国家物资储备局研究项目

**基于互联网大数据的资源能源  
安全风险因素分析方法**

**研究报告**

北京航空航天大学

国家物资储备局

2018年3月

[第一章：资源能源安全风险的相关因素综述 4](#_Toc9638)

[1.2资源能源安全因素的互联网大数据来源 4](#_Toc30819)

[1.2.1 搜索引擎大数据 4](#_Toc6127)

[1.2.2 社交媒体大数据 8](#_Toc9200)

[1.2.3 电商平台大数据 9](#_Toc27183)

[1.2.4 新闻、行业相关网页数据 12](#_Toc6686)

[1.2.5 企业管理大数据 12](#_Toc32254)

[1.2.6 物联网大数据 13](#_Toc17476)

[1.3互联网大数据与传统数据的配合 13](#_Toc29938)

[第二章：资源能源相关互联网大数据分析技术综述 14](#_Toc3282)

[2.1预测分析技术 14](#_Toc27176)

[2.1.1计量经济学方法 14](#_Toc15159)

[2.1.2 时间序列分析方法 15](#_Toc17671)

[2.1.3 人工智能和机器学习的数据挖掘方法 15](#_Toc18441)

[2.1.4 集成学习技术 19](#_Toc22364)

[2.2文本处理相关技术 21](#_Toc13133)

[2.2.1自然语言处理技术综述 21](#_Toc18674)

[2.2.2文本情感分析技术 22](#_Toc30219)

[2.2.3关键词选择技术 25](#_Toc9559)

[第三章 宏观经济指数的大数据监测预测 26](#_Toc12702)

[3.1统计预测价格指数 27](#_Toc17525)

[3.1.1 Billion Prices Project 27](#_Toc4700)

[3.1.2 “阿里指数”aSPI-core与aSPI 29](#_Toc19736)

[3.1.3网络搜索数据与CPI相关性 30](#_Toc18559)

[3.2预测失业率、就业情况 31](#_Toc18600)

[3.3统计预测消费指数 32](#_Toc20869)

[3.4预测经济增长 33](#_Toc14740)

[3.5预测GDP 35](#_Toc23902)

[第四章 行业发展相关因素的监测预测 36](#_Toc29259)

[4.1资源能源与相关产业产品的需求监测预测 36](#_Toc13847)

[4.2相关产业产品价格的监测预测 37](#_Toc6316)

[4.3相关产业产品销售情况的监测预测 39](#_Toc1518)

[4.4资源能源的需求预测 42](#_Toc25178)

[4.5资源能源的价格预测 44](#_Toc30446)

[第五章 金融相关因素的监测预测 48](#_Toc14428)

[5.1股票价格预测 48](#_Toc16517)

[5.1.1 搜索大数据在股票价格预测中的应用 48](#_Toc9442)

[5.1.2 社交媒体大数据在股票价格预测中的应用 50](#_Toc6835)

[5.1.3 专题论坛与新闻数据在股票价格预测中的应用 52](#_Toc17461)

[5.2 投资市场情绪监测 54](#_Toc3010)

[5.3 大宗商品期货相关因素 55](#_Toc31252)

[5.4 其它金融相关因素 57](#_Toc27812)

[第六章 基于互联网大数据的资源能源政策经济因素分析 58](#_Toc30477)

[6.1国家政治和经济稳定性评估技术分析 58](#_Toc8562)

[6.1.1国家政治和经济稳定性评估模型 58](#_Toc13719)

[6.1.2利用社交媒体大数据进行稳定性评估综述 60](#_Toc11928)

[6.1.3利用搜索引擎大数据的评估综述 61](#_Toc5365)

[6.2重要公投与政策预测技术分析 62](#_Toc22708)

[6.2.1研究现状综述 62](#_Toc10275)

[6.2.2利用搜索引擎大数据的预测技术路线 63](#_Toc11325)

[6.3重要选举预测技术分析 64](#_Toc16747)

[6.3.1研究现状综述 64](#_Toc29532)

[6.3.2利用社交媒体大数据的预测技术路线 65](#_Toc18252)

[第七章：资源能源安全风险因素关系网络构建 69](#_Toc5029)

[7.1数据驱动的关系网络构建方法 69](#_Toc1489)

[7.1.1传统统计学的相关分析方法 70](#_Toc23190)

[7.1.2面向大数据的相关分析方法 72](#_Toc1947)

[7.1.3数据挖掘的方法 73](#_Toc14081)

[7.1.4基于因素与数据源关系的网络构建方法 77](#_Toc7198)

[7.2知识驱动的关系网络构建方法 78](#_Toc3788)

[7.2.1基于知识图谱的关系网络构建方法 79](#_Toc7395)

[7.2.2知识图谱在资源能源安全因素关系网络中的应用 82](#_Toc7084)

[参考文献 84](#_Toc20948)

# 第一章：资源能源安全风险的相关因素综述

## 1.2资源能源安全因素的互联网大数据来源

由一期成果，与能源资源风险相关的大数据主要包括互联网大数据、物联网大数据以及产业大数据三个主要类型。互联网大数据具体的数据细分由如下表所示：

表1.1：互联网大数据来源分类

|  |  |
| --- | --- |
| 细分类 | 具体来源 |
| 搜索引擎数据 | 谷歌趋势、百度搜索指数 |
| 社交媒体数据 | Twitter(推特)、Facebook(脸谱)、Microblog(微博)、主题相关论坛（如股票论坛） |
| 电商平台数据 | 跨境电商数据、淘宝 |
| 新闻数据、行业相关网页数据 | 各大新闻媒体（如Thomson Readers新闻）  提供价格的网站、与产品行业相关网站、政府网站、相关项目的网站 |
| 企业管理数据 | 企业内部生产零售数据、所管理企业的交易数据、投资者数据 |
| 物联网数据 | 气象数据、手机通讯数据、地理位置数据、电表电力数据、建筑物等设施数据 |

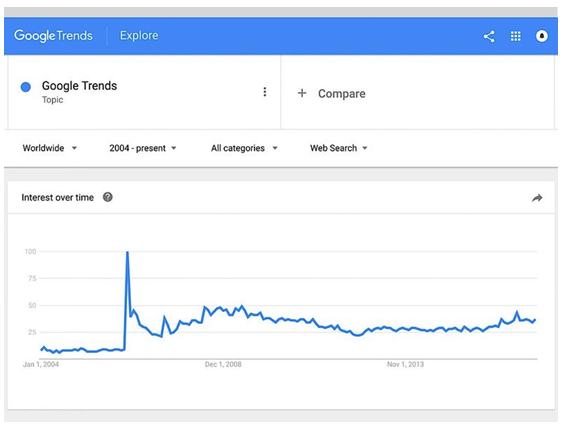
### 1.2.1 搜索引擎大数据

**1）Google Trends 谷歌趋势**

谷歌趋势 (Google Trends)是Google推出的一款基于搜索日志分析的应用产品，它通过分析Google全球数以十亿计的搜索结果，告诉用户某一搜索关键词各个时期下在Google被搜索的频率和相关统计数据。 [http://trends.google.com](http://trends.google.com/" \t "http://www.kguowai.com/html/_blank)

2006年7月31日,Google(谷歌)宣布Google Trends(谷歌趋势)中文版正式上线。Google Trends是2006年5月Google推出的一款基于搜索日志分析的应用产品。Google Trends通过图形化的方式最热门的关键词,同时还展现各种搜索关键词的自2004年开始的历史走势及与此相关的中文新闻报道。Google Trends通过分析Google全球数以十亿计的搜索结果,告诉用户某一搜索关键词在Google被搜索的频率和相关统计数据。在Google Trends中的每一关键词的趋势记录图形显示分为搜索量和新闻引用量两部分,用户可直观的分别看到每一关键词在Google全球的搜索量和相关新闻的引用情况的变化走势,并有详细的城市、国家/地区、语言柱状图显示。

谷歌趋势中的搜索量指数( Search Volume Index,SVI)体现了在一定区域内和一定时间段中针 对某关键词实际搜索数与平均搜索量之间的比例关系，即关键字在一段时间里的相对受欢迎程度。默认情况下，Trends会显示一个词在过去五年中全球范围内的相对受欢迎程度，但也可以调整时间段或选择特定国家/地区。



谷歌的首席经济学家 Hal Varian 认为搜索数据实时地刻画了大众对众多经济领域活动的兴趣，能够预测房屋、汽车和旅游业的销售( Choi＆Varian，2009) ，能够预测文化产品，如电影、视频游戏、歌曲的商业成功 (Goel et al，2010) 。

Google还发布了搜索分析工具[Google Insights](http://www.google.com/insights/search" \t "http://www.williamlong.info/archives/_blank)，主要是针对用户搜索的关键词进行分析，主要是把关键词的搜索量，按时间、地区和内容方面进行分析。在时间分析中，Google最自动加入有关此关键词的重大事件，以供用户，特别是商业用户查询。同时还提供关键词的自动分类功能，共包含605个细分类别。对于广告客户搜索关键字清单、商业发展趋势分析、寻找热门内容等都非常有用，可以使用这个工具帮助用户找到答案。

在推出[Google Insights](http://www.google.com/insights/search" \t "http://www.williamlong.info/archives/_blank)一年之后，开始提供这个工具的国际版，覆盖的语言包含简体中文在内。这个工具已经提供世界各地总共39种语言版本。此外，Google还推出了根据历史数据自动预测的功能，可以帮助用户预测这些变化和趋势，同时增加的还有动态地图功能，可以查看不同地区的搜索变化。

**2）百度搜索指数**

[百度指数](https://baike.baidu.com/item/%E7%99%BE%E5%BA%A6%E6%8C%87%E6%95%B0" \t "https://baike.baidu.com/item/_blank)（Baidu Index）是以百度海量[网民](https://baike.baidu.com/item/%E7%BD%91%E6%B0%91" \t "https://baike.baidu.com/item/_blank)行为数据为基础的数据分享平台，是当前互联网乃至整个数据时代最重要的统计分析平台之一，自发布之日便成为众多企业营销决策的重要依据。

百度搜索指数与谷歌趋势相类似, 用以反映关键词在过去 30 天内的网络曝光率及用户关注度。 它能形象地反映该关键词每天的变化趋势，每天更新一次,并且提供自 2006 年 6 月至今任意时间段的用户关注度数据(陈涛, 林杰. 2013)。 同时,根据不同的关键词,机器自动从百度新闻搜索中获取与该关键词最相关的 10 条热门新闻,并将新闻按时间顺序均匀分布在“用户关注度”的曲线图上,以字母标识,每个字母对应一条新闻。



百度指数是综合反映该关键词在过去1 天用户对它的关注和媒体对他的关注的一个参考值。任意关键词的百度指数都是该关键词在比较期的数值/ 该关键词在基期的数值。比较期的数值和基期的数值是通过当天的用户搜索量和百度新闻中过去 30 天相关的新闻数量相比得来的。 百度指数信息服务出现滞后于谷歌趋势。 但同样体现了某一关键词在特定时间段中被用户和媒体关注的强度。

细分功能如下：

1. 搜索指数

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，科学分析并计算出各个关键词在百度网页搜索中搜索频次的加权和。根据使用百度搜索来源的不同，搜索指数分为PC搜索指数和移动搜索指数。

2、资讯指数

资讯指数以百度智能分发和推荐内容数据为基础，将网民的阅读、评论、转发、点赞、不喜欢等行为的数量加权求和、指数化处理后得出，全面衡量网民对智能分发和推荐内容的被动关注程度。

2. 媒体指数

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，与搜索指数无直接关系。

3. 相关检索词

关键词A的相关检索词是网民搜索A时，同时还搜索过的其他关键词。

4. 上升最快相关检索词

上升最快相关检索词是在特定时间内搜索指数环比上升最快的相关检索词，并用上升箭头以及上升百分比表示相对上一时间上升的具体数值。

5. 需求图谱

需求分布图是针对特定关键词的相关检索词进行聚类分析而得的词云分布。

6. 人群画像

关键词的人群属性，是根据百度用户搜索数据，采用数据挖掘方法，对关键词的人群属性进行聚类分析，给出性别比例、年龄分布、兴趣分布等社会属性信息。

### 1.2.2 社交媒体大数据

在全球范围内，出现了众多较有影响力的互联网社交媒体平台。Twitter（[推特](https://baike.baidu.com/item/%E6%8E%A8%E7%89%B9" \t "https://baike.baidu.com/item/Twitter/_blank)）作为[美国](https://baike.baidu.com/item/%E7%BE%8E%E5%9B%BD/125486" \t "https://baike.baidu.com/item/Twitter/_blank)[社交网络](https://baike.baidu.com/item/%E7%A4%BE%E4%BA%A4%E7%BD%91%E7%BB%9C/420" \t "https://baike.baidu.com/item/Twitter/_blank)及[微博](https://baike.baidu.com/item/%E5%BE%AE%E5%8D%9A/79614" \t "https://baike.baidu.com/item/Twitter/_blank)客服务的网站，是全球[互联网](https://baike.baidu.com/item/%E4%BA%92%E8%81%94%E7%BD%91/199186" \t "https://baike.baidu.com/item/Twitter/_blank)上访问量最大的十个[网站](https://baike.baidu.com/item/%E7%BD%91%E7%AB%99/155722" \t "https://baike.baidu.com/item/Twitter/_blank)之一，是[微博客](https://baike.baidu.com/item/%E5%BE%AE%E5%8D%9A%E5%AE%A2" \t "https://baike.baidu.com/item/Twitter/_blank)的典型应用，它可以让用户更新不超过140个[字符](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6" \t "https://baike.baidu.com/item/Twitter/_blank)的被称作“[推文](https://baike.baidu.com/item/%E6%8E%A8%E6%96%87" \t "https://baike.baidu.com/item/Twitter/_blank)（Tweet）”的[消息](https://baike.baidu.com/item/%E6%B6%88%E6%81%AF/1619218" \t "https://baike.baidu.com/item/Twitter/_blank)。facebook、YouTube、 WordPress（博客）、Instagram（图片分享）、Google+（社交分享）等也有相当的用户量。国内的新浪微博同样是提供微型博客服务类的社交网站。用户可以通过网页、WAP页面、手机客户端、手机短信、彩信发布消息或上传图片。腾讯微博由[腾讯公司](https://baike.baidu.com/item/%E8%85%BE%E8%AE%AF%E5%85%AC%E5%8F%B8" \t "https://baike.baidu.com/item/_blank)推出，提供类[Twitter](https://baike.baidu.com/item/Twitter" \t "https://baike.baidu.com/item/_blank)的微型博客服务。

开放应用程序的API（即Application Program Interface，应用程序接口）可以让开发者在无需访问源码，或理解内部工作机制细节的情况下，调用他人共享的功能和资源。开发者可以通过认证，根据企业规提供的编程接口，限制性的获取自己需要的数据来开发自己的应用。

像国内新浪微博、腾讯微博、百度地图、高德地图等等很多应用都开发了自己的应用接口。开发者在注册完账户，申请自己的相关应用就可以获取平台数据。

TwitterAPI得到了众多研究者的使用，特别是在国外。其使用需要注册一个Twitter账号并获得相应的API key。可以指定获取一定范围内的微博，限制如时间、国家、语言、以及是否带有位置标签等。

微博开放平台开放了包括微博、评论、用户及关系在内的二十余类接口，通过Oauth2.0用户授权后即可在任意开发环境下使用。丰富齐全的功能，可以满足各种类型的产品需求。具体可分为[粉丝服务接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E7.B2.89.E4.B8.9D.E6.9C.8D.E5.8A.A1.E6.8E.A5.E5.8F.A3)、[微博接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E5.BE.AE.E5.8D.9A)、[评论接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E8.AF.84.E8.AE.BA)、[用户接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E7.94.A8.E6.88.B7)、[关系接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E5.85.B3.E7.B3.BB)、[搜索接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E6.90.9C.E7.B4.A2)、[短链接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E7.9F.AD.E9.93.BE)、[公共服务接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l ".E5.85.AC.E5.85.B1.E6.9C.8D.E5.8A.A1)、[OAuth 2.0授权接口](http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI" \l "OAuth2)。

主题相关论坛（如股票论坛）等数据可以通过网络爬虫的方式获取。

### 1.2.3 电商平台大数据

电子商务数据方面，国内领先的电商纷纷推出了自己的数据服务平台，提供数据以及分析产品等服务。

**1）阿里巴巴**

阿里巴巴“天池”大数据研究平台的服务对象是学术界的科研机构。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，基于阿里集团的海量数据离线处理服务ODPS。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。目前其开放的不少数据集已经应用于实际科研中，并为国际顶尖的人工智能会议承办数据竞赛如[IJCAI-2015](https://tianchi.aliyun.com/datalab/dataSet.htm?spm=5176.100073.888.13.491433d86eOZdg&id=1" \o "IJCAI-15 Competition" \t "https://tianchi.aliyun.com/datalab/_blank)等，具有一定的影响力。（Peng Jiang, Yadong Zhu,2015）.

除此之外，阿里巴巴还推出了阿里指数等数据分析服务。阿里指数是了解电子商务平台市场动向的数据分析平台，2012年11月26日正式上线。根据阿里巴巴网站每日运营的基本数据包括每天网站浏览量、每天浏览的人次、每天新增供求产品数、新增公司数和产品数这5项指标统计计算得出。是以阿里电商数据为核心，面向媒体、市场研究员以及社会大众提供的社会化大数据展示平台；提供地域、行业等角度指数化的数据分析，作为市场及行业研究的参考、社会热点的洞察工具。其提供的功能具体如下：

区域指数：通过行业指数，可以了解一个行业的现状，获悉它在特定地区的发展态势，发现热门商品，知晓行业下卖家及买家群体概况。

区域贸易往来：查看当前选择省份与全国各省间交易情况，交易热度及热门类目分布一目了然。

热门类目：不同省份的地域热门交易二级类目可在这里查看；包括东部的浙江、中部的安徽、南部的广东、西南的四川、有地方特色的海南，供用户了解典型地区特色。

行业指数：通过行业指数，您可以了解一个行业的现状，获悉它在特定地区的发展态势，发现热门商品，知晓行业下卖家及买家群体概况。

搜索词排行：即搜索次数的指数化指标，包括衣（女/男装）、食（零食/坚果/特产）、行（户外/ 登山/野营/旅行用品）、用（3C数码配件），供您了解衣食行用相关行业特色。

热门地区：洞察不同类目下，按照交易指数排序，最终提供基于购买热度及销售热度的地区排名；提供最近7天数据的查看。

据了解，目前在中国大陆地区已有1亿家中小企业，而最新的统计数据表明，利用电子商务的中国企业在6000万数量左右。阿里巴巴的用户覆盖了中国目前大部分中小企业，使得阿里指数可能会成为中国电子商务市场的一个关键的晴雨表。

2015年3月2日,阿里巴巴还推出了国内首个面向政府开放的大数据产品——阿里经济云图。各地政府可自助查询当地多维度的电子商务经济数据,为决策服务。功能包括地方经济总览、全景分析、数据监测以及知识服务等功能,数据覆盖全国34个省级行政区、300多个地级市、2 000多个县级行政单位,数据可以细化到区县一级。地方政府可通过阿里经济云图,挖掘出区域内的优势产业、热门商品、潜力企业、消费主力,和相邻地区进行对比,在对比中找到自身的优劣势,成为决策依据。

全球速卖通（AliExpress）是阿里巴巴帮助中小企业接触终端批发零售商，小批量多批次快速销售，拓展利润空间而全力打造的融合订单、支付、物流于一体的外贸在线交易平台，正式上线于2010年4月。全球速卖通面向海外买家，通过[支付宝](https://baike.baidu.com/item/%E6%94%AF%E4%BB%98%E5%AE%9D" \t "https://baike.baidu.com/item/%E5%85%A8%E7%90%83%E9%80%9F%E5%8D%96%E9%80%9A/_blank)国际账户进行担保交易，并使用国际快递发货。截止2013年3月全球速卖通已经覆盖220多个国家和地区的买家；覆盖服装服饰、3C、家居、饰品等等共30个一级行业类目；海外买家流量超过5000万/日；交易额年增长速度持续超过400%；全球网站[alexa](https://baike.baidu.com/item/alexa" \t "https://baike.baidu.com/item/%E5%85%A8%E7%90%83%E9%80%9F%E5%8D%96%E9%80%9A/_blank)排名131，并在快速提升中。该平台积累的进出口交易数据对于资源能源的安全风险分析也有一定作用。

**2）京东**

京东万象大数据开放平台（以下简称“万象平台”）（wxlink.jcloud.com）是京东云在已有的云计算平台基础上围绕数据提供方、数据需求方、数据服务方等多方，构建了以数据开放、数据共享、数据分析为核心的综合性数据开放平台。其服务宗旨是帮助数据的提供方与需求方进行数据对接，解决数据缺失问题，完善数据价值，帮助企业解决数据孤岛的问题，从而提升企业运营效率。目前拥有的数据类型主要包括金融、征信、电商、质检、海关、运营商等领域权威、全面数据，未来会引入各类政府开放的数据，使供需生态更加丰富。

与此同时，京东也推出了与阿里巴巴类似的消费、价格指数。京东金融大数据消费指数于2016年5月31日正式发布，指数涵盖12个行业，直观反映了各行业网上销售及价格趋势。京东金融大数据消费指数包含：家用电器、计算机、手机数码、食品饮料、酒类、服装鞋包、家居家装、珠宝首饰、母婴保健、办公设备、运动户外及个护化妆等12个大消费类行业的价格指数和销售指数。据介绍，其指数主要能够应用于证券投资行业，为证券分析和策略生产提供高质量、独家的信息，提高基本面投资的有效性。

**3）大宗商品交易平台**

除了衣食行用等日常消费品，证券、期货、大宗商品等金融交易也已高度网络化。在证券交易行业，券商对于大数据的重要性已经有了高度共识，特别是在市场行情预测等方面，起步非常早。在期货市场以及大宗商品等场外交易市场，对交易数据的综合分析也得到越来越多企业的重视。我国的大宗商品场外交易市场经过十多年的发展，已形成一个新兴行业。在一些重点领域，如棉花、钢材等，大宗商品交易市场的交易量已经占到全国总交易量的一半以上。一些互联网公司同样聚集了该领域的大量 数据，如金网安泰公司为全国 370 多家大宗商品交易市场（约占到国内近 2/3 合规市场）提供了大宗商品交易平台软件，从而积累了大量原始交易数据。对这些 数据源的后续挖掘和分析，对于宏观经济监测预测也具有重要意义。

### 1.2.4 新闻、行业相关网页数据

新闻与行业相关的网页数据提供了所关注问题的最新资讯，从中可以提取出有价值的事实性数据，以及公众的关注度、情绪等信息。可以被应用于宏观经济指数如网络价格指数的建立，预测产品销量、投资者关注度以及股价等。

