be constructed.

Keywords: stock market; volume; price; prediction; public focus

目　录

第 1 章　引言

1.1　研究的意义及实用价值

股票市场在我国金融体系中扮演重要的组成部分，它不仅反映了国民经济的普遍发展水平，更为经济发展提供了不可或缺的动力。因此股票市场的预测、分析对于经济发展也具有重要意义。因为其有众多影响因素，股票市场是一种复杂系统，并且有很强的非线性[1] 。现有的有效市场理论并不能对一些异常现象进行很好地解释，但随着行为金融理论的提出，人们渐渐意识到投资者的心理和行为，都会对股票市场的波动产生重大的影响。而股票价格预测的关键因素在于大量数据获取的准确率和速度，现有的传统方法，例如市场调研和问卷调查等，并不能满足高速增长的信息总量所带来的需求。

随着互联网的发展，人们越来越多地依赖于网络作为日常生活的场所。普通网民可以在互联网上自由发表信息，不受知识水平、社会背景、甚至是资金容量的限制。在我国社会主义市场经济体制下的金融领域，网络作为新型传播媒介也逐渐进入人们的生活。互联网也成为了投资者分享和传播投资信息的主要渠道。大众投资者的心理和行为信息，都能够从他们在互联网上的言论中获得分析出来。尽管传统认为股票评论大部分为噪声交易信号，但是统计分析研究表明这之中也有大量股票市场的相关信息。因此，分析提取相关情绪和信息，为投资者提供更及时有效的信息，有助于优化改进投资策略，有重要的现实意义。这也是本文的主要研究内容。

对于我国政府和经济部门，股票预测方法可以概览股票市场整体状况，有助于分析股票市场健康度，利于优化调整，促进其高速稳定发展。对于投资者，可以引导投资方向，为决策优化提供重要依据。试图降低投资风险，提高收益率，不仅可以增强投资者投资的科学性，还可以保证市场经济的稳定发展。

进行股票预测研究，还能够为学术界提供理论经验，具有重要意义，也为之后的股票研究提供新的方法和思考。

1.2 研究背景

经典证券理论认为市场参与者是完全理性的，这样市场价格就在真实价值周围波动，保持稳定[2][3]。但是，投资者情绪在股票市场中发挥重要的作用，例如“噪声交易者”会因为价值以外的因素交易而产生深远影响。这些行为受到经济学家广泛关注，例如 Keynes[4] 在解释股市异常的时候将其比喻为“兽性”，而更在更早时候 Hume[5] 称之为“激情推动”。相应地，股票市场中的高度投机行为造成了 1929 年的股票市场爆炸，或者互联网泡沫的发生，而这些现象都很难由简单的基础理论解释。

和经典的完全理性模型相不同，许多新的行为模型试图合理解释发生的这些剧烈波动。De Long 等人[6] 形式化描述了非完全理性交易者，即噪声交易者在市场中扮演的角色。在他们的模型中，有两类投资者：理性套利者和噪声交易者。理性套利者对资产的未来回报持有理性分析，而噪声交易者受外在情感驱动，产生相对于理性分析要么过于乐观要么过于悲观的期待。最终平衡的市场价格反应理性套利者和噪声交易者双方的期待。本质上，De Long 等人[6] 证明了市场价格因为不理性的乐观情绪或者悲观情绪会偏离价值，证明了噪声交易者在交易中起到的作用。

在 De Long 等人[6] 的工作之后，又有大量的经验文献试图研究投资者情绪及其产生的效果之间的关系，例如 Lee 等人[7]、Kumar 等人[8]、Baker 等人[9] 和 Lemmon 等人[10]。传统地，一般人们更多研究情绪直接或者间接的影响。Lee 等人[7] 采用基于市场的间接的方式，他们使用了封闭式基金折价率作为投资者情绪的指标，并且指出封闭式基金折价率是个人积极或者消极情绪相对于更广阔市场情绪的体现。封闭式基金折价率的下降和由噪声交易者持有的股票的回报正相关。

而在我国，对于股票的预测方法大概有几个方面。

第一个方面是基于投资分析的股票预测。一般有基本分析法和技术分析法两类[11]。基本分析法主要是由基本原理出发，依赖众多基础学科知识进行推导。分析影响股票市场变化的因素，并且关注它们与股票市场的相互作用。而技术分析法更多关注股票市场自身变化规律，运用数学、图形和逻辑等方法，从已有数据中归纳总结，分析预测股票市场变化。基本分析法对总体局势把握比较好，但是对于短期波动则稍显不足。技术分析法则在可靠性上得不到保证[12]。

另一个方面则是由时间序列分析而引申出来的股票预测研究。根据技术性的不同主要有建立统计学模型、选择特征指标、和通过图形直观预测分析几种方法[13][14]。时间序列模型里主要包括非平稳模型、回归移动平均模型，和随机游走模型几种。但是，这些线性模型都很难处理股票预测研究中的非线性问题，导致适用性受到局限。

最近还有基于人工智能的股票预测研究。Kimoto 等人[15] 利用神经网络模型的方法，开发出了针对东京股票交易市场的预测系统，并且成功进行预测。人工智能方法相比传统方法有众多优势，比如非线性及自适应性，这些都是传统方法所不能匹敌的。因此人工智能方法已逐渐成为最受欢迎的预测研究方法之一。

以上这些方法面临的主要困难是无法快速有效获得股票预测信息。这些信息既有政治因素、物价水平，更多的是投资者的普遍心理。社会化的媒体的出现引发了人们的思考。近年来，许多股民以社会媒体平台为据点，交流分享股票市场信息，表达观点和意见。这些观点中不仅有现有已知的信息，还有能影响其他投资者的股民普遍心态等有价值得信息。现有研究主要从社会媒体提供的信息的强度和情感极性两个方面研究社会媒体对股票市场研究的作用[16]。

1.3 本文研究内容

本文试图解决社会媒体分析股票行为的一大难点，即快速有效挖掘社会媒体中对股票预测研究有重要价值的信息。针对过去研究过于依赖股票市场的内在规律和历史数据的不足，本文试图强调投资者的投资心理，挖掘传统模型所忽视的重要组成部分。针对过去研究过于关注行为信息而忽略了对于内容的挖掘，本文试图归纳分析社会媒体中股民的情感极性，优化预测分析效果。

本文的主要研究内容为通过互联网论坛上的股民言论及相关活动信息，试图提取出影响股票市场波动的主要因素。研究股票论坛数据和股票指标变量之间的相互影响关系。试图建立股票成交量和价格的预测模型，对股票市场走势进行预测。

第 2 章　数据框架

2.1　数据收集、获取

本文试图从国内关注股较高的股票论坛获得关注度数据，其中包含股票评论、访问量的数量和情感倾向。在相类似的研究中，大多采用了如雅虎、Twitter、或者新浪财经这样的大型网站，东方财富网和金融界等专门提供财经类信息的网站也有大量关于股票市场的关注度数据。

根据网站活跃度和垃圾信息比例等因素，考虑到大型门户网站在用户交流上有一定的局限性，而太小型的论坛又缺乏信息流通度和有效性，最终选取访问量、人均页面浏览量都很大的东方财富网股吧作为主要信息来源。股吧在财经版块设置了新闻、研报和公告等多个版块，针对热门主题和热门个股都有专门的讨论版，信息量充分全面，而且有很好的已经分好类的模块，从各方面都满足本文的需求。

本文另从国信证券金太阳网上交易客户端获取股票市场的交易数据，主要包括个股的开盘价、收盘价、最高价、最低价、成交量以及成交额等。

在本文的研究范围内，主要用到了成交量和对应的收盘价格。至于价格的考虑，在宏观尺度上，开盘价、收盘价、最高价和最低价的影响并不明显。但从信息获取的角度来说，收盘价最能体现当日最新信息，所以使用收盘价作为本文统一使用的价格。

为了排除掉小众个股非统计性的波动因素，以及可能的人为操控因素，选定上证 50 成分股作为主要研究对象。上证 50 成分股有充分的股民和资金来源，能够排除个别资本控盘的非有效市场情况出现。上证 50 成分股又能从一定程度上反映大盘情况，有重要的研究价值和意义。

东方财富网股吧作为主要的股票市场关注度信息来源，提供活跃度分析及情感分析的重要指标。获取上证 50 成分股个股论坛的数据，为了简单明确起见，抓取所有讨论帖的标题和浏览量、评论量、发表日期。注意到讨论区里并不完全都是讨论帖，有很多公告、新闻类型的帖子，这些帖子一般都是全论坛共享或是强制置顶的，对于个股差异没有共享，反而会干扰后续阶段的分析，必须注意剔除这些干扰数据。讨论帖的标题可以作为情感分析的原始材料，讨论帖的浏

代码 2.1　API 返回结果

```
{
  ”re”: [
    {
      ”post_id”: 177688688,
      ...
    },
    ...
  ],
  ”count”: 291447,
  ”rc”: 1,
  ”me”: ”操作成功”
}
```

览量和评论量能作为活跃度分析的原始材料，而讨论帖的发表日期可以作为讨论帖的时间戳。严格意义上讨论帖的发表日期不一定是所有活跃度发生的日期，但是考虑到论坛日均流量很大，老帖很快就会掉出首页，可以粗略认为讨论帖的发表日期即为活跃度发生的时间。由统计上的随机性也可以做出类似的论断，最终活跃度的期望是不受此时间差的影响的。

在实际抓取活跃度数据过程中，发现东方财富网股吧为手机端网页提供了 RESTful API 接口 http://m.guba.eastmoney.com/getdata/articlelist 。可以发送 GET 请求，参数 code 指定股票代码，参数 count 指定返回条数的个数，上限为 200 ，参数 thispage 指定当前页码。即可以固定 count 为 200 ，依次获取每页的所有评论信息，直到服务器返回空值为之，代表已经处理完所有的评论。因为抓取数据的过程中可能有事实的新数据产生，为了防止新帖的出现导致整体向后位移干扰抓取结果，可以按照从前往后的顺序抓取，这样至多抓到重复数据，而不会漏抓。

服务器正常返回的数据结构如代码 2.1 所示。其中以 “...” 表示省略掉的部分。

考虑到获取数据的部分主要都是网络 I/O ，所以是访存密集型而不是计算密集型的任务。决定采用 Node.js 框架编写程序。Node.js 原生带有事件循环机制，对于异步任务的支持非常地好，再加上外部 promise 库，可以发挥显著的作用。在实际中测试，利用异步 HTTP 请求，能在单核的情况下跑满所有网络带宽，达到性能上的极限。

获取服务器请求的程序逻辑如代码 2.2 所示，该部分均使用 CoffeeScript 编

代码 2.2　请求服务器数据

```
parseSinglePage = (symbol, page, redis) ->
  url = makeUrl symbol, page
  requestAsync url
  .then (data) ->
    data = JSON.parse data
    entries = _.map data.re, (post) ->
      stripSingleEntry symbol, post
    dumpEntries entries, redis
    logger.debug 'processed',
      length: entries.length
      symbol: symbol
      page: page
    entries.length
```

写。其中 `requestAsync` 函数处理网络请求的逻辑，负责有关错误处理和重试的部分，在此不再赘述。而 `stripSingleEntry` 函数则对取回来的单个条目进行格式化处理，提取感兴趣的部分，并且整理成方便后续处理的数据结构。最后 `dumpEntries` 函数将处理过后的数据结构放入 Redis 数据库中。

将获取的结果放入 Redis 数据库中的过程，同样也是异步地完成。考虑到获取原始数据的过程中可能发生各种不可预料的网络临时故障或解析错误，此时应该尽早抛出错误，使得程序异常退出，使用外部稳定的数据库可以保证处理进程崩溃后不会丢失数据。

股票市场价格和交易量数据通过国信证券金太阳网上交易客户端获取，该客户端可以导出历史数据成 CSV 格式文件。

至此所有原始数据已经获取完毕。

2.2　数据预处理

为了提取出讨论帖的情感因素，需要用到自然语义分析和情感分析的工具。对于英文的文本分析已经比较成熟了，但对于中文的语义分析还停留在比较表面的阶段。最简单的办法是根据积极情感和消极情感制作两个词表，然后根据词表的匹配程度来决定积极情感和消极情感的明显程度。但是比较开源的已有的中文词表之后发现，大部分词表被没有针对股票市场的特定应用环境，所以很多在股票论坛里带有强烈情感色彩的词语都没有被收录，比如“大涨”、“大跌”都是很常见的极性很强的情感词语。

为了解决词表不够精确的问题，必须修改已有的情感分析模型以适应新的任务。可以利用 GitHub 上的开源项目 twitter-sent-dnn (https://github.com/xiaohan2012/twitter-sent-dnn) 。这是一个利用卷积神经网络，利用 Twitter 数据作为训练集训练得到的分析英语短文情感的模型。使用深度卷积神经网络的好处在于可以挖掘句子内词语之间的复杂的逻辑关系，得到深层的语义上的信息。具体来说，这个模型使用了 Twitter140 的数据作为训练数据。此数据源包含约 160 万已标注的 Twitter 数据供训练使用。在训练数据中，标注的方式是自动进行的，根据 Twitter 中的表情符号作为标注的基准，得到积极情感和消极情感的极性参数。最后去除这些表情符号，作为真正的训练输入。此数据集另有 872 条验证数据和 1821 条测试数据。这两种数据是人工标注的。除此之外，此模型还使用了实时的 Twitter 数据流作为训练数据。

为了把此模型应用在之前抓取的东方财富网股吧数据上，还需解决语言的映射问题。如果直接将中文通过机器翻译变成英文，一般情况下是会保留词语的意思和情感极性，但是会打乱句子的结构。不过在此情感模型的情况下，这并不会造成很大的影响。考虑到此情感模型是基于 Twitter 训练的。Twitter 数据大多是短小破碎的语句，精简但蕴含强烈的语言情感。所以直接通过机器翻译将中文翻译成英文，再通过此情感模型，即可得到较好的结果。

在这里，翻译使用了百度翻译 API 。至本论文成文之时，百度宣布即将发布新的基于深度神经网络的翻译模型，但是并未公开其 API ，所以本文所使用的百度翻译 API 均为现有模型。

基于代码 2.3 ，实现百度翻译 API 的自动调用。比较需要注意的地方在于细节。首先原帖中可能有符号字符或者其他不可见字符，这些字符会干扰翻译引擎工作，必须通过正则表达式将其剔除。翻译引擎对于长度也有严格的限制，考虑一般不会有过长的讨论帖，此处强行限制在 2000 个字符，截断超出的部分。`currentTranslator` 是由轮叫调度得到一个翻译接口。在实际使用中，为了并行使用多个翻译组建加快翻译进度，必须处理好多个翻译组建之间的调度问题。调用过快的话有被封禁的危险。所以本文单独实现一个基于滴漏桶的轮叫调度器，保证有限调度速度更快的翻译组建，同时保证调度的公平性，且以硬上限作为界限，进行指数后退防止接口被禁用。

具体翻译组建的实现较为琐碎，不再赘述。需要注意的是，就算是百度翻译 API 本身也有多种接口，接口之间的数据格式和请求格式稍有不同，出错和

代码 2.3　百度翻译 API 调用

```
translateOne = (key) ->
  Q.ninvoke sourceRedis, 'get', key
  .then (entry) ->
    entry = JSON.parse entry
    Q.ninvoke targetRedis, 'get', entry.id
    .then (targetEntry) ->
      bar.tick 1
      if JSON.parse(targetEntry)?.translation
        process.stdout.write '\rskip'
        return
      else
        src = (entry.title + '\n' + entry.content)
        .replace /<br>/g, ' '
        .replace /[^0-9a-zA-Z\u4e00-\u9fa5]/g, ' '
        .substring 0, 2000
        .trim()
        currentTranslator = round.get()
        process.stdout.write '\r'
        process.stdout.write JSON.stringify(round.status())
        currentTranslator src
        .then (res) ->
          d =
            translation: res
          targetRedis.set "#{entry.id}", JSON.stringify(d)
          logger.debug 'translated',
            src: src
            dst: res
```

