面向财经新闻的文本挖掘系统设计与实现

摘　要

以浏览器-屏幕为媒介传播的财经网站所披露的新闻，涉及到了上市公司的董事意见、产业行情、经营水平和财务战略决策等一系列重要信息，是股民进行投资的重要信息来源。因此大型财经新闻网站的报道，将在一定程度上影响股民的判断与决策。但是，财经新闻的价值到底有多大，如何以批判性思维的角度来辨识财经新闻，是各国学者进行财经研究与分析的热点。

文本挖掘是数据挖掘技术的一个重要组成部分，它通过计算机自动地从不同的文本源中抽取出可用的信息。把这种信息与新现象和假设联系起来，以探索出传统研究手段所研究不到的盲点，是文本挖掘的要点与难点所在。因此，面向财经新闻，构建文本挖掘系统，指导股民进行投资决策，具有理论价值也应用价值。本文通过系统设计与分析后发现，新闻报道中积极词汇往往多于消极词汇，而且新闻报道的情感与股市的涨跌没有正相关关系。此外，本文使用了 ARIMA 模型，基于板块数据对未来 7 天的相关板块趋势进行预测，具有一定借鉴指导意义。

关键词：文本挖掘；财经新闻；投资决策；ARIMA 模型

Financial news oriented text mining system design and implementation

Abstract

The news, which propagated by the media of browser – screen, involves the directors of listed companies of opinion, industry, market, management level and the financial strategy and so on, is important information source of investors to invest. Thus, the reported by the large financial news website will to a certain extent affect people's judgment and decision-making. How much the value of the financial news, and how to identify the perspective of critical thinking in the financial news, is a hotspot of financial research among international scholars.

Text mining is an important part of data mining technology, which can automatically extract from different text source of the information available through the computer. It is the point of text mining and the difficulty, to relate this information with the new phenomenon and assumptions, and to explore the blind spot of traditional research methods research is less than. Therefore, it has the theory value and application value, to construct a system of text mining, guide people to investment decision-making, facing the financial news. In this paper, through the system design and analysis, found that after active vocabulary is often more than the negative words in the news report, and news reports of emotion and downs of the stock market is not positive correlation. In addition, this paper use the ARIMA model, based on the plate data to forecast the future trend of the seven days of related sectors, has certain guiding significance for reference.

Keywords：Text Mining；Financial News；Investment Decision；ARIMA model

目录

1. 绪论

1.1 研究目的和意义

科技日新月异，人类文明随着互联网的普及、信息传播速度的提高而迈进了新纪元。近年来，国家将迎接互联网+时代作为一项重大政策。

在互联网+的时代，网络已然渗入到人类生活的各个层面，极大地影响了人类社会的文化、经济和政治。其中，以浏览器-屏幕为媒介传播的财经网站所披露的新闻，涉及到了上市公司的董事意见、产业行情、经营水平和财务战略决策等各个方面。中国网民人数众多，且当下股民绝大多数为个体户网上自行操作，个体户炒股的决策信息也来源于网上，因此大型财经新闻网站的报道，会影响大量炒股个体户的决策。

在此大背景下，结合互联网+时代下的大量网民访问的财经新闻网站，和计算机学科前沿的文本挖掘技术，研制出面向新闻财经的文本挖掘系统，可以方便地在浩瀚如海的数据中大浪淘金，获得统计信息，便于股民决策，具有理论价值也有应用意义。

1.2 财经新闻综述

1.2.1 简述

财经新闻，也被称为泛经济新闻，广义上是指囊括所有社会经济生活的且跟经济有关的领域，其中涉及到了从生产到消费、由安全生产到服务质量、从宏观到微观、跨越城市到农村等等相关领域。

狭义的财经新闻定义则重点关注资本市场，并用金融资本市场的视角看中国经济。

1.2.2 网络时代下的财经新闻现状

“在信息经济时代,信息作为一种战略性资源,已在国民经济和社会发展中居于主导地位,并深刻地影响着世界各国的经济增长模式和人们的社会生活方式。正如 300 多年前资本和能源取代土地和劳动力一样,信息也正在取代资本和能源而成为创造财富的新型战略资源。”[1]在信息经济时代，网民的数量日益剧增，支付宝等电子银行的出现，以及付款方式的变革，使得货币的流通更加数字化与简便化。另一方面，手机、平板的出现，使得民众阅读新闻的方式发生了巨大的变革。

作为网络时代下的财经新闻，如新浪财经、东方财富网与和讯网等，有着比以往更多的受众(读者)，而这些读者中很大一部分都拥有电子银行，可以方便得进行金融操作，因此当下的财经新闻的影响力比过去更加巨大。

1.2.3 选用新浪财经的理由

1、新浪财经的市场占有率有绝对领先优势：“新浪财经创建于 1999 年 8 月，经过 10 余年的发展壮大，已经成为全球华人的首选财经门户。新浪财经在财经类网站中占有超过三分之一的市场份额，始终保持绝对领先优势，市场占有率为第二名的三倍！”[2]

2、受众中决策者多：“新浪财经成立十余年来，已深深影响广大中产阶级及高端人群，对企业高管和政府经济决策部门人群的覆盖率超过 90%，始终是高价值网民的首选平台。”[2]

3、效率高：由于新浪是一家规模大的互联网公司，其服务器性能好，爬取其新闻时的响应速度快，不会发生服务器崩溃的问题。

1.3 关于 Python 爬虫

1.3.1 Python 简介

Python 是一门流行的编程语言，已被运用到成百上千个实时商业应用上，包括许多大型且重要的系统。[3]

Python 是一种互动的、解释型的和面向对象的编程语言。它包含了模块、异常、动态类型和非常高的级别的动态数据类型与类。Python 将卓越的能力和清晰的语法结合在一起，拥有许多不同系统间的调用和库的接口，并且对于 C/C++可扩展。同时 Python 也是可移植的，他可以运行在 Unix 及类 Unix 系统、Mac、Windows 上。

1.3.2 爬虫简介

网络爬虫(Network crawler)，顾名思义，是指像在大自然中的爬虫一样，能够在由各个链接相互交叉而组成的网络上自动爬取网页内容的程序，是获取互联网上面浩瀚数据的最佳利器，也是搜索引擎的重要组成部分。网络爬虫的主要功能是从万维网下载网页。在正常情况下，网络爬虫可以分类为聚焦爬虫和传统爬虫。

a. 传统爬虫从一个初始 URL 开始，清洗数据得到初始网页上的所有 URL 链接，在爬取网页信息的过程中，不断地从当前页面上获得新的 URL，放入 URL 队列，直到符合用户设定的条件才停止。简单地讲，就是经过源码分析来得到想要的内容。

b．聚焦爬虫的运行过程较为繁琐，要求依据特定的网页分析算法清洗掉与主题无关的 URL，只留下用户需要的链接，并将他们放进未爬取的 URL 队列。然后依据特定的搜索策略，从还未爬取的 URL 队列中选择下一步要抓取的网页 URL，不断重复上述过程，直到达到用户设立的停止条件时才停止。另外，所有已爬取的 URL 都会被系统储存，进行分析、清洗，建立引索，以方便未来的查询和检索。

1.3.3 选用 Python 写爬虫的理由

选用 Python 写爬虫的原因是：

A．Python 可读性强

B．Python 容易学

C．Python 的爬虫库多

1.4 文本挖掘概述

1.4.1 数据挖掘简介

“数据挖掘(英语：Data mining)，又译为资料探勘、数据采矿。它是数据库知识发现(英语：Knowledge-Discovery in Databases，简称：KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法发现、挖掘出隐藏的信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统(依靠过去的经验法则)和模式识别等诸多方法来实现上述目标。”[4]

1.4.2 文本挖掘

“文本数据挖掘(Text Mining)是指从文本数据中抽取有价值的信息和知识的计算机处理技术。顾名思义，文本数据挖掘是从文本中进行数据挖掘(Data Mining)。从这个意义上讲，文本数据挖掘是数据挖掘的一个分支。”正如伯克利大学的 Sims 在其 blog 上所言：“文本挖掘是通过计算机，自动地从不同的文本源中榨取出那些新的信息出来。把榨取出来的信息与新现象和假设联系起来，以探索出传统研究手段所研究不到的盲点，是文本挖掘的要点所在。”

世界上文本挖掘技术已取得了一定的发展。由于中文区别于英文的语法特殊性，想要挖掘出财经新闻的关键字还是一项比较复杂的工程，难点在于对中文的分词，以及词语性质的分类。

