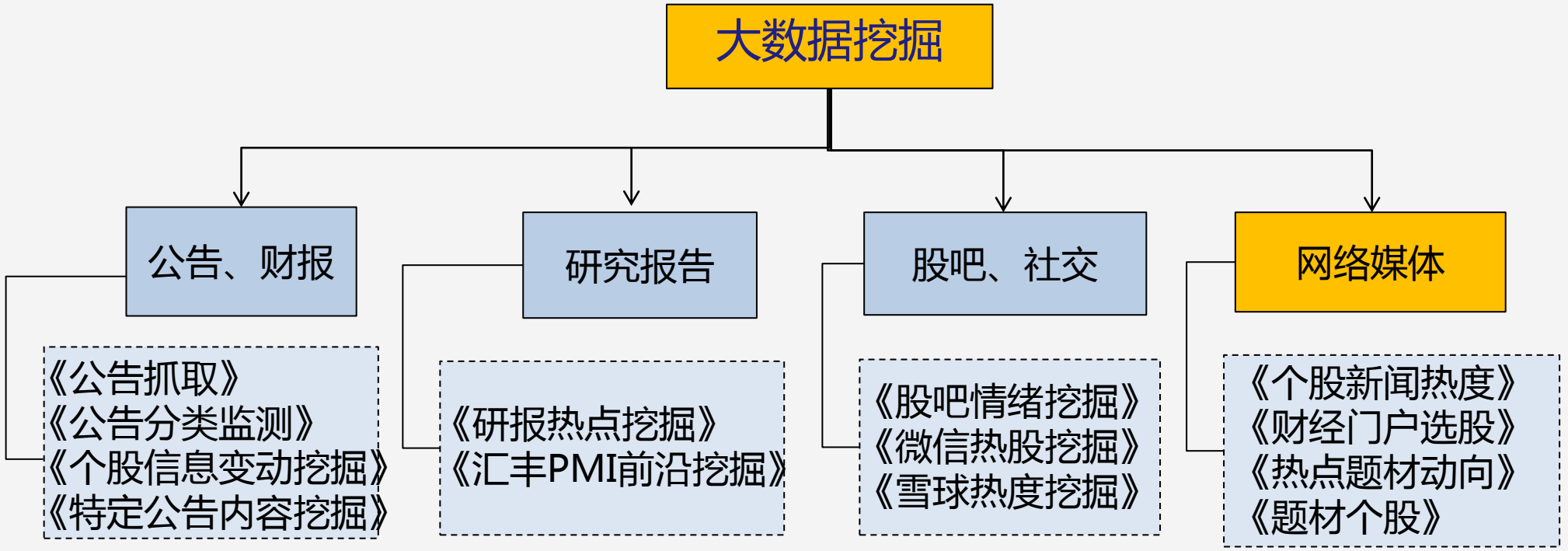
基于大数据挖掘的关联个股投资机会

——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


大数据挖掘
公告、财报
《公告抓取》
《公告分类监测》
《个股信息变动挖掘》
《特定公告内容挖掘》
研究报告
《研报热点挖掘》
《汇丰PMI前沿挖掘》
股吧、社交
《股吧情绪挖掘》
《微信热股挖掘》
《雪球热度挖掘》
网络媒体
《个股新闻热度》
《财经门户选股》
《热点题材动向》
《题材个股》

