# 信息检索—Web 搜索引擎实现文档

1511426 黄静平

December 13, 2017

## 1 前言

一般的搜索引擎架构如 1 所示。所以为了实现一个简易的搜索引擎,我们需要实现爬虫,链接分析,文本构建索引,Page rank,查询接口五部分。我们先通过爬虫爬取网页,然后通过链接分析部分分析处理爬下来的网页内容,同时对文本内容构建索引和 Page rank。前四个部分完成后,我们实现查询接口,处理用户的查询请求。下面我对这五个部分进行一一介绍。

## 2 模块介绍

### 2.1 网页爬虫

爬虫部分使用 Python 实现。具体步骤是我们先给一个种子 url,这里我选择的是南开大学官网首页。然后我们爬取这个网页,分析其内容,提取出页面包含的链接,将它加入到待爬取队列中。这里我设置了一些数据结构,比如待爬取队列,黑名单,白名单,已爬列表等。用来提高爬虫效率,防止爬取重复网页和无效网页。同时我还对重定向之类的问题进行了处理,防止保存下来的数据和实际情况不符。为了减小爬虫过程中对学校网站造成的负担,我的爬虫在爬取一定数量网页后会休息 30 秒。同时为了防止爬到一半因为各种原因程序崩了而导致前功尽弃的情况,我也设置了在爬取一定数目的页面后及时的将数据从内存保存到磁盘中。具体保存的数据包括:已爬取 url 列表,待爬取页面队列,黑名单,页面信息,爬下来的 html等文件等等。这里我为了程序的移植性更好,使用的数据库是 Python 自带的 Sqlite。这里程序有两个入口,一个是对整个爬虫程序重新开始,另一个是读取之前保存的数据,然后从上次停止的地方开始。这个可以由用户自己设置。2是爬虫过程的截图。

### 2.2 链接分析

链接分析是和爬虫一起进行的。爬虫将爬下来的数据交给链接分析模块,链接分析模块使用正则表达式加上 Python 的 BeautifulSoup 包提取出链接和对应的锚文本,然后将链接关系保存到数据库中。为了实现网页快照功能,我们需要将此时爬下来的网页内容进行存储。由于一些网页是通过 javascript 生成的,所以这里我们需要先执行 Javascript 代码才能获取真实页面。这里用到了 BeautifulSoup 中的 extract 函数。前面保存了链接关系,接着我们要保存当前页面的信息,包括当前页面类型 (比如 pdf,word,html 等等),页面 title,页面 id 等等。然后将解析出来的 url 返回给爬虫模块,爬虫模块将新的 url 添加到待爬队列中。

# 搜索引擎体系结构

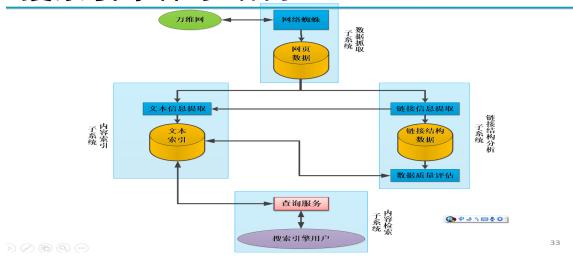


Figure 1: 搜索引擎体系结构.

Figure 2: 爬虫截图

### 2.3 文本索引构建

文本索引构建用的是 Python 的 Pylucene 包。通过这个包,我们爬好的页面就可以构建出索引了,以此提高我们搜索时的效率。

### 2.4 Page rank

Page rank 是对页面重要性的排序。这里我们主要通过链接关系来决定一个页面的重要程度。由于我们之前通过链接分析得到链接关系,将之保存在数据库中,现在我们从数据库中将这些关系读取到内存中,假设有 count 个页面,那么我们就构建一个 count×count 的二维数组 link,其中 link[i][j] 表示页面 j 指向页面 i。但是不同页面的链接关系的价值是不同的,所以我们需要对每个链接关系给出其价值。这里我们使用的方法很简单,遵从一个原则就是"沉默是金"。即如果一个页面只有 1 个链出的链接,那么这个链接关系价值很大;如果一个页面有几千个链出,那么其每个链接关系价值就相对比较小。

