计算机科学 1999Vol. 26№. 1

基于 Intranet 的 Web 信息获取方法和实现

The Method and Implementation of Web Information Search for the Intranet

刘瑞虹　曹东启

（中国科学院软件研究所　北京 100080）

Abstract This paper describes the characteristics of the existing Web information search methods and analyzes the characteristics of the WWW information search by the group users working within an intranet. Based on these, we give out a new Web information searching mechanism for the Intranet. This mechanism can well satisfy the requirements of the WWW information search by the group users working within and Intranet, also can visibly reduce the information transmission amount on Internet. This paper finally gives out how to implement this mechanism by using multi-softbot technology.

Keywords Web, Information search, Softbot, Group users, Group behavior, Intranet

一、引言

自从 1990 年 12 月世界上第一个 Web 软件在 Steven Job 的 NeXT 计算机系统上诞生以来，Web 技术及其应用在世界范围内以惊人的速度迅速扩展，现在其已渗透到了工作生活的各个领域，目前在 Internet 上有几十万个 Web 服务器在世界范围内提供着各种各样的信息服务，根据有关专家预测，到本世纪末，Internet 上的 Web 服务器个数将达到一百万个以上。

面对如此众多的 Web 服务器和其上丰富的 Web 信息资源，如何有效快捷地进行信息查询变得越来越重要，Web 信息的获取也从简单的浏览器漫游查询逐步向定向的信息查询过渡。

目前，已有三种 Web 信息获取方法，且具有各自的特点和适用范围：

直接漫游方式：只有在确切知道要访问的 Web 服务器地址（URL 或 IP 地址）的情况下，才能对此服务器的页面内容进行浏览查询，这种方式在早期 Web 服务器个数较少情况下，非常适用，现在也不失为一种常用的 Web 信息获取方法，但随着 Internet 上 Web 服务器个数的急剧增加，用户已越来越难全部获取这些服务器的地址，即使是获取服务器的地址也很难全部记住这些服务器地址；另一方面，通过直接漫游方式来对某 Web 服务器内容的了解，只能通过一个页面一个页面的浏览显示来实现，没有看到的页面不能判断其是否有所需要的信息内容。

Web 查询树方式：这种方式克服了上述方式存在的问题，即 Web 地址难获取的问题。但由于采用此方式的服务器本身只存放了其它 Web 服务器站点的主题摘要信息，因此用户采用此类方式的工具只能获取所需 Web 站点的地址和主题摘要信息，对于相应 Web 站点的具体页面内容，还要通过其它方式进行进一步的信息查询。

Web 查询索引方式：这种方式完全克服了上述两种方式中 Web 站点地址难自动获取和信息不能定位到具体 Web 页面的问题，但这种方式本身又出现了新的问题，即由于其采用集中式全部词汇的索引方式，造成了此 Web 站点的负载量非常大，如在 Alta Vista 和 Lycos 这些采用 Web 查询索引方式的站点上，已存放了 30 万个 Web 服务器超过 3 千万页面的内容索引（到 1996 年 11 月），这种方式存在的另外一个问题是用户通过使用这类信息获取工具所返回的信息结果随着索引文件内容的增加而不断扩大，这样就容易造成用户不知返回的信息哪些重要，哪些不重要。虽然现在有些工具增加了对返回结果进行优先级排序处理、但其排序的主要依据还是通用意义上的处理结果，如根据页面上查询关键字出现的个数来确定优先级等。

另外现有 Web 信息获取方式所具有的共同特点是假设其用户为单独的个体，这种以个体对象为主服务对象的信息获取方式，具有以下特征：

●获取信息只供一个用户使用，即使是两个在一起的用户，其通过 Internet 对同一远程 Web 服务器进行同样条件的信息查询，Web 服务器要返回两次同样的结果给不同用户。

●不同的用户有不同的主题，几个位置相邻的用户，所关心的信息主题可能大相径庭。

●获取信息的内容范围取决于个体，换句话说，就是获取信息的内容范围和其所在的环境无关，例如一个 ISP 所提供服务环境下的用户，其 Web 信息获取的内容范围完全与相应的用户无关。

二、基于 Intranet 的 Web 信息获取特征分析

随着 Intranet 日益的广泛应用，在 Intranet 上，出现了以本企业工作人员为主体的用户群，此用户群对 Web 信息的获取是以本企业或组织所从事的业务工作为中心，利用 Web 技术所提供的各种软件工具进行数据传输管理、信息获取共享和各工作组之间的协调合作等活动。对于某一特定企业的用户群来说，其在 Intranet 上形成了相应环境下以企业为单位的群体(group)行为操作特征和信息范围，这种群体行为特征的 Web 信息获取具有明显不同于个体行为特征的 Web 信息获取自己所特有的部分，其表现在：

