

互连网上的信息检索

主讲 田捷博士
(研究员, 博士生导师)

Email: tian@dr.com

<http://www.digiark.com/tian>



- 第一节 概论
- 第二节 Web Mining：第二代网络信息处理技术
- 第三节 网上信息搜索的利器：元搜索引擎及其特色
- 第四节 搜索引擎挑战智能化
- 第五节 流行的中文搜索引擎
- 第六节 流行的中文搜索引擎
- 第七节 中国门户网站的“搜索”较量



第一节 概论

- 一、绪论
- 二、历史
- 三、了解查询工具是怎样工作的
- 四、几个搜索概念
- 五、查询检索中的几个要点
- 六、搜索引擎的分类
- 七、搜索引擎的现状
- 八、未来的发展趋势



一、绪论

WWW 资源内容包罗万象，信息浩如烟海，其覆盖面之广，对人类生活影响之大是任何人难以想象的。目前这一庞大的信息库包含了数千万个页面，并且还在以惊人的速度增加。毫不夸张的说，它所包含的信息量足以超过计算机出现以前人类社会所有有文字记载的信息的总和。

到目前为止，我们在Web上漫游依靠的是链接点及推荐的URL来知道目标的可能所在地。在这信息的汪洋大海中，信步漫游自然是一种消闲时的享受，但是如果必须用 Internet迅速地查找某一专题或者进行某种严肃的工作的时候，您需要快速地定位，直达目的地。



这时“漫游”就不是一种高效率的方法，有时候根本就帮不了您的忙了。尤其目前WWW发展突飞猛进，面对无边无际且不断变化着的信息库，快速准确获取自己所需的信息显得尤为重要。这时，指导我们方向的罗盘在哪里？它们就是网上查询工具，或者称做搜索引擎(Search Engine),它们对WWW页面进行分类、查询和检索。



在Internet上有好多种有效的WWW搜索引擎可以用来寻找特定的信息，如Yahoo（<http://www.Yahoo.com/>）InfoSeek（<http://www.infoseek.com/>）等等。熟练掌握查询工具对我们充分利用WWW资源，提高工作效率是必需的。下面我们分别介绍一些目前流行的WWW搜索引擎，供大家使用参考。其中重点讲解国外的Yahoo和其它搜索引擎。



每一查询工具使用一个 Search Engine，它定期地探查 Internet 上的新的信息。当发现有新的信息，它就把它分类或编成索引，并同同一个 URL（统一资源定位器，是信息存储的地址）上的分类目录联系起来。在输入一个特定的查询条目时，查询工具进入索引，找出所有同查询条目相匹配的条目，并显示一个指向存放这些信息的链接点清单。由于用这种方式处理查询，所以查询工具能在数秒钟内完成一项查询工作。

二、历史

1993年，Internet上出现了最早的Web浏览器Mosaic，次年Netscape推出了Navigator，浏览器的发展促使Web得到迅速推广，同时也推动着搜索引擎的发展。1994年初，Internet上出现了包括Lycos在内的第一批Web搜索引擎，同年还成立了Yahoo！，后者成为了近年来最成功的商业目录。现在Internet上已有数千个提供搜索服务的站点，它们不仅要努力改进自己的服务以便能在激烈的竞争中生存下来，还要努力寻求新技术以便能适应Internet的迅速扩张。

三、了解查询工具是怎样工作的

当完成了一项查询，我们也许对查询工具在Internet上彻底搜索所需信息的能力有所了解。实际上，查询工具自身早已事先做完了此事。



每一查询工具使用一个 Search Engine，它定期地探查 Internet 上的新的信息。当发现有新的信息，它就把它分类或编成索引，并同同一个 URL（统一资源定位器，是信息存储的地址）上的分类目录联系起来。在输入一个特定的查询条目时，查询工具进入索引，找出所有同查询条目相匹配的条目，并显示一个指向存放这些信息的链接点清单。由于用这种方式处理查询，所以查询工具能在数秒钟内完成一项查询工作。

查询工具定期漫游Web，编辑长长的索引，它得到了诸如此类的名称，knowbot，robot，web crawler，spider及infobot。这些名称全反映了查询工具的一种本性：它们是自动化的和坚持不懈的。



四、几个搜索概念

在介绍各个搜索站点之前，我们先了解几个搜索概念。

关键字：简单地说，即索引词，是用来检索某一类或某一个信息的提示词。它可以是信息的主题，也可以是作者，也可以是具有某种确定性意义的描述某一特征的词语。

查询工具定期漫游Web，编辑长长的索引，它得到了诸如此类的名称，knowbot，robot，web crawler，spider及infobot。这些名称全反映了查询工具的一种本性：它们是自动化的和坚持不懈的。



布尔逻辑：在Internet的搜索中主要运用and（与）、or（或）、not（非）三种运算。它对快速、准确、有效地搜索信息起极大的作用。



停止词：在输入查询条目时，查询工具不一定关注键入的所有词目。查询工具会忽略掉一些特定的通用词汇，称为 stop word，即停止词。这些词包括计算机常用的名词如computer、Internet以及a、the、what等等一类词。所以，当键入查询条目时，可以只输入一两个具有唯一意义的词汇。

命中符：当查询工具查询其索引时，它对每一条目确定命中符的数目。一个命中符意味着查询工具在索引中找到一个能匹配查询条目内容之一的一个单词。在查询结果清单中，有最多命中符的条目放在最前列。所以当尝试决定用哪一个链接点时，应该从清单的开始向下工作。

五、查询检索中的几个要点

使用许多搜索工具，但当你搜索结果不满意，要么条目过多，要么条目太少的时候，你可以使用以下基本策略，改进您的搜索结果：



1、改变查询范围

缩小查询范围：如果选用格式化查询工具，应该很仔细地挑选查询条目。用一个诸如car这类普通条目来查询，会给出一份怎么用也太长的链接点清单。尝试更专门化的条目，如classic car等，就可以进一步缩小查询范围。

另外，许多查询形式允许设置附加选项从而缩小查询的范围。有些查询工具，设置成找出满足任一查询条件的所有条目。所以，如果输入football game，查询工具会找出一堆同football以及同game有关的文件。所以，要指定查询的是那些同时和football和game都相关的条目。



可以通知查询工具找出同查询条目准确匹配的文件。例如，如果为“basketball”进行一项非准确匹配查询，查询工具可能会列出一个清单上面有诸如“basket weaving”，“basket”等内容的东西。要准确地同basketball 匹配，必须指明要准确匹配。



当查询结果太多或者不是您想要的条目时，您可以将必然相邻的词或词组用引号括起来。例如，用引号加于"space shuttle"，搜索关于outer space和various spaces closer to home的页面，将只返回有关space shuttle的页面。



关键字中应当尽量不用通用或一般性的词语，而用特定的词汇和那些使描述更细化的词。如program这一个词，可以指好多事情，如电视节目或者软件的应用程序，应该去掉它而代之以更具体的词。

也可以从您的搜索结果中浏览，选择与您想找的内容相近的条目，它的题目和摘要可能会给您一些线索，深入挖掘之。

2、拓宽查询范围

搜索结果太少或未找到需要的条目时，可以加入关键字的同义词或近义词来扩大查询结果。例如搜索bed and breakfasts inns in Northern California相关内容太少时，可以试一下bed and breakfasts inns "small hotels" in Northern California.



3、使用逻辑符号

使用逻辑符号可以有效地缩小查询范围。这在我们上面的例子中大家可以看到。



4、使用多个搜索工具

当用一个搜索工具查询条目效果不太理想时，可以尝试换一个别的搜索工具。这是因为各个搜索服务器虽然功能大体相同，但其检索方式，内容分类及其信息资源侧重点还是有所差别的，因此我们利用不同的搜索工具就可能会有不同的结果。

六、搜索引擎的分类

尽管目前存在数量众多的搜索引擎，但根据它们所基于的技术原理，可以把它们分成三大主要类型：基于Robot的搜索引擎、目录（Directory，也叫做Catalog）和Meta搜索引擎。



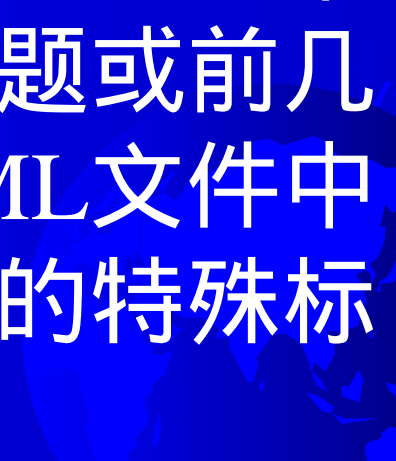
1、基于Robot的搜索引擎

这种搜索引擎的特点是利用一个称为Robot（也叫做Spider、Web Crawler或Web Wanderer）的程序自动访问Web站点，提取站点上的网页，并根据网页中的链接进一步提取其它网页，或转移到其它站点上。Robot搜集的网页被加入到搜索引擎的数据库中，供用户查询使用。Internet上最早出现的搜索引擎就是利用Robot来建立数据库，“搜索引擎”这个词的原义也只是指这种狭义上的基于Robot的搜索引擎。

基于Robot的搜索引擎由三个主要部分构成：Robot、Index和搜索软件。Robot从一个事先制定好的URLs列表出发，这个列表中的URLs通常是从以往访问记录中提取出来的，特别是一些热门站点和"What ' s New"网页，从Usenet等地方检索得到的URLs也常被用作起始URLs，此外，很多搜索引擎还接受用户提交的URLs，这些URLs也会被安排在列表中供Robot访问。Robot访问了一个网页后，会对它进行分析，提取出新的URLs，将之加入到访问列表中，如此递归地访问Web。

Robot作为一个程序，可以用C、Perl、Java等语言来编写，可以运行在Unix、Solaris、Windows、NT、OS2和MAC等平台上。Robot设计是否合理将直接影响它访问Web的效率，影响搜索数据库的质量，另外，在设计Robot时还必须考虑它对网络和被访问站点的影响，因为Robot一般都运行在速度快、带宽高的主机上，如果它快速访问一个速度比较慢的目标站点，就有可能导致该站点出现阻塞甚至当机。Robot还应遵守一些协议，以便被访问站点的管理员能够确定哪些内容能被访问，哪些不能。

Index是一个庞大的数据库，Robot提取的网页将被放入到Index中以便建立索引，不同的搜索引擎会采取不同方式来建立索引，有的对整个HTML文件的所有单词都建立索引，有的只分析HTML文件的标题或前几段内容，还有的能处理HTML文件中的META标记或其它不可见的特殊标记。



基于Robot的搜索引擎一般要定期访问大多数以前搜集的网页，刷新Index，以反映出网页的更新情况，去除一些死链接，网页的部分内容和变化情况将会反映到用户查询的结果中，这是基于Robot的搜索引擎的一个重要特征。



Index在建立索引时，一般会给网页中每个关键词赋予一个等级值，表示该网页与关键词之间的符合程度。当用户查询一个关键词时，搜索软件将搜索Index，找出所有与关键词相符合的网页，有时候这些网页可能有成千上万，等级值的用途就是作为一种排序的依据，搜索软件将按照等级值从高到低的顺序把搜索结果送回到用户的浏览器中。

不同的搜索引擎在计算等级值时使用了不同的方法，但它们都以关键词在网页中出现的位置和频率为基本依据，例如，关键词出现在标题中的网页可能比只出现在其它地方的网页更符合要求，关键词出现在网页的前面可能比只出现在网页的后面更符合要求，同一个关键词出现多次的网页又可能比只出现一两次的网页更符合要求，把这些因素综合起来考虑便可得出一个计算等级值的公式。

不过，绝大多数搜索引擎都没有只按照上述因素来确定计算公式，它们还加入了一些特殊考虑，例如，Excite能检查是否有很多链接指向同一个网页，如果是的话，它就把这个网页的等级值稍微提高一些，理由是这样的网页一般都具有更大的访问量。

。



尽管各个搜索引擎都有一套复杂的等级值计算公式，但仅仅依靠一个数值并不能真正反映出网页的质量，事实上，有些网页在设计时就考虑到了Index的特点，故意使用一些技巧让自己得到很高的等级值，以便能排在查询结果的前列，达到提高访问量的目的。



2、目录

目录与基于Robot的搜索引擎所不同的是，目录的数据库是依靠专职编辑或志愿人员建立起来的，这些编辑人员在访问了某个Web站点后撰写一段对该站点的描述，并根据站点的内容和性质将其归为一个预先分好的类别，把站点的URL和描述放在这个类别中，当用户查询某个关键词时，搜索软件只在这些描述中进行搜索。很多目录也接受用户提交的网站和描述，当目录的编辑人员认可该网站及描述后，就会将之添加到合适的类别中。