重试的逻辑也相应有不同。

基于上述操作，简单地搭建出了实验需要用到的语言情感模型。简单地测试一下，可以发现其性能完全满足情感极性分析的要求。

最后通过一些黏合脚本，将数据导入预处理阶段，并且将处理后的数据序列化成 JSON 格式以供后续阶段分析。部分逻辑如代码 2.4 所示。此处首先剔除了非有效日期。例如周末和其他公休假作为非交易日，并没有交易数据，应该删除这些数据点，防止后续拟合出现问题。此处整合了讨论帖点击量、积极情感数、消极情感数、收盘价格和交易量这几大信息来源，取自 Redis 数据库和交易客户端导出数据等多个渠道。

并且在代码 2.4 中，使用 promise 库，调用 `Q.all` 和 `_.map` 使得任务可以并行执行，是本文中广泛采用的并行方法。需要注意的是，这里指的并行是在单核处理器的情况下，利用异步事件循环机制实现的任务调度，可以利用网络、

代码 2.4　数据组合

```
Q.all _.map(compositeIndex, (symbol) ->
  Q.nfcall fs.readFile, path.join(location.stockFeedDir,\
  "SH#{symbol}.txt"), encoding: 'utf-8'
  .then parser.stockFeed.f
  .then (stockFeedData) ->
    symbolBulletin = bulletinData[symbol]
    pairedData = []
    _.forEach dateRange, (date) ->
      if stockFeedData[date]?.volume?
        clickCount = symbolBulletin[date]?.clickCount\
        ? pairedData[pairedData.length - 1]?.clickCount ? 0
        positiveCount = symbolBulletin[date]?.positiveCount\
        ? pairedData[pairedData.length - 1]?.positiveCount ? 0
        negativeCount = symbolBulletin[date]?.negativeCount\
        ? pairedData[pairedData.length - 1]?.negativeCount ? 0
        volume = stockFeedData[date].volume
        close = stockFeedData[date]?.close\
        ? pairedData[pairedData.length - 1]?.close ? 0
        pairedData.push
          date: date
          volume: volume
          close: close
          clickCount: clickCount
          positiveCount: positiveCount
          negativeCount: negativeCount
    Q.nfcall fs.writeFile, path.join(\
    location.outputDir, "#{symbol}.json"),\
    JSON.stringify(pairedData), encoding: 'ascii'
```

硬盘读取等异步事件加快执行速度。正如上面所示，在单核处理器情况下能利用全部网络带宽，达到性能极限。

本文在代码风格上不仅广泛利用异步事件循环机制，还采取了函数式编程的先进思想，程序逻辑清晰，模块化强。从调试到执行都有很强的复用性。

第 3 章　成交量与讨论帖点击量的关系

3.1　格兰杰因果关系

格兰杰因果关系检验[17] 是一种假设检验的统计方法，检验一组时间序列 x 是否为另一组时间序列 y 的原因。回归分析通常只能得出不同变量间的同期相关性，自回归模型只能得出同一变量前两前后期的相关性，但格兰杰证明了在自回归模型中通过一系列的检验而揭示不同变量之间的时间落差相关性是可行的。需要注意的是，这里所说的原因，抑或时间落差相关性，并无法证明逻辑意义上的因果关系，只能从时间现在的关系上解释过去发生的事件对以后发生的事件有预测作用。

格兰杰因果关系检验的核心假设在于，未来的事件不会对目前和过去的事件产生影响，而过去的事件才可能对现在及未来产生影响。已知时间序列 y 的过去值的情况下，如果另一时间序列 x 的过去值仍能对时间序列 y 有显著的回归相关关系，就意味着时间序列 x 对时间序列 y 有格兰杰因果关系。

它的严格数学定义如下。令 x 和 y 为广义平稳序列。如果要检验零假设 x 非 y 的格兰杰原因，首先引入 y 的落后期建立 y 的自回归模型如下。

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_m y_{t-m} + residual_t. \tag{3-1}$$

简单表示即为

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + residual_t. \tag{3-2}$$

式 3-1 和式 3-2 中的 m 即为落后期。在落后期固定的情况下，使得 $residual$ 极大似然正态分布的系数序列 a_i ，同时为使得预测值

$$\hat{y}_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} \tag{3-3}$$

与标准值 y_t 的 L_2 范数最小的系数序列。取此系数得到自回归模型同式 3-3 。

接着引入 x 的落后期建立增广回归模型如下。

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \sum_{i=1}^{m} b_i x_{t-i} + residual_t. \tag{3-4}$$

同样的方法可以得到系数序列 b_i 。

如果没有任何 x 的落后期被留在模型中，即可以定性地认为从某种意义上 $b_i = 0$ 对于所有的 i 都成立，无格兰杰因果关系的零假设成立。

3.2 落后期计算方法

计算时间序列 x 的落后期，可以使用多种方法。在本文中，主要使用以下四种方法。

1. Akaike information criterion
2. Bayesian information criterion
3. Final prediction error
4. Hannan-Quinn information criterion

其中 Bayesian information criterion 也被称作 Schwarz criterion 。这些选取落后期的模型本身之间差别并不明显，但是在某些情况下会给出不同的选取结果。

Bayesian information criterion 的定义如式 3-5 。

$$\mathrm{BIC} = -2 \cdot \ln \hat{L} + k \cdot \ln n. \tag{3-5}$$

在式 3-5 中，$\hat{L}$ 为模型 M 似然性方程的最大值，如果令 $\hat{\theta}$ 为取得极大似然性的参数，则

$$\hat{L} = p(x \mid \hat{\theta}, M). \tag{3-6}$$

在式 3-5 中，n 为数据点的个数，k 为自由变量的个数。

直观上可以理解，$-2 \cdot \ln \hat{L}$ 项使得模型尽可能拟合，这就要求落后期越大越好，而 $k \cdot \ln n$ 项使得模型不是很大，防止过拟合的情况出现。其他几种选取落后期的方法本质上大同小异，不再详细介绍。

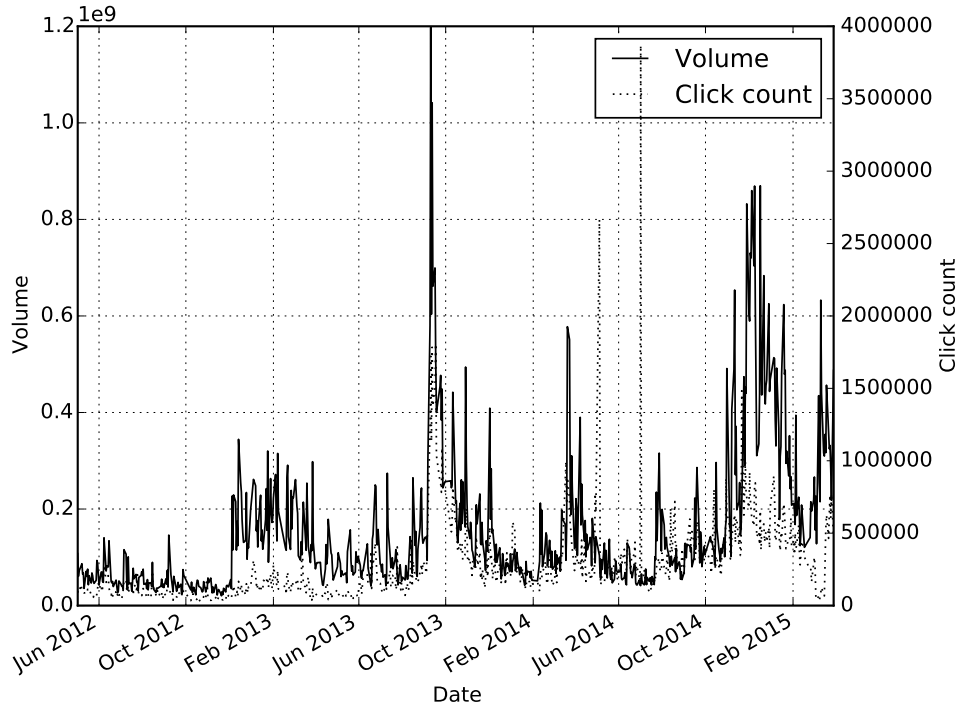

图 3.1　浦发银行 (600000) 成交量与讨论帖点击量关系

3.3　落后期计算

不妨尝试建立基于成交量和讨论贴点击量的向量自回归模型，本文使用了 Python 的数据挖掘框架 Statsmodels[18] 。该框架对于成熟的线性和非线性数据模型都有准确高效的实现，并且自带简单的画图功能。它提供有线性回归模型、时间序列分析等功能，其中包括自回归滑动平均模型等基本模型。

本文在此只上提供了基本的数据框架，提供统一的接口获得处理后的数据。需要注意的是，之前在数据获取的时候为了达到最大的灵活度和速度，采取了 Node.js 框架和 Redis 数据库。在预处理的时候，这部分框架也能非常完美地整合。但在数据计算和模型方面，为了使用已有的开源框架，需要使用 Python 编程。这在本文的数据框架下并不是问题，数据结构是语言无关的，可以方便地在多种框架和语言之间切换，并且接口尽量统一，方便使用。

取浦发银行 (600000) 在 2012 年 5 月 1 日至 2015 年 4 月 1 日之间的过往成交量和对应日期的讨论帖点击量，如图 3.1 所示。在本文的数据框架下，代码如 3.1 所示。

建立向量自回归模型，根据参数选取最佳落后期，见代码 3.2 。落后期选择

代码 3.1　数据框架示例

```
fig, ax0 = plt.subplots()
ax1 = ax0.twinx()
lines = []
d = stats.data.get_merged('600000', 'date', 'volume', 'clickCount')
dates = [datetime.datetime.strptime(i, '%Y-%m-%d') for i in d[:, 0]]
volume = d[:, 1]
click_count = d[:, 2]
ax0.fmt_xdata = matplotlib.dates.DateFormatter('%Y-%m-%d')
fig.autofmt_xdate()
lines += ax0.plot(dates, volume, 'k-', label='Volume')
ax0.set_xlabel('Date')
ax0.set_ylabel('Volume')
lines += ax1.plot(dates, click_count, 'k:', label='Click count')
ax1.set_ylabel('Click count')
labels = [i.get_label() for i in lines]
ax0.grid()
ax0.legend(lines, labels, loc=0)
plt.tight_layout()
plt.show()
```

代码 3.2　落后期选择逻辑

```
data = pandas.DataFrame({'volume': volume,
  'clickCount': click_count})
data.index = pandas.DatetimeIndex(dates)
model = statsmodels.tsa.api.VAR(data)
```

的结果见表 3.1 。标星的数据表示最优的选择。可见根据 AIC 和 FPE 标准，最佳落后期是 4 ，而按照 BIC 标准，最佳落后期是 2 。按照 HQIC ，最佳落后期是 3 。在本文中，鉴于模型的特性，默认情况下使用 HQIC 作为落后期选择的方法。

3.4　假设检验方法

为了检验 x 对 y 的格兰杰因果关系，需要检验 y 的相量自回归模型中没有 x 的过去值的影响。需要检验向量自回归模型中 x 的系数的大小。回顾格兰杰因果关系模型如式 3-7 所示，可以做出零假设 x 对 y 并没有格兰杰因果关系，如式 3-8所示。

表 3.1　浦发银行 (600000) 成交量与讨论帖点击量落后期选择

	AIC	BIC	FPE	HQIC
0	62.45	62.46	1.324e+27	62.46
1	61.23	61.27	3.894e+26	61.24
2	61.08	61.15*	3.366e+26	61.11
3	61.07	61.16	3.322e+26	61.10*
4	61.06*	61.18	3.311e+26*	61.11
5	61.07	61.21	3.324e+26	61.12
6	61.07	61.24	3.325e+26	61.14
7	61.07	61.27	3.334e+26	61.15
8	61.08	61.30	3.358e+26	61.17
9	61.09	61.34	3.393e+26	61.19
10	61.10	61.38	3.423e+26	61.21
11	61.10	61.41	3.445e+26	61.22
12	61.11	61.44	3.478e+26	61.24
13	61.12	61.47	3.489e+26	61.26
14	61.13	61.51	3.526e+26	61.28
15	61.14	61.55	3.554e+26	61.29
16	61.14	61.58	3.574e+26	61.31
17	61.15	61.61	3.611e+26	61.33
18	61.15	61.64	3.592e+26	61.34
19	61.16	61.67	3.631e+26	61.36
20	61.16	61.70	3.649e+26	61.37

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \sum_{i=1}^{m} b_i x_{t-i} + residual_t. \tag{3-7}$$

$$H_0 : b_1 = b_2 = \cdots = b_m = 0 \tag{3-8}$$

而拒绝假设 H_0 意味着 x 对 y 有格兰杰因果关系。

在 Statsmodels 和本文的框架的实现里，一共使用了四种方法进行假设检验。

1. 参数 F 检验
2. 残差平方和 F 检验
3. 残差平方和 χ^2 检验
4. 似然比检验

以残差平方和 F 检验为例，步骤如下。

首先得到约束方程如式 3-9 的最小二乘解。

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + e_t. \tag{3-9}$$

代入最小二乘解得到自回归模型如式 3-10 。

$$\hat{y}_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i}. \tag{3-10}$$

利用自回归模型得到残差如式 3-11 。

$$\hat{e}_t = y_t - \hat{y}_t. \tag{3-11}$$

最后得到残差平方和如式 3-12 。

$$\mathrm{RSS}_0 = \sum_{i=1}^{m} \hat{e}_t^2. \tag{3-12}$$

使用同样的方法，处理非约束方程如式 3-13 所示，得到非约束方程的残差平方和如式 3-14 。

代码 3.3　浦发银行 (600000) 成交量与讨论帖点击量格兰杰因果关系检验

```
statsmodels.tsa.api.stattools.\
grangercausalitytests(d, 5, verbose=True)
```

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \sum_{i=1}^{m} b_i x_{t-i} + u_t. \tag{3-13}$$

$$\mathrm{RSS}_1 = \sum_{i=1}^{m} \hat{u}_t^2. \tag{3-14}$$

计算检验统计量如式 3-15 。

$$S_1 = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/p}{\mathrm{RSS}_1/(m-2p-1)} \sim F_{p,m-2p-1}. \tag{3-15}$$

如果 S_1 大于指定的临界值，则拒绝原假设 H_0 。

在残差平方和 χ^2 检验中，则最后一步检验统计量如式 3-16 。

$$S_1 = \frac{m(\mathrm{RSS}_0 - \mathrm{RSS}_1)}{\mathrm{RSS}_1} \sim \chi^2(p). \tag{3-16}$$

3.5　格兰杰因果关系假设检验

对浦发银行 (600000) 在 2012 年 5 月 1 日至 2015 年 4 月 1 日之间的成交量和对应日期的讨论帖点击量进行格兰杰因果关系检验。程序逻辑如代码 3.3 所示。基于之前的结果，已经知道最佳落后期为 3 ，而在进行格兰杰因果关系检验的时候，需要输入一个最大落后期，然后程序会从 1 到指定的最大落后期都计算统计检验量和对应的 p 值。所以在程序代码中指定最大落后期为 5 。得到输出如代码 3.4 所示。

分析格兰杰因果关系检验中假设检验的结果，其中 p 值随落后期的变化如表 3.2 和图 3.2 所示。理所应当地，随着落后期的增加，成交量自己携带的信息量越来越大，对于点击量信息量的依赖越来越小，p 值单调递增。但是这并不是绝对的规律，因为成交量和点击量都会携带越来越多的信息，要根据最佳落后

代码 3.4　浦发银行 (600000) 成交量与讨论帖点击量格兰杰因果关系检验结果

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=16.3569 , p=0.0001
, df_denom=701, df_num=1
ssr based chi2 test:   chi2=16.4269 , p=0.0001  , df=1
likelihood ratio test: chi2=16.2382 , p=0.0001  , df=1
parameter F test:         F=16.3569 , p=0.0001
, df_denom=701, df_num=1


Granger Causality
number of lags (no zero) 2
ssr based F test:         F=3.9950  , p=0.0188
, df_denom=698, df_num=2
ssr based chi2 test:   chi2=8.0472  , p=0.0179  , df=2
likelihood ratio test: chi2=8.0015  , p=0.0183  , df=2
parameter F test:         F=3.9950  , p=0.0188
, df_denom=698, df_num=2


