华夏英才基金学术文库

# 搜 索 引 擎

## ——原理、技术与系统

李晓明 闫宏飞 王继民 著

科学出版社

北 京

## 内 容 简 介

本书系统地介绍了互联网搜索引擎的工作原理、实现技术及其系统构建方案。全书分三篇共13章内容,从基本工作原理概述,到一个小型简单搜索引擎具体细节的实现,进而详细讨论了大规模分布式搜索引擎系统的设计要点及其关键技术;最后介绍了面向主题和个性化的Web信息服务,阐述了中文网页自动分类等技术及其应用。本书层次分明,由浅入深;既有深入的理论分析,也有大量的实验数据,具有学习和实用双重意义。

本书可作为高等院校计算机科学与技术、信息管理与信息系统、电子商务等专业的研究生或高年级本科生的教学参考书和技术资料,对广大从事网络技术、Web站点的管理、数字图书馆、Web挖掘等研究和应用开发的科技人员也有很高的参考价值。

**图书在版编目(CIP)数据**

搜索引擎:原理、技术与系统/李晓明,闫宏飞,王继民著. —北京:科学出版社,2005
(华夏英才基金学术文库)
ISBN 7-03-014633-6

Ⅰ. 搜… Ⅱ. ①李…②闫…③王… Ⅲ. 因特网-情报检索 Ⅳ. G252.7

中国版本图书馆CIP数据核字(2004)第121546号

责任编辑:巴建芬 姚庆爽/责任校对:陈玉凤
责任印制:钱玉芬/封面设计:陈 敬

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com

印刷

科学出版社发行 各地新华书店经销

*

2005年4月第 一 版 开本:B5(720×1000)
2006年1月第二次印刷 印张:16 1/2
印数:3 001—5 000 字数:312 000

**定价:33.00元**

(如有印装质量问题,我社负责调换〈环伟〉)

# 前　言

随着互联网的不断发展和日益普及,网上的信息量在爆炸性增长,全球 Web 页面的数目已经超过 40 亿,中国的网页数目估计也超过了 3 亿。目前人们从网上获得信息的主要工具是浏览器,而通过浏览器得到信息通常有三种方式:第一,直接向浏览器输入一个关心的网址(URL),如 http://net.pku.edu.cn,浏览器返回所请求的网页,根据该网页内容及其包含的超链接文字(anchor text)的引导,获得自己需要的内容;第二,登录到某个知名门户网站,如 http://www.yahoo.com,根据该网站提供的分类目录和相关链接,逐步"冲浪"浏览,寻找自己感兴趣的东西;第三,登录到某个搜索引擎网站,如 http://e.pku.edu.cn,输入代表自己所关心信息的关键词或者短语,依据返回的相关信息列表、摘要和超链接引导,试探寻找自己需要的内容。

这三种方式各有特点,各有自己最适合的应用场合。第一种方式的应用是最有针对性的,例如,要了解北京大学计算机系网络与分布式系统实验室在做些什么工作,从某个渠道得知该实验室的网址为 http://net.pku.edu.cn,于是直接用它驱动浏览器就是最有效的方式。第二种方式的应用类似于读报,用户不一定有明确的目的,只是想看看网上有什么有意思的消息;当然这其中也可能是关心某种主题,如体育比赛、家庭生活等。第三种方式适用于用户大致知道自己要关心的内容,如"国有股减持",但不清楚哪里能够找到相关信息(即不知道哪些 URL 能给出这样的信息);在这种场合,搜索引擎能够为用户提供一个相关内容的网址及其摘要的列表,由用户一个个试探看是否为自己需要的。现在的搜索引擎技术已经能做到在多数情况下满足用户的这种需要。CNNIC 的信息统计指出,目前搜索引擎已经成为继电子邮件之后人们用得最多的网上信息服务系统。

同时,随着网上信息资源规模的增长,尤其是其内容总体和我们社会的演化发生着越来越密切的联系,研究网上存在的海量信息逐渐成为许多学科关注的一个方向。为此,不少研究人员也有采样搜集特定内容、一定数量网页的需要。

本书以我们设计、实现并维护、运行北大"天网"搜索引擎的实际经验,介绍大规模搜索引擎的工作原理和实现技术。我们要向读者揭示,为什么向搜索引擎输入一个关键词或者短语,就能够在几秒钟内得到那么多相关的文档及其摘要,而点击其中的链接就能够被引导到文档的全文,且其中相当一部分可能正是用户需要的。

我们按照上、中、下三篇展开相关的内容。上篇讲搜索引擎的基本工作原理,要解决的是为什么搜索引擎能提供如此庞大的信息查找服务这一问题,以及它在功能上有什么本质的局限性。这一篇的内容包括网页的搜集过程,网页信息的提取、