如MIT学者创建的Billion Prices Project(BPP)通过采集网络零售商品价格建立零售价格通货膨胀指数时就使用了全球各个地区的行业相关网页中网络价格数据。王雪蓉, 万年红.（2017）在对出口产品销量进行预测时不仅使用了跨境电商的数据，还加入了相关的互联网中各种形式的文档。Ding X, Zhang Y, Liu T, et al.（2015）在预测个股价格和股市波动率时使用了Reuters路透社以及全球商业、金融信息和财经资讯的领先提供商Bloomberg（彭博）的新闻数据。

新闻数据的获取可以采用网页爬虫，也可使用RSS等技术。张铁军（2017）使用RSS实时获取互联网财经新闻,每半小时更新一次。

### 1.2.5 企业管理大数据

企业管理大数据这里包括企业内部生产零售数据、所管理企业的交易数据、投资者数据等。其与资源能源安全风险相关的应用主要集中于两方面。首先在对产业发展进行监测预测时，如分析销量、库存时使用企业内部生产零售数据，李敏波,王海鹏,陈松奎,等（2017）在轮胎销售数据预测时采用了轮胎企业销售数据，并融合了多个不同领域的销售数据源。其次在分析金融市场的相关因素时，使用期货、大宗商品等交易所提供的交易数据、投资人数据，如岳艳涛,章雅婷,张宇等（2015）在分析的橡胶期货交易策略，预测期货价格时利用了期货交易所每秒钟提供两笔交易品种的实时数据(成交量、持仓量、总量、买一价、买一量、卖一价、卖一量)。

企业管理数据与电商数据的不同之处在于其私密性更高，且通常会涉及企业的利益，因此获取的难度更高，需要与相关企业进行深度合作。

### 1.2.6 物联网大数据

物联网就是物物相连的互联网。其核心和基础仍然是互联网，是在互联网基础上的延伸和扩展的网络。其延伸和扩展在于用户端可以是任何物品，即物品与物品之间，进行[信息交换](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF%E4%BA%A4%E6%8D%A2" \t "https://baike.baidu.com/item/_blank)和[通信](https://baike.baidu.com/item/%E9%80%9A%E4%BF%A1/300982" \t "https://baike.baidu.com/item/_blank)，物物相息。

1999 年，美国麻省理工学院 Auto-ID 研究中心的 Ashton 教授最早提出了 物联网的概念：将射频识别（RFID）技术和互联网结合起来，为每个产品建立 全球唯一的标识，即产品电子代码（EPC, Electric Product Code），采用射频识 别技术实现对所有产品的非接触式自动识别，然后通过互联网实现产品信息在 全球范围内的识别和管理。2005 年，国际电信联盟（ITU）发布了《ITU 互联网报告 2005：物联网》， 赋予物联网以新的内涵，报告指出，通过射频识别技术、传感器技术、纳米技术、智能嵌入技术等关键技术，利用互联网将世界上所有的物体都连接到一起， 使世界万物都可以上网。

物联网数据的特性在于，成千上万多种多样的设备产生海量异构、流模式以及地理位置分散的实时数据，且这些设备周期性地发送关于特定监控现象的观察结果，亦或是报告感兴趣的特定／异常事件的发生情况。

与资源能源安全风险相关的物联网大数据包括气象数据、手机通讯数据、地理位置GPS数据、电表电力数据、建筑物等设施的传感器数据等。

## 1.3互联网大数据与传统数据的配合

常见的方案包括使用互联网大数据进行建模，传统数据作为正确性验证。以及互联网大数据与传统数据直接结合用于模型，不加区分。

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。

王娜（2016） 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。

# 第二章：资源能源相关互联网大数据分析技术综述

## 2.1预测分析技术

### 2.1.1计量经济学方法

自回归模型（英语：Autoregressive model，简称AR模型），是统计上一种处理[时间序列](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97" \t "https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B/_blank)的方法，被广泛运用在经济学、信息学、自然现象的预测上。

AR

其公式意为，X的当期值等于一个或数个落后期的线性组合，加常数项，加随机误差。自回归方法的优点是所需资料不多，可用自身变数数列来进行预测。在预测中应用最多的计量经济学模型是多元线性回归、基于经济学系统的联系方程、探讨多个经济变量动态关系的向量自回归（VAR）和向量误差修正模型（VECM）。

自回归只能适用于预测与自身前期相关的经济现象，即受自身历史因素影响较大的经济现象，如矿的开采量，各种自然资源产量等；对于受社会因素影响较大的经济现象，不宜采用自回归，而应改采可纳入其他变数的[向量自回归模型](https://baike.baidu.com/item/%E5%90%91%E9%87%8F%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B" \t "https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B/_blank)。

向量[自回归模型](https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B" \t "https://baike.baidu.com/item/_blank)（Vector autoregression,VAR)是AR模型的推广。VAR模型把系统中每一个[内生变量](https://baike.baidu.com/item/%E5%86%85%E7%94%9F%E5%8F%98%E9%87%8F" \t "https://baike.baidu.com/item/_blank)作为系统中所有内生变量的滞后值的函数来构造模型，从而将单变量自回归模型推广到由多元[时间序列](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97" \t "https://baike.baidu.com/item/_blank)变量组成的“向量”自回归模型。VAR模型是处理多个相关经济指标与预测最容易操作的模型之一，并且在一定的条件下，多元MA和ARMA模型也可转化成VAR模型，因此近年来VAR模型受到越来越多的经济工作者的重视。

### 2.1.2 时间序列分析方法

时间序列分析方法中使用最多的是两类模型，一类是对序列自身变化情况进行刻画的自回归移动平均（ARMA）模型。这一类模型应用广泛，除了标准的ARMA之外，还有适用于非平稳序列ARIMA模型、以及考虑季节效应的季节自回归移动平均SARMA或者SARIMA模型。另一类是讨论序列波动情况的自回归条件异方差（ARCH）模型族。这类模型族中有很多种模型形式，如ARCH模型、GARCH模型、TARCH模型、EGARCH模型、非对称GARCH模型，等等。此外，还有一些预测模型融合了机制转移过程等，例如最常用的马尔科夫机制转换模型（Markov regime switch model）。

ARIMA模型全称为自[回归](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92/53132" \t "https://baike.baidu.com/item/_blank)积分滑动平均模型(Autoregressive Integrated Moving Average Model,简记ARIMA)，是由博克思(Box)和[詹金斯](https://baike.baidu.com/item/%E8%A9%B9%E9%87%91%E6%96%AF" \t "https://baike.baidu.com/item/_blank)(Jenkins)于70年代初提出一著名时间序列预测方法 ，所以又称为box-jenkins模型、博克思-詹金斯法。ARIMA模型的基本思想是：将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列。这个模型一旦被识别后就可以从时间序列的过去值及现在值来预测未来值。现代统计方法、[计量经济模型](https://baike.baidu.com/item/%E8%AE%A1%E9%87%8F%E7%BB%8F%E6%B5%8E%E6%A8%A1%E5%9E%8B" \t "https://baike.baidu.com/item/_blank)在某种程度上已经能够帮助企业对未来进行预测。

### 2.1.3 人工智能和机器学习的数据挖掘方法

**1）逻辑回归**

逻辑回归（logistic回归）是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此[因变量](https://baike.baidu.com/item/%E5%9B%A0%E5%8F%98%E9%87%8F" \t "https://baike.baidu.com/item/_blank)就为是否胃癌，值为“是”或“否”，自变量如年龄、性别、饮食习惯、[幽门螺杆菌](https://baike.baidu.com/item/%E5%B9%BD%E9%97%A8%E8%9E%BA%E6%9D%86%E8%8F%8C" \t "https://baike.baidu.com/item/_blank)感染等。自变量既可以是连续的，也可以是分类的。然后通过logistic回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。

逻辑回归使用 Sigmoid 函数，其形式为：



**2）支持向量机**

支持向量机是由Vapnik领导的AT&T Bell实验室研究小组在1995年提出的一种新的非常有潜力的分类技术，它在解决小样本、非线性及[高维](https://baike.baidu.com/item/%E9%AB%98%E7%BB%B4" \t "https://baike.baidu.com/item/_blank)模式识别中表现出许多特有的优势，并能够推广应用到[函数](https://baike.baidu.com/item/%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)[拟合](https://baike.baidu.com/item/%E6%8B%9F%E5%90%88" \t "https://baike.baidu.com/item/_blank)等其他[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0" \t "https://baike.baidu.com/item/_blank)问题中。

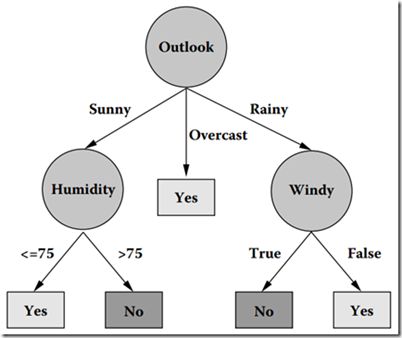
支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的[超平面](https://baike.baidu.com/item/%E8%B6%85%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/_blank)的两边建有两个[互相平行](https://baike.baidu.com/item/%E4%BA%92%E7%9B%B8%E5%B9%B3%E8%A1%8C" \t "https://baike.baidu.com/item/_blank)的超平面。建立方向合适的分隔[超平面](https://baike.baidu.com/item/%E8%B6%85%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/_blank)使两个与之平行的超平面间的距离最大化。其假定为，平行[超平面](https://baike.baidu.com/item/%E8%B6%85%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/_blank)间的距离或差距越大，分类器的总误差越小。

SVM的关键在于核函数。低[维空间](https://baike.baidu.com/item/%E7%BB%B4%E7%A9%BA%E9%97%B4" \t "https://baike.baidu.com/item/_blank)向量集通常难于划分，解决的方法是将它们映射到[高维空间](https://baike.baidu.com/item/%E9%AB%98%E7%BB%B4%E7%A9%BA%E9%97%B4" \t "https://baike.baidu.com/item/_blank)。但这个办法带来的困难就是计算[复杂度](https://baike.baidu.com/item/%E5%A4%8D%E6%9D%82%E5%BA%A6" \t "https://baike.baidu.com/item/_blank)的增加，而[核函数](https://baike.baidu.com/item/%E6%A0%B8%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)正好巧妙地解决了这个问题。也就是说，只要选用适当的[核函数](https://baike.baidu.com/item/%E6%A0%B8%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)，就可以得到[高维空间](https://baike.baidu.com/item/%E9%AB%98%E7%BB%B4%E7%A9%BA%E9%97%B4" \t "https://baike.baidu.com/item/_blank)的分类函数。在SVM理论中，采用不同的[核函数](https://baike.baidu.com/item/%E6%A0%B8%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/_blank)将导致不同的SVM算法。

**3）决策树与随机森林**

决策树是一种基本的分类和回归方法，分类决策树如 ID3、C4.5 算法等，回归决策树如 CART 算法等。在回归分析杠杆影响因素中，需要使用回 归决策树。CART 本身即可用于分类也可用于回归，它假设决策树是二叉树，对 回归树用平方误差最小化准则，对分类树用基尼指数（Gini index）最小化准则， 由生成和剪枝两步组成。在回归树生成过程中，采用启发式的方法选择变量和它 的取值，作为切分变量和切分点，并寻找最优的切分变量和切分点递归地生成最 小二乘树。 决策树是随机森林的基础，用于分类的决策树中，ID3 算法核心是在决策树 各个结点上使用信息增益选择特征，C4.5 算法对 ID3 算法进行了改进，使用信 息增益比来选择特征。

IMG_256



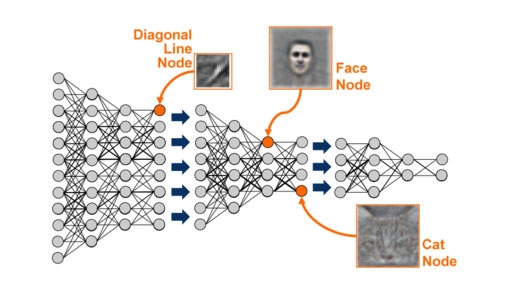
鉴于决策树容易产生过拟合，随机森林采用多个决策树投票机制进行改善。RF 生成首先从样本集中采用 Bootstrap 方式产生 N 个样本，然后选取 A 个特征中的 k 个构建决策树，重复 M 次产生 M 棵决策树后，采用多数投票机制进 行预测。 随机森林的主要思想是每一棵决策树是精通于某一个窄领域的专家（因为从A个特征中选择 k 让每一棵决策树进行学习），这样在随机森林中就有了很多个精通不同领域的专家，对一个新的问题（新的输入数据），可以用不同的角度去看待它，最终由各个专家，投票得到结果。随机森林的优点在于能够处理很高维度的数据并且不用做特征选择，在训练完成后它也能给出不同特征的重要性，并且容易理解，具有很好的抗噪声能力，可以进行并行处理。

**4）深度学习**

[深度学习](https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)的概念源于[人工神经网络](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)的研究。含多隐层的[多层感知器](https://baike.baidu.com/item/%E5%A4%9A%E5%B1%82%E6%84%9F%E7%9F%A5%E5%99%A8" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。

深度学习的概念由Hinton等人于2006年提出。基于深度置信网络(DBN)提出非监督贪心逐层训练算法，为解决深层结构相关的优化难题带来希望，随后提出多层自动编码器深层结构。

[深度学习](https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)是机器学习中一种基于对数据进行[表征](https://baike.baidu.com/item/%E8%A1%A8%E5%BE%81" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)学习的方法。观测值（例如一幅图像）可以使用多种方式来表示，如每个[像素](https://baike.baidu.com/item/%E5%83%8F%E7%B4%A0" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)强度值的向量，或者更抽象地表示成一系列边、特定形状的区域等。而使用某些特定的表示方法更容易从实例中学习任务（例如，[人脸识别](https://baike.baidu.com/item/%E4%BA%BA%E8%84%B8%E8%AF%86%E5%88%AB" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)或面部表情识别）。深度学习的好处是用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取[特征](https://baike.baidu.com/item/%E7%89%B9%E5%BE%81" \t "https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/_blank)。

IMG_256

深度学习在分类问题如图像、NLP 领域 取得了越来越多的成果，其激活函数通常采用 Sigmoid 函数、tanh 函数、ReLU 函数等，输出层常用 Sigmoid 或 Softmax 层，训练算法主要采用梯度下降法如 BP 算法[40]。常见的神经网络模型有深度置信网络（Deep Belief Network, DBN）、 深度玻尔兹曼机（Deep Boltzmann Machine, DBM）、含有卷积层和池化层的卷积神经网络（Convolutional Neural Network, CNN）、循环神经网络（Recurrent Neural Networks, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM） 等。深度学习应用领域跨度较大，如图像识别主流算法 CNN，语音识别和机器翻译，文本模型、时序数据处理以及推荐系统中都有广泛的应用。对极端亏损情况的分类尝试使用合适的网络结构进行分析预测。

深度学习源于人工神经网络的研究。人工神经 网络(artificial neural network, ANN)作为计算工具是由Li L, Leng S, Yang J, et al（2016）引入。之后，Hebb自组织学习规则、感知机模型、Hopfield神经网络、玻尔兹曼机、误差反向传播算法和径向基神经网络等也相继被提出。Xie H, Bian J, Wang M, et al（2014）利用逐层贪心算法初始化深度信念网络，开启了 深度学习的浪潮，指出深度学习的本质是一种通用 的特征学习方法，其核心思想在于提取低层特征， 组合形成更高层的抽象表示，以发现数据的分布规 律。Xie H, Bian J, Wang M, et al（2014）的方法有效地缓解了DNN层数增加所带 来的梯度消失或者梯度爆炸问题。随后Nayak R K, Mishra D, Rath A K, et al（2015）使用自动编码机取代深度信念网络的隐藏层，并通过实验证明了DNN的有效性。同时，研究发现人类信息处理机制需要从丰富的感官输入中提取复杂结构并 重新构建内部表示，使得人类语言系统和感知系统 都具有明显的层结构，这从仿生学的角度，为DNN多层网络结构的有效性提供了理论依据。 此外，深度学习的兴起还有赖于大数据和机器计算性能的提升。大数据是具有大量性、多样性、 低价值密度性的数据的统称，深度学习是处理大数 据常用的方法论，两者有紧密的联系。以声学建模为例，其通常面临的是十亿到千亿级别的训练样本， 实验发现训练后模型处于欠拟合状态，因此大数据需要深度学习。另外，随着图形处理器(graphics processing unit, GPU)的发展，有效且可扩展的分布式GPU集群的使用大大加速了深度模型的训练过程，极大地促进了深度学习在业界的使用。

### 2.1.4 集成学习技术

学习器的结合策略可以分为平均法、投票法以及学习法。

**1）平均法**

数值型数据最常结合策略是用平均法，包括简单平均、加权平均。加权平均法的权重一般是从训练数据中学习而得，现实任务中的训练样本通常不充分或存在噪声，这将使得学出的权重不完全可靠.尤其是对规模比较大的集成来说要学习的权重比较多，较容易导致过拟合.因此，实验和应用均显示出，加权平均法未必一起优于简单平均法。

**2）投票法**

投票法可以分为绝对多数投票法：若某标记得票过半数，则预测为该标记

相对多数投票法：预测为得票最多的标记，若同时有多个标记获最高票，则从中随机选取一个

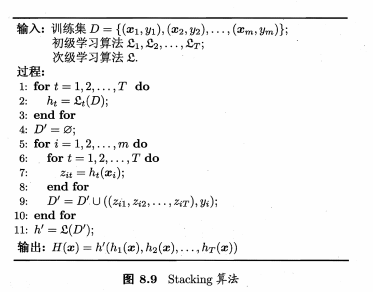
加权投票法：预测为得票最多的标记，初始的标记被赋予不同权值。

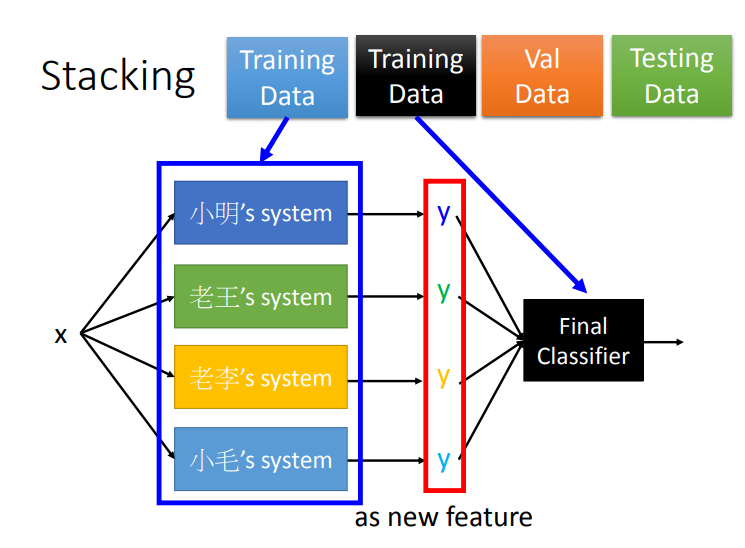
标准的绝对多数投票法提供了"拒绝预测"选项，这在可靠性要求 较高的学习任务中是一个很好的机制.但若学习任务要求必须提供预测结果，则绝对多数投票法将退化为相对多数投票法。因此，在不允许拒绝预测的任务中，绝对多数、相对多数投票法统称为"多数投票法"。

**3）学习法**

当训练数据很多时一种更为强大的结合策略是使用"学习法"，即通过另一个学习器来进行结合。把个体学习器称为初级学习器，用于结合的学习器称为次级学习器或元学习器。

Stacking算法是学习法的典型代表。Stacking 先从初始数据集训练出初级学习器，然后"生成"一个新数据集用于训练次级学习器.在这个新数据集中，初级学习器的输出被当作样例输入特征而初始样本的标记仍被当作样例标记.





次级学习器的输入属性表示和次级学习算法对 Stacking 集成的泛化性能 有很大影响.有研究表明，将初级学习器的输出类概率作为次级学习器的输属性，用多响应线性回归(Multi-response Linear Regression ，简称 MLR) 作为次级学习算法效果较好。

贝叶斯模型平均(Bayes Model A veraging，简称 BMA)基于后验概率来为 不同模型赋予权重可视为加权平均法的一种特殊实现. Stacking BMA 进行了比较.理论上来说若数据生成模型怡在当前考虑的模型中，且数据噪声很少，则 BMA 不差子Stacking; 然而，在现实应用中无法确保数据生成模型一定在当前考虑的模型中，甚至可能难以用当前考虑的模型来进行近似，因此， Stacking通常优于 BMA，因为其鲁棒性比 BMA 更好，而且BMA 对模型近似误差非常敏感.

## 2.2文本处理相关技术

### 2.2.1自然语言处理技术综述

自然语言处理是[计算机](https://baike.baidu.com/item/%E8%AE%A1%E7%AE%97%E6%9C%BA" \t "https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/_blank)科学领域与[人工智能](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD" \t "https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/_blank)领域中的一个重要方向。它研究能实现人与[计算机](https://baike.baidu.com/item/%E8%AE%A1%E7%AE%97%E6%9C%BA" \t "https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/_blank)之间用自然语言进行有效通信的各种理论和方法，综合了语言学、计算机科学、逻辑学、心理学、 人工智能等学科的知识与成果。其主要研究任务包括词性标注、机器翻译、命名实体识别、机器问答、情感分析、自动文摘、句法分析和共指消解等。

自然语言作为高度抽象的符号化系统，文本间的关系难以度量，相关研究高度依赖人工构建特征。而深度学习方法的优势恰恰在于其强大的判别能力和特 征自学习能力，非常适合自然语言高维数、无标签和大数据的特点。当前，如何应用深度学习技术解决自然语言处理(NLP)相关任务是深度学习的研究热点。 2013年，随着词向量word2ve的兴起， 各种词的分布式特征相关研究层出不穷。2014年开始，研究者使用不同的DNN模型，例如卷积网络，循环网络和递归网络，在包括词性标注、情感分析、句法分析等传统NLP应用上取得重大进展。2015年后，深度学习方法开始在机器翻译、机器问答、自动文摘、阅读理解等自然语言理解领域攻城略地， 逐渐成为NLP的主流工具。

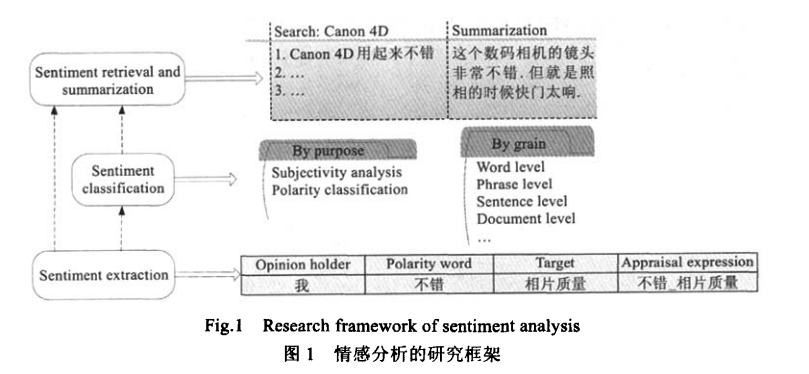
目前，深度学习用于NLP领域的主要步骤可以归结为如下3步：1) 将原始文本作为输入，自学习得到文本特征的分布表示。2) 将分布式向量特征作 为深度神经网络的输入。3) 针对不同的应用需求，使用不同的深度学习模型，有监督的训练网络权重。在未来几年，深度学习将持续在自然语言理解领域做出巨大影响。

### 2.2.2文本情感分析技术

文本情感分析又称意见挖掘,简单而言,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的 过程.最初的情感分析源自前人对带有情感色彩的词语的分析,如,“美好”是带有褒义色彩的词语,而“丑陋”是 带有贬义色彩的词语.随着互联网上大量的带有情感色彩的主观性文本的出现,研究者们逐渐从简单的情感词 语的分析研究过渡到更为复杂的情感句研究以及情感篇章的研究。（赵妍妍,秦兵,刘挺.2010）

按照处理文本的类别不同,可分为基于新闻评论的情感分析和基于产品评论的情感分析.其中,前者处理的文本主要是新闻评论,如情感句“他坚定 地认为台湾是中国不可分割的一部分”,表明了观点持有者“他”对于事件“台湾归属问题”的立场;后者处理的主 要是网络在线的产品评论文本,如“Polo 的外观很时尚”,表明了对评价对象“Polo 的外观”的评价“时尚”是褒义 的.基于产品评论的情感分析可以帮助用户了解某一产品在大众心目中的口碑,对于判断消费者喜好和商家营业情况、行业景气程度等有重要作用.而基于新闻评论的情感分析多用于舆情监控和信息预测中.

