Atitit 信息检索SurfaceWeb和DeepWeb两大部分 暗网

暗网（不可见网，隐藏网）是指那些存储在[网络](https://baike.baidu.com/item/%E7%BD%91%E7%BB%9C/1717974" \t "https://baike.baidu.com/item/%E6%9A%97%E7%BD%91/_blank)数据库里、但不能通过[超链接](https://baike.baidu.com/item/%E8%B6%85%E9%93%BE%E6%8E%A5/97857" \t "https://baike.baidu.com/item/%E6%9A%97%E7%BD%91/_blank)访问而需要通过动态网页技术访问的资源集合，不属于那些可以被标准搜索引擎索引的表面网络。

暗网是[深网](https://baike.baidu.com/item/%E6%B7%B1%E7%BD%91/22089976" \t "https://baike.baidu.com/item/%E6%9A%97%E7%BD%91/_blank)（Deep Web）的一个子集，属于深网的一小部分。

# DeepWeb的内容主要包含4个方面：(

1)通过填写表单形成对后台在线数据库的查询而得到的动态页面；(2)由于缺乏被指向的超链接而没有被搜索引擎索引到的页面，大约占整个比例的21.3%；(3)需要注册或其它限制才能访问的内容；(4)Web上可访问的非网页文件，比如图片文件、PDF和Word文档等。

而在实际中应用中，人们则更关注于DeepWeb中的第一部分内容。其原因不难理解，这部分内容对结构化数据的集成更有意义，可以采用的技术也更丰富。DeepWeb数据集成也主要是指对结构化信息的集成，我们同时把Web中可访问的在线数据库称为[Web数据库](https://baike.baidu.com/item/Web%E6%95%B0%E6%8D%AE%E5%BA%93" \t "https://baike.baidu.com/item/%E6%B7%B1%E7%BD%91/_blank)或WDB

## [Web数据库](https://baike.baidu.com/item/Web%E6%95%B0%E6%8D%AE%E5%BA%93" \t "https://baike.baidu.com/item/%E6%B7%B1%E7%BD%91/_blank)或WDB

# 深网资源

动态内容

未被链接内容

私有网站

Contextual Web

被限制访问内容

脚本化内容

非HTML/文本内容

# 如何解决法

## Sitemap协议（由Google于2005年首次开发并由Google引入

## 数据库导出sql 文本索引法

## 非html法 格式转换法 eml pdf office

## 压缩文件 索引法

## 索引法 文件标题法 everything index

## 导出txt法 qq聊天记录等。。微信聊天记录等

## 截图ocr

暗网（汉语词语、互联网术语）\_百度百科.html