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 | 各大新闻媒体（如Thomson Readers新闻） |
| 行业相关网页数据 | 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 |

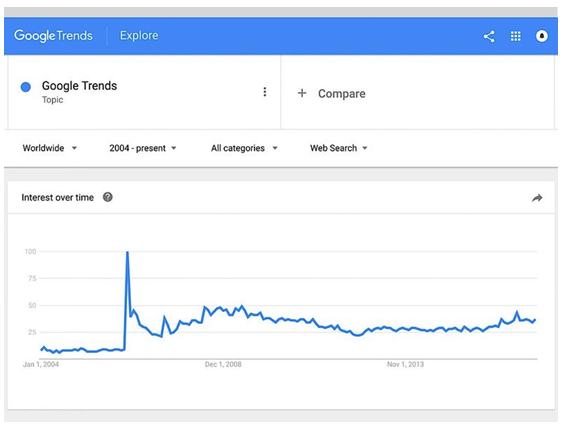
1 搜索引擎大数据

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的使用需要注册一个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。