金融科技（Fintech）和数据挖掘研究（一）——数据挖掘技术框架简介

相关研究

《养老金市场及产品研究（六）——养老金巨头和投顾专家：富达投资的成功之道》2019.02.15

《基于因子剥离的 FOF 择基逻辑系列十七——国内公募权益类基金有 Alpha 吗？》2019.02.14

《量化研究新思维（十四）——20 for Twenty: AQR 20 周年经典文献摘要 1》2019.02.10

分析师:冯佳睿

Tel:(021)23219732

Email:fengjr@htsec.com

证书:S0850512080006

分析师:余浩淼

Tel:(021)23219883

Email:yhm9591@htsec.com

证书:S0850516050004

投资要点:

- **数据挖掘对于投资决策的支持作用越发明显。**随着国内资本市场的发展，证券市场中的可交易标的类型和数量都飞速增长，与市场有关的数据也快速增加。同时，随着市场监管越来越严，机构投资者的比重不断上升，市场有效性也逐步提升。如何快速准确地处理海量数据，并从中得到有价值的信息，是在这样的环境中继续获取超额收益的重要方法。

- **数据挖掘涵盖从数据采集到分析结果评价共 7 个步骤。**即，数据采集、数据清洗、数据特征提取、数据结构化、数据存储、数据分析、分析结果评价。每一个步骤都需要大量的技术进行支撑，由此构成了完整的数据挖掘技术框架。

- **网络爬虫数据的使用需要审慎。**网络爬虫是获取数据的有效方式，然而对网站而言，通常并不欢迎用户利用爬虫获取数据，且绝大部分网站的创立目标是提供浏览服务，而不是数据提供商。因此，在保障服务运营的前提下，网站往往不会确保历史所有信息都可从客户所访问的网页当中获取。网络爬虫更适合作为一种数据抽样工具，而非可获取某领域全部数据的自动化工具。

- **自定义词典可以更好地提升现有自然语言分析库的性能。**自然语言处理往往都是通过神经网络等模型，分析先验的自然语言数据得到。投资所需要分析的文本信息一般是金融媒体新闻、公司公告、研究员报告等行文比较规范的书面语文本，其语法结构较为标准，易于处理。因此，只要准备好尽可能多的投资相关术语、公司名称、标的名称等专有名词字典，便可以很好地帮助自然语言处理系统分析文本，获取所需要的信息。

- **知识图谱可以帮助投资者从全新的视角认知市场。**它是一种直观显示各个实体关联信息的有效数据结构。通过对市场上“实体”和“关系”的定义与分析，帮助投资者从另一视角审视不同上市公司、交易标的之间的相关性，确认公司所处的产业链位置，提升对于市场的认知能力。

- **风险提示。**数据挖掘是从历史先验数据获取经验模型的方法，存在模型失效可能。

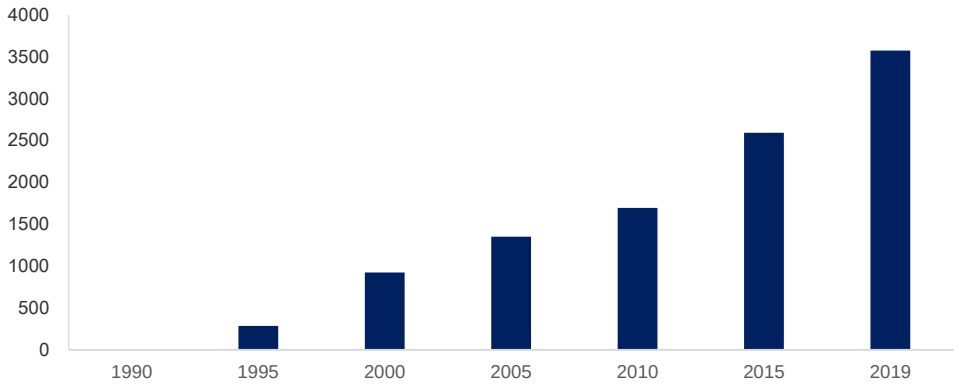
目 录

图目录

1. 投资环境变化提升数据分析的重要性

经过 30 多年的发展，我国的证券市场已从初创逐步走向了全面和丰富。从最早的国债市场，到 90 年代的股票和期货交易市场，再到 2000 年之后的公募基金、各种衍生品，整个证券市场场内、场外的可交易标的数量快速增长。以 A 股为例，从 1990 年到 2019 年，全市场可交易股票数量从 8 只增长到现在的 3500 余只，增长了 400 余倍。

图1 A 股可交易股票数量（只，1990/1 – 2019/1）



资料来源：Wind，海通证券研究所

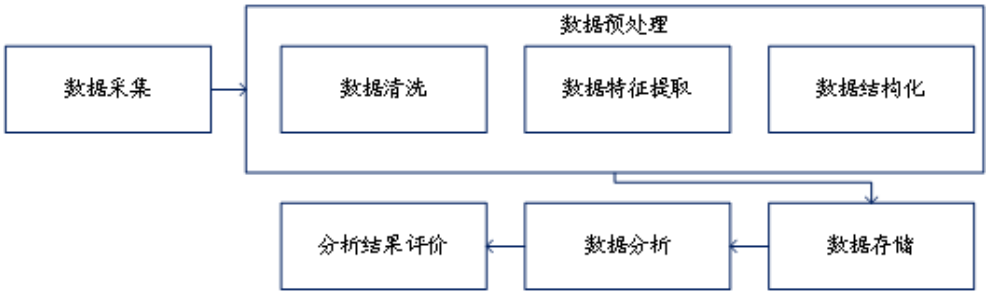
同时，随着信息披露的完善，财经新闻媒体的繁荣以及 2010 年以后社交网络、自媒体等新媒体的出现，与可交易标的相关的信息呈几何式增长。另一方面，市场监管愈发严格，机构投资者比重提升，基于简单的数据分析的投资模式已经很难适应现在的 A 股市场。

因此，如何从市场中获取这些海量数据，如何从这些数据当中过滤提取出有效的、可分析的信息，如何使用适当的方法对于这些信息进行分析汇总，从而帮助投资者进行投资决策，在当下的证券投资过程中正变得越来越重要。

2. 数据挖掘技术的基本流程

伴随着计算机技术，特别是近些年互联网技术的蓬勃发展，数据挖掘相应的技术在计算机技术层面已经日渐成熟，并在非常多的领域投入使用。从其他领域大量的案例来看，数据挖掘主要可以分为以下几个步骤：

图2 数据挖掘技术的基本流程



资料来源：海通证券研究所整理

数据采集：获取数据挖掘所需的原始数据。包括获取传统证券的价量数据、公司财

务数据、宏观经济数据等，也包括市场关注度、情绪指标、产品价格等非传统金融数据。

数据预处理：原始数据往往有不完整、结构不清晰、难以直接处理的问题，因此需要在使用这些数据前，进行一些预处理操作。

数据清洗：对数据进行重新审查和校验，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

数据特征提取：很多数据类型，例如文本、图片等，无法直接用数量模型进行分析处理，便需要从中提取有用的特征信息，并转化为可处理的数据。

数据结构化：好的数据结构可以在提升存取效率的同时，为数据分析提供更多可能。

数据存储：获取到的大量数据，往往需要物理介质进行存储，方便随时读取以进行不同类型的分析。目前常用的数据存储方式还是以数据库为主，因而要求事先将数据处理为数据库可以存储的格式。

