第 3 章　搜索引擎与信息搜索技巧

学习目标

1. 理解搜索引擎的作用、分类、工作原理。
2. 了解常见的搜索技巧，并能够根据需要使用这些搜索技巧提高搜索效率。
3. 熟悉常见的索引型搜索引擎，并能够熟练使用搜索引擎进行信息检索。
4. 了解搜索引擎的未来发展趋势。

搜索引擎是索引型网络检索工具的一种，其特点是可以定时地对收录网站的网页进行检索和更新，以确保其索引数据库信息的新颖、有效和相对完整。此外，搜索引擎还能根据用户输入的查询关键词，并遵从多个查询关键词的相对位置，对网页关键词的接近度进行分析，按照关键词的接近度区分搜索结果的优先次序，筛选与关键词较为接近的结果并提交给用户。

3.1　搜索引擎概述

3.1.1　含义与功能

1. 含义

搜索引擎(Search Engine，SE)是根据一定的策略，运用特定的计算机程序，搜集互联网上的信息，并对信息进行组织和处理，并根据用户需求，将处理后的信息结果返回给用户的应用系统，它是为用户提供检索服务的系统。

目前，搜索引擎泛指网络上以一定的策略搜集信息，对信息进行组织和处理，并为用户提供信息检索服务的工具和系统，是网络资源检索工具的总称。从使用者的角度看，搜索引擎为用户提供了一个查找 Internet 上信息内容的接口，查找的信息内容包括网页、图片以及其他类型的资源。

2. 功能

搜索引擎是高效获取网络信息资源的有力工具，网络用户可以通过搜索引擎查找新闻、

网页、图片、音乐、人物、视频等信息，而各种功能新颖的搜索引擎产品也不断出现，因此搜索引擎的实际应用功能是无法尽数的。下面仅从三方面概括介绍搜索引擎应具备的最基本的功能。

1）及时、全面地搜索网络信息

迅速及时地查找到尽可能多的网络信息，并将新出现的信息收录到自己的索引数据库中，这是搜索引擎技术的首要功能。

2）搜索有效且有价值的网络信息

搜索引擎应提供当前有效的、有价值的网站或网页信息，无效的信息不但没用，还可能造成损害。

3）有针对性地搜索网络信息

网络信息搜索的针对性是指搜索引擎能够通过名词的关联性等技术满足人们对主题内容的深度查找。

由于目前的搜索引擎在技术上存在一定的局限性，无论是信息搜集的及时性、信息甄别的有效性、信息价值评判的合理性方面，还是识别主题的针对性方面，都还不能达到人们的要求，因此，虽然现有的搜索引擎的基本工作原理已经相当成熟，但在质量、性能、服务功能和服务方式上依然存在较大的提升空间。

3.1.2　分类

按照不同的分类原则，搜索引擎可以有多种分类方式。按照工作方式或者检索机制进行分类是最常见的一种分类方式。按照搜索引擎工作方式的不同，可以将搜索引擎分为目录型搜索引擎、索引型搜索引擎和元搜索引擎。

1. 目录型搜索引擎

目录型搜索引擎也称分类索引（Search Index）或网络资源指南（Directory），是一种网站级的浏览式搜索引擎。目录型搜索引擎是由专业信息人员以人工或半自动的方式搜集网络资源站点信息，且采取人工方式对搜集到的网站加以描述，并按照一定的主题分类体系进行编制，形成的一种可供浏览、检索的等级结构式目录（网站链接列表）检索系统。目录型搜索引擎下，用户通过逐层浏览目录的方式，在目录体系的从属、并列等关系引导下逐步细化，寻找合适的类别，直至定位到具体的信息资源。目录型搜索引擎往往根据资源采集的范围设计详细的目录体系，用户检索的结果是网站的名称、网址链接和每个网站的内容简介。

目录型搜索引擎收录的网络信息资源都经过了专业信息人员的鉴别、筛选和组织，并且层次结构清晰，易于查找和导航，质量高，确保了检索工具的质量和检索的准确性。但目录型搜索引擎的数据库规模相对较小，且对新兴学科、交叉学科和某些分类主题的内容收录不够全面，同时由于检索范围只限定在对网站的描述中，因此检索范围非常有限；此外，由于目录型搜索引擎的更新维护速度受系统人员工作时间的制约，更新不及时就可能导致检索内容的查全率不高，因此，目录型搜索引擎比较适用于查找综合性、概括性的主题概念，或对检索的准确度要求较高的课题。

2.索引型搜索引擎

基于关键词检索的索引型搜索引擎是名副其实的搜索引擎，是一种网页级搜索引擎。索引型搜索引擎主要使用一个称作“网络机器人(Robot)”或“网络蜘蛛(Spider)”或“网络爬虫(Crawlers)”的自动跟踪索引软件，通过自动的方式分析网页的超链接，依靠超链接和HTML代码分析获取网页信息内容，并采用自动搜索、自动标引、自动文摘等事先设计好的规则和方式建立和维护其索引数据库，以Web形式给用户提供一个检索界面，供用户输入检索关键词、词组或逻辑组配的检索式，其后台的检索代理软件代替用户在索引数据库中查找出与检索提问匹配的记录，并将检索结果反馈给用户。索引式搜索引擎实际只是一个WWW网站，与普通网站不同的是，索引型搜索引擎网站的主要资源是它的索引数据库，索引数据库的信息资源以WWW资源为主，还包括电子邮件地址、用户新闻组、FTP等资源。

索引型搜索引擎由自动跟踪索引软件生成索引数据库，数据库的容量非常庞大，收录、加工的信息范围广、速度快，能向用户及时提供最新信息。但由于标引过程缺乏人工干预，因此准确性较差，加之检索代理软件的智能化程度不是很高，从而导致检索结果的误差较大。

索引型搜索引擎比较适用于检索特定的信息及较为专深、具体或类属不明确的课题。从搜索结果来源的角度来看，索引型搜索引擎又可进一步细分为两种，一种拥有自己的检索程序，并且构建索引数据库，搜索结果直接从自身的数据库中调用；另一种租用其他搜索引擎的数据库，并按指定格式排列搜索结果。