情感分析可归纳为 3 项层层递进 的研究任务,即情感信息的抽取、情感信息的分类以及情感信息的检索与归纳。



**1）情感信息抽取**

情感信息抽取是情感分析的最底层的任务,它旨在抽取情感评论文本中有意义的信息单元.其目的在于将无结构化的情感文本转化为计算机容易识别和处理的结构化文本,继而供情感分析上层的研究和应用服务.如将情感句“我觉得 Canon 的相片质量不错”转化为如图 1 所示的结构化文本形式.情感信息抽取的内容包括评价词语的抽取和判别、评价对象的抽取、观点持有者抽取以及组合评价单元的抽取，需要使用的技术包括属性抽取、规则/模板法、句法分析、序列标注等。

评价词语又称极性词、情感词,特指带有情感倾向性的词语。词语的情感极性判别主要有基于语料库和基于词典两种方法。基于语料库的方法主要是利用词语之间的连词以及统计特征来判别词语的情感极性。（魏韡, 向阳, 陈千.2011）由连词连接的词语的情感极性 存在某种关联，比如由连词“和”连接的词语的情感极性相 同，由连词“但是”连接的词语的情感极性相反。基于词典的方法是利用WordNet 或 HowNet提供的语义相 似度或者层次结构来判别词语的情感极性。

**2）情感信息分类**

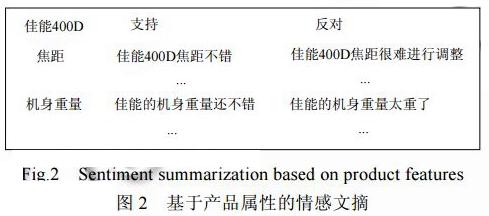
情感信息分类则利用底层情感 信息抽取的结果将情感文本单元分为若干类别,供用户查看,如分为褒、贬两类或者其他更细致的情感类别(如 喜、怒、哀、乐等).按照不同的分类目的,可分为主客观分析和褒贬分析;按照不同的分类粒度,可分为词语级、短语级、篇章级等多种情感分类任务。

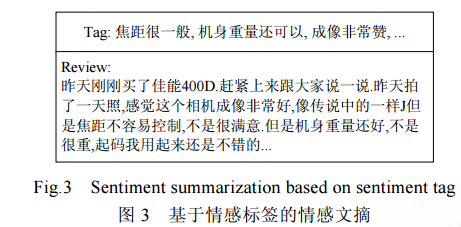
其中最常见的是主观信息的情感分类,包括最常见的褒贬二元分类以及更细致的多元分类，主要方法为基于情感知识的方法以及基于特征分类的方法。前者主要是依靠一些已有的情感词典或领域词典以及主观文本中带有情感极性的组合评价单元进行计算,来获取 主观文本的极性。后者主要是使用机器学习的方法,选取大量有意义的特征来完成分类任务。代表性的研究工作首先分析句子/篇章中的评价词语或组合评价单元的极性,然后进行 极性加权求和.这种方法的重点一般都放在评价词语或组合评价单元的抽取和极性判断方法的研究上。其次将机器学习的方法应用于篇章级的情感分类任务，研究重点在于有效特征的发现以及特征选择和特征融合等问题的研究。

如情绪评估工具Opinion Finder 和 GPOMS。Opinion Finder 根据文本内容可以评估“积极”与“消极”两种情绪；GPOMS 根据文本内容可以评估 “calm(冷静)”、“alert(警觉)”、“sure(确信)”、“vital(活泼)”、“kind(美好)”、“happy(高 兴)”六种情绪

**3）情感信息的检索与归纳**

最后，情感分析技术与用户的交互主要集中于情感信息检索和情感信息归纳两项任务上。情感信息检索要求检索回的文档同时满足两项准则:(1) 主题相关;(2) 具有情感倾向性。传统的基于事实性新闻语料的文摘旨.相比而言,情感文摘的处理对象为某一产品或某一事件的大量用户评论,在提取重要的事实性信息,并去除冗余信息的基础上融入了更多的情感信息。它主要侧重于提取具有明显情感倾向性的主观信息,是对某一产 品或某一事件的评论信息的归纳和汇总。针对产品类评论信息,情感文摘共有两种呈现方式:一种是基于产品属性的情感文摘,另一种是基于情感标签的情感文摘。





### 2.2.3关键词选择技术

在利用搜索大数据进行各项监测预测时，搜索数据关键词的选取是一个重要问题。由于面对的都是海量的搜索数据和关键词，真正具有预测价值的关键词却是需要甄别与筛选的。其主流的方法可以总结为三种。（彭赓, 苏亚军, 李娜，2012）

第一种是采取技术取词法。即利用高性能、大规模的计算设备将一切可能的关键词都纳入到研究范围内。然后计算与主题之间相关性选出核心关键词。例如 Ginsberg 等人利用 800 余台高速计算机在 2003 年~2008 年间 5 000 万个最为常用的搜索词中选择出 45 个与 CDC 发布的流感病人就诊量数据（研究课题）相关性最高的关键词作为预测关键词的来源。

第二种是经验取词法。即由作者运用主观经验确定关键词。例如Askitas N, Zimmermann K F（2009）在网络搜索与失业率相关性时，认为与劳动局或失业机关机构、失业率、人事顾问和德国比较流行的几个职业搜索引擎四类关键词的搜索量将出现变化。因而以这四类关键词为核心合成搜索指标。

第三种是范围取词法。 即先确定一个选词的范围，然后在范围内进行精选。最初的词范围可以是某些工具提供的相关类目关键词，也可以是各个类目中较为重要的关键词。精选的准则通常是度量与主题之间的相关程度。该方法可以看成是首先使用经验法选取词的范围，再用技术取词法选取最终的关键词，是两者的一种折中。例如 Konstantin 在研究网络搜索与美国个人消费增长率的相关性时，首先收集了 Google 提供 的 27 个分类中的前十大搜索词，然后分别做出相关性分析后剔除与个人消费不相关的词，利用剩下 220 个与消费相关的词合成一个指标 。Xu W, Li Z, Chen Q（2012）从 Google Trends 中与失业相关的分类中收集了 500 个左右的关键词作为原始关键词，从这 500个关键词中找出相关系数大于 0.65 的 108 个关键词。然后利用神经网络方法，从这 108个关键词中又筛选出少数几个关键词进行拟合。彭赓, 苏亚军, 李娜（2012）则是先选一个和失业相关的关键词作为初始关键词，由该关键词出发得到Google Trends 推荐的与其热门相关 10 个关键词。进一步搜索这 10 个关键词，得到Google Trends 推荐出其热门相关的大约 100 个关键词。如此往复，剔除重复的关键词后，这些关键词构成基础关键词库，约有403个关键词。

技术选词法精度较高，但是容易受到资源的限制而难以复制应用。直接取词法虽然降低了工作量但是主观性较强，降低了学术研究的科学性。范围取词法可以看为一种折衷。

# 第三章 宏观经济指数的大数据监测预测

国内外经济环境、经济增长情况对于资源能源安全具有显著影响。在经济增长中考虑全球经济增长以及国内情况。经济增长的衡量可以使用宏观经济指数，来反映跨产业的、整体的经济运行情况。宏观经济指数是体现经济情况的一种方式，主要指标包括国民生产总值、通货膨胀与紧缩、投资指标、消费、金融、财政指标等，对于宏观经济调控起着重要的分析和参考作用。由于宏观经济指数可以在全球和国家等多个层面进行统计，因此对全球经济增长和国内情况的监测本质上是一致的，都是对宏观经济指数的监测。

通常情况下，政府统计依赖于传统的调查方式来建立宏观经济指数,这些宏观数据往往滞后于政府宏观调控政策制定或企业经营决策需求。（苏志, 张骐）这种滞后对于资源能源安全监测的影响也是重大的，时效性是风险监测的一个核心要求。因此,使用实时更新的互联网大数据相对于传统数据而言更能满足对于即时统计信息的需求。使用互联网大数据进行宏观经济指数监测预测的工作已经包含如下宏观经济指数：价格指数、失业率、消费指数。在本节中会逐一介绍。

此外，如果对经济增长趋势建立合适的指标体系（申红艳, 吴晨生, 扆铁梅,等），确定影响总体经济增长的因素和其权重，也可以通过对子因素的监测加以组合，形成对经济增长趋势的直接描述。所使用的指标体系可以是目前使用较多的，采用人工专家的方式构建。

## 3.1统计预测价格指数

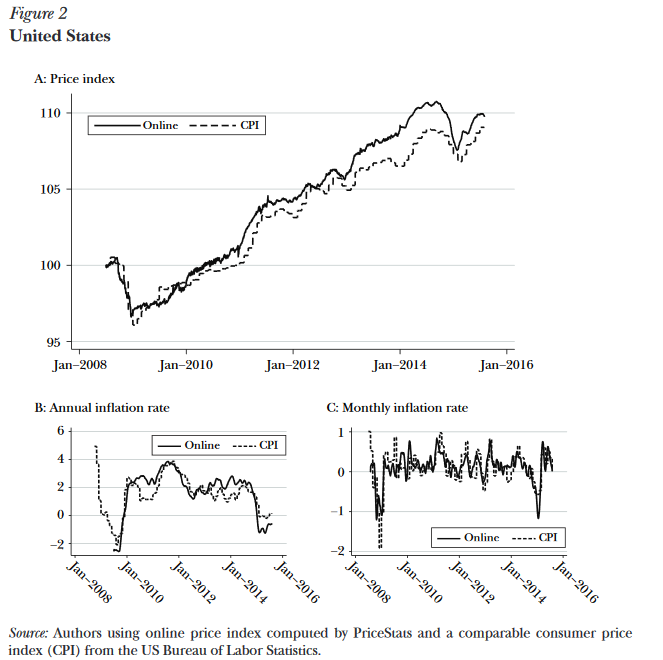
### 3.1.1 Billion Prices Project

MIT学者创建的Billion Prices Project(BPP)通过采集网络零售商品价格建立零售价格通货膨胀指数,为宏观价格和国际价格比较研究提供新数据来源（Cavallo A, Rigobon R,2016）。Billion Prices Project通过互联网中存在的价格信息构建多个国家的居民消费物价指数（CPI）。文中使用多种方法来避免统计上的偏差，如数据的滞后性和汇率的影响。传统的CPI测算方式采用工作人员亲自访问商铺的形式得到物价信息，并且只测算选定的一篮子商品的价格。采样的不频繁和一篮子中物品的选择更新缓慢都导致难以对统计质量进行把控，也不利于新的商品添加。互联网数据在这个统计问题中具有天然的优势，一方面由于价格数据散落在数以千万的网页、网站中，信息量巨大，特别是对于在互联网上出现频繁的商品，一方面由于互联网数据的搜集工具——爬虫技术的普及使得数据的收集迅速、高效、精准，同时对于一篮子中商品组合变化也可以进行快速的调整。截止到2010年，该项目平均每天从50余个国家的300个零售商处采集超过5百万条价格信息，远远超过通过传统方式采集到的数据。

文章使用的数据源为零售商的价格数据。网络零售商的选择采用线上+线下的零售商为主，尽量不选择只有线上销售的零售商以保证数据可以涵盖更多的地区。在商品类型的选择上，尽可能选择原有CPI计算时采用的一篮子商品，而避免过多使用在互联网上过度展现的商品如CD、书籍等。该项目没有采用第三方网站提供的价格信息，而是直接从零售商的线上商城处获得价格，虽然增加了获取信息的难度，但是使得数据的采集范围和质量都更加的可控。数据完成采集后进行清洗，标准化，将来自零售商的商品种类纳入标准的商品分类，并使用相关统计方法对数据进行初步分析。对零售商的数据进行一段时间的考察，分析结果不符合要求的将不予以采用，达到数据质量控制的目的。该方法目前在25个国家都采集到了CPI中所占权重达到70%的商品种类。

文章使用的分析方法为传统的CPI计算方法，包括对所统计商品消费者支出权重的分配等。由于与政府发布的CPI指数具有相同的计算方法，因此可以直接进行比较，其差异仅仅来源于数据源。

该项目目前在CPI计算、通货膨胀率计算上取得了与官方数据十分接近的效果。以美国为例，在2008-2016年间，网络价格计算的CPI、年度及月度通货膨胀率与官方数据始终有很好的贴合性，仅有少数数据有较小的差异，具有实用价值。同时，网络价格计算的CPI还对突发事件有更迅速的反应，当突发事件影响到商品价格时，网络价格CPI的变化要提前官方CPI的变化，在美国某事件的分析中可以提前两个月。在中国、巴西、南非、英国、德国、日本等多个国家的实验中也可以验证网络价格CPI与官方CPI的具有很高的相近度。



网络价格数据在CPI计算中一个主要的问题是采样的零售商数目不够全面，同时采样到的商品种类也不够全面，比如很多种服务的价格在网络上是缺失的，在这一点上要逊于传统方法。另一个问题是目前网络价格中缺乏消费者实际支出的相关信息，这些信息目前只能通过政府发布的官方数据，消费支出的调查，以及相关应用产品获得。

### 3.1.2 “阿里指数”aSPI-core与aSPI

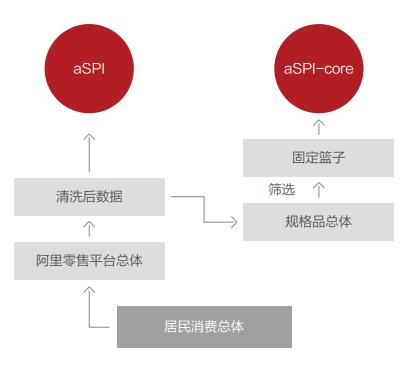
1. “阿里指数”是阿里巴巴构建的数据库。2012年11月26日，阿里指数正式上线，根据阿里巴巴网站每日运营的基本数据包括每天网站浏览量、每天浏览的人次、每天新增供求产品数、新增公司数和产品数这5项指标统计计算得出。其数据大部分是淘宝卖家来阿里巴巴采购留下的，它动态的反应了电子商务平台市场的发展情况，被许多跨国公司以及中国大部分中小企业看成中国电子商务市场的一个关键的晴雨表。（基于淘宝和阿里巴巴平台的电商业态分析）

**aSPI-core指数**

aSPI-core（alibaba Shopping Price Index-core，阿里巴巴网购核心商品价格指数是固定篮子价格指数，通过创新的筛选算法圈定阿里零售平台上近五百个基本分类下接近100000种核心商品作为固定“篮子”，每月追踪该篮子内商品和服务实际网购成交价格变化，以刻画网购主流商品和服务的一般价格波动，从而从网络零售渠道反映宏观物价走势。建立在严格可比的规格品成交价格基础上，数据误差较小，与通常应用的宏观物价指数在基本方法论上具有较高一致性。

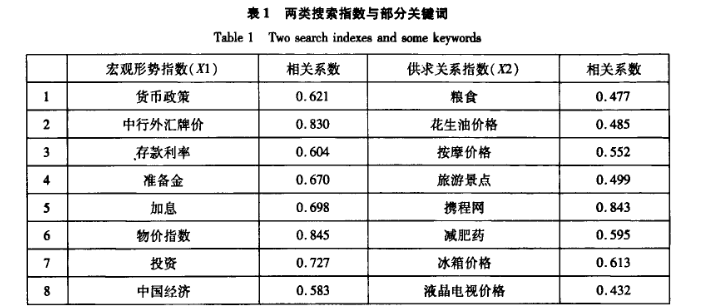
**aSPI指数**

阿里巴巴全网网购价格指数（aSPI）建立在叶子类目每月加权成交均价变动的基础上，以上月成交份额为权重所计算得到的价格指数，用以反映全网总体网购支出价格水平的变化。它同时包含了商品层面一般价格变动和消费者在叶子类目下消费结构变动的信息。 由于现代社会产品更新换代频率加快，许多产品的销售及价格具有生命周期特征，使得固定篮子指数aSPI-core在长期有可能会低估消费支出成本的上升趋势。以生活费用理论为基础的阿里巴巴全网网购价格指数（aSPI），在理论上能克服固定篮子指数面临的这一局限。该价格指数衡量的是消费者获得效用不变条件下，由产品相对价格变动引发消费数量上的替代效应，所带来的最小支出水平的变动。



### 3.1.3网络搜索数据与CPI相关性

张崇, 吕本富, 彭赓,等（2012）研究了Google搜索数据与CPI的相关性。数据选择上采用谷歌搜索解析，关键词搜索总量，2004-2009年的月度数据。表为部分搜索关键词。



分析方法采用回归模型，并使用平稳性检验、协整检验等。实验效果验证Google搜索数据与CPI的确有显著的相关性，模型拟合度达到0.978，预测绝对误差为0. 48。宏观形势搜索指数和供求关系搜索指数相对于CPI的先行周期分别为五个月和两个月；同时模型具有很强的时效性，比国家统计局的数据发布提前一个月左右。同时，搜索数据还对CPI及通货膨胀情况具有一定的预测作用。

## 3.2预测失业率、就业情况

失业率作为一个重要的宏观经济指标，一直受到研究者广泛关注。失业率的预测是网络搜索数据应用于社会经济研究的一个较早的课题。许多研究都表明谷歌趋势中与就业相关的词条查询或招聘查询指数可预测德国、以色列、土耳其、意大利、美国的失业率趋势。而且，部分模型的效果要优于基于专业预测人士调查的失业率预测模型。下表为部分代表性研究的总结。

|  |  |  |  |
| --- | --- | --- | --- |
| 相关研究 | 数据源 | 方法 | 效果 |
| Xu W, Li Z, Chen Q（2012） | 2004-2011谷歌搜索数据与美国失业率，500个失业相关词 | 神经网络 | 对现有模型的预测性能有提升 |
| Askitas N, Zimmermann K F（2009） | 搜索数据与德国失业率 | 传统经济学回归模型加入搜索数据 | 建立了某些关键词的搜关注度与失业率的关联 |
| Suhoy T（2009） | 搜索数据，以色列 | 传统经济学回归模型加入搜索数据 | 预测长期和短期的失业初请人数，绝对误差分别降低 15.74% 和 12.90% |
| Choi H, Varian H（2012） | Google 中有关“工作”和“福利和失业”类别搜索指数，美国的周失业率 | 传统回归模型加入搜索数据 | 模型的拟合度有较大的提高 |
| Francesco D（2009） | 失业相关的Google搜索指数、意大利的季度失业率 | 基于较小的样本建立工作搜索指数、将该指数加入传统预测模型 | 模型的预测效果显著高于传统模型，采样频低仍有较好效果 |
| 彭赓, 苏亚军, 李娜（2012） | Google 推荐的关键词搜索数据 | 改进的逐步回归方法，分层建立预测模型，因果关系检验及有效性检验 | 模型的拟合优度分别达到 0.930、0.935、0.936，三期预测值的 MAPE 分别为 1.20%、 0.89%、0.57%。 |

## 3.3统计预测消费指数

消费者信心指数（Consumer Confidence Index ，CCI）是反映消费者信心强弱的指标，是综合反映并量化消费者对当前经济形势评价和对经济前景、收入水平、收入预期以及消费心理状态的主观感受，是预测经济走势和消费趋向的一个先行指标，是监测经济周期变化不可缺少的依据。

国际上通行的消费指数编制做法，对消费者信心（或情绪）调查采用的是[问卷调查法](https://baike.baidu.com/item/%E9%97%AE%E5%8D%B7%E8%B0%83%E6%9F%A5%E6%B3%95" \t "https://baike.baidu.com/item/_blank)。问卷的设计紧密围绕以下几个方面内容：经济发展形势、家庭收入和就业、[物价水平](https://baike.baidu.com/item/%E7%89%A9%E4%BB%B7%E6%B0%B4%E5%B9%B3" \t "https://baike.baidu.com/item/_blank)、消费或[购买意愿](https://baike.baidu.com/item/%E8%B4%AD%E4%B9%B0%E6%84%8F%E6%84%BF" \t "https://baike.baidu.com/item/_blank)。每一方面由两类问题构成：对现状的看法和对未来的预期。

Vosen S, Schmidt T（2011）使用互联网大数据预测美国总体个人消费水平变化,无论使用样本内还是样本外的预测，其精度都比基于调查数据构建的两种消费者信心指数预测精度要高。其具体使用的数据为使用Google Insights获得的56种与消费相关的关键词与搜索量，关键词包括耐用消费品、非耐用消费品以及服务类的细分商品。分析方法为计量经济学中的VAR模型。



## 3.4预测经济增长

在宏观经济增长监测预测的模型建构方面，国内学者和研究人员进行了许多大 胆的创新和尝试，国内的研究成果就有国家信息中心的“中国宏观经济模型”、中国人民银行的“季度计量经济模型”以及厦门大学的“中国季度宏观经济模型”等。然而传统的宏观经济监测预测模型都是基于同频数据进行的，高频数据必须要降为低频数据。这样会造成数据信息的丢失，进而影响模型预测的准确性。而且国内现有的宏观经济监测预测模型以年度、季度模型为主，周期较长。因而有必要采用大数据的方法在时效性上进行改进。（陈龙, 王建冬等, 2016）

对于利用大数据进行经济增长预测，构建合适的经济增长判断模型，国内学者也提出了一些的看法。

国内学者的一个常用的研究框架是基于“克强指数”，该指数包含三个经 济指标，分别是“工业用电量新增”、“铁路货运量新增”和“银行中长期贷款新增”。“克强指数”源于时任辽宁省委书记的李克强总理会见美国驻华大使时表示，他喜欢通过耗电量、铁路货运量和贷款发放量三个指标来分析和预测辽宁省的经济运行情况。

