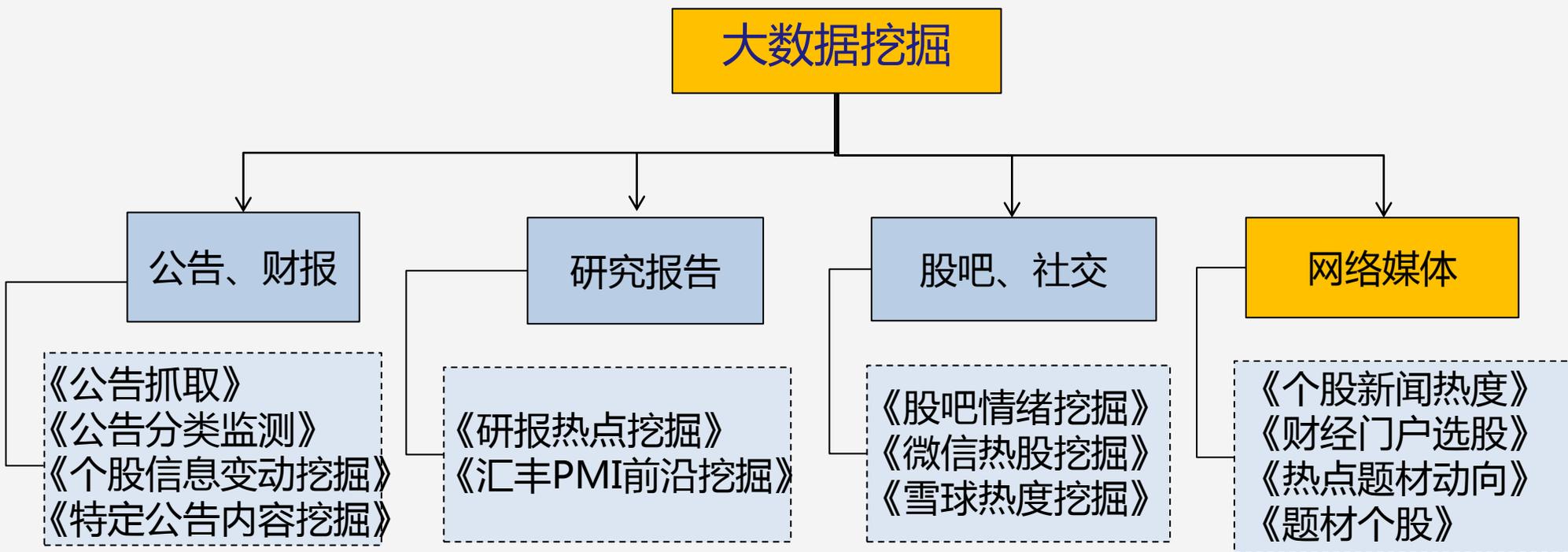


基于大数据挖掘的关联个股投资机会

——互联网大数据挖掘之六

史庆盛 S0260513070004
广发证券金融工程
2015年8月







01

02

03

04

01

| 关注度简介 |



排名趋势

排名	股票代码	关注度变化值	关注度变化幅度	排名趋势
1	601857	23990 ↑		119% ↑
2	600028			-8% ↓
3	6007...			
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				

