

@BRINTBusinessResearchAdaptedAdobe
AdvertisingLawInternetSiteAladinAlbanyNe
tAltaVistaUseNetAltaVistaWebAmazon.co
mBookSearchAmericanMemoryCollection
SearchAmerica'sJobBankSearchIndexAO
LNetFindAppleComputerAquaLinkArchNe
tArchaeologyAskSherlockHousingSearch
AT&T800DirectoryAutoWebInteractiveBell
net.comBigFootBizWebBorlandBytec|netc
dcomcjetSearch.ComCareerCityCareerMo
saicJOBSCBSSportslineCNNDatabaseCN
NfnthefinancialnetworkCollegeNetCommu
nicationsLawCompaqGamingWorldComp
uterNewsDailyCorelCreativeLabsCyberide
rCyclingWWWSiteDATAMANSunManager
sDBCOnlinesequostockquotesDBIUsersArchive
DejaNewsDiscoveryChannelOnlineSearch
Disney.comsearchESpanSimpleJobSearc
hEchoEcilaEducationWorldEdwardLoweD
igitalLibraryElectricLibraryEncartaOnline
EnvironmentalOrganizationWebDirectory
EurekaEuroFerretExciteExciteNewsTrack
erExplorateurExploratoriumWebSearchEx
plorerK12Math/ScienceForeFrontForumO
neOnlineDiscussionForumsGalaxyGolfco
urse.comCourseLocatorGORPGreatOurd
oorecreationPagesHistoryBuff'sSearchHo
over'sMasterListPlusDatabaseHotBotAdv
ancedIBMIBMInfomarketResearchReports
InferenceFINDInfohiwayInfomineInfoseekI
nfospaceBusinessandGovernmentInterne
tArtResourcesInternetAutoExchange'sCla
ssicSportsCarDatabaseInternetCollegeEx
changeInternetMovieDatabaseInternetSho
ppingNetworkJaydeOnlineDirectoryJump
CityKidsHealth.orgLawReasearchLawcra
wlerLibrarians'IndextotheInternetLinkMon
sterLinkStarLiszt,theMailingListDirectory
LycosPictureSearchLycosProLycosSound
sMagellanGreenLightSitesMammaSearch
EnginesMedicalWorldSearchMerrillLynch
WebsiteSearchMetaCrawlerMicrosoftmod
_perlArchivemysqlHypermailArchiveNatio
nalFairHousingAdvocateSearchNetworked
ComputerScienceTechnicalReferenceNe
xialNOAAOceansandAtmosphereDatabas
eNokiaNomadeNorthernLightOneLookDic
tionariesOpenTextOrientation.comAsiaOS
HAPanasonicPCWorldOnlinePlanetSearc
hPoint'sTop5%PolaroidPrincetonReviewP
roductReviewNetPubMedNationalLibraryo
fMedicineQVC/iQVCShopProductSearchR
ec.SkydivingNewsgroupReference.com(M
ailingList)RentAWreckRoughGuideSearch
SavvySearchScienceFictionReviewArchiv
essearchUKShopInternetSkiCentralSocial
ScienceInformationGatewaySportsLineUS
ASpryInternetWizardSunSurfPointSybase
ThePrincetonReviewTheSportingNewsTh
eTechArchiveSearchTheUnitedNationsTi
meMagazineOnlineTime.comforKidsTrave
lWebLodgingSearchTUCOWsUSCodeInte
rnetLawLibraryUSATodayVirtualHospital
WebCrawlerWebCrawlerNewsWhat'sNew
Too!WhatUSeekWhoWhereWindows95Ma
gazineSearchWindows95.comShareWare
WorldWdieWebofSportsSearchWorldWid
eArtsResourcesWWWVirtualLawLibraryW
WWWVirtualLibraryUSGovernmentInformati
onWWWomenYahooYahooimageSurferYa
hooligansYellowPagesOnlinebusinessdir
ectoryYourPersonalNet

Search Tutorial: Guide to Effective Searching of the Internet

Revised and Updated
December 1999



Thanks and Acknowledgements

Thanks for taking the time to learn more about how to effectively use the Internet. We hope sincerely this tutorial helps speed you along the path to better information.

We prepared this tutorial because of our own frustrations in finding a central resource having to do with all things “searching.” We know we’ve missed much of value on the Internet on these subjects, though we’ve tried our darnedest to find all we could. Our apologies to other “power searchers” out there whose valuable work we’ve inadvertently overlooked.

This tutorial was prepared by Michael Bergman of VisualMetrics Corporation, with the super assistance of technical staff including Carol Lushbough, Tom Tiaht, Jerry Tardif and Will Bushee. The authors have attempted to be as accurate and fair as possible; we welcome your suggestions for improvements or informing of us of errors. Please submit all comments to: tutorial@thewebtools.com.

Revised and augmented December 1999 by Tardif, Bergman and Bushee.

Search Tutorial: Guide to Effective Searching on the Internet

Table of Contents

Section 1: Searching with Internet Provided Resources	1
Executive Summary: The Two-Minute Bottom Line	3
Part 1: The Size of the Internet.....	5
Part 2: Internet Search Basics and Why There's a Problem	10
Topic 1: Status of the Internet and Searcher's Frustrations.....	10
Topic 2: Search Engine and Directory Basics.....	10
Topic 3: How Search Engines Rank Documents.....	12
Topic 4: Characteristics of Searchers and What Takes Search Time.....	13
Part 3: Keywords – The Essence of the Search.....	15
Topic 5: Sample Information Problem for this Tutorial.....	15
Topic 6: Query Concepts: What, Where, When, How, Why.....	15
Topic 7: Breaking Down Your Query	16
Topic 8: Focus on Nouns and Objects	17
Topic 9: Word Stemming and Use of Wildcards.....	18
Topic 10: Finding the Right Level.....	19
Topic 11: Synonyms.....	22
Topic 12: Use of Phrases	24
Part 4: Boolean Basics	26
Topic 13: Boolean Overview.....	26
Topic 14: AND Operator.....	27
Topic 15: OR Operator	28
Part 5: Advanced Operators	30
Topic 16: NEAR Operator	30
Topic 17: BEFORE and AFTER Operators	30
Topic 18: AND NOT Operator	31
Part 6: Advanced Construction	33
Topic 19: Use of Parentheses	33
Topic 20: Combining Concepts for Power Searching	35
Topic 21: Punctuation and Capitalization	36
Topic 22: Multiple Queries and Query Refinements	36
Topic 23: Sample Information Problem Revisited.....	36
Part 7: Pitfalls to Avoid.....	38
Topic 24: Avoid Misspellings	38
Topic 25: Redundant Terms.....	38
Topic 26: Ignored Terms and Special Characters	39
Topic 27: Alternate Spellings.....	39
Topic 28: Too Many Terms, Synonyms	39
Topic 29: Improper Boolean or Complicated Construction	40
Part 8: Using Filters	43
Topic 30: Site Filters	43
Topic 31: Size Filters.....	44
Topic 32: Date Filters	44
Topic 33: Specialty Filters and Search Options.....	45
Part 9: Understand Your Engines	46
Topic 34: Some Caveats: The Dynamic Search Business.....	46

Topic 35: Duplication, Coverage and Responsiveness.....	46
Topic 36: Boolean or Not?.....	48
Topic 37: A Comparison of 100 Search Services	49
Topic 38: Features of the Top 10 Search Services.....	50
Topic 39: Specialty Engines	53
Topic 40: Some Other Services to Watch	54
Topic 41: Some Perplexing Behaviors.....	56
Part 10: Specialty Searches.....	58
Topic 42: Product Searches	58
Topic 43: Competitor Intelligence.....	58
Topic 44: Market Research	58
Topic 45: Finding People.....	59
Topic 46: Finding Places	59
Topic 47: Finding Documents.....	59
Topic 48: Finding Recent News.....	59
Part 11: Solutions and the Future of Searching	60
Topic 49: Ruminations on the Future of Internet Searching	60
Part 12: Summary and Further Information.....	62
Section 2: Using a Powerful Desktop Resource – Mata Hari®	64
Topic 50: Mata Hari Product Features	64
Topic 51: What is Fast?	65
Topic 52: Universal Search Power	66
Topic 53: Nearly 600 Search Engines in 60 Topic Categories.....	66
Topic 54: Using Boolean Power with Non-Boolean Search Engines	66
Topic 55: Filtering, Phrases, and Plain Text.....	67
Topic 56: Efficiently Culling Results	67
Topic 57: Local Viewer	68
Topic 58: Query and Engines Folder.....	68
Topic 59: Scores Folder	68
Topic 60: Terms Folder	69
Notes, Links and References.....	70

Section 1: Searching with Internet Provided Resources

Looking for that perfect condo for your ski trip? Needing specifications for a manufacturer's particular piece of equipment? Want discussion and commentary on your favorite, but obscure, author? Trying to find out what your competitors are up to? Seeking recent studies on planets in other solar systems? Needing information on special scholarships for which you might be qualified?

These, and millions of queries covering every conceivable topic, are now being posed daily to the Internet's search services. With about 800 million or more publicly available documents – an amount remarkably doubling every 18 months – the Internet has become a vast, global storehouse of information. The only problem is: how do you find what you're looking for?

Unfortunately, there is no Dewey decimal system or central “card catalog” for the Internet. You must use a search service to find new information. Search services come in one of two main flavors. Each has its place, depending on your information needs.

‘Directories’ use trained professionals to classify useful Web sites into a hierarchical, subject-based structure. Yahoo is the best known and most used of these services. Directories are most useful when looking for information in clear categories, such as makers of yogurt or listings of educational institutions. Each directory uses its own categories and means to screen useful sites and assign them to a single category.

‘Search engines’ work differently. Excite, AltaVista and Infoseek are some of the best known engines. They “index” (record by word) each word within all or parts of documents. When you pose a query to a search engine, it matches your query words against the records it has in its databases to present a listing of possible documents meeting your request. Search engines are best for searches in more difficult topic areas or those which fall into the gray areas between the subject classifications used by directories. But, search engines are stupid, and can only give you what you ask for. You can sometimes get thousands (millions!) of documents matching a query. Also, at best, even the biggest search engines only index up to one third of the Internet's public documents.

So, while three quarters of the users cite finding information as their most important use of the Internet, that same percentage also cite their inability to find the information they want as their biggest frustration. The purpose of this tutorial is to help you end that frustration.

Your ability to find the information you seek on the Internet is a function of how precise your queries are and how effectively you use search services. Poor queries return poor results; good queries return great results. Contrary to the hype surrounding “intelligent agents” and “artificial intelligence,” the fact remains that search results are only as good as the query you pose and how you search. There is no silver bullet.

Most Internet searchers, perhaps including you, tend to use only one or two words in a query. Big mistake! Also, there are very effective ways to “structure” a query and use special operators to target the results you seek. Absent these techniques, you will spend endless hours looking at useless documents that do not contain the information you want. Or you will give up in frustration after search-click-download-reviewing long lists of documents before you find what you want.

All of us need information. But few of us have studied information or library science, and not everyone has used search services or Internet search engines sufficiently to learn all of the

nuances. This tutorial is for those who are learning the ropes about ‘power searching.’ But, even if you’re quite experienced in these areas, you might find some benefit from glancing through these topics.

This tutorial is organized to proceed from the basics to more advanced topics. It is divided into two sections: “Searching with Internet Provided Resources” and “Using a Powerful Desktop Resource – **Mata Hari**”. The first section has 12 parts containing 51 topics and describes the search services, available operators, and the extremely important information on how to compose your queries. The second section contains 11 topics and describes using our tool: **Mata Hari** — which we believe is the most powerful search agent ever developed. As heavy duty searchers ourselves, we had to create **Mata Hari** to automate and expedite the search process for our own needs. A description of its features and how it works is provided for your own assessment of whether or not you can benefit from this powerful tool.

Simple to follow examples are presented in each topic. We’ve written it to be a one-stop reference. Don’t feel you need to work through all of the topics in one sitting. But, if you do take the time to work through this material, we guarantee you’ll reap big dividends in faster and more accurate results. And, you will be on your way to earning the title of an Internet “Power Searcher.”

Documentation is appended at the end [1,2].

Executive Summary: The Two-Minute Bottom Line

To illustrate some of the basic concepts and recommendations covered in this tutorial, let's say we have an interest in recent findings about new planets being discovered outside our solar system. Using the information "contained" in this statement, you can see how an effective query can be built by following these guidelines.

We'll summarize the recommendation, show how the statement is phrased, describe why it's important, and provide a pointer to the specific topic number in the tutorial that covers this recommendation. See the table of contents for relating topic numbers to subject titles.

Recommendation	Example	Why Important?	Topic #
1. Use nouns and objects as query keywords	planet or planets	Actions (verbs), modifiers (adjectives, adverbs, predicate subjects), and conjunctions are either "thrown away" by the search engines or too variable to be useful	6, 7, 8
2. Use 6 to 8 keywords in query	new, planet, planets, discovery, solar, system	More keywords, chosen at the appropriate "level", can reduce the universe of possible documents returned by 99% or more	8,10
3. Truncate words to pick up singular and plural versions	planet* or discover*	Use asterisk wildcard. The wildcard tells the search engine to match all characters after it, preserving keyword slots and increasing coverage by 50% or more	9, Sec. 2
4. Use synonyms via the OR operator	discover* OR find	Cover the likely different ways a concept can be described; generally avoid OR in other cases	11, Sec. 2
5. Combine keywords into phrases where possible	"solar system"	Use quotes to denote phrases. Phrases restrict results to EXACT matches; if combining terms is a natural marriage, narrows and targets results by many times	12
6. Combine 2 to 3 "concepts" in query	"solar system" "new planet" discover* OR find	Triangulating on multiple query concepts narrows and targets results, generally by more than 100-to-1	20
7. Distinguish "concepts" with parentheses	("solar system") ("new planet") (discover* OR find)	Nest single query "concepts" with parentheses. (Overkill for now, but good practice when first learning.) Simple way to ensure the search engines evaluate your query in the way you want, from left to right	19
8. Order "concepts" with subject first	("new planet") (discover* OR find) ("solar system")	Put main subject first. Engines tend to rank documents more highly that match first terms or phrases evaluated	7, 19, 20

Recommendation	Example	Why Important?	Topic #
9. Link "concepts" with the AND operator	("new planet*") AND (discover* OR find) AND ("solar system")	AND glues the query together. The resulting query is not overly complicated nor nested, and proper left-to-right evaluation order is ensured	14, 20, Sec. 2
10. Issue query to full "Boolean" search engine or metasearcher	As above	Full-Boolean engines give you this control; metasearchers increase Web coverage by 3- to 4-fold	3, 35, 36, 38, Sec. 2

By issuing the query in **#9** above to AltaVista, we are able to restrict results from a baseline of 917,754 documents using the query **new AND planet** (actually 1,139,837 if we were to properly include **planets** as well) to a count of 2,036 documents [1].¹ Though that number still seems like a lot, we have reduced our possible universe of results by 400 to 600 times, and four of the first five documents listed give us exactly what we were looking for:

<http://www.got.net/~seasons/new.html>
<http://www.ucar.edu/quarterly/summer97/planet.html>
<http://www.geocities.com/Area51/Nebula/1456/todaysnews.html>
<http://www.npr.org/news/healthsci/indexarchives/1998/May/980529.01.html>

Go ahead; try these queries for yourself!

The ultimate bottom line to getting the best results for your queries is to search multiple services simultaneously using a universal format. Our solution is to provide you full Internet searching power at your desktop via the **Mata Hari®** product [Section 2]

Do you want to be able to get such impressive results for your own queries? Then, welcome. It's now time to start the tutorial.

¹ Here is a measure of document growth on the Web. When we first issued the **new AND planet** (or **planets**) query in April 1998, the Alta Vista counts were 418,934 and 551, 936, respectively. The recommended #9 query also returned 934 results. Thus, in only one year, Alta Vista documents on this subject grew by about 120%!

Part 1: The Size of the Internet

The Internet is a vast place comprised of millions of computers sending information back and forth in packets. It came into being in the early 70s as a U.S. Defense Department network called ARPAnet. This was an experimental network created for military research, initially, for the design and testing of network survival under wartime conditions.

Before long, academic institutions and private companies performing military research were added followed by nonmilitary related communications between other academic facilities. During this period, the net grew slowly and constantly as the need to communicate and share research and technical information increased. In the early 90s, the pace quickened greatly as the personal computer became more affordable and those with access to the net at work also wanted that access from home and smaller offices.

The Mosaic browser developed in 1993 by Marc Andressen at the National Center for Supercomputing Applications at the University of Illinois in Urbana-Champaign was the basis for the graphical web browser commercialized with fantastic success by Netscape. Via the World Wide Web, the Internet became available to the masses and business saw it as the next frontier for commerce.

Though the Internet is comprised of various sections from newsgroups through email services, by far, the most popular and fastest growing is the World Wide Web, cited as most important by two-thirds of the users, followed by electronic mail.

Web Size and Demographics

- There are 5 million Web sites and perhaps 200 million Internet users worldwide. The information below is the latest available, but was compiled at different times for different regions over the last two years [3].

Africa		Russia	1,000,000
Algeria	750	Singapore	500,000
Angola	1750	South Korea	3.1 million
Benin	2000	Sri Lanka	14,000
Botswana	1000	Taiwan	3.01 million
Burkina Faso	700	Thailand	131,000
Burundi	150	Vietnam	6,000
Cameroon	2000		
Cap Verde	50	Europe	
Central African Republic	200	Austria	442,000
Comoros	200	Belgium	1.4 million
Congo-Kinshasa	200	Czech Republic	270,000
Djibouti	300	Denmark	1.1 million
Egypt	40,000	Estonia	152,000
Equatorial Guinea	200	Finland	1.43 million
Eritrea	300	France	2.5 million
Ethiopia	2400	Germany	7.3 million
Gabon	1000	Greece	111,000
Gambia	150	Hungary	200,000
Ghana	4500	Iceland	121,074
Ghana	4500	Ireland	388,176
Guinea	300	Italy	2.6 million
Guinea Bissau	150	The Netherlands	2.3 million
Ivory Coast	2000	Norway	601,000
Kenya	15000	Poland	700,000

Lesotho	200	Portugal	188,000
Liberia	75	Slovak Republic	510,000
Madagascar	700	Spain	2,2 million
Malawi	2000	Sweden	3,3 million
Mali	500	Switzerland	1.2 million
Mauritania	100	Turkey	600,000
Mauritius	1000	United Kingdom	10.6 million
Morocco	20,000		
Mozambique	3500	Middle East	
Namibia	2000	U.A.E.	88,552
Niger	200	Bahrain & Saudi Arabia	46,538
Nigeria	1000	Israel	600,000
Reunion	500	Jordon	20,213
Rwanda	100	Kuwait	42,350
Senegal	2500	Lebanon	43,828
Seychelles	1000	Oman	20,888
Sierra Leone	150	Qatar	17,295
South Africa	900,000	Yemen	2426
Sudan	300		
Swaziland	900	North America	
Tanzania	2500	United States	73 million
Tchad	200	Canada	7.63 million
Togo	300	Mexico	504,900
Tunisia	7000	Central America	7 million
Uganda	3000		
Zambia	3000	South America	
Zimbabwe	10,000	Argentina	170,000
		Bolivia	8,000
Asia/Pacific		Brazil	3.5 million
Australia	4.36 million	Chile	200,000
China	1.5 million	Colombia	120,000
Bangladesh	7000	Costa Rica	50,000
Hong Kong	850,000	Ecuador	5,000
India	500,000	Haiti	2,000
Indonesia	80,000	Paraguay	1,000
Japan	14 million	Peru	65,000
Malaysia	600,00	Uruguay	9,000
New Zealand	561,300	Venezeula	35,000
Philippines	320,000		

- The U.S. accounts for 47% or more of Internet users, though that percentage is dropping.
- Consensus predictions are for 200 million global users by the year 2000 [4].
- Currently, the Yankee Group reports that 25 percent of US households are online.
- Market research by Ovum predicts that, worldwide, the number of people with access to the Internet will quadruple by the year 2005. The firm also predicts that the U.S. market will reach saturation in 2002 and that the Western European and Asian markets will continue to grow, but at a slower pace [5].
- The top 25 Internet sites, 11 of which are search-related portals, measured in thousands of unique monthly users as of November 1999 [6]:

1	AOL Network	54,211
2	Yahoo Sites	41,786
3	Microsoft Sites	38,334
4	Lycos	29,150
5	Go Network	21,981
6	Amazon	15,357
7	Excite@Home	14,975
8	Bluemountainarts.com	12,742
9	Time Warner Online	12,445
10	Go2Net Network	11,472
11	About.com Sites	11,032
12	Real.com Network	10,842
13	eBay	10,553
14	AltaVista Sites	10,485
15	CNET	9,393
16	LookSmart	9,172
17	Snap Sites	9,124
18	ZDNet Sites	8,866
19	Xoom.com Sites	8,457
20	Goto.com	7,347
21	Juno / Juno.com	7,264
22	Macromedia	6,450
23	AT&T Web Sites	6,351
24	Smartbotpro.net	6,214
25	Infospace Impressions	5,922

- While the educated and affluent have been among the first to use the Internet since the boom began, mainstream America has been joining in over the last two years. Conversely, the rural poor are the least connected [7].
- Demographics of the Web have changed dramatically as the user community has evolved from the initial innovators and early adopters; today, demographic changes are stabilizing. In the early years, younger white males dominated the Web, most from educational institutions. Yet with growth, the transition is toward the demographics of the society as a whole (as well as more international). As more people have gotten on the net, especially in the U.S., the demographics of the user population are becoming closer to the population as a whole, as we should expect. There has also been a greater diversification of vocations rather than the computer and education emphasis of the past [8].