组织方式和索引结构,查询提交和响应的过程以及结果产生,等等。这其中,虽然我们假定读者熟悉 URL、HTML、HTTP、CGI、MIME 等基本概念,但在上下文中也给予了必要的介绍,力图保持行文的流畅性。这一部分内容对于需要构建小规模搜索引擎的研究人员会有直接的参考价值。

中篇讨论和大规模实用搜索引擎有关的技术问题。所谓大规模在这里指至少维护超过 1000 万的网页信息,提供相关的查询服务。所涉及的内容包括并行分布处理技术的应用,数据局部性的开发,缓存技术的应用以及搜集的网页在提供服务之前的预处理问题和高效倒排文件的建立技术等。这一部分的讨论有比较强的计算机系统结构的风格,我们将向读者展示计算机系统结构课程中的那些概念是如何生动地体现在一个实际应用系统中的。这一部分内容对构建大规模数字图书馆的技术人员也应该有帮助。

下篇介绍挑战性更强一些的内容。一般地讲,前面所述可以称为是"通用搜索引擎",为最广泛的人群提供信息查询服务是它的基本宗旨。这意味着它的应用模式必须尽量简单,即关键词或查询短语的提交和匹配响应。尽管这已经可以解决许多问题了,但对有些重要的信息需求依然显得力不从心。例如,一个人可能会关心最近半年来网上出现了哪些关于他(她)的信息,一个企业可能要关心它做了一次大规模促销活动后一个月内网上有什么反响,一个政府机构可能会关心在一项政策法规颁布后的网上舆论。面向主题和个性化的信息查询服务就是我们试图描述的一种基本途径。这一部分内容更多地和网上中文信息处理技术有关。更准确地讲,我们要介绍网络与并行分布处理技术和中文处理技术的结合,从而实现大规模、高性能、高质量、有针对性的网上信息查询服务。这一部分内容反过来可能对从事中文信息处理的研究人员有启发作用。

本书的内容是集体智慧的结晶,主要概括了北京大学计算机科学技术系网络与分布式系统实验室自 1996 年以来的研究成果。其中许多段落直接来自同学的博士和硕士论文,他们是雷鸣、赵江华、冯是聪、单松巍、谢正茂、彭波、张志刚、龚笔宏、孟涛、咎红英等。署名作者的主要工作是将这些内容系统化,使其表述的风格统一。我们特别感谢陈葆珏教授,是她在北京大学计算机系开创了搜索引擎这一研究方向,从而使我们能在其后发扬光大,还要感谢刘建国和王建勇,是他们分别带领攻关队伍,实现了天网 1.0 和天网 2.0 版本。感谢黄蕊为本书进行的文字校对。最后,我们要感谢国家九五攻关计划、973 计划和 985 计划的支持,是它们的不断支持使我们得以将天网不断推上新的台阶,实现"让天网和中国网上信息资源规模同步成长"的理想。

作　者

2004 年 5 月于北大燕园

# 图 表 目 录

# 目　　录

## 中篇　对质量和性能的追求

## 下篇　面向主题和个性化的 Web 信息服务

# 第一章　引　　论

信息的生产、传播、搜集与查询是人类最基本的活动之一。考虑以文字为载体的信息,传统上有图书馆、相应的编目体系和专业人员帮助我们很快找到所需的信息,其粒度通常是“书”或者“文章”。随着计算机与信息技术的发展,有了信息检索(information retrieval,IR)学科领域,有了关于图书或者文献的全文检索系统,使我们能很方便地在“关键词”的粒度上得到相关的信息。

我们注意到,上述全文检索系统一般工作在一个规模相对有限、内容相对稳定的馆藏(collection)上,被检索的对象通常是经过认真筛选和预先处理的(如人工提取出了“作者”、“标题”等元数据,形成了很好的“摘要”等),并且系统需要同时响应的查询数量通常都不会太大(如每秒钟 10 个左右)。

1994 年左右,万维网(World Wide Web,简记为 WWW 或 Web)出现。它的开放性(openness)和其上信息广泛的可访问性(accessibility)极大地鼓励了人们创作的积极性。作为一个信息源,Web 和上述全文检索系统的工作对象相比,具有许多不同的特征,它们给信息检索领域带来了新的发展机遇和技术挑战。

规模大。在短短的 10 年左右时间,人类至少生产了 40 亿网页(Google　2004),而人类有文字以来上万年里产生了大约 1 亿本书;中国网上到 2004 年初大致有了约 3 亿网页(天网　2004),而中华民族有史以来出版的书籍大约不过 275 万种。尽管书籍的容量和质量是一般网页不可比的,但在对应的时间背景上考察其文字的总体数量,我们不能不为人类在 Web 上创造文字的激情惊叹!