1.5 本系统架构

1.5.1 基于新浪财经的文本挖掘系统总架构

如图 1-1 所示，本系统架构包含三大模块，分别为爬虫、文本挖掘处理 API 以及基于 B/S 的友好交互界面。

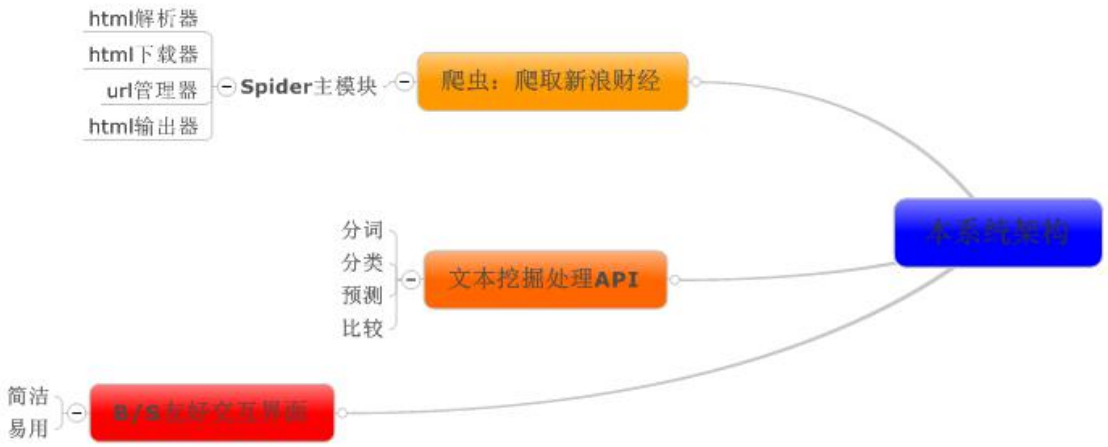

图 1-1：本系统架构目标图

1.5.2 爬虫架构

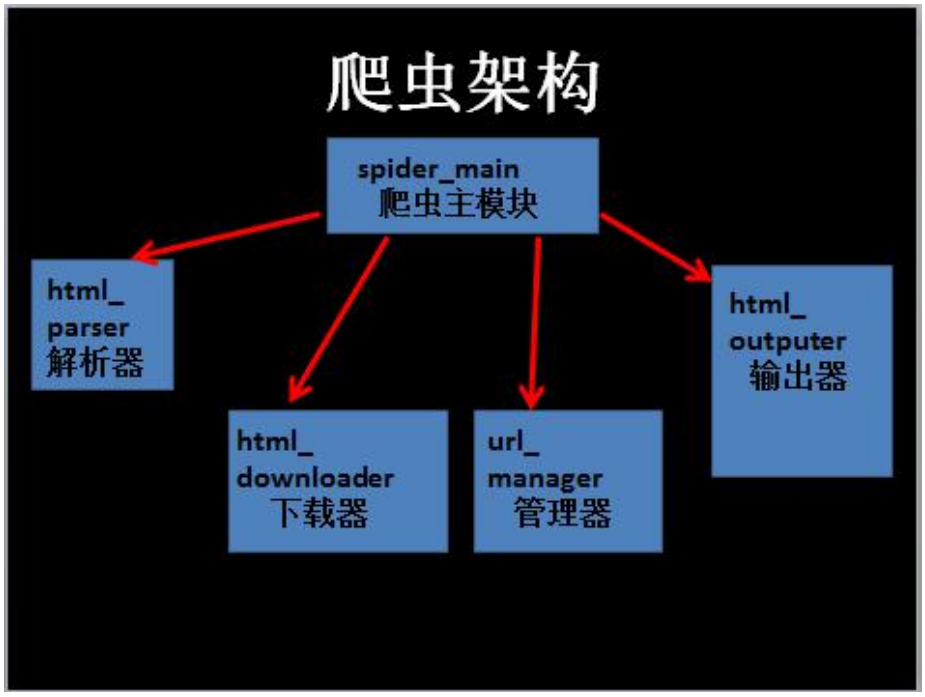

图 1-2：爬虫架构

如图 1-2 所示，其中：

1）在解析器中，设置 timeout 参数，使得链接超时自动重连机制。

2）爬虫入口为新浪财经个股点评网：http://finance.sina.com.cn/column/ggdp.shtml

图 1-3：新浪财经个股点评网

如图 1-3 所示，选择此页面作为爬虫入口的原因为：“个股点评”网站上比较整齐，且其文章内容含有股票名称，比较适合将来的分析。

1.5.3 文本挖掘处理 API 架构

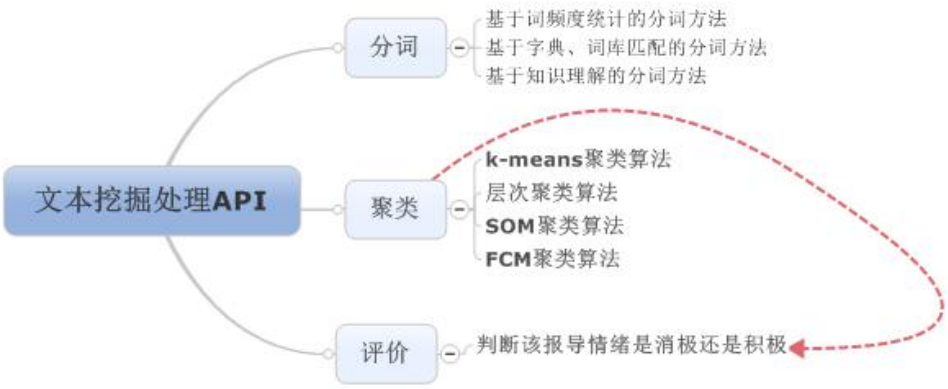

图 1-4：文本挖掘处理 API 架构

如图 1-4 所示，聚类的目的在于判断该篇报导的情绪是否积极，由于中文文本并不是数字，要使用以上聚类方法需要给中文词组分配一个值，而给这个值的过程本身就已经对其分类了，比如“涨”的值为 1，“跌”为-1，这本身就已经对其进行分类。又由于情感分析在本系统中只有积极和消极两类，因此本系统将直接用积极/消极字典进行匹配，则聚类的方法就是将股票报道中的常见词汇加入积极/消极字典中。

2. 数据榨取与分析

2.1 网页源代码分析

爬虫的 URL 入口为：http://finance.sina.com.cn/column/ggdp.shtml

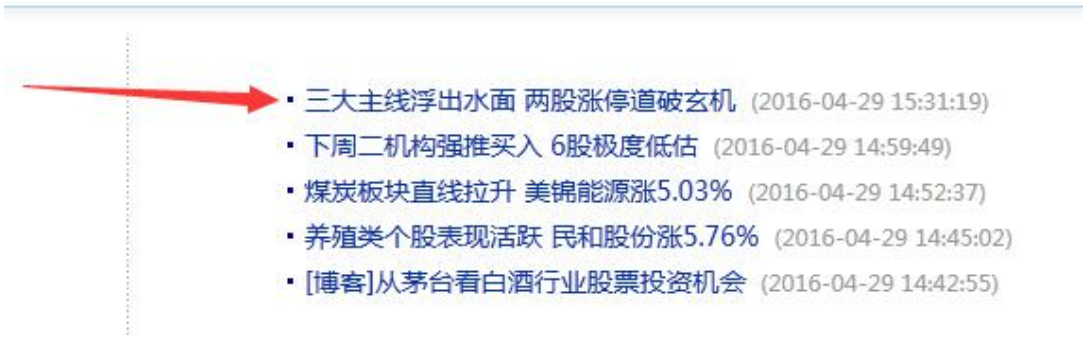

图 2-1：点评网中“正常链接”入口

图 2-1 中所示的链接的源代码为：

图 2-2：“正常链接”源代码

另一种点评为博客，见图 2-3：

图 2-3：点评网中“博客”入口

其源代码如图 2-4 所示：

图 2-4：“博客”链接源代码

可以看到，除了“[博客]”的链接之外，其他的链接(以下称其为“正常链接”)都为 http://finace.sina.com.cn/stock/[字符串]/[数字]/[数字].shtml

因此设置 URL 管理器找到符合下列正则表达式的 URL 入口：

图 2-5：“正常链接”URL 的正则表达式

同理，标题中含有“[博客]”字样的，其对应的正则表达式 URL 为：

图 2-6：“博客”链接 URL 的正则表达式

经过爬虫爬下来的网页源代码为：

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR
<!--[31,656,16] published at 2016-05-23 16:31:46 from #130 by system-->
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<link rel="mask-icon" sizes="any" href="http://www.sina.com.cn/favicon.svg" color="red">
<meta http-equiv="Content-Type" content="text/html; charset=gb2312" />
<title>个股点评_财经_新浪网</title>
<meta name="Keywords" content="财经,新浪网" />
<meta name="Description" content=",财经_新浪网" />
<script type="text/javascript">
//==================================
/*
舌签构造函数
SubShowClass(ID[,eventType][,defaultID][,openClassName][,closeClassName])
version 1.21
*/
function SubShowClass(ID,eventType,defaultID,openClassName,closeClassName){this.version="1.21"
//==================================
</script>

<style type="text/css">
/* 奥运新闻列表页 */

/* 全局 */
body,ul,ol,li,p,h1,h2,h3,h4,h5,h6,form,fieldset,table,td,img,div{margin:0;padding:0;border:0;}
body{background:#fff;color:#000;}
td,p,li,select,input,textarea,div{font-size:12px;}

ul{list-style-type:none;}
select,input{vertical-align:middle;}

a:link{color:#0334ad;text-decoration:none;}
a:visited{color:#0334ad;text-decoration:none;}
a:hover,a:active,a:focus{color:#c00;text-decoration:underline;}
```

图 2-7：待爬取网页的源代码

由图 2-7 可以看到，有很多无用的数据，而我们需要的数据只有：链接、文章标题、文章内容。因此，需要对数据进行清洗、过滤并规则化，才能进行后续工作（如分词、聚类、评价、归纳出预测模型、将数据和预测模型可视化）。

2.2 数据清洗与过滤并规则化

点进“正常链接”后，可以看到：

图 2-8：“正常链接”页面

其中，箭头 1 指向的 title 的源代码为：

图 2-9：title 的源代码

箭头 2 指向的 text 的源代码为：

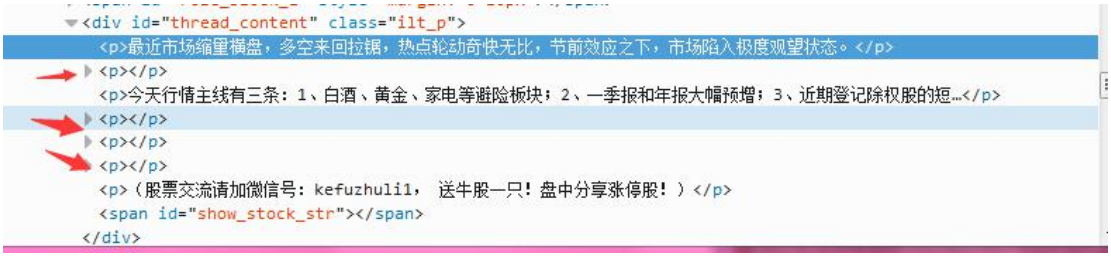

图 2-10：text 的源代码

因此，调用 python 2 的 urllib2 库，获得类为“articalTitle”的 h2 标签的内容。

