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）

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。

A 股市场每日同样有海量的个股新闻信息，这些信息能反映投资者当前对股市的情绪。 如何选择合适的新闻来源及有效变量将极大的影响投资者进行有效的投资决策。采用互联网文本挖掘的方法对财经新闻热度进行挖掘，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。

- **构建 A 股新闻热度择时指标 NQ**

鉴于个股的新闻量能够一定程度上反映网络媒体以及投资者对相应股票的关注程度，因此我们将某日个股的全部新闻总量（News quantity）定义为该日A股市场的股票新闻热度指标NQ，并基于该指标的变化规律来建立A股的择时策略。

经统计个股每日新闻热度增长率与次日大盘指数收益率的相关系数为6.45%。基于此我们猜测当A股新闻热度较高的情况下，大盘次日上涨的概率会更高，因此基于A股热度NQ时间序列构造了布林通道:

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

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

当某日个股新闻量剧增并且突破上轨，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。

- **NQ 指标大盘及行业择时效果突出**

将 NQ 指标应用于沪深 300 指数择时，自 2011 年以来策略年化绝对收益 34.80%，年化超额 51%，分别各有 1/3 的时间策略处于看多、看空及坎平状态；若将 NQ 指标应用于行业指数择时上，同样有突出的表现，尤其在传媒、机械设备及建筑建材等行业。

图 1：新闻热度择时策略原理

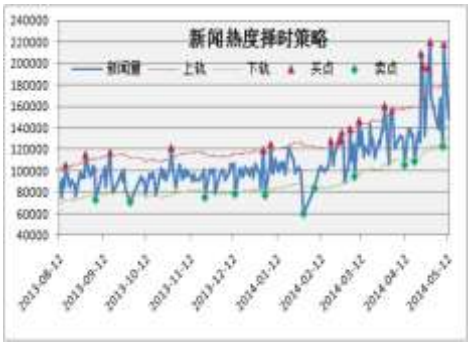


图 2：新闻热度择时策略净值

表 1.新闻热度择时策略表现汇总

年化收益率	37.03%
年化超额收益率	45.87%
胜率	53%
赔率	1.14
累计最大回撤	20%

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:

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”才称得上是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。

