金融工程报告●深度报告 2023 年 12 月 29 日

大数据系列（4）资讯数据分析

核心观点：

- **另类大数据概述。**大数据的属性特征包括体量、速度、多样性和准确性，包括气候信息、卫星图像、数字图片和视频、轨迹记录或 GPS 信号，以及个人数据等。另类数据集合规模有约 1000 多个，买方机构购买另类数据的花费逐年增长，头部对冲基金使用的另类数据比例更大，使用比例最大的另类数据是网页数据、和社会舆情信息等。
- **新闻舆情数据源。**对国内学术新闻库做了介绍。对于其他新闻介绍，包括内容信息字段、所包含的类别、整体时间跨度、数据量大小、数据更新情况等。
- **新闻舆情数据统计。**对 wind 新闻进行详细统计分析，包括新闻数量，股票新闻、情感新闻占比，新闻量时间段差异，新闻源的情况，以及不同新闻类型标签包括研报类等，以及新闻不同重要程度，歧义新闻和人工新闻的，以及不同地区新闻差异等。
- **风险提示：**报告结论基于历史价格信息和统计规律，但二级市场受各种即时性政策影响易出现统计规律之外的走势，文章所引用的第三方相关数据资料等不构成推荐，报告阅读者需审慎参考报告结论。

分析师

吴俊鹏

☎：010-8097631

✉：wujunpeng@chinastock.com.cn

分析师登记编码：S0130517090001

相关研究

《大数据系列(1):舆情事件特征分析》

《大数据系列(2):舆情事件收益分析》

《大数据系列(3):新闻事件收益分析》

目 录

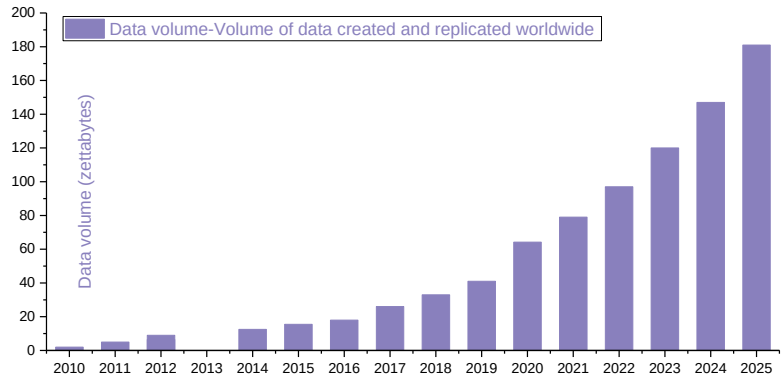
在之前的研究中我们分别通过三篇报告《大数据系列(1):舆情事件特征分析》、《大数据系列(2):舆情事件收益分析》和《大数据系列(3):新闻事件收益分析》尝试将舆情相关数据应用到投资的可能性。本文作为第四篇，从数据源的角度来统计分析新闻舆情的大致情况。首先第一部分阐述大数据（另类数据）的发展应用，第二部分介绍部分现有的平台中关于新闻舆情数据的情况，第三部分以 wind 新闻为例，详细进行统计分析。

