第一章：资源能源安全风险的相关因素综述 20

1.资源能源安全因素概述

2.资源能源安全因素的互联网大数据来源

3.互联网大数据与传统数据的配合

## 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1.4：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 |

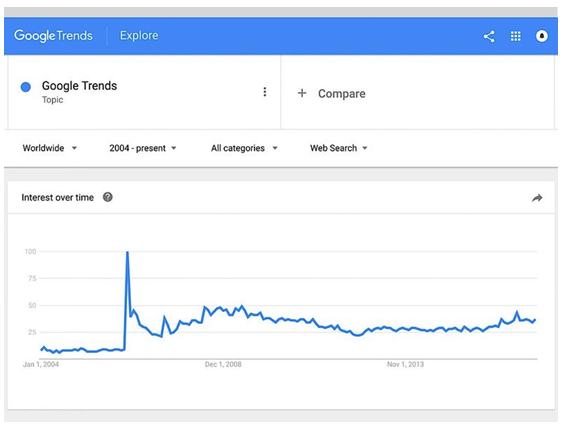
1 搜索引擎大数据

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。

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):7-10.百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[10] 。 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。

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。

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。

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。

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, 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.

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。

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。

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。

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. 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.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. Deep learning for event-driven stock prediction[C]// International Conference on Artificial Intelligence. AAAI Press, 2015:2327-2333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，如张铁军. 一种基于新闻大数据的自动选股方法:, CN107025264A[P]. 2017.使用RSS实时获取互联网财经新闻,每半小时更新一次。

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。

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. 互联网大数据与传统数据的配合

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

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