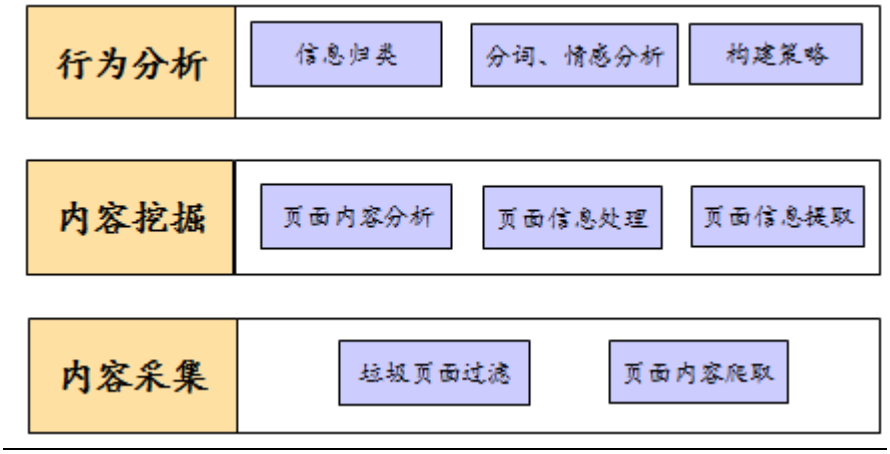
表2. 著名IT公司文本挖掘项目

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：

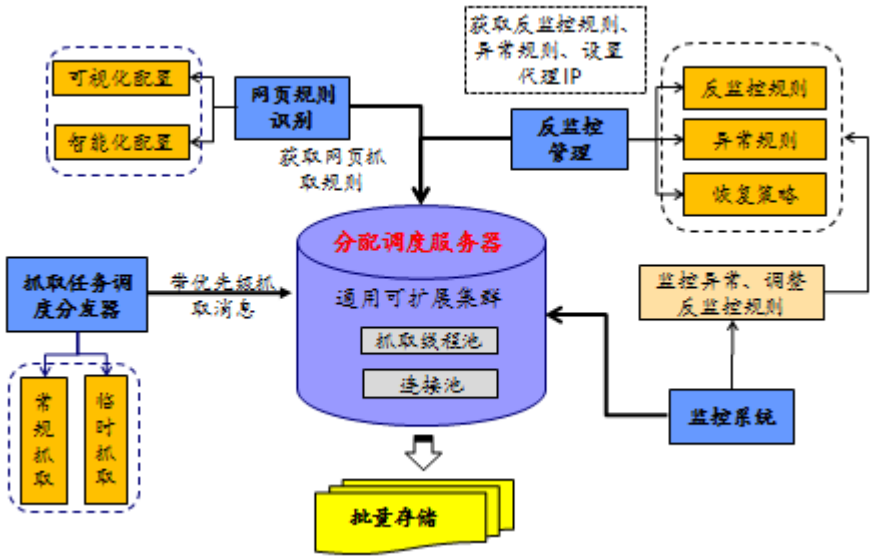
首先是搜索热门网站，对需要提取信息的网站的网页编码格式、网页制作规则进行研究，提取出需要提取的信息的网页编码格式；

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) 找到不同个股的网页url之间的关系，利用url之间的关系，抓取所有个股的网页url；

3) 由于网页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、新闻热度数据来源及抓取

互联网实现了全球的信息共享与交互，使得信息的采集以及传播的速度和规模都达到空前的水平，信息量增长的速度远比人类理解的速度要快，并以海浪式四面八方涌入人类的生活。我们每天所接收到的互联网信息量都在以几何级别的增长，这是一个彻底的“信息大爆炸”时代！

A股市场每日同样有海量的个股新闻信息，这些信息能反映投资者当前对股市的情绪。 如何选择合适的新闻来源及有效变量将极大的影响投资者进行有效的投资决策。

信息渠道：百度爬虫工具、各大财经门户

技术要求：Python / VBA / MATLAB+分布式+多线程+IP欺骗技术等

数据特点及信息价值：非结构化金融数据、半结构化数据，对新闻热度量化挖掘，获取有用价值

2.1 数据来源

评判某个个股某一日的新闻热度的方法有很多，譬如在一些几个热门财经网站上的相关新闻的点击量、浏览次数或者是在相关热门财经网站上的相关新闻的报道数。而如果仅仅局限于几个财经论坛的相关新闻报道的数量或者点击量，这样无疑局限了数据的来源的广泛性。因此在数据的来源上应该尽可能获取到当前市场上的某个个股在某一日的相关新闻的总量，最主要的问题是能够尽可能多地获得该个股在某一日所有的热门新闻网站的相关信息，因此在方法上有两种思路：1、尽可能地搜索到所有的新闻财经网站，然后再在这些网站上搜索个股过去某一个或者是某一段时间的搜索量。这种方法的缺点是效率低，需要识别大量网站的编码规则后然后再提取相关个股的新闻信息，这导致抓取的工作量大，时间长，尤其是想获取个股过去的历史信息时；2、借助搜索引擎的相关新闻搜索功能，只要借助的搜索引擎足够强大，就能搜索到所有的财经上所有的相关个股的财经新闻，这减少了收集财经网站的时间，既加快了效率也增加了数据量。因为搜索引擎的搜索原理一般为采用分布式数据库技术来存储网页缓存，因此能够到相关的个股在各大财经网站上的历史新闻。因此搜索引擎的选择就显得至关重要。当前，我们收集到的有特定的新闻搜索功能的搜索引擎新浪财经新闻搜索、百度新闻搜索、360新闻搜索等等。考虑到新闻搜索的功能的个性化以及新闻搜素量的全面性，我们采用百度搜索引擎的新闻搜索功能，抓取主流财经门户新闻数量。

2.2 新闻搜索引擎结构

百度搜索引擎的新闻搜索功能中的高级搜索功能中，主要包含三大块的功能：搜索结果设定、搜索时间段设定、以及新闻来源的相关设定。搜索结果的设定能够个性化设置需要搜索的内容，例如可以设置搜索结果包含的关键字：全部包含的关键字、任意一个关键字、不包含关键字，同时能够设置关键字是位于新闻的标题还是新闻全文等。在设置搜索的时间段上，也能够个性化地设置需要搜索的时间段，包含历史全部时间以及个性化到某一个时间段或者是某一日的新闻量。在新闻源上

