将爬虫系统的包（svn https://172.16.16.254:8443/svn/Project\_SocialTV/PaChong）copy到需要运行爬虫系统的服务器上，并解压，下面是服务器部署工作流程：

# 1.服务器软件安装

爬虫框架相关的框架要安装软件flask, flask\_wtf, flask\_bootstrap, pymongo, bs4, flask\_sqlalchemy, mysqldb等，安装脚本在"PaChong/tool"目录下的spidersoftware\_setup.sh。

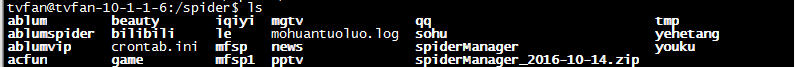
**$cd PaChong/tool && ./spidersoftware\_setup.sh**

Note：若成功，则最后提示“setup spider software OK”；否则提示“"some software setup error, please see setuperror.txt”，看错误log，手动安装失败的插件即可

SVN上的爬虫包与实际服务器上部署的爬虫系统有些出入（也可把二者的内容改成一致，目前先保持现状），说明如下：

1）目录对应关系

实际服务器上部署的爬虫系统的目录如下：



SVN爬虫包的目录



对应如下

|  |  |  |  |
| --- | --- | --- | --- |
| 目录 | SVN对应 | 服务器对应 | 描述 |
| ablumsspider | ablumsspider | spider，除spiderManger外 | 定时爬取模块的目录 |
| spiderManger | spiderManger | spider/ spiderManger | 爬虫管理系统目录 |
| tool | tool | 无对应 | 安装爬虫管理系统的依赖软件 |