Web User Characteristics [8]

- Over 27% of Web users are trained professionals followed by 15% in middle management
- Quality of information is the most important consideration for why users visit a particular Web site
- Two-thirds to three-quarters of all users cite finding information as one of their primary uses of the Internet
- Two-thirds to three-quarters of all users cite the inability to find the information they seek as one of their primary frustrations (second only in frustration to slowness of response)
- Perhaps 31% of those on the Internet are 'heavy' users, online each week for 20 hours or more, with another 34% online for 10-20 hours
- Estimates from VisualMetrics Corporation. are that Internet users (in thousands) break into these on-line categories [9]:

Calendar Year	1997	1998	1999	2000
Power Users I (20+ hrs)	11,100	21,800	42,000	82,600
Power Users II (10-20 hrs)	14,500	24,000	39,100	65,100
Learning Users (5-19 hrs)	17,000	23,700	29,500	32,100
Casual Users (>5 hrs)	<u>7,400</u>	<u>10,500</u>	<u>14,400</u>	<u>20,200</u>
Total Internet Users	50,000	80,000	125,000	200,000

- For active users, links from other Web sites are four times (54%) more effective in creating awareness for a given Web site than via a search engine (13%); newsgroups (10%), direct email (6%) and word-of-mouth (6%) are other effective means for gaining Web site awareness
- Sites are visited more frequently based on an active users' bookmarks (82%) than for search engines; in fact, the average active Web user now has about 40 bookmarks of frequently visited sites

Web Documents and Searching

- Alexa estimates that approx. 1.5 million Web pages are being added daily, with an overall doubling rate for Web documents every 8 months [10]
- The Internet now contains 800 million documents [11]; this is a significant increase from the 320 million estimated in 1998 [12,13]
- International Data Corp. estimates there will be more than 1 billion Web documents by the year 2000 [14]
- Of the major search engines, estimated coverage of the documents on the Internet ranges from a high of 34% to a low of 3% in 1998 [12], with the highest percentage dropping to 16%! In 1999 [11]
- The average user enters 1.5 keywords per search engine search
- Combining multiple search engines in a given search can increase the likelihood of finding the information desired by a factor of 3.5 or more [12]
- Different search formats and conventions make it difficult to search multiple engines at one time
- There are more than 2,500 search services presently on the Web [15,16]. Below, are the twenty or so major Internet search services:

AltaVista
[\[http://www.altavista.com\]](http://www.altavista.com)
 Ask Jeeves
[\[http://www.askjeeves.com\]](http://www.askjeeves.com)
 Direct Hit
[\[http://www.directhit.com\]](http://www.directhit.com)
 Excite [\[http://www.excite.com\]](http://www.excite.com)
 LookSmart
[\[http://www.looksmart.com\]](http://www.looksmart.com)
 Go [\[http://www.go.com\]](http://www.go.com)
 Google
[\[http://www.google.com\]](http://www.google.com)

HotBot
[\[http://www.hotbot.com\]](http://www.hotbot.com)
 Infoseek
[\[http://www.infoseek.com\]](http://www.infoseek.com)
 Inktomi
[\[http://www.inktomi.com\]](http://www.inktomi.com)
 LookSmart
[\[http://www.looksmart.com\]](http://www.looksmart.com)
 Lycos [\[http://www.lycos.com\]](http://www.lycos.com)
 Magellan
[\[http://magellan.excite.com\]](http://magellan.excite.com)
 About.com
[\[http://www.about.com\]](http://www.about.com)

NetFind (AOL)
[\[http://www.aol.com\]](http://www.aol.com)
 Northern Light
[\[http://www.northernlight.com/\]](http://www.northernlight.com/)
 Open Directory
[\[http://dmoz.org\]](http://dmoz.org)

RealNames
[\[http://www.realnames.com\]](http://www.realnames.com)
 Snap! [\[http://www.snap.com\]](http://www.snap.com)
 WebCrawler
[\[http://www.webcrawler.com\]](http://www.webcrawler.com)
 Yahoo [\[http://www.yahoo.com\]](http://www.yahoo.com)

- There are also ‘metasearch’ services that provide a central access point to multiple of these services. Notable names – again, not suggesting endorsement – are:

Dogpile [\[http://www.dogpile.com\]](http://www.dogpile.com)
 Inference FIND [\[http://www.infind.com/\]](http://www.infind.com/)
 Internet Sleuth [\[http://www.isleuth.com\]](http://www.isleuth.com)
 Mamma [\[http://www.mamma.com\]](http://www.mamma.com)
 Metacrawler [\[http://www.metacrawler.com\]](http://www.metacrawler.com)
 SavvySearch [\[http://www.savvysearch.com\]](http://www.savvysearch.com)

- Not all documents on the Web are indexed. Many are not indexed by any search engine or directory at all. Of those indexed, no engine covers even half of the indexed pages (though Alta Vista purportedly comes close). And besides coverage, there is also the question of keeping the links up-to-date. NEC Research Institute tested the degree that links can be out-of-date and, therefore, useless (dead links). Below are the reported coverage statistics [12] and the percent of dead links [17] reported in 1998:

Search Engine	% of all indexed pages	% that are dead links
Alta Vista	47%	2.5%
Northern Light	39%	5.0%
Inktomi 34%	Not Available	
Excite 17%	2.0%	
Lycos 16%	1.6%	
InfoSeek	14%	2.6%

An update of that report in 1999 showed that even the most comprehensive Internet search engines – HotBot, Northern Light and AltaVista – individually catalog at most 16% of the Internet’s documents [11], a significant erosion in coverage from the 35% to 45% estimated in the previous year

- The percentage of Internet users citing one or more of these typical search problems were: searches that turn up broken links (50%); not finding information known to exist (30%); inability to organize collected information (29%); and being unable to find pages already visited (12%) [8]
- More than 98% of active Web users rely on the Internet to find reference material, 30% on a daily basis and a further 40% on a weekly basis [8]
- Use of structured, or ‘Boolean’ queries, while known to help obtain better search results, can be difficult and frustrating for some users to learn.

Part 2: Internet Search Basics and Why There's a Problem

Much is discussed on the Internet regarding its growth and user-driven, decentralized nature. This part overviews the current state of searching and search services on the Internet. The essential arguments are that your time is well spent learning how to issue more effective queries and to understand the basic operations of the search services you employ.

Topic 1: Status of the Internet and Searcher's Frustrations

Many have likened the Internet to a huge, global library. While true in some aspects, it has some unique differences. There is no central "card catalog"; the Internet's growth is outpacing the ability of humans or technology to keep up with it; its sheer size is unknown and perhaps unknowable; and content is (to say the least) of uneven quality. Here's some of what we know (or think we know) about information on the Internet:

- There are 800 million or more documents publicly available on the Internet [11]
- Document growth is, at minimum, doubling each year [18]
- Two-thirds to three-quarters of all users cite finding information as one of their primary uses of the Internet [8]
- Two-thirds to three-quarters of all users cite the inability to find the information they seek as one of their primary frustrations (second only in frustration to slowness of response) [8]
- Of the major search engines, estimated coverage of the documents on the Internet ranges from a high of 34% to a low of 3% [12], with the highest coverages dropping to 16%! in just the past year [11]
- Combining multiple search engines in a given search can increase the likelihood of finding the information desired by a factor of 3.5 or more [12]
- Different search formats and conventions make it difficult to search multiple engines at one time
- Use of structured, or 'Boolean' queries, while known to help obtain better search results, can be difficult and frustrating for some users to learn.

One of the challenges of the Internet is to make its value available to the millions of new users who have had no formal training or experience in query formulation or search strategies.

Topic 2: Search Engine and Directory Basics

The major search services on the Internet are essential starting points for users seeking information. As such, they routinely are some of the most visited locations on the Web.

Search services can be divided into two groups, commercial and non-commercial. Commercial search services go to the effort to catalog information on the Internet to attract attention and advertising revenues. Non-commercial services exist for many different reasons. There are more than 2,500 search services presently on the Web [16]. There are a dozen or more big, major Internet search services. There are also 'metasearch' services that provide a central access point to multiple of these services.

Search services on the Internet come in two main flavors: 1) 'search engines' that index words or terms in Internet documents; and 2) 'directories' that classify Web documents or locations into an arbitrary subject classification scheme or taxonomy. Most of the above are examples of the former; Yahoo, About.com and LookSmart are examples of the latter.

Search engines use 'spiders' or 'robots' to go out and retrieve individual Web pages or documents, either because they've found them themselves, or because the Web site has asked to be listed. Search engines tend to "index" (record by word) all of the terms on a given Web document. Or they may index all of the terms within the first few sentences, the Web site title, or the document's metatags [19]. Due to the ever-changing nature of the Internet, the services must re-sample their sites on a periodic basis. Some of these services re-sample their sites on a weekly or less-frequent basis.

Precision, recall and coverage are limiting factors for most search engines. Precision measures how well the retrieved documents match the query; recall measures what fraction of relevant documents are retrieved [20]. Coverage refers to what percentage of the potential universe of relevant documents is cataloged by the engine. For example, consider a search engine with 10 documents, five of which mention eagles, out of a total universe of 50 potential documents mentioning eagle (45 of which are not indexed by that engine). A query on eagle that returned four documents and two others from this engine would have a precision of 0.66, a recall of 0.80 and coverage of 0.10.

Precision is a problem because of the high incidence of false positives. (That is why you get so many seemingly irrelevant documents in your searches.) This is due to imprecision in the query (searching on eagle and missing the mention of eagles), indexing mistakes by the engine, and keywords entered by the Web document developer that do not actually appear in the document. Coverage is a problem for all engines, with the largest ones only covering at most one sixth to one third of publicly-available documents [11,12,,21].

Search directories operate on a different principle. They require people to view the individual Web site and determine its placement into a subject classification scheme or taxonomy. Once done, certain keywords associated with those sites can be used for searching the directory's data banks to find Web sites of interest.

These distinctions by search service are not clean in all cases. The Excite search engine, for example, uses 'morphological analysis' for determining its keyword matches [12]. While construction of the index is more akin to a search engine, in operation Excite can work like a directory. As other search engines begin classifying information into directory-like clusters, these distinctions are likely to continue to get fuzzier.

For searches that are easily classified, such as vendors of sunglasses, the search directories tend to provide the most consistent and well-clustered results. This advantage is generally limited solely to those classification areas already used in the taxonomy by that service. Yahoo, for example, has about 2,000 classifications (excluding what it calls 'Regional' ones, which are a duplication of the major classification areas by geographic region) in its current taxonomy. When a given classification level reaches 1,000 site listings or so, the Yahoo staff split the category into one or more subcategories. If a given topic area has not been specifically classified by the search directories, finding related information on that topic is made more difficult. Another disadvantage of directories is their lack of coverage because of the cost and time in individually assigning sites to categories.

Most searches of a research or cross-cutting nature tend to be better served by the search engines. That is because there is no classification structure behind the listings; only whether the keywords requested appear in that search engine's index database or not.

The flexibility of indexing every word to give users complete search control, such as provided by AltaVista or OpenText, is now creating a different kind of problem: too many results. In the worst cases, submitting broad query terms to such engines can result in literally millions of potential documents identified. Since the user is limited to viewing potential sites one-by-one, clearly too many results can be a greater problem than too few.

Increasingly, the growth of the Internet is causing the specialization or balkanization of search services. Lawyers, astronomers or investors, for examples, may want information specifically focused on their interest topics. By cataloging information in only those areas, users interested in those topics are better able to keep their search results bounded. Such specialization can also lead to more targeted advertising on those search service sites. Again, though, like the directories, such specialization can limit search results to the boundaries chosen by the service, which may or may not conform to the boundaries sought by the user.

The ultimate challenges to any of these centralized search services, therefore, are to: 1) keep pace with explosive document growth; 2) understand the “boundary” needs of their user communities; 3) provide sufficient “intelligence” to infer what users are really asking for even when their queries don’t specify it; and 4) ensure sufficient coverage to provide one-stop searching. In the race for eyeballs, user retention and repeat visits are key.

Topic 3: How Search Engines Rank Documents

A Web page, or document, can contain various kinds of content (as opposed to display or presentation options like sound, animation or frames), some of which is not shown when you view the document in your browser [2]:

- **Title** – an embedded description provided by the document designer; viewable in the titlebar (it is also used as the description of a newly created bookmark by most browsers)
- **Description** – a type of metatag which provides a short, summary description provided by the document designer; not viewable on the actual page; this is frequently the description of the document shown on the documents listings by the search engines that use metatags
- **Keywords** – another type of metatag consisting of a listing of keywords that the document designer wants search engines to use to identify the document. These too, are not viewable on the actual page
- **Body** – the actual, viewable content of the document.

Search engines may index all or some of these content fields when storing a document on their databases. (Over time, engines have tended to index fewer words and fields.) Then, using proprietary algorithms that differ substantially from engine to engine, when a search query is evaluated by that engine its listing of document results is presented in order of ‘relevance.’ Because of these differences in degree of indexing and algorithms used, the same document listed on different search engines can appear at a much higher or lower ranking (order of presentation) than on other engines.

Though not hard and fast, and highly variable from engine to engine, four factors tend to influence greatly the ranking of a document in a given query:

1. **Order a keyword term appears** – keyword terms that appear sooner in the document’s listing or index tend to be ranked higher
2. **Frequency of keyword term** – keywords that appear multiple times in a document’s index tend to be ranked higher

3. **Occurrence of keyword in the title** – keywords that appear in the document’s title, or perhaps metatag description or keyword description fields, can be given higher weight than terms only in the document body
4. **Rare, or less frequent, keywords** – rare or unusual keywords that do not appear as frequently in the engine’s index database are often ranked more highly than common terms or keywords.

Some engines, notably Excite, attempt to “infer” what you mean in a query based on its context. Thus, the meaning of **heart** can differ if the context of your search is cardiac disease as opposed to Valentine’s Day. The methods by which these inferences are made are statistically based on the occurrence of some words in conjunction with others. Though useful for simpler queries, such inference techniques tend to break down when the subject of the query or its modifiers do not fit expected query relationships. For commonly-searched topics, this is generally not a problem; for difficult queries, it is a disadvantage to standard full-text indexing.

Cottage industries have emerged to help Web site developers place themselves higher in the search engines’ listings (it is clearly more valuable to be within the first few listings sent to a user than be buried hundreds, or thousands, of documents lower). A constant battle is being waged between the engines and those desiring high listings from jimmying the system to “unfair” advantage.

Crude, early attempts to “spam” search engines to get higher listings included adding hidden terms like “sex” that were searched frequently but not the real subject of the document. Other techniques were to use certain keywords repeatedly, such as “cars cars cars cars cars” to get a higher frequency rating. Another was to cram the page with high-interest terms using the same color as the overall Web page, thus “hiding” the added keywords. The leading search engines have caught on to these and now have automated ways to prevent the worst of these spamming techniques.

More subtle techniques, however, are hard to prevent. For example, a listing for ski resorts in Utah could also add hidden tags for “Caribbean” or “beach resort” knowing that wealthy Caribbean travelers may also be looking to take ski vacations. If you as the searcher asked for Caribbean vacations you may logically wonder why you’ve gotten a listing for Utah ski resorts. It is because of such techniques (among others) that you can sometimes get document listings from a search that seemingly have nothing to do with your query.

So, differences in how search services rank documents, how developer’s themselves choose to characterize their Web documents, and just simple errors in how computers process and index these pages can all lead to highly variable ranking results from different search services.

Topic 4: Characteristics of Searchers and What Takes Search Time



Professional information searchers do not have a single style. There is no “correct” way to search on the Internet. Search styles have been described as ranging from ‘ants’ – the carefully planned, methodical search hoping to get exact results on the first try – to



‘grasshoppers’ – intuitively jumping from topic to topic, refining results as more is learned [23]. Only you can determine what your style is.

There is only one meaningful measure for a successful search: getting the results you desire. And within that context, there is only one meaningful basis for judging whether one search strategy or another is superior: whether those results are obtained faster.

Surfing and browsing on the Internet are seductive. One begins with an objective in mind, finds new tidbits of interest, and hours later can wonder where the time has gone. It is often difficult to apply metrics against whether the original search interest was obtained, or whether the whole process was productive or not. So, let's look at some aspects of a typical search. The example assumes a 56.6 KB modem and a relative "fast" time for the Internet [9]. This is perhaps an optimistic mid-range for current users of the Internet. The example is only meant to be illustrative:

Search Step	Est. Process Time (sec.)	No. Repeats	Total Time (min.)	Cumulative Time (min.)
Formulate Query	120	3	6.0	6.0
Issue Search	10	3	0.5	6.5
Get Search Listings from Service (30/query)	10	9	1.5	8.0
Review Documents; Select for Download	12	50	10.0	18.0
Download Document	15	50	12.5	30.5
Review Document	18	50	15.0	45.5
Average Time per Document (90 document example)				0.5

These estimates are likely an underestimate. Information professionals using the Web to do searches in comparison with traditional online search services like Dialog found it took on average 2.4 minutes per document to get acceptable results [24].

Whatever the actual "average" search time is, it will not apply to your circumstances in any case. However, what is the case is that certain aspects of searching can add delays to getting desired results and increase frustration:

- No matter how precise or accurate the query, a large percentage of results returned by search services will not be what you're looking for
- Actual search time in getting candidate listings from services is relatively fast; the one-by-one document download and review is the most time consuming part of the process
- Larger listings of candidate documents from the services require more evaluation time
- Often too little time is spent on search and query formulation; any improvements you can make toward more precise and accurate queries will lead to fewer documents to review and faster overall times to the results you want.

The essential conclusion is that time is well-spent in understanding how to pose a proper query and how to take advantage of the way that search services work. These topics are the focus of the rest of this tutorial.

Part 3: Keywords – The Essence of the Search

Despite all the gobbledygook about things like ‘Boolean’ and query operators, the most difficult – and fundamental – aspect of a search are the keywords used in your query.

A search is inherently looking for information about a **topic**. This part describes how you can proceed from search concepts to identifying the specific keywords – or terms – that will give you the results you’re seeking. We begin by presenting an information problem which will be the basis for progressing through the tutorial’s remaining topics.

Topic 5: Sample Information Problem for this Tutorial

Jan is an office worker in downtown Minneapolis. While on lunch break one fine Spring day, Jan’s eye is caught by a flash in the sky above. Jan sees a bird about the size of a crow diving at high speed and catching in mid-air what appears to be a pigeon. The bird then swoops out of sight. Jan is captivated by the mostly gray and white bird, with the crooked black and yellow beak. Jan has never seen this bird before, and wonders what it is doing in the city. That night, Jan decides to find out more about this mystery bird on the Internet.

Where does Jan begin?

Topic 6: Query Concepts: What, Where, When, How, Why

Mastering the concepts behind a search is not as complicated as may seem at first. The first few searches are perhaps difficult, but, once done, the nuggets behind your information request start becoming clear. Like riding a bike for the first time, it does take some practice.

