PHP抓取地址数据库，步骤如下：

1. 当然是要找到要抓取的链接；
2. 根据自己的需要设计数据库；
3. 设计抓取的程序，主要是链接找出规律，写出正则，还有就是把抓取到的数据放到数据库，顺便把抓取的网页保存起来，虽然感觉没什么用；
4. 开始把设计好的程序跑起来，在跑的过程中，发现问题再修复；
5. 把抓取到的地址库使用起来。

抓取的链接：<http://www.stats.gov.cn/tjsj/tjbz/tjyqhdmhcxhfdm/2016/>

抓取的是中华人民共和国国家统计局的数据，应该会权威一些。这里最新的是2016发布的数据了。

再来说说链接的入口：国家统计局官网：<http://www.stats.gov.cn/> 》统计数据》统计标准：



》统计用区划和城乡划分代码：



》2016年：

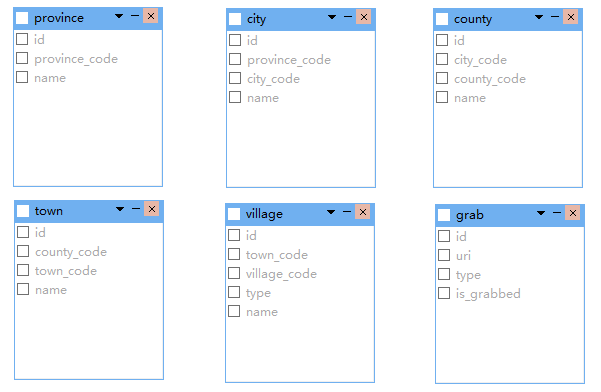


完成这一波操作就可以进入到第一个要抓取的页面了：



随便点点，熟悉熟悉。

然后就到第二步设计数据库了：



grab：链接表，主要是保存要抓取的链接表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段名** | **字段类型** | **默认值** | **索引** | **描述** |
| id | int(11) |  | PRI |  |
| uri | varchar(255) |  | UNI | 要抓取的URI |
| type | tinyint(4) | 0 |  | 0省，1市，2县，3镇，4村 |
| is\_grabbed | tinyint(4) | 0 |  | 是否抓取，0未抓取，1已抓取 |

province：省表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段名** | **字段类型** | **默认值** | **索引** | **描述** |
| id | int(11) |  | PRI |  |
| province\_code | char(12) |  | UNI | 统计用区划代码 |
| name | varchar(20) |  |  | 名称 |

city：市表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段名** | **字段类型** | **默认值** | **索引** | **描述** |
| id | int(11) |  | PRI |  |
| province\_code | char(12) |  |  | 省统计用区划代码 |
| city\_code | char(12) |  | UNI | 市统计用区划代码 |
| name | varchar(20) |  |  | 名称 |

county：县表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段名** | **字段类型** | **默认值** | **索引** | **描述** |
| id | int(11) |  | PRI |  |
| city\_code | char(12) |  |  | 市统计用区划代码 |
| county\_code | char(12) |  | UNI | 县统计用区划代码 |
| name | varchar(20) |  |  | 名称 |

town：镇表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段名** | **字段类型** | **默认值** | **索引** | **描述** |
| id | int(11) |  | PRI |  |
| county\_code | char(12) |  |  | 县统计用区划代码 |
| town\_code | char(12) |  | UNI | 镇统计用区划代码 |
| name | varchar(20) |  |  | 名称 |

village：村表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段名** | **字段类型** | **默认值** | **索引** | **描述** |
| id | int(11) |  | PRI |  |
| town\_code | char(12) |  |  | 镇统计用区划代码 |
| village\_code | char(12) |  | UNI | 村统计用区划代码 |
| name | varchar(20) |  |  | 名称 |

数据库设计完就马不停蹄设计抓取的程序，不过首先还是要抓取内容的规律，我们一步步的点到自己的家乡，发现也就是省市县镇村5个页面，只要分别对这5个页面写出抓数据的正则表达式：

省：/<a\s+href=[\'"](.+?)[\'"]>(.+?)<br\/?><\/a>/

市：/<tr\s+class=[\'"]citytr[\'"]><td><a\s+href=[\'"](.+?)[\'"]>(.+?)<\/a><\/td><td><a\s+href=[\'"].+?[\'"]>(.+?)<\/a><\/td><\/tr>/

