校园网搜索引擎

1. 项目目的：使用Python建立一个适合校园网使用的Web搜索引擎系统，它能在较短的时间内爬取页面信息，具有有效、准确的中文分词功能，能实现对网上新闻的快速检索展示。
2. 项目实现步骤：
3. 网络爬虫爬取这个网站，获得所有网页链接。
4. 得到网页的源代码，解析出想要的新闻内容、标题、作者等信息。
5. 把所有网页的新闻内容做成词条索引，一般采用倒排索引。（倒排索引：倒排表以字或词为关键字进行索引，表中关键字对应的记录表项记录了出现这个字或词的所有文档，一个表项就是一个字表段，它记录该文档的ID和字符在该文档中出现的位置信息）倒排表的结构：

。。。。。

.。。。。。

文档2

文档1

单词2

.。。。。。

文档2

文档1

单词1

1. 搜索时，根据搜索词在词条索引中查询，按顺序返回相关的搜索结果，也可以按照网页评价的排名顺序返回相关的搜索结果。
2. 项目模块

* 信息采集模块：主要是利用网络爬虫实现对校园网信息的抓取；
* 索引模块：负责对爬取的新闻网页的标题、内容和作者进行分词并建立倒排词表；
* 网页排名模块：TF/IDF是一种统计方法，用于评估一字词对于一个文件集或一个词料库中的一份文件的重要程度；
* 用户搜索界面模块：负责用户关键字的输入以及搜索结果信息的返回。

1. 关键技术

* 正则表达式：将网页中的超链接提取出来，Python中的re模块，包含正则表达式的所有功能。
* 中文分词：Python中的jieba包
* 文本分类的关键词提取：jieba可以简便的提取关键词。Jieba.analyse.TFIDF().TF/IDF是一种统计方法，用于评估一字词对于一个文件集或一个语料库中的一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。主要思想：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用于分类。

1. **程序设计步骤**

5.1 信息采集模块---网络爬虫的实现

网络爬虫的实现原理及过程如下：

1. 获取初始的URL。初始的URL地址可以由用户指定的某个或某几个初始爬取网页决定；
2. 根据初始的URL爬取页面并获得新的URL。在获得初始的URL地址之后，首先需要爬取对应的URL地址中的网页，在爬取了对应URL地址中的网页后将网页存储到原始数据库中，并且在爬取网页的同时发现新的URL地址，将已经爬取过的URL地址存放在一个已爬URL列表中，用于去重复判断爬取的进程。
3. 将新的URL放到URL队列中。注意：在第2步中获取了下一个新的URL地址之后会将新的URL地址放到URL队列中。
4. 从URL队列中读取新的URL，并依据新的URL爬取网页，同时从网页中获取新URL，并重复上述的爬取过程。
5. 当满足爬虫系统设置的停止条件时停止爬取。在编写爬虫的时候一般会设置相应的停止条件，如果没有设置停止条件，爬虫会一直爬取下去，直到无法获取新的URL地址为止，若设置了停止条件，爬虫则会在停止条件满足时停止爬取。