计算机科学2000Vol.27№.7

网上信息搜索技术与搜索引擎*)

Network Information Search Technology and Search Engine

姚国祥 罗伟其 沈镇林

(暨南大学信息网络工程研究中心 广州510632)

Abstract As amount of information grows continuously on the internet, people rely on search technology increasingly to find information they need. This paper discusses work principle of search engine, key technology, such as :model of information search, data organization and store, encoding recognition and conversion of Chinese information process and word partition technology of Chinese Characters. We have designed a search engine system for large-scale Web site based on this technology, and introduced function structure of this system. Finally, we discuss the goal of future evolution of search engine.

Keywords Information search technology, Search engine, Search model, Coded indentify

随着 Internet 在全球范围内的迅速兴起,面对纷繁复杂的 Web 空间,如何在浩瀚如海的信息空间里快速找到并取得所需的信息,便成为人们所关注的主要问题。搜索引擎的出现,极大地方便了 Internet 用户,使快速有效地获取信息成为可能。目前网上搜索引擎各种各样,有 Yahoo!、Excite、AltaVista、Lycos、Infoseek、OpenText、WebCrawler、WWW Worm 等几十种。

1 搜索引擎的工作原理

搜索引擎是如何在庞大的 Web 空间收集信息,它们是如何工作的呢?虽然搜索引擎种类不一,但基本的工作原理都是一样的,可以分成这样三个方面:(1)派出网页搜索工具 Spider(蜘蛛)或 robot(机器人)在 Internet 网上搜索信息,并把它们带回搜索引擎;(2)把信息进行分类索引,建立网页数据库;(3)通过 Web 服务器端软件,为用户提供浏览器界面下的信息查询。

1.1 网页搜索工具(Spider)

Spider 实际上是一些基于 Web 的程序,它通过请求站点上的 HTML 文档访问某一站点,它遍历 Web 空间,不断地从一个站点移动到另一个站点,自动建立索引,并加入到网页数据库中。

正是由于 Spider 而使得搜索引擎不同于目录查询。当 Spider 进入某个超文本的时候,它利用 HTML 语言的标记结构来搜索信息及获取指向其他超文本的 URL 地址,可以完全不依赖用户干预实现网络上的自动"爬行"和检索;而目录查询没有自己的 Spider,它的操作需要人的干预。

Spider 每遇到一个新文档,都要搜索它上面的链接。从理论上讲,如果为 Spider 建立一个适当的初始文档集合,由这个文档集出发,遍历所有链接,Spider 将完成对整个 Web 空间的索引。但实际上,这样的目标是不能达到的。首先因为 Web 空间上许多文档是动态生成的,它们根据用户 Form 表单的输入而动态生成。Spider 即使能够找到这些链接,也无法确定它存放的信息。再者,Spider 本身不能访问采用 Cookie、Javascript 或者 Java 技术制作的网页内容。另外,Web 空间非常庞大,Spider 要将整个 Web 空间的网页全部抓取放进一个(或几个)数据库、建立这样规模的数据库也是不可行的。

所以,在 Web 世界里不存在可满足任何要求的 Spider,任何搜索工具都有它的局限性,搜索这个领域合适的 Spider 在另外一个领域中就不一定适合。现在网上出现了一些专用的搜索引擎,这些专用的搜索引擎的 Spider 只对某个特定的领域进行搜索,针对性强,信息的维护和更新也相对容易。如果用户对某个特定领域感兴趣,不妨选择针对这个领域的专用搜索工具。

1.2 网页数据库

Spider 在对 Web 搜索的过程中,将每次搜索的结果(文档名称、URL、概述、链接等信息)存放在网页数据库中。网页数据库一般采用大型的数据库,如 ORACLE、Sybase、Informix 等。

在进行信息检索时,不同的系统会在搜索结果的数量和质量上产生明显的不同。有的系统是把 Spider 发往各个站点、记录下每一页的所有文本内容。也有的系统则首先分析数据库中的地址,以判别哪些站点最受欢迎(一般都是通过测定以往该站点的链接数量),然后再用软件记录这些站点的信息。记录的信息包括

