维普资讯 http://www.cqvip.com

Web 搜索引擎综述*)

Overview of the Web Search Engine

张卫丰　徐宝文　周晓宇　许　蕾　李　东
(东南大学计算机科学与工程系　南京 210096)

Abstract　With the explosive increase of the network information, people can find information more and more difficultly. The occurrence of the Web search engine overcomes this problem in some degree. This paper tells about the history of the search engine, the current state of the search engine. Some guidelines about the search engine are analysed and the related checking methods are also given. In this basis, we introduce the trend of the search engine.

Keywords　Internet, WWW, Search engine, Analysis method

1　引言

互联网络的历史只能从 1960 年代后期算起，从早期的 ARPANet，到目前的 Internet，互联网飞速发展，基于互联网的各种应用也应运而生。互联网上的信息也越来越多，因此迫切要求一种信息检索工具。1991 年，XWAIS 提供了一个界面友好的信息搜索系统，但是这个系统要求很特殊的文件格式。在同一年出现了另外一个信息检索系统 GOPHER，GOPHER 一时成为最为流行的检索工具。由于 GOPHER 基于字符界面，一般的互联网用户还是很少使用它。真正让互联网普及的转机出现在 1993 年。当年美国国家计算机安全协会 NCSA 推出第一个基于 HTML 语言的可以浏览图形的浏览器 Mosaic。它使得普通的用户可以轻松地使用互联网。1994 年，美国网景公司推出免费浏览器 Netscape。这使得以 HTML 为格式的信息迅速膨胀。是年，Yahoo 公司创立，它提供基于目录的信息检索服务。而真正意义上的搜索引擎创建于 1994 年春天的 Lycos，当时 Michael Mauldin 将 John Leavitt 的"网络蜘蛛"(spider)程序接入到其索引程序中。

在随后的几年里，互联网和 Web 技术的进一步发展，网上的信息越来越多，据 1999 年的估计，到 1999 年底，至少有 1600 万台主机联入因特网，网上的网页数量已经达到 10 亿，而且正在以每月近千万的数量增长，甚至有人预言 Web 页面的数量每隔 100 到 120 天要翻一翻。国内外的调查结果都表明，当前互联网上仅次于收发电子邮件的第二大应用就是在网上搜索信息，而这种搜索绝大多数都是通过专门的、高度复杂的搜索引擎实现的。

搜索引擎一词在国内外因特网领域被广泛使用，然而，它的含义却不尽相同。在美国，搜索引擎通常指的是基于因特网的搜索引擎，它们收集因特网上几千万到几亿个网页，并且每一个网页上的每一个词都被搜索引擎所收录，也就是我们所说的全文检索。典型的因特网搜索引擎包括 First Search、Google、HotBot、Infoseek、Nothern light 等。在中国，搜索引擎通常指的是基于网站目录的搜索服务或是特定网站的搜索服务。前者如搜狐、新浪等公司开发的网站搜索服务。后者如 Chinaren(search.chinaren.com)网站上提供的全文检索服务。在下文中所指的搜索引擎均为基于因特网的搜索引擎。

现在大多数的搜索引擎以搜索文字信息为主。随着网络带宽的不断加大，多媒体信息在网上迅速增加，这就对多媒体信息的检索提出了要求。多媒体信息检索主要是指基于音频的检索、基于图片的静态图像检索和基于视频的动态图像检索。现在研究得较多的是图像检索。由于在搜索过程中很难表达图像信息，因此现有的图像搜索引擎通过对图像信息的文字表达来进行检索，文字信息不能充分表达图像信息，而且对于用户来说，不可能在查询时很精确地用文字对图像进行合理的描述，所以查询的精度非常低。由于用户一次搜索反馈的过程一般不会超过 3 次，因此机器学习的过

*)本研究得到国家自然科学基金(60073012)与教育部高等学校骨干教师资助计划资助。张卫丰　博士生，主要从事程序设计语言、软件体系结构、网络语言等方面的研究。徐宝文　博导，主要从事程序设计语言、软件工程、并行程序设计等方面的教学与科研工作。

