

证券研究报告•金融工程深度报告

基于新闻热度的周期、成长、消费风格轮动

一大数据研究之四

重要观点

数据挖掘概述及指标构建

数据挖掘基本步骤包括数据采集、数据预处理、数据存储、数据 分析、知识发现。数据挖掘常见技术包括监督学习(即分类分析)、 无监督学习(即聚类分析)、关联分析、预测分析等。研究中, 统计各个板块成份股新闻数量, 汇总成对应指数新闻热度。

基于新闻热度的风格轮动配置简述

根据周期、成长、消费三大板块新闻热度,在各自板块内,分别 构建新闻热度布林带,当突破上限时,看多指数。根据三大板块 信号, 均等持有看多信号的板块。若三大板块均没有信号, 则空 仓并持有现金,直到风格板块内出现看多信号再加仓。

单一板块策略

2014年01月01日到2017年08月04日期间,周期板块年化收 益 34%, 胜率 76%, 共出现看多信号 35 次; 成长板块年化收益 36%, 胜率 72%, 共出现看多信号 19次; 消费板块年化收益 32%, 胜率 76%, 共出现看多信号 22 次。且三大板块出现信号均持有 20 个交易日。

不同板块新闻热度效果各异

周期、成长、消费持有时间皆为 20 个交易日,说明新闻热度对 三大板块影响的持续时间基本相同,而移动平时时间即向前考察 时间则各不相同,其中周期30个交易日,成长10个交易日,消 费 20 个交易日,说明在三大板块中,对市场影响较大的新闻热 度向前看时间周期最长, 而成长最短。

风格轮动配置明显优于单一板块

分别以周期、成长、消费三大板块作为单一资产; 若三大板块有 买入信号,则第二天买入有买入信号的板块,并等权持有; 若三 大板块皆无买入信号,则空仓等待信号出现再持有对应板块。在 回测期间,风格轮动配置年化收益51%,日胜率57%,最大回撤 17%, sharp 值 1.92。

金融工程研究

丁鲁明

dingluming@csc.com.cn

021-68821623

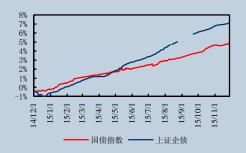
执业证书编号: S1440515020001

研究助理:喻银尤

yuyinyou@csc.com.cn 021-68821600-808

发布日期: 2017年08月16日

市场表现



相关研究报告

17.02.27

大数据研究之三: 新闻情绪选股的多空 17.03.08

差策略

大数据研究之指标构建: 机器学习之贝 17.03.02

叶斯文本分类算法的实现

大数据周报:新闻热度未见突破,大盘

继续谨慎; 新闻情绪选股空头组合相对

沪深 300 指数超额收益为-0.32%

大数据周报:大盘维持谨慎,新闻情绪 17.02.21

选股多头组合跑赢沪深 300 达 0.9%

大数据周报: 大盘相对谨慎, 重点关注 17.02.13

沪深 300 新闻情绪选股多空组合

大数据研究之择时:基于新闻热度的多 16.10.12

空策略

16.09.14 股票行业配置——基于投资时钟理论

16.09.13 基于残差分析的大类资产轮动策略

基本面量化系列之六——"量化基本面" 16.08.09

理论体系及通信行业案例

基本面量化系列之五——投资时钟指 16.06.24

路,量化大类资产轮动破局



目录

— 、	大数据与量化投资	. 3
二、	大数据体系构建	3
	2.1 数据采集与预处理	. 3
	2.2 大数据存储技术	4
	2.3 数据分析与指标构建	. 5
三、	新闻热度在不同风格板块的表现	. 5
	3.1 周期板块	6
	3.2 成长板块	7
	3.3 消费板块	8
四、	新闻热度风格轮动配置原理	9
	4.1 风格配置原理	9
	4.2 回测结果分析	10
五、	总结	11
;	风险提示	12