目录型搜索引擎与索引型搜索引擎在使用上各有优劣。目前，目录型搜索引擎和索引型搜索引擎呈现出相互融合渗透的趋势，很多搜索引擎网站也都同时提供目录和基于自动搜索软件的搜索服务，以便于尽可能地为用户提供全面的检索服务和检索结果。如Google索引型搜索引擎就是借用Open Directory目录型搜索引擎提供分类查询功能的，而Yahho目录型搜索引擎则首先通过与Google等搜索引擎合作，然后通过收购推出了自己的雅虎全能搜以提升搜索功能和扩大搜索范围。在默认搜索模式下，目录型搜索引擎首先返回自己分类目录中匹配的网站，而索引型搜索引擎则默认进行网页搜索，因此，用户一般将索引型搜索引擎的查询称为全网站搜索、全网页搜索，把目录型搜索引擎的查询称为分类目录搜索或分类网站搜索。

3. 元搜索引擎

元搜索引擎(Meta Search Engine，MSE)是一种将多个独立的搜索引擎集成到一起，提供统一的用户查询界面，将用户的检索提问转换成其共享的各个独立搜索引擎能够接受的查询语法，同时提交给多个独立搜索引擎并检索它们的资源库，然后将获得的反馈结果经过聚合、去掉重复信息及综合相关度排序等处理，再将最终检索结果一并返回给用户的网络检索工具。由此可见，元搜索引擎是对搜索引擎进行搜索的搜索引擎，是对多个独立搜索引擎的整合、调用、控制和优化利用。相对于元搜索引擎，可被利用的独立搜索引擎称为源搜索引擎(Source Search Engine)或成员搜索引擎(Component Search Engine)。

元搜索引擎主要由检索请求预处理、检索接口代理和检索结果处理三部分构成。其中，检索请求预处理部分负责实现用户个性化的检索设置要求，包括调用哪些搜索引擎、检索时

间限制、结果数量限制等；检索接口代理部分负责将用户的检索请求翻译成满足不同搜索引擎本地化要求的格式；检索结果处理部分负责所有元搜索引擎检索结果的去重、合并、输出处理等。与独立搜索引擎相比，元搜索引擎一般没有自己的网络机器人及数据库，但在检索请求预处理、检索接口代理和检索结果处理等方面，通常都有自己研发的特色元搜索技术。元搜索引擎的工作过程一般为：用户向元搜索引擎发出检索请求，元搜索引擎根据请求向多个搜索引擎发出实际检索请求，搜索引擎执行元搜索引擎检索，并将检索后的结果以应答的形式传送给元搜索引擎，元搜索引擎再将从多个搜索引擎获得的检索结果汇集整理，通过浏览器展示给用户，如图 3-1 所示。

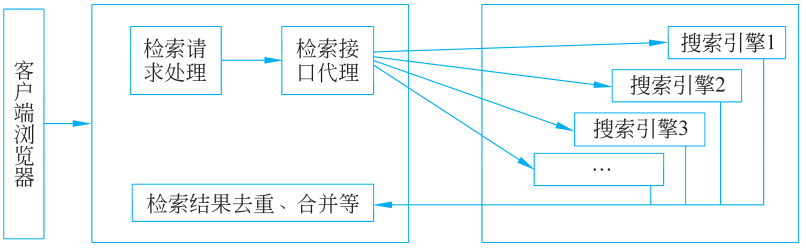

图 3-1　元搜索引擎工作过程示意

集合式搜索引擎(All-in-One Search Page)是元搜索引擎发展进程中的一种初级形态，它是通过网络技术，在一个网页上链接多个独立搜索引擎，检索时须点选或指定搜索引擎，一次输入，多个搜索引擎同时查询，搜索结果由各搜索引擎分别以不同的页面显示。集合式搜索引擎无自建数据库，无须研发支持技术，也不能控制和优化检索结果，其实质是利用网站链接技术形成的搜索引擎集合，而并非真正意义上的搜索引擎。

4. 智能搜索引擎

智能搜索引擎是结合了人工智能技术的新一代搜索引擎，它将信息检索从基于关键词层面提升到了基于知识层面，对知识有了一定的理解和处理能力，能够实现分词技术、同义词技术、概念搜索、短语识别以及机器翻译技术等。智能搜索引擎具有信息服务的智能化和人性化，允许用户采用自然语言进行信息检索，为人们提供了更方便、更确切的搜索服务。

5. 学科信息门户

学科信息门户是将特定学科领域的信息资源、工具和服务集成为一个整体，为用户提供可靠的学科信息导航，也称门户网站或信息门户。学科信息门户通常为用户提供对网上信息的“密集”访问，即将来自不同信息源的信息集合在一个页面上，使用户得以从一个统一的入口检索不同网站信息，无须逐个访问每一个网站。不同于搜索引擎，学科信息门户经过人工选择和标引，保证了信息的质量，对于高校教学和科研工作而言，学科信息门户具有特别的意义，它能使科研人员用较少的精力和时间浏览到高质量的专业信息。

6. 网络版参考咨询工具

各类型传统的工具书几乎都有了网络版，这些网络版参考咨询工具门类丰富，囊括词典、百科全书、人名录、书目、文摘、索引等各种类型，并且提供便利的全文链接，是网络信息

检索工具中的重要一员。

7. FTP 资源检索工具

网络上存在着大量为普通公众提供文件服务的 FTP 服务器，承载着众多的数据资料和免费软件，并允许用户匿名登录、下载和上传信息，极大地扩展了资源共享的空间。

3.1.3 工作原理

为了让用户以最快的速度获取到想要的搜索结果，搜索引擎通常会将待查找的内容以预先整理好的网页索引的形式存储在搜索引擎数据库中。普通的信息搜索不能真正理解网页上的内容，它只能机械地匹配网页上的文字，而真正意义上的搜索引擎通常指收集了互联网上几千万到几十亿个网页，并对网页中的每一个文字（关键词）进行索引，并建立索引数据库的全文搜索引擎，当用户查找某个关键词时，所有在页面内容中包含该关键词的网页都将作为搜索结果被搜索出来，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[11]。

