文本挖掘研究回顾一：互联网数据挖掘系统，行为金融新领域

专题报告

分析师

刘道明 (执业证书编号 :S0930210060005)

021-22169109

liudaoming@ebscn.com

◆事件起因：

最近无论是主动投资还是量化投资领域都对文本挖掘研究产生了高度关注度，作为国内相关研究领域的开创者，我们认为有必要将过去近四年的研究成果进行重新梳理与回顾，最重要的是对过去研究中存在的不足进行反思，为今后无论是我们自身还是同行的相关研究提供参考。

第一篇回顾选用 2010 年 9 月 16 日发的一篇深度报告。该报告是我们进入该研究领域的开篇之作，站在今天的时点看，这篇报告完成了两项有意义的工作：1、建立了我们从文本采集、清洗、结构化，到量化建模、结果输出的大体数据流程和分析体系；2、为市场贡献了独家的股票关注度因子，近四年里，就单因子来讲，关注度因子一直有稳定优异的表现。

由于当时对市场理解的不够深入和研究方法的相对稚嫩，该报告也存在诸多不足：1、情绪指标的构建方法存在较大问题，该指标也在 2011 年我们推出普通投资者情绪指标后被弃用；2、整体的流程与架构虽然方向正确，但在诸多细节处理部分仍存在许多不足之处，致使后期数据更新和模型维护存在较多不便；3、双反转模型并不适合机构投资者和大资金，在之后也被我们弃用。

下面的篇幅中，我们对原报告不作任何修改的重新发布，温故而知新，激励我们在这一领域继续埋头研究，深耕细作。

◆互联网：“营业部自行车”2.0 版

情绪是投资分析框架中非常重要的一环，上个世纪，就有这样的朴素结论：看营业部门口的自行车数量，当自行车很少的时候可以买股票，但当自行车数量很多的时候就得卖股票了。如今，在“交易网络化”和“交通汽车化”的推动之下，互联网毅然接过了“营业部自行车”的大旗，成为情绪指标 2.0 版，而财经网站和股票论坛的火爆使我们有了一个可以直接量化投资者情绪的可靠数据来源。

◆互联网海量数据挖掘系统：业内首创

我们在业内第一个建立了基于互联网的海量数据挖掘系统，完整的系统将包括情绪指标、个股及板块关注度、关键词跟踪等部分。

目前，基于股票论坛新发文章数量指标能作为较好的大市情绪指标，与大盘同步相关性 60%，该情绪指标**波动的变化能提示市场的中短期拐点**。

基于个股的关注度指标验证了“人弃我取、人取我与”投资理念的正确性：单独考虑个股关注度的变化，自 2008 年中以来，以月度考察，**关注度下降最多**的股票构成的组合**显著跑赢关注度上升最多**的股票组合，两年累积收益超过 100%，超额收益 80%以上，且超额收益**完全来自于 Alpha**。以周为单位，选择同时满足**关注度下降最多和表现最差**的股票构成组合，**两年累积收益率 658%**。

基于已经建立的数据挖掘系统，我们将陆续推出一系列产品：大市情绪指标、个股关注度指标、行业及板块关注度指标，从一个完全新的角度形成独立的选时、行业配置、选股的数量化体系，并与传统的数量化体系融合，将国内数量化投资研究推向一个新的领域——行为金融。

1、情绪指标：我们的认识

没有人能够否认，股价由预期推动，预期则受到情绪的巨大影响。在“动物凶猛”的股林之中，我们总有如此的企图：洞悉别人的情绪，做相反的决策。背后含着简单却永恒的哲理：在投资中，赢家总是少数派。

在传统的基于财务基本面的数量化研究开始遭遇瓶颈时，我们希望在行为金融领域另辟蹊径，“情绪”，是不能也不愿绕过的坎。在对“光大证券A股市场恐慌指数”近一年的连续维护过程中，我们对情绪和情绪指标逐渐有了自己的理解，也坚信，从不同角度对情绪的量化，是一项有现实意义的工作。

1.1、情绪指标1.0版：营业部自行车

笔者第一次知道“股票”这两个字，是在1993年，全家凑钱购买青岛啤酒认股证并认股青岛啤酒时，当时，按照家人的说法，买认股证认购新股，没有不赚钱的，而且能赚好几倍。但后来的结果却是：青岛啤酒IPO当日高开低走，此后一路下跌，一直到1994年七月份的最低价3.05元，预期“包赚”最终成为了“大亏”。这是笔者对中国股市的第一印象，不仅不美好，甚至是痛苦。

后来，每每回想起来，大体会有这样的判断：家人入市时点正好是市场情绪最疯狂的阶段，多年积蓄的代价换来对市场乐观情绪的负面认识。

情绪的第二课来源于1994年非常有名的一部电影——《股疯》，大团圆式的结局背后是一场很好的风险教育课。

在市场经历了若干次的暴涨暴跌之后，市场上有了对情绪指标及其应用最最朴素的诠释：看营业部门口的自行车数量，当自行车很少的时候可以买股票，但当自行车数量很多的时候就得卖股票了，这是国内的情绪指标1.0版。

1.2、情绪指标三路径：股票市场、衍生品市场和舆情

“营业部自行车”很好的诠释了什么叫投资者情绪，该如何看待情绪。但是从数量化研究的角度看，则存在以下几个问题：一、难以量化，难以想象有人会每个交易日坚持不懈的数自行车并忠实记录；二、判断标准不确定，因为没有历史的时间序列数据，“多”与“少”的标准就变得很模糊；三、样本问题，在“交易网络化”与“交通汽车化”的双重背景之下，营业部门口的自行车数量已经不再具有代表性。

中国股市发展到今日，已经有了许多种量化情绪的方法，市场上也出现了各式各样的情绪指标，但大体上，按数据来源或者编制途径划分则可分为三类：基于股票市场本身；基于衍生品市场；基于舆情。

1.2.1、基于股票市场

基于股票市场的情绪指标大体是从价、量以及量价配合的角度编制。最著名的莫过于技术指标中的“超买超卖型”指标，我们的“光大证券A股市场活跃度指数EMAX”也是试图从量价配合的角度对情绪做出量化。

直接基于股票市场数据的情绪指标优缺点很明显，优点就是如实的反映市场最新状态，一般都比较简单便于理解；但缺点就是，指标只能反映一个历史已经发生的状况及投资者的应激反应，但不能反映投资者的预期。

1.2.2、基于衍生品市场

随着海外A股相关的衍生品的逐渐活跃与股指期货推出背景下的衍生工具逐渐放开，基于衍生品市场的情绪指标将有很大的用武之地，其中，“光大证券A股市场波动率指数EVX”就是通过衍生品反映了投资者对A股市场未来波动的预期。这类指标不仅与市场最新状况紧密结合，而且能反映投资者对未来的预期。但是在目前状况下，由于国内相关衍生工具的暂时缺乏，相应指标均是不同程度的近似，而一旦相关工具推出并放开，则又有可能因为市场有效而导致预期反映出情绪指标失效。

1.2.3、基于舆情

在互联网高度发达的今日，有一类数据是我们不能忽视的：网络数据。各大财经网站、股票论坛的高度活跃，使得网络舆情成为反映投资者情绪尤其是股民兼网民情绪的最好数据来源。国内目前也有一些相关研究成果，但总体上仍处于起步阶段，相关研究的进展远远落后于网络发展的原因在于数据处理和模型开发的复杂性：互联网数据都是海量级别的，从获取、存储到数据清理都需要丰富的经验和强大的系统支持；来自于互联网的数据形式往往是文本类型的，传统的基于数值型的模型已经不适用，需要基于文本的数据挖掘模型，但目前，由于中文的高度复杂性，语义处理仍未得到较好解决。

1.3、情绪指标的逻辑与评判标准

作为行为金融的重要一环，情绪指标有其重要的逻辑和评判标准。总的来说，无论指标的具体形式如何，都要符合下述条件和标准：应用上的逆向思维与顺势而为；变化上的适度震荡；拐点上的合理提前。

1.3.1、应用上的逆向思维与顺势而为

虽然无论海外经验还是国内经验都证明，长期来看，股票市场的趋势是向上的，但参于者共赢的局面从未有过，多数情况下是，少数参于者成为了暂时的赢家，多数派则成为少数派的赢利之源，这一规律是股票市场本身的金融属性决定的，无论参于者如何变化，此铁律都很难被打破。