```
title_node = soup.find('div', class_="articalTitle").find("h2")
```

图 2-11：获取“artivalTitle”内容的 python 代码截图

同理，对正文，以及“博客链接”的网页做类似处理。

如此清洗后，输出所得到的数据，再去掉换行符、影响 csv 读取的英文引号等干扰字符，并添加日期，形成规范的 csv 文件，其数据结构为：

“TITLE”	“TIME”	“TEXT”	\N

表 2-1：csv 文件的数据结构表

其中，Text 中的换行符和英文双引号已被替换为字符“N”。

效果如图 2-12：

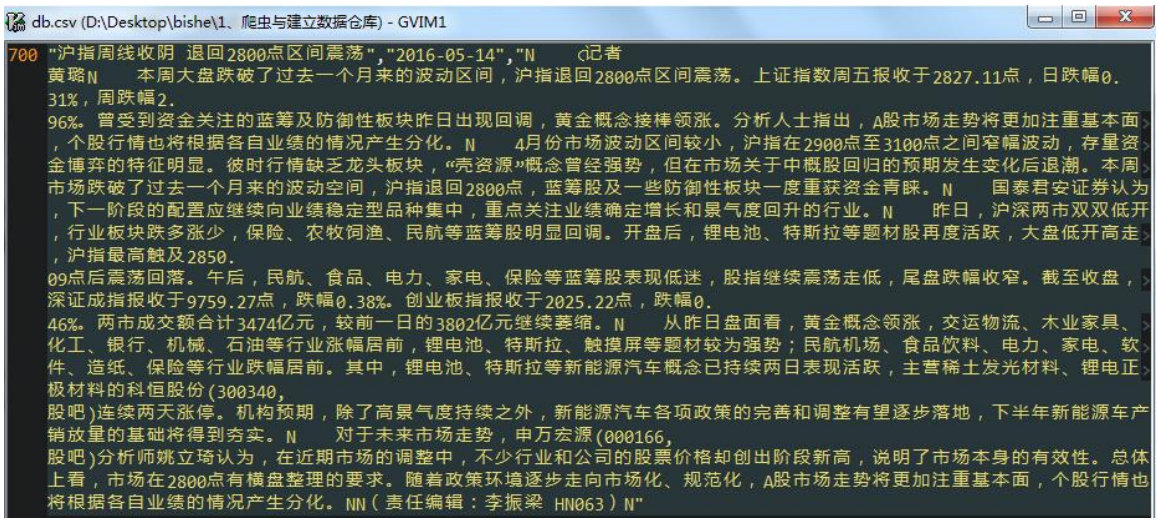

图 2-12：规范的 csv 文件

2.3 中文分词

2.3.1 中文分词的必要性

词(word)是最小的能够独立活动的有意义的语言元素。将文章进行分词，可以更有效率地提取出文章里的主要信息，方便检索和建模。

西方语言，如英文，其单词与单词之间是的自然分界符为空格，很容易进行分词。然而汉语并不像西方语言，中文的词与词之间并没有空格，因此要实现中文的分词处理，关键在于如何确定单词的边界。

2.3.2 分词技术简介

目前为止，分词技术经过了一定的发展，不管是中文汉语、西方英语或者是梵文，其分词方法有三种，第一种是基于人工智能知识理解的分词，第二种是基于数据仓库中词频统计的分词，最后一种是基于关键字字典的匹配分词法。

1）基于人工智能知识理解的分词，它包括三个部分：分词子系统、句法语义子系统和总控部分。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。这种方法的关键在于要让机器具有人类的理解力，因此实现难度很大。

2）基于数据仓库中的词频统计方法，是把相邻字间的信息、词频以及相应的共现特性等应用于分词。因为这些信息是通过真实语料产生的，所以此方法实用性比较好。

3）基于关键字字典的匹配分词法是指通过搜索关键字词典，进行匹配来分词。具体匹配算法有最大匹配法、最小分词方法等。本系统采用最大匹配法，此外，调用的 Jieba 分词模块的原理中也含有此种方法。

2.3.3 LingPipe 的分词效果

LingPipe 是由 alias 公司开发的，基于 JAVA 的一款自然语言处理的软件包。在本系统的研究过程中，使用其对股票点评内容进行分词效果测试。

“后大盘依然维持在高位，软件服务板块涨幅迅速扩大至近 5%，板块内个股共有 13 只涨停。截至记者发稿，涨停的个股有：启明信息、金证股份、卫士通、佳创视讯、超图软件、中科金财、恒生电子、长亮科技、新开普、顺网科技、银之杰、同花顺和北信源。另外，该板块涨幅较大的个股还有：东方通涨 9.79%、美亚柏科涨 9.59%、浪潮软件涨 8.98%。”

上文是一段来自个股点评的主题内容，将其放入 LingPipe 的中文分词 Demo 工具 gui_word_zh_as 中，进行分析。

由图 2-13 可知，股票名称“同花顺”被分成了“同”和“花顺”两个词。因此，此类开源软件并不适合直接对含有股票代码的分词。

2.3.4 基于关键字字典的分词算法

因此，为能对股票点评文章进行分词和研究，本系统采用了基于关键字字典的正向最大匹配分词算法，该算法用 python 实现。原理及测试如图 2-14 所示：

图 2-14：基于关键字字典的分词算法

其中，本系统爬下了 A 股的所有股票名称后，建立了关键字字典，如表 2-2 所示：

A 股股票名称一览表					
国投电力	赣能股份	金山股份	凌云 B 股	凯迪生态	中国核电
豫能控股	甘肃电投	通宝能源	易世达	黔源电力	申能股份
京能电力	红阳能源	华能国际	广安爱众	长源电力	吉电股份
华电能源	粤电力 A	韶能股份	建投能源	宝新能源	银星能源
……	……	……	……	……	……
……	……	……	……	……	……
明星电力	郴电国际	国电电力	天壕环境	新能泰山	哈投股份
浙能电力	*ST 江泉	桂冠电力	三峡水利	湖北能源	天富能源
猛狮科技	海立股份	奋达科技	亿利达	盾安环境	华声股份

表 2-2：A 股股票名称一览表（共约 3000 支股票名称）

运用正向最大匹配分词算法，从字典的第一个字开始，从左到右匹配，查看在数据仓库中是否出现。若出现，则将此关键字剪切、粘贴到新文件中，并将出现次数进行统计，再降序排列，结果如表 2-3 所示：

排名	股票名称	出现次数	排名	股票名称	出现次数
1	国海证券	469	6	天齐锂业	434
2	中国平安	468	7	东北证券	433
3	多氟多	454	8	东方证券	429
4	农业银行	442	9	比亚迪	410
5	纺织服装	438	10	西南证券	401
…	……	…	…	……	…

表 2-3：股票名称出现次数 前十名

2.3.5 python 的 jieba（结巴）分词

为了找出新闻报导中的积极和消极词汇关键字，还需要对报导文本进行分词。由于这部分分词关注的是情感分析，所以可以忽略股票名称，于是就可以使用开源分词工具对文本进行分词。分词后再对其进行频数统计。