1. 构成模块

典型的搜索引擎通常由三大模块组成，分别是信息采集模块、信息组织模块和信息检索模块。

1）信息采集模块

信息采集模块的主要功能是搜索、采集和标引网络中的网站或网页信息。信息采集有人工采集和自动采集两种。人工采集是由专门的信息采集人员跟踪和选择有价值的网络信息资源，并按照一定的方式进行分类、组织、标引并组建成索引数据库。自动采集通过采用一种称为 Robot 的网络自动跟踪索引程序完成信息的采集，由 Robot 在网络上检索文件并自动跟踪该文件的超链接以及循环检索被参照的所有文件。Robot 的具体过程是：首先打开一个网页，然后把该网页的链接作为浏览的起始地址，把被链接的网页获取过来，抽取网页中出现的链接，并通过一定的算法决定下一步要访问哪些链接，同时，信息采集器将已经访问过的 URL 存储到自己的网页列表并打上已搜索的标记，自动标引程序检查该网页并为其创建一条索引记录，然后将该记录加入整个查询表中，信息收集器再以该网页的超链接为起点继续重复这一访问过程，直至结束。

一般搜索引擎的采集器在搜索过程中只取链长比（超链接数目与文档长度的比值）小于某一阈值的页面，数据采集于内容页面，不涉及目录页面。在采集文档的同时记录各文档的地址信息、修改时间、文档长度等状态信息，用于站点资源的监视和资料库的更新。在采集过程中，还可以构造适当的启发策略，指导采集器的搜索路径和采集范围，减少文档采集的盲目性。

为了维护采集页面的新颖性，搜索引擎可以采用定期采集和增量采集两种方式进行内容采集。每次定期采集将替换上一次的内容，因此，该种采集方法也称批量采集。由于每次采集都相当于重新采集一次，因此，对于大规模搜索引擎来说，每次采集的时间通常会花费几周。由于此种方法的开销较大，因此搜索引擎两次采集的时间间隔不会很短（例如

Google 曾每隔 28 天采集一次)。定期采集的好处是系统实现比较简单,但其主要缺点是时新性(freshness)不高,且重复采集带来的额外消耗也较大。增量采集开始时搜集整个网络,但以后只采集那些新出现的网页和在上次采集后有改变的网页(增加到数据库中),以及发现自从上次搜集后已经不再存在的网页(从数据库中删除)。由于除新闻类网站外,许多网页的内容变化并不是很经常的(有研究指出,50%的网页的平均生命周期大约为 50 天),每次搜集的网页量不会很大,所以可以经常性地启动采集过程(例如每天)。增量采集表现出来的信息时新性较高,但其主要缺点是系统实现比较复杂,复杂性不仅体现在搜集过程,还在于下面将要谈到的索引建立过程。

在具体的搜集过程中,抓取一篇篇的网页可以通过不同的方式实现。最常见的一种方式是爬取,即将 Web 服务中的网页集合看作一个有向图,搜集过程从给定起始 URL 集合 S("种子")开始,探查网页中的超链接,沿着网页中的超链接,按照某种搜索策略(如深度优先、广度优先等)不断从 S 中移除 URL(探查过的 URL 将移除),并下载相应的网页,并解析出网页中的超链接,判断该超链接是否已被访问过,并将未访问过的 URL 加入集合 S 中。

需要说明的是,搜索引擎不可能对 Web 上的所有网页都进行完全搜集,为此,通常在某种条件限制下确定搜集过程的结束条件(例如磁盘满、搜集时间超过预期时间等)。那么,在有限的条件下,哪些网页值得搜索呢?对比较重要的网页进行搜索,是一个不言而喻的答案。那么,如何使搜索引擎搜索到比较重要的网页呢?研究表明,这与搜索策略有一定关系,使用图搜索算法中的广度优先搜索策略进行搜索得到的网页集合,要比使用深度优先算法得到的网页集合重要。

广度优先搜索算法实施时的一个困难是:由于 HTML 的灵活性,其中出现 URL 的方式各种各样,要从每一篇网页中提取出包含的所有 URL 是一件很难保证的事情。此外,由于网页站点具有蝴蝶结结构特性(图 3-2),因此广度优先搜索方式搜集到的网页不大会超过所有目标网页数量的 2/3。

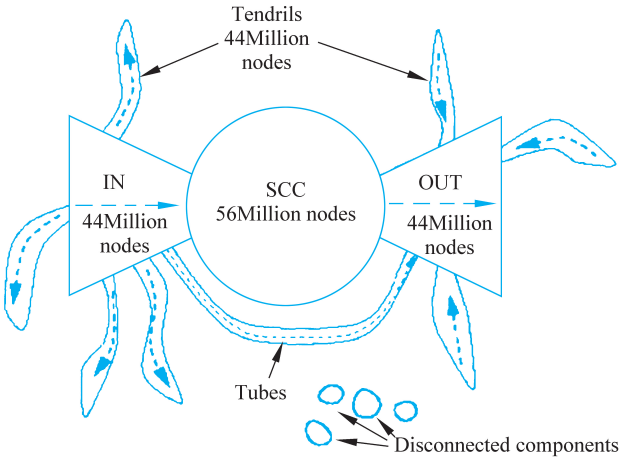

图 3-2　蝴蝶结结构特性

广度优先搜索算法的另一种改进方法是:在第一次全面搜集网页后,搜索引擎将维护一个相应的 URL 集合 S,此后的所有搜索将直接基于该搜索集进行,即每搜索到一个网页,

就判断该集合是否发生变化，如果发生变化且有新的 URL 加入，则抓取新增加的 URL 对应的网页，然后将这些新的 URL 添加到集合 S 中；如果 S 中的某个 URL 对应的网页不存在了，则将该 URL 从 S 中删除。由此可见，这种广度优先搜索方式是一种极端的广度优先搜索策略，即第一层搜索集合是一个比较大的 URL 集合，从该集中的 URL 开始，最多只往下搜索一层。

此外，另一种抓取网页的方法是：让网站拥有者主动向搜索引擎提交它们的网址（为达到宣传自己网站信息的目的，网站作者通常会很乐意进行这种操作），搜索引擎系统在一定时间内（两天到数月不等）定向对提交网址的网站派出“蜘蛛”程序，扫描该网站的所有网页并将有关信息存入数据库中。当前，大型商业搜索引擎一般都采用此种方法。

不同的信息采集方式和不同的自动采集软件采用的标引和搜索策略都各不相同，这对信息检索的质量有着直接影响。自动采集能够自动搜索、采集和标引网络中的众多站点和网页，保证了对网络信息资源跟踪和检索的有效性和及时性；人工采集基于专业性的资源选择和分析标引，保证了资源的采集质量和标引质量。因此，目前许多网络信息资源检索工具都采取了自动采集和人工采集相结合的信息采集方式。