2）目录添加内容

以野荷塘目录为例说明

实际服务器上yehetang目录

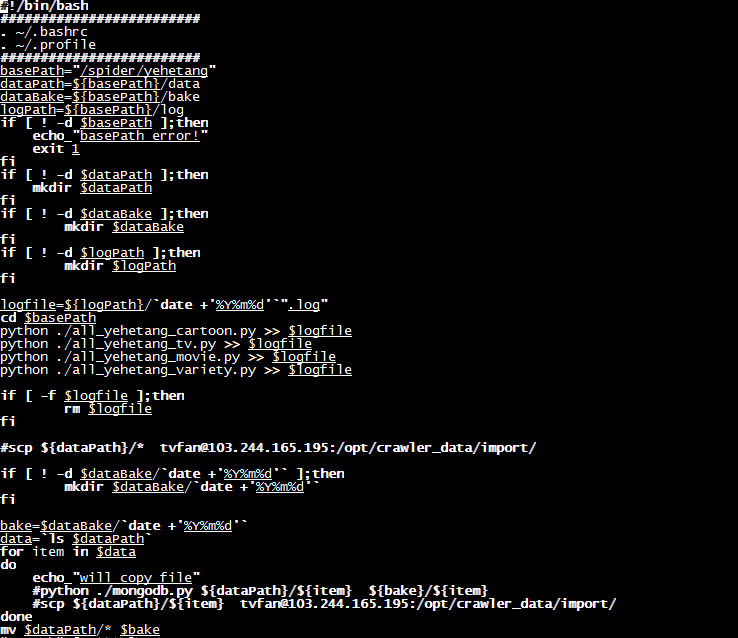


SVN上yehetang目录



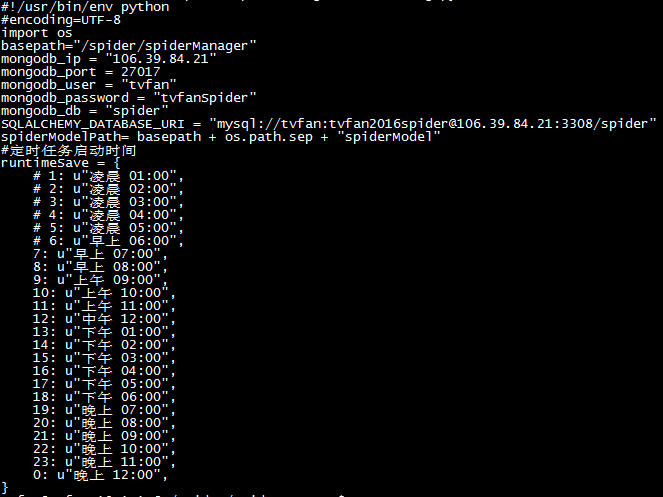
相比之下，实际服务器上主要添加脚本"run.sh"（其它不计或是脚本生成的）

其中“run.sh”即为视频yehetang的爬取脚本，模板如下：



# 2.相关服务器配置文件修改

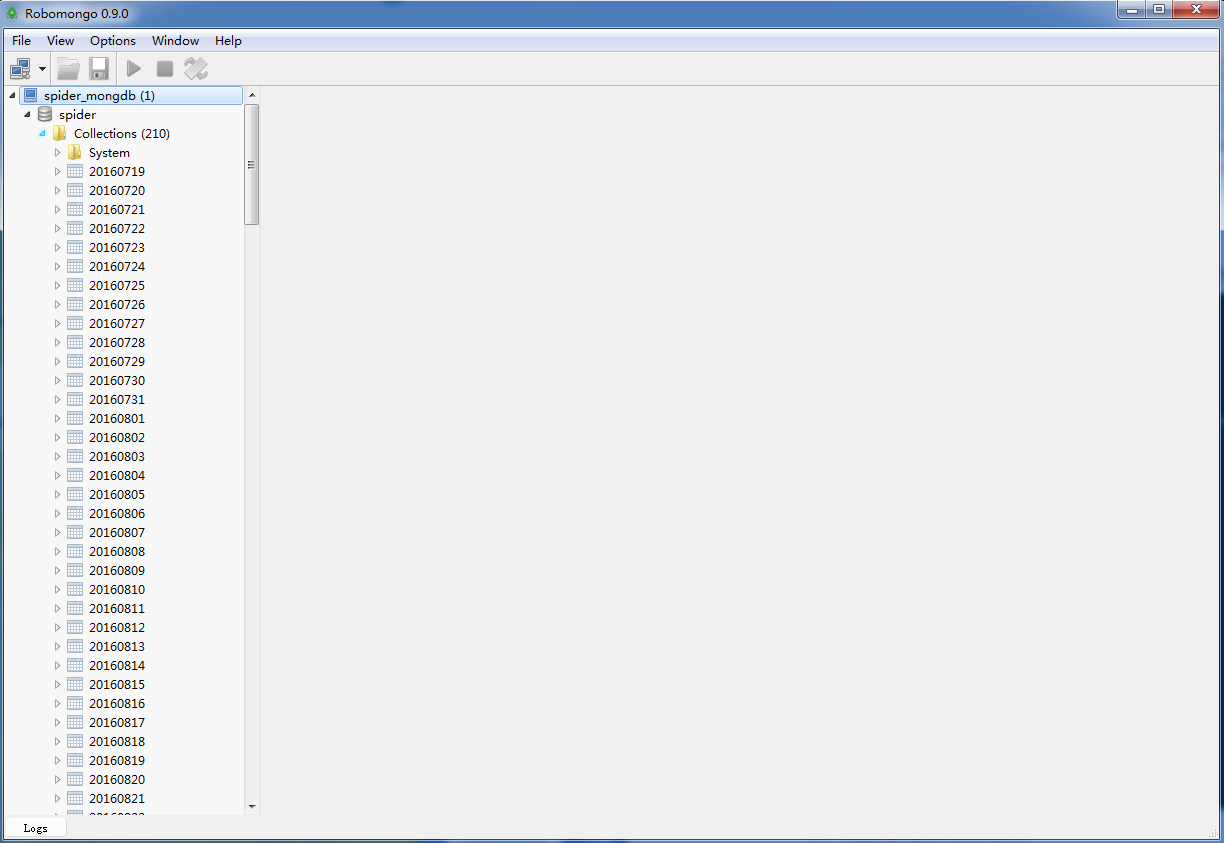
爬虫管理系统在目录"PaChong/spiderManager"中，配置文件config.py，内容如下：