01

02

03

04

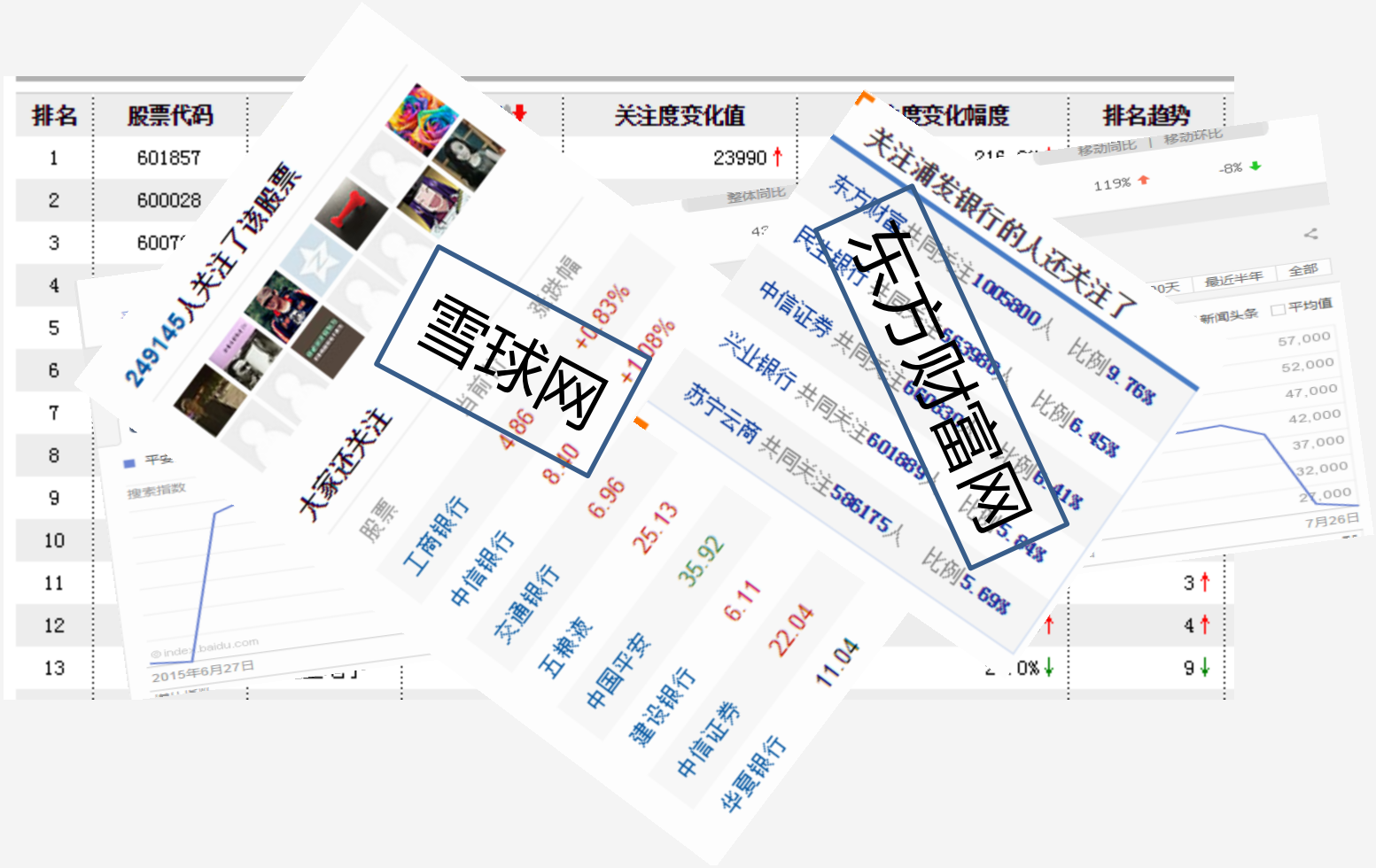
01

| 关注度简介 |

>

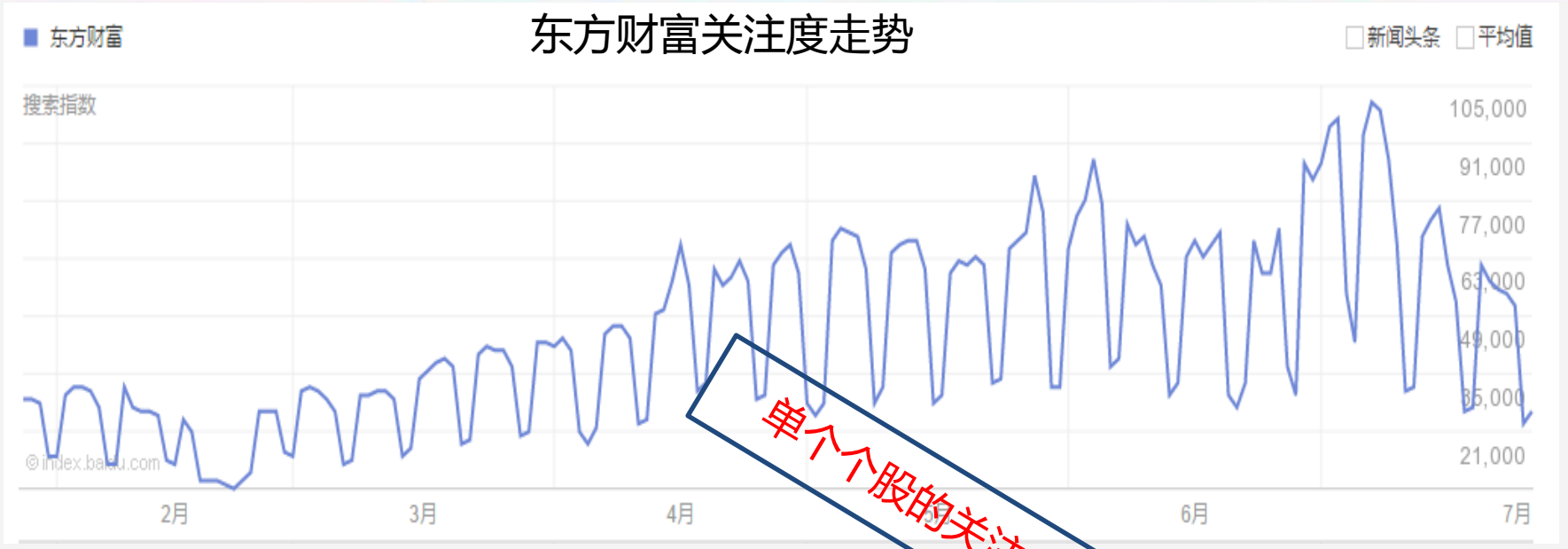
基于大数据挖掘的关联个股投资机会

关注度简介——用户关注度数据

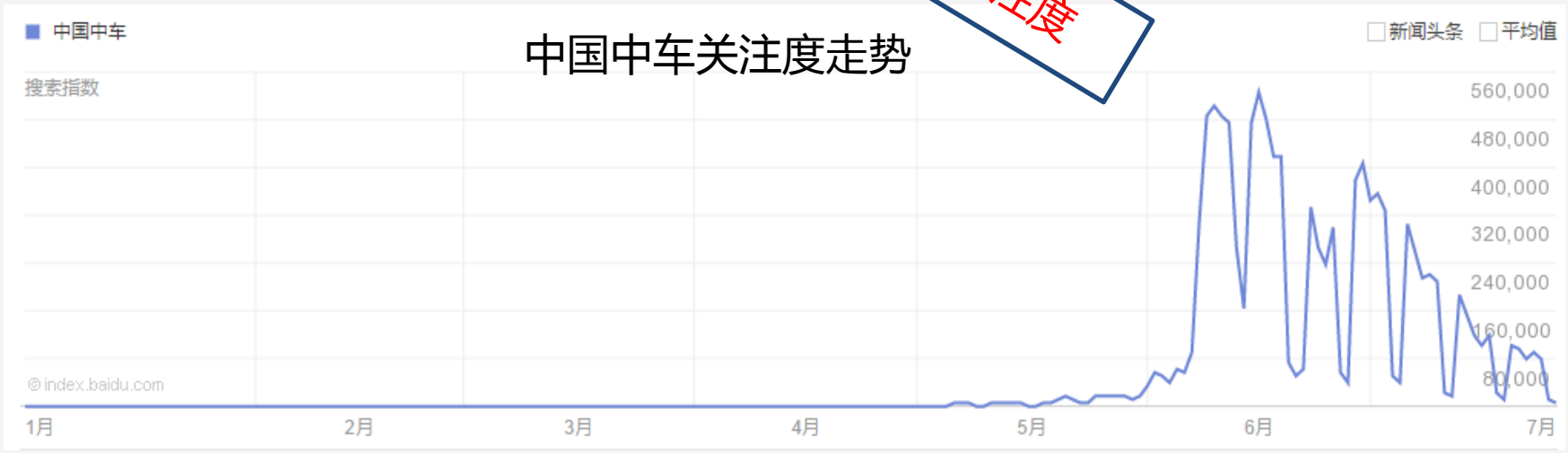


基于大数据挖掘的关联个股投资机会

东方财富关注度走势

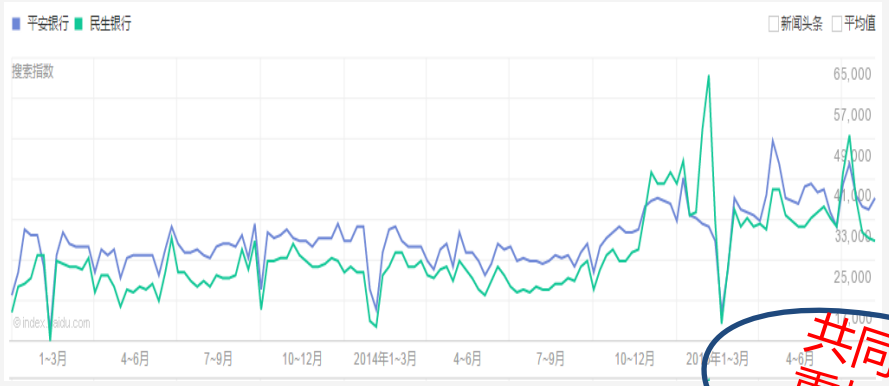