2）信息组织模块

采集到原始网页信息后，还需要进行信息资源的组织，以满足用户方便高效的信息查找需求，该功能通过信息组织模块完成。

信息组织模块又称表查询模块，该模块的核心作用是建立全文索引数据库。搜索引擎信息组织和整理的过程称为建立索引的过程，该阶段实现了将纷繁复杂的网站或网页数据整理成可以被检索系统高效、可靠、方便使用的格式。

通过数据库管理系统组织采集的网络信息资源并建立相应的索引数据库是搜索引擎提供检索服务的基础。不同搜索引擎的数据库的收录范围不一样，数据库中收录的网络信息资源数量存在很大差异，数据库中记录的网络信息资源内容也各不相同。索引数据库中的一条记录既可以对应于一个网站，记录的内容包括网站名称、网址、网站的内容简介等，也可以对应于一个网页，记录的内容包括网页标题、关键词、网页摘要及 URL 等信息。由于数据库的规模和质量直接影响检索的效果，因此需要对数据库数据进行及时的更新和处理，以保证数据库能够准确地反映网络信息资源的当前状况，这样，搜索引擎就能从数据库已保存的信息中迅速找到所需的信息资料了。

信息组织模块对信息组织和处理的对象可以分为两类，分别是对内容信息的处理和对非内容信息的处理。对内容信息的处理主要是对文本内容信息的处理，目的是建立以词项（Term）为中心的文本倒排索引，以提高信息系统的检索效率。网络信息不仅包括内容信息，也包括一定程度上的非内容信息，如链接结构信息、文本结构信息等，这些结构信息在评价数据质量、挖掘数据相关性等方面发挥着十分重要的作用，这些信息就是非内容信息。对非内容信息的处理主要是处理链接结构信息、文本结构信息等，非内容信息的组织与处理最广泛的应用是利用超链接结构分析方法对网络数据质量进行评价。具体而言，信息组织模块需要通过对采集的海量信息执行关键词提取、重复或转载网页的消除、链接分析、网页重要度计算等为后期的用户信息检索做好准备。

（1）关键词提取。

对于普通网页，查看其源代码（可以通过浏览器的“查看源文件”功能查看）就可以发现，

网页中除了包含用户从浏览器中正常看到正常文字内容外，还含有大量的 HTML 标记（以“<”和“>”括起来的部分），如图 3-3 所示。此外，由于 HTML 文档产生来源的多样性，多数网页内容比较个性随意，除包含有意义的文字内容外，还包含许多与主要内容无关的信息，如广告、导航条、版权说明等，如图 3-4 所示。

图 3-3　HTML 网页文档源代码示意

图 3-4　HTML 网页文档示意

为支持后续用户信息查询的需求，搜索引擎需要从网页源代码中提取出能够代表该文档内容的特征信息。从认识和实践角度看，文档中所含的关键词即可作为文档特征的代表，为此，信息处理模块需要做的第一个基本工作就是从网页源代码中提取出其所含的关键词。对中文网页而言，需要根据分词词典，用分词工具从网页文字中切分出分词词典中所含的词语，使用分割成的多个词语近似代表一个网页。对于一般文档和常用的分词词典而言，文档

分词后将可能得到多个结果词;从效果和效率考虑,不应让所有词都出现在分词后的词表中,还需要去除掉诸如“的”“在”等没有内容指示意义的停用词;分词结构中的某一个词,如果在一篇网页中多次出现,直观理解可以得出,该词对于该篇文档而言比较重要。至此,一个网页可以通过约 200 个词语进行标识。

(2) 重复或转载网页的消除。

信息的网络化使得信息的传播和复制变得非常便捷,据统计,互联网上网页的重复率平均约为 4,即当用户通过一个 URL 在网上看到一篇网页时,平均还有另外 3 个不同的 URL 也会给出相同或者相似的内容,这种现象使得用户有了更多获得信息资源的机会,但对于搜索引擎而言,重复网页的存在不仅需要其消耗更多的机器时间和网络带宽资源以采集网页信息,而且重复结果的返回也会为用户的体验带来挑战。为此,消除内容重复或主题内容重复的网页是搜索引擎的一个重要任务。

(3) 链接分析。

从信息检索角度看,如果检索系统仅从文字内容考虑,则可以依据共有词汇假设确定网页的关键词,也可以在此基础上考虑词频和词在文档集合中出现的文档频率等以进一步提高精准度。当考虑 HTML 标记时,关键词的提取还可以进一步改善,如在同一篇文档中,<H1>和</H1>之间的信息很可能就比在<H4>和</H4>之间的信息更重要。此外,互联网范围内,Web 页面之间通过超链接标签对“<A>”和“</A>”实现网页之间的互联互通,如果一个网页与其他网页之间的链接对越多,则说明其重要程度越高,据此可以作为网页重要程度的衡量标准。

(4) 网页重要程度的计算。

实际搜索时,搜索引擎返回给用户的是一个与用户查询相关的结果列表,列表中各条目的排序体现了其重要性。对于重要性的衡量,人们参照了科技文献重要性的评估思想,即“被引用多的就是重要的”。“引用”可以通过网页中的超链接体现。作为 Google 核心技术的 PageRank 就是这种思路的成功体现,重要程度的具体计算方法可以在用户查询前实施,也可以在用户查询时进行。

3) 信息检索模块

经过以上处理,搜索引擎就可以将原始网页集合处理为与网页对应的一组子集元素以及元素之间的内部表示,元素间的内部表示构成了信息检索的直接基础。对每一个子集元素来说,这种表示至少包含原始网页文档、URL 和标题、编号、所含的重要关键词的集合(以及它们在文档中出现的位置信息)及其他一些指标(如重要程度、分类代码等)。其中,系统关键词的集合和文档的编号构成了倒排文件的结构,使得信息检索系统一旦得到一个关键词输入,就可以迅速输出该关键词对应的相关文档的编号,这种根据输入关键词获取网页结果信息的功能是通过信息检索模块实现的。

信息检索模块又称信息查询服务模块,是搜索引擎与用户查询需求的交互界面,是实现检索功能的程序,它实现了将用户输入的检索表达式拆分为具有检索意义的字或词,再访问查询表,如果找到与用户要求内容相符的网站,便采用特殊的算法,通常根据网页中关键词的匹配程度、出现的位置、频次、链接质量等计算出各网页的相关度及排名等级,然后根据关联度高低,按顺序将这些网页链接返回给用户。信息检索模块主要完成以下三方面的工作。

