文章编号:1008-0570(2010)01-3-0136-02

一种新型网络爬虫的设计与实现

Design & Implementation of a New Type Web Crawler

(1.山东理工大学;2.淄博实验中学) 王江红[1] 朱丽君[2] 李彩虹[1]
WANG Jiang-hong ZHU Li-jun LI Cai-hong

摘要: 网络爬虫是当今网络实时更新和搜索引擎技术的共同产物。文中深入探讨了如何应用网络爬虫技术实现实时更新数据和搜索引擎技术。在对网络爬虫技术进行深入分析的基础上,给出了一种用网络爬虫技术实现局域网内服务器和客户端之间网络通信的解决方案。

关键词: Socket; Http; 网络爬虫; 客户端/服务器
中图分类号: TP391 **文献标识码:** A

Abstract: The web crawler is a common product of the network real-time refresh data and search engine technology at present. This article discusses and studies thoroughly how to apply the Web Crawler technique to realize the real-time refresh data and search engine technology. On the basis of deep analysis to the Web Crawler technique, this article gives a kind of solution to realize network communications between the server and the client in the local area network with the Web Crawler technique.

Key words: Socket; Http; Web Crawler; Client/Server

技术创新

1 引言

随着网络的迅速发展,万维网成为大量信息的载体,而万维网可以看作是一个分布式动态快速增长的由各类文档组成的海量信息资源中心,其信息量呈几何指数增长,如何有效地提取并利用这些信息成为一个巨大的挑战。搜索引擎 (Search Engine),例如传统的通用搜索引擎 AltaVista,Yahoo! 和 Google 等,作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。在这样的背景下,人们提出了"网络爬虫技术"的概念并通过一定的技术得以实现。

2 系统开发背景

2.1 搜索引擎

搜索引擎是一种能够通过 Internet 接受用户的查询指令并向用户提供符合其查询要求的信息资源网址系统。它是一些在 WEB 中主动搜索信息(网页上的单词和特定的描述内容)并将其自动索引的 WEB 网站, 其索引内容存储在可供检索的大型数据库中,建立索引和目录服务。

搜索引擎也是目前 Internet 对信息资源进行组织的主要方式。搜索引擎由网上机器人(Spider 或 Rooter)自动在网页上按某种策略进行远程数据的搜索和获取,并生成本地索引。Spider 或 Rooter 是一种软件,它沿着 WWW 文件的链接在网上漫游,记录 URL、文件的简明摘要、关键字或索引,形成一个很大的数据库。这种数据库包括标题、摘要、关键字和 URL、文件的大小、语种以及词出现的频率。引擎系统虽然能在 WWW 信息资源范围内自动发现新的信息,对其所覆盖的资料进行自动更新,并根据检索规则和从其他服务器上得到数据类型对其进行加工处理,自动建立索引,并通过检索接口为用户提供信息查询服务。但是由于系统需将 HTML 文件传送至本地然后分析,大量占用昂贵的网络带宽和 CPU 资源, 资源消耗过大, 增加了被搜索结点的负担;又由于链路效率太低,对一些连接代价很大的获得索引,难免有不能及时加入的新 WWW 地址。此外,由于各搜索引擎标引方式没有统一的规范,有的对网页全文进行索引,有的仅标引网页的标题、RUL、关键段落的前几个单词或文本的前 100 个词,而且生成关键词的技术也不一样,有的支持 MetaTags,接受网页制作者自定义关键词和摘要,有的则不支持 MetaTags,仅仅利用网页的前几行字作为摘要。此外,搜索引擎大多采用自然语言标引和检索,没有受控词表,同义词和近义词得不到控制,词间的关系得不到揭示。因此,搜索引擎的信息组织与标引缺乏控制,信息查询的命中率、准确率、查全率差强人意,往往是输入一个检索式,得到一大堆网页地址,但其中大部分是冗余信息。

2.2 网络爬虫

网络爬虫是一个自动提取网页的程序, 它为搜索引擎从万维网上下载网页,是搜索引擎的重要组成。

传统爬虫从一个或若干初始网页的 URL 开始,获得初始网页上的 URL,在抓取网页的过程中,不断从当前页面上抽取新的 URL 放入队列,直到满足系统的一定停止条件。聚焦爬虫的工作流程较为复杂, 需要根据一定的网页分析算法过滤与主题无关的链接, 保留有用的链接并将其放入等待抓取的 URL 队列。然后, 它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL,并重复上述过程,直到达到系统的某一条件时停止。另外, 所有被爬虫抓取的网页将会被系统存贮, 进行一定的分析、过滤,并建立索引,以便之后的查询和检索。