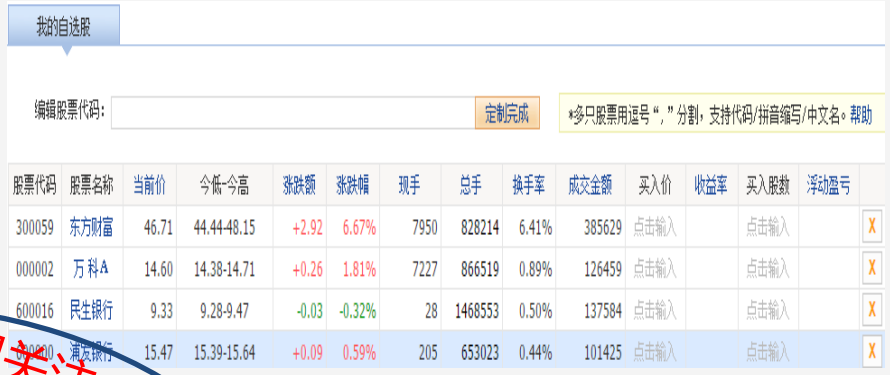
中国中车关注度走势



数据来源：广发证券发展研究中心、百度指数

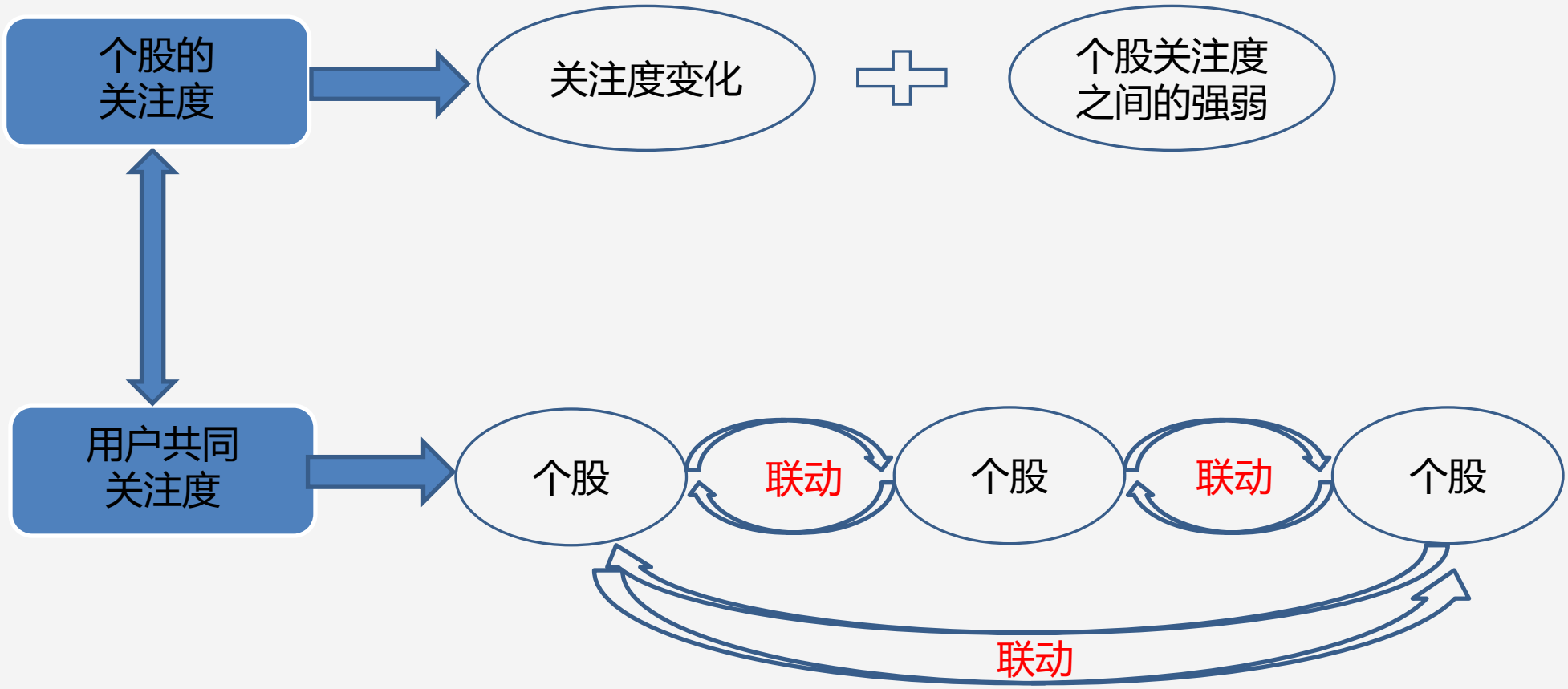
关注度简介——用户关注度数据





共同关注，注重投资者关注度的切换

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


个股的
关注度
关注度变化
个股关注度
之间的强弱
用户共同
关注度
个股
联动
个股
联动
个股
联动

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：

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

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：$\frac{\sum_{i=1}^{n} \xi_{ij} * \xi_{ik}}{\sum_{i=1}^{n} \xi_{ij}}$