### 2.5 查询接口

提供了多种类型的查询服务,包括短语检索、通配符、site、filetype、查询日志等等。所以我们要先通过分析传过来的查询字符串确定用户想要的查询类型,然后根据不同的查询类型获取对应的页面,这里会用之前构建是索引和数据库。然后我们获取一个结果集。我们根据之前通过 Page rank 得到的结果对这个结果集进行排序,然后将排好序的结果返回给前端。前端会采用分页机制展示搜索结果,即我们并不会一次性的将数据全部返回,而是返回部分数据给前端,等用户点击了切换页码时我们再在后台重新搜索。这样就能大大降低后台压力了。当然,因为这里我们只有一万多张页面,所以我直接是一次性搜索出来的。对于用户的每次搜索我们都会将他保存在查询日志中,包括他点击链接。这样可以帮助我们后期对页面重要性进行更新。

## 3 实验成果展示

该实验在本地完成,具体环境是 Windows10 64 位, Java1.8, Python3.6。

## 4 实验复现

对这次实验的复现比较麻烦,因为可能涉及到安装 pyLucene 比较麻烦,同时爬虫和索引构建、链接分析等等都需要耗费较长时间。如果你准备复现的话,请在前面提到过的环境下车时,然后点击根目录下的 run.bat 即可,不过需要等待  $2\sim3$  小时,之后浏览器中打开 http://localhost:8000 即可使用。

## 5 总结

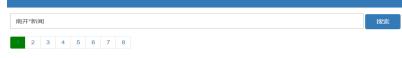
这次实验难度略大,除了需要掌握如爬虫、正则匹配等知识,更需要好的架构设计,因为 多个模块之间关系较为复杂,容易搞混。虽然这次花费了很长时间来做这个搜索引擎,但是其 功能包括准确性等都还不够,如果想要实现一个真正的好用的搜索引擎,仍旧是任重而道远。



Figure 3: 普通短语查询.



Figure 4: 支持快照.



1时 0.793 秒

南升劳训即网域风灾照 南升新间网觉指网学习网大学报校史网BBS【专题】学习贯彻党的十九大精神大会的主题是:不忘初心,牢记使命,高 举中国特色社会主义伟大旗帜,决胜全面建成小康社会。夺取新时代中国特色社会主义伟大旗利,为实现中华民族伟大 复兴的中国梦不懈奋斗南开大学召开会收退取场往几年会议在全国南校思期场往几年会议召开一周年之际,12月5日,南开大学思想政治工作会议在八国台校区学生活动中心召开。校党委书记魏大鹏出席并作深入……

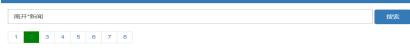
南开新闻网系列专题-专题-新闻专题网页快照 南开新闻网系列专题-专题-新闻专题首页觉悟网学习网大学根校史网BBS您当前的位置:新闻专题正文南开新闻网系列 专题来源:南开新闻网及随时间:2010-03-0710:34【专题】(习近平的七年知青岁月) 迹载【专题】高举中国特色社 会主义伟大城帜为决胜全面小秦社会实现中国参而奋斗【专题】巡视整改进行时2017-07-21【专题】2017年南开大学毕业等2017-06-29【专题】迎接十九大2017-05.....

得力于纳门则似以识别 耐开纳间首次信息公开指南信息公开目录一、学校基本信息二、招生考试信息三、财务、资产及收费 四、人事师资信息 五、教学质量信息穴、学生管理服务信息七、学风建设信息八、学位、学科信息力、对外交流与合作信十、科研及相关 树构信十一、其他最新公开信息终由请会公开校请总览南开箭介历史回路现任领导历届领导商开新闻南开新闻南南开新闻南南开大学 报南开大学电视台服务指南校长信箱南开权历办公电话后勤服务网络服务接待服务指包标30......

時到才前1月9/9/9/以2018 爾开納闻风首贞信息公开指南信息公开目录一、学校基本信息二、招生考试信息二、财务、资产及收费 四、人事师资信 息五、教学质量信息六、学生管理服务信息七、学风建设信息八、学位、学科信息力、对外交流与合作信 十、科研及相 关机构信 十、其他最新公开信息依申请之子权情总实助开筛介历史回时继现任领导的最终等饱开新闻南开新闻网南开大 学校南开大学电视台服务指南校长信箱南开校历分公电话后勤服务网络服务接待服务招投标3......

### 南开新闻 - 南开校友网网页快照

Figure 5: 支持通配符.



