倾听股吧之声，洞察大盘趋势

——互联网大数据挖掘系列专题之（三）

报告摘要:

- **互联网金融大数据强势来袭**

伴随着互联网的快速发展，信息的提供者与使用者之间的界限已经越来越模糊，而在互联网金融时代的背景下，金融信息的来源渠道也越来越丰富多样。投资者有更多的渠道来获取相关的金融信息，过去上百年的金融研究，往往局限于对以数字形式存在的数据的研究，而忽略了对非结构化的金融“数据”的研究，在当前互联网时代下，分析师的研究报告、股吧论坛帖子的信息、新闻媒体的新闻以及微博和维信等非结构化文本信息往往能够反应当前市场上投资者对股市的投资情绪，而这些信息往往又对投资者的投资决策起到潜移默化的作用。

随着互联网的高速发展，投资者越来越倾向于在网络上通过各种股吧论坛来表达自己对市场的观点，同时获取自己所关注的个股信息，因此，在股吧中投资者所发表的文字信息常常隐藏着其对个股及大盘的情绪。

- **构建股吧情绪择时指标 *GS***

使用网络文本挖掘的方法，我们抓取了淘股吧和金融界等热门股吧的股票帖子内容，并采用文本脱水、分词以及情感分析等方法得到每条帖子的“乐观”与“悲观”判断，最后结合该帖子的点击量和回复量来构建股吧情绪指标GS_t:

$$GS_t=(D_t+H_t)\times(G_t-B_t)$$

下面，基于A股情绪指标GS时间序列构造布林通道，当某日股吧情绪指标剧增并且突破上界，则发出看多信号，突破下界则发出看空信号。

布林通道上界：$GS\_UP_t=(GS_t\text{的}M\text{日均值})*(1+N\%)$

布林通道上界：$GS\_DOWN_t=(GS_t\text{的}M\text{日均值})*(1-N\%)$

当某日股吧情绪GSt剧增并且突破上轨，则发出看多信号，次日开始看多大盘指数，突破下轨则发出看空信号，次日开始看空大盘指数。

- ***GS* 指标大盘择时效果突出**

将 NQ 指标应用于沪深 300 指数择时，自 2010 年以来策略年化绝对收益 40.2%，年化超额 48%；分年度来看，策略每个自然年度均获得正的绝对收益，胜率 100%，其中 2014 年截止 5 月份累计收益 9.23%。

- **风险提示:**

股吧言论仅仅是投资者表达情绪的一种方式，且关于投资者的情绪识别具有一定的误差，因此在此基础上搭建的量化策略并不一定完备和准确。

图 1：股吧情绪择时策略原理

图 2：股吧情绪择时策略净值

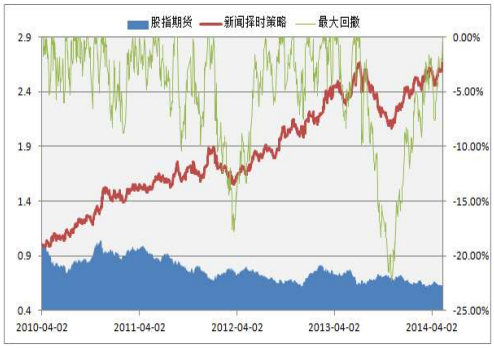


表 1. 股吧情绪择时策略实证结果汇总

年化收益率	40.20%
年化超额收益率	48.90%
胜率	51.50%
赔率	1.2

数据来源：广发证券发展研究中心

分析师： 史庆盛 S0260513070004
020875558888618
sqs@gf.com.cn

相关研究:

基于网络新闻热度的择时策略——互联网大数据挖掘系列专题之（一）	2014-06-25
公告披露背后隐藏的投资机会——互联网大数据挖掘系列专题之（二）	2014-06-26

目录索引

图表索引

一、互联网大数据挖掘体系介绍

1.1 什么是互联网大数据?

随着云时代的来临，大数据（Big data）也吸引了越来越多的关注，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。正如《纽约时报》的一篇专栏中所称，“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉。

实际上，大数据从很早以前就一直存在着，然而只是光只有数据大是没有用处的。世界上最大的数据估计和互联网一点关系都没有，今天我们所普遍关注的数据不仅仅是大，最重要的是这些大数据是以在线的形式存在了，这个恰恰是互联网的特点。所有东西在线这个事情，远远比“大”更反映本质。

像快的打车要用一个交通的数据，如果这些东西不在线，是没有用的。

又好比今天的淘宝数据和大众点评数据，因为他在线了，所以就值钱！写在磁带、写在纸上的数据，根本没有用，也没法用。

反过来讲，互联网也让数据搜集变得非常容易。过去美国谁要做总统，需要做盖勒普调查，去街上拦2000个人，在纸上打个勾，预测就很准了。现在不用做这个事情，只要在twitter上分析每个人发的东西，就可以知道总统会是谁了。

可见，互联网行业的“大数据”才是真正有使用价值和可操作性的“大数据”！这些数据的规模是如此庞大，以至于不能用G或T来衡量，因此也常常称为“巨量数据”或“海量数据”，具有数量巨大、结构复杂、类型众多等特点。那么互联网大数据到底有多大？一组名为“互联网上一天”的数据告诉我们，一天之中，互联网产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2940亿封之多（相当于美国两年的纸质信件数量）；发出的社区帖子达200万个（相当于《时代》杂志770年的文字量）；卖出的手机为37.8万台，高于全球每天出生的婴儿数量37.1万……

目前，互联网上的数据量已经从以往的TB（1024GB=1TB）级别跃升到PB（1024TB=1PB）、EB（1024PB=1EB）乃至ZB(1024EB=1ZB)级别。国际数据公司（IDC）的研究结果表明，2008年全球产生的数据量为0.49ZB，2009年的数据量为0.8ZB，2010年增长为1.2ZB，2011年的数量更是高达1.82ZB，相当于全球每人产生200GB以上的数据。而到2012年为止，人类生产的所有印刷材料的数据量是200PB，全人类历史上说过的所有话的数据量大约是5EB。IBM的研究称，整个人类文明所获得的全部数据中，有90%是过去两年内产生的。而到了2020年，全世界所产生的数据规模将达到今天的44倍。

互联网发展已有二十年，这二十年就是各行各业被互联网冲击的二十年，最先受到冲击的是媒体，然后是零售业、旅游，刚刚开始的是金融业！

受冲击行业的每一次涅槃，也是一次重生！例如：媒体行业诞生了以雅虎为代表的门户，谷歌、百度为代表的搜索引擎，Twitter为代表的社交媒体，至今这个行业的变化还在继续； 在零售业，产生了阿里巴巴、亚马逊、易贝这样的电商公司；在旅游行业则诞生了携程、去哪儿网这样的公司； 而**金融行业的故事则刚刚开始……**

1.2 互联网金融数据获取

伴随着互联网的快速发展，信息的提供者与使用者之间的界限已经越来越模糊。在互联网金融时代的背景下，金融信息的来源渠道也越来越丰富多样。投资者有更多的渠道来获取相关的金融信息，过去上百年的金融研究，往往局限于对以数字形式存在的数据的研究，而忽略了对非结构化的金融“数据”的研究，在当前互联网时代下，分析师的研究报告、股吧论坛帖子的信息、新闻媒体的新闻以及微博和维信等非结构化文本信息往往能够反应当前市场上投资者对股市的投资情绪，而这些信息往往又对投资者的投资决策起到潜移默化的作用。