目录的用户界面基本上都是分级结构，首页提供了最基本的几个大类的入口，用户可以一级一级地向下访问，直至找到自己感兴趣的类别，另外，用户也可以利用目录提供的搜索功能直接查找一个关键词，不过，由于目录只在保存的对站点的描述中进行搜索，因此站点本身的动态变化不会反映到搜索结果中来，这也是目录与基于Robot的搜索引擎之间的一大区别。



商业性质的目录一般都是依靠一群专职编辑来建立和维护的，最出名的商业目录Yahoo！雇用了大约一两百名编辑，他们维护的目录一共收集了上百万个站点。不少学术或研究性质的目录是依靠志愿者来建立和维护的，这些志愿者可能是普通的Internet用户，也可能是一群大学生，还有可能是专家学者，1998年才成立的Open Directory采取了开放管理模式，所有Internet用户都可以申请成为它的志愿编辑，目前Open Directory的编辑人员已超过了14000人。

由于目录是依靠人工来评价一个网站的内容，因此用户从目录搜索得到的结果往往比从基于Robot的搜索引擎得到的结果更具参考价值，Yahoo！能取得成功，与此有着莫大的关系。事实上，现在很多搜索站点都同时提供有目录和基于Robot的搜索服务，以便尽可能地为用户提供全面的查询结果。



3、Meta搜索引擎

Meta搜索引擎也叫做Multiple Search Engine，它的特点是本身并没有存放网页信息的数据库，当用户查询一个关键词时，它把用户的查询请求转换成其它搜索引擎能够接受的命令格式，并行地访问数个搜索引擎来查询这个关键词，并把这些搜索引擎返回的结果经过处理后再返回给用户。



严格意义上来讲，Meta搜索引擎只能算是一种用户代理，而不是真正的搜索引擎。多数Meta搜索引擎在处理其它搜索引擎的返回结果时，只提取出每个搜索引擎的结果中前面10~50条，并将这些条目合并在一起返回给用户，因此最后结果的数量可能会远少于直接在一个搜索引擎上进行查找所得到的数量，这就是为什么很多Internet用户都喜欢使用Meta搜索引擎来查找信息的原因。



Meta搜索引擎实现起来比较简单，但是它也有一定的局限性，例如多数Meta搜索引擎都只能访问少数几个搜索引擎，并且通常不支持这些搜索引擎的高级搜索功能，在处理逻辑查询时也常常会出现错误。



七、搜索引擎的现状

经过了多年的发展之后，现在的搜索引擎功能越来越强大，提供的服务也越来越全面，它们的目标是把自己发展成为用户首选的Internet入口站点，而不仅仅是提供单纯的查询功能。



1、目录和基于Robot的搜索引擎相结合

由于目录和基于Robot的搜索引擎有各自的优点和缺点，目前它们谁也无法完全取代谁，于是很多搜索站点都同时提供这两种类型的服务。




例如Yahoo！主要是一个目录，但它也从有名的搜索引擎服务商Inktomi那里获取网页搜索结果，当用户查询一个关键词时，Yahoo！首先返回从目录中查到的匹配项（由<http://search.yahoo.com/>来完成），如果用户对结果不满意，或者目录中没有匹配项，那么用户还可以继续查找与关键词匹配的网页（由<http://ink.yahoo.com/>来完成）。

国内两个有名的中文搜索引擎搜狐和Yeah也都是这种模式。Infoseek则主要是一个基于Robot的搜索引擎，但它同时也建立了一个由人工编辑的小型目录。



2、 多样化和个性化的服务

现在绝大多数搜索引擎都提供多样化的服务，以吸引更多的用户，商业搜索引擎尤其注重这一点。以Yahoo！为例，用户可以从它的首页中查看新闻、金融证券信息、天气预报、黄页，可以进行网上购物、拍卖、找人，或者使用免费Email和网上寻呼等服务。



近期多个搜索引擎已开始提供个性化的服务，例如Yahoo！的“My Yahoo！”、Infoseek的“Personalized start page”、Lycos的“My Lycos”等，它们允许用户为自己定制起始页面，并选择感兴趣的内容和经常使用的服务放在该页面中。



3、 强大的查询功能

与最早的搜索引擎相比，现在的搜索引擎在查询功能方面已有了很大的改进。除了简单的AND、OR和NOT逻辑外，不少搜索引擎还支持相似查询，例如AltaVista、Northern Light、Lycos等支持短语查询，AltaVista的高级搜索功能支持NEAR逻辑等。



域搜索也是一项很实用的功能，它允许用户把查询范围限制在网页的某个域中，例如标题、URL、图像标记或链接等，AltaVista、Northern Light和Infoseek等搜索引擎都支持对网页的多种域进行搜索。



八、未来的发展趋势

搜索引擎未来的发展面临着两大难题：一是如何跟上Internet的发展速度，二是如何为用户提供更精确的查询结果。

。



近几年来Internet迅速扩张，其上的站点和网页越来越多，预计到2000年，Internet上的文档数量将超过10亿份，而现在最大的两个搜索引擎Northern Light和AltaVista只分别给1.6亿和1.5亿份网页建立了索引，不到Internet现有网页总数的一半，最大的目录Yahoo!也只收集了120万个左右的站点。



另一方面，当搜索引擎的数据库越来越大时，用户查询同一个关键词所得到的结果也就越来越多，然而成千上万的结果对用户并没有什么实际意义，用户关心的是能否迅速在开头几十个结果中找到自己需要的信息。面对着这两个难题，目前很多搜索引擎都在发生一些变化，这些变化中包含着搜索引擎未来的发展趋势。



1、目录占据主导地位

目录与基于Robot的搜索引擎相比更具优越性，这一点已为大多数人所接受，在今后的发展中，目录将会占据主导地位，而基于Robot的搜索引擎将更多地以辅助工具的面貌出现。



2、并行处理技术日趋重要

基于Robot的搜索引擎必须随着Internet的发展不断扩大自己的网页数据库，由此会产生很多技术难题，例如怎样及时地获取新网页和刷新数据库，当数据库增大之后如何保证查询效率不会明显降低等，目前要解决这些问题只有依赖于设计合理的并行处理技术。

Inktomi是一家专门向其它搜索站点提供搜索引擎服务的公司，它的搜索引擎在并行处理技术方面有独到之处，整个系统由100台以上的SUN工作站组成，并可根据需要方便地进行扩展，这些工作站相互之间是平等的，它们能独立地搜集网页，建立数据库。当其它搜索站点传来一条查询指令时，一台工作站被随机选中，它将向其它工作站广播查询指令，这样每台工作站都在自己的库中进行搜索，并把结果返回给选中的工作站，这台工作站再把结果汇总起来返回给用户。

Inktomi的这种并行处理技术能够有效地适应Internet的扩张，现在已有多家搜索站点正在或准备使用它的服务，其中包括Yahoo!、HotBot、MSN Search、AOL NetFind、GoTo等，预计今后Inktomi会争取到更多的客户。



3、特殊搜索引擎越来越多

搜索引擎的另外一个发展趋势是特殊搜索引擎越来越多，这些特殊搜索引擎只收集了某个方面的网站或网页，例如文学、医学、体育、音乐、MP3、软件等等，其中的内容一般都要比通用搜索引擎更好更精，因此很受用户的欢迎。建立特殊搜索引擎的成本要远小于通用搜索引擎，这也促进了它的发展。

第二节 Web Mining


第二代网络信息处理技术

- 一、概述
- 二、网络信息挖掘的步骤
- 三、网络信息挖掘中的关键技术
- 四、总结网络信息挖掘系统



一、概述

随着Internet的飞速发展，网络信息过载（InformationOverload）问题日益突出，以Yahoo为代表的网络信息检索系统出现并迅速发展起来。网络信息检索系统一般由Robot、索引数据库和查询引擎三部分组成。



信息搜集器Robot对WWW进行遍历，尽可能多地发现新的信息；采用全文检索技术对搜集到的信息建立索引存入索引数据库中，能够极大地提高信息检索的速度；查询引擎接收并分析用户的查询，根据较为简单的匹配策略（简单布尔模型或模糊布尔模型）遍历索引数据库，最后将结果地址集提交给用户。由于人工智能研究水平的限制，目前Robot还无法实现信息的准确分类，多数搜索站点都是通过人工方式对信息进行二次处理，信息整理的速度远远落后于网络信息的膨胀。

为了实现个性化的主动信息服务，网络信息挖掘（WebMining）技术成为新的研究热点。网络信息挖掘是指在已知数据样本的基础上，通过归纳学习、机器学习、统计分析等方法得到数据对象间的内在特性，据此采用信息过滤技术在网络中提取用户感兴趣的信息或者更高层次的知识和规律。



网络信息挖掘与网络信息检索所采用的技术有很多相似之处，但又有本质的不同。作为第二代网络信息处理技术，网络信息挖掘技术沿用了Robot，全文检索等网络信息检索中的优秀成果，同时综合运用人工智能、模式识别、神经网络领域的各种技术。网络信息挖掘系统与网络信息检索的最大不同在于它能够获取用户个性化的信息需求，根据目标特征信息在网络上或者信息库中进行有目的的信息搜寻。

二、网络信息挖掘的步骤

1、 确立目标样本：

由用户选择目标文本，作为提取用户的特征信息。

2、 建立统计词典：

建立用于特征提取和词频统计的主词典和同义词词典、蕴含词词典。



3、 特征信息提取：

根据目标样本的词频分布，从统计词典中提取出挖掘目标的特征向量并计算出相应的权值。

4、 调整特征矢量：

根据测试样本的反馈调整特征项权值和匹配阈值。



5、 网络信息获取：

先利用搜索引擎站点选择待采集站点，再利用Robot程序采集静态Web页面，最后获取被访问站点网络数据库中的动态信息。

6、 信息特征匹配：

提取源信息的特征向量，并与目标样本的特征向量进行匹配，将符合阈值条件的信息提交给用户。



三、网络信息挖掘中的关键技术

1、目标样本的特征提取

系统采用向量空间模型(VSM: VectorSpaceModal), 用特征词条及其权值代表目标信息, 在进行信息匹配时, 使用这些特征项评价未知文本与目标样本的相关程度。

特征词条及其权值的选取称为目标样本的特征提取，特征提取算法的优劣将直接影响到系统的运行效果。词条在不同内容的文档中所呈现出的频率分布是不同的，因此可以根据词条的频率特性进行特征提取和权重评价。



一个有效的特征项集应该即能体现目标内容，也能将目标同其它文档相区分，因此词条权重的正比于词条的文档内频数，反比于训练文本内出现该词条的文档频数。



与普通的文本文件相比，HTML文档中有明显的标识符，结构信息更加明显，对象的属性更为丰富。系统在计算特征词条权值时，充分考虑HTML文档的特点，对于标题和特征信息较多的文本赋予较高权重。为了提高运行效率，系统对特征向量进行降维处理，仅保留权值较高的词条作为文档的特征项，从而形成维数较低的目标特征向量。

2、 中文分词处理

西文的句子以空格作为固定的分隔符，而中文中没有，这给中文信息处理带来很大障碍，例如机器无法分辨“白天鹅”到底是“白天”和“鹅”，还是“白的天鹅”，因此在进行词频统计等处理前先要进行词条切分处理。比较简单有效的分词方法是基于大型词库的机器分词法。

通用词库包含了大量不会成为特征项的常用词汇，为了提高系统运行效率，系统根据挖掘目标建立专业的分词表，这样可以在保证特征提取准确性的前提下，显著提高系统的运行效率。



进行词条切分时，先根据标点进行粗切分，然后再分别使用正向和逆向最大匹配法进行细切分。在进行词频统计时，考虑到自然语言的多样性，系统建立并使用相应的同义词库、蕴含词库等辅助词库，以提高信息匹配的准确度。



3、 获取网络中的动态信息

Robot是传统搜索引擎的重要组成部分，它依照HTTP协议读取Web页面并根据HTML文档中的超链在WWW上进行自动漫游，Robot也被称为Spider、Worm或Crawler。




但Robot只能获取Web上的静态页面，而有价值的信息往往存放在网络数据库中，人们无法通过搜索引擎获取这些数据，只能登录专业信息网站，利用网站提供的查询接口提交查询请求，获取并浏览系统生成的动态页面。网络信息挖掘系统则通过网站提供的查询接口对网络数据库中的信息进行遍历，并根据专业知识库对遍历的结果进行自动的分析整理，最后导入本地的信息库。

4、 信息的分类

为了更有效的对信息建立索引，需要对信息进行分类处理，系统采用 Naive Bayes 法实现此功能。Naive Bayes 分类法假设所有词条在文档中的出现概率相对独立并且文档的类别同长度无关，判别原则是将文档 D 指定到使 $P(C_i|D)$ 达到最大概率的 C_i 类中，即求解 $\operatorname{argmax} P(C_i|D)$ ， $P(C_i|D)$ 是给定文档 D 属于文档类 C_i 的概率。

四、总结网络信息挖掘系统

根据用户所提供的目标样本和系统设置，提取目标的特征信息，根据目标特征自动在WWW上搜集资料，然后对所搜集到的资料进行分类整理并导入资料库。系统能够自动运行，不断更新用户的资料库，提供个性化的主动信息服务。



第三节

网上信息搜索的利器： 元搜索引擎及其特色

网上信息资源的膨胀发展，对于资源搜索引擎的检索机制与能力提出了新的要求。这使得搜索引擎的数量在迅速增加，检索方式日益复杂。



专家关于使用搜索引擎的唯一的而且经常的建议，是利用不止一个搜索引擎来解答问题。因为没有哪两个搜索引擎是完全相同的--每一种都有自己的检索特色，都有自己的索引，以不同的方式在网上搜寻网址。出现不同的检索结果丝毫不足为奇。从不同的搜索引擎的检索结果中综合出最为符合要求的答案，对于熟练的检索人员而言，可能不是什么难题，但是对于一般的网上信息搜集者来讲，肯定比较困难。

因此如何准确选择搜索引擎、如何减轻学习与操作负担、如何有效利用多个搜索引擎的“集成”资源与检索能力，就成为制约网络信息检索技术进一步优化和发展的重要问题。正是面对这个挑战，检索工具开发者设计了元搜索引擎(Meta-SearchEngine)。



- 一、什么是元搜索引擎?
- 二、搜索引擎和元搜索引擎的区别
- 三、元搜索引擎的分类
- 四、元搜索引擎的特色



一、什么是元搜索引擎?