一、另类大数据概述

（一）另类数据基本概况

传统的数据更多的集中于一些结构化、定期披露的数据，比如财务报告、机构调研信息等等。

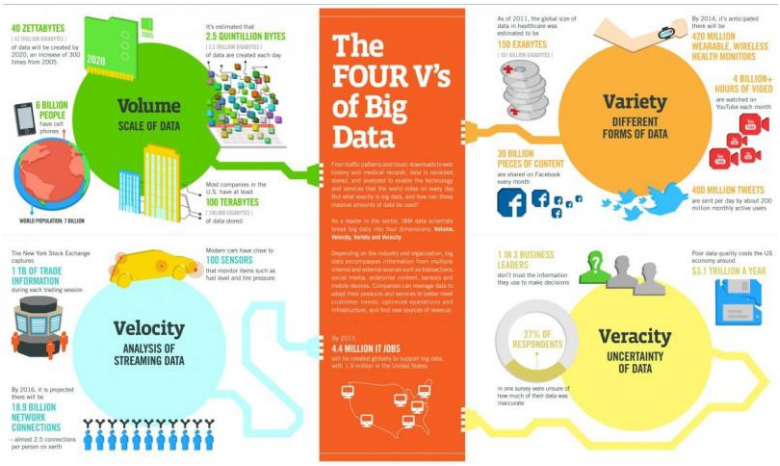
图1：全球数据量大小



资料来源：(IBM,.中国银河证券研究院

随着信息技术的发展，更多维度、非机构化的数据不断产生。大数据的属性特征包括“4V”：体量（数据规模）、速度（高速流数据的处理和分析）、多样性（异构数据）和准确性（数据来源可靠性、真实性），以及包括可扩展性和复杂性等。准确性尤其重要，因为用户可能很难做到评估所使用的数据集是否完整且可信。

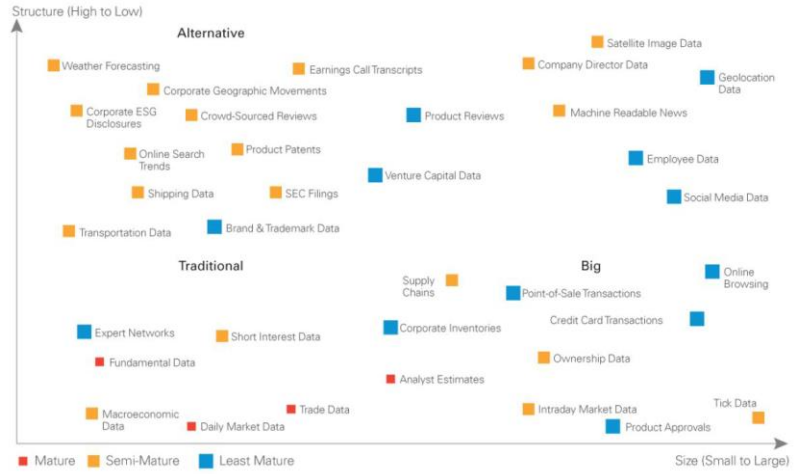
图2：大数据的四个维度



资料来源：OECD（2021）.中国银河证券研究院

大数据包括气候信息、卫星图像、数字图片和视频、轨迹记录或 GPS 信号，以及个人数据(姓名、照片、电子邮件地址、银行详细信息、社交网站上的帖子网络网站、医疗信息或计算机 IP 地址)等。随机机器学习等先进算法的发展，也可以更好的将这样数据应用于包括投资决策等各个方面。

