大数据技术在股票投资上的崛起

基金公司合作互联网公司指数点评

报告摘要:

- **大数据股票指数布局**

上周两家基金公司分别联合互联网机构发布了即将推出指数的新闻，该指数在之后可能会应用在被动管理的基金产品上。具体的编制细节在新闻中并未透露，但以新闻提供的信息分析，两家公司应该均会将互联网公司提供的大数据分析结果应用在传统的量化选股模型中。以指数化的形式进行基金投资及日常管理。在弱市及传统选股因子的超额收益逐渐弱化的条件下，资产管理行业结合互联网企业利用互联网技术探寻基于互联网技术的新的策略因子。

- **互联网技术带来投资研究方式的进步**

大数据的时代来临，给投资研究工作带来了新的体验。虽然并未改变投资和销售的业态，但对之后的投资研究业态带来了进步、对数据及投资策略的分析提出了更高的要求。在投资研究中，大数据带来了覆盖面广、刻画准确、分析高效等三个明显优势，令原有研究框架得到深化，并覆盖更多投资研究领域。

- **互联网技术在投资上的尝试**

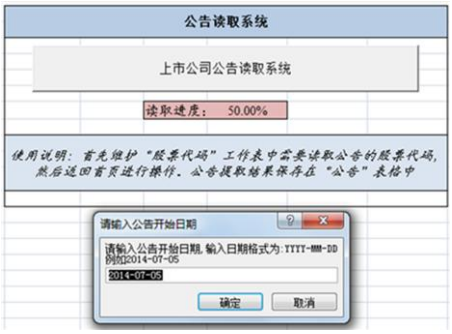
广发证券金融工程小组搭建了一套完善的互联网文本挖掘体系，通过对互联网财经新闻、上市公司公告资讯以及股吧论坛等不同网页数据采取结构分析、文本爬取、数据清理、结构分词以及情感分析等操作，得到对大盘指数、行业板块以及个股未来走势具有参考意义的信息。

- **核心假设风险:**

本文仅对互联网技术对于投资研究的影响做讨论，未提供任何投资建议。

图 1 百度互联网金融指数

图 2 广发公告提取小工具



分析师: **马普凡** S0260514050001
021-60750623
mapufan@gf.com.cn
分析师: **史庆盛** S0260513070004
020875558888618
mapufan@gf.com.cn

相关研究:

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）	2014-06-27
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
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
互联网金融的模式与产品:机构产品，化繁为简	2014-06-24

目录索引

图表索引

一、互联网公司布局股票指数

1.1 两家基金联合互联网企业准备指数

“来自指数业、互联网和金融行业的三家巨头，中证指数公司、百度和广发基金将利用各自优势进行合作，打造了新的互联网金融产品。新产品将利用广发基金在股票研究和量化投资端的优势，在百度大数据搜索因子的基础上，建立科学、稳健的策略指数选样模型。依靠目前国内规模最大、服务最全、最具市场影响力的中证专业指数服务公司运作，之后的产品将陆续在该平台上发布。”----和讯财经

“南方基金和新浪联合推出的财经大数据策略指数将基金公司专业股票研究优势与互联网“大数据”结合，在南方基金量化投资研究平台的基础上，通过对新浪财经“财经大数据”定性与定量分析，找出股票热度预期、成长预期、估值提升预期与股价表现的同步关系，构建策略因子，精选出具有超额收益预期的股票，构建、编制并发布策略指数。从而填补该领域指数市场空白，继而在后期开发跟踪该指数的基金产品，并以互联网为营销渠道进行推广和发行，为基金投资者创造收益价值。”----新浪财经

在上周两家基金公司分别联合互联网机构发布了即将推出指数的新闻，该指数在之后可能会应用在被动管理的基金产品上。具体的编制细节在新闻中并未透露，但以新闻提供的信息分析，两家公司应该均会将互联网公司提供的大数据分析结果应用在传统的量化选股模型中。以指数化的形式进行基金投资及日常管理。在弱市及传统选股因子的超额收益逐渐弱化的条件下，资产管理行业结合互联网企业利用互联网技术探寻基于互联网技术的新的策略因子。

1.2 互联网越发受到金融行业重视

“新产品的大数据源头来自于互联网上用户的搜索数据，通过对特定金融搜索行为进行数据挖掘和分析，作为构建投资模型的重要指标。”据互联网资料显示，此类指数在海外还没有相似的先例。随着国内互联网技术的飞速发展，对于传统的金融领域，从销售到研发，存在着各种创新。

而互联网的影响力也越来越受到金融业的关注。百度的数据显示，在2013年以来互联网金融的搜索频率上升飞快（如图1所示）。在报告《互联网金融的模式与产品》中提到，互联网指数并非对于互联网金融模式的改变。但这里认为在产品研发及投资研究领域，互联网大数据的技术会使得行业获得更为长足的进步。

