

**课 程 报 告**

**搜索引擎发展史**

**院（系）： 软件工程**

**课程名称： 信息检索（公选课）**

**学 号： U201817053**

**姓 名： 粟锦**

**任课老师： 魏巍**

**报告日期： 2019.5.1**

**目录**

摘要 2

1.绪论 2

2.概述：搜索引擎的变化 3

2.1.第一代：分类目录时代 3

2.2.第二代：文本检索时代 3

2.3.第三代：整合分析时代 3

2.4.第四代：用户中心时代 4

3.90年代搜索引擎的发展 5

3.1.Archie(现代搜索引擎的祖先) 5

3.2.90年代搜索引擎的发展 5

3.3.Google搜索算法 6

4.现代搜索引擎的基本概念 7

4.1搜索引擎的含义 7

4.2搜索引擎的组成 7

4.3搜索引擎的工作原理 7

5.搜索引擎未来的发展趋势 9

6.未来展望 11

**参考文献** 11

**摘要**

本文旨在探索回顾搜索引擎从20世纪90年代至今的发展历程，并展望其之后的发展趋势。

**绪论**

随着计算机的普及以及信息工程的日益发展，从网上获取信息已经成为人们的日常生活和工作一个重要途径，互联网如今已成为一个资源不断丰富的平台，整个互联网逐渐成为一个信息量超大的资源存储空间。因此怎样有效而又快捷的从海量数据中获取所需的信息就成为一件困难的事情,而搜索引擎的出现正是为了解决“信息丰富，获取困难”的问题的一种信息检索服务。  
 搜索引擎技术涉及到的领域很多，如：信息检索、数据挖掘、数据库、自然语言理解、中文分词等领域的前沿技术和理论, 具有复杂性和艰巨性。同时有大量的用户在使用搜索引擎,由此产生了一系列商业链具有广阔的经济前景。搜索引擎所具有的研究价值、实用价值以及商业价值是其在当今信息时代获得成功的重要因素。

**概述：搜索引擎的变化**

**（黑体小2加粗居中, 字母、阿拉伯数字为**Time New Roman小2号加粗）

**第一代：分类目录时代**（黑体4号加粗, 字母、阿拉伯数字为Time New Roman4号加粗

目录[搜索引擎](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E" \t "_blank)是以人工方式或半自动方式搜集信息，由编辑员查看信息之后，人工形成信息摘要，并将[信息](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF/111163" \t "_blank)置于事先确定的分类框架中。信息大多面向网站，提供[目录](https://baike.baidu.com/item/%E7%9B%AE%E5%BD%95/96183" \t "_blank)浏览服务和直接[检索服务](https://baike.baidu.com/item/%E6%A3%80%E7%B4%A2%E6%9C%8D%E5%8A%A1" \t "_blank)。目录[搜索引擎](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E" \t "_blank)虽然有搜索功能，但严格意义上不能称为真正的搜索引擎，只是按目录分类的网站链接列表而已。用户完全可以按照分类目录找到所需要的信息。该类搜索引擎因为加入了人的智能，所以信息准确、导航质量高，缺点是需要人工介入、维护量大、信息量少、信息更新不及时。

**第二代：文本检索时代**

在文本检索时代，搜索引擎查询信息的方法是通过用户所输入的查询信息提交给服务器，[服务器](https://baike.baidu.com/item/%E6%9C%8D%E5%8A%A1%E5%99%A8" \t "_blank)通过查阅，返回给用户一些相关程度高的信息。这代的搜索引擎的信息检索模型主要包括例如布尔模型、概率模型或者向量空间模型。通过这些模型来计算用户输入的查询信息是否与网页内容相关程度高低，将相关度高的则返回给用户。采取这种模式的搜索引擎主要是一些早期的搜索引擎，例如像Alta Vista、Excite等等。

**第三代：整合分析时代**

这一代的搜索引擎所使用的方法大概和我们今天的网站的外部链接形式基本相同，在当时，外部链接代表的是一种推荐的含义，通过每个网站的推荐链接的数量来判断一个网站的流行性和重要性。然后搜索引擎再结合网页内容的重要性来和相似程度来改善用户搜索的信息质量。这种模式的首先使用者是Google，Google不仅为首次使用并且大获成功，这一成就在当时引起了学术界和其他商业搜索引擎的极度关注。后来，学术界以此成就为基础，提出了更多的改进的链接分析算法。大多数的主流搜索引擎都在使用分析链接技术算法。

