第 3 章　搜索引擎介绍

搜索引擎是一种非常有效和易于使用的互联网信息检索工具，利用现代互联网搜索引擎本身就可以直接检索到各种各样的信息资源，同时掌握搜索引擎的使用方法也可以为我们提供利用其他专业信息检索工具的必备技能。

自从 1994 年问世以来，搜索引擎逐渐成为人们获取互联网信息资源的主要方式，相关搜索引擎网站也逐渐成为 Web 用户使用互联网时的首选访问站点之一，另外搜索引擎和实时通信、电子邮件等服务也已经成为当今各大门户网站用来吸引用户访问的三大主要方式。2015—2020 年，中国搜索引擎市场规模从 707.5 亿元增长到 1204.6 亿元，年复合增长率 11.23%。截至 2020 年 12 月，我国搜索引擎用户规模达 7.70 亿，占网民整体的 77.8%；手机搜索引擎用户规模达 7.68 亿，占手机网民的 77.9%[①]。这也充分说明了现代移动互联网信息资源检索利用需求的广泛存在。

利用搜索引擎获取互联网信息资源也是网络用户常见的访问操作。本章首先主要介绍搜索引擎在国外和国内的发展历史和现状，通过对此问题的了解，有助于我们认识搜索引擎的特点，如为什么搜索引擎是现在这个样子，为什么该这样使用搜索引擎，这些都是我们需要回答的问题；其次，本章还简单讨论搜索引擎的基本原理，以此来加深读者对搜索引擎的了解，从而更好地帮助读者使用搜索引擎。事实上，有时候我们会觉得搜索引擎不是很好用，其中的原因很复杂，但是有一点是肯定的，如果我们越了解搜索引擎，我们就越能有效地使用搜索引擎。

3.1　搜索引擎的发展

搜索引擎这个名称比较古怪，来自它的英文名称：Search Engine，言下之意，它是一种检索信息的发动机。可以说，整个搜索引擎的发展历史就是互联网的发展历史，因为互联网用户中一直存在着从大量网络信息中获取自己所需信息的需求，而且这种需求随着互联网的快速增加而日渐迫切。

按照检索技术的发展过程，搜索引擎的发展经历过三个主要阶段：第一阶段时间跨度大致为 1990 年到 1998 年，这个时期的搜索引擎主要着力于解决如何快速有效地从大量网页中获取较为完整全面的搜索结果，开始使用爬虫等信息收集方式和使用 Web 目录等信息

① 华经情报网 | 2020 年中国搜索引擎行业发展现状与背景研究. https://baijiahao.baidu.com/s?id=1700797452918714573.

组织方式，代表性的搜索引擎有 Altavista 等；第二阶段时间跨度大致为 1998 年到 2004 年，此时的互联网规模已经相当庞大，检索结果是否完整似乎已经没有太大意义，相反，搜索引擎开始努力地在命中网页的质量及其内容相关度的排序上来提高用户的满意度，基于网页链接分析的算法逐渐被各大搜索引擎广泛采用，Google 就是典型的代表；第三阶段时间跨度为 2004 年至今，各大搜索引擎不断应用先进的技术来改进功能，如增加多媒体信息查询功能、个性化搜索引擎功能等。

尤其是近年来随着机器学习等人工智能方法的不断应用，今天的搜索引擎在用户查询意图理解和结果呈现等方面，都比以前取得了极为明显的进步，整体检索效果越来越好。

3.1.1 国外搜索引擎的发展历史

可以说，如果没有互联网，就没有搜索引擎。但是，在互联网出现之前，很多人所提出的思想和见解却深深地影响了现代搜索引擎的出现和发展。

1945 年，万尼瓦尔·布什（Vannevar Bush）在《大西洋月刊》（*The Atlantic Monthly*）上发表了一篇重要的文章 *As We May Think*（中文译名为《诚若所思》）。虽然那个年代还没有计算机，但是在这篇文章中，作者提到了类似于超文本的思想，同时还指出未来的世界会出现一种独立于人类大脑以外的知识扩展体（Memory Extension），该物体具有无限大的虚拟空间，可以很好地扩展，同时还能提供有效的信息获取方法，作者称为 Memex，如图 3.1 所示。

图 3.1 《大西洋月刊》上的《诚若所思》一文（截取于 2022-4）

万尼瓦尔·布什大胆地预测了未来人类可能会面临的信息处理困境，这是他书中的原话："The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present day interests, but rather that publication has been extended far beyond present ability to make real use of the record."含义为："我们所面临的难题看起来并不是我们从当前兴趣的深度和广度出发发表了不恰当的观点，而是我们现有能力根本不

足以利用这些发表的内容。”

然而，万尼瓦尔·布什并没有在技术上给出实现。20 世纪六七十年代，美国康奈尔大学(Cornell University)的杰勒德·沙顿(Gerard Salton)教授在信息检索技术方面做出了很多贡献，很多技术直到今天还在搜索引擎中得到广泛的应用，如空间向量模型、词频、倒文档频率和相关度反馈等技术，他甚至还研发了 SMART 信息检索原型系统。

3.1.1.1　早期的搜索引擎

相对于其他类型的信息服务类型，互联网使用 WWW 服务的时间比较晚，所以早期的互联网并不存在类似于今天的网页搜索引擎，但是仍然出现了很多类似的网络文件检索工具。

1. Archie

1990 年，加拿大蒙特利尔的麦吉尔大学(McGill University)的三位学生 Alan Emtage、Peter Deutsch、Bill Wheelan 发明了 Archie，据称这个名称来自“Archive(档案文件)”的缩写。当时的互联网已经可以提供诸如 FTP 等文件下载服务，然而用户却缺乏一种直接检索 FTP 文件所在地址的工具。Archie 恰恰可以自动索引互联网上匿名的免费 FTP 文件信息，并提供一种根据文件名称检索文件所在 FTP 地址的方法。因此，Archie 被称为现代搜索引擎的祖先。然而，客观地讲，它并非一个真正的搜索引擎。原因有两个：一是它只能检索 FTP 文件资源，并不能获取诸如网页等其他类型的文件资源，因此它其实是世界上第一个 FTP 搜索引擎；二是它没有机器人(Robot)程序，不能像今天的搜索引擎那样快速有效地抓取互联网上的网页文章内容，相反，它使用的是一个基于脚本的文件名称收集器，并通过正则表达式来匹配用户查询与文件名称来实现检索，并通过文件列表的方式提供信息检索结果。

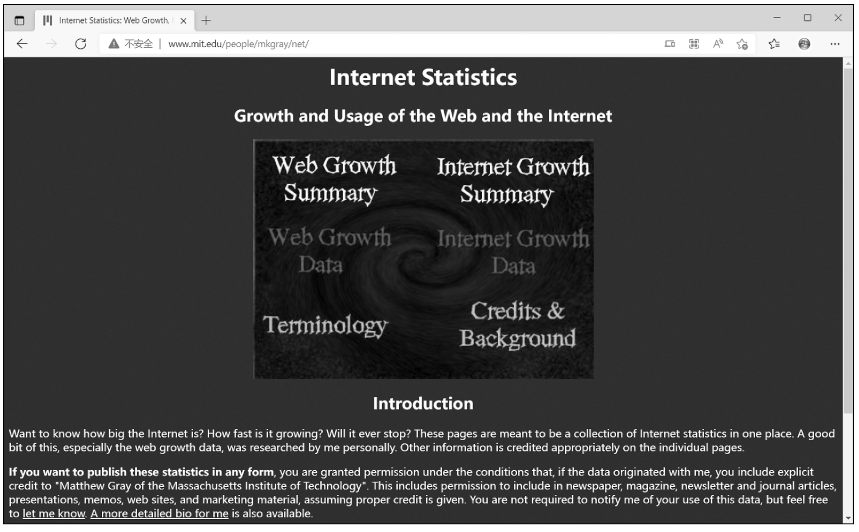
2. World Wide Web Wanderer

现代搜索引擎之所以可以检索网页信息，是因为它有一个被称为机器人(Robot)的程序，所谓机器人程序是指可以连续不断地自动获取互联网上所有网页信息的一种程序。World Wide Web Wanderer 其实并不能算是搜索引擎，它只是世界上第一个机器人程序，由美国麻省理工学院(MIT)的马泰·格雷(Matthew Gray)于 1993 年 6 月开发。它通过自动遍历网络的方法来统计互联网上的服务器数量，所以可以追踪互联网的发展规模，直至后来还可以专门用于获取互联网上网页的 URL 信息。所有遍历得到的信息都被存入自己的数据库，名字叫 Wandex。由于当时对于性能考虑得不是太多，所以这个机器人程序可以在一天内连续对同一网页进行多达几百次的遍历，因而会造成被遍历系统性能的严重下降。虽然马泰·格雷很快修复了这一问题，然而这次事故却给人们带来一个疑问：我的站点如果被别的机器人程序遍历，是不是会引起性能的下降？直到今天，搜索引擎在机器人设计方面仍然存在着这样的挑战。