内容不稳定。除了不断有新的网页出现外,旧的网页也可能会因为各种原因被删除(有研究指出:50% 网页的平均生命周期大约为 50 天(Cho et al.　2000,Cho　2002))。

从原则上讲,读者数和作者数在同一个量级,形式和内容的随意性很强,权威性相对也不高,也不太可能进行人工筛选和预处理。

与生俱来的数字化、网络化。传统载体上的信息,人们目前正忙于将它们数字化、上网(花费极高),而网络信息天生如此。这个特性是一把双刃剑:一方面便于我们搜集和处理,另一方面也会使我们感到太多,蜂拥而至、鱼目混珠。

而作为要在 Web 上提供服务的信息查询系统,如搜索引擎和数字图书馆,通常要具备同时对付大量访问的能力(如每秒钟 1000 个查询),而且响应时间还要足够的快(如 1 秒钟)。

本书旨在介绍构建这类搜索引擎的有关技术。传统的 IR 是其基础,同时本书

也充分讨论了由上述 Web 信息的特征所带来的新问题及其解决方案。

## 第一节　搜索引擎的概念

如上所述,本书的主要内容是介绍搜索引擎的工作原理和实现技术。搜索引擎,在本书指的是一种在 Web 上应用的软件系统,它以一定的策略在 Web 上搜集和发现信息,在对信息进行处理和组织后,为用户提供 Web 信息查询服务。从使用者的角度看,这种软件系统提供一个网页界面,让他通过浏览器提交一个词语或者短语,然后很快返回一个可能和用户输入内容相关的信息列表(常常会是很长一个列表,如包含 1 万个条目)。这个列表中的每一条目代表一篇网页,每个条目至少有三个元素:

1) 标题:以某种方式得到的网页内容的标题。最简单的方式就是从网页的 < TITLE> < /TITLE> 标签中提取的内容(尽管在一些情况下并不真正反映网页的内容)。本书第七章会介绍其他形成"标题"的方法。

2) URL:该网页对应的"访问地址"。有经验的 Web 用户常常可以通过这个元素对网页内容的权威性进行判断,例如,http://www.people.com 上面的内容通常就比 http://notresponsible.net(某个假想的个人网站)上的要更权威些(不排除后者上的内容更有趣些)。

3) 摘要:以某种方式得到的网页内容的摘要。最简单的一种方式就是将网页内容的头若干字节(如前 512 字节)截取下来作为摘要。本书第七章会介绍形成"摘要"的其他方法。

通过浏览这些元素,用户对相应的网页是否真正包含他所需的信息进行判断。比较肯定的话则可以点击上述 URL,从而得到该网页的全文。图 1-1 是 2003 年 8 月 20 日在天网搜索引擎(http://e.pku.edu.cn)上的一个例子,用户提交了查询词"伊拉克战争",系统返回一个相关信息列表。列表的每一条目所含内容比上述要丰富些,但核心还是那三个元素。如果用户主要是想从军事角度关心伊拉克战争,第一条目可能就是很好的选择,不仅摘要看起来军事味道要浓一些,而且从 URL (http://mil.eastday.com)上能看到提供信息的大概是一个专门的军事题材网站。如果用户主要是想关心伊拉克战争对全球经济的影响,则后面的条目可能会更相关些。

这个例子提示了我们一个重要的情况,即搜索引擎提供信息查询服务的时候,它面对的只是查询词。而有不同背景的人可能提交相同的查询词,关心的是和这个查询词相关的不同方面的信息,但搜索引擎通常是不知道用户背景的,因此搜索引擎既要争取不漏掉任何相关的信息,还要争取将那些"最可能被关心"的信息排在列表的前面。这也就是对搜索引擎的根本要求。除此以外,考虑到搜索引擎的应用


