Search 3: Internet Searching: Approaches & Rules

Payam Kabiri, MD. PhD.
Epidemiologist

Department of Epidemiology & Biostatistics
School of Public Health
Tehran University of Medical Sciences

- *“When I took office, only high energy physicists had ever heard of what is called the World Wide Web... Now even my cat has it's own page.”*

Bill Clinton

Internet is like a library

Many have likened the Internet to a huge, global library.

But ?!?

But a Library with many problems

- The Web lacks the *bibliographic control standards* we take for granted in the print world;
- There is no equivalent to the *ISBN* to uniquely identify a document
- There is no standard system, analogous to those developed by the library of congress, of *cataloguing* or classification
- There is no central catalogue including the Web's holdings; in fact, many, if not most, Web documents lack even the name of the author and the date of publication.

User Frustration

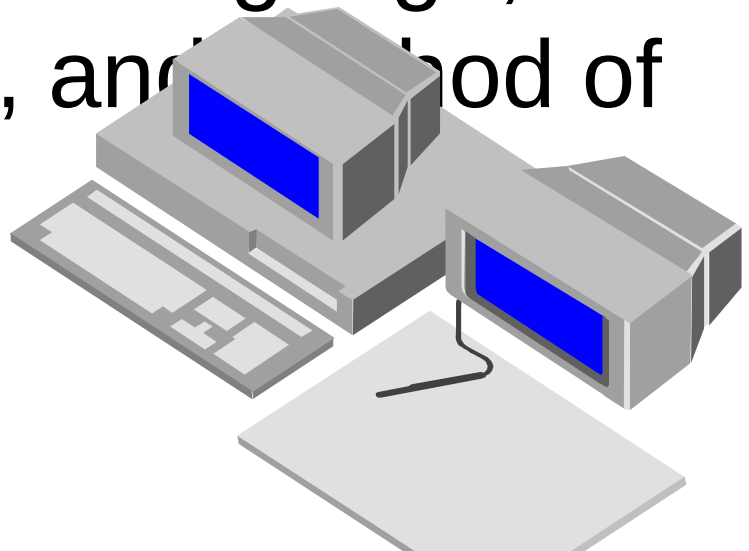
- 2/3 to 3/4 of all users cite finding information as one of their primary uses of the Internet
- 2/3 to 3/4 of all users cite the inability to find the information they seek as one of their primary frustrations (second only in frustration to slowness of response)
- but 2/3 of internet users don't know how to carry out effective internet searches .

Web Growth

- Approximately *+3 milliard* web pages are being added daily, and overall doubling time of web documents is about *8 months*
- The whole number of Websites are more than *200 million* Websites.

Search Tools

- Instead of a central catalogue, the Web offers the choice of dozens of different search tools, each with own database, command language, search capabilities, and method of displaying results.



To find information in the web:

- There are two ways:

1- Using Search Engines

2- Using Directories

Search Engines

Search Engines

- There are more than *2500* search services presently on the web.

How a search engine works ?!?

- Search engines use *Spiders (Crawlers)* or *Robots* to go out and retrieve individual web pages or documents.
- Then they will make index files.

How a search engine works ?!?

- A search engine operates, in the following order
  - Web crawling
  - Indexing
  - Searching

Contents of web-pages

1. ***Title* :** what is seen in the blue bar if the webpage.
2. ***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
3. ***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
4. ***Body* :** the actual, viewable content of the document.

