第3章 搜索引擎及网络信息检索

本章要求掌握搜索引擎中关键词检索的语法规则和 Google 的高级使用方法。由于网络信息以及搜索引擎变化快，各种检索规则也可能随之变化。因此，对本章介绍的检索工具及其应用范例要灵活运用，不能生搬硬套。

3.1 基本知识

3.1.1 基本概念

1. 搜索引擎

搜索引擎是 Internet 上的一种网站，它的主要任务是在 Internet 上主动搜索 Web 服务器信息并将其自动索引，其索引内容存储于可供查询的大型数据库中。一个搜索引擎由搜索器、索引器、检索器和用户接口 4 个部分组成。

1）搜索器

搜索器的功能是在 Internet 上漫游，发现和搜集信息。它常常是一个计算机程序，日夜不停地运行。它要尽可能多、尽可能快地搜集各种类型的新信息，同时因为 Internet 上的信息更新很快，所以还要定期更新已经搜集过的旧信息，以避免死链接和无效链接。

2）索引器

索引器的功能是理解搜索器所搜索的信息，从中抽取出索引项，用于表示文档以及生成文档库的索引表。索引器可以使用集中式索引算法或分布式索引算法。当数据量很大时，必须实现即时索引，否则就跟不上信息量急剧增加的速度。索引算法对索引器的性能（如大规模峰值查询时的响应速度）有很大的影响，一个搜索引擎的有效性在很大程度上取决于索引的质量。

3）检索器

检索器的功能是根据用户的查询在索引库中快速检出文档，进行文档与查询的相关度评价，对将要输出的结果进行排序，并实现某种用户相关性反馈机制。

4）用户接口

用户接口的作用是输入用户查询内容、显示查询结果、提供用户相关性反馈机制。主要的目的是方便用户使用搜索引擎，高效率、多方式地从搜索引擎中得到有效、及时的信息。用户接口的设计和实现使用人机交互的理论和方法，以充分适应人类的思维习惯。

2. 域名

从字面上讲，域名就是 Internet 上某个区域的名字，拥有了域名，就可以定义 Internet

上属于该区域的主机的名字。可以简单将域名理解为任何一个想要和 Internet 连接的个人或机构在 Internet 上的注册地址。

域名在整个 Internet 中必须是唯一的，当高级子域名相同时，低级子域名不允许重复；字母大小写在域名中没有区别；一台计算机可以有多个域名（通常用于不同的目的），但只能有一个 IP 地址。域名服务器实际上就是装有域名系统的主机。当所使用的系统没有域名服务器，只能使用 IP 地址，例如 202.206.242.23，而不能使用域名，如 library.ysu.edu.cn。

完整的域名包括 3 段，如 www.ibm.com 指的是 ibm.com 域内的一台名叫 www 的主机。此例的第 3 段是国际域名，属于顶级域。这类域名包括“.com”代表商业组织，“.edu”代表教育机构或大学，“.org”代表非营利性组织，“.net”代表网络（Internet 骨干网），“.gov”代表非军事性政府组织，“.mil”代表军事性政府组织。最新定义的一系列顶级域名尚未广泛使用，其中“.biz”只对全球企业界开放；“.info”将提供各种信息，对企业和个人开放；“.name”只针对个人开放；“.pro”是针对一些专业人员，如律师、医生和会计师，和“.name”类似，此部分域名也只允许三级域名的注册；“.aero”是专为合法的航运和民航系统定制的，包括航空公司、机场和相关的工业实体；“.coop”向商业合作组织开放，最初将只局限于国家商业合作组织协会或其会员；“.museum”代表得到承认的与文化和科学遗产有关的部门；“.cc”商业国际域名等效于“.com.xx”（xx 代表两个字母的国家代码，如：cn 为中国；jp 为日本）。

中国大陆域名在.cn 这个子域下面，可以将一个域名从后向前解读，如 www.legend.com.cn 就是中国的叫做 legend 的商业机构下的 www 主机。国际或国内域名在使用中没有任何区别。现已开通了中文域名，顶级域名包括“公司”、“网络”、“中国”、“政府”和“公益”。