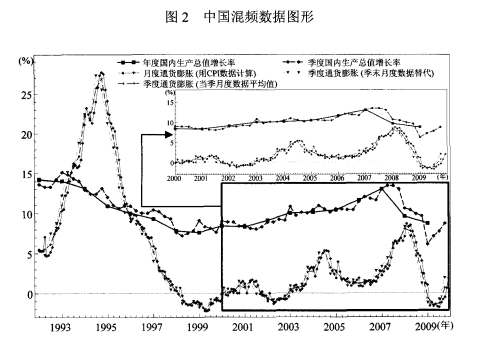
（刘慧，2014）通过构建“克强指数”与经济增长的 VAR 和 VEC 模型，发现“克强指数”的三大指标与经济增长之间 存在长期均衡关系和短期调整机制。（申红艳, 吴晨生, 扆铁梅等，2014）把国内利用大数据进行宏观经济分析中预测经济增长的研究分为三类，与“克强指数”选取的指标相类似。一是用电量与经济增长。大多数研究表明，用电量，尤其是工业用电量与经济增长之间存在长期稳定的均衡关系和因果关系。而且用电量与经济增长之间的这种关系也得到国外学者研究的佐证。二是货运量与经济增长。研究发现，货运量，尤其是铁 5 路货运量与经济增长之间存在交替推拉作用的因果关系。三是银行贷款与经济增长。（刘恩猛, 汪波，2007）发现经济增长和贷款之间存在协整关系和双向因果关系。

“克强指数”也得到了国外研究者和相关机构的认可。2010 年，英国著名政经杂志《经济学人》将“克强指数”视为评估中国 GDP 增长的重要指标。《经济学人》杂志根据这三项指标构造了一个指数，并画出这个指数的时间序列曲线，冠名为“克强指数”。“克强指数”被一些国际机构所认可，如花旗银行就用它来对比工业企业利润，并认为它的解释力更强。

除此之外，国内也完成了一些利用大数据技术监测预测宏观经济增长的相关课题，取得了一定成果。2013 年 7 月全国人大财经委员会向中央提交了《企业发展和宏观经济发展关系分析》报告，该报告根据“企业发展工商 指数”预测了我国宏观经济将企稳回升的趋势，对经济增长采用大样本数据和相关技术进行了预测，而后来的经济发展形势证实了这一指数预测的准确性。“企业发展工商指数”涉及 10 个对宏观经济具有显著先行性的指标，可以提前 1～2 个季度预测宏观经济发展形势。

## 3.5预测GDP

吉林大学的刘汉和刘金全验证了混频数据抽样模型（MIDAS）对中国季度 GDP 的监测和预测能力。混频数据模型（MIDAS）可以利用混频数据，避免高频数据降为低频数据时的信息流失，提高了宏观经济监测预测的准确性。研究发现，出口是造成金融危机阶段中国经济增长减速的主要成因。而且，混频数据模型在短期预测中国宏观经济方面具有比较优势，在实时预报方面具有显著的可行性和时效性（刘汉、刘金全，2011）。该方法的研究对于大数据在GDP的预测上的使用提供了一定基础。



Liu T, Xu X, Fan F.（2016）采用了一个两步的模型对我国的GDP预测进行了探讨，首先仅使用来自于政府部门的结构化统计数据构建模型，选取最优的模型，其次将互联网的搜索行为数据加入前一步的最优模型，再进行挑选即可得到最优模型。

# 

# 第四章 行业发展相关因素的监测预测

互联网大数据对资源能源上下游产业的发展情况也能起到很好的监测预测作用。在不同产业的产品需求、价格和销售情况，以及产业的发展景气水平监测预测上都已经积累了一定的研究内容。主要利用互联网大数据所反映的人们在现实世界里的需求、兴趣、和关注，来和产业相关的社会经济活动联系起来。

## 4.1资源能源与相关产业产品的需求监测预测

相关产业产品的需求预测可以结合实际的销售大数据系统，例如周成（2016）探索了基于大数据的卷烟销量预测系统的设计与实现。年度销量预测方面改进了原有的一元线性规划算法,用包含经济、社会、人口三大因素,GDP、社会消费品总额、城市人均收入、农村人均收入、市人口总数五项指标做线性多元线性规划。月度预测方面提出了基于灰度模型和综合时间序列模型的混合预测模型。预测准确度比用灰度模型或者综合时间序列模型明显提高。用JAVA语言实现了混合预测模型把基于Hadoop的卷烟销售预测系统与现有的正在运行的信息化系统做了集成,很好的与生产系统无缝对接,同时保证不影响现有的运行系统的业务流程。这种研究思路在使用电商大数据进行需求预测时也可以使用，因其具有相似的数据构成。

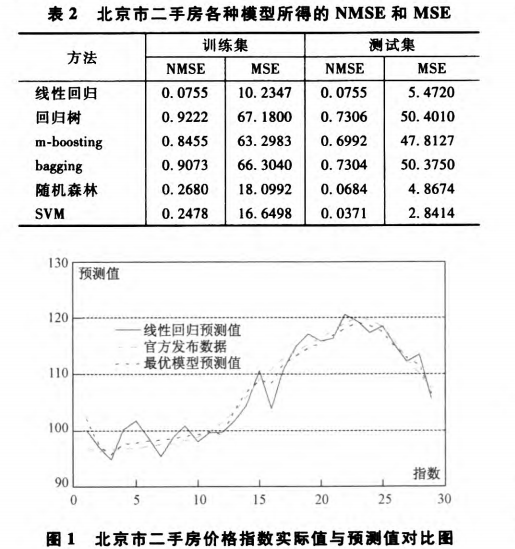
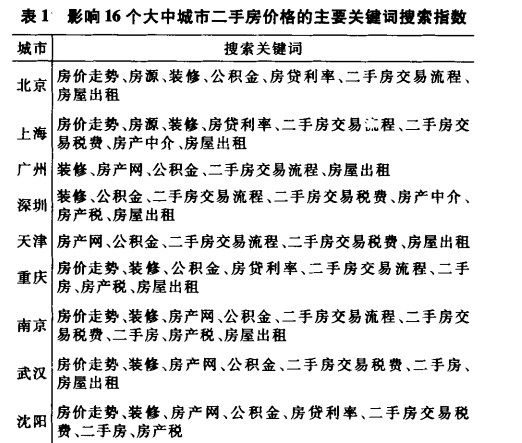
相关产业产品的需求预测也可以直接结合互联网搜索大数据，如王文亮（2012）讨论了现有的网络关键词搜索、电子商务交易量两方面相关的研究，给出了一种通过网络关键词对电子商务交易量的预测方法。（其研究的电子商务交易量与需求含义类似）首先，建立了理论分析框架，揭示了网络搜索数据与电子商务交易量之间存在的相关关系及先行滞后关系。其次通过逐步合成方法合成得到进入回归模型的关键词合成指数。并实证分析中国某电子商务网站的搜索数据和交易量与相应合成指数之间的关系。试验表明电子商务网站的搜索数据与交易量之间具有较高的相关性，加入合成的搜索指数后，模型拟合度达到0.901，模型的预测效果明显提高。通过对电子商务网站搜索数据进行过滤，应用逐步合成法进行预测交易量。实验结果证明模型具有较强的时效性，可以更为及时地预测电子商务交易量，进而帮助电子商务企业完善库存管理、订货计划等。

## 4.2相关产业产品价格的监测预测

目前在产品价格监测预测与大数据的融合上，房地产行业得到了充分的关注，不少学者检验了搜索大数据与房地产价格指数的关系，并使用搜索大数据预测中国各大城市的二手房价格和新房价格，取得了较好的拟合和预测效果，并具有时效性。尽管目前研究在房地产价格上较为集中，但是其分析方法对于其他行业的其他产品价格监测预测也有很好的借鉴和实用意义。房地产行业的发展本身对于资源能源的消耗情况也有很好的指示意义。

杨树新, 董纪昌, 李秀婷（2013）尝试构建网络搜索数据与房地产价格指数之间的关系框架,通过皮尔逊相关系数法,选取了与房屋销售价格指数相关系数在0.55以上的9个关键词,即房价、出租信息、二手房产网、二手房信息、开盘、户型装修、房产网、房源和装修材料；并通过实证检验,得出房屋销售价格指数与提前5个月搜索指数的相关性最大,即对房价感兴趣的购房者通常都会提前5个月左右搜索相关信息。为后续学者采用搜索数据进行房价预测提供了依据。

董倩（2014）以北京、上海、广州、南京、沈阳和西安 6 个大中 城市的二手房价格和新房价格为研究对象,以百度搜索指数为数据基础,首先选出了对二手房价格变动影响最大的 12 个关键词和对新房价格变动影响最大的 8 个关键词;然后采用交叉验证技术,运用线性回归、回归树、随机森林、Bagging、m-Boosting、神经网络、支持向量机和混合线性回归 8 种模型分别对 6 个城市的二手房价格和新房价格进行了拟合和预测，最终在 8 种模型中选择的最优模型成功地预测了 6 个城市的二手房和新房销售价格指数。结果显示:在二手房和新房价格的网络搜索关键词中,宏观经济形势和房产政策是关注的重点。可见,网络搜索数据不但能够较好地预测房价指数,同时能够分析得出经济主体行为的趋势与规律,而且具有一定的时效性, 预测的月度房地产价格指数能够比官方数据发布提前约两周。



王希晶（2016）以北京、上海、广州、深圳、南京、杭州6个城市为研究对象,探索了从关键词选择到复合指数合成再到模型构建及检验的一整套程序和方法,得到了6个城市二手房和新房网络搜索复合指数,并在此基础上利用1000次2折交叉验证法构建了基于网络搜索的二手房和新建商品房价格预测模型。研究结果显示,时差相关分析法筛选并用主成分分析法合成的网络搜索复合指数对房价指数具有较好的拟合和预测效果。分析了加入网络搜索指数后的预测模型对基本预测模型的改进作用有限的原因。

除此之外，也有研究利用数据挖掘算法来监测分析产品价格的异常波动，计算价格边界，实现异常情况的警报。

韩琳, 吴华瑞, 顾静秋（2017）使用K-means 技术解决根据已有海量的农产品市场价格数据来检索异常价格和计算价格边界的问题。采用的数据为农产品市场价格数据，具体包括副食品、瓜果、粮棉油糖、蔬菜和水产品等类别,全国各个省市地区历年的农产品市场价格数据, 例如茄子、生姜、芹菜、山药、大葱、 鸡蛋、牛肉、白鲢活鱼、粳米、菠菜、苹果、和菠萝等11914个品种的1300万条市场价格数据。采用数据挖掘聚类技术K-means技术进行分析, 改进传统 K-means算法并提出了基于K-means 的农产品市场价格异常数据检测方法。实验结果表明, 改进的算法可以很好地筛选出离群点数据, 找到异常价格元素从而获得农产品市场价格边界。目前算法已经应用于农业智能搜索引擎市场价格分析 模块, 效果显著。

## 4.3相关产业产品销售情况的监测预测

大数据在多种相关产业产品的销售情况监测预测上都被验证具有良好的作用。以下是部分代表性研究在不同产品领域上的总结。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 相关研究 | 研究内容 | 数据源 | 方法 | 效果 |
| Goel S,Hofman J M,Lahaie S,et al（2010） | 电影票房收入和游戏销售变化 | 2008-2009Yahoo!网页搜索日志与119部电影的票房数据（互联网电影数据库IMDB） | 选择与IMDB中电影有链接关系的搜索。检测搜索是否指向几个著名的游戏网站  改进的多元线性回归模型 | 搜索数据对销量变化有指示意义，但对现有模型的预测性能提升较小，因此更适用于其它数据来源缺失的情况 |
| Asur, S., & Huberman, B. (2010) | 电影卖座率 | 通过Twitter Search Api获取3个月的与某些电影相关的全部所有词汇 | 微博关键字为电影标题中改进的多元线性回归模型 | 拟合优度R2达到0.9以上，社交媒体的讨论情况对于电影销量有很好的指示作用 |
| Chevalier, J., & Mayzlin, D. (2006) | 亚马逊网站的和BN.com售书情况 | 亚马逊和BN.com的图书销售排名，图书评价，和上架时间、价格等基本信息 | 改进的多元线性回归模型 | 购书用户评价与图书销量排名之间的关系具有一定的相关性，在不同平台上关系不同 |
| Jacques Bughin(2012) | 实时监测预测比利时国家电信公司销量（网络和数字电视的） | Google 搜索数据与社交媒体数据的融合（Twitter, Facebook和其它博客数据） | 加入了短期预测变量的误差修正机制模型(Error Correction Mechanism，ECM) | 销量的 15%可以通过网络社交媒体数据解释，25%可以通过网络搜索数据解释；加入网络社交媒体和网络搜索数据以后，模型的整体预测能力提高了 25%。 |
| Choi H, Varian H（2012） | 汽车销售 | 美国国家统计局公布的机动车及其零部件销售情况，Google 搜索与机动车有关的数据 | 自回归模型（Autoregressive models） | 加入搜索数据后模型在实时预报上有性能提升，样本外的相对提升达到21.5% |
| Barreira, N., Godinho, P., & Melo, P. (2013) | 实时监测预测汽车销量 | 四个国家的与汽车销量相关的Google 搜索数据 | 基本的自回归模型（Autoregressive models） | 部分情况下搜索数据可以帮助解释销量的方差，在实时预报上有优势 |
| 章旭（2017） | 预测汽车销量 | 历史同期销量、前期销量，用户在线评论 | 基于网络大数据和传统统计学时间序列分析的考虑品牌情感的汽车销量预测BOAR模型。通用的销量预测MISF模型,基于MARS变量选择过程和BP神经网络相结合的方式 | BOAR模型平均预测误差为5.93%,比自回归模型降低8.59个百分点,可以准确预测单一汽车品牌的销量。MISF模型的预测误差平均为4.04%,比BOAR模型进一步降低了1.49个百分点。 |
| 李敏波, 王海鹏, 陈松奎,等.（2017） | 轮胎销售数据预测 | 轮胎企业销售数据，多个不同领域的销售数据源 | LASSO(The Least Absolute Shrinkage and Selectionator Operator)方法的多任务学习方法 | 实验数据验证能够提升轮胎销售预测的准确率 |
| 崔东佳（2014） | 品牌汽车销量预测，实施监控 | 百度搜索引擎搜索与奇瑞、大众及宝马三个品牌汽车相关的关键词 | 综合赋权和错位逐步合成方法对搜索到的关键词进行合成，得出搜索指数。建立回归预测模型，进行协整分析和Granger因果检验。 | 相比传统的汽车销量预测方法，该方法具有很高的预测精度（处于低端市场，拟合度为70.0％，中端市场拟合度为95.2％，高端市场拟合度为97.7％，且比统计部门发布提前一个月左右。 |
| 刘晶, 和述群, 朱清香,等（2017） | 线上农产品销量预测 | 涉农电商销售数据 | 结合深度学习算法优势和涉农电商销售数据特点的皇冠模型(ICM)：建立因素评价指标，采用两层自编码网络提取样本特征，训练后用BP微调整个网络 | ICM的分类准确率高达88%,明显高于其他未将数据进行特征学习的浅层分类器,证明了ICM具有较好的增量自学习能力和层次认知能力。 |

销量预测目前已经涉及到的领域包括：文化产业（电影、游戏和唱片等）、电信业、汽车工业以及农业。其中与资源能源安全联系较为密切的是使用能源较多的汽车工业，其它产业虽然没有较为直接的联系，但是其分析方法可以应用于所感兴趣的领域，即资源能源的主要上下游产业。

销量预测主要使用的数据有电商数据，搜索大数据，社交媒体大数据，工业大数据，以及与宏观经济相关的数据等。方法为有较为经典的时间序列分析、灰色模型、也有较新的方法如机器学习、深度学习等。目前应用较多的是经典的统计学方法，同时也有一些使用机器学习等方法进行的尝试。总体来看，销量预测的相关研究都取得了不错的效果，证明了大数据对销量预测的性能有提升作用。

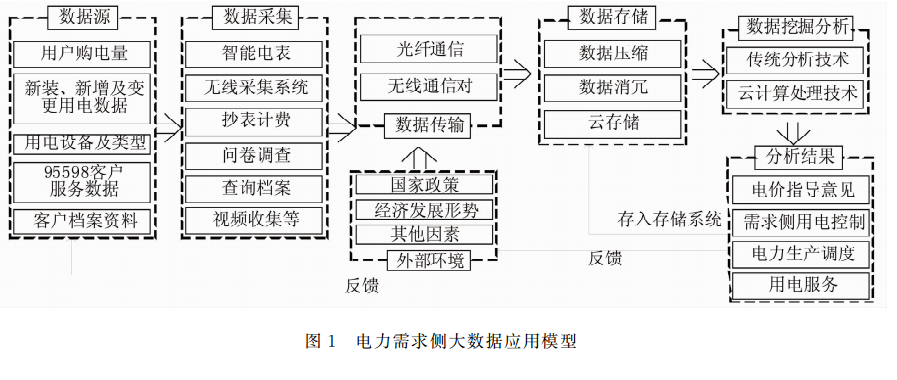
搜索大数据在销量预测中能够取得较好的作用，原因在于网络搜索数据记录了数以亿计的搜索关注与需求，反映了市场主体的行为趋势与规律，可以为研究社会经济行为提供数据基础。例如消费者在购买汽车过程中，首先会有需求准备期，在需求准备的这段时间内，消费者会根据自身的喜好、汽车的价格和性能、以及自身的经济状况来获取潜在购买相关汽车的信息。随着网络技术的快速发展和互联网的普及，消费者更加青睐与从互联网上搜索相关数据和信息，因此而留下的记录可供销量预测等研究使用。

社交媒体大数据则能捕捉到消费者对于产品的评价，考虑消费者情绪因素，也与产品的销量密切相关。电商数据、工业大数据是更为直接的数据来源，而宏观经济的相关数据则是考虑了当前的总体经济环境，在某些情况下有利于更精准地预测。

## 4.4资源能源的需求预测

资源能源的需求和价格往往区别于其他相关行业产品需求和价格，并且也是我们直接关心的对象。

目前资源能源需求与大数据的结合上，国内的一些研究集中在构建大数据在电力需求侧的应用模型，解决电力大数据在数据采集、传输、存储以及分析上的总体框架设计。这种将信息通信和能源电力结合起来，将互联网理念自底向上建立在能源基础设施上，逐步实现信息通信基础设备和能源电力基础设备的一体化的构想也被称为能源互联网。（李士琦. 2017.）如王健, 宋述停, 兰俊美,等.（2014）分析大数据的发展形势和最新研究成果，构建了大数据在电力需求侧的应用模型，以提升电力企业对电力客户的认知水平，以及电力企业的综合管理水平和经营效益。该模型以不低于6次/h 的频率进行数据采集，通过电网现有光纤信道和无线宽带进行数据传输，利用云存储、数据消冗技术、数据压缩技术进行数据存储，利用云计算处理技术与传统数据处理技术结合进行数据挖掘，从而得出电力需求侧分析结果，用于指导电力企业的生产、销售、服务，并通过反馈机制，对国家政策、外部环境等产生积极的影响，从而使该模型不断自我适应。图中为其应用模型，可见经过数据挖掘分析可以得到需求侧用电控制、电价等分析结果。



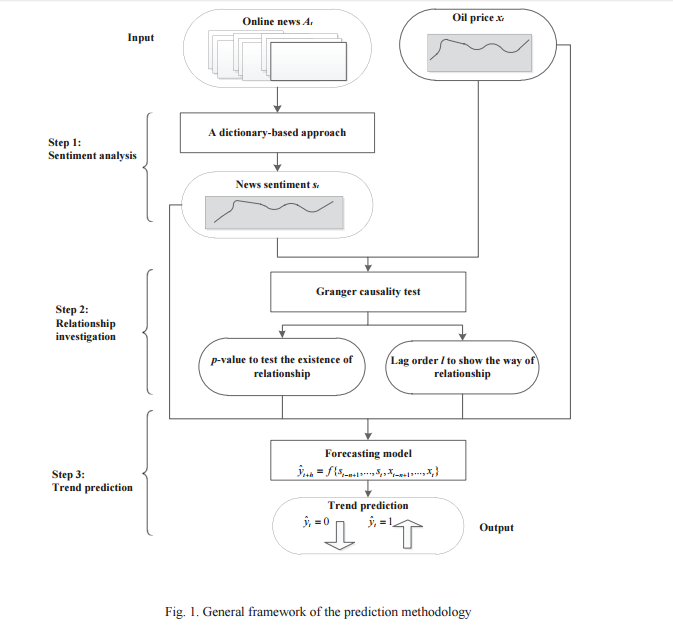
国外的研究也提供了类似的设计，并且进一步提出了基于map/reduce的线性回归等模型，以前期电力消费数据预测能源的消费与需求，测试效果显示预测与真实值十分接近，并且在速度上有优势。

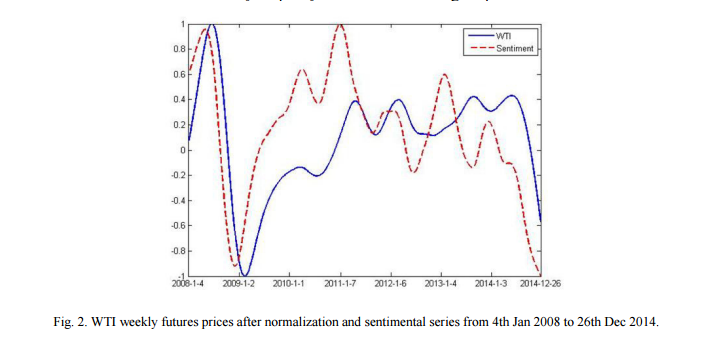
例如，Lee W, On B W, Lee I, et al（2014）设计了能够高效存储、索引、查询海量能源使用大数据的数据管理分析系统。其采用分布式框架并能够整合结构化的和非结构化的数据，且数据可以通过API快速获取。该系统还集成了基于map/reduce的使用正则项的线性回归模型，以预测能源的消费与需求。给与一个特定的时间点，模型能够预测出时间点的能源需求。模型的训练使用了2014年3个月份7百万条含有时间戳和能源消费量的数据，测试效果显示预测与真实值十分接近，而且分布式框架使得训练速度提高到1.72倍。

## 4.5资源能源的价格预测

资源能源的价格预测方面，不少研究进行了国际原油价格走势预测。主要采用新闻情感变化，或搜索数据分析投资者关注度，进而预测油价。模型方面同样是采用机器学习的方法以及计量经济学的经典方法。在油价的预测方面使用传统数据已经有非常多经验，而实验也表明加入了大数据后确实带来了预测效果的提升。