图3：大数据源



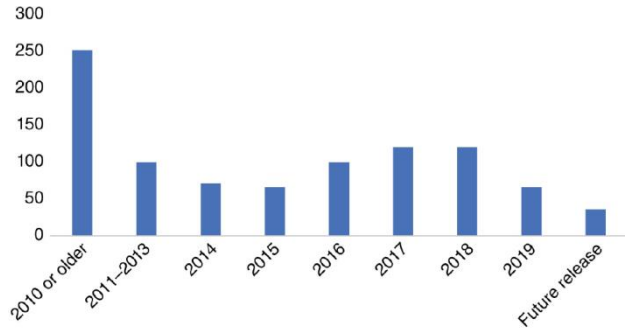
资料来源：OECD（2021）.中国银河证券研究院

（二）金融市场另类数据基本概况

金融市场另类数据投入规模

F.Norrestad2021 年 9 月在 Statista 上发布报告称约 54%的头部对冲基金使用的另类数据集超过 7 个，而其它的对冲基金仅有 8%。

图4：商业发布另类数据集数量

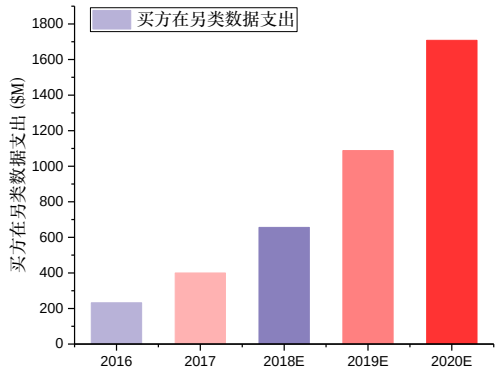


资料来源：ALEXANDER DENEV 等（2020）.中国银河证券研究院

另外据 Neudata 统计，另类数据集合规模有约 1000 多个。而 alternativedata 上统计发布的另类数据源有 445 个（2018 年）

据 alternativedata 网站发布的统计显示，买方机构 2019 年购买另类数据的花费为约 10.88 亿美元，2020 年为 17.08 亿美元。

图5：买方在另类数据支出



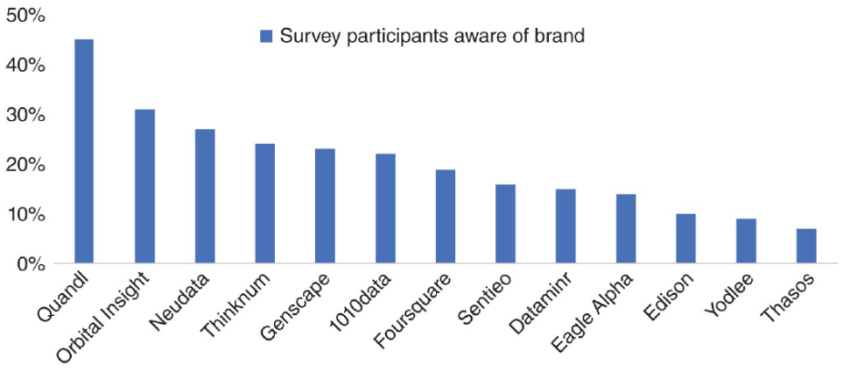
资料来源：alternativedata.org.中国银河证券研究院

根据 Oxylabs 和 Censuswide 的调查，“63%的受访者已经开始使用另类数据来帮助他们进行投资决策”。根据 Statista 的数据，买方机构 2019 年购买另类数据的花费为约 10.88 亿美元，2020 年为 17.08 亿美元。Global Alternative Data Market 在 2020 年调查报告支出全球另类数据市场规模将以 44%的年复合增长率增长，到 2026 年，将达到 111 亿美元。根据 Grand View Research 的报告，2020 年另类数据市场预计将以 58.5%的增长，到 2028 年将达到 693.6 亿美元（来源：Gautam Mitra 等（2023））。

金融市场另类数据品牌和类型

2018 年 Greenwich Associates 对于另类数据的使用等情况对 36 名市场参与者进行调查，排名第一的是 Quandl，而 Orbita Insight 是一家卫星图像数据集的供应商。

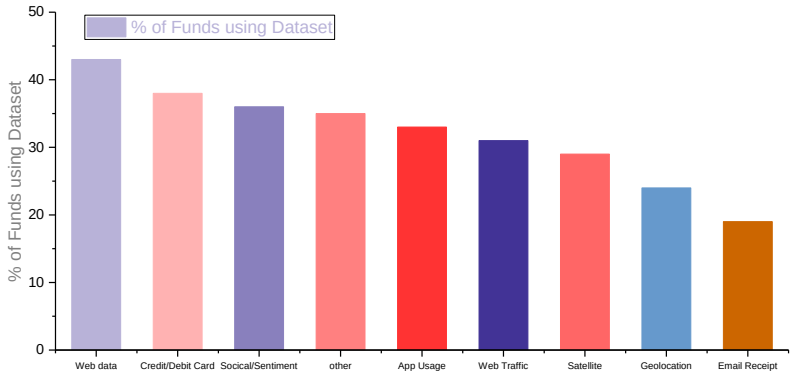
图6：最知名的另类数据提供商



资料来源：ALEXANDER DENEV 等（2020），中国银河证券研究院

此外，alternativedata 提供各种类型另类数据源，包括 App Usage、Credit/Debit Card、Data Aggregator、Data Broker、Email/Consumer Receipts、Geo-location、Other、Point of Sale、Public Data、Satellite、Sell-side、Social/Sentiment、Survey、Weather、Web Data 和 Web Traffic。