*)本文获得国家计委高科技重点科研项目(19982444)和广东省自然科学基金(960212)资助。

从HTML标题到整个站点所有文本内容经过算法处理后的摘要。当然，无论如何，为了保证数据库信息与WEB内容的同步，网页数据库需要定时更新，更新频率决定着搜索结果的及时性。数据库的更新是通过派出Spider对WEB空间的重新搜索来实现的。网页数据库的容量和更新频率根据搜索引擎的不同而有很大的不同。

用于从数据库中提取信息的搜索逻辑是这些工具的另一关键组成部分。搜索引擎应该能够找到与搜索要求相对应的站点，并按其相关程度将搜索结果排序。为了在竞争中占有一席之地，各搜索引擎的速度都变得很快。

1.3 搜索引擎的用户界面

搜索引擎的可见部分就是它的用户界面。用户通过键入查询关键字从网页数据库中提取结果。搜索引擎不仅提供了键入一个或者几个关键字的简单查询，大多数还提供了附加的高级查询选项。对于搜索引擎的查询选项还没有建立统一标准，下面给出几种较为常见的查询选项：

* 布尔操作：AND(与)、OR(或)、NOT(非)。通过这些连接词，可增加目标结果的相关性。

* 短语查询：短语中的关键字按照特定的排列顺序，只有符合这些关键字顺序的文档才作为查询结果提交。

* 关键字的邻近度：只有查询关键字之间满足邻近度的文档，才作为结果提交。不同的搜索引擎定义的邻近度不同，如WebCrawler定义的邻近度为2个单词，而Lycos定义的邻近度为25个单词。

* 媒体搜索：搜索包含Java applets、Shockware等对象的网页。

* 专用搜索：搜索在链接、图像名称、文档标题中的关键字或者URL。

* 多种搜索约束条件：限定搜索条件，如文档创建的时间段、文档使用的语言等。

在搜索引擎的用户界面，键入查询条件进行查询时，我们常常看到会返回很多的结果，那么它们的前后顺序是如何排列的呢？这就是根据文档与查询条件的相关程度。不同的搜索引擎有不同的相关程度计算方法，但最基本的两条是：一是根据所查关键字在文档中出现的次数。二是根据关键字在文档中出现的位置离文档起始位置的距离。

提交给用户的查询结果一般包括文档标题、URL和概述等，有时也包括文档建立的时间和文档大小，对于文档概述部分的形成，大多数搜索引擎采用的办法是：将网页作者提供的META描述作为文档的概述，如果网页没有META数据，则采用文档的前100至200个字作为概述，也有的搜索引擎(如EXCITE等)不使用META标签，而采用一种提炼文档语句的算法来形成概述。

2 搜索引擎的关键技术

2.1 信息检索模型

信息检索模型的建立是搜索引擎服务的核心之一，它决定了系统查找信息的方式，根据搜索查找相关信息方式的不同，可将其分为：布尔逻辑模型、模糊逻辑模型、矢量空间模型和概率模型。

布尔逻辑模型是最早应用和较为普遍使用的检索模型，也是其他检索模型的基础。标准布尔逻辑为二元逻辑，即对应于文件特征的二元变量。所检索出的文档与查询相关或无关，只有布尔值为真的文件才作为查询结果提交给用户。布尔逻辑检索模型的优点是结构简单、容易实现，缺点较多，主要是布尔逻辑表达式关系复杂，查询建立困难，容易出错，无文件相关性排序等。

模糊逻辑模型是以逻辑真值为[0,1]的模糊逻辑为基础的，在查询结果处理过程中引入模糊逻辑运算，将所检索的信息和用户的查询要求进行模糊逻辑比较，按相关性的优先次序排序查询结果，因而克服了布尔逻辑查询无序的缺点。