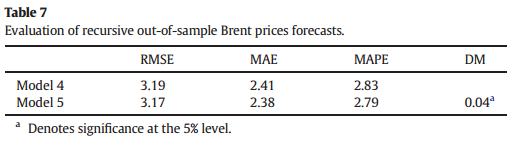
徐振敬（2016）提出了基于情感分析的国际原油价格走势预测模型。其模型主要基于石油相关新闻的分析,通过采用领域关键词词典的方法,得到新闻的情感序列,再通过格兰杰因果检验的方法,得到情感序列和石油价格序列的相关性和滞后期,最后通过机器学习的方法(支持向量机、决策树、逻辑回归和神经网络)预测石油价格的走势。同时,为了验证新闻情感对于石油价测具有预测能力,选取了美国西得克萨斯轻质原油(WTI)和路透社原油新闻作为研究对象进行案例分析。结果表明,新闻情感和原油价格之间确实存在着格兰杰因果关系,即新闻情感的变化会引起石油价格的变化,这说明新闻情感对于石油价格的走势具有预测能力。实验结果还表明,对于大部分预测模型而言,新闻情感的引入一般能极大地提高石油价格走势的预测准确率。

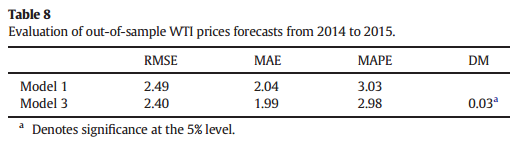




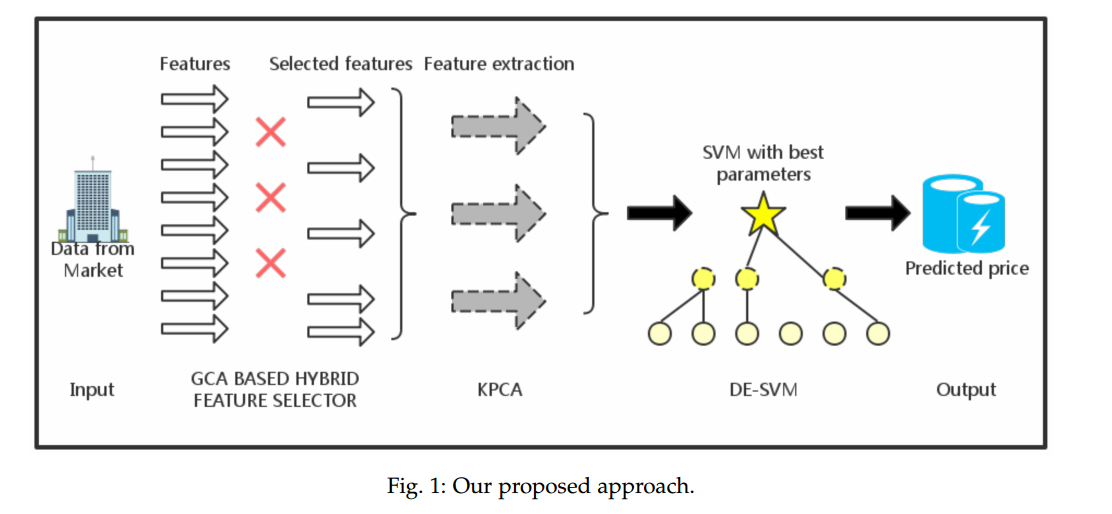
李新（2015）研究了时代下，基于海量互联网数据分析及建模的理论基础、研究框架及关键模型、技术和方法。首先，对提出了基于海量数据的投资者关注对原油价格的“非对称”影响关系的马尔科夫状态转移模型。采用网络挖掘-特征选择-广义动态因子模型(WM-FS-FHLR)提取互联网搜索关键词，进行特征选择筛选变量，并构建反映互联网关注的“一致指数”。建立了投资者关注与国际原油价格收益及波动关系的EGARCH模型，建立包含投资者关注与原油价格收益的VAR模型。

其次，通过集成时间序列模型、多元回归模型、和具有不同核函数的支持向量回归模型等方法研究如何在原有数据的基础上，通过引入海量数据构建的投资者关注来提高国际原油价格预测精度，通过样本外滚动预测等方法综合评价该方法的预测精度，并探讨了该框架在其他研究问题中的可扩展性等。其所提出的基于海量数据的预测模型为传统原油价格预测引入新的数据，并为更及时和准确的原油价格预测提供多种方法和技术。





在电力价格预测上，也有研究对以往的算法进行改进以更好利用电力大数据提高精确度。Wang K, Xu C, Zhang Y, et al（2017）使用电网中的海量电价数据进行电力价格预测，为解决数据中存在的冗余的特征问题，采用了三种模型相结合的方式。首先，将随机森林模型与Relief-F算法相结合，得到了一种基于灰色关联分析 （GCA）的去除冗余的混合特征筛选器。其次，使用基于核函数的主成分分析(KPCA)方法对特征进行进一步提取，降维。最后，提出了基于差分进化算法(differential evolution，DE)的支持向量机(SVM)分类器，对电价进行预测。实验证明其提出的方法比原有方案有更好的效果。



# 第五章 金融相关因素的监测预测

## 5.1股票价格预测

股市预测作为一直以来金融领域研究的热点，搜索、社交媒体、专题论坛等大数据都在股价预测上有所应用，同时也有许多经济学、统计学以及人工智能的相关算法被应用于其中。下面从股价预测的不同数据来源进行相关研究的梳理。

### 5.1.1 搜索大数据在股票价格预测中的应用

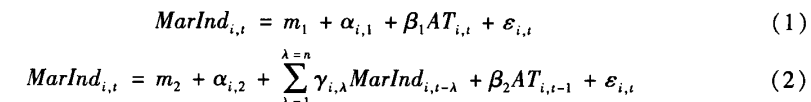
在数据来源上，搜索大数据在股票价格预测中得到了一定的使用。近年来，学术界对金融的投资者关注行为领域产生了浓厚的兴趣，因为股票价格的巨大变化似乎由投资者关注单独驱动。现有研究把投资者关注分为关注( Attention) 和 有限关注( Limited Attention) 。有限关注是指因为投资者的时间和精力有限，不可能考虑所有的股票，对影响股票基本面的信息反应不足，从而对股票市场的价格和成交量产生影响的情况关注有限。关注是指投资者因特定的引人关注的事件，对相应股票产生偏离其基本面的过度反应。对有限关注的研究已经有了很多理论与实证成果，相较而言，因为很难直接观察到投资者关注，所以尽管已经有理论较好地解释了投资者关注对资产价格的影响机制，但实证检验依然困难。

投资者大多通过网络进行股票交易( 特别是在中国①) ，网络用户通常使用搜索引擎来获得信 息。当用户使用搜索引 擎 来 搜 索 某 支 股 票 的 名 称 时，用户对某支股票的关注就直接地显示出来，因此，网络总搜索量是投资者关注直接和明确的衡量指标，搜索引擎中金融相关关键词搜索量可代表投资者搜索股市信息行为密集度。这是利用搜索大数据进行股价预测的一个理论依据。

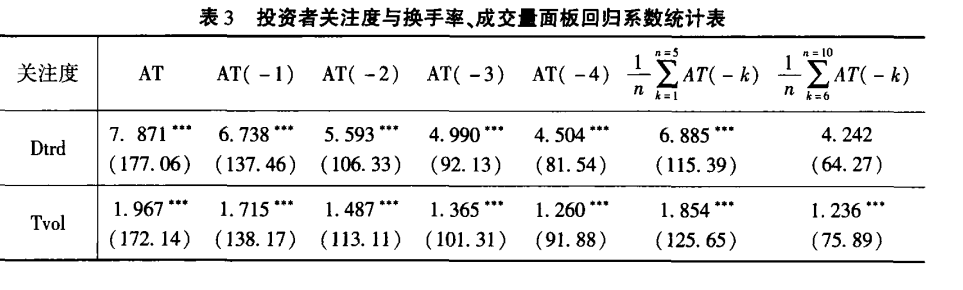
]宋双杰, 曹晖, 杨坤. （2011）用个股公开发行前谷歌搜索量衡量的投资者关注度对于个股市场热销程度、首日超额收益和长期表现有一定的解释力，可以解释首日超额收益的23% ，长期累积收益率的 10% 以上，结果均十分显著。每支股票每周的搜索量指数。也有研究用与谷歌趋势类似的百度指数构建投资者关注度,发现其对股票市场有正向价格影响。俞庆进, 张兵（2012）使用的数据为创业板196支股票的百度搜索指数。投资者关注度采用股票简称和股票代码的百度指数相加，取对数得到。



对市场指标和投资者关注度进行了单变量回归，使用模型（1）等验证了百度指数代表的投资者关注度的确与股票市场指标具有显著的相关性，投资者关注度越高，股票的市场表现就越活跃。使用模型（2）在普通的AR模型上增加投资关注指数的滞后项，验证了历史的关注度一定程度上能预测未来的表现，凸显了百度指数区别于市场指标的独特价值。



通过对投资者关注和股票的成交量、换手率的相关性分析，得出了滞后的投资者关注度对上述指标具有显著的正影响。进一步明确了投资者关注本身对股票市场表现有正影响，明确了是投资者先关注某只股票，进行搜索，然后在市场买卖股票，引起了市场指标的变化。而不是某只股票的市场表现突出，引发了投资者的搜索和关注。

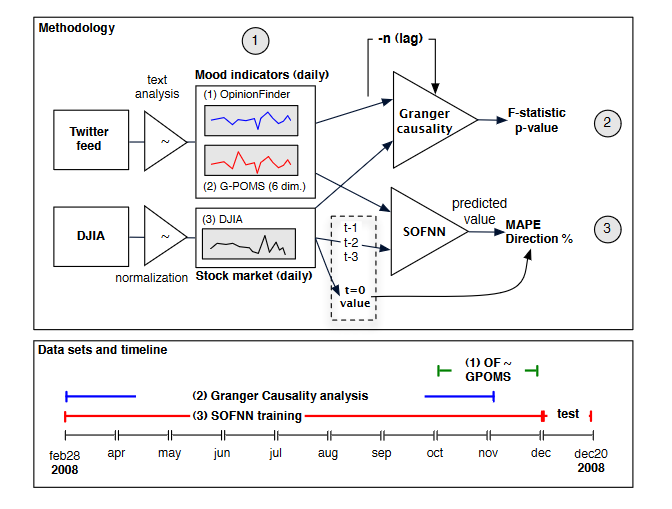


这种投资者关注指标的分析还可以解释更复杂的市场现象，如分析显示，关注度对市场造成正向的价格压力，且这种压力会很快反转，并且非交易日的股票关注与下一交易日集合竞价时的股票价格跳跃显著相关。

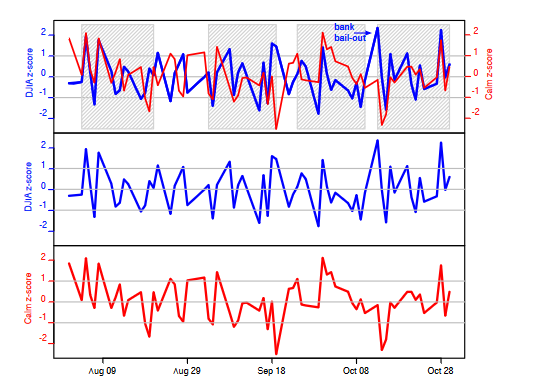
Preis T,Moat H S,Stanley H E.（2013）研究发现金融相关关键词的谷歌搜索量在道琼斯指数涨之前下跌的规律,据此规律建立炒股策略比传统炒股策略收益高。

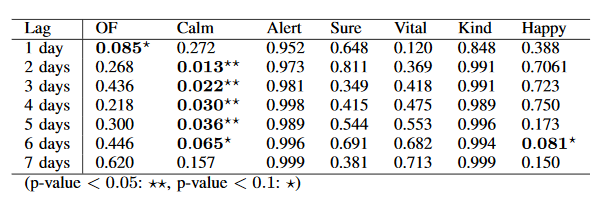
### 5.1.2 社交媒体大数据在股票价格预测中的应用

微博等社交媒体舆情对预测股价也有显著作用。近些年,以推特和微博为代表的网络社交平台上产生用户生成数据具有空前的规模性和群体性,其中的舆情与股票的市场交易活动也紧密关联。 例如 Bollen J, Mao H, Zeng X.（2010）发现基于 Twitter(推特)平台表达的公共情绪可以用来预测股市变动。这一研究的数据来源于两方面：一是 Yahoo·金融发布的道琼斯工业平均指数(DJIA)的收盘价；二是 2008 年 3 月到 10 月，270 万推特用户推送的 970 万条消息。这 970 万条消息经过情绪评估工具——Opinion Finder 和 GPOMS 被赋值。Opinion Finder 根据文本内容可以评估“积极”与“消极”两种情绪；GPOMS 根据文本内容可以评估 “calm(冷静)”、“alert(警觉)”、“sure(确信)”、“vital(活泼)”、“kind(美好)”、“happy(高 兴)”六种情绪，计算出每一天包含某种情绪的微博数量占全部微博的比重，就可以得到衡量公众情绪的时间序列，与股价的时间序列相对应。

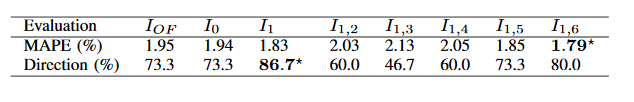


股价与情绪间相关性首先采用加入滞后影响的格兰杰因果检验进行分析，结果发现，在道琼斯工业平均指数（DJIA）和 GPOMS 中的“calm （冷静）”情绪之间存在较为显著的相关性，且情绪对于股价具有一定的预测能力。





进一步股价的预测采用融合了情绪信息的模糊神经网络 Fuzzy Neural Network (SOFNN)。实验发现， “calm（冷静）”情绪可以很好地预测道琼斯工业平均指数在未来 2 到 6 天的涨跌情况，而且这种每日预测的准确率高达到 87.6%。



研究还发现包含观点和情感的微博舆情模型比非舆情微博模型能更准确的预测股票价格变化（Oh C,Sheng O R L，2010）熊市预测比牛市预测更精确（Zhou S,Shi X,Sun Y,et al.2013） 在中国,“牛市冶”,“利空冶”等股市相关词汇在微博里出现的频率可以短期预测上证指数变化,而负面词汇有更强的预测能力（O ' connor B,Balasubramanyan R,Routledge B R, et al.2010）。

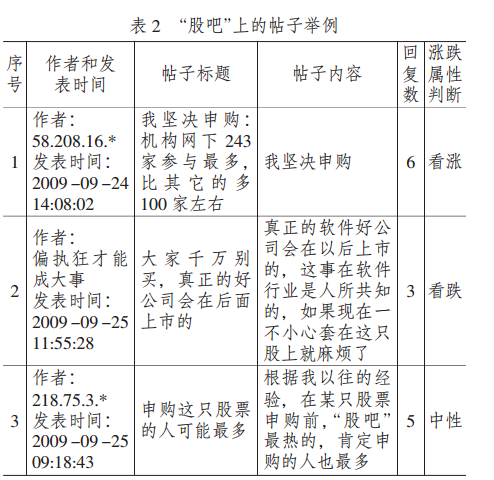
### 5.1.3 专题论坛与新闻数据在股票价格预测中的应用

股票专题论坛则是更加集中了投资者感兴趣的话题和讨论，比传统的社交媒体针对性更强。论坛里有关股票收益的讨论比专业分析师的预测更能体现收益预期,根据这种非正式信息进行操盘的收益也更高。

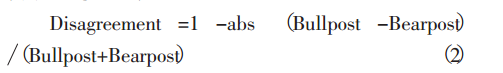
林振兴（2010）通过对投资者网络讨论信息进行整理分析，构建了反映投资者情绪的三个指标，即投资者关注度、投资者乐观情绪与投资者意见分歧，研究了投资者情绪与 IPO 抑价之间的关系。

公司首次公开发行的股票通过交易所上市之 后，二级市场的交易价格通常要高于股票的发行价 格，理论界和实务界将这一现象称为 IPO 抑价（IPO Underpricing）。大量的实证研究证实，IPO 抑价现是全球股市普遍存在的现象，其中，马来西亚和中国股市的 IPO 抑价幅度超过了 100%，尤其是中国股市的 IPO 抑价幅度更是高达 256.9%。

采用了86支创业板上市公司的股票价格，以及网络“股吧”的讨论。对目前我国用户访问量最大的互联网财经互动社区之一，我国用户黏性最高的互联网财经互动社区之一——东方财富网旗下频道“股吧”关于创业板个股的发帖讨论进行整理，汇总为投资者情绪指标。

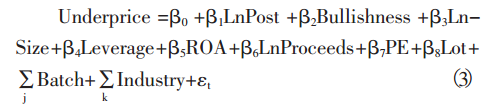


用样本“股吧”在招股 日和公开上市前一日之间的发帖数量来衡量投资者 对于个股的关注程度，人工判断投资者乐观、悲观情绪，并用以下公式来度量投资者的意见分歧（其中，Bullpost 是看涨的帖子数量，Bearpost 是 看跌的帖子数量）。

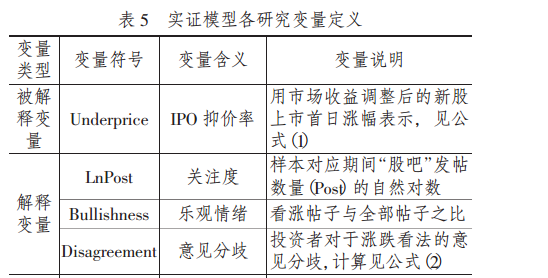


最终建立的实证模型同时控制了样本公司相关财务变量的影响，所有的财务数据均来自 Wind 数据库。包括公司规模（Size，总资产）、财务杠杆（Lever－ age，负债和总资产之比）和盈利能力（ROA，总资产 收益率）；以及新股发行的主要指标，包括募资 金额（Proceeds，实际募资金额）、市盈率（PE，发行价 格与发行后每股净利润之比）以及中签率（Lot，发行 股数和有效申购股数之比）。

建立模型如下，



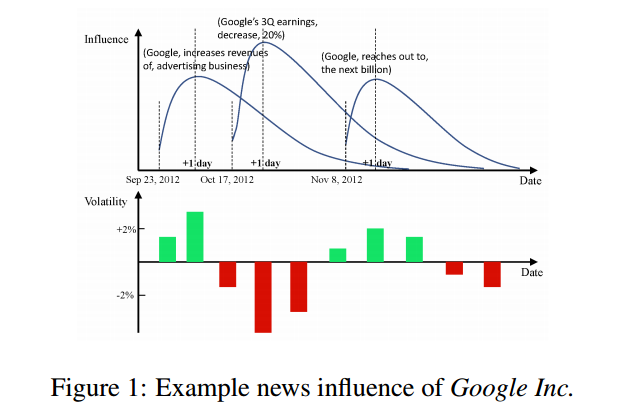
其中解释和被解释变量定义如下，包含了待解释的IPO抑价率，和所研究的情绪指标。

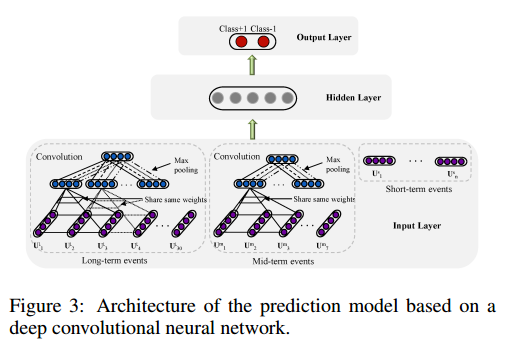


结果表明，应用“股吧”讨论帖子数量作为投资者关注度的代理指标，投资者关注度与 IPO 抑价显著正相关，即投资者关注度越高，IPO 抑价率 越高。并且投资者关注度同样与样本公司 IPO 首日的换手率有显著的正相关关系，即投资者关注度越高， IPO 首日交易量越大。而投资者乐观情绪与投资者意见分歧与IPO抑价和上市首日交易量没有显著关系。

同时, 股票论坛里异常的发帖活动和观点变化与股票异常成交量和回报率有显著关系（Tumarkin R,Whitelaw R F.2001） 。从股票论坛中提取中小投资者观点,构建的投资者情绪指标与传统指标存在一定关联性（Das S R,Chen M Y.2007）。

在新闻数据的应用上，Ding X, Zhang Y, Liu T, et al.（2015）提取新闻词向量并使用 CNN 对 S&P 500 指数及个股价格进行预测， 另外，他们还使用 NLP 技术处理事件文本信息和知识图谱提供的实体、属性等信息预测股市波动率。在 S&P 500指数和个股预测上取得了比现有最好模型6%的提高，模拟实验也显示据此模型建立炒股策略比现有炒股策略收益高。

.



## 5.2 投资市场情绪监测

2012 年，国泰君安推出了“个人投资者投资景气指数”（简称 3I 指数），该指数系国泰君安研究所对海量个人投资者样本进行持续性跟踪监测，对账本投资收益率、持仓率、资金流动情况等一系列指标进行统计、加权汇总后得到的综合性投资景气指数。旨在通过对中小投资人真实投资交易行为的量化解读，更好地了解投资人对市场的预期以及当前的风险偏好等信息。

　　在样本选择上，选择资金100万元以下、投资年限5年以上的中小投资者，样本规模高达10万，覆盖全国不同地区，具有相当的代表性。对于沪深股市等个人投资者为主的市场更有针对性。在参数方面，主要根据中小投资者持仓率的高低、是否追加资金、是否盈利这几个指标，来看投资者对市场是乐观还是悲观。

“3I指数”每月发布一次，以100为中间值，100—120属于正常区间，120以上表示趋热，100以下则是趋冷。从实验数据看，从2007年至今，“3I指数”的涨跌波动与上证指数走势拟合度相当高。

图为2012年1月至2014年6月3I指数运行走势，可以看出个人投资者的投资景气度处于“低迷”（80以下）。



此外，基于上市公司经营报表统计、券商投行研报看涨看跌指数（可按地域/行 业/经营领域细分）的分析，也可以为投资市场情绪提供重要参考依据。

## 5.3 大宗商品期货相关因素

商品期货是买卖双方在未来某个约定的日期按约定价格进行交易的标准化协议。如何对海量数据进行分析并对期货价格等因素做出预测,以期获取稳定的收益成为商业与学术界关注的重点研究问题。

李新（2015）构建反映投资者对国际原油市场关注度的衡量指标，探讨投资者关注与国际原油市场上不同类型的投资者交易持仓关系。采用网络挖掘-特征选择-广义动态因子模型(WM-FS-FHLR)提取互联网搜索关键词，进行特征选择筛选变量，并构建反映互联网关注的“一致指数”。引入生物信息领域的因果分析方法Partial Directed Coherence(PDC)分析因果关系对所构建的指数进行综合评价。该指数是一个基于互联网数据所构建的综合指数，可用于更全面地衡量原油市场中部分投资者的关注，更及时且准确的表征原油市场的动态变化。

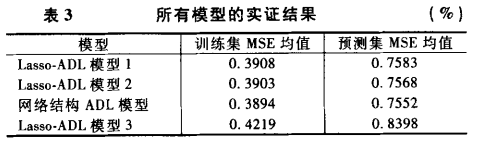
进一步构建多组VAR模型来研究投资者关注指数对非商业投资者、商业投资者及散户的持仓关系，从而分析投资者关注指数与投机活动的领先性。

岳艳涛, 章雅婷, 张宇,等（2015）利用期货交易所每秒钟提供两笔交易品种的实时数据(成交量、持仓量、总量、买一价、买一量、卖一价、卖一量)，首先确定影响价格的因素。通过成交价与7个因素的散点图和相关系数分析,发现成交价与B1价S1价相关性较高,然后利用ADF单位根和Johansen检验法对数据原序列、一阶差分序列进行平稳性和协整检验,采用Granger因果关系检验确定具有滞后性的影响因素。

其次对期货价格进行了两部分的预测。首先对橡胶价格进行短期预测,建立自回归与分布滞后模型,用Eviews6.0和SAS9.1软件进行求解和相关性检验。同时利用小波神经网络,对历史数据进行挖掘和模拟分析,进行200和500次训练,预测结果的相关系数达到0.9493,比较预测值与真实值可知模型对短期预测精度高。然后对波动价格长期预测建立阻尼衰减趋势指数平滑模型,并对价格进行预测与误差分析。

进一步，通过期货价格预测还可以建立投资收益模型,以期获得最优收益。从期货交易买卖角度对问题进行分析建立动态规划模型,利用遗传算法,求解预测价格下的交易策略，同样取得了较好的精度和稳定性。

王娜（2016）为了研究大数据是否有助于预测碳排放权价格,本文讨论了结构化数据和非结构化信息对预测碳价所起的作用.结构化数据选取国际碳现货价格、碳期货价格和汇率,非结构化信息选择百度搜索指数和媒体指数.考虑到当解释变量很多时,平等对待每一个解释变量是不合理的,提出了网络结构自回归分布滞后(ADL)模型,在参数估计和变量选择的同时兼顾了解释变量之间的网络关系.实证分析表明,网络结构ADL模型明显优于其他模型,可以获得较高的预测准确性,更适合基于大数据的预测.



