

【信息管理】

网络信息检索

张秋霞¹, 闫秀萍²

(1. 长安大学 人文社科部, 陕西 西安 710064 2. 长安大学 经济管理学院, 陕西 西安 710064)

摘要: 在介绍 www 信息检索的组成和工作原理基础上, 分析了网络信息搜索引擎的组配检索方法和高级检索技巧, 并对网络信息检索工具存在的问题和解决方法进行了探讨。

关键词: 信息检索; 检索工具; 搜索引擎; 因特网; 组配检索

中图分类号: G250.72 **文献标识码:** A

Internet Information Retrieval

ZHANG Qiu-xia¹, YAN Xiu-ping²

(1. Department of Humanities and Social Science, Chang'an University, Xi'an 710064, China;

2. College of Economics and Management, Chang'an University, Xi'an 710064, China)

Abstract On the basis of introducing the composition and principle of www information retrieval, the tips of Internet information searching is analyzed and the skills of information retrieval are advanced. The problems and the possible solutions of Internet searching engine and other retrieval tools are discussed.

Key words information retrieval; retrieval tools; searching engine; Internet; faceted retrieval

一、引言

信息检索是指从文献集中查找出所需信息的程序和方法。所谓文献集合是指有组织的文献整体。它可以是数据库的全部记录,也可以是某种检索工具,还可以是某个文献收藏单位收藏的全部文献,当然也可以是某个单位通过 Internet 发布的各类信息集合。网络信息检索是指对利用 Internet 信息发布技术,通过 Internet 发布的信息进行的检索,主要利用搜索引擎、网络机器人和门户网站等来完成。

近几年来,随着 Internet 的迅速发展,网上信息以爆炸性的速度不断丰富和扩展,其信息数量之大、类型之多,已经给人们的工作、学习和生活方式带来了巨大影响。为了充分发挥网络信息的重要作用,并能迅速在上百万个网站中快速有效地查找到想要的信息,必须对网络信息的检索方式进行研究,并掌握网络信息检索的基本方法和高级技巧。

二、互联网络信息检索工具的基本原理

随着 www 站点的增长,Internet 上的信息数量和种类越来越多,为了解决信息利用的难题,互联网建立了许多专门的信息检索工具——搜索引擎,使用户可以通过关键词或分类的方法找到所需信息。一般来说,搜索引擎的工作原理如下:

(一)在网上搜寻所有相关信息

网络信息搜索引擎的信息来源主要有两种渠道:一种是由信息发布单位通过搜索引擎的信息登记界面,将自己的信息登记到搜索引擎的数据库中;另外一种搜索引擎利用信息搜索工具——搜索机器人(Robots),自动在网络中进行信息搜索。

Robots(又称为 Spider Worm Crawler 等)是一个能利用 HTTP 协议读取 Web 页面并沿着 HTML 文档中的超链在 www 上进行自动漫游的程序。

Robots 必须做到自动与服务器连接,然后将网页下

载。而漫游功能的实现是有后继和前趋问题的,所以 Robots 的运行是周而复始的,它永不停歇地在网际漫游,在下载网页的同时,为以后要去往何方打好基础。漫游功能通过 HTTP 协议提供的请求指令来实现。另外,Robots 还需具有分析文档的功能,识别文档,对文档进行分类

(二)对搜索到的信息进行加工分类,建立搜索引擎数据库

在搜索到相关网络信息后,必须对网络信息进行加工处理,包括利用自动标引技术对搜集的信息进行标引,按学科分类、图书分类等分类标准进行人工分类标引等,并把相关信息记入数据库,建立索引数据库

(三)通过 Web 页面接受用户的查询请求,并将搜索引擎数据库中查询到的信息返回给用户

查询服务部分是搜索引擎的用户界面,用来接收用户的查询请求,根据用户的查询请求对数据库进行检索,并将检索结果集按相关度反馈给用户。

大多数的搜索引擎在查询服务界面,除了提供关键词检索服务外,还将所搜集到的信息按各自的分类标准进行分类,以提供分类检索方法

三、互联网络信息检索工具的特点

目前各种各样的中西文搜索引擎有很多,比较先进的搜索引擎有 Suhu Yahoo Excite Infoseek Lycos AltaVista 等。每个搜索引擎都有其各自的特点,有的以查询速度快见长,有的以数据库容量大占优,但总而言之,与传统信息检索工具相比较,网络信息检索工具有如下特点:

(一)支持全文检索 (Full Text Search)

全文搜索引擎(如美国的 Altavista^[3] (www.Altavista.com)) 将站点的每一项都抓取过去,其优点是查询全面而充分,用户能够对各网站的每篇文章中的每个词进行搜索。当全文搜索引擎遇到一个网站时,会将该网站上所有的文章(网页)全部获取下来,并收入到引擎的数据库中。只要用户输入查询的“关键字”在引擎库的某篇文章中出现过,则这篇文章就会作为匹配结果返回给用户。从这点上看,全文搜索真正提供了用户对 Internet 上所有信息资源进行检索的手段,给用户以最全面最广泛的搜索结果。全文搜索的缺点是提供的信息虽然多而全,但由于没有分类式搜索引擎那样清晰的层次结构,有时给人一种繁多而杂乱的感觉。这类搜索查询也称为关键词检索

(二)支持目录式分类检索

目录式分类搜索引擎的优点是将信息系统分门归类,当遇到一个网站时,它并不像全文搜索引擎那样,将网站上的所有文章和信息都收录进去,而是首先将该网站划分到某个类下,再记录一些摘要信息 (Abstract),对该网站进行概述性的简要介绍。最具代表性的目录式分类搜索引擎是 Yahoo 网站,另外还有搜狐 (www.sohu.com) 常青藤等。目录式分类搜索引擎提供一份按类别编排的国际互联网网站目录。各类下边排列着属于这一类别的网站名、网址,就像电话号码簿一样。这类搜索引擎往往还伴有网站查询功能,也称之为网站检索,即提供一个文字输入框,用户可以在文字框中输入要查找的字、词或短语,搜索引擎会查找相关的站名、网址和内容。

(三)能够区分搜索结果的相关性 (Pertinency)

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。这里的相关程度是指搜索关键字在文档中出现的频度,当频度越高时,它则认为该文档的相关程度越高。但由于目前的搜索引擎还不具备智能,除非用户知道要查找的文档的标题,否则排列第一的结果未必是“最好”的结果。所以有些文档尽管相关程度高,但并不一定是用户最需要的文档。

(四)检索方法多样 查找手段完备

有些性能完善的搜索引擎不仅能检索 Internet 上的文献,还能查找公司和个人的信息;不仅能检索 Web 页面,还提供对新闻组内文章的查找;不仅能输入单词、词组或句子进行检索,还能指定多个单词之间的逻辑组配及其位置关系;不仅能以词语查询有关主题的页面信息,也能以特定的域名、主机名、URL 等查找有关信息;此外,还可以对被检索文献发表的语种、日期等进行限制。

(五)其他性能

一个优秀的搜索引擎产品还必须查询速度快,具有较好的可维护、可更新性能。系统必须稳定可靠,具有完整的容错、备份、崩溃修复机制,系统即使出错,也可以得到迅速的恢复。

四、互联网络信息检索方法

互联网络信息检索的基本方法有两种:一是通过分类搜索引擎检索;二是通过关键词检索

分类搜索引擎可以清晰方便地查找到某一大类信息,比较符合传统的信息查找方式,尤其适合那些希望了解某一方面(或范围)信息,并不严格限于查

询关键字的用户。但目录式搜索引擎的搜索范围较全文搜索引擎要小许多,尤其是当用户选择类型不当时,这样有可能遗漏某些重要的信息源。分类搜索引擎的分类方法有学科分类和图书分类两种。学科分类由各搜索引擎将搜集来的信息按照学科类型分门别类地进行排列,大多数搜索引擎都提供这种检索方法,只是它们采用的分类标准各不相同。大多分类搜索引擎不提供图书分类搜索,因为图书分类的分类标准来源于图书分类法的基本大类,如我国的《中国图书馆图书分类法》国际上通用的《国际十进分类法》和《杜威十进分类法》等,要求相对比较严格。CERNET网络中心的网络指南针提供图书分类搜索。