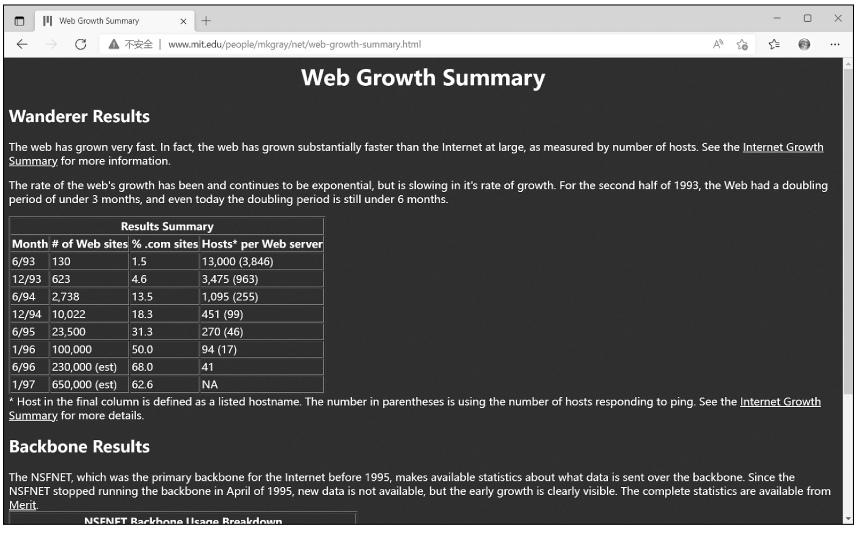
今天依然可以看到 MIT 网站上有关的网络统计历史数据，网址为 http://www.mit.edu/people/mkgray/net，如图 3.2 所示。

3. Veronica 和 Jughead

之所以把这两个搜索引擎放在一起，是因为它们的功能很相似，出现的时间也很接近。



(a)



(b)

图 3.2 MIT 网站关于 World Wide Web Wanderer 的信息（截取于 2022-4）

Veronica 由美国内华达大学（University of Nevada）的系统计算服务小组（System Computing Services Group）于 1991 年开发。和 Archie 不同的地方在于，Veronica 只对存在于 Gopher 上的普通文本文件进行查询。随后出现的 Jughead 也具有类似的作用，据称这个名称来自“Jonzy's Universal Gopher Hierarchy Excavation and Display”（Jonzy 的统一 Gopher 层次性挖掘和显示工具）。有趣的是，后人常常把 Archie 称为搜索引擎之父，而把 Veronica 称为搜索引擎之母。

不过，这些工具都已经不复存在，然而人们依然可以在互联网上看到一些遗留下来的服务，如图 3.3 所示。

4. ALIWEB

ALIWEB 是个划时代的搜索引擎，借助它人们首次可以对 WWW 网页进行全文查询。

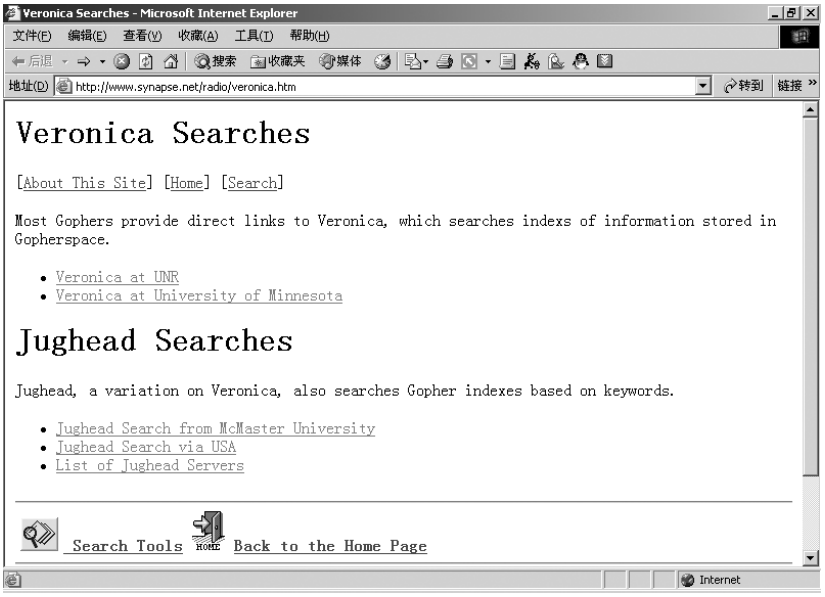


图 3.3　某站点展示的几个大学所提供的 Veronica 服务(截取于 2007-9)

它是由马汀·考斯特(Martijn Koster)于 1993 年 10 月开发的,名称含义是“类似于 Archie 的 Web 索引”(Archie-Like Indexing of the Web),它相当于 Archie 的 Web 版本。但是,ALIWEB 没有自己的机器人程序,相反它却要求愿意被 ALIWEB 收录的网站网管主动提交自己网站的网页索引信息,这样做的好处在于克服了机器人程序带来的带宽消耗,同时网管可以自主地描述网页内容。但缺点也是显而易见的,很多网管并不知道如何来做这个事情,甚至都不知道是否需要这样做,所以 ALIWEB 的网页数据库规模一直不大。ALIWEB 的网址为 http://www.aliweb.com,今天依然还在运行,主页界面如图 3.4 所示。

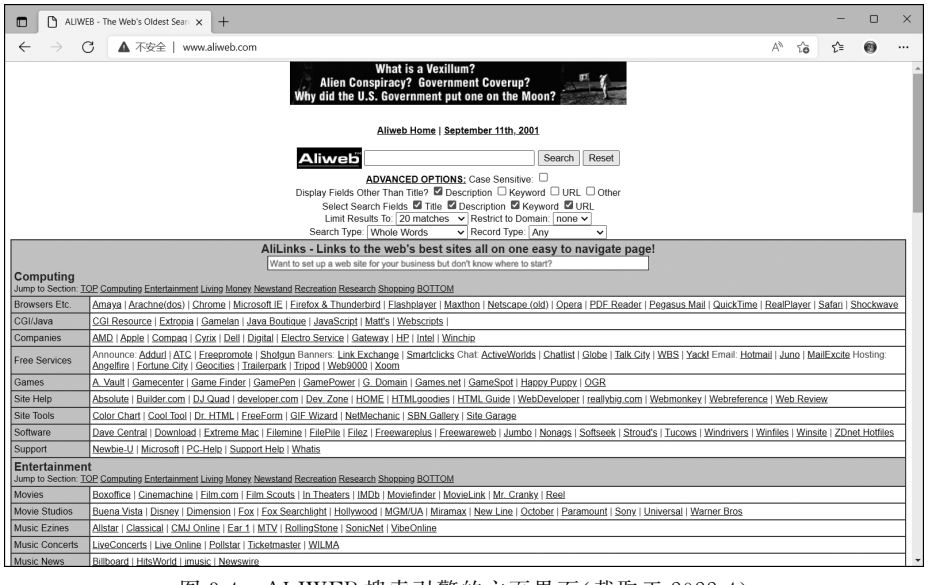


图 3.4　ALIWEB 搜索引擎的主页界面(截取于 2022-4)

虽然它很古老,但是它所提供的检索功能却非常强大,例如它在引号中提供的“子串部

分匹配(Substrings)”功能连 Google 和百度都不能提供(它们只能提供全词匹配)。

后来,马汀·考斯特并没有停止对搜索引擎技术的研究,他还成为机器人拒绝协议(Robots Exclusion)标准的主要设计者。通过机器人拒绝协议,网站可以告知搜索引擎哪些信息可以被搜索引擎机器人程序遍历,而哪些不可以遍历,据此人们就可以更好地在信息公开性和保密性之间取得一种平衡。这个协议现在已经成为现代搜索引擎的标准之一。

马汀·考斯特的个人主页网址为 http://www.greenhills.co.uk/,如图 3.5 所示。

图 3.5 马汀·考斯特的个人主页(截取于 2022-4)

3.1.1.2 基于爬虫的搜索引擎

爬虫(Crawler)是从搜索引擎机器人程序发展而来的。虽然两者在功能上很相似,但是爬虫程序却可以通过分析遍历来的网页中含有的网页链接信息,自动获取下一步需要遍历的网页,这个过程可以自动地持续地进行下去。爬虫是个非常形象的称呼,也有人称之为蜘蛛(Spider),它们都是一个意思,Web 单词本来的意思就是“蜘蛛网”,因此它们真像互联网上的蜘蛛爬虫,自由地跑来跑去,抓取所能获得的各种网页信息①。