One of the bigger mistakes you can make in preparing a query is not providing enough keywords. On average, most users submit 1.5 keywords per query [25]. This number is insufficient to accurately find the information you are seeking. Thus, a central task in query formulation is for you to identify a sufficient number of appropriate keywords.

If you are new to searching, the first task we recommend when formulating a search is writing down what information you are seeking. This is best done – go ahead, use some paper and a pen – in the form of some questions. Before doing a search, it is important to bound your topic as completely yet succinctly as possible. After experience is gained, you can skip writing things down and plunge right into it.

Formulating a query is akin to solving a mystery. Some pieces of information are available, but if sufficient information were available the answer would be known and there would be no need to seek more. This is the essence of a query: missing information. It is up to you, the searcher, to define your snare – the query (quarry? pun intended) – sufficiently to trap that missing information and solve the mystery.

As any good detective would, it is useful to begin by listing what you do know according to these standard categories. Jan lists these for the mystery bird:

- **WHO / WHAT?** – gray and white bird, about the size of a crow; yellow and black beak
- **WHERE?** – downtown office buildings in the City of Minneapolis
- **WHEN?** – daylight in the Spring
- **HOW?** – fast flyer, hunting pigeons (?) as prey
- **WHY?** – hunting bird; why never seen before? blown off course? is it migrating?

Of course, not all of these five categories will apply to a given query, and the specifics will obviously vary for your desired topic. But it is useful to keep these five categories in mind – the what, where, when, how and why – when analyzing the major components.

TIP

Always keep in mind the **who, what, where, when, how and why** in formulating your query.

Topic 7: Breaking Down Your Query

Let’s take the five responses to the query tests in **Topic 5** apart (yours will differ substantially, but the same ideas apply). First, there are many common words in these responses that are prepositions, conjunctions or common verbs. These include: **and, about, the, of, a, in, as, if, not, why, never, before, is and it**. These common words are referred to as “stoplist” words: they are essential to the connecting tissue in language, but they are filler in any search request. **All** search engines ignore them because they have minimal information value and are found commonly in all language. Search services include on the order of 600 of these common words in their “stoplists”; if you use them in a query they are ignored. Therefore, you should ignore them as well.

Okay, removing such words from our responses leaves these remaining words:

TIP

Never use articles, pronouns, conjunctions or prepositions – the connecting tissue in language – in your queries.

- | | | |
|---------------|--------------------|------------------|
| gray | downtown | flyer |
| white | office | hunting |
| bird | buildings | pigeons |
| size | city | blown |
| crow | Minneapolis | off |
| yellow | daylight | course |
| black | Spring | migrating |
| beak | fast | |

Now, let’s further classify these terms into three categories, similar to diagramming a sentence (but made simpler for our purposes). Let’s use the classifications of objects/nouns, actions/verbs and modifiers/qualifiers (adjectives, adverbs and predicate subjects). And, let’s now re-list these words by these categories:

- | <u>Objects</u> | <u>Actions</u> | <u>Modifiers</u> |
|-----------------------|-----------------------|-------------------------|
| bird | blown | gray |
| buildings | migrating | white |
| city | not seen | size |
| Spring | | crow |
| daylight | | yellow |
| | | black |
| | | beak |
| | | downtown |
| | | office |
| | | Minneapolis |
| | | fast |
| | | flyer |
| | | hunting |
| | | pigeons |
| | | off |
| | | course |

Not all of these categories are equally useful in a query.

Topic 8: Focus on Nouns and Objects

Almost without exception, the central keywords in your queries will be nouns. Though sometimes adverbs and adjectives can help refine your search, the key pivot point is a noun, or series of nouns. Why is this?

The most precise terms we have in language are for tangible, concrete “things” or objects. Actions and modifiers are very diverse, easily substitutable, and generally not universally applied in any given description. For, example, take the concept of “fast”. A thesaurus will give 75 or more different words for fast. Here are some counts from AltaVista [1] for numbers of Web documents containing these terms:

fast	4,014,683
speed	3,472,826
quick	3,071,867
rapid	1,191,356
fleet	527,714
swift	280,851
breakneck	11,826

Or, alternatively, take a modifying concept like ‘color’. Again, here are the AltaVista document counts:

color	4,452,189
red	6,406,755
yellow	2,347,845
blue	5,708,757
gray	1,137,864
grey	811,921
slate	143,430
white	6,742,020

TIP

The keywords in your queries will most often be nouns – and then likely no more than 6 or 8 of them.

Note three aspects about these lists. First, some modifiers are also nouns like truck ‘fleet’, the bird ‘swift’, pool table ‘slate’ or Justice ‘White’. Second, a concept like speed or color can be described in lots of ways (most of which are not shown). Third, you generally don’t know how others would describe the same thing. In our example of Jan’s mystery hunting bird [Topic 5], would someone else describe it as “fast”, “quick” or “like a bolt from the sky”? Would someone else describe the bird as “gray”, “grey”, “slate-gray” or “smoky”?

The same kind of ambiguity and substitutability applies to actions or verbs. Does the bird “fly”, “soar”, “swoop” or “glide”, or any of the other dozens of ways the act of flying can be described?

As a general rule, try to avoid using action terms and mostly try to avoid using modifiers in your queries. Where exceptions to these guidelines may make sense is when a modifier helps to precisely define your object, such as in “Limburger cheese.”

We've thus gone through a process that has led us to these possible objects as the focal points for constructing our query terms:

bird
buildings
city
Spring
daylight

The obvious main subject is **bird**. The next few topics will concentrate on it; we'll return to the other objects as we later refine our final query.

Topic 9: Word Stemming and Use of Wildcards

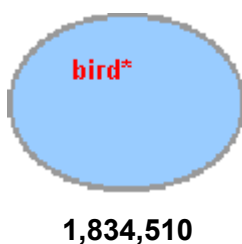
One of the first mistakes in query formulation is not using word stemming – or truncation – sufficiently. Let's look at this question in regards to our subject, **bird**. Accounting for singular and plural cases of an object is easy to overlook; but, if done, can act to unduly restrict the universe of documents in which you will be conducting your search. Using AltaVista again, here are the document counts for the single and plural versions of **bird**:



By using either only **bird** or **birds** as our subject, we would eliminate half or so of the potential documents that we'd like to use as our search basis. We could use both **bird** and **birds** as query terms, but that takes up valuable keyword slots. The better way to handle this problem is through truncation.

Truncation is applying a wildcard character after the first few letters in a term (the "stem"). The asterisk (*) is the almost universally accepted truncation wildcard. Generally, you must also have a minimum of three characters at the beginning of the word as your stem basis. Once marked for truncation, then any matching characters after that will be picked up in the search query. Some search engines do stemming and truncation for you if you pick the right option on the search form. Some engines don't support stemming or truncation at all. In any case, using the asterisk wildcard will generally be ignored or you'll get a query format error if the search engine doesn't support it.

Remember, ANY words with characters after the stem will be matched to your query term if the search engine supports truncation. Thus, if we stem **bird***, our search will match on the words **bird**, **birds**, **birding** and **birdbrain**. Posing **bird*** to AltaVista we now get these document counts:



Note the document count is a bit lower than the total for the individual words **bird**, **birds**, **birding** and **birdbrain**. There are minor errors in how search engines retrieve word stems. But they are of a smaller magnitude than ignoring singular and plural cases altogether in the query, and seem to be a minor price to pay for being able to eliminate another keyword (**birds**, in addition to **bird**) from the search.

As you first begin to use truncation you need to be aware of unintended consequences. In the case of the stem **bird*** there are relatively few unwanted words (**birdbrain**) picked up in the search. But let's look at another of the objects, **city**, in our mystery bird sample problem.

TIP
Truncation, or word stemming, keeps your keyword count down and makes for simpler queries.

To stem and pick up the plural form of **city**, **cities**, we would need to specify **cit***. But look at some of the words this stem specification would match:

citadel	citations	cited	citizen	citizenship	citriculture	citronella
citadels	cite	cities	citizenry	citrate	citrine	citrus
citation	cites	citify	citizens	citric	citrone	city

The **cit*** stem clearly picks up way too many unwanted words.

Stemming tends to work best when the actual stem is longer, when plurals are represented by an added '-s' (as opposed to '-ies' or other forms), and the stem itself is not a root to many other common words. With just a little thought, however, truncation is easy and can pay useful dividends in properly scoping your query with a minimum of keywords. We highly recommend its use.

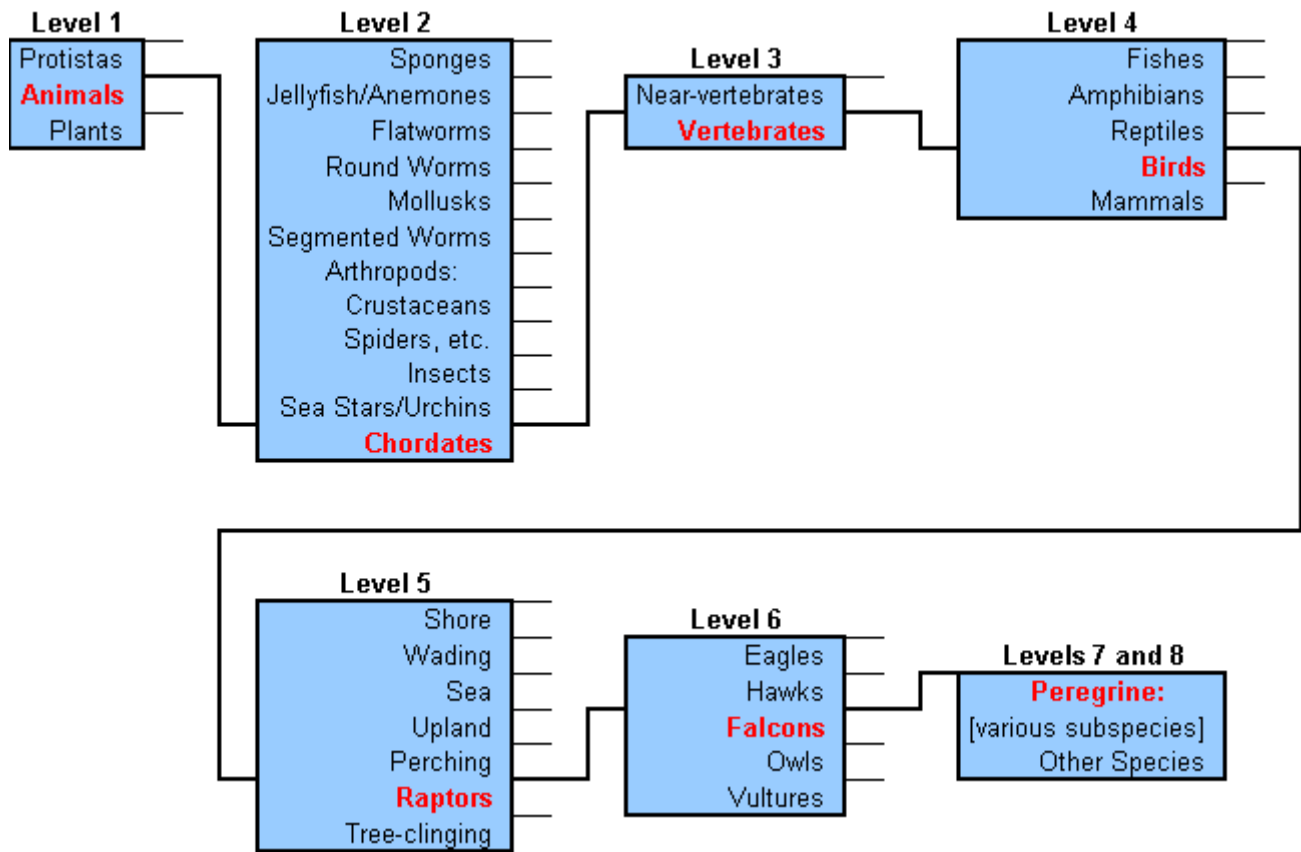
Topic 10: Finding the Right Level

Perhaps you've already noticed, but our query subject **bird*** is contained on more than 1 million documents (in AltaVista alone). It would be a little difficult to review all of those documents at one sitting.

THE MOST CRITICAL PROBLEM IN ALL QUERIES IS FINDING THE RIGHT LEVEL OF SPECIFICITY FOR THE SUBJECT QUERY TERM(S). Too broad a keyword specification, and too many results are returned; too narrow a specification, and too few are returned.

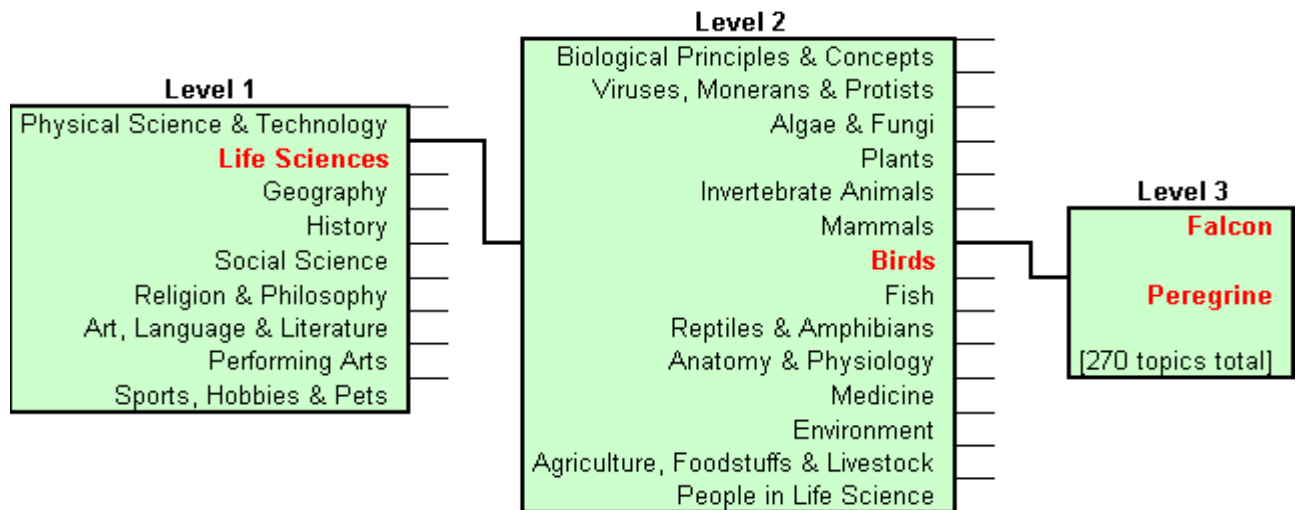
All information is classifiable and amenable to structure. We are all familiar with dictionaries, which classify words alphabetically. However, an alphabetical structure is not of much use to query formulation. But there are many other classification schemes used for information which CAN help find the right level, or specificity, for your keywords. A few examples appropriate to our mystery bird search are presented in this topic.

Our first example classification presents the structure of the animal kingdom [26]:



'Level' Example Using the Kingdom of Life

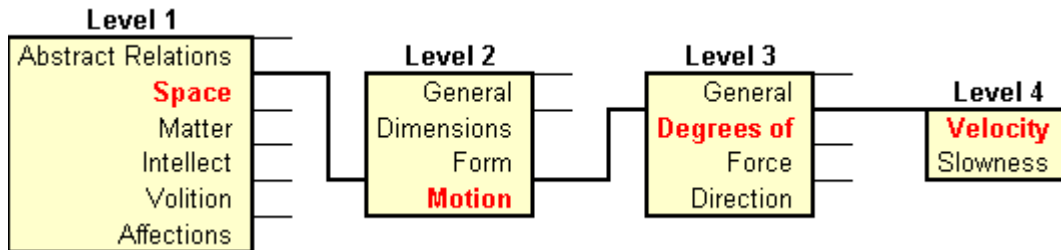
As we will see, our initial keyword term of **bird*** is at least three levels off of where it should be. Using **bird*** as is would lead to massive results sets from the search engines and virtually no likelihood that we will find the information we're looking for.



'Level' Example Using Encarta

Another way to classify information is shown by the encyclopedia, (the above example is from Microsoft’s Encarta 96 [27] – the actual encyclopedia doesn’t matter; we’re only illustrating a point).

As a very different example, the chart below shows how the word “fast” is placed within the structure of a thesaurus [28]:



'Level' Example Using Thesaurus

As noted, search ‘directories’ also apply a classification structure for how they organize and present Web sites. The structure for the largest and best known of these directories, Yahoo, with some 2000-odd individual categories, is shown on the next page.

Like the first animal phylum example above, **bird*** in the Yahoo! example is about three or four levels off from where our subject keyword should be.

Finding the right level may involve your personal knowledge and experience, doing a preliminary search or consulting other references. In the case of Jan and the mystery bird, looking in a bird book was sufficient to match pictures with the bird seen as a **peregrine falcon**.

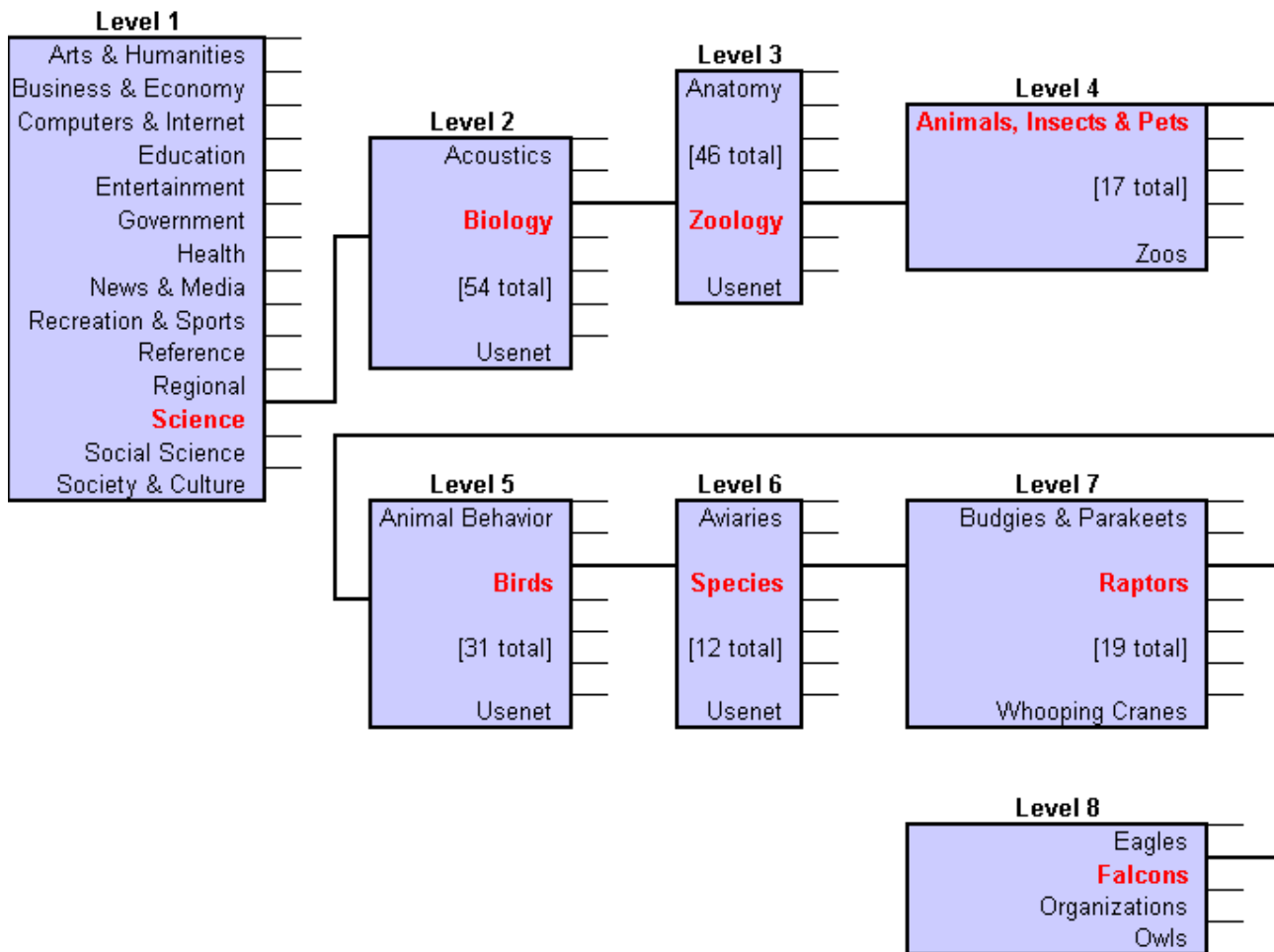
The time spent in finding how to characterize your subject at the proper level is definitely well spent, as these document counts from AltaVista illustrate:

bird*	1,834,510
falcon*	340,707
peregrine falcon*	14,510

By identifying our mystery bird as a **peregrine falcon**, we’ve narrowed the search by 99%! Remember, at 30 seconds to 2.5 minutes per document reviewed, the effort spent in zeroing in on the bird of interest has saved us tremendous overall search time.