也能够个性化设置新闻的来源，例如说人民网、新浪网等。在弄清楚百度的新闻搜索架构后，在采用网络爬虫时候，还需要弄清楚两点：第一为针对不同个股的新闻搜索的网址规律即url的规律，或者是通过程序模拟鼠标操作的方法，直接获取到设定搜索词后的网站；第二为根据查询的网址返回的搜索信息找出需要提取的信息的网页代码规律。

2.3 数据结构

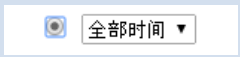
如（二）中关于百度新闻搜索的引擎结构阐述的，想要获取到特定的信息，第一步先对网址结构进行分析，找出不同搜索的url的内在规律。例如在百度新闻搜索中的高级搜索网址: http://news.baidu.com/advanced_news.html中的搜索结果框的包含以下全部关键词中输入"平安银行"，在时间中框中选择2014-6-21到2014-6-21，关键词框中点击在新闻全文中，其他选项保持默认，点击百度以下，在网页加载完毕后可以看到，返回的网页url为:
http://news.baidu.com/ns?from=news&cl=2&bt=1403366400&y0=2014&m0=6&d0=22&y1=2014&m1=6&d1=22&et=1403452799&q1=%C6%BD%B0%B2%D2%F8%D0%D0&submit=%B0%D9%B6%C8%D2%BB%CF%C2&q3=&q4=&mt=0&lm=&s=2&begin_date=2014-6-22&end_date=2014-6-22&tn=newsdy&ct1=1&ct=1&rn=20&q6=

通过不同的关键词搜索的测试，可以发现返回的不同的url之间的内在规律。例如固定字段为：http://news.baidu.com/ns?from=news&，固定的参数名称分别为：cl、bt、y0、m0、d0、y1、m1、d1、et、q1、submit、q3、q4、mt、lm、s、begin_date、end_date、tn=newsdy、ct1、ct、rn、q6等，而不同的参数之间根据不同的输入，采取不同的值。因为百度新闻搜索引擎网址中的中文字符采用的编码格式为GB2312的，因此在模拟操作获取数据时，需要将输入的部分信息进行编码转换，例如q1、q3、q4、submit等后面连接的参数值，例如"平安银行"转换为GB2312后为%C6%BD%B0%B2%D2%F8%D0%D0，而submit参数后值为固定的字符值，将%B0%D9%B6%C8%D2%BB%CF%C2转换为utf-8编码后可以看出为“百度一下”这四个中文字符。

具体参数说明如下表:

表 3. 网址 URL 参数说明一览

前提	参数	取值	说明
	from	news	
	submit	百度一下	
	q6	空值	
	cl	2	
	q1	转换后的 GB2312 编码	包含以下**全部**的关键词对应的输入字符
	q3	转换后的 GB2312 编码	包含以下**任意一个**关键词

	q4	转换后的 GB2312 编码	**不包含**以下关键词
	s	1	表示选中这个 
	s	2	表示选中这个 
s=1	mt	0	全部时间
s=1	mt	24	最近一天
s=1	mt	168	最近一周
s=1	mt	720	最近一月
s=1	mt	8640	最近一年
s=2	begin_date		开始日期
s=2	end_date		结束日期
s=2	bt	(1970-01-01 到 begin_date 的秒数) - 28800	
s=2	et	(1970-01-01 到 end_date 的秒数) - 28800 + 86399	
s=2	y0	begin_date 的年	
s=2	m0	begin_date 的月	
s=2	d0	begin_date 的日	
s=2	y1	end_date 的年	
s=2	m1	end_date 的月	
s=2	d1	end_date 的日	
	tn	newsdy	在新闻全文中
	tn	newstitledy	仅在新闻的标题中
	ct1	1	按焦点排序
	ct1	0	按时间排序
	rn	10	每页显示 10 条
	rn	20	每页显示 20 条
	rn	50	每页显示 50 条
	rn	100	每页显示 100 条
	lm		当 s=1 时跟 mt 的值一样，当 s=2 时为空值
	ct		1-ct1

数据来源：广发证券发展研究中心

在找到不同url之间的规律后，就可以根据url中不同参数之间的规律构成搜索的网

址，利用程序模拟浏览器操作，获取需要得到信息的页面，然后再分析相应的网页信息，从中提取出需要的信息。本专题考虑的是个股新闻热度指标，因此可以考虑某一时段或者是当日个股在各大财经网站上出现的新闻报道的数量，从响应后的网址可以得到个股的搜索结果，例如，搜索"平安银行"在2014年6月1日至6月21日的新闻量，从新闻搜索中返回的界面中可以看到在这个时间段内共有10600篇相关的新闻量。