采用互联网文本挖掘的方法对这些非结构化文本形式存在的金融信息进行挖掘，并从中提取出对投资决策有利的信息是该专题报告研究的重点。

图3：互联网数据来源

数据来源：广发证券发展研究中心

基于上述三类数据来源，我们将采用互联网文本挖掘技术来获取相关的信息。

“文本信息挖掘”的概念最早由Ronen Feldman博士提出，并倡导将非结构化的内容转变为有价值的商业智能行业中，即文本驱动商务智能概念。目前许多IT巨头已经纷纷在不同领域针对大数据开展了文本挖掘的项目。

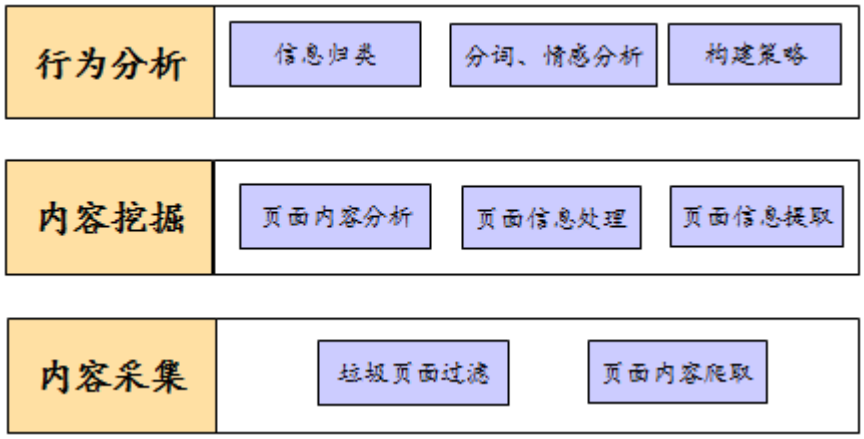
表2. 著名IT公司文本挖掘项目

公司	文本挖掘项目
谷歌	X Lab，谷歌大脑项目
微软	TextFlow项目
脸书	深度学习研究小组预测用户行为
百度	创建深度学习研究院
腾讯	成立中文处理研究室，从事文本挖掘研究

文本信息挖掘是通过分析用户数据，从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示等步骤。移动互联网数据具有数据量大、数据结构复杂、数据内容分散等特点，呈现出爆炸性增长的趋势。因此，为了从浩如烟海的数据中提取出有效信息，必须选择合适的数据挖掘策略。

信息挖掘是一个复杂的过程，需要进行大量的数据采集和运算等。按照基本功能，可以将整个信息挖掘流程划分成**内容采集**、**内容挖掘**和**行为分析**3个环节。

图4：文本信息挖掘流程



数据来源：广发证券发展研究中心

（一）内容采集

进行互联网文本数据挖掘的基础是数据的真实性和有效性，内容采集主要包括以下两个方面。

a）页面内容爬取。这是将网页的内容通过爬虫获取的部分，分析页面代码格式，进行网页代码的编码转换等，尽可能获取自己需要的信息。

b）页面垃圾过滤。页面中不可避免地会存在大量的垃圾信息，这些信息严重干扰到对后期信息挖掘的准确性，页面垃圾过滤机制会找出包括广告在内的段落以及其他对内容挖掘无效的部分，并将其清除，不进入内容挖掘部分。

（二）内容挖掘

主要是对需要的特定信息进行提取，该阶段处理后的文本数据是后期进行分词、情感分析的基础;

（三）行为分析

整个文本挖掘过程的重点和难点是分词和情感分析，对于**海量文本信息**而言，程序的处理速度也是至关重要的一点。

1.3 互联网大数据抓取体系

对于**大规模的互联网信息抓取**而言，单线程的数据获取是一项非常局限的事情，因此多线程、分布式的信息抓取平台是必须搭建的。因此数据的抓取平台的搭建是一项基础性的工作。

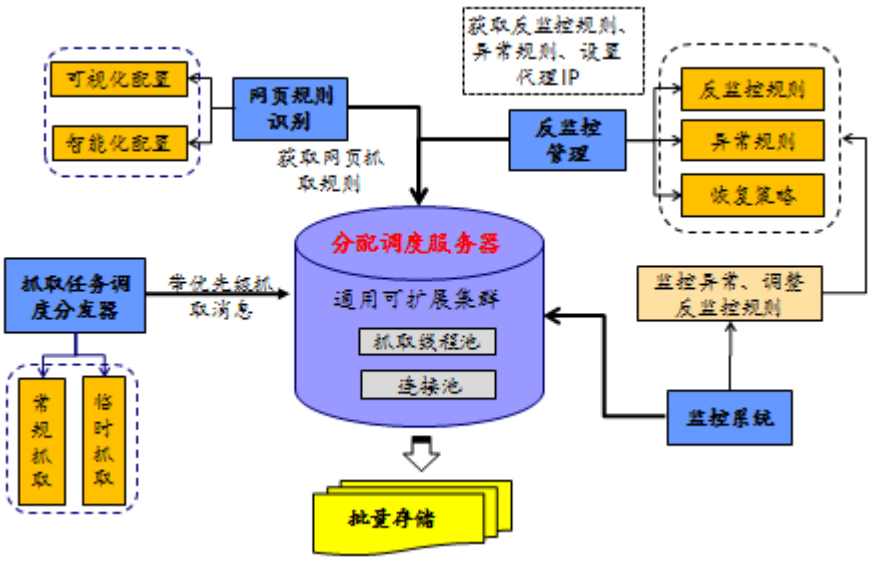
整体上，我们的大数据抓取平台可分为三部分:

首先是搜索热门网站，对需要提取信息的网站的网页编码格式、网页制作规则进行研究，提取出需要提取的信息的网页编码格式;

第二是对需要提取信息的网站进行分布式配置，多线程爬取特定的信息，并将信息存储到指定的数据库中;

第三，监控数据爬取过程，防止网站数据异常带来的数据抓取缓慢等问题，主要是防止对指定网站的频繁访问导致IP被限制访问的问题。

图5：互联网数据抓取体系



数据来源：广发证券发展研究中心

上图刻画了在大规模的互联网数据爬取过程中的整体框架，对于单个网站的股吧论坛的数据抓取框架可以简化成以下几个步骤：

1) 首先对需要批量爬取信息的网页结构进行研究，主要是研究需要提取的特定信息的网页格式进行研究，这步骤可以借助在IE浏览器下打开需要提取信息的网页，然后调用出开发人员工具，找到需要批量提取的特定信息的网页代码规则，或者利用Firefox浏览器下的XPATH工具以及View Source Chart工具；

2) 找到不同个股的股吧论坛url之间的关系，利用url之间的关系，抓取所有个股的股吧url；

3) 由于股吧抓取的数据量往往非常庞大，出于效率的考虑，往往采用分布式、多线程的方式进行，若条件允许需要同时几十部甚至上百部电脑多个线程地抓取需要的内容。此时，就需要一台主程序电脑控制其他电脑的程序运转；在数据抓取过程中由于网络不稳定、电脑突然死机等问题，需要对抓取的过程进行监控，记录下每个进程运行的过程，防止数据抓取的缺漏，同时处于对网站安全的考虑，需要实时监控程序的运行，防止网络访问限制问题，设置适当的断线重连机制，IP切换机制等；

