【Smartxt 创新策略】文本大数据因子的选股能力浅析

东方证券 ORIENT SECURITIES

研究结论

- 投资是人类的活动，但是传统因子库缺少直接对投资者行为进行预判的工具；
- 运用大数据的手段，可以获得近似描述投资者行为的工具，比如本文介绍的关注因子；
- 关注因子主要衡量股票的传播深度，包括投资者提及的频率和投资者看到的频率；
- 2008~2016 回测表明，关注因子在 IC、区分度和单调性方面表现良好；
- 回测表明，关注因子与市值因子相关性不大。

风险提示

- 基于大数据的因子属于创新因子，投资风险较大，不建议一般投资者参与。
- 东方创新 Smartxt 相关报告均采用互联网文本挖掘技术，基础数据由包括计算机爬虫、文本处理系统、智能匹配系统在内的东方创新 Smartxt 网站自动生成。与传统研究报告相比，本报告的信息来源更加广泛与多元化。特别地，系统挖掘出的个股因子，所体现出的是市场对该股票的一般性看法。尽管对文本信息进行了数据清洗，我们仍无法确保消除全部系统噪音。
- 报告结论仅供参考，特此声明。

报告发布日期 2016 年 12 月 29 日

证券分析师 **冯剑**
021-63325888-4311
fengjian@orientsec.com.cn
执业证书编号：S0860515080003

联系人 **张亚南**
021-63325888-6117
zhangyanan@orientsec.com.cn
吴鸣远
021-63325888-6160
wumingyuan@orientsec.com.cn

相关报告

【Smartxt 创新策略】探索举牌话题传播对股价真实影响	2016-12-08
【Smartxt 创新策略】机构调研行为交易策略详解	2016-08-30
【Smartxt 创新策略】机构调研行为交易策略之增强版	2016-09-28
【Smartxt 创新策略】寻找高送转话题传播对股价的真实影响	2016-08-25
【Smartxt 高级应用】主题交叉搜索实现	2016-09-29
【Smartxt 基础应用】利用大数据抓住突发事件投资机会	2016-07-18
【Smartxt 基础应用】巧用调研信息，获取超额收益	2016-07-25
【Smartxt 基础应用】如何利用大数据寻找概念龙头	2016-07-01
【Smartxt 基础应用】如何利用大数据做主题轮动	2016-06-30

深度报告

【投资策略·证券研究报告】

目录

Smartxt.cn 是东方证券金融创新团队自主研发的开放式证券智能搜索工具，目前正在公测。

Smartxt.cn 网站目前拥有十亿级文本数据库，全面覆盖官媒、新闻、从业人员言论、公司公告、调研信息、互动平台、股票论坛、移动聊天工具等信息源。我们的目标是，建设人人可用的金融文本大数据平台，让用户可以方便地运用大数据帮助投资决策，带动证券投研领域的金融科技进步。

东方创新 Smartxt 平台主要功能目前包括主题股票互查、调研寻踪、预期探索、智能公告、智能预警、群聊助手。详情请参阅本报告后附的“Smartxt.cn 简介”。

一、文本大数据的投资价值

投资是人类的活动，是投资者以其逻辑认知来指导的行为。所以，一切对于未来股价波动的分析预测，归根到底都离不开对投资者行为的判断。

遗憾的是，对于传统手段而言，投资者行为是一个黑箱，只能通过其结果进行事后的回顾，根据历史规律进行谨慎的外推。

比如，技术分析流派使用的量价指标，属于投资者行为的产生的结果，在时间上是滞后的，并不能直接对投资者行为进行预判，只是希望类似行为能够重复或持续；

又如，基本面流派使用的财务指标，本身并不会导致股价波动，只有投资者意识到的有利或不利因素，才会引致交易行为，从而导致股价波动。所以基本面分析转而判断财务指标有多“好”，才有可能引起投资者关注。