图 1-1 2003 年 8 月 20 日在天网上检索“伊拉克战争”的结果

环境是 Web,因此对大量并发用户查询的响应性能也是一个不能忽略的方面。

作为对搜索引擎工作原理的基本了解,这里有两个问题需要首先澄清。第一,当用户提交查询的时候,搜索引擎并不是即刻在 Web 上“搜索”一通,发现那些相关的网页,形成列表呈现给用户;而是事先已“搜集”了一批网页,以某种方式存放在系统中,此时的搜索只是在系统内部进行而已。第二,当用户感到返回结果列表中的某一项很可能是他需要的,从而点击 URL,获得网页全文的时候,他此时访问的则是网页的原始出处。于是,从理论上讲搜索引擎并不保证用户在返回结果列表上看到的标题和摘要内容与他点击 URL 所看到的内容一致(上面那个“伊拉克战争”的例子就是如此),甚至不保证那个网页还存在。这也是搜索引擎和传统信息检索系统的一个重要区别。这种区别源于前述 Web 信息的基本特征。为了弥补这个差别,现代搜索引擎都保存网页搜集过程中得到的网页全文,并在返回结果列表中提供“网页快照”或“历史网页”链接,保证让用户能看到和摘要信息一致的内容。

## 第二节 搜索引擎的发展历史

早在 Web 出现之前,互联网上就已经存在许多旨在让人们共享的信息资源了。那些资源当时主要存在于各种允许匿名访问的 FTP 站点(anonymous FTP),

内容以学术技术报告、研究性软件居多,它们以计算机文件的形式存在,文字材料的编码通常是 PostScript 或者纯文本(那时还没有 HTML)。

为了便于人们在分散的 FTP 资源中找到所需的东西,加拿大麦吉尔大学(University of McGill)计算机学院的师生于 1990 年开发了一个软件,Archie。它通过定期搜集并分析 FTP 系统中存在的文件名信息,提供查找分布在各个 FTP 主机中文件的服务。Archie 能在只知道文件名的前提下,为用户找到这个文件所在的 FTP 服务器的地址。Archie 实际上是一个大型的数据库,再加上与这个大型数据库相关联的一套检索方法。该数据库中包括大量可通过 FTP 下载的文件资源的有关信息,包括这些资源的文件名、文件长度、存放该文件的计算机名及目录名等。尽管所提供服务的信息资源对象(非 HTML 文件)和本书所讨论搜索引擎的信息资源对象(HTML 网页)不一样,但基本工作方式是相同的(自动搜集分布在广域网上的信息,建立索引,提供检索服务),因此人们公认 Archie 为现代搜索引擎的鼻祖。

值得一提的是,即使是在 10 多年后的今天,以 FTP 文件为对象的信息检索服务技术依然在发展,尤其是在用户使用界面上充分采用了 Web 风格。北大天网文件检索系统就是一个例子(见 http://bingle.pku.edu.cn)。不过鉴于本书写作定位的关系,后面将主要讨论网页搜索引擎的相关问题。

以 Web 网页为对象的搜索引擎和以 FTP 文件为对象的检索系统一个基本的不同点在于搜集信息的过程。前者是利用 HTML 文档之间的链接关系,在 Web 上一个网页一个网页地“爬取”(crawl),将那些网页“抓”(fetch)到本地后进行分析;后者则是根据已有的关于 FTP 站点地址的知识(如得到了一个站点地址列表),对那些站点进行访问,获得其文件目录信息,并不真正将那些文件下载到系统上来。因此,如何在 Web 上“爬取”,就是搜索引擎要解决的一个基本问题。在这方面,1993 年 Matthew Gray 开发了 World Wide Web Wanderer,它是世界上第一个利用 HTML 网页之间的链接关系来监测 Web 发展规模的“机器人”(robot)程序。刚开始时它只用来统计互联网上的服务器数量,后来则发展为能够通过它检索网站域名。鉴于其在 Web 上沿超链“爬行”的工作方式,这种程序有时也称为“蜘蛛”(spider)。因此,在文献中 crawler、spider、robot 一般都指的是相同的事物,即在 Web 上依照网页之间的超链关系一个个抓取网页的程序,通常也称为“搜集”。在搜索引擎系统中,也称为网页搜集子系统。

现代搜索引擎的思路源于 Wanderer,不少人在 Matthew Gray 工作的基础上对它的蜘蛛程序做了改进。1994 年 7 月,Michael Mauldin 将 John Leavitt 的蜘蛛程序接入到其索引程序中,创建了大家现在熟知的 Lycos,成为第一个现代意义的搜索引擎。在那之后,随着 Web 上信息的爆炸性增长,搜索引擎的应用价值也越来越高,不断有更新、更强的搜索引擎系统推出(本章第三节会有介绍)。这其中,特别

引人注目的是 Google(http://www.google.com),虽然是个姗姗来迟者(1998 年才推出),但由于其采用了独特的 PageRank 技术,使它很快后来居上,成为当前全球最受欢迎的搜索引擎(作者 2003 年初访问印度,就听到总统阿卜杜勒·卡拉姆讲他经常用 Google 在网上查找信息)。