3 系统需求

3.1 管理需求

因为网络的快速发展和广泛应用, 迅速获得网络资源上信息的要求也与日俱增,网站信息的实时更新速度也随之提高。然而,网络上的信息大多都是无组织的,并且由于网络的分布式特性,很难对它进行信息和知识管理。我们普通的手工临时管理并

王江红: 副教授

不能将这些庞大与繁杂的数据进行有效地截取和保存，也不具备整体跟踪能力,这势必要求开发一个全新系统来调节与管理,网络爬虫技术也由此诞生。

搜索引擎以一定的策略在互联网中搜集、发现信息,对信息进行理解、提取、组织和处理,并为用户提供检索服务,从而起到信息导航的目的。搜索引擎提供的导航服务已经成为互联网上非常重要的网络服务,搜索引擎站点也被美誉为“网络门户”。尽管基于海量多媒体信息的语音、图形、视频搜索引擎技术已成为搜索引擎领域的研究热点,但是基于 Web 的全文本搜索引擎仍然是使用最为广泛的，如信息量较大的专业门户网站的站内信息检索、基于互联网的特定信息搜集等等。

3.2 目标需求

本系统的开发目的就是通过网络爬虫技术，实现搜索引擎从自己想要访问的网上下载网页，再根据已下载的网页继续访问其它的网页,并将其下载,直到满足用户的需求。

根据现实中不同用户的各种实际需求，本系统实现网络中的搜索引擎,这需要达到如下几个目标:

1. 通过 socket 套接字实现客户端对服务器的访问，客户端向服务器发送自己设定好的请求。

2. 通过 http 将 Web 服务器上协议站点的网页代码提取出来。

3. 根据一定的正则表达式提取出客户端所需要的信息。

4. 采用深度优先搜索从网页中某个链接出发,访问该链接的网页,并通过递归算法实现依次向下访问。

5. 采用广度优先搜索从网页中某个链接出发,访问该链接网页上的所有链接,访问完成后,再通过递归算法实现下一层的访问。

一般情况下，基于 Web 的全文搜索引擎均由页面搜集器、页面索引器、页面检索器等三个主要部分组成,如图 1 所示。

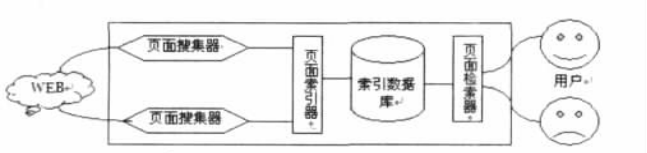

图 1 基于 Web 的全文搜索引擎系统架构

基于对以上例图的分析，其中页面搜集器和页面索引器是搜索引擎最为核心的模块,也是本系统的核心关键,结合本系统的自身的特点,故将本系统设计成一个基于后台运行的系统,它会根据预先设定的网址去查询对应的网页，根据该网页中的链接继续去访问。网络爬虫访问页面的过程就是对网上信息遍历的过程,从而满足客户的需求。

4 系统的设计与实现

4.1 系统的结构框架

在对本系统进行总体架构分析基础上,得到系统的流程图:图 2。

4.2 系统的模块划分

本系统可划分为五个模块:socket 功能模块、http 功能模块、正则表达式功能模块、深度搜索功能模块和广度搜索功能模块。

Socket 功能模块:它是网络爬虫依赖的背景知识,存在于知识管理系统的结构中，客户通过 socket 套接字与服务端建立起连接。

http 功能模块:客户端必须定义一组 URL 来确定要浏览的地址,当客户机与服务器建立连接后,发送一个请求给服务器,如果服务器接到请求后,给予相应的响应信息,这样就可以将网页上代码提取出来。