1994 年在搜索引擎发展历史上发生了很多具有里程碑意义的事件,各种基于爬虫的搜索引擎纷纷出现,彻底改变了人们获取互联网信息的习惯。

1. JumpStation、The World Wide Web Worm 和 RBSE

最初产生的著名搜索引擎有三个:一是英国苏格兰大学(Scotland University)开发的 JumpStation,它可以自动收集网页的标题等信息,但是随着网页数量的增加,该搜索引擎却不能很好地适应这种变化,性能变得很差,最终停止了运行;二是美国科罗拉多大学(University of Colorado)的奥利弗·麦克布莱(Oliver McBryan)开发的 The World Wide

① 爬虫程序要想抓取所有的互联网网页信息,需要有个假设前提,那就是互联网的所有网页都相互链接。事实上这并不可能。不过,探讨这个问题意义不是很大,尤其在互联网网页数量规模已达万亿级的今天,人们更关心的是能否快速地找到一些最想要的信息资源而非全部的信息资源。

Web Worm，字面意思是“万维网蠕虫”，它可以自动收集网页的标题和 URL 等信息，而且它也是第一个解析超文本信息的搜索引擎；三是美国航空航天局（NASA）开发的 RBSE，意思是基于存储库的软件技术设备（The Repository-Based Software Engineering）。它是第一个能够索引 Web 网页正文的搜索引擎，也是第一个能够在搜索结果排列中引入查询词语相关度概念的搜索引擎。与前两种搜索引擎不同，它不再简单地只根据找到匹配网页信息的先后次序来排列搜索结果，而是利用网页链接分析重新设计新的结果网页排序算法，因此可以把用户最想要的相关网页放置在搜索引擎结果的最前面。

现在这些搜索引擎都早已停止了服务，但是后来的搜索引擎基本上都采用了基于爬虫的网页信息获取方法。

2. Excite

Excite 是一个非常有代表性的搜索引擎，它是由美国斯坦福大学（Stanford Unversity）6 名本科生在 1993 年 2 月研发的一个项目 Architext 发展而来。最初这些学生认为可以通过对网页中的词语关系进行统计分析来提高搜索的效果，因此他们在引入风险投资后就研发了 Architext 系统。到了 1993 年中期，他们发布了一个供网络管理员可以在自己网站上使用的查询软件版本，称为 Excite for Web Servers。到 1999 年，Excite 被一个名叫@Home 的宽带运营商以 65 亿美元收购，因此搜索引擎也改名为 Excite@Home。从此，Excite@Home 开始侧重于宽带市场，在搜索引擎方面也就没有更新的技术出现。好景不长，Excite@Home 于 2001 年 10 月破产，2002 年 5 月被 InfoSpace 公司以 1000 万美元收购。今天，Excite 仍然还在运营，不过它已经改用 Dogpile 来提供元搜索引擎服务。Excite 主页网址为 http://www.excite.com，检索界面如图 3.6 所示。

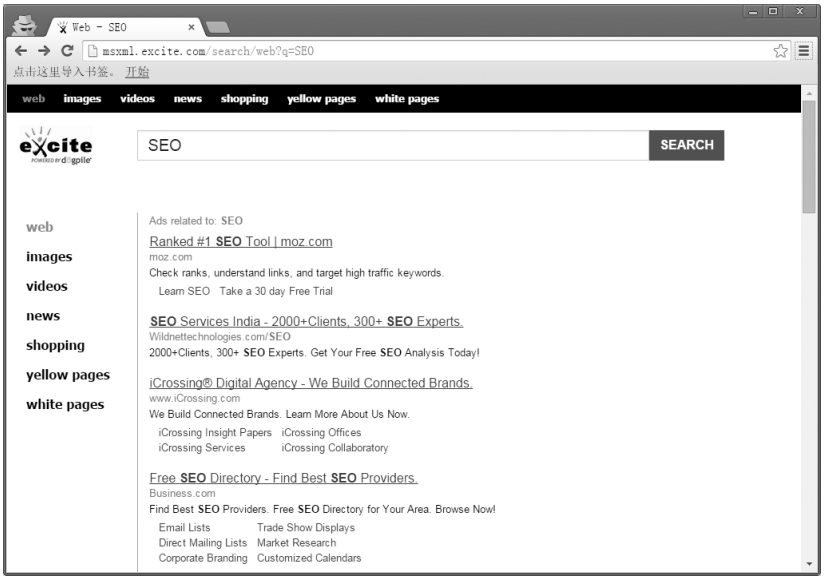


图 3.6　在 Excite 搜索引擎检索“SEO”的相关结果（截取于 2015-3）

其中，它还专门提供了黄页（Yellow Pages）检索和白页（White Pages）检索，前者是指电话号码检索，后者是指电子邮箱检索。

Excite 搜索引擎有两点非常引人注目：一是在商业上，它最早提出“免费让人搜索，用

广告收入来补贴”的搜索引擎盈利模式，这在当时是比较新的理念；二是在技术上，Excite 一直以概念搜索闻名。所谓概念搜索，是指搜索引擎可以理解用户检索词语的语义含义，并进行自动语义扩检[①]来推荐更多的查询内容。当然，受限于技术的发展，这种概念检索的功能并非十分强大。图 3.7 展示了在 Excite 中查询“Apple”的界面，在窗口的右边显示了一组扩展的查询词语，如 Apple Store(苹果用品商店)，甚至还有 Banana 等水果类词语。

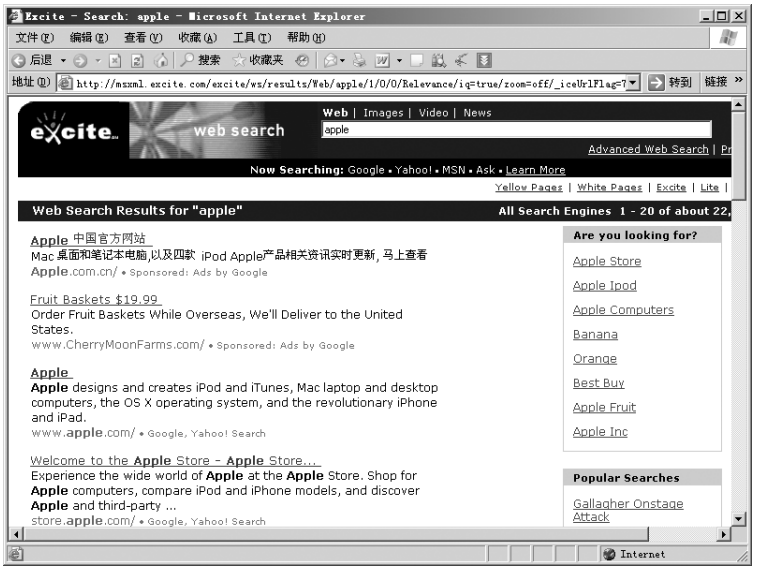


图 3.7　Excite 所提供的概念检索(截取于 2010-3)

3. WebCrawler

WebCrawler 是美国华盛顿大学(University of Washington)计算机科学系的学生布赖恩·平克顿(Brian Pinkerton)于 1994 年 4 月 20 日创建，虽然它最早只是从一个非正式学术研讨会上的小型项目发展而来，最初亮相时只包含来自 6000 个服务器的网页内容，但它却是世界上第一个可以对遍历网页的全部文字内容进行索引和检索的搜索引擎[②]。在此之前，搜索引擎只能提供网页 URL 和网页摘要来供用户查看查询结果，其中网页摘要一般来自人工评论或者是由程序自动抽取网页正文的前若干词语组成。

1995 年，美国在线收购了 WebCrawler。1997 年，Excite 又把 WebCrawler 买走，此时的美国在线就开始使用 Excite 作为它自己搜索项目 NetFind 的技术提供商。随着 Excite 的风光不再，今天的 WebCrawler 已改用 Dogpile 来提供元搜索引擎服务，主页如图 3.8 所示。

4. Lycos

Lycos 的名字来自拉丁文单词 Lycosidae(狼蛛)，狼蛛和一般蜘蛛最大的区别就是不结网，而是直接追随猎物捕食。这个形象有力的名称表达了 Lycos 遍历网页的强大能力，事实上，它也是搜索引擎历史上的代表作之一。它由美国卡内基-梅隆大学(Carnegie Mellon