第三代搜索引擎的典型特征就是：智慧整合第二代返回的信息为立体的界面。让用户能轻易地一眼进入到最相关的分类区域去获取信息。[1-2]

**第四代： 用户中心时代**

第四代所用的搜索引擎技术是互联网上面用的最普遍的。主要是以用户为中心。当客户输入查询的请求时候，同一个查询的请求关键词在用户的背后可能是不同查询要求。例如用户输入的是“苹果”，那么作为一个想要购买iPhone的用户和一个果农来说，那么要求就是大大的不一样。甚至是同一个用户，所查询的关键词一样，也会因为所在的时间和所在的场合不同而返回的结果不同。所有主流搜索引擎，都在致力于解决同一个问题：怎样才能从用户所输入的一个简短的关键词来判断用户的真正查询请求。

**90年代搜索引擎的发展**

1. **Archie(现代搜索引擎的祖先)**

1990年，由于当时并没有出现www（万维网），所以ftp（文件传输协议）软件成为了共享文件的主要工具。但是要共享文件，必须要一个ftp服务器。当时检索ftp数据也要必须使用ftp客户端，然而很多数据都是零散分布在各个不同的地方的。然而就催生了搜索引擎的祖先—Archie。1990年，加拿大蒙特利尔的[麦吉尔大学](https://baike.baidu.com/item/%E9%BA%A6%E5%90%89%E5%B0%94%E5%A4%A7%E5%AD%A6" \t "_blank)（McGill University）的三位学生Alan Emtage、Peter Deutsch、Bill Wheelan发明了Archie。

当时的Internet可以提供诸如FTP等文件信息服务，然而用户却缺乏一种直接查询FTP文件所在地址的工具。而Archie恰恰可以自动索引Internet上匿名的免费FTP文件信息，并提供一种根据文件名称查询文件所在FTP地址的方法。因此，Archie被称为现代搜索引擎的祖先。

然而，客观的讲，它并非一个真正的搜索引擎。原因有两个：1.它只能搜索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。2.它没有机器人（Robot）程序，不能象今天的搜索引擎那样快速有效的抓取Internet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查询，并通过文件列表的方式提供信息查询结果。

**2.90年代搜索引擎的发展**

1993年Matthew Gray开发了 World Wide Web Wanderer，这是第一个利用HTML网页之间的链接关系来检测万维网规模的“机器人（Robot）”程序。开始，它仅仅用来统计互联网上的服务器数量，后来也能够捕获网址（[URL](https://baike.baidu.com/item/URL" \t "_blank)）。

1994年4月，[斯坦福大学](https://baike.baidu.com/item/%E6%96%AF%E5%9D%A6%E7%A6%8F%E5%A4%A7%E5%AD%A6" \t "_blank)（Stanford University）的两名博士生，美籍华人Jerry Yang（[杨致远](https://baike.baidu.com/item/%E6%9D%A8%E8%87%B4%E8%BF%9C" \t "_blank)）和David Filo共同创办了Yahoo。随着访问量和收录链接数的增长，Yahoo目录开始支持简单的数据库搜索。因为Yahoo！的数据是手工输入的，所以不能真正被归为搜索引擎，事实上只是一个可搜索的目录。

1994年初，[华盛顿大学](https://baike.baidu.com/item/%E5%8D%8E%E7%9B%9B%E9%A1%BF%E5%A4%A7%E5%AD%A6" \t "_blank)（University of Washington ）的学生Brian Pinkerton开始了他的小项目WebCrawler。1994年4月20日，WebCrawler正式亮相时仅包含来自6000个服务器的内容。WebCrawler是互联网上第一个支持[搜索文件](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E6%96%87%E4%BB%B6" \t "_blank)全部文字的全文搜索引擎，在它之前，用户只能通过URL和摘要搜索，摘要一般来自人工评论或程序自动取正文的前100个字。

1994年7月，[卡内基·梅隆大学](https://baike.baidu.com/item/%E5%8D%A1%E5%86%85%E5%9F%BA%C2%B7%E6%A2%85%E9%9A%86%E5%A4%A7%E5%AD%A6" \t "_blank)（Carnegie Mellon University）的Michael Mauldin将John Leavitt的spider程序接入到其索引程序中，创建了Lycos。除了[相关性排序](https://baike.baidu.com/item/%E7%9B%B8%E5%85%B3%E6%80%A7%E6%8E%92%E5%BA%8F" \t "_blank)外，Lycos还提供了前缀匹配和字符相近限制，Lycos第一个在搜索结果中使用了网页自动摘要，而最大的优势还是它远胜过其它搜索引擎的数据量。