由于新闻标题中的字的受众多于新闻主体中的字的受众，因此我们设定标题中的关键字的热度值为 3，新闻主体中的关键字的热度值为 1。

如表 2-4 所示：

所属位置	热度值
“Title”	3
“Text”	1

表 2-4：热度值设置一览表

经过分词、统计并排序后，得出以下表格，共有 160490 个词，一部分词预览如表 2-5 所示：

关键字	热度值	关键字	热度值
,	797533	数据	13756
的	415731	下跌	13684
\b	393164	机构	13602
。	329623	行情	13502
、	205944	高	13392
N	195483	可能	13320
\t	151382	股票	13254
,	101455	责任编辑	13138
……	…	……	…

表 2-5：关键字热度值部分预览表

2.4 股票板块

2.4.1 聚类

聚类是指将总数据集合(包括数字或者文字)中，拥有类似特质的单元组成的新的子集合。对于股票名称来说，其总数据集合是股票名称，若按照常规的聚类方法，需要将各个股票名称进行赋值，再通过各种分析方法进行聚类。然而，考虑到股票的赋值要么为股票代码，要么从特定的标准进行赋值，前者意义不大，后者人为倾向比较重，在对每一支股票进行赋值的时候，实际上已经知道那支股票所代表的特性了。

2.4.2 股市板块

然而，股票名称有一个通用的、合理的且工程量小的赋值方法，那就是股票板块这个属性。只要对每支股票进行板块归属，即可组成“具有类似特质的单元组成的子集合”。而这也是本文本挖掘系统的研究要点所在。

股票的板块分类有行业板块、概念板块和地区板块三种。本系统主要研究按行业分类

的股票板块的涨跌与财经新闻的联系。

按行业分类的股票板块有 33 类，如表 2-6 所示：

股票分类一览表
保险，电力，电器，电子信息，房地产，纺织服装，钢铁，工程建筑，供水供气，化工化纤，机械，计算机，建材，交通工具，交通设施，教育传媒，旅游酒店，煤炭石油，酿酒食品，农林牧渔，其他行业，券商，商业连锁，通信，外贸，医药，仪电仪表，银行类，有色金属，运输物流，造纸印刷，非银金融，其他金融

表 2-6：股票分类一览表

2.4.3 建立数据结构

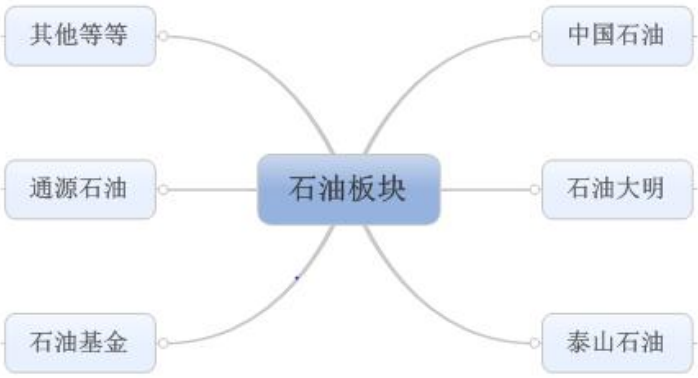

图 2-15：石油板块与股票名称部分关系图

如图 2-15 所示：我们建立了石油板块里股票名称与石油板块的联系，石油板块包含了：中国石油、石油大明、泰山石油、石油基金、通源石油等等。

“煤炭石油”板块
瑞茂通,通源石油,靖远煤电,中海油服,*ST 新集,中国石化,西山煤电,恒泰艾普,潜能恒信,中国神华,*ST 百花,广汇能源,大有能源,昊华能源,平庄能源,海越股份,伊泰 B 股,云煤能源,*ST 黑化,阳泉煤业,*ST 山煤,中煤能源,开滦股份,中国石油,泰山石油,冀中能源,永泰能源,龙宇燃油,美锦能源,*ST 云维,陕西煤业,海默科技,平煤股份,广聚能源,潞安环能,兰花科创,上海能源,盘江股份,兖州煤业,郑州煤电,杰瑞股份,恒源煤电,美都能源,洲际油气,宝泰隆,陕西黑猫,石化油服,露天煤业,安源煤业,*ST 神火,大同煤业,金瑞矿业,国际实业,惠博普,仁智油服,新大洲 A,安泰集团,煤 气 化,吉艾科技,*ST 新亿,准油股份,山西焦化

表 2-7：“煤炭石油”板块所含全部股票名称一览表

如表 2-7 所示，是“煤炭石油”板块所含有的所有股票名称。

根据行业分板块，可以分为 33 个板块，逐一建好这些数据结构，在统计出各个股票的词频之后，就可以得出各个板块的总词频，从而方便进行预测。表 2-8 为各个板块的股票名称一览表：

电力	*电器*	*供水供气*	*化工化纤*	*交通工具*
内蒙华电	海信科龙	国中水务	史丹利	日上集团
皖能电力	骆驼股份	首创股份	利民股份	长安汽车
湖南发展	蒙发力	渤海股份	*ST 亚星	广汇汽车
深圳能源	南都电源	巴安水务	万华化学	福达股份
……	……	……	……	……

表 2-8：各个板块的股票名称部分一览表

2.4.4 结果

表 2-9 为各股票统计结果一览表：

股票名称	热度值	所属板块
美盛文化	112	教育传媒
凯撒旅游	83	旅游酒店
万盛股份	20	化工化纤
天山生物	29	农林牧渔
海特高新	109	交通工具
……	……	……

表 2-9：各股票统计结果一览表

2.5 评价

2.5.1 新闻报导的情感分析

新闻报导中，不仅仅会涉及股票名称和股票代号，还会出现一些对该股的看法，比如看涨或是看跌。

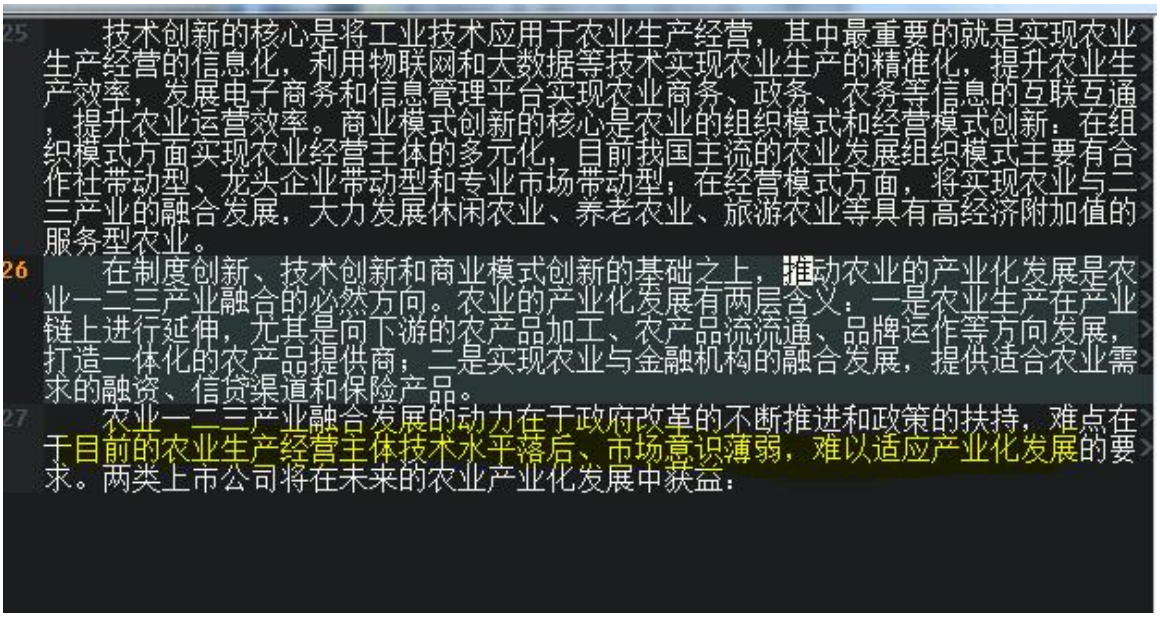

图 2-16：财经新闻报道内容

如图 2-16 中黄色高亮文本所示，“意识薄弱”、“难以适应”等为消极字眼，而“获益”为积极字眼，其他则为中性词。

2.5.2 股票报导的常用情感关键字字典

经过人工寻找前 1000 个词，得出股票报导的常用情感关键字字典为：(如表 2-10 所示)

消极关键字字典	*积极关键字字典*