大家还关注

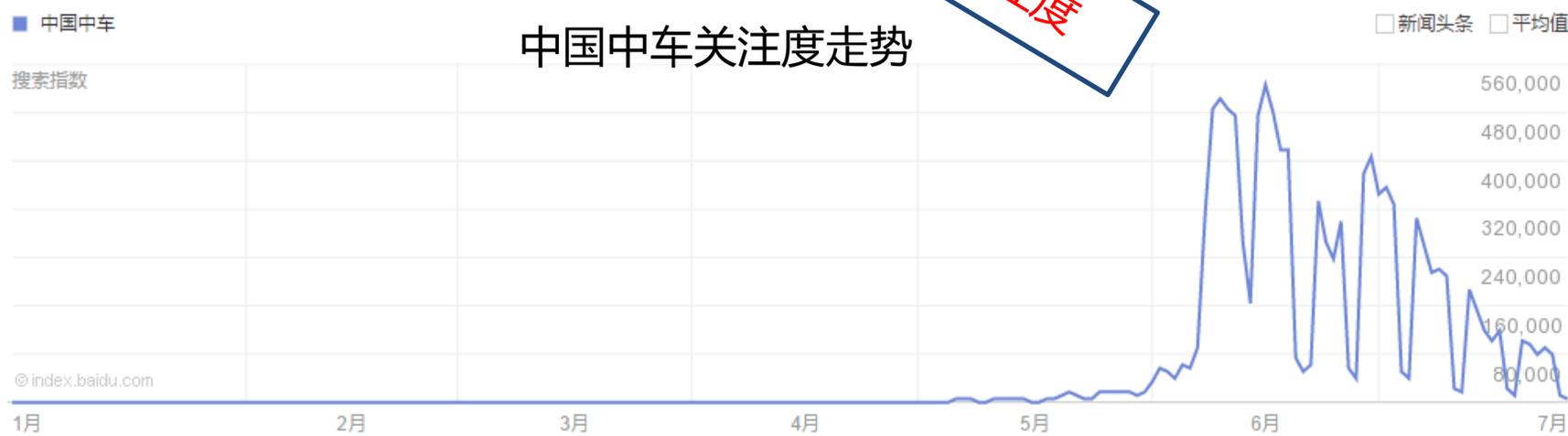
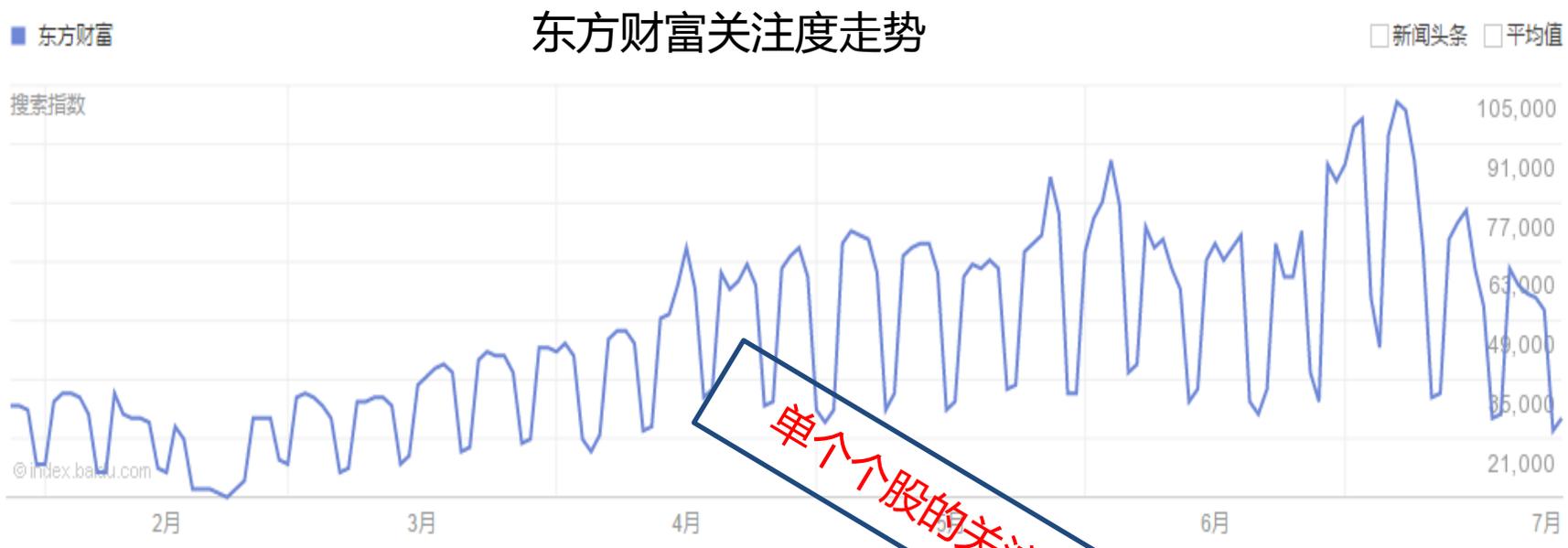
股票	关注度	比例
工商银行	4.86	比例9.76%
中信银行	8.40	比例6.41%
交通银行	6.96	比例5.94%
五粮液	25.13	比例5.69%
中国平安	35.92	
建设银行	6.11	
中信证券	22.04	
华夏银行	11.04	

关注浦发银行的人还关注了

共同关注	比例
东方财富共同关注1005800人	比例9.76%
民生证券共同关注600390人	比例6.41%
兴业银行共同关注601689人	比例5.94%
苏宁云商共同关注586175人	比例5.69%

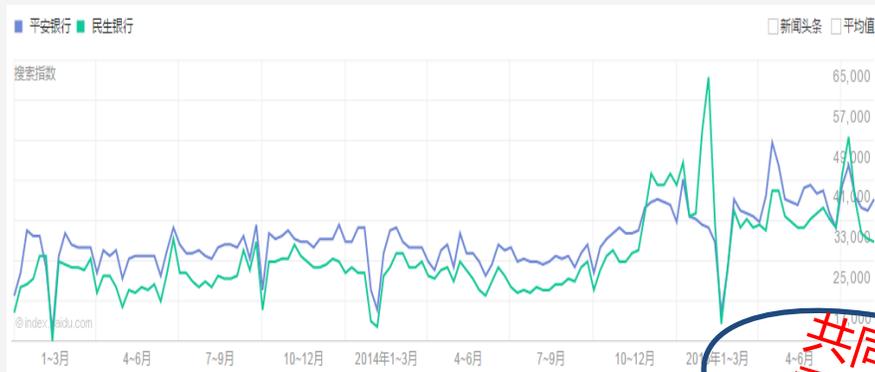
搜索指数

2015年6月27日



基于大数据挖掘的关联个股投资机会

数据来源：广发证券发展研究中心、百度指数



我的自选股

编辑股票代码: 定制完成 *多只股票用逗号“,”分割,支持代码/拼音缩写/中文名,帮助

股票代码	股票名称	当前价	今低-今高	涨跌幅	涨跌幅	现手	总手	换手率	成交金额	买入价	收益率	买入股数	浮动盈亏
300059	东方财富	46.71	44.44-48.15	+2.92	6.67%	7950	828214	6.41%	385629	点击输入		点击输入	X
000002	万科A	14.60	14.38-14.71	+0.26	1.81%	7227	866519	0.89%	126459	点击输入		点击输入	X
600016	民生银行	9.33	9.28-9.47	-0.03	-0.32%	28	1468553	0.50%	137584	点击输入		点击输入	X
600010	浦发银行	15.47	15.39-15.64	+0.09	0.59%	205	653023	0.44%	101425	点击输入		点击输入	X

共同关注，注重投资者关注度的切换

自选股

我的自选 资金流向 DDE决策 盈利预测 财务数据 多股同列 盈亏一览

大字 小字 批量设置:

提醒	代码	名称	相关链接	最新	涨幅	涨跌	总手	现手	买入价	卖出价	换手	金额	市盈率(动)	最高	最低	开盘	笔记	删除
	300379	东方财富	股吧 资金研报	70.66	7.26%	4.78	3.35万	230	70.61	70.66	5.71%	2.31亿	-	70.88	66.02	69.00		X
	510300	300ETF	股吧 资金研报	3.953	2.07%	0.08	685.89万	45	3.95	3.95	-	27.06亿	-	3.98	3.90	3.90		X
	300059	东方财富	股吧 资金研报	46.71	6.67%	2.92	82.82万	7950	46.70	46.71	6.41%	38.56亿	99.85	48.15	44.44	44.65		X

沪深自选 自选 港自选 美股自选 股价提醒 ^{NEW} 休市 8月9日 星期天 01:02:32 行情时间: 08-07 15:04:02

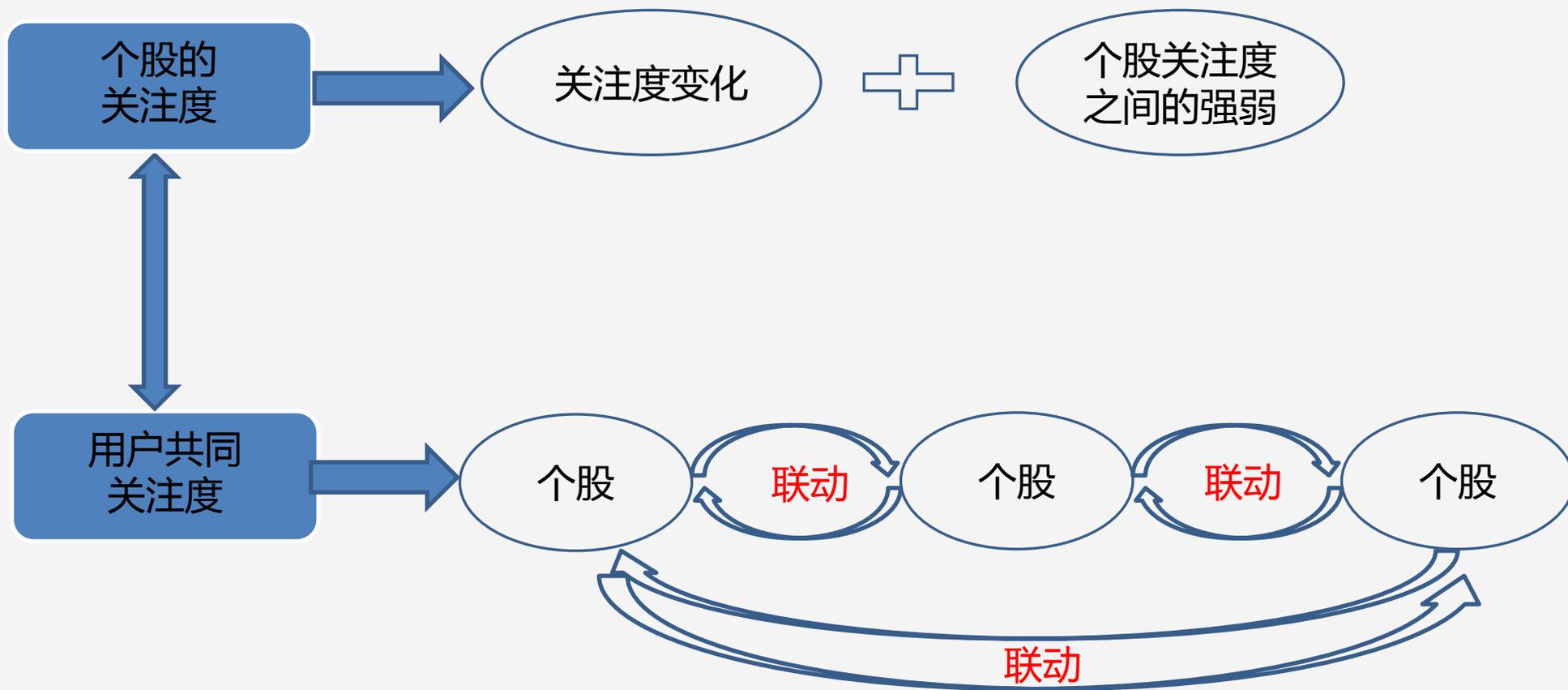
投资组合 我的组合

当日走势 基本面 盈利一览 交易记录 默认排序中,点击表头可更改排序状态

名称	代码	最新价	涨跌额	涨跌幅	买入/卖出价	成交量(手)	成交额(万)	换手率	昨收/今开盘	最高/最低价	股票吧	操作
广发证券	sz000776	18.37	+0.17	+0.93%	18.36/18.37	364955	67280	0.62%	18.20/18.20	18.69/18.19	股票吧	X
海通证券	sh600837	16.49	+0.11	+0.67%	16.50/16.51	1359710	224995	1.68%	16.38/16.40	16.78/16.30	股票吧	X
中国中车	sh601766	15.42	+0.17	+1.11%	15.43/15.44	2442719	376796	1.07%	15.25/15.30	15.59/15.25	股票吧	X
中国平安	sh601318	33.89	+1.00	+3.04%	33.87/33.88	1524162	515694	1.41%	32.89/33.34	34.24/33.15	股票吧	X
中国重工	sh601989	14.29	+0.41	+2.95%	14.29/14.30	8012037	1127129	4.46%	13.88/14.05	14.36/13.70	股票吧	X

数据来源：广发证券发展研究中心、百度指数、搜狐网、新浪网、东方财富

基于大数据挖掘的关联个股投资机会



用户关注个股数据

- 时效性强(每日更新)
- 消除个体关注偏差(海量数据)
- 反应用户的注意力(有限注意力)

共同关注个股指标

- 共同关注人数(反映用户关注绝对量)
- 共同关注比例(反映用户关注相对占比, 消除基准影响)

