关于搭建测试环境的说明

1 测试方法

为了显示的相关实验的改进效果，搭建了测试环境，相关说明如下。

在改进前和改进后，分别统计爬取同一网站所有目标的url所花费的时间，并将爬取到的内容输出到result.txt文档中，最后根据统计结果，证明改进方案是否成功。

1.1 爬取对象

北航计算机学院官方网站

主站url为：http://scse.buaa.edu.cn/buaa-css-web/initAction.action

1.2 统计方式

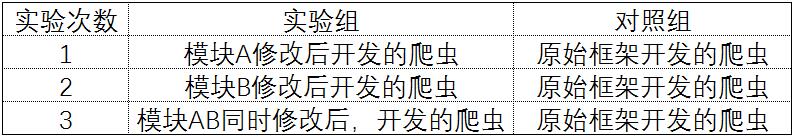
分别统计记录爬取开始和爬取结束的时间。

为了排除偶然因数的影响，会进行12次实验，去掉最长时间和最短时间，计算余下10次爬取过程的平均时间，提高统计的可靠性。

1.3 实验对照组

本次实验，采取控制变量法，进行对照实验，主要会分成3组对照实验进行，具体对照内容，如表1所示

表1 对照实验说明



2 爬虫主体

爬虫主体，采取scarpy框架（优化前 优化后）,基于python2.7进行开发，爬虫主体目录如图1所示（备注：其中的test\_time.c和test\_time与爬虫主体无关，将在3小节中介绍）

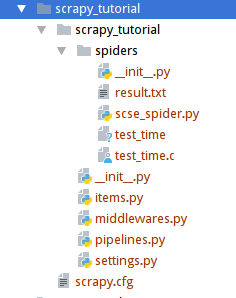


图1 爬虫目录缩略图

2.1 相关技术

利用正则表达式与框架支持的css选择器，对url链接进行过滤排除筛选

2.2 实现原则

此次爬取内容，仅统计了站内的所有相关网页的url信息，而没有进行进一步的内容析取，因为这不影响实验的最终结论。

每个url只会爬取一次，不会进行重复的爬取。

非计算机学院的url会通过正则表达式排除在外，保证相关变量的统一性。

3 统计程序

时间统计程序采用C语言，得益于C语言的高效特性，尽可能排除其它无关因素的影响。在C语言内，通过stdlib.h库，调用系统shell接口，调用爬虫进行爬取，通过time.h库，记录爬取开始和爬取结束的时间，计算时间差，实现1.2节所描述的实验方法。

得到实验结果，实验结果样例如图2所示：

C:\Users\Administrator\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Screenshot from 2017-05-02 22-39-58.png

图2 统计结果输出样例

4 实验环境

操作系统：Ubuntu 16.04 LTS

框架：scrapy 3.1

语言：python 2.7，C语言

CPU：Intel Core i7-6700

内存：8G

系统配置配置一览如图3所示：

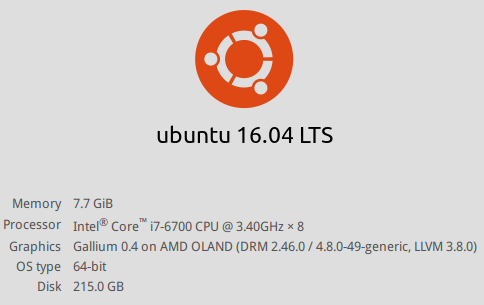


图3 实验平台硬件信息