图形目录

图	1:	中信建投金融工程新闻数据来源3
图	2:	中信建投金融工程爬虫系统框架体系图4
图	3:	中信建投金融工程大数据存储5
图	4:	周期板块新闻热度6
图	5:	周期板块新闻热度择时策略净值7
图	6:	成长板块新闻热度7
图	7:	成长板块新闻热度择时策略净值8
图	8:	消费板块新闻热度8
图	9:	消费板块新闻热度择时策略净值9
图	10:	风格板块轮动配置策略净值10
		表格目录
表	1:	风格板块轮动配置策略结果统计10

大数据与量化投资

IBM 最早定义大数据的 5V 特点: Volume (大量)、Velocity (高速)、Variety (多样)、Value (价值)、Veracity (真实性)。当今社会,大数据所带来的信息风暴正在深刻的影响着我们的生活、工作和思维,大数据将开启一 次重大的时代转型。

传统量化投资主要包括量化选股、量化择时、股指期货套利、商品期货套利、统计套利、算法交易,资产 配置,风险控制等。传统的量化投资研究的数据来源一般是公司的财务指标、交易行情数据、政策宏观方面的 投资信息等。而随着量化投资这一领域的快速发展,这些传统数据中所包括的大部分投资信息已经被专业投资 者所挖掘,想要从这些信息中获取收益难度将越来越大。大数据将为量化投资这一领域创造前所未有的可量化 的新的维度,为量化投资提供了新的研究视野。如何把大数据这一金矿从数据转变为知识则充满挑战和困难, 大数据将驱动量化投资的创新。

、 大数据体系构建

2.1 数据采集与预处理

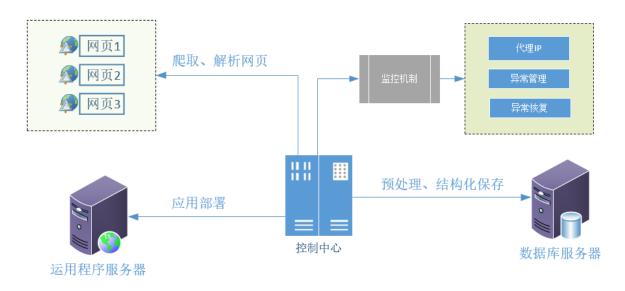
大数据的源头质量,直接决定我们指标质量,决定着我们的策略优劣性。目前,国内的相关数据来源主要 为第一类上交所,第二类财经新闻网站,第三类社交媒体,第四类为关注数据等。我们目前数据主要爬取新浪 财经个股相关新闻,包括 200 多家媒体在内的所有个股新闻。大数据采集则是通过网络爬虫或网站公开 API 等 方式从上述相关网站上获取我们所需要的数据信息,将非结构化数据从网页中爬取下来,并解析相关信息,将 其存储为统一的本地数据文件,并以结构化的方式存储在我们的数据库中。





数据来源:新浪财经

图 2: 中信建投金融工程爬虫系统框架体系图



数据来源:中信建投证券研究发展部

数据预处理指直接从网页爬取的数据并不能直接用于使用,而是需要经过一定的预处理,以保证数据质量和数据安全。因为在大数据应用中,数据来源非常广泛,数据质量良莠不齐,更需要预处理过程。数据预处理主要是去除无法解析的错误网页,删除重复的数据,去除无效的数据等;将不同的数据源爬取到的数据统一存储,建立数据仓库。

2.2 大数据存储技术

2014年1月1号到2016年9月26日,已经有200多万条个股新闻数据,共45g多,虽然现在不算超级大数据,但随着我们系统的逐渐完善,数据来源的多样化,数据存储一定会成为较大的瓶颈。为了满足大数据访问的效率与要求,大数据处理需要合理地存储与组织各种数据,以减少网络和存储I/O开销,提升系统性能;大数据存储目前我们主要是采用分表和分区技术。