1）爬虫管理系统的路径

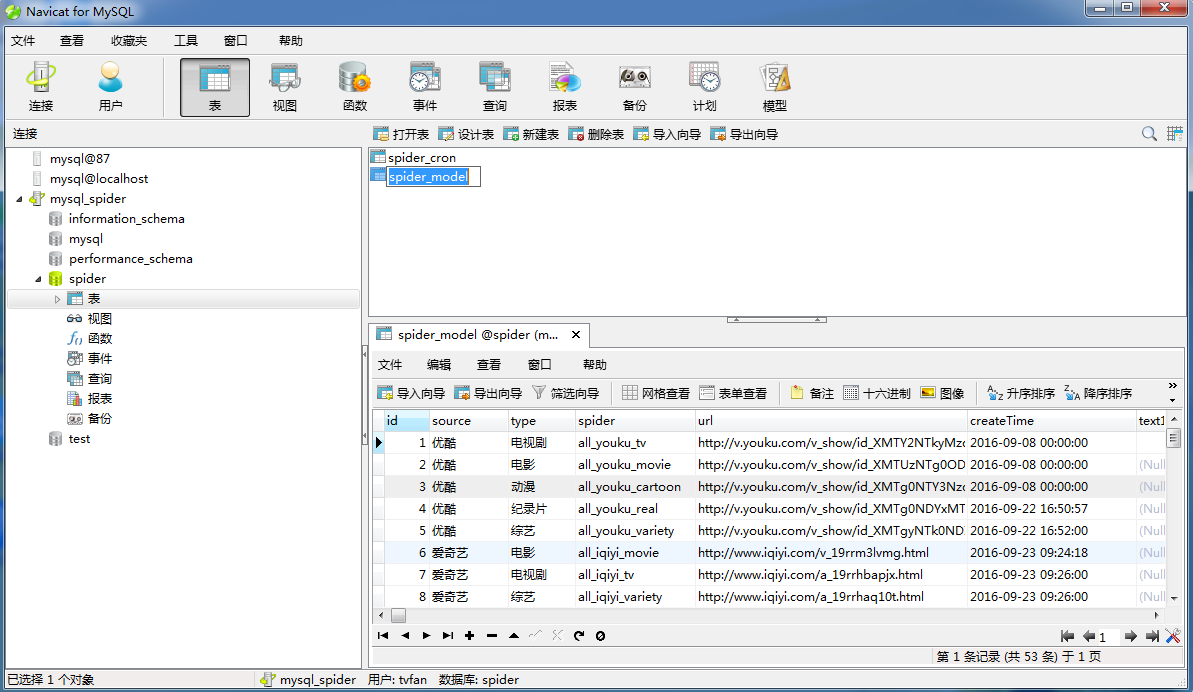
2）mongodb数据库配置

包括mongodb的ip，port以及登陆的user，password和db名，ip、port及user、password、db真实有效，否则不能进行统计工作（不影响爬取系统运行）