The critical point about finding the right “level” in your keywords is that words at levels higher than where you should be return way too many results; those at levels lower than where you should be return too few or no results. This phenomenon is due to the fact that “things” at lower levels tend to “rollup” and sum into “things” at higher levels.

Philosophers, epistemologists, taxonomists, linguists and others can argue for centuries about “proper” ways to classify information. That is not our concern. Rather, the point is that keyword objects can be placed into a structure at various levels. Always keeping forefront whether your query subject is at the right level or not in those structures can bring big benefits in faster, and more accurate, searches.



'Level' Example Using Yahoo

Topic 11: Synonyms

Let's assume, however, that Jan was not able to match the bird book pictures with the mystery bird to identify it as a peregrine falcon. How can we use the query concepts identified in **Topic 6** to better hone in on what type of bird it is?

One useful place to begin is with synonyms. Jan knows the mystery bird is a hunting bird. Jan lists other synonyms that come to mind for **hunting bird**. We provide AltaVista document counts for these synonyms:

hunting bird*	2,663
bird* of prey	29,313

Jan, however, suspects neither of these terms is the "correct" synonym. Attacking this problem from another angle, Jan writes down specific kinds of birds of prey:

hawk
eagle
owl

Using these three keywords, Jan's search immediately turns up a number of sites referring to raptors, the technical term for hunting birds. Jan finds a great site on raptors that also has pictures that positively identifies the mystery bird as a peregrine falcon [29]. Jan also learns that vultures are raptors, too.

The best synonyms provide relatively complete coverage for the subject at hand and are "pitched" for the right informational objective. In Jan's case, it was needing to identify a specific bird, and a more technical term like "raptor" fit the bill. Were Jan's interest more oriented to references in novels, perhaps "hunting bird" or "bird of prey" would have been more appropriate.

An illustration of a good synonym with proper coverage is:



**Good Synonyms Provide
Good Coverage**

TIP

You can use synonyms both to find the right "level" for your query subject and to ensure proper coverage.

Good coverage is not always possible. Where not possible, provide a couple of alternate terms (that is, synonyms). But, remember, always play the numbers game. Your query terms are limited so choose them carefully.

Having determined the mystery bird to be a **peregrine falcon**, Jan considers whether synonyms for this term are also worthwhile. Based on what Jan has learned, these are the possible synonyms and document counts from AltaVista:

peregrine falcon*	14,510
<i>Falco peregrinus</i>	2,845
duck hawk*	188
all three combined	18,073

Again, note the three synonym counts do not exactly sum due to indexing gaps by the search engines. This example is a good instance where multiple synonyms do not buy enough increased coverage to be warranted. **peregrine falcon** is the most used description of this bird; adding the other terms increases coverage by only about 15%.

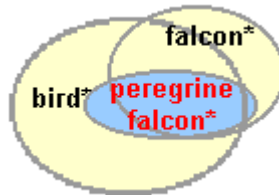
You need not get actual document counts from search engines in order to weigh such choices in your own queries. Simply use good judgment of what you're gaining – if anything – by adding more synonyms to your query subjects. Common sense should be a sufficient guide.

A thesaurus, a dictionary, personal knowledge or a preliminary Internet search can all be worthwhile places to find synonyms for the major subject(s) in your query. Generally, you should not waste the time thinking about synonyms for other terms in your queries, unless you know them to have very poor coverage.

Topic 12: Use of Phrases

Your most powerful keyword term is the phrase. Phrases are combinations of words that must be found in the search documents in the EXACT order as shown. You denote phrases within closed quotes (“**peregrine falcon***”). Some search services provide specific options for phrases, some do not allow them at all, but almost all will allow you to enter a phrase in quotes, ignoring the quotations if not supported.

Why phrases are powerful is illustrated below:



Phrases Target Results

Again, using AltaVista document counts, the ability of phrases to zero in on desired results is clear:

bird*	1,834,510
falcon*	340,707
peregrine falcon*	14,510

Phrases should be used where the constituent terms are naturally married. Other examples would be “rain in Spain”, “Gettysburg Address”, “solar system” or “big bad wolf”. Where two or more words are necessary to capture the subject, but may not always be next to one another in the same order, the **AND** or **NEAR** Boolean operators should be used [Part 4] .

[When using phrases, it is important to consider nuances of the phrase that wouldn't normally be of concern. For example, the spaces between words are as important as any other character. If you include a double space between any two words in the query and the phrase typically has only one, the search will fail. Also, sometimes two dashes are used together on Web documents to approximate an en- or em- dash. If you include only one dash, the search engine may miss all those documents that use two. There is variability in the way certain search engines treat spaces, dashes, and the like. If you suspect there may be a problem, consider submitting your phrases in different ways to capture these variations.]

TIP

Always look for natural phrases in your query concepts – they are one of the most powerful weapons available.

In addition to “**peregrine falcon***”, Jan also uses “**endangered species**” to help focus the search. Jan chose “**endangered species**” because information gained in identifying the mystery bird indicated that peregrine falcons were at risk of extinction in the 1970s due to DDT effects. Jan suspects that the answer to the why question of the search is the rarity of the bird and not migration or being blown off course. “**endangered species**” is a logical construct for a phrase because the terms are almost always used together to discuss organisms at risk of extinction.

Part 4: Boolean Basics

Despite its intimidating name, Boolean search techniques are really quite simple to learn and can add tremendous effectiveness to your searching. While working through this part, most of you will recognize constructs that were taught to you in high school math.

“Boolean” searching draws its name from George Boole, a mathematician and logician from the 19th century. He developed Boolean algebra, which is the basis for this form of structured search technique. Boolean algebra is also of prime importance to the design of modern computers.

Most information on the Web is highly unstructured. Boolean search techniques were first applied by information professionals to traditional search services like Dialog or Lexis-Nexis. Boolean techniques, while not supported by all Internet search services, provide a way for you to bring structure to this unstructured environment.

Without Boolean techniques, you are stuck with doing a lot of free-text searching, meaning, looking for documents that contain words you think will be in the document you are seeking. Sheer document volume makes free-text searching difficult and prone to failure. Boolean techniques give you the power to narrow your search to a reasonable number of potentially useful documents thereby increasing your likelihood of success.

Topic 13: Boolean Overview

Boolean logic is used to construct search statements using logical **operators** and specified **syntax**. These are combined into **Boolean expressions**, which always are either true or false when evaluated.

The shopping list of operators and syntax available to Boolean searching (though not supported by all Boolean search services) is:

- **AND** – terms on both sides of this operator must be present somewhere in the document in order to be scored as a result
- **OR** – terms on EITHER side of this operator are sufficient to be scored as a result
- **AND NOT** – documents containing the term AFTER this operator are rejected from the results set
- **NEAR** – similar to **AND**, only both terms have to be within a specified word distance from one another in order to be scored as a result
- **BEFORE** – similar to **NEAR**, only the first (left-hand) term before this operator has to occur within a specified word distance **before** the term on the right side of this operator in order for the source document to be scored as a result
- **AFTER** – similar to **NEAR**, only the first (left-hand) term before this operator has to occur within a specified word distance **after** the term on the right side of this operator in order for the source document to be scored as a result
- **Phrases** – combined words or terms that must appear directly **adjacent** to one another and in the phrase order for the source document to be scored as a result
- **Wildcards** (stemming) – beginning characters that must match the same beginning characters in a document’s words in order for it to be scored
- **Parentheses** – nested operators that are evaluated in an inside-out, then left-to-right order of precedence.

Example uses of these operators are based on the sample tutorial problem of finding information on the peregrine falcon discussed in **Topics 5 - 12**.

The underlying premise of Boolean logic is set theory. The **AND** operator is equivalent to the set **intersection** operation; the **OR** operator is equivalent to the **union** set operation. To help explain these concepts, specific topics below use so-called Venn diagrams. Don't worry about the fancy name. The diagrams are color-coded to indicate the result of an operation. The universe of possible results is shown in yellow on these diagrams; the accepted results in blue.

One way to decide when to use the **AND** or **OR** operators is to test whether your keywords are different concepts, or a just different ways (synonyms) to say the same thing. For different concepts, use **AND**; for synonyms, use **OR**

Boolean search syntax needs to follow a precise structure. Queries constructed using Boolean syntax do not look like real sentences. The **AND** and **OR** Boolean operators, in particular, sometimes seem to mean the opposite of what they do in natural language. Searching based on simple sentences and phrases is a different construct known as **natural text searching**.

Topic 14: AND Operator

AND means "I want *only* documents that contain **both** words." **AND** logic focuses, coordinates and narrows a search. The connector **AND** narrows a search, retrieving only those records containing at least one term or phrase from each concept. The **AND** operator is a binary one; that is, it operates on the terms or phrases on both sides of it. It is the same concept as **intersection** in set theory.



Example of AND Operator

Using AltaVista document counts, the results of the query "**endangered species**" **AND** "**peregrine falcon***" is:

endangered species	143,786
peregrine falcon*	14,510
endangered species AND peregrine falcon*	3,217

Note the **AND** operator says nothing about where the terms or phrases are located in the document with respect to one another, nor whether their linkage makes sense or not. This operator only requires that the terms or phrases immediately on both sides of the **AND** must both appear in the document.

The **AND** operator can be used to chain a number of required terms or phrases together, all of which must be present in order for the outcome to be a successful result. For example, the query **London AND "Big Ben" AND "Buckingham Palace" AND Trafalgar** would only return documents that contained all four terms or phrases.

The **AND** operator is also a very useful qualifier. For example, AltaVista counts for **falcon*** total 340,707. Some of these references are to cars, others to various companies, falconry or a sundry of products using the name falcon. To zero in on the falcon bird, a search phrase of **falcon AND bird*** removes these extraneous references. The AltaVista document count now becomes 36,939.

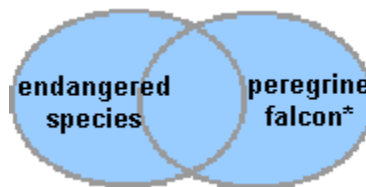
TIP

AND should be your most frequently used Boolean operator.

False “results” can be common using the **AND** operator. For example, let’s apply Jan’s query of **endangered species AND peregrine falcon*** to a large document discussing unusual birds. In one section it could discuss the 200 mph diving speed of peregrine falcons; in another the extinction of the dodo bird. A positive result would be scored for this document, even though there is no discussion about the endangered status of peregrine falcons. One of the reasons these false positives occur on the Internet is the occurrence of large Web documents that simply list links or references to other documents and contain HUGE numbers of terms. They often produce false results.

Topic 15: OR Operator

OR means “I want documents that contain *either* word; I don't care which word.” **OR** broadens a search and makes it less focused. It is equivalent to the union operator in set theory. Again, using our peregrine falcon example, the results set for this operator looks like:



Example of OR Operator

The document counts from AltaVista using this **OR** operator are:

endangered species	143,786
peregrine falcon*	14,510
endangered species OR peregrine falcon*	103,896

These results illustrate some interesting facts. First, the **OR** operator is NOT equivalent to a sum. Documents which contain both phrases still get counted as a single document. Second, we would expect at minimum the **OR** operator to result in a total number of documents no smaller than the count of documents for the largest term or phrase in the operation (*i.e.*, “**endangered species**” with 143,786 counts). Yet our result set is smaller than this. Why?

Strictly speaking the results shown should not happen. The reason they do is based on internal decisions search engines make in evaluating queries and to keep search performance snappy. See further **Topic 41** for some perplexing behavior of the Alta Vista search engine.

TIP

Use **OR** to string together synonyms; be careful about mixing it in with **AND** !.

The **OR** operator can be used to chain a number of terms or phrases together, any one of which must be present in order for the outcome to be a successful result. For example, the query **London OR “Big Ben” OR “Buckingham Palace” OR Trafalgar** would return all documents that contained one or more of these four terms or phrases. As with the **AND** operator, there is no assurance that any of these terms or phrases are logically or conceptually linked in any of the results documents.

Unless used in parenthetical clauses (most useful for synonyms) or as a fishing expedition as part of preliminaries to a search, we do not recommend the use of the **OR** operator. Overuse of the **OR** operator can cause results sets to grow too large to be useful.

Nonetheless, the **OR** operator is one of the two main operators within Boolean syntax. It should be used in a controlled way to expand your results set, most often as part of a parenthetical argument.

Part 5: Advanced Operators

There are four additional Boolean operators that provide more fine-grained control than the basic **AND** and **OR**. These operators are less frequently used and are not all supported by search services with basic Boolean capabilities.

Topic 16: NEAR Operator

Remember for the **AND** operator that the terms or phrases on both sides of the operator can appear *anywhere* in the document in order to get a successful result. One example above described how a successful result for “**peregrine falcon***” **AND** “**endangered species**” could be obtained, even though the falcon reference was to 200 mph diving speeds and the endangered species discussion was many pages later dealing with the dodo bird. The **NEAR** operator is designed specifically to avoid such false results.

The **NEAR** operator requires the two phrases or terms to be within a specified word count of one another to be counted as a successful result. Generally, most search engines that support the **NEAR** operator have a set value of a ten word maximum distance. A few [see **Topic 38**] allow you to specify a word distance of your choice if you supply an additional argument. Some engines also use **ADJ** (for adjacent) as the equivalent operator to **NEAR**.

The **NEAR** operator does not care which of the phrases or terms on either side of the argument comes first or not, just that the two phrases or terms are within the specified distance.

The **NEAR** operator is a great way to ensure that your search terms occur within the same sentence or same paragraph. It is a very useful way to remove large, comprehensive Web sites that have a reference to everything under the sun, but not specific information of use to your search.

The **NEAR** operator can have drawbacks, however. It is possible to overlook the definitive document on endangered peregrine falcons, for example, if in one section of the document it uses peregrine falcon but elsewhere when its endangered status is discussed it only uses the word peregrine. It is very difficult in all cases to foretell how document authors will use, repeat or link such terms.

Another drawback is the relatively few search services that support this operator. This problem can be overcome when using third-party search tools such as **Mata Hari** that work on the results of search engines but support this operator themselves.

But, if your terms can pass the test of confidently appearing within a sentence or so of one another, we recommend you consider the use of the **NEAR** operator.

Topic 17: BEFORE and AFTER Operators

The **BEFORE** and **AFTER** operators work in the exact same manner as the **NEAR** operator, only you can now specify which terms or phrases need to come first or second. In the case of the **BEFORE** operator, the first term or phrase **MUST** occur before the second term or phrase within the specified word distance. In the case of the **AFTER** operator, the first term or phrase **MUST** occur after the second term or phrase within the specified word distance.

TIP

Use **NEAR** as an alternative to phrases and an improvement to **AND**, **but only** when you know the concepts are closely linked.

These operators do provide even greater control to your searches. But their drawbacks are even more severe than the **NEAR** operator. First, not only must your terms appear within the word distance, but you also must get the order right. Second, to our knowledge, only two major search engines support these operators [see **Topic 38**].

For these reasons we've included these operators here for the sake of completeness, but we do not recommend that you seriously consider using them. If you become an Internet 'power searcher' and you decide you disagree with this recommendation, then your skills have surpassed the purpose of this tutorial anyway.

Topic 18: AND NOT Operator

AND NOT removes any documents that contain that term or phrase. **AND NOT** is a unary operator; that is, it only works on the term or phrase that immediately follows the operator. It does not evaluate terms or phrases on both sides of the operator.

Most of the major search services support the **AND NOT** operator. It is sometimes called **BUT NOT** or **NOT** and sometimes denoted by placing a minus sign (-) before the term or phrase to be removed. NOTE: Technically **NOT** is the unary operator. For example,

NOT falcon

would exclude all documents that use the word **falcon**. The problem arises in the middle of a query. While some search engines allow **NOT** by itself, such as:

falcon NOT car

which would return documents using the word **falcon** but not **car**, the statement is technically ambiguous as to how to treat **falcon**. As a result, most engines require matching **NOT** in the middle of a query with **AND** or **OR** (**OR NOT** is rarely used). This removes the ambiguity and is the form we've adopted herein for use within the middle of a query.]



Example of AND NOT Operator

Again using AltaVista document counts, here are the results for this operator:

endangered species	143,786
peregrine falcon*	14,510
endangered species AND NOT peregrine falcon*	98,307

As discussed for other operators, search services indexing or retrieval accuracy or is not 100%.

AND NOT is a very powerful command that should be used with care. **AND NOT** works to narrow a search, subtracting all citations that contain the specified term or phrase.

AND NOT is completely non-discriminatory; it only takes one instance of a word or phrase to eliminate a document from your results set. As one source describes it, think of **AND NOT** logic sort of like peeling a potato [30]. A peeled potato is **potato AND NOT peel**. There's only one trouble: some of the good part of the potato goes with the peel. So, use the **AND NOT** operator with as much care as you would a paring knife, and only when you're absolutely sure you want to exclude a term or phrase from your results.

TIP

AND NOT is a powerful operator, use with care! A single instance will cause a document to be excluded.

Generally, we do not recommend using **AND NOT** in the beginning iterations of a search. See what results are obtained in the early steps before applying this operator, if at all. Then, apply it incrementally to make sure you're not stripping away too much of the fruit.

A good example of where this might apply is with the **falcon*** search noted for the **AND** operator. The term **falcon*** returns references to cars, products, companies and place names, in addition to birds.

Successively applying **AND NOT** to **car***, **product*** and **compan*** is another approximation to the search **bird* AND falcon***. On the other hand, using **AND NOT** with **place*** could be going too far by eliminating references to falcon bird sightings that occur in various places.

Though in this example we have a good **AND** qualifier in **bird*** for our interest in peregrine falcons, a suitably encompassing word such as **bird*** may not apply to other search topics. In these cases, **AND NOT**, judiciously applied, can be an alternate way of getting to the same end.

Part 6: Advanced Construction

This part builds on the Boolean operators and basic search concepts previously discussed to show how they can be combined into effective, complete queries. Much of the discussion concerns how to construct proper syntax. This part ends with a reprise of our sample search problem for Jan's mystery bird [see **Topic 5**]. The guidance below, however, should be generally applicable to most engines that support structured, Boolean syntax:

Standard Syntax	Meaning	Alternative Syntax	If Not Supported
AND	both required	+	ignored
OR	either required	blank	all support
AND NOT	exclude following	-, BUT NOT, NOT	ignored
NEAR	required within set word distance	ADJ	ignored
BEFORE	first required before within distance		ignored
AFTER	first required after within distance		ignored
()			ignored
" "	treat as phrase	checkbox option	
*	stem word	checkbox option	ignored

Topic 19: Use of Parentheses

Search services that support structured (Boolean) syntax do not always read from left to right like we do. Instead, they read "inside-out", in order of the nested levels of arguments set off by parentheses. Each bounded argument set off by parentheses is called a **Boolean expression**. (The entire query is also assumed to have parentheses around it, whether you put them in or not.) This is the same concept drummed home in high school math in how to evaluate an algebraic expression.

Learning how to construct this Boolean syntax structure is easy. You only need to remember four things:

1. You define a Boolean expression through use of an open parenthesis ['('] to begin it, and a closed parenthesis [')'] to end it
2. Make sure the first search concept you want evaluated is at the inner-most level of your Boolean expressions; followed by subsequent expressions in your desired order
3. Make sure you have a balanced (equal) number of open and close parentheses in your entire query
4. Expressions at the same "level" are read in order, from left to right.

It is really worth your time to master these simple rules. It adds immensely to your control over your queries and their ability to return the results you desire.