① 扩检是指扩展检索，意即对当前检索词语的语义进行分析，找到更为一般的或者与此相关的其他检索词语提供给用户做进一步查询时使用。

② 在当时，强大的全文索引能力引发了巨大的访问流量，据称当时的华盛顿大学校园网络几乎因此崩溃。

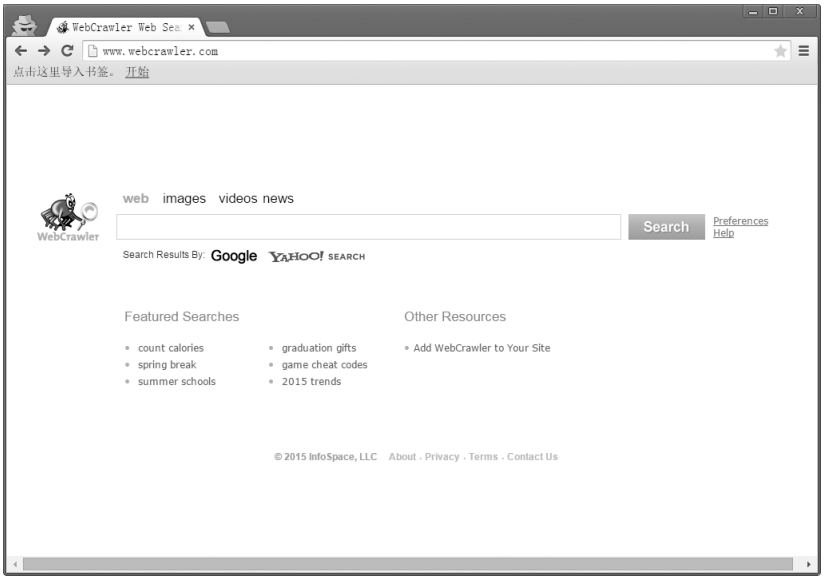


图 3.8　WebCrawler 搜索引擎的主页界面(截取于 2015-3)

University)的博士生迈克尔·墨登(Michale Mauldin)于 1994 年 7 月在匹兹堡创建,和其他美国搜索引擎不太一样的地方在于,它是早期唯一诞生于美国东部的搜索引擎,而其他的搜索引擎则都在西部的硅谷创建。

从技术上看,Lycos 能够提供网页结果排序、查询词语的前缀匹配、邻近位置词语查询和自动网页摘要等一系列功能。在 1994 年 10 月,用户通过当时最为流行的航海者浏览器查询"Surf[①]"相关结果时,Lycos 是排名第一的搜索引擎结果。正如 Lycos 名字暗示的那样,Lycos 遍历网页的能力非常强,这是它的最大特点,而这一点在互联网刚开始发展的年代无疑非常吸引人。据报道,1994 年 7 月 20 日,Lycos 就可以遍历 54 000 篇网页,到了 1994 年 8 月则达到 39.4 万篇,1995 年 1 月达到 150 万篇,1996 年 11 月更达到6000 万篇网页,超过了当时任何一款搜索引擎所能收集的网页量。

但是,客观地讲,Lycos 的搜索引擎技术并不是最好。不过,Lycos 在商业上却做得不错,如很早就开始投资做社区网站,网络广告也经营得不错,这些成功掩饰了 Lycos 技术的不足。Lycos 后来似乎意识到了这一点,它收购了一家广受好评的搜索引擎 Hotbot,而 Hotbot 后台使用的是 Inktomi 搜索引擎的技术,Lycos 希望通过此次收购来提升自己的技术水平。但是,这也使得 Lycos 一直需要维持着两个搜索引擎的技术平台。可能是 Inktomi 的技术确实比较先进,直到最后它全面改用 Inktomi 的搜索技术。不过,由于受到 Yahoo!和 Google 的竞争,Lycos 逐渐衰落,最终在 1999 年 4 月停止了服务,改由 Fast 搜索引擎来提供服务,主页如图 3.9 所示。

5. Infoseek

Infoseek 也诞生于 1994 年。Infoseek 的起点比较高,因为它所使用的搜索技术来自于美国马萨诸塞大学(University of Massachusetts),而在全美高校中,马萨诸塞大学的信息

① Surf 是指冲浪,这里意指所谓的网上冲浪,通常上网的用户也被称为"冲浪者(Surfer)"。

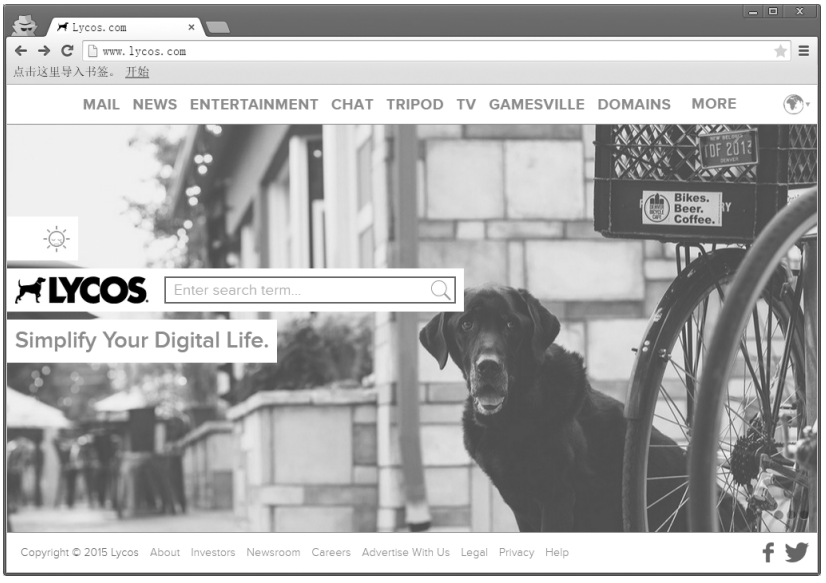


图 3.9 Lycos 搜索引擎的主页界面(截取于 2015-3)

检索技术名数一流。但是在设计完成之后,设计师还是发现它无法适应如此多的互联网网页处理要求,因此聘请一位名叫威廉·张(William I. Chang)的中国台湾设计师进行了改进,改进后的技术平台被称为 Ultraseek。该平台较前者而言,不仅在处理速度上,而且在检索结果的相关度方面,都比较优秀。事实上,后来的 Infoseek 也确实因为相关度算法好而闻名。同时,它还允许网站管理者提交自己的网页来进行实时索引,该项功能非常吸引人,不过,Infoseek 也同时成了搜索造假者(Search Spammer)[①]的天堂,很多网站管理者利用此项功能来恶意提升自己网站的搜索结果排名和被搜索的次数。

Infoseek 不断增强用户界面的友好性,同时提供大量附加服务以吸引用户使用。最为重要的是,1995 年 12 月,Infoseek 连说服带花钱,让网景(Netscape)公司不再使用 Yahoo! 作为默认的搜索服务提供商,也就是说,当用户单击航海者浏览器的搜索按钮时,默认弹出 Infoseek 的搜索引擎。但是,随着 1999 年被迪士尼(Disney)公司收购,Infoseek 最终沦落为 Go.com 网站做娱乐方面的索引和搜索服务,从此在技术方面的革新越来越少。在 2001 年 2 月,Infoseek 终于停止了自己的搜索引擎,改用 Overture 的搜索服务。有趣的是,百度创始人李彦宏也曾经在 Infoseek 从事过技术工作,但于1999 年回国创立了百度。更为有趣的是,那个当时改进 Infoseek 的工程师 William I. Chang 就在工作中认识了李彦宏,并于 2006 年 12 月 6 日加盟了百度,成为百度首席科学家。Infoseek 的网址为 http://go.com,现在已经完全关闭了搜索服务,原先的搜索引擎主页如图 3.10 所示。

6. AltaVista

AltaVista 可以被看成早期搜索引擎中的 Google,它不论是在软件功能上还是硬件条件上都达到了那个时代的顶峰,在很多方面对现代搜索引擎都产生了深刻的影响。

---

① 所谓搜索造假者,是指一些恶意的网站管理者通过故意修改网页内容来设法提升自己网页在搜索引擎命中结果中的位置,或者使得用户在输入一些常见词语进行检索的时候,可以很方便地找到那些网站管理者自己的网页。显然,这种行为并不公平,而且会极大地影响搜索引擎自身的声誉。

图 3.10　Infoseek 搜索引擎的主页界面(截取于 2010-3)