以上技术应用于小型大数据还可以完美解决,但是超级大型数据则无能为力。目前有以下几种典型的大数据存储技术解决方案,第一种采用 MPP 架构的新型数据库集群,重点面向行业大数据,采用 Shared Nothing 架构,通过列存储、粗粒度索引等多项大数据处理技术,再结合 MPP 架构高效的分布式计算模式,具有高性能和高扩展性的特点,在企业分析类应用领域获得极其广泛的应用。第二种是基于 Hadoop 的技术扩展和封装,围绕 Hadoop 衍生出相关的大数据技术,应对传统关系型数据库较难处理的数据和场景。第三种是大数据一体机,这是一种专为大数据的分析处理而设计的软、硬件结合的产品,由一组集成的服务器、存储设备、操作系统、数据库管理系统以及为数据查询、处理、分析用途而特别预先安装及优化的软件组成,高性能大数据一体机具有良好的稳定性和纵向扩展性。



图 3: 中信建投金融工程大数据存储



数据来源: 中信建投证券研究发展部

2.3 数据分析与指标构建

数据挖掘常用的几种方法为:分类分析、聚类分析、关联分析、预测分析、异常分析等等。分类分析是首先从已有的数据中选出已有的分类,且把所有的没有分类的要进行分类的数据按照这些已规定好类别分别进行分类。聚类分类不属于预测性的问题,该算法主要解决的是把一群给定的对象划分成若干个组的问题。划分样本的依据是聚类问题的核心点。聚类分析主要是解决当要分析的数据缺乏描述信息或者是无法组织成任何分类模式时用于样本的聚类分析。关联分析中主要技术是对象相关度或者他们之间的关系。预测分析主要包括一无线性回归,多元线性回归,Markov 预测模型等。本文主要是统计新闻数量得到新闻热度指标。

三、新闻热度在不同风格板块的表现

在各自板块中,策略思想主要来源于股市技术分析的常用工具之布林带思想。

新闻热度指利用数据挖掘技术,分析财经网站个股新闻,统计各板块成份股每日新闻数量并构建成当天板 块新闻热度指标。

布林带指标设置:

中轨线: 新闻热度 M 日移动平均值

上轨线: 中轨线*(1+N%)

下轨线: 中轨线* (1-N%)

择时策略:

若当日信号突破上轨,则发出看多信号,次日以开盘价开多仓,同时将上次信号以次日开盘价平仓,开仓之后,若没有新信号,持有 nday 天以 nday+1 天开盘价平仓。

3.1 周期板块

以中信周期板块指数作为择时标的,当当日新闻热度突破前 30 个交易日平均新闻热度的 1.8 倍,则次天开仓买入。2014 年 01 月 01 日到 2017 年 08 月 04 日期间,周期板块年化收益 34%,胜率 76%,共出现看多信号 35 次。

图 4: 周期板块新闻热度



图 5: 周期板块新闻热度择时策略净值



数据来源:中信建投证券研究发展部

3.2 成长板块

以中信成长板块指数作为择时标的,当当日新闻热度突破前 10 个交易日平均新闻热度的 1.5 倍,则次天开仓买入。2014 年 01 月 01 日到 2017 年 08 月 04 日期间,成长板块年化收益 36%,胜率 72%,共出现看多信号 19 次。

图 6: 成长板块新闻热度

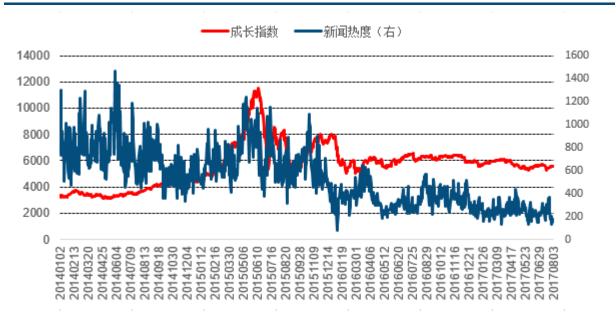


图 7: 成长板块新闻热度择时策略净值



数据来源: wind 资讯,中信建投证券研究发展部

3.3 消费板块

以中信消费板块指数作为择时标的,当当日新闻热度突破前 20 个交易日平均新闻热度的 1.5 倍,则次天开仓买入。2014 年 01 月 01 日到 2017 年 08 月 04 日期间,成长板块年化收益 32%,胜率 76%,共出现看多信号 22 次。