Though some search services support quite a few layers of nested Boolean expressions, in practice the amount of nesting you need or is even desirable is quite low, likely no more than three at most. To show a three-level example, consider the following dummy query:

THIRD expression (SECOND expression (FIRST expression evaluated) evaluated) evaluated

Note, you do not need to put parentheses around the entire query; the outermost layer is evaluated last in any case. But, even when you think the computer is going to do what you want, it is always safer to use parentheses if there is even a chance of confusion. Parentheses will also help you read your own searches.

In the absence of any nesting, or with expressions at equivalent levels, the order of query interpretation is from left to right. For example:

FIRST expression AND SECOND AND THIRD AND FOURTH

or,

(FIRST main subject) AND THIRD expression AND (SECOND expression)

TIP

Don't heavily "nest" your parentheses. Remember, *keep it simple*

AS A GENERAL RULE, YOU SHOULD ALWAYS PLACE YOUR MAIN SUBJECT TO BE EVALUATED FIRST. This is because many search engines determine the rank order of document results by relevance, with first query terms to be evaluated ranked higher. This rule can be a bit tricky until you get used to it. For example, taking the last query example above, but forgetting the initial set of parentheses shown, produces the following:

SECOND main subject AND THIRD expression AND (FIRST expression)

Using the form above, if you placed your main query subject first in your query expecting it to be evaluated first, you would get the unintended consequence of having it evaluated second.

TIP

Don't assume an evaluation order. Specify the order you want by using parentheses.

Finally, Boolean operator precedence is enforced by most search engines with **AND** and **AND NOT** being evaluated before **OR**. If you have doubts of operator precedence, consult the help system for the search engine being used. Our recommendation: eliminate ambiguity as to how a given engine treats operator precedence by explicitly putting your expressions into parentheses in the evaluation order you desire.

The **OR** operator should generally be used solely within nested expressions, and then mostly to capture synonyms.

For example, you may recall from our sample problem of Jan's mystery bird [Topic 5] that Jan wanted the concept of having seen the bird in the city as part of the query. Also recall there is a problem with picking up too many unwanted words when city is truncated as **cit***. A good way to handle this problem is with a nested Boolean expression using **OR**. Thus, to capture both the singular and plural forms of city, Jan would write:

(city OR cities)

This expression now covers the singular and plural without inadvertently adding undesired words (such as 'citizen' or 'citrus') to the query term list.

Whenever you mix Boolean operators in a query you should always use parentheses to force the evaluation order you want. This helps avoid unintended consequences. For example, the following query without parentheses.

hawks AND eagles OR falcons AND owls OR vultures

Is actually evaluated as:

(hawks AND eagles) OR (falcons AND owls) OR vultures

The result of this expression is not very useful. The expression does not require any one term. You could end up with pages containing only vultures or only owls and falcons or only hawks and eagles. This is most likely not the way you intended it.

Lastly, there are times when parentheses are not needed. This is when all operators are either **AND** or **OR** in the query. For example,

hawks AND eagles AND falcons AND owls AND vultures

or,

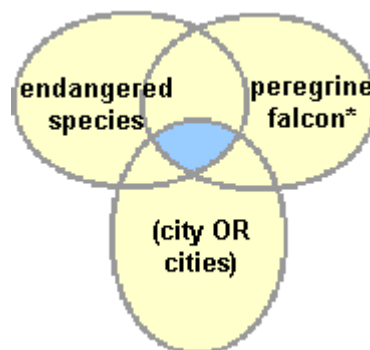
hawks OR eagles OR falcons OR owls OR vultures

The former requires all five types of bird to be included in a successful document; the latter only one. Additional examples of possible pitfall query syntax is shown in **Topic 29**.

Topic 20: Combining Concepts for Power Searching

A good rule of thumb when searching for relatively hard-to-find information on the Internet is to juxtapose three “concepts” in your query (we’ve also used the term **Boolean expression** to represent a “concept”). The first concept should be your subject, defined at the proper level [**Topic 10**], with synonyms or phrases as appropriate to provide adequate yet accurate subject coverage. The other two concepts should correspond to two of the when, where, how and why concepts discussed in **Topic 6**.

Each of these concepts should be provided as a Boolean expression with the **AND** operator connecting all three. In the case of Jan’s mystery bird example, the resulting query can be represented as:



Combine Concepts in Query

Note how this acts to restrict your final results space. Posing this query to AltaVista in the form:

("peregrine falcon*") AND ("endangered species") AND (city or cities)

produces a results set of 1,721 documents. This number may sound like a lot, but remember we began with millions, and as Jan discovers, the first twenty of which (at least) directly respond to the desired results [31]. The actual results from this search are discussed in **Topic 23**.

You should generally not need to exceed three concepts in a successfully constructed query; four is unusual. If you find you can't narrow them to two or three, double check to be sure all the concepts are necessary and all are at the right level.

TIP

Try to link three concepts together in your queries, joining with the **AND** operator.

Topic 21: Punctuation and Capitalization

Not all search engines handle punctuation equivalently. When in doubt, you should consult the help file of the search engine you are using.

Most search engines are insensitive to whether you use upper, lower or mixed case in your queries. If you use lower case, most engines will match on both upper and lower case. For general searches, it is the safest form to use. Where the engine does support upper or mixed case, if you use upper case characters the engine assumes you want an exact match. Most engines also do not care if you use upper or lower case for Boolean operators.

For the few engines that do support capitalization, you can use this fact to advantage in finding proper names or place names. See **Topic 38** for the capitalization features of major services.

Topic 22: Multiple Queries and Query Refinements

Strictly speaking, only one current Internet search tool supports multiple, simultaneous queries [see **Section 2**]. However, a number of the search services support being able to pose additional queries to a previous results set [see **Topic 38**].

These can be very valuable techniques to you as a searcher. It enables you first to cast a fairly broad query, and then successively hone in on desired results. With the search services, you can also use your browser's back arrow to try a search, evaluate, and, if you don't like the results, to back up and start over again.

As you first begin trying more advanced query techniques, we highly recommend that you start with those services that support query refinement. It gives you a way to test out ideas and put into action some of the concepts discussed here.

Topic 23: Sample Information Problem Revisited

In **Topic 5**, we met Jan, who encountered a mystery hunting bird. Through successive refinement of the subject, Boolean expressions and query syntax, Jan found a listing of 1,721 Web documents, the most highly ranked of which met the desired results [**Topic 20**].

Here's what Jan discovered [32,33,34,35]:

- The mystery bird was a male, peregrine falcon. Nearly lost to extinction, in at least the Eastern U.S., the bird was making a stunning comeback through a combination of breeding-and-release programs and a cleaner environment free of DDT

- Peregrine falcons had found a natural home in downtown cities, where the building ledges gave them protection as their natural cliff habitats had, and where there were plenty of delectable pigeons to feed on
- Breeding pairs of peregrine falcons were now found in such urban areas as Cincinnati, Dayton, Columbus, New York City, Cleveland, Toledo, Chicago, Milwaukee, Toronto, Montreal, Philadelphia, Wilmington, Baltimore, Washington, DC, Salt Lake City and Pittsburgh
- From a base of zero in the 1970s, there are more than 1,000 breeding pairs now known East of the Rocky Mountains
- Live-cams showing peregrine falcon nests on building ledges are now being beamed 24 hrs per day over the Internet from Toronto, Montreal, Columbus and Pittsburgh
- Jan's sighting in Minneapolis was the first recorded in that city
- Tremendous additional information was gained about great viewing sites for peregrine falcons at nature preserves and general information about the species.

Jan came to understand that the recovery of peregrine falcons was one of the great environmental success stories of the past two decades. Jan is presently setting up Minneapolis' own live-cam to monitor the new breeding pair in that city. Jan is also a local celebrity and resident authority on peregrine falcons.

Part 7: Pitfalls to Avoid

This part describes many of the common errors made by Internet searchers. Some are within the control of you, the searcher. Others are due to the rapid growth of the Internet and the inherent limitations to search services on the Internet.

Topic 24: Avoid Misspellings

You know, it's so obvious that it is most often not mentioned: Searchers on the Internet are atrocious spellers. See for yourself. The two links below enable you to monitor in real time the queries being issued on the Internet. Observe for yourself bad spelling, not to mention bad query construction. (WARNING: the links you are about to see may contain graphical sexual content; another common feature of searching on the Internet.):

http://webcrawler.com/voyeur_wc [36]
http://www.metaspay.com/spy/filtered_a.html

It is not the purpose of this tutorial to rap people on the knuckles if they misspell words. But, in your query and searching, if you misspell your keywords, you are immediately penalized. Let's do a little exercise to test this with the terms **query** and **searching** used in the previous sentence. Again, our document counts are based on AltaVista:

query	712,851
queryy	2,156
qerry	37
kwerrie	2
searching	1,507,951
serching	1,541
searchng	192
seerching	5
sherching	8

Clearly, Web developers also misspell words on their own documents (don't we all!). (Note: some of the misspelled instances above refer to this search tutorial on-line.)

Computers and indexing algorithms are inherently stupid. If the Web developer misspells a word, it is entered as such on the database. If the searcher issues a misspelled query term, that is what is searched for. So, recognize that computers are stupid and guard against these mistakes yourself. Sloppy entry of query terms will cost you time and cause you frustration.

Topic 25: Redundant Terms

Think of constructing a query as being in a card game. You have only so many cards (terms) to play to get a winning hand, or successful results from your query. Using redundant terms "burns" one of your cards, and diminishes your prospects for success.

TIP

Limit your keywords to six to eight. Check to make sure you're not duplicating "levels" in your terms.

Redundant terms mostly arise from combining terms from multiple "levels" dealing with the same concept [see **Topic 10**]. For example, in Jan's search case, the subject of the query became **peregrine falcon***. Were Jan to also add **bird*** to the query it would repeat information – at the wrong level to boot.

You can generally spot redundant terms by asking the question, "Is this term already covered by another term?" If the answer is yes, pick the term at the appropriate level and discard the other one.

Topic 26: Ignored Terms and Special Characters

There is an emerging class of words that are becoming like stoplist terms – often ignored by the search engines because of their ubiquity on the Internet. Examples include: computer, Internet, Web, sex and software. These words, and others like them, are not always ignored. It appears that at high-demand search times that some of the engines choose to ignore processing them.

Should you experience such behavior, one solution, if you indeed need to use such ignored terms in your query, is to make sure that you place these words in quotes or make them part of a phrase. The ignored behavior appears to be limited to use of such terms as individual words in queries, and then only at some times of the day.

You should also be aware that none of the search services covered in this tutorial handle special characters such as: ~ ! @ # \$ % ^ & () = | { } ' " < > ? / , . _ , or non-English language characters such as the cedilla (ç) or umlaut (Ö) (or many others). Depending on the engine, the special characters are generally ignored, but are sometimes treated as a space. Generally, too, the characters of - and + have reserved meanings for **NOT** and **AND**, respectively.

Topic 27: Alternate Spellings

English has become the standard language for Internet communications. However, some of the largest user domains on the Internet come from a background of traditional public school (U.K.) English. There are perhaps 50 countries around the world whose English is traditional, and not based on usage and spelling in the United States.

As a searcher, you should be aware that many common terms – colour/color, organise/organize, behaviour/behavior – may differ in spelling between these two forms. If you suspect that a keyword in your queries may have alternate spellings, we advise you to treat these alternates in the same way you handle synonyms: list both forms in an **OR** Boolean expression.

Topic 28: Too Many Terms, Synonyms

We have recommended throughout this tutorial two overall guidelines for the size of your queries:

- Limit the key concepts (e.g., Boolean expressions) to three or fewer; under rare occasions this guideline can increase to four
- Keep the actual terms in your queries to no more than six to eight.

These guidelines are not just a goad to refine query construction, content and syntax. They are also driven by experience that indicates that at high numbers of term counts search engine behavior can become erratic and unpredictable.

It is difficult to judge the latter point, since each search service closely guards how it indexes, retrieves and scores queries. Attraction of eyeballs has become a highly-competitive factor of the Internet; many are vying to gain advantage in where they are listed on search engine results; and there are real technical demands to serve all search requesters in real time at peak demand periods.

The fact that search service rules are today opaque is unlikely to change any time soon. As users, we are left with observing engine behavior, reading the public help documents, and gleaning insights from others on the Web who have been focused on similar questions. This is not really an attractive state of affairs. Absent definitive and public disclosure by the search services of how they handle these matters, room for misinterpretation and misunderstanding looms large.

Such disclosure is unlikely to happen in the foreseeable future. Searching has become big business on the Web, and as the starting point for most users and most searches, will likely remain so. In this competitive, market share-and revenue-driven environment, the incentives for major search services to disclose more than they already are doing is minimal, and will possibly even diminish.

Topic 29: Improper Boolean or Complicated Construction

Part 6 describes advanced construction of Boolean queries. This topic elaborates on four pitfalls that you may encounter:

- Excessive nesting or terms, which search services may not process in all instances and which may not achieve what you want the query to do
- Unintended results from combining the **AND** and **OR** operators
- Improper (and unintended) use of the **AND NOT** operator
- Unbalanced parentheses.

Let's reprise a complicated form of our standard mystery bird query, only this time focusing on citations in those cities which are known to have Internet live camera shots of falcon breeding pairs. The number shown after the query is the number of documents identified by AltaVista. Let's say our first query is as follows:

("peregrine falcon*") AND ("endangered species" OR extinct*) AND ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [477 counts]

Whew! That's a complicated query. Let's also say that we are ambivalent about whether the endangered species status or the listing of cities both need to be in our results set. We could thus change the query as follows:

("peregrine falcon*") AND ("endangered species" OR extinct*) OR ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [41,191 counts]

Whoa! Why did the results set zoom to more than 41,000? First, because of the precedence order of evaluating nesting, the query above is really being evaluated as follows:

((("peregrine falcon*") AND ("endangered species" OR extinct*)) OR ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [41,191 counts]

This really amounts to both sides of our query being evaluated independently, and then combined:

**("peregrine falcon*") AND ("endangered species" OR extinct*) [3,794 counts]
((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [44,074 counts]**

Clearly, this is not what we intended. We can try to fix the evaluation order by changing the nesting order by now bracketing around the two concepts for which we didn't have a preference, endangered species status or presence in one of the named cities:

("peregrine falcon*") AND (("endangered species" OR extinct*) OR ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding)) [6,359 counts]

The main point of these examples is that combining **AND** and **OR** operators in long, complicated queries can lead to undesirable results and some difficulty in figuring out what is being evaluated first.

A more important point is to slim down your query terms and make your construction simpler. Taking the first query above, let's do that. We first get rid of **extinct***; we think it is covered pretty well by "**endangered species**". We then decide to eliminate the **breeding** term because we deem it to have much lower informational value than the other query concepts. Finally, we will put all of the concepts at the same evaluation level by linking them with **AND** operators and putting each within its own parenthetical listing. Our streamlined query now becomes:

("peregrine falcon*") AND ("endangered species") AND (Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) [825 counts]

Now, our results set has become acceptably low, and the query is easier to read and understand.

As one final refinement, we could choose to use the **NEAR** operator to make sure these key query concepts are indeed related in our source documents. Here are the results of trying that:

("peregrine falcon*") NEAR ("endangered species") AND (Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) [164 counts]

("peregrine falcon*") AND ("endangered species") NEAR (Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) [87 counts]

("peregrine falcon*") NEAR ("endangered species") NEAR (Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) [4 counts]

We see that we have indeed narrowed our results, in one case to nearly zero! [One of the four is this on-line tutorial; the remaining three are the same story shown in different locations.] That is because the concepts of falcon and endangered status are more closely related than the cities in which the birds might be found. We could choose to go with the first **NEAR** query, but really the results set from the simple **AND** construction still had low enough counts and the first listings met our desired results. Our conclusion: Use of the **NEAR** operator may be just a bit too fancy in this case.

So, despite the fact that Boolean queries can become quite complicated with different operators that you can use, the better rule is **Keep it Simple**. As long as you try to combine two or three query concepts at the same level linked via the **AND** operator, you should be just fine in getting meaningful results.

TIP

Avoid complicated nesting with too many parentheses; they can sometimes give results you did not intend.

A different kind of unintended consequence can arise from the use of the **AND NOT** operator. To illustrate this, let's take this query as our starting example:

hawk* AND eagle* AND falcon* AND raptor* [4,929 counts]

We see, however, that we violated one of the rules of mixing redundant terms at different levels. Hawks, eagles and falcons are all raptors. So to test what happens when we pull the **raptor*** term out, we try the **AND NOT** operator:

hawk* AND eagle* AND falcon* AND NOT raptor* [16,922 counts]

But, wait, why didn't our document count go down? It went way up! Didn't we remove a term from our query?

This is a good illustration of a common misperception about operators and the universe upon which they operate. In fact, based on the left-to-right evaluation rule (absent nesting), the universe upon which the **AND NOT** operator was working in this query is:

hawk* AND eagle* AND falcon* [22,873 counts]

Thus, some 5,900 of these source documents do not contain the words raptor or raptors.

Lastly, unbalanced parentheses can be a common mistake in query formulation. All of the leading search engines that support Boolean queries test for this and give you a bad syntax error should you forget an open or close parenthesis. However, if you keep your nesting simple as we recommend, you should minimize occurrences of this mistake.

org – non-profit organizations

These major domains are now being expanded to include:

arts – entities emphasizing cultural and entertainment activities

firm – businesses, or firms

info – information service providers

nom – for those wishing individual or personal nomenclature

rec – emphasizing recreation/entertainment activities

store – businesses offering goods for purchase

web – entities emphasizing activities related to the World Wide Web

- 4** Country domains (also known as geographical or ISO3166 domains) are the top-level domains maintained by every country and territory in the world. These domains are organized by locality, and are useful to organizations and business that wish to operate overseas OR want to protect their company or brand identity. Like generic domains, country domains are accessible to any user of the Internet. Country domains have two-letter designators, e.g. **.fr** for France, **.uk** for the United Kingdom, **.au** for Australia, **.us** for the United States (not generally used), etc. There are over 230 top-level geographical domains, of which about 190 currently accept domain registrations. You may obtain a complete listing of these abbreviations from [37]
- 5** All information prior to this point identifies how to get to the given physical location where the Web documents reside. Field **5** represents the path and specific Web pages at that location internal to that site. This field can contain useful information, such as **howitworks**, but is sometimes quite cryptic and often can be quite long. Note that absent a designation in this field you are generally directed to the home, index or main page of the given site. Also note that some engines that support site filtering do not allow you to search in this field.

Generally, fields **2**, **3** and **4** are the most useful to use when restricting sites. **5** is subject to much variation and is not always supported. We recommend that you only use it when you have advance information or specification of the given document(s) for which you are looking.

When using site filters, you need to be careful that you don't enter too broad a specification. For example, using '**com**' as a site filter specification would result in including sites with the '**.com**' domain as well as sites such as **commonplace.edu**, **commercial.net** or **markettips.org/commercialization.html**. Attentive use of periods ('.') and slashes ('/') can help narrow your restrictions for those search engines that support the site filtering feature.

Topic 31: Size Filters

Presently, no major search services are known to filter documents by size. There are third-party products, however, that can do so [see **Section 2**].

Topic 32: Date Filters

Date filters can be especially useful when doing research on time-sensitive information. Depending on the engines that support this feature, you can restrict retrievals to documents modified since a certain date or within a range of dates.

Date filtering provides a good argument for keeping a record of your exact query and its date for very important searches. Then, should you want to see what results have been updated or added to the Internet since your last search, you can simply re-submit the initial query and select the appropriate date restriction.

There is a caveat to date filters, however. The dates shown used by the engines are (generally) the date the page was indexed, not created. (Date created fields are available to Web developers, but not all use them. Also, not all engines read this field, anyway.) Some search engines are running days to weeks behind in indexing pages. To prevent possible gaps in your date searches, you may want to consider moving the start date back by three weeks or so from the absolute date you want to filter.

Topic 33: Specialty Filters and Search Options

In the competitive race to provide more features, many search engines are providing specialty filters and search options. For a listing of these features by major services, see **Topic 38**; for a listing of our specialty options see **Section 2**. Here, however, we describe what options are available. Please note these options are supported by only a limited number of services. Also note that these features may be described slightly differently by different services; consult their specific help files.