3）sql数据库配置

sql数据库的用户名，密码，及数据库名，数据库的信息真实有效，表里存储临时爬取模块的信息



4）spider模块路径

临时爬取的所有模块的路径

5）定时任务启动时间

可选的定时任务启动时间表，供在临时爬取成功用于保存结果时用

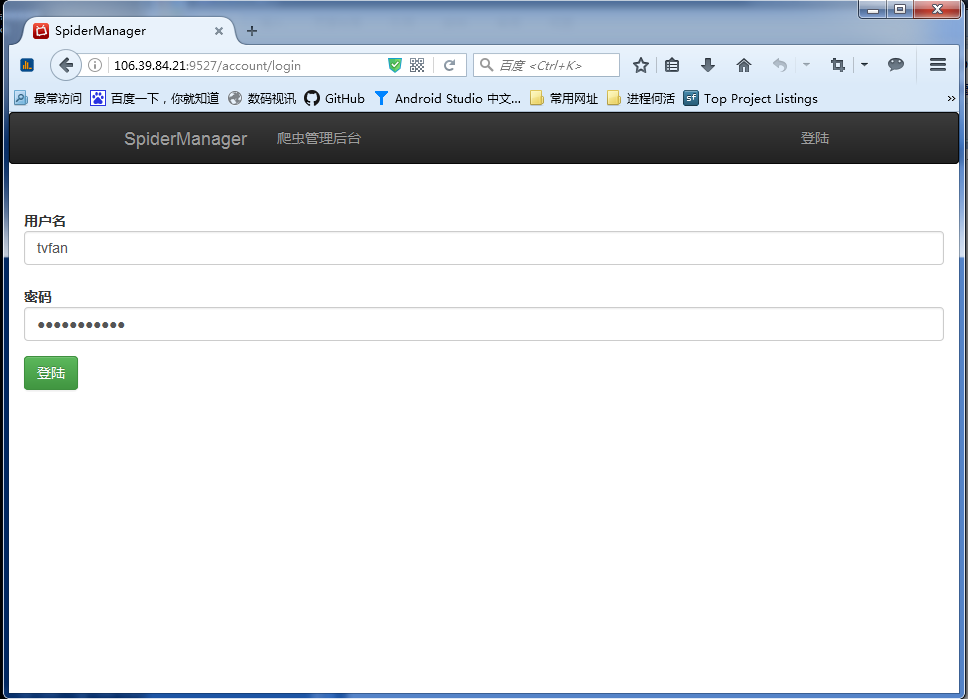
# 3.运行爬虫管理系统

## 3.1系统运行

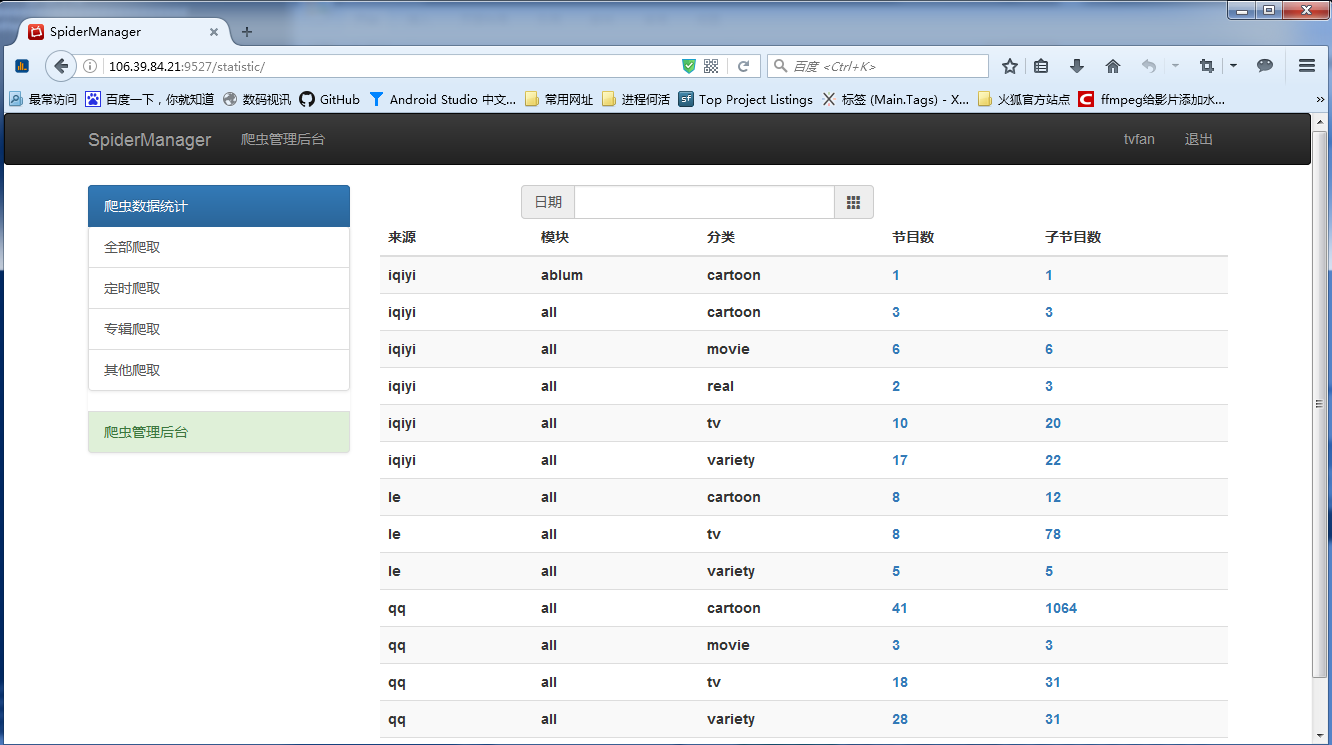