数据分析：一般指对数据进行建模处理的过程。常用的模型有：传统线性回归或最近较为流行的机器学习等。数据分析模型的选取应当根据数据本身的特征和想要的分析结果的特性来确定。

分析结果评价：对于投资者而言，数据挖掘的最终目标是辅助进行投资决策。因此，数据挖掘的结果能否帮助投资者选取有更好表现的投资标的、规避潜在的风险，是评判数据挖掘结果的核心标准。

以上是对数据挖掘处理流程的简介，下文将介绍上述步骤中所涉及的一些比较关键的技术与解决方案。

3. 数据挖掘过程中的技术方案

3.1 数据采集技术

对于投资者而言，辅助投资决策所需要的数据通常包括投资标的价量信息，公司公开的财务报表，宏观经济数据等。随着国内金融市场的逐步完善，这些“传统”数据，已有如 Wind、朝阳永续等数据供应商通过各种渠道向投资者提供。因而，无论是数据的质量和丰富程度还是获取的便利性，都能在很大程度上得到保证。

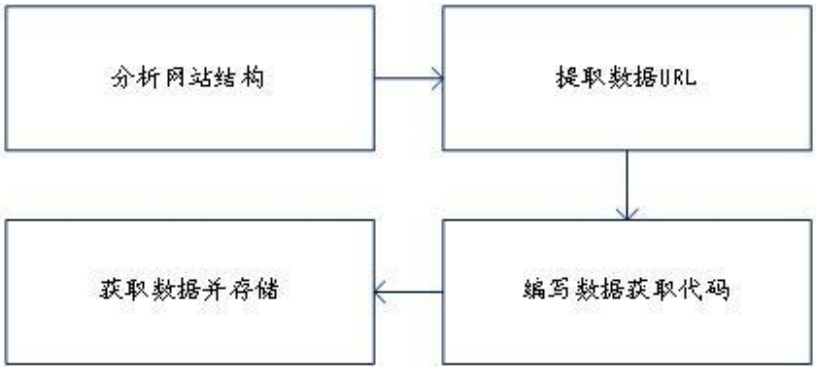
然而，对于面向整个市场的数据供应商来说，出于成本效益的考虑，第一，不会为单独的客户提供他所需要的一些特异性数据；第二，不会第一时间提供某些新出现的反映市场信息的数据。因此，如何以较低的成本获取满足自己需求，且传统数据供应商又无法提供的数据，对于投资策略的构建具有重要意义。

举例来说，网络上对于某些公司或者领域的关注热度可能和该公司的股价有一定相关性；很多电商平台上的产品特性和价格特征对分析该行业或企业的未来业绩有很大的帮助作用。但这类型数据往往具有标的众多、数据量大的特征，很难通过人工的方式采集。当前十分流行的网络爬虫技术，恰好是获取这类互联网公开数据的有效方式。

网络爬虫，即一种模仿浏览器部分功能，通过 HTTP 协议获取 HTML 脚本，解析得到所需信息的常用网络数据获取技术。该技术通常以构建搜索引擎而闻名。实际上，网络爬虫可分为通用爬虫与聚焦爬虫，相比搜索引擎常用的通用爬虫，获取特定网站上特定信息的聚焦爬虫往往是更加常用的一种数据获取工具。

聚焦爬虫的具体操作流程如下图所示。

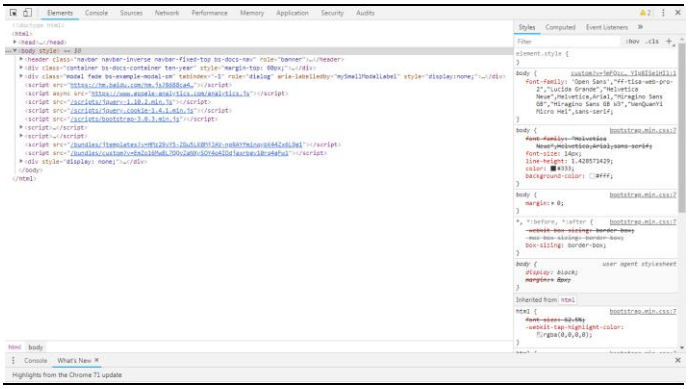
图3 聚焦网络爬虫构建流程图



资料来源：海通证券研究所整理

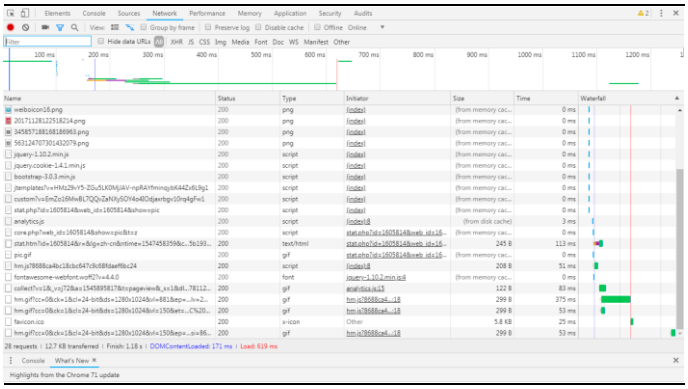
分析网站结构：构建聚焦网络爬虫的第一步需要对所“爬”的网站进行静态结构分析，即整个网页由哪些元素构成，每个元素又是怎样的一种数据，可以通过怎样的方式获得。常用的浏览器，如 Chrome，Firefox 等都有代码调试功能，可以展现浏览器从访问网站 URL 到显示所见网站的全部过程。本文以 Chrome 为例，展示如何通过这种工具，从网站中分析得到所需的信息。详细过程如以下三图所示。

图4 Chrome 调试窗口 Element 模块



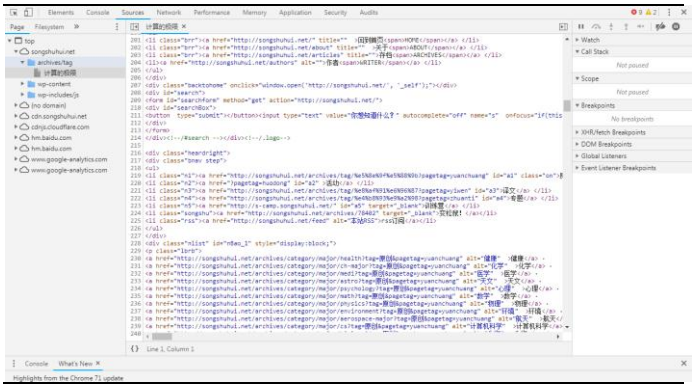
资料来源：Chrome 浏览器截屏，海通证券研究所

图5 Chrome 调试窗口 Network 模块



资料来源：Chrome 浏览器截屏，海通证券研究所

图6 Chrome 调试窗口 Sources 模块



资料来源：Chrome 浏览器截屏，海通证券研究所

其中，图 4 为 Element 模块，显示整个网页由哪些部分构成，所需数据属于该网页的哪个部分。图 5 为 Network 模块，显示构成该网页的每个元素分别通过哪些 URL 从网站获取，即，可提取出所需数据的对应 URL。图 6 为 Sources 模块，显示整个网页的 HTML 源代码。