Granger Causality
number of lags (no zero) 3
ssr based F test:         F=2.9458  , p=0.0322
, df_denom=695, df_num=3
ssr based chi2 test:   chi2=8.9265  , p=0.0303  , df=3
likelihood ratio test: chi2=8.8702  , p=0.0311  , df=3
parameter F test:         F=2.9458  , p=0.0322
, df_denom=695, df_num=3


Granger Causality
number of lags (no zero) 4
ssr based F test:         F=2.3264  , p=0.0550
, df_denom=692, df_num=4
ssr based chi2 test:   chi2=9.4267  , p=0.0513  , df=4
likelihood ratio test: chi2=9.3638  , p=0.0526  , df=4
parameter F test:         F=2.3264  , p=0.0550
, df_denom=692, df_num=4


Granger Causality
number of lags (no zero) 5
ssr based F test:         F=2.1061  , p=0.0629
, df_denom=689, df_num=5
ssr based chi2 test:   chi2=10.6987 , p=0.0577  , df=5
likelihood ratio test: chi2=10.6178 , p=0.0595  , df=5
parameter F test:         F=2.1061  , p=0.0629
, df_denom=689, df_num=5
```

表 3.2　浦发银行 (600000) 成交量与讨论帖点击量格兰杰因果关系检验 p 值

落后期	SSR χ^2 test	Params F test	LR χ^2 test	SSR F test
1	0.0001	0.0001	0.0001	0.0001
2	0.0188	0.0179	0.0183	0.0188
3	0.0322	0.0303	0.0311	0.0322
4	0.0550	0.0513	0.0526	0.0550
5	0.0629	0.0577	0.0595	0.0629

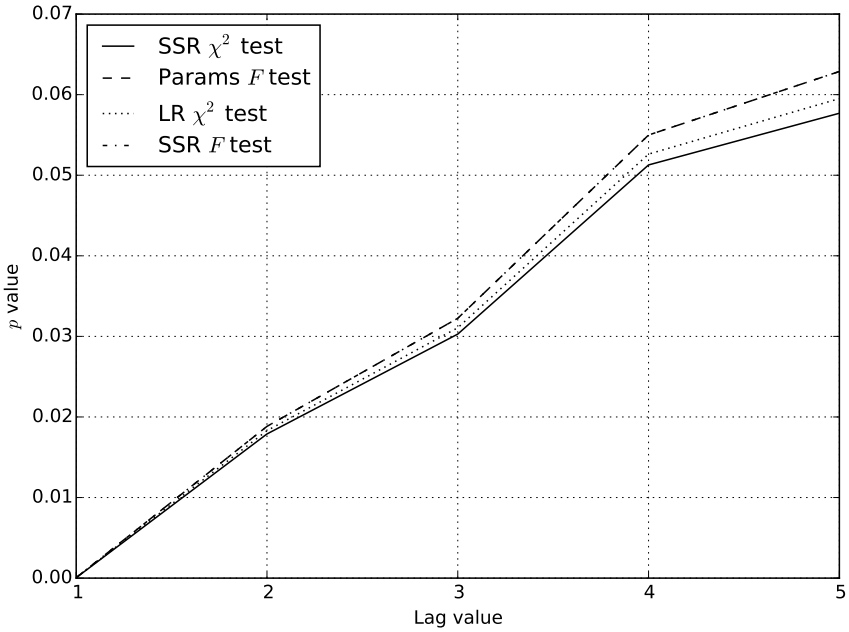

图 3.2　浦发银行 (600000) 成交量与讨论帖点击量格兰杰因果关系检验 p 值

期实际情况变化来分析 p 值的变化。如果选择显著性水平 $\alpha = 0.05$，则在落后期为之前已经选择的最佳落后期 3 的时候就已经可以拒绝零假设 H_0，认为讨论帖点击量对于成交量有格兰杰因果关系。

不妨对于上证 50 所有股票进行格兰杰因果关系检验。使用同样的方法计算 50 支股票的四种检验下的 p 值。结果如下所示。

1. 似然比检验结果如图 3.3 所示
2. 参数 F 检验结果如图 3.4 所示
3. 残差平方和 χ^2 检验结果如图 3.5 所示
4. 残差平方和 F 检验结果如图 3.6 所示

可以发现有如下几条规律。首先对于某一特定股票而言，四种检验方法的

差别并不是很大，从对于零假设 H_0 的验证的角度来说是没有区别的。从一个角度上增加了结果的可信度。其次不妨取显著性水平 $\alpha = 0.05$，在这种情况下，大部分股票都能得出拒绝零假设的结论。50 支股票中只有 9 支股票是认为讨论帖点击量对于成交量是没有格兰杰因果关系的，其他 41 支股票都是有关系的。从理论上证明了这种格兰杰因果关系的普遍性。关于这 9 支股票的特殊情况，可能有以下几种可能。

1. 该股票本身存在异常情况，不符合普遍规律，例如受到庄家操控等人为因素影响
2. 该股票讨论区存在比较极端的讨论帖，干扰统一分析格兰杰因果关系

对于第二种情况，在格兰杰因果关系中某些数据点出现的异常值能很敏感地影响假设检验的结果。例如有些之前的讨论帖被频繁顶至首页，导致总点击量很高。但是这些点击量并不如在数据框架中所假设的是短时间内产生的，而是长时间的积累，但这些点击量都被计入讨论帖发帖日期的点击量上。这样的数据点即使只有两三个，也会对检验结果造成比较深刻的影响。为解决这些问题，数据的预处理显得尤为关键如何辨别并且抹平这些极端数据点，但是又不忽视正常情况下出现的极大极小值，是非常细节并且值得深入探讨的。

需要注意的是，在格兰杰因果检验的过程中，落后期的选择十分重要。对于同一支股票，落后期的不同会导致假设检验得到的 p 值产生很大的不同。可以理解的是落后期的不同使得讨论帖点击量带来的影响的作用产生极大的分歧。所以选择对的落后期是格兰杰因果关系检验的关键。

而且之所以如此关注落后期的选择，因为一个足够精简的模型是十分必要的。选择过大的落后期会导致模型模拟和，对于后续的预测和分析是十分不利地。

3.6 异常分析

不妨单独观察 p 值最高的股票中国石化 (600028)。其落后期的选择如表 3.3 所示。显而易见的是四种落后期选择的指标产生了极大的分歧，预示着此数据带来的不对称性使得落后期的选择不稳定。

得到其格兰杰因果关系检验的 p 值随落后期变化，如图 3.7 所示。可以看出其 p 值一直都在 0.2 以上，对于这种极端情况，无论选择什么样的落后期都无法

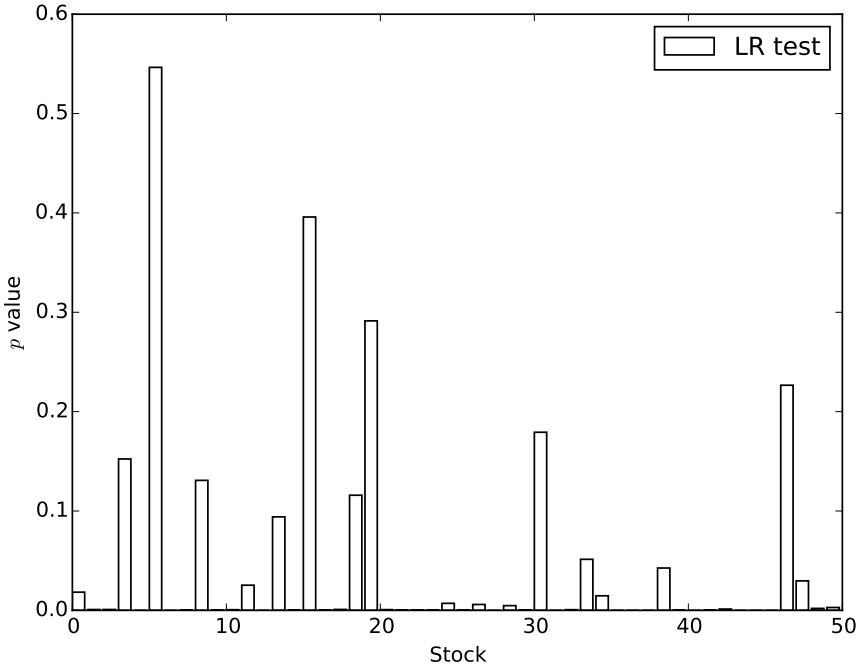

图 3.3　上证 50 成交量与讨论帖点击量似然比检验

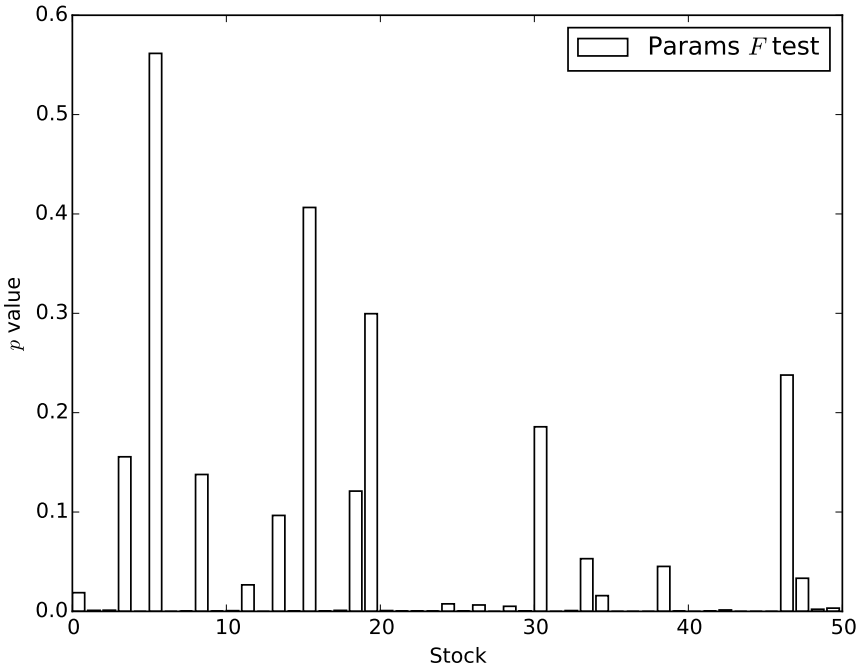

图 3.4　上证 50 成交量与讨论帖点击量参数 F 检验

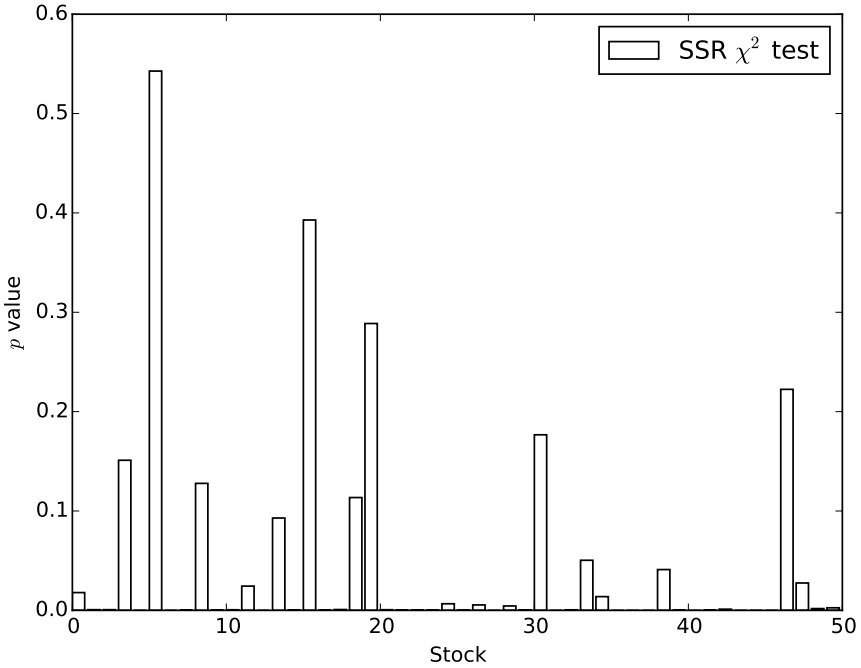

图 3.5　上证 50 成交量与讨论帖点击量残差平方和 χ^2 检验

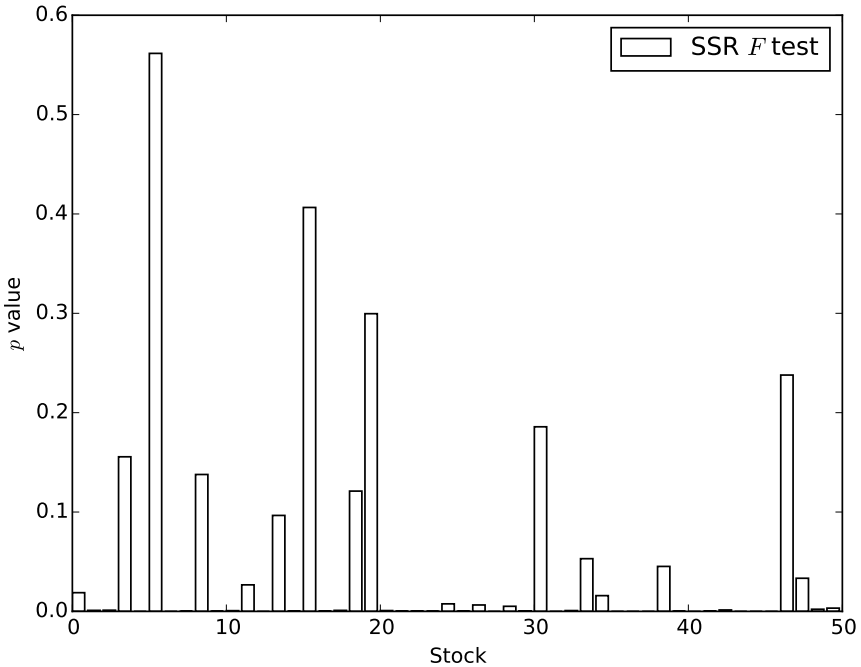

图 3.6　上证 50 成交量与讨论帖点击量残差平方和 F 检验

表 3.3　中国石化 (600028) 成交量与讨论帖点击量落后期选择

	AIC	BIC	FPE	HQIC
0	61.60	61.61	5.661e+26	61.61
1	59.68	59.72	8.286e+25	59.69
2	59.60	59.67*	7.654e+25	59.63
3	59.59	59.68	7.542e+25	59.62
4	59.59	59.71	7.566e+25	59.63
5	59.59	59.73	7.561e+25	59.64
6	59.56	59.73	7.322e+25	59.62
7	59.55	59.74	7.250e+25	59.62
8	59.48	59.70	6.774e+25	59.56*
9	59.48	59.73	6.764e+25	59.57
10	59.48	59.76	6.801e+25	59.59
11	59.48	59.79	6.822e+25	59.60
12	59.49	59.82	6.829e+25	59.61
13	59.49	59.85	6.890e+25	59.63
14	59.49	59.87	6.838e+25	59.64
15	59.49	59.90	6.839e+25	59.65
16	59.49	59.92	6.839e+25	59.66
17	59.49	59.95	6.849e+25	59.67
18	59.44	59.92	6.495e+25	59.62
19	59.43*	59.94	6.440e+25*	59.63
20	59.43	59.97	6.457e+25	59.64

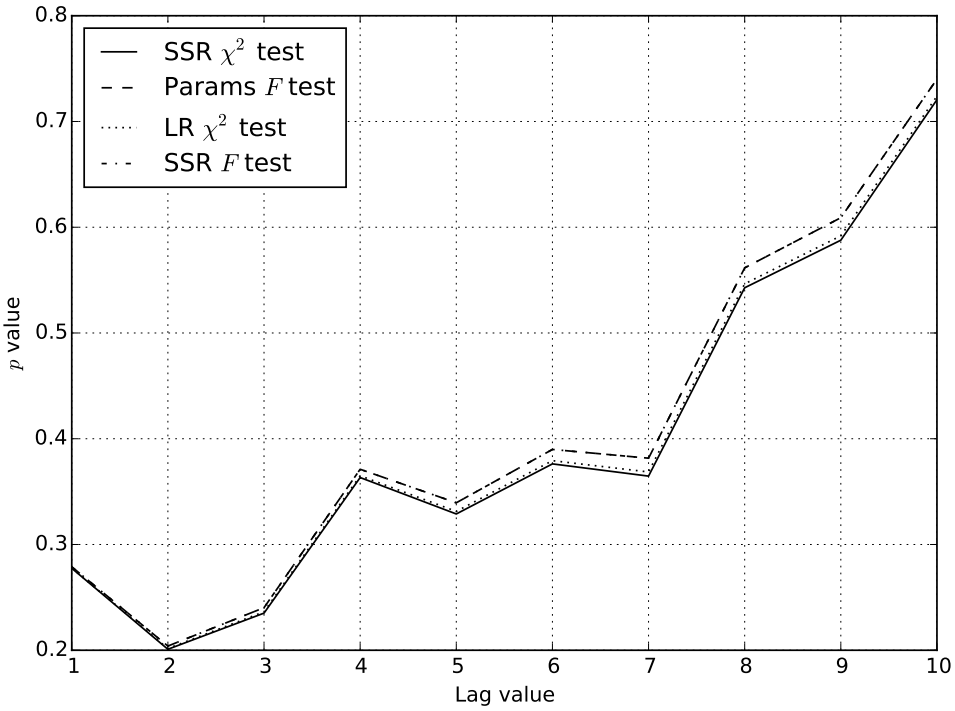

图 3.7　中国石化 (600028) 成交量与讨论帖点击量格兰杰因果关系检验 p 值

验证其格兰杰因果关系。

继续深究原因，得到其成交量与讨论帖点击量折线图如图 3.8 所示。比较明显的是在 2013 年 8 月左右的时间点，有一次点击量的突然增加，但是并没有对应成交量的大幅度变化。根据之前的分析，有可能是确实讨论区产生了如此异常的反应，也有可能是讨论帖归类的原因导致了过去的老贴被过多累加点击量。至此，可以发现该股票产生异常高的 p 值的直接原因。其他几支 p 值异常的股票也有各自不同的原因，深层次可能有更加统一地解释。但是对于绝大多数股票来说，都是符合格兰杰因果关系检验的结果的，不妨更多关注于普遍情况，利用其格兰杰因果关系。

3.7　成交量趋势预测

试图利用向量自回归模型做最简单的预测模型。首先给出向量自回归模型的严格构造。它描述了在同一样本期间内的 n 个内生变量可以作为它们过去值的线性函数。可以写成如式 3-17 所示。

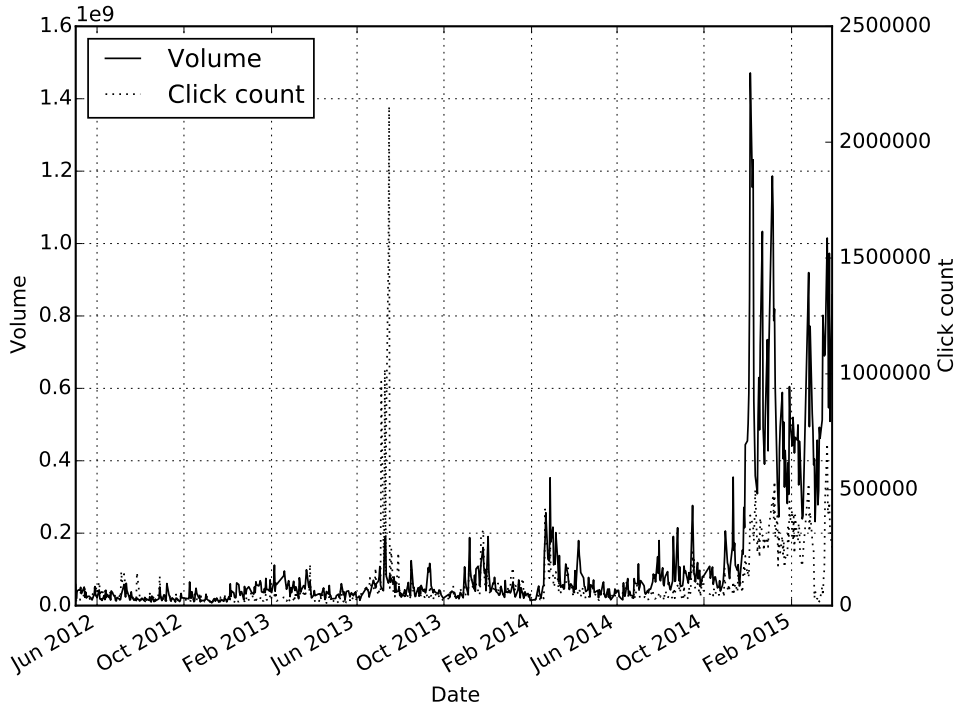

图 3.8　中国石化 (600028) 成交量与讨论帖点击量关系

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t. \tag{3-17}$$

其中 c 是 $n \times 1$ 的常数向量，A_i 是 $n \times n$ 矩阵，e_t 是 $n \times 1$ 的误差向量。误差项满足以下几条性质。

1. $E(e_t) = 0$
2. $E(e_t e_t') = \Omega$ （正定矩阵）
3. $E(e_t e_{t-k}') = 0$

在自回归模型的参数 A_i 都已知，并且没有外生变量的情况下，在最小误差平方和的意义下最好的线性预测函数是基于向量自回归模型本身的，进行提前 1 步的预测，即在已有时间 t 信息的基础上预测 y_{t+1} ，如式 3-18 所示。

$$y_{t+1|t} = c + A_1 y_t + A_2 y_{t-1} + \cdots + A_p y_{t-p+1}. \tag{3-18}$$

要预测更长的提前量，可以使用预测链式法则，如式 3-19 所示。

$$y_{t+h|t} = c + A_1 y_{t+h-1|t} + A_2 y_{t+h-2|t} + \cdots + A_p y_{t-p+h|t}. \tag{3-19}$$

其中当 $j \le 0$ 的时候 $y_{t+j|t} = y_{t+j}$。

为了获得自回归模型的系数，首先需要把数据整理成 pandas 支持的数据帧。然后选择落后期，并且根据落后期计算优化得到向量自回归模型。最后使用该模型来进行预测。

在预测的过程中，主要有两方面需要特别注意。

1. 预测步长的选择
2. 时间序列的预处理

预测步长的选择意味着每经过多少次预测就要重新计算模型。从理论上来说似乎是每预测一次都新建一个模型比较好，包含的信息更多，但在实际应用场景下，并不是如此。另外时间序列的预处理也是十分重要的。不难看出股票成交量的变化并不是在任何时间都保持同样的期望的。这些因素都有可能影响最后模型的产生。

需要注意的是，在对已有数据进行回测的时候，切忌使用所有数据算出模型，再用同样的数据计算。这样不符合实际情况下的因果关系，也不利于评价模型的泛化能力。一定要使用交叉验证的办法，才能评价模型在实际场景中的可用性。

不妨使用招商银行 (600036) 作为测试数据，试图构建预测模型。其成交量与讨论帖点击量的历史关系如图 3.9 所示。先通过向量自回归模型得到系数，逻辑如代码 3.5 所示。

落后期的选择使用了 HQIC 方法。得到结果如表 3.4 和表 3.5 所示。表 3.4 展示的是关于成交量的回归模型的系数。对称地表 3.5 展示的是关于讨论帖点击量的回归模型的系数。

由于成交量的预测更加关键，暂时只关注成交量的回归系数。以 y_t 代表 t 时刻的成交量数据，以 x_t 代表 t 时刻的讨论帖点击量。可以得到由招商银行 (600036) 在 2012 年 5 月 1 日至 2015 年 4 月 1 日之间的数据产生的预测模型。该模型如式 3-20 所示。可以看出 t 时刻成交量的预测对于 t 时刻的讨论帖点击量是没有依赖的，这是正确的行为，只能依赖过去几天的讨论帖点击量。

观察式 3-20 系数绝对值的大小，可以看出过往时间讨论帖点击量对应的系数都较大，这是因为成交量一般都在 10^8 量级，而讨论帖点击量一般在 10^5 量

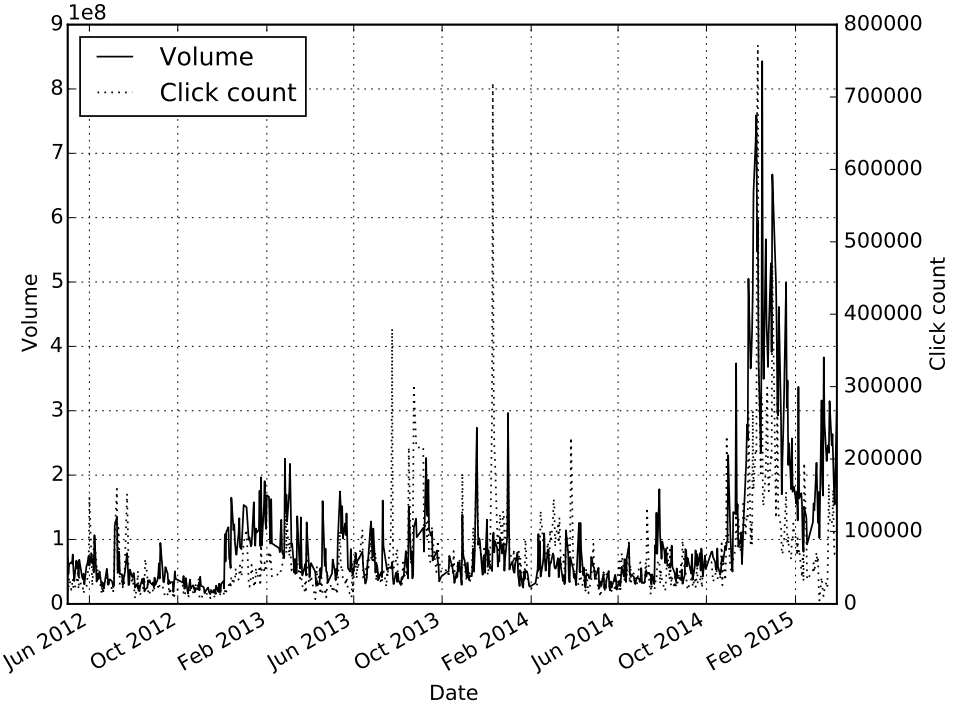

图 3.9　招商银行 (600036) 成交量与讨论帖点击量关系

代码 3.5　向量自回归模型系数计算