在前面的工作准备完成，可以运行爬虫管理系统（运行一次即可)

**$cd spiderManager && ./run.sh**

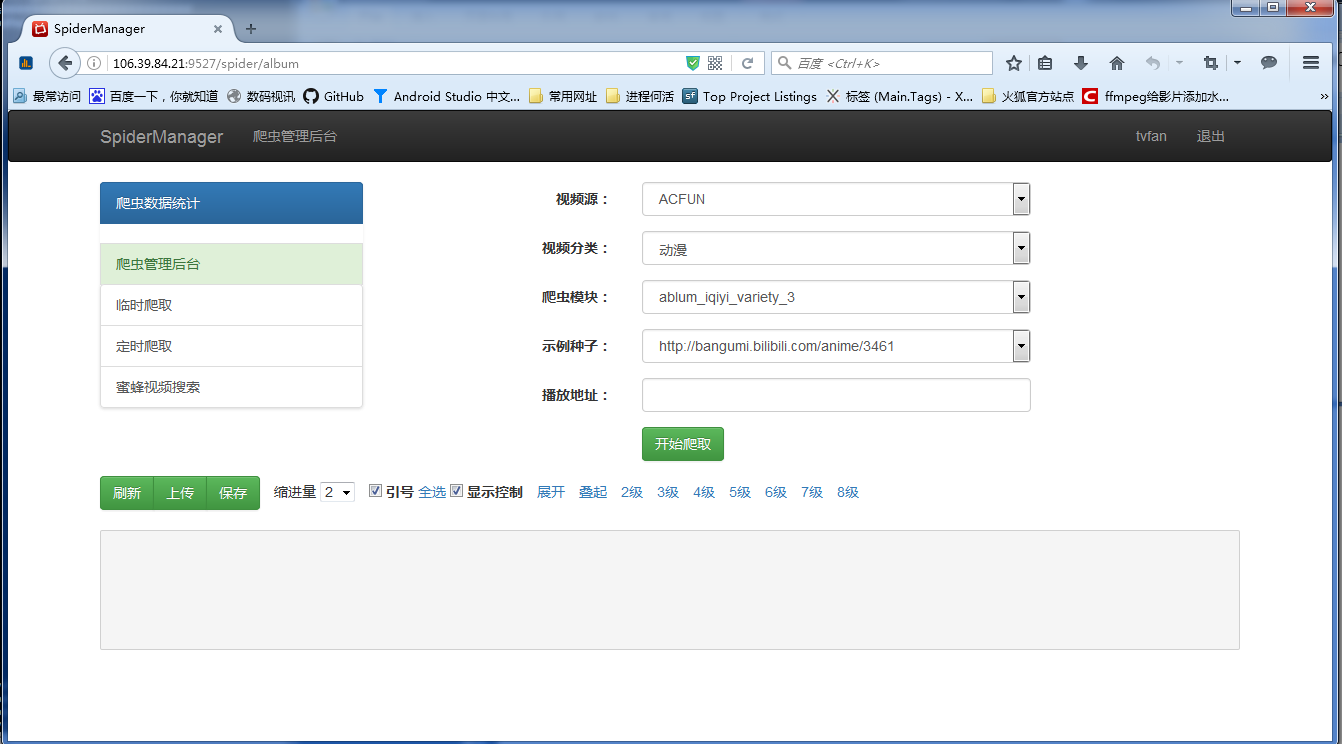
这样就可以访问爬虫管理系统了，访问"http://106.39.84.21:9527"，用户名：“tvfan”，密码：“befun@tvfan”，界面如下：