4) 对于抓取到的信息统一批量存储到指定的数据库中，构成后续分析的底层的数据库。

二、股吧数据结构及特点

随着互联网的高速发展，投资者越来越倾向于在网络上通过各种股吧论坛来表达自己对市场的观点，同时获取自己所关注的个股信息，因此，在股吧中投资者所发表的文字信息常常隐藏着其对个股及大盘的情绪。

而股吧信息具有其独有的特征，其挖掘也面临着更多的困难。

2.1 股吧数据特点

（一）热门股吧

淘股吧、金融界、东方财富网等;

（二）数据类型

通过对大规模的文本类型数据的挖掘，获取到股民对个股涨跌情绪指标，从而获取到对相应行业指数、大盘指数涨跌预期，构建相应的策略，跟踪个股、行业指数、大盘指数涨跌;

（三）数据特点

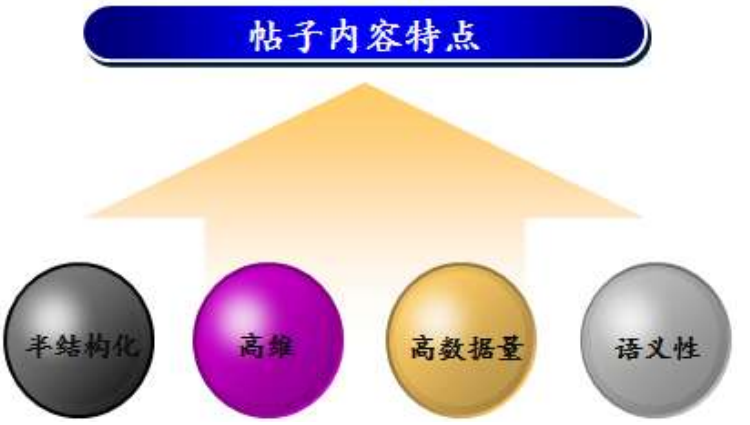
（1）半结构化：文本数据既不是完全无结构的也不是完全结构化的。例如文本可能包含结构字段，如作者、长度、发帖时间、回复量、点击量等，也可能包含大量的非结构化的数据，如摘要和内容。

（2）高维：文本向量的维数一般都可以高达上万维，一般的数据挖掘、数据检索的方法由于计算量过大或代价高昂而不具有可行性;

（3）高数据量：一般的文本库中都会存在最少数千个文本样本，对这些文本进行预处理、编码、挖掘等处理的工作量是非常庞大的，因而手工方法一般是不可行的

（4）语义性：文本数据中存在着一词多义、多词一义，在时间和空间上的上下文相关等情况。

图6：股吧数据特点



数据来源：广发证券发展研究中心

2.2 股吧数据结构

下面以“金融界”的股吧论坛为例来解析常见的股吧论坛网页结构：

（一）个股股吧列表

在金融界股吧论坛上，有存储个股股吧列表的主网址，网址为：***http://istock.jrj.com.cn/forumlist.shtml***，在该主网址上能够提取到当前市场上所有个股在该网站上的股吧论坛的网址，该主网址的结构主要可以分为：沪市A股、深市A股、沪深300、港股、主题吧等板块。根据需要提取的信息，可以分别提取某一板块上的个股股吧论坛链接。

（二）个股股吧论坛链接：

从主网址上能够提取到所有的个股的股吧论坛对应的链接，每个个股都有对应的唯一的股票代码的股吧链接，例如平安银行的个股股吧论坛链接为：http://istock.jrj.com.cn/list,000001.html。因此每个个股股吧的url结构为：

http://istock.jrj.com.cn/list,股票代码.html。

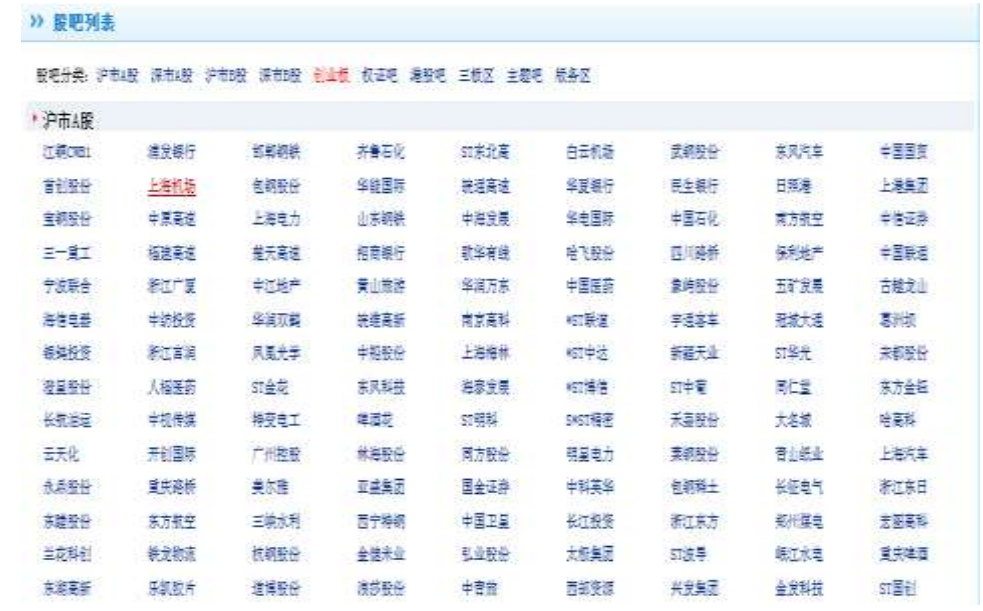
（三）个股股吧论坛页数链接：

因为每个个股股吧论坛的历史信息量非常大，因此个股股吧论坛历史信息存储在不同页数上，例如平安银行截止至2014年6月20日，一共有1444页的个股论坛信息，帖子数为144342条。因此想要获取到历史的个股股吧信息还需要获取到每一页的股吧论坛链接，例如平安银行的股吧上第二页帖子对应的链接为http://istock.jrj.com.cn/list,000001,p2.html，因此每个个股股吧帖子对应页数的url结构为http://istock.jrj.com.cn/list,股票代码,p帖子页数.html

（四）个股帖子对应链接：

在每一页的个股论坛帖子上能够获取到每个帖子的点击量、回复量、帖子标题、发帖者等相关信息。此外，如果想要获取到每个帖子的内容以及回复内容，还需要知道每个帖子对应的帖子链接，这个可以通过每个帖子在网页结构中的href属性直接获得该帖子对应的链接地址。

图7：金融界股吧主网址示例



数据来源：广发证券发展研究中心

图8：金融界个股股吧首页（平安银行示例）

数据来源：广发证券发展研究中心

图9：金融界个股股吧帖子内容（平安银行示例）

数据来源：广发证券发展研究中心

就仅仅从金融界个股股吧论坛的帖子数量角度上看，可以看出数据量非常庞大。例如截止至2014年6月20日，平安银行的历史帖子页数一共有1444页，帖子量为144342条，如果单单计算金融界股吧主网址上的沪市A股以及深市A股的股票量一共有2489只。平均每个个股的帖子总量为25000条，总用的帖子量就用62225000条帖子。这些帖子量还是不包括一些主题吧的帖子的数量，因此仅从金融界论坛上，个股的股吧帖子数量规模上看，数据量还是非常巨大的。

