**SITE\_ANALYZER设计文档**

黄明 赖海腾

1. **功能分析**

本次作业完成**SITE\_ANALYZER**设计，实现功能如下：

1. 爬去整个网站的页面，去掉无效页面也链接；
2. 提取网站页面的链接图；
3. 分析页面的链接关系，求page\_rank并输出要求文件。
4. **架构设计**

本系统整体架构如下图1所示，由爬取模块(crawler)，线程池(tpool)，网络图(webgraph)，链接队列(url\_queue)和主模块(main)五个主要部分组成，其中爬取模块又用到了网络模块(netwokr)，获取网页(http\_client)，提取链接(link\_parser)这3个小模块。其中：

线程池：提供n个线程，每个线程执行的都是爬取模块；

爬取模块：发送http请求，从网站获取网页，并提取其中的链接，将所得的链接分析后插进网络图和链接队列中；

网络图：存储网站的链接和边集；

链接队列：为未来得及处理的链接提供队列缓存，线程从队列中获取链接进行爬取；

主模块：调用上面4个模块，完整整体功能。

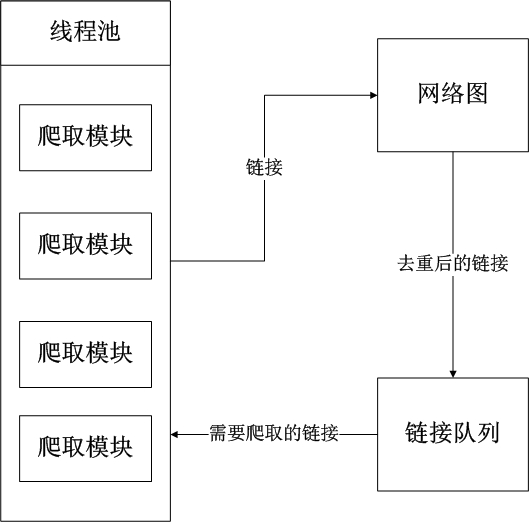


图1 架构图

**模块之间的接口：**

线程池中的爬取模块获取网页，从网页中提取链接，将链接插入到网络图中，网络图将去重后的链接插入到链接队列中，爬取模块从链接队列中获取需要爬取的链接进行爬取。

1. **核心模块设计**

本系统的核心模块主要包括网络图模块、爬取模块和提取链接模块，因为队列模块只是设计队列的一般操作（初始化，出队入队），线程池模块也只是对线程的创建回收，网络模块用于建立连接，http\_client模块用于接收服务器返回内容，这些模块都属于编程语言方面的内容，所以这里就不赘述了，下面将主要介绍网络图模块、爬取模块和提取链接模块这3个模块的设计和核心流程。

* 1. **网络图模块**

本模块主要是用于存储链接集合和链接之间形成的边集，提供了初始化图(init\_webg)，插入点(insert\_vertex)，判断一个点是否在图内(has\_vertex)，插入边和查询点大小以及销毁图和将图的信息输出到文件的这些基本操作。

1、web\_graph节点定义

struct web\_graph

{

struct node \*vertex\_set;

char \*\*all\_url\_list;

int \*\*edge\_set;

int all\_url\_list\_size;

int \*ind; //这个一维数组用来统计顶点的入度

int \*existed\_page;//这个一维数组用来统计页面是否存在（若返回200就记为1）

int existed\_page\_size;//统计一共有多少个page是存在的

int edge\_set\_size;

};

图2 webg\_graph节点

1. vertex\_set:

点集是一个hash\_table，主要用了struct node去构造这个hash\_table，struct node定义如下图3所示：

struct node

{

char \*url;

int number;//url编号，由1开始

struct node \*next\_ptr;

};//hash\_table的节点

图3 node节点

node里面的url用于存储url的具体内容，node里面的number用于存储url的编号（由1开始编号），hash\_table可以用于快速判断一个url是否在点集内，并且可以快速得到url到点编号的这个映射关系；

1. all\_url\_list:

这个表存储了所有爬取到的所有url，并根据这些url分配到的编号按顺序存储，由这个表可以得到编号到url的一个映射关系；all\_url\_list\_size则存储这个表的大小；

1. edge\_set:

存储的是每个url的出度和链出的url的编号；edge\_set\_size用于存储边集的大小；

1. ind:

用于存储所有点的入度；

1. existed\_page:

用于记录页面是否存在，当一个url对应的http响应状态是200，则对应这个url编号的位置置为1，否则置为0，从而可以得出一个url对应的页面在服务器上是否存在；existed\_page\_size用于存储所有链接对应的页面有多少是存在的。

* 1. **爬取模块**

爬取模块是整个系统的运行控制模块，从链接队列中获取任务（需要爬取的链接），根据链接获取网页，并分析网页，将网页中的链接以及对应的边信息插入到图中，将去重后的链接插入到队列中，主要的流程如下图所示：