由此，逆向思维就变得非常重要，“人弃我取，人取我与”是大体的方法：在情绪过度悲观时买入，在情绪过度乐观时卖出。但问题在于，需要有比较明确的取值范围表明什么位置算是过度乐观，什么位置算是过度悲观，因为只有在极限位置，逆向操作才有意义，而在中间态时是需要顺势而为的。

1.3.2、变化上的适度震荡

市场的瞬息万变影响着投资者的情绪，情绪则通过行为反馈于市场，从本质上说市场与情绪形成一个简单的开环回馈系统。从这个角度上说，在单边市场中，市场与情绪会互相加强，情绪指标只在极限值区域才有较高的价值，而在中间态时指导意义则会打折。

震荡市中则不同，“波段操作，高抛低吸”是获利的法门，在这种市场环境之下，情绪指标应当表现的适度震荡，以保证对于市场中短期波动的敏感性。

1.3.3、拐点上的合理提前

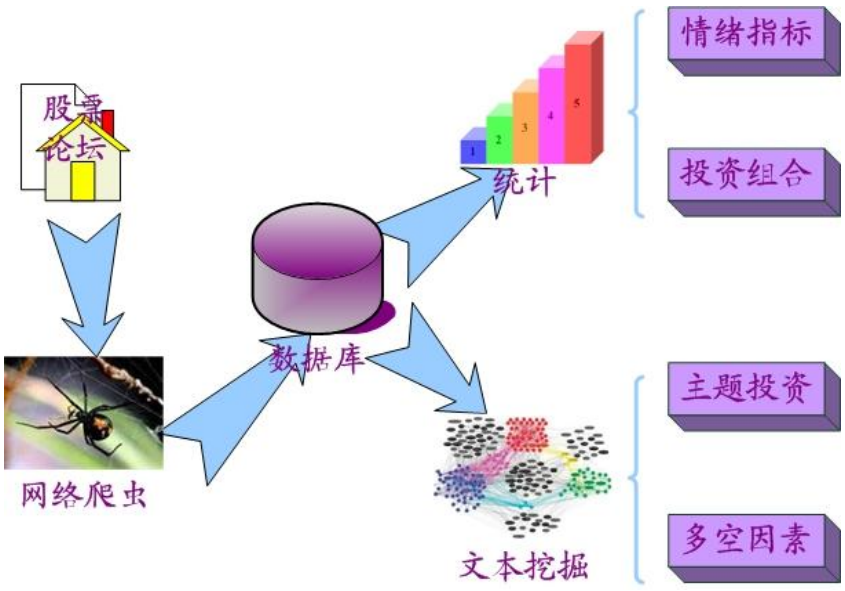
如前所述，在单边市场中，市场与情绪是互相加强的，在大牛大熊市里，情绪都有可能突破历史的极限值。比如，如果我们以 06 年之前的历史乐观情绪高点来衡量 07 年的单边上涨，就可能错失大段的市场上涨；而如果以之前的情绪低点来衡量 08 年的单边下跌，则又可能因过早入场而蒙受巨大损失。在市场出现突破历史经验的情况时，虽然有难度，但我们还是希望情绪指标能够提早于市场出现拐点。这种拐点的提前性在逻辑上有其存在的可能：市场的高点往往不是情绪的最高点，因为只有当逆向思维、反向操作的力量累积到一定程度，才有可能改变市场固有的趋势，而这种力量的积累一定是能在情绪中看到的，问题就是在于我们如何捕捉这种情绪。

2、互联网数据：不仅是“营业部自行车”2.0 版

在互联网高度发达的今日,基于网络数据的挖掘将是行为金融学研究领域中一个重要的方面。与之前我们经常使用的宏观经济、财务和行情数据不同，网络数据无论以哪种类型呈现，后期处理都很复杂，需要依靠完整的系统和丰富的经验，这也是目前基于网络数据的研究远远落后于互联网发展速度的重要原因。

我们依靠先前在互联网海量数据挖掘领域的经验，建立了较为完整的从数据获取、数据清理到文本挖掘、统计分析流程，我们认为互联网数据如实的反映了方方面面的情绪，而完整的分析体系将在两个层面（数据型统计、文本挖掘）、四个方向（市场情绪、基于关注度变化的股票组合、主题投资、利多利空因素对市场影响程度测算）上为投资提供有效指导。

图 1：网络数据挖掘系统流程



资料来源：光大证券研究所

2.1、股票论坛：信息集聚地

在互联网顺利进入 2.0 时代，并向 3.0 大踏步迈进时，大量网络数据由众多的用户自发、交互的生成，而不再是由媒体按照传统的采、编、发布的流程生成。其背后涵盖的信息量是惊人的，不仅包括各种消息，更包含了方方面面不同立场不同思维方式的人群的不同观点与情绪。其中，越来越发达的股票论坛是此类数据的重要集聚地。

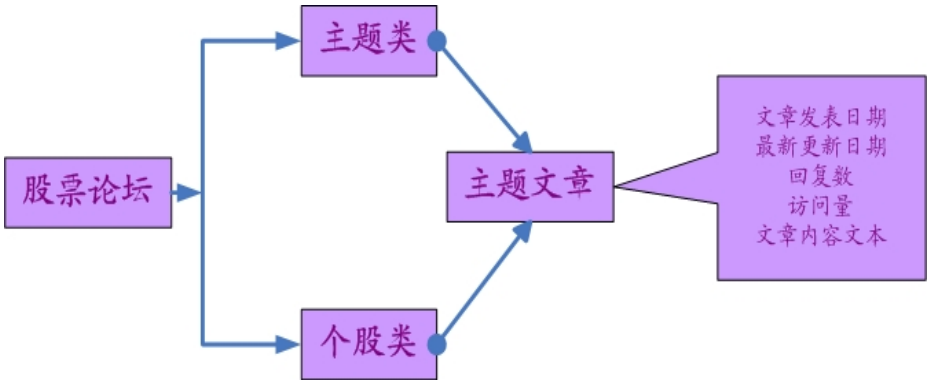
现阶段，股票论坛的主要形式是某一用户按某一主题发帖，众多用户回复、访问。将所有的主题文章数据通过“网络爬虫”抓取到本地分析系统后，经加工整理，可以得到几部分的数据：

- **数值类：**文章属于某一特定主题或者某个股；文章发表日期；文章最新更新日期；文章回复数量；文章访问数量
- **文本类：**文章标题；文章内容；文章回复内容

基于这些数据应用不同的统计分析，能够在我们前述两个层面四个方向对投资做出指导：

- **情绪指标：**通过每一文章的发表时间、回复量和访问量，可以每天得到三个指标，股票论坛总访问量、股票论坛总回复量、股票论坛当日总新发文章数量。这三个指标都将是很好的市场情绪指标。
- **投资组合：**对所有文章根据某一主题或者某一个股进行汇总统计，可以得到基于单一主题或者单一个股的总访问量、总回复量、总新发文章数量数据。我们对历史数据的回朔表明，以一段时间新发文章数降幅最大的个股构成的投资组合表现较好。
- **主题投资：**基于文章内容的文本挖掘程序可以提取每篇文章的关键词，加以统计后，可以得到每个关键词的出现频度。这些统计数据，基于主题投资，最大的意义是能够在一个新关键词刚冒头时便捕捉到，加以研究后提前布局。
- **利多利空因素：**同样是基于前述关键词出现频度的统计数据，能够直观反映影响市场的利多利空因素的关注程度：当某一多空因素对应的关键词出现频度高时，显然该因素对市场仍有较大的影响；而当某一因素对应的关键词出现频度低时，则表示该因素对市场的影响已经有限。

