

1 网络信息资源概述

1.1 网络信息资源概述

1.1.1 网络信息资源的概念

网络信息资源是指以电子资源数据的形式将文字、图像、声音、动画等多种形式的信息存放在光、磁等非印刷质的介质中,并通过网络通信、计算机或终端等方式再现出来的信息资源的总和^①。Internet 是目前世界上最大的信息宝库,它已成为全球范围内传播和交流科研教育信息、经济信息、商业和社会信息的最主要的渠道。若要有效地获取网络资源,首先需要了解网络信息资源的特点和种类。

1.1.2 网络信息资源的特点

1) 信息量大,信息层次多,检索空间拓宽

中国互联网络信息中心 CNNIC^②为了深入、广泛地开发利用我国的互联网络信息资源,对我国的互联网络信息资源有一个全面、深入的了解和掌握,开展了中国互联网络信息资源数量调查。调查采用抽样调查和计算机网上自动搜索技术,调查内容包括我国域名总数、网站总数、网页总数、在线数据库总数等四大方面的数据,并涉及各指标的分类、地域分布等。据 CNNIC 中国互联网络发展状况统计报告(第十五次)数据,截止到 2004 年底,中国内地上网用户总数为 9 400 万,上网计算机数 4 160 万;CN 下注册的域名数 43 万、网站数 69 万,网络国际出口带宽总数达到 74 429 M。我国大陆的 IPv4 地址数达到了 59 945 728 个。在网络应用上,人们的上网用途进一步多元化发展,不论是在关注的信息内容还是在使用功能上,涉及范围都更深更广,几乎覆盖了整个 Internet 网络,为访问和获取广泛分布在世界各地的、成千上万台服务器和主机上的各学科、各领域、各地域、各种语言的信息资源提供可能。

2) 信息传播速度快,变化频繁,检索方式交互式作业

伴随网络化和数字化而产生的大数量、多类型、多媒体、非规范、跨时间、跨地理、跨行业、跨语种的网络信息资源,是对原来以相对集中和规范为标志的传统数据库资源的突破性发展。Internet 信息源分散无序,变化频繁,更迭消亡难以预

^① 华薇娜. 网络学术信息资源检索与利用. 北京: 国防工业出版社, 2002

^② <http://www.cnnic.net.cn>

测,其多功能和简单易用的友好用户界面要求应运而生。几乎所有网络信息检索工具都具有交互式作业的特点,都能够从用户命令中获取指令,即时响应用户的要求,执行相应操作,并且有良好的信息反馈功能。

3) 自由发布,直接交流,新的检索园地

Internet 提供了自由发表个人见解的广阔空间和获取非出版信息的丰富机会,反映了许多研究成果背后的原始数据或第一手资料,并扩大了人际交流的范围,提供了更多的直接交流机会。Internet 提供了在正式的出版和发表渠道之外的发表个人见解的空间,是检索新观点、不成熟或不正确观点,未成定论的理论、假说、概念等的一个新园地。

1.1.3 网络信息资源的种类

网络信息资源包罗万象,可以由多种标准划分,按其所采用的网络协议标准可作如下划分^①。

1) WWW(World Wide Web)网络信息资源

是 20 世纪 90 年代瑞士欧洲粒子物理研究中心工作人员为了利用超文本方式传递信息而发明的,它能方便、迅速地发布和浏览分布于网络各处的文字、图像、声音和多媒体超文本信息,发展迅速,已成为目前 Internet 上的主流资源。

2) FTP 网络信息资源

FTP(File Transfer Protocol)文件传输协议是 Internet 上历史最悠久和应用比较广泛的网络工具。用户通过协议连接到 Internet 上运行 FTP 协议的任一服务器,可访问大量的程序和文件,并将其下载到自己的计算机上,也可把自己的文件传送给远程的主机,相当于在 Internet 上的两个主机间拷贝文档。Archie 是第一个自动索引互联网上匿名 FTP 网站文件的程序,可搜索 FTP 文件名列表,但用户必须输入精确的文件名,然后 Archie 会告诉用户哪一个 FTP 地址可以下载该文件。

3) Telnet 网络信息资源

用户利用 Internet 上的远程登录协议 Telnet,可将自己的计算机作为 Internet 上的某一个主机的远程终端与该主机相连,从而使用该主机的硬件、软件和信息资源。

4) Gopher 网络信息资源

Gopher 是一种菜单式的网络服务,各个 Gopher 服务器之间相互连接,用户通过一级级菜单的指引来浏览和选择所需资源。

5) 用户服务组网络信息资源

用户通讯和服务组是 Internet 上受欢迎的信息交流和讨论形式,它们都是由一组对某一特定主题有共同兴趣的网络用户组成的电子论坛,包括新闻组

^① 华薇娜.网络学术信息资源检索与利用北京:国防工业出版社,2002.12

(Usenet Newsgroup)、电子邮件群(Listserv)、邮件列表(Mailing List)和专题讨论组(Discussion Group)等。

1.2 网络信息资源的组织

信息技术的发展,为网络知识信息的组织提供了技术支持,使人们能够突破传统的信息结构的线性组织模式,进入超文本、超媒体信息组织模式的新时代。

1.2.1 网络信息组织原理

网络信息组织就是网络信息有序化的过程,它是在宏观和总体上处理大量的信息,改变它们的结构和存在形式,使其具有一种人为的、科学的、合理的结构。在这个结构中,个别的用户能根据自己特定的信息需求,通过网络检索工具所提供的检索手段,获取自己所需的信息。信息有序化的基本原理是:设有若干信息,将这些信息按照某种特征值排列成一个序列,而且需要使用信息的用户能将自己的信息需要转换成相应的特征值,并在排列后的信息序列中找到自己所需的信息。