图7：基金不同类型另类数据使用比例

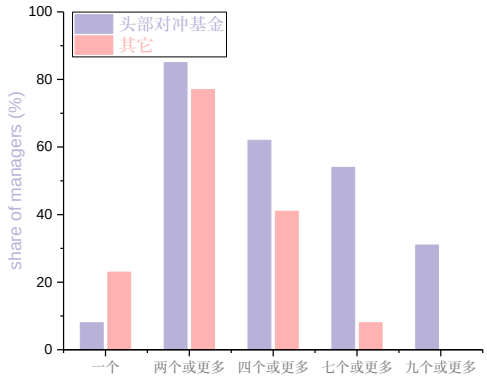


资料来源：alternativedata.org.中国银河证券研究院

据 alternativedata 统计，约 43%基金公司使用“Web Data”，“Social/Sentiment”使用比例为 36%。

金融市场另类数据使用情况

图8：对冲基金使用另类数据源情况

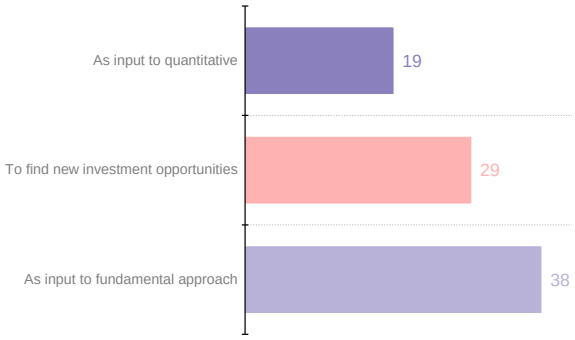


资料来源：Gautam Mitra 等（2023）,中国银河证券研究院

F.Norrestad2021 年 9 月在 Statista 上发布报告称约 54%的头部对冲基金使用的另类数据集超过 7 个，而其它的对冲基金仅有 8%。

对于另类数据使用者来说，19%的情况应用与量化分析，29%情况下用于寻找新的投资机会，38%应用与基本面的分析等。

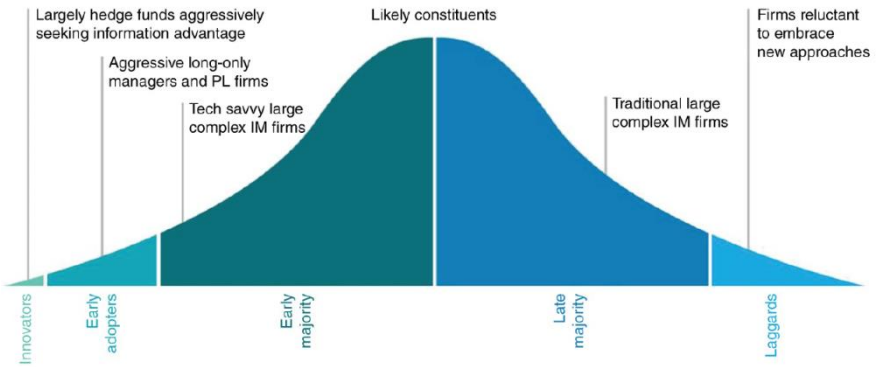
图9：对冲基金使用另类数据源应用（%）



资料来源：Gautam Mitra 等（2023），中国银河证券研究院

对于另类数据使用者来说，首先面临的就是价格成本；其次不同类型的数据覆盖面也是不同的，进而其可应用的投资范围也是不同的，影响策略的容量；第三，通常在使用常规数据进行投资时，通常都需要对相应策略进行回溯测试，而大部分的另类数据通常可用的历史数据较少。种种原因，都会影响到另类数据的实际使用。下图就很好的反应了不同类型投资者对另类数据使用的不同阶段。