（1）查询方式和匹配。

查询方式是指检索系统允许用户提交查询的形式。一般认为，对于普通网络用户来说，最自然的查询方式就是想要什么就输入什么，但该方式是一种相当模糊的说法，如用户输入“中国人民解放军战略支援部队信息工程大学”，则检索系统可能认为用户想检索中国人民解放军战略支援部队信息工程大学目前向外发布了哪些信息，但用户也许是想看看今年的招生政策，或想了解外界目前对中国人民解放军战略支援部队信息工程大学的评价，由此可见，用户需求和系统的理解差距较大。在其他情况下，用户可能关心的是间接信息，如用户输入“珠穆朗玛峰的高度”，则“8848 米”应该是他需要的，但按照传统信息检索系统的检索方法，该数据不可能包含在分词结果中；当用户输入“惊起一滩鸥鹭”时，很可能是想知道该诗的作者是谁，或希望能获得该诗的其他语句。

尽管如此，用一个词或者短语直接表达信息需求，希望网页中含有该词或者该短语中的词，依然是主流的搜索引擎查询模式，这不仅是因为它的确代表了大多数的情况，还因为它比较容易实现。

（2）结果排序。

得到满足用户查询需求的相关文档集合后，还需要以一定的形式将结果集合呈现给用户。多数搜索引擎采用列表形式进行展现。列表是一种按照某种标准确定列表中元素排列顺序的方法。通常，搜索引擎是依照搜索结果与查询词之间的相关性确定搜索结果的排列顺序的，这是一种有效的顺序排列方式。但事实上，有效地定义相关性本身就是一件很困难的事情，从原理上看，它不仅与查询词有关，还与用户的检索背景及查询历史有关，不同查询需求的用户可能会输入同一个查询词，同一个用户在不同的时间输入的相同的查询词可能针对不同的信息需求。为了形成一个合适的顺序，早期的搜索引擎采用传统信息检索领域很成熟的基于词汇出现频度（词频）的排序方法。基本词频思想是：一篇文档中包含的查询词越多，则该文档的排序就应越靠前。随后，又提出了基于文档频率的排序方法，其基本思想是：若一个词在越多的文档中出现过，则该词用于区分文档相关性的作用就越小。以上方法都具有合理性，但由于网页编写的自发性和随意性，仅针对词的出现决定结果文档的顺序，在网页上进行信息检索结果展示表现出了明显的缺点。为此，人们提出了基于 PageRank 的排序方法，即通过为每篇网页建立独立于查询词的重要性指标，将它和查询过程中形成的相关性指标结合在一起，形成一个最终的排序。该方法是目前搜索引擎给出查询结果的主要排序方法。

（3）文档摘要。

搜索引擎给出的结果是一个有序的条目列表，每一个条目有 3 个基本元素：标题、网址和摘要。其中，摘要需要从网页正文中生成。一般来讲，从一篇文字中生成一个恰当的摘要是自然语言理解领域的一个重要课题，目前已经取得了傲人的成果。传统搜索引擎常采用两种生成摘要的方法，一种方法是静态生成方式，即独立于查询，按照某种规则，事先在预处理阶段从网页内容中提取出一些文字，如截取网页正文开头的 512 字节或者将每一个段落的第一个句子拼起来，此种方式生成的摘要预先存放在检索系统中，一旦相关文档被作为检索结果选中，搜索引擎就读出该文档预先生成的摘要并返回给用户。显然，此种方式对查询子系统来说是最轻松的，不需要做另外的处理工作。但该方式的最大缺点是摘要与查询无关，即当用户输入某个查询时，他一般希望摘要中能够突出显示和本次查询直接对应的文

字，希望摘要中出现和其关心的文字相关的句子。因此，人们提出了动态摘要方式，即搜索引擎在响应查询时，根据查询词在文档中的位置提取出网页周围的文字，在显示时将查询词高亮显示，这是目前大多数搜索引擎采用的方式（图 3-5）。此种方式下，为了保证查询的效率，需要查询分词时记住每个关键词在文档中出现的位置。

图 3-5　查询结果页面中查询词高亮显示示意

2. 工作过程

搜索引擎的基本工作过程可以划分为 3 个阶段，即信息发现、建立索引库和信息查询与排序。

1）信息发现

搜索引擎首先需要按照一定的方式在互联中搜集相关网页信息，并把获得的信息保存下来以建立索引库。需要注意的是，搜索引擎搜集的信息不只包含网页内容信息，还包括用户搜索习惯等其他信息。

2）建立索引库

搜索引擎对信息发现结果获得的网页相关信息进行提取和组织，建立索引库。该阶段，搜索引擎需要首先进行数据分析与标引处理，对已经收集到的资料按照网页中的字符特性等进行分类，建立搜索原则。如对于中文词语“软件”，搜索引擎必须为其建立一个索引，当用户查询该词时，搜索引擎知道去哪里调取相关资料。需要说明的是，对于网页内容而言，不同搜索引擎对字符的处理方式（如大小写处理、中文的分词等）存在不同，每个搜索引擎都有自己的存档归类方式，这些方式往往影响着未来的搜索结果。标引完成后，搜索引擎需要进行数据组织，负责形成规范的索引数据库或便于浏览的层次型分类目录结构，通常需要计算网页的优先等级，该原则在 Google 中非常重要。一个接受很多网页链接的网页，搜索引擎必然在所有的网页中将其排序进行提升。

3）信息查询与结果排序

在该阶段，搜索引擎会驱动检索器，根据用户输入的查询关键词在索引库中检索结果文档，并对检索结果和查询词进行相关度评价，以对将要输出的结果进行排序，然后将查询结

果返回给用户。搜索引擎不仅负责帮助用户用一定的方式检索索引数据库，获取符合用户需要的网络资源信息，还负责提取检索用户的相关信息，利用这些信息提高检索服务的质量，即信息挖掘。信息挖掘在个性化服务中起着关键的作用。