- **People's Names** – only provided by Yahoo as a specific option (use of Four11); can be accomplished with other services that support mixed capitalization. Also, though not a specific option, AltaVista will search for any name entered in place of a URL. In addition, there are special engines on the Internet specifically for finding people, such as Switchboard. See the section on specialty engines, **Topic 39**.
- **Depth** – provides the ability to retrieve additional pages from a given site; 'depth' represents the nested levels to retrieve
- **Anchor** – finds pages that contain the specified word or phrase as contained in a link. For example, 'Click here to download' could be text associated with a link. If specified with this option, documents that contain this phrase would be scored as a result
- **Applet** – identifies documents with Java applets corresponding to the name provided
- **Domain** – finds documents restricted to the country or generic domain specified
- **Host**– finds documents on the specific computer specified for 'host'
- **Image** – identifies documents with images (graphics) corresponding to the filename specified
- **Link** – finds documents with links to the URL specified as the argument
- **Title** – identifies documents that contain the word or phrase specified in their titles
- **URL** – finds documents whose URLs match the word or phrase specified
- **File/Media Types** – identifies documents which contain the file or media type specified; useful, for example, in finding documents with audio or video
- **Business Document Types** – restricts retrieval of documents to those matching the document types of press releases, product reviews or job listings.

Part 9: Understand Your Engines

Effective searching requires understanding how best to utilize the features of your search services. But, Internet searching is a highly-competitive, dynamic area. New search engines are cropping up continually, others are folding or being acquired, and feature sets change almost daily in order to keep pace.

This part is a comprehensive overview of the state of search services on the Internet as of Spring 1999, updated from the first version a year previous. When published, it was already possibly dated. The authors therefore take no responsibility for the accuracy or completeness of the information herein. Hey, we're just doing the best job we can. But we do make mistakes

Topic 34: Some Caveats: The Dynamic Search Business

Searching on the Internet extends from the quick question, for which a lot of information is known to exist, to serious and purposeful research on esoteric topics. Casual users simply surfing or posing the quick question likely do not need an understanding of query syntax and construction nor search engine features and operation. This tutorial is definitely geared to those who want to spend the time to get more enjoyment and results from serious searches.

As of early 1997, some 600 search services were known to exist on the Internet. Recent citations have noted as many as 1,800 and one Web site, www.beaucoup.com, includes references to more than 1,400 [15]; our own estimate is 11,000 on the low side and likely many times more that amount [16,38]. Recent major engines including Galaxy, Magellan and WebCrawler, have gone out of business or been acquired by competitors. Major partnerships have been formed and some apparently separate engines, such as AOL NetFind, are branded implementations of other services (in this case, Excite). Entirely new services, such as Direct Hit, have also begun in the past year and achieved early prominence. The industry is clearly in flux.

This dynamism makes it impossible to keep absolutely current on the state of Internet search services. The information presented herein is a best-faith effort to provide an accurate snapshot of its state as of Spring 1999. The authors or VisualMetrics Corporation make no representations as to the accuracy or completeness of the information presented.

The authors do not intend endorsement by virtue of whether a search service is listed herein. The decision as to which engines to include as major ones comes from one of the more authoritative Web sites on search engines, www.searchenginewatch.com [39]. The engines included in that service were used to define which search services in this tutorial were classified as "major."

Additional updates of this tutorial are likely. We welcome you to identify errors or provide us additional, useful information. These updates and corrections will be reflected in future versions.

Topic 35: Duplication, Coverage and Responsiveness

Best estimates of the number of publicly-available documents on the Internet are 800 million [11]. The fact that the numbers available are simply estimates and differ greatly is an indication of how little is truly known about the size of the Internet and the completeness with which search services cover it.

The same *Science* article by Steve Lawrence and Lee Giles of the NEC Research Institute from which the larger estimate was drawn also is the reference for much of the information on search engine duplication and coverage.

Lawrence and Giles (L&G) analyzed coverage of 575 mostly scientific or technical queries posed by researchers at their institute in December 1997. Krishna Bharat and Andrei Broder (B&B) of the Digital Systems Research Center recently conducted a similar study with nearly equivalent methodology [21]. Here are their findings for coverage of the Internet by six of the major services, all of which do full-text indexing (in other words, a directory service like Yahoo was not included in their analysis):

Search Engine	% Combined Coverage		% Coverage of Total Web	
	B & B	L & G	B & B	L & G
HotBot	48%	58%	42%	34%
AltaVista	62%	47%	50%	28%
Northern Light	---	33%	---	20%
Excite	20%	23%	17%	14%
Infoseek	17%	17%	15%	10%
Lycos	---	4%	---	3%

The combined coverage figure refers to what percentage of searches were successfully returned by that engine. Because none of the engines comprehensively covered the Internet, the percent coverage of the total Web represents the authors' estimate of gaps and overlap.

TIP

You should always use multiple search services for your important queries.

One of the main conclusions of both studies is that no search engine indexes more than about one-third to one-half of the publicly-available documents on the Internet. By applying these figures to the known documents these services have indexed as of late 1997, the authors were able to come up with their estimates of 200 million to 320 million total documents on the Web. Even still, the authors believed their size estimate to be a lower bound, expecting the "true size of the Web to be much larger" than their methodology suggests [12].

Three additional conclusions from the L&G study deserve mention. First, submitting queries to multiple search engines greatly increases the amount of results obtainable. They estimated that combining queries to the six engines studied increased the likelihood of finding results by a factor of 3.5.

Second, they found surprisingly little duplication between the engines. With the largest two engines, HotBot and AltaVista, the number of duplicates was only 18% [40].

And, third, they found that "dead links," that is pages listed on the search engines but no longer in existence, ranged from 1.6% to 5.3%. Though not universally true, there tended to be a correlation of engines that indexed more documents, such as HotBot, with a higher incidence of dead links. This result should not be surprising, in that significant effort must be expended to maintain a larger database, and the room for error and untimeliness is higher.

Of course, size is not all that matters on the Internet. Many search engines justifiably make the argument that better and more accurate beats bigger. As a searcher, your interests should be on the quality of results. What perhaps is most disturbing, then, is that many quality results may not be indexed by the major engines in use. This possible lack of coverage is likely not a concern if the search topic is one of a broad, widespread nature. But, if looking for technical information or that which is inherently not part of the mainstream, these results are not comforting.

There is perhaps a serious methodological flaw at the heart of the *Science* article analysis. Recall two things: first, the subject of the analysis was technical queries; and, second, the nature of how items get listed initially by search engines.

Full-text search engines get their listings in one of two ways. Either a site developer submits one or more Web addresses asking the engine to index it (in which case it is then scheduled for a later full-site indexing). Or, the 'spiders' used by the engines to find new content on the Web encounter the site and then include it. Spiders depend on linkages from prior sites to identify new ones. Information tucked away in the nooks and crannies of the Internet – in other words, some of the most specific information you may be trying to obtain – may have few if any links to them. Without links, or without prior notification by the developers, spiders will only chance upon new sites.

Because businesses tend to actively seek listings on search engines, it is not at all clear that the lack of coverage implied by the *Science* article would apply to this sector. By focusing on technical searches, the authors could therefore have significantly overestimated the lack of coverage on the Internet. Whether coverage is better or worse for different subject areas or for different focuses on the Web is unknown at this time.

As professional information searchers have come to well understand, individual search engines can return outstanding results that are found on no other engines [41]. For this reason, and the reason of inadequate coverage by those engines, you should always submit your important queries to multiple search engines.

Topic 36: Boolean or Not?

For serious searching, perhaps the most important first choice facing you is choice of search engines. Which search engines better cover the topics you are interested in? Which support the search features that will enable you to find what you want?

Not all searches are created equal. The increasing ability of some search engines to take your requests in context, and then enable you to narrow results based on your first attempt, is a promising development. Certainly being able to type in a few words and then begin receiving documents of value bodes well for common-topic searches. We ourselves use this approach when quick searches are needed.

We doubt, however, the ability of search engines in the near term to improve on this process for complicated searches or for hard-to-find information. Not only is coverage of such topics weak for a given engine, but the ability to anticipate refinements is weakened by the need to categorize information into levels insufficiently specific to the difficult query.

Thus, for difficult search topics, we still must recommend the use of search engines with full Boolean support. Only you know what information you are seeking (even though it may be ill-defined or abstract). With full Boolean searching, you have complete control to find what you seek.

This recommendation, however, exacerbates the lack of coverage of any given search engine. By definition, hard-to-find information is not well-indexed, meaning you will likely need to use more than one search engine to get the robust results you desire.

TIP

Use search engines with full-text indexing and Boolean support for your most demanding queries.

Topic 37: A Comparison of 100 Search Services

A listing comparing major features of 100 of the largest search services on the Web is shown below. For a larger listing of about 2,500 to 3,000 search services, see [16].

Search Service	URL Address	Boolean Operators	Results/ Page	Multiple Pages ?	Max. Listings
@BRINT - Business Research	www.brint.com	AND,OR,(,),"		No	
AlbanyNet	www.albany.net	---	10	Yes	
AltaVista UseNet	www.altavista.digital.com	---	10	Yes	200
AltaVista UseNet Advanced	www.altavista.digital.com	AND,OR,(,),NOT,NEAR,",*	30	Yes	
AltaVista WEB	www.altavista.digital.com	---	10	Yes	200
AltaVista WEB Advanced	www.altavista.digital.com	AND,OR,(,),NOT,NEAR,",*	10	Yes	200
American Memory Collection Search	lcweb2.loc.gov	---	20	Yes	
America's Job Bank Search Index	www.ajb.dni.us	---	200	No	200
AOL NetFind	www.aol.com	AND,OR,(,),NOT,,"	40	Yes	
AquaLink	www.aqualink.com	---			40
ArchNet Archaeology	spirit.lib.uconn.edu	AND,OR,(,),NOT,,"	200	No	200
BizWeb	www.bizweb.com	#	200	No	200
c net News	www.news.com	AND,OR,(,),NOT,,"	25	Yes	500
c net Search.Com	www.search.com	---	10	Yes	
c net Shareware.com	www.shareware.com	---	100		
CBS Sportsline	cbs.sportsline.com	---	25	Yes	
CNN Database	www.cnn.com	---	10	Yes	
CNNfn - the financial network	www.cnnfn.com	---	100		
CollegeNET	collegenet.com		200	No	200
Computer Gaming World	cgw.gamespot.com	---	50	Yes	
DejaNews	www.dejanews.com	AND,OR,(,),NOT,NEAR,",*	50	Yes	
Discovery Channel Online Search	www.discovery.com/whatsonline/search.html	---	10	Yes	
Encarta Online	find.msn.com	AND,OR,(,),NOT,NEAR,",*	50	Yes	
Environmental Organization Web Directory	www.webdirectory.com	---			
EuroFerret	www.euroferret.com/	---	10	Yes	
Excite	www.excite.com	AND,OR,(,),NOT,,"	20	Yes	
Excite News Tracker	excite.com	AND,OR,(,),NOT,,"	10	Yes	
Explorer-K-12 Math/Science	unite.ukans.edu	---			
Forum One - Online Discussion Forums	www.forumone.com	---			
Galaxy	www.einet.net	---	20	Yes	
HotBot	hotbot.com	---	100	Yes	
HotBot Advanced	www.hotbot.com	AND,OR,(,),NOT,,"	100	Yes	
IBM Infomarket-Research Reports	www.infomarket.ibm.com	---	15	Yes	
Inference FIND	www.inference.com/infind/	AND,OR,(,),NOT,,",*		No	
Infohiway	www.infohiway.com	---	30	Yes	
Infomine (Internet Enabling Tools)	lib www.ucr.edu/search/ucr_enb search.html	AND,OR,(,),"#			
Infoseek	www.infoseek.com	AND,OR,NOT,"	50	Yes	200
Internet ArtResources	artresources.com/search.htm l-ssi	AND,OR,(,),NOT,,",*		No	
Jayde Online Directory	www.jayde.com	---	50	No	50
Lawcrawler	www.lawcrawler.com	AND,OR,(,),NOT,NEAR,",*	10	Yes	
Librarians' Index to the Internet	sunsite.Berkeley.EDU	---	200	No	200
LinkMonster	www.linkmonster.com	AND,OR,(,),NOT,,",*	200	Yes	
LinkStar	www.linkstar.com	---	10	Yes	

Search Service	URL Address	Boolean Operators	Results/ Page	Multiple Pages ?	Max. Listings
Liszt, the Mailing List Directory	www.liszt.com	AND,OR,(,NOT,"			
Lycos Pro	www.lycos.com	AND,OR,(,NOT,NEAR,"	40	Yes	
Magellan	www.mckinley.com	AND,OR,(,NOT,,,"	10	Yes	
Mamma Search Engines	mamma.com	---	10	Yes	
Metacrawler	www.metacrawler.com	---	30	Yes	
Microsoft(r)	www.microsoft.com/search/default.asp	AND,OR,(,NOT,NEAR,"*	10	Yes	
Northern Light	www.northernlight.com	---	25	Yes	
OneLook Dictionaries	www.onelook.com	---		No	
Open Text	www.opentext.com	---	10	Yes	None
Orientation.com - Asia	www.orientation.com	AND,OR,(,NOT,NEAR,"*			
PC World Online	www.pcworld.com	---		No	
PlanetSearch	www.planetsearch.com	---	10	Yes	
Point's Top 5%	www.pointcom.com	---	10	Yes	
Product Review Net	www.productreviewnet.com	---		No	
PubMed - National Library of Medicine	www.ncbi.nlm.nih.gov	AND,OR,(,NOT,"*			
Reference.com (Mailing List)	www.reference.com	AND,OR,(,NOT,NEAR,"*	200	No	200
SavvySearch	www.savvysearch.com	---			
Science Fiction Review Archives	julmara.ce.chalmers.se	---	20	Yes	
searchUK	www.searchuk.com	AND,OR,(,NOT,"*			
Social Science Information Gateway	/www.sosig.ac.uk	AND,OR,NOT,*			
Spry Internet Wizard	www.sprynet.com	---		No	
Surf Point	www.surfpoint.com	---	30	Yes	
The Sporting News	www.sportingnews.com	---			
The United Nations	www.un.org	---			
Time Magazine Online	www.pathfinder.com/time	---	10	Yes	
WebCrawler	www.WebCrawler.com	AND,OR,(,NOT,,,"	25	Yes	
WebCrawler News	search.excite.com/wc	AND,OR,(,NOT,,,"	200	Yes	
What's New Too!	newtoo.manifest.com	---	25	Yes	
Windows 95 Magazine Search	www.win95mag.com	---		No	
World Wide Arts Resources	world-arts-resources.com	'#		No	
WWW Virtual Law Library	www.law.indiana.edu	---	20		
WWW Virtual Library-US	iridium.nttc.edu	---	100		
Government Information					
WWWomen	www.wwwomen.com	---			
Yahoo	search.yahoo.com	---	20	Yes	Varies
Yahooligans	www.yahooligans.com	---	25	Yes	

Topic 38: Features of the Top 10 Search Services

Based on March 1999 rankings from Media Metrix [42], and including service characterizations from Search Engine Watch [39] and Search Engine Showdown [43], the chart below compares features for the major search services on the Web. Included in this listing are search engines (SE) and directories (D). For a further description of search service types, see Topic 2; for a description of the features listed, see Topic 33. Specific notes on some of the services are appended at the end of the table.

Some of the listed features are coded. These codes represent our judgment as to the completeness of a feature compared to other services in the listing:

means the feature is deemed to be as complete as others

means the feature is not as complete as others offered or does not provide full functionality

A blank means that service does not offer the feature shown.

As before, we do not imply endorsement nor claim complete accuracy for the features presented. You are always advised to consult the online help topics for any given services. Features and sometimes syntax change on a periodic basis.

	Y A H O O	I N F O S E E K (G O)	E X C I T E	L Y C O S	A L T A V I S T A	S N A P	H O T B O T	A B O U T . C O M	L O O K S M A R T	G O T O
--	-----------------------	--	----------------------------	-----------------------	---	------------------	----------------------------	---	---	------------------

GENERAL

Ranking by User Base	1	2	3	4	5	6	7	8	9	10
Type	D	SE	SE	SE	SE	D	SE	D	D	SE
Unique Visitors/mo (Mill)	31.0	21.2	16.7	16.1	10.5	9.8	7.4	5.8	4.8	4.1
Size (Mill pages)	1.2	45	55	50	150	110	110	0.4	1.0	110
Pageviews/Visitor	176	22	64	23	21	9	13	---	10	---

STRUCTURED QUERIES

Complete Boolean			y	y	y	y	y	y		
Stemming	y	y		y	y	y	y		y/n	y
Case Sensitive		y			y	y/n	y/n			y/n
Phrases	y	y	y	y	y	y	y	y		y
AND	y	y	y	y	y	y	y	y	y	
OR	y	y	y	y	y	y	y	y		y
NOT	y	y	y	y	y	y	y	y		
NEAR				y	y			y		
BEFORE				y						
AFTER				y						
' ()			y	y	y	y	y			

INDEXING

Separate Names/Titles		y								
Metatag		y		y/n	y	y	y			y
Title	y	y	y	y	y	y	y	y		y
Body		y	y	y	y	y	y			y
ALT Tags		y		y	y					
Comments		y				y	y			y

RESULTS/RANKINGS

Relevancy	5	2	1	4	3	2	2	5	5	2
Results Clustering		y	y/n	y		y	y			y
Suggest Related Terms			y		y	y	y			y
Find Similar Pages		y	y	y						

	Y A H O O	I N F O S E E K (G O)	E X C I T E	L Y C O S	A L T A V I S T A	S N A P	H O T B O T	A B O U T . C O M	L O O K S M A R T	G O T O
User Specified					y					
FILTERS										
Date	y/n				y	y	y			y
File/Media Types					y/n	y	y			y
LANGUAGE CHOICES										
Language				y	y	y	y			
Special Characters					y					
SPECIAL SEARCH OPTIONS										
Automatic Phrase Attempt					y					
People's Names	y					y			y	
Text	y	y	y	y	y	y	y			y
Depth						y/n	y			
Anchor					y					
Applet					y			y		
Domain		y			y	y		y		
Host		y	y		y			y		
Image					y	y/n	y/n			
MP3 (music)	y			y					y/n	
Link		y			y	y	y			
Title	y	y		y	y	y	y			
URL	y	y		y	y					
NOTES	YA	IS	EX	LY	AV	SN	HB	MC	LS	GT

The key for how the services determine relevance is: 1 -- 3/4 star review; 2 -- metatag keywords; 3 -- title keywords, popularity; 4 -- none; 5 - in title, higher in category tree.

All of the directory services link to a standard search engine if their own listings do not satisfy the query. The directories and their associated engines are: Yahoo! - Inktomi; About.com - AltaVista; and LookSmart - AltaVista. There are also differences between the services that license the Inktomi engine: Snap, HotBot and GoTo. While all of these index and score pages in a similar manner, the options presented to the user can differ quite substantially, with HotBot providing the most power, GoTo the simplest interface.

Specific service notes are:

YA – people searching uses the Four11 specialty engine

IS – need to use commas to separate phrases and hyphenate words that need to appear next to one another; word within brackets are found if within 100 words of one another

EX – employs ‘morphological analysis’ to suggest refinement words for the keywords entered into a search

LY – can specify **NEAR**, **BEFORE**, **AFTER** word distances. Lycos announced in April 1999 that it was going to switch its search engine service to a directory structure using Netscape’s Open Directory format, an unprecedented move

AV – specialized functions for usenet searches; advanced searching turns off relevancy ranking (can hand enter); allows translation from different languages

SP - uses special search options through advanced search page with dropdown lists

HB – special search options through what HotBot calls meta words

MC – paid experts provide listings in about 600 topic areas

LS – uses AltaVista as source engine; presents results with category options for each entry; entries reviewed by editors

GT – simplest interface of all of the services.

The options shown in the table are often noted by different terms by the services that support them, and usually involve special syntax rules. Sometimes, too, the descriptions of how these features operate is difficult to find from the main pages of the services. Directly consult each service’s home page; and, then, try consulting advanced or power searching, the help sections or the frequently asked questions (FAQ) areas to read about the special operators and their rules.