维普资讯 http://www.cqvip.com

程也不能超过 3 次就让用户得到所需要查找的信息。微软中国研究院的研究人员提出通过机器学习的方法让用户在 3 次反馈之内得到比较精确的结果[6]。

2 基于因特网的搜索引擎的构成

搜索引擎根据用户的查询请求，按照一定的算法从索引数据库中查找对应的信息返回给用户。为了保证用户查找信息的精度和新鲜度，搜索引擎需要建立并维护一个庞大的索引数据库。一般搜索引擎主要由网络蜘蛛、索引与搜索引擎软件等部分组成。

· 网络蜘蛛　也称"爬行者(Crawler)"，是一个功能很强的程序，它会定期根据预先设定的地址去查看对应的网页，如网页发生变化则重新获取该网页，否则根据该网页中的链接继续去访问，网络蜘蛛访问页面的过程是对互联网上信息遍历的过程。为了保证网络蜘蛛遍历信息的广度，一般事先设定一些重要的链接，然后对这些链接进行遍历。在遍历过程中不断记录网页中的链接，不断遍历下去，直到访问完所有的链接。

· 索引　网络蜘蛛将遍历得到的页面存放在临时数据库中，为了提高检索的效率，需要建立索引。索引一般按照倒排文件的格式存放。如果有时索引不能及时更新，网络蜘蛛带回的新信息就不能被使用搜索引擎的用户查到了。因此，

新信息更新周期＝网络蜘蛛停止的时间＋网络蜘蛛遍历的时间＋索引建立的时间

· 搜索引擎软件　该软件用来筛选索引中无数的网页信息，挑出符合查询要求的网页并把它们分级排序，与查询关键字关联越大的排得越前，然后将分级排序后的结果显示给查询用户。

根据专家的评测，目前主要的搜索引擎返回的相关结果的比率不足 45%，而且由于机制、范围、算法等的不同，导致同样一个检索请求在不同搜索引擎中的查询结果的重复率不足 34%。因此，要想获得一个比较全面、准确的结果，就必须反复调用多个搜索引擎[15,16]。元搜索引擎的出现，在一定程度上解决了这些问题[18]。

3 搜索引擎的主要指标及其分析

搜索引擎的主要指标有响应时间、查全率、查准率、受欢迎程度、建立索引的方法和相关度等等。所谓查全率是指一次搜索结果集中符合用户要求的数目与和用户查询相关的总数之比；所谓查准率是指一次搜索结果集中符合用户要求的数目与该次搜索结果总数之比；相关度是指用户查询与搜索结果之间相似度的一种度量。响应时间、查全率、查准率和受欢迎程度为搜索引擎的主要评价指标，建立索引的方法和相关度是搜索引擎有代表性的技术指标。搜索引擎的技术指标决定了搜索引擎的评价指标。好的搜索引擎应该具有较快的响应速度和高的查全率和查准率，当然这些都需要搜索引擎的技术指标来保障。下面将从搜索精度、搜索引擎的受欢迎程度、搜索引擎建立索引的方法以及相关度这几个方面对典型的搜索引擎进行比较。

3.1 搜索引擎的精度

搜索引擎的查准率是个复杂的概念，一方面表示搜索引擎对搜索结果的排序能力，另一方面却体现了搜索引擎对垃圾网页的抗干扰能力。对搜索结果的排序能力主要取决于搜索引擎采用的排序(rank)算法的优劣。互联网中大多数网页是比较正式的，但是有些人为了赢利(由于他可以通过用户访问他的站点、他可以从一些点击付费网站获取利益)，通过非法途径欺骗搜索引擎使得自己有较高的排名，从而使得使用该搜索引擎的用户获得无用的信息。例如，在"莱温斯基事件"炒得沸沸扬扬的时候，很多人想了解有关该事件的最新报道，但是通过搜索引擎找到的却是很多无关紧要的东西，这主要是有些站点的垃圾信息影响了正常搜索结果的排序。对付垃圾网页的办法主要是让网络蜘蛛具有判别是否是垃圾网页的能力。目前根据 searchenginewatch[4] 最近的统计报告，在搜索结果中信息的质量情况如表 1 所示。