提取数据 URL：通过上述过程对网站进行分析后，便可以确定所需要的数据包含在网页的哪个部分中，这个部分的内容又是通过访问哪个 URL 获得的。相较于通用爬虫需要对 HTML 当中所有 URL 进行解析不同，聚焦爬虫的目标仅仅是获得所需的特定数据，因此只需访问相应的 URL 即可。

编写数据获取代码：与大家通常的认知不同，编写网络爬虫代码并没有特定的编程语言限制。从某种意义上讲，支持 TCP|IP 协议的所有高级语言都可以用以编写爬虫代码。目前主流的爬虫语言是 JAVA 和 Python，主要得益于这两个语言所构建的爬虫函数库 Jsoup 与 Scrapy 较为易用和流行。但如果仅就某些特定使用场景来说，R 语言、VBA 语言都有相应的函数库，可以达到爬虫的效果。

获取数据并存储：运行写好的代码，便可以获得想要的数据进行存储。以下四图分别展示了两个网络爬虫的实际例子。其中，图 7 和图 8 是从东方财富“股吧”论坛上，爬取上证 50 权重股近期热度信息的示例。图 7 是原网站样式，图 8 是获取到的统计结果。图 9 和图 10 是从“天猫”Only 专卖网站，获取近期 Only 产品信息的示例。前一个标的热度的例子完全由 EXCEL VBA 编写完成，而后一个产品信息的例子由于数据形式和内容较为复杂，获取过程需要访问同一页面的多个 URL，因此采用 JAVA 编写得到。

图7 爬虫实例：“股吧”论坛截图

资料来源：东方财富“股吧”，海通证券研究所

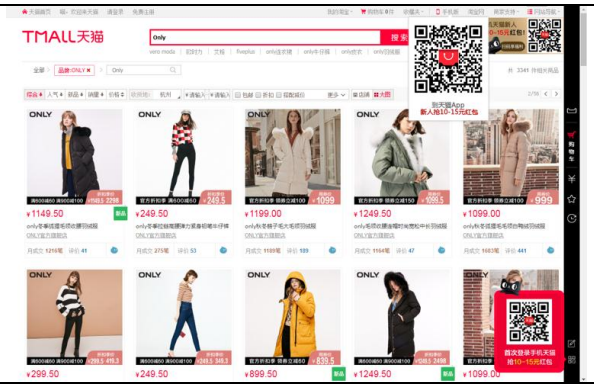
图8 爬虫实例：“股吧”热度获取结果

指数代码： 000016.SH　　提取

上证50	近一月发帖数
600000.SH	652
600016.SH	643
600019.SH	1180
600028.SH	1322
600029.SH	1090
600030.SH	1858
600036.SH	628
600048.SH	948
600050.SH	890
600104.SH	1652
600196.SH	869
600276.SH	1038
600309.SH	2005
600340.SH	1020
600519.SH	3526
600547.SH	988
600585.SH	1249
600606.SH	1166
600690.SH	936
600703.SH	679

资料来源：东方财富“股吧”，海通证券研究所

图9 “天猫”网站 ONLY 专卖截图



资料来源：“天猫”网站，海通证券研究所

图10 “天猫”网站 ONLY 专卖产品信息获取结果

品牌	品名	店名	价格	月销量	评论数	月销售额
Only	only秋冬高腰加绒深色修身牛仔长裤	ONLY官方旗舰店	309	1460	119	451140
Only	only冬季磨毛薄绒高腰紧身黑瘦牛仔裤	ONLY官方旗舰店	274.5	970	80	266265
Only	only冬季高腰磨毛薄绒修身牛仔裤	ONLY官方旗舰店	384.3	466	44	179083.8
Only	only冬季高腰弹力紧身牛仔裤	ONLY官方旗舰店	274.5	367	121	100741.5
Only	only夏季简约高腰修身牛仔裤	ONLY官方旗舰店	274	613	1552	167962
Only	only冬季黑色高腰紧身牛仔裤	ONLY官方旗舰店	274.5	1174	90	322263
Only	only限售秋冬高腰弹力修身牛仔裤	ONLY官方旗舰店	249	3583	1468	892167
Only	only冬季舒适高腰修身显瘦牛仔裤	ONLY官方旗舰店	247	1301	606	321347
Only	only冬季抽绳显瘦字母时尚连帽羽绒服	ONLY官方旗舰店	1249.5	1108	30	1384446
Only	only冬季高腰显瘦微喇牛仔裤	ONLY官方旗舰店	299.5	528	108	158136
Only	only冬季不规则加绒加厚紧身牛仔裤	ONLY官方旗舰店	274.5	396	28	108702
Only	only秋冬简约修身卷边低腰牛仔裤	ONLY官方旗舰店	199.5	878	53	175161
Only	only冬季少女高腰宽松牛仔哈伦裤	ONLY官方旗舰店	384.3	595	66	228658.5
Only	only冬季时尚连帽毛绒宽松羽绒服	ONLY官方旗舰店	1539.3	210	3	323253
Only	only冬季高腰宽松牛仔九分裤	ONLY官方旗舰店	274.5	398	89	109251
Only	only冬季弹力高腰紧身铅笔牛仔裤	ONLY官方旗舰店	274.5	697	202	191326.5
Only	only冬季简约时尚高腰显瘦直筒牛仔裤	ONLY官方旗舰店	199.5	1203	129	239998.5
Only	only冬季简约高腰弹力舒适牛仔裤	ONLY官方旗舰店	384.3	383	126	147186.9
Only	only夏季宽松蕾丝刺绣圆领套头卫衣	ONLY官方旗舰店	199	1419	2047	282381
Only	only秋冬可拆卸加绒连帽卫衣	ONLY官方旗舰店	249	1225	1230	305025
Only	only冬季连帽可拆卸领子毛领羽绒服	ONLY官方旗舰店	1399.3	411	13	575112.3
Only	only冬季简约时尚显瘦拼接高腰牛仔裤	ONLY官方旗舰店	249	1111	581	276639
Only	only秋冬时尚高腰不规则牛仔裤	ONLY官方旗舰店	249	1229	1018	306021
Only	only秋冬百搭显瘦直筒牛仔裤	ONLY官方旗舰店	224	1205	264	269920