元搜索引擎，通过一个统一用户界面帮助用户在多个搜索引擎中选择和利用合适的(甚至是同时利用若干个)搜索引擎来实现检索操作，是对分布于网络的多种检索工具的全局控制机制。



元搜索引擎的出现，对于那些需要连续地使用不同的搜索引擎重复相同的检索的人来说，是一个福音。使用元搜索引擎同时对几个搜索引擎进行检索，获得分级编排的检索结果。检索人员就象采用在国际联机检索中常用的，利用411文档进行一次多库预检一样。仅从一个搜索界面，检索人员可以选取几个搜索引擎，然后构建检索式。

二、搜索引擎和元搜索引擎的区别

我们可将元搜索引擎看成具有双层客户机 / 服务器结构的系统，用户向元搜索引擎发出检索请求。元搜索引擎再根据该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请求后将检索结果以应答形式传送给元搜索引擎，元搜索引擎将从多个搜索引擎获得的检索结果经过整理再以应答形式传送给实际用户。当然，某些元搜索引擎具有略微不同的机制。

搜索引擎与元搜索引擎的主要区别在于搜索引擎拥有独立的网络资源采集标引机制和相应的数据库，而元搜索引擎一般没有自己独立的数据库，却更多地是提供统一联接界面(或进一步地提供统一检索方式和结果整理)，形成一个由多个分布的、具有独立功能的搜索引擎构成的虚拟整体，用户通过元搜索引擎的功能实现对这个虚拟整体中各独立搜索引擎数据库的查询显示等一切操作。

元搜索引擎中各独立搜索引擎被称为“目标搜索引擎”，或者“成员搜索引擎”，它们各自保持其原来的局部数据模式和自己的检索指令；元搜索引擎给出一个全局外部模式，用以接受用户检索输入和结果输出。不过，有些元搜索引擎给出的全局外部模式不够完善。



三、元搜索引擎的分类

在可以检索的目标搜索引擎、检索提问的处理方式以及如何编译和显示结果方面，元搜索引擎有着很大的差异。有些元引擎一个接一个的搜索目标搜索引擎，另一些则同时进行搜索。有些搜索引擎将检索提问转变成目标搜索引擎的提问语言，而有一些则原封不动的发送给目标引擎。



按功能划分，元搜索引擎包括多线索式搜索引擎和All-in-One式搜索引擎；按运行方式的差异可分为在线搜索引擎和桌面搜索引擎。



1、多线索式元搜索引擎

多线索式元搜索引擎，指利用统一的检索界面，实现对多个独立搜索引擎索引数据库进行检索，并将检索结果以统一格式显示的网络检索工具。Metacrawler、Savvysearch、Profusion等都属于多线索式元搜索引擎。这类元搜索引擎一般具有以下特征：

统一检索界面：元搜索引擎提供统一界面，提供对各搜索引擎特点介绍和选择机制，但所有目标搜索引擎构成一个逻辑整体，元搜索引擎检索界面构成唯一的全局外部检索模式，用户通过这个全局界面实现对多个或任意一个搜索引擎的检索。



检索指令转换：在具有唯一全局外部检索模式情况下，系统可提供统一的全局指令语言，并自动地实现元搜索引擎指令与其目标搜索引擎指令的转换，用户使用同一指令语言检索不同的搜索引擎的索引数据库。

统一结果集的组织与显示：元搜索引擎提供全局组织器，对各目标搜索引擎返回的结果进行处理，形成全局结果集，并以统一格式显示，主要涉及数据格式转换、去重、统一排序等。

2、 All-in-one方式

All-in-One方式，是指元搜索引擎界面以任意顺序或分类罗列多个(一般都是数十个)搜索引擎，元搜索引擎本身主要提供各类搜索引擎的介绍信息和物理连接机制。用户可通过这类元搜索引擎了解有关的搜索引擎、联入所选择的搜索引擎。但元搜索引擎没有统一的全局外部模式，而是以各搜索引擎的检索模式和数据格式直接面对用户。

例如 All-in-one元搜索引擎
([WWW.albany.net / allinone.html](http://WWW.albany.net/allinone.html))。这种 All-in-one方式的元搜索引擎确切地说只是搜索引擎的罗列，它们具有以下特点：

仅仅提供一个简单的界面来帮助用户选择和使用各搜索引擎；只能选择一个搜索引擎进行检索；对各目标搜索引擎检索界面的复制可能是部分的或全部的；直接利用所选搜索引擎的显示格式呈送给用户。

3、 桌面元搜索引擎

以上各类元搜索引擎都是通过网上调用方式在线使用，还有另外一类元搜索引擎可直接在用户计算机上运行，相当于用户自己拥有一个元搜索引擎，称之为桌面元搜索引擎。这些桌面元搜索引擎可从网络上下载。



桌面元搜索引擎是一个包括多个成员搜索引擎的完整系统，它们往往允许用户自定义检索式运行的搜索引擎集合(例如一个或全部目标搜索引擎)，甚至可由用户添加新的搜索引擎，例如EchoSearch和WebCompass。这些桌面元搜索引擎不仅可以实现对多个搜索引擎的并行检索，而且也能提供重要的后期处理功能。例如用户定义结果排序方式、删除重复记录等功能。

四、元搜索引擎的特色

1、目标引擎的数量和名称

确定一个元引擎能够检索多少目标引擎以及哪些目标引擎，听起来象是一个简单的问题。确实，对于某些元引擎来讲，浏览一下页面首页的checkbox，就可以得出答案。但是对于其他许多元搜索引擎而言，这些细节隐藏在 Help 页面中，有时根本就没了这些内容。多不见得意味着好。有些元引擎可以让使用者选择被使用得最多的前8-10个目标引擎，诸如Lycos、HotBot、Alta Vista、Excite 等等。

最好的引擎提供一个简便的浏览列表，使用者可以从中进行选择。

SavvySearch允许检索者选择不同的目标引擎，还可以自行确定这些引擎的使用顺序，这些设置可以保存起来，以备将来之用。ProFusion提供9个目标引擎，使用者可以从中进行选择，或者要求系统提供"best 3"、"fastest 3"来完成检索。Dogpile的定制检索可以让使用者确定目标引擎的使用顺序，然后同时对3个目标引擎进行检索，完成之后还可以按照检索者的要求继续对下3个引擎进行检索。

在选择元搜索引擎时，可以考虑那些明确列出了目标引擎、并容许使用者自行组合使用目标引擎的元搜索引擎。通过 cookies 实现的保存用户的检索设置，就象 SavvySearch所提供的那样，是一项相当不错的功能，其他的元搜索引擎可能很快就会提供这样的功能。当然，由于SavvySearch提供了100个目标搜索引擎，与其它仅仅提供5-10个目标引擎的元搜索引擎相比，这样的功能确实显得更为重要。



2、其他资源和专门的引擎

除了能检索目标引擎的数据库之外，许多元搜索引擎还能搜索网页的其他部分。通常，可以选择Web或Usenet，或者选择newswires、DejaNews或其他资源。比较频繁的是，主题或者分类类目可以提供这样的机会，使我们能在某一主题领域进行搜寻，而不必使用一个宽泛的搜索引擎来对网页的全部内容进行查找。



许多元搜索引擎被很好的用作特殊搜索的导航器，而不仅仅只是一个泛泛的查询工具。我们可以关注一下众多的专题搜索引擎，然后再来看看元搜索引擎首页所提供的局部的、专题的或特殊的链接，就可以明白这个道理。



3、 检索提问

许多元搜索引擎可以容许我们构建自己的提问式，其语法结构与流行的搜索引擎类似，大多数还有布尔逻辑选项。在使用某一个元搜索引擎时，最重要的问题就是，该搜索引擎是否能够把我们的检索提问“翻译”成目标引擎所遵循的语法结构，或者仅仅只是原样照搬。

一些元搜索引擎(我们称之为伪元搜索引擎，因为他们仅仅只是把众多目标引擎集合到一起，而并没有将他们的检索功能集成)显示一系列提问框。每一个提问框对应一个元搜索引擎。我们必须逐个输入检索提问，而且还必须使用相应的语法规则，然后点击"Submit"按钮，分头检索不同的目标引擎。Beaucoup Search Engines提供了14个这样的选项。

我们应该利用那些能将我们的检索提问连释成目标引擎的检索语言的元搜索引擎，否则就会大大降低检索的效能，无法发挥提问特色的优势。



4、其他检索选项

在线路堵塞或服务器繁忙的情况下，是否有一个“超时”的选项框，能够让我们确定，多长时间以后，例如10、15或30秒，元搜索引擎才放弃对目标引擎的搜寻？

元搜索引擎是按照目标引擎列表框中的顺序对目标引擎进行检索，还是同时进行检索？在处理时间以及结果的返回方式上，对于这个问题的答案是不同的。

如果我们在构建检索式时有布尔逻辑或自然语言 / 主题词的选项，我们应该找一找还有什么其他参数。



5、 结果选项

大多数元搜索引擎按照关联度来排列检索结果，其他的显示特色则相差很大。比较普遍的方式是，将来自不同目标引擎的检索结果集中到一起，并显示每条结果的来源。很多元搜索引擎对结果进行去重。在1999年，这看来是一件简单的事，但是传统的联机检索人员应该记得在过去对不同文档的命中记录进行去重是多么的差强人意。Cyber411按照目标引擎来组织结果，但进行去重处理，为我们提供了一个干净、实用的结果列表

一些结果列表仅仅显示指向目标页面的简单标题，也有一些显示题目和描述，与目标引擎的结果显示方式类似。有些结果列表还显示每一条结果在目标引擎的命中结果中的排序位置，如 # 10 on Excite 或 # 1 on HotBot。这种特色很不错，但是未必有必要，因为元搜索引擎对结果按照关联度顺序显示。



有一两个元搜索引擎，如 InferenceFind，按照类目排列命中结果。而且显示非常简单的标题，我们没有足够的信息来决定点击哪一个链接。VerioMetasearch的高级检索功能允许我们指定来自8个目标引擎的10条命中结果的权重。返回的结果则显示其排序及分数。



最好的元搜索引擎显示集成的、信息充分的按照关联度算法排序的结果。最低限度地，结果列表应该包含标题、URL、描述以及结果的来源。要避免使用那些直接调用目标引擎的结果显示页面的元搜索引擎，除非你想顺着元搜索引擎逐个目标引擎的浏览检索结果。



6、创建自己的网络搜索服务模式

确定一个首选的元搜索引擎与选择一个独立的搜索引擎一样复杂。但是一旦选择了一个符合自己的检索特色的元搜索引擎，并且按照自己的要求进行定制，我们就可以高效地检索一系列搜索引擎，如果需要的话，可以对其中一两个目标引擎进行扩展检索。

在返回结果的精确性方面，元搜索引擎不会好过独立的搜索引擎。但是由于它能够简便的检索多个独立的搜索引擎，应该成为我们进行网络搜寻的经常之选。

第四节

搜索引擎挑战智能化

传统的搜索引擎不能适应信息技术的高速发展，新一代智能搜索引擎作为一种高效搜索引擎技术在当今网络信息时代日益引起人们的关注。



一、搜索引擎面临的挑战

1. 网络信息量迅猛增加，人工无法对它们进行有效的分类、索引和利用。Internet 为实现多种交互和交易提供了电子通信平台。对企业而言，他们需要处理比以往更多的信息，传统处理方式已经不能承受如此重负而使得决策缓慢、效率低下。对普通用户而言，简单的关键词搜索，返回的信息数量之大，往往让用户无法承受。

