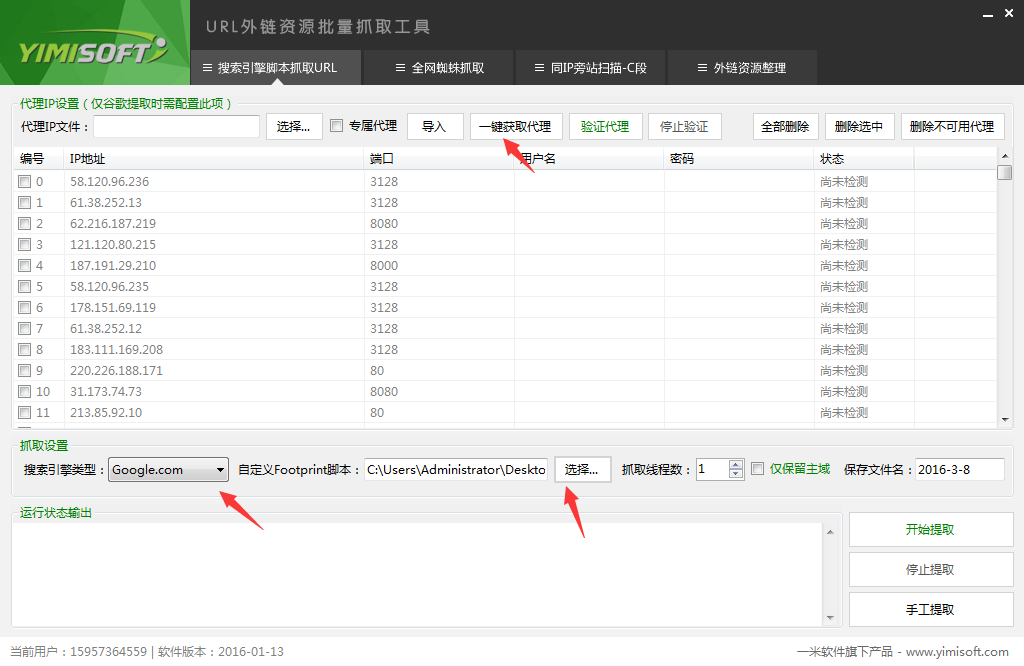
**URL外链资源批量抓取工具使用教程**

**一:搜索引擎脚本抓取URL**



**1.除了google.com抓取需要使用代理外，其他搜索引擎均不需要代理获取**

官方自带每天500条免费高匿名代理，点“一键获取代理”即可获得，如需要导入自己购买的专属代理，则选择代理IP文件，勾选“专属代理”，导入即可

专属代理的文本导入格式为：

IP:端口:用户名:密码，如：192.168.1.22:8080:testuser:pass

一行一个整理在一个TXT文本中导入

**2.选择搜索引擎类型和脚本规则**

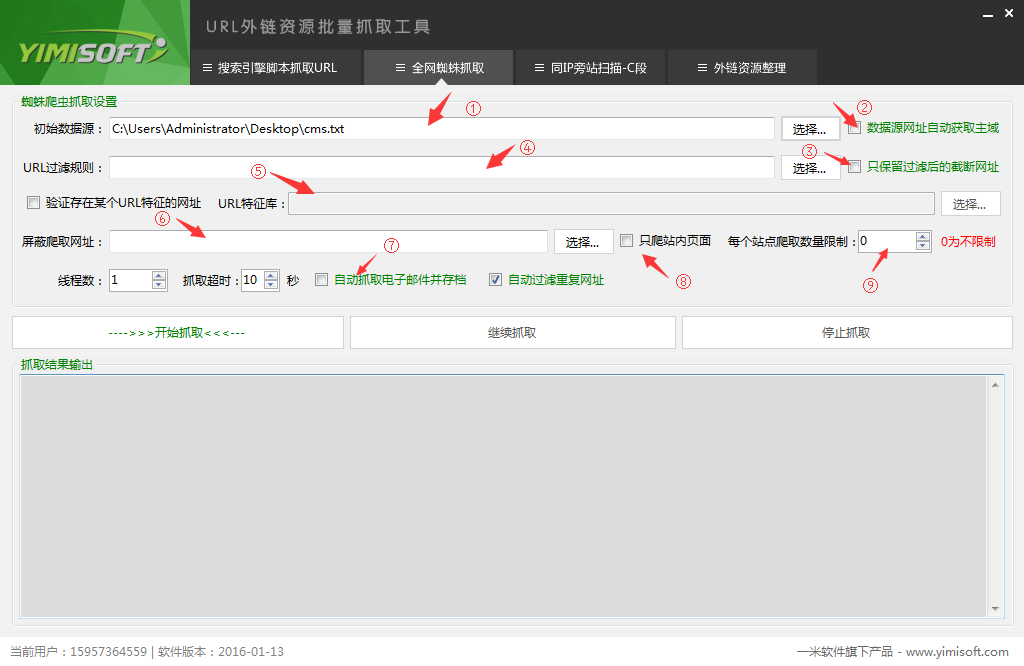
脚本规则（footprint）一般为带引号的关键词，现在谷歌直接搜索命令抓取了已经限制的很严了，这也是一种替代办法，比如"wp-content"+"/themes/sommerce" 作为一个脚本规则，可以多个一行一个存放在一个TXT文本中，通用性比较好，其他搜索引擎也都支持