图6：新闻数据抓取首页

数据来源：广发证券发展研究中心

近年来随着互联网的普及，关于个股的财经新闻业与日俱增，截止最近，每日关于A股股票的财经新闻多达4万条左右。

图7：新闻数据抓取内容

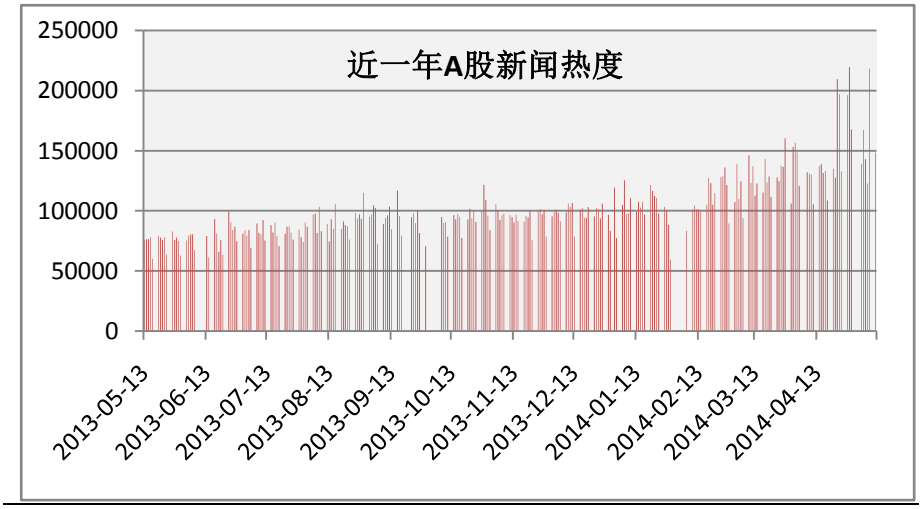
数据来源：广发证券发展研究中心

三、新闻热度指标及策略构建

互联网实现了全球的信息共享与交互，使得信息的采集以及传播的速度和规模都达到空前的水平，信息量增长的速度远比人类理解的速度要快，并以海浪式四面八方涌入人类的生活。我们每天所接收到的互联网信息量都在以几何级别的增长，这是一个彻底的“信息大爆炸”时代！

近年来随着互联网的普及，关于个股的财经新闻业同样与日俱增，截止2014年，每日关于A股股票的财经新闻（仅统计部分主流媒体）保守估计为4万条左右。

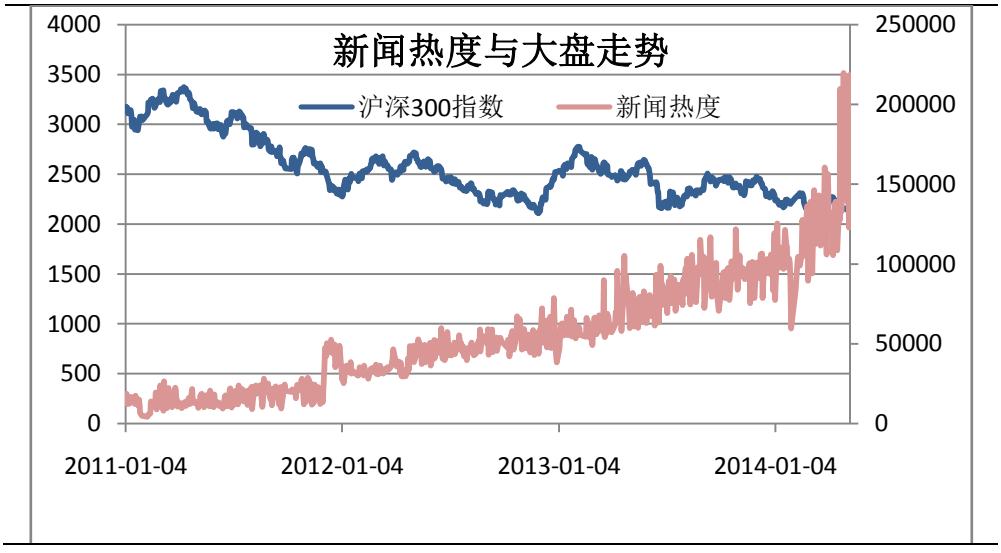
图8：新闻热度近期变化



数据来源：广发证券发展研究中心

鉴于个股的新闻量能够一定程度上反映网络媒体以及投资者对相应股票的关注程度，因此我们将某日个股的全部新闻总量（News quantity）定义为该日A股市场的股票新闻热度指标NQ，并基于该指标的变化规律来建立A股的择时策略。