关联个股选股策略

- 动态考虑基准个股所在行业与关联个股所在行业相关性
- 考虑基准个股与关联个股的涨跌幅、成交量等关系

数学定义

假设市场共有 n 个关注者以及 m 只个股，
市场关注度矩阵 AM 如下所示：

$$AM = \begin{bmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1m} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2m} \\ \xi_{i1} & \vdots & \xi_{ij} & \vdots \\ \xi_{n1} & \xi_{n2} & \xi_{nj} & \xi_{nm} \end{bmatrix}$$

其中矩阵中元素为布尔变量， ξ_{ij} 表示第 i 个关注者对股票 j 的关注，关注则为 1，否则为 0。
 关注股票 j 的总人数为： $\sum_{i=1}^n \xi_{ij}$ ，关注股票 j 同时关注股票 k 的总人数为： $\sum_{i=1}^n \xi_{ij} * \xi_{ik}$ ，关注股票 j 同时关注股票 k 占关注股票 j 总人数比例为： $\frac{\sum_{i=1}^n \xi_{ij} * \xi_{ik}}{\sum_{i=1}^n \xi_{ij}}$



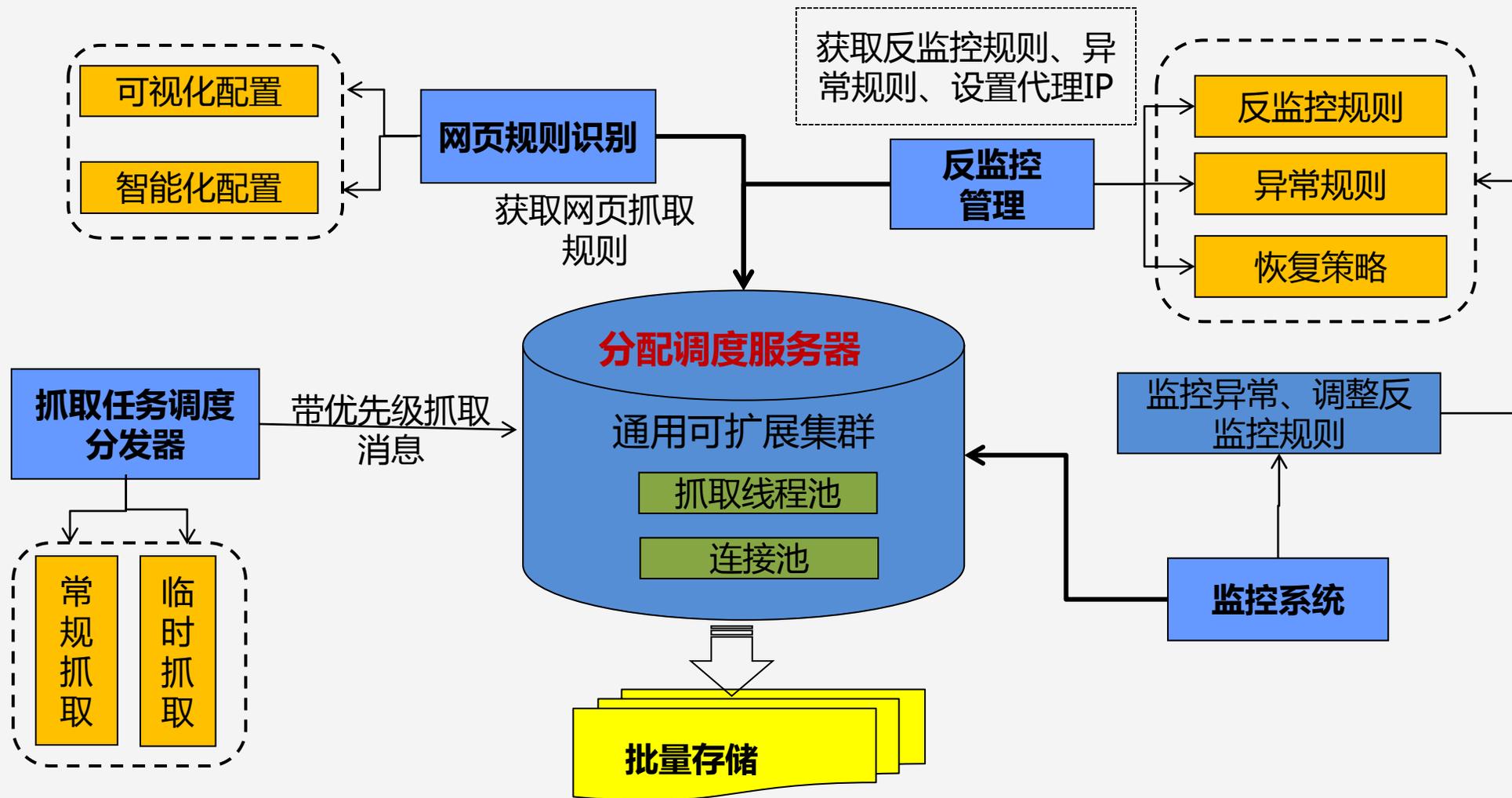
基于大数据挖掘的大众量化投资云平台



02

| 策略构建 |



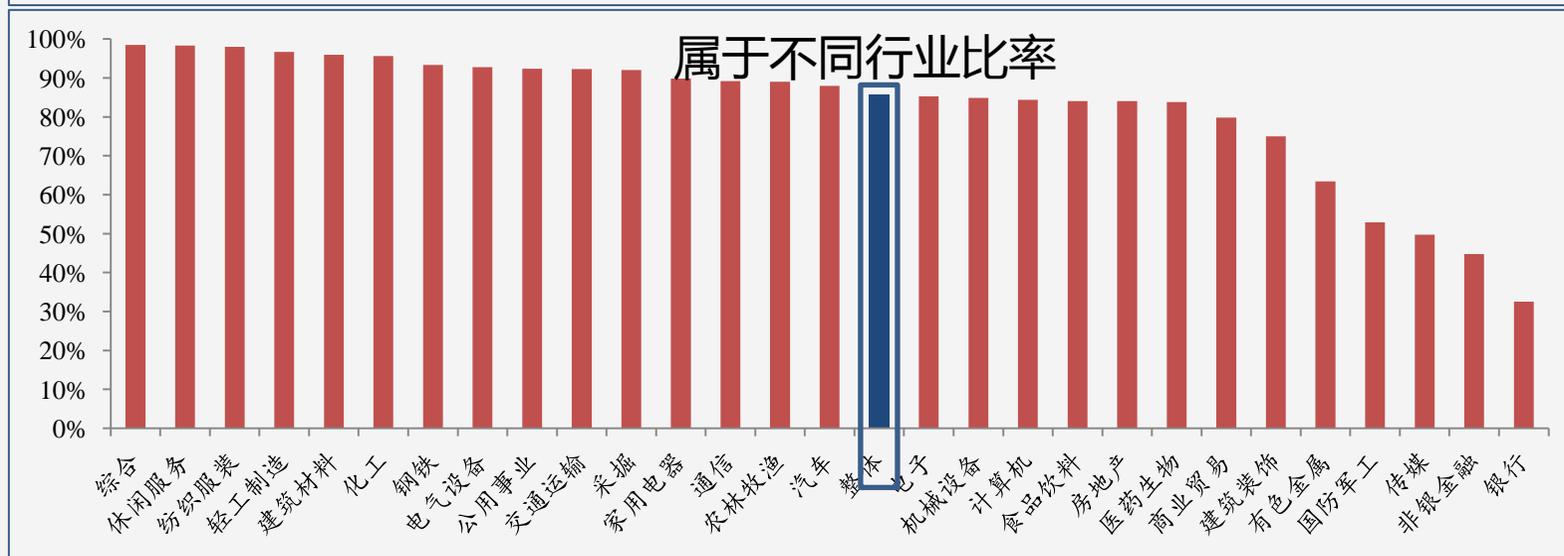
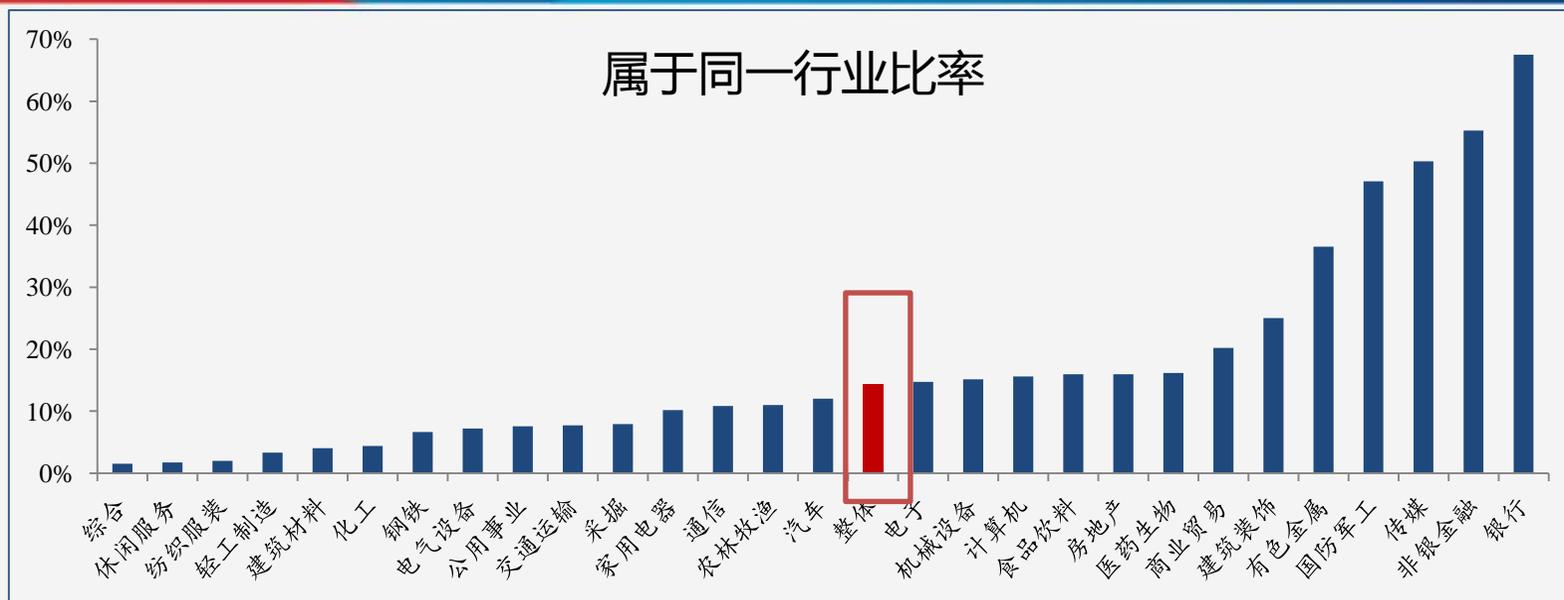


基于大数据挖掘的关联个股投资机会

基准个股	关联个股	共同关注绝对量	共同关注占比
平安银行	民生银行	502457	17.31%
	浦发银行	497372	17.14%
	万科A	476150	16.41%
	兴业银行	453587	15.63%
	中信证券	452849	15.60%
广发证券	中信证券	415985	22.82%
	海通证券	311162	17.07%
	中国中车	302324	16.59%
	中国平安	263957	14.48%
	中国重工	260114	14.27%
东方财富	浦发银行	1005800	5.25%
	中国中车	744233	3.89%
	乐视网	616784	3.22%
	中国重工	593744	3.10%
	苏宁云商	574106	3.00%