●获取的信息可供企业内用户共享使用：由于企业的 Intranet 上具有功能强大的数据库管理服务器和代理(proxy)服务器，因此有足够的空间来存储供企业内部使用的共享信息，用户只要在防火墙(firewall)内的 Intranet 上就可以获取所需的业务信息。

●存在有企业内部共同关心的信息：一个企业具有特定的业务活动，因此也就具有此企业内部员工所共同关心的主题。例如一个计算机公司企业，其所共同关心的主题绝大部分是和计算机有关的信息。

●获取信息的范围取决于企业环境：企业为满足自身业务活动所需要获取的信息范围远小于 Internet 环境所提供的总信息量。这包含两方面的含义，一方面企业正常运行所需要获取的信息范围是受自身业务活动限制的，另一方面企业也不希望自己的员工利用上班时间到 Internet 上随意漫游进行与企业业务无关的 Web 信息获取。

为了形式化描述这些特征，首先引进“包”的概念。

包(bag)是集合的扩展。和集合相似，包也是在某个域中的元素的集合，但是在包中的元素可以重复出现多次。下面的 B_1～B_3 都是包的例子。

$B_1=\{a,a,b\}$，$B_2=\{a,b,c\}$，$B_3=\{b,b,b\}$

B_2 是集合，B_1，B_3 则不是集合。

将元素 x 在包 B 中出现的次数记为 $\#(x,B)$，在上述例子中，$\#(a,B_1)=2$，$\#(b,B_3)=3$。

包 B 的基数的定义为 $|B|=\sum_x \#(x,B)$。$|B_3|=3$。

包 B 对应的集合记为 B'，即对包中的重复元素进行合并。如：$B_1'=\{a,b\}$，$B_2'=\{a,b,c\}$。

显然有 $|B|\geqslant B$。

域 D 是一个包中各元素的集合。D^n 则是包的空间，它是元素在 D 中且每个元素出现不超过 n 次的所有包的集合。

例如：当 $D=\{a,b,c\}$，$D^2=\{\{a,b,c\},\{a,a,b\},\{a,a,c\},\cdots\cdots\}$。

如果用 C_i 表示某企业中用户 i 对 Web 信息获取的基本命令包(bag)，用 R_i 表示 C_i 命令包所对应获取的 Web 信息包(bag)，那么基于企业 Intranet 的 Web 信息获取特征可用以下公式表示：

$$\sum_{i=1}^{n}|C_i| > |\bigcup_{i=1}^{n}C_i| \tag{1}$$

$$\sum_{i=1}^{n}|R_i| > |> |\bigcup_{i=1}^{n}R_i| \tag{2}$$

$|\bigcup_{i=1}^{n}R_i| \propto$ Internet 上 Web 的信息量(其中 n 表示用户总个数) (3)

从公式(1)中可以导出 $|\bigcup_{i=1}^{n}C_i|$ 与 $\sum_{i=1}^{n}|C_i|$ 的比率小于 1，此比率越小说明企业 Intranet 上有关 Web 信息获取操作的群体特征越强，反之则越弱；从公式(2)中可以导出 $|\bigcup_{i=1}^{n}R_i|$ 与 $\sum_{i=1}^{n}|R_i|$ 的比率小于 1，此比率越小说明企业 Intranet 上有关 Web 信息获取结果的群体特征越强，反之则越弱；公式(3)表明，为满足企业 Intranet 上有关 Web 信息获取所需要的信息量 $|\bigcup_{i=1}^{n}R_i|$ 远小于 Internet 上的总 Web 信息量。

如何充分认识 Intranet 环境下 Web 信息获取特征，并利用这些特征来构造基于 Intranet 的 Web 信息获取机制，随着 Intranet 应用的日益广泛而变得越来越重要。

从前面对现有 Web 信息获取方式的分析可以看出，已有的 Web 信息获取工具不能充分满足基于企业 Intranet 的信息获取要求，为了解决此问题，我们在有关课题中提出了采用分布式软件机器人体系结构的方法来构造企业 Intranet 环境下的 Web 信息获取系统，使其一方面充分利用已有的 Web 信息获取技术和软件工具，另一方面把 Intranet 环境下的 Web 信息获取操作作为一个群体行为来考虑，以使充分满足企业 Intranet 的 Web 信息获取特征的要求，同时实现 Web 信息获取的优化处理，我们采用的主要方法包括：