1995年，一种新的搜索引擎形式出现了——元搜索引擎（A Meta Search Engine Roundup）。用户只需提交一次搜索请求，由元搜索引擎负责转换处理，提交给多个预先选定的独立搜索引擎，并将从各独立搜索引擎返回的所有查询结果，集中起来处理后再返回给用户。

1995年9月26日，[加州伯克利分校](https://baike.baidu.com/item/%E5%8A%A0%E5%B7%9E%E4%BC%AF%E5%85%8B%E5%88%A9%E5%88%86%E6%A0%A1" \t "_blank)助教Eric Brewer、博士生Paul Gauthier创立了[Inktomi](https://baike.baidu.com/item/Inktomi" \t "_blank)，1996年5月20日，Inktomi公司成立，强大的[HotBot](https://baike.baidu.com/item/HotBot" \t "_blank)出现在世人面前。声称每天能抓取索引1千万页以上，所以有远超过其它[搜索](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2" \t "_blank)引擎的新内容。HotBot也大量运用cookie储存用户的个人搜索喜好设置。

1995年12月，DEC的正式发布AltaVista。AltaVista是第一个支持[自然语言搜索](https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E6%90%9C%E7%B4%A2" \t "_blank)的搜索引擎，第一个实现[高级搜索](https://baike.baidu.com/item/%E9%AB%98%E7%BA%A7%E6%90%9C%E7%B4%A2" \t "_blank)语法的搜索引擎（如AND、 OR、 NOT等)。用户可以用[AltaVista](https://baike.baidu.com/item/AltaVista" \t "_blank)搜索新闻组（Newsgroups）的内容并从互联网上获得[文章](https://baike.baidu.com/item/%E6%96%87%E7%AB%A0" \t "_blank)，还可以搜索图片名称中的文字、搜索Titles、搜索Java applets、搜索ActiveX objects。

1997年8月，Northernlight搜索引擎正式现身。它曾是拥有最大数据库的搜索引擎之一，它没有Stop Words，它有出色的Current News、7100多出版物组成的Special Collection、良好的高级搜索语法，第一个支持对搜索结果进行简单的自动分类。

1996年8月，sohu公司成立，制作中文网站分类目录，曾有“出门找地图，上网找搜狐”的美誉。随着互联网网站的急剧增加，这种人工编辑的分类目录已经不适应。

**3.Google搜索算法**

[谷歌](https://baike.baidu.com/item/%E8%B0%B7%E6%AD%8C" \t "_blank)算法始于[PageRank](https://baike.baidu.com/item/PageRank" \t "_blank)，这是1997年[拉里·佩奇](https://baike.baidu.com/item/%E6%8B%89%E9%87%8C%C2%B7%E4%BD%A9%E5%A5%87" \t "_blank)([Larry Page](https://baike.baidu.com/item/Larry%20Page" \t "_blank))在[斯坦福大学](https://baike.baidu.com/item/%E6%96%AF%E5%9D%A6%E7%A6%8F%E5%A4%A7%E5%AD%A6/278716" \t "_blank)读博士学位时开发的。佩奇的创新性想法是：把整个互联网复制到本地数据库，然后对网页上所有的链接进行分析。基于入链接的数量和重要性、及[锚文本](https://baike.baidu.com/item/%E9%94%9A%E6%96%87%E6%9C%AC/370930" \t "_blank)对网页的受欢迎程度进行评级，也就是通过网络的集体智慧确定哪些网站最有用。PageRank具有其优势，但这种过度依靠外链分析单一算法也具有弊端，那就是很多站长采取作弊手法来增加网站的外链，因此网络上有很多垃圾外链。为了应对这种情况谷歌13年更新了其核心算法，那就是[蜂鸟算法](https://baike.baidu.com/item/%E8%9C%82%E9%B8%9F%E7%AE%97%E6%B3%95/12007878" \t "_blank)（Hummmingbird)。在此套算法中，PageRank仍旧起很大作用，但是已经不是唯一的排名机制！随着时间的推移，外链在排名中所起的作用将逐渐衰落！