从总体上看，Lasso—ADL模型2训练集MSE均值和预测集MSE均值都低于Lasso-ADL模型l，网络结构ADL模型同样如此。LLasso-ADL模型2和网络结构ADL模型都使用了非结构化信息，而LassoADL模型1并没有使用。所以，加入非结构化信息可以改善预测效果。

## 5.4 其它金融相关因素

对于其它金融相关因素如市场流动性、投资者交易行为等也有利用大数据展开的相关研究，大数据对于金融市场运行规律的揭示起到了一定作用。

王晓琳（2016） 关于投资者行为如何影响金融市场进行了探讨。首先，探讨了投资者关注对分析师盈余预测行为的影响，运用标准普尔500股票的谷歌搜索趋势作为投资者关注的代理变量，发现搜索量与分析师预测修正、分析师总人数和分析师预测分歧相关。其次，发现我国股指期货市场流动性日内效应与周内效应，得出了期货市场流动性的日内模式在上午呈现反“J”形变化模式，下午呈现倒“U”型变化模式，期货流动性的周内模式呈现倒“V”形变化模式。表明中国股票指数期货市场流动性的周内模式和日内模式存在着显著的周期性。并且，基于指令失衡这一关键指标作为投资者交易行为的代理变量，分析了投资者交易行为与价格行为之间的关系，结果发现在中国农产品期货市场指令失衡具有收益预测的能力。过度指令失衡引起的价格行为并不符合传统库存理论的解释，中国期货市场的价格行为主要由投机者的羊群效应导致。

运用多种统计学与计量经济学中常用的分析方法，包括：时间序列回归分析、协整检验、VECM模型估计、格兰杰因果检验与Fama-MacBeth截面回归模型等，具有前沿性、系统性、市场性和监管导向性。

# 第六章 基于互联网大数据的资源能源政策经济因素分析

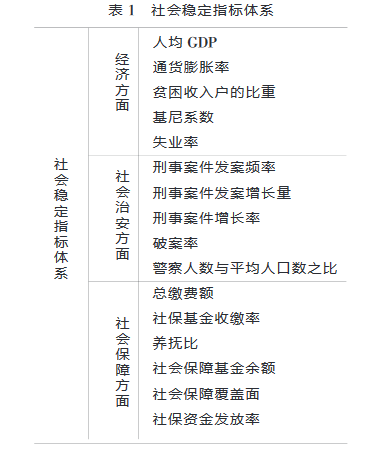
## 6.1国家政治和经济稳定性评估技术分析

### 6.1.1国家政治和经济稳定性评估模型

国家政治和经济稳定是国家社会稳定的重要因素。资源能源生产国的政治和经济稳定性直接关系到资源能源来源的安全和风险，资源能源产品的出口国的政治和经济稳定性也直接关系到我国资源能源的需求和消费。

上海《社会稳定指标体系》课题组（2002）指出，用于衡量社会政治、民生稳定的一般状态指标可以由以下指数定义：社会痛苦指数、公共安全指数、精神卫生指数、社会紧张指数和利益保障指数，其中具体包括通货膨胀率、城镇登记失业率、刑事案件立案率、重大交通事故发生率、每万人精神病患者比率及增长率、信访率、劳动合同纠纷受理件数及增长率等23个指标。同时，还考虑了社会稳定突变状态指标，包括社会冲突、腐败、分裂指数；以及社会稳定解释性指标，包括经济总量、生活水平、贫富差距、社会保障、政治透明度等指数。

夏帆（2004）从三个方面构建社会政治稳定指标体系。包括经济方面的人均GDP、通货膨胀率、失业率、基尼系数等指标；社会治安方面的刑事案件发案频率、增长率、破案率等指标；社会保障方面的社保基金总缴费额、社保基金收缴率、社会保障覆盖面等指标。见下图。



柳长兴，秦琴（2011）从主观、客观两个方面考虑社会稳定水平。客观因素中包括基尼系数、失业率、CPI、刑事案件发案率、公共财政投入、社会保障覆盖率等指标。主观因素则从总体生活、政治生活、经济生活、法治与权益质量等满意度维度指标进行测算，建构了社会稳定的主观综合指标体系，设计科学合理的抽样模型，用问卷调查和社会科学统计分析方法调查计算重庆市社会稳定总体性水平。

综上所述，社会政治稳定水平可以由主观和客观的指标进行衡量，而客观指标更受到研究者的关注。其中受到普遍认可指标的有：经济指标如通货膨胀率、失业率；社会治安指标如刑事案件、重大案件发案频率；社会保障指标如社会保障覆盖面等。

### 6.1.2利用社交媒体大数据进行稳定性评估综述

互联网社交媒体提供给人们随时随地发布、传播信息的服务，其信息传播的广泛性和快捷性使得其成为人们政治、社会生活的新的平台。2010年[阿拉伯世界](http://baike.baidu.com/item/%E9%98%BF%E6%8B%89%E4%BC%AF%E4%B8%96%E7%95%8C" \t "http://baike.baidu.com/_blank)爆发的民主运动“阿拉伯之春”被认为是一场“Twitter革命”。有媒体评论，“阿拉伯之春”属于“谙熟互联网、要求和世界其它大部分地区一样享有基本民主权利的年轻一代”，互联网、社交媒体的重大作用意义可见一斑。

Howard et al 指出，在阿拉伯之春事件中，社交媒体上一定类型的活动加速了大规模抗议等事件的发生。

Sullivan.2011分析了突尼斯同摩尔多瓦、伊朗的抗议活动都可被称为“Twitter革命”的原因。抗议活动期间，传统媒体完全被政府控制失去了传媒功能，无法从中获取事态的最近进展，因此人们转向了YouTube上的每日活动视频, 以及 Twitter、Facebook等社交媒体上的博客信息。特别是由于网络站点封锁和网络攻击等原因，政府难以对互联网上的抗议信息进行控制，导致相关信息广泛传播，网络上抗议呼声一时高涨。

Bruns, Axel,Highfield, Tim, &Burgess, Jean E.(2013）也考察了在埃及和利比亚的冲突中，英语和阿拉伯语用户在Twitter上的发言情况。文章通过统计微博数量分析得出结论，阿拉伯语用户中，埃及的Twitter活跃用户数相对较高，总微博数量也达到一定规模，因此可以被用于观察事件参与者的情感倾向，同时也证实了，以Twitter为代表的网络社交媒体在埃及的国家抗议活动中起到了一定传播、组织和发布信息的作用。

上述研究均分析了“阿拉伯之春”在社交媒体上的强烈反响，探讨了社交媒体在事件中起到的作用。正因如此，社交媒体上的动态也可以作为分析、预测实际事件走向的依据。

Xiaofeng Wang et al. 2012使用Twitter微博文本数据对肇事逃逸案件的发生进行预测，将犯罪预测的监督学习框架和社交媒体数据进行结合。文章整体采用有监督学习框架，使用一定量的微博文本和对应时间段发生的肇事逃逸案件信息进行学习，需要采集微博数据，和真实的案件发生数据。其中使用了较为精确的自然语言处理工具来对微博文本进行特征提取，首先NLP（natural language processing）工具抽取一条微博中的事件内容，再使用半监督的基于事件的主题分类模型确定关键的主题词，并根据主题词来进行预测。单纯采用时间维度的分析并不能有效提取案件的发生规律，预测效果近似随机猜测，而该文中的方法则能显著提升预测性能。

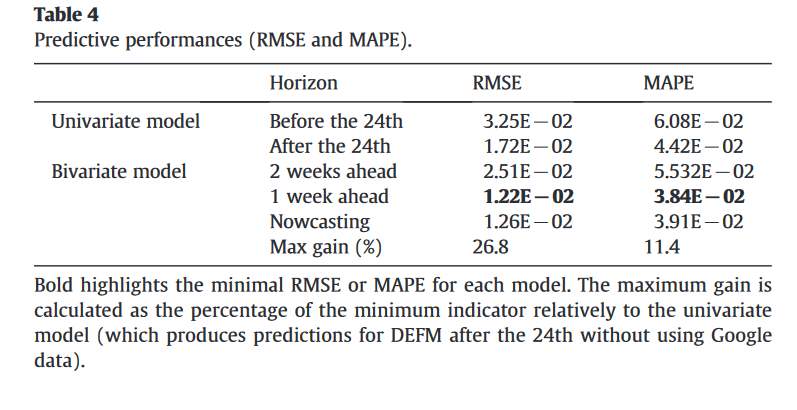
目前研究仅针对了肇事逃逸案件，如能获得实际的案件信息，也可用于其他类型的案件，特别是对国家政治稳定性有重大影响的案件，如抗议示威、大规模游行等案件，对于评估社会稳定突变状态指标，或刑事案件发案率等社会治安指标都有一定作用。

### 6.1.3利用搜索引擎大数据的评估综述

由于越来越多的经济活动都转变为在网络上进行，或者在网络所提供的种种服务中留下交易痕迹。因此搜索引擎在评价经济指标上的作用得到了研究者的重视，一系列研究都表明以Google Trends为代表的搜索引擎大数据对阐释经济市场的波动震荡，和做出精确地经济评估及预测有明显的作用。

Askitas and Zimmermann (2009) 发现了Google Trends搜索量和德国的失业率之间有强正相关性。Choi and Varian(2009b) 对美国，Suhoy (2009) 对以色列，D'Amuri (2009)对意大利 ，以及D'Amuri and Marcucci (2009) 对美国的失业率进行考察时都有相似的结论。

Fondeur and Karamé (2013) 验证了 Google Trends数据能否用于预测法国青少年（15-25岁）的失业人口数量。文章使用了政府发布的周度和月度失业人口数据，结果显示，使用了 Google Trends数据的预测模型效果要远远优于未使用的模型。在两年内的失业数据上验证，使用了 Google Trends数据的预测模型对预测质量整体提高了27%，对预测的准确度平均提高了40%。（见下图）



利用搜索引擎大数据的评估主要从两个思路来提升性能，一是搜索关键词的选择，另一是建模方法的使用。在关键词选择中，Fondeur and Karamé (2013)简单选择了EMPLOI（意为工作、失业），文章认为，该关键词可以作为劳动力市场供需双方查找信息的入口。Askitas and Zimmermann (2009)选择了四组，每组1-8个关键词，如HR咨询、招聘网站，以及直接使用失业、就业等词语，从4个角度对失业相关的搜索入口进行了考察。Choi and Varian(2009b) 选择了jobs、welfare & unemployment两个关键词。D'Amuri and Marcucci (2009) 仅选择了jobs一个关键词。可见，直接使用失业、就业等词语，选择较为简单而少数的关键词是目前采用较多的方法。在建模方法上，Fondeur and Karamé (2013)选择了扩散的Kalman滤波和最大似然估计法，以得到非参数滤波不能达到的更好的数据贴合度。

## 6.2重要公投与政策预测技术分析

### 6.2.1研究现状综述

英国脱欧、苏格兰独立等国际上的一些重要的公投事件对于国际政治、经济、金融和外交环境都有着密切的联系，对资源能源产品的生产和出口也有直接或间接的影响。公投结果的预测方法受到了广泛关注。

Ronald MacDonald 等提出了一种使用Google搜索引擎数据和计量经济学思路分析苏格兰独立公投的方案，使用Google Trends中与公投相关的周搜索量数据建立计量经济学VAR模型，在2013年8月至2014年9月公投前2天的数据上预测赞成票率为45%，与实际结果44.7%十分接近。利用Google Trends日搜索量数据对公投前最后一个月的影响预测结果的因素进行分析，定量地探讨了期间重大事件和其他因素对结果的影响程度，为合理解释公投过程和民众心理提供了新的视角和依据。

传统的公投预测使用的新闻报道等信息是对提供给投票者的信息进行分析。 这种信息是投票者被动接受，并不能从中了解投票者的情感倾向。而利用Google Trends 数据则可以分析投票者主动希望了解的信息。

### 6.2.2利用搜索引擎大数据的预测技术路线

利用Google等社交媒体数据进行选举预测的技术路线可以总结如下：数据采集、预测方法应用和效果评价。下面针对前两个部分进行详细阐释。

1. **数据采集**

利用Google Trends等搜索引擎工具提供的搜索量查询接口，设定一定的关键字和采集时间范围以得到搜索量数据，还可以设置地理位置的限制。

Ronald MacDonald 等在利用Google Trends对苏格兰公投进行预测时，采用单一的关键字“Alex Salmond”，采集时间为2013年8月至2014年9月公投前2天，同时设置了搜索请求的地理位置必须在苏格兰境内。

其中搜索关键字的选择很大程度上决定了预测的质量，在Ronald MacDonald等在苏格兰公投的案例中，提出了选择搜索关键字的两个原因，为关键字的选择提供了一定思路：

1. Alex Salmond为公投时期苏格兰首席大臣，并且担任苏格兰民族党党首（主张苏格兰独立），因此是投票者最可能在Google 上搜索以了解苏格兰独立相关信息的关键词。
2. 在苏格兰议会选举中，“Alex Salmond”的搜索量就和最终的投票表决结果有高度的正相关性。如其在2011年2月到5月间的搜索量和投票结果有0.6的相关性。
3. **预测方法应用**

Ronald MacDonald 等使用Google Trends中该关键词的周搜索量数据建立计量经济学VAR模型，并在考虑了意外事件波动时对结果进行了适当修正。VAR模型及建立过程采用了Brooks and Tsolacos(1999) 与Vozlyublennaia (2014)中实用的方法。采用的两个主要变量分别是“可能的赞同票数”（来自6家主要选举公司的数据平均）和Google Trends搜索量。修正结果时考虑了变量的不确定性进行动态预测。Doornik and Hendry (2013)也采用了类似的模型和流程流程。

## 6.3重要选举预测技术分析

### 6.3.1研究现状综述

资源能源生产国、资源能源产品的消费国、资源能源运输重要通道的所在国，根据其政治体制的不同，大都会周期性的产生其国家或地区的政治领导人。新领导人的政治倾向和政策选择将会对资源能源的安全风险产生巨大的影响。重要选举的预测也逐渐从仅仅依靠传统的民意调查到同时利用互联网大数据提供的社会舆论信息。

在选举预测问题中Twitter数据得到了广泛应用。O’Connor et al.2010最早使用Twitter数据代替传统的民意调查，根据预先得到的词汇与其情感的对照表给每条微博的情感评分，再对正面、负面的微博数量统计。将微博中针对某个候选人的情感倾向评分与选票结果进行关联度分析，探讨了微博数据的选举预测作用。Tumasjan et al.2010对Twitter数据中提及选举各方政党的微博数量进行统计，将结果用于预测2009年德国联邦选举，预测的胜率误差为1.65%，效果优于传统民调。Bermingham and Smeaton 2011利用Twitter数据，并使用了多种语义分析方法，在2011年爱尔兰大选中进行测试，预测的胜率误差为3.67%，略优于民调结果。Metaxas et al.2011使用Twitter中参选者被提及数量以及语义分析的方法对2010美国各州选举进行预测，在6个州中两种方法均有3个州正确预测了获胜方。

也有研究者探讨了搜索引擎数据在选举预测中的使用。C. Lui 2011关注了 Google搜索量数据对于2008和2010美国国会选举的预测作用。他们根据候选人之间竞争的激烈程度将全部选举划分为若干竞选单元，发现使用Google搜索量数据的预测在部分单元中得到了比随机猜测更好的效果，最好结果出现在2008年的某个单元，达到了81%的准确度。[17] C. Lui, P. T. Metaxas, and E. Mustafaraj, “On the predictability of theu.s. elections through search volume activity,”

### 6.3.2利用社交媒体大数据的预测技术路线

作为目前研究选举预测问题的最常见思路，利用Twitter等社交媒体数据进行选举预测的技术路线可以总结如下：数据采集、数据清洗、预测方法应用和效果评价。下面针对前三个部分进行详细阐释。

**1）数据采集**

主要利用Twitter等社交媒体提供的数据访问接口，如Twitter search API来获取选举前一段时间内全部或随机比例的微博用户发布的与选举相关的微博信息，从选民个人发布的相关内容中推测个人及选民整体的选举意向。其中需要确定采集的时间，过滤微博的关键字及其他参数配置。

现有相关研究在数据的采集时间上有较多差别，虽然采集的截止日期大多选择了选举日前一天，但是开始时间却各不相同，时间跨度从选举前一周到选举前三年均有实验。目前时间选择对预测结果的影响尚没有确切的结论。如Tumasjan et al.2010在预测2009年德国联邦选举时，选取了选举前7周到选举前1周的数据。而Tjong Kim Sang& Bos 2012在预测2011年荷兰参议院选举时则只选取了选举前1周到选举日前一天的数据。

同时，大部分研究都对微博内容进行了一定过滤，限制内容的题材为与待预测的选举相关,通常会选择候选人的姓名、参选政党的名称作为过滤的关键词。Metaxas et al.2011在预测2010年美国选举时采用了候选者姓名在gardenhose数据库中过滤相关微博条目。Skoric et al.2012在预测2011年新加坡大选时同时采用了候选人的姓名和参选政党的名称过滤相关微博条目。

1. **数据清洗**

由于选民的微博是目前多数研究采用的数据来源，因此以下讨论的数据清洗也主要针对这种情况。

* 保证数据纯粹性

保证数据的纯粹性要求发布微博等社交信息的用户是可能的参选者。鉴于这种信息并没有在用户个人资料中明确显示，因此需要通过其他信息来推断。目前研究中所采用的限制措施有两种，即限制地理位置和限制微博内容的发布语言。

限制地理位置时可以通过控制Twitter search API的有关参数，Gayo-Avello 2011在预测2008年美国大选时就采用了设置参数的方法将微博数据来源限制在所希望了解的郡县中。也可以通过限制微博发布者的地理位置达到上述目的，Skoric et al.2012在预测2011年新加坡大选时限制了发布者的位置必须在新加坡境内以去除不具有选举资格的用户发表的内容。

限制发布语言主要应用在德语、荷兰语等全球化程度不太高的语种上，如Tumasjan et al.2010 以及Jungherr et al.2011在预测2009年德国联邦选举中都限制了微博必须使用德语，以减少数据集中非德国民众发布的微博信息，在一定程度上达到了与限制地理位置同样的效果。

* 去除数据偏差

去除偏差即减少数据的不平衡性，使之可以成为统计意义上的代表样本。Twitter等社交媒体中采集到的用户群并不一定符合人口统计学上的均衡性，而这种偏差导致了数据样本的片面性，特别是在选举这种需要考虑全局民意的情况下会造成对预测结果较为严重的的人为干扰。解决这种数据偏差的一个直观思路是对每个用户所属的人口统计学上的社会阶层进行划分和判断，根据用户所在阶层的选举活动参与程度得出其发布内容的重要性，给予不同用户的信息不同的的权重。

然而目前的互联网社交媒体中没有强制用户登记他们完善的真实信息（如性别、年龄等），而从用户发布的内容中推测这些个人资料又涉及大量工作，因此使得将用户划分为不同的投票阶层十分困难。但仍然有研究者做出了一定尝试，Gayo-Avello 2011 利用用户登记的真实姓名、郡县等信息，以及线上的公开查询记录得到了2500名用户的年龄，以此估计出他数据集中用户的年龄范围在18-44周岁之间。接下来，他根据在2004年大选中选民的年龄分布情况对数据集中的用户信息给予不同的权重。这两步措施使得预测误差从13.10%降低到了11.61%，可见进行了数据去偏差后对性能带来了显著提升。

* 数据去噪声

数据去噪声表示去除与实际选民的选举意向无关的内容，如没有信息量或灌水、炒作的条目。数据中噪声的存在会降低预测方法的使用效率，更有可能对预测过程造成干扰，使预测结果出现错误。尽管目前的选举预测研究中没有研究者采用过实际的数据去噪声方法，但一些研究者已经意识到了数据去噪的重要性和必要性。

Metaxas et al.在2011年的研究中指出，恶意灌水和政治宣传者都会利用软件创造一些虚假的社交媒体账户，并通过这些假账号夸大、散布他们的政治主张，刻意营造其政治势力，这种行为污染了每个数据集观察者的视听。因为社交媒体所呈现的信息在人群中有广泛的影响力，这一平台对恶意灌水和政治宣传者有重要意义，使他们很可能在其中混淆视听，达到干扰舆论的效果。

更进一步，他们进行了实验以验证基于词极性的语义预测方法在被灌水信息上的鲁棒性表现，结果显示，仅仅依靠词极性的语义预测方法并不能识别出被灌水信息和正常微博信息的微妙区别，并且会对这些虚假的信息作出错误的判断。更加表明了，在选举预测过程中进行数据去噪声的步骤应该是必须完成的。

1. **预测方法应用**

使用Twitter进行选举预测方法的核心即是从微博中推测选民意向，可以先推测单个用户的意向再对整体进行统计，也可以跳过单个用户，直接以整体意向为单位统计、推测。预测方法在范围上可分为在国家、州、郡县等层面，在预测结果上可分为预测获胜率或预测获胜方。