3.1.2 搜索引擎的优点和缺点

搜索引擎现在已经成为网络信息检索最重要的指路标，几乎达到了无所不搜的地步。正确使用搜索引擎，可以检索到本书第 4～8 章所列的事实数据、图书、期刊、学位论文、专利等各类信息的题录或者部分原文，还能检索文字、图像、声音、动画等不同格式的文件。

但是目前的搜索引擎普遍存在着以下缺点：

（1）质量参差不齐，信息的分类加工欠规范，各搜索引擎在检索指令的输入格式与输入内容上存在差异并难以兼容，缺乏通行易用的检索方法与技巧。

（2）没有统一的网络信息分类标准，令网络用户无所适从，而且网络信息分类难以与传统的文献分类融合，与常见的学科及知识体系之间缺乏必要的内在联系，使得网络信息的分类体系对知识面或学科的覆盖率达不到要求，对专业性较强的深度信息的查全率较低。

（3）建立资源索引时针对性不强，搜索速度慢，死链接过多，重复信息及无效信息过多。

（4）对资源不具有选择和价值判断的能力，排序结果不理想，难以搜索动态网页，查全率下降。据调查，功能最强大的搜索引擎最多能覆盖 16%的网络信息资源，依照网络信息呈几何级数增长的趋势，搜索引擎覆盖的信息资源量还将有所下降。因此搜索引擎还无法完全代替第 4～8 章介绍的专门检索工具。

要解决这些难题，搜索引擎将向智能化、精确化、交叉语言检索、多媒体检索、专业化等适应不同用户需求的方向发展。现在已经出现了自然语言智能答询。自然语言的优势在

于，一是使网络交流更加人性化；二是使查询变得更加方便、直接、有效。

例如，用关键词查询计算机病毒，用 virus 这个词来检索，结果中必然会包括各类病毒的介绍、病毒是怎样产生的诸多无效信息，而输入自然语言“How can kill virus of computer?”，职能化搜索引擎在对提问进行结构和内容分析之后，或直接给出提问的答案，或引导用户从几个可选择的问题中进行再选择，将怎样杀病毒的信息提供给用户，提高了检索效率。

3.1.3 搜索引擎的类型

依据不同的原则，网络搜索引擎可划分成不同的类型。

1. 根据搜索引擎的数据检索机制划分

1）主题型搜索引擎

主题型搜索引擎将不断收集到的网上页面及地址信息以数据库的形式组织存储。查询时用户向其提问框中输入关键词，搜索引擎便会从数据库中检索与之相匹配的相关记录，按一定的排序返回给用户，如图 3-1 所示。主题搜索引擎的优点是查询全面、充分、直接、方便，用户能够对各网站的每篇文章中的每个词进行搜索，而且可使用布尔逻辑检索、短语检索等高级功能。但全文搜索的缺点是提供的信息虽然多而全，但由于没有分类型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。代表性的主题型搜索引擎是 Google(www. google. com)和百度(www. baidu. com)网站。

图 3-1　主题型搜索引擎 Google 的界面

2）分类型搜索引擎

通过用户浏览层次类型目录来寻找所需信息。分类一般按主题分类，并辅之以年代、地区等分类，如图 3-2 和图 3-3 所示。

分类搜索引擎可以使用户清晰方便地查找到某一大类信息，这符合传统的信息查找方式，尤其适合那些希望了解某一范围内信息，并不严格限于查询关键字的用户。但分类型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，可能遗漏某些重要的信息源。代表性的目录式分类型搜索引擎是 Yahoo(www. yahoo. com)、搜狐(dir. sohu. com)、新浪(dir. sina. com. cn)。