关键词检索是直接由搜索引擎提供的检索对话框中输入要检索的关键词进行的检索。输入的关键词可以是单个词汇,也可以是多个词汇,通过组配的方法进行比较复杂的检索。

关键词检索是网络信息检索的主要方法。下面详细说明关键词检索的方法和技巧。

关键词组配检索是根据关键词之间的逻辑关系,利用逻辑运算符把关键词连接起来,构成检索表达式进行的检索。正确的掌握和利用此方法是有效提高网上信息资源检索利用的关键。逻辑运算主要有三种:“逻辑与”、“逻辑或”和“逻辑非”。

逻辑与(通常用“AND”或“*”表示)。检索式为: A AND B或 A* B。可用来表示其所连接的两个检索项的交叉关系,也即交集部分,表示让系统检索同时包含检索词 A和检索词 B的信息集合。

逻辑或(通常用“OR”或“+”表示)。检索式为: A OR B(或 A+ B)。表示让系统查找含有检索词 A B之一,或同时包括检索词 A和检索词 B的信息。

逻辑非(通常用“NOT”或“-”表示)。检索式为: A NOT B(或 A- B)。表示检索含有检索词 A而不含检索词 B的信息。即将包含检索词 B的信息集合排除掉。

网络信息搜索引擎对逻辑运算的支持可以分为如下几种:

(一)支持完全的逻辑运算

支持该种检索功能的检索工具有: Altavista^[3]、Excite Lycos^[3]、Snap AOL Netfind WebCrawler MSN Hotbot Netscape Search 等。在这些检索工具中,用户在搜索框中输入检索词和逻辑运算符。比如,查找不包含美国在内的高校教育方面的信息,其检索式为: (university OR college) AND education

NOT(America OR US OR United States OR us OR USA)。必须注意的是网上检索工具基本上都提供一般检索和高级检索两种检索方式,上述检索必须在高级检索方式中才能进行;而有的搜索引擎其逻辑运算符是隐藏于搜索框附近的菜单中,用户无需输入任何运算符,只需用鼠标点击该菜单上的按钮,即可选择使用。

(二)部分支持逻辑运算

有的搜索引擎只支持三种逻辑运算中的一种或两种。如 Look Smart只支持 OR运算(通过缺省方式实现),不支持 AND NOT运算; Google只支持 AND和 NOT运算,不支持 OR运算;在 AOL的一般检索方式中,只能进行 AND和 AND NOT运算,不支持 OR运算。

(三)用运算符代替逻辑运算

每一种逻辑运算都有一个运算符,很多搜索引擎用运算符代替逻辑运算,连接关键词构成检索表达式,而且大多数的搜索引擎将“或”运算作为缺省设置。支持此种检索方式的检索工具有 AOL Netfind AltaVista Lycos Excite HotBot Infoseek MSN Search Netscape Search WebCrawler Yahoo等。

(四)词语检索

网络中的许多信息,单纯依赖关键词检索和逻辑检索很难检索到。在传统的计算机文献检索中,采用了“邻近(Near)运算”^[2],用来表示关键词之间的关系。网络中的许多检索工具,如 AltaVista等,引进了这种传统的检索方式,实现了词语检索。如查找美国前总统 Bill Clinton的有关资料,如果直接用其名字检索,就会把许多其它 Bill的资料检索出来,此时可以用词语检索: Bill NEAR Clinton来检索,从而提高查准率。

(五)截词检索

截词检索是指利用不完整的词或词根进行的检索。截词检索主要用来提高检索的查全率,扩大检索范围,其缺点是检索结果的准确率降低。绝大多数网络检索工具都支持截词功能。有的是自动截词(如 Lycos),有的是在一定条件下才能截词(如 Alta Vista)。在允许截词的检索工具中,一般是指右截词,部分支持中间截词(如 Archie),有的需要使用通配符,如“*”、“?”等,如要查找计算机辅助设计或制造,可以使用截词检索:“CA*”或“CA?”。

五、网络信息检索中存在的问题

尽管国际互联网检索工具的发展已具有一定规

模和达到一定层次,然而,作为一个整体,还存在查准率差的问题。网络信息检索,尤其是万维网信息的检索,经常会检得成千上万条无用信息。总的来说,Internet搜索引擎存在如下问题:

(一)缺乏网络信息质量控制^[1]

任何个人团体,只要具备上网条件,知道如何使用超文本标识语言,就可以把任何信息放到网上。这些信息经过种种检索工具的标引,就可供用户查询,中间没有任何形式的质量控制。未经质量控制的信息,必然影响检索结果的查准率。

(二)网络检索工具的功能尚待完善^[4]

与传统计算机检索工具相比,网络检索工具尚不能修改原有检索结果,每次检索都是重新开始,不能对原有结果加以利用。由于网络文件的结构特殊(如超文本),且不按传统意义(如著者或篇名)的字段进行检索。目前还没有任何一个网络检索工具可在检索功能上与传统计算机化的检索工具相媲美。

(三)缺乏检索词汇控制^[1]

几乎所有的网络检索工具都采用自然语言标引和检索,其必然结果是同义词和近义词得不到控制,词间相互关系得不到揭示,最终影响检索效果。

(四)自动标引的局限性^[2]

自动标引虽然省时省力,但不可避免地给检索带来一些问题和困难。这些问题和困难最突出地表现在自动标引不可能像人工标引那样进行智能甄别和选择,而只能依赖关键词词频等标准判断网络文件的价值。

(五)逻辑运算无统一标准

搜索引擎中有的用 AND OR NOT;有的用“+”、“-”号代替 AND NOT,而将逻辑或 OR 设为缺省值;有的则是 AND NOT 两种符号都采用。

(六)支持功能不统一

有的搜索引擎具备完整的逻辑检索功能,有的则只支持部分逻辑检索功能,比如有的检索工具能与圆括号()结合进行复杂的课题检索,而有的检索工具则不能。

(七)使用途径不统一

有的搜索引擎必须在其高级检索方式中才能使用(如只能用 AND 而不能用“+”,只能用 NOT 而不能用“-”);有的必须在一般检索方式中才能使用,有的则可在两种检索方式中混合使用。

为了提高 WWW 搜索引擎的检索质量,在搜索引擎的开发中应注意完善搜索引擎的功能,增加检索途径和限定提高查准率;同时在信息标引时采用词频和词表相结合的办法,加强对检索词汇的控制,并提高标引速度。

参考文献:

- [1] 曾民族. 网络信息检索现状和性能评价[J]. 情报学报, 1997, (2).
- [2] 王云. 内容标示语言与统一内容定位[J]. 计算机世界, 1998, (50).
- [3] 刁倩. Internet 上的英文搜索引擎[J]. 计算机工程, 1999, (7).
- [4] 杨惠. 基于 WWW 的多媒体信息搜索系统[J]. 计算机系统应用, 1999, (8).

(上接第 73 页)

来说这本身是一个进步,后勤职工应当适应这个转变。改革要经历一个痛苦的过程,必然伴随着利益的调整。后勤职工应努力学习,认真提高自身素质,不忘服务育人的宗旨。在改革大潮中经风雨,见世面。对于学校来讲,在一段时间内,后勤虽然剥离,仍是高校的一个组成部分,应当从政策上、资金上给予一些扶持和帮助,用通俗的话说叫“扶上马,送一程”,能够使得高校后勤产业在高校后勤改革中稳妥的走下去,最终为高校的发展起到服务、促进和稳定的作用。

李岚清副总理在总结了全国高校后勤社会化工作时指出,推行高校后勤社会化一定要坚持“政府主导,政策扶持,社会参与,八方抬‘教’”的 16 字方针。

同时要求各高校应加强对此项工作的领导,在政府的指导下有计划有目标的逐步实施。由于中国经济秩序还有待整顿,不能一推了之。对于后勤干部的配备,产权的明晰,甲乙方的职责、分工等,学校都应当过问、把关;对于后勤产业的产品质量,服务质量,育人意识,价格体系等都应该施行合理的监督和评价,并及时给予指导和调整。在运行初期,后勤产业仍是学校的一个组成部分,学校还应给予一定的资金和政策上的支持及倾斜,使其健康稳定地持续发展。

参考文献:

- [1] 钟顺虎. 高校后勤系统重组研究[M]. 西安: 陕西人民出版社, 2001.