2. 网络信息组织的无序性。Internet用户面对的是非常多的随机的未组织的信息。从如此庞杂的信息海洋中取出对用户最有用的信息是搜索引擎面临的一项挑战，而信息的有序化组织也是搜索引擎高效工作的前提。



3. 信息有用性评价困难。一些站点在网页中大量重复某些关键字，使得容易被某些著名的搜索引擎选中，以期借此提高站点的地位，但事实上却可能没有提供任何对用户有价值的信息。这些情况更加加深了评价信息有用性的难度。



4 . 网络信息日新月异的更变。人们总是期望挑出最新的信息，然而网络信息时刻变动，实时搜索几乎不可能。就是刚刚浏览过的网页，也随时都有更新、过期、删除的可能。好的搜索引擎必须在速度和效率上进行仔细的权衡。



5 . 信息媒体多样化。迄今为止，搜索对象主要是文本。多媒体技术的发展，对搜索引擎提出了更多的要求。人们期望引擎不仅能挑出自己需要的文章，还能挑出自己所关心的图片、电影、音乐等。



6. 带宽等其他因素的制约。搜索引擎的关键问题之一就是如何将网络信息收集与整理，也就是如何将网络信息有序化。为此，搜索引擎需要定期不断地访问网络资源。然而，遍历如此庞杂的网络本身就是一件非常困难的事情。目前网络带宽不足，网络速度不够理想，使得搜索引擎搜索网络资源的速度较慢。

举个例子，假定网络连接良好，连接一个网站需要3秒钟，那么处理一个B类网段如清华大学的166.111.0.0 ~ 166.111.255.255共65536个IP地址就需要 $65536 \times 3 \text{秒} = 1996608 \text{秒} = 54.61 \text{小时}$ ，也就是说，即使是在网络速度比较快的情况下，把整个B类网段遍历一次，就需要两天多的时间。



二、智能搜索引擎的特征

智能搜索引擎设计追求的目标是：根据用户的请求，从可以获得的网络资源中检索出对用户最有价值的信息。一般而言，智能搜索引擎有3个主要的特征：



1、 网络蜘蛛的智能化

网络蜘蛛通过启发式学习采取最有效的搜索策略，选择最佳时机获取从Internet上自动收集、整理的信息。众所周知，信息动态更替无时无刻不在进行，即使在搜索过程中，文档会被添加、删除、改变。因此，智能引擎有一个设计网络蜘蛛，自动完成在线信息的索引。

搜索引擎能在Internet或Intranet的任何地方工作，能尽可能地挖掘和获得信息。网络蜘蛛既可收集特定站点的信息，又能遍历整个Internet，对整个Internet进行索引。为了提高搜索速度，智能搜索引擎可以同时启动多个引擎并行工作，将各个引擎的搜索结果整合，作为一个整体存放到数据库中。




此外，智能搜索引擎具有跨平台工作和处理多种混合文档结构的能力。譬如既能处理HTML（HyperText Markup Language，超文本标志语言），又能处理SGML（Standard for General Markup Language，通用标志语言标准）和XML（eXtended Marked Language，扩展标志语言）文档以及其他类型的文档，譬如Word、WPS等。



同时，智能搜索引擎还具有高的召回率和准确率。所谓召回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。所谓准确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。

最后，智能搜索引擎应该可以支持多语言搜索，允许用户可以用中文输入查询英文或其他语言的信息。



2、为特定用户提供相关信息

智能搜索引擎能通过观察用户的行为，了解用户的兴趣爱好，另外能通过不断的训练学习增长智能。每次用户对引擎返回的信息进行评价，智能引擎根据用户的评价调整自己的行为。智能搜索引擎还能对搜索结果进行合理的解释。



智能搜索引擎具有主动性，可以在任何特定的时候（如用户最关心的信息发生了某种变化的时候）用各种方法与用户取得联系，这些方法包括电子邮件、电话、传真、寻呼机、移动电话等。搜索引擎还可根据用户特定时刻的位置信息，选择恰当的方法跟用户通信。




3、搜索引擎人机接口的智能化

智能搜索引擎可以通过自然语言和用户交互。它采取诸如语义网络等智能技术，通过汉语分词、句法分析以及统计理论有效地理解用户的请求，甚至能体会出用户的弦外之音，最大程度地了解用户的需求。



三、智能搜索引擎的技术

要想真正实现如上所述的智能搜索引擎还有大量的工作要做。一种比较实际的做法是将智能技术跟传统搜索引擎结合，逐步实现智能化。下面就是搜索引擎在向智能化迈进的过程中可以采取的一些技术。



三、智能搜索引擎的技术


1、 汉语分词技术

我们知道，关键词查询的前提是将查询条件分解成若干关键词，同时一些关键词表示文档。对英文而言，一个单词就是一个词。但中文就没有这么简单，主要问题是中文词与词之间没有界定符，需要人为切分。

此外汉语中存在大量的歧义现象，对几个字分词可能有好多种结果。简单的分词往往会歪曲查询的真正含义。譬如查询条件为“中国人”，若不能正确地分词，按“中国”、“人”、“中国人”等3个关键词去搜索，这样搜索结果的质量就可想而知了。因此，可以根据语料库进行总结，获得每个词的出现概率以及词与词的关联信息，就可能有效地排除各种歧义，大幅度提高分词的准确性，从而准确地表述查询请求和文档信息。

2、短语识别

用短语描述查询请求的情况很常见。譬如查询条件“北京的气温”，“北京”和“气温”存在一定的关系，但如果不将“北京”和“气温”联合起来作为一个短语查询那么除了选出关于“北京的气温”的文档之外，还将查出有关“北京”和“气温”的文档。因此，短语识别也是智能化引擎所关注的一个技术。



3、处理同义词

处理同义词的一种方法是人工构造同义词表。对专用领域的搜索引擎，这种方法是非常有效的。另外一种方法是从语料库中自动取得同义词关系。给出一个查询的关键词，引擎能主动“联想”到与其同义或意思相近的词。



4、文档信息压缩

存储文档信息的Word矩阵如果不经过压缩处理，将需要巨大的存储容量。采取诸如SVD（Singular Value Decomposition，奇异值分解）等矩阵分解技术，构造出新的基向量组，从该向量组中挑出若干主要基向量构成新的向量空间，将原来的向量向新的向量空间投影，这样便可以大大减少存储量，有效地提高处理速度。

在搜索过程中经常还会碰到这样的情况，用户本身有时候也不清楚自己究竟需要获取什么信息。因此，让用户对返回的结果进行选择，挑出真正所需，然后根据用户挑出的文档，对查询条件进行修正，进行二次查询是一种非常有效的手段。此外，还有一些其他的人工智能方法，譬如可以将知识库和推理机应用到搜索引擎中等，所有这些都是实现信息时代对搜索引擎智能化挑战的有效手段。

第五节

流行的中文搜索引擎

传统的搜索引擎不能适应信息技术的高速发展，新一代智能搜索引擎作为一种高效搜索引擎技术在当今网络信息时代日益引起人们的关注。



一、悠游中文搜索引擎

<http://www.goyoyo.com.cn/main/indexgb.htm>(
北京)

<http://www.goyoyo.sh.cn/>(上海)

<http://www.goyoyo.cq.cn/>(重庆)

<http://www.goyoyo.com.hk/>(香港)

<http://www.goyoyo.com/>(美国)



"悠游"能自动收集英文、中文国标码和大五码的网页，用户使用支持一种中文码的浏览器，便可搜索和阅读所有网页。GOYOYO 有超智能的Robot系统，能对全球的新网页和每日更新的资料进行分秒不停的自动搜索，并可以自动分类识别。目前"悠游"拥有全球40多万个中文网页资料，有5个站点，不过它们并不是拷贝性质的镜像站点，内容各有不同，所以都不要错过。

二、网络指南针

<http://compass.net.edu.cn:8010>

网络指南针收集了CERNet、ChinaNet、CSTNet、ChinaGBN等网络上大量的信息资源，WWW网页达20万页。网络指南针向用户提供中、英文信息查询服务，以及中、英文翻译和拼音转换功能，用户可以选择各种输入方法进行查询。



三、 搜索客

<http://www.cseek.com>

CSEEK的内容广泛，速度也不慢，在技术上融合了国内外成熟的信息采集技术和检索技术，收集了全球各地的中文网络资源，并兼容中文简体(GB)和繁体(BIG-5)两种检索方式，保证检索技术的稳定性和检索信息的广泛性。根据国际标准和中文网络资源的具体实际，Cseek还为检索者编制了详细的分类目录，帮助用户限定检索区域，提高检索效率。

四、YEAH搜索引擎

<http://www.yeah.net>

YEAH中文搜索引擎的内容繁多，五花八门，包含有4000万个链接。其速度飞快，分类索引一目了然，查询语法简便而非常实用，颇具高手风范。它以所包含的关键词数目愈多愈优先的原则排列查询结果，所列网页至少包含一个关键词。

五、天网中英文搜索引擎

<http://pccms.pku.edu.cn:8000/gbindex.htm>

天网WWW资源检索系统是中国教育和科研 计算机网示范工程应用系统课题之一,并被列入CERNet"九五"攻关项目,由北大计 算机系网络研究室设计开发。



这是一个WWW资源索引和查找服务系统，它提供一种检索Web资源(主要是中国教育和科研计算机网上的Web资源)及Newsgroup的手段。目前天网大约收集了60.6万个WWW页面(国内)和9.6万篇Newsgroup文章。本系统主要提供的是针对内容的检索，目前还没有对URL名检索的支持。

六、CEI信息导航

<http://infonavi.cei.gov.cn>

CEI信息导航是目前国内惟一大型的专业化导航应用系统，拥有庞大的信息资源数据库和先进的系统软件。CEI信息导航在国内起步较早，作为“金桥工程”信息部分的主体技术之一，在表层和底层技术方面均相对领先，在CEInet开放环境成功运行一年有余，证明其技术先进、信息资源丰富，在国内仍以高访问率居领先地位。

七、若比邻

<http://www.robot.com.cn>

中科院计算机网络信息中心、中国互联网络信息中心（CNNIC）推出的全新设计的中文搜索引擎“若比邻”具有三大特点：全、易、快。查询站点全,适用面广,查询方式全。操作简便,易学易用。用户可按目录列表及关键词查询等方式查询,由于采用了先进的检索算法和对信息的合理分类,用户在“若比邻”上查询信息时,将不再无奈地等待。

八、北极星

<http://www.beijixing.com.cn>

北极星是由中国科技信息研究所网络中心与巴西AITECH公司合作开发的大型Internet中文搜索引擎，面向中文和国内的信息资源，功能比较齐全，利用它可以进行布尔逻辑检索、中英文自动翻译检索和分类检索等功能。



九、搜狐

<http://www.sohoo.com.cn/>

作为爱特信公司创办的大型中文网络系统，“搜狐”是针对目前国际互联网上中文信息日渐丰富而信息查找却愈加困难的实际情况，根据中国人的文化传统专门为中国用户量身设计的网络分类式查找引擎。



查询结果包括满足条件的目录及站点，信息量大，分类清晰，不愧为网路神探。“搜狐”引擎从中国文化的角度进行了非常精细的分类，而不单纯是机器搜索。分类搜索与关键字检索并重，是搜狐区别于其他中文搜索引擎的重要地方。



十、我是野虎中文搜索引擎

<http://www.5415.com/>

"我是野虎"中文搜索引擎的数据库内包括来自大陆、台湾、香港、新加坡和美国的网址，并备有自行开发的索引式搜索引擎协助使用者进行查询。同时，该网站将其网罗的10余万个网址分门别类，让使用者可以依照网址类别一层一层地逐步找到想看的网站或网页。

十一、雅虎中文

<http://gbchinese.yahoo.com/>

Yahoo!中文版并非英文版的全文翻译，而是针对大陆、香港和台湾的访问者的需要、兴趣与习惯将信息集中起来供中文读者使用。它收录了全球Internet上众多的中文站点，为全球中文读者提供中文Internet 导航服务。



此中文站点在外观上类似于Yahoo!的其他站点，但在体育、文化、艺术和医学等栏目中增加了来自中国内地、香港地区、台湾地区富有地方特色的内容。用户还可以通过Yahoo!中文站点链接到当地新闻机构，如新华社以及西方新闻机构等。



十二、欧姆龙

<http://www.omron.online.sh.cn/>

上海欧姆龙计算机有限公司开发的SEARCH '97网上中文搜索工具应用了美国Verity公司的SEARCH'97全文搜索引擎技术。其搜索方式是基于词的搜索，以高速准确的中文分词算法为基础，符合中文的语言习惯，如在查找“文学”时不会找出“天文学”的文章。