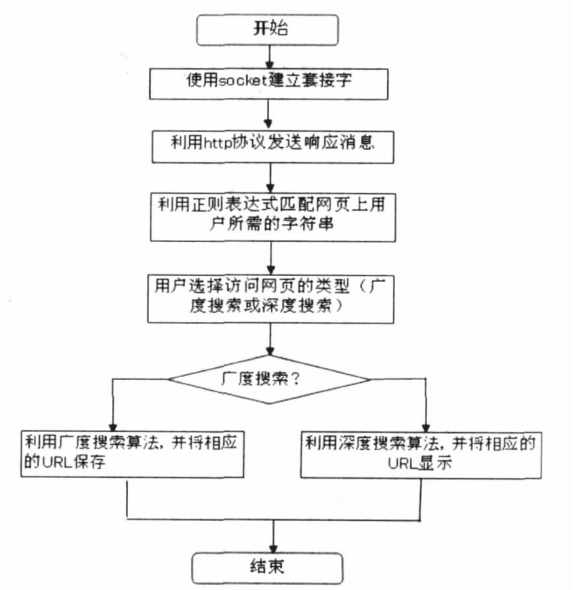

图 2 系统的流程图

正则表达式功能模块：正则表达式描述了一种字符串匹配的模式,可以用来检查一个串是否含有某种子串、将匹配的子串做替换或者从某个串中取出符合某个条件的子串等。客户端接收到服务端网页上代码后，正则表达式作为一个模板，将 URL 字符模式与所搜索的字符串进行匹配,然后将相应的 URL 提取出来。

深度搜索功能模块:本系统开始时对网页上的所有 URL 均未曾访问过。通过正则表达式在网页提取出来的第一个 URL 为初始出发点(源点)V,并将其标记为已访问过;然后依次从 V 出发搜索 V 相连接的网页的链接地址 W。若 W 未曾访问过,则以 W 为新的出发点继续进行深度优先遍历,直至满足用户的需求。若此时仍有未访问的顶点,则另选一个尚未访问的顶点作为新的源点重复上述过程，直至在满足用户需求的前提下将网页中所有顶点均已被访问为止。

广度搜索功能模块:本系统开始时对网页上的所有 URL 均未曾访问过。 爬虫从初始页面 p0 的 URL 开始,通过正则表达式检索页面 p0 并抽取页面中的所有 URL，将它们添加到 URL 队列中。然后,爬虫按某种次序从队列中获得 URL,重复上述过程,直到满足客户端的要求。

5 结束语

未来主题网络爬虫的研究主要是围绕如何提高链接主题预测的准确性,降低计算的时空复杂度,以及增加主题网络爬虫的适应性这几个方面展开。

提高链接价值预测的准确性一直是近年来研究的焦点,将各类评价相结合，尤其是分类器的主题相关度预测和基于在线练习,反馈的主题爬行方法值得进一步研究。将目前信息检索领域中的概念检索理论应用于链接价值的计算，是一个新的尝试方向。网络爬虫的爬行具有重复性,如何将 WEB 的动态规律与先前搜索的统计结果相结合,以提高全价值计算的准确性,是一个值得研究的问题。降低网络爬虫在搜索过程中的计算复杂性,也是有待进一步研究的问题。目前的网络爬虫通常采用固定的

(下转第 142 页)

技术创新

根据上述提出的数据结构模型，对某一应用菜单的数据文件和控制流程文件的字节码存储,用表 1 来描述。

表 1 菜单数据存储表

文件 ID	数据信息	数据值	流程 ID	命令标识	EF; 数据信息	下一流程信息
1000(自定义)	00(偏移量) 0D(长度)	0F0B010…	2001(自定义)	24	(ID)1000 00 12	2002 00 1C
	0D(偏移量) 0B(长度)	0F0B010…		21	(ID)1000 12 1E	2002 1C 46

对于表 1,可拆分为两大部分,左边三列代表一个数据存储文件,最后四列代表一个控制流程文件。对于控制流程文件的第一行字符串对应一级菜单第一选项，相应的主动命令是 SELECT ITEM(24);第二行对应第二选项,相应的主动命令是 DISPLAY TEXT(21)。当用户选择第一项,程序根据控制流程文件中的 EF1 数据信息，在相应 ID 的数据存储文件中获取数据并组成完整的主动命令，最后由下一级流程的操作信息决定菜单的跳转方向。

3.3 菜单生成的仿真测试

形象地说,OTA 技术就好像一个空中的写卡器。当用户要增加、修改、变更和删除增值业务时,无须作任何硬件更换,由 USIM 卡的应用描述层将变更数据封装打包后,USAT 应用开发包向手机发送主动命令 SEND SMS，随后通过空中下载功能,利用从 OTA 平台发来的信息修改 USIM 卡中远程更新文件的数据,就能得到所需的业务。

为了验证在应用接口层中提出的菜单数据存储结构及对 OTA 动态更新菜单过程的描述,我们利用 OTA 菜单模拟器工具对本设计进行求证,如图 4,当用户选择一级菜单中的“天气预报”时,手机向 OTA 服务器发送下载请求,通过 SMS 传输方式下载应用数据后,手机显示应用相关内容。