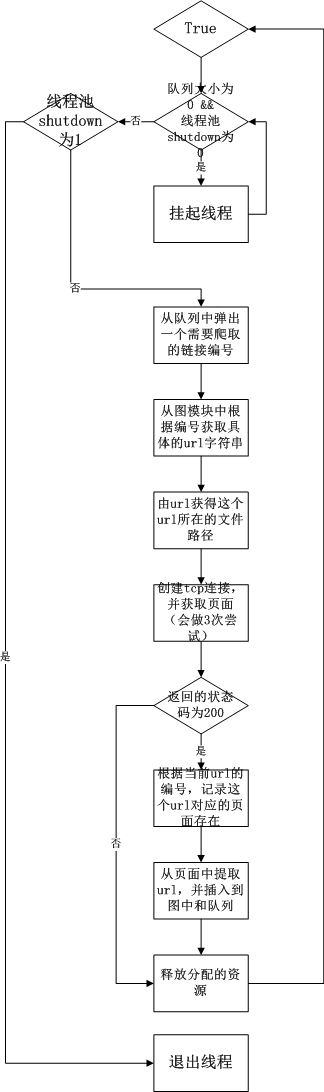


图4 爬取模块流程图

* 1. **提取链接模块**

从网页中提取链接，是根据自动机对网页中的字符逐个分析从而获得链接的，所用自动机如下图5所示：

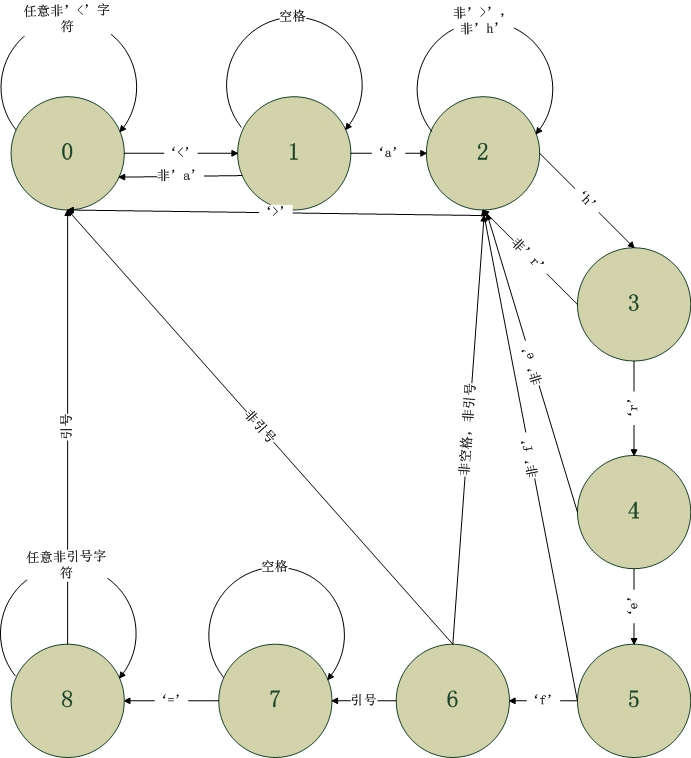


图5 提取链接自动机

只需要提取状态8中的字符就可以将链接提取出来。

1. **CDF图**

下面两个图为入度的CDF图，因为入度0-1分布很快就达到概率1，所以将图分成了两个，第一个图6为前面一部分图像，可以较为清晰地看见曲线变化，第二个图因为是全局的，所以曲线变化非常不明显。

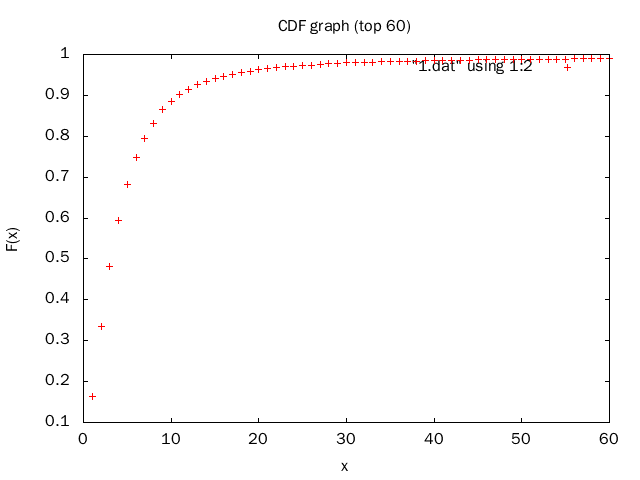


图6 IND\_CDF\_FIRST60

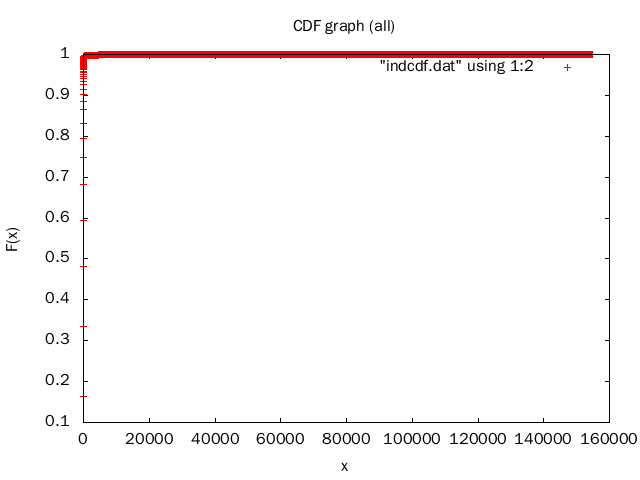


图7 IND\_CDF\_ALL