十三、常青藤

<http://www.tonghua.com.cn/>

"常青藤搜索引擎"作为一个智能的中文搜索引擎，汇集了大陆、台湾、香港、澳门、新加坡等中文地域的网络资料，每条记录、每个网址均精心筛选，力争覆盖当地信息资源，并全部手工编辑，精雕细琢。网上用户可采用搜索框检索和目录检索两种方式查找自己所需的信息。

十四、司南中文网上信息检索

<http://www.yippee.com.cn/>

Yippee收录了以大陆为主、包括世界各个国家和地区在内的中文WWW网页。如果您在Yippee主页上的查询框中输入关键词并点击查询，系统将检索整个Yippee信息库。如果您要查找的信息有明确的范围，可以首先进入到相应的分类中，选择“只在本类中查询”的选项后，就地进行查询，这样可以大大节省阅读查询结果的时间。

十五、 哇塞

<http://www.whatsite.com.tw/>

"哇塞"是一个老资格的中文搜索引擎，内容涵盖了大陆、台湾、香港、新加坡、日本、韩国等地区的网上中文资料。同时，"哇塞"还能让没有中文系统的用户也能查看中文资料。"哇塞"的速度相当快，收纳的网页有40万个以上，分类检索简洁明了。"哇塞"领先推出了全球第一个"三合一"的中文网页目录，即使用者可利用大五码、GIF中文目录及英文查寻任何需要的资料。

十六、 蕃薯藤

<http://www.yam.com.tw/>

"蕃薯藤台湾网际网路资源索引"是台湾资深的中文搜索引擎，多年积累使其家底厚实，但近来连接速度较慢。"蕃薯藤"的查询语法强大并且精确，让人查询起来相当得心应手。



十七、 COO台湾索引

<http://www.coo.com.tw/>

COO台湾索引主要以台湾的网站为中心，实际上它收纳的网站相当丰富，查询速度也令人满意。它在台湾可以说是“后起之秀”，近来非常活跃，提供的服务也不仅仅局限于搜索引擎。COO的分类索引和查询结果似乎不太精确，有时同一网站会反复出现，其实这几乎是众多搜索引擎的通病。

十八、添达香港搜索器

<http://www.hksrch.com/>

400多个目录，内容包罗娱乐、电脑、财经、地产、商业、社会文化、政府、教育及个人网页等，应有尽有。收录9000多个香港网址，完全支持中英文搜索，简单易用。



十九、华页指南查询

<http://www.c3s.org.sg/>

主要查询新加坡的中英文资料。支持逻辑运算符AND、OR和括号嵌套,如((AAAANDBBB)OR(CCCORDDD))AND EEE。关键词之间加空格相当于OR。



二十、AltaVista中文查询

http://altavista.digital.com/av/oneweb/query_euccn.html

AltaVista源自于西班牙语中的“虚拟空间”。DEC公司的AltaVista搜索引擎每天要检索27万多个Web服务器、450万个主页、1.4万个新闻组的300万篇文章，每日接受超过290万次的访问，是全球知名度极高的搜索引擎之一。

AltaVista全球索引的最成功之处在于它从技术上实现了在整个网络中支持多语言代码，能够帮助用户更简单、快捷地查询任何语言的信息，这其中当然也包括中文信息。



二十一、中文查寻引擎

<http://www.searchchina.com/>

EIN是欧洲发展最迅速的网络公司之一，中文查寻引擎是由EIN的专业设计者为网络中的中文用户设计的，所有信息均由该公司专业中文编辑精心收集编排。



二十二、其它中文搜索引擎

此外，中文搜索引擎还有

华好网景(<http://www.chinaok.com>)、

四通利方中英文查询器(<http://www.richsurf.com>)

八闽搜索器

(<http://netcity.fz.fj.cn/guidenew/default.htm>)

网路神搜索引擎(<http://webpro.com.cn/search.htm>)

网络直通车

(<http://www.online.xa.sn.cn/new/net/daohang/index.htm>)



下列网站有的有分类导航功能，有的引用了别的搜索工具进行搜索，也可帮您查找网上中文信息。其中国内的有：

网上城市中文搜索引擎

(<http://netcity.fz.fj.cn/guidenew/searchengin/searchengin1.htm>)

东方网景导航(<http://www.east.cn.net/search>)

瑞得站点导航(<http://www.rol.cn.net/station>)

中国导航搜索器(<http://www.chinavigator.com.cn>)

千里眼中国资源搜索器(<http://www.chinese-resources.com>)、和讯时代站点导航(<http://www.homeway.com.cn>)

台源网络站点导航

(http://tnet.beijing.cn.net/gis/collection/internet/link_main.htm)

深圳之窗信息导航

(<http://www.szptt.net.cn/info7.html>)

深圳万用网搜索器(<http://newsnet.szptt.net.cn>)

中国信息导航(<http://www.chinainfo.gov.cn>)等等

而国外则有大名鼎鼎的Netscape 中国指南

(<http://www.netcenter.com/zh/cn/index.html>)。

它还能利用Verity公司的内置功能TopicSearch，实现方便的概念搜索，如用户在查找“文艺”时，能找出有关的“美术”、“雕塑”等文章。同时，用户还能使用and、or、not、in等多种操作符，达到精确搜索的目的。



第六节

流行的英文搜索引擎

一、Yahoo



1994年4月，Stanford大学的两位电子工程学博士研究生 David Filo和Jerry Yang（杨致远）开始编制一个Internet上他们感兴趣节点的目录，这就是最原始的 Yahoo。不久，他们即发现目录变得很长且不易控制。因此，他们将这一原始的Yahoo 转换成一个数据库，设计了用来有效地定位，确认Internet上信息的软件，并开始为同一圈子内的数千Internet访问者提供服务。

1995年初，应Netscape公司的Marc Andreessen 之邀，Yahoo从斯坦福大学Filo和Yang的计算机上移到Netscape公司更大型的主机上。随着WWW的飞速发展，Yahoo也不断成熟壮大。虽然人们总认Yahoo一词别有含义，如一种分层数据库 Yet Another Hierarchical Oracle。



但两位研创者坚持认为他们选择这一名称仅仅因为他们认为自己是Yahoo（乡巴佬）。其实，乡巴佬（Yahoo）并不土，而是一个热门的大众化的WWW分类目录和搜索工具，是目前Internet上最流行最受欢迎的一个WWW搜索引擎。对一般信息的查询方便迅速、结果满意，值得经常应用。



Yahoo (<http://www.yahoo.com>)由美国多家公司和个人资助，拥有全世界最广泛的用户群。它覆盖范围广，连接速度快，数据容量大，简便易用，并且是全免费的。它提供了两种风格的的查找方式：列表式目录链接方式和关键词查询。



列表式目录链接方式：启动浏览器进入WWW后，在地址栏键入<http://www.yahoo.com>即能到达Yahoo主页。页面顶部是一些常用链接，如黄页、寻人、城市地图等。底部是Yahoo自身的一些常用链接。中部是主体，按内容分为十四大类：艺术、商业和经济、计算机和互联网、教育、娱乐、政府、健康、新闻、休闲和运动、参考信息、区域、科学、社会科学和社会及文化。

每一大类下面还分多个小类，如在计算机和互联网下有互联网、WWW、软件、多媒体等；在休闲和运动下有奥林匹克、体育运动、游戏、旅游、汽车等。



上述主页上的每一分类均是链接词，可用鼠标点击任一链接词而进入相关领域。目录链列表按树形结构组织，可以从点击根链开始，不断深入，最终到达所需的Web页、新闻组、FTP站和其它可由Web访问的资源。这种列表式分层搜索易于控制，适合浏览性的查找，但因层次内容太多会感到速度太慢，为此Yahoo提供了另一种选择，即关键字匹配查询。



关键字查询方式： 在Yahoo的主页或任一个查询结果返回页的顶部和底部，都有一个文字输入框，在这里可以输入关键字进行快速查找。当你在此填入指定的关键字，单击右侧的Search按钮后，Yahoo就会从Yahoo目录、Yahoo网点、Yahoo网上事件和谈话、最新新闻等四个方面找出相匹配的记录。查询结果返回的是与关键字匹配的记录列表，最前面的是Yahoo目录链，其后是Yahoo网点，网点通常由以链接形式出现的标题和简介组成。

Yahoo不仅能将与所选主题词相关的主页列出，对其作出简短的介绍并提供链接点，而且在搜索结果页面底部还提供了导向其他搜索工具的链接点，如 Alta Vista、Open Text Infoseek、WebCrawler、Lycos等，以便对Yahoo搜索结果不满意时方便地启动其它搜索工具对同一主题词进行搜索。

如果在Yahoo目录和网点中都没有相匹配的内容，Yahoo则自动利用Alta Vista查询机进行整个Web范围的文档查找。如想获得与关键字匹配的^{最新}新闻和网上事件的列表，可以单击该页上部的目录条上的相应链接。在上述查询方式中，如果只输入单独的用空格分开的关键字，Yahoo则使用所谓的智能查询方式进行查找，处理时忽略单复数形式，自动猜测词义等。

如果要自己控制查询，可以使用下列两种方式：

1.单击输入框右侧的Option链接进入查询选项页，进行选项设置。

2.用Yahoo自定义的修饰法对关键字加以限制。第一种方式进入后，各种选项一目了然。



禁止或要求出现某词在单词前加+，要求匹配记录必须含有该词；单词前加-，禁止匹配记录含有该词。

限制单词出现位置 在单词前加t：限定该词只能位于记录标题上；单词前加u：限定该词只能位于URL地址名中。



词组 将一系列单词用双引号括起，则引号内的词组作为一个整体查询。

通配符 与DOS通配符类似，在单词右边加*将返回含有起始字母为该单词的记录。

组合查询 上述四种可以组合使用。

下面我们以一个实例讲解利用Yahoo进行搜索的过程。进入Yahoo主页，画面如下页：



Yahoo标题



常用链接

[Yahoo! Chat](#)



[Weekly Picks](#)

 [options](#)

[Yellow Pages](#) - [People Search](#) - [Maps](#) - [Classifieds](#) - [News](#) - [Stock Quotes](#) - [Sports Scores](#)

[Arts and Humanities](#)

[Architecture](#), [Photography](#), [Literature...](#)

[News and Media \[Xtra!\]](#)

[Current Events](#), [Magazines](#), [TV](#), [Newspapers...](#)

分类列表

[Business and Economy \[Xtra!\]](#)

[Companies](#), [Investing](#), [Employment...](#)

[Recreation and Sports \[Xtra!\]](#)

[Sports](#), [Games](#), [Travel](#), [Autos](#), [Outdoors...](#)

[Computers and Internet \[Xtra!\]](#)

[Internet](#), [WWW](#), [Software](#), [Multimedia...](#)

[Reference](#)

[Libraries](#), [Dictionaries](#), [Phone Numbers...](#)

[Education](#)

[Universities](#), [K-12](#), [College Entrance...](#)

[Regional](#)

[Countries](#), [Regions](#), [U. S. States...](#)

[Entertainment \[Xtra!\]](#)

[Cool Links](#), [Movies](#), [Music](#), [Humor...](#)

[Science](#)

[CS](#), [Biology](#), [Astronomy](#), [Engineering...](#)

[Government](#)

[Military](#), [Politics \[Xtra!\]](#), [Law](#), [Taxes...](#)

[Social Science](#)

[Anthropology](#), [Sociology](#), [Economics...](#)

关键词查



其它向导工具

[My Yahoo!](#) - [Yahooligans! for Kids](#) - [Beatrice's Web Guide](#) - [Yahoo! Internet Life](#)
[Weekly Picks](#) - [Today's Web Events](#) - [Chat](#) - [Weather Forecasts](#)
[Random Yahoo! Link](#) - [Buy Yahoo! Stuff Here!](#)

National Yahoos [Canada](#) - [France](#) - [Germany](#) - [Japan](#) - [U.K. & Ireland](#)

Yahoo! Metros [Atlanta](#) - [Austin](#) - [Boston](#) - [Chicago](#) - [Dallas / Fort Worth](#) - [Los Angeles](#)

Get Local [Miami](#) - [Minneapolis / St. Paul](#) - [New York](#) - [S.F. Bay](#) - [Seattle](#) - [Wash D.C.](#)

[How to Include Your Site](#) - [Company Information](#) - [Contributors](#) - [Yahoo! to Go](#)

我们先以列表式目录链接查询有关体育的内容。

在目录中单击“Recreation and Sports”超链接。Netscape显示Recreation and Sports页。在页的顶部是标准的Yahoo标题，在它下面是到各种娱乐和体育有关的主题的超链接。



YAHOO!