**3.选择线程数**，建议5-20为佳，如果勾选“仅保留主域”，则抓取的数据为网站主域名。文件名可用默认日期形式，也可以自定义，点开始抓取即可

抓取的数据自动保存在软件目录下的“Search”文件夹下

**二：全网蜘蛛抓取**

全网蜘蛛抓取根据初始的网站数据，不断从上面抓取网页链接（包括内页和外部链接），并一层层的爬下去，无限爬取，从而分离出符合需求的网址数据（可以理解为小型的谷歌或者百度爬虫），获取的URL保存在软件的“Spider”文件夹下



1. 为初始数据源，主要是网站主域名，一行一个整理在一个TXT，也可以直接程序中勾选“数据源网址自动获取主域”选项，可以直接利用第一步获取的URL主域名作为初始数据源
2. 勾选此项可自动获取数据源网址中的主域名
3. 勾选此项为数据保存为截断后的地址
4. URL过滤规则，如果您想要过滤出符合该URL规则的，把想过滤出的url小段网址填写在TXT中一行一个，比如抓取到的网址为[http://www.xxx.com/wp-content/uploads/asf.html， 那么 /wp-content/](http://www.xxx.com/wp-content/uploads/asf.html，%20那么%20/wp-content/) 或 wp- 或 /wp-content 都符合
5. 验证存在某个URL特征的网址，特征库为具体根目录后的网址，具体到文件名，比如抓取到的网址为[http://www.xxx.com/wp-content/uploads/asf.html，验证网址为 /license.txt](http://www.xxx.com/wp-content/uploads/asf.html，验证网址为%20/license.txt)

那么程序会自动去验证<http://www.xxx.com/license.txt> 这个网址存在不存在，如果存在则保留

⑥ 屏蔽爬取网址，不想爬取的网址比如某些大站如百度，youku，sina等等，软件会自动跳过，提高爬取效率。一行一个整理在一个TXT中。

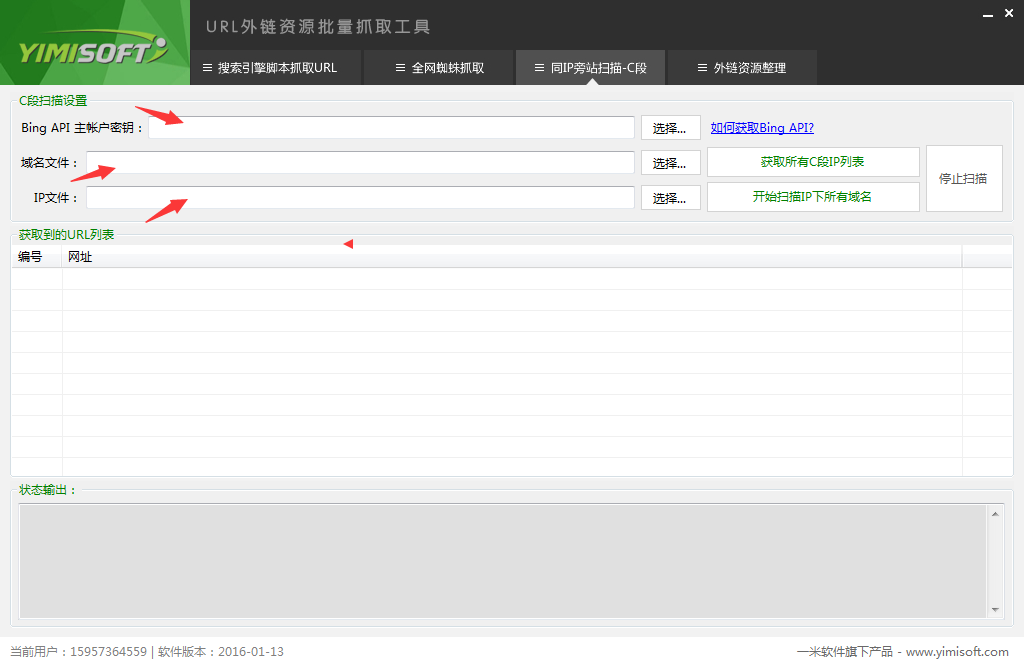
⑦ 自动提取电子邮件并存档，网址中存在的邮件地址都会保存下来，保存在软件目录下的“E-mail”文件夹下