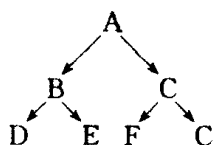
信息有序化的结构特点可分为下列四种基本情况^①:

1) 线性结构

对于信息集中的信息单元,根据某种特征排序,每个信息单元只能有一个前驱信息,且也只能有一个后继信息, $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \cdots$ 如字典、词典、按文摘号排列的文摘等。

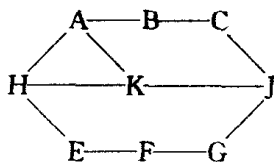
2) 树状结构

对于每个单元只能有一个前驱信息,但可以存在一个或多个后继信息。如雅虎的等级体系分类目录。



3) 网状结构

对于每个信息单元都可有一个或多个前驱信息和后继信息。如 Internet 中通过超链接显示的信息单元之间关系的网络检索工具。



^① 孙建军. 网络信息资源搜集与利用. 南京: 东南大学出版社, 2000

4) 复合结构

上述三种结构是任何检索工具最基本的内部结构。任何复杂的检索工具都可以由这三种基本结构复合而成。

1.2.2 网络信息组织的分类方法

1) 超文本信息组织

超文本信息组织方式是网络信息资源组织的一种主要方式。它以节点和链接在网上相关信息间建立关联,从而将众多的网络文本信息组织成某种网状结构,用户可从任一节点开始,随机、自然地浏览查找自己所需信息,并将之从网上下载。

2) 网络信息检索系统的信息组织

Internet上存在着巨量的独立文本文件的超文本链接体,如何将它们有效地进行组织,有效地进行深层开发,有效地逼近用户的真正需求,是网络信息资源组织的又一重要研究课题。目前,大多数网络信息资源的检索系统都包含两大模块,一是网络信息索引模块,一是网络信息查询模块。索引模块作用是建立网上信息资源数据库,查询模块作用是利用索引数据库,获取所需信息。

1.2.3 多媒体信息的组织

传统信息检索大都以结构化文本内容来组织索引工具^①,而网络信息除文本信息外,有图表、图形、图像、声音、影视等信息。为此,多媒体信息内容特征抽取更加复杂化。基于文本描述的多媒体信息的组织必须把它们转换成文本信息或者添加文本说明,而基于内容的多媒体信息组织则是以相似匹配来代替精确匹配^②,因为相同内容的多媒体信息的表现形式可能不同,并在时效、自动化实现、成本等方面具有实用价值的要求。

多媒体信息组织的技术核心是基于内容的图像检索技术,即从图像库中查找含有特定物质的图像(也包括从连续的视频图像中检索含有特定物体的视频片段)。因此,基于内容的图像检索系统,其处理对象可以分为静止图像检索和活动视频检索。前者用于检索的特征主要有颜色、纹理、草图、形状等,后者一般采用脱胎于序列图像分析的技术^③。

目前,网上大部分多媒体搜索引擎都属于基于文本描述的,如 Lycos、Google、新浪等,它们都支持图像、声音、影像等媒体类型,且都支持布尔检索及加减检索,但检索范围基本上限制在 WWW 范围内。基于内容描述的多媒体搜索引擎相对少见,如 IBM 公司的 QBIC,上海交大的音乐数据库检索系统。

① 曾民族. 网络信息检索现状和性能评价. 情报学报, 1997(2)

② 邓成剑. WWW 多媒体搜索引擎研究. 图书情报知识, 2003(1)

③ 李爱红. 网络搜索引擎的比较研究. 中国信息导报, 1999(1)

1.2.4 信息组织的方法

1) 主题树方式

主题树方式通过人工发现信息,区别信息并进行分类,人工建立结构化的 Internet 网址主题类目和子类目,子类目下可进一步细分,最底层类目内按字顺或其他标识进行排序,用户可利用这个相当详尽的等级分类目录体系结构对网上特定主题信息进行浏览查询。Yahoo、新浪就是这种组织方式的典型代表。

2) 计算机索引数据库方式

计算机索引数据库方式与主题树方式最大的不同是它的标引,数据库非人工建立,而是由被称为 Spider 或 Robot 的计算机程序创建的。Spider 根据网络协议在网上漫游,不断发现各类新的网址及网页信息,经抽取、排序与归并,建立网络索引数据库,数据库按一定方式、结构存储满足特定处理系统需要的相关信息,通常为网络信息的地址、相关信息的描述及供计算机识别的字段标识符。当用户输入检索词后,检索软件将在数据库中找出相匹配的记录,为用户提供检索服务。AltaVista、百度就是这种组织方式的典型代表

1.2.5 数字图书馆的信息组织

基于数字化资源的图书馆,往往嵌入到传统图书馆系统或上层机构信息系统中,将跨时空检索和传递特定数字化资源作为主要任务。数字图书馆的功能形式、技术方式和操作管理机制往往取决于资源库的内容格式、元数据格式、知识组织体系和特定软硬件平台,利用专用代理模块来实现 Web 用户的检索浏览。

