传统数据与大数据的融合

刘 涛雄和徐晓飞(2015)¨。使用了PC端百度搜索指数 讨论了互联网搜索行为对宏观经济预测产生的影 响，得出非结构化数据有助于提高预测宏观经济的 准确性，但不能替代政府统计数据，且要使用合适的 预测模型，由此提出了“两步法”，即先使用政府统 计数据进行初步预测，再加入百度搜索指数。

本文选取的在线大数据不仅 包括了代表Pc端和移动端网民对碳价关心程度的百度整体搜索指数，还包括了反映网络媒体对碳价 关注程度的媒体指数，力求用更全面的数据信息预 测碳价。在借鉴“两步法”思想的同时，不限制新增变量的个数，旨在选出最优模型。

媒体数据

非结构化信息采用了百度整体搜索指数和媒体指数。百度搜索指数是以网民在百度的搜索量为数 据基础，以关键词为统计对象，计算出的每个关键词在百度网页搜索中搜索频次的加权和。百度媒体指数是以各大互联网媒体报道的新闻中，与关键词相关的、被百度新闻频道收录的数量，采用新闻标题包 含关键词的统计标准，与百度搜索指数无直接关系，两者均可在百度指数网站获得。虽然经过标准化处理，但本质上是非结构化的，属于非结构化数据¨

交易数据

大数据分析方法提取和挖掘的数据

在对大宗商品价格波动风险的影响因素分析中，除了供需关系、库存、利率和汇率等结构化数据可以通过市场数据得到外，通过市场调研报告得到的市场投机活动、市场联动数据，以及地缘政治与突发事件这样的非结构化数据是不易直接获取的，但是人为的监测不但无法做到迅速的反应，而大数据分析方法对于非结构化数据的监测和预测都有着良好的效果，因为大数据分析方法可以快速的识别文本信息中的关键信息，并进行信息的整理。因此，大数据分析方法很好地弥补了市场数据的空白，使得价格波动风险相关影响因素的数据来源更加全面而准确。

常见的互联网大数据获得方式通常有两种：

网络爬虫

一种是使用网络爬虫软件对开放的网络数据进行爬取。网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列,直到满足系统的一定停止条件。爬虫对于网络的页面的搜索方式包括广度优先搜索、深度优先搜索和最佳优先搜索等。网络爬虫已经成为了互联网搜索引擎的重要组成部分，目前商用的爬虫软件和专业的网络爬虫公司也非常容易获取。使用网络爬虫工具进行互联网数据收集的优势在于成本低、定制方式灵活、可控性好等，其劣势在于能够获取的数据一方面取决于爬虫策略的设计和软件本身性能，另一方面也受限于网络数据的开放程度。对于一些并没有在互联网上进行公开的数据，如电商平台的交易数据等，使用爬虫软件无法获得。

互联网公司

互联网数据另外一种重要的获得方式是通过和一些互联网公司合作而直接获取。这种获取方式的优势在于有可以获得非常全面的网络全量数据（严格意义上说，只有通过这种方式获得数据才具有大数据的全面特性，而网络爬虫获得的网络数据只有在积累到达一定规模之后才能称之为大数据）。使用互联网公司内部数据的缺陷在于数据使用的业务协调成本较高。互联网公司中的各个部门在服务定位上通常为内向型服务，即只关心公司内部制定的KPI考核体系所制定的工作内容（如点击率、安装率、转化率等），对于无法列入KPI的社会服务和政府部门所关心的国家战略需求关心较少，因此在得不到相应管理层强有力的协调干预的情况下，很难获得互联网公司的有效配合。但是，一旦获得了互联网公司实质性的合作支持，那么除了可以获得必要的数据接入之外，还极有可能获得低成本甚至是免费的技术服务支持。此外，由于互联网公司掌握了非常全面的客户数据，里面必然会涉及大量的用户隐私，因此互联网公司出于社会责任和自身安全的考虑，对于数据的对外共享也始终都是非常谨慎的（尽管在公司内部对于这些数据的使用是非常充分的），这也进一步增加了数据的协调难度。

目前，互联网公司的数据根据公司的类型大体上也可以分为三种类型，即搜索引擎数据、电商数据、以及社交媒体数据。

（1）搜索引擎数据。由于网络爬虫是搜索引擎收集数据的基本方式，因此搜索引擎数据本质上可以看作一个比较全面的网络爬虫数据集。百度是搜索引擎领域国内最具代表性的公司，针对大数据的使用，百度在百度研究院成立了大数据实验室（BDL），此外还有深度学习实验室和人工智能实验室两个相关的企业研究机构，这三个实验室并称百度研究院三大实验室。百度除了具有搜索引擎中收集到的海量互联网网页数据之外，还具有基于百度云平台、LBS平台的用户地理位置数据。搜索引擎数据在资源能源安全领域最重要的应用是可以密切追踪与目标品种价格等因素关系较为密切的关键词搜索信息，使用这些关键词的搜索量与出现量能够对一些相关指标进行趋势先行预测。目前，百度大数据实验室使用百度收集的全量预测数据开发了一系列预测应用。