Top: Recreation

[Options](#)

Search all of Yahoo Search only in Recreation



[Yahoo! Net Events: Recreation](#) - today's chats, games and programs.
[Sub Category Listing](#)

[Amusement and Theme Parks](#)
[Animals, Insects, and Pets](#)
[Automotive](#) (2841) NEW!
[Aviation](#) (1079) NEW!
[Chat](#) (6)
[Cooking](#)
[Dance](#)
[Dating](#)
[Drugs](#)
[Employment](#) (7)
[Events](#) (10)

[Fashion](#)
[Fishing](#)
[Games](#) (14852) NEW!
[Hobbies and Crafts](#) (2416) NEW!
[Home and Garden](#) (417) NEW!
[Hovercraft](#) (14)
[Motorcycles](#) (767) NEW!
[Outdoors](#) (3799) NEW!
[Sports](#) (20890) NEW!
[Toys](#) (534) NEW!
[Travel](#) (4653) NEW!

图 1—2

单击Sports超链接，出现有关体育的链接列表



YAHOO!   

FREEBIES *hosted by Cindy Pawford*

[Happy Puppy! The #1 Game Site on the Internet!](#)

Top: Recreation: Sports

[Options](#)

Search all of Yahoo Search only in **Sports**

[Image Surfer](#) **NEW!** - Try browsing the images in this category with this new feature.

[Current Sports Headlines](#)
[Scoreboard](#) 
[Yahoo! Net Events: Sports](#) - today's chats, games and programs.
[Indices](#) (34) **NEW!**

Amateur (6)	Lumbering (3)
Archery (72) NEW!	Magazines@
Art@	Mailing Lists (17)
Athletes (23)	Martial Arts (553) NEW!
Auto Racing (960) NEW!	Medicine@
Badminton (38)	Motorcycle Racing (77) NEW!

单击Sports超链接，出现有关体育的链接列表

[Dog Racing](#) (6)
[Dogsledding](#)@
[Employment](#) (5)
[Equestrian](#) (237) NEW!
[Events](#) (350) NEW!
[Extreme Sports](#) (21)
[Fantasy Leagues](#) (23)
[Fencing](#) (146) NEW!
[Fishing](#)@
[Flying Discs](#) (142) NEW!
[Footbag \(Hacky Sack\)](#) (10)
[Football \(American\)](#) (1464) NEW!
[Football \(Australian\)](#) (47) NEW!
[Football \(Gaelic\)](#) (5) NEW!
[Gambling](#)@
[Golf](#) (562) NEW!
[Gymnastics](#) (273) NEW!
[Handball](#) (41)
[High School](#) (29)
[History](#) (33) NEW!
[Hockey](#) (2652) NEW!
[Horse Racing](#)@
[Jai-Alai](#) (7)
[Jump Rope](#)@
[Korfball](#) (21)
[Lacrosse](#) (127)
[Skating](#) (319)
[Skiing, Snow](#) (305)
[Skydiving](#)@
[Snowboarding](#) (95)
[Snowmobiling](#)@
[Soccer](#) (841) NEW!
[Softball](#) (150)
[Software](#) (3)
[Squash](#) (29)
[Stadiums and Venues](#) (47)
[Surfing](#) (160) NEW!
[Swimming and Diving](#) (346) NEW!
[Table Tennis](#) (36)
[Technology](#) (2)
[Tennis](#) (347) NEW!
[Track and Field](#) (183) NEW!
[Triathlon](#) (114) NEW!
[Trivia](#) (11)
[Tug-of-War](#) (6)
[Volleyball](#) (266) NEW!
[Wakeboarding](#)@
[Walking](#)@
[Water Polo](#) (39)
[Waterskiing](#) (54)
[Weightlifting](#) (131)
[Windsurfing](#) (126)

单击football (american) ，则出现一个页面，中间为相应的链接点，下部为相关的内容链接点。

[Top: Recreation: Sports: Football \(American\)](#)

[Options](#)

Search all of Yahoo Search only in **Football (American)**

[Yahoo! Net Events: Football \(American\)](#) - today's chats and programs.

[Indices](#) (2)

[Arena Football League](#)@

[Canadian Football League \(CFL\)](#)@

[Chat](#) (5) **NEW!**

[Coaching](#) (5)

[College and University](#) (311)

[Companies](#)@

[Fantasy Leagues](#) (65)

[History](#) (2)

[Leagues](#) (986) **NEW!**

[National Football League \(NFL\)](#)@

[News and Media](#) (7)

[Players](#) (2)

[Regional](#) (22)

[Software](#) (3)

[Stadiums](#) (9)

[Technology](#) (1)

[Youth](#) (35)

[Usenet](#) (2)

如此一直搜索，即可找到想浏览的任何链接点。
接下来，我们以关键字搜索功能查找信息。

回到Yahoo主页，或在任一页面中有Search的文本输入框中，键入要搜索的关键字。在这里可以使用前面介绍的任何布尔参数。比如，我们键入“+football +american”。



回车或单击Search按钮，搜索结果列出一份清单。可根据自己的需要单击超链接点。

[Categories](#) - [Sites](#) - [AltaVista Web Pages](#) | [Headlines](#) | [Net Events](#)

Found 27 Category and 623 Site Matches for +football +american.

Yahoo! Category Matches (1 - 20 of 27)

[Recreation: Sports: Football \(American\)](#)

[Business and Economy: Companies: Sports: Football \(American\)](#)

[Business and Economy: Companies: Sports: Fantasy: Football \(American\)](#)

[Recreation: Games: Gambling: Sports Gambling: Football \(American\)](#)

[Recreation: Sports: Football \(American\): Leagues: World League of American Football \(WLAF\)](#)

[Business and Economy: Companies: Computers: Software: Games: Gambling: Football \(American\)](#)

[Business and Economy: Companies: Sports: Fantasy: Software: Football \(American\)](#)

或者键入“football american”，单击Search按钮旁边的option选项，

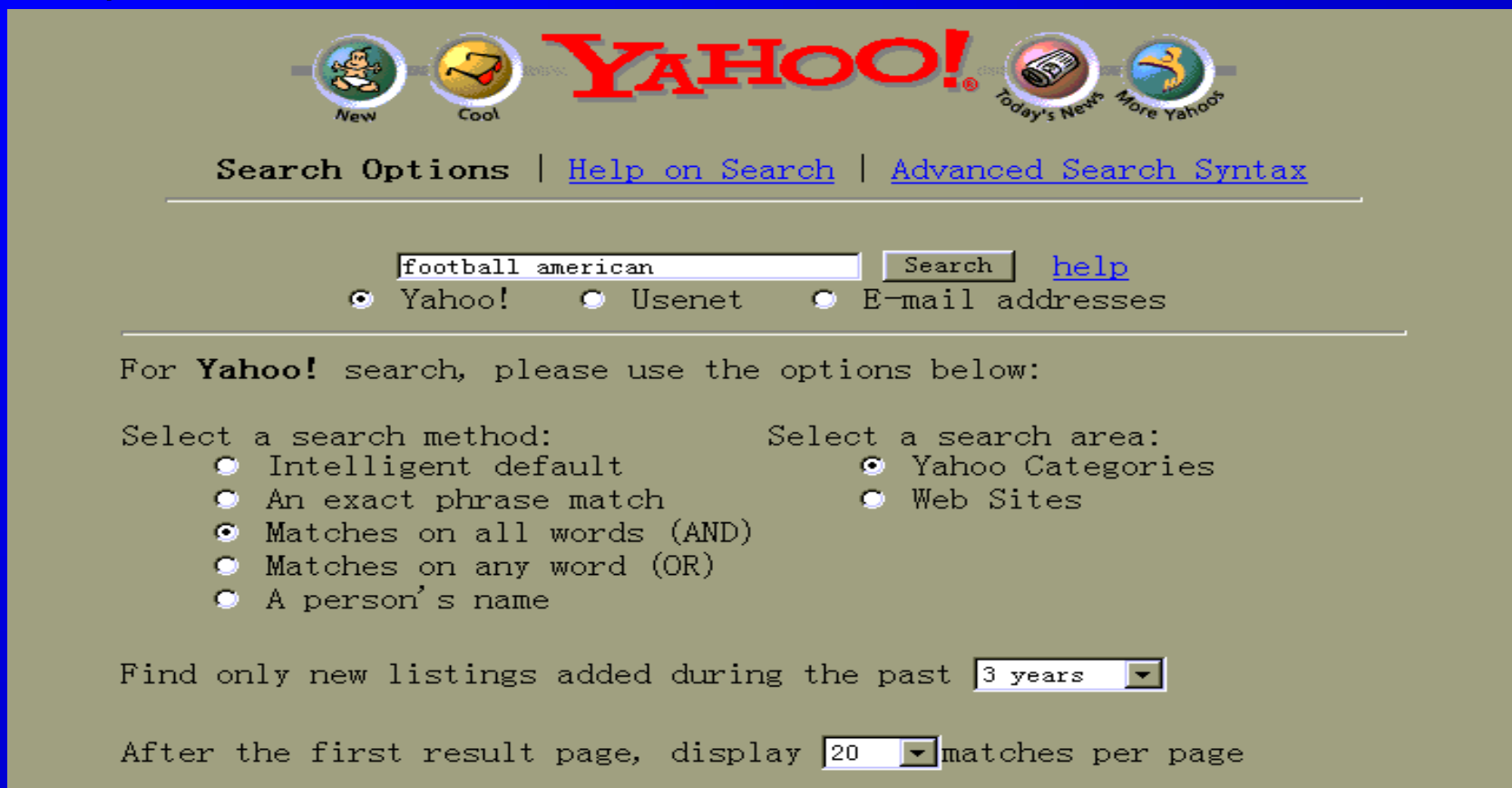
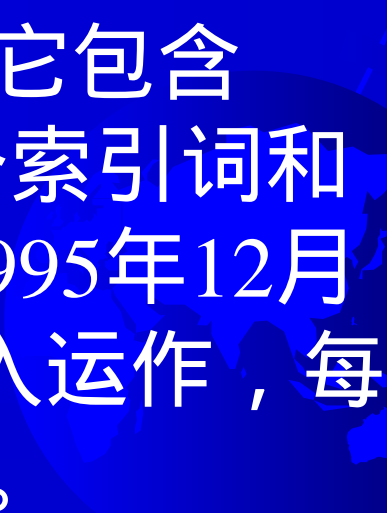


图 1 — 7

可按提示选择不同的搜索方式。我们选Match on all words(AND)其结果和上面的是一样的。

二、Alta Vista (<http://www.altavista.com>)

美国DEC (Digital Equipment Corporation)的Alta Vista一直被认为是Internet上最大的查询工具，它包含2200万Web页面中的110亿个索引词和13000个新闻组。该工具于1995年12月在DEC的Palo Alto研究室投入运作，每天被访问约800万次,如图1-8。



 [Software](#)

 [Add URL](#)

 [Contests!](#)

 [Ad Info.](#)

 [Web Sites](#)

 [About Us](#)

There's a brand new [Howdy!](#) in store.
Over 120 backgrounds, tons of audios and fun poems turn boring email
into multimedia Internet Postcards! It's email with an
attitude...it's ME-Mail!
ME-Mail someone today!

Search and Display the Results

Search with Digital's Alta Vista [[Advanced Search](#)]
AltaVista Technology, Inc. is not affiliated with
Digital Equipment Corporation, AltaVista Internet
Software, Inc. or the AltaVista Internet Search Service.
The AltaVista Internet Search Service may be found at
<http://www.altavista.digital.com>.

 **Add your URL to the most
popular Search Engines**

 **Look up your
Domain Name.**
Web Cow! Domain Name Registration

What's New... Check out [Howdios!](#)
Howdios! are background, audio or text add-on packages that plug
right into your Howdy! software. Start making new postcards faster
than you can say Howdy!

利用AltaVista进行搜索非常容易。只要键入一个问题或者一些关键词，单击Submit，然后AltaVista就会返回搜索结果。

对大多数searches，我们用的都是SimpleSearch。而对一些熟练的用户，可以使用高级搜索（Advanced Search）。用日期范围、包含或排除词语的组合来缩小搜索范围。如图1-9所示。

Search

the Web

for documents in

any language

Ranking:

From:

21/Mar/86

To:

search

refine

[Help](#) . [Preferences](#) . [New Search](#) . [Simple Search](#)

[Our Network](#) | [Add/Remove URL](#) | [Feedback](#) | [Help](#)
[Advertising Info](#) | [About AltaVista](#) | [Jobs](#) | [Text-Only](#)

digital

Digital Equipment
Corporation

AltaVista Search
Gets Personal

[Disclaimer](#) | [Privacy Statement](#)

Copyright 1997 All Rights Reserved

还可以使用Refine（细化）来进一步缩小搜索范围，得到想要的信息。Refine工具分析满足关键字的文档内容，并将相关的词组附加显示，称为主题词，用来细化您的查询。主题词是动态生成的，以命中符数来排序。这一相关主题的动态生成，附加到每次搜索结果之后，使您每次总能有所斩获。

Refine提供了两种主题排列的方式：一种是列表式，每个主题有一个下拉式菜单，方便您的查询，如图1-10。

Refine your search by requiring a few relevant topics, excluding irrelevant ones, and ignoring the others.



search

refine

- 93%** Football, nfl, sports, espn, sportszone, espnet, sportsline
- 72%** Audionet, realaudio, vxtreme
- 67%** Basketball, hockey, baseball
- 62%** Bullfighting, boxing, cricket, polo, collectibles
- 54%** Category, streamworks, realvideo, monsterbit, netshow, kvr, internetv, onlive
- 50%** League, teams, stats

图 1-10

三、Webcrawler (<http://www.webcrawler.com>)

这是Internet上出现较早的大型WWW搜寻工具，1994年由美国 University of Washington推出，1995年归属于America Online公司。其搜索功能丰富，手段灵活，覆盖面庞大。可一次检索世界各地 25万个服务器上的160万余个文档，每天平均被访问 300多万次。另外它还经常公布通过Webcrawler被访问次数最多的25个WWW地址。以供用户参考，如图1-12。



Search

search

guide

services

fun

help

Search

ex. [sagittarius horoscope](#)

[options](#)

[Arts](#)

[Games](#)

[Recreation](#)

[Business](#)

[Health](#)

[Reference](#)

[Chat](#)

賤

[Internet](#)

[Romance](#)

[Computers](#)

[Kids](#)

[Science](#)

[Education](#)

[Life](#)

[Sports](#)

[Entertainment](#)

[News](#)

[Travel](#)

classifieds2000

[Classified Ads](#)

MUSIC BOULEVARD

[Music Store](#)

QUOTE.COM

[Stock Quotes](#)

DEJA NEWS

[Newsgroups](#)



services: Leo lover got you down?
[Rate your Romance](#) and find your zodiac
match!