图10：另类数据应用曲线



资料来源：ALEXANDER DENEV 等（2020），中国银河证券研究院

关于另类数据使用的情况我们先介绍到这里。下面对于基金机构对于另类数据中使用最多的新闻舆情信息数据（web data 相关）进行阐述。

二、新闻舆情数据源

（一）学术新闻库

《全国报刊索引数据库》

《全国报刊索引数据库》是上海图书馆主管主办，其包括目次库、篇名库、会议库、西文库以及其他专业数据库为一体，收录了自 1833 年至今的报刊篇目信息，收录报刊数量达 15000 余种（包括港台地区），是查找晚清时期、民国时期和解放初期文献的重要工具，是目前国内唯一揭示中文报刊资源时间跨度最大，报道报刊品种最多的报刊数据库产品。数据库目前年更新数据量 350 万余条。

全国报刊索引平台为个人用户提供的产品如下：（一）近代期刊、（二）现代期刊-现刊索引数据库和图片数据库，现刊索引数据库从 1980 年开始。

《慧科新闻搜索研究数据库》

慧科中文媒体资讯及商业情报数据库（WiseNews (慧科新闻) 已更改为 WisersOne (慧眼舆情)），开始于 1998 年，逾 3 9 亿篇文章存档，历史新闻可追溯至 1998 年，该数据库不仅涵盖了丰富的平面媒体资源，同时也收录源自 1200+报刊、10000+网站、1500+社交媒体的新闻资讯，其中港澳台地区主流媒体超过 95% 覆盖率，以及欧美、新加坡、泰国等地核心媒体，并以每日数百万篇的幅度增长。

图11：慧科新闻

资料来源：www.wisers.com.cn.中国银河证券研究院

《中国资讯行-中国经济新闻库》

1995 年，中国资讯行率先在互联网上建立自己的信息平台。中国资讯行（China InfoBank）是香港专门收集、处理及传播中国商业信息的企业，为世界各地各行各业的公司和研究机构提供信息。其数据产品包括 14 个大型专业数据库，内容涉及 19 个领域，197 个行业。14 个在线数据库拥有逾 100 亿汉字，每日增加逾 200 万汉字。

表1：中国资讯行数据库

更新资料库:	数据库名称	库记录数	最后更新日期
1	中国经济新闻库	7658361	20231228
2	中国商业报告库	812287	20231228
3	中国法律法规库	740960	20231228
4	中国统计数据库	931454	20231206
5	中国上市公司文献库	524210	20231228
6	中国医疗健康库	34521	20231228
7	INFOBANK 环球商讯库	1258675	20231228
参考资料库:			
1	中国人物库	17552	20000622
2	English Publications	193426	20020629
3	中国中央及地方政府机构库	163	20070129
4	中国拟建在建项目数据库	7980	20080214

5	中国企业产品库	279322	20070129
6	香港上市公司资料库（中文）	10432	20010131
7	名词解释库	1550	20000623

资料来源：www.infobank.cn，中国银河证券研究院

中国经济新闻库：数据库收录了中国范围内及相关的海外商业经济信息，以消息报导为主，数据源自中国千余种报章与期刊及部分合作伙伴提供的专业信息，按行业及地域分类，共包含 19 个领域 197 个类别，数据库每日更新。截至 2023 年 12 月 28 日，中国经济新闻库记录 7658361 条。

