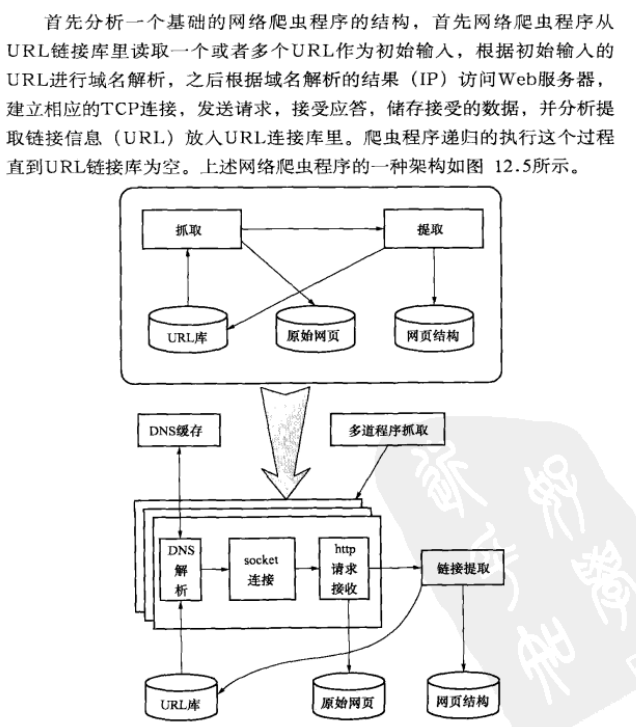
## 搜索引擎简介

包括网络爬虫模块、索引模块、搜索模块。考虑响应时间、关键词搜索和结果排序

### **网络爬虫**：

1. 通过web页面的解析，根据页面间连接关系抓取这些页面



1. 网页抓取算法：

深度、广度优先算法、基于内容算法、HITS（Hyper link Induced topic search）、PageRank（Google 基于每个页面的backlink生成PageRank，利用超级链接结构改进搜索结果）

### 索引模块：

1. 对抓取的数据进行预处理建立关键字索引
2. 分词、倒排文档（词汇表和文档集合表）

### 搜索模块：

1. 根据数据库索引和用户关键字，得到结果。
2. 结果显示（记录用户行为）+网页快照