图 1-12

WebCrawler支持所谓的“自然语言搜索（natural language searching）”。因此它允许用户敲入日常英语而不必理会复杂的搜索语法。但在这种方式下，它将默认为您对所有敲入的词都感兴趣，从而列出大量的清单。然而如果您用高级搜索，它会支持丰富的布尔逻辑运算。

下面列出一些利用逻辑运算进行搜索的例子

。



运算	举例	说明
AND	gardening AND vegetables	同时包含 gardening 和 vegetables 的页面
OR	whales OR cetaceans	包含 whales 或者 cetaceans 或者两个都有的页面，这是一个默认方式，所以无需专门指明
NOT	science NOT fiction	包含 science 但不包含 fiction 的页面
NEAR	arthritis NEAR/25 nutrition	包含两个词，但前后相邻必须在 25 个词之内的页面；若只写 NEAR，不标明数字，则默认两

运算	举例	说明
ADJ	bal ADJ warming	两个词按此顺序相邻出现的页，如包含 global warming 的页面。
"..."	"1996 World Series Champions"	包含"1996 World Series Champions"这一词组的页面，效果等同于 1996 ADJ World Series Champions
(...)	Homer NOT (Simpson OR Alaska)	包含 Homer 但不包括 Homer Simpson 或者 Homer Alaska 的页面。

WebCrawler也允许用户设置一些搜索结果的显示开关，您可以选择简短格式或者详细格式来显示查询结果。简短格式列出符合查询条件的Web资源的标题，而详细格式提供标题，摘要，URLs，命中符数目及显示相关页面的选项。

WebCrawler还有一个很有特色的Shortcuts（捷径）功能：如图1-13所示。

。



搜索文本框

Shortcuts

结果统计

查询结果

结果排序

后面的搜索结果

hotels in San Francisco

Search Results **Shortcuts**

Top 10 of 94568 for **hotels in San Francisco**
Show [summaries](#) for these results.

97%	General Information About Hotels in San Francisco
96%	Beresford Hotels - San Francisco
95%	Discount San Francisco Hotels
94%	california beer page - microbrewery links, beer links,...
94%	Street Sheet July 1995
94%	San Francisco Hotels
94%	San Francisco Non-Bookable Hotels, Inns and Resorts
94%	San Francisco Hotels, Bed & Breakfast Inns, Resorts
94%	hotels accomodations lodging san francisco bay area...
94%	The Finest Internet Free Hotel Reservation Service


guide
WEB SITE REVIEWS
[U.S. Travel Guides](#)
[Lodging](#)


SEE A MAP OF:
[San Francisco, CA](#)

图 1-13

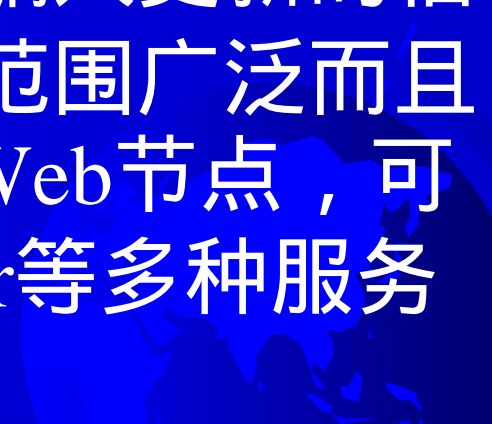
当您敲入关键字，按search钮，WebCrawler返回查找到的Web页的清单及其到达相关资源的Shortcuts。比如查找hotels in San Francisco，会返回上图的页面。单击底部Guide to the Results Page可以看到对页面不同部分的描述。包括：



- 1、搜索文本框，可以进一步细化查询；
 - 2、查询结果；
 - 3、结果统计，显示找到多少文档并给出查看条目摘要的选项；
 - 4、结果排序，按照相关程度从高到低排列；
 - 5、后面的搜索结果，可以查看其它搜索结果；
 - 6、Shortcuts，您可以从这儿连接到其它相关资源。
- 

四、Lycos(<http://www.lycos.com>)

这是Carnegie-Mellon大学著名的查询工具，号称“Internet上最流行的搜索机”。到1996年1月，Lycos已经编目和索引了近1100万个URL，且每天都在增长，编入更新的信息。其特点是功能强大，查询范围广泛而且很彻底，差不多覆盖了91%的Web节点，可进行包括WWW、FTP、Gopher等多种服务的搜索。



最近该工具加入了一种可以大大加快数据搜索速度的技术，称为CentiSpeed，它每秒处理4000个查询要求。

使用Lycos搜索Internet的步骤：

连接Lycos的主页，左边是常用超链接，中间是列表式分类目录，上部有两个Search for的搜索文本框。要搜索football american的信息。Search文本框是一个搜索选项下拉菜单，包含Web，sound，picture，personal homepages及前5%(TOP5%)等一些分类，如图1-14。

LYCOS®

Your
Personal
Internet Guide

[YELLOW PAGES](#)

[COMPANIES ONLINE](#)

[CLASSIFIEDS](#)

[LYCOS PRESS](#)

[FREE SOFTWARE](#)

[LINK TO LYCOS](#)

[ABOUT LYCOS](#)

Get Lycos or get lost

[TOP5%](#) [CityGuide](#) [Pictures & Sounds](#) [PeopleFind](#) [StockFind](#) [RoadMaps](#) [UPS™Services](#)

Search for:

[Click here](#) New! **LYCOS PRO™** Custom Search

[Go Get It®](#)



[NEWS](#)

[SPORTS](#)

[MONEY](#)

[TRAVEL](#)

[TECHNOLOGY](#)

[HEALTH](#)

[SCIENCE](#)

[EDUCATION](#)

[LIFESTYLE](#)

[CULTURE](#)

[SHOPPING](#)

[KIDS](#)

[BUSINESS](#)

[ENTERTAINMENT](#)

[CAREERS](#)

[FASHION](#)

[GOVERNMENT](#)

[AUTOS](#)

Need Help?
[Start Here.](#)

ON LYCOS NOW

[Win Cold, Hard CASH!](#)

[Swiss War Debt](#)

[Shop Til You Drop](#)

[Lycos Special Features](#)

[Plan a Trip Online](#)

[Business Resources](#)

[Lycos Germany](#) ▪ [Lycos Sweden](#) ▪ [Lycos France](#)

[Lycos UK](#) ▪ [Lycos Italy](#) ▪ [Lycos Netherlands](#)

[Lycos Spain](#) ▪ [Lycos Switzerland](#)

Copyright 1997 Lycos®, Inc.

All Rights Reserved.

Lycos is a trademark of
Carnegie Mellon University

[Questions & Comments](#)

[New Search](#) ▪ [TopNews](#) ▪ [TOP 5%](#) ▪ [City Guide](#) ▪ [StockFind](#)

[PeopleFind](#) ▪ [Companies Online](#) ▪ [Road Maps](#) ▪ [Software](#) ▪ [About Lycos](#) ▪ [Help](#)

[Add Your Site to Lycos](#) ▪ [Advertise with Lycos](#) ▪ [Business Development](#) ▪ [Jobs4You](#) ▪ [New2Net](#)

另外，Lycos也支持自然语言查询和全部的布尔逻辑运算。而且增加了Before, Far两种运算。具体语法和例子可以从Lycos主页上的Help项中学习。

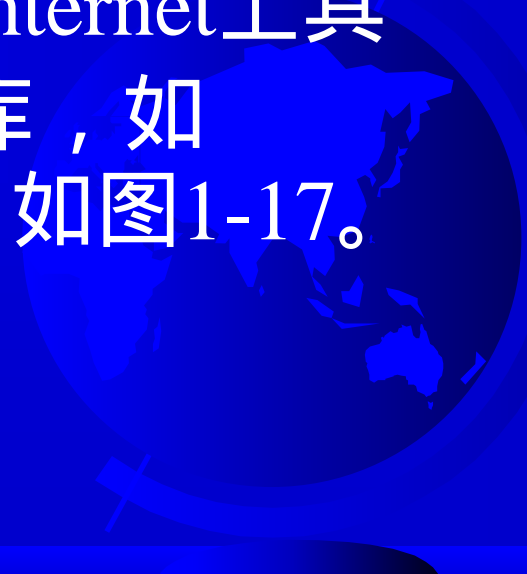


五、InfoSeek(<http://www.infoseek.com>)

这是一个著名的有偿查询工具，每月查询100条收费\$9.95。在Internet World 1996年5月的各种搜索工具评比中名列第一。在其主页上有一个演示版，可以进行免费搜索，但只返回十条信息。



InfoSeek类似于Lycos和Open Text，用户可以通过设置“调整搜索模式”的Web浏览器表单发出关键字查询。除了下面介绍的免费查询外，InfoSeek还提供一种收费订阅搜索服务。收费搜索包括在Internet上其它任何地方不可用的一些数据库，如Computer Select数据库的文档，如图1-17。





infoseek®

proof of intelligent life on the netsm

BigYellow

Yellow Pages Search



UPS[®] Services

Type a specific question, "phrase in quotes" or Capitalized Name.

the Web

seek

[Tips](#)

[news center](#)

[Personalized News](#)

[World](#)

[Business](#)

[Technology](#)

[Sports](#)

[smart](#)

[info](#)

[People](#)

[Business](#)

News Flash: [Microprocessor Price and Acquisition Wars Heat Up](#)

Click a topic to explore the Web's largest directory:

[Arts & Entertainment](#)

[books](#), [music](#), [games](#), [movies...](#)

[Internet News](#)

[intranet](#), [HTML](#), [web publishing...](#)

六、Open Text (<http://www.opentext.com>)

它宣称是“Internet上最快、功能最强的搜索工具”。它支持全文本搜索，可以访问近100万个Web页面。

使用Open Text步骤：

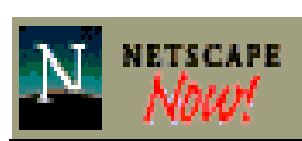
(1) 进入Open Text页面，首先显示一个简化的搜索文本框。文本框下方允许用户直接从简化的菜单使用布尔逻辑的“或”（OR）和“与”（AND）以及引号功能，如图1-18。

Search the World Wide Web for pages that contain...

Search this Site for pages that contain...

- this exact phrase:
- all of these words:
- any of these words:

URL: <http://www.opentext.com> Tel: +1.519.888.7111 Fax: +1.519.888.0677



Navigate Open Text in XSpace! [Download](#) HotSauce plug-in; restart your browser; [click here](#)

(2) 在Search文本框中输入football american，选择Any of these Words，即“或”功能，单击Search开始搜索，如图1-19。





Instant online communication

(click here for free download and service)



[Click Here to Visit Site](#)

The Open Text Index found **365,556** pages containing: **football Or american**

You can [improve your result](#) or [start a new search](#).

pages 1-10 (of 365556)

1. [PRO SHOP EXCALIBUR FOOTBALL FOOTBALL - THE DICK BUTKUS PRO SHOP - FOOTBALL...](#) (score: 4321, size: 15.0k)

From: <http://www.wvcd.com/excalibur/index.html>

Join Today and SAVE! EXCALIBUR MARKETING GROUP. DICK BUTKUS Autographed - Footballs - Photos - Jerseys - Helmets - Trading Cards FRAN TARKENTON Autographed - Footballs - Photos - Jerseys - Helmets DEACON JONES Autographed - Footballs - Photos - Jerseys - He

[\[Visit the page\]](#)

2. [American Heart Assoc., American Lung Assoc. and American Cancer Society...](#) (score: 3309, size: 6.4k)

Open Text返回一个页面，列出含有football和american的文档的超链接。注意一下返回的文档个数。然后尝试使用Open Text Power Search功能调整该搜索。

单击back按钮回到Open Text Index页。单击页底部的Power Search超链接，如图1-20。





Instant online communication
(click here for free download and service)



[Click Here to Visit Site](#)

- ▶ Search
- ▶ Discover
- ▶ Interact
- ▶ Help

Add Your
URL



Search for	<input type="text" value="football american"/>	within	<input type="text" value="anywhere"/>	▼
<input type="text" value="and"/>	<input type="text"/>	within	<input type="text" value="anywhere"/>	▼
<input type="text" value="and"/>	<input type="text"/>	within	<input type="text" value="anywhere"/>	▼
or				
but not				
near				
followed by				
		<input type="button" value="Search"/>	<input type="button" value="Reset"/>	

IDC Rates Open Text Livelink [#1 Leader](#) in Intranet Document Management.

