**configs详解——之成员**

爬虫的整体框架是这样：首先定义了一个$configs数组, 里面设置了待爬网站的一些信息, 然后通过调用$spider = new phpspider($configs);和$spider->start();来配置并启动爬虫.

$configs数组中可以定义下面这些元素

**name**

定义当前爬虫名称

**String类型 可选设置**

举个栗子:

'name' => '糗事百科'

**log\_show**

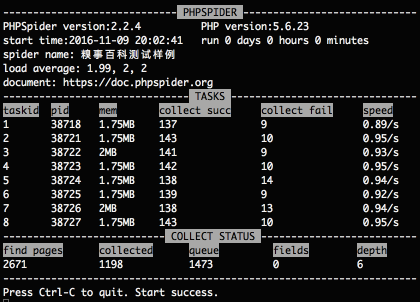
是否显示日志  
为true时显示调试信息  
为false时显示爬取面板

**布尔类型 可选设置**

**log\_show默认值为false，即显示爬取面板**

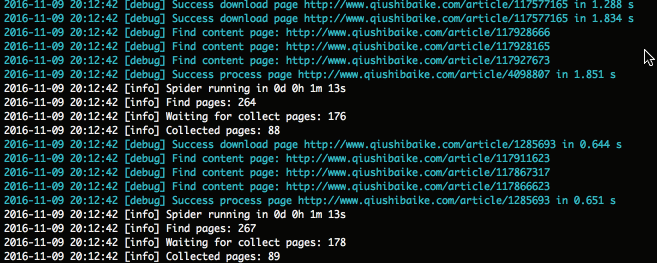
栗子1:

'log\_show' => false



栗子2:

'log\_show' => true



**注意：显示爬取面板时，也可以通过tail命令来查看日志**

tail -f data/phpspider.log

**log\_file**

日志文件路径

**String类型 可选设置**

**log\_file默认路径为data/phpspider.log**

举个栗子:

'log\_file' => data/qiushibaike.log

**log\_type**

显示和记录的日志类型  
普通类型: info  
警告类型: warn  
调试类型: debug  
错误类型: error

**String类型 可选设置**

**log\_type默认值为空，即显示和记录所有日志类型**

栗子1:  
显示错误日志

'log\_type' => 'error'

举个栗子:  
显示错误和调试日志

'log\_type' => 'error,debug'

**input\_encoding**

输入编码  
明确指定输入的页面编码格式(UTF-8,GB2312,…..)，防止出现乱码,如果设置null则自动识别

**String类型 可选设置**

**input\_encoding默认值为null，即程序自动识别页面编码**

举个栗子:

'input\_encoding' => 'GB2312'

**output\_encoding**

输出编码  
明确指定输出的编码格式(UTF-8,GB2312,…..)，防止出现乱码,如果设置null则为utf-8

**String类型 可选设置**

**output\_encoding默认值为utf-8, 如果数据库为gbk编码，请修改为gb2312**

举个栗子:

'output\_encoding' => 'GB2312'

**tasknum**

同时工作的爬虫任务数  
需要配合redis保存采集任务数据，供进程间共享使用

**整型 可选设置**

**tasknum默认值为1，即单进程任务爬取**

举个栗子:

开启5个进程爬取网页

'tasknum' => 5

tasknum在[如何实现多任务爬虫](https://doc.phpspider.org/development_skills/multitasking_crawler.html)中作详细介绍。

**multiserver**

多服务器处理  
需要配合redis来保存采集任务数据，供多服务器共享数据使用

**布尔类型 可选设置**

**multiserver默认值为false**

举个栗子:

'multiserver' => true

multiserver在[如何实现多服务器集群爬虫](https://doc.phpspider.org/development_skills/crawler_cluster.html)中作详细介绍。

**serverid**

服务器ID

**整型 可选设置**

**serverid默认值为1**

举个栗子:  
启用第二台服务器

'serverid' => 2

**save\_running\_state**

保存爬虫运行状态  
需要配合redis来保存采集任务数据，供程序下次执行使用  
注意：多任务处理和多服务器处理都会默认采用redis，可以不设置这个参数

**布尔类型 可选设置**

**save\_running\_state默认值为false，即不保存爬虫运行状态**

举个栗子:

'save\_running\_state' => true

**queue\_config**

redis配置

**数组类型 可选设置**

保存爬虫运行状态、多任务处理 和 多服务器处理 都需要redis来保存采集任务数据

举个栗子:

'queue\_config' => array(

'host' => '127.0.0.1',

'port' => 6379,

'pass' => '',

'db' => 5,

'prefix' => 'phpspider',

'timeout' => 30,

)

**proxy**

代理服务器  
如果爬取的网站根据IP做了反爬虫, 可以设置此项

**数组类型 可选设置**

栗子1:  
普通代理

'proxy' => array('http://host:port')

栗子2:  
验证代理

'proxy' => array('http://user:pass@host:port')

**注意：如果对方根据IP做了反爬虫技术，你可能需要到**[阿布云代理](http://www.abuyun.com/)**申请代理通道 或者 第三方免费代理IP，然后在这里填写代理信息**

**interval**

爬虫爬取每个网页的时间间隔  
单位：毫秒

**整型 可选设置**

举个栗子:  
设置爬取时间间隔为1秒

'interval' => 1000

**timeout**

爬虫爬取每个网页的超时时间  
单位：秒

**整型 可选设置**

**timeout默认值为5秒**

举个栗子:

'timeout' => 5

**max\_try**

爬虫爬取每个网页失败后尝试次数  
网络不好可能导致爬虫在超时时间内抓取失败, 可以设置此项允许爬虫重复爬取

**整型 可选设置**

**max\_try默认值为0，即不重复爬取**

举个栗子:

'max\_try' => 5 // 重复爬取5次

**max\_depth**

爬虫爬取网页深度，超过深度的页面不再采集  
对于抓取最新内容的增量更新，抓取好友的好友的好友这类型特别有用

**整型 可选设置**

max\_depth**默认值为0，即不限制**

举个栗子:  
采集知乎好友时只采集到5级深度

'max\_depth' => 5

**max\_fields**

爬虫爬取内容网页最大条数  
抓取到一定的字段后退出

**整型 可选设置**

max\_fields**默认值为0，即不限制**

举个栗子:

'max\_fields' => 100

**user\_agent**

爬虫爬取网页所使用的浏览器类型  
phpspider::AGENT\_ANDROID  
phpspider::AGENT\_IOS  
phpspider::AGENT\_PC  
phpspider::AGENT\_MOBILE

**枚举类型 可选设置**

栗子1:  
使用内置的枚举类型

phpspider::AGENT\_ANDROID**, 表示爬虫爬取网页时, 使用安卓手机浏览器**  
phpspider::AGENT\_IOS**, 表示爬虫爬取网页时, 使用苹果手机浏览器**  
phpspider::AGENT\_PC**, 表示爬虫爬取网页时, 使用PC浏览器**  
phpspider::AGENT\_MOBILE**, 表示爬虫爬取网页时, 使用移动设备浏览器**

'user\_agent' => phpspider::AGENT\_ANDROID

栗子2:  
使用自定义类型

'user\_agent' => "Mozilla/5.0"

栗子3:

随机浏览器类型，用于破解防采集

'user\_agent' => array(

"Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_11\_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/52.0.2743.116 Safari/537.36",

"Mozilla/5.0 (iPhone; CPU iPhone OS 9\_3\_3 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13G34 Safari/601.1",

"Mozilla/5.0 (Linux; U; Android 6.0.1;zh\_cn; Le X820 Build/FEXCNFN5801507014S) AppleWebKit/537.36 (KHTML, like Gecko)Version/4.0 Chrome/49.0.0.0 Mobile Safari/537.36 EUI Browser/5.8.015S",

)

**client\_ip**

爬虫爬取网页所使用的伪IP，用于破解防采集

**String类型 或 数组类型 可选设置**

栗子1:

'client\_ip' => '192.168.0.2'

栗子2:

随机伪造IP，用于破解防采集

'client\_ip' => array(

'192.168.0.2',

'192.168.0.3',

'192.168.0.4',

)

**export**

爬虫爬取数据导出

type：导出类型 csv、sql、db  
file：导出 csv、sql 文件地址  
table：导出db、sql数据表名

**注意导出到数据库的表和字段要和上面的fields对应**

**数组类型 可选设置**

栗子1:  
导出CSV结构数据到文件

'export' => array(

'type' => 'csv',

'file' => './data/qiushibaike.csv', // data目录下

)

栗子2:  
导出SQL语句到文件

'export' => array(

'type' => 'sql',

'file' => './data/qiushibaike.sql',

'table' => '数据表',

)

栗子3:  
导出数据到Mysql

'export' => array(

'type' => 'db',

'table' => '数据表', // 如果数据表没有数据新增请检查表结构和字段名是否匹配

)

**db\_config**

数据库配置

'db\_config' => array(

'host' => '127.0.0.1',

'port' => 3306,

'user' => 'root',

'pass' => 'root',

'name' => 'demo',

)

**domains**

定义爬虫爬取哪些域名下的网页, 非域名下的url会被忽略以提高爬取速度

**数组类型 不能为空**

举个栗子:

'domains' => array(

'qiushibaike.com',

'www.qiushibaike.com'

)

**scan\_urls**

定义爬虫的入口链接, 爬虫从这些链接开始爬取,同时这些链接也是监控爬虫所要监控的链接

**数组类型 不能为空**

举个栗子:

'scan\_urls' => array(

'http://www.qiushibaike.com/'

)

**content\_url\_regexes**

定义内容页url的规则  
内容页是指包含要爬取内容的网页 比如http://www.qiushibaike.com/article/115878724 就是糗事百科的一个内容页

**数组类型 正则表达式 最好填写以提高爬取效率**

举个栗子:

'content\_url\_regexes' => array(

"http://www.qiushibaike.com/article/\d+",

)

**list\_url\_regexes**

定义列表页url的规则  
对于有列表页的网站, 使用此配置可以大幅提高爬虫的爬取速率  
列表页是指包含内容页列表的网页 比如http://www.qiushibaike.com/8hr/page/2/?s=4867046 就是糗事百科的一个列表页

**数组类型 正则表达式**

举个栗子:

'list\_url\_regexes' => array(

"http://www.qiushibaike.com/8hr/page/\d+\?s=\d+"

)

**fields**

定义内容页的抽取规则  
规则由一个个field组成, 一个field代表一个数据抽取项

**数组类型 不能为空**

举个栗子:

'fields' => array(

array(

'name' => "content",

'selector' => "//\*[@id='single-next-link']",

'required' => true,

),

array(

'name' => "author",

'selector' => "//div[contains(@class,'author')]//h2",

)

)

上面的例子从网页中抽取内容和作者, 抽取规则是针对糗事百科的内容页写的

field在[configs详解——之field](https://doc.phpspider.org/configs-field.html)中作详细介绍。