Intranet 搜索索引自动生成工具 WebIndex

沈达阳 于斌 林作铨 (汕头大学计算机科学研究所 515063)

摘要:本文在分析几种典型 Internet 搜索引擎的基础上,设计实现了一种面向 Intranet 的信息收集 agent,着重阐述了 WebIndex,一个基于 IICA 体系的 Intranet 搜索索引自动生成工具。该系统克服了现有产品的一些缺点,在系统的可伸缩性、可移植性和中文处理能力方面有明显的改进。

关键词:Intranet agent 搜索引擎 索引生成器

针对 Intranet 用户,我们设计实现了一种 Intranet 上的信息收集 agent。与众多同类 agent 不同的是,我们利用自己的搜索索引自动生成器 - WebIndex,通过对有限 Intranet 信息源的访问来建立索引。它既能提供和 Yahoo 类似的高度相关性的信息,使用户能够得到局限于某一网络范围的信息,避免通用搜索引擎所反馈的,十分庞杂的信息干扰,又能自动更新和扩展,可以对指定的 Intranet 范围内进行自动搜索,这些对 Intranet 系统的信息处理和检索具有重要的意义。

一、现有搜索索引的建立方法和存在问题

目前,通用 Internet 搜索引擎所用搜索索引的建立方法,主要有以下两种:

1. 人工维护的搜索索引:如 Yahoo! 和一些虚拟图书馆系统,利用大量的人力浏览 Internet 页面,对其进行分类,因此,建立的搜索索引覆盖面较窄,但比较精确,一般只适用于相对小型、静态的信息(如大学的 Web 信息),因为编辑人员不可能跟上庞杂,快速增长的信息(如个人主页或美国经济新闻,某个主题的学术文章等)。

2. Internet 搜索蜘蛛(Internet spider)产生的搜索索引:(如 AltaVista, Lycos, Excite 等),这些自动生成的索引覆盖面很广,但在精确性方面却很差。用户往往被迫人工从庞杂的反馈中,过滤出所需的信息。

还有另外一种搜索引擎 - 元搜索引擎。所谓元搜索引擎,实际上是一种本身不具备搜索索引,而依靠其他原始引擎的索引或搜索接口,来完成其搜索任务的引擎。元搜索引擎的研究重点在于,如何在众多原始引擎提供的信息中,以友好的人机界面,精确地挑选出用户所需的信息。因此,虽然元搜索引擎往往能提供比通用引擎相关性更好,覆盖面更广的搜索结果,但由于它依赖于其他的搜索索引,对于希望为自己的 Intranet 系统建立搜索索引的用户,或者对于希望搜索局限于某个 Intranet 范围的信息的用户,它都是无能为力的[3]。

同时,有一些 Web 服务器,也提供一些简单的 Intranet 搜索产品,如 Microsoft IIS4.0 中的 SiteSearch,它可以提供 Intranet 页面信息的简单搜索。但是 SiteSearch 只能搜索它所在服务器的页面,所以它不能适应包含多台 Web 服务器的 Intranet 体系。

二、信息收集 agent 的整体设计

1. WebInde 所索引的 Intranet 信息结构

WebInde 的应用目标,是通过一些搜索蜘蛛为一定 Internet 范围内的信息建立搜索索引。对搜索蜘蛛而言,Intranet 就是作为它活动范围的一些 Web 服务器,这些服务器的 URL 满足一定的条件。以汕头大学的校园网为例,有下列 Web 服务器:

服务器名称	URL
汕头大学网络中心服务器 1	Http://www.stu.edu.cn/
汕头大学网络中心服务器 2	http://xjshi.stu.edu.cn/
汕头大学计算机科学研究所服务器 1	http://www.ics.stu.edu.cn/
汕头大学计算机科学研究所服务器 2	http://stuics2.tu.edu.cn/
汕头大学智能和模式识别实验室服务器 1	http://stuaipr2.stu.edu.cn/
汕头大学智能和模式识别实验室服务器 2	http://www.aipr.stu.edu.cn/
汕头大学数据研究所服务器 1	http:www.maths.stu.edu.cn/
汕头大学医学院服务器 1	http://www.med.stu.edu.cn/

从上表可以发现,所有服务器的 URL(Uniform Resource Locator,是一个用于标识任何 Internet 信息的字符串)都是以"stu.edu.cn"结尾的,说明它们都是属于汕头大学校园网的服务器。搜索蜘蛛就利用这样的特征,来

限定它漫游的范围。因此，WebIndex 的蜘蛛并不需要提供 Intranet 中所有服务器的 URL，只要提供一些 URL 的识别特征，它就能够实现在指定范围内漫游的任务。如果没有提供任何特征，它就根据常识，从给定的漫游起始 URL 中抽取。例如：