Search Tip: [Use the Open Text Index to search Usenet Newsgroups!](#)

[Click here to find out what's new at the Open Text Index!](#)

Search: [Simple Search](#) | [Power Search](#) | [Current Events](#) | [Newsgroups](#) | [Email](#) | [Other Languages](#)

3

图 1-20

Open Text返回Open Text Index-Power Search页，它含有让用户调整搜索的选项。

在Search for文本框中，输入football，然后在含有And的下拉列表中单击，单击Followed By。



在第二个Search for文本框中，输入american。单击Search按钮开始搜索。

这次Open Text返回同时含有football和american的文档，如图1-21。





Instant online communication
(click here for free download & service)

To get all the
TUICY details

[Click Here to Visit Site](#)

The Open Text Index found 135 pages containing: **football american**

You can [improve your result](#) or [start a new search](#).

pages 1-10 (of 135)

1. [Sports Events Worldwide Calendar - Traveler's Guide to Great Sporting Events](#) (score: 103, size: 30.7k)

From: <http://www.digidiscounts.com/sports.html>

Chapter 9: Great Sporting Events and Games. Select a category, or scroll thru the list. The sports are listed alphabetically, with North American cities listed first, followed by Europe, Asia, Oceania, Australia, South America and Africa. Aerial | Auto Raci

[\[Visit the page\]](#)

2. [History of Football](#) (score: 103, size: 36.4k)

From: <http://sachen.k12.ny.us/sports/football/fthis.html>

Football Football, American, distinct type of football that developed in the United States in the 19th century from soccer (association football) and rugby

图 1-21

Open Text还有其它一些搜索调整技术，如加权。

七、World Wide Web Worm (<http://www.cs.colorado.edu/home/mcbryan/wwwwww.html>)

搜索结果返回数量大，而且通常有很高的准确性，同样是一个一流的WWW搜索工具，如图1-22。



WWW - WORLD WIDE WEB WORM

Please note our new address: <http://www.goto.com>



[Best of the Web '94](#) - Best Navigational Aid. [Oliver McBryan](#)

Serving 3,000,000 URL's to 2,000,000 folks/month.

[Instructions](#), [Definitions](#), [Examples](#), [Failures](#), [Register](#), [WWW Paper](#).

1. Search all URL references

a. AND - match all keywords

b. OR - match any keyword

5 matches


Keywords:

图 1-22

八、Excite

(<http://www.excite.com>)

后来加入Internet搜索工具行列的一个WWW节点，由Architext Software公司开发，可以通过纯文本和关键字两种方法搜索Web和Usenet新闻组，如图1-23。





News



Stocks



TV



Weather

Get Ready for Some Football!

Excite Search

Search

[Search Tips](#)
[Power Search](#)

[People Finder](#) 柯 [Email Lookup](#) 柯 [Yellow Pages](#) 柯 [M](#)
[Stock Quotes](#) 柯 [Book Flights](#) 柯 [Newsgroups](#) 柯 [Share](#)

Channels by Excite

[Business & Investing](#) [My Channel](#)

[Careers & Education](#) [News](#)

[Computers & Internet](#) [People & Chat](#)

[Entertainment](#)

[Politics](#)

[Games](#)

[Shopping](#)



[Health](#)

[Sports](#)

[Lifestyle](#)

[Travel](#)

Exciting Stuff



**Win the NEW
VW Beetle!**

**Free Email Address
by Excite**

**Instant Paging
Excite PAL**

Other Services: [Excite Direct](#) [Bookmark Excite](#) [Free Email](#) [Horoscopes](#)
[Instant Info](#)

Global Excite: [France](#) [Germany](#) [Japan](#) [Sweden](#) [U.K.](#)



图1-23

九、HotBot

(<http://www.hotbot.com>)

这是著名的HotWire Web节点和在线杂志的开发者Hotwire Ventures公司最新推出的一种检索工具。据称，其覆盖面超过了目前Internet上最大的搜索工具Alta Vista，如图1-24。





GREATEST TALE EVER TOLD CONTEST

AMAZON.COM BOOKS

[Earth's biggest bookstore! Amazon.com](http://Amazon.com)



Click here

To secure your network

The WIRED Search Engine



SEARCH

the Web 或 all the words

Return 10 results with

full descriptions

SAVED SEARCHES

Tip: Set the number of results -- up to 100 per page!

OPEN ALL

MODIFY

DATE

LOCATION

MEDIA TYPE

PAGE TYPE

图1-24

Internet网上98%的信息均采用英文书写和传播，在信息高速公路上，迫切需要有一个“中文路标”来给上“路”的中国人指点迷津。


这样，中文搜索引擎应运而生。但这样的工具还不是很多。这儿给大家介绍两个。



第七节

中国门户网站的“搜索”较量

经历了上市狂潮的洗礼之后，中文门户网站仿佛一度归于沉寂。但是如果稍加留意，不难发现，众多门户网站正转向另一个层面的较量，这就是从一味的“眼球大战”向以技术作为网站安身立命之本的回归。




近来，ChinaRen、搜狐、网易相继对其搜索引擎进行了升级。众所周知，搜索引擎历来是门户网站的看家功夫之一，如此频繁的门户网站的技术升级，俨然酝酿着新一轮的门户之争。



一、 搜狐、ChinaRen "相中"百度

8月28日，搜狐宣布与百度在线网络技术公司合作，对搜狐搜索引擎全面升级，升级后的搜狐检索系统在原先的类目检索、网站检索、新闻检索、中文网址检索功能之外，新增加了一项网页检索功能。



搜狐CTO叶忻介绍，中文检索有自己的独特性，而百度在中文网页检索方面有自己的专长，其主要技术原理是通过不同网页对某个网页的连接次数，决定该网页在检索结果中的排序。百度在线公司是去年底才进入中国的一家网络技术公司，由美国著名的风险投资公司投资建立。



而几乎在同一时间，网易已选定大名鼎鼎的搜索引擎Google作为其全球合作伙伴。

叶忻对此评论说，网易此举是针对搜狐近两个月来在技术上的一系列市场举动而做出的反应，“我想这是他们的一种防御性工作。Google的英文检索非常出色，在检索的相关性方面十分出色，而百度在中文检索方面也是很好的。实际上，我们也考虑过Google，但最终还是选定了百度。”

现在，搜狐的搜索引擎有两大部分，一是分类检索，这是搜狐的核心技术；另一大部分是全球网站检索，叶忻承认这是以前搜狐比较落后的部分：“通过这次合作，搜狐的全球网站检索功能提升到了国内领先的地位，而且检索速度得到了提高。”



其实，早在搜狐采用百度技术之前，第二代门户网站之一的ChinaRen也已采用了该技术。ChinaRen在百度产品的基础上做了二次开发，以表现个性化与满足网民的需要。ChinaRen开发了智能的切词技术，能够识别自然语言的提问，因此用神通广大的“孙悟空”来命名。



ChinaRen CTO 杨宁介绍，百度的搜索使ChinaRen每天增加了几十万的访问量，而这些流量正是网站创造财富的基础，而且搜索引擎能够达到更多的潜在客户。ChinaRen在搜索中放置的旗帜广告也收到了良好的经济效益。



至于大家都选择在近期升级，杨宁认为，更多的动因还是各网站自己的需求。面对当前呈爆炸性增长的市场，原来网站有许多设计已不能满足大访问量、多用户的需求，因此技术上的升级是必然的。



二、网易选定Google

8月中旬，网易与美国搜索引擎开发商Google公司联合宣布，网易将选择Google作为其门户网站163.com搜索引擎的优先合作伙伴和供应商，并在15日内完成系统整合，之后正式开始向用户提供搜索引擎服务。根据双方协议，Google公司将为163.com提供中文和全球互联网范围的网络目录服务。

网易联合首席技术执行官兼副总裁许杰良介绍，网易自己的搜索引擎产品是1998年开发的，而现在已经不能满足中国网民更高层次的需求，为此，网易引进了Google的搜索技术。Google能够提供多语言的检索，能够检索10亿个以上的网页，其中包括2400万个中文网页。

。



网易公司此次推出基于Google技术的搜索引擎只是技术升级的一个步骤，很快还将有新的产品推出，包括邮件系统与数据库系统等。在搜索引擎方面，网易还将推出一个开放式的目录，吸引社会上各界的专家一起来做搜索，由专家重新筛选与分类的搜索将大大提高查准率。



三、新浪一直在提升搜索引擎

针对搜狐、网易在搜索引擎方面的举措，记者采访了新浪网CTO严援朝。严援朝认为，搜狐、网易这次主要都是对其搜索引擎的全文检索功能（又称网页检索）做了改善。实际上，网民对该部分搜索引擎功能的需求通常只占到20%还弱。

所以，这不会对搜索引擎的整体性能产生本质的改变。因为网站检索才是搜索引擎技术的核心。全文检索服务只占搜索引擎技术的一个很小的部分，真正要把搜索引擎做好，关键是依靠网站自身投入更多精力，把网站搜索和其他搜索做得更好，而网站检索数据库主要依靠各个门户网站自己的长期积累来完成。

至于这两大门户网站为何都选择在近期改进其搜索功能，严援朝猜测，可能是与近期一些媒体关于搜索引擎排名的报道有关，尽管新浪自己也不知道其评测的依据，但结果都是新浪占优，其他几家门户网站可能感觉到了压力，所以都在做一些改善性的工作。



严援朝说，新浪网在搜索引擎方面一直都在默默地改进，所以能够在原来落后的情形下发展到现在的境界，改善搜索引擎性能是新浪网长期的工作，而不是短期的行为。



新浪目前仍使用台湾OpenFind的搜索技术,它在全球为新浪提供支持,也表示愿意做出新的改进。另外,新浪从支持本地新公司的角度,也在试用百度的产品,如果表现不错,可以在使用OpenFind的基础上,采用它的一些的技术。



四、门户之争 向技术回归

其实，搜狐近来在技术方面的升级动作远不止搜索引擎这一项，比如采用CacheFlow的解决方案加速其网站的浏览速度；推出JAVA聊天室、网上寻呼SOQ等。搜狐的这种举动，在很大程度上代表了上市后的中文门户网站在竞争策略上的急剧转变。

7月15日，张朝阳在搜狐上市之后就说过：“各个门户上市后，从以往的‘游击战’转向‘正规军’之间的较量。”为了打好这场“正规军”之间的战争，搜狐第一条最主要的策略就是不断推出强势产品，像搜狐电子邮件、JAVA聊天室、留言板、搜索引擎等。



而网易CTO丁磊不久前也撰文认为，“互联网诞生至今已经30多年了，而WWW技术出现只有10年，只有当这个技术推广开来，互联网才产生了真正意义上的进步。



因此，今天人们考察互联网行业的发展时，不能仅仅看其运作模式的改变，也不能仅仅关注它是如何炒作的，更重要的是要关注它在技术上的跃迁，只有技术才能决定网络的未来。”



ChinaRen CTO 杨宁则认为，对于现在的网站来说，商业模式是很容易拷贝的，但技术则不同，所以技术成为区分竞争者的标志，技术是一个网站运营的保证。ChinaRen今年已陆续推出了主页大巴、聊天室、校友录，依靠这些独特的功能吸引了大量的网民，这也是ChinaRen杀入最近CNNIC评选前10名的重要原因。

雅虎公司中国区总经理张平和说，除了资金之外，是什么支持网站真正成功？核心问题是技术。当你有5000个服务器分布在全球各地时，如何支持，就已经不是钱的问题，也不是运作理念的问题，而是硬碰硬的技术。



所以，表面上看来十分凑巧的这场搜索引擎的升级较量，实际上已经兆示出网站在经营路线上的某种回归。门户网站的竞争最终仍归结到网站综合实力、服务质量的较量，而这恰恰都需要有技术为保障。由此推断，技术上的较量可能会是下一波门户大战的主战场。