爬取管理系统登陆界面



爬虫统计数据界面



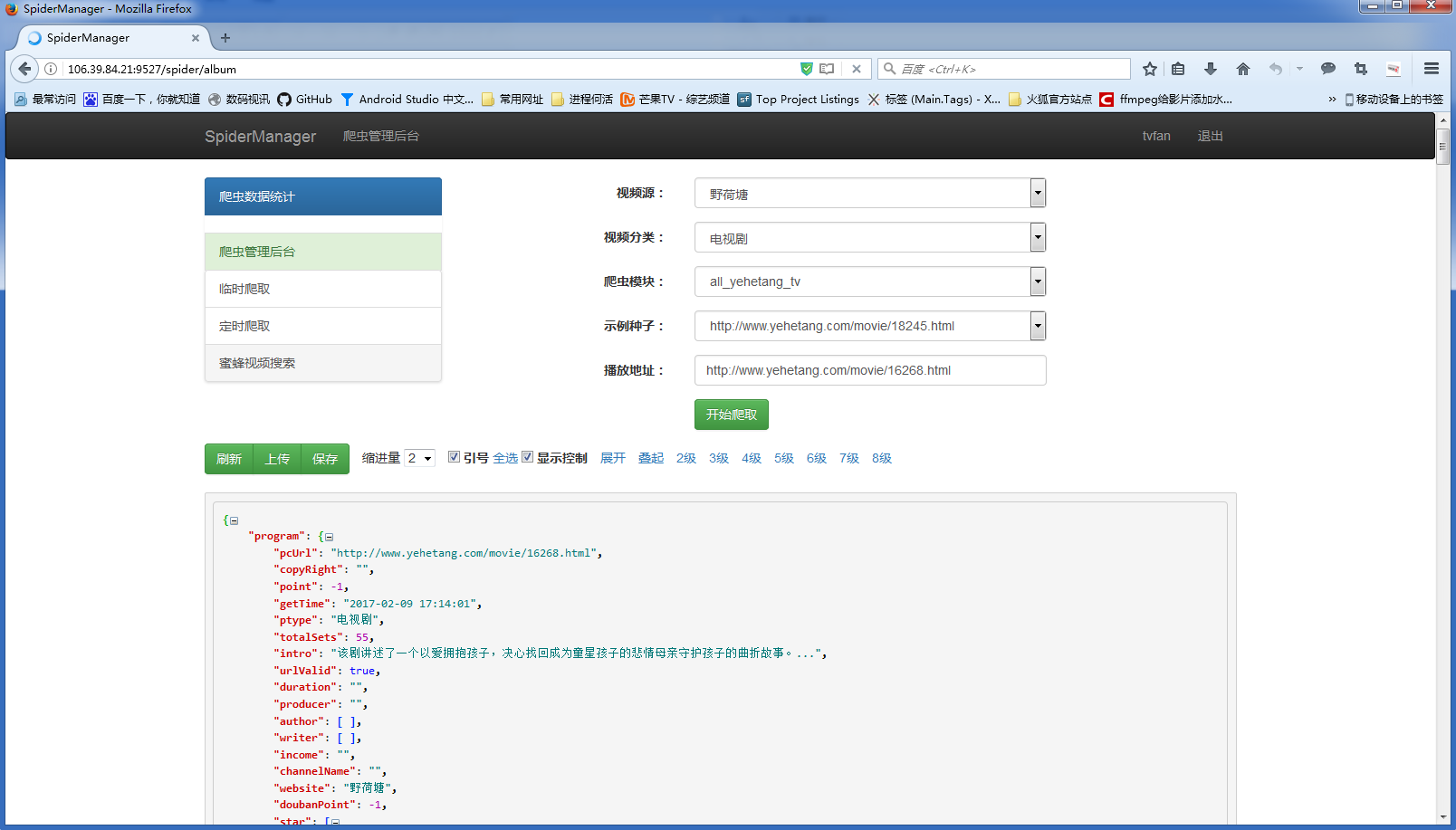
临时爬取界面

## 3.2系统说明

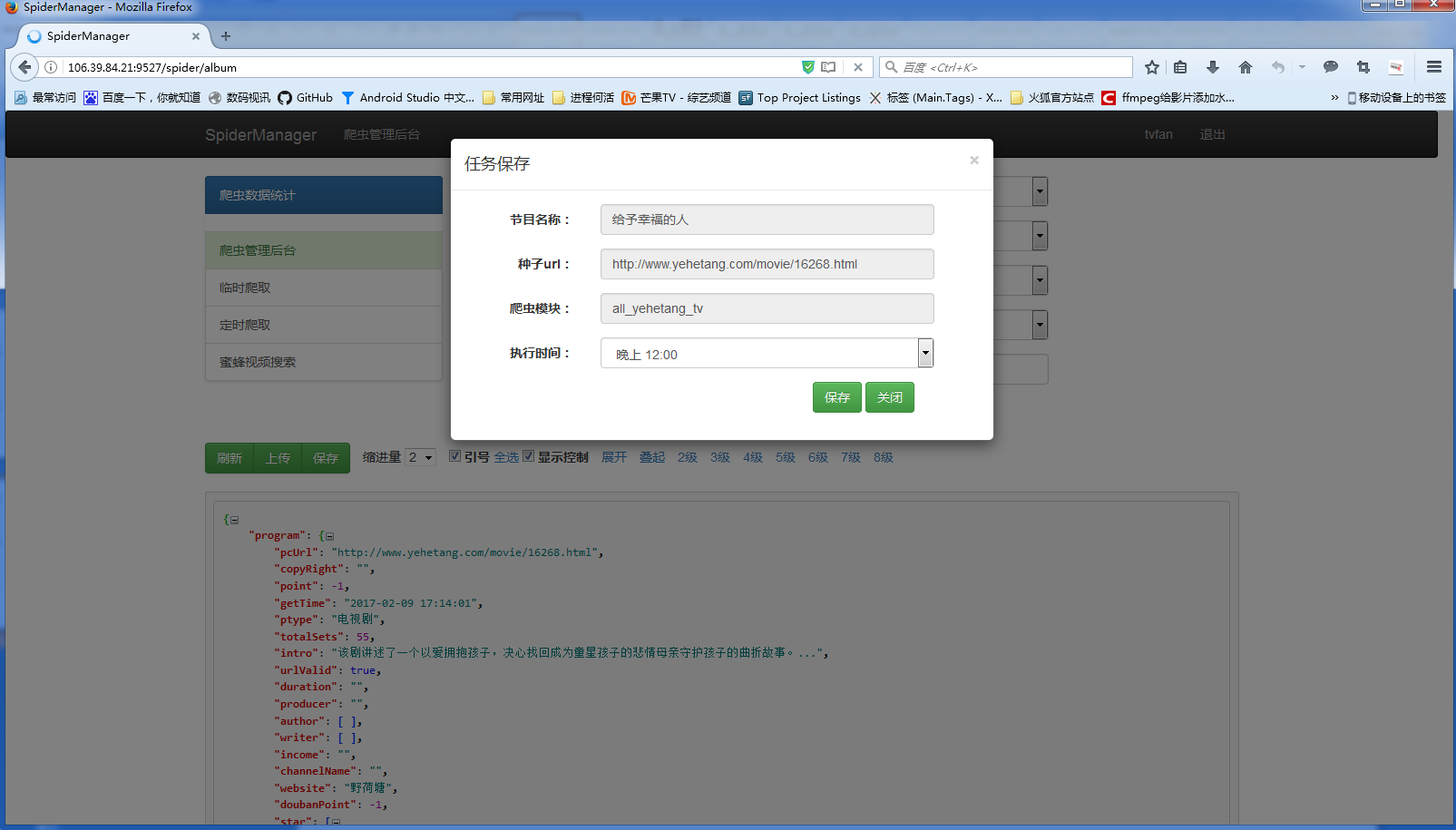
爬虫管理后台说明如下：

1）临时爬取

爬取已有爬虫模块的资源，依次选取视频源、视频分类（个别的还坱要选择爬虫模块），在播放地址输入要爬取视频的种子，格式类似示例种子，点击爬取，若一次爬取失败可再次点击开始爬取。点击上传按钮会把爬取到的视频信息保存下来并入爬虫库。点击保存按钮会出现任务保存的对话框，可以修改执行的时间，爬取任务会在每天的这个时间点进行爬取入库，叵要编辑保存的定时任务，可在点击定时爬取并编辑。