在中国,据我们所知,对搜索引擎的研究起源于“中国教育科研网”(CERNET)一期工程中的子项目,北京大学计算机系的项目组在陈葆珏教授的主持下于 1997 年 10 月在 CERNET 上推出了天网搜索 1.0 版本。该系统在这几年里不断发展,目前已成为中国最大的公益性搜索引擎(http://e.pku.edu.cn)。在这之后,几位在美国留学的华人学者回国创业,成立了百度公司,于 2000 年推出了“百度”商业搜索引擎(http://www.baidu.com),并一直处于国内搜索引擎的领先地位。我们看到慧聪公司也在中国推出了一个大规模搜索引擎(http://www.zhongsou.com),用起来感觉也不错,但往后发展如何,还有待时间的考验。

当我们谈及搜索引擎的时候,不应该忽略另外一个几乎是同期发展出来的事物:基于目录的信息服务网站。1994 年 4 月,斯坦福(Stanford)大学的两名博士生,David Filo 和杨致远(Gerry Yang)共同创办了 Yahoo! 门户网站,并成功地使网络信息搜索的概念深入人心。1996 年中国出现了类似的网站,“搜狐”(http://www.sohu.com)。在许多场合,也称 Yahoo! 之类的门户网站提供的信息查找功能为搜索引擎。但从技术上讲,这样的门户中提供的搜索服务和前述搜索引擎是很不同的。这样的门户依赖的是人工整理的网站分类目录,一方面,用户可以直接沿着目录导航,定位到他所关心的信息;另一方面,用户也可以提交查询词,让系统将他直接引导到和该查询词最匹配的网站。图 1-2 就是我们在搜狐上查询“伊拉克战争”的结果。和图 1-1 相比,不难看到其风格是很不相同的。在需要区别的场合,我们可以分别称“自动搜索引擎”和“目录搜索引擎”,或者“网页搜索引擎”和“网站搜索引擎”。一般来讲,前者的信息搜索会更全面些,后者则会准确些。在没有特殊说明的情况下,本书中所讨论的“搜索引擎”不包括 Yahoo! 和搜狐这样的搜索方式。

随着网上信息越来越多,单纯靠人工整理网站目录取得较高精度查询结果的优势逐渐退化——对海量的信息进行高质量的人工分类已经不太现实。目前有两个发展方向。一是利用文本自动分类技术,在搜索引擎上提供对每篇网页的自动分类,这方面最先看到的例子是 Google 的“网页分类”选项,但它分类的对象只是英文网页。在中文方面,文本自动分类的研究工作有很多,但我们知道的第一个在网上提供较大规模网页自动分类服务的是北大网络实验室冯是聪和龚笔宏等人的工作(冯是聪 2003),他们于 2002 年 10 月在天网搜索上挂接了一个 300 万网页的分类目录。另一个发展方向是将自动网页爬取和一定的人工分类目录相结合,希望形成一个既有高信息覆盖率,也有高查询准确性的服务。

互联网上信息量在不断增加,信息的种类也在不断增加。例如,除了我们前面


图 1-2　2003 年 8 月 20 日在搜狐上检索“伊拉克战争”的结果

提到的网页和文件，还有新闻组、论坛、专业数据库等。同时上网的人数也在不断增加，网民的成分也在发生变化。一个搜索引擎要覆盖所有的网上信息查找需求已出现困难，因此各种主题搜索引擎、个性化搜索引擎、问答式搜索引擎等纷纷兴起。这些搜索引擎虽然还没有实现如通用搜索引擎那样的大规模应用，但随着互联网的发展，我们相信它们的生命力会越来越旺盛。另外，即使通用搜索引擎的运行现在也开始出现分工协作，有了专业的搜索引擎技术和搜索数据库服务提供商。如美国的 Inktomi，它本身并不是直接面向用户的搜索引擎，但向包括 Overture（原 GoTo）、LookSmart、MSN、HotBot 等在内的其他搜索引擎提供全文网页搜集服务。从这个意义上说，它是搜索引擎数据的来源。

搜索引擎出现虽然只有 10 年左右的历史，但在 Web 上已经有了确定不移的地位。据 CNNIC 统计，它已经成为继电子邮件之后的第二大 Web 应用。虽然它的基本工作原理已经相当稳定，但在其质量、性能和服务方式等方面的提高空间依然很大，研究成果层出不穷，是每年 WWW 学术年会[①]的重要论题之一。

① International WWW Conference Committee，网址 http://www.iw3c2.org.