```
data = pandas.DataFrame({
    'volume': volume,
    'clickCount': click_count
})
data.index = pandas.DatetimeIndex(date)
model = statsmodels.tsa.api.VAR(data)
results = model.fit(ic='hqic')
```

表 3.4　招商银行 (600036) 成交量回归系数

	coefficient	std. error	t-stat	prob
const	6673808.985347	2839014.831623	2.351	0.019
L1.clickCount	44.314506	40.706032	1.089	0.277
L1.volume	0.459554	0.039591	11.607	0.000
L2.clickCount	-164.745012	46.790235	-3.521	0.000
L2.volume	0.220279	0.043528	5.061	0.000
L3.clickCount	15.514133	47.333759	0.328	0.743
L3.volume	0.141555	0.044510	3.180	0.002
L4.clickCount	31.436235	47.375577	0.664	0.507
L4.volume	0.039574	0.044540	0.889	0.375
L5.clickCount	191.049542	47.130236	4.054	0.000
L5.volume	-0.033348	0.044100	-0.756	0.450
L6.clickCount	-163.072352	41.064196	-3.971	0.000
L6.volume	0.137050	0.040193	3.410	0.001

表 3.5　招商银行 (600036) 讨论帖点击量回归系数

	coefficient	std. error	t-stat	prob
const	10449.057853	2781.641575	3.756	0.000
L1.clickCount	0.577187	0.039883	14.472	0.000
L1.volume	-0.000036	0.000039	-0.925	0.356
L2.clickCount	-0.127870	0.045845	-2.789	0.005
L2.volume	0.000205	0.000043	4.807	0.000
L3.clickCount	0.078802	0.046377	1.699	0.090
L3.volume	0.000038	0.000044	0.868	0.385
L4.clickCount	0.054840	0.046418	1.181	0.238
L4.volume	0.000002	0.000044	0.040	0.968
L5.clickCount	0.067477	0.046178	1.461	0.144
L5.volume	-0.000032	0.000043	-0.729	0.466
L6.clickCount	-0.071944	0.040234	-1.788	0.074
L6.volume	-0.000012	0.000039	-0.314	0.754

级，所以讨论帖点击量对应项的系数会更大。但是这带来一个问题，在不同时期成交量和讨论帖点击量的绝对值变化是非常大的，有宏观上的涨跌趋势。当在不同时期成交量和讨论帖点击量的量级比例发生变化时，回归系数并不能应对这个变化带来的影响。因此后续需要平衡不同时期的整体趋势，以使得不同时期的数据可以放在一条时间序列里回归分析。

$$\begin{aligned} y_t = {} & 6673808.875347 \\ & + 0.459554 y_{t-1} \\ & + 0.220279 y_{t-2} \\ & + 0.141555 y_{t-3} \\ & + 0.039574 y_{t-4} \\ & - 0.033348 y_{t-5} \\ & + 0.137050 y_{t-6} \\ & + 44.314506 x_{t-1} \\ & - 164.745012 x_{t-2} \\ & + 15.514133 x_{t-3} \\ & + 31.436235 x_{t-4} \\ & + 191.049542 x_{t-5} \\ & - 163.072352 x_{t-6}. \end{aligned} \tag{3-20}$$

正如之前提到，式 3-20 是不能够用于回测的，因为它包含了整条时间序列的信息。为了检验向量自回归模型是否适用，需要在仅提供 t 时刻之前的信息的情况下，得到向量自回归模型，然后预测下一个时间点的成交量。由于并没有设置常数参数，可以认为该模型并没有针对特定股票优化，也不会出现过拟合，甚至无法交叉验证的情况。

如图 3.10 所示，是对招商银行 (600036) 历史成交量的回测。其中虚线为预测值，实线为历史真实值。可以看出大部分波动都被一定程度上预测。2015 年 1 月左右的剧烈波动则稍有延迟。

可以计算预测的均方根误差。均方根误差的定义如式 3-21 所示。其中 $\hat{y}_i$ 为预测值，而 y_i 为实际值。

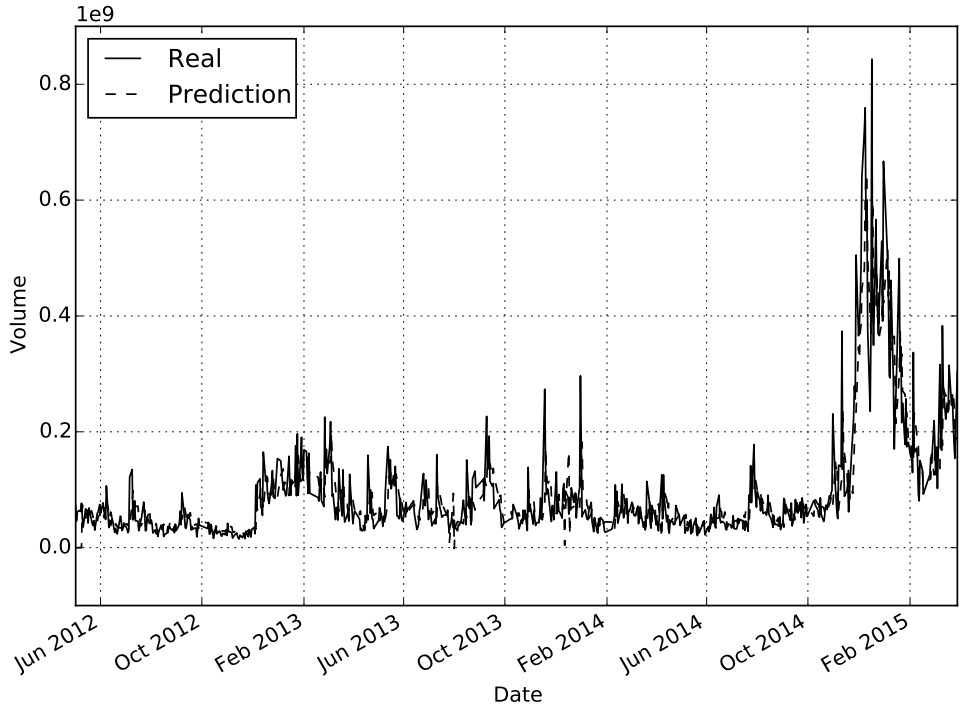

图 3.10　招商银行 (600036) 成交量预测

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}. \tag{3-21}$$

为了平衡不同模型之间的数值差异，可以使用标准均方根误差，如式 3-22 所示。

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{y_{max} - y_{min}}. \tag{3-22}$$

计算图 3.10 对应的预测的标准均方根误差，得到结果为 5.857×10^{-2} 。即从某种意义的平均误差的角度来说，只有 5% 左右。

之前进行预测计算的时候步长选取为 1 ，即每次建立模型只进行之后 1 天的预测。为了减少计算复杂性，可以试图增长预测计算的步长。如果取步长为 n ，则 t 时刻的预测值 $\hat{y}_t$ 由式 3-23 中时刻对应项组合得到。

$$\left(\left\lfloor\frac{t}{n}\right\rfloor - 1\right)n, \cdots, \left\lfloor\frac{t}{n}\right\rfloor n - 1. \tag{3-23}$$

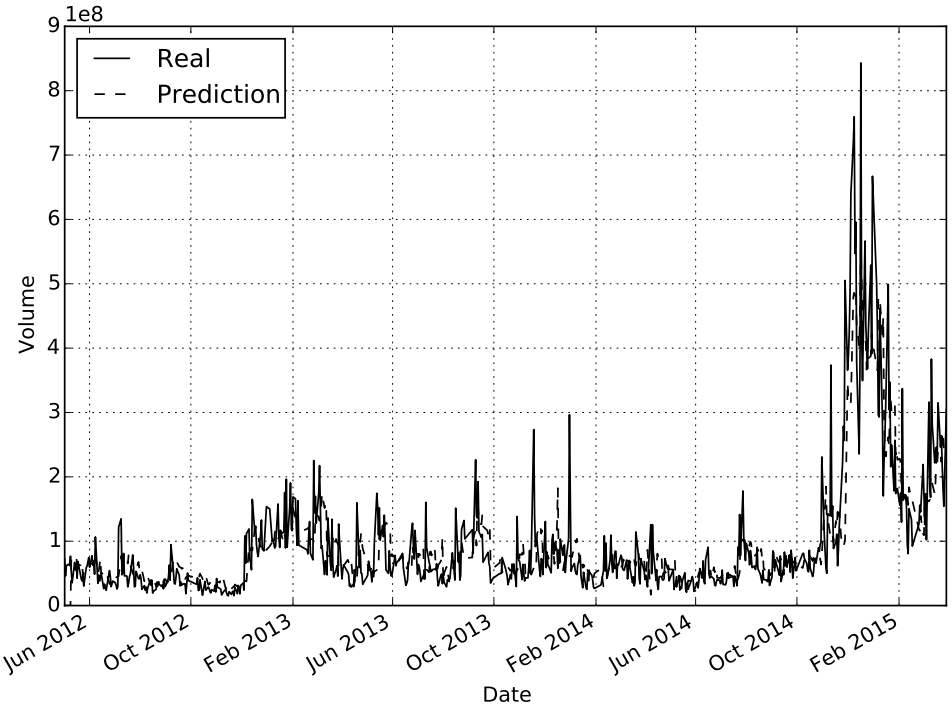

图 3.11　招商银行 (600036) 步长 5 时成交量预测

不妨取步长 $n = 5$，得到结果如图 3.11 所示。可以看出虽然步长是之前的 5 倍，但是精度并没有很严重的损失。计算得到其标准均方根误差为 6.634×10^{-2}，是步长为 1 对应标准均方根误差的 1.133 倍。

3.8　滑动窗口量比

之前面对的一个主要问题是成交量在时间上的期望并不是不变的。例如在招商银行 (600036) 的成交量上，如图 3.9 所示，2015 年 1 月以后的成交量是之前的数倍。而且不同时间之间成交量波动剧烈。这样进行线性回归分析的时候，必定导致各时间段的权重不一致，更加偏向于成交量更高的部分。然而对于股市来说，成交量的高低携带同等重要的信息，为了消除这种不平等，可以引入量比的概念，通过滑动窗口并且在滑动窗口内进行量比的计算，可以把所有数据归一化到统一的尺度上。

滑动窗口量比的具体操作步骤如下。

1. 选取窗口大小为 w
2. 选取需要计算的时刻 t

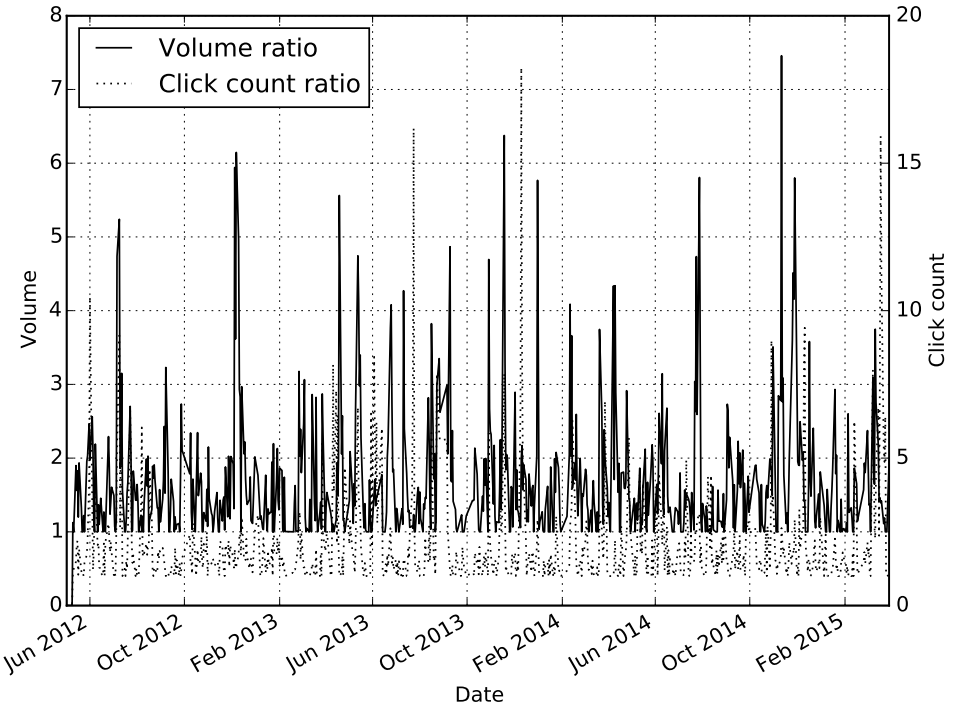

图 3.12　招商银行 (600036) 滑动窗口大小 5 时量比

3. 取 $[t-w+1, t]$ 时间区间内的最小值，即

$$m = \min_{\forall i \in [t-w+1,t]} (y_i)$$

4. 取 y_t 与 m 的比值，即

$$\hat{y}_t = \frac{y_t}{m}$$

不妨以窗口大小 $w = 7$ 为例，对招商银行 (600036) 做滑动窗口量比之后得到如图 3.12 所示的结果。

由于是和最小值的比值，所有的量比之后的值满足 $y_i \geq 1$，但是少数点任然出现比值等于 0 的情况，这是由于最小值中出现 0，除法得到 NaN 的原因。这和某些数据点不存在有关，并不会对结果造成影响。

如果对量比计算之后的结果进行预测分析，使用之前的方法，将数据源替换为量比计算之后的数据，可以得到结果如图 3.13 所示。

计算得到标准均方根误差为 1.045×10^{-1}。由于最大最小值横跨的区间明显比之前的结果要小，所以直接比较标准均方根误差是没有意义的。但是不难看出，此时的模型对于拐点的把握和反应都相比之前好了很多。很明显这是由于

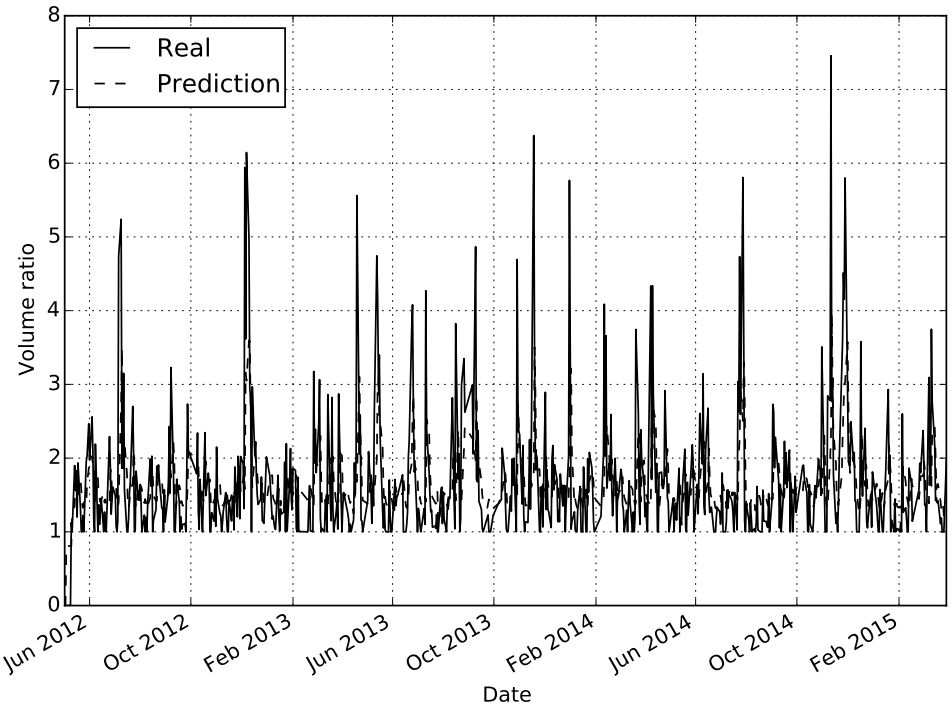

图 3.13　招商银行 (600036) 滑动窗口大小 5 时成交量预测结果

预处理之后的数据在不同时期的绝对值范围区间相似。之所以这样的原因，可以认为股市变动主要由两大方面因素相互作用。

1. 宏观经济形势
2. 股民心理

宏观经济形势主要体现在大盘走势上，如果经济形势比较好，大盘走势就会上涨，相应股票可能短期内有波动，但是总体趋势上往上涨的。但是股民心理波动相对频繁，会在短期的市场变化上产生明显的影响，这种心理波动虽然受到大盘走势的影响，但更多是一种独立的因素。表现在短期内股价的频繁波动。从一定程度上可以用均线来分析模拟宏观经济形势。所以使用量比计算相对削弱了宏观经济形势对成交量的影响，反而关注股民心理部分。

尝试不同的滑动窗口大小，在此不再展示单独的结果。对于上证 50 里的所有股票，在滑动窗口为 $w \in [2, 10]$ 的范围内分别得到标准均方根误差，以此来显示不同滑动窗口大小之间的关系，也显示不同股票之间对于滑动窗口量比有不同反应的差异。

例如对于招商银行 (600036)，结果如表 3.6 所示。滑动窗口大小对于结果标

表 3.6　招商银行 (600036) 不同滑动窗口大小对比

窗口大小	标准均方根误差
2	0.09899992956586048
3	0.10308134849238397
4	0.09082552122214050
5	0.09635710957749012
6	0.10324352591264475
7	0.10453272329607123
8	0.09928063376309698
9	0.10280568105864621
10	0.10639861314404000

准均方根误差的影响并不是单调的。在实际使用的时候可以根据个股不同选取最佳的窗口大小，以获得最好的效果。

对于上证 50 所有股票，其最佳窗口大小的选取和对应的标准误差均方根如表 3.7 所示。

表 3.7　招商银行 (600036) 不同滑动窗口大小对比

股票代码	最佳窗口大小	标准误差均方根
600000	4	0.06507378279547231
600010	3	0.046787640382425631
600015	8	0.08102542723555374
