要在图 9-1 所示的 Web 上爬行，使用哪个根集比较好呢？与在真实的 Web 中一样，没有哪个文档最终可以链接到所有其他文档上去。如果从图 9-1 的文档 A 开始，可以到达 B、C 和 D，然后是 E 和 F，然后到 J，然后到 K。但没有从 A 到 G，或从 A 到 N 的链路。

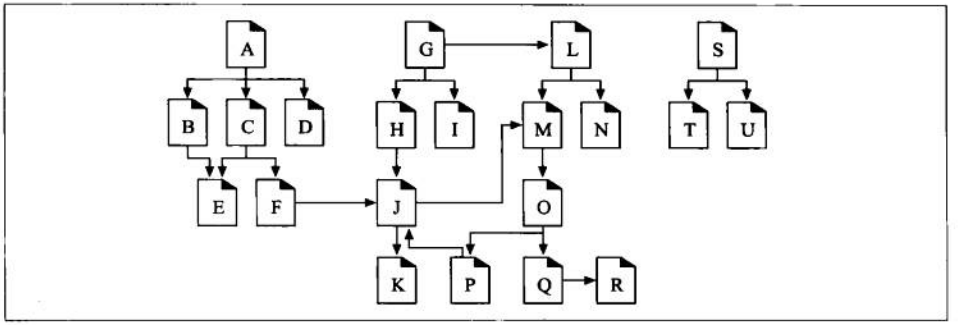

图 9-1　根集要能够到达所有的页面

这个 Web 结构中的某些 Web 页面，比如 S、T 和 U，几乎是被隔离开来的——它们是孤立的，没有任何链接指向它们。可能这些孤独的页面是一些新页面，还没人找到它们。或者可能是一些非常老的或不显眼的页面。

总之，根集中并不需要有很多页面，就可以涵盖一大片 Web 结构。在图 9-1 中，要抵达所有页面，根集中只需要有 A、G 和 S 就行了。

通常，一个好的根集会包括一些大的流行 Web 站点（比如 http://www.yahoo.com）、一个新创建页面的列表和一个不经常被链接的无名页面列表。很多大规模的爬虫产品，比如因特网搜索引擎使用的那些爬虫，都为用户提供了向根集中提交新页面或无名页面的方式。这个根集会随时间推移而增长，是所有新爬虫的种子列表。 216

9.1.2　链接的提取以及相对链接的标准化

爬虫在 Web 上移动时，会不停地对 HTML 页面进行解析。它要对所解析的每个页面上的 URL 链接进行分析，并将这些链接添加到需要爬行的页面列表中去。随着爬虫的前进，当其发现需要探查的新链接时，这个列表常常会迅速地扩张。[2] 爬虫要通过简单的 HTML 解析，将这些链接提取出来，并将相对 URL 转换为绝对形式。2.3.1 节讨论了如何进行这种转换。

注 2：我们会在 9.1.3 节开始讨论爬虫是否需要记住它们到过何处。在爬行过程中，这个已发现 URL 列表会不断扩张，直到已经对 Web 空间进行了彻底的探查为止，这时爬虫就会到达一个不再发现新链接的状态了。