图 2：股票论坛数据组织形式



资料来源：光大证券研究所

2.2、新发文章数：大盘情绪指标

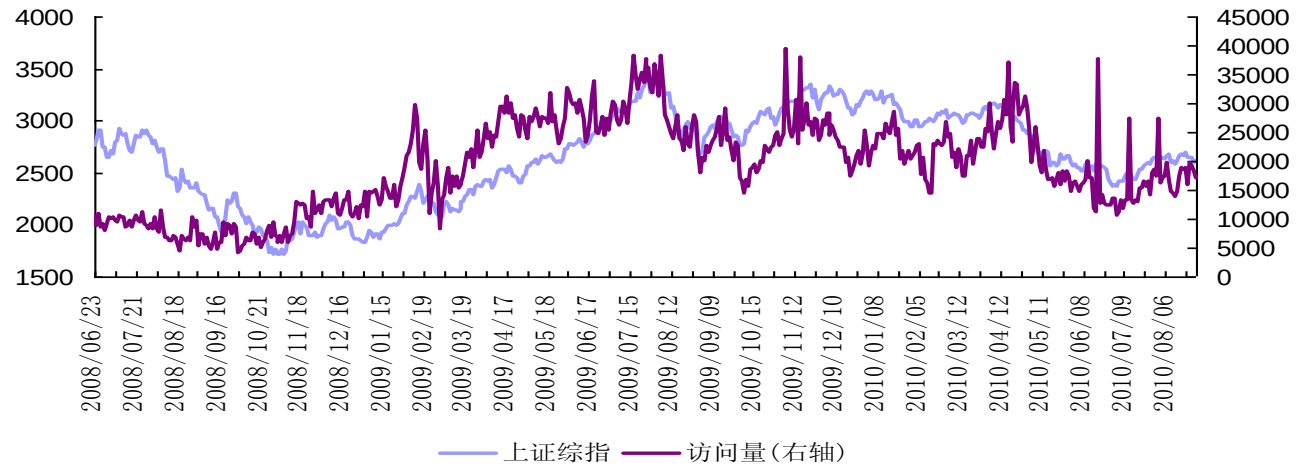
将股票论坛每天的访问量、文章回复量、新发文章数汇总统计后，能够得到三个人气指标：论坛每日总访问量、论坛每日总回复量、论坛每日总新发文章数。

从实际结果看，三个指标都是明显的同步指标，指标高低点基本与市场高低点对应。从三指标与上证综指的同步相关性看，访问量、文章回复量、新发文章数三指标与上证综指的相关系数分别为 66.37%、59.77%、59.14%。

但从单日的变化率来看，三指标的波动性要远远强于大盘的波动性：自我们数据样本起始的 2008-6-23 日至今，上证综指每日变化幅度的标准差为 2.06%，而访问量、文章回复量、新发文章数三指标每日变化幅度的标准差则分别达到 19.76%、21.87%、10.61%。

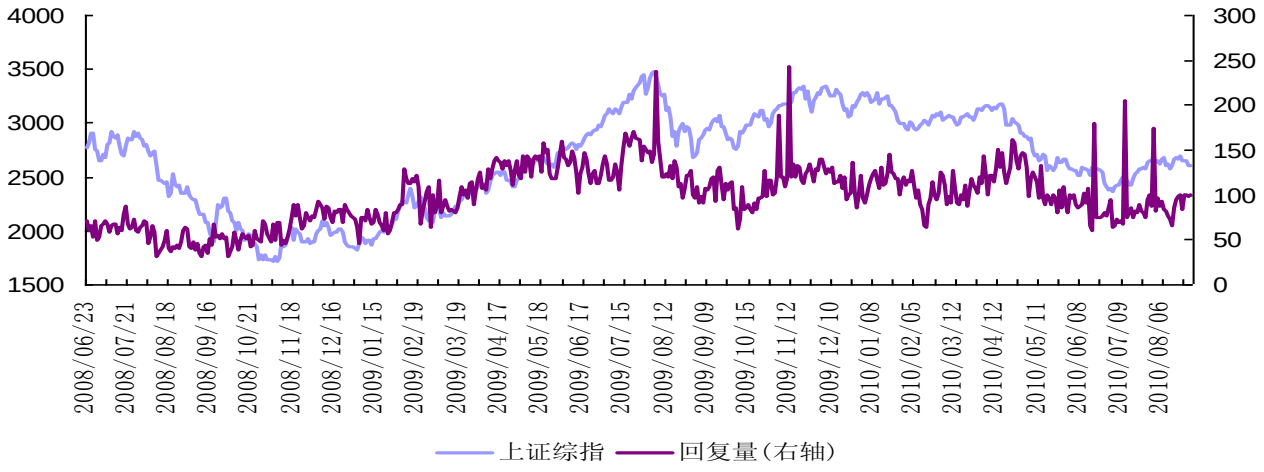
从与大盘相关性角度以及波动性角度综合考虑，访问量指标在此三个指标中效果最佳。但如果考虑数据的有效性，目前，我们倾向于选择新发文章数。主要原因在于：在回朔历史数据的过程中，我们按每篇文章的发布时间、回复数量和访问量进行统计，其中，发布时间是一个唯一确定值，而回复数量和访问量都是表示的自文章发布之日起至获取数据当时的总存量，例如，我们抓取了 2010 年 8 月 3 日发布的一篇帖子的数据，无论我们是在其后的哪一天得到了这样的数据，8 月 3 日都是一个唯一确定值，在处理上此文章也唯一确定的被统计到 8 月 3 日的数据中；而随着时间的推移，该文章的访问量和回复量是一直变化的，显然，8 月 3 日该文章的访问量和回复量一般都会小于 9 月 3 日该文章的相应值，但由于数据上的限制，在统计时，8 月 3 日发布的文章的访问量和回复量都被统计到 8 月 3 日的数据中，这就引入了不小的误差，在回朔历史数据时，我们在同一天获取了近两年的数据，这会显著的高估历史值。未来，随着系统的完善，我们可以做到每天更新数据，访问量和回复量指标未来可能存在的这种误差将被消除，在持续跟踪后，我们会对三指标进行重新的审视。

图 3：论坛每天访问量指标



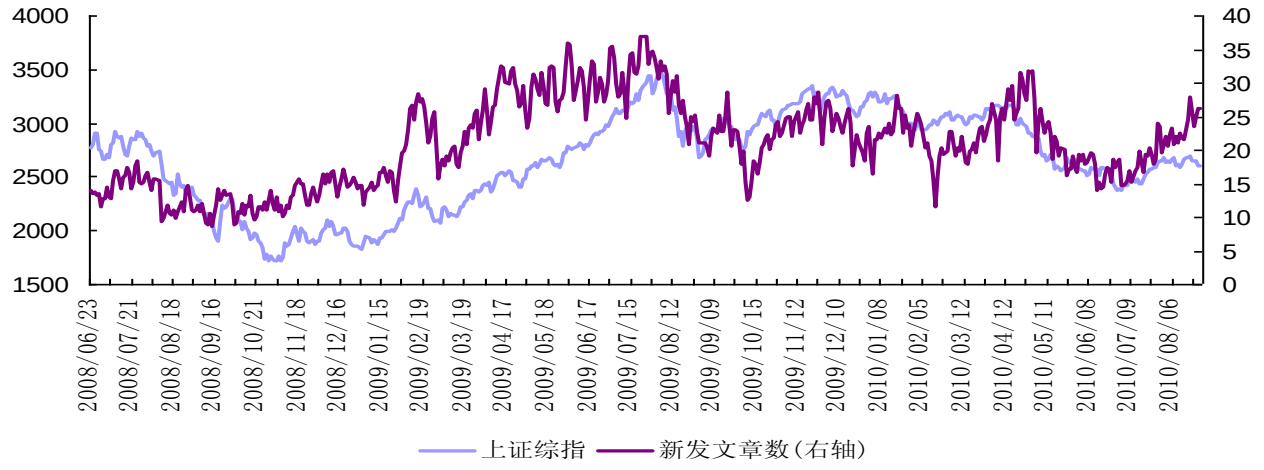
资料来源：光大证券研究所 单位：万次

图 4：论坛每天文章回复量指标



资料来源：光大证券研究所 单位：万次

图 5：论坛每天新发文章数量指标



资料来源：光大证券研究所 单位：万篇

单纯考虑新发文章数指标与大盘的同步性，我们认为该指标的绝对值很难对市场走势的预测提供有效参考。真正有效的参考来自于该指标更高的波动性，我们计算了新发文章数指标十天的年化历史波动率（每个交易日计算过去十个交易日新发文章数变化幅度的标准差，并乘以 252 的开根进行年化），发现：该指标在一定范围内表现出较好的震荡性，且高低拐点往往领先于市场的中短期拐点（一般周期在一个月左右）。