下跌，震荡，超过，没有，减持，暴跌，波动，回落，大跌，亏损，谨慎，通胀，不足，过剩，弱势，冲击，回调，恐慌，低迷，债转股，跳水，担忧，严重，暂停，悲观，跌幅，下降，低，跌，下行，贬值，减少，降低，下滑，跌破，退市，跌停，杀跌，股灾，压力，炒作，债务，熊市，萎缩，无法，退出，下调，缺乏，难，泡沫，减仓，过度	反弹，发展，增长，上涨，新，高，多，涨幅，超过，有望，买入，提升，涨停，增持，利好，增速，较大，最大，建设，收益，重组，推进，回升，超，积极，上升，涨，看好，大涨，信心，解禁，有效，偏好，修复，推荐，活跃，升级，新增，牛市，加强，回暖，爆发，分红，乐观，大量，有利于，促进，开展，降准，恢复，优质，暴涨

表 2-10：股票报导的常用情感关键字字典

2.5.3 股市大盘的情感分析

经过统计分析，可得数据仓库中，所有报导的总情感。如表 2-11 所示，是 2015-12-28 至 2016-5-13 期间内的报导的情感词汇统计表：

	A	B	C	D	E
1	Time	积极	消极	sum	
2	2015/12/28	20	26	-6	
3	2015/12/29	54	37	17	
4	2015/12/30	72	46	26	
5	2015/12/31	157	102	55	
6	2016/1/1	202	100	102	
7	2016/1/2	148	30	118	
8	2016/1/3	533	373	160	
9	2016/1/4	2033	1778	255	
10	2016/1/5	2182	1899	283	
11	2016/1/6	1425	866	559	
12	2016/1/7	1257	1141	116	
13	2016/1/8	1252	1099	153	
14	2016/1/9	748	491	257	
15	2016/1/10	615	469	146	

表 2-11：2015-12-28 至 2016-5-13 期间内的情感词汇报导的统计表

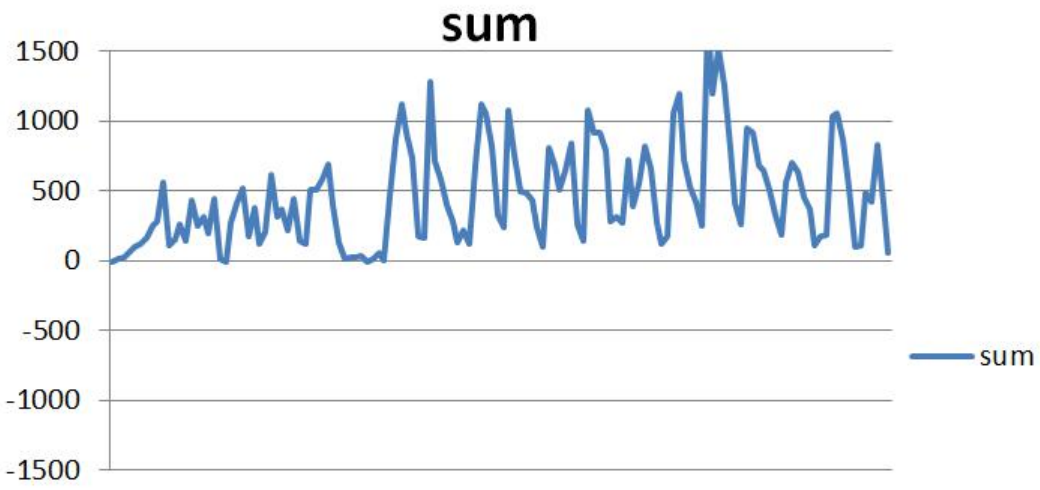

图 2-18：2015-12-28 至 2016-5-13 期间内的报导的情感词汇总和的时间曲线

从表 2-11 和图 2-18 可以看出，sum 绝大多数为正数，说明新闻报道中积极词汇绝大数情况下是多于消极词汇。

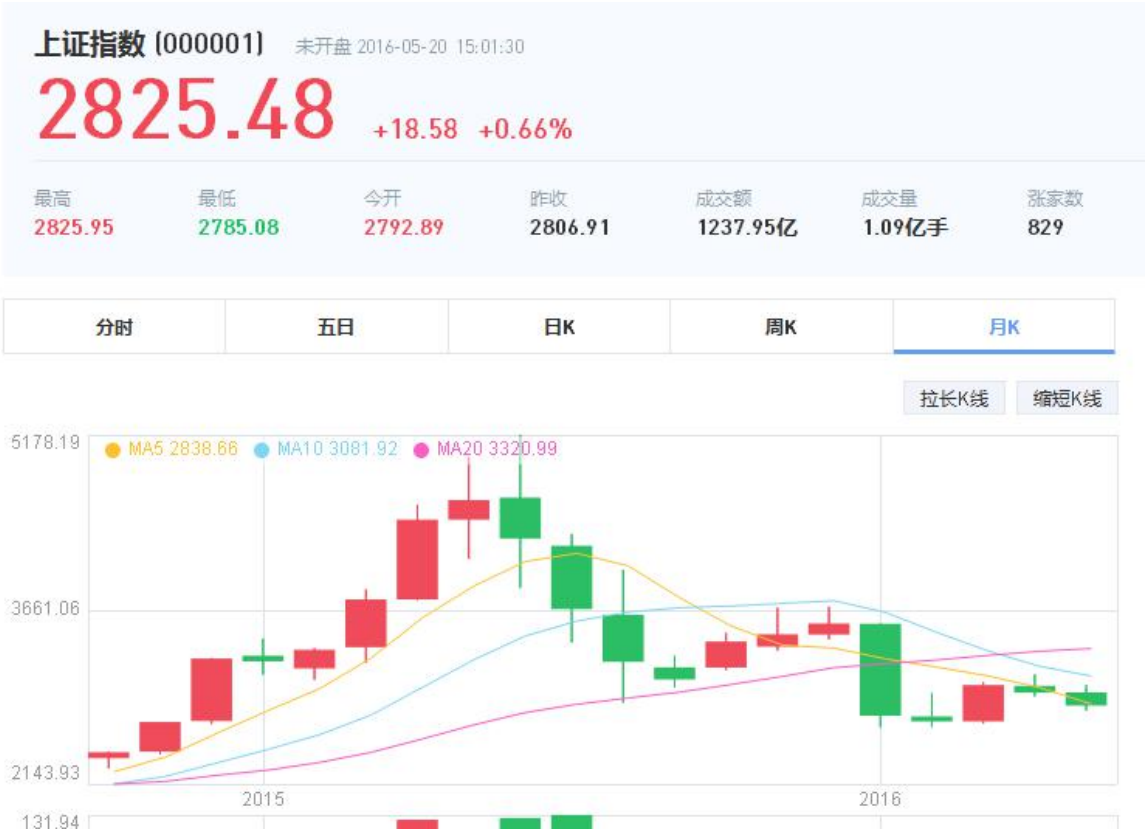

图 2-19：上证指数月 k 图

由图 2-19 可看出，2016 年以来，大盘都是在降，因此可以得出新闻报道中的情感的积极性与走势不存在正相关关系！（至于有没有负相关关系还未知）

2.6 股市预测

2.6.1 股市预测理论

本文本挖掘系统的股市预测理论，基于统计学、新闻传播学和心理学。

其基础有四点：

1）相关性假设：财经新闻可以预测股票，其理论基础是财经新闻与股票之间具有相关性，若无相关性，则一切都无从谈起。本系统假定了财经新闻与股市走势之间具有相关性，从而进行后续研究。若最后得出好的结论，则此假设自然成立。若数据处理过程中出现他们不具有相关性的现象，那么由反证法可知，他们没有相关性的结论。

2）统计学理论：数据仓库中，出现越多次的股票代码和板块，说明对其的预测越可靠（不管用什么分析方法，对出现次数多的股票进行预测，肯定比出现次数少的股票预测来得可信）。

3）新闻传播理论：一个股票代码出现出现越多次，说明受众越多，对市场、股票走势的影响更大。

4）心理学理论：作者拥有炒股经验，也采访过股民的炒股心态，因此决定忽略“报导的消极还是积极”这个维度。本文本系统认为不管报导是消极还是积极，都能使股票上涨。

2.6.2 为什么消极的报导也预测股市上升？

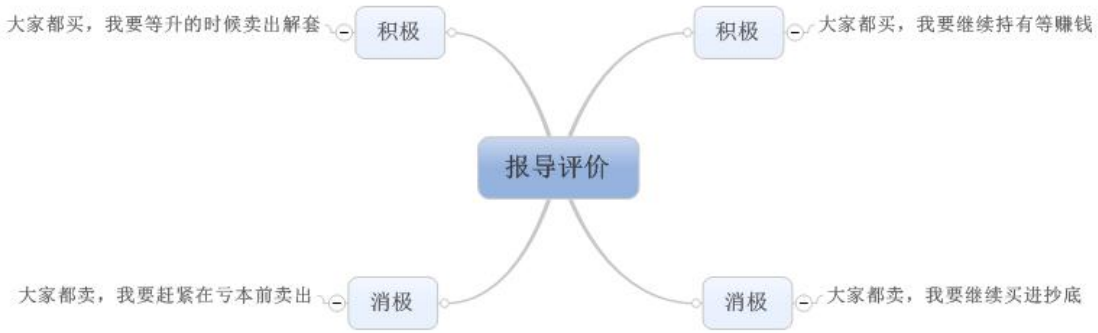

图 2-20：报道评价对股民的决策影响

如图 2-20 所示，不管报导是积极还是消极，总会有不同的人根据自身的情况做出截然不同的股票操作。而在大量数据下，这些微小的差异应该被消除，就如掷硬币一样，当统计量达到一定程度时，就可以默认为正反几率各为 0.5。

而且经过 2.5.3 的分析可以看到，积极的词语总是大于消极的词语。因此，在本系统中，我们假定，曝光度就成了预测股票上涨的唯一因素。

2.6.3 ARIMA 模型方法简介

首先，先确定数据的差分，得到平稳序列。

其次，找到合适的 ARIMA 模型，关键是确定 ARIMA(p,d,q)中的参数，其中，d 为 difference，代表差分的阶数；p 和 q 的值则由 pacf 图和 acf 图来确定。

最后，运用 Ljung-Box 进行检验。

以下是 ARIMA 的理论基础：ARIMA 模型，它是时间序列分析系统里面的一个非常著名的预测算法，具体介绍如下：

1）差分运算

这是一个方法性工具，它可以使我们的模型表达和序列分析更加简洁、方便。所在介绍具体的模型之前，为更好地理解我们所建立的预测模型，我们引入差分运算这一工具。

a. p 阶差分

相距一期的两个序列值之间的减法运算称为 1 阶差分运算。记∇x_t为x_t的 1 阶差分：

$$\nabla x_t = x_t - x_{t-1}. \tag{5.1}$$

对 1 阶差分后序列再进行一次 1 阶差分运算称为 2 阶差分。记$\nabla^2 x_t$为x_t的 2 阶差分：

$$\nabla^2 x_t = x_t - x_{t-1}. \tag{5.2}$$

依此类推，对 p-1 阶差分后序列再进行一次 1 阶差分运算称为 p 阶差分。记$\nabla^p x_t$为x_t的 p 阶差分：