3. 工作机制

搜索引擎的工作机制就是采用高效的蜘蛛程序，从指定的 URL 集合开始，顺着网页上的超链接，采用深度优先算法或广度优先算法对整个互联网上的网页资源进行遍历，并将网页信息抓取到本地数据库，然后使用索引器对数据库中的重要信息单元（如标题、关键字、摘要等）或全文进行标引，以供查询导航，当用户提出查询请求时，搜索引擎的检索器就将用户通过浏览器提交的查询请求与索引数据库中的信息通过某种检索技术进行匹配，并将检索结果按某种方式进行排序并返回给用户。搜索引擎的工作机制示意如图 3-6 所示。

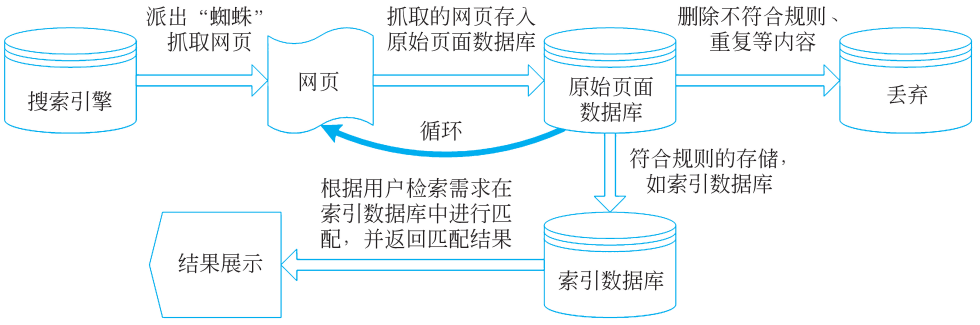

图 3-6　搜索引擎工作机制示意

基于以上分析，可以给出搜索引擎的体系结构图，如图 3-7 所示。

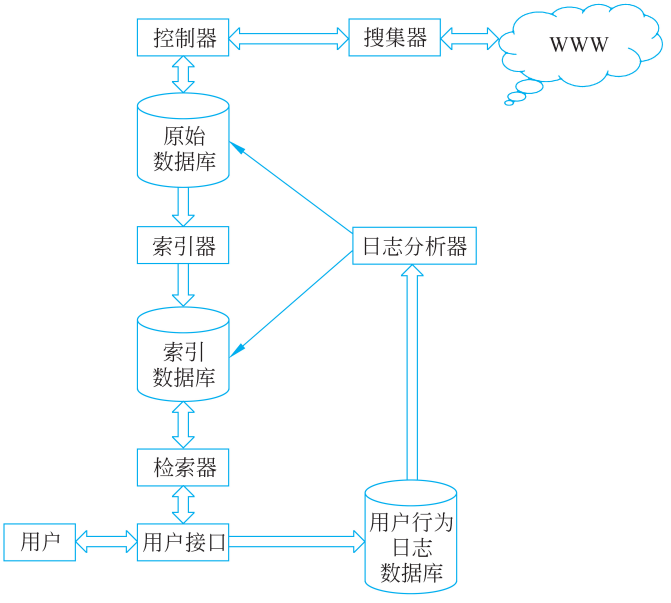

图 3-7　搜索引擎体系结构示意

在搜索引擎体系结构示意图中，大部分模块和前面的原理描述直接对应，这里需对控制器模块进行说明。对于需要向大规模搜索引擎稳定地提供网页数据的爬虫来说，其需要持续、海量地搜集网页，此时需要综合考虑效率、质量和“礼貌”的问题，这就是控制器需要实现的功能。所谓效率，是指如何利用尽量少的资源（计算机设备、网络带宽、时间等）完成预定的网页搜集量；所谓质量，是指在时间有限和搜集网页数量有限的情况下，需要尽量做到比较“重要”的网页被搜索到或不漏掉很重要的网页；所谓“礼貌”，是指爬虫在进行网页抓取时，需要尽量避免在短时间内频繁地从某网站抓取过多的网页内容，以免影响该网站的正常工作或被该网站实施“反爬取”。

3.1.4 搜索原则

互联网的快速发展促进了搜索引擎市场的繁荣，越来越多的搜索引擎应运而生。由于目标定位、系统性能等方面的差距，各搜索引擎的数据采集范围、采集方式、排序机理等均存在一定区别。如何根据自己的需要，借助搜索引擎快速准确地搜索到个人需要的目标信息，对于用户来说非常重要。下面介绍一些提高检索效率的原则方法，以帮助用户提高搜索效率。

1. 分析搜索对象，选用适当的搜索引擎

搜索引擎品种多样，工作方式不同，信息来源更有差异，每种搜索引擎都有其不同的特点，也有其局限性，只有选择合适的搜索引擎才能得到最佳的目标结果。选择哪个搜索引擎，用户应根据个人的具体查询需求而定，一般的选择规则是：如果查找目标不太具体明确，则可以使用综合性索引型搜索引擎，如 Google 搜索引擎、Baidu 搜索引擎、Bing 搜索引擎等，这些综合性搜索引擎是通过网页的完全索引搜索信息的，可提供的信息广且全；如果用户已明确搜索主题，检索目标为需要从总体或要全面地了解某一个主题内容，则可以使用目录型搜索引擎，如新浪搜索引擎，此类网站中的分类目录提供的内容很大程度上是由人工编辑整理的，系统性强、可靠性高。

2. 确定搜索途径，尝试不同的搜索方式

目前，大多数搜索引擎都提供两种搜索途径，一是分类浏览，二是关键词检索。根据不同的检索目的确定正确的检索途径，才能达到预期的检索效果。分类浏览适合于对信息所属知识类目有大概了解的用户，或是对某类信息想要有初步认识的用户，而关键词检索则对于细节性问题的查准率较高。利用分类浏览方式搜索信息的一般步骤是：首先使用搜索引擎或者其他方法浏览某一类别，得到一个大致的范围，然后在得到的搜索结果网址中选择一些具有代表性的网址，进入这些网站进行浏览，再从跟踪的网页中单击相关超链接，从而进一步发现更多的网址和信息。

3. 准确提炼，尽量使用搜索关键词而不是句子

搜索信息前，一定要明确搜索目标，并且尽量将大而泛的搜索需求转换为小而精的搜索目标。除此之外，当前大部分搜索引擎采用的是基于关键词匹配法的搜索技术，如果输入的

是句子，则搜索引擎会对输入的句子基于一定的规则进行分词，根据分词后的结果展开搜索。系统分词可能会导致所分词语与目标检索对象之间存在一定差距，为此，用户搜索时应尽量将自己的搜索目标抽取为若干关键词，再通过一定的逻辑关联词关联后进行搜索。

