9.2 机器人的HTTP

机器人和所有其他 HTTP 客户端程序并没有什么区别。它们也要遵守 HTTP 规范中的规则。发出 HTTP 请求并将自己广播成 HTTP/1.1 客户端的机器人也要使用正确的 HTTP 请求首部。

很多机器人都试图只实现请求它们所查找内容所需的最小 HTTP 集。这会引发一些问题；但短期内这种行为不会发生什么改变。结果就是，很多机器人发出的都是 HTTP/1.0 请求，因为这个协议的要求很少。

9.2.1 识别请求首部

尽管机器人倾向于只支持最小的 HTTP 集，但大部分机器人确实实现并发送了一些识别首部——最值得一提的就是 `User-Agent` 首部。建议机器人实现者们发送一些基本的首部信息，以通知各站点机器人的能力、机器人的标识符，以及它是从何处起源的。

在追踪错误爬虫的所有者，以及向服务器提供机器人所能处理的内容类型时，这些信息都是很有用的。鼓励机器人实现者们使用的基本识别首部包括如下内容。

- `User-Agent`
  将发起请求的机器人名字告知服务器。
- `From`
  提供机器人的用户 / 管理者的 E-mail 地址。[9]
- `Accept`
  告知服务器可以发送哪些媒体类型。[10] 这有助于确保机器人只接收它感兴趣的内容（文本、图片等）。
- `Referer`
  提供包含了当前请求 URL 的文档的 URL。[11]

9.2.2 虚拟主机

225 机器人实现者要支持 `Host` 首部。随着虚拟主机（参见第 5 章）的流行，请求中不

注 9：一种 RFC 822 E-mail 地址格式。

注 10：3.5.2 节列出了所有 `Accept` 相关的首部；机器人可能会发现，如果它们对特定版本感兴趣的话，发送 `Accept-Charset` 之类的首部是很有帮助的。

注 11：有些站点管理者会尝试着记录机器人是如何找到指向其站点内容的链接的，对这些人来说，这个首部非常有用。