图 8: 消费板块新闻热度



图 9: 消费板块新闻热度择时策略净值



数据来源: wind 资讯,中信建投证券研究发展部

周期、成长、消费持有时间皆为 20 个交易日,说明新闻热度对三大板块影响的持续时间基本相同,而移动平时时间即向前考察时间则都不相同,其中周期 30 个交易日,成长 10 个交易日,消费 20 个交易日,说明在三大板块中,对市场影响较大的新闻热度向前看时间周期最长,而成长最短。

四、新闻热度风格轮动配置原理

在中信风格板块内配置,和投资单一板块不同,绝大多数投资期,无论所处的经济状况如何,都存在相对 收益较高的板块。配置本身强调资产的低相关性,但本文主要从风格板块出发,所以此配置强调如何抓住每 个板块的绝对收益。

在互联网大数据时代,我们获取信息的途径更加广泛和便捷,我们不再局限于传统的基本面数据和技术面数据。如何利用这些信息则是非常具有挑战性的问题。资本市场变幻莫测,至今没有任何理论能完全解释并预测股票未来的趋势。互联网大数据的到来,则为我们提供了新思路。本文中,我们利用新闻这一大数据来源来配置不同风格的资产(板块),以获取每个板块的收益。达到投资收益最大化。

4.1 风格配置原理

得到周期、成长、消费三大板块新闻热度,在各自板块内,分别构建新闻热度布林带,当突破上限时,看 多指数。根据三大板块信号,均等持有看多信号的板块。若三大板块均没有信号,则空仓并持有现金,直到风 格板块内有看多信号再加仓。

4.2 回测结果分析

数据说明:

数据区间: 2014-01-01 至 2017-08-04。

板块数据:周期、成长、消费三大板块新闻热度指数。

配置标的:周期、成长、消费三大板块指数。

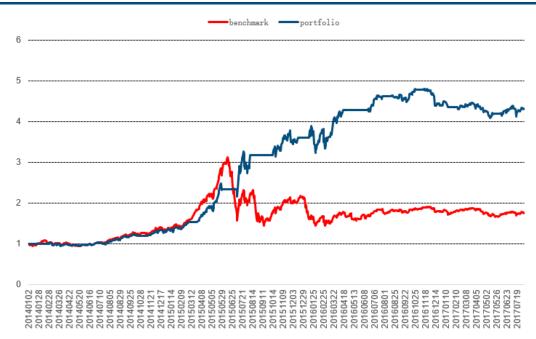
持有组合:若三大板块或二大板块有信号,则第二天等权持有有信号的板块;若只有一个板块有信号,则满仓持有这一板块;若三大板块皆没有信号,则空仓并持有现金。

结果分析:

表 1: 风格板块轮动配置策略结果统计

回测期间	2014-01-01	2014-01-01 至 2017-08-04				
初始净值	1	最终净值	4. 29			
年化收益	51%	年化超额收益	29%			
最大回撤	17%	夏普比	1.92			
数据来源:中信建投证	正券研究发展部					

图 10: 风格板块轮动配置策略净值



五、总结

传统量化投资主要包括量化选股、量化择时、股指期货套利、商品期货套利、统计套利、算法交易,资产配置,风险控制等。传统的量化投资研究的数据来源一般是公司的财务指标、交易行情数据、政策宏观方面的投资信息等。大数据将为量化投资这一领域创造前所未有的可量化的新的维度,为量化投资提供了新的研究视野。如何把大数据这一金矿从数据转变为知识则充满挑战和困难,大数据将驱动量化投资的创新。