在十天年化历史波动率指标的基础上，我们通过下述方法筛选出拐点，对于 T 日，考察 T-4 日自 T 日的波动率值 V，定义波动标志 X：

$$IF:(V_{T-2}>V_{T-3})\wedge(V_{T-3}>V_{T-4})\wedge(V_{T-2}>V_{T-1})\wedge(V_{T-1}>V_T)\wedge(V_{T-2}>1.8),X=1$$

$$IF:(V_{T-2}<V_{T-3})\wedge(V_{T-3}<V_{T-4})\wedge(V_{T-2}<V_{T-1})\wedge(V_{T-1}<V_T)\wedge(V_{T-2}<1.2),X=-1$$

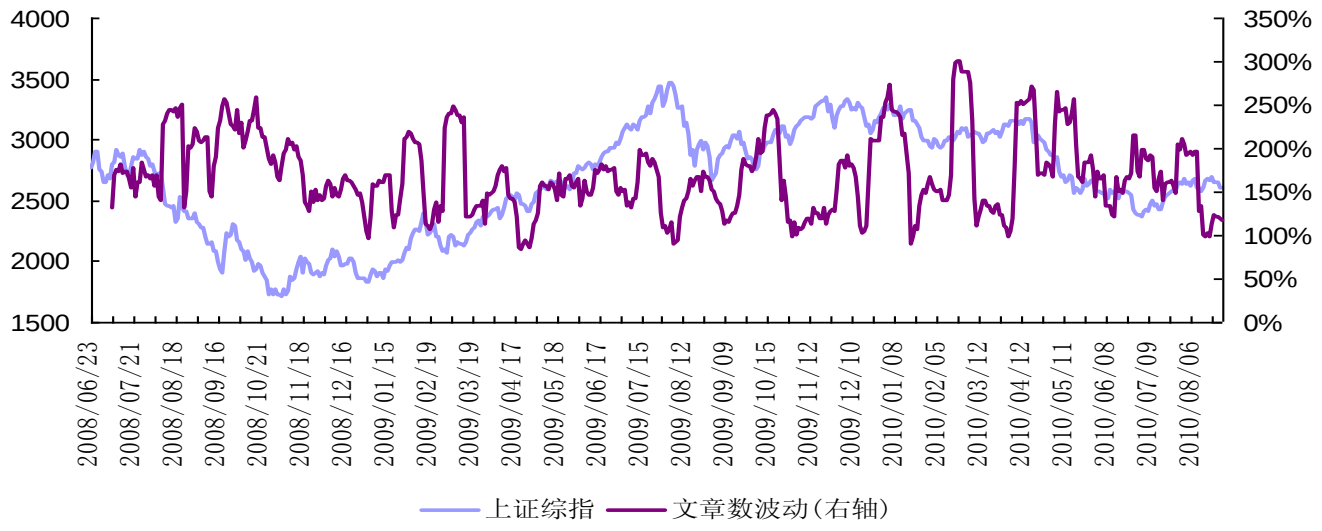
其中，“∧”代表“与”逻辑运算，即“∧”两边条件同时满足，阈值 1.8 与 1.2 的参数来自于历史统计的经验值。波动标志 X 定义公式背后的逻辑是：考察包含当天的过去五个交易日，如果呈现的模式是波动率先高后低，中间一个交易日是最高点，且波动率值高于阈值 1.8，则将当天定义为高波动拐点日，滞后实际拐点两个交易日，低波动拐点日的定义则相反。

如果单纯的以高波动拐点作为卖出信号，低波动拐点作为买入信号，考察其后一个月时间市场的涨跌表现，自 2008-6-23 日至今，共发出信号 25 次，成功率 72%。

如果不考虑高低波动点的不同，单纯的以波动拐点作为市场拐点信号，考察拐点前十个交易日市场表现，如果之前市场上涨，则波动拐点作为卖出信号，如果之前市场下跌，则波动拐点作为买入信号，也考察信号发出后一个月市场的涨跌表现，自 2008-6-23 日至今，同样发出信号 25 次，成功率同样是 72%。

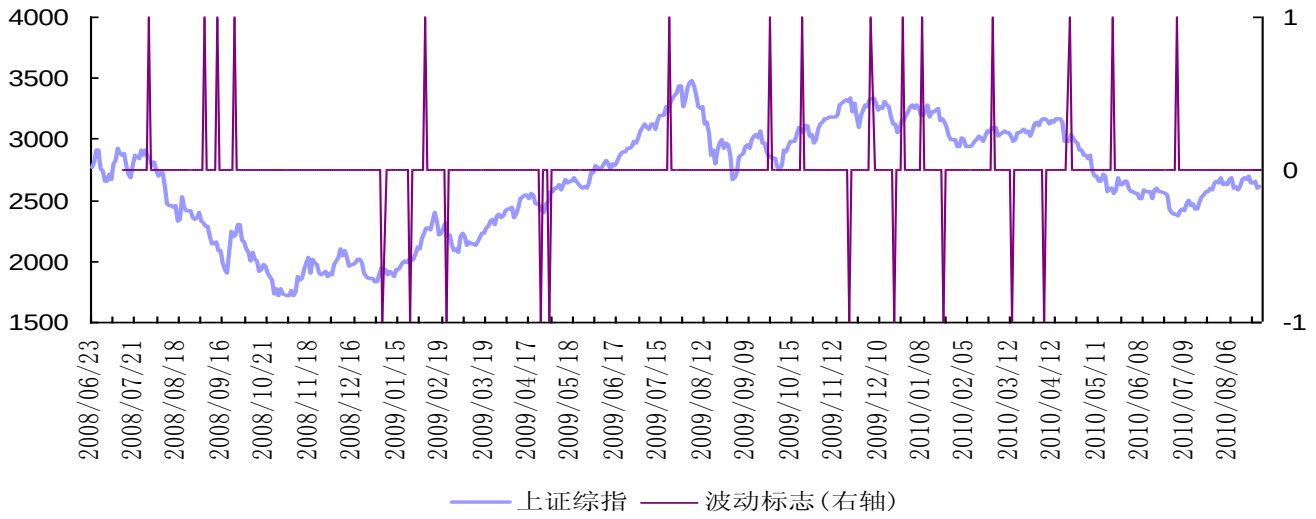
上述两种判断方式，虽然成功率相同，但是具体的成功与失败的情况并不相同，考虑将两种模式结合，即：如果在高波动拐点前十个交易日市场上涨，则为卖出信号，如果在低波动拐点前十个交易日市场下跌，则为买入信号，考察其后一个月市场涨跌表现，成功率为 9 次中的 8 次。当然这种模式有过度拟合之嫌，真正的有效性有待未来不断的跟踪验证。

图 6：新发文章数十天年化历史波动率



资料来源：光大证券研究所

图 7：新发文章数十天年化历史波动率高低拐点



资料来源：光大证券研究所

一个好的指标不仅需要较好的历史拟合结果，更需要背后强大的经济学与金融学逻辑支撑，基于论坛新发文章数量的情绪指标以及在此基础上利用波动性拐点作为买卖信号的逻辑如下：

股票论坛新发文章数量反映了投资者对股票市场的关注程度，在融资融券规模很小、股指期货门槛较高的限制之下，股票现货市场的最主要盈利方式仍是做多，投资者的情绪依然会延续上涨时活跃、下跌时低迷的状态，这就表现在新发文章数指标在走势上与市场基本同步；

A 股有涨跌停板的限制，同时由于市场与统计学两方面的特性（市场在上涨时经常是连续阳线，下跌时更多出现大阴大阳线；统计学上对波动率的计算是一段时间市场每日变化幅度的标准差，意味着上涨时标准差小，下跌时标准差大），市场在下跌时的波动会明显高于上涨时期；新发文章数量则不同，单日的变化幅度较大，且上升和下降呈现巨大的不对称性（可能从 100 篇突然上升到 1000 篇，涨幅就是 900%，但即使从 100 篇降到 0 篇，跌幅最多就 100%，可谓上不封顶、下有极限），这意味着市场上涨时文章数波动高，下跌时文章数波动低；