4. 适当使用搜索运算符

大多数搜索引擎允许用户使用布尔逻辑运算符 AND、OR、NOT 及与之对应的“＋”（限定搜索结果中必须包含的词汇）、“－”（限定搜索结果中不能包含的词汇）等逻辑符号提高搜索结果的精确度。

需要强调的是，布尔逻辑运算符在不同的搜索引擎中的含义和使用方法略有不同，为此，除非用户明确地知道运算符在某个搜索引擎中是如何使用的，否则可能会因使用错误而导致搜索结果有误。对于多数用户，不建议直接使用布尔逻辑运算符，可以选用搜索引擎自带的高级搜索功能实现（图 3-8）。

图 3-8　百度高级搜索示意

5. 巧用搜索小技巧

用户除了可以借助逻辑运算符形成逻辑表达式提升搜索准确性外，还可以使用一些其他符号或技巧提升搜索效率。

1）双引号运算符

使用双引号（西文符号）运算符可以实现整词检索。双引号括起来的部分，搜索引擎会认为是一个不可分割的最小单位，会作为一个不可分割的整体进行搜索。

2）书名号运算符

使用书名号可以实现书籍搜索和整词检索。在查找音乐、电影、电子书时，可以通过在检索关键词的两边加上“《》”大大提升检索的准确率，如图 3-9 和图 3-10 所示。此外，对于使用书名号括起来的内容，搜索引擎一般不会再进行拆分，会将其作为一个整体进行搜索。

3）限制查询范围

搜索引擎限制查询范围的能力越强，其就能越精准地找到所需的目标信息。如搜索粉色玫瑰花，可以在百度图片搜索框下输入“玫瑰花”，另外在“全部颜色”下拉选项中选择“粉色”选项，如图 3-11 所示。

图 3-9 整词检索示意(1)

图 3-10 整词检索示意(2)

图 3-11 限制查询范围搜索示意

6. 培养高效的搜索习惯

网络信息资源检索是一种需要通过大量实践才能发展起来的技能，真正的搜索者不会在搜索不到满意的结果时就马上离开搜索引擎，他们会思考，会回顾，并通过不断地练习和总结培养快速高效地找到所需内容的搜索能力。

3.2　目录型搜索引擎

目录型搜索引擎的工作流程与索引型搜索引擎的工作流程基本相似，通常由信息采集模块、信息组织模块和信息查询和展示模块三个基本模块组成，但其中的信息采集和信息组织主要由人工完成。常见的目录型搜索引擎有 Yahoo、Galaxy、搜狐、新浪、Open Directory Project、Infoseek、The WWW Vitual Library、BUBL LINK、AOL Search 等。本书以 Open Directory Project、搜狐、新浪为例简要说明目录型搜索引擎的基本使用方法。

3.2.1　Open Directory Project

1. 概述

Open Directory Project(开放式分类目录搜索系统，ODP)是互联网上最大的目录型分类检索系统，它是由来自世界各地的志愿者共同维护与建设的最大的全球目录社区，Google、Netscape、Dogpile、Thunderstone、Linux 等搜索引擎都在使用 ODP 的目录体系，用户可以通过 URL“http://www.odp.org/”进行访问，其主页面如图 3-12 所示。

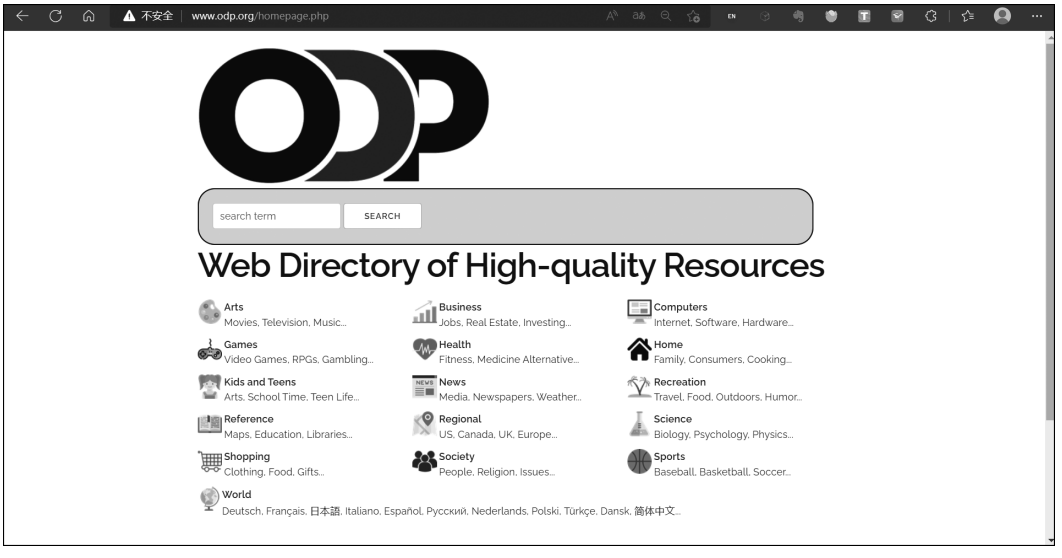

图 3-12　目录型搜索引擎示意

Open Directory Project 默认以领域为类别进行内容分类，分为 Arts、Business、Computers、Games、Health、Home、Kids and Teens、News、Recreation、Reference、Regional、Science、Shopping、Society、Sports 15 个类别，支持德语、法语、日语、中文等 80 多种语言。

2. 检索功能

Open Directory Project 支持分类目录检索和关键词检索两种检索方式，即用户可以通过单击网站首页中提供的类目名称进行纵深查看，也可以在主页的检索框中输入检索词进行目标内容检索。

3.2.2 新浪

1. 概述

新浪网是一家服务于中国及全球华人社群的网络媒体公司，成立于 1998 年 12 月，为全球用户提供全面及时的中文资讯、多元快捷的网络空间以及轻松自由地与世界交流的先进手段。

新浪网包含分频道的中文新闻和内容、社区和社交服务以及基于新浪搜索和目录服务的网络导航能力，同时通过移动应用，如新浪新闻、新浪财经和新浪体育以及移动门户提供针对移动端用户订制的新闻资讯及娱乐内容。新浪网的重要频道包括新浪新闻、新浪财经、新浪科技、新浪微博。用户可以通过 URL“https://www.sina.com.cn/”访问新浪网，其首页如图 3-13 所示。

