可配置 Web Robot 的研究与实现

郑莉霞[1]，刘连芳[1,2]
（1.广西大学 计算机与电子信息学院，广西 南宁 530004；
2.广西南宁平方软件公司，广西 南宁 530004）

摘 要：针对个性化搜索需要多种 Web Robot 支持工作的需求，在分析 Web Robot 工作原理的基础上，利用面向对象的分析设计方法，提出了一种可实现增量开发的 Web Robot 的系统模型，并经过了充分的实验验证。实验结果表明应用此模型可以灵活控制 Web Robot 的搜索策略，通过合理扩展可适用于不同类型的 Web 资源搜索，能够根据不同的个性化搜索需求灵活定制相应的 Web Robot，有效地节约了开发 Web Robot 的各项成本。

关键词：Web Robot；搜索引擎；Web 信息采集
中图分类号：TP311.52 **文献标识码**：A **文章编号**：1673-629X(2007)06-0083-04

Research and Realization on Web Robot

ZHENG Li-xia[1], LIU Lian-fang[1,2]
(1. School of Computer and Electronic Information, Guangxi University, Nanning 530004, China;
2. Pingsoft New Technology Co., Ltd, Nanning 530004, China)

Abstract: Based on a thorough analysis of the working principle of Web Robot and through object-oriented analysis and design, proposes and demonstrates a Web Robot systematic model that enables the incremental software development to meet the needs of multifunctional Web Robot that supports individual searching. Experimental results indicate that the model could be used to achieve a flexible control of the searching strategy of Web Robot and are adaptable to various types of Web resource searching upon suitable extension. According to different needs of individual searching, corresponding specific Web Robot could be readily tailor-made, which effectively reduces the cost of the Web Robot development.

Key words: Web Robot; searching engine; Web information collection

0 引 言

Web 信息规模的持续性增长使搜索 Web 信息变得日益困难，为了处理这类问题，研究者开发了许多自动检索 Web 页面信息的软件程序，这些程序被称为：网上机器人（Web Robot）、蜘蛛（Spider）、爬行者（Crawler）、网络代理（Web agent）、漫游者（Wanderer）、蠕虫（Worm）等。这里主要讨论“Web Robot”。

Web Robot 是一种通过 HTTP 协议获取远程站点中的 Web 文档信息，并根据其中的超文本链接递归遍历整个 WWW 信息空间的软件程序[1,2]。Web Robot 的研究开始于 20 世纪 80 年代末，1993 年诞生了世界上第一个 Web Robot，被命名为 Wanderer[3]。随着 WWW 的广泛应用，Web Robot 日益引起研究者的关注。目前，Web Robot 广泛应用于各类提供 Web 服务的系统中，成为实现这些 Web 服务的前提和基础。

搜索引擎是 Web Robot 最广泛应用的领域之一，几乎所有的搜索引擎都要依赖 Web Robot 收集 Web 页面信息后，并进行后续的超链分析处理，最后创建索引才能向用户提供有效的搜索服务[2,4,5]。近些年来，由于网络资源的多样化等发展趋势，导致网络搜索服务也朝着多样化、专业化、个性化的方向发展，这直接要求 Web Robot 做出反应以适应新的环境需求。以具有‘搜索老大’之称的 Google 为例，其使用的 Web Robot 从最初用于抓取普通网页的 GoogleBot，发展分化成具有不同功能的多种 Web Robot，有用于分析投放广告网页的 MediaBot、用于分析图片的 ImageBot，用于抓取各种 RSS Feed 的 FeedFetcher-Google 等等。每开发一种 Web Robot 都需要投入大量的人力、物力和财力，因此需要一种有效的增量开发方式，以节省开发成本。

文中应用面向对象的设计方法，在分析普通 Web

收稿日期：2006-08-30
作者简介：郑莉霞（1980-），女，江西上饶人，硕士研究生，研究方向为数据库理论、Web 搜索；刘连芳，研究员，研究方向为数据库、Web 搜索、超媒体、中文信息处理。

Robot 工作的基础上，将 Web Robot 的整体架构划分成结构独立的三层架构，并根据此三层架构实现了可灵活配置的 Web Robot 系统，实验表明，我们所设计的可配置 Web Robot 可根据搜索环境的变化需求，通过开发适当的插件和处理规则，能够以增量方式实现特定 Web 资源的搜索。

1 可配置 Robot 的分析设计

Web Robot 是各类搜索服务实现的前提和基础，其以用户指定的种子链接作为起始遍历点，模拟人类访问 Web 的方式，利用像 HTTP 这样的协议读取相应的页面信息，然后解析页面中包含的链接，并以此作为新的访问起点并递归实现同样的漫游，直到无满足条件的 URL 存在为止。如果将 WWW 抽象成以 Web 页面为顶点，以超链接为有向边的巨型网 $G(V,E)$，那么 Robot 的遍历搜索过程即可以抽象成树搜索方式，也就是以某个树结点为起点，以有向边为方向，可以在此树中实现深度优先遍历或广度优先遍历，最终遍历得到树中所有的资源。其中广度优先遍历算法广泛应用于 Web Robot 中[6]。