● 在 Intranet 环境下设立统一的 Web 信息的 Cache 机制，使企业内的用户尽可能少地访问外部 Web 站点。

● 设立用户代理机制，一方面可大大减少用户需记忆的 Web 站点信息，另一方面可对用户的 Web 信息获取请求进行集约化处理。

● 设立 Web 信息获取过程的优化机制，在保证获取 Web 信息内容正确的前提下，减少 Internet 上的信息传输量。

● 设立 Web 信息推(push)和频道(channel)定制机制，使用户尽可能快地获取所定制的 Web 信息内容。

为了实现这些方法，我们采用多软件机器人体系结构来实现构造基于 Intranet 的 Web 信息获取系统。

三、多软件机器人体系结构

软件机器人作为人工智能(AI)研究的一个十分活跃的领域，是分布式人工智能(DAI)的一个基本术语，同时也是人工智能的一个原语。综观国内外研究人员在这方面开展的研究工作，可以发现不同的研究人员都赋予软件机器人以不同的结构、内容和能力。我们认为它是一种抽象实体，能作用于自身和环境，并对环境做出反应。每个软件机器人是一个自包含的软件对象程序，用于完成相应系统内一些特定任务，多个软件机器人一起构成一个完整的应用系统。软件机器人具有以下能力：

(1)自主能力：可以在没有人或其它软件机器人的直接干预的情况下运行，并且对自己的行为和内部状态有某种控制能力；

(2)社交能力：可以和其它软件机器人(或人)通过某种语言进行交互；

(3)反应能力：能观察其环境，并在一定时间内作出反应，以改变该环境；

(4)预动能力：不仅简单地对其环境做出反应，也能够通过接受某些启动信息，体现出目标定向的行为来。

每个软件机器人的基本构成包括：常识、专业知识(角色专业知识)、方法、推理机制、属性、语言(通讯协议)。在软件机器人之间的通讯方面，根据言语行为(speech act)理论，把软件机器人之间的通讯内容分为通知(inform)、请求(request)、提供(offer)、接受(accept)、拒绝(reject)、竞争(compete)和帮助(assist)等行为原语，来完成软件机器人相互之间的信息交流。两个软件机器人之间的相互作用可用下图表示：

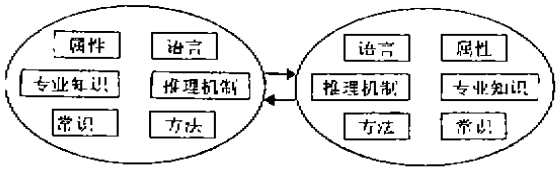

图 1　软件机器人相互作用

多软件机器人体系结构是由多个软件机器人组成的集合，形成一个计算机网络环境下完成有关特定任务的实体群。其中每个软件机器人通过统一的接口与其它软件机器人进行交互访问，一个系统内的多个软件机器人不一定都运行在同一种计算机上，大部分情况下是异构计算机环境。

四、基于 Intranet 的 Web 信息获取系统

针对 Intranet 环境下 WWW 信息获取的具体要求，我们采用多软件机器人体系结构进行了基于 Intranet 的 Web 信息获取软件系统的设计。由于 WWW 信息环境和人类社会环境有许多相似之处，如果能把人类社会活动中的角色技术应用到 Web 信息的获取上，可带来许多优势。在系统中，每个软件机器人作为一个相对独立的个体承担着不同的角色，并具有自己相应的任务，为了克服上述已有 Web 信息获取方式在 Intranet 环境下的不足和实现上述有关机制，我们在系统中引入了以下软件机器人角色：

用户代理软件机器人：驻留在本地服务器上，负责接受用户的信息获取请求命令并对请求进行分析处理，采用算法把多个用户的请求进行合并。

调度员软件机器人：驻留在本地服务器上，负责

接受用户代理机器人发出的信息获取请求，并根据具体情况调度有关软件机器人来完成 Web 信息的获取操作。

资料管理员软件机器人：驻留在本地服务器上，负责管理维护相应 Intranet 上专用的 Web 信息 Cache，接受并执行调度员机器人发出的 Web 信息查询请求和收发员软件机器人发出的对获取 Web 信息的存储请求，并能根据获取的用户信息和领域知识，对 Cache 中的 Web 信息进行优先级排队处理。

