## 1聚焦爬虫介绍

聚焦爬虫（又被称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁，自动索引，模拟程序或者蠕虫。

　　随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。搜索引擎(Search Engine)，例如传统的通用搜索引擎AltaVista，Yahoo!和Google等，作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。但是，这些通用性搜索引擎也存在着一定的局限性，如：

　　(1) 不同领域、不同背景的用户往往具有不同的检索目的和需求，通用搜索引擎所返回的结果包含大量用户不关心的网页。

　　(2) 通用搜索引擎的目标是尽可能大的网络覆盖率，有限的搜索引擎服务器资源与无限的网络数据资源之间的矛盾将进一步加深。

　　(3) 万维网数据形式的丰富和网络技术的不断发展，图片、数据库、音频/视频多媒体等不同数据大量出现，通用搜索引擎往往对这些信息含量密集且具有一定结构的数据无能为力，不能很好地发现和获取。

　　(4) 通用搜索引擎大多提供基于关键字的检索，难以支持根据语义信息提出的查询。

　　为了解决上述问题，定向抓取相关网页资源的聚焦爬虫应运而生。聚焦爬虫是一个自动下载网页的程序，它根据既定的抓取目标，有选择的访问万维网上的网页与相关的链接，获取所需要的信息。与通用爬虫(general purpose web crawler)不同，聚焦爬虫并不追求大的覆盖，而将目标定为抓取与某一特定主题内容相关的网页，为面向主题的用户查询准备数据资源。

　　1 聚焦爬虫工作原理及关键技术概述

　　网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列,直到满足系统的一定停止条件，如图1(a)流程图所示。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止，如图1(b)所示。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

　　相对于通用网络爬虫，聚焦爬虫还需要解决三个主要问题：

　　(1) 对抓取目标的描述或定义；

　　(2) 对网页或数据的分析与过滤；

　　(3) 对URL的搜索策略。

　　抓取目标的描述和定义是决定网页分析算法与URL搜索策略如何制订的基础。而网页分析算法和候选URL排序算法是决定搜索引擎所提供的服务形式和爬虫网页抓取行为的关键所在。这两个部分的算法又是紧密相关的。

　　2 抓取目标描述

　　现有聚焦爬虫对抓取目标的描述可分为基于目标网页特征、基于目标数据模式和基于领域概念3种。基于目标网页特征的爬虫所抓取、存储并索引的对象一般为网站或网页。根据种子样本获取方式可分为：

　　（1） 预先给定的初始抓取种子样本；

　　（2） 预先给定的网页分类目录和与分类目录对应的种子样本，如Yahoo!分类结构等；

　　（3） 通过用户行为确定的抓取目标样例，分为：

　　a) 用户浏览过程中显示标注的抓取样本；

　　b) 通过用户日志挖掘得到访问模式及相关样本。

其中，网页特征可以是网页的内容特征，也可以是网页的链接结构特征，等等。

http://drops.wooyun.org/tips/3915

## 2分析目前比较流行的开源爬虫程序

### 2.1pyspider

此开源程序利用python语言实现，下面列举一些学习此开源软件的目的以及目标：

1. 掌握python编程技术。
2. 页面的请求和响应的处理。
3. URL去重的方法。
4. URL相似度的处理。
5. 其中的并发操作。

### 2.2整体架构



#### 2.2.1源码目录结构

在根目录中可以找到的文件夹有：

•data，空文件夹，它是存放由爬虫所生成的数据的地方。

•docs，包含该项目文档，里边有一些markdown代码。

•pyspider，包含项目实际的代码。

•test，包含相当多的测试代码。

这里我将重点介绍一些重要的文件：

•.travis.yml，一个很棒的、连续性测试的整合。你如何确定你的项目确实有效？毕竟仅在你自己的带有固定版本的库的机器上进行测试是不够的。

•Dockerfile，同样很棒的工具！如果我想在我的机器上尝试一个项目，我只需要运行Docker，我不需要手动安装任何东西，这是一个使开发者参与到你的项目中的很好的方式。

•LICENSE，对于任何开源项目都是必需的，（如果你自己有开源项目的话）不要忘记自己项目中的该文件。

•requirements.txt，在Python世界中，该文件用于指明为了运行该软件，需要在你的系统中安装什么Python包，在任何的Python项目中该文件都是必须的。

•run.py，该软件的主入口点。

•setup.py，该文件是一个Python脚本，用于在你的系统中安装pyspider项目。

Pyspider文件夹下面是包含整个程序最重要得实现文件，有database、fetcher、libs、processor、result、scheduler、webui，很清楚的看到对应于框架图中的模块。

程序入口run.py,下面认真分析一下：

主要是进行一些初始化工作，大概猜测有如下工作要进行，a.多个线程之间的数据怎么去共享b.数据库的初始化。

#### 2.2.2webUI实现机制

如何实现？利用了哪些框架？如何改造？

#### 2.2.3 scheduler实现机制

#### 2.2.4 fetcher实现机制

#### 2.2.5 processor实现机制

#### 2.2.6 processor实现机制