图 3-13 新浪首页

2. 检索功能

与 Open Directory Project 一样，新浪也提供了目录检索和关键词检索两种检索方式，检索时也支持使用逻辑运算符。

1）分类检索

分类检索是指从分类目录首页开始，按照树状主题分类逐层单击查找所需信息资源的

一种检索方法。

2）关键词检索

关键词检索是利用所需信息的主题词(关键词)进行信息内容查询的一种方法。使用关键词方法进行检索时，只要在新浪分类目录页面的检索框中输入关键词，然后在资源列表中选择查询的资源类型(网页、MP3、新闻标题、图片等)后单击“搜索”按钮即可开启检索过程(图 3-14)。

图 3-14　使用关键词进行内容搜索

3. 检索结果

新浪网站的检索结果包括网页、新闻、视频、音乐、图片、地图、网址等多种形式，用户可以在检索结果页面中通过单击检索结果列表的超链接进入某一检索结果进行查看(图 3-15 和图 3-16)。

图 3-15　新浪检索结果页面示意(1)

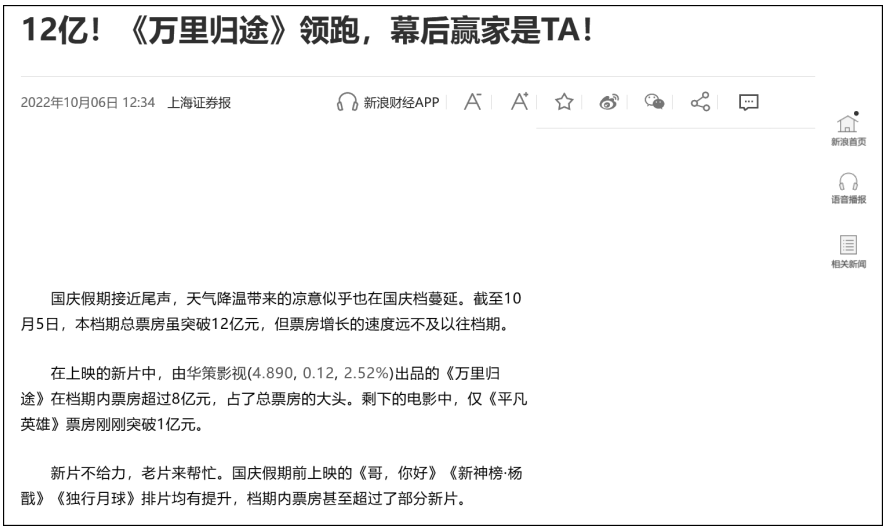

图 3-16 新浪检索结果页面示意(2)

3.2.3 搜狐

1. 概述

搜狐公司成立于 1996 年，1998 年 2 月推出了中国第一个全新的中文网络资源目录系统。站点的全部内容采用人工分类编辑，并充分考虑用户的查询习惯，确保了分类体系和网站信息的人性化特点以及网络资源目录的准确性、系统性和科学性，是目前中国影响力最大和国内用户首选的目录型网络资源检索工具，全面收录了各式各样的网络资源，其目录导航式搜索引擎完全由人工完成，大类设置采用了按学科和按主题相结合的方式。用户可以通过“https://www.sohu.com/”访问搜狐网，其首页如图 3-17 所示。

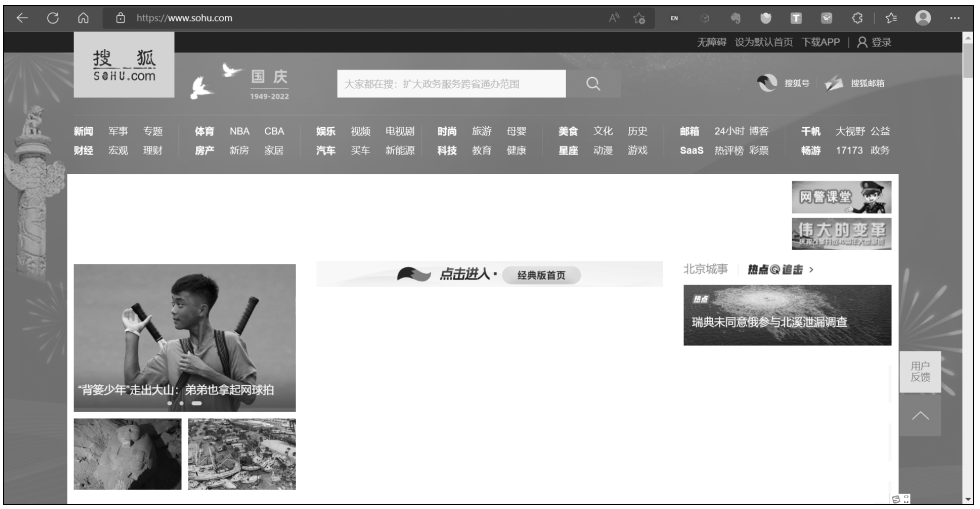

图 3-17 搜狐网首页示意

2. 检索功能

与上述两种目录搜索引擎类似，搜狐搜索引擎主要有分类目录导航检索和关键词检索两种检索方式。

1）分类目录导航检索

分类目录导航检索是按照信息所属的类别，使用分类目录，层层单击进入查找所需的信息，查询结果会提供有关该主题的全部网站。因此，使用分类目录导航检索的关键是要考虑清楚待查询信息的所属类别。

2）简单检索

实施简单检索时，用户只需要在分类目录主页的检索框中输入查询的关键词或者关键词的逻辑组合，就可以检索到相关的信息。

3. 检索结果

搜狐搜索引擎会根据分类类目及网站信息与关键词（组）的相关程度排列出相关的类目和网站，相关程度越高，排列位置越靠前（图 3-18）。

图 3-18　搜狐检索结果页面示意

4. 搜狗搜索引擎

2004 年 8 月，搜狐推出了第三代互动式搜索引擎——搜狗（http://www.sogo.com），它采用人工智能新算法分析和理解用户可能的查询意图，对不同的搜索结果进行分类，对相同的搜索结果进行聚类，在用户查询和搜索引擎返回结果的人机交互过程中，引导用户更快速、更准确地定位自己关注的内容。该技术已全面应用到搜狗网页搜索、音乐搜索、图片搜