图9：新闻热度与大盘走势



数据来源：广发证券发展研究中心

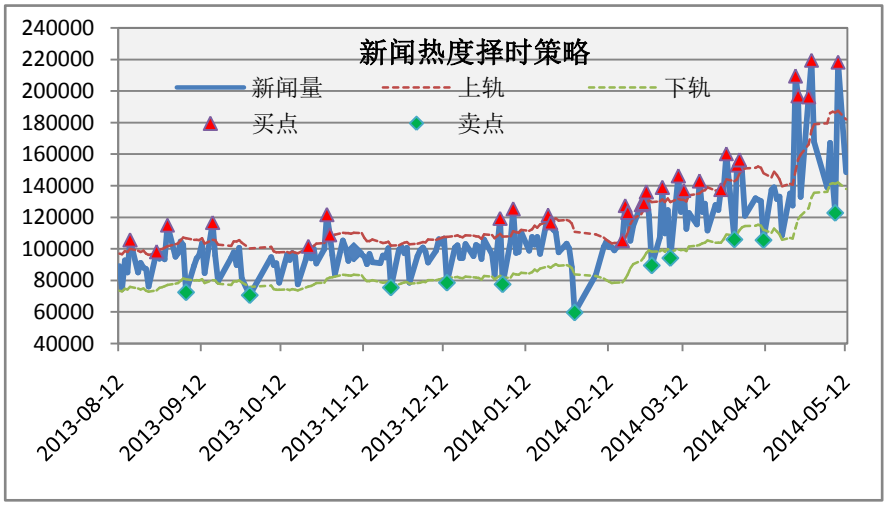
上图为2011年以来沪深300指数的走势与A股新闻热度对比图，经统计个股每日新闻热度增长率与次日大盘指数收益率的相关系数为6.45%。基于此我们猜测当A股新闻热度较高的情况下，大盘次日上涨的概率会更高，因此基于A股热度NQ时间序列构造了布林通道：

布林通道上界：NQ_UP_t=(NQ_t的M日均值)*(1+N%)

布林通道上界：NQ_DOWN_t=(NQ_t的M日均值)*(1-N%)

当某日个股新闻量剧增并且突破上轨，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10：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：2011/1/4-2014/5/12。

个股数据：沪深300指数成分股每日新闻量。

择时标的：沪深300指数日频数据。

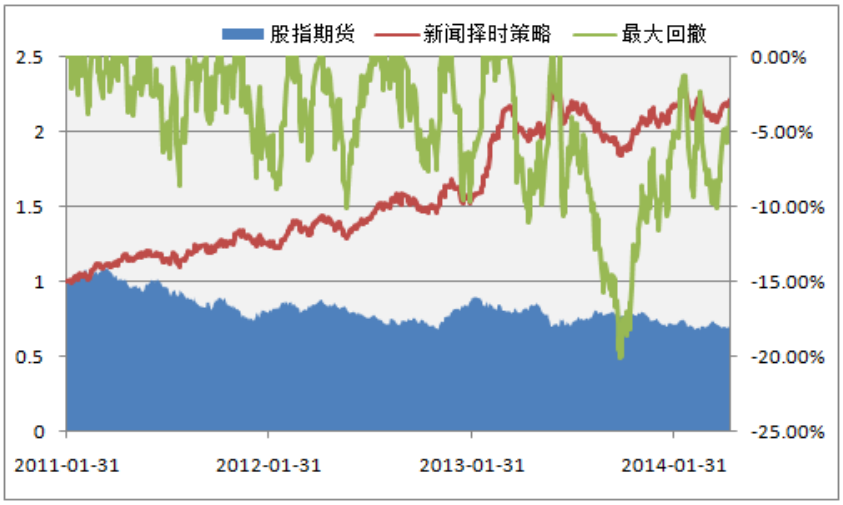
开平仓：若当日信号为多/空，则于开盘价开多/空仓，同时将上次信号平仓。

策略参数：M=20；N=10。

4.2 实证结果

新闻热度指标NQ对沪深300指数的择时效果如下所示：

图11：新闻热度大盘择时策略净值



数据来源：广发证券发展研究中心

表4. 新闻热度大盘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
37.03%	45.87%	53%	1.14	20.14%