资料来源：“天猫”网站，海通证券研究所

在通过爬虫获取数据时，有一点非常值得注意，即网站的设立通常以提供浏览服务为目标。为保证真实使用网站服务的用户的访问体验，网站通常并不欢迎用户利用爬虫获取数据。因此，很多网站都做了反爬虫限制，如，限制 IP 地址短时间内的访问次数，需要重复验证等降低数据自动获取速度的方式。

更重要的是，由于绝大部分网站存在的目标并不是成为数据提供商，所以用网络爬虫技术搜集数据本质上是一种搭便车行为。很多时候，普通用户并没有浏览全部数据的需求，而网站也往往不保证历史所有信息都可以从页面中获取。因此，网络爬虫得到的数据作为一种自动的，实时的抽样工具非常有效。倘若希望通过数据对于某些问题的全貌进行统计论证，那么所得到的结论往往有失公允。

3.2 数据特征提取技术

从存储形式来区分，可以将数据分为结构化和非结构化两类。

结构化数据：即，行数据。可以直接存储在数据库里、用二维表进行存储的数据。

常用的金融数据，如，价量信息、财务信息、标的属性信息等，都属于结构化数据，可直接存储于关系数据库中读取并使用。结构化数据的特征是易于处理，数量信息可直接输入计量模型，非数量信息一般表意明确简洁，也可以方便地用枚举等方法量化和统计。

非结构化数据：无法直接用二维表结构来逻辑表达实现的数据，如，文本、图片、视频等。相比结构化数据，非结构化数据往往需要经过更加复杂的预处理，才能提取到有效的数量信息，进而被存储应用。

对于非结构化数据的信息提取，优先处理其可结构化的属性信息是比较有效的做法。直接对文本、图片提取数量信息，技术复杂度较高，很难保证百分之百的准确性。如果可从属性信息的提取过程中获得有效的数量信息，将是一种事半功倍的做法。

当非结构化数据没有足够的属性信息时，直接从非结构化数据本身提取数量信息就是唯一的办法。相较于图片、视频，投资领域中较为常见且有效的非结构化数据多为金融文本，如公司公告原文，公司新闻原文等。很多公司的重大事件往往隐藏在这些文本信息当中，利用自然语言处理技术可对文本中投资者所需要的信息进行提取。

图11 负面清单公司筛选示例

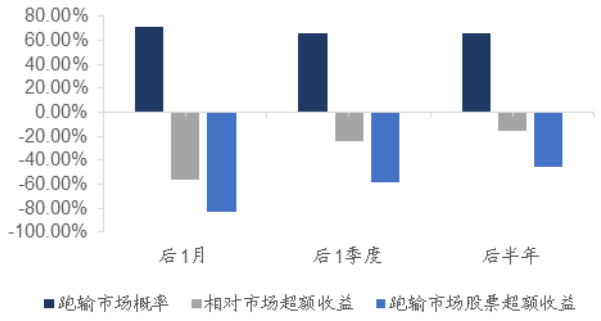
代码	名称	处罚公告日期	备注	指数信息
300005.SZ	探路者	20190215	监管关注	
600078.SH	澄星股份	20190214	公开处罚	
000553.SZ	安道麦A	20190213	公开处罚 责令整改	沪深300
600288.SH	大恒科技	20190213	监管关注	
600165.SH	新日恒力	20190211	公开谴责	
300356.SZ	光一科技	20190211	公开批评	
601969.SH	海南矿业	20190201	公开批评	中证500
600515.SH	海航基础	20190201	公开批评	
600119.SH	长江投资	20190130	公开批评	
601127.SH	小康股份	20190130	监管关注	中证500
002592.SZ	八菱科技	20190128	监管关注	
601099.SH	太平洋	20190126	责令改正	
000007.SZ	全新好	20190126	其他	
300093.SZ	金刚玻璃	20190125	其他	
300362.SZ	天翔环境	20190125	其他	
300313.SZ	天山生物	20190125	其他	
002259.SZ	ST升达	20190125	其他	
600226.SH	升华股份	20190124	监管关注	
300325.SZ	德威新材	20190124	公开处罚	
603301.SH	振德医疗	20190124	公开处罚 责令改正	
000806.SZ	银河生物	20190124	其他	
002450.SZ	ST康得新	20190123	其他	沪深300 中小板指
000565.SZ	渝三峡A	20190123	公开处罚 责令整改	
002102.SZ	ST冠福	20190121	其他	
000588.SZ	粤水电	20190119	其他	

资料来源：Wind，海通证券研究所

如上图所示，通过公司公告文本的分析，可以确认出违法违规、被监管机构问询、被证监会立案等相关事件最早的披露时点，从而在行政处罚事先告知阶段就知晓可能被处罚的标的。

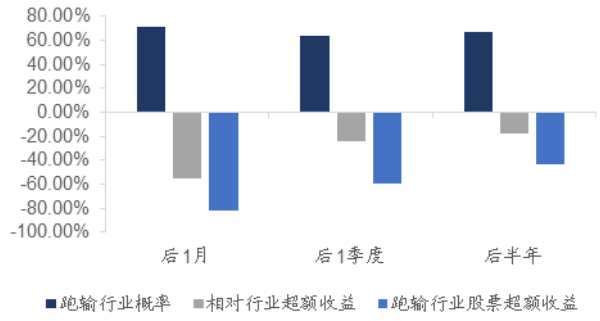
本文以证监会立案调查事件为例，考察公司被文本分析第一次被扫描到相关公告后，股价在未来一个月、一个季度及半年内，相对 Wind 全 A 指数和所在行业指数的表现。结果如以下两图所示。

图12 证监会立案事件标的相对市场表现



资料来源：Wind，海通证券研究所

图13 证监会立案事件标的相对行业表现



资料来源：Wind，海通证券研究所

上图统计了 2001 年至今，发生证监会立案调查事件的股票，事件发生之后一个月，一个季度，以及半年内的表现情况。其中，在剔除长时间停牌股票后，事件发生后一个月跑输 Wind 全 A 指数的概率达到 71.65%，相对市场平均年化超额收益为-56.36%，跑

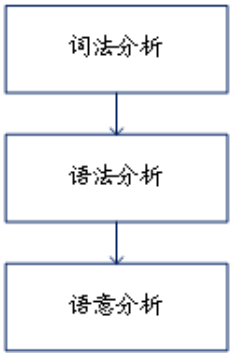
输指数的股票的平均年化超额收益为-82.44%。若与股票所在行业的指数比较，跑输概率为 71.34%，相对于行业指数的平均超额收益为-55.24%，跑输行业指数的股票的平均超额收益为-82.06%。

不过，从上图也可发现，随着时间的推移，这类股票跑输市场与所在行业指数的概率、负向年化超额收益的幅度，均有所收窄。这意味着尽可能早地发现被立案调查的股票，便可以尽可能多地减少该事件带来的负面影响。

此外，从公司发布被证监会立案调查的公告，到最终调查结果出炉，往往会历经一年甚至以上的时间。在此期间，也会连续发布调查进展的相关公告。但是，一般的金融信息数据源在调查结果出炉之前，并不会对该股票的这一事件进行标注，直到调查尘埃落定。此时再进行相应的操作，显然为时已晚。由此可见，利用文本挖掘方法，可以及早发现此类重大风险事件，帮助投资者有效规避。