图1：百度互联网金融指数

数据来源：百度，广发证券发展研究中心

二、互联网技术带来投资研究方式的进步

大数据的时代来临，给投资研究工作带来了新的体验。虽然并未改变投资和销售的业态，但对之后的投资研究业态带来了进步，对数据及投资策略的分析提出了更高的要求。在第三章也有三个利用互联网技术帮助实现投资决策及构建投资策略的例子。

2.1 大数据在投资研究上的优势

A. 覆盖面广

互联网的信息以及计算方法使得可供研究的领域大大拓宽。譬如传统的方式分析师对于股票的覆盖相当有限，并且一些数据获取较难的公司无法相应的覆盖。互联网的技术以及大数据使得分析师对股票的覆盖程度提高。

B. 刻画准确

数据量的增大使得分析的结果更加准确。在分析时，从数据推断的结论应该随着数据样本量的不断增加变得更加可信。互联网的信息和数据处理能力都呈现出几何级数的增长，对问题的刻画也变得更加精确。

C. 分析高效

互联网的技术及算法使得以往的计算效率得到提升。云计算及互联网的机器学习算法使得之前无法在短时间实现的大量数据计算可以很快的得到相应的结果。

2.2 投资研究应用大数据的形式

A. 深化原有研究

互联网数据的处理能力大幅提升、提供的信息大量增加，在传统分析框架之内，可以提高研究的深度和广度。如利用互联网数据分析某公司业绩，假设公司共有100

个直属分销商外加500个渠道商，40%分布在国内另外60%分布在海外。一个卖方分析师面对15家该类公司时，利用财务模型精确推断每家公司的业绩则变得相当困难。又如面对集中财报数据，每个分析师覆盖股票数量有限，难以对于每一家的财报数据给出及时的点评及合理的分析（可见3.1例子）。应用互联网数据及技术，可以深化之前的相关研究，在有限人力的基础上提供更深入的研究成果，让上述两个问题取得更好的解决方案。

B. 覆盖更多领域

在传统分析框架之外，互联网技术也提供了更多问题的分析方式。如社交网络数据或特定分类下群体社交网络上体现的投资情绪，可以推测市场上未来一段时间受到情绪影响较大的风险溢价等。又如新闻事件等无法用数据刻画的公司影响因素，利用互联网大数据可以计算出市场关注度等策略因子值（可见3.2例子）。

三、互联网技术在投资上的尝试

在之前的《互联网大数据挖掘系列专题》中，我们提供了三种互联网技术应用于股票投资的方法。

3.1 互联网读取公告数据

策略的目的是通过对公告进行重新分类，并结合公告的披露频率以及公告日个股股价涨跌情况来对相应个股进行涨跌预测，针对上市公司最新披露的公告信息进行事件选股策略构建时，我们主要考虑了以下几方面的信息:

（1）考虑公告类型;

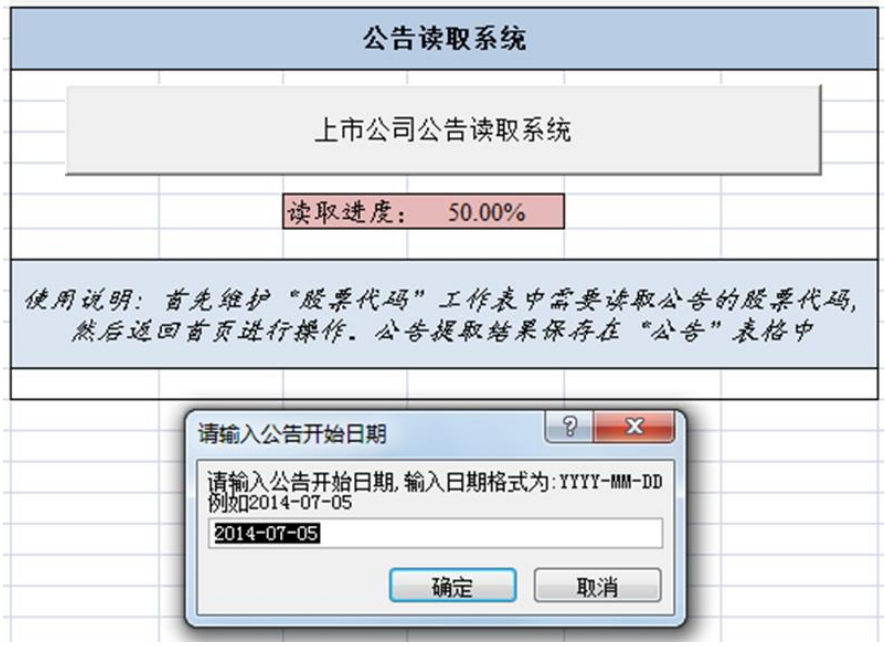
（2）考虑长期未出公告;