多数情况下，市场上涨时，文章数波动不断上升，市场催生情绪，情绪则支撑市场不断走高，当市场中短期上涨接近尾声时，文章数波动率呈直线加速上升，但是边际递减，率先出现拐点，意味着没有新的乐观情绪支撑市场，从而，市场在短暂的惯性上涨之后开始进入调整，这是文章数波动的高点稍稍提前于市场的原因；

而在市场下跌时，文章数量下降，同时因为下降比例有最低限度，表现在波动上则是文章数波动率也不断下降，当文章数量下降到一定限度之后，从绝对值来看情绪降到底部，但是持续时间并不能有效判断，此时，波动率非常敏感，因为一旦有文章数量的微小上涨，就会引起波动率的迅速上升，低波动拐点产生，情绪拐点也产生；

上述的几点解释，都是针对大多数情况而言，事实上，在所有测试的高低波动拐点信号中，我们也发现了一些异常情况，例如 2010 年 4 月份以后，几次高波动的拐点都恰恰对应了市场的买点。我们认为，这种现象很有可能就是股指期货的推出在一定程度上改变了现货市场规律的证据：股指期货推出之后，市场情绪的宣泄速度显著加快、但时间缩短，往往在极端情绪刚刚产生时，市场就转向了。反映在文章数的变动上，就是市场在下跌刚进入最后的恐慌阶段时，文章数会突降，引起波动突升，此时市场却见到了中短期的底部。

上述解释如同文章数情绪指标的判断方式一样，很不完美，但一件事物总是在不断的跟踪、观察、修正中逐步完善，我们认为文章数情绪指标至少给了我们另一种观察市场的方法和角度，随数的波动去感受投资者的情绪，把握市场的脉搏，值得我们长期的关注。

3、个股关注度：量化选股新因子

在量化选股模型中，如果加入个股关注度，将会提供一个逻辑上与传统因子独立的新因子，问题是，如何量化的表达关注度指标。研究报告数量是一个最直接能想到的指标：研究报告数量多，说明股票受关注度高，数量少甚至没有，说明受关注度低。但是，无论是选择高关注股还是低关注股，在投资逻辑上都能解释，具体哪种模式比较合适，是需要大量的统计验证的，我们认为，研究报告数量这个指标在目前并不具备大样本统计的条件：

研究报告总数这个值在时间上具有一定的前后连贯性（考虑对报表期效应剔除），但是针对特定股票的值没有时间连续性，从时间序列上来看，这一指标更多的呈现“0-1”的跳跃形态。在这种数据特性下，一方面指标没有合理的区分度，将高中低有效的分组；另一方面，对市场的影响往往要考虑边际，报告数量的边际难以衡量。

为此，基于互联网数据挖掘系统，我们能得到更合理的关注度指标：个股的论坛新发文章数。相对于研究报告数量：80%以上的有效股票的值具有较好的时间连续性，指标区分度较好，边际也容易衡量。

实际统计测试结果表明，在不考虑其它因子的情况下，仅依据关注度和股价因子的选股效果就已经令人满意。

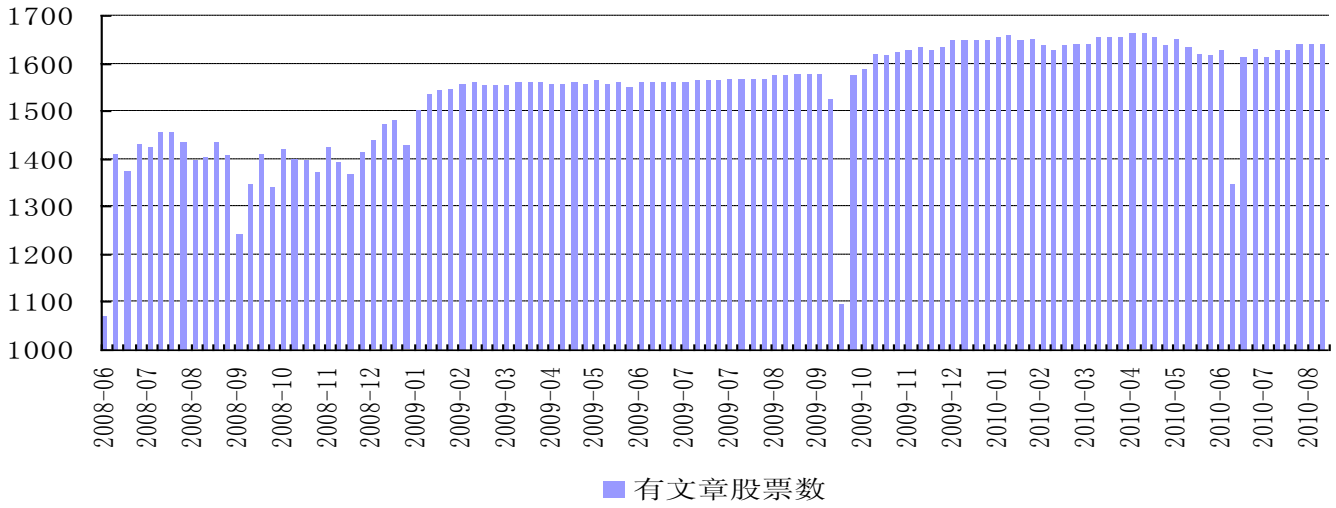
3.1、个股新发文章数量:关注度指标

前文已经提及，互联网论坛的文章往往对应到某一个特定主题或者是具体的某一支股票。将数据汇总之后，容易得到一段时间里某一支股票的新发文章数量，以此作为个股关注度的指标，将具有较高的参考意义。

一个疑问是，是不是会存在大量的没有文章的股票，经过统计，80%以上的股票每周都会有新的文章发出，而剩余近 20%的毫无关注度股票则多数是“ST”、长期停牌等一般选股模型都会剔除的个股。

上述统计也表明，绝大多数股票的关注度具有较好的时间连续性，将关注度作为选股的一个因子，从数据上是可行的，逻辑上也成立。

图 8：每周有新发文章的股票数量超过 80%



资料来源：光大证券研究所

3.2、关注度因子:选择关注度显著下降股

在使用关注度因子时，我们主要考虑关注度的变化，主要是因为相较流通市值小的股票，流通市值较大的股票总体上有更高的关注度绝对值，这种情况下，关注度绝对值参考意义降低，而变化则能很好的反映投资者情绪的改变。

目前，在考虑关注度因子时，只使用关注度和股价表现两个因素，这可以很好的反映关注度因子的效果，其它因子，会在未来模型深入时逐渐加入。

现阶段，基于关注度选股的有效模型有两个，分别对应的操作周期是 20 天和 5 天，无论是哪种模型，一个确定性的结论是：整体上，一个阶段关注度相对于上一阶段显著下降的股票在下一阶段的表现会显著超过关注度显著上升的股票。

20 天周期模型

不考虑创业板和长期停牌股票，在 20 天周期的关注度选股模型中，股票总样本数量为 1519 只，样本期为 2008 年 6 月 23 日至 2010 年 8 月 11 日。

20 天周期模型是为了考察一段时间关注度的变化与未来一段时间股价变化之间的关系。假设目前是 T 日，按每只股票分别计算（T－39）日至（T-20）日的总文章数量 X_1，（T-19）日至 T 日的总文章数 X_2，则过去两期关注度的变化 $\Delta X = (X_2 - X_1)/X_1$（如果 X_1 为 0，即前一期没有文章发出，则当期该股票不在考虑之列）。将所有样本股票的关注度变化按照从小到大，由负到正排序，并按照不同的百分位进行组合，构成十个投资组合，以等权重配置考察（T+1）至（T+20）的组合收益率。

例如，将关注度变化从小到大排序后，以关注度变化的（2%，5%，10%，25%，50%，75%，90%，95%，98%，100%）为百分分割位，则每期可得到十个组合，分别代表关注度上升比例最小的 2%的股票（0%－2%）、

关注度上升比例次小的 2%-5%的股票（2%-5%）…，直至关注度上升最多的 2%的股票（98%-100%）。

根据上述方法，样本有 26 个 20 天周期，扣去前两个周期用以计算第一期关注度变化，则有 24 个收益考察周期，每个周期都能得到按上述百分位切分表的十个投资组合，对每个组合每个周期的收益率进行考察，结果如下：