图 14 展示了文本分析的三个步骤或层次。

图14 文本分析层次流程图



资料来源：海通证券研究所整理

词法分析：传统意义上的分词器。即，将字符序列转换为单词（Token）序列的过程。结果可应用于实体、行为的识别，情感分析等领域。

语法分析：按照源语言的语法规则，从词法分析的结果中识别出相应的语法结构。获取语法结构往往是为了准确识别语句表达的含义做准备。

语意分析：根据一套变换规则将语法结构映射到语义符号（如逻辑表达式、语义网络、中间语言等），得到文本表达含义的过程。从原理上说，语意分析可以得到文本中的全部表达信息，可用于提取文本中数量信息的含义，自动构建知识图谱等。

从词法分析到语意分析，后一步过程依赖于前一步过程的分析结果。因此，前一步分析的准确性会对后一步分析的效果产生极大的影响。而且，从分词到语意分析，技术难度将成倍增加。从应用角度来看，很多基于文本的数据特征提取，依靠语法分析结果便可达到，例如，文本中的特定实体匹配、统计等。因此依据不同的应用场合与目标，进行不同程度的文本分析是实际操作中更为高效的做法。

目前，中文文本分析已经有了长足发展，从最简单的分词到语意分析都有非常多的工具可以使用。例如，NLPIR 自然语言处理库，FudanNLP 自然语言处理库等。这些开源的自然语言处理工具，可以很好地帮助投资者对于投资中的相关文本进行词法分析、词性标注、语法结构分析等工作，最终从文本数据中获取想要的信息。

无论是词法分析还是语意分析，上述提及的自然语言处理都是通过神经网络等模型，分析先验的自然语言数据得到的。所以，为了提升自然语言处理库的使用范围，自然语言处理一般都会采用人民日报等标准的、涉及领域较广的媒体数据作为训练样本。这种设定对标准的新闻类语言会有较好的处理效果，但对某些特定领域，处理的准确率可能会有所下降。因此，一般的自然语言处理库都会提供自定义字典等扩展功能，以提升在特定领域的应用效果。

投资中所需要分析的文本信息一般来自于金融媒体的新闻、公司公告、研究员报告等行文比较规范、标准的书面语文本。因此，只要准备好尽可能多的投资相关术语、公司名称、标的名称等专有名词字典，便可以很好地达到处理投资相关文本数据的目标。

3.3 数据结构化方法

数据的逻辑结构决定了处理的效率及从数据存储系统中存取的效率。好的数据逻辑结构可以在提升处理效率的同时，扩展数据处理的可能性，挖掘更丰富的数据特性。

传统的数据结构化方法为二维结构，即，将所有的数据转化为二维矩阵进行存储。这种数据结构与关系数据库的数据存储方式相对应。同时，矩阵也是 MATLAB 等主流科学计算语言的基本数据结构。这种逻辑上的共通，方便了对数据进行存储预处理的逻辑构建。

随着 MongoDB 等非结构化数据的兴起，数据可以直接以“对象”的形式，在数据存储系统与逻辑处理过程中传递。与简单的矩阵方式不同，“对象”的形式除了保存数据信息之外，还保留了数据与数据之间的层次关系，能对进一步的数据分析提供更多信息。

除以上常用数据结构化方法之外，采用图的方式，构建知识图谱，也是一种重要的数据结构方式。

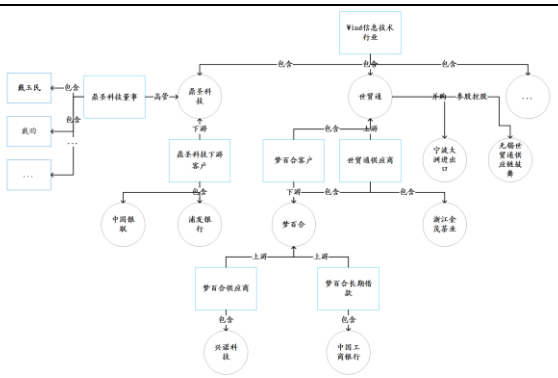
知识图谱又称科学知识图谱，在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列不同的图形。目的是用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。

知识图谱可以存储并展示一切“实体”与“关系”的关联性信息。在投资领域的应用中，“实体”，广义上包括可交易标的、公司、经济指标、产品价格、抽象分类等等。“关系”，指实体之间任何可能存在的关联方式，包括 A 是 B 的股东、正相关关系、因果关系等等。

通过上文的数据提取技术，就可以从各种数据源获取投资过程中所需要研究的各种“实体”和“关系”，通过不同“关系”，将不同“实体”链接起来。例如，以上市公司为“实体”，不同上市公司之间可能存在各种“关系”。通过构建公司之间关联关系的知识图谱，可以直观地看到不同公司之间的一些潜在关联属性，从而对产业链上下游、主题概念板块等公司间的潜在相关性有更深入的了解。

下图以鼎圣科技（839660.OC），世贸通（834896.OC），梦百合（603313.SH）三个公司为示例，构建了一个简单的知识图谱。从这三家公司“实体”出发，通过构建这三个公司“实体”与各种行业分类，供应商等“实体”的“关系”，可以清晰地发现这三家公司通过什么样的“关系”产生了关联。

图15 上市公司基本信息知识图谱示例



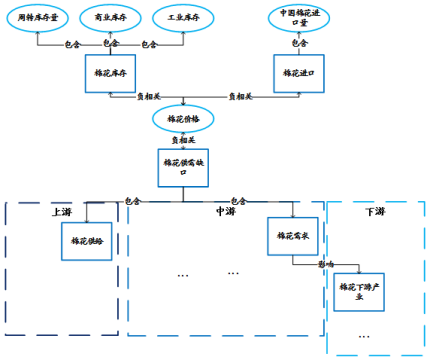
资料来源：Wind，海通证券研究所

挖掘产业链结构，确认企业在产业链当中的位置是知识图谱的一个重要作用。如上文所述，一切的概念、分类、资源品均可被标注为实体，从而透过知识图谱的构建，由

下而上地将资源品到终端产品的所有公司串联起来，得到某一产业的产业链结构。

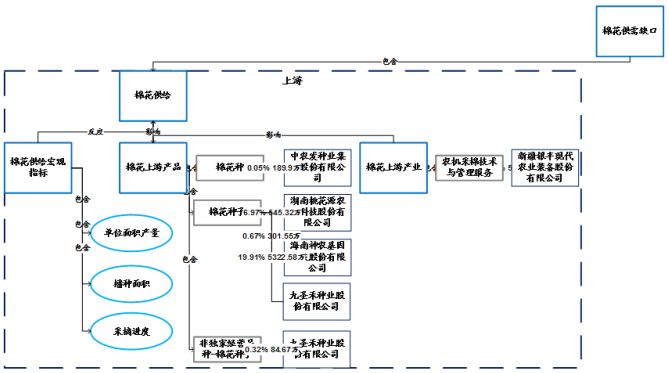
下图是以棉花为例构建的产业链示例。通过链接到影响棉花价格的供需关系“实体”，查询到与供需有关的商品“实体”，再链接到公司主营业务“实体”，可以确认不同公司在棉花产业链中的上下游位置。如，棉花的上游包括棉花种子的生产销售，农业采棉技术的服务管理等，而下游则包括棉纱、棉布等商品的生产和销售。通过棉花种子经营，棉纱经营，棉布经营的有关主营业务“实体”，就可以将有相关主营业务的公司“实体”与棉花的供给和需求“实体”链接，从而得到这些公司在棉花产业链当中的位置。