向量空间检索模型是将查询和文件构成一个多维向量空间，其中信息文件和用户查询都利用这个向量空间中的向量来表示。通过欧几里德距离和余弦法作相似性比较，根据向量空间的相似性，排列查询结果。利用向量空间模型检索，系统与用户交互性强，检索结果按相关顺序排序，但该模型在文件量大和信息复杂环境下受到限制。

概率模型基于概率排队理论，即当信息文件按相关概率递减原则排列时可以获得最大的检索性能，有效地解决了在布尔检索模型中的不确定性问题。

以上四种检索模型各有优缺点和适用范围，目前使用较为广泛的是布尔逻辑模型和向量空间模型，我们设计的搜索引擎，正是采用了这两种信息检索模型，并将它们有机地结合起来。

2.2 数据组织与存贮

由搜索程序Spider在Internet各站点上搜索回大量的信息，将这些HTML格式的信息文件取到本地之后，由处理程序进行加工，将其辅助部分去掉，并按一定的策略将其中可用于查询的信息存贮到数据库中，形成本地查询数据库，以后当用户检索时，就不必到远程站点去获取HTML格式文件了。为了获得数据存贮和访问效率的高性能输出，必须针对不同的数据对象特点和查询要求来选择数据库及其存贮结构，由于需索引的信息量非常庞大，还应该对索引文件进行优化，优化技术不仅具有较高的压缩比，而且还支持多种检索算法。在设计数据结构的同时，还要考虑数据组织和数据获取以及数据检索等模块之间的并行操作和通信等问题。

2.3 中文信息处理

中文搜索与英文搜索的不同，主要表现在两大方面：首先是中文信息编码繁杂，其次是汉字切分词困难，这是制约中文信息搜索发展的两大主要障碍。

Internet上的英文编码较为单一，为ASCII码，而中文编码有GB码、BIG5码、HZ码、大字符集

ISO2022-CN、ISO-10646等，对于不同站点的HTML格式文件可能会采用不同的汉字编码标准，在存贮这些信息之前，必须对它们进行预处理，首先是识别其编码标准。编码识别主要方法有顺序排除法，编码边界范围确定法，汉字词表对照法和地区域名法，大多数情况下是将多种方法配合使用。在识别了汉字编码之后，还要将它们转换为统一的内部编码，我们一般将它们转换为GB2312编码，这符合我国的汉字编码标准，在编码转换(GB码则不需转换)之后，还要对句子进行分词，然后再存入数据库。

英文的每个单词之间都有明显的分隔符号(如空格、标点符号等)，而中文则没有，因此对中文句子中词的切分便是中文信息搜索的一个重要问题，要进行分词，首先要建立词典，包括通用词典、专业词典、专用地名词典、人名姓氏词典等，在词典的基础上，分词的方法有按词典正向最大词组匹配、逆向最大词组匹配、最佳匹配法、联想-回溯法、全自动词典切词法等。近年还出现了智能专家系统的分词方法和基于统计和频度分析的分词方法。许多情况下，对于同一个句子，不同的分词方法可能会产生不同的分词结果，这就是分词岐异。因此需要根据上下文知识或词法分析，来解决分词岐异，以便获得有用的词语，建立更为有效的信息索引数据库。

3 一个大型站点的搜索引擎—JNUQ

随着Internet在中国的发展，不少大型的信息站点都相继涌现了。与此同时，在网络上原本比较贫乏的中文资源也日益丰富起来，然而，基于中文的搜索引擎还为数很少。因此，建立一个多用途可调式(Scalable)的中文网络资源搜索系统就显得越来越重要了。JNUQ(暨南大学查询)就是建立这样一个网络资源搜索引擎，并以之建构信息服务站，提供给国内用户一个方便的中文网络资源搜索服务。

3.1 系统的硬件平台

国外的知名搜索引擎都采用计算能力很强、容量很大的计算机作为搜索引擎的硬件平台，但是对于中国来说，拥有计算能力强、容量大的计算机系统的站点只有寥寥几家，因此，我们可以采用大容量、分布式的结构以弥补单台机器的容量和计算能力的问题。