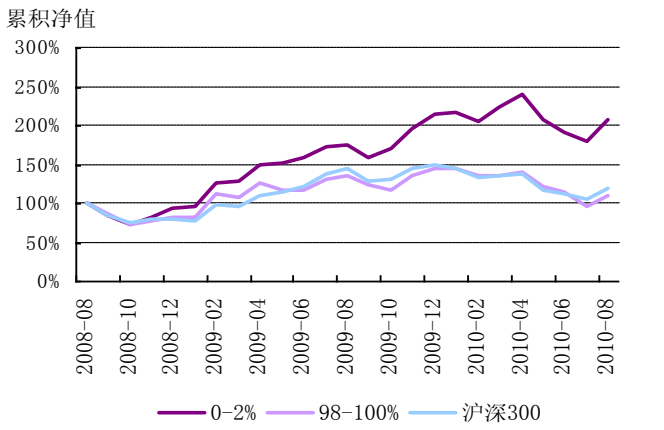
表 1：关注度下降组合显著战胜关注度上升组合

组合	累积收益率	Beta	Alpha (年化)	R Square	战胜基准比率	绝对收益比率	最大单次收益	最大单次亏损
0-2%	107.58%	1.04	36.14%	0.86	66.67%	70.83%	32.05%	-16.81%
2-5%	110.34%	1.09	36.73%	0.89	70.83%	66.67%	31.57%	-15.30%
5-10%	117.14%	1.07	39.36%	0.87	75.00%	66.67%	31.77%	-18.37%
10-25%	98.62%	1.09	32.70%	0.89	70.83%	70.83%	33.37%	-16.18%
25-50%	84.76%	1.07	27.60%	0.90	70.83%	66.67%	32.41%	-16.60%
50-75%	71.91%	1.05	22.82%	0.91	66.67%	62.50%	31.07%	-17.23%
75-90%	59.21%	1.04	18.19%	0.89	66.67%	58.33%	29.62%	-16.70%
90-95%	53.18%	1.01	15.82%	0.87	66.67%	58.33%	33.37%	-16.15%
95-98%	50.33%	1.03	14.92%	0.84	54.17%	58.33%	34.25%	-16.54%
98-100%	9.76%	1.08	-2.58%	0.83	58.33%	50.00%	34.46%	-16.06%

资料来源：光大证券研究所 基准为沪深 300，考察期从 2008 年 8 月 14 日至 2010 年 8 月 4 日

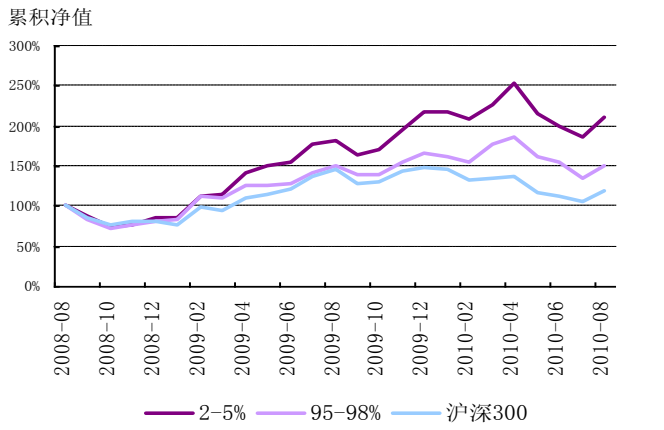
显著的结论是：关注度上升比例最小（下降比例最大）的组合显著跑赢关注度上升比例最大的组合。平均看，关注度上升比例最小的 10%的股票组合两年来录得超过 110%的累积收益率，而且各组合 Beta 均接近于 1，超额收益主要来自于 Alpha。

图 9：主要组合累积净值对比



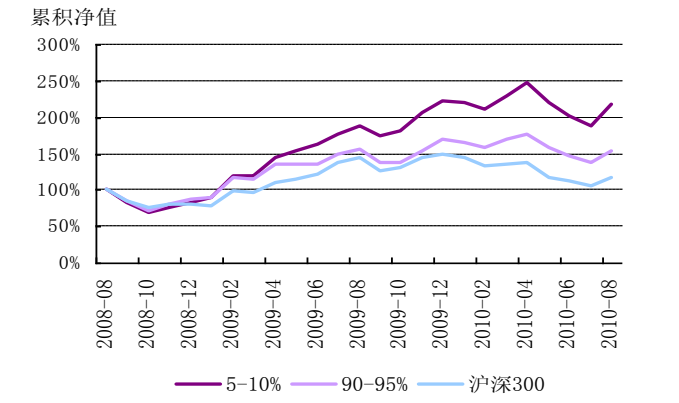
资料来源：光大证券研究所，基准为沪深 300，考察期从 2008 年 8 月 14 日至 2010 年 8 月 4 日。

图 10：主要组合累积净值对比



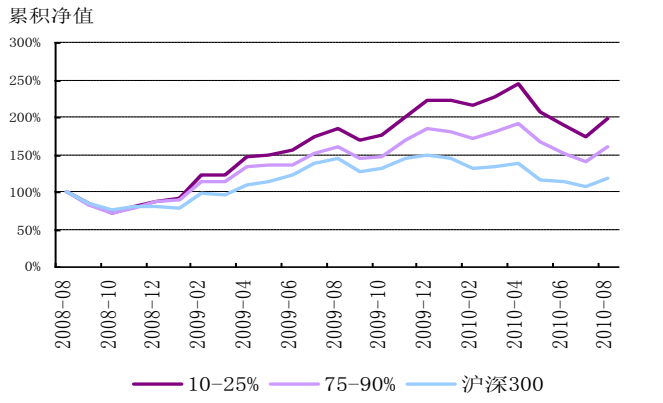
资料来源：光大证券研究所，基准为沪深 300，考察期从 2008 年 8 月 14 日至 2010 年 8 月 4 日。

图 11：主要组合累积净值对比



资料来源：光大证券研究所，基准为沪深 300，考察期从 2008 年 8 月 14 日至 2010 年 8 月 4 日。

图 12：主要组合累积净值对比



资料来源：光大证券研究所，基准为沪深 300，考察期从 2008 年 8 月 14 日至 2010 年 8 月 4 日。

5 天周期“双反转”选股模型

5 天周期模型，我们采用了“双反转”的策略：同时满足关注度上升比例最小和股价表现最差两个条件。

具体的，对于关注度上升比例最小的条件：假设目前是 T 日，按每只股票分别计算（T－9）日至（T-5）日的总文章数量 X_1，（T-4）日至 T 日的总文章数 X_2，则过去两期关注度的变化 $\Delta X = (X_2 - X_1)/X_1$（如果 X_1 为 0，即前一期没有文章发出，则当期该股票不在考虑之列）。将所有样本股票的关注度变化按照从小到大，由负到正排序，取前 5%的股票，即关注度上升比例最小的 5%。

股价表现最差条件：假设目前是 T 日，考察每只股票在（T-4）日至 T 日的收益率 R，按照 R 由低到高、由负到正排序，取前 5%的股票，即当前五日表现最差的 5%。