$$\nabla^p x_t = \nabla^{p-1} x_t - \nabla^{p-1} x_{t-1}. \tag{5.3}$$

b. k 步差分

相距 k 期的两个序列值之间的减法运算称为 k 步差分运算。记$\nabla_k x_t$为x_t的 k 步差分：

$$\nabla_k x_t = x_t - x_{t-k}. \tag{5.4}$$

2）ARIMA 模型介绍

具有如下结构的模型称为求和自回归移动平均（autoregressive integrated moving average）模型，简称为 ARIMA(p,d,q)模型：

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, s \neq t. \\ Ex_s\varepsilon_t = 0, \forall s < t \end{cases} \tag{5.5}$$

在（5.5）式中，$\nabla^{\mathrm{d}} = (1-B)^d$；$\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$，为平稳可逆 ARIMA(p,q)模型的自回归系数多项式；$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$，为平稳可逆 ARMA(p,q)模型的移动平滑系数多项式。

求和自回归移动平均模型这个名字的由来是因为 d 阶差分后序列可以表示为：

$$\nabla^{\mathrm{d}} x_t = \sum_{i=0}^{d} (-1)^i C_d^i x_{t-i}. \tag{5.6}$$

式中，$C_d^i = \frac{\mathrm{d}!}{i!(d-i)!}$，即差分后序列等于原序列的若干序列值的加权和，而对它又可以拟合自回归移动平均（ARMA）模型，所以称它为求和自回归移动平均模型。

从(5.1)式中可以看出，ARIMA 模型的实质就是差分运算与 ARMA 模型的组合。这说明任何非平稳序列如果能通过适当的阶数，进行差分运算，实现差分后平稳，就可以对差分后序列进行 ARMA 模型拟合了。而 ARMA 模型的分析方法非常成熟，这就意味着对差分平稳序列的分析也将是非常简单、可靠的。

3）Ljung-Box 检验

Ljung-Box 检验是一种用于检验某个时间段内的一些观测值是不是随机的独立观测值。如果观测值并非彼此独立，一个观测值可能会在 k 个时间单位后与另一个观测值相关，形成一种称为自相关的关系。自相关可以削减基于时间的预测模型（例如：时间序列图）的准确性，并导致数据的错误解释。

Ljung-Box Q(LBQ)统计量将检验最多滞后 k 的自相关等于零的原假设，亦即数据值在某以滞后数 q 之前是随机和独立的。如果 LBQ 大于特定临界值，则一个或多个滞后的自相关可能显著不同于零，这说明该段时间内各个值并不是独立和随机的。本文主要采用此方法来检验下文中的 ARIMA 预测模型的构建是否有效、合适。

2.7 数据可视化

2.7.1 数据可视化

数据可视化，是一种关于数据视觉表现形式的科学技术，也是一种视觉艺术。所谓“一图胜千言”，特别对于大数据文本来说，更需要形象生动地展示出结果。

2.7.2 股市预测的可视化

图 2-21 和图 2-22 是 138 天以来，关于板块报导的热度值权重的示例图：

图 2-21：数据仓库中股票板块热度 wordcloud 可视化图

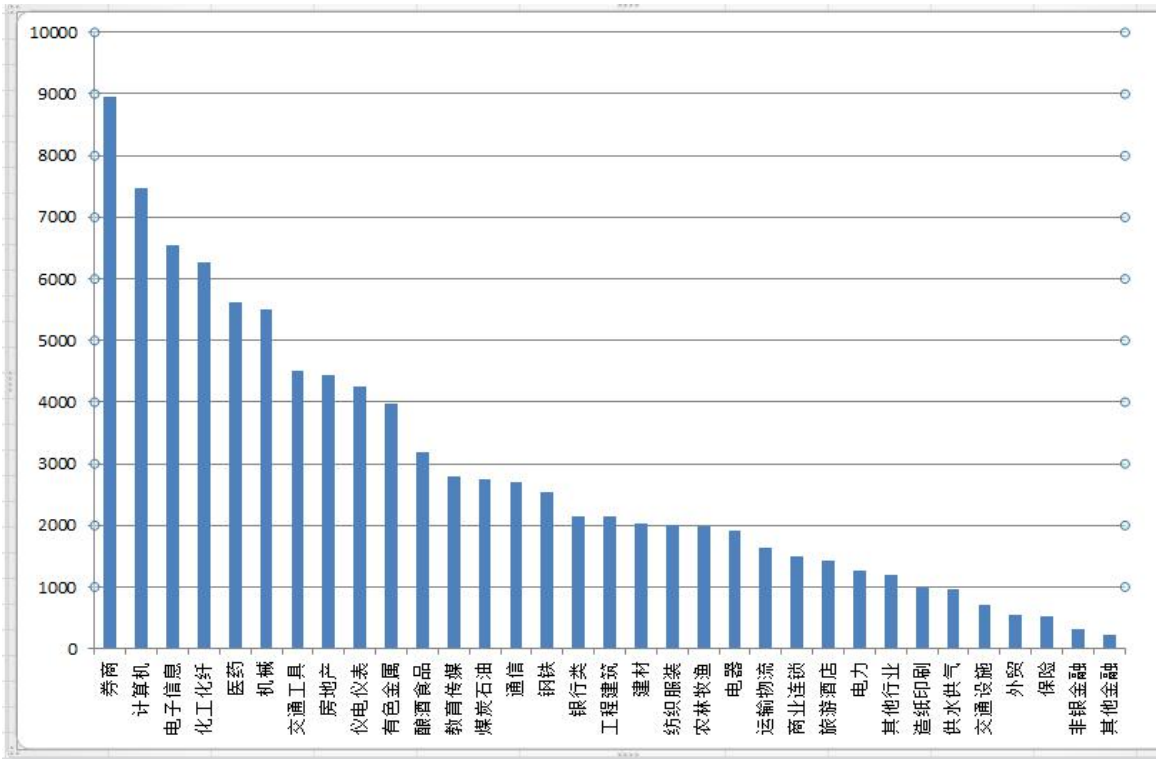

图 2-21：数据仓库中股票板块热度条形图

3. 股市行情验证与反馈

3.1 对未来七天的预测

3.1.1 股市板块锁定

由图 2-20 和图 2-21 可以看到，券商、计算机和电子信息所占的权重最大。鉴于专业和兴趣，选择计算机板块为研究点进行深入研究。

3.1.2 预测过程

首先用 RStudio 读取处理过的数据，即 138 天以来，每一天股票新闻中出现的计算机板块的热度值。如图 3-1 所示：

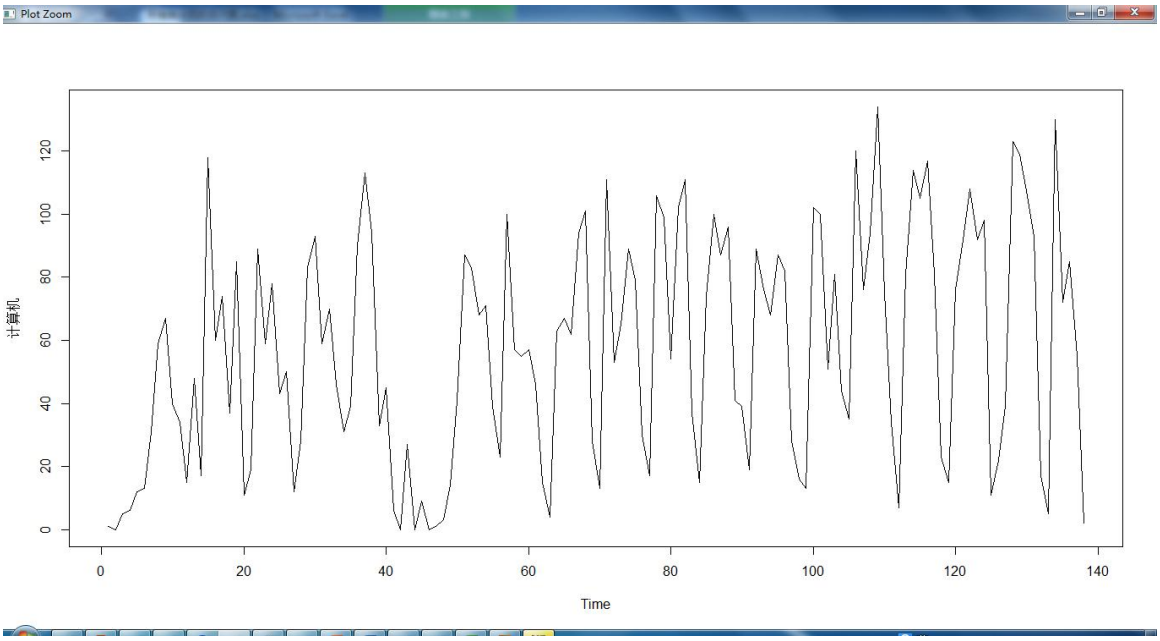

图 3-1：138 天以来，每一天股票新闻中出现的计算机板块的热度值

由图 3-1 可知，数据不太平稳，因此对其进行差分，重复差分，直到平稳：

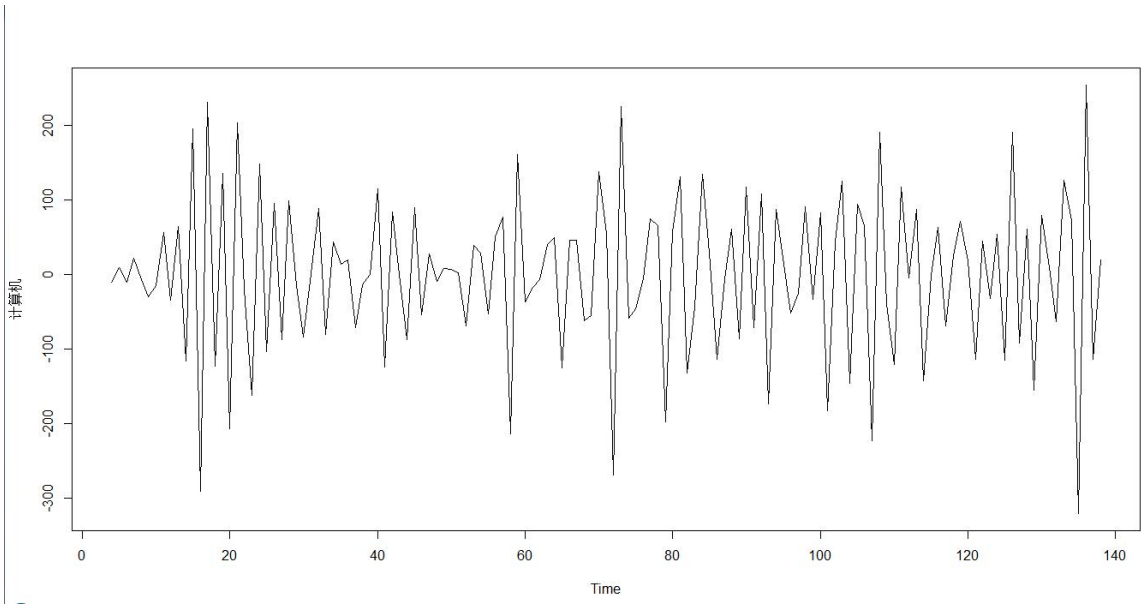

图 3-2：3 阶差分图

如图 3-2 可知，3 次差分后数据较为平稳。随着横坐标（时间）的推移，时间序列的水平和方差大致保持不变。因此，我们需要对数据进行 3 次差分以得到平稳序列。

通过测试不同的 p 和 q，得到 ARIMA(12,3,4)，建模，得到预测图为：

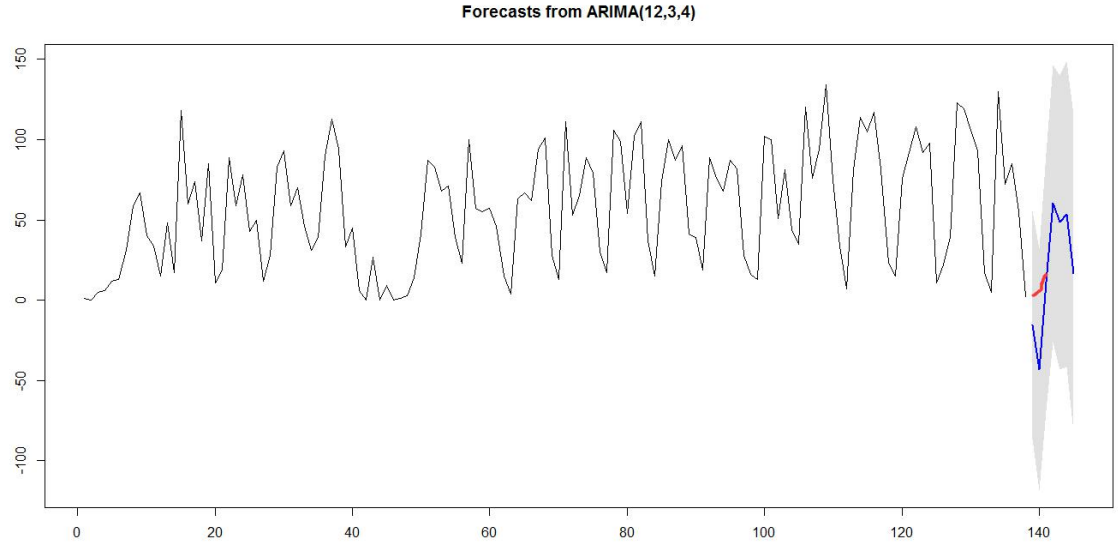

图 3-3：计算机板块未来走势预测效果图

由于预测值不能为负，因此红色部分为修补线，在灰色背景内，因此是在假设下可发生的情况。

预测结果为：计算机板块的股票在财经新闻中出现的次数会先增加后减少，而且极值点没有以往的极值点高，说明跌涨幅不大。

图 3-4 为预测值：