从技术角度，预测的方法主要有以下两类：

* 使用关键词提及数量

简单统计微博中提及某一选举人或政党的数量，预测的准确率随着总微博数量的增大而提高。其优点在于：易于实现；可以在实际临近投票结果前应用；不仅可以用于预测选举的总体情况，并且基于用户倾向选举的对象会在其微博中被最多提及还可以用于预测单个用户的选举倾向；并且被认为是与传统民意调查的思路接近而具有很高的准确率。

Tumasjan et al.2010 以及Jungherr et al.2011在预测2009年德国联邦选举中都采用了微博关键词计数的方法。Tumasjan et al.2010 对提到参选各党派的微博进行统计，发现各党派微博数量的排名与其在选举中得票数的排名相同，并且有相似的比例。简单地采用微博数量进行预测可以达到误差1.65%的效果，与6家专业选举机构发布的预测结果比较，表明预测水平近似。这一结果与Véronis (2007)指出观点相符合的，即在2007年法国总统选举中，简单地使用参选者提及数统计可以获得比传统民调更好的结果。

Tjong Kim Sang& Bos 2012在预测2011年荷兰参议院选举时遵循了与Tumasjan et al.2010类似的思路，他们根据微博用语的特点，同时对政党名称的缩写及全称进行计数，按照政党名称的提及数量占比预测其获得的席位，与2家选举公司的结果对比，达到了相似的结果。进一步，他们分析了数据集中存在的偏差和噪声问题，只保留了每个用户的第一条、且只提及一个政党的微博，获得了更好的效果。

但由于其未考虑任何语义信息，忽视了文本中不同的感情倾向而受到质疑。Gayo-Avello 2011在预测2008年美国大选时的实验显示，单纯使用微博数量统计的预测效果甚至远远低于随机猜测。

* 使用语义情感分析

只有少数研究采用机器学习的方法对文本进行正面、负面的情感分类如Bermingham and Smeaton 2011 ，Tjong Kim Sang and Bos 2012。

多数研究仅使用了词汇特征，根据词汇与其所对应感情（词汇极性）的对照来推测整条微博的情感和选举倾向。这种方法简单，易于实施，但也被证实有许多分类错误，精确的语义情感分析仍然是一个困难的挑战。

O’Connor et al.2010在预测2008年美国大选时，首先筛选出包含与选举相关关键词的微博条目，如包含obama和mccain的条目。再根据词汇极性对照表OpinionFinder（Wilson et al.2005），只要一条微博信息中出现了正面或负面情感的词汇，就相应地认为该条微博针对该关键词也具有正面、负面的倾向。 为了衡量某天的正面情感的强度，定义分值为当天内该关键词正面微博条目数/负面微博条目数。

Gayo-Avello 2011在预测2008年美国大选时，采用了两种基于词极性的方法。一种方法对微博中词极性进行统计，如果正面词汇多于负面词汇，则认为该条微博也有正面倾向。通过统计某个用户对于所有候选者的正面倾向评分，可以得出分数最高者即为该用户的选举意向。一种为Vote & Flip方法，首先对句子中不同极性的词汇进行统计，再应用一系列复杂规则来判断整个句子的极性，对传统的简单情感推测进行了规则上的扩充。

该研究还采用了另一种semantic orientation（语义源分析）方法。该方法在句子中寻找有特定情感倾向的短语，这些短语预先定义的短语可以借助搜索引擎获得，也可以采用那些已经表现出自己选举意向的微博中短语。其基本思路与词极性法相似，而将最小的情感推测单元扩充到了短语层面。

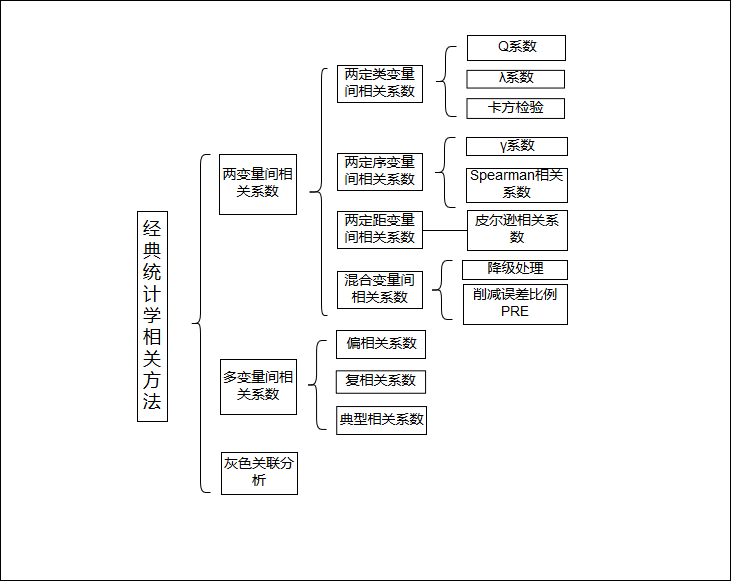
# 第七章：资源能源安全风险因素关系网络构建

## 7.1数据驱动的关系网络构建方法

数据驱动的关系网络构建方法从数据的视角挖掘资源能源安全风险因素之间的关系。通过因素变量的历史数据来自动挖掘出可能蕴含在数据中的规律，揭示因素间的关系和变化原因。数据驱动的方法可以分为传统的统计学方法，面向大数据的相关分析方法，数据挖掘的方法，以及基于因素与数据源间关系的推测方法。

### 7.1.1传统统计学的相关分析方法

相关分析在经典的统计研究中已经被广泛讨论，针对不同类型的变量和随机向量形成了种类繁多的相关系数。统计学的相关分析方法针对变量的数值序列进行分析，通过考虑其服从的统计模型的方法来衡量变量间的相关性。具体而言，它通过定义相关系数的计算方法给出相关性的定量度量，是一种典型的数据驱动的分析方法。统计学相关分析方法可以分为两变量间相关系数、多变量间相关系数，同时，灰色关联分析作为一种使用较为广泛的模型是对经典的方法的很好补充。



1. **两变量间相关系数**

* 两定类变量间相关系数

定类变量又称为名义变量，这种变量的值代表一个类别，没有大小、次序之分，比如性别这个变量的取值为男、女这样的类别。计算两定类变量间相关系数通常需要使用列联表，也称交互作用表。其典型的相关系数为Q系数（只能计算2\*2的列联表），λ系数（可以计算任意两个定类变量将相关性），以及卡方检验等。其中最常用的是λ系数法，其优点是计算简单，前提宽松，且具有削减误差比例意义。

* 两定序变量间相关系数

定序变量也称等级变量，其取值有序的意义，即其取值区分等级和序列。例如高校教师职称分为助教、讲师、副教授、教授四个等级。两定序变量间相关性定义常用的概念包括同序对、异序对。常用系数有γ系数、Spearman系数（统计上可看为皮尔逊相关系数的特例），这两个系数本身或其变形具有削减误差比例意义，因此较为常用。其他针对定序变量的相关系数还有dy系数，肯德尔系数等。

* 两定距变量间相关系数

定距变量即数值变量，其取值有数值上的差别，可以进行加减乘除运算。度量其相关性的常见方法是皮尔逊相关系数法，其主要思想是将两向量中心化后计算向量夹角的余弦值。该方法在几何、代数、三角等角度上都有良好的数学解释，且其平方具有削减误差比例意义，因此被广泛使用。也有学者针对其容易受到大样本、异常点影响的缺点进行改进。

* 混合变量间相关系数

主要采用两种方法。其一是直接采用削减误差比例法，其主要思想是如果已知X对预测Y有很大的帮助，则X与Y相关性较大。通过计算已知X预测Y产生的误差和不知道X预测Y产生的误差来进行判断。其二是采用降级处理，其主要思路是把其中一方的变量类型降级，如计算定序和定类变量相关性可将定序变量降为定类变量处理，其缺点为信息损失。

1. **多变量间相关系数**

常见的多变量间相关系数有偏相关系数、复相关系数和典型相关系数，且都是两定距变量间线性相关系数（皮尔逊系数）的推广。偏相关系数衡量的是多个变量中固定其他变量时任意两个变量的相关性。复相关系数衡量一个变量对多个变量的相关性。

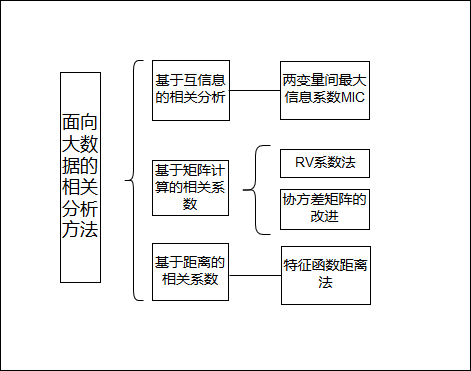
典型关系数衡量多个变量对多个变量的相关性，是一种较为通用的多元统计方法。典型相关分析的主要思路是寻找两个投影方向，使得中心化后的两个变量在投影后皮尔逊相关系数最大。即寻找两个线性变换，将原始数据压缩到一个低维子空间，使得压缩后的数据在子空间线性相关性最大。典型相关分析在电子通信、生物信息、经济管理等领域都有广泛应用。

1. **灰色关联分析**

灰色系统理论以部分信息已知部分信息未知的“小样本”“贫信息”不确定系统为研究对象，主要通过对部分知信息的生成开发提取有价值的信息，实现对系统运行行为、演化规律的正确描述和有效监控。灰色关联分析是灰色系统理论的重要组成部分之一，而且是灰色系统分析、建模、预测、决策的基石。其基本思想 是根据事物或因素的序列曲线的相似程度来判断其关联程度的。若两条曲线的形 状彼此相似，则关联度大；反之，关联度就小。灰色关联分析在经济、社会、农业、矿业、交通等各个领域都有广泛应用，如小麦生育期气候因素与产量的关联分析[39]、中国碳排放强度与产业结构的关联分析（李健）、 区域经济系统的分析[41]等等。（孙玉刚）

### 7.1.2面向大数据的相关分析方法

针对经典统计学不适用于高维、非线性大数据的缺陷，许多面向大数据特征的相关分析思路被提出。其中应用较多的包括基于互信息的方法、基于矩阵计算的方法、基于距离的方法。



基于互信息的方法采用信息论的思想，使用信息熵来度量系统的不确定性。通过计算已知X后Y的信息量变化程度，如果变化程度越小则说明X与Y的相关性越弱。一个经典的该方法使用案例是两变量间最大信息系数的定义（MIC），实验表明MIC比传统的皮尔逊系数更能细致刻画变量间的相关性，且对非线性的大数据相关性有很好的挖掘能力。

基于矩阵计算的方法重点考虑用矩阵论的思想进一步推广，发展经典的统计相关分析方法。其典型方法例如RV系数、协方差矩阵的改进。RV系数是一个针对两个随机向量的更具有泛化性的相关系数，可以证明多种经典统计相关分析方法都可看为RV系数广义模型的特例。协方差矩阵的改进相关研究通过简化协方差矩阵表，降低计算复杂度，更能适应大数据的高维特性。

基于距离的相关系数通过分布的特征函数与距离相关联，得到相关系数。这种方法的优点在于可以度量的相关性不限于线性关系，同时可以度量任意两个不同维度的随机向量间相关性。被广泛应用与机器学习，时间序列分析，核化学等分析领域。对于大数据的高维、非线性特性都有较好的适应性。

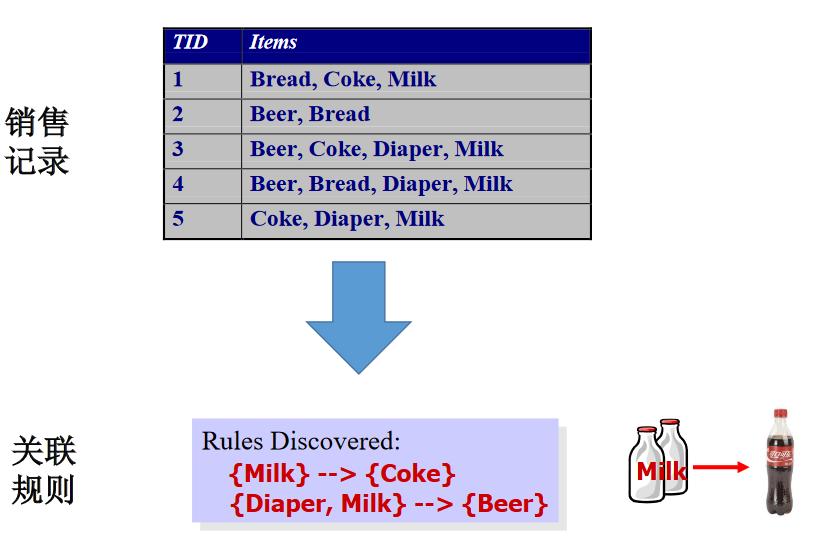
### 7.1.3数据挖掘的方法

数据挖掘技术是针对海量数据的处理而发展起来的一种数据处理技术 ，它通过对历史数据和实时数据进行分析 ，最后总结出预见性结论。

关联规则挖掘是一种在大型数据库通过识别数据库中的频繁模式发现变量之间关系的方法，是数据挖掘最先研究的问题之一，也是数据挖掘的一个主要研究方向。通过将资源能源安全相关因素的监测历史数据建立数据库，就可以采用关联规则挖掘的方式挖掘出不同因素间的关联关系。

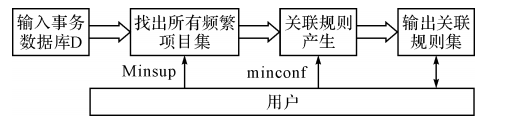
**1）关联规则挖掘简介**

关联规则挖掘最早是为了发现超市交易数据库中不同的商品之间的关系。关联规则可以反映出隐藏在数据库中有价值的知识。关联规则是形如X→Y的蕴涵式，存在支持度和信任度。X和Y分别称为关联规则的先导和后继。简言之，关联规则就是描述一个事物中物品甲的出现对物品乙的出现的影响有多大的一种规则。



关联规则挖掘的主要思想是通过用户指定最小支持度和最小置信度，挖掘出强关联规则。代表的方法有Apriori算法、FP-树频集算法、多层次关联规则算法、多值属性关联规则算法等等。其基本实现是一个两步的过程 ：1） 根据最小支持度找出事务数据库 D 中所有的频繁项目集 ；２） 由频繁项目集和最小置信度产生强关联规则。

经典的 Apriori方法的核心思想是利用1-频繁项目集生成２-频繁项目集 ，然后根据２-频繁项目集生成３-频繁项目集 ，以此类推 ，直至生成所有的频繁项目集 ，然后从频繁项目集中找出符合条件的关联规则。技术实现上使用层次顺序搜索的循环方法（又称做逐层搜索的迭代方法）产生频繁项集。 Apriori算法在众多领域都有广泛应用，也有许多针对不同场景和需求的改进算法。下图展示了关联规则挖掘的基本过程。



**2）关联规则挖掘的应用**

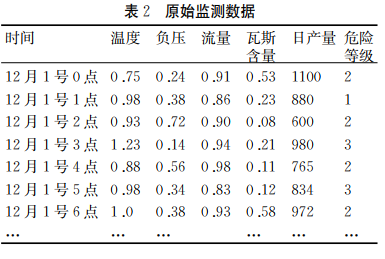
关联规则数据挖掘方法应用十分广泛。在网络安全领域，入侵检测系统 (IDS)运用关联规则生成入侵检测标准规则库，通过判断网络数据与规则库的匹配度产生告警数据流。隐私保护中，隐私保护关联规则算法基于数据干扰和分布重构思想对原始数据进行保护，避免商业敏感数据、个人隐私数据泄露。利用关联规则还可进行蛋白质结构预测，根据蛋白质序列的特点可以对这些序列数据进行量化和处理，再将关联规则算法用于序列数据集上寻找蛋白质序列中的关联关系。在遥感导航定位等科学领域，关联规则算法也被用于分析事件出现的原因，解释其与指标之间的关系。以下给出了关联规则算法用于社会经济相关问题一些的案例，能更具体阐明其用途。

（张鑫，陈燕）使用Apriori算法对经济指标关联进行了分析。选取1999～ 2012年大连市的GDP、CPI、失业率、港口货物吞吐 量、港口旅客吞吐量、航空旅客吞吐量增量六大指标来进行关联挖掘分析。得出的关联规则例如在最小支持度为0.5和最小可信度为0.7的水平 下，GDP增量范围在[0，0．17)且航空旅客吞吐量增量范围在[0，0．19)时，CPI增量％范围在[一2， 1．7)。得出了港口货物吞吐量、航空旅 客吞吐量、港口旅客吞吐量等与经济发展三大指标存在潜在关联，为利用吞吐量来预测经济发展趋势提供理论依据。

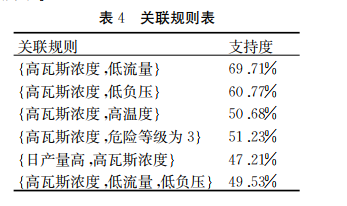
（王晨绫，刘伟）基于Apriori算法对微信热点事件的关联进行了分析与研究。使用微信平台萃取出来的大量热点事件，剔除无效数据，将Apriori算法应用在获取的大量微信热点事件上，分析出影响微信事件成为热点事件的原因。得出结论为微信热点事件的形成与事件主体是否知名，事件发布平台是否知名，事件是否有知名平 台转发，事件阅读数，事件转发数等因素成正相关。

（贾慧娟, 卫晨）使用Apriori算法对煤矿安全数据进行了分析，为了挖掘的方便及有效性 ，首先对数据进行预处理，把连续的数值转化为离散的等级值，再输入到顶尖数据挖掘平台（TipDM ）中的Apriori算法模型，分析了影响煤矿安全预警的因素（如瓦斯浓度 、矿压 、ＣＯ 浓度 、温度 、流量 、水 、火等）与矿井安全之间的内在关联信息。得出结论为影响煤矿安全隐患的主要因素 是瓦斯含量和日产量。同时，数据库管理人员可以利用挖掘结果对可能出现由瓦斯 、ＣＯ 浓度超限等工况信息引起的安全隐患直接监控指挥 、实时控制并及时进行处理。

以下是该研究中使用的原始数据、处理后数据以及输出结果的部分示例。



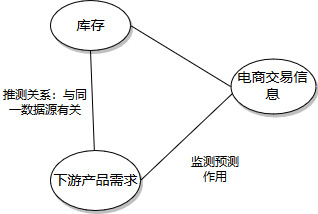




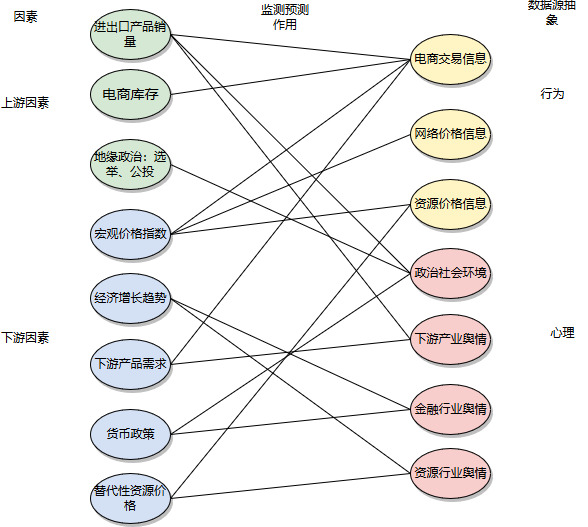
可以从中得出的关联规则如瓦斯浓度高 → 低流量的置信度约为 ７０％ ，瓦斯浓度高 → 低负压的置信度约为 ６１％等，可见通过Apriori算法建立了因素之间，因素与危险等级之间的关系。该研究对于进行资源能源安全因素关系挖掘和风险评定都有很好的借鉴意义。

### 7.1.4基于因素与数据源关系的网络构建方法

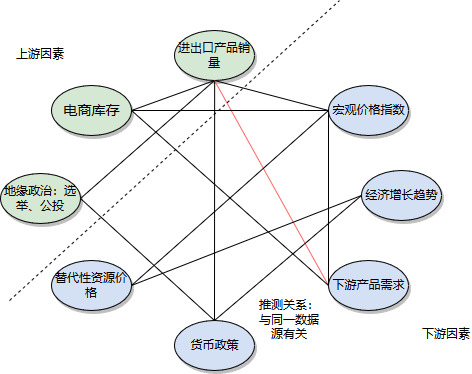
由于使用大数据进行资源能源安全风险因素的监测预测时，往往会出现不同的因素对应同样的，或是具有同样含义的数据源。例如有研究表明，产品库存的预测可以采用淘宝网电商数据（商品历史库存价格、购买量、收藏量、访问量等），同时需求的预测也可以采用电商数据，即库存和需求都可以使用电商数据进行监测预测。正因为因素与数据源之间有相关关系，数据源才能作为监测预测的信息源，而两因素与同一数据源有相关关系，可以推断出这两因素间也存在某种相关关系。可以使用这种“与同一数据源相关”的关系搭建出因素的相关性网络。如下图所示。



下图为选取了部分上下游相关因素构建的与数据源之间的关系图。展示因素与数据源抽象之间的关系：监测预测作用。



下图展示了上述因素之间的关系边的来源：两个因素具有推测关系：与同一数据源有关。边的权重：两因素共同相关的数据源个数。红色边表示权重为2，其余边权重为1。



以上例子的构建针对可监控因素（四级因素）可以容易地扩展到针对三级因素：只要其包含的四级因素中有一个与某一数据源有关，则建立三级因素到该数据源的边，边的权重为所有包含的四级因素边的权重加和。针对二级因素，一级因素也可以按照同样方式构建。

这种方法的一个明显的优势在于，可以充分利用使用大数据进行因素监测预测的模型和数据，直接在此基础上推测关系，而不需要建立额外的模型，再搜集历史数据去分析因素间的相关性。

## 7.2知识驱动的关系网络构建方法

知识驱动的方法从知识的视角构建关系网络。其网络表示的关系是基于人类知识的语义、推理等规律性认识。其中基于经验和规则的方法是直接将领域专家的知识加入到网络中，表现出因素之间在业务逻辑上的关系。而基于知识图谱的方法则是自动挖掘出互联网文本中的因素间语义关系，即不断挖掘新的知识加入网络，更具智能性。

### 7.2.1基于知识图谱的关系网络构建方法

知识图谱本质上是一种语义网络，其结点代表实体(entity)或者概念(concept)，边代表实体/概念之间的各种语义关系。资源能源安全因素关系网络可以看作一个特定领域的知识图谱。其中的点代表资源能源安全因素，边代表因素之间的语义关系，边的权值可以为因素间关系的关联强度。因而可将关系网络的构建问题转化为知识图谱的构建问题。这样既可以完成对给定的因素建立关系网络，也可以完成新因素的自动扩充，关系网络的不断更新迭代。