Another useful resource, though based on relatively small sample sizes, is Greg Notess’ Search Engine Showdown [43]. This site reports dead link percentages, unique hits, overlap and some other different statistics.

Topic 39: Specialty Engines

Specialty engines have the advantage of cataloging information particular to a narrow topic area, thus potentially increasing coverage versus the general search services. This advantage, however, often comes at the cost of not providing you with the search options and flexibility that the general services provide.

One of the most complete catalogs of Internet search engines is found at www.beaucoup.com, listed below for English-oriented services, by its breakdown of about 1,400 search engines and major topic area [44]:

Category	Count	Category	Count
General	76	School Listings/Student Aids	31
Multiple/Meta	28	Educational Resources	28
Radio/TV	18	Music/Sounds	35
Publications	38	Arts/Graphics	46
Regional - Global	5	New Sites/Reviews	18
Regional - Americas	42	Science/Nature/Technology	49
Regional - Europe	124	Business Directories	51
Regional - Asia/Aust/Africa/+	73	Email/Domains/People	16
Software - Windows	30	Computer/Programming	41
Software - Other	15	Webmasters	20
Reference	64	Internet/WWW	32
Language	25	Social/Political/Environment	30
Literature	25	Politics/Government/Law	64
Health and Fitness	18	Finance/Consumer	51
Foods and Diet	29	Malls/Classifieds	14

Category	Count	Category	Count
Medicine	33	Large Corporations	67
Hobbies/Rec/Pets/Games	59	Potpourri	46
Employment Listers	45		
Corporate Employers	63	Total	1,449

Be aware some of these services catalog information that is not normally spidered or indexed by the general search services.

Not shown on the table above are searches in languages other than English. For example, major search alternatives are provided in the languages of Dutch, Spanish, German, Japanese, French, and specialty search services are offered in perhaps another 30 languages or so. For regional alternatives, Yahoo alone provides 12 different country-based search services and another 12 focused specifically on individual U.S. metropolitan areas. Similar diversification is occurring with other major search services.

These specializations are natural and reflect the huge size of the Internet (plus, obviously, the fact that English is not the only language used on the Web!). This specialization trend is likely to continue.

Other search engine directories that are comprehensive listings of other specialty engines on the Web are:

<http://www.searchpower.com>
<http://www.123go.com/drw/search/search.htm>
<http://www.dreamscape.com/frankvad/search.html>
<http://www.finderseeker.com/>

The latter in fact seems to have the largest listing; you will need to poke around some to get them. To page individually through about 26 pages (about 2,500 engines), try this URL, and then continue paging using the 'Next Page' option at the bottom of each page:

<http://www.finderseeker.com/cgi-bin/search.cgi?disp=99&sp=1&cat=&key=&country=>

Depending on the topics of your searches, you are encouraged to test out and try these listing services.

Topic 40: Some Other Services to Watch

There are a number of other search services that bear watching, either because of new and unique search technology, or because of partnering or plans that may cause them to become big players:

- **AOL Netfind** – this service, a branded implementation of Excite, deserves note because of AOL's largest installed user base on the Web. See <http://www.aol.com/netfind/>.
- **AskJeeves** – the premise of this service is to accumulate questions posed by searchers to obtain specific information. AskJeeves already has a database of many million questions; if the question that you pose is new, they add it to their listing. If information has not already been compiled by AskJeeves staff in response to previous questions, the service provides a "smart" question query to leading search engines such as AltaVista, Yahoo!, Infoseek and

Webcrawler. Though a long-time Web presence, AskJeeves has recently gotten prominence. Compaq is a major investor and Dell has a branded service called AskDudley. Recent major financing suggests this service is one to watch. See <http://askjeeves.com>.

- **BrightPlanet** – there have been Internet rumors for some time about a major, new search “filter” service that combines pre-qualified results in a directory structure (like Yahoo!) with automation for populating it. The service reportedly intends a subject structure of nearly a million topics, with only the highest quality results from the Web on it. The service has not been commercially announced, but the address of <http://brightplanet.com> has login protection, suggesting unveiling of this service in the near future.
- **DirectHit** – Direct Hit tracks which search results links users click and how long users stay at each site. The longer a user stays at a site, the higher it's ranked. As such, DirectHit is more of a “popularity” engine than a general search service. Management of DirectHit has been quoted as indicating they intend on adding a general search component to the site. See <http://www.directhit.com>.
- **FAST** – begun in Norway, this new service has the most ambitious plans reported to date in indexing the entire Web. Already the site has 80 million documents indexed, with claims to reach 200 million by Summer 1999 with an eventual goal for 1 billion documents. This may be a player to watch for those who desire complete Web indexing. Dell is a partner with FAST. Lycos has also partnered with FAST for MP3 music searches. See <http://web.fast.no/whatsnew/>.
- **Google!** – Originally developed at Stanford, Google has gone commercial. According to the company, it uses a complicated mathematical analysis, calculated on more than a billion hyperlinks on the web, to return high-quality search results. This analysis allows Google to estimate the quality, or importance, of every web page it returns in conjunction with standard text retrieval criteria. The key difference of Google's approach is the premise that the more influential sites that link to a result, the better it is. See <http://www.google.com>.
- **iAtlas** – a proclaimed “filter” service, iAtlas claims to add filter capabilities for segments/topics, dates, addresses and document sizes to the basic Inktomi (same as HotBot) search capability. Testing of the filtering capabilities proved disappointing; however, the service was just announced in the past few months and may be experiencing some start-up pains. IAtlas wants to act like Inktomi in branding for specific vertical markets. See <http://www.iatlas.com>.
- **InferenceFIND** – this metasearcher categorizes results by anticipated categories; searches are time limited. InferenceFIND metasearches WebCrawler, Yahoo, Lycos, AltaVista, InfoSeek, and Excite. Results are presented as links only with no descriptions. The clustering tends to be mostly by domain and country. See <http://www.infind.com/>.
- **Metacrawler** – part of the Go2Network, this metasearcher metasearches Lycos, Infoseek, WebCrawler, Excite, AltaVista, Thunderstone, About.com., Looksmart, and Yahoo. Results are combined with duplicates removed and presented in the standard format with description. See <http://www.metacrawler.com>.
- **MSN** – Microsoft has most recently chosen the Inktomi engine to power its MSN search site. See <http://home.microsoft.com/>.

- **Netscape** – the Netcenter portal lists a number of large search engines: Excite, Infoseek, Lycos, Snap and LookSmart. The default service is Excite. Perhaps the most interesting part of this site, however, is its directory, called Open Directory. The goal of the Open Directory Project is to produce the Internet's most comprehensive directory by relying on a vast army of knowledgeable volunteer editors. These “editors” submit links into a Yahoo! like structure; it has proven very popular and as of April/May 1999 Lycos made the unprecedented choice to abandon its normal search engine structure for this directory. For these various search services see: <http://home.netscape.com/escapes/search/ntsrchrnd-1.html>
- **Northern Light** – one of the largest document collections on the Web, can create custom folders, relevant to current search, that can be updated as more searches take place. Northern Light presently catalogs about 140 million documents, including content from proprietary information providers covering 5,400 different titles and 7 million documents that can be obtained for a per document charge. See <http://www.northernlight.com/>.
- **OneView** – this unique service is based on accumulating bookmarks from individual submitters and then placing them in a comprehensive subject structure. The concept is intriguing because bookmarks often tend to be pre-screened, quality, comprehensive sites. Complete access to the site requires a free sign-in. The site is from Germany; much of its current documentation is in German. See <http://www.oneview.com/>.

Topic 41: Some Perplexing Behaviors

Search engines may not always perform as indicated on their help pages. These differences are due to constant changes in how they handle their service, strange quirks relating to their scoring and indexing methodologies, errors made by the developers of Web pages, and decisions the service may make to speed performance at high-traffic volume periods. We can illustrate some of these quirks using our standard AltaVista search source [1].

Let's take the most simple example we've used throughout this tutorial: posing the query **query**. Depending on the time of day, during a 24 hour period AltaVista returned results ranging from 671,424 counts to 712,851 counts. This indicates that at times of high traffic limits are placed internally on the search. In the worst cases, AltaVista or other engines may even provide a message that the server is busy, and prompt you to return at a later time to obtain results. Most often when this error occurs, a quick re-issue of the query will obtain results.

Here's another example of quirky behavior when comparing results counts from two different formats (text vs. phrase) using our basic query, and using the simple and advanced search forms on AltaVista:

Form/Query	“bird**”	bird*
Simple Form	3,935,170	3,935,170
Advanced Form	1,836,261	1,836,261

The advanced form returned only half of the results of the simple (this query was posed at a low-demand time). Sometimes, too, the unquoted version on the advanced search turns up no results matching the query. Clearly, both time of day and search form can affect results.

Order on a seemingly equivalent pair of queries can also produce slightly different results. Let's compare these two queries:

“bird” AND falcon**

vs.

falcon AND "bird*"

The first query produces a results count of 9,581; the second a count of 9,228.

Of course, counts are not what you the searcher wants when you search. The actual results pages for these examples were quite similar. But it's useful to realize that how an engine operates exactly may not be clear or consistent.

As **Topics 21** and **26** note, you must also exercise care in the use of capitalization, special characters and special terms. What you are thinking you are posing as a query may not be evaluated as such by specific engines.

This is not meant to be a criticism of the search engines, or of AltaVista in particular. Many anomalies occur because of improperly formatted Web pages. And, after all, engines are indexing millions of pages in very short periods of time and need to provide snappy response in all instances. The fact they do accurately index very high percentages is remarkable. But, you, as a searcher, should be aware results are not foolproof.

Part 10: Specialty Searches

This part provides a compendium of tips for specialty search topics. Most of the individual topics below simply offer bulleted suggestions for ways to first approach these searches. Once used, you will likely find for yourself additional 'power searching' tips.

Topic 42: Product Searches

Here are some tips on finding product information:

- Use product-oriented specialty search engines (see [15,16])
- Make sure and use the actual product name in your search; use if possible a search engine that supports mixed upper and lower case
- Join appropriately stemmed search terms using the product and known company name. For example, to find information about **Mata Hari** from VisualMetrics Corporation, do not enter specific version information or full company titles; these can overly restrict your results. Instead, try **mata hari AND visualmetrics**
- Try limiting your searches with the **.com** filter; this will eliminate references from non-business sites (also note that some countries, such as Australia, United Kingdom and Canada, also use the **.com** site domain for commercial sites before the country domain)
- For product-related announcements, use the domain or url search options [see **Topic 33**].

Topic 43: Competitor Intelligence

Here are some tips on finding information about competitors:

- Job listing or employment sites can be a first indicator of whether competitors are growing or not. Try searching at the individual company's site and job listing engines and monitor trends over time
- Alternatively, but less useful, is to search resume posting services to see if many employees are bailing out. Because employees in this position are generally reluctant to announce their intentions, this tactic is generally less useful than company hiring trends. One important exception: When the company itself has internally announced a staff reduction. Sudden blips in resume postings can be a valuable early indicator
- Many of the major search engines contain sections entitled 'Company Profiles' or a similar category. Try restricting searches to these categories
- One useful way to discover partners of competitors is to use the link search option, providing the source company's name in the link text [see **Topic 33**]
- Generally, all company sites have a press releases section, where new advances or partnerships are often announced. Using the domain option [see **Topic 33**]. Conjoin your specific text query with a stemmed press* in the domain. Since often press releases are kept by news services for longer periods than on individual company sites, you may also want to make sure your query includes the company name as well
- Archive your useful queries and repeat over time. Search engines that contain a 'CGI-bin' name in the query produced can be saved and used again later
- Monitor business news sources [see **Topic 46**].

Topic 44: Market Research

Here are some tips on doing market research:

- For comparative market information, first try combining words or phrases that you know appear for the leading market-share companies or products. For example, in cereals, try conjoining “Rice Krispies” and “Captain Crunch”; for computer software information, try conjoining “IBM” with “Microsoft”. These should not be the ending of your query, but the narrowing beginning
- Consider using search engines that support the link, host or domain filters [see **Topic 33**].

Topic 45: Finding People

Here are some tips on finding people on the Internet:

- Use specialty engines
- Use search services that support people searches (Yahoo!, Snap, LookSmart)
- Use search services that support mixed upper and lower case
- Be careful, first names are often not reliable; many individuals use initials or diminutive forms for first names (“Mike” v. “Michael” v. “M.”), or may be cited by others in different ways.

Topic 46: Finding Places

Here are some tips for finding information about geographical locations on the Internet:

- Try limiting your searches by country domains
- Use regional Yahoos
- Used mixed case when searching for proper place names
- Consider using the geographic-specialty search engines
- Try using the location options in HotBot’s SuperSearch mode.

Topic 47: Finding Documents

Here are some tips for finding documents:

- There is a tremendous storehouse of information not actually catalogued by search engines because the documents are not distributed as Web pages. When looking for such information, consider using meaningful document title names plus common extensions for such files (*e.g.*, .PDF for Adobe Acrobat, .DOC for MS Word documents, etc.) in your queries
- Use the Anchor option in AltaVista, (such as **anchor:”pdf”** for Acrobat files) matched with an appropriate query dealing with your topic of interest
- Use engines like Snap or HotBot that support specifying file types in your searches.

Topic 48: Finding Recent News

Normal search engines and services are generally poor sources for recent news. Some of them, however, (Excite and HotBot as two examples) have separate search options for news postings that tend to work in the same ways and with the same features as the standard engines.

There are very useful magazine and daily periodical resources on the Web, notably including Time Warner’s Pathfinder, USA Today, Wall Street Journal, San Jose Mercury News, etc., etc. Most of the major magazines and dailies now have a Web presence.

Another useful source for news are the 20,000 newsgroups on the Web. While news per se is not covered in a traditional way, opinions and links to breaking news sources can often be found. DejaNews and AltaVista’s Usenet are good starting points.

Part 11: Solutions and the Future of Searching

This tutorial has spent considerable time on all aspects related to how to search on the Internet and the search services available. What does the current state of Internet searching suggest for the future? And, are there easier ways than needing to learn all of the nuances of various search services?

Topic 49: Ruminations on the Future of Internet Searching

We do not see the demise or “death” of search engines, as some pundits have argued. Major search engines will continue to be one of the most important first access points to the Internet. The sheer growth and chaos of the Internet assures this. But there will also be twin, divergent forces toward consolidation on the one hand and specialization on the other.

The first-generation of Internet search services are facing an untenable impasse. Sheer document volumes – approaching 1 billion in the near future – doesn’t bode well for either traditional search engines or search directories. The largest search engines today approach nearly 150 million documents. Results are overwhelming and force users to study tutorials such as this one in order to query for meaningful results. This problem will only get worse. Directories, however, with their (supposedly) higher-quality content selected by human reviewers are also falling more rapidly behind. For example, perhaps 1.5 million new documents are being added daily to the Web, surely a volume that can not be hand-screened by humans at acceptable cost. And, spot checks of Yahoo! already indicate that 20-25% of existing links are no longer in existence or out of date.

Machine indexing, classification and qualification are the only meaningful techniques whereby the quality content of the Internet can be maintained in a central location. Results presentations also need to be well-organized in a directory structure of much greater richness and depth than what current directories offer. In the words of Clifford Lynch in a recent *Scientific American* article [45]:

“One sometimes hears the Internet characterized as the world’s library for the digital age. This description does not stand up under even casual examination. The Internet — and particularly its collection of multimedia resources known as the World Wide Web — was not designed to support the organized publication and retrieval of information, as libraries are. But if the Internet is to grow and thrive as a new means of communication, something very much like traditional library services will be needed to organize, access and preserve networked information. Even then, the Net will not resemble a traditional library, because its contents are more widely dispersed than a standard collection. Consequently, the librarian’s classification and selection skills must be complemented by the computer scientists’ ability to automate the task of indexing and storing information. Only a synthesis of the differing perspectives brought forth by both professions will allow this new medium to remain viable.”

We see the continued specialization and balkanization of search engines on two levels. The first level, involving the major search services, will see consolidation and specialization at very different ends of a spectrum according to the needs of various user communities.

At the broadest consumer level, one thrust will be to provide more “intelligence” to infer simple query needs. This will enable natural language querying. Services that emphasize this strategy will attempt to become “one-stop” destinations, offering much more than searching, as a means to keep visitors longer. Virtually all of the directory services now fit in this category, with Excite and others moving in this direction as well. One might call this the McDonald’s or Pepsi approach to

establishing a broad, branded consumer service. Absolute coverage of the Web's content will be less of a driver; listing positions will be based on payments, popularity and advertising support. Query simplicity will be emphasized over user control and elaborate syntax.

At the other end of the major search service spectrum will be full-text engines, with full Boolean support and many filter options, to serve the information-intensive user community. Great emphasis will be placed on expanding the Web's coverage by these services. The revenue model here may be advertising revenues from firms targeting this demographic, or providing demonstrations of advanced technology (Digital's original motivation in establishing the AltaVista service now owned by Compaq). Attempts like Northern Light to provide "special" information on a subscription basis may work well for business users; we have doubts whether this is a sustainable revenue model for educators, students and others with strong information needs.

The second level, an opening created by today's inability for the major services to cover the Web, will likely be the fastest growing category. This level is the specialization of engines by major topic area – law, science, medicine, business, etc. – to serve those specific communities. We see much consolidation occurring in each of these niches, even while the importance of the niche expands. Adding proprietary content, and the possible aggressive entries of traditional search database firms such as Dialog and Lexis-Nexis, should keep specialty topic searches an area of ferment for some time to come.

The bridging "glue" to tie all of these disparate pieces together will be the metasearchers, either Web-based services or dedicated desktop tools. It is quite conceivable, indeed likely, that Web-based metasearchers will partner with the specialty topic services to broaden their current offerings beyond the six or so major search engines that they now cover. This would free the specialty services to focus on the topics they already understand, while giving the consumer more of a single-entry point into the Web's entire content. The role of the metasearchers will be to provide de facto standardization to Internet searches.

Part 12: Summary and Further Information

The information professionals at the University of California at Berkeley recommend a graduated approach to Web searching [46]. Here's their stepwise sequence of steps to follow, which we generally endorse for beginning searchers:

1. ANALYZE your topic to decide where to begin
2. Pick the right starting place
3. Learn as you go & VARY your approach with what you learn
4. Don't bog down in any strategy that doesn't work
5. Return to previous strategies better informed.

As you gain experience, you can begin cutting out the middle steps. By the time you're doing real heavy lifting with your queries, you really only need spend some time first getting your query right and then cutting to the bottom line with a full Boolean search using phrases and three or so concepts linked through the **AND** operator and multiple search engines.

Here's a recap of some of the recommendations made throughout this tutorial:

- Spend time BEFORE your search to **analyze what it is you're looking for**
- **Use nouns** in your queries – the who/what, when, where, how and why; avoid conjunctions, verbs, adverbs and adjectives
- Use **keywords at the right "level"** of specificity: precise, but not overly restrictive
- **Use phrases** where natural; they are your most powerful weapon
- **Use** structured ("Boolean") syntax, especially the '**AND**' operator
- Constrain your search by using **two** or **three** related, but narrowing, **concepts** in your query
- BUT, generally, keep overall query length limited to **six to eight keywords** maximum
- **Use advanced search options** and specialty features when appropriate [Topic 48]
- **Use multiple search engines for your most important queries** – research shows accuracy improves many-fold [Topic 48]
- For **difficult searches**, use only search engines that support **Boolean syntax**, or tools or metasearchers that do [Topic 48]
- For specific topic searches, consider search engines tailored to those topics
- Save time by **learning your search engines** and advanced, 'power searching' techniques [Topic 48].