表 1　搜索引擎结果中的质量

搜索引擎 / 搜索质量	AltaVista	Excrte	HotBot	Infoseek	Lycos
优良	0%	20%	20%	40%	20%
好	20%	30%	10%	40%	10%
较差	30%	0%	30%	10%	10%
垃圾信息	0%	20%	10%	0%	10%
非所要信息	50%	30%	30%	10%	50%

3.2 搜索引擎受欢迎的程度

搜索引擎的受欢迎程度体现了用户对搜索引擎的偏爱程度。知名度高、性能稳定和搜索质量好的搜索引擎将备受青睐。

根据 Nielsen/NetRatings(http://www.nielsen-netratings.com/)在 2000 年 6 月份的统计资料显示，搜索引擎受欢迎的程度如图 1 所示。

统计资料[4]显示，AltaVista 在 2000 年 9 月的日均访问量为 5 页面，Inktomi 在 2000 年 4 月的日均访问量为 4700 万页面，Google 在 2000 年 8 月的日均访问量也达到了 4000 万页面(其中 1400 万来自于 Google.com 本身，其它访问量来自于它的合作伙伴，如 Yahoo、Netscape Search 等)。

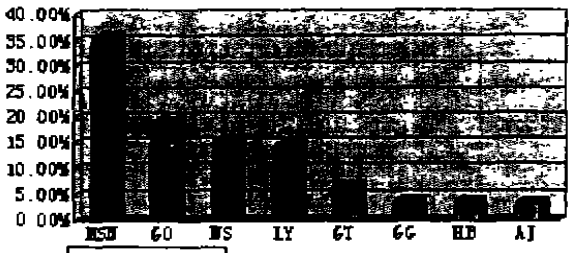

图 1 搜索引擎受欢迎程度比较

搜索引擎的受欢迎程度将随着它的知名度和服务水平的变化而动态改变。搜索引擎的服务水平跟它所采集的信息量、信息的新鲜度、搜索的响应速度和查询的精度紧密关联。随着各种新的搜索技术的出现，智能化的、高精度的、支持图像检索的搜索引擎将越来越受欢迎。

3.3 搜索引擎建立索引的方法

"网络蜘蛛"将信息存放到搜索引擎本地数据库中。为了加快检索的速度，搜索引擎要对这些数据库中的信息建立倒排索引。建立倒排索引的时候，不同的搜索引擎有不同的选项.

·全文索引和部分索引 有些搜索引擎对于信息库中的页面建立全文索引，有些只建立摘要部分，或者每个段落前面部分的索引，还有些搜索引擎（如Google）建立索引的时候，同时考虑超文本的不同标记所表示的不同含义。如粗体、大字体显示的东西往往比较重要；放在锚链中的信息往往是它所指向页面的信息的概括，所以用它来作为它所指向的页面的重要信息。Google、Infoseek 还在建立索引的过程中收集页面中的超链接。这些超链接反映了收集到的信息之间的空间结构。利用这些结果信息可以提高页面相关度判别时的准确度。

·是否过滤无用的词汇 由于网页中存在着许多无用的单词（stop words），例如英文单词"a"、"an"、"the"等，中文中的"的"、"啊"等。这些词汇不能明确表达该网页信息，所以有些搜索引擎保存有一个无用词汇表，在建立索引的时候将不建立这些词汇的索引。我们可以通过简单的方法来测试搜索引擎是否过滤掉了无用词汇。选择针对性的几个无用词汇，然后将这些无用词汇作为检索的关键词提交给搜索引擎，如果搜索引擎检索不到结果，则该搜索引擎具有过滤无用词汇的功能，但是有些搜索引擎在利用用户的查询串进行检索之前，先过滤掉了查询串中的无用词汇。这样通过我们的方法就不能检测出搜索引擎建立索引时是否过滤掉了无用词汇。