图16 棉花产业链整体示意图



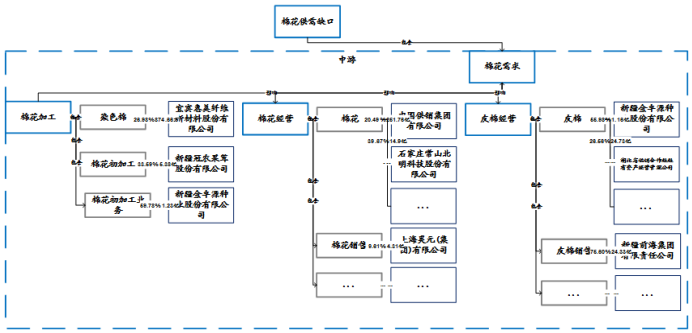
资料来源：Wind，海通证券研究所

图17 棉花产业链上游示意图



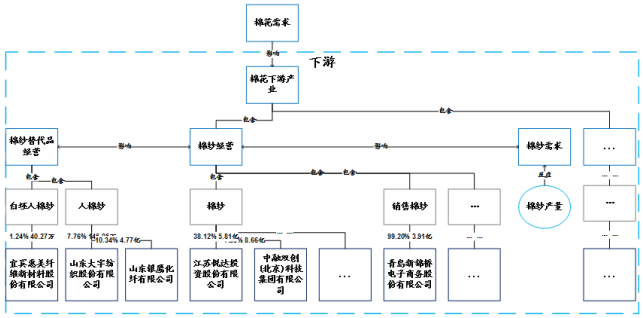
资料来源：Wind，海通证券研究所

图18 棉花产业链中游示意图



资料来源：Wind，海通证券研究所

图19 棉花产业链下游示意图



资料来源：Wind，海通证券研究所

知识图谱的构建是一种典型的自下而上，自组织构建的过程。在这个过程中，往往需要依据有关先验信息，如棉花的上游产业，下游产品分别包括哪些对象。先验信息的获取与分析是能否构建有效知识图谱的关键所在。

前文所提到的数据处理与数据分析技术，特别是自然语言处理技术，便是获取先验知识的关键。一般来说，结构化数据中的先验知识最容易被获取，如上例中，主营业务实体与公司实体之间的关联可以直接从财务报表中获取。而棉花上下游的产品和产业各有哪些，就不那么容易获取了。如果没有足够的数据来源或者分析技术不够成熟，可能需要通过手工方式进行构建。

3.4 数据存储方式

所有的数据在获取并预处理之后，都需要存入公共的数据存储系统，方便在数据分析时随时调用。目前主流的数据存储系统主要有关系数据库、非关系数据库、列数据库、内存数据库等。

关系数据库：一种以表格为载体的数据存储工具。以列表示数据类型，以行表示一个数据实体。关系数据库是目前应用最广泛的一种数据存储形式，较常用的包括 SQL Server，Oracle，MySql 等。

列数据库：相较于传统关系数据库，虽然数据在逻辑上依然以行和列表示，但在物理存储中用的是唯一字符串的方式。传统关系数据库的优势是可以快速定位到某个实体，而列数据库则在读取某一列数据时有更高的效率。因此，列数据库一般会被用来存储高频行情等海量的时间序列数据。

非关系数据库：如，面向海量文档存储的 MongoDB，面向高性能并发读写的 Redis 等。存储的基本数据结构将不仅仅是表格，更多的是以键值对（Key-Value）形式构建。

内存数据库：与传统基于硬盘存储的数据库不同，该数据库的数据存储于内存当中。用户可以更加方便地存取，是提升某些热度较高数据存取效率的重要工具。

从某种意义上说，几乎所有的逻辑结构都可以用不同的形式存储于不同类型的数据库系统当中。以知识图谱为例，从数据结构的角度来看，其本质上是一张图，实体即为图的顶点，关系即为图的边。在存储当中，以“关系”，即图的边作为存储核心，每一个“关系”都会囊括两个“实体”。因此，关系、实体 A、实体 B 即可作为一行存入传统的关系型数据库。使用时，读取关系型数据库中满足要求的“关系”，即可构建出一张完整的知识图谱。

由于逻辑数据结构均可以用某种方式映射到不同类型的数据库系统，因此数据库系统的稳定性，易用性是提升效率的关键。新型数据库，如 MongoDB 等，的确提供了非常多的易用特性。但在处理存取逻辑关系明晰，数据结构简单固定的数据时，其维护难度、存取效率并不一定优于传统的关系型数据库。所以，从应用角度出发，了解所有主流数据库系统的优缺点和目前发展状况，是保证数据存取效率的关键。

3.5 数据分析方法

数据挖掘所使用的分析方法主要有以下几类：

分类：对已经分好类别的样本进行建模，用所构建的模型对未标记类别的样本进行分类的过程。

聚类：将拥有相似特性的样本汇聚到一类当中，从而自动地将所有样本分为 N 类的过程。

估计：通过给定的输入数据，得出未知变量的过程。

预测：通过分类或估计给出的模型，对变量进行外推的过程。

相关性分组或关联规则：分析不同事件、实体之间相关关系的过程。

描述和可视化：对数据挖掘结果的一种表示方式。

上述每一种数据挖掘方法，目前都有相应的经典模型可供选用。从传统的 OLS 回归、LASSO 回归，到新进的 KNN、KMeans 等简单机器学习算法，再到 SVM、人工神经网络、深度学习等最近较热的复杂机器学习算法，都是进行数据挖掘的重要工具。

在选择算法时，数据规模、数据特性和目标的匹配性是重要标准。以 SVM 与人工神经网络为例，一般情况下，人工神经网络有更好的分类或者预测效果。然而随着数据维度和数据量的增长，人工神经网络的训练复杂度也会快速攀升。如果对于预测结果的精度没有那么高的要求时，反而简单的数据挖掘算法可能更有效率。

4. 总结与讨论

随着计算机技术的飞速进步，中国证券市场的数据量呈爆发式增长。合理地利用数据挖掘工具，并从数据中提取到对投资最有益的信息，是一件非常重要且有意义的工作。

本文介绍了数据挖掘技术框架中，从数据采集到数据处理和数据分析的全部过程，

并对其中的技术要点做了简要介绍。同时，本文也给出了在不同步骤中需要注意的关键性问题，希望能给投资者在构建自己的数据挖掘工具的过程中提供参考。

未来，海通量化团队将继续利用金融科技（Fintech）与数据挖掘技术拓展金融工程的认知和应用边界。并以知识图谱为核心，构建丰富的数据分析工具，为投资者提供不一样的市场解读视角与方式。