数据来源：广发证券发展研究中心、互联网

注：数据截止至2015年7月29日

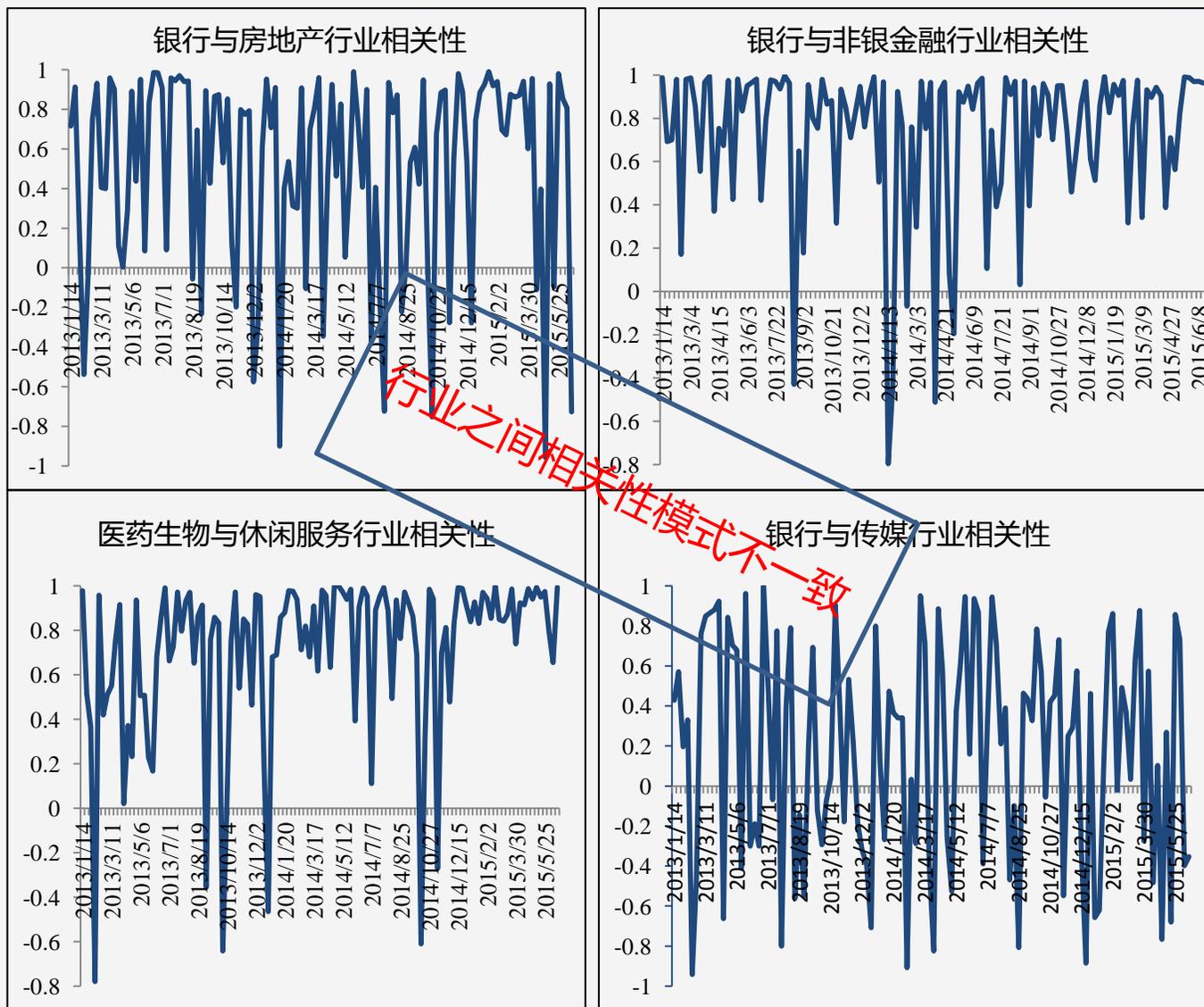


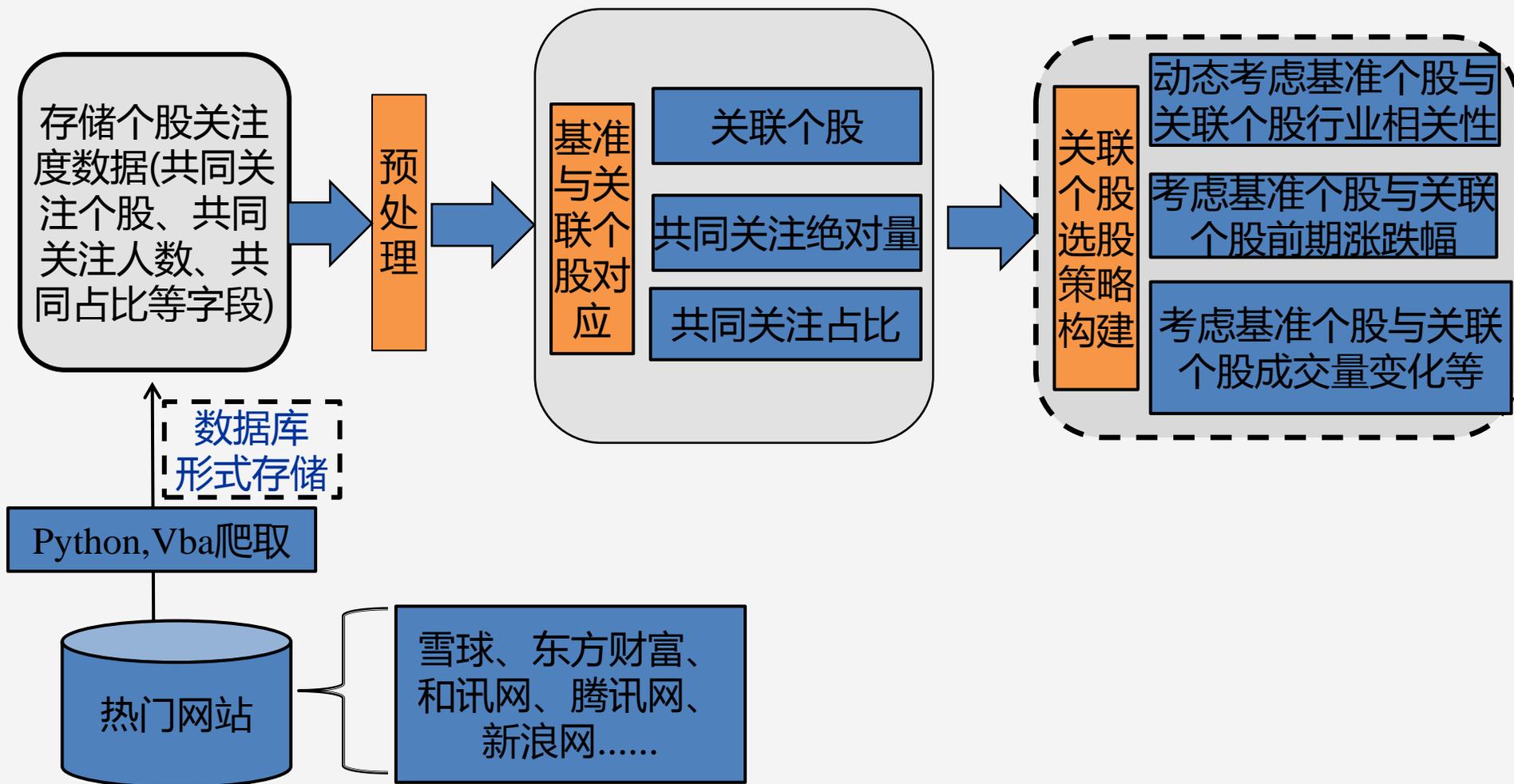
基准行业	共同关注行业 占比最高行业	比例	共同关注行业 占比最高行业	比例	共同关注行业 占比最高行业	比例
银行	银行	67.50%	非银金融	17.50%	房地产	6.25%
房地产	非银金融	17.92%	商业贸易	16.53%	房地产	15.97%
医药生物	传媒	17.51%	非银金融	17.40%	医药生物	16.20%
休闲服务	非银金融	18.86%	商业贸易	18.29%	银行	11.43%
综合	商业贸易	19.62%	非银金融	13.58%	有色金属	10.94%
建筑材料	非银金融	17.68%	商业贸易	17.10%	银行	10.72%
家用电器	商业贸易	18.11%	非银金融	14.34%	银行	12.08%
汽车	商业贸易	16.76%	非银金融	15.48%	汽车	12.02%
食品饮料	非银金融	18.55%	食品饮料	15.94%	商业贸易	13.91%
电子	传媒	22.70%	商业贸易	15.27%	电子	14.73%
计算机	传媒	37.06%	计算机	15.59%	商业贸易	12.79%
交通运输	建筑装饰	21.14%	非银金融	16.59%	商业贸易	12.95%
轻工制造	商业贸易	16.70%	非银金融	14.51%	传媒	9.01%
公用事业	非银金融	18.15%	建筑装饰	12.78%	银行	12.41%

数据来源：广发证券发展研究中心、互联网

基准行业	共同关注行业 占比最高行业	比例	共同关注行业 占比最高行业	比例	共同关注行业 占比最高行业	比例