### 用时 0.0005 秒

南开大学电视新闻2017年12月1日--视频网页快照 南开大学电视新闻2017年12月1日--视频高页觉悟网学习网大学报校史网BBS综当前的位置: 视频正文南开大学电视新闻2017年12月1日末源: 南开新闻网发随时间: 2017-12-0116:18编辑: 資子官博专题学习贯彻党的十八届六中全会精神10月31日下午,南开大学党委召开全委扩大会议、传达学习贯彻党的十八届六中全会精神。聚集2016年网络安全宣传周今年的网络安全宣传周将在9月19日-25日举......

南开新闻网专题系列(2009年10月-2010年7月)--新闻专题网页快照 南开新闻网专题系列(2009年10月-2010年7月)--新闻专题返回首页南开大学办公招生就业教务科技图书信息门户影视 软件杀霉网关邮箱旧版南开奥闻综合校园专题评论学术人物新闻中心校报觉悟电视广播BBS成博南开新闻网系列专题南 开新闻网南开大学2012年度新闻能公园农业现集和专家评议、"认真学习规"党的十八精神"等新闻当选2012年南开大学年 度新闻。西南村那些事儿在南开大学的西南方,有一个被称为"西南村……

南开新闻网2012年专题--新闻专题网页铁照 南开新闻网2012年专题--新闻专题应回首页南开大学办公招生就业教务科技图书信息门户影视软件杀霉网关邮箱旧版南 开要调综合校园专题评论学术人物新闻中心好探觉哲电视广播图SS微博姆开新闻网系列专题南开新闻网陷开大学2012年 度新闻经过两及投票和专家评议、"认真学习惯觉它的十八精神专等和同型选2012年南开大学年度新闻。南开学子塞眼社 会实践塞假期间,南开学子前往延安、井冈山等地,开展多种形式的社会实践。

### 南开新闻网2011年专题--新闻专题网页快照

H3773mH392201平(交融。 milla(交融30x26mi 南开新闻网2011年专题-新闻专题返回首页南开大学为公招生就业数务科技图书信息门产影视软件杀毒网关邮箱旧版南 开要闻综合校园专题评论学术人物新闻中心校报觉馆电视广播BBS欧博南开新闻网系列专题南开新闻网南开大学2012年 度新闻经过两及投票和专家评议,"认真学习假觉您的十八精神"等新闻量达2012年南开大学年度新闻。南开学子塞假社 会实践塞假期间,南开学子前往延安、井冈山等地,开展多种形式的社会实践。……

南开新闻网2010年系列专题--新闻专题网页快照 南开新闻网2010年系列专题--新闻专题医回首页南开大学办公招生就业教务科技图书信息门户影视软件杀毒网关邮箱旧版南于要调缔合校园专题评论学术人物新闻中心校校觉错电视广播BS欧博南开新闻网系列专题南开新闻网南开大学2012年康新临经过网友校费和专家评议、"认真学习贯彻验的十八精神等新闻当选2012年南开大学年度新闻。南开学子举题社企学生使国期间。本证学之始技术文 计时间接触

Figure 6: 支持分页查询.

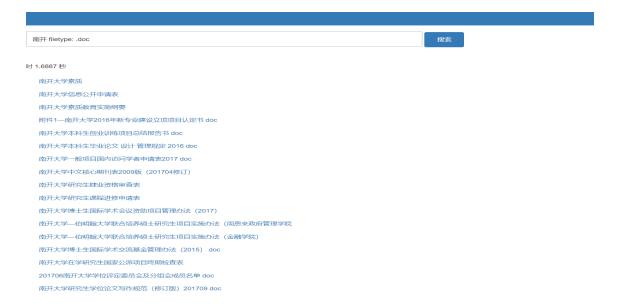


Figure 7: 支持按文件类型查找.

```
📕 query_log.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
                      query for 学生 留学
12/13/2017 09:26:
                                            used time: 0.8287 s
                      query for 学生*新闻
12/13/2017 09:26:
                                         used time: 0.1619 s
                      query for 南开*新闻
12/13/2017 09:27:
                                            used time: 0.4979 s
                      query for 南开*新闻
12/13/2017 09:27:
                                             used time: 0.0005 s
                      query for 南开 filetype: .pdf used time: 1.735 s
12/13/2017 09:27:
                      query for 南开 filetype: .doc used time: 1.815 s
12/13/2017 09:27:
                      query for 南开 filetype: .doc used time: 1.6887 s
12/13/2017 09:27:
```

Figure 8: 查询日志.

# 6 额外说明

因为不是很懂 latex 的排版原理,所以这次文档的排版中的图片位置比较奇怪,望谅解。