1.1 可配置 Web Robot 的分析与设计

为了根据实际需求灵活配置 Web Robot 的搜索策略，应用面向对象的分析方法[7]，将 Robot 的核心工作任务抽象成两类：一类为遍历工作，其根据指定的超链接，通过协议获取其页面信息；另一类为解析工作，其根据遍历所得到的 Web 页面信息，解析其中包含的超链接并以此链接作为起点实现递归遍历工作。为了灵活控制 Robot 的行为，需要设计适宜的规则扩充系统，并在任务执行之前根据用户指定的规则做出是否执行 Robot 各项任务的判断。

根据以上分析，将 Robot 的任务进行再次抽象，一类为决策型的工作任务，其根据用户定义的各项规则，做出执行相应任务的判断，其由遍历任务决策（DecideOnWalkingTask）和解析任务决策（DecideOnParsingTask）组成；通过执行决策型的工作任务，会产生核心工作任务的两个队列，一类为遍历获取页面信息的工作任务（WalkingHttpURLTask），另一类则为解析获取页面中引用的超链接任务（InterpreteHtmlTask），并递归调用相应决策以执行任务。这四组任务的执行流程如图 1 所示。决策任务的实现可根据需要，由用户指定规则给予控制，同时，不同用户可根据实际需要自定义规则以实现任务执行的有效控制，其设计如图 2 所示。

Robot 作为搜索引擎实现的基础，其主要作用不仅在于获得原始页面信息，同时还需要在获取过程当中做出即时的数据统计，对链接响应做出不同的处理等等，这些信息的统计和捕获是后续超链分析的基础，是搜索引擎实现的前提。

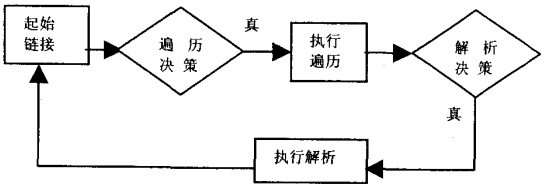

图 1 核心任务工作流程图

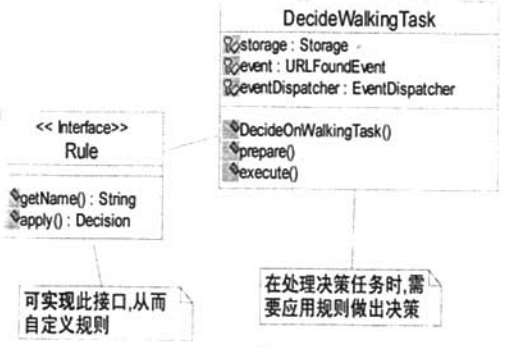

图 2 决策处理

根据以上分析，若要灵活控制工作任务的执行过程和结果处理方式，需要对工作任务进行再次分解，得到每个工作任务的执行环节，并确保这些环节是不可再分的最小单元。由此，对每个工作任务的执行环节进行了详细分析，并将不可再分的环节抽象为事件，通过对事件进行统一调度，根据用户定制的事件处理配置情况，实现不同事件的灵活处理。

事件处理由用户指定配置的插件负责，用户可根据自己的需求，针对不同的事件进行自定义插件的开发。其具体设计如图 3 所示，通过扩充插件接口，可以灵活配置 Robot 的事件处理能力。

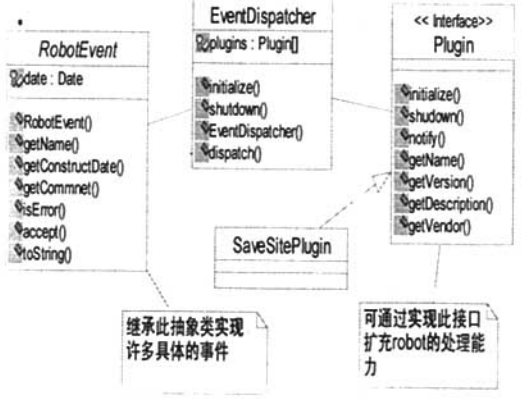

图 3 Robot 事件处理

为了真正做到 Robot 的灵活配置，将整个 Robot 工作环境的配置都置入到配置文件当中，配置文件的

呈现格式为:'AttributeName = AttributeValue',可根据需要重复使用,用户利用这种形式定义系统配置,并将其保存为文本文件以备初始化整个 Robot 的工作环境。