总之，传统手段并不能够对投资者行为进行判断，只是用一些间接的方式进行推测。

诚然，人类的投资行为在事前是隐密而无法直接观测的。幸而有了互联网和大数据的协助，可以使我们以一定的概率逼近投资者的真实行为。

依托于 Smartxt 平台丰富的文本贮备和结构化框架，我们整理出一系列文本大数据因子，用以推断投资者真实行为。

本文主要介绍其中的关注因子。

二、关注因子的回溯测试

（一）因子简介

整个 Smartxt.cn 平台的基础测度是统一的，正如我们前期报告中一再说明的，词频是我们运用的基础度量方式。

在整个文本数据库中，按天汇总各股票出现的频数，包括股票简称、股票代码、股票昵称，经整理而成股票的关注因子。

关注因子衡量两方面的信息：

1、某只股票被投资者提及的频率；

2、某只股票被投资者看到的频率；

我们把这两种信息统称为传播深度。

直观上理解，传播深度的不断增强，代表着对于某只股票的认知在投资者群体中的不断普及。如果认知将以一定的概率转化为投资行为，传播深度事实上就构成了对行为的近似描述。

从这个意义上理解，关注因子能够一种对投资者行为的直接观测。

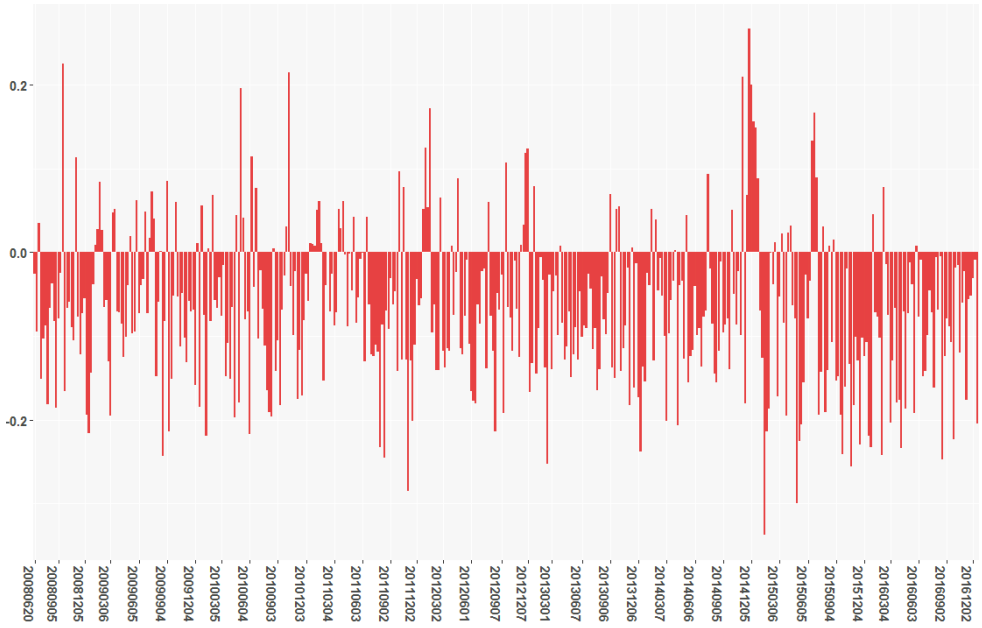
我们下面来分析这种行为产生的价格波动的方向。

（二）因子 IC

- IC 代表着因子对未来股价波动的预测能力；
- IC 的绝对值越高，说明这种预测能力越强；
- IC 为正，说明因子值越大，股票未来涨幅越大；IC 为负，说明因子值越小，股票未来涨幅越大；
- 各时期 IC 正负的一致程度越高，因子质量越好；