## 第三节 一些著名的搜索引擎

为了让感兴趣的读者有目的的试一试,我们整理了一些当前主流的搜索引擎,包括网址、首页面图片及其介绍。在这些搜索引擎中,排在最前面的几个提供多语言的支持,可以满足不同母语读者的需求。

主流搜索引擎的选定参考了(Sullivan 2004),为使读者有感性认识,特别加入了每个网站的相关页面。主流搜索引擎是指非常有名,或者被广泛使用的搜索引擎。

Google, http:// www. google com


四次荣获 Searchenginewatch(Searchenginewatch 2004)读者选举出的"最杰出搜索引擎"称号的 Google 作为在网络上搜索页面的首选是无愧于这个称号的。它基于搜集器[①] 的服务既保证了能够覆盖广泛的网页,同时在查询效果上也表现得极其优秀。

为了方便的检索到所需网页,Google 提供几种可供选择的方法。利用 Google 首页搜索框上面的标签,可以容易的检索网络上的网页、图像、网上论坛、新闻和

① 自动搜索引擎的搜集子系统。

Open Directory 提供的经过人工整理后的网页目录。

Google 还因为提供许多其他特性而闻名,如网页快照,保证您在存有网页的服务器暂时出现故障时仍可浏览该网页的内容,或者可以浏览到不是最新版的该网页的内容;拼写检查,如果您查询词包含错误的拼写,它会提示正确的查询词;股票行情查询;街区地图查询等特殊功能。更多的特性可以查看 Google 的帮助大全。此外,Google 工具条因为提供了方便存取 Google 和它的特性而为其赢得了一定的声誉。

Google 除了提供无需付费的排序结果,还有自己的竞价排名程序。与其他提供此项服务的公司一样,依据点击才有花费,竞价排名程序在 Google 的返回结果中放置广告。Google 还提供自己的无需付费的排序结果给其他一些搜索引擎。

Google 最初起源于斯坦福大学的 BackRub 项目,当时是由学生 Larry Page 和 Sergey Brin 主要负责。到了 1998 年,BackRub 更名为 Google,并且走出校园成为一个公司。

AllTheWeb, http://www.alltheweb.com


作为一个优秀的基于搜集器的搜索引擎,AllTheWeb 提供广泛的网络覆盖与显著的相关性。除了提供网页查询,AllTheWeb 还提供新闻、图像、视频和音频的检索。AllTheWeb 于 1999 年 5 月推出,先是由 FAST 运作;2003 年 4 月 Overture 收购了 AllTheWeb;后来 Yahoo! 买下了 Overture,现在的 AllTheWeb 由 Yahoo! 运作。

Ask Jeeves, http://www.askjeeves.com

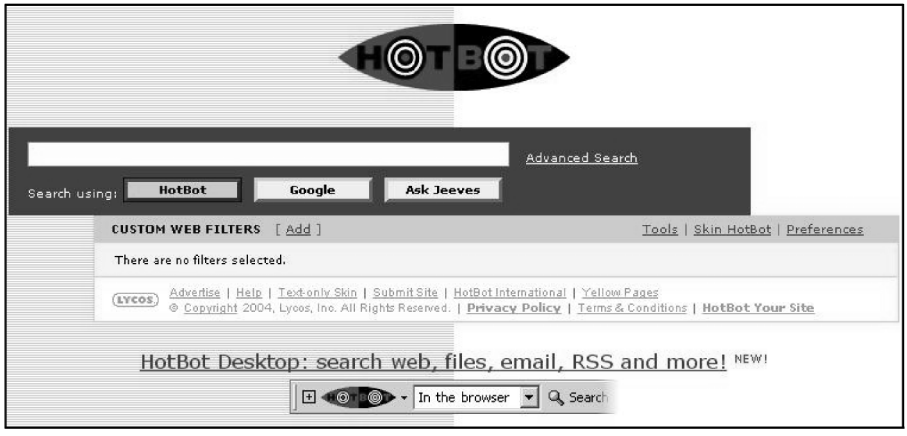

Ask Jeeves 最初获得名声是在 1998 和 1999 年。作为自然语言搜索引擎,能够让用户通过输入问题来得到查询结果,并且所得到的结果看起来好像是对的。

事实上,技术并不是 Ask Jeeves 运行很好的原因。在幕后,公司曾经指定 100 个编辑人员监视查询日志。然后这 100 个人上网查找与最常用查询词最相关的网页链接。目前,Ask Jeeves 仍然在使用人来参与结果的查找,但是现在编辑只有 10 个人左右。尽管如此,通过人的参与提供答案仍然是一个卖点,尤其对于那些新接触网络的人,他们会想使用 Ask Jeeves。对于通常的查询,人工选择的匹配结果让人感觉非常的相关。如果显示出来,这些结果出现在查询结果页面的最上端。除了人工参与外,Ask Jeeves 还利用基于搜集器的技术提供查询结果给用户。这些结果来自它所拥有的 Teoma 搜索引擎。