如果将比较热门的网站，例如淘股吧、东方财富网、新浪财经以及和讯网上的个股股吧帖子信息采集下来的话，仅仅是帖子的数量规模就是亿级别的数量级。如果再加上帖子的点击量、回复量、发表时间等信息，数据量就更加庞大了。以上仅仅是从数量上看出帖子的规模非常庞大，如果从存储的大小的角度上看，也可以看出，单单是存储也是需要非常大的空间的。一个汉字占两个字节，1KB就是1024个

字节，即512个汉字就能存储1KB大小。这样的话，假设每个帖子内容、回复量、点击量、回复内容、帖子标题等信息加起来平均有1kb的大小，大约有60G左右的数据量，这仅仅金融界上的沪深A股的个股股吧的数据量大小，而且是截止至2014年6月20日的估计，这些热门财经网站，每天的访问量以及发帖量也是非常巨大的。所以对于股吧内容的文本信息的挖掘要求非常高，想要在短期提取相关热门网站上所有个股的股吧信息，需要对程序的设计有高的要求，必须采取分布式、多线程等高效的方法才能完成短期内大量数据的抓取，这也对电脑的配置等硬件提出了高的要求。

三、股吧数据挖掘及策略构建

3.1 核心技术

股吧数据的挖掘及策略的构建主要包含了一下三个部分的核心技术：

图10：股吧情绪指标构建核心技术

数据来源：广发证券发展研究中心

（一）数据获取阶段

数据获取阶段是进行后续分析的基础阶段。

对于热门的网站，譬如说金融界、和讯网、淘股吧、东方财富网等，这些网站每天的访问量非常巨大。例如，东方财富网的股吧论坛，据统计，平均每天发的新贴量超过了40000个，而每天的在线人数达到了数百万之多。因此，如果要抓取全市场个股在若干个热门网站的股吧论坛上过去的所有帖子，考虑到效率问题，就必须采取分布式、多线程的抓取模式。而抓取到的数据量非常大，目前，数据的存储仓库主要采用SQL Server数据库，而个股的存储信息包括，帖子标题、点击量、回复量、帖子内容、发帖时间、帖子作者、回复内容、回复时间等维度。

数据抓取过程实现主要是通过JAVA语言实现的。由于对热门网站的频繁访问，出于网站的安全性考虑，网站往往会对频繁访问的IP等相关信息进行记录，从而限制访问。这样的话，就降低了抓取的效率。一般的处理方法是通过延时的方法，但是这样也是降低了效率问题，治标不治本。比较有效的方法是通过代理IP的方法，设置一定量的代理IP，切换IP的方法来实现突破频繁访问限制，这也是我们未来改进的方向之一。

（二）数据处理过程

数据处理过程是进行策略构建的关键阶段。

数据提取完后，首先需要将数据进行预处理，将其中的“垃圾”信息，去除掉，例如网页代码。这样可以降低“垃圾”信息对后续分析的影响，达到“降噪”的作用。数据预处理完后，可以将信息进行分类，一类为数值型信息，例如帖子的发表时间、帖子的回复量、点击量、回复时间等，一类为文本型信息，例如帖子的标题、帖子内容、回复内容等。数值型的信息的挖掘相对比较简单，而对于文本型信息的挖掘，则是当前学术、业界研究的难点。

而对于文本型信息的挖掘也是本专题报告研究的重点。对于文本型信息的挖掘，我们主要采用的是语义分析的方法。语义分析的大概流程为：首先对帖子标题、内容、帖子内容、回复内容等进行分词，因为金融文本信息的特殊性，想要达到较理想的文本信息挖掘效果，在文本分词上需要一些额外的信息。分词的词库需要足够庞大，尽量能够包含所有个股可能会用到词语，例如名词、形容词、副词以及相对应的行业用语；基于分词后的结果，再基于情感分词的方法进行情感的判别。而分词的常见模型有隐马模型和N最短路径等，也可以借助比较成熟的分词系统来实现分词功能，如IKAnalyzer等。在情感分词方法上，我们分别采用了关键词匹配算法，后续我们还会尝试更先进的方法，如朴素贝叶斯估计算法以及支持向量机的算法等。数据处理过程的实现也主要采用JAVA语言实现。

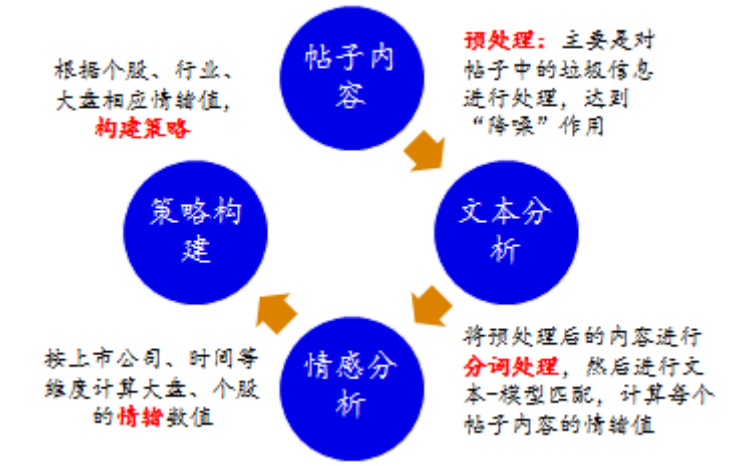
（三）策略构建

策略构建是进行文本信息挖掘的目的。

在完成数据处理后，个股的信息主要包含两类，一类是基于数值型信息挖掘的结果，一类是基于文本型信息挖掘的结果。对于个股相关帖子的点击量、回复量的变化可以构建个股的热度指标，结合个股某一时段或者当日的帖子的整体的情感值的变化，可以构建个股某一时段或者当日的看涨、看跌指标。进而可以结合相关行业的个股，构建行业热度、行业看涨、看跌指标，进一步获取到大盘的热度、看涨、看跌指标。而基于看涨、看跌指标的变化，可以构建个股的择时、行业择时、择股策略。策略的构建实现的理论基础为现代投资组合理论，我们的策略实现手段主要采用JAVA、MATLAB以及VBA语言实现。

3.2 实现步骤

图11：股吧情绪指标构建步骤



数据来源：广发证券发展研究中心

（一）文本数据抓取

以金融界股票论坛为例：

流程解读：

1、首先分析需要提取的帖子内容的架构。

金融界股票论坛的帖子内容的提取可以分为三个层次。首先根据主网址，得到当前市场上各个个股在论坛上的股吧网址，其次根据个股的股吧网址，得到个股每天的发帖内容、相应帖子的点击量、回复量、以及帖子标题，最后根据帖子的标题链接，得到帖子的发帖时间、帖子内容以及回复内容等相关信息。

2、具体的实现手段：

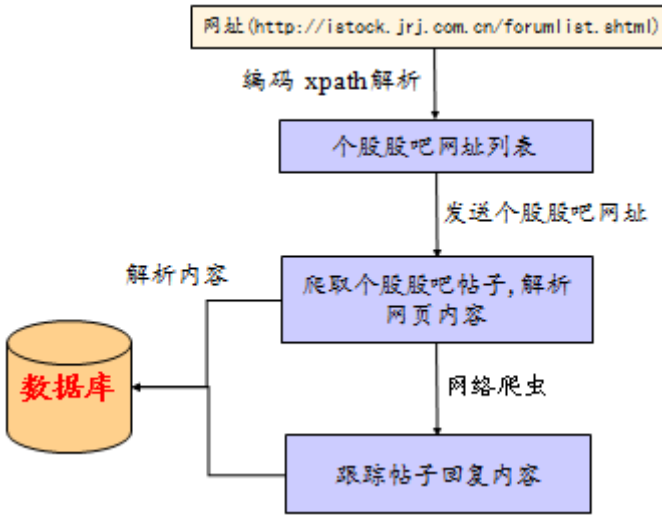
1) 设置股吧的主网址：http://istock.jrj.com.cn/forumlist.shtml