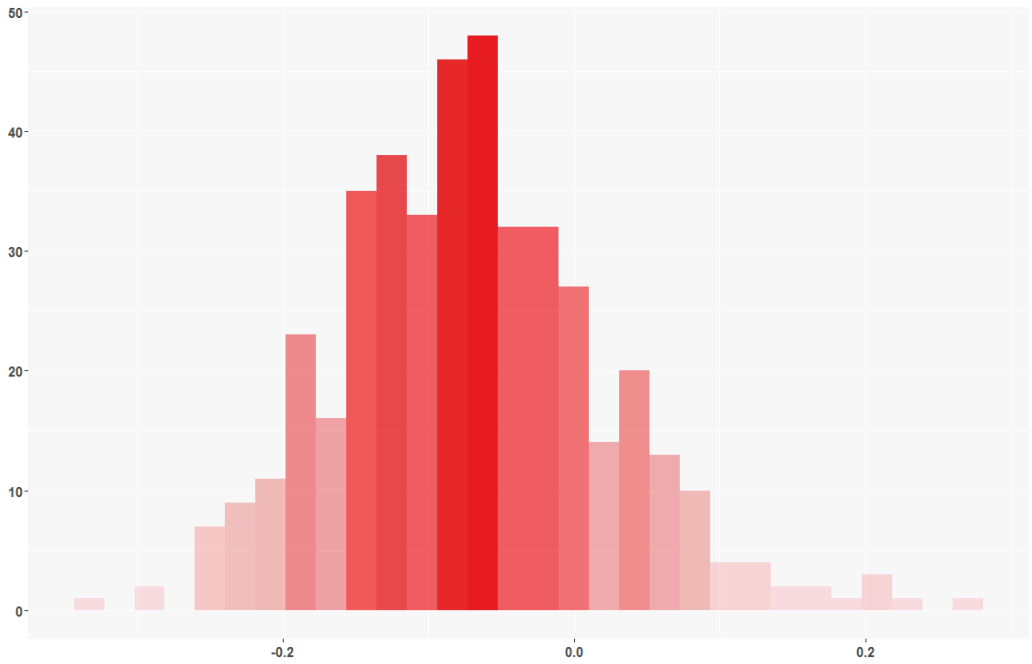
我们来看一看关注因子的 IC，股票池是全市场股票等权，计算的频率是周，也就是说 IC 是用上周的关注和本周的涨幅来计算的。

图 1：关注因子 IC（全市场，周频，2008～2016）



资料来源：东方证券研究所，Smartxt.cn

图 2：关注因子 IC 分布（全市场，周频，2008～2016）



资料来源：东方证券研究所，Smartxt.cn

图 1 向我们展示了关注因子的 IC，时间跨度是 2008 年 6 月到 2016 年 12 月 7 年。图 2 展示了因子 IC 的分布。

- 7 年 IC 绝对值为 0.07，绝对值高于一般因子，关注因子预测能力良好；
- IC 为负，说明关注因子值越小，股票未来涨幅越大；
- 各时期 IC 基本一致为负，因子质量良好；