常驻记者软件机器人：长期驻留在 Internet 的某 Web 服务器上，受本地服务器上调度员软件机器人的控制，当其接到调度员软件机器人发来的信息获取命令时，在信息源所在地进行信息搜索处理，然后把获取 Web 信息结果通过网络送给收发员软件机器人。其作用在于减少网络上原始信息的传输量。

派驻记者软件机器人：当需要对某远程 Web 服务器的 Web 信息内容进行非经常性搜索时，调度员软件机器人可送一个派驻记者软件机器人到相应的 Web 服务器上进行实时采访，把这些 Web 服务器上内容发生的变化或出现的新信息及时进行整理并返回给收发员软件机器人。

收发员软件机器人：驻留在本地服务器上，负责接收常驻记者软件机器人和派驻记者软件机器人发回的信息，通知资料管理员软件机器人对获取的 Web 信息进行存储，并对获取信息进行分发预处理。

专题播送软件机器人：运行在本地服务器上，负责定时或定条件把获取信息采用推(push)技术送给有关用户。

信息分发软件机器人：运行在本地服务器上，负责根据用户的捕获请求情况把相应的获取信息传送给用户。

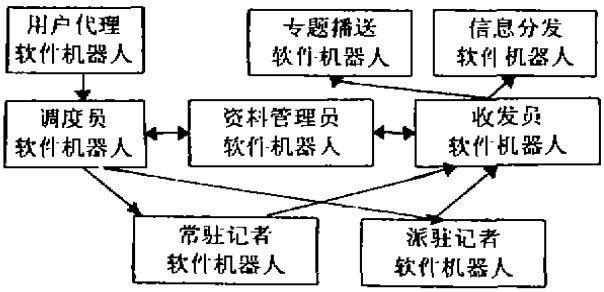

图 2　系统结构示意图

这种多软件机器人结构示意图如图 2 所示。在这种结构中，对于远程 Web 服务器，可预安装一个常驻记者软件机器人，或通过 Internet 实时送一个派驻记者软件机器人运行在相应的 Web 服务器上，因此调度员软件机器人和记者软件机器人是一对多的关系，其它的软件机器人运行在 Intranet 范围内的本地 Web 服务器上，通过网络信息交换构成一个智能化有机整体。

五、实现方法

为了实现上述基于 Intranet 的多软件机器人体系结构的 Web 信息获取系统，其重要前提是能够获得 Internet 环境下的资源共享。这里所说的资源主要包括：信息、网络带宽和计算能力。对于 Internet 环境下的信息和网络带宽资源的可共享性是显而易见的，但对于计算资源的共享是否能实现？

计算机出现的最初目的是进行科学计算，而目前世界上与 Internet 网相联的绝大多数计算机大多数都没有达到满负荷运行(处于空闲状态)，如果能够把这些计算能力充分利用起来，将是一件很有意义的工作。

那么如何实现 Internet 环境下的计算资源共享机制，一种方法是在具有计算资源可共享且愿意共享的计算机上设立一个标志信息接口，其它计算机通过访问此标志信息接口，来获取此计算机是否提供计算资源共享以及可提供共享计算资源的性能指标；另一种方法是建立相关计算机之间相互提供计算资源。

基于上述二种方法，即可实现 Internet 环境下的计算资源共享。文献[6]描述了 Internet 环境下进行计算资源共享的实例，其主要方法是基于 WWW 环境下的超文本传输协议(HTTP)和 WWW 各节点上驻留的守护进程(HTTPD)，当一台机器想在另一台机器上运行本地程序时，一方可调用另一方的有关 Web 页面，此页面可通过对话框形式提交有关参数并放送给提供共享计算资源的另一方服务器，此服务器根据用户提交的参数，采用 CGI(Common Gateway Interface，公共网关接口)到被提供的机器上取回参数中描述的源程序并存放起来，然后对源程序进行校验、编辑及运行，最后把运行结果存放在一个文件中供用户查询或通过 E-mail 方式传送给用户。这属于上述第一种方法所进行的计算资源共享。

我们在有关课题开发过程中，同时采用了上述

两种方法来实现 Internet 环境下的计算资源共享。在系统具体实现上，采用 Java 作为编程语言，通过构造软件机器人类(class)的方法来实现系统的功能。软件机器人分类及其继承关系如下：