通过市场情绪分析、财经文本分析、新闻热点捕捉、主题挖掘等从这些大量的新闻中挖掘出有效信息。利用数据挖掘技术,即利用各种方法分析我们需要处理的数据,发现隐藏在海量数据背后的知识和规律。挖掘步骤简单的可以概括为 a.前期数据的准备 b.从这些数据中寻找他们的规律 c.把寻找到的规律表示出来,这 3 个步骤。前期数据的准备是从这些相关的数据源中以一定的规则挑选我们所需的数据,然后整合成我们用于数据挖掘的数据集;寻找这些数据的规律是利用数据挖掘相关的方法将这些数据集所含的规律挖掘出来;把寻找到的规律表示出来是利用比如图表等可视化的技术尽可能以用户可以理解的方式展示出来挖掘出来的规律。

在中信风格板块内配置,和投资单一板块不同,绝大多数投资期,无论所处的经济状况如何,都存在相对收益较高的板块。配置本身强调资产的低相关性,但本文主要从风格板块出发,所以此配置强调如何抓住每个板块的绝对收益。

得到周期、成长、消费三大板块新闻热度,在各自板块内,分别构建新闻热度布林带,当突破上限时,看多指数。根据三大板块信号,均等持有看多信号的板块。若三大板块均没有信号,则空仓并持有现金,直到风格板块内有看多信号再加仓。2014年01月01日到2017年08月04日期间,周期板块年化收益34%,胜率76%,共出现看多信号35次;成长板块年化收益36%,胜率72%,共出现看多信号19次;消费板块年化收益32%,胜率76%,共出现看多信号22次。且三大板块出现信号均持有20个交易日。

分别以周期、成长、消费三大板块作为单一资产;若三大板块有买入信号,则第二天买入有买入信号的板块,并等权持有;若三大板块皆无买入信号,则空仓等待信号出现再持有对应板块。在回测期间,风格轮动配置年化收益 51%,日胜率 57%,最大回撤 17%,sharp 值 1.92。

后续研究方向:

该策略新闻来源比较单一,目前主要来源于新浪财经网站,来自 200 多家媒体的新闻。其次,指标比较单一,该策略仅对板块成份股的新闻数量进行了统计,虽然具有一定的代表性,但并不全面。后续可以考虑从多网站多指数进行研究。

以上相关结论仅来自 2014 年 1 月 1 日到 2017 年 8 月 4 日期间的统计分析得出,时间相对并不太长,不能完全代表未来及 14 年以前的过去。



六、风险提示

以上结论来自历史统计,并不代表未来一定有这种规律。

大数据预测的前提是数据大而全,并且数据质量可靠。由于数据来源有限,目前主要用新浪财经的个股新闻来做研究,虽然具有代表性,但并不能完全代表市场

分析师介绍

丁鲁明: 同济大学金融数学硕士,中国准精算师,现任中信建投证券研究发展部金融工程方向负责人,首席分析师。9 年证券从业,历任海通证券研究所金融工程高级研究员、量化资产配置方向负责人;先后从事转债、选股、高频交易、行业配置、大类资产配置等领域的量化策略研究,对大类资产配置、资产择时领域研究深入,创立国内"量化基本面"投研体系。多次荣获团队荣誉:新财富最佳分析师 2009 第 4、2012 第 4、2013 第 1、2014 第 3 等;水晶球最佳分析师 2009 第 1、2013 第 1 等。

研究助理 喻银尤: 021-68821600-808 <u>yuyinyou@csc.com.cn</u>

复旦大学硕士,通过 CFA 三级,两年上交所相关部门工作经验。专注于多因子构建和 大数据择时,选股,风格轮动等相关策略研究。

社保基金销售经理

研究服务

彭砚苹 010-85130892 pengyanping@csc.com.cn 姜东亚 010-85156405 jiangdongya@csc.com.cn

机构销售负责人

赵海兰 010-85130909 zhaohailan@csc.com.cn

北京非公募组

张博 010-85130905 zhangbo@csc.com.cn
朱燕 010-85156403 zhuyan@csc.com.cn
李祉瑶 010-85130464 lizhiyao@csc.com.cn
李静 010-85130595 lijing@csc.com.cn
赵倩 010-85159313 zhaoqian@csc.com.cn
周瑞 18611606170 zhourui@csc.com.cn
刘凯 010-86451013 liukaizgs@csc.com.cn