AltaVista 由美国数字设备公司(Digital Equipment Corporation，DEC)研发，于 1995 年 12 月面世。从硬件条件来看，DEC 公司本身就是生产计算机设备的公司，凭借 DEC 强大的 Alpha 芯片运算能力，AltaVista 可以运行在当时最为先进的计算机上，因此运行速度非常快。从软件功能上看，AltaVista 搜索引擎的功能也非常多，如 AltaVista 第一个允许用户使用句子来进行自然语言查询，第一个支持和实现了布尔查询，能对不同格式的文档、多媒体信息甚至多国语言的网页进行查询。同时，AltaVista 还是第一个允许用户自主增删网页索引信息的搜索引擎，更新的信息最快可以在 24 小时内上线。另外，AltaVista 还能查询有链接指向某个特定网页的所有其他网页，该功能称为链入检查(Inbound Link Check)，这个功能有助于网站管理者了解自己站点受人关注的程度，显然，这种被其他网页建立的链接越多，自己网页的受欢迎程度相对也就越高。在界面上，AltaVista 还提供了大量的易用帮助提示信息以方便用户使用。

1997 年，AltaVista 发布了一个图形演示系统 LiveTopics，它采用一个图形化的界面来整理搜索引擎的返回结果，从而方便用户找到所需内容，界面如图 3.11 所示。

这些技术都令人刮目相看。然而，由于管理混乱和竞争者的不断增多，进入 21 世纪以后 AltaVista 逐渐走了下坡路。2003 年 2 月 18 日，Overture 收购了 AltaVista，随后 Yahoo!又收购了 Overture，AltaVista 因此成为了 Yahoo!搜索系统的实验平台，也为 Yahoo!推出自己的搜索引擎打下了必要的技术基础。AltaVista 的网址为 http://www.altavista.com，今天只能打开 Yahoo!的搜索界面，原先的搜索引擎主页如图 3.12 所示。

7. Inktomi

Inktomi 的正确念法是 Ink-to-me，它来自美洲印第安人传说中的一个蜘蛛魔法师，据说给人类带来了文化和知识。Inktomi 是由美国加州大学伯克利分校(University of California，Berkeley)计算机教授埃里克·布鲁尔(Eric Brewer)和他的博士生保罗·高瑟(Paul Gauthier)于 1996 年 1 月创建。他们是研究并行处理的专家，也希望以 Inktomi 来证

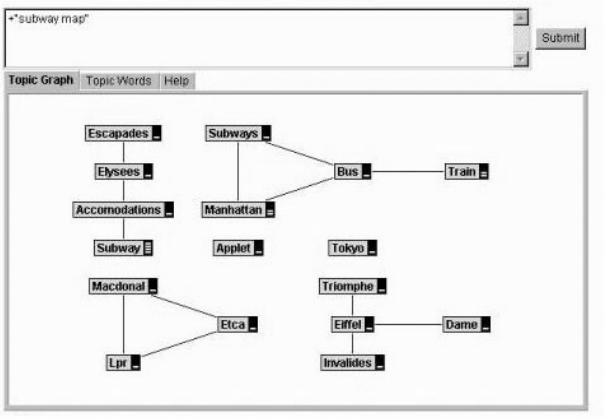


图 3.11　LiveTopics 系统的界面

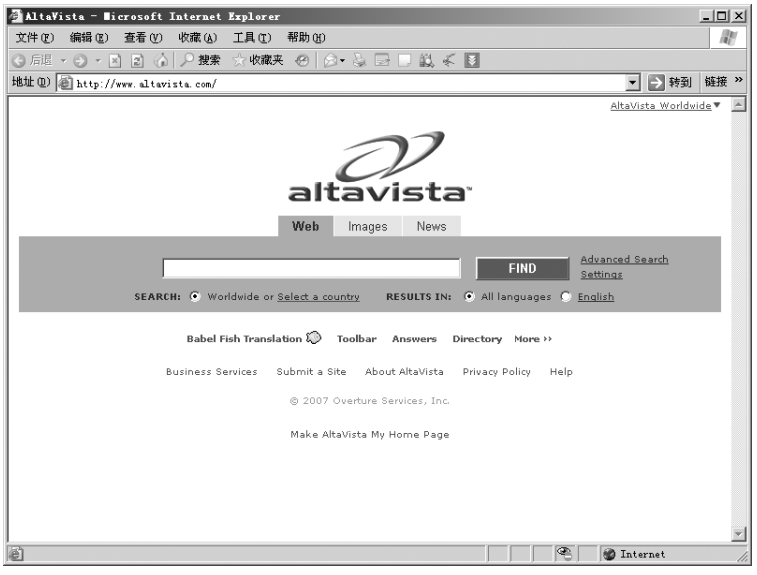


图 3.12　AltaVista 搜索引擎的主页界面(截取于 2015-3)

明他们所提出的并行算法是有效的。但是，此时的互联网搜索引擎已经群雄并起，要想和它们正面交锋，难度很大。所以，Inktomi 创建者决定只做技术提供商，并在 1996 年5 月 20 日开始为 Hotbot 提供服务。事实证明，Hotbot 很受欢迎，它声称每天能遍历1000 万篇以上的网页，同时还大量运用 cookie 来储存用户的设置信息以提供个性化的查询服务。在商业运行模式上，Inktomi 还提出了很多直到今天依然还在沿用的概念，如 Search Submit(付费提交)、Index Connect(付费索引)、Web Portal Solution(Web 门户解决方案)和 Enterprise Search(企业搜索)等。到了 1999 年，Inktomi 达到了鼎盛，成为诸如 Yahoo!和微软 MSN 搜索引擎在内近一百多个大网站的搜索后台技术提供商。

随后，Hotbot 被 Lycos 收购，Yahoo!也转用 Google 作为搜索技术提供商，这对 Inktomi 是个巨大打击，不断流失的客户和影响力开始使得 Inktomi 走向下坡路。Inktomi 于 2002 年 12 月 23 日还是被当年抛弃它的 Yahoo!以低价收购。在此之前，Yahoo!一直在

使用 Altavista 作为后台技术提供商。现在该服务已经关闭[①]，关闭前的最后主页如图 3.13 所示。

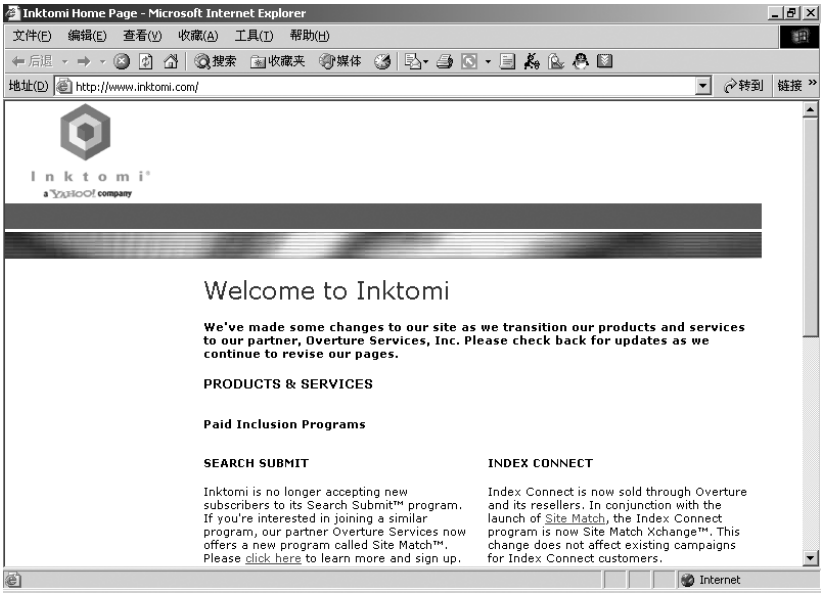


图 3.13　Inktomi 搜索引擎的主页界面(截取于 2007-9)

3.1.1.3　基于分类目录的搜索引擎

前文所述的搜索引擎多是采用爬虫方式来获取网页信息，同时在检索界面上多采用输入检索关键词的方式来获取网页结果，通常我们称这种方式为全文检索(Full-text Retrieval)，因为网页只要在任何位置上含有用户的检索词语就可以被命中。与此相对的，还有另外一种有效的信息检索形式，那就是 Web 目录(Web Directory)，也称为“分类目录”或者“网页目录”。它采用层次性的目录组织体系，将所收集的网页分门别类地归入不同的子目录，用户按照目录提示可以逐层定位找到自己所需的内容。采取此类方法实现的搜索引擎和信息检索站点也有很多。

1. Virtual Library

发明 WWW 访问方式的蒂姆·伯纳斯·李于 1991 年利用 WWW 方式组织过一个 Web 目录站点，称为虚拟图书馆(Virtual Library)，于是它被看成世界上最早的 Web 目录站点。不像一般的商业站点，这个站点是由一群志愿者维护的，志愿者分别根据自己所了解的学科知识领域给出相应目录下的推荐网页结果，所以体系不大，但是收录的网页质量却较高，主页如图 3.14 所示。

