<<搜索引擎基础教程>>

图书基本信息

书名：<<搜索引擎基础教程>>

13位ISBN编号：9787302220497

10位ISBN编号：7302220492

出版时间：2010-7

出版时间：清华大学出版社

作者：袁津生，李群

页数：320

<<搜索引擎基础教程>>

内容概要

本书从教学的角度出发，对搜索引擎的原理及开发技术进行了全面的介绍，内容包括搜索引擎的基本原理、网页抓取技术、信息预处理技术、信息索引技术、信息查询技术和多媒体信息检索技术。另外，本书还对搜索引擎开发技术进行了详细的讨论。

本书适合高等院校计算机科学与技术专业及相关专业的高年级学生和研究生阅读参考，也适合相关领域的工程技术人员参阅。

<<搜索引擎基础教程>>

书籍目录

章节摘录

插图：在抓取过程中，可以使用多个爬虫来合作抓取，这样可以进一步降低每个爬虫的用于记录历史抓取情况的哈希表大小。
如果有n个爬虫，则可将哈希表继续压缩到原有大小的行分之一。
如果n个爬虫分别运行在不同的机器上，那么每个机器被哈希表占用的内存用量将非常少；通常保持抓取历史记录所需要的内存在百兆字节左右是恰当的。
通过不重复抓取的方法初步解决了死循环的问题，即抓过的不再抓。
然而实际操作中还有这样的一个问题，如果任意两个网页存在链接，则链接它们的最短路径值为17。
这样，爬虫无论用何种遍历方法都不能保证一定会按照最佳路径抓取每一个网页，因为任何一个网页都可能从多个种子站点开始广度优先被遍历到。
为了防止爬虫无限制的广度优先抓取，必须在某个深度上进行限制。
到达这个深度后就应该停止抓取，这个深度的取值就是万维网直径长度。
当在最大深度上停止时，那些深度过大的未抓网页，总是期望可以从其他种子站点更加经济地到达。
例如，种子站点B和C在抓取到深度为17的时候，立即停止抓取，把抓取剩余网页的机会留给从种子站点A出发的进行抓取工作的爬虫。
此外，深度优先策略和广度优先策略的组合可以有效地保证抓取过程中的封闭性。
即在抓取过程（遍历路径）中总是在抓取相同域名下的网页，而很少出现其他域名下的网页。

版权说明