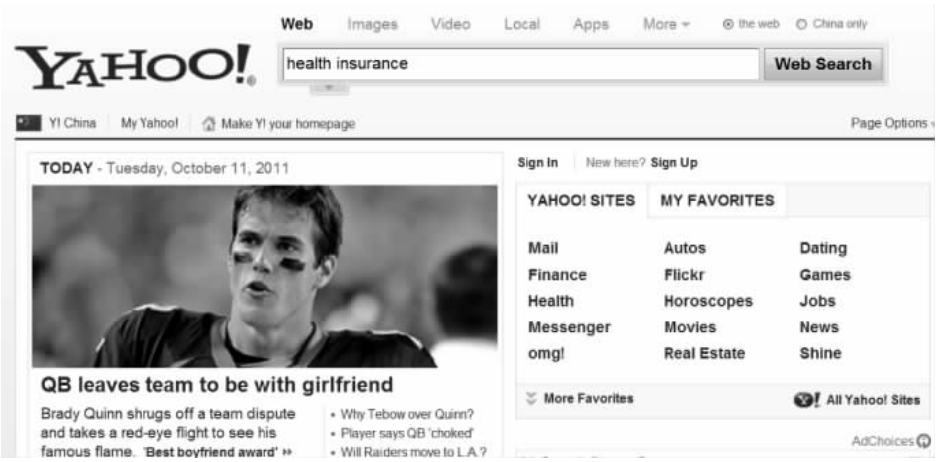

图 3-2 Yahoo 的目录界面

图 3-3 Yahoo 的分类显示界面

3）混合型检索工具则兼有以上两种类型的特点

图 3-4 所示是雅虎的检索界面，既有检索窗口，又有分类浏览目录。

2. 按检索内容划分

分为综合型、专题型和特殊型。综合型搜索引擎在采集标引信息资源时不限制资源的主题范围和数据类型，又称为通用型检索工具。例如，常见的 Google、新浪、搜狐和网易，这些搜索引擎网罗百科，信息种类繁多。

图 3-4　雅虎的检索界面

专题型搜索引擎专门采集某一主题范围的信息资源，并用更为详细和专业的方法对信息资源进行标引描述。例如，针对生物专利的搜索引擎、科技信息搜索引擎 Scirus、重点学科导航系统、学科信息门户。学科信息门户是指利用网络技术向用户提供某一学科领域各类网上资源和各种信息，提供对这一学科信息资源的“一站式”检索途径。学科信息门户实际上就是本学科领域网络信息资源的“信息超市”。重要学科门户网站如中国科学院国家科学数字图书馆学科门户（tsg. csdl. ac. cn）、CALIS 重点学科网络资源导航门户、ISIHighlycited. com（Institute for Scientific Information，ISI，美国科技信息研究所）。

特殊型检索工具是指那些专门用来检索图像、声音等特殊类型信息和数据的检索工具，如查询地图的检索工具 MapBlast（www. mapblast. com）、查询图像的检索工具 webseek（www. webseek. com）等。图 3-5 所示是地图搜索引擎的界面，其搜索到的地图如图 3-6 所示。

图 3-5　Go2map. com 的界面

图 3-6 Go2map.com 中检索到的重庆市北碚区地图

3. 按搜索引擎数据来源划分

分为单独型和集中型。单独型搜索引擎拥有独立的采集标引机制和独立的数据库，例如搜狐；集中型搜索引擎（如 3721 网）没有自己的数据库，利用一个统一的界面，查询其他单独型搜索引擎的数据库。

3.1.4 主题搜索引擎的关键词语法规则

1. 自动将关键词拆分进行模糊查询

目前自动将关键词拆分进行模糊查询的搜索引擎有 Google、百度、雅虎（www.yahoo.com）、3721（www.3721.com）等。例如，输入“西南大学”，首先会检索到关于西南大学的网页，然后会自动扩展到包含“西南交通大学”、“西南财经大学”、“西南政法大学”等名称的网页。这种搜索方式使查询结果的信息量和信息覆盖面非常大，但重复信息和无效链接较多，查询效率不高。

2. 按关键词进行精确查询

按关键词进行精确查询的有新浪、搜狐、网易、找到啦、中华网和常青藤等。这种搜索方式的搜索结果与查询目标的相关程度较高，尤其新闻查询的准确性更高。

为了扩大检索范围，实行精确查询的部分搜索引擎，也添加了自动拆分词组的功能。

3. 检索式的运算符号

如果想要得到最佳的搜索效果，就要使用搜索的基本语法来组织要搜索的条件。

1）使用逻辑运算算符

搜索引擎基本上都支持“与”、“或”、“非”、括号或引号等运算符号，但是不同的搜索引擎使用的运算符号不完全相同，常见的有 AND、OR、NOT 以及“＋”、“－”、“&”、“^”等逻辑符号。