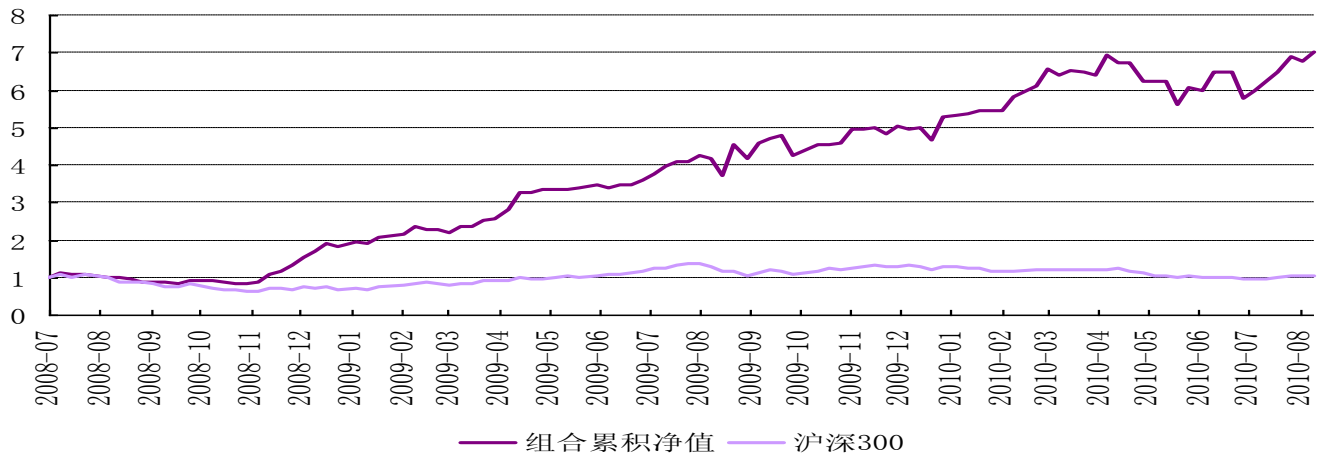
将同时满足上述两个条件的股票选出，构成等权重组合，考察（T+1）日至（T+20）日的市场表现。自 2008-7-3 日至 2010-8-11 日，累积 103 个收益率考察期，有 18 个周期没有股票入选，这 18 个周期我们假设持有现金，而考虑累积收益率的话，按照上述方法选出的组合，在两年多时间里，累积收益率为 658%，与沪深 300 组合之间的 Beta 是 0.78，但是 R Square 只有 0.38，显然，说明该方式做出的组合与市场之间的相关性随着时间的变化较大。

表 2：关注度与股价双反转组合

组合	累积收益率	Beta	Alpha (年化)	R Square	战胜基准比率	绝对收益比率	最大单次收益	最大单次亏损	平均股票数量
双反转	658.00%	0.78	295.00%	0.38	67%	66%	26.29%	-11.38%	2.12

资料来源：光大证券研究所 基准为沪深 300

图 13：关注度与股价双反转组合净值表现



资料来源：光大证券研究所　　基准为沪深 300

表 3：双反转模型历史组合 日期代表选出股票的 T 日，收益率考察（T+1）日至（T+5）日

日期	股票一	股票二	股票三	股票四	股票五	股票六	R1	R2	R3	R4	R5	R6	组合收益
2008-07-03	岳阳兴长						13.26%						13.26%
2008-07-10	兴化股份	熊猫烟花	鼎立股份	中粮屯河			-2.37%	-9.18%	-3.51%	-6.61%			-5.42%
2008-07-17													
2008-07-24	四川湖山	拓邦股份	东方银星	成商集团			-3.89%	5.51%	-4.98%	-5.70%			-2.27%
2008-07-31	泛海建设	滨海能源					-2.83%	-6.36%					-4.60%
2008-08-07													
2008-08-14	中航三鑫	大元股份	中电广通				-9.08%	-2.25%	-4.51%				-5.28%
2008-08-21	国兴地产						-9.69%						-9.69%
2008-08-28	华天科技	宁夏恒力					1.55%	1.74%					1.64%
2008-09-04													0.00%
2008-09-11	天茂集团	顺鑫农业					-2.46%	3.34%					0.44%
2008-09-19	宁波富邦						17.83%						17.83%
2008-09-26	绿大地						3.10%						3.10%
2008-10-10	北京旅游	航天科技	江特电机	中国卫星	熊猫烟花	南京化纤	-18.07%	-10.09%	-16.23%	-1.06%	20.96%	-6.96%	-5.24%
2008-10-17		民和股份	安徽合力					1.40%	-6.56%				-2.58%
2008-10-24													0.00%
2008-10-31	航天机电						2.55%						2.55%
2008-11-07	思源电气	国金证券					33.70%	18.88%					26.29%
2008-11-14		运盛实业						20.62%					20.62%
2008-11-21	黑猫股份	中国船舶					28.44%	-2.17%					13.13%
2008-11-28	天山纺织		中航精机	延华智能	贵航股份		13.70%		19.22%	12.96%	28.42%		18.57%
2008-12-05	中通客车	安源股份		西藏城投			-2.08%	9.84%		19.28%			9.01%
2008-12-12	重庆港九	老白干酒	新华锦				24.18%	9.76%	4.97%				12.97%
2008-12-19	兴蓉投资	西藏城投	南京熊猫				-12.24%	6.09%	-9.57%				-5.24%
2008-12-26	厦门港务	康强电子	东睦股份				1.46%	7.79%	12.93%				7.39%
2009-01-06	酒鬼酒	东湖高新					0.35%	-2.68%					-1.17%
2009-01-13	大通燃气	万力达	中国软件	天地科技			1.75%	7.15%	6.23%	14.33%			7.36%
2009-01-20	丽珠集团	海螺型材	科陆电子				2.43%	3.77%	7.95%				4.72%
2009-02-03	深国商	江山化工	新海股份		银座股份		10.68%	14.31%	10.62%		9.00%		11.15%

2009-02-10	鱼跃医疗						-3.27%					-3.27%
2009-02-17	利欧股份	江南化工	哈药股份				-5.75%	-2.97%	6.59%			-0.71%
2009-02-24	金山开发						-3.37%					-3.37%
2009-03-03	云铝股份		世荣兆业	栋梁新材			6.36%		2.42%	13.06%		7.28%
2009-03-10												0.00%
2009-03-17	中兴商业	南京港	成商集团				3.44%	14.12%	1.15%			6.24%
2009-03-24	北方国际	国统股份					1.83%	3.95%	0.00%			2.89%
2009-03-31	光华控股	华东数控	浙江广厦	西南药业	工大首创		31.92%	1.61%	15.91%	0.25%	-3.13%	9.31%
2009-04-08		南纺股份	世茂股份					8.02%	18.66%			13.34%
2009-04-15												0.00%
2009-04-22	海隆软件	证通电子	昌九生化	金证股份			0.26%	9.58%	2.63%	-2.24%		2.56%
2009-04-29												0.00%
2009-05-07												0.00%
2009-05-14	欧亚集团	浪潮软件					-0.65%	3.11%				1.23%
2009-05-21	天保基建	新海股份	成发科技				1.04%	3.27%	2.21%			2.17%
2009-06-01	大洋电机						-7.40%					-7.40%
2009-06-08	精功科技	荣信股份	爱使股份				0.14%	2.08%	2.97%			1.73%
2009-06-15	新都酒店	盐湖集团	得润电子	常铝股份			-1.39%	3.85%	0.79%	-0.62%		0.66%
2009-06-22		新民科技	林海股份	浪潮软件			0.00%	2.72%	3.16%	-1.78%		1.37%
2009-06-29	闽福发 A	普洛股份	华联股份	达安基因	交大昂立	上海辅仁	5.66%	2.60%	8.87%	9.60%	2.82% 1.01%	5.09%
2009-07-06	鑫茂科技	正邦科技					5.10%	5.04%				5.07%
2009-07-13	珠江实业						2.72%					2.72%
2009-07-20	古井贡酒	阳光城	天地科技		王府井		6.87%	-2.76%	0.89%		-2.23%	0.55%
2009-07-27	北巴传媒						3.39%					3.39%
2009-08-03	安源股份						-1.65%					-1.65%
2009-08-10	莱茵生物	桂东电力					-12.43%	-9.11%				-10.77%
2009-08-17	宝石 A	太原刚玉	新嘉联	天目药业			13.32%	46.28%	22.59%	5.21%		21.85%
2009-08-24	陆家嘴						-8.42%					-8.42%
2009-08-31	平煤股份						10.30%					10.30%
2009-09-07	万丰奥威	诺普信					4.78%	1.33%				3.05%
2009-09-14	安纳达	美罗药业	安琪酵母	信雅达	南京熊猫	鲁抗医药	0.47%	6.80%	-2.29%	1.78%	-0.76% 1.41%	1.23%
2009-09-21	有研硅股	红星发展					-11.43%	-11.33%				-11.38%
2009-09-28	高金食品	中航三鑫	贵研铂业	双钱股份	三爱富	上海普天	9.03%	9.52%	10.85%	-0.70%	7.26% 8.39%	7.39%
2009-10-13												0.00%
2009-10-20	南京中商						0.37%					0.37%
2009-10-27		梅花伞						9.34%				9.34%
2009-11-03												0.00%
2009-11-10	瑞贝卡						0.53%					0.53%
2009-11-17												0.00%
2009-11-24	奥特迅	亚宝药业	双良节能	欧亚集团			5.72%	3.09%	0.34%	7.94%		4.27%
2009-12-01	准油股份						-1.67%					-1.67%
2009-12-08	海印股份	成发科技	江中药业	五洲明珠			5.14%	0.25%	-0.83%	1.09%		1.41%
2009-12-15	卧龙地产	园城股份					-6.27%	-7.97%				-7.12%
2009-12-22	昆百大 A	新赛股份					8.14%	18.16%				13.15%
2009-12-29	中材科技						1.09%					1.09%
2010-01-06	三房巷		龙元建设				4.41%		-1.56%			1.42%
2010-01-13	苏泊尔						1.91%					1.91%
2010-01-20												0.00%
2010-01-27												0.00%