5. 风险提示

数据挖掘是从历史先验数据获取经验模型的方法，存在模型失效可能。

信息披露

分析师声明

冯佳睿 金融工程研究团队
余浩淼 金融工程研究团队

本人具有中国证券业协会授予的证券投资咨询执业资格，以勤勉的职业态度，独立、客观地出具本报告。本报告所采用的数据和信息均来自市场公开信息，本人不保证该等信息的准确性或完整性。分析逻辑基于作者的职业理解，清晰准确地反映了作者的研究观点，结论不受任何第三方的授意或影响，特此声明。

法律声明

海通证券股份有限公司研究所

路 颖 所长
(021)23219403 luying@htsec.com

高道德 副所长
(021)63411586 gaodd@htsec.com

姜 超 副所长
(021)23212042 jc9001@htsec.com

邓 勇 副所长
(021)23219404 dengyong@htsec.com

荀玉根 副所长
(021)23219658 xyg6052@htsec.com

涂力磊 所长助理
(021)23219747 tll5535@htsec.com

宏观经济研究团队
姜 超(021)23212042 jc9001@htsec.com
于 博(021)23219820 yb9744@htsec.com
李金柳(021)23219885 ljl11087@htsec.com
联系人
宋 潇(021)23154483 sx11788@htsec.com
陈 兴(021)23154504 cx12025@htsec.com

金融工程研究团队
高道德(021)63411586 gaodd@htsec.com
冯佳睿(021)23219732 fengjr@htsec.com
郑雅斌(021)23219395 zhengyb@htsec.com
罗 蕾(021)23219984 ll9773@htsec.com
沈泽承(021)23212067 szc9633@htsec.com
余浩淼(021)23219883 yhm9591@htsec.com
袁林青(021)23212230 ylq9619@htsec.com
姚 石(021)23219443 ys10481@htsec.com
吕丽颖(021)23219745 lly10892@htsec.com
周一洋(021)23219774 zyy10866@htsec.com
联系人
张振岗(021)23154386 zzg11641@htsec.com
颜 伟(021)23219914 yw10384@htsec.com
梁 镇(021)23219449 lz11936@htsec.com

金融产品研究团队
高道德(021)63411586 gaodd@htsec.com
倪韵婷(021)23219419 niyt@htsec.com
陈 瑶(021)23219645 chenyao@htsec.com
唐洋运(021)23219004 tangyy@htsec.com
宋家骥(021)23212231 sjj9710@htsec.com
皮 灵(021)23154168 pl10382@htsec.com
徐燕红(021)23219326 xyh10763@htsec.com
谈 鑫(021)23219686 tx10771@htsec.com
王 毅(021)23219819 wy10876@htsec.com
蔡思圆(021)23219433 csy11033@htsec.com
联系人
庄梓恺(021)23219370 zzk11560@htsec.com
谭实宏(021)23219445 tsh12355@htsec.com

固定收益研究团队
姜 超(021)23212042 jc9001@htsec.com
朱征星(021)23219981 zzx9770@htsec.com
周 霞(021)23219807 zx6701@htsec.com
姜珮珊(021)23154121 jps10296@htsec.com
杜 佳(021)23154149 dj11195@htsec.com
联系人
李 波(021)23154484 lb11789@htsec.com

策略研究团队
荀玉根(021)23219658 xyg6052@htsec.com
钟 青(010)56760096 zq10540@htsec.com
高 上(021)23154132 gs10373@htsec.com
李 影(021)23154117 ly11082@htsec.com
姚 佩(021)23154184 yp11059@htsec.com
周旭辉 zxh12382@htsec.com
联系人
唐一杰(021)23219406 tyj11545@htsec.com
郑子勋(021)23219733 zzx12149@htsec.com
王一潇(021)23219400 wyx12372@htsec.com

中小市值团队
张 宇(021)23219583 zy9957@htsec.com
钮宇鸣(021)23219420 ymniu@htsec.com
孔维娜(021)23219223 kongwn@htsec.com
潘莹练(021)23154122 pyl10297@htsec.com
联系人
程碧升(021)23154171 cbs10969@htsec.com
相 姜(021)23219945 xj11211@htsec.com

政策研究团队
李明亮(021)23219434 lml@htsec.com
陈久红(021)23219393 chenjiuhong@htsec.com
吴一萍(021)23219387 wuyiping@htsec.com
朱 蕾(021)23219946 zl8316@htsec.com
周洪荣(021)23219953 zhr8381@htsec.com
王 旭(021)23219396 wx5937@htsec.com

石油化工行业
邓 勇(021)23219404 dengyong@htsec.com
朱军军(021)23154143 zjj10419@htsec.com
联系人
胡 歆(021)23154505 hx11853@htsec.com
张 璇(021)23219411 zx12361@htsec.com

医药行业
余文心(0755)82780398 ywx9461@htsec.com
郑 琴(021)23219808 zq6670@htsec.com
联系人
贺文斌(010)68067998 hwb10850@htsec.com
范国钦 02123154384 fgq12116@htsec.com
梁广楷(010)56760096 lgk12371@htsec.com
吴佳栓(010)56760092 wjs11852@htsec.com

汽车行业
王 猛(021)23154017 wm10860@htsec.com
杜 威(0755)82900463 dw11213@htsec.com
联系人
曹雅倩(021)23154145 cyq12265@htsec.com

公用事业
吴 杰(021)23154113 wj10521@htsec.com
张 磊(021)23212001 zl10996@htsec.com
戴元灿(021)23154146 dyc10422@htsec.com
联系人
傅逸帆(021)23154398 fyf11758@htsec.com

批发和零售贸易行业
汪立亭(021)23219399 wanglt@htsec.com
李宏科(021)23154125 lhk11523@htsec.com
联系人
史 岳 sy11542@htsec.com
高 瑜(021)23219415 gy12362@htsec.com
谢茂萱 xmx12344@htsec.com

互联网及传媒
郝艳辉(010)58067906 hyh11052@htsec.com
孙小雯(021)23154120 sxw10268@htsec.com
毛云聪(010)58067907 myc11153@htsec.com
联系人
陈星光(021)23219104 cxg11774@htsec.com

有色金属行业
施 毅(021)23219480 sy8486@htsec.com
联系人
李姝醒(021)23219401 lsx11330@htsec.com
陈晓航(021)23154392 cxh11840@htsec.com
甘嘉尧(021)23154394 gjy11909@htsec.com

房地产行业
涂力磊(021)23219747 tll5535@htsec.com
谢 盐(021)23219436 xiey@htsec.com
杨 凡(021)23219812 yf11127@htsec.com
金 晶(021)23154128 jj10777@htsec.com

