

# 高校图书馆搜索引擎设计思考

[作者] 郭海明

[单位] 潍坊学院南校区图书馆

[摘要] 面对当前网络信息环境中缺少针对高校信息用户的网络信息搜索引擎,文章在分析了高校信息用户需求特点的基础上,就如何设计一个方便、高效、适合高校信息用户的搜索引擎作了较为深刻的思考。具体包括高校图书馆搜索引擎的工作流程分析、输入模式设计、整体结构模式设计及相关技术分析等。

[关键词] 高校图书馆,搜索引擎,网络信息

近年来,Internet 已成为信息和计算机领域最为热门的一项技术,Internet 的普及使人们可以突破空间、地域的限制,方便地共享网络信息资源,这给现代图书馆的发展带来了良好的机遇和发展前景。但在实际使用中,WEB 网上的海量信息也给信息用户的信息查询带来了极大的困难。鉴于此,各种搜索引擎应运而生,如 Yahoo、Alta、Vista、搜狐、网易等。这些综合型搜索引擎,面向的是所有的网络信息用户,带有普遍性。它们不能有针对性地准确有效地满足高校图书馆信息用户的需求。作为高等教育的主干网的中国教育科研网,目前也只能满足 5.9% 的用户的信息需求。这显然与教育科研网的地位不相符合,其中一个重要原因就是教育科研网上缺少一个强有力的能搜索整个教育网及 Internet 网上信息资源的搜索引擎。因此建设一个高效、方便的高校图书馆搜索引擎是高校图书馆做好信息服务工作的当务之急。笔者对此思考如下:

## 1 高校图书馆信息用户分析

高校图书馆信息用户是指那些通过高校图书馆接受信息服务的人类个体或群体,其中主要是教师、学生和科研人员。他们获取网络信息资源的途径主要是通过 Internet 和中国教育科研网。具体内容包括各高校的情况介绍、图书馆特色介绍、联机书目数据库(OPAC)、特色数据库和研究生毕业论文数据库等。

据中国互联网络发展状况统计报告(2002.1)统计,网络信息用户文化水平在大专以上学历的占了 59.8%,而这部分用户主要是在学校。他们受过较高层次的教育,同时普遍受过信息素质教育培训,他们上网的主要目的是为获取教学、科研、学习相关的信息。这类信息用户有着区别于其它类信息用户的明显的特点: 信息需求的相对稳定性。主要表现在师生对教学信息需求上,特别是专业核心教学信息资源的需要上。 信息需求的集中性。高校信息用户常常是按照教学计划、教学大纲进行教学实践。各学科各专业的师生往往形成了一个相对集中的信息用户群体,而且往往集中于某学科专业方面的信息。 信息需求的阶段性、节律性。这是由于高校教学实践是分阶段的,而且是有规律、有节奏地变化的。 信息需求的专业性。高校是教学、科研的基地,各项工作专业性极强。因此对信息的需求往往要求的专业层次较深。

因此,高校图书馆网络信息资源搜索引擎在设计时一定要根据自身用户的特点和需求出发,针对各高等学校的教学、专业与科研特色设计,而且它的设计可以有一定的专业性和学科性。

## 2 高校图书馆搜索引擎的工作流程分析

搜索引擎是一组对网络信息资源进行采集、标引,并根据用户检索要求进行查找的软件。高校图书馆搜索引擎和一般搜索引擎一样,其工作流程包括以下三个步骤:

(1) 信息资源采集流程。在图书馆集成网络系统后台,搜索引擎有一个人工或自动的网络信息采集工具。若为自动方式,则是一个被称为“机器人”(ROBOT)的软件根据 WEB 页面的超链接对 Internet 上的网络信息资源进行搜索、采集,并将采集到的资源添加到自己的数据库中。

(2) 信息资源标引流程。当 ROBOT 把信息资源采集到数据库后,搜索引擎会对它们进行一定的人工或自动标引建立索引数据库或将它们放入到已组织好的分类体系目录下以方便将来的检索。自动标引一般都是根据词表进行机器标引。我认为高校图书馆搜索引擎的信息资源标引应采取自动和人工标引相结合的方式,先分类后按主题,这样将有利于用户的使用。

(3) 用户信息检索流程。在图书馆集成网络系统前台,用户将自己的检索要求提交给搜索引擎,搜索引擎根据用户的要求在后台数据库中进行查找、匹配,并将检索到的结果以 WEB 页面的方式返回给用户。这就需要我们设计一个适应高校信息用户的友好的用户界面。