在系统初始化时,可以生成指向配置文件的输入流,并通过 java. util. ResourceBundle 类将配置文件以键值对的形式导入至系统中[7],由此可根据这些定义的键值对实例化指定进行事件处理的插件、配置 Robot 遍历和解析时应用的处理规则,进行存储配置和并发工作控制配置等,从而实现 Robot 运行环境的构建,等待用户的操作指令。

1.2　系统总体架构

由以上分析,可将可配置的 Robot 系统的总体架构分为三层:用户服务提供接口层(SPI, Service Provider Interface)、应用程序接口层(API, Application Programming Interface)及核心工作层(Core)。整体框架如图 4 所示。

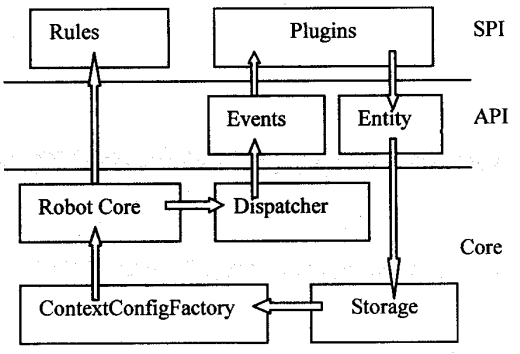

图 4　系统总体框架

通过实现服务提供接口层中的规则和插件接口,可以灵活地扩充控制 Robot 工作的行为规则和事件处理机制。同时,系统的核心部分的驱动来自于可方便定义的配置文件,通过环境配置工厂读取系统配置文件,实现整个 Robot 工作系统的灵活配置。

1.3　算法描述

可配置的 Robot 实现算法如下:

1) 构建指向环境配置文件的输入流,读取配置文件;

2) 根据配置文件说明初始化 Robot 的工作环境:

①根据配置文件中的插件说明,实例化需要设置的插件;

②构造插件调度器实现事件调度管理;

③进行存储配置;

④进行规则控制配置;

⑤根据两类工作任务构造两个可并行工作的线程池;

3)根据用户指定的种子 URL,启动相应的工作任务,并将处理结果放置到两个可并行工作的队列中(遍历和解析队列);

4)根据线程池的线程调度情况,并发处理相应工作队列中的工作任务,并根据情况,及时进行事件处理;

5)检查相应工作队列,实现递归处理或终止。

2　实验结果分析

根据上述设计分析,采用 Java 平台实现了可配置的 Web Robot 系统,系统采用 MySQL 数据库存储采集信息,实现可灵活配置、高度并发的 Web 信息搜索。

我们在处理器为 Intel(R) Pentium(R) M1.6GHz,内存储器为 256M,操作系统为 WinXP 的 PC 机上进行了单机模拟实验,在平均链接带宽为 2MBPS 的情况下,针对两个可并发处理的线程池中的不同线程数的比例,在指定起始链接为 http://www.sohu.com,分别进行了'普通搜索'和'Web 图像资源搜索'两组实验,实验结果说明,采用以上所述的方案可行有效,图 5 为 Robot 的工作性能测试结果,从实验结果来看,walker 与 thinker 线程池的大小比例为 10:4 时,性能最佳。这主要是因为遍历与解析工作的工作量比例不同造成的,通过调整线程池大小的比例,使工作负载较平衡,以此可改善并发工作的性能。此外,整个系统的工作性能还受到网络带宽的影响。

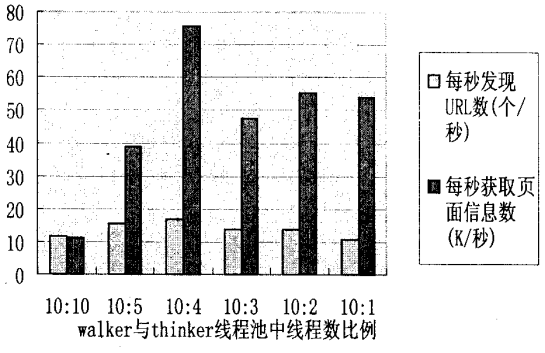

图 5　Robot 工作性能

图 6 显示在线程池大小比例为 10:4 时,在 Robot 系统中配置图像资源搜索插件后的部分搜索结果,从图中可以看到,利用图像搜索插件机制可以将 Web 图像以主机域名为目录,按站点原始目录的形式将图像镜像到本地存储器中。