1.3 网络信息资源检索技术

网络检索工具正分别向综合型和专业型发展,导致检索语言也出现了向简单和复杂两极发展的趋势。对于综合型检索工具,由于用户对所需信息的深度要求不高,用户界面的易用性将成为评价这类检索工具的极其重要的指标。而对于专业型的检索工具,由于用户对所需信息的深度、内容的精确性和相关性要求较高,使用一定的信息资源检索技术,可提高搜索的精确度。

1.3.1 布尔逻辑检索

布尔逻辑检索是检索系统中应用最为广泛的检索技术,是最简单、最基本的匹配模式,几乎所有的网络信息资源检索系统都支持布尔逻辑检索。

1) 逻辑与,网络检索符号一般为“* ”或“ and”

它反映的是词与词之间的交叉组配。如“ A * B”,表示检索结果中同时会有 A

2) 逻辑或,网络检索符号一般为‘+’或‘or’。

它反映的是词与词之间的并列组配。如“ $A+B$ ”表示检索结果中含有 A 检索词或含有 B 检索词的信息都会命中。逻辑或可扩大检索范围,提高查全率。

3) 逻辑非,网络检索符号一般为‘-’或‘not’。

它用于从原来的检索范围中排除不需要的内容。如“ $A-B$ ”表示检索结果中必须含有 A 检索词但不得含有 B 检索词的所有信息。

基本布尔逻辑运算的执行顺序: not, and, or, 但可根据需要用括号重新规定顺序。在布尔逻辑检索中,另有布尔逻辑提炼符()。括号的作用是使括在其中的操作符先起作用,或者说先被计算机解释,逻辑非、逻辑与、逻辑或紧跟其后。

1.3.2 加权检索

加权检索并不是所有系统都能提供的检索技术。它通过判定检索词或字符串在满足检索逻辑后对文献命中与否的影响程度,给予一定的数值(权重)加以区别,同时给出检索命中界限进行限制。界限值也称为阈值,其作用是限制检索结果的输出量。进行加权检索时,利用检索词查找数据库,每条命中记录将其所包含的检索词根据检索时所设定的权值,分别计算命中记录的权值之和,当已检出记录的权数之和超过或达到阈值时,才为‘命中’信息,否则为‘落选’。

1) 词加权检索,最常见的加权检索方法

在检索式的构造过程中,检索者根据对用户检索需求的理解,为需求选定检索词,同时每一个检索词(概念)给定一个数值(权重)表示其针对本次检索的重要程度。检索时先判断检索词在文献记录中是否存在,对存在检索词的记录计算其所包含的检索词权值总和,通过与预先给定的阈值比较,权值之和达到或超过阈值的记录,视为命中记录,命中结果的输出按权值总和从大到小排列输出。

2) 词频加权检索

根据检索词在数据库中出现的频次来决定该检索词的权值,而不是由检索者指定检索词的权值。该法消除了人工干预因素,但这种加权检索方式必须建立在全文或文摘型数据库基础之上,否则词频加权将没有意义。

(1) 绝对词频加权指检索时累计检索词在记录中出现的次数(权值),检索的记录权值之和由记录包含的所有检索词在记录中出现的次数总和决定。

(2) 词频与平均词频相结合确定权值。使用某检索词在检索集中出现的频率与含该词的文献量之比,可得到检索集中该词在每篇文献中的平均出现次数,称为该检索词的平均词频,它可以用于系数完善绝对词频的计算,即用绝对词频与平均词频乘积的总和作为该词的权值。

(3) 逆文献频率确定权值。仅仅依赖词频或对词频增加一个系数,不能对在每一篇文献内都出现的词和只在一些文献中出现的词进行区别,可以将平均词频

1.3.3 截词检索

截词检索在信息系统中,尤其是在联机检索、网络检索系统中应用十分广泛,其主要是利用检索词的词干或不完整的词型进行检索。

1) 精确匹配(对称不截断)

指将检索词作为一个完整的词来检索,记录中具有同一属性的规定字段的词必须与检索词完全相同才算命中。

2) 任意截断检索

指检索词串与被检索词实现部分一致的匹配,被截取的部分用“*”等来表示,截断形式有前截断(后方一致)、后截断(前方一致)、前后截断(任意一致)等三种情况。

3) 有限截断检索

指检索词串与被检索词只能在指定的位置可以不一致的检索,每一个被屏蔽的字符都用一个“?”来替代(汉字应用两个“?”来替代),表示在其前面和(或)后面最多可以有多少字符。

1.3.4 位置检索

位置检索是全文检索系统中主要的检索技术,是规定了检索词在原始文献中相对位置的限定性检索。它大致包括四种级别的检索,即记录级检索,限定检索词在数据库的同一记录中;字段级检索,限定检索词在数据库记录的字段范围内;子字段或自然句级检索,限制检索词在同一子字段或自然句中;词位置检索,限定检索词的相互位置满足某些条件。

1.3.5 全文检索

与其他检索技术相比,全文检索技术的新颖之处在于可以使用原文中任何一个有实际意义的词作为检索入口,而且得到的检索结果是该文献而不是文献线索。全文检索技术中的“全文”,表现在它的数据源是全文,检索对象是全文,采用的检索技术是全文,提供的检索结果也是全文信息。

全文检索系统的内涵可概括为具有全文数据库,系统除布尔逻辑检索外还具有文本检索功能(位置检索、字符串检索、截词检索等),允许用户的自然语言检索,直接获得原文的有关章、节、段、句等信息。