## 3 高校图书馆搜索引擎的输入模式设计

搜索引擎的输入界面及输入方式往往决定了整个搜索引擎系统的索引方式、检索方式甚至体系结构,所以在对搜索引擎的整体结构设计之前,有必要先对搜索引擎的用户界面进行设计。

(1) 输入界面的设计。根据高校图书馆信息用户需求的特点,输入界面设计如下:

附图

说明:在输入想要搜索的关键词后,OPAC 等五个键作为搜索开始键,用户可以根据自己的需要对不同的数据库进行搜索,OPAC 是指各高校的联机书目数据库,特色数据库是指各高校自建的有本校特色的数据库,研究生论文库是各高校的重要资源,本地资源库是为了方便检索而自建的关于各高校及图书馆的介绍,网页信息与一般搜索引擎是一样的。

(2) 输入方式选择。目前搜索引擎的输入方式主要有两大类:关键词输入和分类目录,考虑到此高校图书馆搜索引擎的特殊性,本搜索采用新型的关键词输入方式。

(a) 一般关键词输入方式。这是目前最常用的输入方式,用户只要输入描述自己需求的一个或多个关键词,系统便可据此检索出结果,该方式简单,直观,便于数据库检索,但存在以下缺点:语义空间过大,易发生歧义;某些需求难以表达;难以体现用户个性化特征;无法检索视频和音频多媒体元素。

(b) 关键词重构输入方式。调查表明,用户在使用搜索引擎时,在输入关键词后,如检索不到满意的结果,常在原来的基础上进行修改,然后再次查询,此过程可能重复多次,因此设计了关键词重构输入方式。此方式有明显的优点:克服用户不知如何用关键词语法表达信息需求的问题,帮助用户理清思路;简化用户的输入过程。但也存在一定的缺点:检索范围有限;难以评价系统的检索性能;缺乏个性化的设置。

(c) 关键词调整输入方式。此方式允许用户在输入关键词的同时,可以选择本次检索的个性偏好,然后系统在不改变用户意愿的前提下,根据个性偏好,适当调整或修改关键词,使之更准确地描述用户需求。

## 4 高校图书馆搜索引擎整体结构模式设计

搜索引擎整体结构模式设计如下：

附图

各模块的说明：

(1) 网页信息。网页信息的搜索是基于 ROBOT 的搜索引擎，主要是由 ROBOT、INDEX 和智能查询软件构成。通过蜘蛛或蠕虫程序，从事先制定好的 URL 列表自动访问 WEB 站点，分析提取网页中超文本的 URL，将其加入列表，并据此进一步访问其它站点，INDEX 是一个索引数据库，ROBOT 采集到的网页信息全部在此，智能查询软件提供用户访问的查询界面和服务端的查询程序，当用户查询一个关键词时，搜索引擎将搜索所有的与关键词相符合的网页，按照一定算法生成网页结果返回用户浏览器。如下图：

搜索引擎机器人 (ROBOT) 结构如下：

附图

(2) 本地资源。本地资源是一个各高校图书馆自建的数据库，在 WEB 页上可以直接对其查询。先按预定的用户信息需求收集信息，创建数据库，然后创建一个新的 ASP 文件，加入读取数据库编写相关 HTML 代码，复制 ASP 文件和数据库到 WEB 服务器，然后调试。查寻模式如下：

附图

(3) 数据库。先把 OPAC 及各数据库的位置存入列表，用户提出检索请求后，搜索引擎选择一些数据库进行检索，转化用户的查询条件并发送给这些数据库的 WEB 查询系统，被检索的数据库向搜索引擎返回结果，搜索引擎对这些检索结果进行融合整理后返回给用户。具体模式如下：

附图

## 5 高校图书馆搜索引擎设计的相关技术分析

### 5.1 OPAC 及数据库选择方法

据统计，各高校图书馆大约有 1000 左右的数据库。可以根据用户查询条件的学科领域或地域等，来选择在这些领域内检索效果较好的数据库进行检索。考虑到各学校 OPAC 内的数据大量重复，在默认情况下，只选择几个大的和专业性强的图书馆的 OPAC 作为主要检索入口。如北大、清华、中国地质大学图书馆等。

其它数据库的选择可采用：(1) 定性方法。此方法只用很粗略的信息来代表每个数据库的内容。通常，这种数据库用几个关键字或几个句子，好处在于这些信息相对比较容易获得而且只需要很少的存储空间，但这种简短的描述很难全面地表达数据库的内容，使用这种方法的结果有可能把有用的数据库漏掉。(2) 定量方法。此方法和定性方法的主要区别在于前者使用的衡量数据库有用性的标准更加明确实用。它使用数据库有用性的一种标准是“数据库中对于给定查询潜在有用的文件的数量”。很明显，这个数量清楚地反映了数据库对于给定查询的有用性，另一种有用性的定量衡量标准是一个数据库中 与给定查询最为相似的文件的全局相似度。一方面，该标准表明了能够指望从一个数据库中得到最好的结果是什么；另一方面，对于给定的查询，它可用来最优化地给数据库排序进而从所有数据库检索得到最相似的 M 个文件。(3) 基于学习的方法。此方法基于从以前提交的查询中得到的检索经验来预