600016	10	0.083725100995798496
600018	9	0.021331092998438765
600028	10	0.065520033284345899
600030	3	0.04676993370312172
600036	4	0.090825521222140507
600048	3	0.092944567536801775
600050	2	0.075557010306984654
600089	9	0.056564360423647861
600104	4	0.071729737949525488
600109	9	0.049883031237732363
600111	3	0.076252016540148862
600150	6	0.066390846190124822
600196	2	0.079346865488974724

续下页

续表

股票代码	最佳窗口大小	标准误差均方根
600256	4	0.078877440949510047
600332	2	0.043799048623684056
600372	8	0.050747852952962171
600406	4	0.086307070150588028
600518	7	0.037577657365749979
600519	8	0.064246325320964734
600585	5	0.077798367779057992
600637	3	0.04154556056467603
600690	4	0.068584442960259284
600703	2	0.079706103509941728
600832	7	0.039982071870845022
600837	3	0.051166167363317992
600887	7	0.033803112756368082
600999	2	0.068583930621682124
601006	3	0.06280839910132438
601088	2	0.062492194672069429
601118	10	0.056133427558822405
601166	5	0.087799296273357794
601169	10	0.058330683650961855
601288	2	0.058701752262776942
601299	7	0.02265766121005593
601318	3	0.090806705123322756
601328	8	0.048585173177144163
601398	3	0.06260252509397557
601601	4	0.063536597449525986
601628	3	0.057380354111528677
601668	2	0.086962000811413601
601688	7	0.037033208067903868
601766	9	0.027915740616423042
601818	4	0.049852371490162604
601857	10	0.064288946183429771
601901	9	0.068175951565399767
601989	5	0.0048737705383712707
601998	4	0.04780899852798054

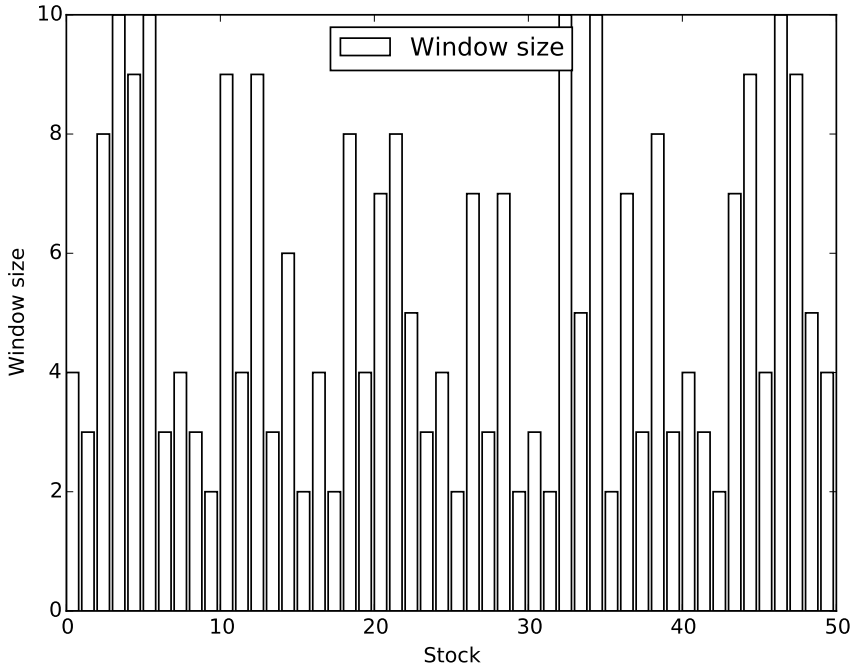

图 3.14　上证 50 最佳窗口大小

如果将最佳窗口大小以柱状图的形式表示出来，可以得到图 3.14 。与图 3.3 、图 3.4 、图 3.5 和图 3.6 对比可以看出，最佳窗口大小比较大的几只股票，意味着他们在预测的意义上结果并不稳定，对应的也正好是在格兰杰因果关系检验时因果关系不明显的几只。建立两种计算之间的联系，可以作为指标发现跟大盘整体形势和股民心理耦合度比较低的几只股票，再单独分析。

在所有的计算中，均使用的线性模型。使用线性模型的好处有结构简单，关系明显，易于分析验证。因为参数个数比较少，可以通过枚举的方法得到不同系数之间的联系，并且尝试获得更好的结果。但是在时间序列上，或许非线性模型可能可以带来更好的性能。首先注意到不同时间对于当前的影响不应该是线性的关系，更多是类似指数衰弱的关系。而成交量和讨论帖点击量也不仅仅是线性组合的关系。最后还有其他的信息来源，比如讨论帖发帖量，或者相关的信息可以引入进来。如果考虑非线性模型，则模型的选择十分重要。因为选择空间太大，不太可能一一枚举。

第 4 章　价格与讨论帖情绪的关系

4.1　与讨论帖点击量格兰杰因果关系检验

在已经得到讨论帖点击量与成交量的格兰杰因果关系并且结果较为理想的情况下，考虑讨论帖点击量与价格的格兰杰因果关系。再以浦发银行 (600000) 为例，其股票价格和讨论帖点击量的关系如图 4.1 所示。可以看出，与讨论帖点击量与成交量的关系图 3.1 对比，股票价格与讨论帖点击量并没有很强的直观上的关系。考虑到向量自回归模型中由众多项的线性组合得到，虽然并不能以直观上的关系作为最终评判的标准，但是至少可以发现结果不会比之前更好。

利用之前进行格兰杰因果关系检验的思路和推导出来的方法，先选择最佳落后期，得到结果如表 4.1 所示。使用 HQIC 方法选取的落后期 3 ，进行格兰杰因果关系，得到结果如代码 4.1 所示。分析可知，在最佳落后期 3 的情况下，p 值均已大于 0.7 ，并不能显著地表明有格兰杰因果关系。而且在其他落后期的选

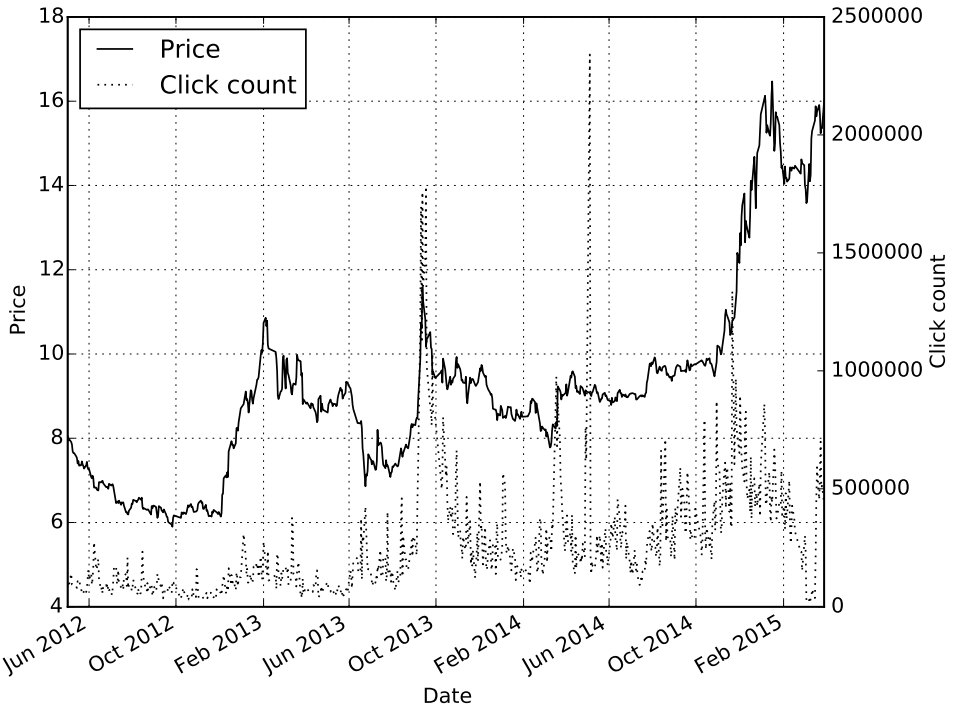

图 4.1　浦发银行 (600000) 价格与讨论帖点击量关系

表 4.1　浦发银行 (600000) 价格与讨论帖点击量落后期选择

	AIC	BIC	FPE	HQIC
0	26.30	26.32	2.647e+11	26.31
1	20.95	20.99	1.257e+09	20.97
2	20.88	20.95*	1.171e+09	20.91
3	20.86*	20.96	1.152e+09*	20.90*
4	20.87	20.98	1.153e+09	20.91
5	20.87	21.01	1.155e+09	20.92
6	20.87	21.04	1.159e+09	20.94
7	20.87	21.07	1.162e+09	20.95
8	20.88	21.10	1.164e+09	20.96
9	20.89	21.14	1.177e+09	20.98
10	20.88	21.16	1.172e+09	20.99
11	20.88	21.19	1.174e+09	21.00
12	20.89	21.22	1.186e+09	21.02
13	20.90	21.26	1.193e+09	21.04
14	20.91	21.29	1.204e+09	21.06
15	20.92	21.33	1.214e+09	21.08
16	20.93	21.36	1.227e+09	21.10
17	20.93	21.39	1.230e+09	21.11
18	20.94	21.43	1.238e+09	21.13
19	20.95	21.46	1.249e+09	21.15
20	20.95	21.49	1.257e+09	21.16

择上，p 值都相对较高，可以断言在此情况下，讨论帖点击量并没有给收盘价格带来更多的附加信息，无法推导出它们之间的格兰杰因果关系。

从直观上进行分析，可以看出在 2015 年 1 月以后，收盘价格呈现大幅度上涨趋势，这时却没有看到讨论帖点击量成比例的反应。对比 2013 年 9 月附近的大涨，当时的讨论帖点击量就有很明显的成比例的大幅增长。然后在 2014 年 5 月附近讨论帖点击量出现一个高峰，而收盘价格并没有相应的动作。所以不难从直观上也发现它们之间的因果关系较弱。

接着考虑所有上证 50 股票，对其进行格兰杰因果关系检验。得到如下结果。

1. 似然比检验结果如图 4.2 所示
2. 参数 F 检验结果如图 4.3 所示

代码 4.1　浦发银行 (600000) 价格与讨论帖点击量检验结果

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=1.1545  , p=0.2830
, df_denom=701, df_num=1
ssr based chi2 test:   chi2=1.1594  , p=0.2816  , df=1
likelihood ratio test: chi2=1.1585  , p=0.2818  , df=1
parameter F test:         F=1.1545  , p=0.2830
, df_denom=701, df_num=1


Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.5832  , p=0.5584
, df_denom=698, df_num=2
ssr based chi2 test:   chi2=1.1748  , p=0.5558  , df=2
likelihood ratio test: chi2=1.1738  , p=0.5561  , df=2
parameter F test:         F=0.5832  , p=0.5584
, df_denom=698, df_num=2


Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.4344  , p=0.7284
, df_denom=695, df_num=3
ssr based chi2 test:   chi2=1.3165  , p=0.7252  , df=3
likelihood ratio test: chi2=1.3152  , p=0.7255  , df=3
parameter F test:         F=0.4344  , p=0.7284
, df_denom=695, df_num=3


Granger Causality
number of lags (no zero) 4
ssr based F test:         F=1.0045  , p=0.4044
, df_denom=692, df_num=4
ssr based chi2 test:   chi2=4.0703  , p=0.3966  , df=4
likelihood ratio test: chi2=4.0585  , p=0.3981  , df=4
parameter F test:         F=1.0045  , p=0.4044
, df_denom=692, df_num=4


Granger Causality
number of lags (no zero) 5
ssr based F test:         F=0.8507  , p=0.5140
, df_denom=689, df_num=5
ssr based chi2 test:   chi2=4.3215  , p=0.5041  , df=5
likelihood ratio test: chi2=4.3082  , p=0.5059  , df=5
parameter F test:         F=0.8507  , p=0.5140
, df_denom=689, df_num=5
```