1.3.6 超文本检索

超文本检索技术是一种相对新型的检索技术,有着传统检索技术所没有的优势,凭借其独特的直观性、易用性、方便性,必将成为检索系统的主流。

1) 超文本概述

超文本是对原有单向线性工作、单值媒体或单值排列的一种扩充、一种开拓,实质上是对‘文本’的一种扩充^①。它既是一种信息单元的组织和检索形式,也是一种利用计算机软件获取信息的技术。所谓文本信息包括文字信息与数字信息等,而超文本信息则包括了文本信息、图像信息与声音信息等,也即一切可用的现代电脑存储的信息。超文本强调了信息对象之间的关系,方便了人机交互,增加了用户对系统的了解,符合人们联想式的泛读与思维习惯。

2) 超文本检索的特点

超文本检索技术利用计算机技术、通信技术和人工智能的知识表达技术,将包含文字、图像、图形、声音、动画等多种形式的电子信息按其相互之间的关联和可能出现的连续性,进行非线性编排。用户可按照自己的意愿来组织这个相互关联的网络信息系统,同时只要两个信息单元之间存在着直接或间接的关联,就可以从其中之一出发顺着关系链到另一个信息单元。

3) 超文本检索系统的分类

对超文本检索系统的分类还没有一致的标准,根据 Carison 的观点,超文本检索系统可以分为基于浏览的检索系统、基于提问的检索系统以及智能检索系统。

(1) 基于浏览的(Navigation - based)检索系统 这是最一般的超文本检索系统模式,表达了现有超文本系统信息存储与检索的特点。基于浏览的超文本系统的最主要检索方式,即浏览检索方式,指通过跟踪信息节点间的链路,在网络中移动获得所需信息。通过对系统包含内容的浏览,可以了解信息单元本身的内容、数据库的组织方式、信息之间的相关程度、信息的查询路径等,用户可以据此调整检索策略甚至检索目标。

(2) 基于提问的检索系统(Query - based) 可以看作一种双层结构模型,其中辅助数据集合和语义链构成检索系统的概念层,信息集合与结构链构成检索系统的文献层,连接链则贯穿于两个层面之间,具有实现两个层次的切换沟通作用。

(3) 智能型检索系统 它实质上是基于导航浏览和提问系统的逻辑功能的延伸。

1.3.7 智能化检索

智能化检索是集中融合传统检索技术和人工智能技术建立的新一代智能化检索,能以自然语言方式接受检索课题,并像人工那样进行课题的分析与设计,全过程自动地完成课题的检索。从理论上讲,一个智能型情报检索系统由知识获取及加工系统、情报资料获取及加工系统、知识库(智能化检索系统的核心部分之一,主要存放来自知识获取及加工系统、情报资料获取与加工系统的分析方法集、检索策略集、情报资料集即各种数据库等)、知识库管理系统、搜索机(智能化检索系统的

^① 孙建军. 网络信息资源搜集与利用. 南京: 东南大学出版社, 2000

又一核心部分，负责接受用户提问，借助知识库完成相应的检索，并以适当的方式将检索结果提供给用户。输入／输出接口几部分组成。

1.4 化学化工网络信息资源概述

1.4.1 化学化工网络信息资源

网络化学化工信息资源分布在世界各地的计算机上，以文件、数据库、公告板、自录文档和超文本档案等各种形式存在。具体而言，网络化学化工信息资源存在于几种服务器，如电子邮件服务器、网络新闻服务器、FTP 服务器、Gopher 服务器、Web 服务器上。目前，Internet 上绝大多数化学化工信息都是以 WWW 方式组织、存储和提供服务的，这些资源主要有化学信息数据库、化学专利、网络期刊、网络会议、网络教育资源、化学软件资源、讨论组和邮件列表等。

1.4.2 化学化工网络信息资源的利用

1) 化学化工网络信息资源的访问方法

(1) 人的资源的访问：我们可以通过 E-mail、网络论坛 Usenet、Maillists、Iphone 参加 Internet 的各种通信组以及 Usenet 专题讨论组。

(2) 计算机系统资源的访问：一般地说，要求使用计算机系统的 Internet 用户，都可以通过远程登录到达某台目标计算机，只要这台计算机允许登录者使用并建立了管理者的账号。

(3) 计算机上的信息资源的访问：首先人们对 Internet 上各种信息资源做出标识，用统一资源定位即 URL (Uniform Resource Locator) 实现各个资源的统一编号。这样人们可以在浏览器的 URL 地址栏中，直接键入 URL 地址即可访问到所需的资源。

2) 化学化工网络信息资源的查询方法

目前获取 Internet 化学化工信息资源可以利用的工具主要分为两类：一类是 Internet 通用资源搜索引擎，如 Yahoo、AltaVista 等；另一类就是 Internet 化学化工专业站点，即宏站点，它把网上许多有关化学化工的信息加以组织，形成一个专业型的导航站点，如 ChIN、Chemweb 等。

2 搜索引擎与网络信息资源导航

搜索引擎对收录的网页编制的摘要和评述大都很简略,缺乏很好的专业参考价值,这使得根据它提高查询效率,筛选出所需要的数据库信息是一件费力的工作。面向专业的化学化工宏站点一般都提供了详细的主题目录,这些主题目录一般在大类目下分成若干小类目,类目之间按照等级系统排列,然后将人工或巡视软件搜集或选择的网页主题连接起来,用户可以方便地通过主题目录的指引,逐层浏览,直到找出有关的信息为止。

