## 2.资源能源与相关产业产品的需求、价格

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。

2.1相关产业产品需求的监测预测

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。

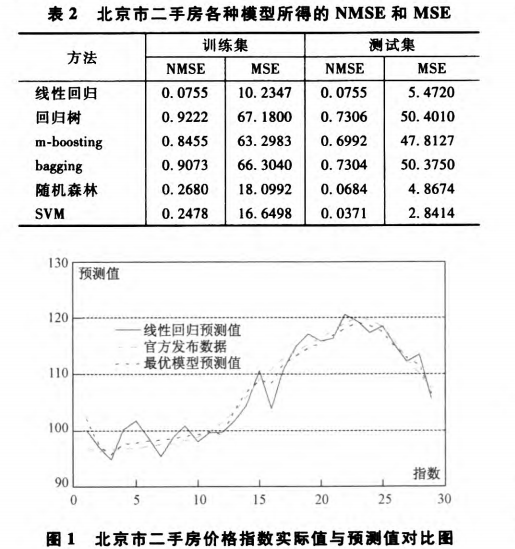
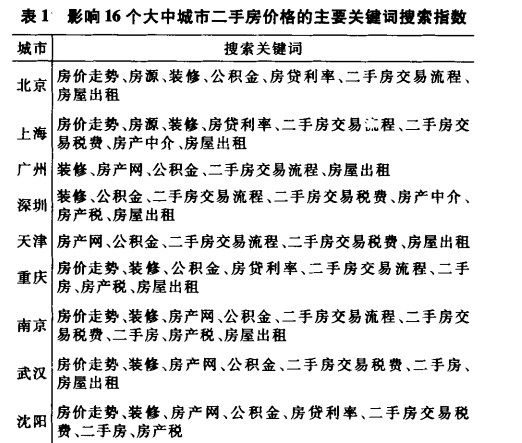
相关产业产品的需求预测也可以直接结合互联网搜索大数据，如王文亮（2012）讨论了现有的网络关键词搜索、电子商务交易量两方面相关的研究，给出了一种通过网络关键词对电子商务交易量的预测方法。（其研究的电子商务交易量与需求含义类似）首先，建立了理论分析框架，揭示了网络搜索数据与电子商务交易量之间存在的相关关系及先行滞后关系。其次通过逐步合成方法合成得到进入回归模型的关键词合成指数。并实证分析中国某电子商务网站的搜索数据和交易量与相应合成指数之间的关系。试验表明电子商务网站的搜索数据与交易量之间具有较高的相关性，加入合成的搜索指数后，模型拟合度达到0.901，模型的预测效果明显提高。通过对电子商务网站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

2.2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。

2.3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微博，关键字为电影标题中所有词汇 | 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 | 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。

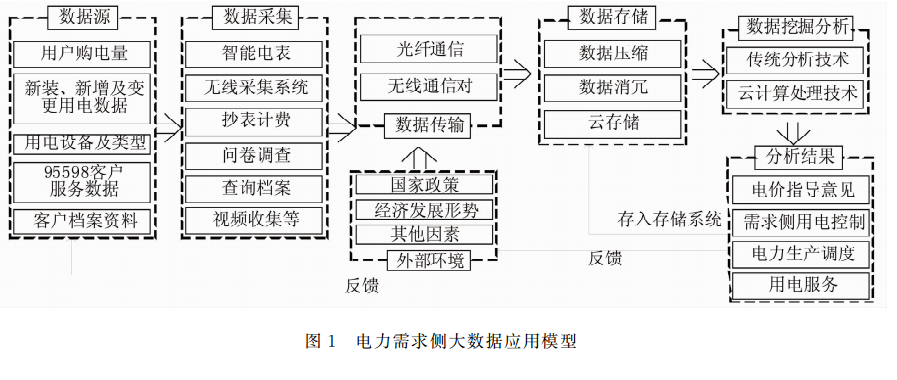
搜索大数据在销量预测中能够取得较好的作用，原因在于网络搜索数据记录了数以亿计的搜索关注与需求，反映了市场主体的行为趋势与规律，可以为研究社会经济行为提供数据基础。例如消费者在购买汽车过程中，首先会有需求准备期，在需求准备的这段时间内，消费者会根据自身的喜好、汽车的价格和性能、以及自身的经济状况来获取潜在购买相关汽车的信息。随着网络技术的快速发展和互联网的普及，消费者更加青睐与从互联网上搜索相关数据和信息，因此而留下的记录可供销量预测等研究使用。

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。

2.2.1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，海洋石油工程股份有限公司，能源互联网大数据分析技术的研究）如王健, 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.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。



