## 关于针对已下载网页过滤的过滤器的改进

#### 改进说明

1. Scrapy实现说明。

Scrapy采用Python内置的集合（set）实现内存过滤器，同时提供文件过滤器让用户选择。两者并不排斥，文件过滤器的设置是为了可以停止爬虫（结束爬虫进程），然后在之后的某一个时间内重新接着上次的爬取­­继续时，可以直接从文件中读取已经爬取的url。实际运行期间的过滤都是在内存过滤器（set集合实现的过滤器）进行的。

1. Scrapy实现分析以及改进说明。

默认的Python集合需要存储网页本身的url，同时为了实现管理集合以及其他一些针对集合的操作，Python需要保存的内容除了URL之外还有许多，因此，当爬取大量网页的时候占用内存会比较大。而分析发现，我们实际上只需要判断一个url是否已经出现过，而至于url是什么或者何时出现并不在乎。其次，我们也不会对已经爬取的url进行索引等操作，所以Python内置的set结合本身的许多操作相对来说是冗余的。

#### 实现方案

1.方案说明。

根据以上分析，可以使用布隆过滤器来实现对已下载网页的过滤。

2.布隆过滤器介绍

布隆过滤器（英语：Bloom Filter）是1970年由布隆提出的。它实际上是一个很长的[二进制](https://zh.wikipedia.org/wiki/%E4%BA%8C%E8%BF%9B%E5%88%B6)向量和一系列随机[映射函数](https://zh.wikipedia.org/w/index.php?title=%E6%98%A0%E5%B0%84%E5%87%BD%E6%95%B0&action=edit&redlink=1)。布隆过滤器可以用于检索一个元素是否在一个集合中。它的优点是空间效率和查询时间都远远超过一般的算法，缺点是有一定的误识别率和删除困难。