HotBot, http://www.hotbot.com

HotBot 提供便于访问三个搜索引擎(HotBot、Google、Ask Jeeves)的入口,但是不同于元搜索引擎[①],它不能将各搜索引擎的返回结果综合显示。

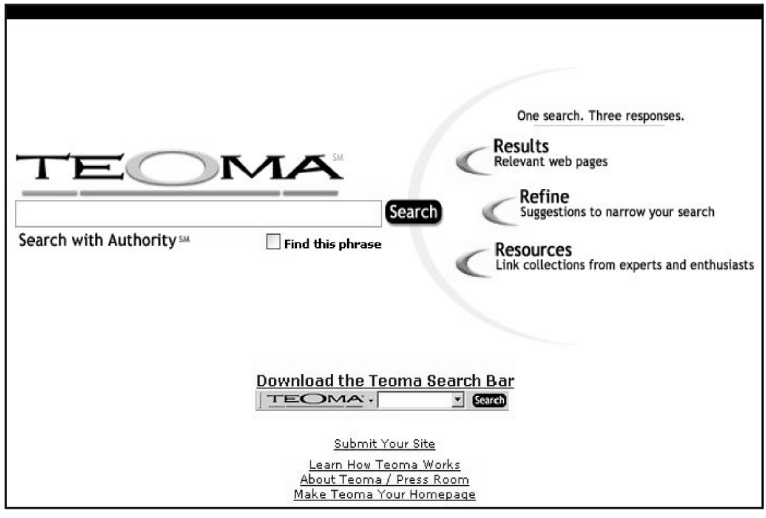
① 元搜索引擎又称集合型搜索引擎,是将多个独立的搜索引擎集合在一起形成的检索工具,即搜索引擎之搜索引擎。




HotBot 在 1996 年初次登场,因为其庞大的由 Inktomi 提供的基于搜集器的检索页面和质量成为搜索者喜欢的引擎。特别是它的不同寻常的颜色和接口,还为它赢得了有经验的网民的注意。

1999 年,HotBot 因为采用 Direct Hit 的 clickthrough 结果作为排序列表获得了恶名。Direct Hit 当年出现时是一个很热的搜索引擎。不幸的是,Direct Hit 的结果与同期登场的 Google 不能相比。HotBot 的声望开始下降。

Teoma, http:// www. teoma. com



Teoma 是基于搜集器的搜索引擎,2001 年 9 月被 Ask Jeeves 收购。它索引的网页比同样基于搜集器的竞争对手 Google 的少。然而对于通常的查询检索,索引网页多少并不会产生很大的分别,自从 2000 年 Teoma 出现,就因为它很好的网页相关性赢得了称赞。一些人喜欢 Teoma 的"相关检索"特性,您先输入一个简单词语搜索,然后,Teoma 会为您提供其他相关搜索词作为参考。"专家推荐资源"部分也是 Teoma 的一个特色,指导用户去访问不同主题的链接。

Lycos, http:// www. lycos com


Lycos 是一个资格最老的搜索引擎,1994 年开始提供服务。在 1999 年 4 月它停止了自己基于搜集器的结果,取而代之的是利用 LookSmart 人工整理的常用查询分类结果和其他基于搜集器的搜索引擎,如 Yahoo!、Inktomi 等搜集器提供的结果。那么用户为什么不直接使用其他的搜索引擎而还要使用 Lycos 呢? 你也许

是喜欢 Lycos 提供的一些特性。

在搜索框的下方 Lycos 会建议其他的与用户检索主题相关的查询词,也许正是用户想看和感觉更确切的查询词。在这之下,就是 Lycos 提供的与其他搜索引擎一样的既相关又广泛覆盖的结果。

Lycos 属于 Terra Lycos 公司,它是在 2000 年 10 月由 Lycos 合并了 Terra 网络公司后形成的公司。Terra Lycos 公司还有 HotBot 搜索引擎。

WiseNut, http://www.wisenut.com


与 Teoma 类似,WiseNut 是基于搜集器的搜索引擎,在 2001 年出现的时候吸引了大家的注意力。WiseNut 的结果也有很好的相关性,并且有很大的数据库,几乎像 Google、AllTheWeb 和 Inktomi 一样大。然而,WiseNut 的数据库更新很慢,查询结果经常是几个月前的内容。LookSmart 在 2002 年 4 月并购了 WiseNut。