2. Galaxy

1994 年 1 月，Galaxy 在美国得克萨斯大学(University of Texas)创建，最早的名称是

① Inktomi 的最终失败从一方面反映了搜索引擎必须要正视的问题，那就是究竟应该直接面对用户树立品牌还是甘当幕后英雄。事实证明，要想取得市场的成功，搜索引擎必须及时转型，尽快走到台前。后来的 Google 和百度则采取了正确的转型路线，成为现代搜索引擎的巨头之一。

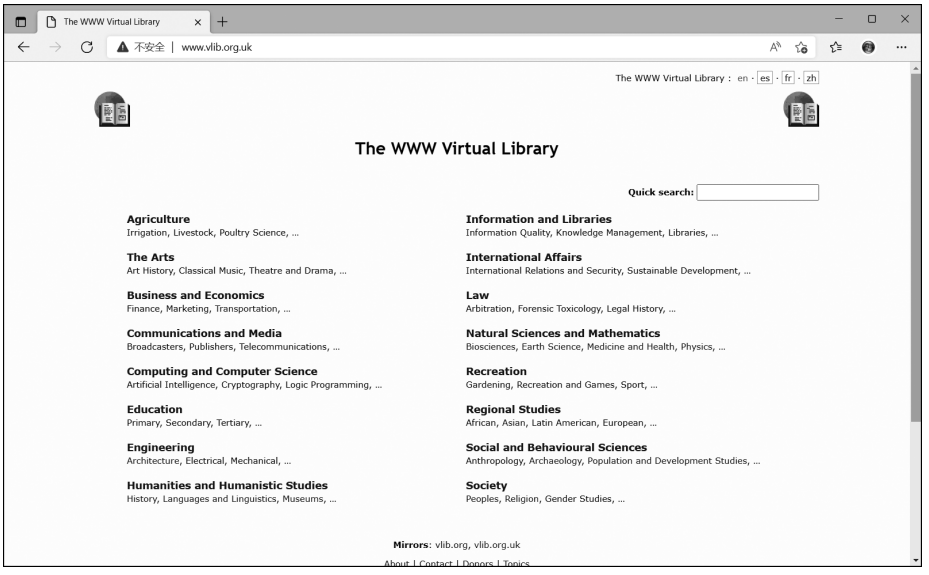


图 3.14　蒂姆·伯纳斯·李创立的虚拟图书馆(Virtual Library)(截取于 2022-4)

EINet Galaxy。在创建之初,Galaxy 主要面向电子商务的大型目录指南服务。1995 年 4 月,Galaxy 由一个研究项目转变为一个商业项目,1997 年被网络安全公司 CyberGuard 收购,1998 年 9 月,CyberGuard 又把 Galaxy 卖给美国健康网(AHN.com),1999 年 5 月,Fox/News 公司介入 Galaxy。直到 2000 年 5 月,几经变故的 Galaxy 终于成为一个独立的站点,由 TradeWave 公司负责。

Galaxy 是一个著名的 Web 目录搜索引擎,这个目录体系首先按照主题分类,各主题目录再依字母顺序排列,大主题下分有小主题,因此是个较为综合全面的 Web 目录体系。同时,在内容上包含了较多的学术性和专业性知识,内容非常丰富。同时,Galaxy 除了可以提供 Web 网页检索功能外,还能提供当时还在流行的 Telnet 和 Gopher 环境下的信息检索功能。其实 1994 年互联网的规模还很小,小到似乎没有必要去建立 Web 目录,而事实上 Galaxy 创建的一个主要原因也就是提供一种 Gopher 信息的目录检索功能,而 Gopher 采用的层次型菜单结构非常需要同时也非常适应 Galaxy 所提供的目录体系。它的网址为 http://www.galaxy.com,主页如图 3.15 所示。

目录型搜索引擎近几年的发展都受到了很大的影响,目前 Galaxy 已经停止服务。

3. Yahoo!

Yahoo!(雅虎)和 Google、Bing[①] 已经成为全球三大著名搜索引擎。事实上,它是这三者当中资格最老的一个。

20 世纪 90 年代初,美国斯坦福大学电机研究所攻读电机工程博士学位的美籍华人杨致远(Jerry Yang)和大卫·费罗(David Filo)与其他学生一样,开始喜欢上刚出现的互联网。不过,他们都有一个特殊的爱好,那就是经常将自己收集到的一些较好的网页内容链接在自己的个人网页上。渐渐地,他们自己的网页在斯坦福大学内部开始小有名气,人们称呼

① 微软公司早期推出的搜索引擎也很著名,如 MSN Search、Live Search 等,2009 年微软公司推出了 Bing(中文名称是“必应”),并同时停止了原有的那些搜索引擎服务。

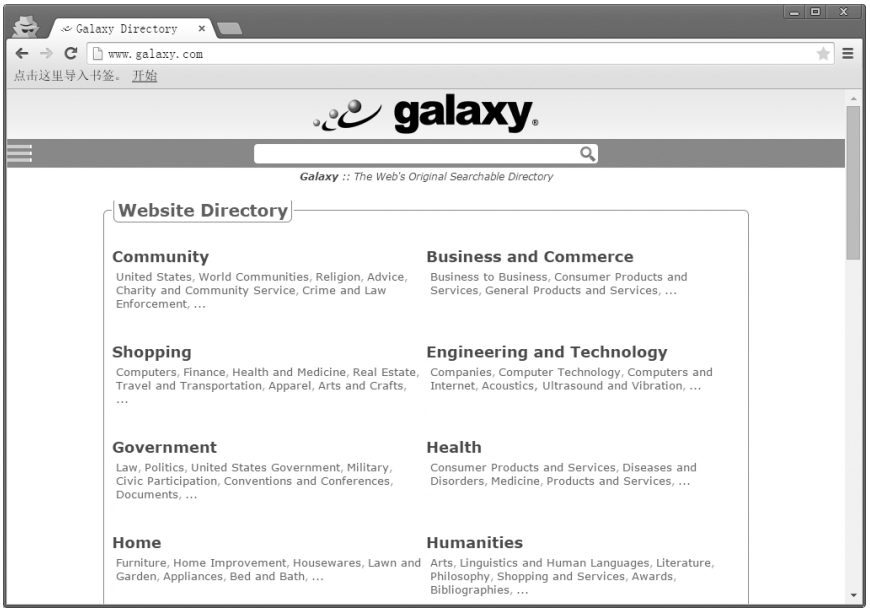


图 3.15　Galaxy 搜索引擎的主页界面(截取于 2015-3)

他们的网页为“杰里和大卫的万维网向导”(Jerry & David's Guide to the World Wide Web)。根据这些已有的经验和前期的基础,杨致远和大卫·费罗于 1994 年 4 月使用学校的工作站创建了一个网页目录查询系统,称为 Yahoo![①]。刚开始,这个网页目录就已经收集了超过 1000 个不同站点的网页信息。较基于爬虫的早期搜索引擎而言, Yahoo!所收集的网页内容能够含有人工编撰的说明信息,可以极大地方便用户的使用,而基于爬虫的搜索引擎只能通过采集网页 URL 和标题之类的简单内容来作为网页内容的提示信息,显然不论是在网页体系的组织上,还是在网页内容的说明上,都难以做到和 Yahoo!同样的效果。

事实上,Yahoo!的成功离不开它的幸运。当时有一家著名 Web 浏览器公司网景(Netscape)生产一种称为航海者(Navigator)的 Web 浏览器软件,该软件非常流行,人们都在使用它去访问 Web 网络。为了增强网络信息检索的快捷性,该浏览器的创始人马克·安德森(Marc Andreessen)看中了 Yahoo!,并且在 1995 年 1 月把航海者浏览器上一个最为重要的网络检索按钮默认指向了 Yahoo!目录。可以说,借助航海者浏览器的平台,Yahoo!很快在互联网上树立了名声。1995 年 4 月,Yahoo!还吸引了曾经给 Apple、Oracle 和 Cisco 投资过的 Sequoia 公司接近 200 万美元的投资。此时,Yahoo!已经成为互联网上的一个重要的门户站点。

然而,通过人工组织方式获取的 Web 目录结构不可能适应网络快速增长的发展要求,因此,Yahoo!先后使用了诸如 Altavista 和 Inktomi 等搜索引擎来为自己提供基于关键词的