·给定“http://www.stu.edu.cn/”为起始 URL，蜘蛛就默认识别特征为“stu.edu.cn”，搜索范围就限定在汕头大学校园网。

·给定“http://ww.stu.edu.cn/chinese/xueyuan/”为起始 URL，蜘蛛就默认识别特征为“stu.edu.cn/xueyuan/”，搜索范围就限定在汕头大学学院主页。

·给定“http://www.edu.cn/”为起始 URL，蜘蛛就默认识别特征为“edu.cn/”，搜索范围就限定在中国科研和教育网(CERNET)。

可见，对 WebIndex 的蜘蛛而言，Intranet 的概念十分灵活，这使 WebIndex 生成的索引系统具有较好的可伸缩性。

2. WebIndex 的搜索蜘蛛

搜索蜘蛛(Spider，也称为 Web Wanderers，Web Crawlers，Web Robots 等)实际上是一个程序，它从一批给定的 URL 开始，根据一定的算法在 Interent 上漫游，对它从服务器上取回信息(一般是 HTML 文本)进行分析，抽取出新的 URL，作为以后漫游的目标。

漫游的基本算法有两种：宽度优先算法和深度优先算法[2]。虽然 IICA 实现了这两种算法及其变种，但由于 WebInde 索引的目标是 Intranet，对蜘蛛来说，也就是某个 WebIndex 采用宽度优先的算法，同时采用许多启发式知识来限制搜索的内容和深度，如协议，服务器名称，目录名称和目录深度等。

在 IICA 体系中，搜索蜘蛛是一个以 Java 线程形式运行的 Agent，它包含 3 记忆状态：ToDo(将要处理的工作)，Done(已经完成的工作)和 Doing(它正在做什么工作)：

·ToDo 队列：蜘蛛以宽度优先的方式搜索 Web 服务器，由于并发线程数量有限，它往往不能同时完成几个页面所有 Internet 连接的分析工作，ToDo 是一个先进先出队列，其目的是，保存所有没有被完成的连接的 URL，以便蜘蛛下一步工作的需要。

·Done 哈希表：Internet 的 Web 页面往往是互相连接，因此蜘蛛在漫游过程中会碰到许多已经经历过的 URL，Done 哈希表的目的是记录蜘蛛经历过的所有 URL，从而避免 URL 的重游。由于经历过的 URL 很多，需要利用哈希表在漫游的过程中，提供快速的检索。

·Doing URL：记录蜘蛛正在处理的 URL，在 IICA 体系中，蜘蛛有一个函数指针，指向默认的处理函数，其他驱动蜘蛛的 Agent 可以把它喜欢的处理方法，以函数的形式传送给蜘蛛，从而改变蜘蛛对该 URL 的处理行为。

同时，它也包含了一些辅助性的 Agent，如：

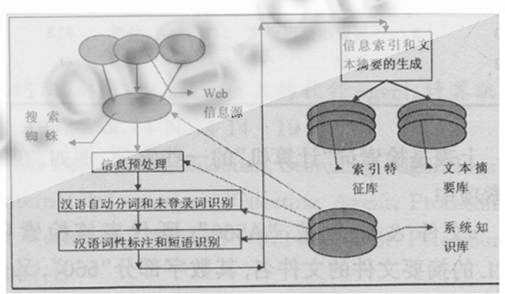
·Robotics 检查器：根据 Robot Exclusion Standard(这是一个蜘蛛搜索的工业规范)，查找 Web 服务器的 Robot Exclusion 限制，以规范蜘蛛的漫游行为。

·HTML 分析器：分析 HTML 文本，抽取出蜘蛛所需的信息。

3. 搜索过程中索引的建立方法

和英文的 Internet 搜索索引相比，中文 Internet 信息检索系统的发展相对比较慢，目前已有的中文搜索引擎大部分还处于“字”的索引阶段。不仅效率较低，而且信息检索的精度很差。究其原因，是中文信息检索有自身的特点：中文词之间没有标记(如英文中的空格)，也没有显式的标记来辨识人名、地名等未登录词(如英文中的大写字母)，另外，中文的词性标注难度也较高(没有英文动词时态变化这样的形态特征)[4]。

因此，如果要对中文词建索引，首先，要对中文文本进行自动分词和未登录词的辨识，其次，对词法和短语的进一步处理也有助于提高索引的有效性。WebIndex 建立索引的过程如下图：



