信息搜索引擎综述

许晋军 苏新宁 (南京大学信息管理系 210093)

摘要:本文在简要描述了因特网上信息搜索工具的基础上,论述了信息搜索引擎的工作原理以及改进与发展的方法。

关键词:信息检索 因特网 搜索引擎

自从 1993 年美国克林顿政府提出:"国家信息基础结构"计划以来,"信息高速公路"的热浪迅速席卷全球。世界各国竞相支持,相继推动自己的信息化建设规划,迎接即将到来的 21 世纪信息社会。中国在信息基础设施的建设中也采取了积极的态势。短短两年的时间,便出现了中国科学院网络中心的 NCFC 网、国家教委的 CERNET 网。邮电部的 CHINA NET 和金桥网等许多大型网络。

对因特网(Internet)的重视与发展使得各类网络信息资源以惊人的速度增长。如何快捷而准确地检索出自己所需信息,避免陷入浩瀚的信息海洋而不知所措已成为摆在人们面前的一个不大不小的难题。

1. 信息搜索引擎

因特网起源于美国。美国的信息产业在近 20 年取得了长足的发展,如今已相当发达,信息搜索工具更是层出不穷。客观上基于英文的超文本信息检索系统起步较早,近年来又得到了迅速的发展,技术上也相当成熟,市场上业绩不凡。如 Altavista, Yahoo! Web Crawler, lycos, Infoseek 等著名的检索工具已经广为人们熟知。

(1)AltaVista 信息搜索引擎

早在 1995 年,美国 DEC 公司就在加州 Palo Alto 的研究室中开发出了 Altavista。它是一种能够对整个 Internet 资源进行索引的工具,基于 DEC 公司的 Alpha 硬件平台,以令人难以置信的查询速度解决了世界上最大的信息源 Internet 网络的检索难题。我们利用 AltaVista 可以查到任何一个在 WWW 或 Usenet(新闻组)中出现的词。事实上 Alta Vista 已经成为一个非常有效的查找信息的工具,并且具有着广阔的商业应用前景。Alta Vista 全部 Web 索引总容量为 60GB,在强大的软件和 Internet 网以及 DEC 公司的 Alpha 技术支持下,可以让每一个用户在世界上任何地方的任何一台连上互连网络的微机查询 Internet 网上的信息。

(2)Infoseek 信息搜索引擎

Infoseek 于 1995 年 2 月由美国 Infoseek 公司推出，是 Web 上第一家收费的查询系统。从一开始的只对 110 万 URL(统一资源定位)进行索引到 1996 年 4 月启用新的 Ultra Smart/Ultraseek 服务之后，已能对 8000 多万个 URL，包括 Web，过去四周的 News Group(新闻组)，FAQ；文件传输站点(FTP)、Gopher 等进行全文检索。

作为一家收费的商业站点，Infoseek 对网络信息查询服务除了主要的 UltraSmart/Ultraseek 外，还有 News Center，Smart Info Infoseek Spotlight Program 以及 Big Yellow(黄页)等多种项目。其中，News Center 主要可供查询来自路透社、有线商业网等的有线新闻以及主要来自 7 个新闻机构的当日重要新闻，如 CNN，纽约时报，华盛顿邮报等，内容包括世界各地的时事新闻、商业信息、技术信息、体育信息、娱乐信息等各个方面。

2.信息搜索引擎的工作原理

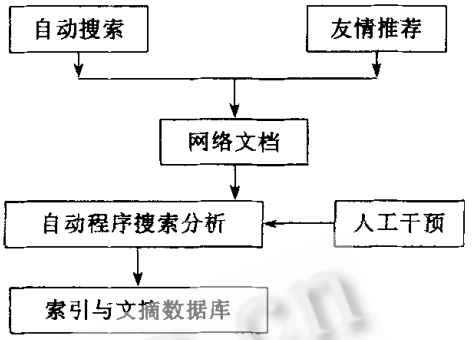
网络上信息搜索引擎基本上都是以全文检索技术为支持的。全文检索是信息检索发展过程中的一个重要分支。它迅速发展于 70 到 80 年代，90 年代以来得到了广泛的应用。信息的繁冗复杂促使信息整序的研究进一步深化，抽取能够表现出信息内在特征的值然后建立索引，在用户检索时再与需求进行匹配成为全文检索的核心任务。

完成信息搜索引擎的任务，需要两个过程。其一是在服务器方也就是服务提供者对网络信息资源进行搜索分析标引的过程；其二是当用户方提出检索需求时，服务器方搜索自己的信息索引库然后发送给用户的过程。前者可以称作信息标引过程，后者可以称作提供检索过程。

(1)信息标引过程

信息标引过程是服务方对信息资源进行整序的过程。目前主要采用两种方式，一种是网络自动漫游方式，由计算机程序自动去搜索资源，另一种是友情推荐方式，由信息发布方或者用户将有用信息的网络地址(URL)填入搜索清单，然后再由机器程序对指定地址进行搜索。

工作方式如下图所示：

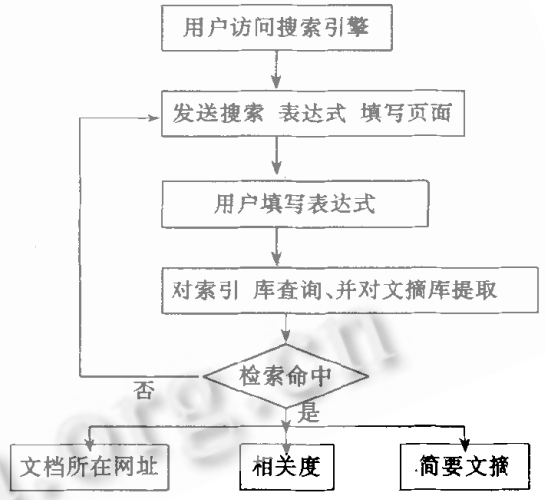


计算机自动程序定期在网络上漫游，对各种文档进行索引分析，将结果记录进数据库，人工也可以干预此过程期望更为准确地表达出文档原意。有的搜索引擎(如中文搜索引擎中北京大学的“天网”，就会在检索结果中同时附上一段简要文摘)在搜索过程中使用了自动文摘的技术生成了文摘数据库。