① 关于 Yahoo!这个名称的来历也是众说纷纭,很多人认为它是“另一个层次性的民间先知”(Yet Another Hierarchical Officious Oracle)的缩写词,这可能借鉴于 UNIX 系统中一个表示网络查询技术的缩略语 YACC(Yet another compiler compiler,另一个编译器代码生成器)。但是,根据杨致远等人的说法,Yahoo 的“Ya”来自杨致远的姓,他们曾利用韦氏词典设想过 Yauld、Yammer 和 Yardage 等一系列可能的名字。之所以选中 Yahoo,是因为在《格利佛游记》中 Yahoo 是一种粗俗和不懂世故的人形动物,它具有人的种种恶习,他们反其意而用之,认为在强调平等权利的互联网上大家都是乡巴佬,为了增加褒义色彩,又在后面加上一个感叹号,于是就有了 Yahoo!。

全文检索服务。2002 年 10 月 9 日，Yahoo!开始不再使用 Web 目录作为主要搜索工具，而是使用另外一家后起之秀 Google 来为自己提供关键词查询服务，并成为真正的全文搜索引擎。正如当年 Yahoo!借助航海者成功一样，Google 最终也借助 Yahoo!成名，并敢于和 Yahoo!抗衡。此时的 Yahoo!只能通过收购的方式来获得较快的发展，2002 年 12 月 23 日收购 Inktomi 搜索引擎，2003 年 7 月 14 日收购包括 Fast 和 Altavista 在内的 Overture 公司。直到 2004 年，雅虎中国在中国内地终于推出了自己独立研发的搜索引擎“一搜”。2004，雅虎中国推出独立的搜索门户网站一搜网，“一搜天下小”是当时的广告语。后来又改名为雅虎全能搜，2013 年雅虎中国正式关闭并退出了中国市场。雅虎搜索引擎的主页如图 3.16 所示。

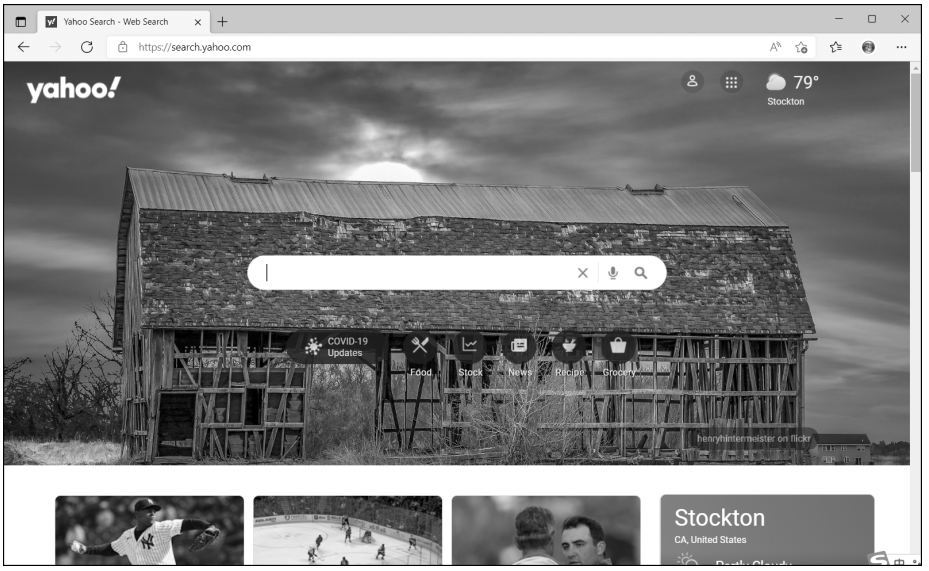


图 3.16　雅虎搜索引擎的主页界面（截取于 2022-4）

不过，Yahoo!的 Web 目录仍然还是一个重要的网络信检索工具，它的设计结构经过不断的调整，已经非常成熟和易于使用。Yahoo!的 Web 目录网址为 http://business.yahoo.com，网页如图 3.17 所示。

当然，这个 Web 目录也渐渐地融入了更多的特点。尤其是随着名声的增大，Yahoo!早已开始对收录的商业站点收费，2007 年的收录报价是每年 299 美元。但是，对于那些真正著名的站点而言，Yahoo!还是采用免费收录的方法。

受限于发展不利，Yahoo! 于 2016 年最终被 Verizon 收购，并于 2017 年 1 月更名为 Altaba。

4. ODP

ODP 是 Open Directory Project（开放目录项目）的简称，是由瑞奇·斯克伦塔（Rich Skrenta）于 1998 年和合伙人一起创办的。这个目录体系结构不仅可以提供一种 Web 网页目录的检索方法，而且这个目录体系的内容还是由全球各地的志愿者集体编撰而成，至今已经成为全球最大的 Web 目录，因此那些本来需要等待被 Yahoo!目录收录的网站现在终于找到了新的地方。更为重要的是，人们还可以免费地下载整个目录体系，以供自己的科学研

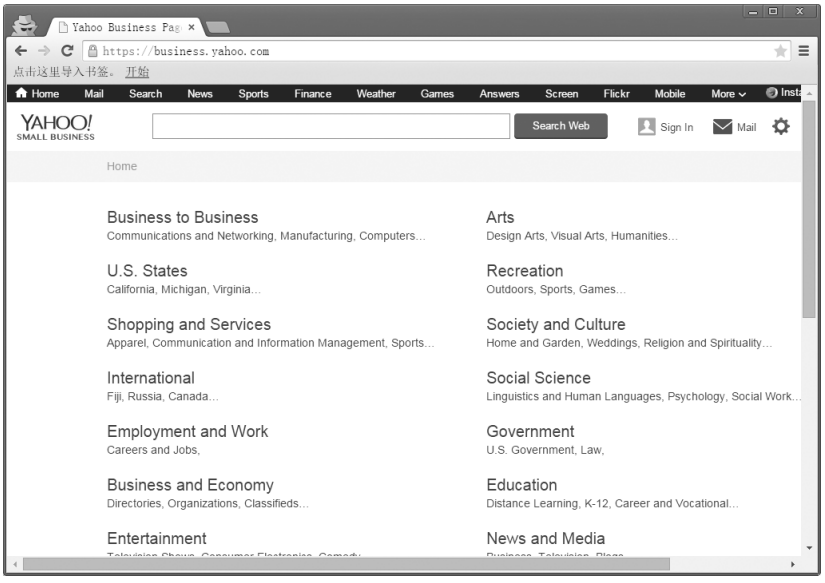


图 3.17　Yahoo!的 Web 目录主页界面(截取于 2015-3)

究。1998 年 11 月,网景(Netscape)公司收购了 ODP。随着网景公司自己在同年同月被美国在线(AOL)以 45 亿美元收购,ODP 后来归入了 AOL 的名下。ODP 的网址为 http://www.dmoz.org,主页如图 3.18 所示。

图 3.18　ODP 的 Web 目录主页界面(截取于 2015-3)