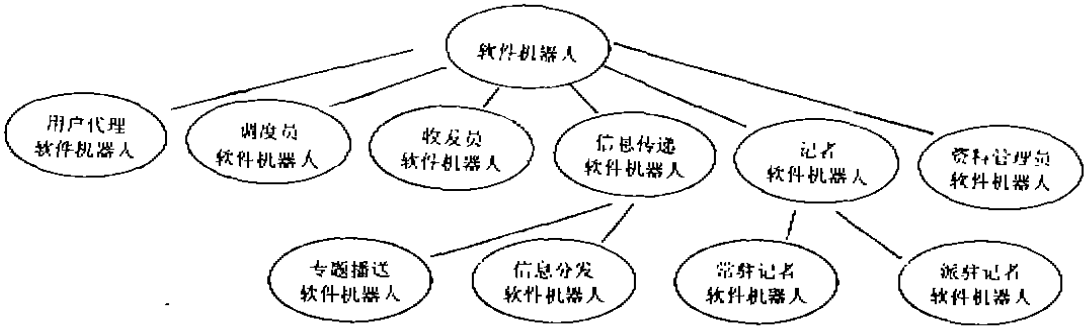

图 3　基于 Intranet 的多软件机器人分类及其继承关系

每种软件机器人在系统运行过程中可存在一个或几个实例，根据系统的运行状态情况而定。不同软件机器人根据自身角色的不同而具有自己独有的专业知识、方法和推理机制。下面着重描述调度员软件机器人的结构和运行方式(图 4)。

调度员软件机器人具有一个资源目录库，它存放着两类信息：①站点信息：站点描述、站点地址 URL、相应站点提供的搜索方法和具体形式；②站点分类信息：分类描述、对应的站点。调度员软件机器人利用资源目录库来确定搜索查询的信息源地址，具体方法为：首先对用户代理软件机器人提交的信息获取命令进行规格化处理，将其转换成规格化命令，然后执行规格化的命令，对资源目录库进行自动查找匹配，以确定信息源地址，最后根据模型算法，计算确定信息捕获搜索路径，并自动优选最佳搜索路径。

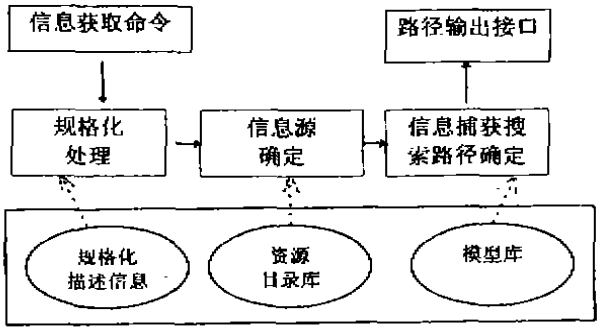

图 4　调度员软件机器人逻辑结构图

模型算法的主要判断条件是：资料管理员软件机器人是否拥有需获取的信息；被搜索的信息源是否有搜索引擎；被搜索的信息源是否允许派驻记者软件机器人。

另外在处理获取信息结果的收发员软件机器人中，我们采用了智能概念抽取等技术来对获取的信息结果进行结构化处理；在负责存储管理获取信息结果的资料管理员软件机器人中，采用主题词典技术来实现关键词的同义关系、上下位关系、相关关系查询。

参 考 文 献

1　Perrochon L. A Quick Tutorial On Searching and Evaluating Internet. IEEE Communications Magazine, June 1997

2　Kosmynin A. From Bookmark Managers to Distributed Indexing: An Evolutionary Way to the Next Generation of Search Engines. IEEE Communications Magazine, June 1997

3　Musella D, et al. Step by Step Toward the Global Internet Library. IEEE Communications Magazine, May 1997

4　Kevin C A, Ammar M H. Multicast Group Behavior in the Internet's Multicast Backbone (Mbone), Georgia Institute of Technology

5　Hills M. Intranet as Group Ware, John Wiley & Sons, Inc.

6　刘欣然，等．基于 WWW 的计算资源发布．小型微型计算机系统，1997，18(5)

7　王怀民，陈火旺，高洪奎．面向智能主题的程序设计．计算机学报，1994，17(5)

8　沈达阳，林作铨．Internet 上的软件 Agent．计算机科学，1997，24(4)

9　姚郑，高文．面向 Agent 的程序设计风范．计算机科学，1995，22(6)

10　孙淑玲　环球网 WWW 及其应用．中国科技大学出版社

11　冯玉琳，黄涛，倪彬．对象技术导论．科学出版社