9.1.3 避免环路的出现

机器人在 Web 上爬行时，要特别小心不要陷入循环，或环路（cycle）之中。我们来看看图 9-2 中所示的爬虫。

- 在图 9-2a 中，机器人获取页面 A，看到 A 链接到 B，就获取页面 B。
- 在图 9-2b 中，机器人获取页面 B，看到 B 链接到 C，就获取页面 C。
- 在图 9-2c 中，机器人获取页面 C，会看到 C 链接到 A。如果机器人再次获取页面 A，就会陷入一个环路中，获取 A、B、C、A、B、C、A……

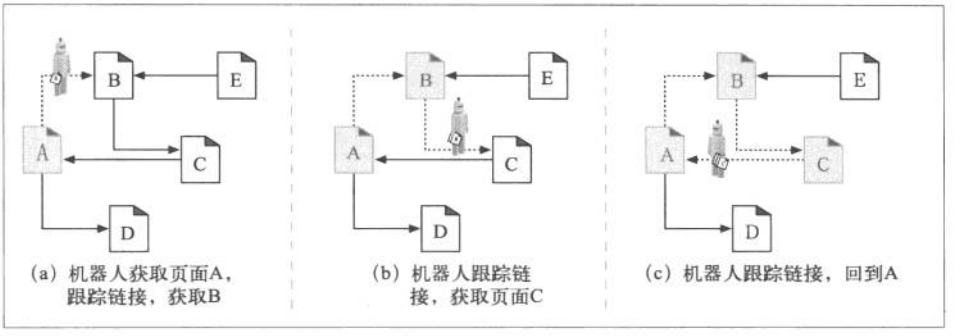

图 9-2 在 Web 的超链中爬行

机器人必须知道它们到过何处，以避免环路的出现。环路会造成机器人陷阱，这些陷阱会暂停或减缓机器人的爬行进程。

9.1.4 循环与复制

至少出于下列三个原因，环路对爬虫来说是有害的。

- 它们会使爬虫陷入可能会将其困住的循环之中。循环会使未经良好设计的爬虫不停地兜圈子，把所有时间都耗费在不停地获取相同的页面上。爬虫会消耗掉很多网络带宽，可能完全无法获取任何其他页面了。
- 爬虫不断地获取相同的页面时，另一端的 Web 服务器也在遭受着打击。如果爬虫与服务器连接良好，它就会击垮 Web 站点，阻止所有真实用户访问这个站点。这种拒绝服务是可以作为法律诉讼理由的。
- 即使循环自身不是什么问题，爬虫也是在获取大量重复的页面 [通常被称为“dups”（重复），以便与“loops”（循环）押韵]。爬虫应用程序会被重复的内容所充斥，这样应用程序就会变得毫无用处。返回数百份完全相同页面的因特网搜索引擎就是一个这样的例子。

217