CnOpenData

CnOpenData 是覆盖经济、法律、医疗、人文等多个学科维度的综合型数据平台，并持续提供个性化数据定制服务，现拥有 200+个专题数据库，涵盖专利数据（1.1 亿+量级）、工商注册企业数据（2 亿+量级）、上市公司数据、土地数据、政府数据、新冠疫情数据、分地区数据、交通数据、气象数据等十大数据系列。

其中涉及新闻的有：CCTV 新闻联播文本数据、CNN 新闻文本数据、华尔街日报新闻文本数据、中国财经新闻报纸文本数据、A 股上市公司新闻舆情数据、中国新闻关联信息数据、中国新闻热度统计数据。CNN 新闻文本数据，CnOpenData 收集了 22 年来的 CNN 各专题节目的新闻文本内容，字段简洁；华尔街日报新闻文本数据包括标题、副标题、所属板块、类别、作者、发布时间等字段，时间区间为 2010 年-2022 年；CnOpenData 推出中国财经新闻报纸文本数据，覆盖国内多个财经新闻报纸的文本信息，包含站点名称、发文时间、板块名称、标题、作者、正文等字段，包括证券时报、证券日报、上海证券报、每日经济新闻、经济日报、大智慧、人民日报、华尔街见闻、东方财富网、腾讯财经、深圳商报、南方日报、都市快报、金融界、凤凰财经等；A 股上市公司新闻舆情数据是 CnOpenData 与通联数据合作的数据库，包括 A 股上市公司新闻情感表、A 股上市公司新闻关联表和 A 股公司新闻指数统计表。

此外，其还有中国各省份官方报纸数据，CnOpenData 将四川、黑龙江、青海、广西、湖北、福建、上海、吉林、辽宁、内蒙古、贵州等各地区官方报纸中的报道详细信息进行了汇总整理，最终形成中国各省份官网报纸数据；中国各地区政府工作报告文本数据，覆盖全国 29 个省份及对应的 304 个地级市，时间跨度近 20 年，包含了国务院、省级政府、市级政府在内的三级政府部门的工作报告文本文件；谣言数据，收集整理了较真查证平台上的辟谣信息，包含谣言内容和时间、辟谣结论和时间，以及辟谣人员及其单位，为谣言相关研究提供了数据资源。

《CSMAR 国泰安》

财经数据库收集了 1993 年以来，宏观经济、股票、债券、基金等经济社会及金融市场各方面的新闻披露情况，包括新闻的基本信息、新闻行业信息等内容。数据总记录数 8920920 条（20023-12-28），日度，数据开始时间为 1986 年 04 月 12 日。对于新闻进行了细致的分类：新闻、研究报告、市场评论、法律法规、财经证券知识、投资者关系互动平台等。新闻分类中主要分类有时事闻，金融新综合消息属资讯股票市场咨询等。

事件研究，包括公司事件、灾害突发事件、市场事件等。灾害突发事件：收集了国家地震局、气象局、国家安全监督管理总局、民政部等发布的发生在中国的各类突发事件，如地震、台风、安全生产事故、社会安全事件等资料。

其中地震事件表发震时间、经度纬度、地区代码等，总记录数有 11525 条，数据频率为日度，数据开始时间为 1949 年；台风事件表本表台风序号、登陆地、登陆时等级、城市代码等，数据总记录数为 853 条，数据开始时间为 1949 年；自然灾害事件影响情况及损失表(2014-2020)，内容包括事件分类、影响地区、直接经济损失、房屋损毁、伤亡人数、受灾人口等，数据时间段为 2014 年到 2020 年；安全生产事故事件表，包括发生日期、地点、事件级别、伤亡人数、影响公司等，数据总记录数为 1723 条，

数据频率为日度，数据开始时间是 1993-12-15；社会安全事件表(1970-2020)，包括发生地、犯罪者、死亡人数、目标类型、区域、攻击类型等，数据总记录数为 207682 条，数据时间段为 2014 年到 2020 年。此外，还有法律、法规、规则信息表，包括名称、正文、文案分类、施行日等，数据总记录数为 3772 条，数据开始时间为 2001 年。