综上，我们认为关注因子在股价预测能力方面，属于良好水平。

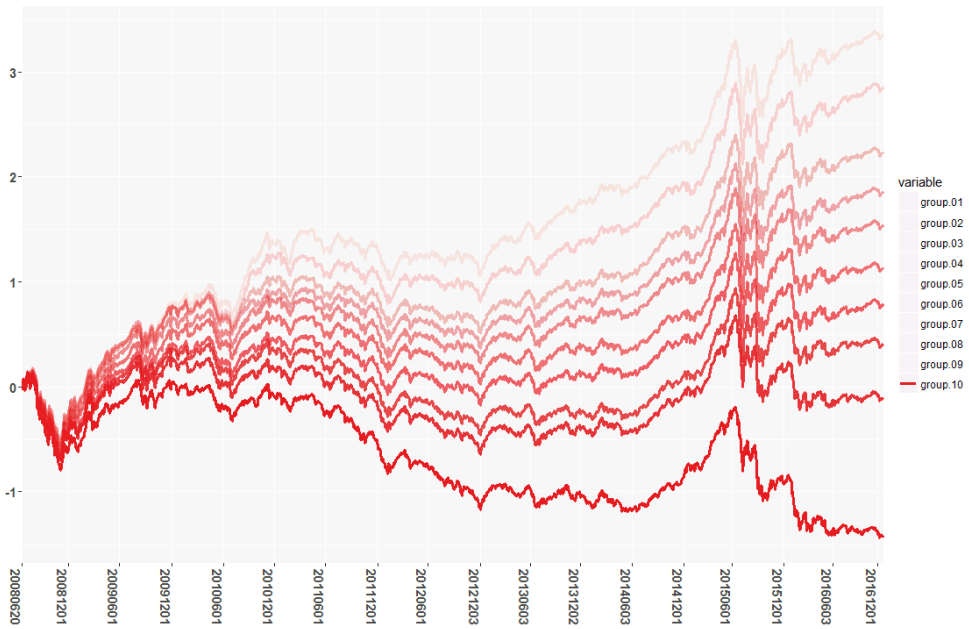
（三）因子区分度及单调性

因子实际上就是选择股票的一种指标。

除了分析这种指标对股价的预测能力，还要看指标在区分涨幅不同的各类股票的时候是否泾渭分明。

我们依据各期因子值由低至高将股票分为数量相同的 10 组，按周重新分组，组内等权，各组的累计收益率如图所示。

图 3：关注因子分组收益率（全市场，周频，2008～2016）



资料来源：东方证券研究所，Smartxt.cn

用不同的透明度标识 1~10 组，最浅色是第 1 组，因子值最低；最深色是第 10 组，因子值最高；中间依次为第 2~9 组。

我们可以非常清楚地看出：

- 7 年时间内，关注因子能够持续地将股票划分成为涨幅显著不同的 10 组，各组之间涨幅差距明显，关注因子区分度良好；
- 依因子升序排列的 1~10 组，累计涨幅呈现降序，7 年时间各组排序未见交叉，关注因子单调性良好；

综上，我们认为关注因子在区分度和单调性方面，属于良好水平，

（四）关注因子与市值因子的讨论

这是个不可回避的话题，因此我们会用报告的一半篇幅来加以阐述。

按照以往的习惯，本节首先讨论的话题是，关注因子和市值因子，是否存在逻辑上的必然关联。下一章我们再研究两者之间是否存在事实上的关联。

存在一种先入为主的观点：大市值股票的关注度高，小市值股票的关注度低。其依据也比较直接：大股票股东人数多，小股票股东人数少。

我们对这种观点提出两个疑问：

- 持股市值是否会在投资者中呈现均匀分布。简单点就是说，是不是各个股东持有的市值基本相同。答案是非常明显的，前人还发明了“持股集中度”这个指标来刻画这种不均匀的程度。
- 股东人数与传播深度是否正相关。两者的区别是十分明显的：股东可以不参与传播，参与传播的不一定是股东。是否引发传播的关键因素，可能不是市值大小，而是题材、板块、公司事件、涨幅等其它因素。

总之，大市值股票的股东人数不一定多；即便股东人数多，也不一定导致股票关注度高，因为股东可以不参与传播，参与传播的也不一定是股东。因此，关注因子和市值因子，并不存在逻辑上的必然关联。

下一章我们来看两者之间是否存在事实上的关联。

三、关注因子 VS 市值的实证研究

（一）方法讨论

因子之间的相关性，有很多现成的方法。

但是对于 A 股市场，一旦涉及市值因子，问题就有点棘手——因为市值的表现太好了。

从基础方法上理解，要证明两个因子之间存在或不存在关联，首先会观测其中一个因子发生变化时，另一个因子是否同步变化。

具体到关注因子和市值因子上面，就是说，我们要观测，当市值因子表现好的时候，关注因子的表现如何，以及市值因子表现差的时候，关注因子的表现又是如何；然后来对比前后两种情况下的表现之间是否存在差异，进而分析两个因子的相关性。

但是问题在于，A 股历史上绝大部分时间，市值因子的表现都非常好。我们缺少用于对比的，市值因子表现差的时间段。短期的风格转换是存在的，但是足以进行统计分析的时间段却是没有。

