

| | |
|--------------------------------|-----|
| 第9章 Web 机器人 | 225 |
| 9.1 爬虫及爬行方式 | 226 |
| 9.1.1 从哪儿开始：根集 | 226 |
| 9.1.2 链接的提取以及相对链接的标准化 | 227 |
| 9.1.3 避免环路的出现 | 228 |
| 9.1.4 循环与复制 | 228 |
| 9.1.5 面包屑留下的痕迹 | 229 |
| 9.1.6 别名与机器人环路 | 230 |
| 9.1.7 规范化 URL | 230 |
| 9.1.8 文件系统连接环路 | 231 |
| 9.1.9 动态虚拟 Web 空间 | 232 |
| 9.1.10 避免循环和重复 | 233 |
| 9.2 机器人的 HTTP | 236 |
| 9.2.1 识别请求首部 | 236 |
| 9.2.2 虚拟主机 | 236 |
| 9.2.3 条件请求 | 237 |
| 9.2.4 对响应的处理 | 238 |
| 9.2.5 User-Agent 导向 | 239 |
| 9.3 行为不当的机器人 | 239 |
| 9.4 拒绝机器人访问 | 240 |
| 9.4.1 拒绝机器人访问标准 | 241 |
| 9.4.2 Web 站点和 robots.txt 文件 | 242 |
| 9.4.3 robots.txt 文件的格式 | 243 |
| 9.4.4 其他有关 robots.txt 的知识 | 246 |
| 9.4.5 缓存和 robots.txt 的过期 | 246 |
| 9.4.6 拒绝机器人访问的 Perl 代码 | 246 |
| 9.4.7 HTML 的 robot-control 元标签 | 249 |
| 9.5 机器人的规范 | 251 |
| 9.6 搜索引擎 | 254 |
| 9.6.1 大格局 | 255 |
| 9.6.2 现代搜索引擎结构 | 255 |
| 9.6.3 全文索引 | 255 |
| 9.6.4 发布查询请求 | 257 |
| 9.6.5 对结果进行排序，并提供查询结果 | 258 |
| 9.6.6 欺诈 | 258 |
| 9.7 更多信息 | 258 |
| 第10章 HTTP-NG | 261 |
| 10.1 HTTP 发展中存在的问题 | 262 |
| 10.2 HTTP-NG 的活动 | 263 |