·是否使用 Meta 标记中的信息 网页中的 Meta 标记用来标注一些非显示性的信息。有些网页将页面的关键词等信息放在其中，便于建立索引的过程中提高这些词汇的相关度。Go 与 Inkotomi 在建立索引的时候考虑了页面中的 Meta 标记。

·是否对图像标记中的替换文本（ALT text）或者页面中的注解做索引 由于现有的搜索引擎对于图像的检索技术还不成熟，大多数搜索引擎不支持图像的检索。在超文本的结构页面中，图像标记中往往存放着图像的替换信息。这些信息说明了该图像标记对应的图像的基本信息。如果能够对这种图像替换信息进行索引，将可以进行某种程度上的图像检索。AltaVista、Google 专门针对图像的替换文本建立了索引。

页面中的注解（comments）信息主要是页面的设计者在页面设计过程中添加的一些附加信息。它不在浏览器中显示。Inkotomi 在索引过程中建立对注解信息的索引。

·是否支持词干提取技术 词干提取（stemming）技术是指搜索引擎在建立索引的过程中，只对词汇的词干部分建立索引。如对单词"computers"、"computing"等单词"compute"的词性变换形式，统一建立单词"compute"的索引。我们可以通过向搜索引擎提交同一词汇的不同变化形式，如果得到相同的搜索结果，那么该搜索引擎已经使用了词干提取技术。有些搜索引擎在用户提交查询请求的时候，首先对查询请求中的词条进行词干提取。然后再送给后台的搜索程序。有些搜索引擎在对查询请求中的词条进行词干提取后，将词干的所有变化形式一起作为搜索请求提交给搜索程序。

根据上文所说的检测方法结合 searchenginewatch（www.searchenginewatch.com）的报告，我们可以得到如表 2 所示的统计信息。

3.4 搜索引擎相关性考虑

搜索引擎在决定关键词与页面的相关性的时候主要考虑关键词在页面中的位置和频率，即所谓的"位置/频率"法[3,5]。如果关键词出现在页面的头部，或者在它的标题（title）标记中，那么显然该页面比较重要。如果关键词在页面中重复多次出现，显然该页面跟那关键词的相关度越高。

近年来，出现了一些计算相关度的新方法。这些方法的出现大大提高了搜索的精度。Excite、Google 等充分挖掘超文本本身的结构特点，考虑页面之间的链接关系对页面相关度的影响，它们基于这样一种直觉：如果一个页面被重要的页面所指向，那么被指向的页面

维普资讯 http://www.cqvip.com

也相对重要。Hotbot 和 Loycos 则考虑用户的点击行为对页面相关度的影响；Go 和 Inkotomi 考虑了 Meta 标记对页面相关性的作用，一些混合结构（搜索引擎和目录共存）的搜索引擎可能会把那些目录中已存在的站点的网页靠前，因为一个站点要足够好才能放到目录中，所以就应该有机会比那些不被列在目录中的站点的网页靠前。

表 2 搜索引擎索引方法比较

主要技术指标 / 是否采用相关技术	全文索引	过滤无用词汇	Meta 标识	注解	替换文本	词干提取
是	所有的搜索引擎	AIta Vista Excite Inktomi Lycos Google	大部分都使用，除了下面的几个搜索引擎	Inktomi	AltaVista Go Google Lycos	Go Lycos Northern Light
否		FAST Go NLight	FAST Google Lycos NLight	其它	Excite FAST Inktomi NLight	

注：有些搜索引擎不对无用词汇索引

4 搜索引擎的发展趋势

新一代的搜索引擎应该在自然语言处理、数据挖掘和机器学习技术、基于内容的多媒体查询技术、多通道用户界面（语音、自然语言、多媒体）方面有所突破[6]。有人指出基于关键字的查询很难表达很多复杂的概念[1]，而且常常得到太多不相关的结果（浪费时间和精力）。