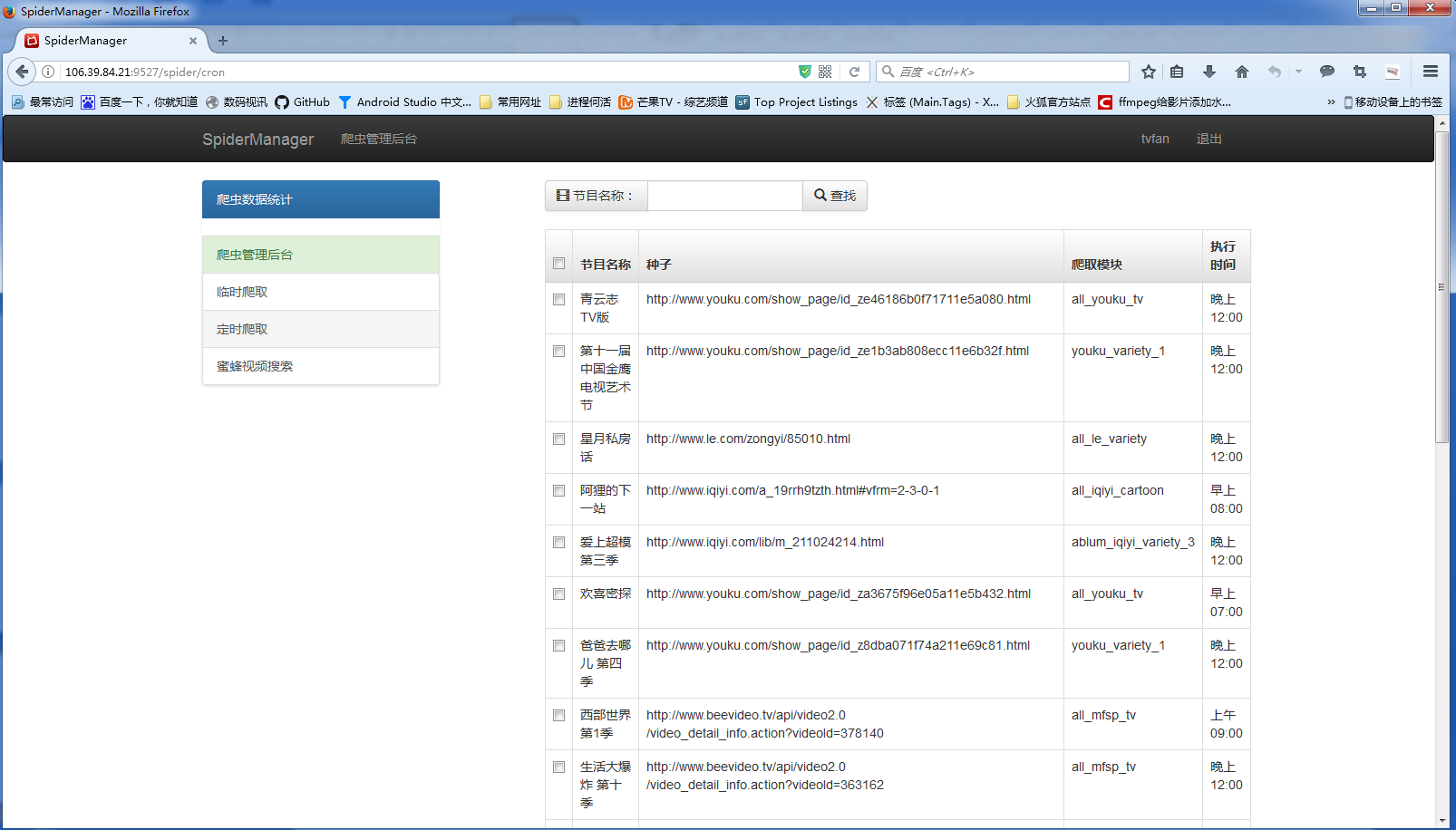
临时爬取界面



临时爬取结果保存

2）定时爬取

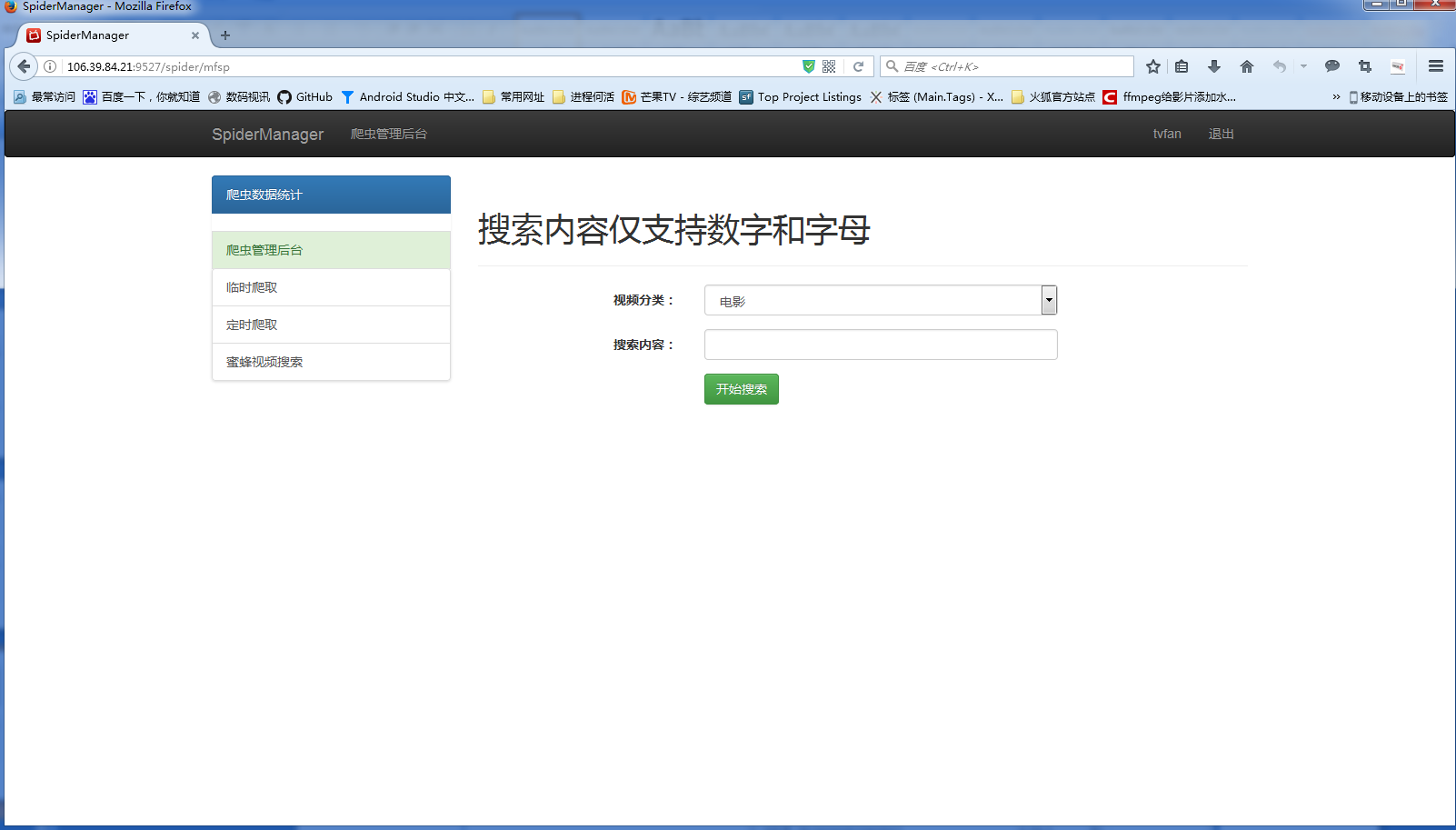
临时爬取保存来下就变成定时爬取，定时爬取会列出定时任务，定时任务可以进行编辑，有修改、增加和删除，增加会自动跳到临时爬取（临时爬取的保存会增加一条定时爬取任务）