Overture, http://www.overture.com

最初叫 GoTo,2001 年更名为 Overture。Overture 是一个非常流行的竞价排名搜索引擎,它提供广告给许多搜索引擎排在检索结果的上方。Overture 在 2003 年 3 月购买了 AllTheWeb,2003 年 4 月收购了 AltaVista。Yahoo 在 2003 年 10 月购买了 Overture。


Vivisimo, http://www.vivisimo.com


Vivisimo 于 2000 年 6 月由卡耐基-梅隆大学(CMU)推出,作为不同于基于搜

集器的元搜索引擎,有自己的独到之处。它把其他搜索引擎的返回结果利用自动聚类的办法来满足不同类型客户的需要。在搜索引擎上,任何人搜索同一个词的结果都是一样。这样明显不能满足访问者。科学家搜索“星球”,可能是希望了解星球的知识,但普通人可能是想找“星球大战”电影,但搜索引擎所给的都是一样的结果。如何满足这些不同类型的访问者,需要对搜索结果进行个性化处理。搜索结果排序从单一化到个性化,Vivisimo 已经迈出了一步。

Baidu(百度), http:// www. baidu. com


百度于 2000 年推出,是目前在中国最成功的一个商业搜索引擎,主要提供中文信息检索,并且为门户站点提供搜索结果服务。搜索范围涵盖了中国内地、香港、台湾、澳门、新加坡等华语地区以及北美、欧洲的部分站点。拥有的中文信息总量达到 1 亿 2000 万网页以上,并且还在以每天几十万页的速度快速增长。

Tianwang(天网), http:// e pku. edu. cn

于 1997 年 10 月开始提供服务,是中国最早的搜索引擎。它由北京大学网络与分布式系统实验室开发并维护运行,搜集了中国范围内大量的网络信息资源,尤其

是较全面地覆盖了中国教育科研网(CERNET)内的资源。天网目前索引的信息资源除已经超过3亿的网页外,还包括2000多万各种非网页类型的文件,是目前世界上最大的中文搜索引擎之一。在系统功能上,天网除提供通常的关键词和短语检索外,还有自动网页分类目录。本书所介绍的技术内容主要就是以天网为背景展开的。

# 上篇　Web搜索引擎基本原理和技术

上篇的主要目的是向读者介绍典型Web搜索引擎的基本工作原理,并通过一个实例具体展示该工作原理中各个环节的一种实现方法,以期使读者很快从技术上对搜索引擎系统的全貌有一个透彻的了解。

我们首先指出,所谓“搜索引擎”,说到底是一个计算机应用软件系统,或者说是一个网络应用软件系统。从网络用户的角度看,它根据用户提交的类自然语言查询词或者短语,返回一系列很可能与该查询相关的网页信息,供用户进一步判断和选取。为了有效地做到这一点,它大致上被分成三个功能模块,或者三个子系统,即网页搜集、预处理和查询服务。第二章详细分析了这三个部分的主要功能和其中需要关注的种种问题。应该指出,在实践中这三个部分是相对独立的,它们的工作形成了搜索引擎工作的三个阶段,通常分别都由人工启动。同时我们注意到,在早期的搜索引擎中,系统处理的网页数量少,预处理部分的工作比较简单,只是涉及汉语的分词(英文还没有这个问题)和建索引,因此也有将分词合并到网页搜集过程中,将建索引归到查询服务子系统中,从而整个系统看起来只有两个模块的安排[①]。

在介绍了基本原理后,第三、四、五章分别就上述三个阶段中的技术要求给出了一种实现方案。为了便于读者对搜索引擎技术在短时间内有一个全面实在的掌握,这三章的基本写作风格是提出问题,给出解决思路,然后是对应程序实现要点的注释和讲解。如果要掌握每一个细节(如需要开发一个搜索引擎),要求读者对C++程序设计语言比较熟悉。

---

① 1997年10月我们在CERNET上发布的天网1.0版本就是这种结构,每抓来一个网页就立即在内存分词,然后将得到的结果存入数据库中,供建索引程序直接使用。

然而,从了解现代搜索引擎技术原理的需求来看,中篇和下篇并不很依赖这三章的内容,因此对 C++ 程序设计语言不熟的读者可以跳过这三章,直接阅读中篇和下篇,或者不一定要一句句探究那些程序段落的逻辑。程序代码可以在(Tse 2004)下载。

对于希望动手构建搜索引擎的读者来说,掌握了这一篇的内容,直接用我们提供的实例代码,应该能够很快(如一周)构建出一个可用的小型通用搜索引擎。同时我们指出,一个实用的大规模搜索引擎还有许多其他重要问题要解决,集中在效率和质量两个方面,我们安排在中篇讨论。