Useful tips for how to govern the accuracy and scope of your searches are:

Search Action	Search Scope	Results Likelihood	Comments
Focused Keywords	narrows	higher	yes; but can be too focused
Broad Keywords	broadens	lower	low yield
Use of Synonyms	broadens	higher	watch for long query sizes
Additional Keywords	broadens	higher	good, if related well
More Query 'Concepts'	narrows	higher	should not exceed 3 to 4
Fewer Query 'Concepts'	broadens	lower	single concept or keyword MAJOR search mistake
Use of Phrases	narrows	higher	exact word order critical
Use of Wildcards	broadens	higher	recommend; watch short

Search Action	Search Scope	Results Likelihood	Comments
Multiple Queries	broadens	higher	stems
Simple Text Search	broadens	lower	useful when search uncertain quick; same as all OR operators
Structured (Boolean) Search	narrows	higher	takes time to master
AND Operator	narrows	higher	highly recommended
OR Operator	broadens	lower	only for synonyms; be careful when using with AND
NEAR Operator	narrows	higher	excellent alternative to phrases
AND NOT Operator	narrows	higher	useful in limited circumstances
Use of Parentheses	depends	depends	great when done well; tricky to do; keep simple
Redundant Keywords	broadens	lower	use care and remove
Alternate Spellings	broadens	higher	not common; be aware
Filters	narrows	depends	can be useful or too narrow

Fondren Library at Rice University has also published useful tips on Internet search strategies [47]. For advanced topics, and a resource that is increasingly focusing on Web-related topics, you may want to consult *Searcher: The Magazine for Database Professionals* [48].

Finally, for issues relating to search engines, their capabilities, market share and how they work, two excellent resources are VirtualPromote [19] and Search Engine Watch [39]. You may also enjoy checking out Steve Steinberg's fascinating article for Wired on the nature of search services and the general topic of why knowledge organization matters [49].

Section 2: Using a Powerful Desktop Resource – Mata Hari®

To overcome search frustrations, we at VisualMetrics developed a unique search agent that combines the best features of metasearchers, document indexers, filters and databases in a single desktop package. **Mata Hari™** is the name of our tool. The remaining portions of this section describe the product's features, speed and automation of many of the points raised in this tutorial.

Topic 50: Mata Hari Product Features

Mata Hari is a search agent that provides one access point to Internet search services with complete query power and the ability to speak all dialects of "search engine." It is an advanced desktop search agent that uniquely combines these features:

- **metasearching** – Mata Hari is a universal translator for all dialects of "search engine". It can access nearly 600 search engines which are categorized into 60 topic areas for easy selection. Mata Hari can simultaneously search up to 150 engines at one time using a single and simple query format. Mata Hari ends the search service Tower of Babel
- **easy to learn and use** – A simple slider can be used to specify search settings. Using the slider automatically adjusts many search parameters and requires no learning curve for new users. Individual control of each parameter is available for advanced searchers that want to optimize large, complex searches
- **full-text-indexing** – after removing all duplicates, Web documents that meet the query specifications are fully indexed by Mata Hari. This means that all of the power within the Mata Hari product via full Boolean searches and robust filter capabilities can be applied to every single search service on the Web, whether those services natively support these features or not
- **features and filters** – a complete set of user-controllable filters can limit searches based on time, document size and date and site characteristics and name
- **fast retrievals** – a multi-threaded design enables Mata Hari to establish up to 120 simultaneous connections to the Internet, resulting in extremely fast document retrievals and background or offhours searching
- **multiple scoring methods** – unlike individual services or other metasearch products, Mata Hari provides five different ways to score documents for relevance, including a "more like this" (reranking) option that scores results based on documents containing the results you want
- **local, desktop database** – all documents returned by an initial search are stored on the desktop. Mata Hari is thus perfectly suited to the strategy of successively narrowed searches that information professionals recommend. Moreover, these subsequent searches are done at the much higher speeds of the local computer, eliminating the delays and bottlenecks of Internet retrievals
- **advanced set manipulation** – single to multiple terms, engines, queries, scores, or documents can be used for narrowing and manipulating results, using either **AND** or **OR** operators. And, after selections have been made, you can re-generate the terms lists applicable only to those results
- **publish and share results** – you can send a partial or complete list of results to friends and colleagues as a Web page, or you can distribute your databases or search configuration specifications. Mata Hari results can also be output as a comma-delimited file. You can determine to export all results or only those selected. These files can then be imported into most databases and spreadsheets.

Once you learn how to use Mata Hari you've learned how to squeeze the most out of the Internet. And, for those in a hurry and new to searching, just use Mata Hari's simple text searches.

You can learn more about Mata Hari or download a fully-functional copy for a free 30-day trial at:

<http://thewebtools.com/>

Topic 51: What is Fast?

The only reason for searching the Internet is to find Web pages of value. The only meaningful measurement is the total elapsed time from information request to acquiring acceptable results. There is no faster tool than Mata Hari using this standard. Here's why:

The complete cycle to acquiring desired information has several steps. Let's review how you perform it manually:

1. Pose a query to a search engine to get a list of potential Web page links that may or may not meet requirements – wait for the search engine to return the first page of the link list
2. Evaluate potential links and click on one that appears to have merit – wait for it to download for viewing in your browser
3. Evaluate the returned Web page for acceptability
4. If unacceptable, go back to the search engine listing of potential links and repeat from step 2
5. If partially acceptable, remember or bookmark the Web page and repeat from step 2
6. If no links appear acceptable, refine query, and pose again to search engine – return to step 1
7. If the results from the specific search engine are not acceptable, repeat entire process for each additional search engine you decide to try.

This is a time-consuming process which can get old fast! Specific search engines or desktop search tools may appear faster than Mata Hari because they only perform one or a few of these steps. But the remaining steps are not automated and can literally cause hours to be wasted waiting, downloading, and viewing pages on a one-by-one basis.

Instead, Mata Hari does many of these steps automatically and in PARALLEL. Also, it is able to operate in the background while you do other tasks. Multiple "threads" or connections are used to get Web page listings from multiple search engines simultaneously. As soon as these listings are returned from the search engines, Mata Hari begins downloading the actual Web pages (minus images) from multiple sites – again, simultaneously. Mata Hari can also simultaneously submit multiple queries to multiple search engines. Web page links are evaluated at two stages for whether they are "acceptable", and if so, are then scored. If the Web page link is a duplicate or does not first meet its site filter criteria, it is immediately rejected so no processing time is lost downloading a useless Web page.

Potential misconceptions can arise when not comparing search tools on an apples-to-apples basis. For example, on Mata Hari's search progress screen during the search, the "total" value represents all Web page links identified during that search. In competitor products, the equivalent value is often presented as "results". Yet those "results" DO NOT represent the documents desired, they are only those links returned by the source search engines as "potential documents". This is because those links include many duplicates, many dead links (for Web pages that have been moved or no longer exist), and even mis-indexed pages (content having nothing to do with the query -- this can occur by human indexing error or by Web developers including metatags having nothing to do with the page topic.) Because of its powerful screening capabilities, only "accepted" results in Mata Hari reflect the desired search criteria.

Something we think you'll find interesting is to check out the rate at which Mata Hari downloads new links and documents simultaneously from the Internet – not through just a big pipe, but even from a 28.8K modem. There are many ways and third-party tools to monitor your connection to the Internet while searching with Mata Hari. Run whatever tool you have; you'll be amazed at the rate Mata Hari is delivering data to your system with any size pipe.

Topic 52: Universal Search Power

Mata Hari brings to the user a universality that has been lacking. Once it becomes obvious that just a few of the largest search engines cover, at most, one third of the Internet with the remainder covering far less, it is equally clear that the net can only be adequately searched through the use of multiple search services.

Universal Search Power means that functions within Mata Hari – phrase searching, plain-text searching, extensive Boolean operators, use of special search characters, or flexible filter options – can be used with all selected search engines whether they natively support the function or not.

There are a number of desktop tools that access multiple engines and several metasearch engines on the net that do the same. But, the biggest problem has not been acquiring a sufficient number of returned documents, it has been sifting through all returned documents to get what you're searching for. Mata Hari accesses multiple engines to survey the net, provides universal functions across these engines, and then provides powerful tools to sort it all out quickly.

Topic 53: Nearly 600 Search Engines in 60 Topic Categories

Searching with more than one search engine is very easy. While you could just select every engine in the list, testing indicates that is not the most efficient way to use Mata Hari or any other metasearch tool.

For all of the valuable information available on the net, there is probably at least ten times as much chaff. And to make matters worse, spamming through the use of inappropriate metatags by overly enthusiastic Web page developers often causes documents with content unrelated to the search query to be identified. Including search engines that specialize on unrelated topic areas only adds more chaff with scant chance that any additional information of value will be identified. So, including Web engines when you're searching usergroups or an engine specializing in jobs when you're looking for information about early civilizations will add nothing but more clutter to the documents returned. In addition, valuable bandwidth is wasted communicating with and downloading useless pages from these engines.

Selecting search engines directly related to your query is VERY effective in returning a much larger percentage of quality results and in conducting the search more quickly. It's very easy to select the appropriate engines from Mata Hari's 60 topic categories. Often, you'll just use the whole category.

Topic 54: Using Boolean Power with Non-Boolean Search Engines

You've already learned how use of Boolean operators can quickly winnow down the number of returned documents to those that meet your search goals. But, there are two barriers to taking full advantage of this ability: loose Boolean interpretation by many Boolean engines and the inability to use Boolean power with topic-related, non-Boolean engines.

Unfortunately, many of the Boolean engines do not strictly adhere to true Boolean constructs. The reason is that many individuals not familiar with Boolean techniques also use those search engines. To facilitate such use and assure they get something returned, the engines instead use a loose interpretation of Boolean. Sadly, to those familiar with Boolean, the loose interpretation defeats the entire purpose and returns extra junk.

For example, if you conduct a search for “chevy AND pickups” you should receive only those documents that have both terms on their pages. In reality, you will also get documents that have only “chevy” and others that have only “pickups”. This undermines the entire concept and power of using Boolean queries.

However, with Mata Hari, if you check the “Strict Boolean Adherence” checkbox, you will receive only those pages that truly adhere to Boolean and only return documents with both terms of your query. This will work with ALL the search engines you select whether they are Boolean engines or not. This feature is very powerful in filtering the returned documents.

Topic 55: Filtering, Phrases, and Plain Text

Mata Hari offers additional ways to filter documents and keep the amount of chaff to a minimum. When you know other characteristics of the information you seek, you can often use that knowledge to reduce the clutter. For example, you can exclude documents in other languages or by date. Mata Hari can filter documents by several parameters:

1. URL – The exclude criteria can be defined as any or all portions of a given site’s URL. You can exclude documents from selected countries or type (.org, .mil, .edu, etc.). As many sites or site specifications as you would like can be picked and added to an exclude group.
2. Page size – You can determine the minimum and maximum size of all pages to be accepted.
3. Time – You can accept only those pages that have been created or modified within a set number of days that you select.

Because these filtering options are part of Mata Hari’s Universal Search Power, they work with all selected search engines whether all the engines support these features or not.

Similarly, not all search engines support use of phrases or plain text. But, you can use phrases and plain text with any and all engines you select through Mata Hari. Phrases are particularly important because they are so powerful in excluding undesired documents. In fact, other than properly identifying your search topic, phrases are probably the next most powerful feature to use to cull useful documents from all the rest.

Topic 56: Efficiently Culling Results

Now you’ve surveyed much of the Web, simultaneously using a hundred Internet search engines and directories. This is all well and good, but comprehensive surveys mean comprehensive results. When it comes time to sort through the returns to find the information you’re after, you need powerful tools to make the job fast and efficient. Mata Hari provides these, too.

The main screen is primarily comprised of three folders and makes use of set theory and advanced information processing techniques. This allows you to sort, cull, quickly review, and study the results in many different ways: You can display only those pages that came from one engine or query, or from combinations of engines or queries. Or you can display the terms and their counts from any page(s) you select. Or display only the pages that contain the terms you select. Essentially, you can slice and dice the returned information and view it from any angle you wish.

Topic 57: Local Viewer

The Local Viewer integrated into Mata Hari allows you to quickly view and scan through returned URLs. It is an interim display point for URL page contents and is sort of a mini-browser that works on and off-line enabling results to be quickly viewed within the Mata Hari application without needing to invoke the full browser. When you first bring it up, it will display the URL that is highlighted on one of the folders. Clicking on a different URL will immediately cause it to be displayed.

Because the Local Viewer displays Web documents directly from the database, it is very fast. In fact, you can click through documents as fast as you can read and can usually review many pages in just a few minutes.

Topic 58: Query and Engines Folder

On the Query and Engines folder, you are able to display results depending on which queries and search engines you select as active. This allows you to determine which Web pages resulted from which query and from which search engines. It is a very powerful way to get a feeling for the kinds of topics each search engine covers and which engine is best for a particular type of search.

When you place a check in the checkbox next to one or more of the search engines used, the Web pages displayed will be only those pages that were returned from the engines checked. Accordingly, checking or unchecking a query also determines which Web pages are displayed—AT LEAST one query and one engine must be selected to see results. When you check one or more, Mata Hari displays only those Web pages that were returned by those selected engines and queries.

One use of the Query and Engines folder is to identify which search engines produce the most results for you. Selecting each engine in turn will identify which engines brought back the most results. This analysis can also show that some engines mostly duplicate the results of others. Through such tests you can refine the search engines you typically use to those that are most productive, thereby reducing your overall connect time to get valuable results and the amount of garbage returned.

Topic 59: Scores Folder

Mata Hari automatically uses five different methods for scoring Web page documents. In addition, it provides a sixth method allowing you to rerank documents for relevance to one or more documents you select. These methods are sometimes only slight variations of one another, but can result in substantial scoring differences when comparing documents. Mata Hari scores and reranks documents that have been returned by a search and are now resident in its local database.

To use reranking, you inspect the retrieved documents (Web pages) with the Local Viewer. When you find a Web page that is very close to or exactly what you've been looking for, mark it, and rerank. That will cause Mata Hari to evaluate all the other pages against the marked page as if it were a template and will score the other pages accordingly. Then, look at the reranking scores – those with the highest scores will be at the top and most like the page you've marked. This ability is invaluable for quickly finding similar pages out of the many hundreds returned. Mata Hari does the comparison and reranking in seconds, something that would take a person hours to do manually. Sometimes you may find two or three pages that each have part of what you're looking for, but none have everything. Just mark both or all three pages and rerank. Mata Hari will score all the documents based upon a composite of the marked pages. You can mark up to 32 pages per reranking.

Topic 60: Terms Folder

The Terms folder provides powerful union and intersection options to manipulate the returned Web pages and the terms on them. Also, the counts of those terms are displayed for whatever page or pages you've highlighted. The collective functionality is extremely powerful as it lets you identify the frequency of occurring terms on the Web pages you've received from a search. This can help you identify terms related to your queries which you had not previously considered and can help you to easily locate certain terms on the Web pages from which they were derived.

Let's say you posed a query about "Montana" and "resorts". After viewing a few results pages, you now realize you would like to narrow your results to resorts that offer flyfishing and golf. By highlighting the three terms of "resort", "flyfishing" and "golf", selecting the intersection option, and clicking on the **Show pages by highlighted terms** button, Mata Hari now displays only those pages that contain ALL THREE terms. If you select the union radio button instead, the Web pages shown will be those upon which ANY of the highlighted term(s) appear. Or you can go the other way: highlight one or more URLs, select the intersection option, click on the **Show terms for highlighted pages** button, and the terms shown will ONLY be those which are contained on ALL the selected URLs.

Notes, Links and References

- 1** Term counts used in this tutorial are based on the AltaVista Advanced Search option [<http://www.altavista.com/cgi-bin/query?pg=aq>]. Actual term counts were obtained by posing the query indicated using the 'Give me only a precise count of matches.' checkbox on this page. Term counts were originally taken on April 24, 1998; they were completed updated on May 6, 1999.
- 2** The directory engine used for this tutorial is Yahoo! [<http://www.yahoo.com/>].
- 3** Nua, April, 1999.
- 4** Nua; T. Miller, "Interactive Demographics," presented at the ANA Conference, New York City, April 8, 1997.
- 5** CyberAtlas, December, 1998.
- 6** MediaMetrix, November 1999, <http://www.mediametrix.com/TopRankings/TopRankings.html>.
- 7** CyberAtlas, November 1998.
- 8** Georgia Tech's 10th Graphic, Visualization and Usability Survey, October 1998.
- 9** TWTC, The WebTools Company Business Plan, May 1999, 43 pp.
- 10** Internet World, Sept. 28, 1998.
- 11** S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," *Nature* magazine, v. 400, pp. 107-109, July 1999.
- 12** S. Lawrence and C.L. Giles, "Searching the World Wide Web," *Science* magazine, v. 280, April 3, 1998, pp. 98-100.
- 13** See <http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm>
- 14** IDC Web Index, July 1997.
- 15** <http://www.beaucoup.com/>.
- 16** Other sources for search engine listings besides beaucoup.com are <http://www.searchpower.com>, <http://www.dreamscape.com/frankvad/search.html> and <http://www.finderseeker.com/>. VisualMetrics believes the combined listings from these services to exceed 3,000. However, other known, specialty search sites are NOT included in these listings. Perhaps across the entire Internet there are 10,000 or more useful sites with search capabilities.
- 17** Additional NEC Research information may be found at: <http://www.neci.nj.nec.com/homepages/lawrence/websize.html>.
- 18** US Department of Commerce, "The Emerging Digital Economy," April 15, 1998.
- 19** One of the best discussions about metatags can be found on the Virtual Promote site: <http://www.virtualpromote.com/metatag.html>.
- 20** G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison Wesley, 1989.
- 21** See <http://www.research.digital.com/SRC/whatsnew/sem.html>.
- 22** If you'd like to see these hidden tags on a given Web document, save the document while viewing in your browser using the 'Save As' option on your browser's 'File' menu option. Then, view that document with a text editor or Wordpad. You will see these hidden fields shown in HTML brackets (e.g., <Description= ...>).
- 23** These terms were coined by Barbara Quint; see further: <http://www.hut.fi/~ipaavola/FLI/Courses/Database/quest01.html>.
- 24** This article presents results of searching on conventional search databases, such as Dialog, versus the Internet: <http://www.infotoday.com/searcher/feb98/story1.htm>.
- 25** B. Pinkerton, "Finding What People Want: Experiences with WebCrawler" found at: <http://info.webcrawler.com/bp/WWW94.html>.
- 26** See <http://niko.unl.edu/bs101/notes/lecture19.html>.
- 27** Encarta 96 Encyclopedia, Microsoft, c. 1995.
- 28** The Original Roget's Thesaurus, St. Martin's Press, 1962.

-
- 29 University of Minnesota Raptor Center, <http://www.raptor.cvm.umn.edu/>.
- 30 See <http://altos.rnd.runnet.ru/internet/logic.html>.
- 31 Based on using the ranking option for all search keywords in the order shown in the query.
- 32 See <http://raysweb.net/specialplaces/> .
- 33 See <http://www.deev.com/falcons/live-falcons.html>.
- 34 See <http://www.dnr.state.oh.us/odnr/wildlife/publications/peregrine/falcon.html>.
- 35 See <http://www.wwfcanada.org/facts/peregrin.html>.
- 36 This link requires Java; most browsers have it
- 37 A complete listing of country two-letter domain codes can be found at:
<http://www.thrall.org/domains.htm>.
- 38 This estimate comes from using the 'anchor' search option in AltaVista by issuing the search **anchor:"method=" AND anchor:"action="**, which returns 11,000 hits on this service. Results were then spot checked to determine that search forms appeared on the results pages; most did. Since AltaVista incompletely indexes the Web, the total amount is likely much higher.
- 39 Search Engine Watch is a very useful, authoritative site on all aspects of search engines. It is found at: <http://www.searchenginewatch.com>. Highly recommended.
- 40 This calculation is based on information presented in Figure 2 of reference [11].
- 41 See further, S. Feldman, "Just the Answers, Please: Choosing a Web Search Service," Searcher magazine, May 1997. This article is no longer online. However, other useful references may be found at <http://www.infotoday.com/searcher/>.
- 42 See <http://mediamatrix.com/TopRankings/TopRankings.html>. Some domains include other services than search; however, as many large services have also become portals, the rankings are still comparable.
- 43 See <http://www.notess.com/search/stats/>.
- 44 This service has added nearly 500 engine listings in the past year, most in the 'regional' categories.
- 45 Clifford Lynch, *Scientific American*, December 1998.
- 46 "Recommended Search Strategy: Search With Peripheral Vision," issued by the Teaching Library Internet Workshops from UC Berkeley, found at:
<http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Strategies.html>.
- 47 See <http://www.rice.edu/Fondren/Netguides/strategies.html>.
- 48 See <http://www.infotoday.com/searcher/default.htm>.
- 49 The Steinberg article may be found at: <http://www.wired.com/wired/archive/4.05/indexweb.html>.