实际上，本系统的实验平台仅是一台配备了PⅡ450处理器的机器，配有128MB的内存和9GB的SICI硬盘，物理结构如图1所示。

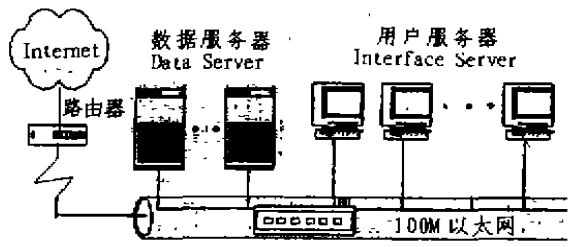

图1 系统的物理结构图

3.2 系统的软件平台

系统测试时使用的操作系统是Linux核心版本2.0.35，使用SUN的C编译器和Perl5.004解释器。当然，也可以运行在配备了Solaris 2.6的Sun Sparc机器上，或者BSD系统上。之所以采用Unix系统的原因是，当系统的资源有限时，Unix系统的性能似乎比相同配置的Windows NT系统要好，表现为可以同时服务更多的用户，占用更少的系统资源和更加稳定。

3.3 搜索引擎系统的功能结构

JNUQ系统采用模块化、分布式的结构，共包含以下的子系统：

(1)资料搜集子系统；(2)资料分析管理子系统；(3)WWW界面软件子系统；(4)索引/查询子系统；(5)虚拟代理服务器子系统。系统主要结构如图2所示。

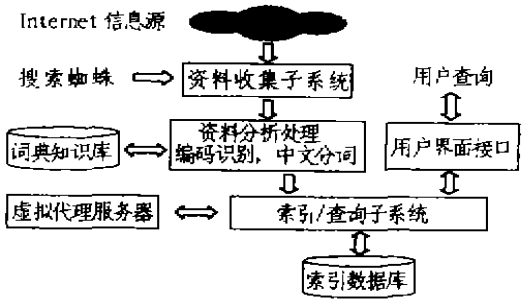

图2 系统功能结构图

全套服务可运行在一台机器或者多台机器上提供服务，各个服务单元可分别独立地、平稳地扩展升级。系统可平行/纵向扩展，各台服务器可分别增加CPU、内存和硬盘进行扩展，或者可通过增加不同类型的服务器或升级某一服务器消除瓶颈，保证系统的服务能力。

3.3.1 *资料搜集子系统*(Data Gatherer Sub System) 用来搜集网络上的信息或内部的信息，是传统搜索引擎的Spider(蜘蛛)或者是Robot(机器人)的同义词。由于传统的搜索引擎采用自动的Spider和Robot带有很大的盲目性，造成了取回的资料的有用率很低，以及对网络的流量是一个很大的负担，因此本系统采用的是半自动的资料收集器，以最大程度地提高取回的资料的有用率，以及尽可能地减少网络负担。

3.3.2 *资料分析管理子系统* 用来过滤分析摘要转换或管理资料，并可去除重复多余的资料，例如：HTML文件中的标记，是应该在索引之前就要去掉的。另外，资料分析管理子系统，还把一些摘要信息和新加入的站点以及多人访问的站点加入到数据库中。

3.3.3 *WWW界面软件子系统* 是一些界面程序，目前是采用CGI接口，向广大查询的用户提供一

个基于 Web 的查询环境。使用 CGI，是考虑到 CGI 的通用性，但是由于 CGI 是多进程的结构，可能会带来性能的降低。考虑到系统的性能并不在这里，因此使用 CGI 仍然是一个不错的选择。

3.3.4 索引/查询子系统 是整个搜索引擎系统的效率关键，只有拥有高效率的索引/查询子系统，才可能有高效率的搜索引擎，它提供对资料的索引和搜索的功能。由于采用了 agrep 较先进的搜索库和压缩技术，因此索引/查询子系统的效率还是比较高的。

3.3.5 虚拟代理服务器(Virtual Proxy)子系统 提供虚拟的高速缓存空间，并可用来架构层次式的信息搜索与信息分布并可嵌入智能型代理(Intelligent Agent)，提供方便的信息过滤与获取的功能。