⑧ 只爬站内页面，勾选此项为只爬网址库中每个网站内的页面

⑨ 每个站只爬取数量限制，如果只想获取每个站的主域名，这里可以填写的少点，这样可以每个站爬到设定数量后即跳过爬取下一个站点，提高执行效率，如果写0则不限制爬取，爬完一个站再跳到下一个站点

**三：同IP旁站扫描-C段**

我们知道，同一个服务器下或者同一个IP下可能存在很多同类型的站点，或者用同一个类型的cms做的站点，尤其是国外很多服务器提供商都默认自带可以一键安装wp或者joomla等开源cms，那么可以使用此功能全部扫描出来，利用bing的api接口，是目前最新最全的扫描工具



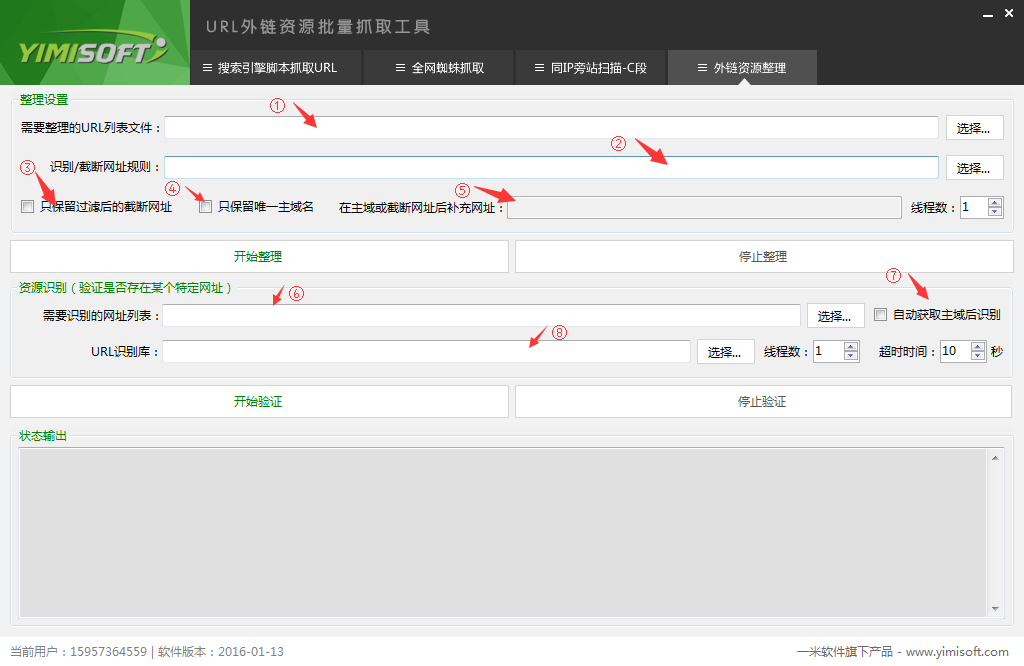
Bing api主账户秘钥申请是免费的，具体申请方法请看网站介绍，这里就不重复叙述

主账户秘钥一行一个保存在一个TXT中调用，程序会轮换检测过去，一个账户秘钥每月限额是5000次

域名文件为需要检测C端旁站的域名文件，一行一个域名存放在一个TXT中，可以为之前两步获取的主域名

点击“获取所有C段IP列表”按钮程序会自动ping出所有域名的IP地址并保存在软件目录的“C-scan”文件夹下，并自动显示路径在IP文件 文本框中，点“开始扫描IP下所有域名”按钮执行C段扫描

**四：外链资源整理**



1. 为需要整理的URL网址，一个TXT中一行一个
2. 为识别/截断的网址规则，如网址为 <http://www.xxx.com/wp-content/uploads/asd.html,网址规则为/wp-content/> 则符合规则，该网址将保留
3. 勾选此项为保留过滤后的截断网址，如②中所举例子则将只保留 <http://www.xxx.com>
4. 勾选此项为只保留唯一主域名
5. 在主域或截断网址后补充网址，填写此项为当勾选③或④时，将在这些网址后再增加填写的URL追加在尾部

**整理的URL自动保存在软件目录的“Checkurl”文件夹下**

1. 为需要识别的网址列表，一行一个保存在一个TXT中
2. 自动获取主域后识别，勾选后将返回所有网址的主域名
3. URL识别库，格式为：URL|识别字符 一行一个存放在一个TXT中，验证识别某个具体网址，比如 /test.html 那么程序会在每个主域后加上这个URL进行识别如 <http://www.xxx.com/test.html>

**整理的URL如果存在则自动保存在软件目录的“Exitsurl”文件夹下，不存在的在“Failurl”文件夹下**