信息预处理包括：

·分析信息源的类型，区分协议类型(目前支持 http 和 ftp 两种)，目录和文件，以及文件的类型(目前支持 HTML 文本，一般 TEXT 文本和 GIF、JPG 两种图形格式)[2]。

·把 HTML 文本分析为标记和文本的向量空间形式[2]。

对 HTML 文本的文本空间和一般 TEXT 文本进行

汉语自动分词。其目的是：

·减少索引所需的空间(一般来说，词索引量比字索引量要少)

·提高索引的效率(词索引的检索速度比字索引要快得多)

·提高检索的精确度(歧义可以引起小于 3% 的分词错误，未登录词有可能引起 7% 的分词错误)[4]

4. 索引特征库的结构

一般全文检索系统典型的倒排检索索引的结构如下

检索词 1	URLi	Wi	URLi+1	…	URLm	Wm
检索词 2	URLj	Wj	URLj+1	…	URLn	Wn
…						
…						
…						

在某一检索词所对应的检索项中，可以查到包含该检索词的所有信息源的 URL 以及相关的权重。由于 WebIndex 不但要求能提供相关检索词信息源的 URL，也要提供该信息源的摘要，而且，由于一个 URL 往往被多个索引项所使用，非常浪费系统资源，所以，我们改进了以上的结构。把倒排索引分为三部分：基本索引、URL 索引、摘要文件。基本索引的结构如下：

检索词：计算机

文件 & 地址	权重	前导词	前导地址	后继词	后继地址
. . .					
A560	3	常识问题	0	研究	223
A560	3	超协调限制逻辑	133	学报	40
A560	3	可废除推理研究	247	科学	818
A570	1	青年	0	学术	65
. . .					

上表是检索词“计算机”的一些索项，例如，在第一个检索项中：

·文件 & 地址域：“A560”，既代表该检索项所在 URL 的摘要文件的文件名，其数字部分“560”，又是该索引项在 URL 索引中的偏移地址，利用该域就可以得到该检索项所对应 URL 的名称和摘要。

·权重域：代表该检索词在该 URL 出现的权重(不一定是简单的出现频度，而是根据 HTML 标记调整)。

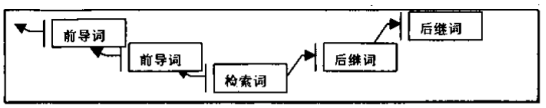
·前导词：该检索词在此 URL 出现时，出现在其前面的词或短语

·前导词地址：此前导词所对应检索项的偏移地址

·后继词：该检索词在此 URL 出现时，出现在其后面的词或短语

·后继词地址：此后继词所对应检索项的偏移地址

根据后四个域，对于某个检索项，可以复原出 URL 所对应信息源的文本的片段，如下所示：



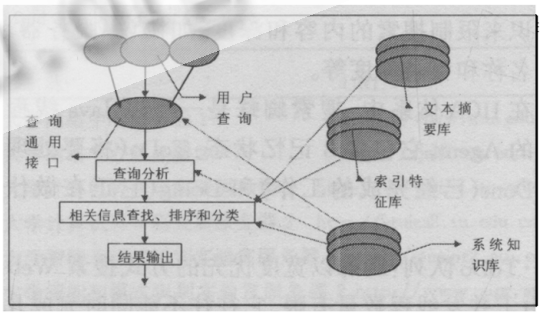
这使 WebIndex 在没有该 URL 全文的情况下，能够对该 URL 所对应的信息源实现全文检索。

5. 信息收集 Agent 的结构

在 Internet 信息检索领域，许多搜索引擎在英文信息检索技术方面发展较为迅速。如 Hotbot，可以利用向量空间表示线索信息内容，并将自然语言处理应用于信息检索，大大提高了信息查询的准确性，Infoseek，Lycos 在检索界面都增加了一些和自然语言有关的搜索功能等。本系统主要针对中文的特点，做了一些改进，其结构如下：

·信息检索模型 WebIndex 搜索引擎采用布尔逻辑和向量空间相结合的信息检索模型，先把和用户查询有关的结果映射到一个向量空间，再根据相关信息项的权重进行排序和分类(分类的工作目前尚未实现)。

·查询分析，包括：中英文查询的区分，码表的转换(对中文简体而言，是 GB 码和 Unicode 的转换 0，对要查询的中文字符串进行分词处理(这部分功能目前尚未实现)，以及查询的扩展处理。



6. 性能的评价和改进

要判断 web 搜索工具的性能是比较困难的，不过，任何对信息抽取技术的评价都是基于对它的精确率和召回率的测量。精确率测量在搜索过程中返回的相关信息的比例，召回率测量的是存在的相关信息被找到的比例。一般来说，增加召回率必须付出降低精确率的代价，反之

