网络搜索引擎一般分为3个组成部分

网络爬虫spider

整理搜索到的网页,建立倒序索引

提供查询服务

宽度优先 BFS

使用多叉树来表示网站

使用todo列表来记录接下来要爬取的网页

使用visited来记录已经爬取过的网页

注意爬虫黑洞(循环连接), 可使用限制深度的爬取,但是不保险

信息提取

HTML中使用正则表达式

Pdf

Doc

Video/aideo

网页中的噪声 降噪

网页中很多是网页模板或者广告其他非该网页主题相关的内容

使用统计学

分析web图

Pagerank 分析页面的重要性

PR(n) = 1-d + d\*( PR(T1)/ C1 + …. PR(Tn)/Cn)

其中PR(n)是该网页的重要性pagerank值

d是一个阻尼系数

PR(T1)是与n相关的网页1的pageranke值

C1是与n相关的网页1的链出

去掉重复的网页 去重

查询部分 分类聚类

优秀爬虫特性

高性能

可扩展性 可通过增加机器和爬虫数量来提高性能

健壮性 能防止html编码不规范 爬虫陷阱等

友好性 robots协议 Robots Exclusions protocol

robots.txt

<meta name=”robots” content=”noindex”> 禁止索引网页内容

<meta name=”robots” content=”nofollow”> 禁止抓取网页链接

分布式爬虫

1主从分布式爬虫

2,对等分布式爬虫

倒排索引

文档document

文档集合 document collection

文档编号 document ID

单词编号 word id

倒排索引 inverted index

单词词典 lexicon

倒排列表 posting list

倒排文件 inverted file

典型倒排索引的表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 单词ID | 单词 | 文档频率 | 倒排列表 |
| 1 | 谷歌 | 5 | (1;1;<1>),(2;1;<1>),…(5;1;<1>) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

第一行所示表示“谷歌”这个单词，单词ID为1，出现了5次（文档频率）， 倒排列表信息: (1;1;<1>) 表示在文件ID为1的文件中出现了1次，出现的目标行数是1

单词词典 用于维护文档集合中出现过的所有单词的相关信息，同时用来记载某个单词对应的倒排列表在倒排文件中的信息位置。在支持搜索时，根据用户的查询词，去单词词典中查询，就能够获取相应的倒排列表，并依此作为后续排序的基础。