```
> Stockforecast #查看预测值
    Point Forecast    Lo 99.5   Hi 99.5
139      -15.31705  -86.40256  55.76847
140      -43.15928 -118.26048  31.94192
141       11.30198  -68.64488  91.24883
142       60.18867  -26.22255 146.59989
143       48.45862  -43.24480 140.16205
144       53.44995  -41.55311 148.45302
145       16.66932  -82.29858 115.63723
```

图 3-4：ARIMA 对未来 7 天的预测值

自相关检验：如图 3-5 所示，可得出在 1-20 阶内没有一次是超出置信边界的，因此置信度为 95%.

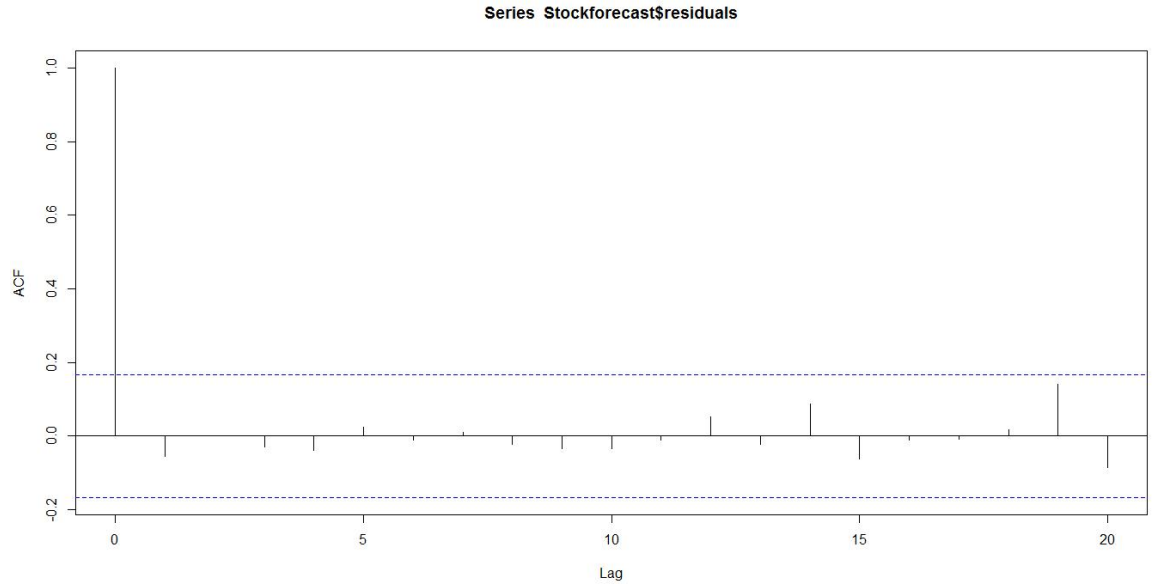

图 3-5：自相关检验图

Ljung-Box 检验：