2017 年 3 月,该网站停止服务,原有的历史内容转移到 http://dmoztools.net/,并不再更新,界面如图 3.19 所示。

5. 专业的 Web 目录站点

如果读者细心,就会发现上述这些 Web 网页目录的结构有时科学性并不强,如图 3.20 显示了部分 hao123 中文分类目录的内容。

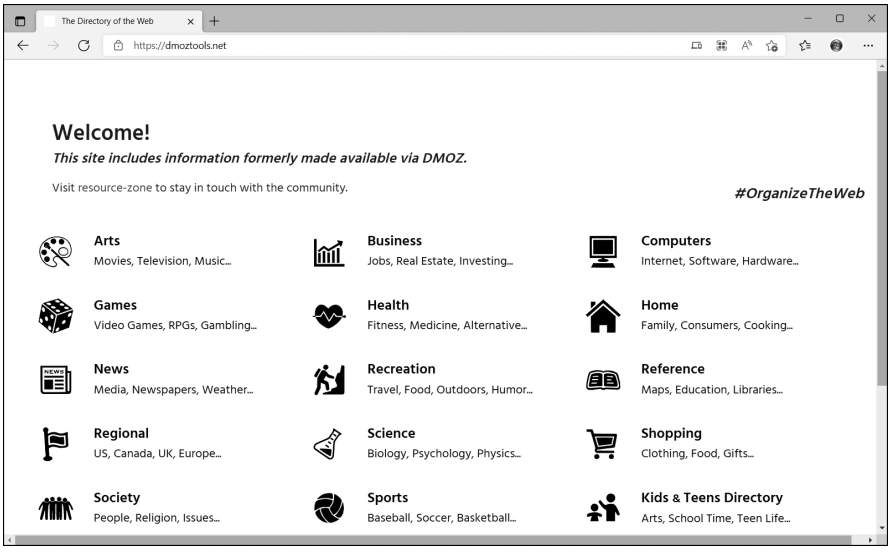


图 3.19　ODP 的 Web 目录主页界面（截取于 2022-4）

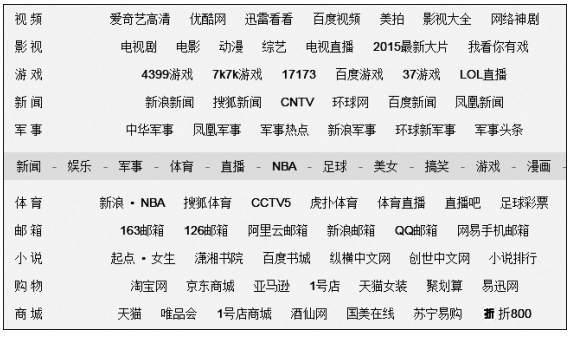


图 3.20　部分 hao123 中文 Web 目录的截图（截取于 2015-3）

显然，“商城”属于“购物”，无论如何将两者并列作为同一个目录下的子内容项并不合适，更不必说“影视”是否应该放入“视频”目录下了。

但是我们要注意，这些 Web 目录并不在意科学性，相反，它们更加在意易用性。一般的 Web 用户可能并不十分了解目录的层次结构，他们往往希望能够在最短的时间内找到自己所要的目录项，所以这些目录往往是集中了最为流行常见的目录项，并且以一种极为方便和直观的方式来展示目录结构，尽可能使用一级目录来呈现最常见的分类。

然而，对于那些诸如图书馆员等从事专门信息资源管理的专家而言，他们可能并不满意这样的结构，为此还有一些更为专业的搜索引擎 Web 目录。

克伦·施耐德（Karen G. Schneider）创办的“图书馆员 Internet 索引”（Librarians' Internet Index，LII）就是一个专门面向图书馆员的专业 Web 目录站点，该目录的结构具有较为完善的组织，科学性强，质量较高。一般而言，那些具有收费收录（Paid Inclusion）服务的 Web 目录，通常都不具备这些特点。2010 年 1 月，它和“互联网公共图书馆”（Internet Public Library，IPL）合并为 ipl2，网址为 http://www.ipl.org/div/subject，主页如图 3.21 所示。该网站现已被关闭，但还能提供服务，只是不再提供数据更新。

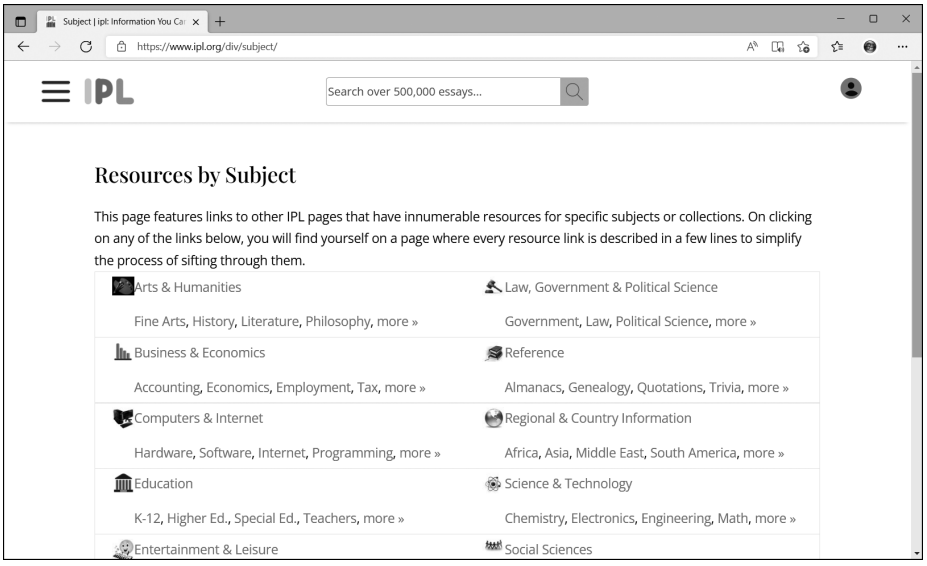


图 3.21　ipl2 的 Web 目录主页界面(截取于 2022-4)

6. 主题 Web 目录

上述这些综合 Web 目录还有很多。不过我们也应该注意到它们所面临的共同问题,如果目录小,价值不大,难以吸引用户使用,如果目录大,相关的人工整理成本太高,维护困难。所以,与综合 Web 目录不同,主题 Web 目录采取了不同的设计策略,它只面向诸如商业和经济等特定领域,从而取得了目录规模和运转成本之间的一个平衡。通常人们把这些主题 Web 目录也称为垂直 Web 目录(Vertical Web Directory)、面向局部的 Web 目录(Locally Oriented Directory)。

1995 年成立的 Looksmart 也是一家 Web 目录站点,早期曾经通过增加网页收录数量和规模来和 Yahoo!目录竞争。虽然早期并不十分成功,但是 2002 年 Looksmart 发明的一种新型的盈利模式为它的快速增长提供了基础。在此之前,几乎所有的 Web 目录站点都是采取较为固定的付费收录策略,如每月只需付多少钱可以收录到哪个目录中等。然而,Looksmart 采取了按单击付费(Pay Per Click)的收费方法[①],也就是说,用户单击该收录网页次数越多,相应的收录费用也就越高,这对被收录网页而言,显然是一种很好的激励措施,愿意为较高的点击率而支付更多的费用。

在内容上,Looksmart 不仅在自己的目录结构中收录网页索引,而且还根据主题分门别类地收录不少很有价值的内容资源,应该是个很不错的主题 Web 目录。然而,这些收录的内容相关性却因为各种原因而逐渐变差,后来在很大程度上又损害了 Looksmart 的声誉。而且在商业上发生了一连串的失败,给 Looksmart 带来了越来越多的不利影响。1998 年,Looksmart 以 2000 万美元收购一家非盈利的 Web 目录站点 Zeal 来扩展自己的目录规模,但是到了 2006 年 3 月 28 日,Looksmart 却关闭了这个 Zeal 目录。2002 年 3 月,Looksmart

① 按单击付费(Pay Per Click)的搜索引擎盈利模式最早是由 Goto 搜索引擎提出的,它允许网站管理者实时进行查询结果的排序,客户可以花钱购买排序的位置,通过拍卖的形式将相关网站放在前面,但同时明确标出这个查询结果是付费的。这种方式给它带来了巨大的收益。2001 年,Goto 更名为 Overture。

还试图通过收购 WiseNut 搜索引擎来获得发展，结果也不理想。不过，最大的问题还不止这些。Looksmart 曾经通过加盟诸如 MSN 等门户站点，通过付费收录方式来获利。然而，Looksmart 一直以来建立的良好信誉却随着这个合作而逐渐变差，而且 Looksmart 在商业上也逐渐越来越依赖于微软的 MSN 搜索引擎。到了 2003 年，微软公司宣布放弃与 Looksmart 的合作，对于 Looksmart 来说，这无疑是个致命的打击。后来，Looksmart 开始改用一个称为 Furl 的社会化书签(Social Bookmarking)管理站点来期望获得新的访问流量增长，现在它主要为广告商提供按单击付费的搜索网络平台服务。

近年来，随着 Web 用户对日常生活信息检索需求的快速增长，很多专门提供生活信息分类目录的网站逐渐受到人们的关注，如国内的“58 同城”(网址为 http://www.58.com)、国外的 craigslist(网址为 http://www.craigslist.org)等，它们都可以提供较为完整的生活信息分类目录，同时为了提高易用性，大部分目录只有一级，用户只需单击一次即可看到相关下级记录信息。

尤其在电子商务领域，基于分类目录的商品检索方式发展更为成熟，不仅能提供越来越灵活的商品类目组织体系，而且形成了和关键词检索进行有效结合的新型检索提示方法，如图 3.22 所示。

图 3.22 结合分类体系的即时类别提示

同时，由于商品属性具有较强的通用性，因此商品分类目录检索直到今天依然还在使用，如淘宝的商品分类目录检索，网址为 https://www.taobao.com/tbhome/page/market-list。

综上所述，Web 目录确实具有不少优点，所以在搜索引擎领域中一直都是一个不可或缺的角色。连 Google 公司都曾经利用 ODP 目录推出了自己的 Web 目录站点，不过已经在 2011 年 7 月份宣布关闭该服务。

然而，这种方式并非现代搜索引擎的主流。由于 Web 网页目录需要大量的人工编撰工作，所以维护成本很高，缩放性很差。而且网页目录规模通常都不大，相对于关键词查询而言，可以认为虽然关键词查询可能查准率不高，但具有更高的查全率，而网页目录查询则具