Internet 搜索引擎和资源导航系统都是对 Internet 上的信息资源及其 URL 建立索引,并通过基于 Web 的界面向用户提供信息资源的查询服务。

2.1 搜索引擎概述

在已开发的众多网络信息查询工具中,搜索引擎后来居上,日益显示其技术优越性和可用性,发展势头似乎压倒其他网络信息查询工具,成为目前 Internet 网络信息检索方式的主流。

2.1.1 搜索引擎的概念

搜索引擎(Search Engine)泛指在数据库系统中查找信息的工具,是用来对网络信息资源进行管理和检索的一系列软件。我们通常所说的搜索引擎大部分是指以 Web 站点存在,提供网址、网页、文章搜索及综合服务。搜索引擎是 Internet 上的一个网站,其主要任务是在 Internet 上主动搜索一定范围内的服务器信息,并将其自动索引(一般将各站点的服务器按主题内容组织成等级结构),其索引内容存储于可供查询的大型数据库中,用户可以依照这个目录逐层深入,直至找到所需信息;用户也可在它们的各种程序中输入关键词查询,搜索引擎就会在自己的数据库中告诉用户包含该关键词的所有网址,并提供通向该网站的链接。

随着网络技术的飞速发展,Internet 数据库积累的信息和主题越来越多,怎样快速、高效、经济地查询某个主题的所有信息就成了十分热门的话题,搜索引擎正成为 IT 行业兵家必争之地。

2.1.2 搜索引擎的工作原理与过程

搜索引擎的工作原理可以阐述为自动收集网页的 Spider 系统程序通过扫描定范围内每一篇文章或网页中的每一个词(从互联网上抓取网页),建立以词为

单位的倒排文件(建立索引数据库),检索程序根据检索词在每篇文章或网页中出现的频率,对包含这些检索词的文章或网页进行排序,最后输出检索结果。换句话说,就是根据用户的查询请求,将含有特定单词的文章或网页列出来。

一个完整的 Internet 搜索引擎主要由机器人(Robot, Spider, Wanderer, Crawler)、索引库(Index 或 Catalog)和用于检索索引库的检索软件(Search Engine Software)三部分组成^①。搜索引擎的工作过程为机器人自动在 Web 上按某种策略进行远程数据的搜索和获取,并生成本地索引,由此产生一个数据库。索引数据库内保存机器人访问过的 Web 网页的拷贝。用户按照 HTTP 协议,向 Web 服务器发出申请,检索软件就根据用户的查询要求在索引库中筛选满足条件的网页记录,并依照排序得分依次给出查询结果。

2.1.3 搜索引擎的分类

(1) 按信息的组织方式可分为范畴层次搜索引擎(List-based Search Engines)(或称目录式分类搜索引擎)、词语搜索引擎(Word-based Search Engines)两大类。而范畴层次搜索引擎通常带有关键词索引,词语搜索引擎通常带有范畴层次索引,以弥补分类法和主题法各自的不足。

(2) 按覆盖范围可分为通用搜索引擎和专业性搜索引擎。

(3) 按引擎功能可分为常规搜索引擎(或称独立搜索引擎)和多元搜索引擎(或称宏搜索引擎)两大类。常规搜索引擎有单一的功能,搜索时通常只检索自己的数据库,并根据数据库内容反馈出相应的查询信息或链接站点。多元搜索引擎是多个单一引擎的组合,对同一个查询可同时在多个搜索引擎上查询,或组合指定的搜索引擎查询,或在单一搜索引擎上查询,但可综合输出查询结果,也可分别输出每个搜索引擎的查询结果。

2.1.4 搜索引擎的评价指标、存在问题及发展对策

PageRank 是 Internet 上网页重要性的评估标准,它是基于网页中相互链接关系的分析得出的,哪个网页被链接得多,哪个网页的 PageRank 值就高。但搜索引擎的评估标准应达到查全率和查准率的和谐统一,因此搜索引擎的技术专家提出通过搜索引擎的相关性定律、人气质量定律、自信心定律来评价搜索引擎,将搜索引擎的具体评价指标分为数据库规模和内容、索引方法、检索功能、检索结果、用户界面和其他等六个指标。相关性无疑是评价一个搜索引擎性能最重要的指标。动辄成千上万的结果,很少有人能连续地翻看几页。如果第一页没有中意结果,用户就可能换用另外的搜索引擎。所以,搜索引擎都会针对输入的关键词,挑选出最相关、最重要、最可能满足用户需求的网页排在前面。

^① 李梦龙等. Internet 与化学信息导论北京:清华大学出版社 2001

目前,较好的搜索引擎都是既支持目录式分类结构,又支持全文检索,同时区分搜索结果的相关性。但由于搜索引擎尤其是中文搜索引擎缺乏统一的协调和合作,查询方式和服务较为单一,逻辑检索有限,分类范畴的搜索引擎过多,易造成漏检,而对于关键词提问的申请,多为盲目匹配,筛选较费时,索引数据库的更新不及时,结果显示简单等。

搜索引擎的历史虽然不长,但发展趋势必将从综合性向专业化发展,从单一性向多元化方向发展,以英文为主向多语言方向发展。没有哪两个搜索引擎是完全相同的,每一种都有自己的检索特色,都有自己的索引,以不同的方式在网上搜寻网址。出现不同的检索结果丝毫不足为奇,应从不同的搜索引擎的检索结果中综合出最为符合要求的答案。