```
> Box.test(Stockforecast$residuals,lag=20,type="Ljung-Box")

        Box-Ljung test

data:  Stockforecast$residuals
X-squared = 8.1586, df = 20, p-value = 0.9908

>
```

图 3-6：Ljung-Box 检验结果图

由图 3-6 可看出 p-values 的值为 0.9908，故我们有 99%的置信度可以相信在滞后 1-20 阶中没有明显证据说明预测误差是非自相关的。下面我们接着来检查预测误差是否是平均值为零，方差为常数的正态分布，我们可以做预测误差的时间曲线图和直方图：

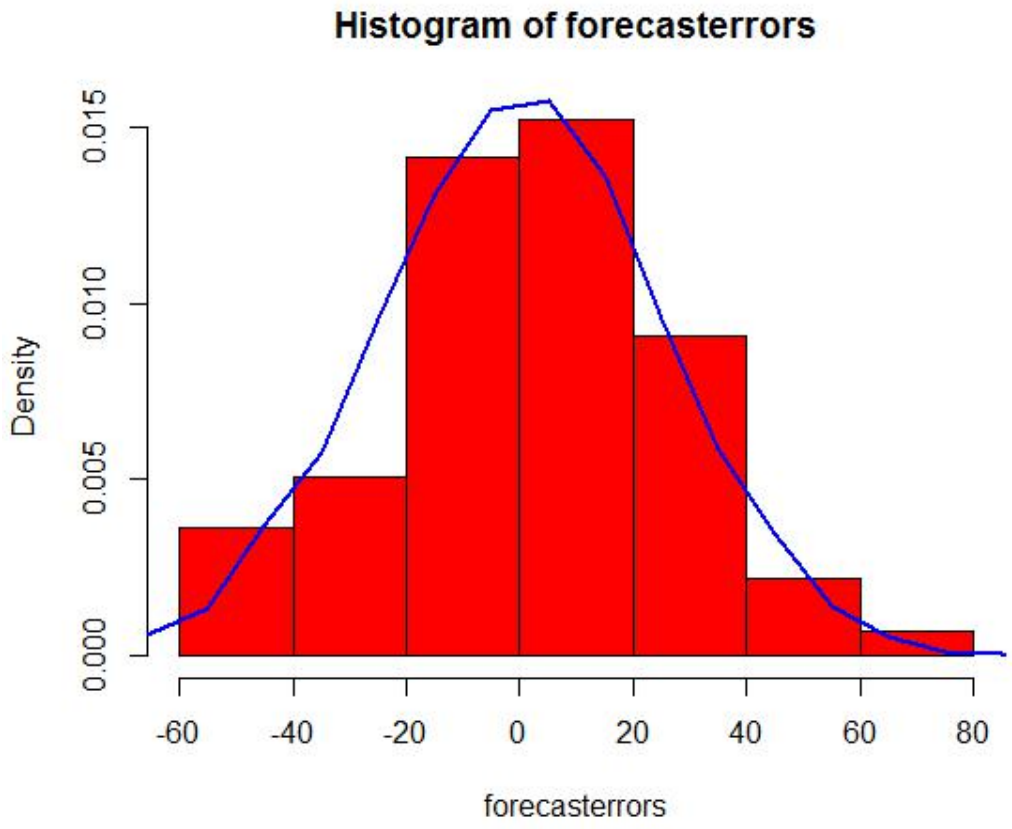

图 3-7：预测误差直方图（带正态分布曲线）

显然，图 3-7 呈现均值为 0 ，方差恒定的正态分布，可见预测算法已无法再继续改进。

3.2 七天后的实际情况

3.2.1 实际的新闻报导中计算机板块热度值

预测的最低值、预测平均值、实际值和预测的最高值如下表所示：

Lowest	Forecast	Real Value	Highest
-86.40256	-15.31705	23	55.76847
-118.26048	-43.15928	17	31.94192
-68.64488	11.30198	49	91.24883
-26.22255	60.18867	86	146.59989
-43.2448	48.45862	93	140.16205
-41.55311	53.44995	63	148.45302
-82.29858	16.66932	57	115.63723

表 3-1：未来 7 天各种预测值和实际值一览表

由表 3-1 作图如下：

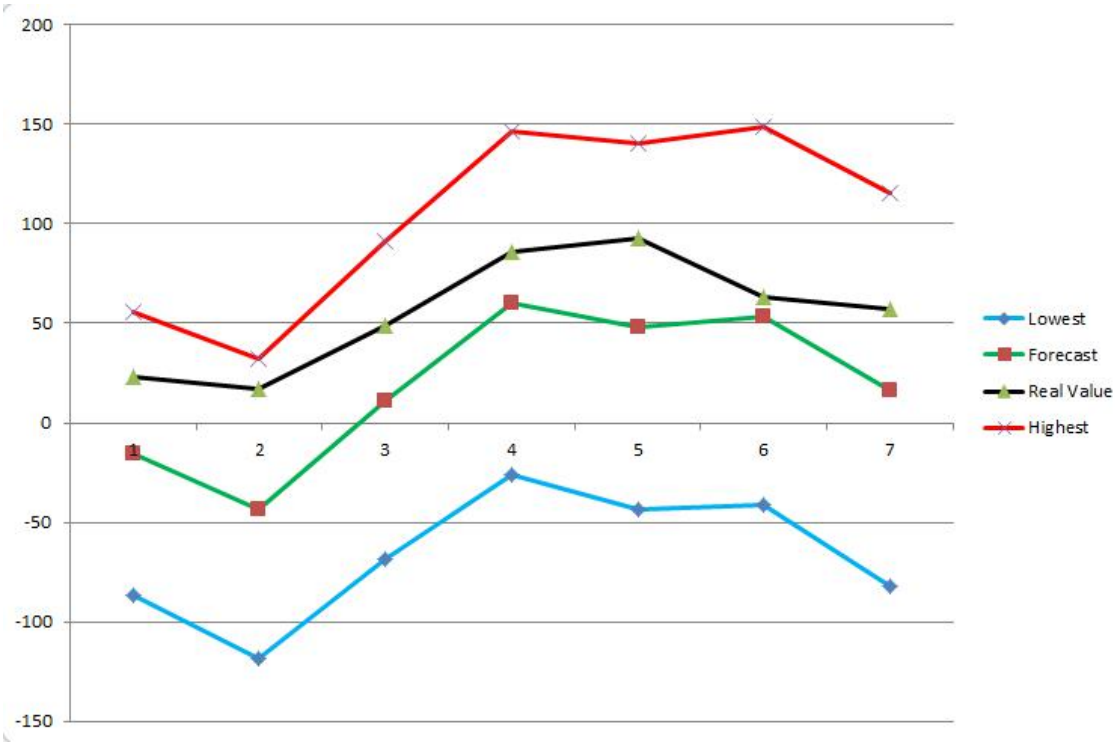

图 3-8：未来 7 天各种预测值和实际值折线图

由图 3-8 可知，虽然实际值跟预测值有偏差，但是实际值都在预测区间内，且升降趋势大概一致，都是先降再升最后再降。

3.2.2 实际的计算机板块股市走势

源数据为 2015-12-28 至 2016-5-13 日共 138 天的新闻报道，预测的是 2016-5-14 至 2016-5-20 号的股市。其中，14 号和 15 号是周末，没开市。这可能解释了预测新闻报导时出现负数的原因，但有待进一步验证。而 16-20 日（5 天）的计算机板块走势图为：

图 3-9：16-20 日计算机板块实际走势图

由图 3-9 可以看到，从 16 号到 20 号间，以每日的收盘价作为观测点，并对股价进行处理，减去 4000，再除以 10，使其的数量级与股票热度值同阶，可以看到其趋势为：

图 3-10：16-20 日计算机板块收盘价趋势图

将其与预测值一起作图得：

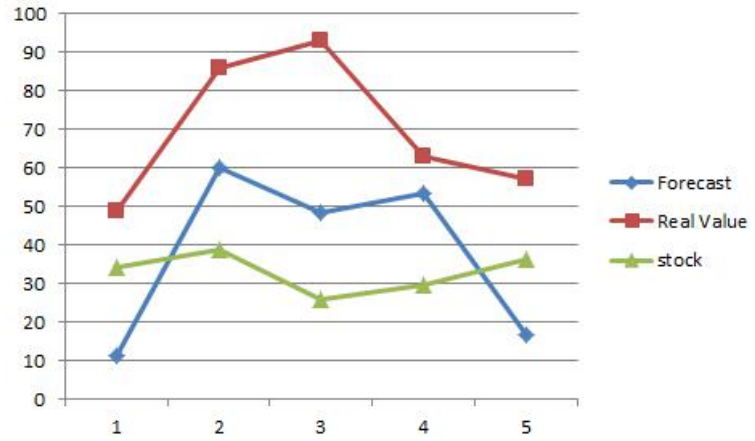

图 3-11：16-20 日计算机板块收盘价趋势与预测值、实际值的趋势对比图

由图 3-11 可以看到，其与预测的那五天的走势趋势基本相符合（前三段的涨跌情况是一致的，最后一天不一致），精确到每一天会增加不命中的概率，毕竟源数据太少，预测的也太少。

而以 5 天为单位，可以看到，无论是预测的股票板块热度值，还是实际的股票板块热度值，还是股票板块走势，都是涨的。如图 3-12：

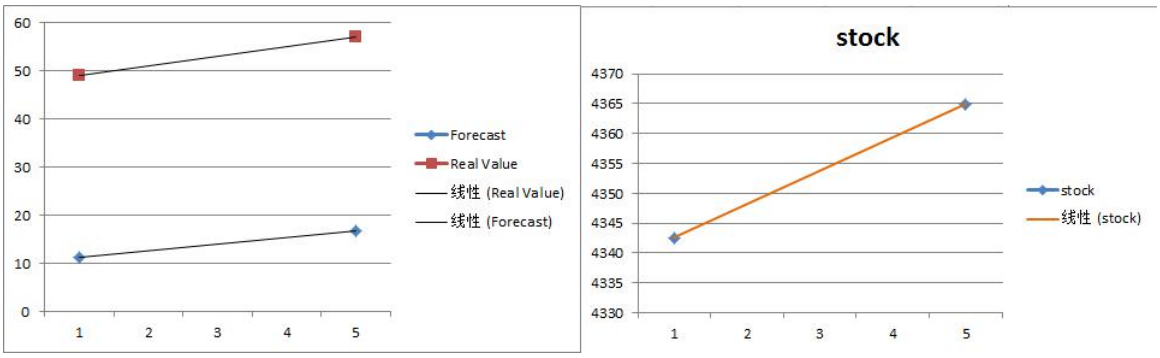

图 3-11：以 5 天为单位的股票板块热度值和实际走势的趋势

3.3 分析预测准确度

3.3.1 涨跌之预测与实际比较

通过对比分析，可以知道，按天预测与实际的误差不超过 20%。而按 5 天来预测，则实际的涨跌与预测的一致。

由于源数据的来源不全面和股市的复杂性等原因，本系统无法对股市进行每一天、每一只股的具体时间段的涨跌进行预测，而只能对股票板块进行未来一星期（5 天开市）的预测。

通过验证，得到了与预测一致的涨的结果。

4. 结论

4.1 本文本挖掘系统的科学性与实用性总结

本文本挖掘系统通过具有预处理网页源代码的爬虫，爬取了各大财经新闻的报导，并在统计学、新闻传播学和心理学的基础上，对文本进行处理与分析，最终通过数据可视化技术，将预测展示出来。

将预测与实际股市做了对比和分析，得到较为准确的结果。整个预测的流程如图 4-1 所示：

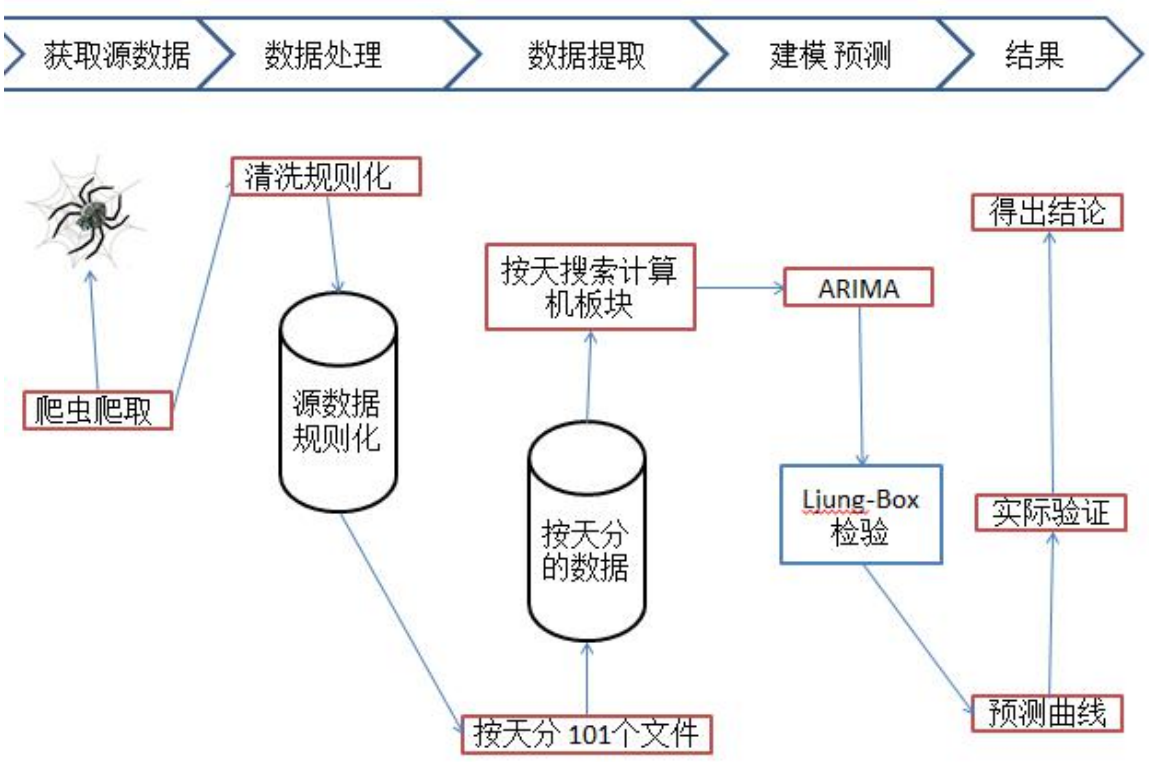

图 4-1：预测的流程图

4.2 本文本挖掘系统得出的结论

本文本挖掘系统所得出的结论有：

1） 新闻报道中积极词汇往往多于消极词汇。

2） 报道的情感的积极性与股市的涨没有正相关关系。（至于有没有负相关关系还未知）

3） 未来七天新闻报道关于计算机板块的热度值，在预测区间内，因此这种预测方法可行。

4）按 5 天（一个星期）来预测，实际的涨跌与预测的一致。而按每天预测的话则会有误差。此结论由于验证的次数少，还未证得其一般性。

5）新闻报道中，关于券商板块的报道最多，其次是计算机板块。

6）相关性假设成立，即新闻报道与该股票的涨跌有相关性。

4.3 不足与展望

虽然得到了比较准确的结果，但是本系统仍然存在不足，如：没有完成友好的交互界面；没有搭建出安全且完整的网站体系；各功能未统一为一体；财经新闻的来源不够广和深，譬如没有包括外文的财经新闻报道；

而且由于时间不够，对股票涨跌的验证只有一次，板块只选择了计算机，因此基数不够多，导致还未能够证明出其普遍性；

展望未来，可以尝试去研究预测值出现负数的原因是否是因为休假日；还可以去研究股票报道的情感“积极”性与股票的“涨”之间有无“负相关性”。

在以后的时间里，可以继续优化改进！

参考文献：

[1] 李本乾 李彩英. 财经新闻. 东北财经大学出版社，2006.7

[2] 百度百科.新浪财经. [2013.2.4].http://baike.baidu.com/link?url=WQ3uTtvUsDQWF2NI_MOKg5f-WZndGLYBBanRglCCyi7YZTTi7dAgKBWPy0bCGSE3laykNIz3XxGklsKuot4Oua

[3] Python 官网. What is Python. (2016.4.29) https://docs.python.org/2/faq/general.html#what-is-python

[4] 中国科学技术协会. 百度百科. 数据挖掘. (2016.4.30). http://baike.baidu.com/link?url=D7EGlNyELBY_wIDKVBwBW_RCT8jIbo0QYfaKWD4sttpWR4CIy9dRfjRqRtIsnK0vH4v6CxC9Gs4w8kbklRLVUq

[5] 陈茜，连婉琳. 基于文本挖掘技术的互联网股票新闻的情感分类. 中国市场. 2015 年 24 期.

[6] 王珊. 数据库系统概论[M]. 北京：高等教育出版社，2000. 302-307.

[7] 姜大志，吴志健，康立山，汤铭端，李康顺. 基因表达式程序设计的 GRCM 方法. 系统仿真学报. 2006，18(6)：1466-1468.

[8] Ruey S. Tsay(著), 王辉，潘家柱(译). 金融时间序列分析. 机械工业出版社. 2006.

[9] [美] 博克斯 等 著；王成璋 等 译. 时间序列分析：预测与控制. 机械工业出版社. 2011.

[10] [美] 埃里克·西格尔 著；周昕 译. 大数据预测：告诉你谁会点击、购买、死去或撒谎. 中信出版社. 2014.

[11] 赵华 著. 时间序列数据分析：R 软件应用. 清华大学出版社. 2016.

[12] 李勇，王文强 著. Web 程序员成功之路：Python Web 开发学习实录. 清华大学出版社. 2011.

[13] [爱尔兰] Igor Milovanovic 著；颛清山 译. Python 数据可视化编程实战 [Python Data Visualization Cookbook]. 人民邮电出版社. 2015.

[14] [美] 大卫·比斯利（David Beazley），布莱恩·K.琼斯（Brian K.Jones） 著；陈舸 译. Python Cookbook（第 3 版）中文版. 人民邮电出版社. 2015.

[15] [美] Robert I. Kabacoff 著；高涛，肖楠，陈钢 译. R 语言实战 [R in Action:Data Analysis and]. 人民邮电出版社. 2013.

[16] 张良均，云伟标，王路，刘晓勇 著. R 语言数据分析与挖掘实战. 机械工业出版社. 2015.

[17] [美] Jiawei Han，[美] Micheling Kamber，[美] Jian Pei 等 著；范明，孟小峰 译. 数据挖掘 概念与技术（原书第 3 版） [Data Mining Concepts and Techniques Third Edition]. 机械工业出版社. 2012

[18] [美] 米切尔（Ryan Mitchell） 著；陶俊杰，陈小莉 译. Python 网络数据采集. 人民邮电出版社. 2016.

[19] [印尼] 伊德里斯（Ivan Idris） 著；韩波 译. Python 数据分析. 人民邮电出版社. 2016.

[20] [德] 西蒙·蒙策尔特 等 著；吴今朝 译. 基于 R 语言的自动数据收集：网络抓取和文本挖掘实用指南. 机械工业出版社. 2016.

致谢

这次毕业论文能够得以顺利完成，并非我一人之功劳，是所有指导过我的老师，帮助过我的同学和一直关心支持着我的家人对我的教诲、帮助和鼓励的结果。我要在这里对他们表示深深的谢意!

谢谢！