数据来源：广发证券发展研究中心

策略自2011年以来，取得了约120%的绝对收益，年化37%，最大累积回撤发生在2013年11月份，回撤幅度约20%，其余时间策略的最大累积回撤基本都在10%以内。

平均日度胜率约53%，日度赔率1.14倍，可见新闻热度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，4年均获得正收益，胜率100%。

策略分年度表现如下：

表5. 新闻热度大盘择时策略年度表现

	2011	2012	2013	2014(截止 5/12)
收益率	28.0%	26.7%	26.8%	8.0%

数据来源：广发证券发展研究中心

将相同的策略用于行业指数择时效果如何呢？下面采用行业新闻热度对申万一级行业共28个进行择时测试，结果如下：

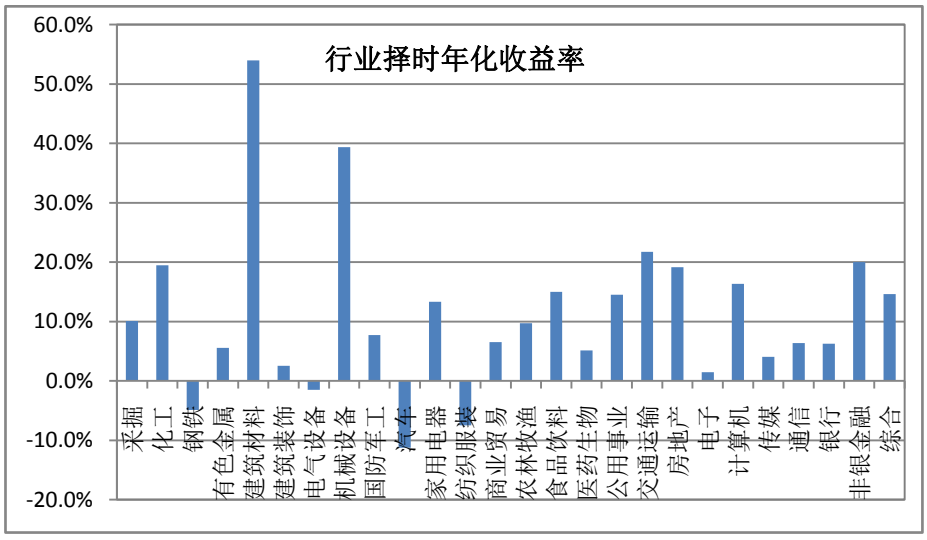
表6. 新闻热度行业择时策略表现汇总

行业名称	年化收益率	年化超额收益率	胜率	赔率	累计最大回撤
采掘	10.1%	23.4%	51.1%	1.06	23.4%
化工	19.5%	30.9%	50.4%	1.14	25.2%

钢铁	-4.9%	10.1%	48.7%	1.02	25.9%
有色金属	5.6%	22.9%	51.5%	1.00	36.0%
建筑材料	**53.9%**	**69.4%**	**0.5**	**1.15**	**24.3%**
建筑装饰	2.5%	16.9%	49.3%	1.07	33.4%
电气设备	-1.5%	15.1%	49.8%	1.02	36.6%
机械设备	**39.4%**	**55.9%**	**50.2%**	**1.20**	**23.1%**
国防军工	7.7%	25.4%	52.0%	0.99	34.0%
汽车	-11.2%	-2.4%	46.7%	1.06	51.3%
家用电器	13.3%	16.2%	52.2%	1.01	22.9%
纺织服装	-7.5%	2.9%	50.0%	0.96	44.3%
商业贸易	6.5%	18.7%	50.1%	1.06	24.3%
农林牧渔	9.7%	21.3%	50.5%	1.06	45.3%
食品饮料	15.0%	27.2%	51.4%	1.05	46.1%
医药生物	5.1%	8.1%	49.9%	1.06	44.9%
公用事业	14.5%	19.6%	52.9%	1.01	21.2%
交通运输	**21.7%**	**34.3%**	**54.0%**	**1.00**	**27.7%**
房地产	19.1%	25.4%	51.1%	1.08	33.9%
电子	1.5%	7.3%	48.4%	1.10	37.9%
计算机	16.3%	19.7%	50.8%	1.07	35.9%
传媒	**4.1%**	**-11.4%**	**49.7%**	**1.06**	**37.7%**
通信	6.4%	12.8%	51.4%	1.00	36.8%
银行	6.2%	14.2%	50.7%	1.05	27.8%
非银金融	20.0%	29.8%	50.2%	1.12	30.0%
综合	14.6%	21.9%	52.4%	1.01	24.8%