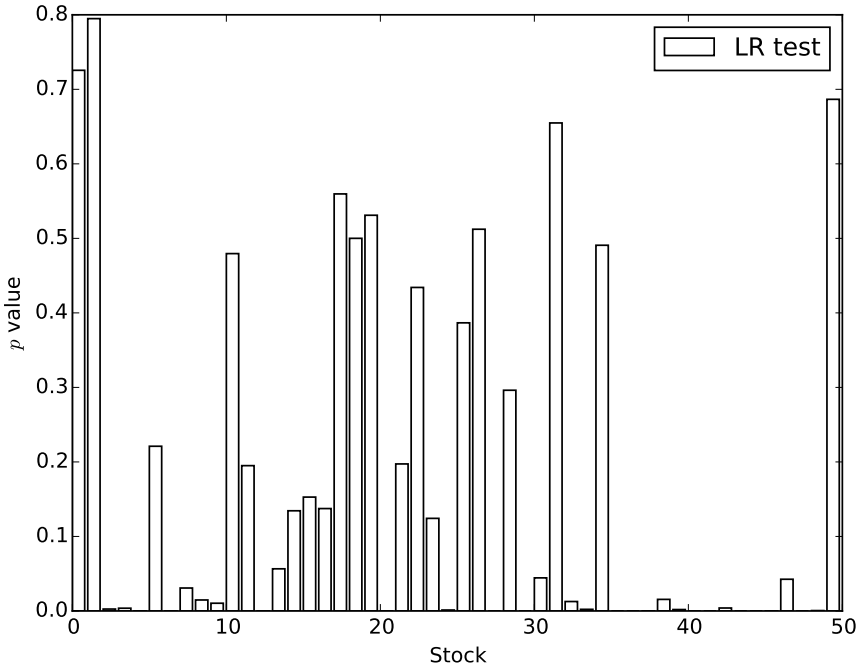

图 4.2　上证 50 价格与讨论帖点击量似然比检验

3. 残差平方和 χ^2 检验结果如图 4.4 所示
4. 残差平方和 F 检验结果如图 4.5 所示

不难看出整体上格兰杰因果关系没有成交量与讨论帖点击量之间的强烈。在显著性水平 $\alpha = 0.5$ 的情况下有 20 支股票没有显示出很强的格兰杰因果关系。间接地来看，成交量与讨论帖点击量有着相对过强地格兰杰因果关系，但是成交量与价格之间并没有很明显的关系，也可以认为讨论帖点击量对收盘价格直接提供的信息不够充足。

4.2　情感分析

为了更好地得到讨论帖点击量和价格之间的关系，需要引入情绪分析。目前情绪分析可以在多个维度上进行，提取出不同的情感成分并且给出相应的评分。例如比较经典的做法是分成冷静、警觉、肯定、重要、善良和高兴这几个类别。分别分析讨论帖内容中的这几个感情成分的评分，再利用线性或者非线性模型得到与价格的关系。

为了简单起见，本文中只对情感做一个维度的评分，即评判是积极还是消

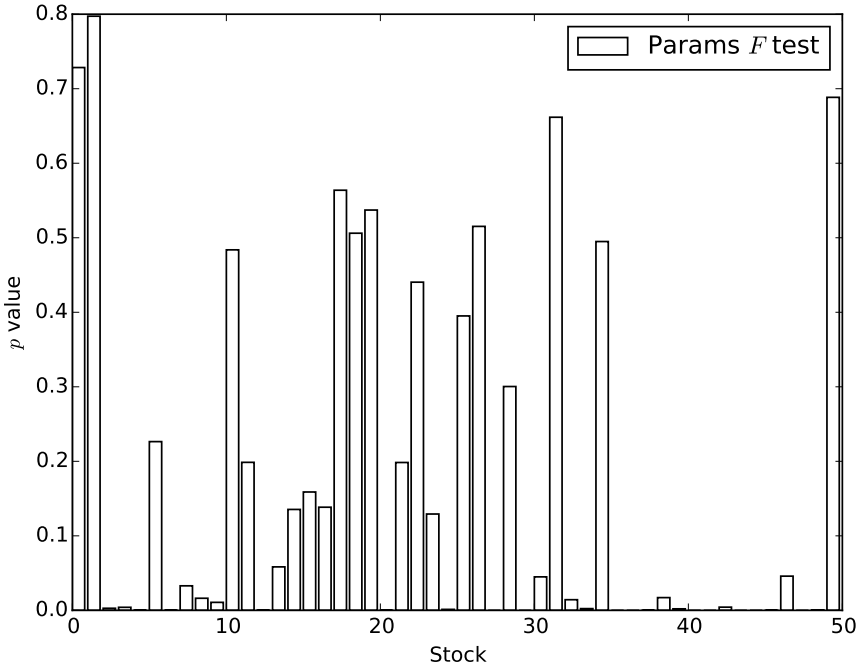

图 4.3　上证 50 价格与讨论帖点击量参数 F 检验

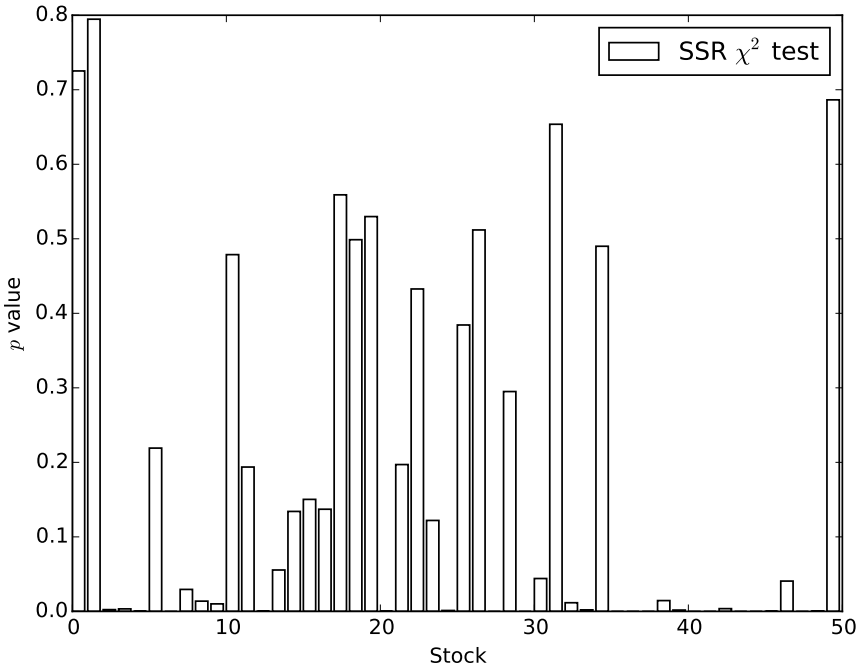

图 4.4　上证 50 价格与讨论帖点击量残差平方和 χ^2 检验

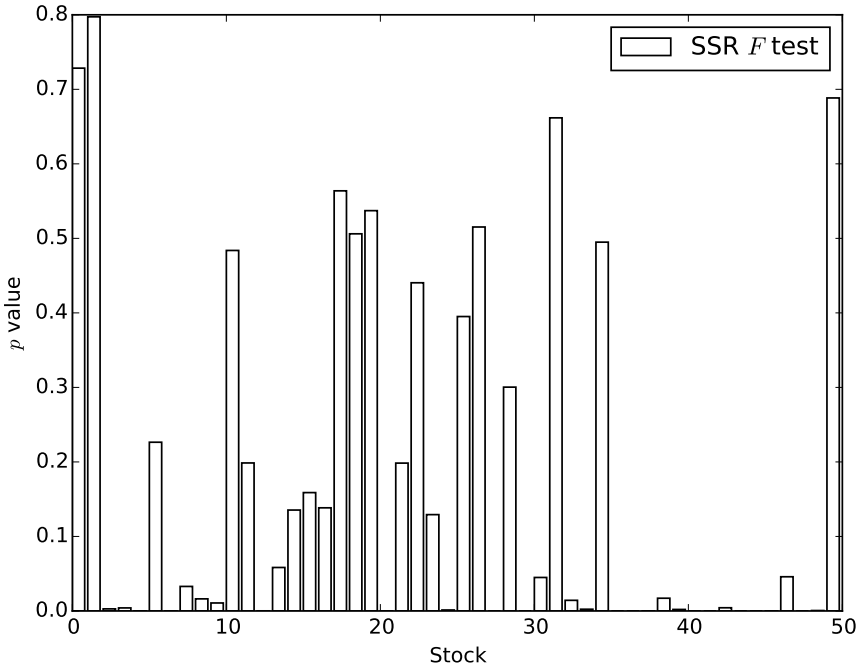

图 4.5　上证 50 价格与讨论帖点击量残差平方和 F 检验

代码 4.2　情感分析逻辑

```
def analysis_worker(i):
    source_redis = redis.StrictRedis(db=1, password='***')
    target_redis = redis.StrictRedis(db=2, password='***')
    key = target_redis.rpop('tasks')
    while key != None:
        text = source_redis.get(key)
        sent = sentiment_score(json.loads(text)['translation'])
        target_redis.set(key, sent)
        logger.info('{} remaining'.format(target_redis.llen('tasks')))
        key = target_redis.rpop('tasks')
```