为了让大多数用户方便地使用搜索，要求搜索引擎具有处理自然语言输入的功能，而且作为面向全球服务的搜索引擎必须面对不同语言的用户。即未来的搜索引擎应该具有满足对多种语言输入的功能。Ask Jeeves 巧妙地将用户提问转化为系统已知的问题，然后对已知的问题进行回答，这样就降低了对自然语言理解技术的依赖性。Google 自动检测用户所在的位置，然后给出对应风格的文字界面。

随着语音识别技术和多媒体技术的发展，未来的搜索引擎应该可以利用语音作为输入，可以搜索的内容也不再局限于文字信息，而可以拓展到多媒体信息。

由于人们各自感兴趣的领域不同，各自对词意的理解也不尽相同，不同的用户对同一个检索请求得到的检索结果有不同的评价。一个理想的搜索引擎应该对不同的用户在相同的检索请求下有不同的检索结果，即对用户具有自适应能力。这可以通过两种方案实现：

其一，需要系统在检索请求提交数据库之前智能化地调整查询表达式和查询域，既查询预处理；

其二，在查询结果返回的时候，智能化地对搜索结果进行预处理后再返回给用户。

我们根据用户的对搜索结果的选择情况，提出了通过自动采集用户的兴趣，然后对用户的搜索结果进行过滤以提高用户搜索精度的方法[14]。文[1]中提出由于普通的搜索引擎的倒排索引都是以关键词为基础的。Pinkerton 通过实验指出，用户使用搜索引擎平均输入的关键词的个数为 1.5 个[2]，这将使得用户得到成千上万看似相关的结果。Chia Hui Chang 提出通过对搜索结果建立聚集的方法和记录用户查询的模式来提高对用户的精度[1]。通过对搜索结果中抽取重要的词汇以及利用用户兴趣的反馈来修正查询，对搜索结果建立目录层次空间来缩小用户的搜索空间。

综上所述，未来搜索引擎将有如下主要发展趋势：

· 自然语言，精度更高　自然语言的输入将更加方便用户的使用，更易于用户与搜索引擎的交互。自然语言更能贴切地表达用户的查询需求，这将有利于提高查询的精度。现有的一些搜索引擎如 Infoseek 和 Google 通过对网上的超链结构进行分析来提高搜索结果的精度。Directhit（现被 Ask Jeeves 收购）则通过分析用户的点击行为来提取用户的兴趣，将搜索引擎与网站目录相结合也是提高搜索结果表达的一种有效手段。

· 多种语言搜索　多语言搜索可以是集中式的多语言搜索，也可以是分布式的（按照不同语言的分布来分布搜索引擎，即搜索引擎的本地化）多语言搜索。集中式多语言搜索的搜索引擎将多种语言的处理和搜索引擎索引数据放在同一个地方上。分布式多语言搜索引擎将搜索引擎按照语言习惯、地理位置分布在不同的区域，一个搜索引擎负责处理一种或类似的几种语

维普资讯 http://www.cqvip.com

言。

·善解人意，学习个人喜好　搜索引擎通过不断地学习，来掌握用户的喜好，通过对用户搜索习惯、用户兴趣的掌握，达到改进搜索结果的目的。

·多通道输入和多媒体输出　用户可以通过声音、图像、视频作为查询的输入，查询的内容也不再局限于文字信息，而是多媒体信息。

·个性化和本地化　新一代搜索引擎应该考虑人的性别、年龄，地域等方面的差别，给出个性化的搜索结果。随着因特网在全球的迅速普及，综合性的搜索引擎已经不能满足很多非本地区用户的信息需求。近来，Yahoo、Excite 等公司不断推出各国、各地区的本地搜索网站，搜索的本地化已经是必然趋势。

5　我们的工作

近年来，我们在对 Web 技术作了初步研究的基础上[7~12]，又尝试将数据挖掘技术，Agent 技术和遗传算法等应用于智能 Web 搜索引擎的研究中，并取得了初步成果[11~13,17]。