数据来源：广发证券发展研究中心

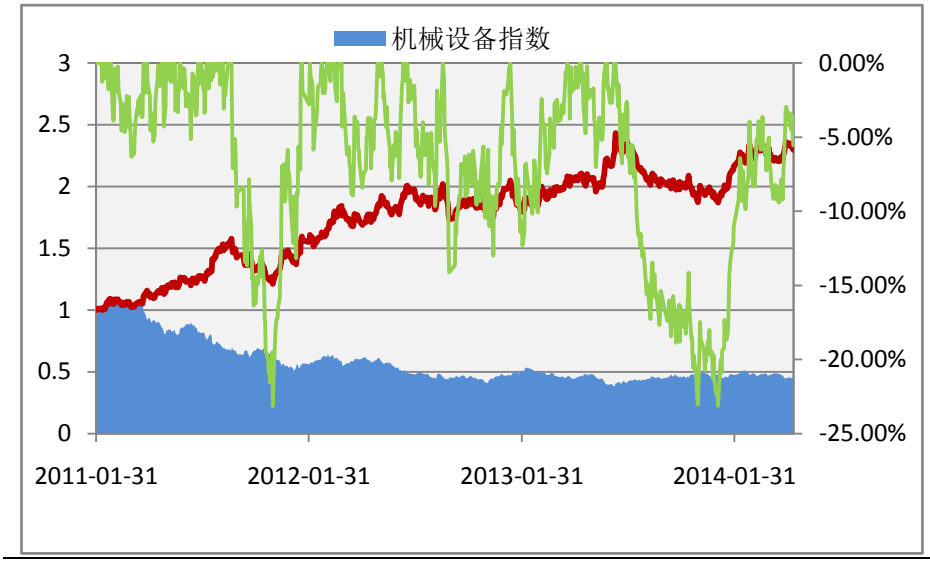
图12：新闻热度行业择时策略表现



数据来源：广发证券发展研究中心

以机械设备及建筑材料行业为例，下面为应用新闻热度指标对该两个行业进行择时的效果：

图13：新闻热度“机械设备”行业择时效果



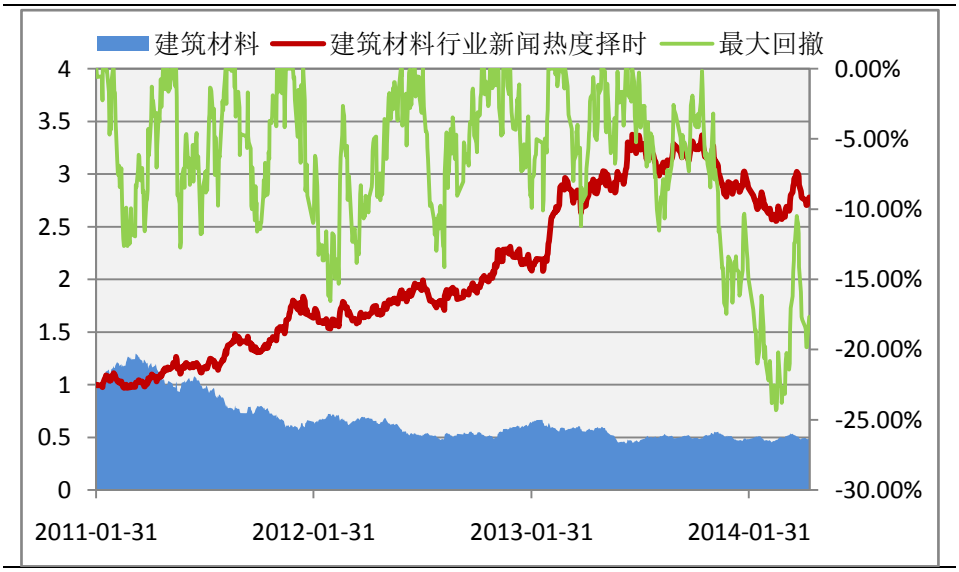
数据来源：广发证券发展研究中心

表7. 新闻热度“机械设备”行业择时效果汇总

年化收益率	年化超额收益率	胜率	赔率	累计最大回撤
29.6%	64.5%	50.2%	1.2	18.2%

数据来源：广发证券发展研究中心

图14：新闻热度“建筑材料”行业择时效果



数据来源：广发证券发展研究中心

表8. 新闻热度“建筑材料”行业择时效果汇总

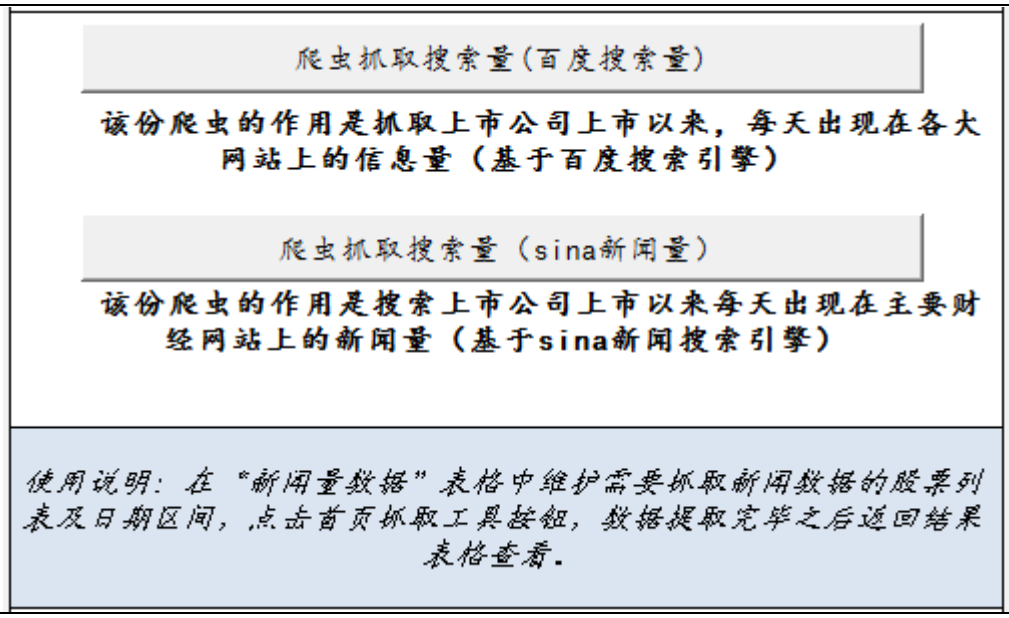
年化收益率	年化超额收益率	胜率	赔率	累计最大回撤
29.9%	21.6%	51.8%	1.15	22.9%

数据来源：广发证券发展研究中心

4.3 工具推介

俗话说："工欲善其事，必先利其器"。数据的获取是进行后期分析的基础阶段，数据的质量好坏很大程度地决定了后期进行策略构建效果的优劣。基于对百度新闻搜索引擎以及新浪网新闻搜索引擎的研究以及VBA、HTTP以及HTML的相关技术知识，我们开发出了个股百度新闻搜索量获取以及新浪新闻搜索量查询的工具，该工具能够查询上市公司在上市以来每天出现在各大财经网站上的新闻量。