极情感。所有具体分类的情感都可以归纳为积极和消极两大类，所以如此近似并不会导致某种情感信息的完全丢失，只是使得不同细分情感无法区分。逻辑如代码 4.2 所示。该部分代码可以使用进程级并行，由 Redis 数据库处理并行访问的一致性和性能问题，单个进程负责调用之前训练好的神经网络模型，即 `sentiment_score` 函数，获取情感极性。

至于翻译的部分已经在数据框架中提到，在此不再赘述。从数据框架的角度，还有从论坛数据中提取整合情感分析文本的部分，属于预处理中比较琐碎

的部分，也不再赘述。

在分析情感的时候，对于标题、正文和回复可以有不同的处理方法。本文将标题和正文连接在一起进行情感分析。考虑到回复中平均蕴含的信息量比较小，暂时不把回复包括到情感分析里面。但是为了更加全面地考虑这个问题，可以尝试对回复进行情感分析。

首先把每日的讨论帖归类，逐条进行情感分析，得到在区间 $[0, 1]$ 之内的情感极性评分。对情感极性评分进行二值化，以 0.5 作为分界线，高于 0.5 认为是积极情感，低于 0.5 则认为是消极情感，最后对积极情感和消极情感分别计数累加，作为当天的积极情感和消极情感成分。

分析每天的积极情感和消极情感的比例，可以近似得到当天有多少的讨论帖点击量是积极或者是消极的。按这种方法讨论帖点击量分成两组不同的数据。

再以浦发银行 (600036) 为例，得到其价格与讨论帖点击量中的积极成分的关系如图 4.6 所示。对比图 4.1 所示的价格与讨论帖点击量所有成分总和的关系，可以发现提取积极情感成分后，2013 年 9 月和 2015 年 1 月附近的快速上涨对应的讨论帖点击量结果更加明显，而相应的 2014 年 5 月附近之前虚高的讨论帖点击量，此处趋势被相对削弱。相比之前的分析，能更加清晰地体现出股民情感所造成的讨论帖点击量和股票收盘价格之间的关联性。

4.3 格兰杰因果关系检验

基于以上情感分析的内容，再进行格兰杰因果关系检验。不同的是，将研究提取过后的讨论帖点击量中的积极情感成分与收盘价格之间的关系。检验方法相同。以浦发银行 (600000) 为例，首先选择最佳落后期，结果如表 4.2 所示。虽然最佳落后期还是 3 ，但是可以发现得到的值比之前的略小，说明模型更加精简，拟合程度更高。

使用最佳落后期进行格兰杰因果关系检验，得到结果如代码 4.3 所示。在最佳落后期 3 的情况下，p 值由之前的平均 0.7269 降到了平均 0.5970 ，为之前的 82% 。虽然仍然不能表明有很明显的格兰杰因果关系，但在其他非最佳落后期的情况下，p 值有更加明显的优化下降。

可以看出，将情感分析因素引入之后，可以获得和之前完全不同的信息量。

这是符合直觉的结果。既然已知单纯的讨论帖点击量与成交量有很明显的

表 4.2　浦发银行 (600000) 价格与积极讨论帖点击量落后期选择

	AIC	BIC	FPE	HQIC
0	23.97	23.99	2.582e+10	23.98
1	18.57	18.61	1.157e+08	18.58
2	18.52	18.58*	1.102e+08	18.54
3	18.50	18.59	1.083e+08	18.54*
4	18.50	18.62	1.082e+08	18.55
5	18.50*	18.64	1.080e+08*	18.55
6	18.50	18.67	1.084e+08	18.57
7	18.50	18.70	1.086e+08	18.58
8	18.51	18.73	1.088e+08	18.59
9	18.52	18.77	1.100e+08	18.61
10	18.51	18.79	1.095e+08	18.62
11	18.51	18.82	1.098e+08	18.63
12	18.52	18.85	1.108e+08	18.65
13	18.53	18.89	1.114e+08	18.67
14	18.54	18.92	1.127e+08	18.69
15	18.55	18.96	1.134e+08	18.71
16	18.56	18.99	1.147e+08	18.73
17	18.56	19.02	1.149e+08	18.74
18	18.57	19.06	1.159e+08	18.76
19	18.58	19.09	1.169e+08	18.78
20	18.59	19.13	1.179e+08	18.79

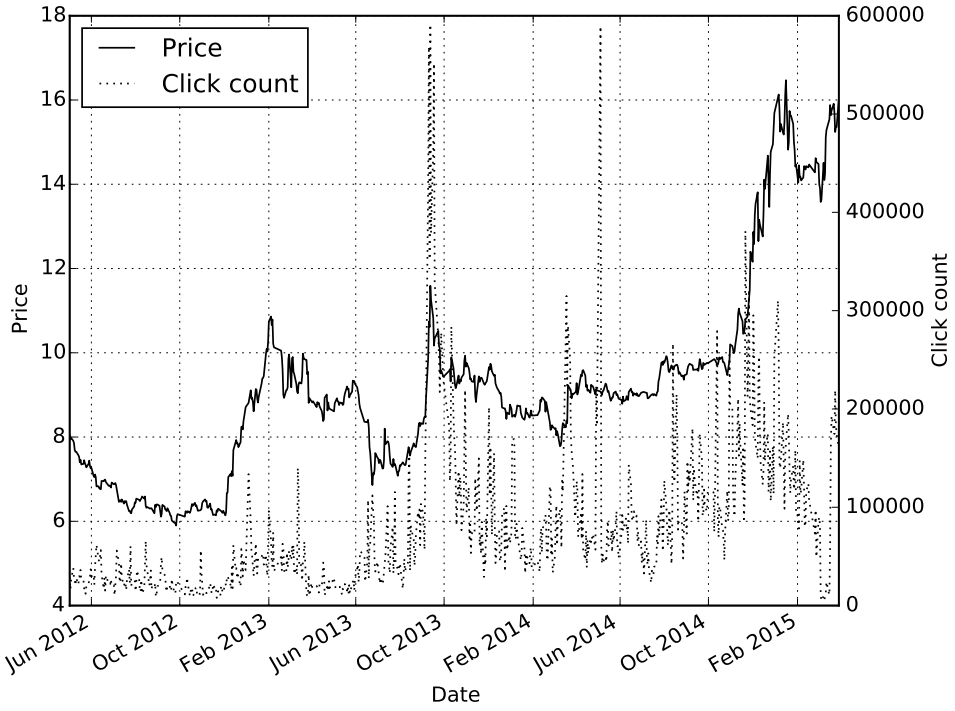

图 4.6　浦发银行 (600000) 价格与积极讨论帖点击量关系

相关关系，则可知要获得收盘价格的相关信息，必须依赖讨论帖点击量中更加细分之后的成分。而要获得此成分，进行一维地情感分析无疑是最简单的方法。无论是单独地使用讨论帖情感极性比例而不考虑讨论帖点击量总量，还是引入更多维度的情感信息，都只会使模型变得更加片面，而不能反映出最本质地股民情感因素。

并且不难看出，股民情感因素对于价格的波动无疑起着举足轻重的作用。进一步验证了最初的猜想。

再次对上证 50 所有股票进行格兰杰因果关系检验。得到结果如下。

1. 似然比检验结果如图 4.7 所示
2. 参数 F 检验结果如图 4.8 所示
3. 残差平方和 χ^2 检验结果如图 4.9 所示
4. 残差平方和 F 检验结果如图 4.10 所示

在之前 p 值比较大的地方，可以看出加入情感分析之后都有显著地效果。对于之前格兰杰因果关系就已经很敏感的部分，则效果不是那么明显。

要想获得更加全方面的提升，需要更加细粒度的情感分析。比较好的做法

代码 4.3　浦发银行 (600000) 价格与积极讨论帖点击量检验结果

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=1.2235  , p=0.2690
, df_denom=701, df_num=1
ssr based chi2 test:   chi2=1.2288  , p=0.2676  , df=1
likelihood ratio test: chi2=1.2277  , p=0.2679  , df=1
parameter F test:         F=1.2235  , p=0.2690
, df_denom=701, df_num=1


Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.6407  , p=0.5272
, df_denom=698, df_num=2
ssr based chi2 test:   chi2=1.2906  , p=0.5245  , df=2
likelihood ratio test: chi2=1.2894  , p=0.5248  , df=2
parameter F test:         F=0.6407  , p=0.5272
, df_denom=698, df_num=2


Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.6251  , p=0.5990
, df_denom=695, df_num=3
ssr based chi2 test:   chi2=1.8941  , p=0.5947  , df=3
likelihood ratio test: chi2=1.8915  , p=0.5952  , df=3
parameter F test:         F=0.6251  , p=0.5990
, df_denom=695, df_num=3


Granger Causality
number of lags (no zero) 4
ssr based F test:         F=1.7066  , p=0.1467
, df_denom=692, df_num=4
ssr based chi2 test:   chi2=6.9150  , p=0.1404  , df=4
likelihood ratio test: chi2=6.8812  , p=0.1423  , df=4
parameter F test:         F=1.7066  , p=0.1467
, df_denom=692, df_num=4