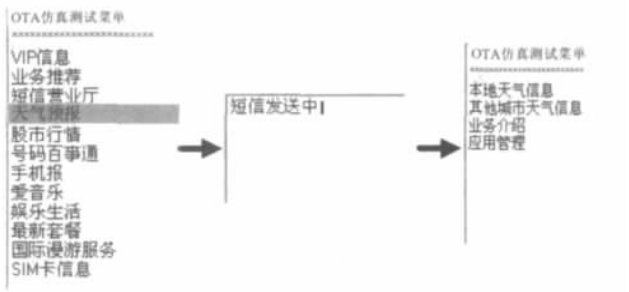

图 4 OTA 仿真菜单测试

通过各大公司开发的菜单模拟软件，对本论文中提出的 USIM 卡三层结构模型与数据存储结构进行测试,结果证明能够正确显示菜单并执行操作,并且利用 OTA 技术也能实现业务菜单的动态更新。

4 结语

3G 网络和 OTA 动态下载技术对比国外，中国仍处于起步阶段,尤其在 3G 网络下使用 OTA 技术实现手机视频业务、网上证券业务及应用下载业务等仍存在巨大的发展空间。因此,本研究具有深远的意义。对于 USIM 卡中的 COS(操作系统),利用本文提出的三层应用模型，能更好说明卡与网络是如何交互应用数据的。更重要的是,在此基础上按照自定义的字节编码规则开发出一套能有效存储和组织主动命令信息的软件。另外本文的研究已成功应用到华大电子股份有限公司和清华同方公司的芯片上,并且得到相关部门的鉴定可用于生产。

本文作者创新点:对 USIM 卡中的 COS(操作系统)提出一个三层系统结构模型,描述了 USIM 卡如何与服务终端交换数据。并提出菜单数据在卡中如何高效存储与组织的解决方法。

参考文献

[1]3GPP TS 31.111 Smart cards; Application Toolkit [s]; Physical and logical Characteristics 2007.03

[2]GSM 11.14-2000.Specification of the SIM Application Toolkit for (SIM-ME) interface[s]

[3]张岩,高立新.OTA 技术在手机卡菜单更新中的应用研究[J].邮电设计技术,2004,35(1):39-43.

[4]张鲁国,马自堂.智能卡操作系统中存储管理设计[J].微计算机信息,2005,8-3:18-19

[5]薛杉.3G 环境 USIM 卡的远程管理[D].北京:北京邮电大学,2007

[6]中国移动集团.STK 卡梦网短信业务菜单 OTA 下载实现方案(二阶段)[P].中国移动通信企业标准:QB-QT-003-2003

作者简介:周允强(1983-),男(汉族),广东省中山人,广东工业大学研究生,硕士,主要从事智能卡开发研究。

Biography:ZHOU Yun-qiang (1983-),Male (Han Race),Guangdong Province, Graduate, Master, Research area: Smart Card Research.

(510006 广东广州 广东工业大学计算机学院) 周允强 李代平 刘志武 周 林 郑汉的

(School of Computer Science, Guang dong University Of Technology,Guangzhou 510090, China) ZHOU Yun-qiang LI Dai-ping LIU Zhi-wu ZHOU Lin ZHENG Han-de

通讯地址:(510006 广州 广东工业大学计算机学院) 周允强

(收稿日期:2009.01.14)(修稿日期:2009.04.14)

(上接第 137 页)

搜索策略,缺乏适应性,如何提高网络爬虫的自适应性也有待进一步的研究。本文作者创新点是:运用正则表达式提取客户端所需要的信息,采用深度和广度优先搜索从网页中某个链接出发,访问该链接网页上的所有链接，再通过递归算法实现下一层的访问,从而实现实时更新数据和搜索引擎技术,在此基础上给出了一种用网络爬虫技术实现局域网内服务器和客户端之间网络通信的解决方案。

参考文献

[1]严蔚敏,吴伟民. 数据结构(C 语言版)[M].清华大学出版社. 2006.

[2]谢希仁. 计算机网络[M]. 大连理工大学出版社. 2006

[3]徐远超等.基于 Web 的网络爬虫的设计与实现[J].微计算机信息,2007,7-3:119-121。

作者简介:王江红(1967-),男(汉族),山东淄博人,山东理工大学计算机学院副教授,学士,主要从事计算机教学与研究。

Biography:WANG Jiang-hong(1967-),Male(the Han nationality), Shandong, Working in College of computer science & technology of Shandong University of Technology, Associate professor, Bachelor, The computer teaching and researching.

(255049 山东淄博 山东理工大学) 王江红 李彩虹

(255090 山东淄博 淄博实验中学) 朱丽君

通讯地址:(255049 山东理工大学计算机学院) 王江红

(收稿日期:2009.02.11)(修稿日期:2009.05.11)