经济指数预测：监控中小企业景气状态，预测国家宏观经济走势。

景点预测：提供全国5A风景区未来2日人流及舒适度预测，为短期旅游出行提供参考。

疾病预测：预测流行病的城市发病情况，分析医疗结构的受关注程度。

赛事预测：预测欧洲五大联赛和世界杯各场比赛的结果。

高考预测：预测大学报考的难度、专业报考难度、以及作文命题可能的关键词。

电影票房预测：预测电影票房和观影人数等。

（2）电商平台数据。电商数据包括了各个主要行业的在线交易数据和询盘数据，这两种数据对于相关领域的指数预测都有着非常大的帮助。此外，一些电商平台本身还具有支付平台的功能，这种支付平台收集到的数据可以对宏观经济趋势和个体经济状况做出较为准确的描述，也可以用作资源能源安全预测的输入变量。目前国内最具代表性的电商平台有阿里巴巴和京东两家，这两家公司在数据开放上都做了一定的工作。

阿里巴巴“天池”大数据研究平台

天池大数据研究平台的服务对象是学术界的科研机构。阿里巴巴集团希望和学术界展开更紧密的合作研究，更好的发掘大数据的价值。针对当前学术界面临的两个问题，一是缺少有价值的真实的商业数据；二是缺少具有强大功能的计算平台支持复杂的数据处理。阿里巴巴集团于2014年正式推出“天池”平台，“天池”平台基于阿里集团的海量数据离线处理服务ODPS，向学术界提供科研数据和开放数据处理服务。天池平台目前向用户开放的活动主要有三类：开放式数据研究、课题合作、竞赛活动。开放三类科研数据集，包括用户购买成交记录、商品购买评论记录、商品浏览日志记录等，所有数据均只能在“天池”平台中使用。

京东万象平台/宙斯平台

和天池的科研平台性质不同，京东万象平台则是一个第三方数据交易平台。万象平台本身不使用任何数据，而是主要帮助数据提供方、数据需求方进行数据对接，帮助企业解决数据孤岛的问题，从而提升企业运营效率。平台本身会对接多维度的丰富数据，保证数据的安全性与接入效率，是企业数据输出与流入的最佳渠道。与此同时，万象平台还提供企业与企业之间数据互联服务，解决客户内部资源多系统之间的整合问题。目前万象覆盖的数据包括个人和企业征信报告，黑名单数据，失信数据等。另外，京东的JOS（宙斯平台）也提供了数据开放功能。京东JOS数据开放服务，将商家关心的数据开放给ISV及商家。ISV和京东商家可以利用这些服务，搭建最适用于商家店铺的数据分析平台，通过数据分析，及时了解店铺销量上升及下降的原因，从而调配资源，实现店铺效益最大化。JOS现已开放经营数据、销售数据、流量数据，后面会继续开放广告数据、客户数据、营销数据、及店铺日志数据。

（3）社交媒体数据。社交媒体数据包括微博等开放社交媒体数据和微信、QQ等私密社交媒体数据。由于涉及个人隐私等问题，通常情况下私密社交媒体数据是完全不对外开放的，即便在互联网公司内部，有关用户私密社交内容（如聊天记录等）也是不允许被访问和使用的。对于开放社交媒体数据，社交媒体公司都会开放相应的数据获取API，用户可以使用这些API直接访问和抓取社交媒体数据。不过通常情况下这种通过API的数据获取方式是根据用户权限的不同而受到一定限制的，例如权限较低的用户在单位时间内只能进行极其有限的访问次数，每次只能抓取固定数量的数据等。社交媒体数据最为广泛的应用即对舆情的监控。通过对社交媒体数据的自然语言处理，可以获得目前网络环境当中社会舆论比较关注的话题以及对于某一些关键词人们在情感上正面或负面的反映。使用这些舆情数据可以把握一个地区群体的思维倾向与社会动态，以及掌握人们对于市场的预期判断和信心指数等。

除了通过和互联网公司合作获取网络数据之外，网络运营商也是获得网络数据的一条渠道。目前移动网络和宽带网络均由移动、联通和电信三家运营商控制，换言之上述介绍所提到的各类互联网数据都必须经过三大运营商的网络服务器的传输才能够完成用户和网络服务商之间的数据交换，因此三大运营商有能力获取互联网上的几乎全量数据的。目前，三大运营商内部也都成立了相关的大数据事业部，并且建立了相关的数据服务平台。例如，中国移动在各个省级分公司建立了数据收集与交换平台，能够将全省的网络数据汇集起来；中国联通则在亦庄建立了全国范围的数据收集与交换平台，实现了全国数据的交换共享；中国电信则成立了专门的大数据分公司，专门研究运营商数据的应用于变现。此外，工信部国家互联网应急响应中心也掌握了五大微博为主的社交媒体全量数据。同这些专业管理机构合作可以以相对较低的协调成本获得多家互联网公司的集成数据。

通过上述对互联网大数据获取渠道的介绍可以看出，通过网络爬虫+互联网相结合的方式，就可以从网络中快速获取地缘政治与突发事件信息，以及各大公司报告中的关键信息。不过正如本节最开始所提到的，目前互联网数据在内容上整体偏向用户端，这就使得其所收集到的信息比较偏向零售、消费、个人信贷等产业终端领域。对于大宗海外进出口、大型企业的产品库存等数据，依然需要重点依赖传统的统计数据或是其它渠道进行数据收集。因此，互联网大数据只能用作能源资源安全预警预测领域偏向终端消费和公共舆情领域的一个方面的数据输入。