2) 根据主网址对应的个股提取出每个个股在金融界的股吧对应的链接，这里采用的是深度优先的方法，即每次先提提取每个个股股吧链接，然后进入个股股吧，提取个股股吧的相关信息。

3) 根据2中得到的个股股吧网址，首先提取每个帖子的标题，回复量、点击量、帖子作者等相关信息，然后再根据每个个股的帖子标题提取出对应的链接。

4) 根据3中得到的个股帖子标题对应的链接，得到帖子标题对应的帖子内容、发帖时间、回归内容等相关信息。

图12：金融界论坛股吧数据抓取流程



数据来源：广发证券发展研究中心

图13：帖子内容提取范例

stockid	author	title
sz000001	波浪王1	魏被带走时，家中发现上亿现金
sz000001	未飞的候鸟	亏的生活不能够继续了 求包养
sz000001	brown	大盘在做无量反抽
sz000001	花开花谢	黄奇帆：向海外转移财产应征税。应重税
sz000001	zlxuan9915	今天大盘破2000！
sz000001	yuandufushu	那些年被贱卖的银行股
sz000001	大众情人	大蓝筹起来了，0001打头，倒挂最多的是000338潍柴动力
sz000001	越好越来	从一瓶啤酒的价值分析平安银行熔发的影响
sz000001	yzh10	有必要澄清的三大市场模糊认识
sz000001	郭手	周小川：逐步允许境外公司境内发行股票
sz000001	yfclang	11月1日：做全多仓位92%
sz000001	卖电子书	10月23日 市场要闻 公告信息 收集！！！
sz000001	山魂永存	银行股破冰之旅，带领大盘远航。
sz000001	东莹	每日龙头与今日总结——东呈团队
sz000001	中原鹿正肥	光大事件暴露——期货T+0 现货T+1 的重大缺陷！
sz000001	职业炒手99	声明 本ID 和炒手 无任何瓜葛 贴个账户希望能模
sz000001	云端攀猴	中国经济“上、下限”
sz000001	漫舞烟花	从这一单开始，我满怀憧憬
sz000001	跳跳虫	声讨兴业银行管理层，无视股东的利益，你们是谁的狗？！
sz000001	一万点	期货，一边是天堂，一边是地狱。
sz000001	1996入市	友情提醒：大盘反弹到位，3月大跌！！！
sz000001	Velaciela	2245点的中国股市——十年一场大梦！
sz000001	acman	证监会：6月底完成财务核查之后IPO才会重启
sz000001	山南水北	以民生银行为代表的很多银行其实很垃圾

数据来源：广发证券发展研究中心

基于提取后的内容，我们再对相关维度进行后续的处理、分析。

（二）词库维护及分词处理

分词效果的好坏很大程度上决定了后续进行情感分析效果的好坏。金融文本信息多样化，内容形式多样化，这给分词带来了一定的难度。因此如果需要让分词的效果尽量好，词库的构建是至关重要的一步。根据每个个股所在行业以及与个股相关的名词、副词、动词、否定词等以及各个行业的专有名词等构建分词需要的词库。基于词库的信息，然后再根据各个个股所在的行业的不同，调用不同的行业词库，对个股的帖子标题、内容等信息进行分词。因数据量庞大，分词过程需要耗费大量的时间，在进行实际的分词时，采用多线程方式的处理，提高效率。

我们搭建的自定义金融词库包含“乐观”和“悲观”两大类词汇共**约1000个**，并且随着我们所采集样本的不断增多，词库也在逐渐扩充。

图14：自定义金融词库范例

乐观					
担任	即将	世界第一	翻倍	入选	共同进步
鼓掌	共识	资源整合	埋伏	技术服务	实将
股权	达标	优质	香悖悖	合伙人	世界第一
特等奖	动力	划时代	很好	提升	资源整合
更适合	继续	超预期	发扬光大	获益匪浅	优质
宣战	招标	点燃	点睛	必能	划时代
引入	藏龙卧虎	感受	精彩	联盟	超预期
可获	特色	更适合	大吼	出炉	点燃
宣战	更大	宣战	成长型	远景	感受
远大于	变化	引入	满足	兼备	便宜
耀眼	科学技术	可获	丰收	批准	入场
发言	供应量	必会	获得	威慑	可期
自我改造	获颁	远大于	期望	报送	点亮
维持	打造	耀眼	冲天	退税率	爽
先锋	相机	发言	成长	实质	反超
必会	关注	自我改造	相信	符合	展示
维持	先锋	盈利	概念	风骚	标志性

悲观					
磨损	被忽视	空虚	失败	新低	最重
婊子	更难	惊悚	提心吊胆	谴责	死缓
骗子	障碍	爆炸案	关灯	被封	冤
拖后腿	出货	窦娥	茫茫	分配不公	最糟
无法自拔	丢失	担忧	遭遇	吓人	破产
死定了	被判	伪科学	悲情	爆炸性	沦陷
变相	潜能	竟然	骗人	隐患	腐烂
没有反应	套用	杀跌	撤离	吐槽	莫名其妙
失效	担心	道德风险	跌跌	清仓	无稽之谈
矛盾	没道理	苦寒	操盘	惨痛教训	妖精
隐藏	创新低	夭折	适可而止	垃圾	坑爹
埋葬	套了	隐蔽	行不通	报忧	怀疑
不务正业	切忌	低谷	茫然	潜质	恐
套现	最冷	更惨	悲哀	崩塌	失眠
严重	死不	大顶	哭	腐败	砸低
孰不可忍	猥琐	委屈	初审	下降	惨不忍睹
吃不上	惨烈	大逆转	减仓	没赚	砸坑

数据来源：广发证券发展研究中心

基于构建好的词库信息，对每个个股历史帖子信息，包括帖子标题、帖子内容、回复内容等相关维度进行分词处理。分完词后，然后对分词后的每个帖子的标题、内容、回复内容等进行情感上的判别，主要判别出帖子反映的信息是乐观的还是悲观的。