目前，该工具处于继续完善阶段，主要完善的方向为更加个性化的查询功能，例如查询特定时间段内的搜索量、将搜索到的文本结果批量提取并存储特定的数据库或者根据用户需要保存为特定的文本格式等。

图15：A股新闻热度抓取工具展示



数据来源：广发证券发展研究中心

若您对该抓取工具感兴趣，欢迎来电与我们交流，或者发邮件向我们索取。

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。投资者关注的信息来源越来越丰富多样，信息的提供者与使用者的界限也越来越模糊。而影响投资决策的因素也越来越复杂，近年来，随着行为金融的发展，股市中个股的新闻“热度”也越来越引起人们的注意。如何衡量个股的受关注程度是当前行为金融领域研究的重点，而A股大盘及行业的受关注程度则影响其未来走势。

本专题报告在互联网大数据背景下，利用互联网个股的新闻量的变化构建了A股热度指标，并构建了大盘及行业量化择时策略。历史回测结果表明，该策略的收益也是相当可观的。

后续我们将继续推出其他互联网大数据挖掘系列专题，将分别从上市网络公告及热门股票论坛等角度挖掘可能存在的投资机会，敬请关注！

风险提示

新闻热度仅仅是投资者投资情绪的一个方面表现，而A股的走势受到各种综合因素共同影响。

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