我们通过数据挖掘技术来提取用户的兴趣，然后根据用户的兴趣来过滤搜索引擎的返回的结果。这使得用户所得到的搜索结果善解人意，可以满足用户的个人喜好。

用户的反馈信息对于提高搜索引擎的精度非常重要。我们通过建立用户对搜索结果的反馈机制[14]，使得搜索引擎不断学习用户的兴趣信息，以此来提高搜索引擎的对用户的自适应能力。

元搜索引擎通过调用其它独立搜索引擎来实现搜索。而选择哪些独立的搜索引擎是个非常棘手的问题。遗传算法是通过模似生物的进化过程来实现问题优化的一种方法。我们通过遗传算法实现了对独立搜索引擎的合理调度。

我们将来的工作将主要集中在搜索引擎搜索请求的预处理和搜索结果的后处理上。通过对搜索引擎的机器学习，来达到对这些搜索引擎更进一步的掌握，通过对用户搜索习惯、用户兴趣的学习，达到改进搜索结果的目的。

结束语　本文介绍了搜索引擎的发展历史，讨论了搜索引擎的基本工作原理，分析比较了搜索引擎的几个关键指标并给出了检测这些指标的方法。在此基础上，我们分析了搜索引擎面临的问题和将来的发展趋势。我们还介绍了我们在搜索引擎方面所做的工作。未来的搜索引擎应该信息量更大，搜索速度更快、搜索精度更高和能够满足用户个性化的要求。

感谢英国 De Montfort 大学 Hongji Yang、Staffordshire 大学 Kecheng Liu、台湾东海大学 William C. Chu 以及香港大学 William Song 等教授与我们在 Web 搜索引擎研究方面的友好合作。

参考文献

1　Chang C H, et al. Customizable Multi Engine Search Tool with Clustering. Sixth International World Wide Web Conference. Available at: Http://www6.nttlabs.com/hypernews/get/paper53.html

2　Pinkerton B. Finding What People Want: Experiences with the WebCrawler. In: Second Intl. WWW Conf.'94, July 1994, Chicago, USA, Oct. 1994. http://info.webcrawler.com/bp/WWW94.html

3　Available at: http://www.searchenginewatch.com/webmasters/rank.html

4　Available at: http://www.searchenginewatch.com/webmasters/tips.html

5　Available at: http://www.searchenginewatch.com/sereport/98/08-clicks.html

6　Available at http://www.microsoft.com/china/research

7　张卫丰，徐宝文. Web 搜索引擎框架研究. 计算机研究与发展，2000，37(3)：376～378

8　张卫丰，徐宝文，周晓宇. Web 页面中的计数器研究. 小型微型计算机，2000，21(10)：1096～1099

9　张卫丰，徐宝文，周晓宇. Web 页面中元素间交互技术研究 计算机工程，2000，26(8)：62～64

10　张卫丰，徐宝文，许蕾. Web 页面安全性技术初探. 计算机工程与应用，2000，36(10)：158～161

11　张卫丰，徐宝文，许蕾. 利用 Agent 个性化搜索结果. 小型微型计算机，已录用

12　Zhang Weifeng, Xu Baowen, Yang Hongji, Chu W C. A Genetic Algorithm Based General Search Engine. In: Proc. of IEEE MSE'2000

13　Zhang Weifeng, Xu Baowen, Chu W C, Yang Hongji. Data Mining Algorithms for Web Pre-Fetching. In: Proc. of The Workshop on the World Wide Web Semantics (WebSem'2000)

14　张卫丰，徐宝文. 带反馈自适应 Web 搜索引擎研究. 待发表

15　Selberg E, Etzioni O. Multi-Engine Search And Comparison Using The MetaCrawler. In: Proc. of the Fouth World Wide Web Conference'95, Boston USA. Dec. 1995

16　Dreilinger D. Integrating Heterogeneous WWW Search Engines May 1995 Ftp://132.239.54.5/savvy/report.ps.gz

17　Xu Baowen, Zhang Weifeng, Chu W C, Yang Hongji. Application of Data Mining in Web Pre-Fetching. In: Proc. of IEEE MSE2000

18　张卫丰，徐宝文，等 元搜索引擎研究. 计算机科学，2001，28(8)