图15：论坛帖子分词案例

sz000542	cosmking	tcl|智能|电视|车|联网|锂电池|智能|穿戴|网络|教育|数字|医疗
sz000542	华山论股	13年|净利润|大涨|165|两市|最低价|的|智能家居|智能|电视|电视|游戏|移动|智能|终端|概念
sz000542	晴月光	我|狂|买|智能手机|股|的|理由
sz000542	赛亚	大|商|股份|的|未来|会|怎样|06年|历史|帖
sz000542	来的来去的去	1月|15日|可|穿戴|设备|委员会|wdc|正式|成立
sz000542	热点财经	同|洲|电子|发布|手机|操作系统|960|欲|挑战|ios|宏|卓
sz000542	卖电子书	1月|08日|nbsp|市场|要闻|nbsp|公告|信息|nbsp|收集
sz000542	卖电子书	11月|11日|nbsp|市场|要闻|nbsp|公告|信息|nbsp|收集
sz000542	来的来去的去	tcl|旗下|alcatel|在|英国|发售|最|廉价|手机|nbsp|仅|5|英镑
sz000542	millanbob	我|不知道|他|牛|不|牛|但他|的|客户|不是|一般|的|牛
sz000542	三点	都|跟|李|东|生|学|tcl|大型|企业集团|搞|曲线|tpo|美的|电器|控股|股东|美的|集团|拟|以
sz000542	化城宝所	华强北|康盘
sz000542	卖电子书	6月|7日|nbsp|市场|要闻|nbsp|公告|信息|nbsp|收集
sz000542	tianxiayuxue	港股|tcl|通讯|与|a股|的|tcl
sz000542	卖电子书	12月|7 |nbsp|市场|要闻|nbsp|公告|信息|nbsp|收集|nbsp|nbsp|nbsp|nbsp|nbsp|nbsp|
sz000542	娱乐为主	太|赫兹|地沟油|检测|概念股|一览
sz000542	北京小杜	10月|24日|回顾|太|赫兹|概念股
sz000542	jj711	太|赫兹|地沟油|检测|概念股|一览
sz000542	闲逛中	隔墙|看|物
sz000542	czw3388	太|赫兹|概念|上市公司
sz000542	翻番股	最|便宜|的|太|赫兹|概念股|tcl|000100|抢
sz000542	uncle9	今天|炒作|的|新|题材|太|赫兹|技术|概念|个股
sz000542	s320	港股|tcl|多媒体|tcl |通讯|暴涨|nbsp|nbsp|为|什
sz000542	ares0256	14日|复|盘|作业
sz000542	luzg	王老吉|谁弟|崛起|背后|的|策划|内幕|nbsp

数据来源：广发证券发展研究中心

（三）情感分析

情感分析是进行后期策略构建的关键一步，策略构建的效果的好坏很大程度上取决于情感分析结果的质量的好坏。中文语义博大精深，对中文语义进行情感分析是当前学术界、业界研究的重点，然而目前并没有非常理想的解决方案。本专题的情感分析方法采用关键词匹配法，此外，当前情感分析较成熟的方法还有支持向量机以及朴素贝叶斯估计等方法，在后续研究中我们将尝试结合其他方法来进行文本

情感分析。

图16：情感分析步骤（关键词匹配法）

1、对分词后的帖子主题查找对应的情感词，记录消极还是积极及位置；

2、查找程度词，为程度词设置权值，乘以情感值；

3、查找否定词，若数量为奇数，乘以-1，若为偶数，乘以1；

4、判断分词结尾是否有感叹号，有叹号则往前寻找情感词，有则相应的情感值加上一个权重值；

6、计算并记录所有帖子主题、内容情感值，存进相应数据库；

7、计算个股、行业、大盘指数当天情绪指标；

数据来源：广发证券发展研究中心

图17：单个句子情感分析案例

这|股票|的|前景|极|好，成长性|也|比较|不错。

1、情感词：好、不错，积极情感+2；
2、程度修饰词：极、比较，“极”程度较深，赋予权重值4，“比较”程度较浅，赋予权重值为2，所以情感分值为1*4+1*2=6；
3、否定词：在词库中没有匹配到指定的否定词；
4、判断分词后结尾是否有感叹号：没有匹配到感叹号；
5、区分句子情感：根据匹配到的词语以及句子情感，该个帖子情感分值为6，情感倾向为乐观；

数据来源：广发证券发展研究中心

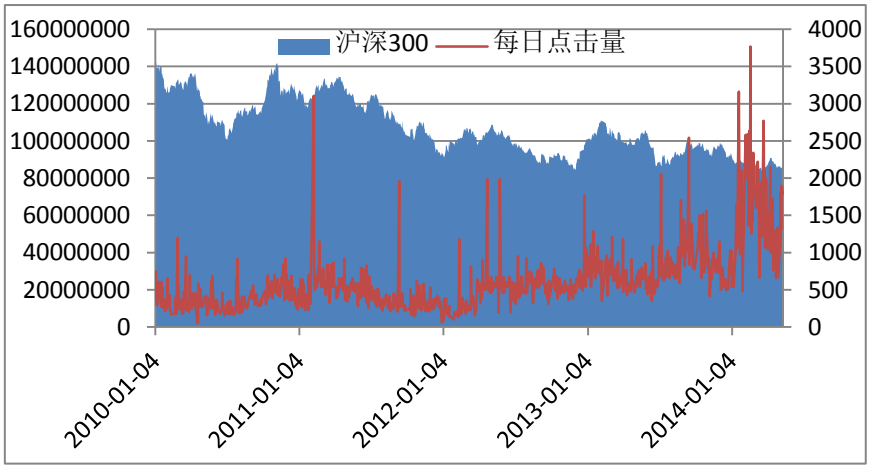
基于关键词匹配的算法，可以得到每个个股每个帖子的情感值，并基于情感值的相对大小，得到每个个股每个帖子的乐观、悲观情绪指标。该算法主要利用的是匹配搜索的原理，根据特定词语的不同权重得到帖子的情感值。

图18：某日股吧帖子情感分析案例

代码	作者	帖子标题	态度
sh600579	lry168	行情|仍将|犹犹豫豫|前行	悲观
sh600853	Q家738533979	600853|龙|建|股份|26日|早|盘|将|高|开|获利|5|以上	乐观
sh600697	大海蛇	商业|股票|将|飙升	乐观
sh600462	股海戏金蟾	600462 |今日|摘|帽|跌停|每股|收益|1.13元|市盈率|5.5倍	悲观
sh600697	daimeizi66	浙|江东|日|期待|第|十个|涨停	乐观
sh600697	三班长	基金|扎堆|的|票|票	乐观
sh600579	寻牛	有|只|股|2|个月|没|跌了	乐观
sh600793	职业流氓	澄清|公告|超导|截念|下周|一|全部|跌停	悲观
sh601177	挣钱很难的	第一次|碰到|跌停|牌|了	悲观
sz002714	咸蛋操人	拿什么|拯救|nbsp|我|跌停|的|a股	悲观
sz000953	kengaoxm	这样|的|垃圾|公司|投资者|是不是|应该|赔到|钱|st|河|化|预|盈|变|预|亏|股价|跌停|r	悲观
sz000542	华山论股	13年|净利润|大涨|165|两市|最低价|的|智能家居|智能|电视|电视|游戏|移动|智能|终端|	乐观
sz000660	森林焱炎	明天|大涨	乐观

数据来源：广发证券发展研究中心

图19：股吧帖子点击量变化

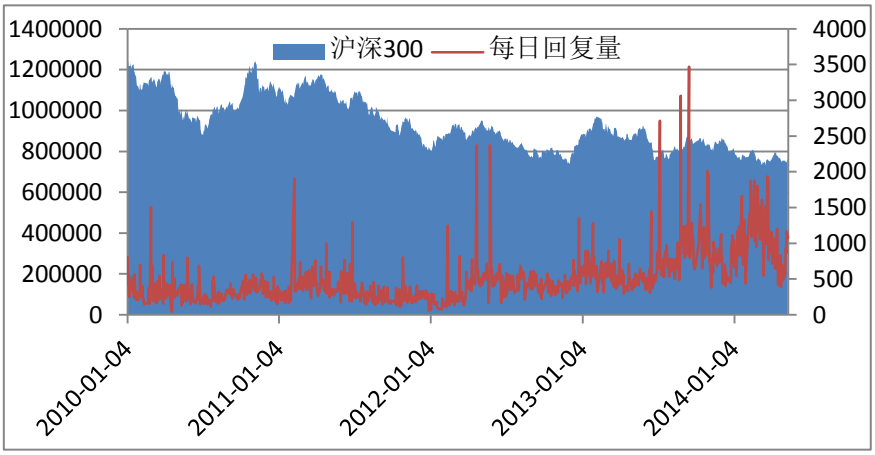


数据来源：广发证券发展研究中心

上图显示，截止2014年5月份，上述几大股吧中每日的帖子点击总量在5-8千万左右，其中大多数帖子均为零回复，每日帖子总回复量约40万，为点击量的1%，可见大多进入股吧的网名都是抱着打酱油的心态来随便看看而已，遇到感兴趣的帖子题目或观点就会点进去，但如果帖子毫无意义，或者某些标题党写的帖子内容空无新意，则鲜有回复。