（3）考虑公告披露当日个股表现。

历史回测结果表明，财务报表、股东大会、增发公告以及重大积极事项等公告策略的选股效果相当突出，均能获得较高的超额收益。

基于对巨潮资讯公告披露平台以及深圳证券交易所信息披露平台的网页格式的研究，我们开发出了上市公司公告读取系统，该系统能够批量读取上市公司任意时间段内的公告信息，并将与公告相关的信息，譬如公告标题、公告日期、公告下载地址等信息提取出来。目前该开发工具处于继续完善阶段，改善的方向有根据用户需求，选择是否将公告正文下载下来、根据提取出的公告标题直接将公告归类等。（界面可见图2）

图2：A股公告抓取工具展示



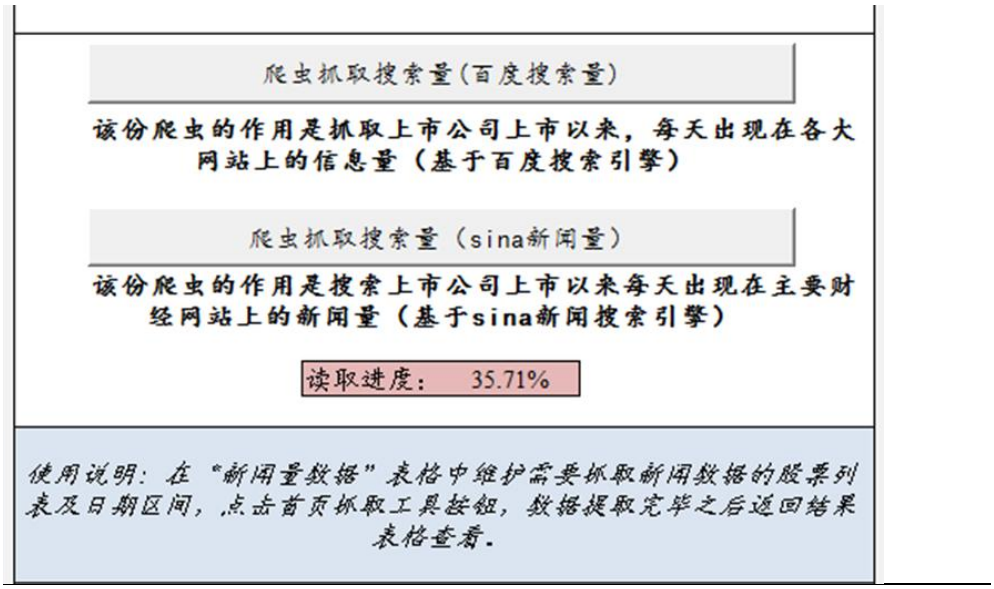
数据来源：广发证券发展研究中心

3.2 把握新闻热度

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。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、看空及看平状态；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。

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。（界面可见图3）

图3：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

3.3 投资者投资情绪

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。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:

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

将GS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%。

在对股吧论坛等复杂的文本数据进行挖掘和识别过程中，非常关键的一步是需要对文本内容进行关键词识别及语义识别，我们开发出了能够针对批量文本文件的内容进行关键词统计及上下文获取的小工具。（界面可见图4）

图4：文本关键词识别工具展示

1	2	3	4	5	6
关键字输入		文件名称	关键字	匹配数量	链接字符
关注	关键词识别	2008年上海市虹口区	关注	0	未匹配到数据
重大			重大	5	重大市政工
利好			利好	0	未匹配到数据
意义			意义	0	未匹配到数据
	说明：关键字数量不少于两个	2011年保亭黎族苗族	关注	1	关注民生，
			重大	5	重大历史机
			利好	1	利好大势，
			意义	0	未匹配到数据
		2011年北京市大兴区	关注	2	关注的生活
			重大	27	重大战略机
			利好	0	未匹配到数据
			意义	2	意义十分重
		2011年北京市昌平区	关注	1	关注的热点
			重大	11	重大突破，重
			利好	0	未匹配到数据
			意义	2	意义。意义重
		2011年北京市朝阳区	关注	0	未匹配到数据
			重大	15	重大项目和
			利好	0	未匹配到数据
			意义	0	未匹配到数据
		2011年安顺市政府工	关注	0	未匹配到数据

数据来源：广发证券发展研究中心

风险提示

本文仅对互联网技术对于投资研究的影响做讨论，未提供任何投资建议。

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