AND 在中文搜索引擎都可以用空格代替；NOT 有时可以用减号代替，格式如“关键词 A－关键词 B”，减号前面要有空格。OR 有时用“|”表示，如在百度搜索引擎的格式是“关键词 A | 关键词 B”。Google 直接用 OR 表示，格式是“关键词 A OR 关键词 B”。

2）使用位置算符

AltaVista使用位置算符“NEAR/n”，n是两个词之间的单词的数目，如“Microsoft NEAR/5 Internet”表示在Microsoft和Internet这两个关键字之间的单词数目不得超过5个。如果不输入数字，表示两个词挨在一起。为了控制挨在一起的两个词之间的顺序，可以使用ADJ(adjacent)位置算符，如“Microsoft ADJ Internet”表示Microsoft必须在Internet之前。

3）使用字段限定

搜索引擎的字段限定方法俗称高级搜索。

(1) intitle。限定网页标题，title是网页的标题，intitle的意思是所有搜索结果的title中都要包含“关键词A”。例如，检索清华大学主页，排除仅仅含有介绍“清华大学主页”词组的其他网页，可以输入“intitle 清华大学”。

(2) site。限定在某类网站或某个网站内搜索。例如，“论坛搜索引擎 site：sowang.com”，是在sowang.com这个网站内搜索“论坛搜索引擎”的网页。

(3) filetype。限定文件类型。网上存在大量非网页格式的资料，如Word文件、pdf文件、ppt文件、xls文件等。用法是“关键词A filetype：文件格式后缀名”，如“个人年终总结 filetype：doc”，搜索结果全都是Word文件的个人年终总结。

(4) inurl。限定域名，inurl常见的使用方式是“关键词A inurl：英文字符B”。例如，“搜索引擎 inurl：ssyq”，是检索在URL中含有ssyq的网页中关于“搜索引擎”的信息。

3.2 典型的搜索引擎

1. Google

1）概述

Google(www.google.com)搜索引擎由斯坦福大学博士生Larry Page与Sergey Brin于1998年9月发明，Google公司于1999年创立。2000年7月份，Google替代Inktomi成为Yahoo公司的搜索引擎，同年9月份，Google成为中国网易公司的搜索引擎。1998年至今，Google已经获得30多项业界大奖，是易用性最强的搜索网站。但是Google最大的问题是死链接率比较高，中文信息的更新慢，不能及时淘汰已经过时的链接。虽然通过“网页快照”功能，可以减少目标页面不存在的现象，但Google的“网页快照”功能在中国有时无法访问。

2）Google的搜索语法

Google的基本检索算符是空格、减号和OR。

逻辑“与”(AND)用空格代替。用减号“－”表示逻辑“非”。注意，这里的“－”号是英文字符，而不是中文字符的“－”。此外，操作符与关键字之间不能有空格。

Google不支持通配符，如“＊”、“?”等，关键字后面的“＊”或者“?”会被忽略掉。Google对英文字符大小写不敏感，GOD和god搜索的结果是一样的。Google的关键字可以是词组(中间没有空格)，也可以是句子，但是用句子做关键字，必须加英文引号。

2. 百度

百度(www.baidu.com)是中国领先的搜索技术提供商，在国内提供搜索功能的大型网站中，有许多网站采用的是百度的搜索技术。除了提供网页搜索以外，百度还支持新闻、MP3、Flash的搜索，另外还有搜索特定关键字的“主题搜索”和网站目录导航功能。在显示

搜索结果时，有与 Google 的“网页快照”功能相同的“百度快照”，搜索速度快，返回结果的准确性也相当高。百度搜索的网站分类目录比较落后，缺乏应有的检索功能，并且死链接率也较高。

3. 搜狐

搜狐(search. sohu. com)的全部内容采用人工分类，适合人们的思维习惯。搜狐中文检索系统兼容传统的搜索引擎的所有标准语法和逻辑操作符。搜狐的网页搜索准确性较高，但搜索结果中重复的问题比较严重，虽然重复的结果未必是同一个网页，但都属于同一个网站。另外，搜狐的搜索结果中没有标出关键字，查阅起来非常不便。