2.1.5 搜索引擎的常用检索命令

大部分搜索引擎都为用户提供了各种各样的检索命令和菜单以帮助用户提高他们的检索命中率,下面的检索命令一般的搜索引擎都支持它们,只不过具体每一种搜索引擎所使用的检索命令多少有所不同,或默认选项不同,或有些检索命令前后有空格。

- “+”检索命令 希望搜索引擎返回的结果必须包含所键入的关键词,可以在关键词前使用‘+’命令
- “-”检索命令 希望搜索引擎返回的结果不能包含一些关键词,可以在关键词前使用‘-’命令
- 英文双引号或括号检索命令 希望搜索引擎返回的结果必须精确包含某些词组,可以使用双引号或括号将词组引起来
- “title :或 t :”标题检索命令 希望搜索引擎返回网页的标题必须包含所要检索的关键字,就可以使用这些标题检索命令
- “site 或 host”域名检索命令 希望搜索引擎检索某一具有指定域名的网站或某一域下的网站,或将该网站或某一域下的网站的网页从结果集中排除

2.2 综合性搜索引擎

随着互联网的迅速发展,检索所有新出现的网页变得越来越困难,CNNIC^①最新调查报告显示,搜索引擎是用户得知新网站的最主要途径,网站访问量82.2%以上来源于搜索引擎。搜索引擎作为查询网络信息资源的一个工具,预先对网页中的每一个词(即关键词)进行索引,建立索引数据库的全文搜索引擎。当用户查找某个关键词的时候,所有在页面内容中包含了该关键词的网页都将作为

^① <http://www.cnnic.net.cn>, 2005-03-08

搜索结果被搜出来。在经过复杂的算法进行排序后，这些结果将按照与搜索关键词的相关度高低，依次排列。目前，搜索引擎已普遍使用超链分析技术，除了分析索引网页本身的内容，还分析索引所有指向该网页链接的 URL、Anchor Text、甚至链接周围的文字。研究表明，搜索引擎是目前最重要、效果最明显的网站推广方式，也是最为成熟的一种网络营销方法。

2.2.1 中文综合性搜索引擎

人们常常把搜索引擎称之为上网第一站、网络之门，随着我国互联网络的发展，上网用户越来越多，对中国人的网络之门的争夺战也越演越烈。世界著名搜索引擎 Yahoo、AltaVista 相继推出中文版搜索引擎，Netscape 不久前也声称推出中文版的 Netcenter，中国人的网络之门正逐渐被外国人占据。但应该看到最了解中国人需求的还是我们中国人自己，中国人应该利用自己的文化优势、资源优势牢牢地占住自己的网络之门。从 20 世纪 80 年代中期起，随着中文网络信息规模的日益庞大，中文搜索引擎如雨后春笋般发展起来。中文的搜索引擎包括中国大陆、港澳台地区、亚洲其他地区的网络制作商的成果及相当数量的海外搜索引擎的中文版。

中文搜索引擎指南网站^①介绍了七种主要中文搜索引擎和三种其他中文搜索引擎(如表 2-1)。本书还给出了国内用户经常使用的其他中文搜索引擎。

表 2-1 中文搜索引擎指南网站介绍的中文搜索引擎

主要中文搜索引擎	Google(http://www.google.com/intl/zh-CN/) 百度(http://www.baidu.com) 一搜 http://www.yisou.com/ 中国搜索 http://www.zhongsou.com/ 搜狗搜索 http://www.sogou.com/ 新浪网搜索引擎 http://cha.sina.com.cn/ 网易搜索引擎 http://so.163.com/
其他中文搜索引擎	TOM 搜索引擎 http://i.tom.com/ 21CN 搜索引擎 http://search.21cn.com/ QQ 搜索引擎 http://search.qq.com/

1) Google(<http://www.google.com/intl/zh-CN/>)

Google 的使命就是要提供网上最好的查询服务,促进全球信息的交流,开发出世界上最大的搜索引擎,提供最便捷的网上信息查询方法,每天需要提供 2 亿次查询服务。现在,Google 通过对 40 多亿网页进行整理,可为世界各地的用户提供适合需要的搜索结果,而且搜索时间通常不到半秒。Google 秉持着开发“完美的

^① <http://www.sowang.com/>, 2005-04-01

搜索引擎’的信念,就如公司创始人之一 Larry Page 所定义的那样,可以‘确解用户之意,切返用户之需”。为了实现这一目标,Google 不受现有模型的限制,坚持不懈地追求创新,开发了自己的服务基础结构和具有突破性的 PageRank 技术,使得搜索方式发生了根本性变化。Google 利用相互链接的 PC 机来快速查找每个搜索的答案,这一创新技术成功地缩短了响应时间,提高了可扩展性,降低了成本。传统的搜索引擎在很大程度上取决于文字在网页上出现的频率,Google 搜索技术所依托的软件可以同时进行一系列的运算,且只需片刻即可完成所有运算,而且使用 PageRank 技术检查整个网络链接结构,并确定哪些网页重要性最高,然后进行超文本匹配分析,以确定哪些网页与正在执行的特定搜索相关。在综合考虑整体重要性以及与特定查询的相关性之后,Google 可以将最相关、最可靠的搜索结果放在首位。