所以，本文选择了另一种方法。

既然在时间这个维度上，无法进行对比，就应当在另一个维度上进行对比。

我们选择的方法是，固定市值范围。即在不同的市值范围内对关注因子进行回测，观测其表现是否存在重大差异。

（二）沪深 300 股票池的因子 IC

因为上一章的回测使用的是全市场等权，当时的因子表现是以中小市值为主。

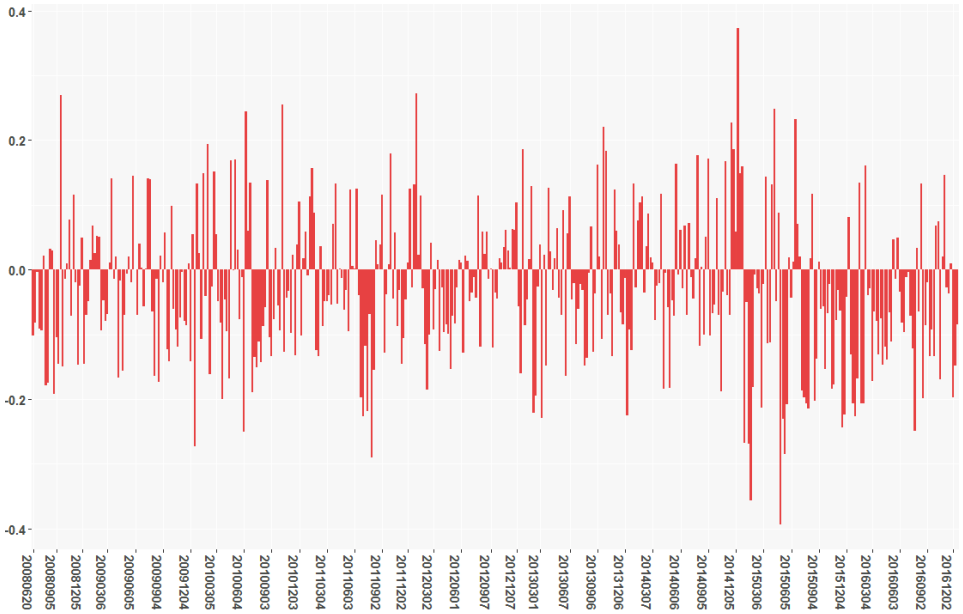
本章只需选择大市值股票作为股票池，使用类似的方法进行回测，然后进行前后对照即可。

我们选择沪深 300 作为股票池。

请读者注意，用沪深 300 股票池进行因子回测，本身的条件就非常苛刻，类似于一种极限挑战。因为股票因子回测需要一个相对差异化的股票池，这样才能表现出因子的功效。而在沪深 300 这个高度一致化的股票池中，因子表现必然会有很大退化，