4. 新浪网

新浪网(search. sina. com. cn)是一家为世界各地中国人提供全面 Internet 信息服务的国际性公司。它提供分类检索(主要针对网站)和关键词检索两种查询方法。它的关键词检索中可用冒号、空格、逗号、加号和“&”等。

5. 其他英文搜索引擎

1) Yahoo(www. yahoo. com)

搜索主题范围广，当不十分清楚自己究竟要找什么的时候，用 Yahoo 最好。

2) HotBot(www. hotbot. com)

擅长限定媒体类型、日期的复杂搜索，结果高度准确。

3) Altavista(www. altavista. com)

适合于刨根问底式的搜索、多语种搜索。

4) Excite(www. excite. com)

对泛泛搜索较擅长，附加内容多，但准确性低。

5) Infoseek(www. go. com)

准确性最高的搜索引擎之一。如果知道某个网站肯定存在，但不知道具体地址，用 Infoseek 较好。

6) Lycos(www. lycos. com)

执行复杂搜索的功能强大，不过准确性差，适合于查找 USENET 或按媒体类型搜索，也适用于购物。

6. 其他中文搜索引擎

目前在一些主要的中文搜索引擎中，所提供的搜索方式各不相同，以下列举部分。

提供网站、类目搜索的主要有雅虎和 3721 网；提供网站、网页、类目搜索的有中华网(www. china. com)和网易(www. 163. com)；提供网站、网页、新闻搜索的有新浪和搜狐等。而上海热线(www. online. sh. cn)和广州视窗(www. gznet. com)这两个地方性网站的搜索引擎只提供网站内部(网页和新闻)搜索。

雅虎知名度高且信息量大；雅虎与新浪搜索速度快；搜狐与新浪使用比较方便。搜狐与网易的搜索准确性上略强于其他网站，同时 21CN(www. 21cn. com)在亲朋推荐使用等方面选择比例高于其他网站。

7. 特殊型搜索引擎

网络上的信息资源丰富多样，为了查寻所需要的资料，用户往往使用 Infoseek、Yahoo

和 Excite 一类的检索引擎，但是，要查找一些专门的信息，如人名录、软件、新闻组、邮件列表、图像、视频、音频等，则必须使用特殊的检索工具。博客搜索引擎是其中最引人注目的。

博客（weblog，Blog）是一种在线网络出版形式，版面通常由单栏文本帖子按倒时间顺序不断更新排列构成，并能提供一些个人化的链接。Blog 搜索引擎的原理和 Google、Yahoo 等搜索引擎基本相同，都是由 spidering，、indexing 和 search 三大部分构成。不同的是 Google、Yahoo 等搜索引擎面向整个 Internet，处理的是 html 文件；Blog 搜索引擎专门面向 Blog，处理的是 xml 格式的 rss 和 atom 文件。与其他网络应用（如电子邮件、万维网等）相比，博客更具社会沟通的潜质，能为主流媒体提供新闻和公众观点来源，为教育业和商业创造知识共享的环境，同时能为个人提供一个自我表达和自我价值实现的平台。

比较常用而且支持中文的 Blog 搜索引擎主要包括 technorati（technorati. com），Feedster（www. feedster. com），icerocket（www. icerocket. com），bloglines（www. bloglines. com），blogpulse（www. blogpulse. com）等；而中文 Blog 搜索引擎还处于发展的初期，无论是知名度还是用户的使用率都比较低，主要有 Grassland（www. grass. org. cn），FeedSearch（www. feedsearch. net），feedss（feedss. com）等。目前，许多 Web 搜索引擎也开发出 Blog 搜索引擎产品，如 Google Blog Search（blogsearch. google. com）。