北京公募组

黄玮 010-85130318 huangwei@csc.com.cn 黄杉 010-85156350 huangshan@csc.com.cn 任师蕙 010-85159274 renshihui@csc.com.cn 王健 010-65608249 wangjianyf@csc.com.cn 罗刚 15810539988 luogang@csc.com.cn

上海地区销售经理

陈诗泓 021-68821600 chenshihong@csc.com.cn
邓欣 021-68821610 dengxin@csc.com.cn
黄方禅 021-68821615 huangfangchan@csc.com.cn
戴悦放 021-68821617 daiyuefang@csc.com.cn
李岚 021-68821618 lilan@csc.com.cn
肖垚 021-68821631 xiaoyao@csc.com.cn
吉佳 021-68821600 jijia@csc.com.cn
朱丽 021-68821600 zhuli@csc.com.cn
杨晶 021-68821600 yangjingzgs@csc.com.cn
谈祺阳 021-68821600 tanqiyang@csc.com.cn

深广地区销售经理

胡倩 0755-23953859 huqian@csc.com.cn 张苗苗 020-38381071 zhangmiaomiao@csc.com.cn 许舒枫 0755-23953843 xushufeng@csc.com.cn 廖成涛 0755-22663051 liaochengtao@csc.com.cn



评级说明

以上证指数或者深证综指的涨跌幅为基准。

买入:未来6个月内相对超出市场表现15%以上;

增持:未来6个月内相对超出市场表现5-15%:

中性: 未来6个月内相对市场表现在-5-5%之间;

减持:未来6个月内相对弱于市场表现5-15%;

卖出: 未来6个月内相对弱于市场表现15%以上。

重要声明

本报告仅供本公司的客户使用,本公司不会仅因接收人收到本报告而视其为客户。

本报告的信息均来源于本公司认为可信的公开资料,但本公司及研究人员对这些信息的准确性和完整性不作任何保证,也不保证本报告所包含的信息或建议在本报告发出后不会发生任何变更,且本报告中的资料、意见和预测均仅反映本报告发布时的资料、意见和预测,可能在随后会作出调整。我们已力求报告内容的客观、公正,但文中的观点、结论和建议仅供参考,不构成投资者在投资、法律、会计或税务等方面的最终操作建议。本公司不就报告中的内容对投资者作出的最终操作建议做任何担保,没有任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺。投资者应自主作出投资决策并自行承担投资风险,据本报告做出的任何决策与本公司和本报告作者无关。

在法律允许的情况下,本公司及其关联机构可能会持有本报告中提到的公司所发行的证券并进行交易,也可能为这些公司提供或者争取提供投资银行、财务顾问或类似的金融服务。

本报告版权仅为本公司所有。未经本公司书面许可,任何机构和/或个人不得以任何形式翻版、复制和发布本报告。任何机构和个人如引用、刊发本报告,须同时注明出处为中信建投证券研究发展部,且不得对本报告进行任何有悖原意的引用、删节和/或修改。

本公司具备证券投资咨询业务资格,且本文作者为在中国证券业协会登记注册的证券分析师,以勤勉尽责的职业态度,独立、客观地出具本报告。本报告清晰准确地反映了作者的研究观点。本文作者不曾也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

股市有风险,入市需谨慎。

地址

北京中信建投证券研究发展部

中国北京 100010

东城区朝内大街 2 号凯恒中心 B座 12 层

电话: (8610) 8513-0588 传真: (8610) 6518-0322

上海中信建投证券研究发展部

中国上海 200120

浦东新区浦东南路 528 号上海证券大厦北塔 22 楼 2201 室

电话: (8621) 6882-1612 传真: (8621) 6882-1622