**现代搜索引擎的基本概念**

**1.搜索引擎的含义**

搜索[引擎](https://baike.baidu.com/item/%E5%BC%95%E6%93%8E/2043" \t "_blank)（Search Engine）是指根据一定的[策略](https://baike.baidu.com/item/%E7%AD%96%E7%95%A5/4006" \t "_blank)、运用特定的[计算机程序](https://baike.baidu.com/item/%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A8%8B%E5%BA%8F/3220205" \t "_blank)从[互联网](https://baike.baidu.com/item/%E4%BA%92%E8%81%94%E7%BD%91" \t "_blank)上搜集[信息](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF" \t "_blank)，在对信息进行[组织](https://baike.baidu.com/item/%E7%BB%84%E7%BB%87/10200" \t "_blank)和[处理](https://baike.baidu.com/item/%E5%A4%84%E7%90%86/4869736" \t "_blank)后，为用户提供[检索](https://baike.baidu.com/item/%E6%A3%80%E7%B4%A2" \t "_blank)服务，将用户检索相关的信息展示给用户的[系统](https://baike.baidu.com/item/%E7%B3%BB%E7%BB%9F/479832" \t "_blank)。搜索引擎包括[全文索引](https://baike.baidu.com/item/%E5%85%A8%E6%96%87%E7%B4%A2%E5%BC%95/1140318" \t "_blank)、[目录索引](https://baike.baidu.com/item/%E7%9B%AE%E5%BD%95%E7%B4%A2%E5%BC%95/1166180" \t "_blank)、[元搜索引擎](https://baike.baidu.com/item/%E5%85%83%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E/205513" \t "_blank)、[垂直搜索引擎](https://baike.baidu.com/item/%E5%9E%82%E7%9B%B4%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E/210198" \t "_blank)、集合式搜索引擎、门户搜索引擎与免费链接列表等。

1. **搜索引擎的组成**

搜索引擎一般由搜索器、[索引器](https://baike.baidu.com/item/%E7%B4%A2%E5%BC%95%E5%99%A8" \t "_blank)、检索器和[用户接口](https://baike.baidu.com/item/%E7%94%A8%E6%88%B7%E6%8E%A5%E5%8F%A3" \t "_blank)四个部分组成

搜索器：其功能是在互联网中漫游，发现和搜集信息。

索引器：其功能是理解搜索器所搜索到的信息，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。

检索器：其功能是根据用户的查询在索引库中快速检索文档，进行相关度评价，对将要输出的结果排序，并能按用户的查询需求合理反馈信息。

用户接口：其作用是接纳用户查询、显示查询结果、提供个性化查询项。

1. **搜索引擎的工作原理**

**第一步：爬行**

搜索引擎是通过一种特定规律的软件跟踪网页的链接，从一个链接爬到另外一个链接，像蜘蛛在蜘蛛网上爬行一样，所以被称为“蜘蛛”也被称为“机器人”。搜索引擎蜘蛛的爬行是被输入了一定的规则的，它需要遵从一些命令或文件的内容。

**第二步：抓取存储**

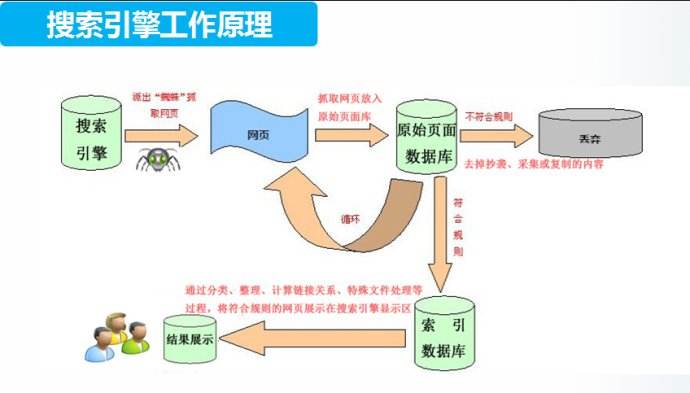
搜索引擎是通过蜘蛛跟踪链接爬行到网页，并将爬行的数据存入原始页面数据库。其中的页面数据与用户浏览器得到的HTML是完全一样的。搜索引擎蜘蛛在抓取页面时，也做一定的重复内容[检测](https://baike.baidu.com/item/%E6%A3%80%E6%B5%8B" \t "_blank)，一旦遇到权重很低的[网站](https://baike.baidu.com/item/%E7%BD%91%E7%AB%99" \t "_blank)上有大量抄袭、采集或者复制的内容，很可能就不再爬行。