Google 默认的连接符是 AND,如果要从几个关键字中指定任何一个,可以使用电脑编程里的管道符号‘|’来表示 OR 的功能,如果要在搜索结果中不包括某个关键字,使用‘-’(减号/短横线)来实现。

Google 已经可以支持 13 种非 HTML 文件的搜索,除了 PDF 文档,还可以搜索 Microsoft Office (doc, ppt, xls, rtf)、Shockwave Flash (swf)、PostScript (ps) 和其他类型文档。新的文档类型只要与用户的搜索相关,就会自动显示在搜索结果中。Google 在访问网站时,会将看过的网页复制一份网页快照,以备在找不到原来的网页时使用,单击‘网页快照’时,将看到 Google 将该网页编入索引时的页面。Google 依据这些快照来分析网页是否符合需求,在显示网页快照时,其顶部有一个标题,用来提醒这不是实际的网页。符合搜索条件的词语在网页快照上突出显示,便于快速查找所需的相关资料。尚未编入索引的网站没有‘网页快照’,另外,如果网站的所有者要求 Google 删除其快照,这些网站也没有‘网页快照’。

Google 于 2004 年 11 月推出了学术性文献搜索新服务^①,帮助用户搜索学术性文献,如搜索技术报告、论文、摘引等。Google Scholar 还能够在 Google 的索引中搜索一个特定的主题,覆盖范围包括从医学、物理学到经济学、计算机科学在内的很多领域。Google 于 2004 年 12 月宣布已和美国哈佛大学、斯坦福大学、密歇根大学及英国牛津大学的图书馆,以及纽约公共图书馆等五家图书馆签约协作,将这些图书馆里的数千万册藏书和文档转成数码文件,供网友免费检索,正式向 Google 建构全球仿真图书馆的构想迈出第一步。

2) 百度(<http://www.baidu.com>)

百度搜索引擎(如图 2-1)使用了高性能的‘网络蜘蛛’程序,具有全球独有的‘超链分析’技术,亚秒级的迅捷速度,庞大的服务器群,接受来自全球各个国家的中文搜索请求,是目前世界上最大的中文搜索引擎,总量超过 3 亿页以上,并且还

^① <http://scholar.google.com>

在保持快速增长。中国所有提供搜索引擎的门户网站中，超过 80%以上都由百度提供搜索引擎技术支持。它根据中文网页的特点，自动在整个互联网上收集资源，同时还能够消除一些已经不能够再用的网页。由于后台应用了高效的信息索引算法，大大提高了检索时的响应速度和承受大访问量时的稳定性。可以承受所有高峰的负荷而不会在性能方面有任何偏差。

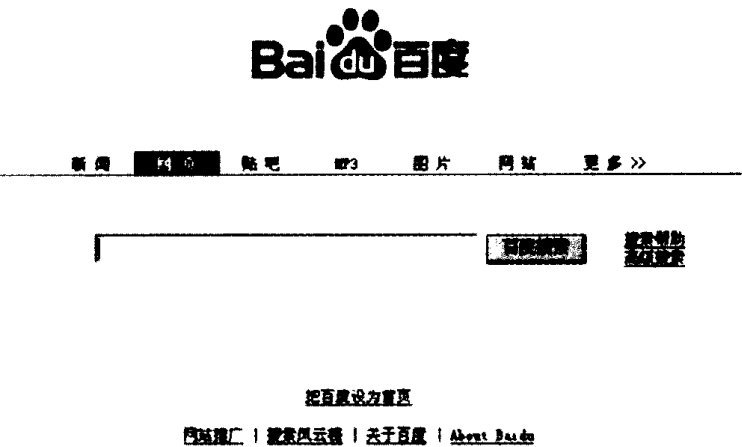


图 2-1

百度具有自己独特的语法结构，支持多种高级检索语法，使用户查询效率更高、结果更准。现已支持 “+”(AND)、“-”(NOT)、“|”(OR)、“site:”、“link:”等搜索语法，还将继续增加其他高效的搜索语法。类似于“and”的功能在使用时，可在两个关键词之间加空格，查询时百度会在多个以空格隔开的词语之间自动添加“+”。百度支持“-”功能，用于有目的地删除某些无关网页，但减号之前必须留一空格。百度的“or”功能可使用“A | B”，来搜索。百度的高级搜索中分别可按搜索结果“包含以下全部的关键词”“包含以下的完整关键词”“包含以下任意一个关键词”“不包括以下关键词”来显示，同时可按时间、地区、语言、文档格式、关键词位置等进行限定搜索。

用户提供查询后，系统根据用户的检索词和查询选项返回查询结果，并支持二次检索。检索结果的智能相关度算法，采用了基于内容和基于超链分析相结合的方法进行相关度评价，能够客观分析网页所包含的信息，从而最大限度保证了检索结果相关性。检索结果输出支持内容类聚、网站类聚、内容类聚+网站类聚等多种方式，支持用户选择时间范围，根据网页级别排列出优先顺序，提高用户检索效率。同时检索结果能标示丰富的网页属性(如标题、网址、时间、大小、编码、摘要等)，并突出用户的查询串，便于用户判断是否阅读原文，每一项结果基本上列出标题、网页(站)介绍、网址及附带的全新功能等相关信息。

3) 一搜 <http://www.yisou.com/>)