3　总结与展望

随着 Web 信息的不断增长和搜索引擎应用的不断深入,大量的个性化信息搜索引擎的诞生需要对 Web

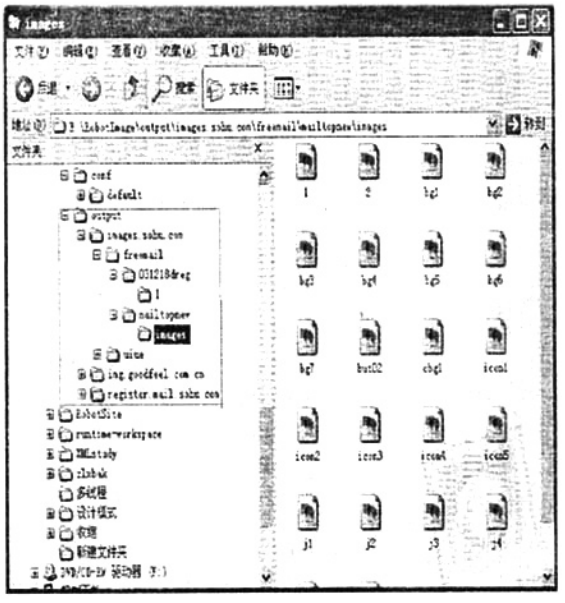

图 6 图像搜索插件工作结果

Robot 工作模式进行灵活的配置和调整，以使信息的采集具有针对性，从而降低搜索服务运营的整体开销。我们所设计的可配置 Robot 系统，在继承常规 Robot 系统的基础上，通过应用面向对象的设计思想，构建了可扩充的服务提供接口，可以使 Robot 根据实际需求进行灵活的控制。实验表明此系统的设计方案可行有效，通过开发所需要的插件和规则，即可以实现不同 Web 资源的搜索，使其灵活应用于各种搜索环境，实现了增量式开发，节省了开发成本。

目前，此可配置 Robot 系统还存在许多需要改进之处，如：缺乏动态性能的监测和调整、未能构建分布式存储机制等，在下一个阶段，将针对这些方面展开工作，使系统得到完善。

参考文献：

[1] Koster M. Robots in the web: threat or treat? [EB/OL]. 1995-04[2006-07-02]. http://www.robotstxt.org./wc/threat-or-treat.html.

[2] Chau M, Chen HsinChum. Personalized and Focused Web Spiders[EB/OL]. 2003[2006-06-20]. http://citeseer.ist.psu.edu/548327.html.

[3] Gray M. Internet Growth and Statistics: Credits and Background[EB/OL]. [2006-06-20]. http://www.mit.edu/-mkgray/net/background.html.

[4] Chau M, Zeng D, Chen Hinchun. Personalized spiders for web search and analysis[C]// Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries. New York, USA: ACM press, 2001:79-87.

[5] Arasu A, Cho Junghoo, Garcia-Molina H, et al. Searching the Web[J]. ACM Transactions on Internet Technology, 2001(1):2-43.

[6] Menczer F, Pant G, Srinivasan P. Topical web crawlers: Evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology, 2004(4):378-419.

[7] 埃克尔. Java 编程思想[M]. 陈昊鹏，等译. 北京：机械工业出版社，2005.

(上接第 82 页)

association rules to correlations[C]// Proceeding of the ACM-SIGMOD Conference 1997. New York: ACM Press, 1997:265-276.

[3] Savasere A, Omiecinski E, Navathe S. Mining for strong negative associations in a large database of customer transaction[C]// Proceedings of the IEEE 14th International Conference on Data Engineering. Los Alamitos: IEEE-CS, 1998:494-502.

[4] Wu Xindong, Zhang Chengqi, Zhang Shichao. Mining both positive and negative association rules[C]// Proceedings of the 19th International Conference on Machine Learning(ICML-2002). San Francisco: Morgan Kaufmann Publishers, 2002:658-665.

[5] 周 欣，沙朝锋，朱央勇，等. 兴趣度——关联规则的又一个阈值[J]. 计算机研究与发展，2000，37(5):627-633.

[6] Srikant R, Agrawal R. Mining quantitative association rules in large relational tables[C]// Proceedings of the 1996 ACM SIGMOD international conference on Management of Data. Montreal, Canada: [s.n.], 1996:1-2.

[7] Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations[C]// Proceedings of the 1997 ACM SIGMOD international conference on Management of Data. Tucson, USA: [s.n.], 1997:265-276.

[8] Brin S, Motwani R, Ullman J D, et al. Dynamic itemset counting and implication rules for market basket data[C]// Proceedings of the 1997 ACM SIGMOD international conference on Management of Data. Tucson, USA: [s.n.], 1997:255-264.

[9] Aggarwal C C, Yu P S. Online generation of association rules, Data Engineering[C]// 1998 Proceedings of the IEEE 14th International conference on Data Engineering. Orlando, Florida, USA: [s.n.], 1998:402-411.

[10] 董祥军，王舒静，宋瀚涛，等. 负关联规则的研究[J]. 计算机工程与应用，2004，40(11):978-981.