**大作业方案设计与实现**

**学号：2016202101**

**姓名：梁钧凯**

**班级：图灵实验班**

**大作业名称：**信息学院页面检索系统

**方案设计：**

1. 编写程序，对一个网站页面的标题，正文以及URL使用正则表达式进行提取，并用分词工具进行分词。（难点：正则表达式、分词工具的使用）
2. 编写程序，对信息学院所有网站进行爬取。方法：在Ubuntu系统下使用wget命令下载网页，对网页中的URL进行分析，把info.ruc.edu.cn域名下的网站加入队列，BFS进行爬取。（难点：多线程技术的使用）
3. 使用步骤1的程序对得到的网站进行标题和正文的提取，并以此建立倒排索引，将倒排索引保存到一个文件中。（要点：倒排索引中必要的信息的存储方式）
4. 编写程序，对每一个搜索的query构建空间向量，在倒排索引中找到结果，并使用tf-idf及一定的方法对结果排序，并输出结果。（难点：空间向量的构建，评估方法的选择）
5. 编写网页，以及cgi程序获取query，将每一个query输入步骤4得到的程序，并将结果也返回，得到一个小型的信息学院页面检索系统。（难点：cgi与网页的链接，cgi与4中程序的交互）

注：分词工具使用清华大学自然语言处理与社会人文计算实验室提供的THULAC工具包。（<http://thulac.thunlp.org/>）

**核心数据结构：**

1、对于页面，构造一个类，其中包含string存放其源代码，标题，正文；以set存放其URL列表；以vector存放其分词结果。

2、对于倒排索引，以链表形式存放与每个词相关的页面。

**核心算法及代码：**

查询算法：

query(string&input)

{

将输入切词;

for(输入中词语)

获取倒排索引;

for(倒排索引)

获取相关文档及评估相关数据;

for(相关文档)

打分;

sort(相关文档);

return result;

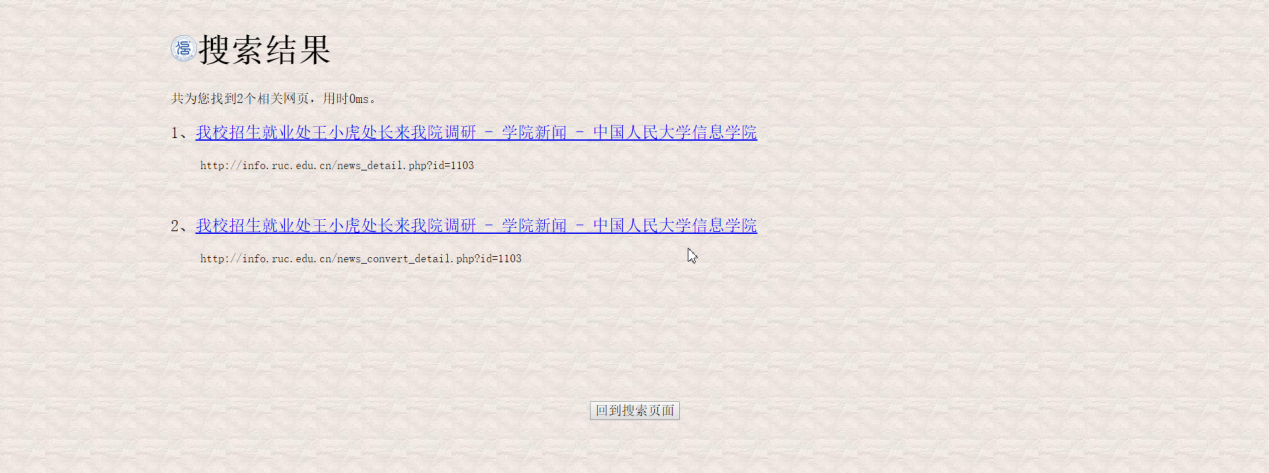
}

**效果展示：**

检索内容输入界面：



检索结果界面：



效率：对于绝大多数的查询可以在100ms内返回结果。

正确率：据已进行测试为100%。

**作业感受：**

通过这次大作业，了解到更多计算机可以做的事情。学到了正则表达式，多线程技术，信息检索等之前未接触过的知识。在遇到困难，解决困难的过程中积累了很多解决问题以及debug的经验。学会了写一些小型的项目，对于编译的原理，程序运行的原理，网页的原理有了进一步的认识。