其他一些常用的特种搜索引擎如下所示。

- www. ctr. columbia. edu/webseek 采用了先进的特征抽取技术，是基于内容特征的搜索引擎。按目录方式组织数据。
- www. lib. sjtu. edu. cn/music. htm 提供高级的音频搜索功能。
- www. whowhere. com 提供人际交流的桥梁。
- www. alibaba. com 阿里巴巴强大的数据库及搜索引擎，隐藏着无数商机、合作伙伴和产品。
- www. go2map. com 中国地图搜索引擎，可查询中国各大城市的信息，但速度很慢。
- bingle. pku. edu. cn 天网 FTP 搜索，擅长寻找软件、图像、电影和音乐等文件。
- www. books. com 擅长热门书籍查找。
- www. humorsearch. com 擅长搜索笑话。
- www. medsite. com 擅长搜索医疗信息。
- 商机搜索：阿里巴巴（www. alibaba. com）拥有强大的数据库及搜索引擎，隐藏着无数商机、合作伙伴和产品。
- 论坛搜索：奇虎论坛搜索（search. qihoo. com）是目前最好的论坛搜索引擎。收录论坛数量多，索引范围广。
- 旧文档搜索：中国 Web 信息博物馆（www. infomall. cn/）专查网站的历史页面。
- 地图搜索：图吧（mapbar. com）提供中国国内 200 多个大中型城市的地图查询服务，同时提供博客地图、手机地图等特色功能。
- 图书搜索：中搜图书（book. httpcn. com/search）中国最大的电子图书搜索引擎，提供数万本电子图书（E 书）完全免费下载！
- 软件搜索：软件吧（www. soft8. net）中国首家专业软件搜索引擎，能轻松地找到几乎所有的软件和驱动。数据量大，排序合理。
- FTP 搜索：天网 FTP 搜索引擎（bingle. pku. edu. cn）擅长寻找软件、图像、电影和音

乐等文件。

- BT 资源搜索：飞客 BT 搜索引擎(fkee.com/)是目前来说最好用的 BT 资源搜索引擎。速度快，摘要信息丰富！
- 人肉搜索：猫扑网(dzh2.mop.com)是国内最早也是目前最大的人肉搜索网站(论坛)之一，于 1997 年 10 月建立，日平均浏览量一亿五千万，有注册用户 2200 万。

3.3 搜索引擎的检索技巧

3.3.1 常规的检索技巧

1. 分类查询

有时候不能准确地确定搜索的是什么或搜索的主题范围很广，如想知道关于法律学校、球类运动以及金融方面公共基金 mutual funds 的信息，如果利用主题搜索引擎 Altavista 检索，结果误检一个名字叫做 Mutual Funds 摇滚乐团的主页。所以应该首先考虑使用 Yahoo 一类的分类搜索引擎，以便对于主页的内容加以区分。在 Yahoo 的搜索框中输入 mutual funds，这种检索的实质是分类途径和主题途径结合。在返回的结果中，有 18 种与该论题有关的大类，其中有一类叫做“Business and Economy：Companies：Financial Services：Investment Services：Mutual Funds”，这是最符合要求的一类。单击该论题，出现了一些更深入的子论题和一些与该论题有关的网站。在这些网站中有 Morningstar.Net、Quicken.com 和 Mutual Funds Interactive 等，这些都和 mutual funds 有关，但也不要忽略那些子论题，其中有一类叫做 Reference and guides，单击它会出现一些更基本的网站。

2. 关键词查询

一般来说，首次检索时不要把条件限制得过于严格，最好是检索出一些结果后再使用其他限定条件来检索，即在结果中做二次检索。百度的搜索界面如图 3-7 所示，单击“在结果中找”按钮就是二次检索。

图 3-7 二次检索

此外使用太专业、生僻的词汇(如一些产品名称、产品品牌、公司名称、人名及专业名词)可能检索不到结果，不恰当的限定条件也导致有用的信息被过滤掉，因此要谨慎使用。

下面是初学者搜索时容易犯的错误。

(1) 输入错别字。

(2) 关键词太常见。例如，以“大学”、“论文”作为关键词，可能会检索出成千上万的网页，所以建议加限定条件。此外，虚词是常见词汇，没有检索意义，可能被一些搜索引擎列为禁止使用的单词而被过滤。关于禁用词，参考 10.2 节的实例。

(3) 滥用多义词。要小心使用多义词，如搜索 Java，要找的信息究竟是太平洋上的一个岛、一种著名的咖啡还是一种计算机语言？最好使用“岛”、“咖啡”或“计算机语言”等词语在二次检索时限定 Java 的多义性。