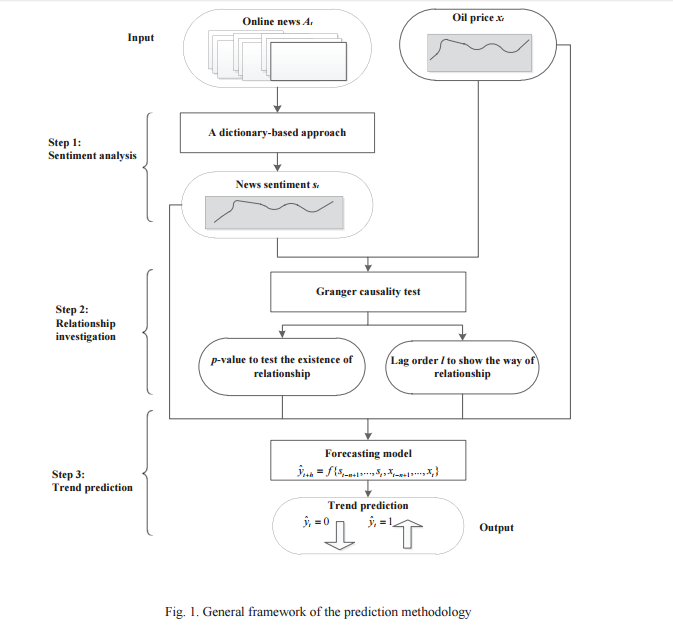
国外的研究也提供了类似的设计，并且进一步提出了基于map/reduce的线性回归等模型，以前期电力消费数据预测能源的消费与需求，测试效果显示预测与真实值十分接近，并且在速度上有优势。

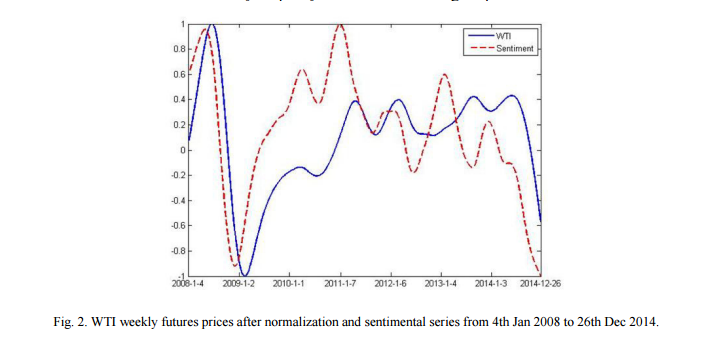
例如，Lee W, On B W, Lee I, et al（2014）设计了能够高效存储、索引、查询海量能源使用大数据的数据管理分析系统。其采用分布式框架并能够整合结构化的和非结构化的数据，且数据可以通过API快速获取。该系统还集成了基于map/reduce的使用正则项的线性回归模型，以预测能源的消费与需求。给与一个特定的时间点，模型能够预测出时间点的能源需求。模型的训练使用了2014年3个月份7百万条含有时间戳和能源消费量的数据，测试效果显示预测与真实值十分接近，而且分布式框架使得训练速度提高到1.72倍。

2.2.2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。

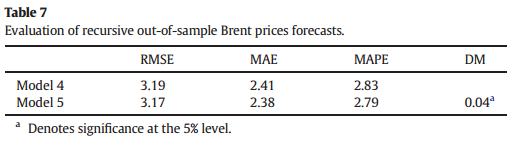
徐振敬（2016）提出了基于情感分析的国际原油价格走势预测模型。其模型主要基于石油相关新闻的分析,通过采用领域关键词词典的方法,得到新闻的情感序列,再通过格兰杰因果检验的方法,得到情感序列和石油价格序列的相关性和滞后期,最后通过机器学习的方法(支持向量机、决策树、逻辑回归和神经网络)预测石油价格的走势。同时,为了验证新闻情感对于石油价测具有预测能力,选取了美国西得克萨斯轻质原油(WTI)和路透社原油新闻作为研究对象进行案例分析。结果表明,新闻情感和原油价格之间确实存在着格兰杰因果关系,即新闻情感的变化会引起石油价格的变化,这说明新闻情感对于石油价格的走势具有预测能力。实验结果还表明,对于大部分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

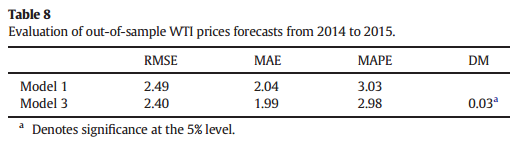




李新（2015）研究了大数据时代下，基于海量互联网数据分析及建模的理论基础、研究框架及关键模型、技术和方法。首先，对提出了基于海量数据的投资者关注对原油价格的“非对称”影响关系的马尔科夫状态转移模型。建立了投资者关注与国际原油价格收益及波动关系的EGARCH模型，建立包含投资者关注与原油价格收益的VAR模型。

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。