互联网


BIG
DATA


注意力


量化投资

01

02

03

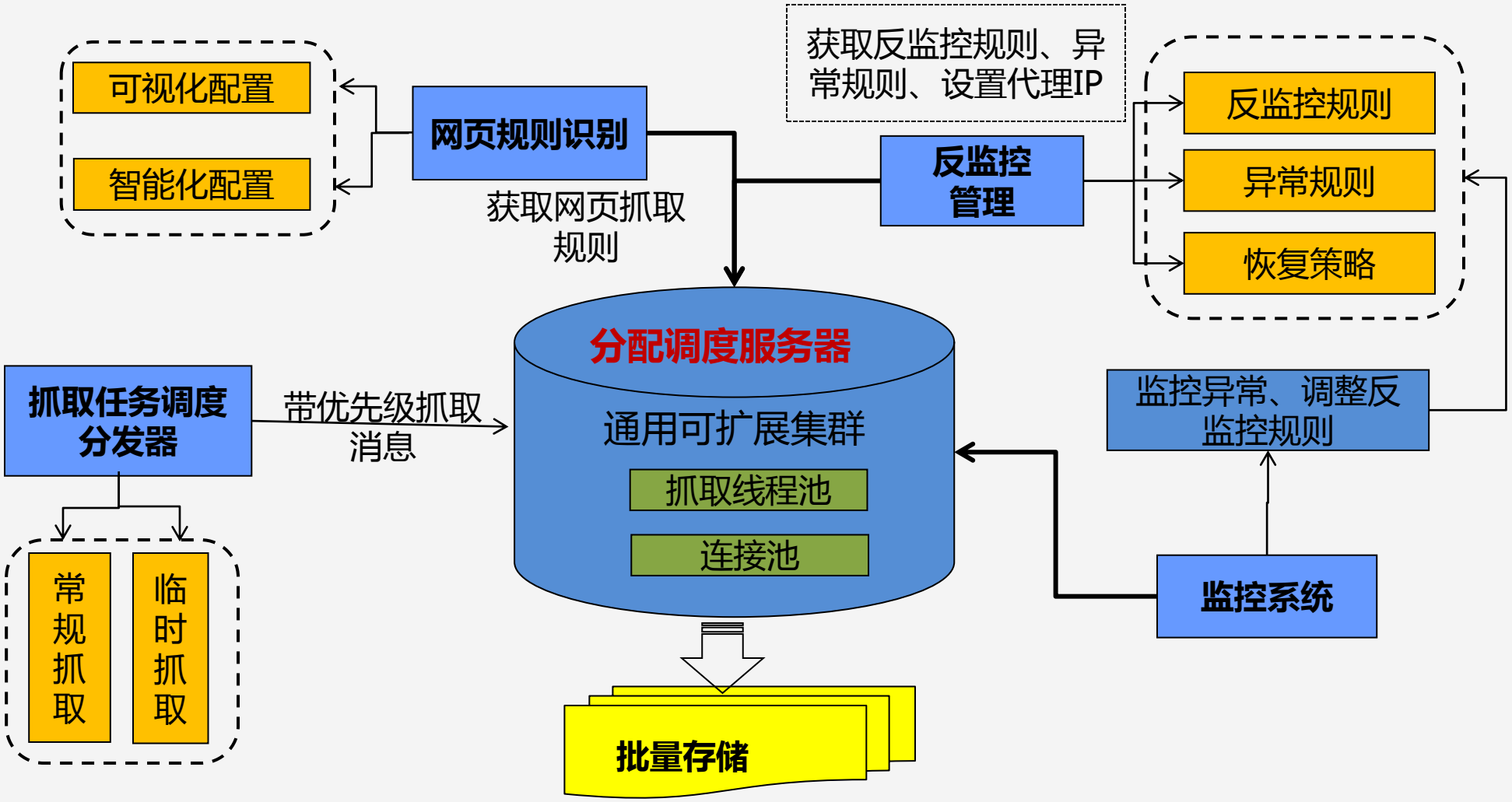
04

02

| 策略构建|

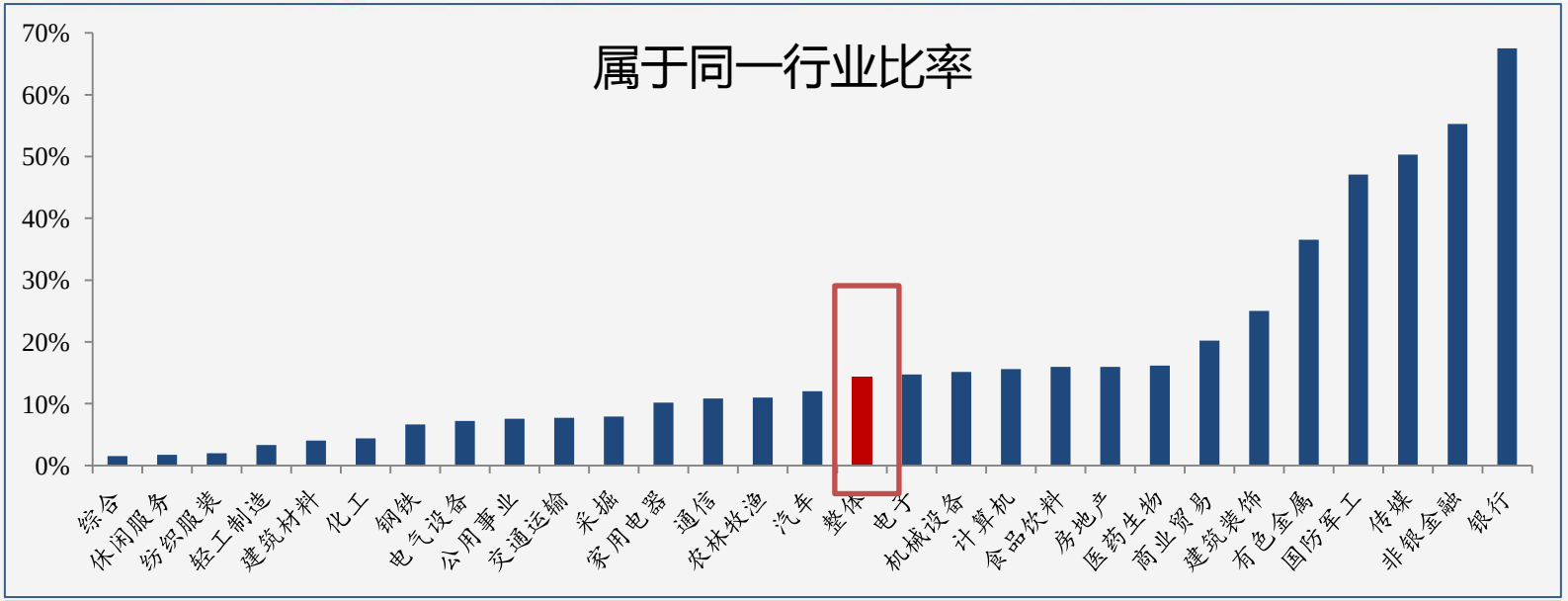
>

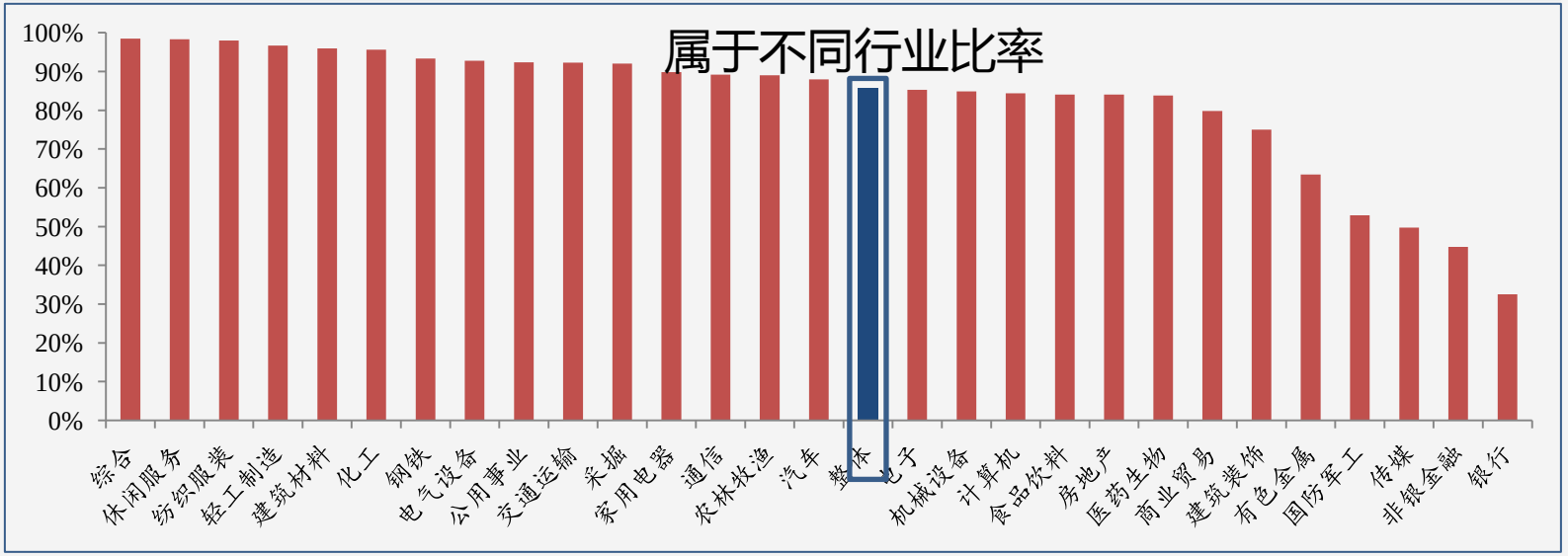
基于大数据挖掘的关联个股投资机会


可视化配置
智能化配置
网页规则识别
获取网页抓取规则
获取反监控规则、异常规则、设置代理IP
反监控管理
反监控规则
异常规则
恢复策略
分配调度服务器
通用可扩展集群
抓取线程池
连接池
抓取任务调度分发器
带优先级抓取消息
常规抓取
临时抓取
批量存储
监控异常、调整反监控规则
监控系统