测数据库对于新查询的有用程度文献。

## 5.2 查询结果融合与排序技术

从各个数据库返回来的检索结果需要经过融合才能反馈给用户,这个过程应包括以下程序:删除重复,将结果合并到一个完整的列表中,按相关性排序。参照元搜索引擎的方法,MetaCrawler 和 ProFusion 为消除重复设计了比较好的算法,为了更好地进行融合,一个共同的想法就是:搜索引擎亲自对找到的结果进行分析处理,这样做的好处是可以直接对文档的相关性和可用性做出判断,并可以借此提供一些新的功能。

## 5.3 中文切词技术

目前的切词方法很多,归纳起来有两大类:基于词典与规则的方法和基于统计的方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词,第二类基于统计的分词方法则将汉语基于字和词的统计信息,如相邻字间信息、词频及相应的共现信息等应用于分词。下面介绍几种常用的方法:逐词遍历法。逐词遍历法将词典中的所有词按由长到短的顺序在文章中逐字搜索,考虑到文章结束,也就是说,不管文章有多长,词典有多大,都要将词典遍历一遍。正向最大匹配法。其目的是将最长的复合词分离出来,先假定最大复合词长度为 L,进行匹配,如果词典里有这样的字,则匹配成功,否则,去掉最后一个词,继续下去,直到成功为止。基于频度的方法,这种方法不依靠词典,而是将文章中任意两个字同时出现的频率进行统计,次数越高的就可能是一个词,这种方法容易将专用名词提取出来。

## 5.4 网页信息的采集策略

全国普通高校的总数据 2001 年统计是 1234 所,因此,搜索引擎可以尽可能地对中国教育科研网内的所有站点的 WEB 页的内容建立索引。WebCrawler 提出的一个基本策略是采集的文档要来自尽可能多的站点。其遍历算法如下:每当一个新的站点上的文档被发现,该站点加入到一个要进行采集的站点列表中,在继续文档采集之前,要从每个这些新发现的站点返回一篇文档建立索引,所有站点被访问之后,采集过程继续在这些已经发现的站点中进行,直到又发现了新的站点,然后重复上面的过程。

## 5.5 索引技术

就建立索引的内容来说,有的只对标题建立索引。Webcrawler 的方法是采用向量空间模型对内容和标题都进行索引。Google 建立的是全文索引,它对文档中出现的每个词建立索引,全文索引的好处在于简单和便于实现,并且检索时不会错过每个包含检索关键词的文档。一般它的实现方法是:首选使用一个切词工具对文档进行分解,生成一个词序列,切分词工具必须能够处理文档作者可能犯下的各种 HTML 标记错误或语法错误,接着用一个无关词列表对这个词序列过滤,剔除无意义的词,然后用文档向量表示方法计算词的权值,最后建立反向索引。

结束语:目前,尽管出现了上百种搜索引擎,但专业搜索引擎技术还亟待发展,特别是

象面向高校信息用户的搜索引擎还未正式出现。我们应当在发展搜索范围全能的搜索引擎的同时加快设计专用的搜索引擎。这一类型的搜索引擎有着广泛的发展空间,是下一阶段搜索引擎的研究热点。

## 参考文献

- 1 孟卫一,吴宗寰.集成搜索引擎的文本数据库的选择[J].计算机研究与发展,2001(4):15-20
- 2 彭洪汇,林作金.Internet 上的搜索引擎和元搜索引擎[J].计算机科学,2002(9):1-3
- 3 李志蜀,李果.中文搜索引擎的原理剖析及开发实现技术[J].计算机应用研究,2001(11):97-99
- 4 王剑,邵志清.大规模中文搜索引擎的架构和设计技术[J].计算机科学,2002(1):26-29
- 5 陈敏等.一种 WWW 搜索引擎的设计与实现[J].计算机工程与应用,2002,(7):148-150
- 6 夏祖奇等.基于分类目录的元搜索引擎模型的提出与实现.情报学报,2003(1):27-31
- 7 <http://www.searchenginewatch.com>
8. <http://www.archive.org/>
9. <http://www.isc.org/>
10. <http://www.robotstxt.org/>