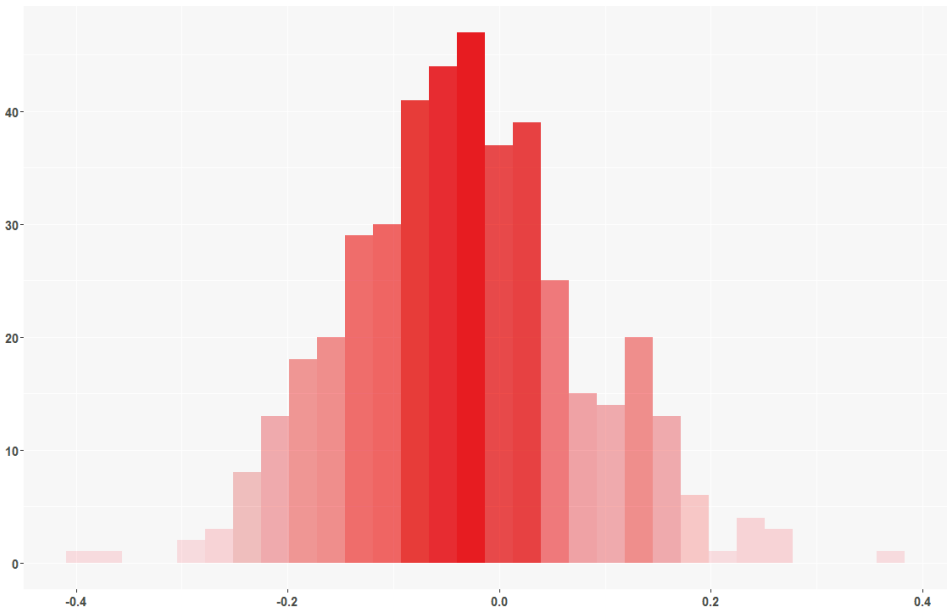
我们需要对比的是，因子在两个股票池的表现，是否存在重大的、本质上的差异即可。

图 4：关注因子 IC（沪深 300，周频，2008～2016）



资料来源：东方证券研究所，Smartxt.cn

图 5：关注因子 IC 分布（沪深 300，周频，2008～2016）



资料来源：东方证券研究所，Smartxt.cn

图 4 向我们展示了沪深 300 股票池关注因子的 IC，时间跨度是 2008 年 6 月到 2016 年 12 月 7 年。图 5 展示了因子 IC 的分布。

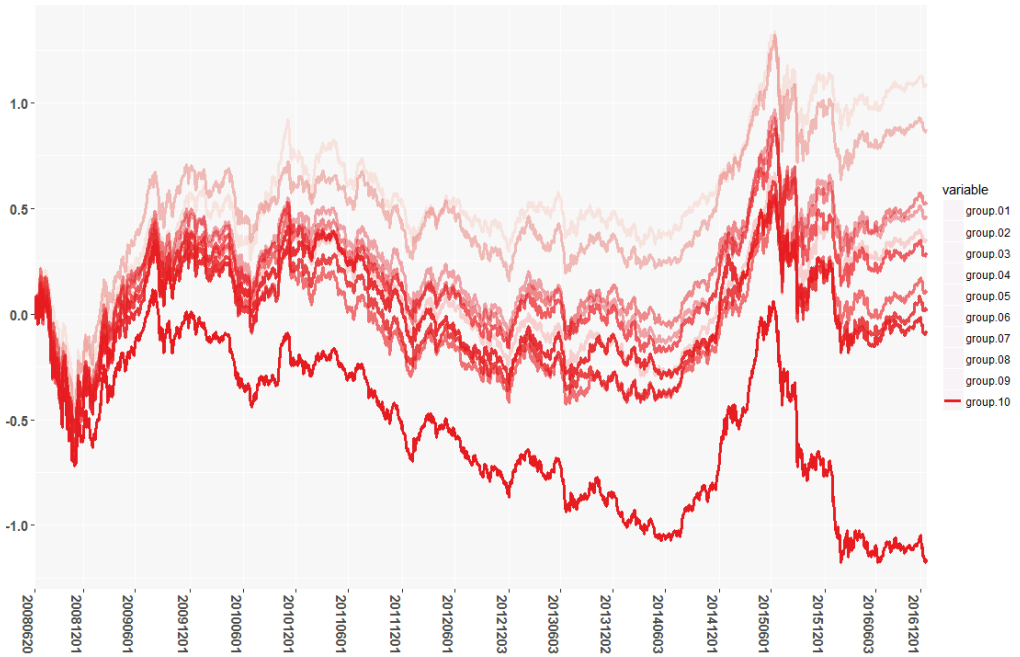
- 7 年 IC 绝对值为 0.03，绝对值高于同等条件下的一般因子，关注因子预测能力良好；
- IC 仍然为负，说明关注因子值越小，股票未来涨幅越大；
- 各时期 IC 存在一定的正负反转，考虑到股票池约束，仍可接受；