亦然。目前，WebIndex 和其他许多基于 Robot 的搜索索引一样，通过对信息源相对检索词的权重排序来提高检索结果的精确度，一般可以保证第一个查询反馈(20 个 URL)的精确率在 90% 以上。召回率的提高要复杂得多，需要进行查询扩展的处理。其中，包括同义词扩展(如查询"计算机"时，也同时查询它的其他同义词，如"电脑"，"微机"等)和语义蕴涵扩展(如查询"动物"时，查询其他属于此类的主题词，如"猫"，"狗"等)。这将在以后的工作中，利用《汉语同义词词林》来实现。

三、信息收集 agent 的实现和实验结果

信息收集 Agent 主要由两部分组成，搜索蜘蛛和引擎界面。

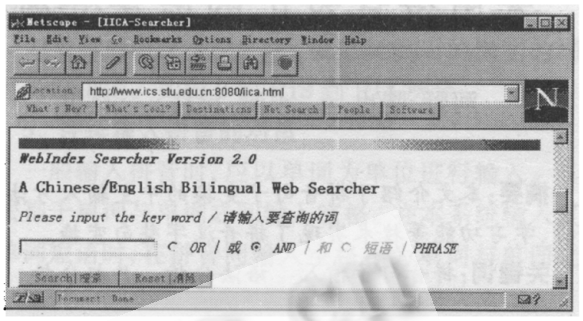
搜索蜘蛛是采用 JDK1.15，在 SUN Solaris2.5 操作系统上实现。目前运行在 SUN Sparc20 工作站上，由于 Java 的平台无关性，很容易移植到别的平台。目前，我们分别以本单位(汕头大学计算机科学研究所)的 Intranet，汕头大学网络中心，汕头大学的校园网，中国科研和教育网(CERNET)为指定的搜索范围，搜索蜘蛛的搜索结果如下：

网络范围	搜索深度	页面个数	索引大小(千字节)	索引/页面比例
计算机科学研究所	3	37	1,029	27.8
汕头大学网络中心	3	177	6,324	34.0
汕头大学	3	1224	3,4404	28.1
CERNET	2	2580	40,748	15.8

从上表可以看到，WebIndex 所建立的索引的效率还是比较高的，当页面的个数达到 Internet 级别时，其索引/页面比例就开始有较大幅度的下降。

搜索引擎是利用 JavaWebServer 的 Servlet 技术实现的。Servlet 是一种服务器端的 Applet(faceless object)，它按照 JavaWebServer 的接口规范，从而可以和 JavaWebServer 融为一体。其运行效率要比传统的 CGI 技术高得多，而其实现的方便程度和可移植性，又比 Microsoft - IIS 或 Netscape Entersprise Server 中的动态链接库或 API 好。目前，JavaWebServer 有 Solaris，Windows NT，Windows95 几个版本，这保证 WebIndex 的搜索引擎可以运行在以上几个流行的 Internet 服务器平台上。

搜索引擎的 Internet 用户界面如下：



四、本系统的特点和以后工作的设想

从上文可见，由于采用了 IICA 提供的灵活的搜索蜘蛛，Java 语言，和相关的 Web 服务器 JavaWebServer，本系统在可伸缩性和可移植性方面具有较大的优势，其次，采用了大量中文自然语言的处理技术，对中文而言，在索引的精确度和效率方面，也有一定的优势。

目前，本系统已经提供的搜索索引生成，维护和查询的基本框架，下一步的工作将从以下两个方面同时进行：

·通过引入中文自然语言处理技术，在文档分类，摘要生成和查询分析几个方面，改进用户的查询效率和反馈信息的精确度和可用性。

·利用 MetaSearch 技术，把分别负责几个信息领域的 WebInde 搜索引擎集成起来，在不降低系统信息的精确度的前提下，探讨进一步提供 WebIndex 通用性的方法。

参考文献

[1] 沈达阳，林作铨，Internet 上的软件 agent，计算机科学，1997 Vol.24 No.4 14 - 19

[2] Shen Dayang, Lin Zuoquan, Searching Algorithms of Internet Information COllecting Agent, Proceedings of Workshop DAIMAS - 97, Russia St. Petersburg, 1997,8

[3] 陈智健，Internet/Intranet 上信息查询的研究与实现，汕头大学硕士论文，1998

[4] Sun Maosong, Shen Daynag, CSeg&Tag1.0:A practical word segmenter and POS tagger, ANLP'97, the Fifth Conference on Applied Natural Language Prcessing, USA Washington D.C., ACL, 1997.3

(来稿时间：1998 年 9 月)