电子行业
陈 平(021)23219646 cp9808@htsec.com
尹 苓(021)23154119 yl11569@htsec.com
谢 磊(021)23212214 xl10881@htsec.com
联系人
石 坚(010)58067942 sj11855@htsec.com

煤炭行业
李 淼(010)58067998 lm10779@htsec.com
戴元灿(021)23154146 dyc10422@htsec.com
吴 杰(021)23154113 wj10521@htsec.com
联系人
王 涛(021)23219760 wt12363@htsec.com

电力设备及新能源行业
张一弛(021)23219402 zyc9637@htsec.com
房 青(021)23219692 fangq@htsec.com
曾 彪(021)23154148 zb10242@htsec.com
徐柏乔(021)23219171 xbq6583@htsec.com
张向伟(021)23154141 zxw10402@htsec.com
联系人
陈佳彬(021)23154513 cjb11782@htsec.com

基础化工行业
刘　威(0755)82764281　lw10053@htsec.com
刘海荣(021)23154130　lhr10342@htsec.com
张翠翠(021)23214397　zcc11726@htsec.com
孙维容(021)23219431　swr12178@htsec.com
联系人
李　智(021)23219392　lz11785@htsec.com

计算机行业
郑宏达(021)23219392　zhd10834@htsec.com
杨　林(021)23154174　yl11036@htsec.com
鲁　立(021)23154138　ll11383@htsec.com
于成龙　ycl12224@htsec.com
黄竞晶(021)23154131　hjj10361@htsec.com
联系人
洪　琳(021)23154137　hl11570@htsec.com

通信行业
朱劲松(010)50949926　zjs10213@htsec.com
余伟民(010)50949926　ywm11574@htsec.com
张　弋 01050949962　zy12258@htsec.com
张峥青(021)23219383　zzq11650@htsec.com

非银行金融行业
孙　婷(010)50949926　st9998@htsec.com
何　婷(021)23219634　ht10515@htsec.com
联系人
李芳洲(021)23154127　lfz11585@htsec.com

交通运输行业
虞　楠(021)23219382　yun@htsec.com
罗月江（010）56760091　lyj12399@htsec.com
联系人
李　丹(021)23154401　ld11766@htsec.com
党新龙(0755)82900489　dxl12222@htsec.com

纺织服装行业
梁　希(021)23219407　lx11040@htsec.com
联系人
盛　开(021)23154510　sk11787@htsec.com
刘　溢(021)23219748　ly12337@htsec.com

建筑建材行业
冯晨阳(021)23212081　fcy10886@htsec.com
联系人
申　浩(021)23154114　sh12219@htsec.com

机械行业
余炜超(021)23219816　swc11480@htsec.com
耿　耘(021)23219814　gy10234@htsec.com
杨　震(021)23154124　yz10334@htsec.com
沈伟杰(021)23219963　swj11496@htsec.com
周　丹　zd12213@htsec.com

钢铁行业
刘彦奇(021)23219391　liuyq@htsec.com
刘　璇(0755)82900465　lx11212@htsec.com
联系人
周慧琳(021)23154399　zhl11756@htsec.com

建筑工程行业
杜市伟(0755)82945368　dsw11227@htsec.com
张欣劼　zxj12156@htsec.com
李富华(021)23154134　lfh12225@htsec.com

农林牧渔行业
丁　频(021)23219405　dingpin@htsec.com
陈雪丽(021)23219164　cxl9730@htsec.com
陈　阳(021)23212041　cy10867@htsec.com
联系人
孟亚琦　myq12354@htsec.com

食品饮料行业
闻宏伟(010)58067941　whw9587@htsec.com
成　珊(021)23212207　cs9703@htsec.com
唐　宇(021)23219389　ty11049@htsec.com

军工行业
蒋　俊(021)23154170　jj11200@htsec.com
刘　磊(010)50949922　ll11322@htsec.com
张恒晅　zhx10170@htsec.com
联系人
张宇轩(021)23154172　zyx11631@htsec.com

银行行业
孙　婷(010)50949926　st9998@htsec.com
解巍巍　xww12276@htsec.com
林加力(021)23214395　ljl12245@htsec.com
联系人
谭敏沂(0755)82900489　tmy10908@htsec.com

社会服务行业
汪立亭(021)23219399　wanglt@htsec.com
陈扬扬(021)23219671　cyy10636@htsec.com
许樱之　xyz11630@htsec.com

家电行业
陈子仪(021)23219244　chenzy@htsec.com
李　阳(021)23154382　ly11194@htsec.com
朱默辰(021)23154383　zmc11316@htsec.com
联系人
刘　璐(021)23214390　ll11838@htsec.com

造纸轻工行业
衣桢永(021)23212208　yzy12003@htsec.com
曾　知(021)23219810　zz9612@htsec.com
赵　洋(021)23154126　zy10340@htsec.com

研究所销售团队

深广地区销售团队
蔡铁清(0755)82775962　ctq5979@htsec.com
伏财勇(0755)23607963　fcy7498@htsec.com
辜丽娟(0755)83253022　gulj@htsec.com
刘晶晶(0755)83255933　liujj4900@htsec.com
王雅清(0755)83254133　wyq10541@htsec.com
饶　伟(0755)82775282　rw10588@htsec.com
欧阳梦楚(0755)23617160
oymc11039@htsec.com
宗　亮　zl11886@htsec.com
巩柏含　gbh11537@htsec.com

上海地区销售团队
胡雪梅(021)23219385　huxm@htsec.com
朱　健(021)23219592　zhuj@htsec.com
季唯佳(021)23219384　jiwj@htsec.com
黄　毓(021)23219410　huangyu@htsec.com
漆冠男(021)23219281　qgn10768@htsec.com
胡宇欣(021)23154192　hyx10493@htsec.com
黄　诚(021)23219397　hc10482@htsec.com
毛文英(021)23219373　mwy10474@htsec.com
马晓男　mxn11376@htsec.com
杨祎昕(021)23212268　yyx10310@htsec.com
张思宇　zsy11797@htsec.com
慈晓聪(021)23219989　cxc11643@htsec.com
王朝领　wcl11854@htsec.com
邵亚杰 23214650　syj12493@htsec.com
李　寅 021-23219691　ly12488@htsec.com

北京地区销售团队
殷怡琦(010)58067988　yyq9989@htsec.com
郭　楠 010-5806 7936　gn12384@htsec.com
吴　尹　wy11291@htsec.com
张丽萱(010)58067931　zlx11191@htsec.com
杨羽莎(010)58067977　yys10962@htsec.com
杜　飞　df12021@htsec.com
张　杨(021)23219442　zy9937@htsec.com
何　嘉(010)58067929　hj12311@htsec.com
李　婕　lj12330@htsec.com
欧阳亚群　oyyq12331@htsec.com

海通证券股份有限公司研究所
地址：上海市黄浦区广东路 689 号海通证券大厦 9 楼
电话：（021）23219000
传真：（021）23219392
网址：www.htsec.com