此外，其也提供了社交媒体的数据，中国社交媒体研究数据库通过对网络股票贴吧的股评文本进行抓取，然后利用人工智能模型进行判断，整理出各上市公司股评的情绪和观点态度，并对用户属性和发帖属性细分，按照粉丝量、评论数、净评论数及发帖时间、发帖终端等维度进行量化统计，为客户提供按上市公司、时间和发帖者特征分类为基础的多维度统计的量化舆情数据。投资者情绪统计表，数据开始时间为 2020-10-01，数据总记录数为 3938993 条。文本情绪统计表数据总记录数为 6131082 条,数据开始时间也为 2020 年 10 月 1 日。

社交媒体除了投资者情绪统计表和文本情绪表、其它还有投资者情绪筛选统计表（发帖时间）、投资者情绪筛选统计表(发帖终端)、投资者情绪筛选统计表(粉丝量)、投资者情绪筛选统计表(评论数)、投资者情绪筛选统计表（净评论)、投资者情绪筛选统计表（发帖时间及发帖终端）、投资者情绪筛选统计表（发帖时间及粉丝量等级）、投资者情绪筛选统计表（发帖时间及评论数）、投资者情绪筛选统计表（发帖时间及净评论）、文本情绪筛选统计表（发帖时间）、文本情绪筛选统计表（发帖终端）、文本情绪筛选统计表（粉丝量)、文本情绪筛选统计表（评论数)、文本情绪筛选统计表（净评论)、文本情绪筛选统计表（发帖时间及发帖终端)、文本情绪筛选统计表（发帖时间及粉丝量等级)、文本情绪筛选统计表(发帖时间及评论数)和文本情绪筛选统计表(发帖时间及净评论)。其中内容有股票代码、发帖日期、按照不同发帖终端分类筛选的发帖人影响力、发帖人吧龄、帖子数量、阅读量、点赞量、看涨帖子数量、中立帖子数量、看跌帖子数量、投资者情绪指数、情绪一致指数等统计指标。

（二）其它新闻库

再上一部分介绍了部分国内新闻舆情学术研究相关的数据库。在实际咨询服务以及投资等相关应用中，相关财经等类型的门户网站、第三方金融资讯等数据供应商业都自己生产相关新闻数据。

图12：选股宝板块关联新闻

资料来源：xuangubao.cn.中国银河证券研究院

对于大型搜索引擎网站，提供百度新闻，通过生产整理出各类新闻等，当然也包括财经新闻等，其通过相应的百度指数产品供使用。此外，大型门户网站，比如 sina 其除了有专门的新闻版块页面外，在财经版块，除了行情等相关资讯外，对于相关的财经新闻也进行 24 小时滚动播报。对于一些专门的财经媒体及终端，也有专门的市场快讯等，比如选股宝，对于新闻中所涉及的股票、行业、概念板块都会进行相应的标注。