**1）通用知识图谱**

知识图谱的概念是由谷歌公司提出的。2012年 5月17日，谷歌发布知识图谱项目Knowledge Graph ，并宣布以此为基础构建下一代智能化搜索引擎。其目的是为了让用户能够更快更简单的发现新的信息和知识，将搜索结果进行知识系统化，任何一个关键词都能获得完整的知识体系。该项目始于2010 年谷歌收购Metaweb公司，并籍此获得了该公司的语义搜索核心技术，其中的关键技术包括从互联网的网页中抽取出实体及其属性信息，以及实体间的关系。

社交网站 Facebook 于 2013 年推出了 Graph Search[126]产品，其核心技术就是通过知识图谱将人、 地点、事情等联系在一起，并以直观的方式支持精 确的自然语言查询，例如输入查询式：“我朋友喜欢的餐厅”“住在纽约并且喜欢篮球和中国电影的朋友”等，知识图谱会帮助用户在庞大的社交网络中找到与自己最具相关性的人、照片、地点和兴趣等[7]。 Graph Search提供的上述服务贴近个人的生活，满足了用户发现知识以及寻找最具相关性的人的需求。（知识图谱技术综述）

微软的Bing Search[116]和Twitter[118]等大型社交服务站点达成了合作协议，在用户个性化内容的搜集、 定制化方面具有显著的优势。

中文知识图谱目前也有不少相关研究。早期的中文知识库主要采用人工编辑的方式进行构建，例如中国科学院计算机语言信息中心董振东领导的知网(HowNet) 项目。由于中文知识图谱的构建对中文信息处理和检索具有重要的研究和应用价值，近年来吸引了大量的研究。例如在业界，出现了百度知心、搜狗知立方等商业应用。其中，搜狗的知立方[119]是国内搜索引擎行业的第一款知识图谱产品，它通过整合互联网上的碎片化语义信息，对用户的搜索进行逻辑推荐与计算，并将最核心的知识反馈给用户。百度将知识图谱命名为知心[120]，主要致力于构建一个庞大的通用型知识 网络，以图文并茂的形式展现知识的方方面面[7]。图为搜狗知立方在实际搜索中的应用，扩展用户搜索结果，发现更多内容。



在学术界，清华大学建 成了第1个大规模中英文跨语言知识图谱XLore④、 中国科学院计算技术研究所基于开放知识网络 (OpenKN)建立了“人立方、事立方、知立方”原型系统。中国科学院数学与系统科学研究院陆汝钤院士 提出知件(Knowware)的概念、上海交通大学构建并发布了中文知识图谱研究平台zhishi．me④、复旦 大学GDM实验室。推出的中文知识图谱项目等[2一， 这些项目的特点是知识库规模较大，涵盖的知识领 域较广泛，并且能为用户提供一定的智能搜索及问答服务。

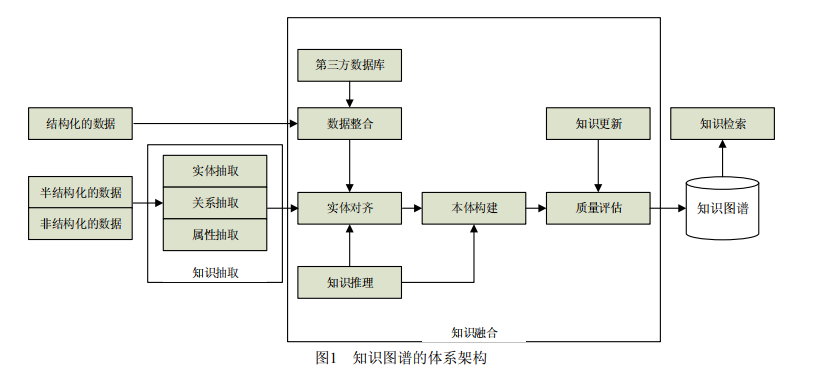
**2）垂直行业知识图谱应用**

行业知识图谱也称垂直知识图谱，其知识来源针对某些特定行业，如医疗行业、金融行业等。垂直知识图谱的构建和应用也有广泛的研究。

例如在金融行业中，反欺诈是一个重要的环节。它的难点在于如何将不同税务子系统中的数据整合在一起。通过知识图谱，一方面有利于组织相关的知识碎片，通过深入的语义分析与推理，可对信息内容的一致性充分验证，从而识别或提前发现欺诈行为；另一方面，知识图谱本身就是一种基于图结构的关系网络，基于这种图结构能够帮助人们更有效 地分析复杂税务关系中存在的潜在风险[127]。知识图谱还可以在精准营销方面，通过链接的多个数据源，形成对用户或用户群体的完整知识体系描述，从而更好地去认识用户的行为。

在医疗行业，耶鲁大学拥有全球最大的神经科学数据库 Senselab[128]，然而，脑科学研究还需要综合从微观分子层面一直到宏观行为层面的各个层次的知识。 因此，耶鲁大学的脑计划研究人员将不同层次的与脑研究相关的数据进行检索、比较、分析、整合、 建模、仿真，绘制出了描述脑结构的神经网络图谱， 从而解决了当前神经科学所面临的海量数据问题，从微观基因到宏观行为，从多个层次上加深了人类 对大脑的理解，达到了“认识大脑、保护大脑、创 造大脑”的目标。

**3）知识图谱的构建**



知识图谱的构建有两种方式，其一是自顶向下法：将百科中已有的结构化数据整理，直接加入知识库。其二是自底向上法：从公开采集的数据中自动抽取知识模式加入。自顶向下法在知识图谱的早期应用中被广泛采用，而自底向上法则是目前的研究重点以及未来发展的趋势。下面主要介绍自底向上法的构建过程。

* 信息抽取

从各种类型的数据源中提取出实体(概念)、属性以及实体间的相互关系，形成本体化的知识表达。信息抽取的主要技术方法是实体抽取、关系抽取、属性抽取。

* 知识融合

在获得新知识之后，需要对其进行整合，以消除矛盾和歧义，比如某些实体可能有多种表达，某个特定称谓也许对应于多个不同的实体等。知识融合中涉及的主要技术方法是实体链接（实体消歧、共指消歧）、知识合并（外部数据库、关系数据库）等。

* 知识加工

对于经过融合的新知识，需要经过质量评估之后(部分需要人工参与甄别)，才能将合格的部分加入到知识库中，以确保知识库的质量。新增数据之后，可以进行知识推理、拓展现有知识、得到新知识。知识加工中涉及的主要技术方法有本体构建（实体并列关系相似度计算、实体上下位关系抽取以及本体的生成）、质量评估（模型评估、综合评估）等。

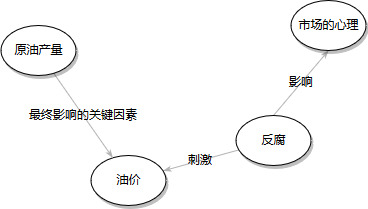
在知识加工环节中，知识推理是一个备受关注的研究领域。知识推理是指从知识库中已有的实体关系数据出发，经过计算机推理，建立实体问的新关联，从而拓展和丰富知识网络。通过知识推理，能够自动地从现有知识中发现新的知识，更加具有人类智能。如已知(乾隆，父亲，雍正)和(雍正，父亲，康熙)，可以得到(乾隆，祖父，康熙)或(康熙，孙子，乾隆）。再如（虎，科，猫科)和(猫科，目，食肉目)，可以推出(老虎，目，食肉目)。知识推理的相关方法可以分为基于逻辑的推理，包括一阶谓词逻辑、描述逻辑、基于规则的推理。以及基于图的推理，包括基于神经网络模型或Path Ranking算法等。如Socher等人将知识库中的实体表达为词向量的形式，进而采用神经张量网络模型进行关系推理。

### 7.2.2知识图谱在资源能源安全因素关系网络中的应用

可以用下面这句话展示基于知识图谱的关系网络构建过程。

*“不过，虽然沙特的反腐刺激了油价，但原油的产量才是最终影响油价的关键因素，随着反腐行动的平息，其对市场的心理影响也将减弱，油价短期存在冲高回落的可能。”*

从中可以抽取出实体：反腐行动、油价、原油产量、市场心理。其中存在的语义关系对包括（反腐行动、刺激、油价）、（原油产量、最终影响的关键因素、油价）、（反腐行动、影响、市场心理）。根据这句话可以画出如下的关系图谱。



总体而言，基于知识图谱的方法将资源能源安全因素关系网络看作一个特定领域的知识图谱。采用特定领域的知识图谱构建方法，技术包括文本挖掘、自然语言处理、知识推理等。采用的数据源是与资源能源安全相关的互联网、以及传统文本。主要的架构是从数据源中使用实体抽取方法挖掘新的影响因素，使用关系抽取、属性抽取方法挖掘文本中的因素间的语义关系，再对挖掘出的实体和关系进行融合、消除矛盾和歧义，进行质量评估，将达到质量要求的实体和关系加入到知识图谱中。还可以进一步进行知识推理实现语义关系即知识的自动扩展。

## 参考文献

1. Asur S, Huberman B A. Predicting the Future with Social Media[J]. Social Science Electronic Publishing, 2010, 7(2):492-499.
2. Bagnoli M,Beneish M D,Watts S G. Whisper Forecasts of Quarterly Earnings per Share [ J]. Journal of Accounting and Economics,1999,28 (1):27-50.
3. Barreira N, Godinho P, Melo P. Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends[J]. Netnomics, 2014, 14(3):129-165.
4. Bermingham A, Smeaton A F. On Using Twitter to Monitor Political Sentiment and Predict Election Results[J]. 2011.
5. Bertsimas D, Kallus N, Hussain A. Inventory Management in the Era of Big Data[J]. Production & Operations Management, 2016, 25.
6. Bogomolov A, Lepri B, Staiano J, et al. Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data[C]// International Conference on Multimodal Interaction. ACM, 2014:427-434.
7. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1):1-8.
8. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 1, 1–8.
9. Bruns A, Highfield T, Burgess J. The Arab Spring and Social Media Audiences[J]. American Behavioral Scientist, 2013, 57:871-898.
10. Bughin J. Google searches and twitter mood: nowcasting telecom sales performance[J]. Netnomics Economic Research & Electronic Networking, 2015, 16(1-2):87-105.
11. Cavallo A, Rigobon R. The Billion Prices Project: Using Online Prices for Measurement and Research[J]. Journal of Economic Perspectives, 2016, 30(2):151-178.
12. Chevalier J A, Mayzlin D. The Effect of Word of Mouth on Sales: Online Book Reviews.[J]. Social Science Electronic Publishing, 2006, 43(3):345-354.
13. Chevalier, J., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of Marketing Research, XLIII, 345–54.
14. Choi, H., & Varian, H. (2012). Predicting the present with google trends. Economic Record, 88(S1), 2–9.
15. Choi,H.,&Varian,H.(2009). Predicting Initial Claims for Unemployment Benefits.TechnicalReport,Google.http://research.google.com/archive/papers/initialclaim
16. Chong O, Sheng O. Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement[C]// International Conference on Information Systems, Icis 2011, Shanghai, China, December. DBLP, 2011.
17. D’Amuri,F., & Marcucci, J. (2009). Google it! Forecasting the US unemployment rate with a Google job search index. MPRA paper, University Library of Munich, Germany.
18. Das S R, Chen M Y. Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web[J]. Management Science, 2007, 53(9):1375-1388.
19. Das S R,Chen M Y. Yahoo! for Amazon:Sentiment Extractionfrom Small Talk on the Web [J]. Management Science,2007,53(9):1375-1388
20. Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction[C]// International Conference on Artificial Intelligence. AAAI Press, 2015:2327-2333
21. Ding X, Zhang Y, Liu T, et al. Deep learning for event-driven stock prediction[C]// International Conference on Artificial Intelligence. AAAI Press, 2015:2327-2333
22. Ettredge M, Gerdes J, Karuga G. Using web-based search data to predict macroeconomic statistics[J]. Communications of the Acm, 2005, 48(11):87--92.
23. Fadiya,S.O.,Saydam,S.,&Zira, V. V. (2014). Advancing big data for humanitarian needs. Procedia Engineering, 78, 88-95.
24. Fondeur Y, KaramÉ F. Can Google data help predict French youth unemployment? ☆[J]. Economic Modelling, 2012, 30(1):117–125.
25. Funke M, Mehrotra A, Yu H. Tracking Chinese CPI inflation in real time[J]. Empirical Economics, 2015, 48(4):1619-1641.
26. Goel S, Hofman J M, Lahaie S, et al. What can search predict[J]. 2010.
27. Grolinger K, Capretz M A M, Seewald L. Energy Consumption Prediction with Big Data: Balancing Prediction Accuracy and Computational Resources[C]// IEEE International Congress on Big Data. IEEE, 2016.
28. Grolinger K, L’Heureux A, Capretz M A M, et al. Energy Forecasting for Event Venues: Big Data and Prediction Accuracy[J]. Energy & Buildings, 2016, 112:222-233.
29. http://finance.jrj.com.cn/2014/07/07152017555905.shtml
30. http://index.1688.com/
31. http://news.cntv.cn/20120711/114745.shtml
32. http://www.gnnt.com.cn/web/index.html
33. http://www.thebillionpricesproject.com/
34. Jacques Bughin, Google searches and twitter mood: nowcasting telecom sales performance
35. Lee W, On B W, Lee I, et al. A big data management system for energy consumption prediction models[C]// Ninth International Conference on Digital Information Management. IEEE, 2014:156-161.
36. Li J, Xu Z, Xu H, et al. Forecasting Oil Price Trends with Sentiment of Online News Articles[J]. Procedia Computer Science, 2016, 91:1081-1087.
37. Li L, Leng S, Yang J, et al. Stock Market Autoregressive Dynamics: A Multinational Comparative Study with Quantile Regression[J]. Mathematical Problems in Engineering, 2016: 1-15.
38. Li X, Ma J, Wang S, et al. How does Google search affect trader positions and crude oil prices?[J]. Economic Modelling, 2015, 49:162-171.
39. Lin H, Liu X, Wang X, et al. A fuzzy inference and big data analysis algorithm for the prediction of forest fire based on rechargeable wireless sensor networks[J]. Sustainable Computing Informatics & Systems, 2017.
40. Liu T, Xu X, Fan F. Forecasting Chinese GDP using Online Data[J]. 2016.
41. Lui C, Metaxas P T, Mustafaraj E. ON THE PREDICTABILITY OF THE U.S. ELECTIONS THROUGH SEARCH VOLUME ACTIVITY[J]. 2012.
42. Macdonald R, Mao X. An Alternative way of Predicting the Outcome of the Scottish Independence Referendum: The Information in the Ether[J]. Sire Discussion Papers, 2015.
43. Metaxas P T, Mustafaraj E, Gayo-Avello D. How (Not) to Predict Elections[C]// IEEE Third International Conference on Privacy, Security, Risk and Trust. IEEE, 2011:165-171.
44. Nayak R K, Mishra D, Rath A K, et al. A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices[J]. soft computing, 2015:670-680.
45. O ' connor B,Balasubramanyan R,Routledge B R, et al. From Tweets to Polls:Linking Text Sentiment to Public Opinion Time Series [C]. Proceedings of the the International AAAI Conference on Weblogs and Social Media,2010
46. Oh C,Sheng O R L. Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price DirectionalMovement [C]. Proceedings of the Thirty Second InternationalConference on Information Systems,2011.
47. Ozozen A, Kayakutlu G, Ketterer M, et al. A combined seasonal ARIMA and ANN model for improved results in electricity spot price forecasting: Case study in Turkey[C]// Portland International Conference on Management of Engineering and Technology. IEEE, 2017.
48. Peng Jiang, Yadong Zhu, Yi Zhang, Quan Yuan, Life-stage Prediction for Product Recommendation in E-commerce. To appear in Proceedings of the 21th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2015.
49. Preis T,Moat H S,Stanley H E. Quantifying Trading Behavior in Financial Markets Using Google Trends [ J]. Scientific reports,2013,3 (1684):1-6.
50. Rajasekaran T, Sruthi J, Revathi S, et al. FOREST FIRE PREDICTION AND ALERT SYSTEM USING BIG DATA TECHNOLOGY[C]// Iciems. 2015.
51. Suhoy, T. (2009). Query indices and a 2008 downturn: Israeli data. Technical report, Bank of Israel, 2009-4.
52. sUS.pdf.Accessed7July 2013.
53. ta to predict macroeconomics statistics. Communications of the ACM, 48(11), 87–92.
54. Tosi D, Marzorati S, Rosa M L, et al. Big data from cellular networks: How to estimate energy demand at real-time[C]// IEEE International Conference on Data Science and Advanced Analytics. IEEE, 2015.
55. Tumarkin R,Whitelaw R F. News or Noise? Internet Postings and Stock Prices [J]. Financial Analysts Journal,2001:41-51.
56. Vosen S, Schmidt T. Forecasting private consumption: survey‐based indicators vs. Google trends[J]. Journal of Forecasting, 2011, 30(6):565–578.
57. Vosen, S., & Schmidt, T. (2012). A monthly consumption indicator for Germany based on internet search query data. Applied Economics Letters, 19(7), 683–687
58. Wang K, Xu C, Zhang Y, et al. Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid[J]. IEEE Transactions on Big Data, 2017, PP(99):1-1.
59. Wang X, Gerber M S, Brown D E. Automatic Crime Prediction Using Events Extracted from Twitter Posts[M]// Social Computing, Behavioral - Cultural Modeling and Prediction. Springer Berlin Heidelberg, 2012:231-238.
60. Wu F, Chen C, Guo X, et al. Prediction of Crop Yield Using Big Data[C]// International Symposium on Computational Intelligence and Design. IEEE, 2016:255-260.
61. Wu L, Brynjolfsson E. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales[J]. Social Science Electronic Publishing, 2014.
62. Xie H, Bian J, Wang M, et al. IS TECHNICAL ANALYSIS INFORMATIVE IN UK STOCK MARKET? EVIDENCE FROM DECOMPOSITION-BASED VECTOR AUTOREGRESSIVE (DVAR) MODEL[J]. Journal of Systems Science & Complexity, 2014, 27(1): 144-156.
63. Xu W, Li Z, Chen Q（2012） A Neural Network Based Forecasting Method For the Unemployment Rate Prediction Using the Search Engine Query Data
64. Zhang Z. Study on inventory control model based on the B2C mode in big data environment[C]// International Conference on Modern Economic Technology and Management. 2017:53-58.
65. Zhou S, Shi X, Sun Y, et al. Stock Market Prediction Using Heat of Related Keywords on Micro Blog[J]. Journal of Software Engineering & Applications, 2015, 06(3):37-41.
66. Zhou S,Shi X,Sun Y,et al. Stock Market Prediction Using Heatof Related Keywords on Micro
67. 陈龙, 王建冬, 窦悦. 基于互联网大数据的宏观经济监测预测研究:理论与方法[J]. 电子政务, 2016(1):18-25.
68. 陈涛, 林杰. 基于搜索引擎关注度的网络舆情时空演化比较分析——以谷歌趋势和百度指数比较为例[J]. 情报杂志, 2013(3):7-10.
69. 董倩、孙娜娜、李伟.基于网络搜索数据的房地产价格预测[J]》.统计研究,2014(10):81-82
70. 李士琦. 能源互联网大数据分析技术的研究[J]. 数码世界, 2017(6):100-100.
71. 李新. 基于海量数据的投资者关注与国际原油价格研究[D]. 中国科学院大学, 2015.
72. 林振兴. 网络讨论、投资者情绪与IPO抑价[J]. 山西财经大学学报, 2011(2):23-29.
73. 刘恩猛, 汪波. 天津市经济增长和贷款关系的实证分析[J]. 北方经济, 2007(22):66-67.
74. 刘汉、刘金全.中国宏观经济总量的实时预报与短期预测——基于混频数据预测模型的实证研究[J].经济研究.2011(3):4-6
75. 刘慧.“克强指数”与经济增长的动态关系研究: 基于 VAＲ 和 VEC 模型的实证分析[J].商业时代.2014(1):11-13.
76. 刘涛雄、徐晓飞，《大数据与宏观经济分析研究综述》载于《国外理论动态》2015年第1期，第61页
77. 刘涛雄和徐晓飞(2015) 碳价预测
78. 柳长兴, 秦琴. 经济高速增长与社会转型背景下的重庆市社会稳定指标体系研究[J]. 重庆邮电大学学报(社会科学版), 2011, 23(2):96-104.
79. 彭赓, 苏亚军, 李娜. 失业率预测研究--基于网络搜索数据及改进的逐步回归模型[J]. 现代管理科学, 2013(12).
80. 上海《社会稳定指标体系》课题组. 上海社会稳定指标体系纲要[J]. 社会, 2002(12):8-11
81. 申红艳, 吴晨生, 扆铁梅,等. 大数据时代宏观经济分析面临的机遇与挑战[J]. 经济研究参考, 2014(63):19-25.
82. 宋双杰,曹 晖,杨 坤. 投资者关注与 IPO 异象———来自网络搜索量的经验证据 [J]. 经济研究,2011(增 1):145-155.
83. 苏志, 张骐. 基于网络大数据的社会经济预测综述[J]. 情报杂志, 2015(7):18-21.
84. 王健, 宋述停, 兰俊美,等. 电力需求侧大数据应用模型的建立[J]. 电力与能源, 2014, 35(3):283-286.
85. 王娜. 基于大数据的碳价预测[J]. 统计研究, 2016, 33(11):56-62.
86. 王晓琳. 金融投资者行为及其对股票与期货市场的影响研究[D]. 哈尔滨工业大学, 2016.
87. 王雪蓉, 万年红. 基于跨境电商可控关联性大数据的出口产品销量动态预测模型[J]. 计算机应用, 2017, 37(4):1038-1043.
88. 魏韡, 向阳, 陈千. 中文文本情感分析综述[J]. 计算机应用, 2011, 31(12):3321-3323.
89. 夏帆. 浅议社会稳定指标体系的构建[J]. 湖北经济学院学报, 2005, 3(2):60-63.
90. 销量预测研究 [J]. 管理学家:学术版,2011(1):12-24.
91. 熊浩、孙有望，《我国货运量与经济增长关系的协整分析》，载于《系统工程》2010 年第 9 期，第 46-50 页.
92. 徐振敬. 基于情感分析的国际原油价格走势预测研究[D]. 北京化工大学, 2016.
93. 杨莎, 余伟, 李石君,等. 基于Web大数据挖掘的证券价格波动实时影响研究[J]. 计算机科学, 2015, 42(04):166-171.
94. 俞庆进,张 兵. 投资者有限关注与股票收益———以百度指数作为关注度的一项实证研究 [ J]. 金融研究,2012 (8):152 -165
95. 袁庆玉,彭 赓,刘 颖,等. 基于网络关键词搜索数据的汽车
96. 岳艳涛, 章雅婷, 张宇,等. 基于大数据分析的橡胶期货交易策略研究[C]// 2015.
97. 张崇, 吕本富, 彭赓,等. 网络搜索数据与CPI的相关性研究[J]. 管理科学学报, 2012, 15(7):50-59.
98. 张铁军. 一种基于新闻大数据的自动选股方法:, CN107025264A[P]. 2017.
99. 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8):1834-1848.