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





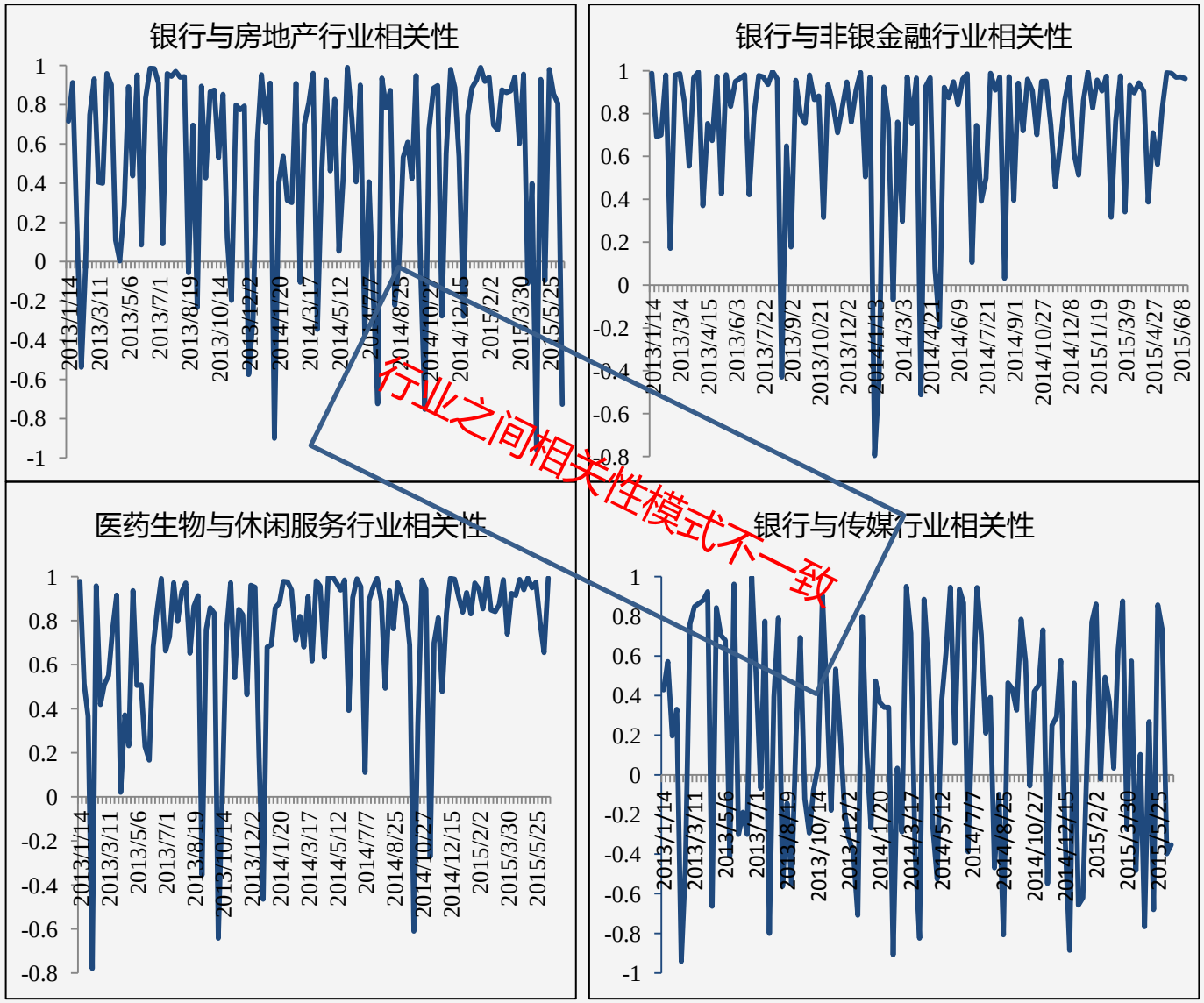
数据来源：广发证券发展研究中心、互联网

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
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%

