# 失信人数据抓取

## 1、百度数据

### 1.1、抓取方式

链接url为：

https://sp0.baidu.com/8aQDcjqpAAV3otqbppnN2DJv/api.php?resource\_id=6899&query=失信被执行人名单&cardNum=0001&iname=&areaName=&pn=0&ie=utf-8&oe=utf-8&format=json

|  |  |  |
| --- | --- | --- |
| 参数 | 值 | 说明 |
| resource\_id | 6899 |  |
| query | 失信被执行人名单 | 程序中赋值为编码后的字符 |
| cardNum | 0001 | 证件号码身份证号码/组织机构代码，传4位的时候返回失信个人，传入单个字符的时候，返回失信企业。 |
| iname |  | 被执行人姓名/名称 |
| areaName | 区域名称 | 区域包括："北京", "天津", "河北", "山西", "内蒙古", "吉林", "黑龙江", "上海", "江苏", "浙江", "安徽", "福建", "江西", "山东", "河南", "湖北", "湖南", "广东", "广西", "海南", "重庆", "四川", "贵州", "云南", "西藏", "陕西", "甘肃", "青海", "宁夏", "新疆", "香港", "澳门", "台湾"。程序中赋值的时候编码处理 |
| pn | 0 | 起始条数，非当前页数。百度接口每次返回50条数据，翻页为pn+50. |
| Ie,oe | utf-8 | 输入输出编码 |
| format | json | 返回json格式数据 |

百度接口每次查询最多返回2000条数据，超过2000条的数据不返回。

个人数据 cardnum为四位数，0000-9999+最后一位x

企业数据cardnum为一位数，0-9+x

### 1.2、库表

企业和个人的数据均存在表cred\_dishonesty 中。

### 1.3、数据问题

1、个人查询条件大于2000条的时候，增加区域筛选条件。

2、企业总数据大于2000条，暂未解决

3、个人数据 身份证号码 不完整

4、企业数据已将组织机构编码中的"-"去除后入库

## 2、法院数据

### 2.1、抓取方式

法院网址：http://shixin.court.gov.cn/

获取查询结果，个人数据，根据cardnum的脱敏方式，补全身日信息查询，将脱敏数据补齐。企业数据根据名称大于4位的方式根据区域省份查询入库。

查询url：

http://shixin.court.gov.cn/findd?pName=\_\_&currentPage=267&pProvince=0&pCardNum=\_\_\_\_\_\_\_\_\_\_1116\_\_\_\_&pCode=5561

|  |  |  |
| --- | --- | --- |
| 参数 | 值 | 说明 |
| pName | \_\_ | 被执行人姓名/名称 |
| currentPage | 267 | 当前页数 |
| pProvince | 0 | 区域代码 |
| pCardNum | \_\_\_\_\_\_\_\_\_\_1116\_\_\_\_ | 证件号码身份证号码/组织机构代码 |
| pCode | 5561 | 验证码 |

根据此请求返回的html结果，解析出当前页码的数据id，每页10条数据。在根据每条记录的id，获取记录详细信息：http://shixin.court.gov.cn/findDetai?id=2704954&pCode=6078

|  |  |  |
| --- | --- | --- |
| 参数 | 值 | 说明 |
| id | 2704954 | Id号码 |
| pCode | 5561 | 验证码 |

返回的为json格式的数据，可直接解析入库。

### 2.2、库表

cred\_dishonesty\_person -- 个人数据存储表

cred\_dishonesty\_ent -- 企业数据存储表

cred\_dishonesty\_log -- 每个查询条件一个线程时的日志表，包括endpage和startpage。执行时 从startpage跑到endpage

cred\_dishonesty\_pagelog --每个页面一个线程时的日志记录表。包括当前页数。每次启动从第1页跑到endpage，此表中已存在的页数跳过，不存在的发送网络请求。

### 1.3、数据问题

1、数据重复问题，不同页数包含重复数据

2、百度和法院数据，互有多少。

cred\_dis\_baidu\_more 为百度多出来的个人数据 增加sremark字段

cred\_dis\_baidu\_more\_ent 为百度多出来的企业数据

cred\_dis\_fayuan\_more 为法院多出来的个人数据

cred\_dis\_fayuan\_more\_ent 为法院多出来的企业数据

|  |  |
| --- | --- |
| sremark值 | 说明 |
| 1 | 百度接口有 且在 法院网站能查到但是没有入库的数据 以及 百度接口新获取的增量数据 ---待法院表处理的数据 |
| 2 | 百度接口曾经有 现在百度接口和法院网站都查不到的数据 |
| 0 | 标示为1的客户 重新处理后入库法院表的数据。 |
|  |  |