通信	传媒	26.56%	商业贸易	16.07%	通信	10.82%
机械设备	机械设备	15.12%	商业贸易	13.68%	非银金融	13.20%
农林牧渔	商业贸易	19.02%	非银金融	17.07%	农林牧渔	10.98%
建筑装饰	建筑装饰	25.00%	非银金融	18.53%	商业贸易	12.35%
商业贸易	商业贸易	20.21%	非银金融	18.96%	银行	17.29%
化工	商业贸易	17.60%	非银金融	14.64%	有色金属	10.08%
有色金属	有色金属	36.53%	非银金融	14.29%	商业贸易	12.86%
传媒	传媒	50.29%	商业贸易	13.53%	非银金融	11.18%
纺织服装	商业贸易	19.00%	非银金融	17.50%	银行	9.75%
采掘	非银金融	17.14%	商业贸易	13.33%	银行	13.02%
非银金融	非银金融	55.26%	银行	21.05%	商业贸易	6.84%
电气设备	商业贸易	15.87%	非银金融	14.84%	传媒	11.87%
钢铁	建筑装饰	30.30%	非银金融	16.97%	银行	9.09%
国防军工	国防军工	47.10%	非银金融	11.61%	机械设备	10.97%

数据来源：广发证券发展研究中心、互联网





基于大数据挖掘的关联个股投资机会

- ◆ 在历史回测期，定期地计算基准个股中共同关注个股的关注人数以及共同关注占比，选取共同关注的个股中关注度最高的前N只个股，动态地考虑基准个股与对应的共同关注个股所在的行业的相关性，根据行业之间的相关性以及基准个股与关联个股前一段时间的涨跌幅等因素，选择满足条件的关联个股为多头组合，同时以基准个股为空头组合；
- ◆ 基于构建的多空组合，在下一个交易日以开盘价做多多头组合，以开盘价做空空头组合，考虑涨跌停因素的影响；
- ◆ 初始资金为1，资金等权投资；
- ◆ 周频调仓；