因此，我们在分析帖子是否隐藏投资者情绪信息的同时，也将格外地注重该帖子的点击量和回复量，点击量尤其是回复量较高的帖子，意味着帖子观点受到更为广泛的关注甚至认可，因此该帖子所隐藏的信息比如得到有效的放大。

图20：股吧帖子回复量变化

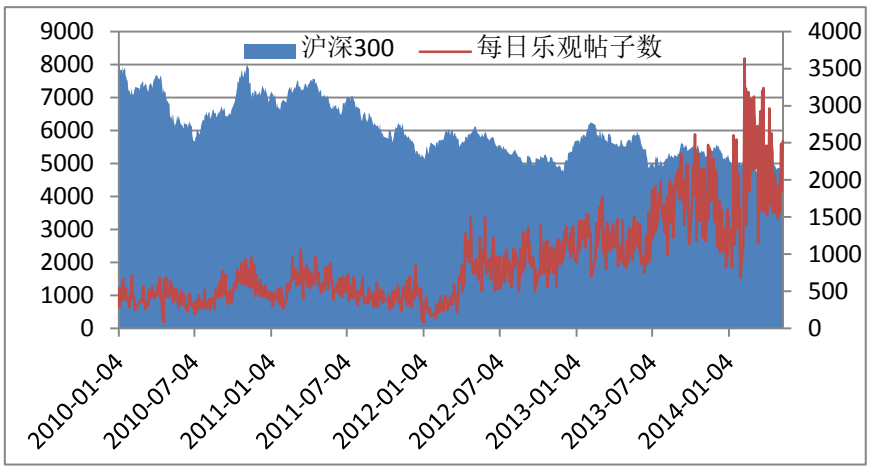


数据来源：广发证券发展研究中心

下面我们再来对比每日帖子的乐观和悲观数量情况，下图显示，帖子内容识别为乐观的数量远远高于识别为悲观的数量，这显然也是可以理解的，持有股票的人自然会更加关注相应的标的，从而也更加倾向于在论坛上发表有利于自己持有股票的言论，而已经卖出股票的投资者则更容易沦为围观者，默默等待下一次低点买入的机会。

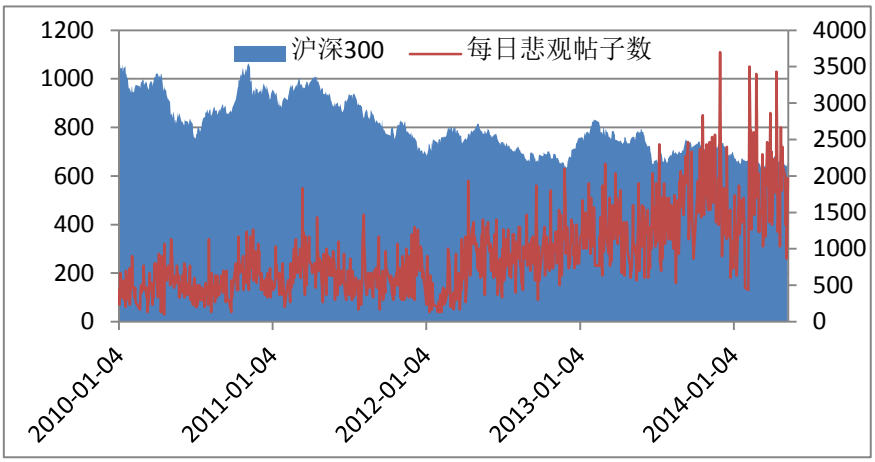
当然，两者数量之间的差别并不会影响我们基于该数据来构建投资者股吧情绪指标（Guba Sentiment，简称GS），因为我们更加关注的是该情绪指标短期内的变化趋势。

图21：乐观帖子数量变化



数据来源：广发证券发展研究中心

图22：悲观帖子数量变化



数据来源：广发证券发展研究中心

下面我们分别基于股票的帖子点击量、回复量、乐观帖子数以及悲观帖子数来构建股吧情绪指标GS_t：

假设，

股票S在t日，

点击量为：D_t　回复量为：H_t

乐观帖子数为：G_t　悲观帖子数为：B_t

则构造股票S在t日的情绪指标：

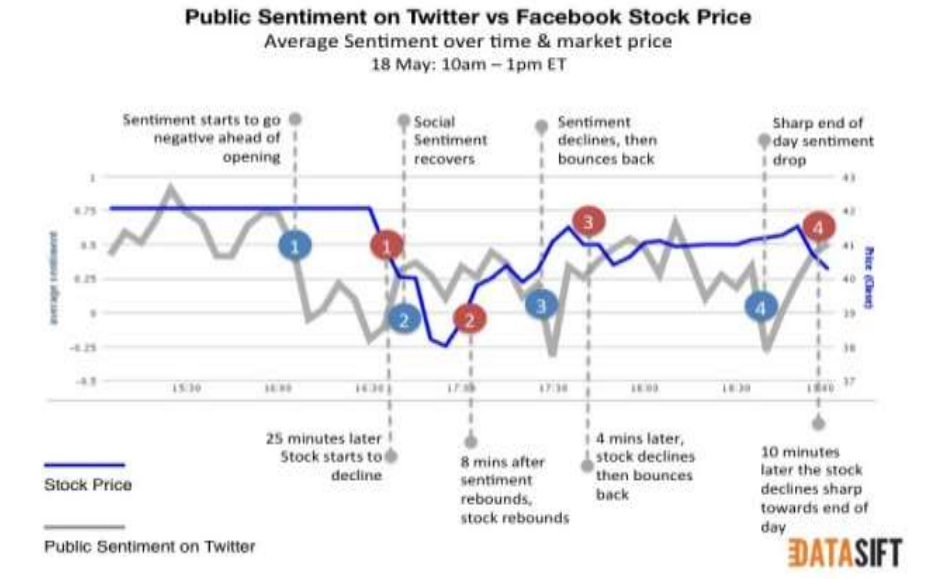
$$GS_t=(D_t+H_t)\times(G_t-B_t)$$

3.3 策略构建

上述我们从热门股吧论坛中采用文本挖掘的方法抓取到了一些大数据，并借此构造了反映投资者情绪的指标，那么具体应该如何将这些大数据以及相应的指标应用到金融投资中呢？

我们先来看一个有趣的案例：在Facebook（脸谱） IPO当天，著名的社交媒体监测平台DataSift在开盘前对Twitter上的投资者对Facebook的股价预期产生的大数据进行情感倾向监测，并将大数据分析的结果用于预测当日Facebook股价波动。结果显示，在Facebook开盘前Twitter上的情感显著转向负面，而25分钟之后Facebook的股价便开始大幅下跌。而随着Twitter上的情感转向正面时，Facebook股价在8分钟之后也开始了回弹。最终当股市接近收盘、Twitter上的情感转向负面时，10分钟后Facebook的股价又开始下跌。最终的结论是：Twitter上每一次情感倾向的转向都会影响Facebook股价的波动。

