中南大學学报 自然科学版 英 文 版

教育部 主管
中南大学 主办

论文摘要

中南大学学报(自然科学版)

ZHONGNAN DAXUE XUEBAO(ZIRAN KEXUE BAN)

Vol.32 No.3 Jun.2001

[PDF全文下载] [全文在线阅读]

文章编号：1005-9792(2001)03-0325-03

Web Robot技术及其Java实现

谭淑英，刘丽华

(中南大学信息科学与工程学院，湖南长沙 410083)

摘 要：WWW环球信息呈指数级增长，使WWW成为全球最大的信息系统，研究其中的信息搜索工具具有现实意义.Web Robot是搜索引擎中的核心部分，它从给定的统一资源地址开始分析，递归地搜索新的Web文档.作者论述了Web Robot的工作原理以及机器人排斥标准，用Java实现了Web文档的下载、超链提取、新超链的可用性判断和访问站点的安全性检查，为提高Web Robot的效率提出了2种解决途径，即采用Java多线程处理技术和集群式Robot.此外，给出了用Java多线程处理技术提高效率的算法，对网站建设和信息搜索工具的开发具有一定的参考价值.

关键字：Robot；机器人排斥标准；JAVA多线程；信息搜索

The Web robot technique and implementation with Java

TAN Shu-ying,LIULi-hua

（College of Information Science and Engineering, Central South University, Changsha 410083, China）

Abstract: Due to its exponential growth, WWW is turned into the largest information system. It is becoming a hot point to study the information retrieval tool on WWW. Web Robot is the kernel of searching engine. It starts its analysis from the given URL, and searches new Web files recursively. In this paper the working principle of Web Robot and SRE (Standard for Robots Exclusion) are introduced, and the Web file download, hyperlink extracting, usability judging of new hyperlinks and accessing security checking of Web site are implemented with Java language. In order to improve the robot's searching efficiency, two methods are put forward. One is Java multi-thread processing , and the other is a cluster of robots working together. And an algorithmusing multithread to improve the efficiency is given. This paper has certain reference value for the Web site constructor and the developer of information retrieval tool.

Key words:robot; SRE; Java multi-thread; information retrieval

地 址：湖南省长沙市中南大学　　邮编： 410083
电 话： 0731-88879765 传真： 0731-88877727
电子邮箱： zngdxb@mail.csu.edu.cn 湘ICP备09001153号