先获取的百度数据，后获取的法院数据

其中法院多出来的数据为更新数据。百度能查到。

百度多出来的数据中，在cred\_dis\_baidu\_more 表中增加了remark字段，字段值为1的时候为在法院页面能查到的数据，其中60%左右可查到

3、法院数据排序混乱，无法增量获取新更新数据

4、法院个人数据晚上十点半开始更新数据，页面数据获取不准确

6、企业数据已将组织机构编码中的"-"去除后入库

## 3、代码问题

### 1、部分验证码请求返回错误

部分请求,大部分第一次请求http://shixin.court.gov.cn/image.jsp时，返回的不是image类型的数据，是js重新运算cookie后重新请求的代码。



解压缩后代码：

### 2、链接重置

5个线程并发访问，在多线程导致网站无法访问，修改httputilpool模式后，client不回收，会出现代码链接能保持访问，页面打不开的情况。

通过Burp软件拦截网络拦截发现不发送image请求，可以使用同一验证码无线访问。

但是在代码访问中，并没有请求image，会返回验证码错误的页面。

### 3、ip被封

使用代理的时候，小部分代理会篡改请求数据，120.52.72.24整个段

有稳定代理可以用同一代理多线程请求数据。但是代理返回数据较慢，容易超时报错。

cred\_dishonesty\_proxy 为代理表，其中isusered为2是不可用的代理，0为未使用，1为正在使用的代理.

ProxyHandler 类 可以抓取http://www.youdaili.net/ 页面的代理地址入库

### 4、补全百度证件号码效率低

补全百度证件号码效率很低，平均每个线程每分钟补全两条数据。

## 4、剩余工作

### 1、企业数据重新抓取

企业数据之前按照名称：'\_\_' 证件号码：'\_\_\_\_\_\_\_\_-\_' 按照省份分别抓取，但是很大一部分数据无法匹配，可以改成 名称：'\_\_\_\_' 证件号码：'\_\_\_\_\_\_\_\_\_' 按省份抓取，认为企业名称都大于4位，可能会抓取到部分姓名较长的个人数据。

### 2、增量获取数据

因为法院数据不规则，无法按时间增量获取新增数据，需要从百度接口获取相应信息后从法院补齐脱敏信息。

百度增量数据可用**update\_time** 字段过滤获取

### 3、删除超时作废数据

部分数据会超时删除，建议长时间段做全量扫描

## 5、主要类文件说明

### 1、com\fayuan

com\Main.java:主方法，调用pageHandle多线程抓取数据

com\MainForGetLog.java:日志信息主方法类，调用DBLogHandler完善日志表cred\_dishonesty\_log信息

com\MainForBaiduMore.java：处理百度比法院多出来的数据

com\MainForLess.java：处理每部分页数获取总条数不足10条的异常数据

### 2、com\fayuan\handler

com\fayuan\handler\DBLogHandler.java：日志信息处理 通过MainForGetLog调用，完善个人和企业的日志表cred\_dishonesty\_log信息 包括最大页数和总条数

com\fayuan\handler\DishonestyService.java：法院数据获取的serveice方法，包括数据库交互，处理网络请求等

com\fayuan\handler\HelpBatch.java：预备在主线程中启一个线程，每段时间执行一次，获取httputilpool的执行情况。后来用作判断时间点，晚上十点半到凌晨一点程序不执行。

com\fayuan\handler\PageHandler.java：main方法通过调用此类实现多线程，按页执行请求并处理返回数据入库

com\fayuan\handler\ProxyHandler.java：代理信息处理，代理数据抓取入库

### 3、com\fayuan\util

com\fayuan\util\CheckNumber.java:解析图片验证码

com\fayuan\util\CheckNumber.java：时间相关工具类

com\fayuan\util\HttpUtil.java：抓取网站数据工具类

com\fayuan\util\HttpUtilPool.java：httputil的连接池

com\fayuan\util\GetDateException.java: 自定义异常

com\fayuan\util\NetWorkException.java:自定义异常

### 4、com\fayuan\dao

src\com\fayuan\dao\ConnectionPool.java： 数据库链接池 增加了获取次数，等待时间以及获取不到的异常抛出功能

src\com\fayuan\dao\ConnUtil.java:数据库链接工具类 增加修改了单例模式 ，初始化链接的时候创建ConnectionPool

### 5、com\baidu

com\baidu\GetData.java:主方法，获取企业和个人数据

### 6、com\baidu\handler

com\baidu\handler\Dishonesty\_ENT.java：企业数据处理类

com\baidu\handler\Dishonesty\_Person.java：个人数据处理类

com\baidu\handler\BaiduEnum.java：相关常量类

com\baidu\handler\ServiceForBaidu.java：抓取百度数据的相关service类

## 6、相关SQL

