网络爬虫

0. 查询煤炭及其化工产品

煤炭（动力煤、焦煤）、煤焦化（焦炭）、粗苯精制、煤焦油深加工、甲醇及下游。

1. 打开网站首页，在url后加上/robots.txt查看robots协议

例如：金投网

Robots协议链接：https://www.cngold.org/robots.txt

协议内容：

User-agent: \* · 允许所有搜索引擎访问网站

Disallow: /templets/\* · templets及其子目录内容不可抓取

Disallow: /errorpage/\* ·errorpage同上

Disallow: /special/\* · special同上

Disallow: /top/\* · top同上

Disallow: /c/2013-07-07/\* · c/2013-07-07同上

Disallow: .php?\* · /\*禁止访问网站的php文件\*/

Disallow: .asp$ · /\*禁止访问网站的asp文件\*/

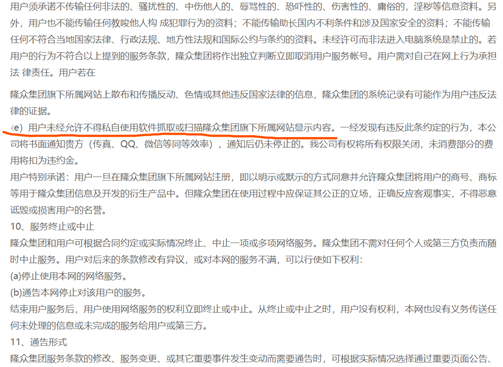
Sitemap: <http://www.cngold.org/sitemap_index.xml> · 用于告诉搜索引擎自己网站的所有网页及其重要信息。

2. 各网站爬虫可行性分析

(1) 金投网，目前可爬，爬虫请求时增加timeout，防止造成网站崩溃。

(2) 曲合期货网，目前可爬，要求同上。

(3) 隆众资讯，含有各种煤化工产品信息，但首先需要会员，其次需要联系对方获取授权。



(4) 生意社 化工产品：工业萘、活性炭 未经授权不允许爬取信息。

3. 爬虫守则---安全守法.

(0) 绝对不能爬取国家机关政务网站信息与其它敏感信息。

(1) 爬虫速度不要太快，不能给对方服务器造成太大压力。

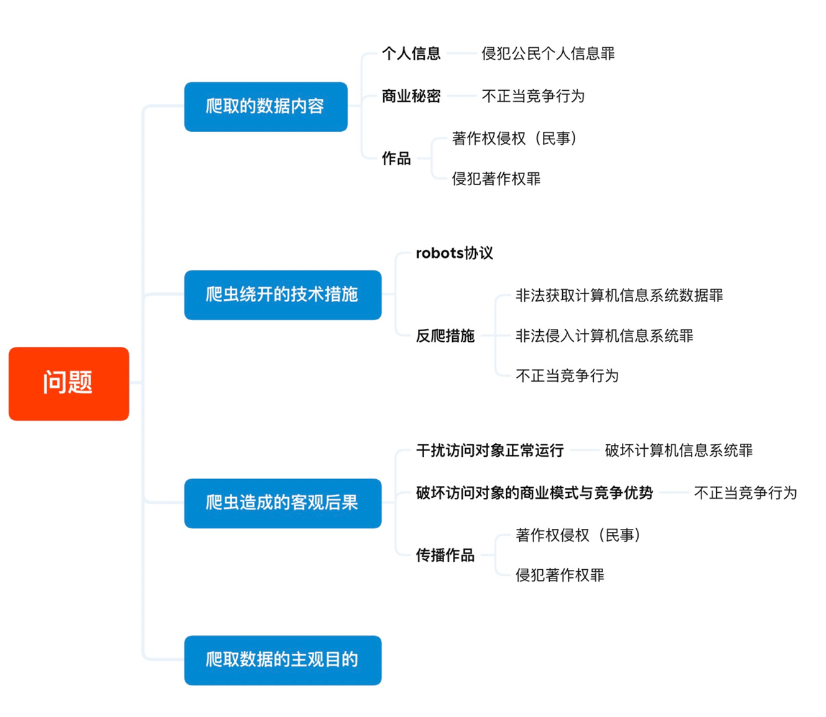
(2) 爬虫不要伪造VIP，绕过对方的身份验证。

(3) 不要爬取公民的身份信息。

(4) 爬取的数据不能用于盈利。

(5) 爬虫是模拟人，不要做人不能做到的事情。

4. 爬虫产生的问题及对应的法律规范



5. 爬虫违法违规案例汇总

<https://blog.csdn.net/qq_38759383/article/details/125194662>