Granger Causality
number of lags (no zero) 5
ssr based F test:         F=1.3891  , p=0.2262
, df_denom=689, df_num=5
ssr based chi2 test:   chi2=7.0564  , p=0.2165  , df=5
likelihood ratio test: chi2=7.0211  , p=0.2191  , df=5
parameter F test:         F=1.3891  , p=0.2262
, df_denom=689, df_num=5
```

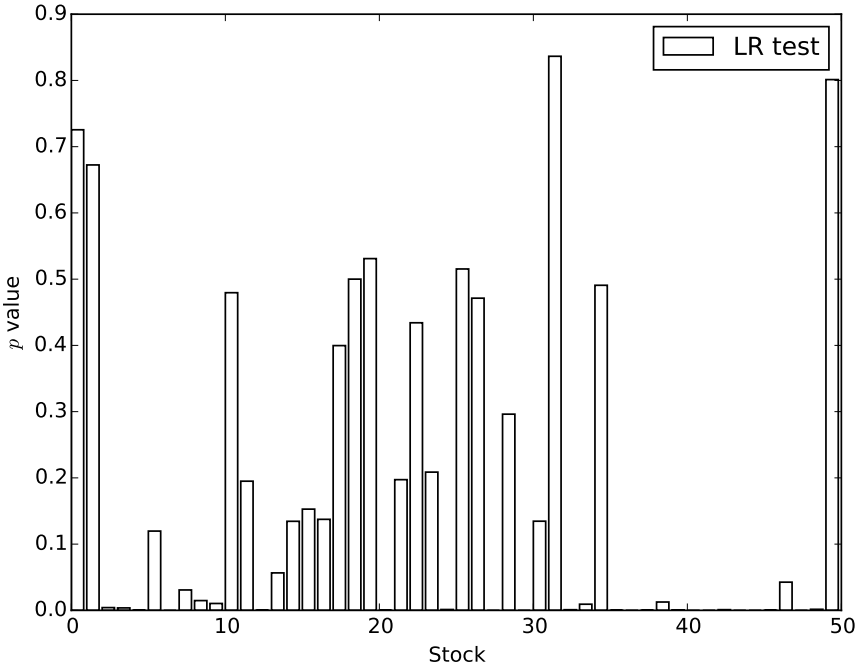

图 4.7　上证 50 价格与积极讨论帖点击量似然比检验

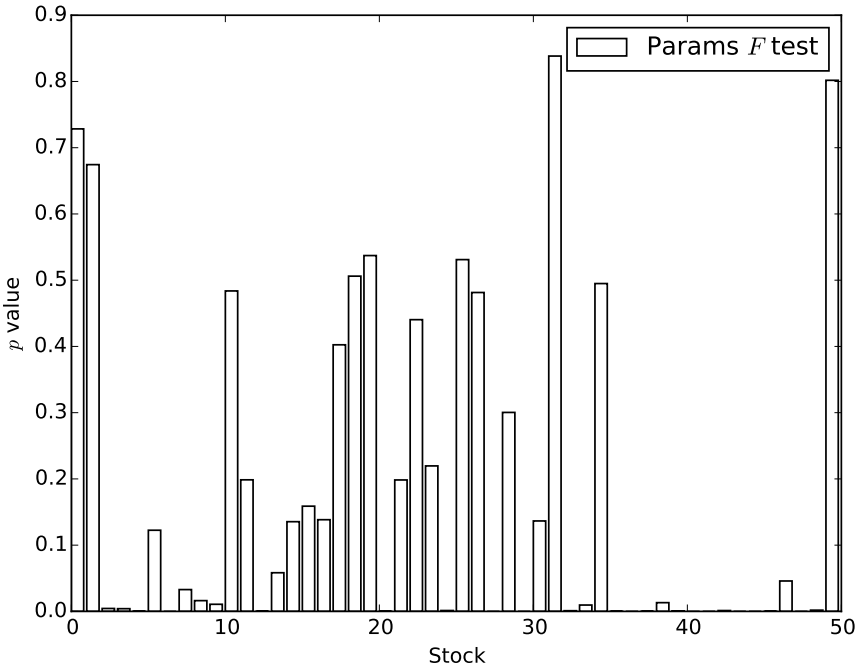

图 4.8　上证 50 价格与积极讨论帖点击量参数 F 检验

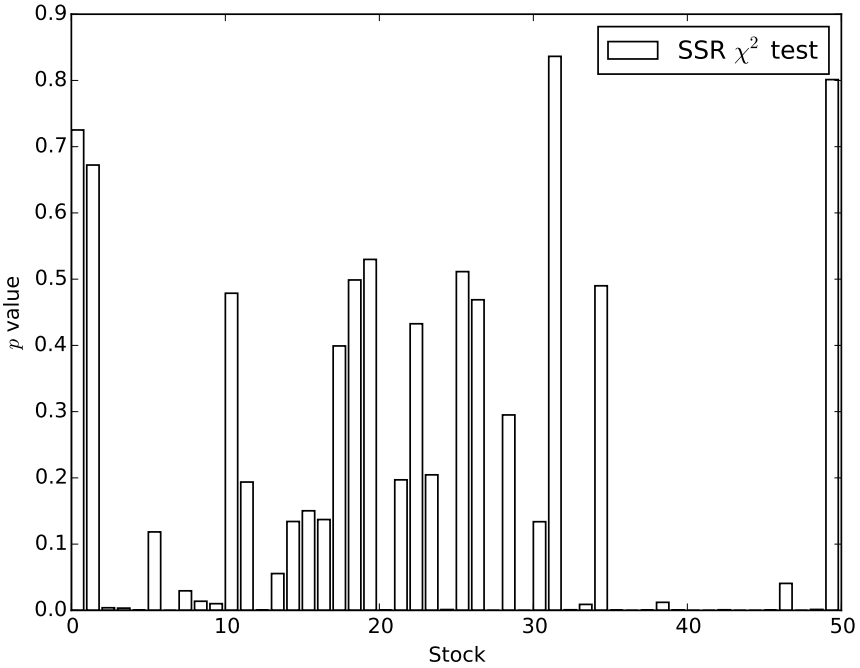

图 4.9　上证 50 价格与积极讨论帖点击量残差平方和 χ^2 检验

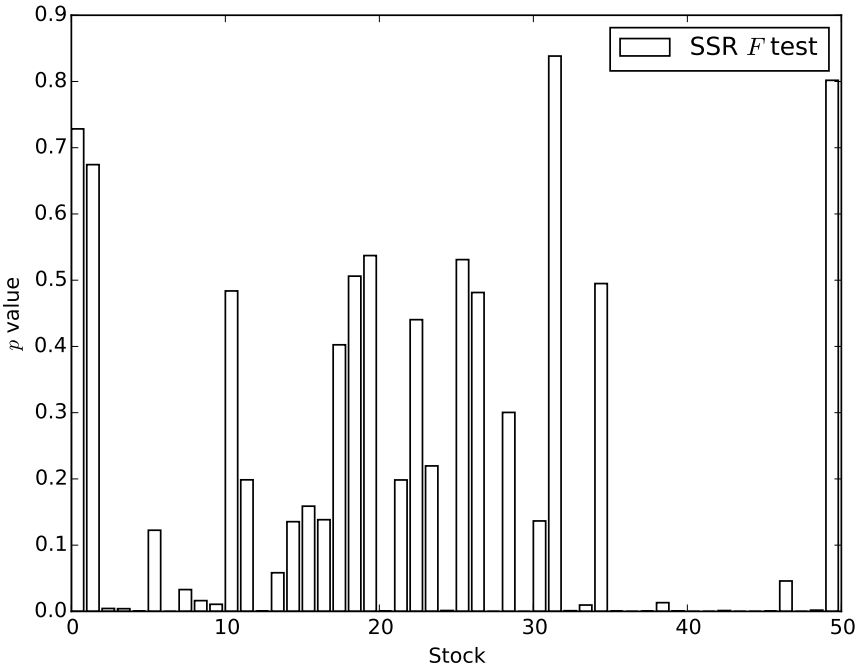

图 4.10　上证 50 价格与积极讨论帖点击量残差平方和 F 检验

是加入更多的维度的信息，分析更多不同的情感类型。或者可以引入非线性的方法。现在只是取积极情感在总讨论帖点击量中的比例，来获得积极讨论帖点击量的数量。可以考虑使用更加复杂的模型，使得得到的点击量的数量更加具有区分度。

4.4 价格趋势预测

得到积极情感成分对应的讨论帖点击量与价格的关系之后，可以试图构建使用积极情感成分进行价格预测的模型。同样使用线性回归模型，如式 4-1 所示。

$$y_{t+1|t} = c + A_1 y_t + A_2 y_{t-1} + \cdots + A_p y_{t-p+1}. \tag{4-1}$$

与之前使用讨论帖点击量预测成交量趋势所不同的是，这里使用的是讨论帖点击量中的积极情感部分，所以逻辑如下所示。

1. 获取每日讨论帖点击量中的积极情感部分
2. 使用积极情感部分和价格建立向量自回归模型
3. 使用模型预测后续价格

以招商银行 (600036) 为例，基于上述步骤构建价格预测模型，并在历史数据上进行测试。在结果中，如果落后期为 l ，则前 l 个时间点是没有预测结果的。为了使得结果更加直观，将无效部分去除，得到结果如图 4.11 所示。

从图上直观上可以发现吻合度非常高。计算得到标准均方根误差为 2.546×10^{-2} 。

接着在上证 50 所有股票上使用相同的方法进行测试。在此不再一一展示结果，只展示各只股票预测的标准均方根误差如图 4.12 所示。

不难发现，所有股票的均方根误差都比较接近，位于 $[0.01, 0.035)$ 区间内。可以认为预测模型基本有效，能对股票价格有一定的解释性，并且这种有效形是普遍的。最后需要注意在构建模型的过程中，并没有使用随机算法或者手动设定参数，所有参数的选择都是通过确定性算法得来的，这使得模型有更强的适用性，减少过拟合的可能性。

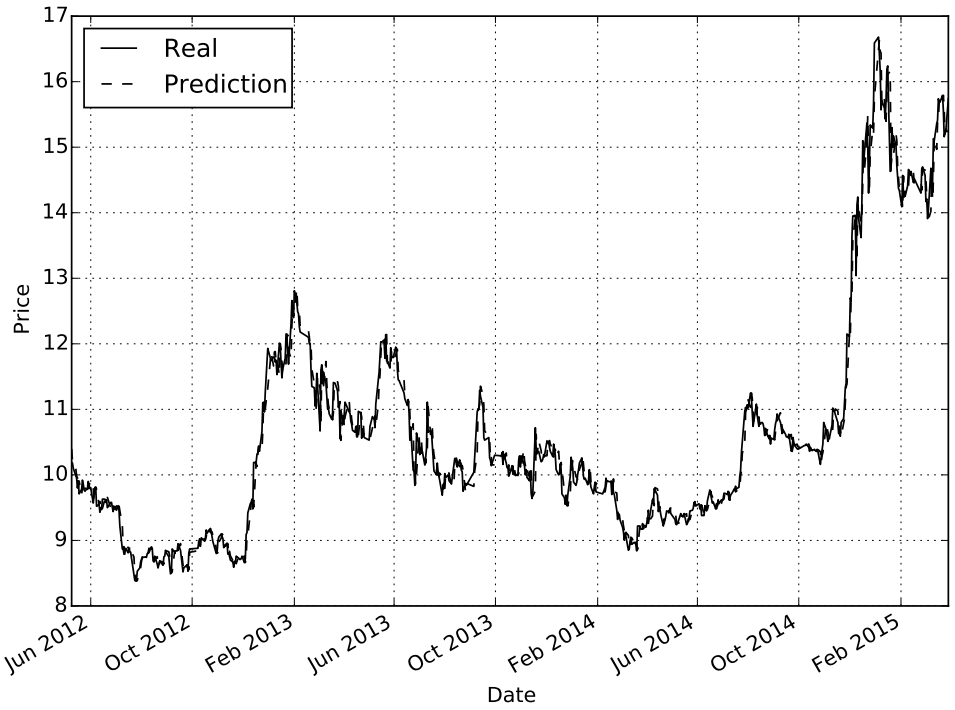

图 4.11　招商银行 (600036) 价格预测

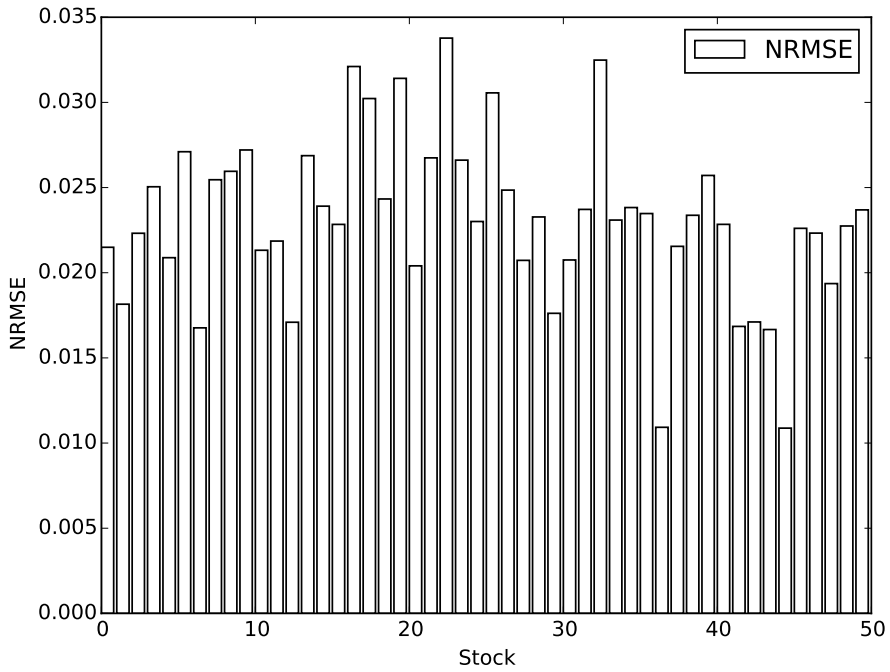

图 4.12　上证 50 价格预测标准均方根误差

第 5 章　结论

由成交量、价格分别和讨论帖点击量的格兰杰因果关系检验可以发现，绝大部分上证 50 成分股的成交量方面都显示强烈的格兰杰因果关系。可以认为是由股民的关注度很大程度上决定了股票的交易量，证实了股民情感在股票交易中起到的重要作用。

而相应地价格与讨论帖点击量的因果关系则稍有不同，部分股票仍有较强的格兰杰因果关系，部分股票则相对较弱。为了进一步挖掘它们之间的联系，引入情感极性分析的办法，对于之前相关性较弱的股票取得了一定的进展。发现讨论帖点击量中有更加细分的组成部分，如果能单独分离这些部分，可以获得更加准确的与收盘价格之间的格兰杰因果关系。继续证实了股民情感在股票交易中的作用。

本文还有很多不足，例如并没有对许多更高级的模型进行尝试。本文仅仅对最基本的线性模型进行分析，得到数值上有指示意义的结果，证实因果关系的确切存在。要使得关系更加明显，准确度更高，必须使用更加细致的分类方法和更高级的模型。

插图索引

表格索引

参考文献

[1] 陈兴, 孟卫东, 严太华. 基于 T-S 模型的模糊神经网络在股市预测中的应用. 系统工程理论与实践, 2001, (02)

[2] Sharpe W F. Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, 1964, 19:425–442

[3] Lintner J. Security prices, risk and maximal gains from diversification. Journal of Finance, 1965, 20:587–616

[4] Keynes J M. The General Theory of Employment, Interest, and Money. London: MacMillan, 1936

[5] Hume D. An Enquiry concerning Human Understanding. Oxford University Press, 1748

[6] De Long J B, Shleifer A, Summers L H, et al. Noise trader risk in financial markets. Journal of Political Economy, 1990, 98:703–738

[7] Lee C, Shleifer A, Thaler R. Investors sentiment and the closed-end fund puzzle. Journal of Finance, 1991, 46:75–109

[8] Kumar A, Lee C M. Retail investor sentiment and return comovemonts. Journal of Finance, 2006, 61:2451–2486

[9] Baker M, Wurgler J. Investor sentiment and the cross-section of stock returns. Journal of Finance, 2006, 61:1645–1680

[10] Lemmon M, Portniaguina E. Consumer confidence and asset prices: Some empirical evidence. Review of Financial Studies, 2006, 19:1499–1529

[11] 吴云勇, 范树杰. 证券投资分析方法研究. 中国市场, 2012, 27

[12] 许泱. 基于神经网络的股票市场预测研究 [D]. 华中科技大学, 2008

[13] 李妙用. 基于混合神经网络的股票预测研究系统 [D]. 华南理工大学, 2010

[14] Baba N, Suto H. Utilization of artificial neural networks and the td-learning method for constructing intelligent decision support systems. European Journal of Operation Research, 2000, 122(2):501–508

[15] Kimoto T, Asakawa K, Yoda M, et al. Stock market prediction system with modular neural networks. Neural Networks, 1990., 1990 IJCNN International Joint Conference on, 1990. 1–6 vol.1

[16] Das S R, Chen M Y. Yahoo! for amazon: Sentiment extraction from small talk on the web. Management Science, 2007, 53(9):1375–1388

[17] Granger C W J. Investigating causal relations by econometric models and cross-spectral methods. Econometrica (The Econometric Society), 1969, 37(3):424–438

[18] The statsmodels development team. Statsmodels[EB/OL]. [2015-06-01]. http://statsmodels.sourceforge.net/

致　谢

衷心感谢导师余宏亮副教授对本人的精心指导。他的言传身教将使我终生受益。他严谨细致、一丝不苟的作风一直是我工作、学习中的榜样。他循循善诱的教导和不拘一格的思路给予我无尽的启迪。

感谢众多朋友对我的支持和帮助，他们用各自领域的专长帮助指导我，带我迅速了解相关领域基础知识，为我指明正确的道路。

还要感谢我的家人，让我在漫长的人生旅途中使心灵有了虔敬的归依，而且也为我能够顺利的完成毕业论文提供了巨大的支持与帮助。

感谢 ThuThesis，它的存在让我的论文写作轻松自在了许多，让我的论文格式规整漂亮了许多。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

附录 A　外文资料的调研阅读报告或书面翻译

There has been a number of attempts to predict future stock price movements. The implication of the efficient market hypothesis (EMH), which was first introduced by Fama [1], was more or less accepted by all these approaches. The semi-strong form of EMH assumes all publicly available information is present in the prices of the market, and new information could be incorporated rapidly. So it is impossible to predict price fluctuation and do better than an ordinary buy-and-hold strategy.

However there are new approaches that utilize investors' sentiment as an index of price fluctuation. Baker and Wurgler [2] states the importance of incorporating sentiment into models to predict future returns. They point out six different proxies to identify various investor sentiment, and show their effects on price fluctuation.

In the mean time, the internet evolved into something O'Reilly [3] called Web 2.0. Users could freely share content through all kinds of services. They could share photos, files, blogs, or through online bulletin boards and social networks. Researchers then started to focus on anaylizing those user-generated content. There are attempts at monitoring the big crowds in online media (Pang and Lee [4]). Others authors also tried to find the connection between mass opinion and prediction of stock prices.

Wysocki [5] was the first on to systematically work on this problem. He captured data from Yahoo! message boards, and detected some positive correlations between firm characteristics and message board posts. Based on the result, he tried to predict future stock market activities. His prediction was confirmed over a large scale on overnight posting activities, but showed no result on daytime counterpart.

However, on the other side, Dewally [6] found out that information on social media do not contain actual valuable information, because of its accuracy. In spite of this finding, the author proposed a trading strategy based on social media data. Dewally [7] also analyzed stock recommendation from online newsgroups in one of his earlier research paper. The finding spawns two different time periods, and found out that positive recommendations are often a result of strong stock performance happened before. It does not contain information that is predictive.

Sentiment analysis from text sample is a very difficult task. One reason for this is that text is often ambiguous. It cannot be assigned a single meaning without considering the context. Not to mention there are cases that text is without any sentiment at all. As a result, text sentiment analysis often has a high error rate. To alleviate this problem, people often restrict the number of target classes. For text containing financial information, it is often categorized into either positive or negative, focusing on the "buy" or "sell" side sentiment polarity. As an addition, some researchers might include a third sentiment, a neutral one. All others are not considered. Pang and Lee [4] gave an review of sentiment analysis methods and some of the problems faced.

Sentiment analysis on stock market is often carried out in two ways. The simple method is to count individual word occurences and use a naive Bayesian classifier. It is also called bag-of-words method. The other method involves part-of-speech tagging and recognition. It requires more complex language models to detect contextual information. Wilson et al. [8] extracted contextual polarity based on sentences. Das and Chen [9] use combination of different algorithms to classify bulletin board posts. It forms a knowledge-discovery architecture. In this context, adjectives and adverbs are tagged and analyzed. It assumes that in a sentence, adjectives and adverbs contain more information about the whole context, and hence higher significance in the classification process.

Most recently, with the development in the area of neural network, sentiment analysis is also seeing a new opportunity. A simple neural network can often capture the structure and individual words simultaneously in a speech, and offering a much better solution compared to traditional methods.

Also, there are attempts to include other mood dimensions. Based on the work of Wilson et al. [8], OpinionFinder (OF) provides indexing of various aspects of subjectivity within the speech. There is also another tool called Profile of Mood States (POMS), along with its derivative GPOMS. It has six classifiers: calm, alert, sure, vital, kind, and happy.

I have so far covered various data sources and method to analyze those data sources. There are surely many more new methods and tools out there, with far exceeding better results. There are also different methods to identify the relationship

between the interpreted data and stock market feed. A simple approach would be using time-lagged vector autoregressive models. Neural networks could also be utilized in the stage, providing non-lineariaty and much flexibility.

The research area is still growing rapidly and a number of new unprecedented methods is emerging. Future research should not restrict themselves to the existing methods and models.

References

[1] Fama Eugene F. Efficient Capital Markets: A Review Of Theory And Empirical Work. The Journal of Finance 25.2 (1970): 383.

[2] Baker Malcolm P. and Wurgler Jeffrey A. Investor Sentiment And The Cross-Section Of Stock Returns. SSRN Journal. (2006)

[3] O'Reilly Tim. What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. Communications & Strategies 17.1.

[4] Pang Bo and Lee Lillian. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2.1-2 (2008): 1-135

[5] Wysocki Peter D. Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards (November 1998). University of Michigan Business School Working Paper No. 98025.

[6] Dewally Michaël. The Informational Value of Earnings Whispers. American Jounal of Busisness 23.1 (2008): 37-52

[7] Dewally Michaël. Internet Investment Advice: Investing with a Rock of Salt. Financial Analysts Journal 59.4 (2003): 65-77

[8] Wilson Theresa, Mooney Raymond J., Wiebe Janyce and Hoffmann Paul. Recognizing contextual polarity in phrase-level sentiment analysis. Human Language Technologies Conference (2005): 347-354

[9] Das S. R., Chen M. Y. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science 53.9 (2007):1375-1388