有些大型 Web 机器人会使用机器人“集群”，每个独立的计算机是一个机器人，以汇接方式工作。为每个机器人分配一个特定的 URL“片”，由其负责爬行。这些机器人配合工作，爬行整个 Web。机器人个体之间可能需要相互通信，来回传送 URL，以覆盖出故障的对等实体的爬行范围，或协调其工作。

Witten 等人编写的 *Managing Gigabytes: Compressing and Indexing Documents and Images*（《海量数据管理——文档和图像的压缩与索引》）[4]，*Morgan Kaufmann* 出版社出版，是实现大规模数据结构的很好的参考书。这本书讲的全是管理大量数据所需的各种诀窍和技巧。

9.1.6 别名与机器人环路

由于 URL“别名”的存在，即使使用了正确的数据结构，有时也很难分辨出以前是否访问过某个页面。如果两个 URL 看起来不一样，但实际指向的是同一资源，就称这两个 URL 互为“别名”。

表 9-1 列出了不同 URL 指向同一资源的几种简单方式。

表9-1 同一文档的不同URL别名

	第一个URL	第二个URL	什么时候是别名
a	http://www.foo.com/bar.html	http://www.foo.com:80/bar.html	默认端口为 80
b	http://www.foo.com/~fred	http://www.foo.com/%7Ffred	%7F 与 ~ 相同
c	http://www.foo.com/x.html#early	http://www.foo.com/x.html#middle	标签并没有修改页面内容
d	http://www.foo.com/readme.htm	http://www.foo.com/README.HTM	服务器是大小写无关的
e	http://www.foo.com/	http://www.foo.com/index.html	默认页面为 index.html
219 f	http://www.foo.com/index.html	http://209.231.87.45/index.html	www.foo.com 使用这个 IP 地址

9.1.7 规范化URL

大多数 Web 机器人都试图通过将 URL“规范化”为标准格式来消除上面那些显而易见的别名。机器人首先可先通过下列步骤将每个 URL 都转化为规范化的格式。

(1) 如果没有指定端口的话，就向主机名中添加“:80”。
(2) 将所有转义符 %xx 都转换成等价字符。
(3) 删除 # 标签。

注 4：本书中文版已由科学出版社出版。（编者注）