选股宝在特定板块页面，对于其关联的新闻在右侧展示（如上图）对于此外，选股宝也会发布“原创热文”。

第三方财经资讯供应商也有相应的资讯提供。下面分别介绍常用的财经数据资讯供应商对于新闻舆情等信息的展示。

Choice

Choice 是东方财富旗下的金融数据平台。其资讯展示页面如下图说示。

图13：Choice 财经快讯板块

资料来源：Choice，中国银河证券研究院

Choice 将资讯分为：全部资讯、头条资讯、财经聚焦、国际资讯、公司资讯、choice 早班车、每日必读、早盘内参等。另外，还单独列出公众号资讯。

图14：Choice 央行动态新闻板块

央行动态

刷新 导出ex

序号	时间	标题	来源	作者
34	17:53	债市收盘 | 央行单日净投放1790亿元 国债期货收盘多数下跌 地产债多数上涨	财联社	-
35	17:53	央行副行长：完善境外银行卡受理环境 丰富移动支付产品供给	澎湃新闻	-
36	17:48	债市收盘 | 央行单日净投放1790亿元，国债期货收盘多数下跌，地产债多数上涨	C-Bond资讯	财联社 李响
37	17:30	央行：支付业务新分类方式有利于防范监管空白	澎湃新闻	-
38	17:21	央行副行长张青松：我国移动支付普及率达到86%，居全球第一	界面新闻	-
39	17:18	欧洲央行管委Holzmann：不保证2024年会降息	财联社	-
40	17:17	央行最新发声：适当提高支付机构注册资本要求	大河财立方	-
41	17:16	欧洲央行管委Holzmann：不保证2024年会降息	C-Bond资讯	-
42	17:06	央行：将抓紧制定《非银行支付机构监督管理条例》实施细则	中国证券报·中证网	-
43	17:02	央行：将抓紧制定《非银行支付机构监督管理条例》实施细则	澎湃新闻	-
44	17:02	央行就支付市场答澎湃：切实防范业务异化、资金挪用、数据泄漏等风险	澎湃新闻	-
45	16:53	央行副行长张青松：《非银行支付机构监督管理条例》出台 标志着支付行业发展进入崭新阶段	界面新闻	-
46	16:46	财联社12月28日电，据泰国央行，截至第三季度末，泰国家庭债务占GDP比例略微上升至90.9%。	财联社	-
47	16:44	央行副行长张青松：《非银行支付机构监督管理条例》出台 标志着支付行业发展进入崭新阶段	财联社	-
48	15:53	巴以爆发新一轮大规模冲突；美联储全年加息100个基点……2023年十大国际财经新闻 | NBD年度新闻榜	每日经济新闻	-
49	14:57	日本央行减少购债操作加剧政策转向猜测	C-Bond资讯	-
50	14:18	美联储和欧洲央行“背对背” 美元指数下跌逼近100	华夏时报	冉学东

资料来源：Choice，中国银河证券研究院

其中财经媒体包括：报刊头条、新闻联播精选、中国证券报、上海证券报、证券时报、证券日报、第一财经、21 世纪经济报道、每日经济新闻、金融时报、南方都市报、经济日报、经济参考报、澎湃新闻、财联社、南方财经网等，可以单独展示。

此外，还汇总了央行动态。包括新闻事件、标题、来源还有作者，需要注意的是这里对于同一个事件会有不同“来源”同时报告（如图序号 42 和 43）。这里还有其他国家央行新闻信息等。

对于政府机构的相关新闻 Choice 也进行了单独的分类，包括全部政府资讯、国务院、发改委、国资委、央行、财政部、工信部、商务部、农业农村部、住建部、国土部、交通部、人保部、生态环境部、水利部、科技部、文旅部、卫健委、市场监管总局、国税总局、外管局、统计局、邮政局、海关总署、广播电视总局、证监会、银保监局、上交所、深交所、中金所、上海期货交易所、大连商品交易所、郑州商品交易所、自然资源部等。

Choice 按照行业聚焦、热点概念、省市地区等也对新闻进行了分类展示。

iFinD

iFinD 是同花顺公司旗下的金融数据终端。其资讯展示页面如下图说示。

图15：iFinD 财经快讯板块

资料来源：iFinD，中国银河证券研究院

其新闻板块分为热点新闻、每日速递、舆情预警、金融市场、行业新闻、概念新闻、国内宏观（国务院、央行、财政部、商务部、外管局、证监会、发改委、工信部、住建部、海关总署、统计局等）、全球宏观、新闻媒体（上海证券报、21 世纪经济、证券日报、证券时报、每日经济、中国证券报等）、市场要闻等。国内宏观板块包括国务院、央行、财政部、商务部、外管局、证监会、发改委、工信部、住建部、海关总署、统计局等；新闻媒体中对上海证券报、21 世纪经济、证券日报、证券时报、每日经济、中国证券报等进行单独展示。

大智慧