定时爬取界面

3）蜜蜂视频

蜜粉视频没有视频网站，能过搜索进行爬取

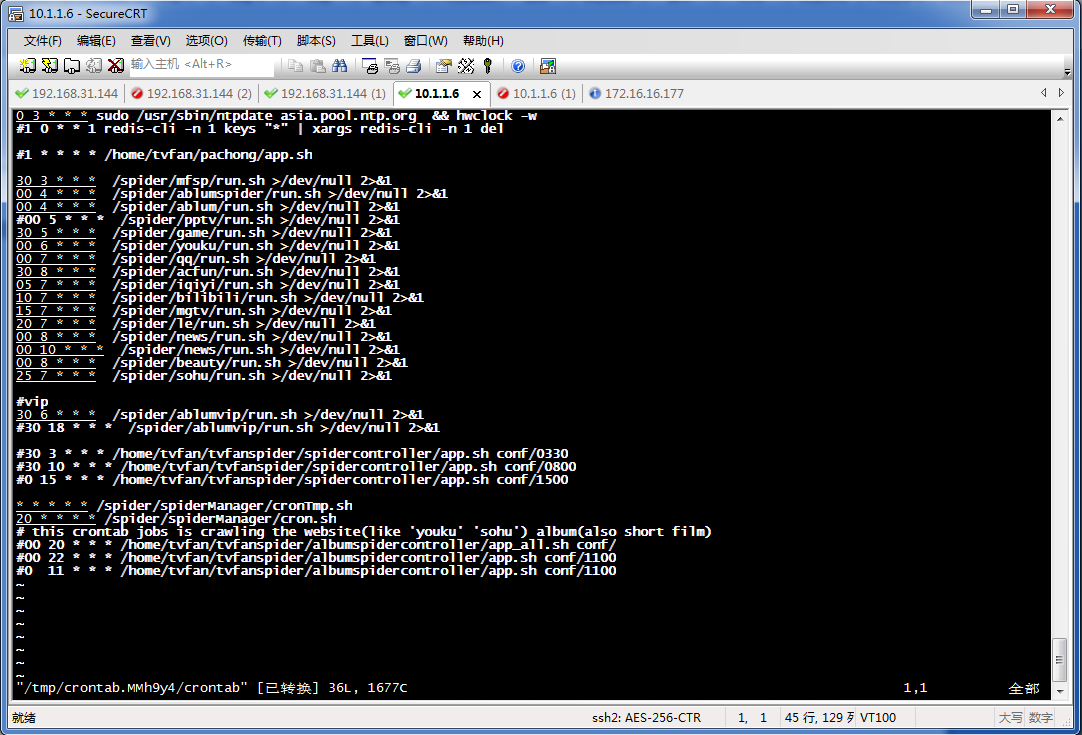


蜜蜂视频搜索界面

# 4定时爬取任务

此定时任务与爬虫管理系统中的定时任务不是同一概念，这是视频cp的定时爬取任务

通过"crontab -e"打开定时任务界面并可以编辑，其中的"run.sh"脚本也即是各cp爬取的脚本，在“服务器软件安装”有说明。



定时爬取任务界面