**第三步：预处理**

搜索引擎将蜘蛛抓取回来的页面，进行各种步骤的预处理。

**第四步：排名**

用户在搜索框输入关键词后，排名程序调用索引库数据，计算排名显示给用户，排名过程与用户直接互动的。但是，由于搜索引擎的数据量庞大，虽然能达到每日都有小的更新，但是一般情况搜索引擎的排名规则都是根据日、周、月阶段性不同幅度的更新。

****

**搜索引擎未来的发展趋势**

1. **社会化搜索**

    传统搜索技术强调搜索结果和用户需求的相关性，社会化搜索除了相关性外，还额外增加了一个维度，即搜索结果的可信赖性。对某个搜索结果，传统的结果可能成千上万，但如果处于用户社交网络内其他用户发布的信息、点评或验证过的信息则更容易信赖，这是与用户的心里密切相关的。社会化搜索为用户提供更准确、更值得信任的搜索结果。

1. **实时搜索**

 实时搜索最突出的特点是时效性强，搜索核心强调“快”，用户发布的信息第一时间能被搜索引擎搜索到。

1. **移动搜索**

  随着智能手机的快速发展，基于手机的移动设备搜索日益流行，但移动设备有很大的局限性，比如屏幕太小，可显示的区域不多，计算资源能力有限，打开网页速度很慢，手机输入繁琐等问题都需要解决。

1. **个性化搜索**

 个性化搜索的核心是根据用户的网络行为，建立一套准确的个人兴趣模型。而建立这样一套模型，就要收集与用户相关的信息，包括用户搜索历史、点击记录、浏览过的网页、用户email信息、收藏夹信息、用户发布过的信息、博客、微博等内容。比较常见的是从这些信息中提取出关键词及其权重。

1. **地理位置感知搜索**

目前很多手机已经有GPS的应用了，这是基于地理位置感知的搜索，而且可以通过陀螺仪等设备感知用户的朝向，基于这种信息，可以为用户提供准确的地理位置服务以及相关搜索服务。

1. **跨语言搜索**

对于一个全球性的搜索引擎来说，具备跨语言搜索功能是必然的发展趋势，而其基本的技术路线一般会采用查询翻译加上网页的机器翻译这两种技术手段。

1. **多媒体搜索**

目前搜索引擎的查询还是基于文字的，即使是图片和视频搜索也是基于文本方式。那么未来的多媒体搜索技术则会弥补查询这一缺失。多媒体形式除了文字，主要包括图片、音频、视频。

1. **情境搜索**

  所谓情境搜索，就是能够感知人与人所处的环境，针对“此时此地此人”来建立模型，试图理解用户查询的目的，根本目标还是要理解人的信息需求。比如某个用户在苹果专卖店附近发出“苹果”这个搜索请求，基于地点感知及用户的个性化模型，搜索引擎就有可能认为这个查询是针对苹果公司的产品，而非对水果的需求。

**未来展望**

如今，互联网的飞速发展已经将“秀才不出门，便知天下事”变成了可能，对此搜索引擎功不可没。随着科技的不断进步，搜索引擎势必将变得更加个性化，更加精准有效，更好地服务于人类社会。

**参考文献** (黑体小2号加粗居中)

1]百度百科：Archie(现代搜索引擎的祖先)，<https://baike.baidu.com/item/Archie/1272913?fr=aladdin>

[2]甲今文（昵称）：搜索引擎发展的历史过程与发展现状，<https://www.cnblogs.com/jiajinwen/archive/2012/03/07/2384427.html>

[3]百度百科：谷歌搜索引擎算法，<https://baike.baidu.com/item/%E8%B0%B7%E6%AD%8C%E6%90%9C%E7%B4%A2%E7%AE%97%E6%B3%95/6543372>

[4]百度百科：搜索引擎，https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E/104812?fr=aladdin

[5]百度百科：搜索引擎发展史，<https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E%E5%8F%91%E5%B1%95%E5%8F%B2/2422574>

[6]百度文库：搜索引擎的社会和经济影响（2014）, https://wenku.baidu.com/view/64826585afaad1f34693daef5ef7ba0d4a736ded.html