(2)提供检索过程

提供检索的过程就是根据用户检索需求表达式来进行查找与输出结果的过程。它建立在对网络信息标引的索引库与文摘库之上。

工作原理如下图所示：



用户通过检索表达式页面的填写反映出自己的检索意向，向系统送交请求。系统答复后，用户可以根据具体情况(包括相关度、文摘等所能反映出的状况)再决定是否访问资源所在地。信息搜索引擎在整个信息检索过程中起到了指南和向导的作用，无疑大大地方便了人们的检索。

3.信息搜索引擎系统功能的探讨

信息搜索引擎既然称为 Internet 网络的信息导航员，就应当全心全意地为人们服务。分析了系统的特点与功能后，可以从以下几个方面改进和提高其性能。

(1)自动漫游与友情推荐相结合

搜索引擎一般来说都是通过称做“WebRobot”、“We-

bCrawler"、"Spider"一类的自动程序在网络上漫游，从时间上和空间上反映出网络资源的浩瀚，这已经有效地提高了人们的工作效率。同时系统引用了用户友情推荐的方式无疑在既方便用户登录自己信息的同时又减少了系统搜索的很大盲目性。系统具有与环境交往的过程与功能，如果能够得到进一步发展与用户的认真使用，必将在提高查全率的基础上也提高了查准率。

(2)主题与分类的结合细化

传统信息检索经过长期理论与实践的发展，在内容上基本形成了主题与分类两个不同角度。分类以特性检索为特点，主题以特性检索为优势，两者反映了人类思维的两个不同侧面并相互弥补。分类方面有长期研究的结果——分类法，主题方面也有重要贡献——主题词表，二者各自依赖于自身的配套表，为检索结果的高准确度奠定了扎实的基础。另一方面也由于词表对照的不便，使普通用户颇感不便。

网络上信息检索目前大都以关键词为入口，这是一种类似于主题检索的方法，关键词的任意性大大方便了用户但也导致了过分依赖匹配用户需求外在表达方式的缺点。考虑在关键词的基础上适当采用主题词表中"用"、"代"关系很有可能取得良好效果，使检索系统更富智能性。分类方法便于用户难以表达具体主题时使用，通过一层一层地细化，找到用户感兴趣的内容。即使不很确切，分类方法也缩减了检索范围。这时如果再让用户输入自己所需的关键词，就会有更多的可能性命中用户需求。Infoseek 检索系统就很好地将分类检索与主题检索结合在一起，供不同层次的用户选择使用。

(3)提供定制检索，开展定题信息服务

信息搜索引擎通过大量辛苦地劳动形成了索引库期待着用户来检索，这是一种被动服务的方式。如何才能变得主动呢？在传统的情报系统中有一种称作 SDI 的服务方式。SDI 即 Selective Dissemination of Information 的缩写，意思是定题情报服务。

定题情报服务指某一机构所从事的这样一种情报服务，它把自任何来源取得的情报新要素发往该机构的一些点上，把这些情报新要素应用于目前工作的概率最高，或者对此情报的需要最迫切的地方。

信息搜索引擎引进定题情报服务，可以采用一种称作"信息推送"的技术。通过事先与用户达成的协议将需求表达式预先存放于搜索系统中。当检索到满足用户需求的资源后就累积起来并定时地发送给定题用户。用户收到通知或检索结果后审阅并确定是否查看资源所在地址(也可以是服务方将原文获得后保存在本地供用户索取或者发入用户的邮箱)。

(4)文摘支持

网络上资源多而复杂，用户在通过搜索引擎得到网络地址后；更希望能够看到有关文档的简要内容而不仅仅是一个对关键词的匹配度排序，自动文摘正好能满足用户的需求。

虽然机器自动编写文摘还没有非常实用的研究成果，但已经有了一些成熟的思路与方法。对于用户来说，文摘的推出只具有参考价值，就足够了，并不象传统情报加工中文摘的质量要求那么严格。

分析手工编写文摘的方法可以用机器来模仿，仿人算法在实践中证明的是一种比较有效的方法。浏览全文抓住文献的中心思想与关键内容，用精简的语言书写比较规范化的文摘，是文摘员的基本方法。针对计算机而言，由于人工智能语言理解尚没有取得突破性进展，暂时可以用词频统计。字频统计等方法大致确定文献的关键内容。对文章的段首段尾给予较重的权值，文献的标题、副标题、段落小标题、文章的起始段、结束段都具有比较重要的意义。抽取文章关键字和较重要的句子，组织成一篇简短的摘要对于计算机来说是很善长的工作，对于用户来说则有相当的参考价值。

4.结束语

因特网上资源异常博大而丰富，对于用户来说喜忧参半，检索特定信息成为一个研究的热点话题，并且带动了全球网络信息检索技术的空前活跃与发展。有关 WWW 网络信息查询工具的研究热点尤显突出。就目前来说这方面的实际运营成果也有数百个之多。可以说已成为因特网信息检索方式的主流，并大有发展成为网络标准检索工具的优势。

参考文献

[1] http://www.altavista.digital.com
[2] http://www.Yahoo.com
[3] http://www.sohoo.com
[4] http://www.Pku.edu.cn

(来稿时间:1998 年 10 月)