3.4 系统可提供的搜索服务

本搜索系统具有多功能的特性，提供的信息搜索服务涵盖了 Internet 里最常用的一些网络资源，主要有 WWW 搜索，BBS 搜索和 FTP 搜索。

* WWW 搜索引擎，WWW 主页搜索引擎，计划索引国内中文 WWW 站的主页，提供全文检索的功能，这部分资料大约每隔一个月就应该更新一次。在下一阶段计划扩大索引范围，遍及亚太地区国家，并引入智能型检索的功能。

* BBS 搜索引擎，BBS 文章查询系统，针对国内之 BBS 讨论区，凡是有透过 News 转信的文章均加以索引，以提供检索服务，此部分之资料每日更新索引资料库，以月份区隔，提供最近三个月的文章查询。

* FTP 搜索引擎，其功能和著名的 Anonymous FTP 档案搜索引擎 Archie 类似，能在 FTP 服务器站点按文件名、文件类型、文件大小及生成日期等属性进行搜索。

除上列常用服务外，还有 USENET News 文章查询、科学文献索引查询和针对 WWW 里的多媒体构件搜索，

JNUQ 搜索引擎，可以很容易调整来作不同种类资料的查询应用，在使用时也极富弹性，例如：可弹性地决定对哪些档案及档案的哪一部分做索引，另外，搜索引擎可让用户弹性地订制查询结果输出的方式。

3.5 长远目标以及未来展望

JNUQ 的长远目标是发展一套搜索软件来提供查询服务，所关心的乃是整个网际网络的信息分布、复制、搜索，撷取、过滤、管理和信息服务器的架构与功能。

建构多用途的网络信息搜索引擎，提供各种常用信息的搜索服务，其目标在于追求强大的查询功能，并达到多语言、无国界的检索功能。

发展一套多用途、可调式的信息搜索管理软件，以供一般企业、学校、或事业单位内的计算中心或信息中心，来建立其信息查询管理中心。此查询系统，一方面可以提供外界用户来查询所提供的信息服务，另一方面也可以让内部用户查询内部的资料，或用来搜索取得网络上的信息。

建立一套层次式的信息搜索架构与分散式的索引模式，让以上下同层次的搜索系统互相分工合作，以提供用户透明的层次式的信息搜索。这不仅可以避免很多不必要的网络带宽浪费，而且让用户更快速地找到并取得所需要的信息。

发展一套层次式、虚拟的 WWW 服务系统，通过虚拟代理(Virtual Proxy)与虚拟缓存的骨架和智能型代理(Intelligent Agent)技术来妥善解决信息分布、复制、搜索，提取、过滤等根本问题。

参 考 文 献

1 常桂然，张晓辉. Web 信息检索服务系统与搜索引擎. 计算机科学，1998，25(5)

2 荣传湘，等. 中英文 WWW 搜索引擎中数据获取的设计与实现. 小型微型计算机系统，1999(5)

3 韩耀伟，潘金贵，等. 一个基于 Mobile Agent 的课件搜索系统的框架设计. 小型微型计算机系统，1999(8)

4 蒋盛，马范援. 中英文 WWW 搜索引擎的信息处理. 计算机工程，1999，4

5 史延春. 中文文字 ASCII 码识别与应用系统开发. 计算机工程，1999，10

6 单守斌，等. 中英文信息检索系统的设计与实现. 华南理工大学学报，1998，26(Suppl)

7 金燕，等. WWW 上的全文信息检索技术. 计算机应用研究，1999，1

8 刘培俊. NT4.0下使用 IDC、HTX 开发 WWW 数据库检索程序. 微电脑世界，1997(10)

9 Gudivada V N. Information Retrieval the World Wide Web. IEEE Internet Computing, 1997,(5):58～68

10 Udell J. Search Again, Byte, Jan. 1997

11 Dwight J, et al. Special Edition Using CGI Second Edition

12 Asbury S. CGI Programming Guide, Oct. 1997