图23：Twitter情感分析预测Facebook IPO走势



数据来源：广发证券发展研究中心, 36Kr

上述Twitter基于大数据对Facebook股价的预测案例显然是成功的，其利用投资者在论坛上的言论进行情绪指标构造的方法也值得我们借鉴和学习。

上述我们已经同样基于热门股吧论坛，对投资者的投资观点进行了挖掘和分析，并得到每个个股每个帖子的情感值，对个股某个阶段或者当日的所有帖子的情感值进行处理，可以得到个股在某个阶段或者当日的乐观、悲观情绪，从而得到行业、大盘在某个阶段或者当日的投资者的整体乐观、悲观情绪，并根据这些情绪指标构建策略。

股票S在t日，

点击量为：D_t　　回复量为：H_t

乐观帖子数为：G_t　　悲观帖子数为：B_t

则构造股票S在t日的情绪指标：

$$GS_t=(D_t+H_t)\times(G_t-B_t)$$

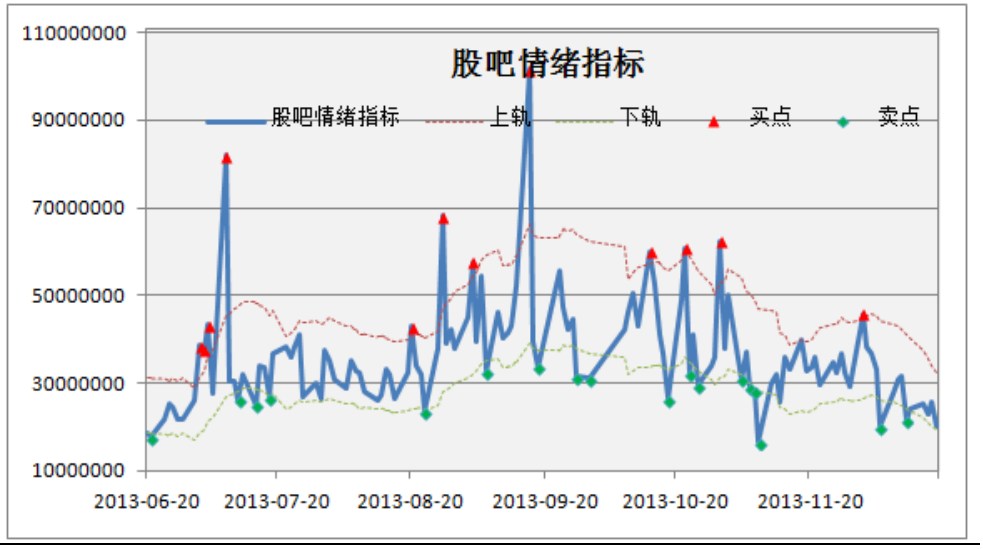
下面，基于A股情绪指标GS时间序列构造布林通道，当某日个股却情绪指标剧增并且突破上界，则发出看多信号，突破下界则发出看空信号。

布林通道上界：$GS_UP_t=(GS_t$的M日均值$)*(1+N\%)$

布林通道上界：$GS_DOWN_t=(GS_t$的M日均值$)*(1-N\%)$

当某日股吧情绪GS_t剧增并且突破上轨，则发出看多信号，次日看多大盘指数，突破下轨则发出看空信号，次日看空大盘指数。

图24：策略构建原理



数据来源：广发证券发展研究中心

四、实证分析

4.1 数据说明

样本区间：2010/4-2014/5。

个股数据：全部A股每日股吧信息。

择时标的：以沪深300指数作为大盘指数进行择时，采用股指期货进行模拟交易。

开平仓价格：若当日信号为多/空，则于开盘价开多/空仓，当日收盘价平仓；若当日信号为平，则不开仓。

交易费用：按照股指期货交易费用，双边万分之2。

策略参数：M=10；N=30。

4.2 实证结果

股吧情绪指标GS对沪深300指数的择时效果如下所示：

图25：股吧情绪择时策略净值

数据来源：广发证券发展研究中心

表3. 股吧情绪择时策略实证结果汇总

年化收益率	年化超额收益率	胜率	赔率	累计最大回撤
40.2%	48.9%	51.5%	1.20	22.3%

数据来源：广发证券发展研究中心

策略自2010年4月份以来，取得了约160%的绝对收益，年化40%，最大累积回撤发生在2013年底，另一次较大回撤发生在2012年3月底，约17%，其余时间策略的最大累积回撤基本都在10%以内。

平均日度胜率仅约52%，但日度赔率1.2倍，可见股吧情绪指标一旦预测正确则获得的收益高于预测错误所遭遇的亏损幅度。

策略分年度表现如下表所示，5年均获得正收益，年度胜率100%。

表 4. 股吧情绪择时策略年度表现

	2010	2011	2012	2013	2014(截止 5/12)
收益率	45.50%	26.30%	18.61%	11.84%	9.23%

数据来源：广发证券发展研究中心

五、总结

在互联网大数据以及互联网金融时代背景下，信息量的快速增加为投资者进行投资决策提供了越来越多的信息。随着文本信息的快速增加，人们也开始注意到文本型信息对投资决策的重要性。但是，如何对海量金融文本信息进行挖掘，提取出

对投资决策有用的信息，是当前投资者面对文本型信息的难点。当前，越来越多的量化型投资基金也开始对文本挖掘表示较高的关注，如何对海量的文本信息进行挖掘，提取出有用的信息，也是当前量化型基金面对的难题。

本专题报告对热门股吧论坛的个股的历史帖子信息进行研究，利用文本挖掘技术，对帖子信息进行挖掘，并利用情感分析的结果构建了量化策略。历史回测结果表明，该策略的收益也是相当可观的。

风险提示

股吧言论仅仅是投资者表达情绪的一种方式，且关于投资者的情绪识别具有一定的误差，因此在此基础上搭建的量化策略并不一定完备和准确。

广发证券—行业投资评级说明

买入: 预期未来 12 个月内，股价表现强于大盘 10%以上。

持有: 预期未来 12 个月内，股价相对大盘的变动幅度介于-10%～+10%。

卖出: 预期未来 12 个月内，股价表现弱于大盘 10%以上。

广发证券—公司投资评级说明

买入: 预期未来 12 个月内，股价表现强于大盘 15%以上。

谨慎增持: 预期未来 12 个月内，股价表现强于大盘 5%-15%。

持有: 预期未来 12 个月内，股价相对大盘的变动幅度介于-5%～+5%。

卖出: 预期未来 12 个月内，股价表现弱于大盘 5%以上。

联系我们

	广州市	深圳市	北京市	上海市
地址	广州市天河北路 183 号大都会广场 5 楼	深圳市福田区金田路 4018 号安联大厦 15 楼 A 座 03-04	北京市西城区月坛北街 2 号月坛大厦 18 层	上海市浦东新区富城路 99 号震旦大厦 18 楼
邮政编码	510075	518026	100045	200120
客服邮箱	gfyf@gf.com.cn			
服务热线	020-87555888-8612			

免责声明