综上，我们认为关注因子在沪深 300 股票池的 IC 表现，与全市场股票池的 IC 表现，并未出现本质差异。从 IC 角度分析，关注因子与市值因子关联性不强。

（三）沪深 300 股票池的因子区分度及单调性

除了股票池由原来的全市场换为沪深 300，其余条件均相同。

图 6：关注因子分组收益率（全市场，周频，2008～2016）



资料来源：东方证券研究所，Smartxt.cn

在区分度上我们可以看到，虽然不如全市场股票池那么泾渭分明，但是关注因子在沪深 300 股票池中依然表现出良好的区分能力和单调性，低关注度的股票表现高高在上，而高关注度的股票表现则持续落后。

我们仍然要提醒读者，现在是把沪深 300 股票池分为 10 组，每组只有 30 只股票，关注因子能够达到如此程度实属不易。

综上，我们认为关注因子在沪深 300 股票池的区分度和单调性表现，与全市场股票池的表现，并未出现本质差异。从区分度和单调性角度分析，关注因子与市值因子关联性不强。

四、结论

- 投资是人类的活动，但是传统因子库缺少直接对投资者行为进行预判的工具；
- 运用大数据的手段，可以获得近似描述投资者行为的工具，比如本文介绍的关注因子；
- 关注因子主要衡量股票的传播深度，包括投资者提及的频率和投资者看到的频率；
- 2008~2016 回测表明，关注因子在 IC、区分度和单调性方面表现良好；
- 回测表明，关注因子与市值因子相关性不大。

五、风险提示

基于大数据的因子属于创新因子，投资风险较大，不建议一般投资者参与。

东方创新 Smartxt 相关报告均采用互联网文本挖掘技术，基础数据由包括计算机爬虫、文本处理系统、智能匹配系统在内的东方创新 Smartxt 网站自动生成。与传统研究报告相比，本报告的信息来源更加广泛与多元化。特别地，系统挖掘出的个股因子，所体现出的是市场对该股票的一般性看法。尽管对文本信息进行了数据清洗，我们仍无法确保消除全部系统噪音。

报告结论仅供参考，特此声明。

附：smartxt.cn 简介

Smartxt.cn 是东方证券金融创新团队自主研发的开放式证券智能搜索工具，目前正在公测。Smartxt.cn 网站目前拥有十亿级文本数据库，全面覆盖官媒、新闻、从业人员言论、公司公告、调研信息、互动平台、股票论坛、移动聊天工具等信息源。六大特色功能，让大数据成为投资的驱动力。

1、任意词搜股票

任意主题可自动匹配股票；输入股票简称反查相关主题；支持多主题复合查询。开放式搜索是 Smartxt.cn 的一大特色，新造词也可以直接查询。运用智能搜索，节约前期研究时间，提高投研效率。

2、调研寻踪

登录网站查看近期机构调研信息。通过对 A 股 1449 只股票，40983 人次调研的归纳发现，无论牛熊，集中调研的股票均显著强于大盘。模拟交易 3 年最高 7 倍，详见深度报告《机构调研行为交易策略之增强版》。

3、预期探索

传统事件投资收益越来越少。Smartxt.cn 直接探索投资者预期，获得预期传播阶段的收益。投资范例见深度《寻找高送转话题传播对股价的真实影响》、《探索举牌话题传播对股价真实影响》。

4、智能公告

证券研究领域，机器解放人力的时代已开启！智能公告使用语义识别技术，自动提取公告摘要，实时更新，支持全字段搜索，分行业分类型展示，自选股公告提醒，全文阅读通道，打造公司公告的智能化入口。

5、智能预警

常实用的智能新闻处理功能，可实现黑天鹅事件超前预警。高速后台确保信息实时传递，智能算法多空瞬间识别。主动、量化、风控、两融、期货，黑天鹅预警已成为各类客户最欢迎的功能。

6、群聊助手

数不尽的微信群，错过的信息远比看到的多。现在，只需加入群聊助手，就可以实现多个群的“聚合式”展示，全文搜索、自选股信息提醒，你能想到的都有。有价值信息再也不怕错过。