一搜(如图 2-2)是雅虎公司基于全球领先的 YST(Yahoo Search Technolo-

gy)技术在中国推出的独立搜索门户,于 2004 年 6 月 21 日正式发布,具有简洁专业、海量、客观精准、国际化、稳定等特质,是值得依赖的专业搜索门户。目前设立了网页、图片、MP3 和网址四个频道,可以搜索全球 50 亿网页,5.5 亿图片,1 000 万音乐,网页搜索支持 38 种语言,居国际领先水平。

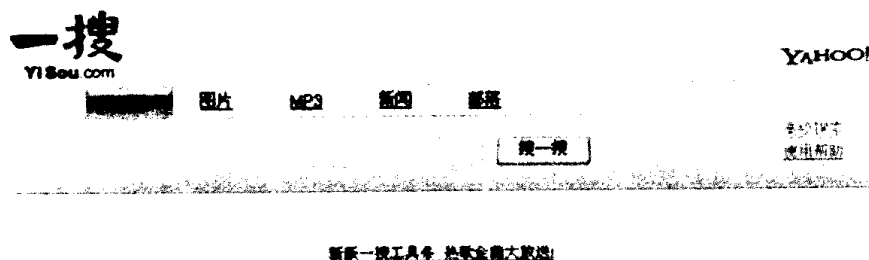


图 2-2

一搜具有先进的搜索分析与排序技术(而非人工干预)保证了结果的客观与精准。搜索利用文本、来源、相关链接及其他特定文件的特征进行技术相关性分析,以决定文件与查询的相关性。YST 不仅对网页内容分析,而且分析链接结构及查询意图(linkflux 技术),这种新的搜索方式更加适合越来越复杂、变化多样的网络。

一搜语法功能中包含关键字的查法,是在要加入的词前输入一个空格(如,化合物 反应),或在要加入的词前输入半角的加号“+”(化合物+反应)去除关键字的查法,与包含关键字相反,想要去除一个关键字,需要在这个词前输入减号“-”;但在减号之前必须留一空格(化合物-反应)。

一搜搜索中文结果的网页,默认的设置是简体中文和繁体中文结果,也可选择搜索所有网页或搜索简体网页。一搜搜索的默认设置是包含用户输入的所有关键字,且不区分英文字母大小写。网站排列位置会被一些因素影响,包括某个网站被链接的数量、日访问量以及网站或网页的网址、名称和内容等,搜索结果中列出的位置会根据其与关键词的相关性强弱排列。

4 中国搜索(<http://www.zhongsou.com/>)

2003 年 12 月 23 日,刚刚上市的慧聪国际集团重拳出击,原慧聪搜索正式独立运作成立了中国搜索。与一般的搜索引擎相比,中国搜索具有网页覆盖率高、数据更新快、支持中文模糊查询、强大的个性化查询、智能查询、内容相关性分析、便利的专业信息查询等优势,被公认为第三代智能搜索引擎的代表(如图 2-3)。

中国搜索的搜索引擎提供方便快捷的查询界面,仅需输入搜索内容并敲一下回车键(Enter),或单击“搜索”按钮即可帮助您找到相关的内容。关键词可以是任何中文、英文、数字,或中文、英文、数字的混合体,不区分英文字母大小写。中国搜索用“+”或直接输入一个空格来表示“与”、“/”表示“或”及“-”来表示“非”的关

系,例如 在搜索文本框中输入“ 软件硬件 ”,表示‘ 与 ’关系,即选中的文章中必须同时包含‘ 软件 ’和‘ 硬件 ’这两个搜索词。在搜索文本框中输入“ 软件／硬件 ”,表示“ 或 ”关系,即文章中只要包含‘ 软件 ’和‘ 硬件 ’这两个搜索词之中的一个词就会被查到,如果输入的是‘ 软件 —硬件 ’,则搜索结果全部包含‘ 软件 ’而没有‘ 硬件 ’。中国搜索的高级检索功能可以对检索词、检索范围、搜索内容进行限定。



图 2-3

5) 搜狐(<http://www.sohu.com.cn>)和搜狗搜索(<http://www.sogou.com/>)

“ 搜狐 ”是爱特信公司创办的中国人自己的网络神探,全部内容用人工分类,将网络机器人“ 抓 ”回来的网站划分到各个类别及子类下,并将各个网址抽象为一般摘要性信息,作为该网站的概括介绍(如图 2-4)。它提供分类查询和关键词全文检索两种方式,适合人们的思维习惯,查准率较高。“ 搜狐 ”收集的信息资源范围以中国为主,加入了部分英文网络,但网络提要或简或无。“ 搜狐 ”中文检索系统在进行全文检索时,利用独特的中文分词功能,对关键词在指定的文章中精确定位和评估,以免发生歧义冗余的检索结果;对关键词在文章里出现的次数、位置和每个关键字的数值进行综合评估,对每个检索结果进行合理的排序;具有拟人的思维模式,即针对不同的思维定向,自行定义搜索专题。



图 2-4

搜狐分类搜索引擎网站具有‘ 纵向成枝,横向成网 ’的分类类目体系,即根据网站的主题,首先把网站分为十八大类,再按细分主题层层分下去,形成一个“ 树枝状 ”分类结构,再根据不同用户的使用习惯和不同的分类方式,把不同类目下相关的类目链接起来,形成搜狐的‘ 网状 ’分类体系。其次搜狐分类结合图书分类法、企业／行业分类法和产品分类法的特点,在图书分类法指导下,根据网站发展的实际情况,不同类目运用不同的方法便于用户查找。现在搜狐还在关注网站资源发展