2010-02-03												0.00%
2010-02-10	沙隆达 A	鸿博股份					3.65%	7.70%				5.67%
2010-02-24												0.00%
2010-03-03	东阿阿胶	广发证券	凯迪电力	太化股份	国阳新能	江中药业	-3.79%	-1.71%	-5.70%	-1.70%	-1.10%	-1.28%	-2.55%
2010-03-10	禾嘉股份	九龙电力	大厦股份	士兰微	友好集团		2.09%	2.74%	1.32%	3.54%	1.49%		2.24%
2010-03-17	欧亚集团						-0.72%						-0.72%
2010-03-24	江苏宏宝	三房巷	合肥三洋				-1.53%	-1.16%	-0.96%				-1.22%
2010-03-31	兴发集团	大湖股份			文山电力		11.92%	8.22%			9.63%		9.92%
2010-04-08	苏州高新						-2.53%						-2.53%
2010-04-15													0.00%
2010-04-22	丹化科技						-7.52%						-7.52%
2010-04-29													0.00%
2010-05-07													0.00%
2010-05-14	芭田股份						-9.95%						-9.95%
2010-05-21	露天煤业	金飞达					2.03%	13.33%					7.68%
2010-05-28	得润电子	北京城建	亿利能源	华发股份			3.65%	-2.29%	-2.28%	-2.86%			-0.94%
2010-06-04	西藏发展	飞马国际	柳化股份	湘邮科技			5.25%	14.62%	11.53%	1.24%			8.16%
2010-06-11	安妮股份						-1.12%						-1.12%
2010-06-23	雪莱特	中新药业	国电南瑞	千金药业			-16.76%	-8.42%	-8.18%	-10.66%			-11.00%
2010-06-30	华联股份						3.72%						3.72%
2010-07-07	西南合成	鲁润股份					5.67%	2.89%					4.28%
2010-07-14	世荣兆业	荣信股份	绿大地				4.54%	2.64%	4.29%	0.00%			3.82%
2010-07-21	深深宝 A	华润锦华	三特索道	长春一东			9.17%	4.74%	5.75%	5.21%			6.22%
2010-07-28	西昌电力						-1.45%						-1.45%
2010-08-04	深长城	国电南自	美克股份	三普药业			-1.56%	0.05%	15.48%	0.35%			3.58%

资料来源：光大证券研究所

分析师声明

负责准备本报告以及撰写本报告的所有研究分析师或工作人员在此保证，本研究报告中关于任何发行商或证券所发表的观点均如实反映分析人员的个人观点。负责准备本报告的分析师获取报酬的评判因素包括研究的质量和准确性、客户的反馈、竞争性因素以及光大证券股份有限公司的整体收益。所有研究分析师或工作人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

行业及公司评级体系

买入—未来 6-12 个月的投资收益率领先市场基准指数 15%以上；
增持—未来 6-12 个月的投资收益率领先市场基准指数 5%至 15%；
中性—未来 6-12 个月的投资收益率与市场基准指数的变动幅度相差-5%至 5%；
减持—未来 6-12 个月的投资收益率落后市场基准指数 5%至 15%；
卖出—未来 6-12 个月的投资收益率落后市场基准指数 15%以上；
无评级—因无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使无法给出明确的投资评级。
市场基准指数为沪深 300 指数。

分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

特别声明

光大证券股份有限公司（以下简称“本公司”）创建于 1996 年，系由中国光大（集团）总公司投资控股的全国性综合类股份制证券公司，是中国证监会批准的首批三家创新试点公司之一。公司经营业务许可证编号：z22831000。

公司经营范围：证券经纪；证券投资咨询；与证券交易、证券投资活动有关的财务顾问；证券承销与保荐；证券自营；为期货公司提供中间介绍业务；证券投资基金代销；融资融券业务；中国证监会批准的其他业务。此外，公司还通过全资或控股子公司开展资产管理、直接投资、期货、基金管理以及香港证券业务。

本证券研究报告由光大证券股份有限公司研究所（以下简称“光大证券研究所”）编写，以合法获得的我们相信为可靠、准确、完整的信息为基础，但不保证我们所获得的原始信息以及报告所载信息之准确性和完整性。光大证券研究所可能将不时补充、修订或更新有关信息，但不保证及时发布该等更新。

本报告根据中华人民共和国法律在中华人民共和国境内分发，仅供本公司的客户使用。

本报告中的资料、意见、预测均反映报告初次发布时光大证券研究所的判断，可能需随时进行调整。报告中的信息或所表达的意见不构成任何投资、法律、会计或税务方面的最终操作建议，本公司不就任何人依据报告中的内容而最终操作建议作出任何形式的保证和承诺。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问或金融产品等相关服务。投资者应当充分考虑本公司及本公司附属机构就报告内容可能存在的利益冲突，不应视本报告为作出投资决策的唯一参考因素。

在任何情况下，本报告中的信息或所表达的建议并不构成对任何投资人的投资建议，本公司及其附属机构（包括光大证券研究所）不对投资者买卖有关公司股份而产生的盈亏承担责任。

本公司的销售人员、交易人员和其他专业人员可能会向客户提供与本报告中观点不同的口头或书面评论或交易策略。本公司的资产管理部和投资业务部可能会作出与本报告的推荐不相一致的投资决策。本公司提醒投资者注意并理解投资证券及投资产品存在的风险，在作出投资决策前，建议投资者务必向专业人士咨询并谨慎抉择。

光大证券股份有限公司研究所 销售交易部 机构业务部

上海市新闸路 1508 号静安国际广场 3 楼 邮编 200040

总机：021-22169999 传真：021-22169114、22169134

销售交易部	姓名	办公电话	手机	电子邮件
北京	郝辉	010-58452028	-	haohui@ebscn.com
	黄怡	010-58452027	-	huangyi@ebscn.com
	梁晨	-	-	liangchen@ebscn.com
	刘公直	010- 58452029	-	liugongzhi@ebscn.com
上海	严非	021-22169086	-	yanfei@ebscn.com
	周薇薇	021-22169087	-	zhouww1@ebscn.com
	徐又丰	021-22169082	-	xuyf@ebscn.com
	李强	021-22169131		liqiang88@ebscn.com
	奚亦扬	021-22169091		xiyy@ebscn.com
	张弓	021-22169083		zhanggong@ebscn.com
	罗德锦	021-22169146	-	luodj@ebscn.com
深圳	黎晓宇	0755-83553559	-	lixy1@ebscn.com
	黄鹂华	0755-83553249	-	huanglh@ebscn.com
	李潇	0755-83559378	-	lixiao1@ebscn.com
	张亦潇	0755-23996409	-	zhangyx@ebscn.com
	王渊锋		-	wangyuanfeng@ebscn.com
机构客户业务部	**姓名**	**办公电话**	**手机**	**电子邮件**
	濮维娜(总经理)	021-62152373	13611990668	puwn@ebscn.com
上海	张辉	021-22167108	13611990668	zhanghui1@ebscn.com
	计爽	021-22167101	18621181721	jishuang@ebscn.com
	吉喆	021-22169129	18918212345	jizhe@ebscn.com
北京	朱林	010-59046212	18611386181	zhulin1@ebscn.com
	徐放	010-56513051	18618469955	xufang@ebscn.com
国际业务	陶奕(副总经理)	021-62152393	18018609199	taoyi@ebscn.com
	戚德文(执行董事)	021-22169152	13585893550	qidw@ebscn.com
	顾胜寒	021-22167094	18352760578	gush@ebscn.com