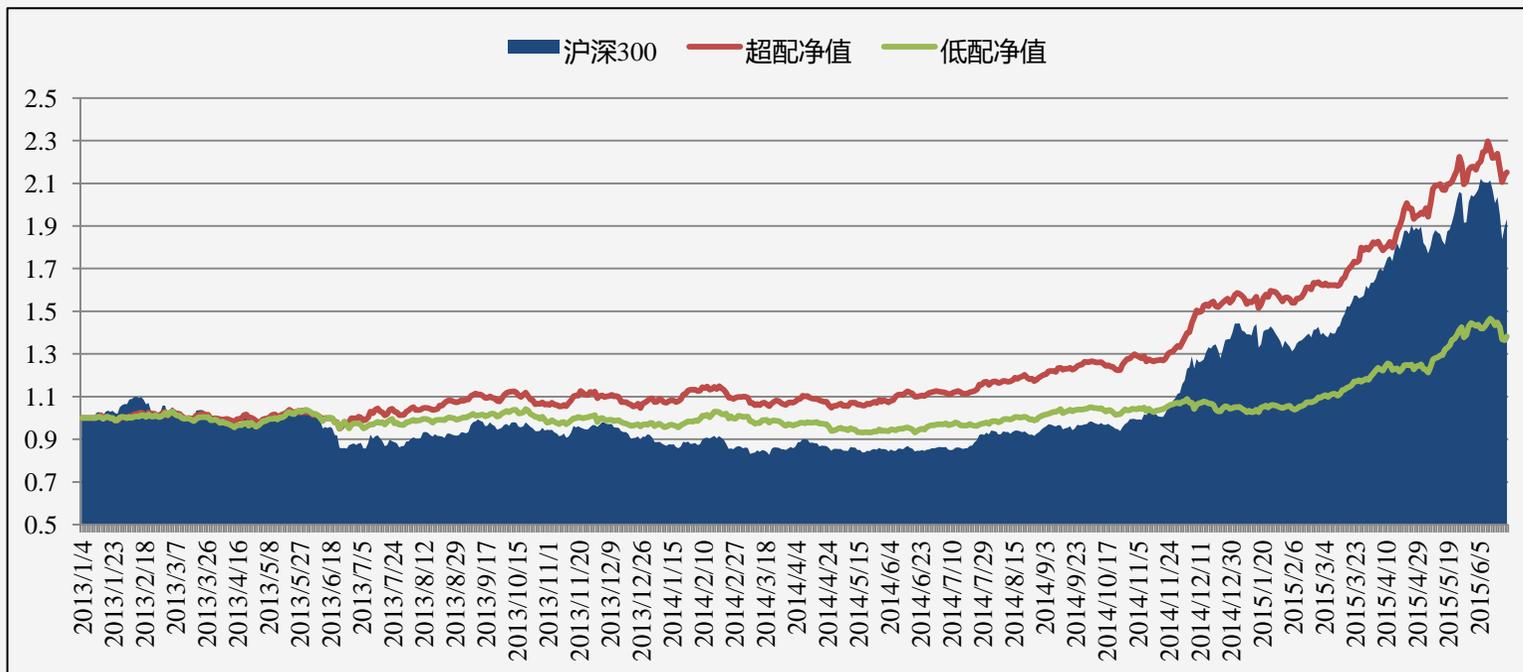


03

| 实证分析 |



- ◆ **个股数据**：2013年1月1日至今全市场个股开盘价、收盘价、成交量等数据，其中开盘价、收盘价用向后复权数据；
- ◆ **行业数据**：2013年1月1日至今申万一级行业指数收盘价；
- ◆ **关注度数据**：2013年1月1日至今关注度数据；



指标	对冲净值	超配净值
累计净值：	1.55	2.15
累计收益率：	54.78%	115.18%
年化收益率：	20.26%	38.21%
信息比	1.76	2.51
日胜率：	50.93%	57.70%
周胜率：	55.65%	55.65%
最大回撤：	-7.99%	-8.79%

基于大数据挖掘的关联个股投资机会



04

| 总结及未来研究方向 |



总结

- ◆ 基于互联网大数据构建的基准个股与关联个股构建的**共同关注度指标**能够作为一个选股因子；
- ◆ 基于基准个股与关联个股之间的联动构建的选股策略在历史回测期内**效果显著**；

未来研究方向

- ◆ 基于共同关注度，加入**更多的因素**考虑个股之间的联动构建策略；
- ◆ **个股的关注度变化**以及**行业整体关注度变化**研究选股以及行业配置策略；
- ◆ 个性化需求定制；

专题策略报告

有代表性的研究报告如下：

- <<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(一)>>
- <<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(二)>>
- <<倾听股吧之声，洞察大盘趋势—互联网大数据挖掘系列专题(三)>>
- <<那些年一起追过的财经小编选股策略—互联网财经频道文本挖掘策略>>
- <<上市公司披露信息变更隐含的投资机会—事件驱动策略之(十四)>>
- <<基于热点概念的文本挖掘选股策略-互联网大数据挖掘系列专题(四)>>等

互联网文本挖掘工具

有代表性的工具如下：

- 1、A股新闻热度搜索工具；
- 2、A股上市工具公告抓取工具；
- 3、上市公司信息变更抓取；
- 4、文本信息批量识别及处理；
- 5、汇丰PMI实时监测工具；
- 6、个股研报热点监测工具；
- 7、特定公告实时监测工具；
- 8、财经小编选股工具；

.....

基于大数据挖掘的关联个股投资机会

Thanks !
谢谢

地址: 广州市天河北路183号大都会广场 P.C.510075 电话: 020-87555888 传真: 020-87553600 WWW.GF.COM.CN