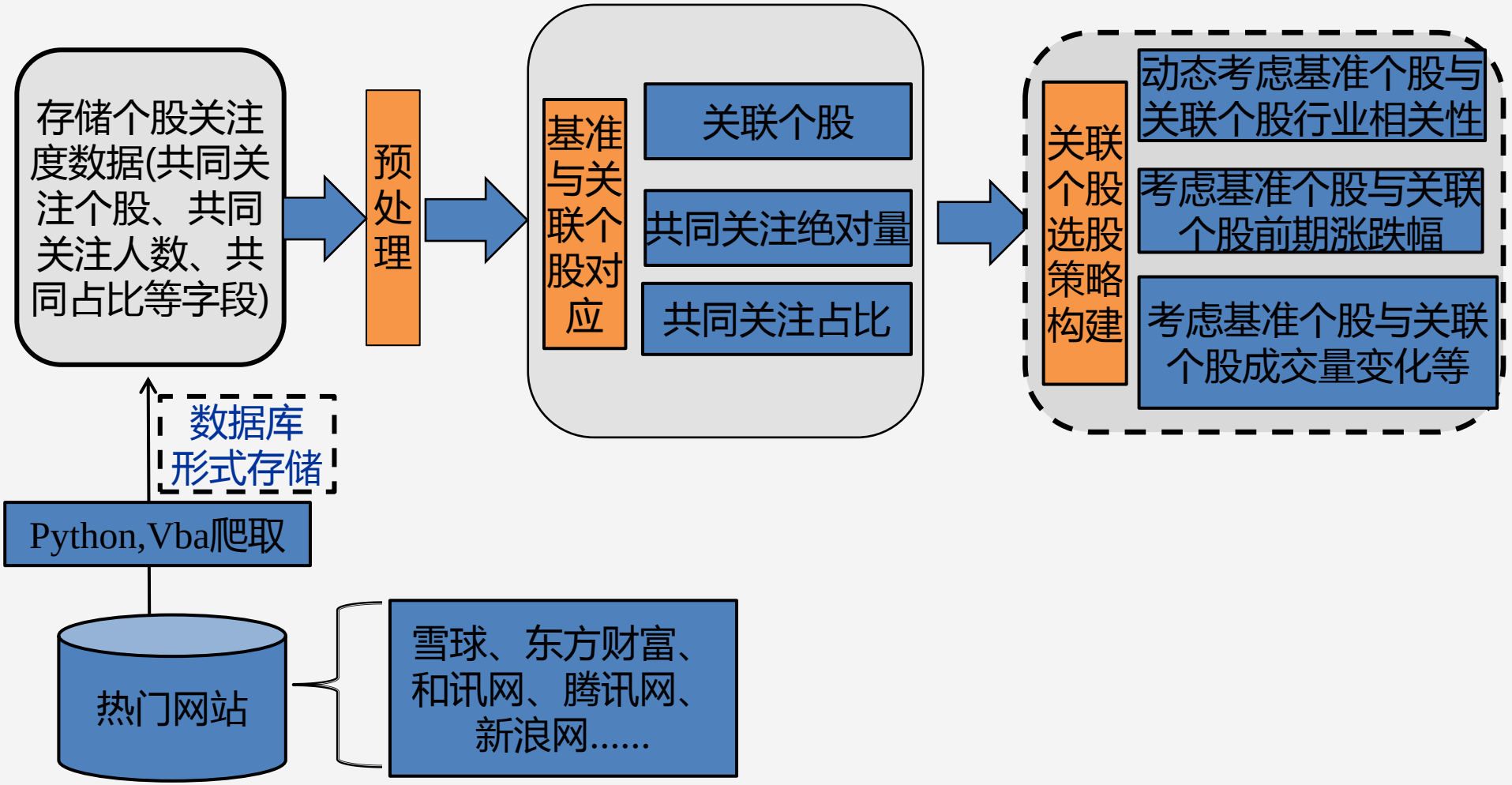
数据来源：广发证券发展研究中心、互联网

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



数据来源：广发证券发展研究中心、Wind


存储个股关注度数据(共同关注个股、共同关注人数、共同占比等字段)
预处理
基准与关联个股对应
关联个股
共同关注绝对量
共同关注占比
关联个股选股策略构建
动态考虑基准个股与关联个股行业相关性
考虑基准个股与关联个股前期涨跌幅
考虑基准个股与关联个股成交量变化等
数据库形式存储
Python,Vba爬取
热门网站
雪球、东方财富、和讯网、腾讯网、新浪网……

- **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；**

01

02

03

04

03

实证分析

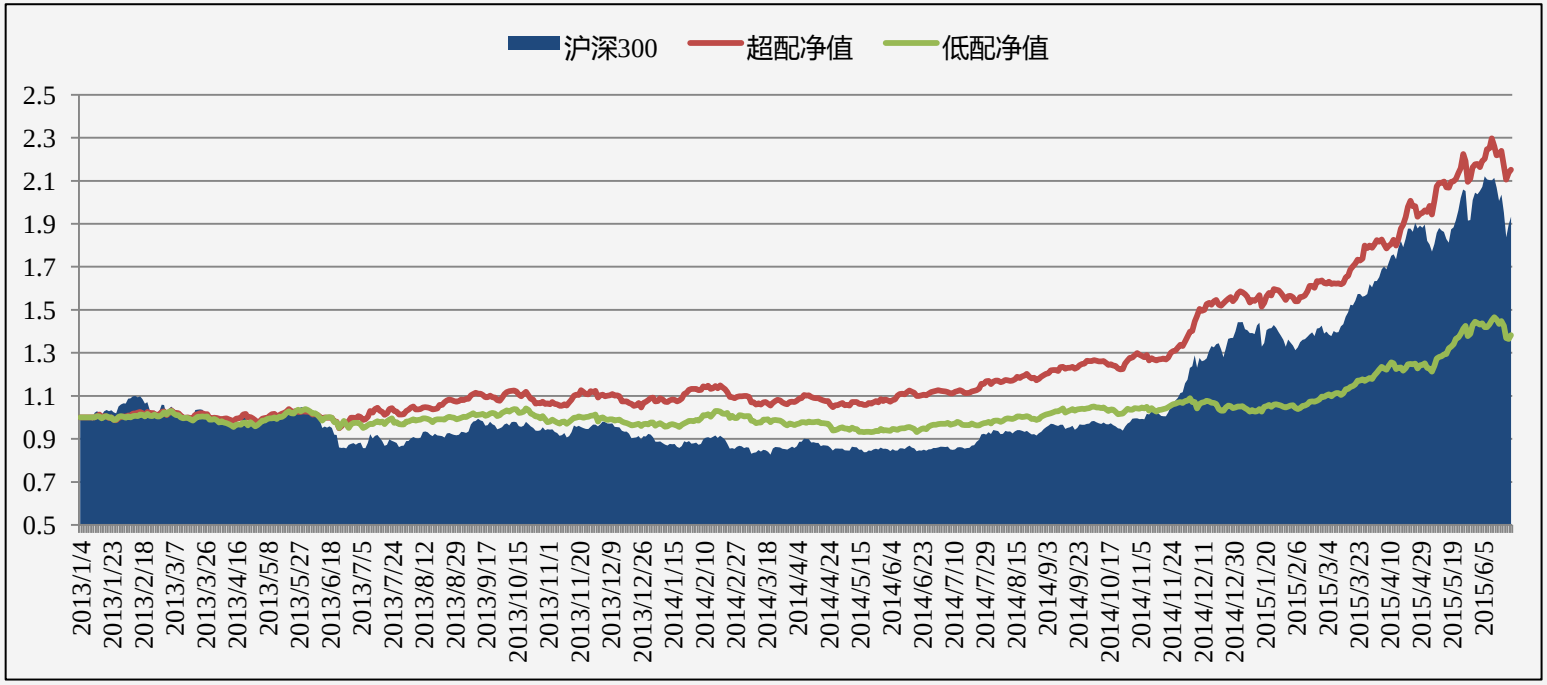
>

基于大数据挖掘的关联个股投资机会

◆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%

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；

……

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