篇幅所限，未能全面介绍东方创新 Smartxt.cn 所提供的所有功能。目前网站正在公测，正式开放敬请期待。部分功能需要登录使用，各项实用新功能将不断推出。如需安排路演，提前了解平台特性及功能并参与测试，请客户联系东方证券研究所销售同事。

分析师申明

每位负责撰写本研究报告全部或部分内容的研究分析师在此作以下声明：

分析师在本报告中对所提及的证券或发行人发表的任何建议和观点均准确地反映了其个人对该证券或发行人的看法和判断；分析师薪酬的任何组成部分无论是在过去、现在及将来，均与其在本研究报告中所表述的具体建议或观点无任何直接或间接的关系。

投资评级和相关定义

报告发布日后的 12 个月内的公司的涨跌幅相对同期的上证指数/深证成指的涨跌幅为基准；

公司投资评级的量化标准

买入：相对强于市场基准指数收益率 15%以上；

增持：相对强于市场基准指数收益率 5%～15%；

中性：相对于市场基准指数收益率在-5%～+5%之间波动；

减持：相对弱于市场基准指数收益率在-5%以下。

未评级——由于在报告发出之时该股票不在本公司研究覆盖范围内，分析师基于当时对该股票的研究状况，未给予投资评级相关信息。

暂停评级——根据监管制度及本公司相关规定，研究报告发布之时该投资对象可能与本公司存在潜在的利益冲突情形；亦或是研究报告发布当时该股票的价值和价格分析存在重大不确定性，缺乏足够的研究依据支持分析师给出明确投资评级；分析师在上述情况下暂停对该股票给予投资评级等信息，投资者需要注意在此报告发布之前曾给予该股票的投资评级、盈利预测及目标价格等信息不再有效。

行业投资评级的量化标准：

看好：相对强于市场基准指数收益率 5%以上；

中性：相对于市场基准指数收益率在-5%～+5%之间波动；

看淡：相对于市场基准指数收益率在-5%以下。

未评级：由于在报告发出之时该行业不在本公司研究覆盖范围内，分析师基于当时对该行业的研究状况，未给予投资评级等相关信息。

暂停评级：由于研究报告发布当时该行业的投资价值分析存在重大不确定性，缺乏足够的研究依据支持分析师给出明确行业投资评级；分析师在上述情况下暂停对该行业给予投资评级信息，投资者需要注意在此报告发布之前曾给予该行业的投资评级信息不再有效。

东方证券 ORIENT SECURITIES

免责声明

本证券研究报告（以下简称“本报告”）由东方证券股份有限公司（以下简称“本公司”）制作及发布。

本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告的全体接收人应当采取必要措施防止本报告被转发给他人。

本报告是基于本公司认为可靠的且目前已公开的信息撰写，本公司力求但不保证该信息的准确性和完整性，客户也不应该认为该信息是准确和完整的。同时，本公司不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的证券研究报告。本公司会适时更新我们的研究，但可能会因某些规定而无法做到。除了一些定期出版的证券研究报告之外，绝大多数证券研究报告是在分析师认为适当的时候不定期地发布。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。

本报告中提及的投资价格和价值以及这些投资带来的收入可能会波动。过去的表现并不代表未来的表现，未来的回报也无法保证，投资者可能会损失本金。外汇汇率波动有可能对某些投资的价值或价格或来自这一投资的收入产生不良影响。那些涉及期货、期权及其它衍生工具的交易，因其包括重大的市场风险，因此并不适合所有投资者。

在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任，投资者自主作出投资决策并自行承担投资风险，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均为无效。

东方证券研究所

地址： 上海市中山南路 318 号东方国际金融广场 26 楼
联系人： 王骏飞
电话： 021-63325888*1131
传真： 021-63326786
网址： www.dfzq.com.cn
Email： wangjunfei@orientsec.com.cn