9.6.1 大格局

Web发展的初期，搜索引擎就是一些相当简单的数据库，可以帮助用户在Web上定位文档。现在，Web上有数十亿可供访问的页面，搜索引擎已经成为因特网用户查找信息不可缺少的工具。它们在不断地发展，以应对Web庞大的规模，因此，现在已经变得相当复杂了。

面对数十亿的Web页面，和数百万要查找信息的用户，搜索引擎要用复杂的爬虫来获取这数十亿Web页面，还要使用复杂的查询引擎来处理数百万用户产生的查询负荷。

我们来考虑一下产品级Web爬虫的任务，它要获取搜索索引所需的页面，它要发出数十亿条HTTP请求。如果每条请求都要花半秒钟的时间（对有些服务器来说可能慢了，对另一些服务器来说可能快了[27]），（对十亿份文件来说）就要花费：

0.5秒 × （100 0000 000）/（60秒/天）×（60分/小时）×（24小时/天）

如果请求是连续发出的，结果差不多是5700天！很显然，大型爬虫得更聪明一些，要对请求进行并行处理，并使用大量机器来完成这项任务。但由于其规模庞大，爬行整个Web仍然是件十分艰巨的任务。

9.6.2 现代搜索引擎结构

现在的搜索引擎都构建了一些名为“全文索引”的复杂本地数据库，装载了全世界的Web页面，以及这些页面所包含的内容。这些索引就像Web上所有文档的卡片目录一样。

242

搜索引擎爬虫会搜集Web页面，把它们带回家，并将其添加到全文索引中去。同时，搜索引擎用户会通过HotBot（http://www.hotbot.com）或Google（http://www.google.com）这样的Web搜索网关对全文索引进行查询。Web页面总是在不断地发生变化，而且爬行一大块Web要花费很长的时间，所以全文索引充其量也就是Web的一个快照。

现代搜索引擎的高层结构如图9-7所示。

9.6.3 全文索引

全文索引就是一个数据库，给它一个单词，它可以立即提供包含那个单词的所有文档。创建了索引之后，就不需要对文档自身进行扫描了。

注27：这取决于服务器的资源、客户端的机器人，以及两者之间的网络状况。