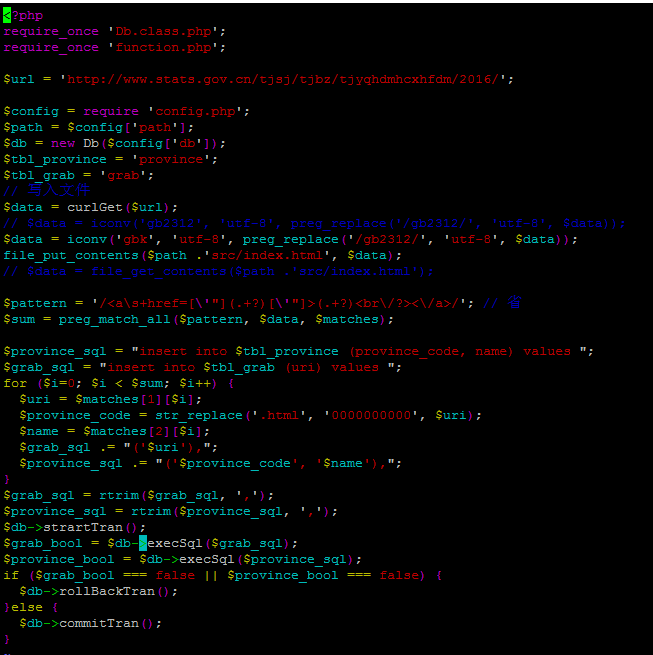
县：/<tr\s+class=[\'"]countytr[\'"]><td><a\s+href=[\'"](.+?)[\'"]>(.+?)<\/a><\/td><td><a\s+href=[\'"].+?[\'"]>(.+?)<\/a><\/td><\/tr>/

镇：/<tr\s+class=[\'"]towntr[\'"]><td><a\s+href=[\'"](.+?)[\'"]>(.+?)<\/a><\/td><td><a\s+href=[\'"].+?[\'"]>(.+?)<\/a><\/td><\/tr>/

村：/<tr\s+class=[\'"]villagetr[\'"]><td>(.+?)<\/td><td>(.+?)<\/td><td>(.+?)<\/td><\/tr>/

使用到的正则函数是：preg\_match\_all，如果不太懂用法的话可以自行查看一下官方文档。

province.php代码如下：



grab.php代码简要说明：

1. 从grab表取出一条没被抓取的记录；
2. Curl get获取链接内容，将其转换成utf-8编码，存放到指定路径；
3. 通过正则获取数据；
4. 将数据写入相应的数据表，并更新grab表的is\_grabbed字段，这里使用了事务特性，请在创建数据表的时候使用innodb引擎。

程序根据grab表未被抓取数进行循环。

运行grab.php最好直接通过PHP命令执行，不要通过web服务器，以免浪费性能：

php –f /path/grab.php & # 将其作为一个后台进程运行

tail –f /path/log.txt # 运行时可以看日志，如果倒数第二个数字大于0，而最后一个数字是0的话，说明程序有问题了：

Ctrl+c # 停止tail命令

ps aux|grep grab.php # 查看进程id

kill -9 进程id # 杀死执行进程

找出原因并修改之，重新运行。

文件执行顺序：先执行province.php 文件，确认grab表和province表都有数据了，再开始跑grab.php脚本。

程序源码还有一些说明和地址库都打包好了，打包文件包括：

1. 数据库结构：structure.sql
2. 完整的数据库：db\_address.sql
3. Readme文件，在使用程序之前最好先阅读此文件：readme.txt
4. 日志文件，在运行grab.php文件时会不断写入日志到此文件，可以tail –f动态读取文件内容：log.txt
5. 记录所有抓取到的页面：src.tar.gz
6. 记录抓取的页面的目录：src
7. 配置文件：config.php
8. 数据库操作类：Db.class.php
9. 自定义函数：function.php
10. 抓取省脚本文件：province.php
11. 抓取市县镇村脚本文件：grab.php
12. 测试文件：test.php

接下来到了第四步就是将程序跑起来，跑的时候难免会有各种各样的问题，停下来解决后重新跑起来就行。

大概几天后就跑完了，最后一步就行将地址库使用起来。

The end.THX.