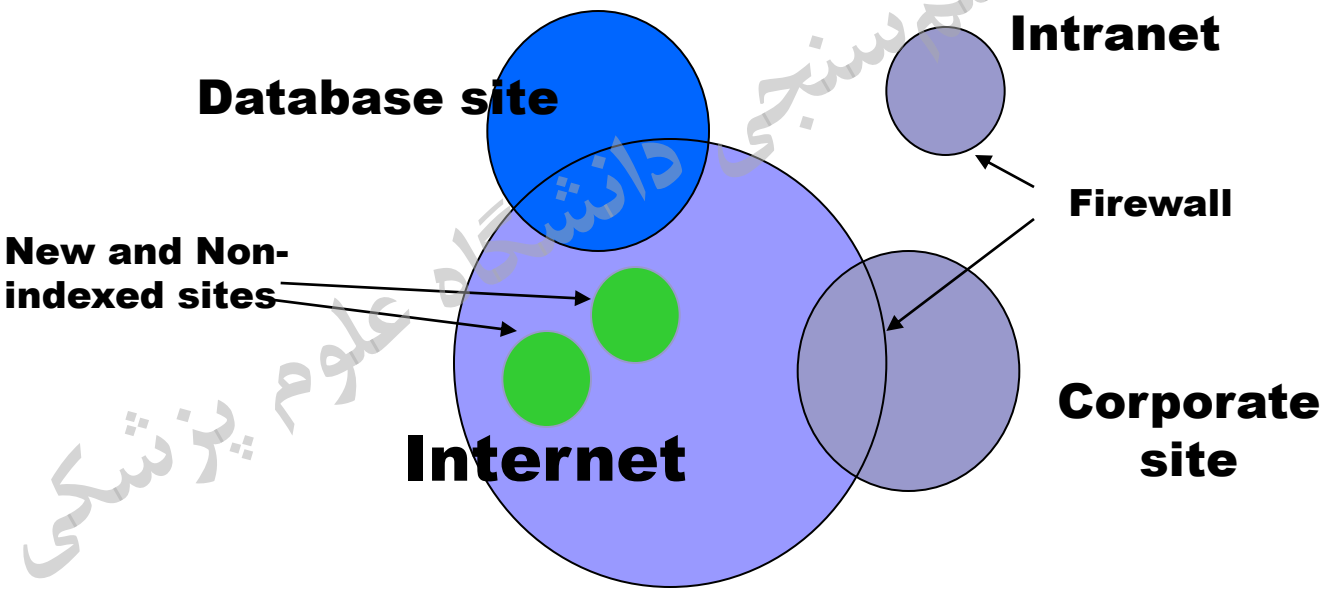
Ranking of documents

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.

But none of them come close to indexing the entire Web…!

- Content of Adobe PDF and formatted files
- The content in sites requiring a log in
- Intranets; pages not linked from anywhere else
- Commercial resources with domain limitations
- Sites that use a robots.txt file to keep files and/or directories off limits
- Non-Web resources

Why isn't it all indexed?



Limiting factors for search engines

Recall, precision, and coverage are limiting factors for most search engines.

- Coverage refers to what percentages of the potential universe of relevant documents is cataloged by the search engine.
- *Recall* measures what fraction of relevant documents retrieved
- *Precision* measures how well the retrieved documents match the query

Example

- For example consider a search engine with 10,000,000 documents, five of which mentions halzoun out of a total universe of 50 articles about halzoun (45 documents not indexed in this search engine).
- For a query about halzoun that returned 4 documents and 2 of other documents :
  - Precision = 0.66 (4/6)
  - Recall = 0.8 (4/5)
  - Coverage = 0.1 (5/50)

Updating the indexes

- Beside coverage there is also question of keeping the links *up to date*.

Coverage statistics and dead links *1998*

Search engine	% of all indexed pages	% that are dead links
Alta Vista	47	2.5
Northern Light	39	5
Inktomi	34	Not available
Excite	17	2
Lycos	16	1.6
InfoSeek	14	2.6

Examples of search engines

- AltaVista http://www.altavista.com
- Excite http://www.excite.com
- FAST http://www.alltheweb.com
- Google http://www.google.com
- HotBot http://www.hotbot.com
- Northern Light http://www.northernlight.com

http://www.jostejoogar.com

Health specific search engines

- Medstory
  - http://www.medstory.com/
- Eheath Sites
  - http://www.ehealthsites.com/
- Med Explorer
  - http://www.medexplorer.com/
- Mayo Clinic Health Oasis
  - http://www.mayohealth.org
- Medical World Search
  - http://www.mwsearch.com

Health specific search engines

- TextMed
  - http://www.textmed.com/
- OnHealth
  - http://www.onhealth.com
- MedHunt
  - http://www.hon.ch/MedHunt/
- Md Tools
  - http://www.mdtool.com/
- Nurse Web Search
  - http://www.nursewebsearch.com/

Search engines for search engines

Search Engine Colossus

- http://www.searchenginecolossus.com/

Search Engine Watch

- http://searchenginewatch.com

Search Engine Showdown

- http://www.searchengineshowdown.com/

Popular multi-threaded search engines (MetaSearchEngines)

- Dogpile http://www.dogpile.com
- Metacrawler http://www.metacrawler.com
- Search.com http://www.search.com
- Inference FIND http://www.infind.com
- Internet Sleuth http://www.isleuth.com
- Mamma http://www.mamma.com

Academic Search Engines

- Google Scholar
- http://scholar.google.com

- Scirus
- http://www.scirus.com

Benefits of search engines

1. Because many searches are not very well defined, indexes will often be the *best starting point*.
2. Indexes, as they cover most (or at least more) words on a given page will offer a *richer list of returns*.
3. Indexes are usually larger because of the much lower overhead in adding pages to the search engine (more sensitivity)

Problems with Indexes

- 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* (less specificity).
- 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.

To find information in the web:

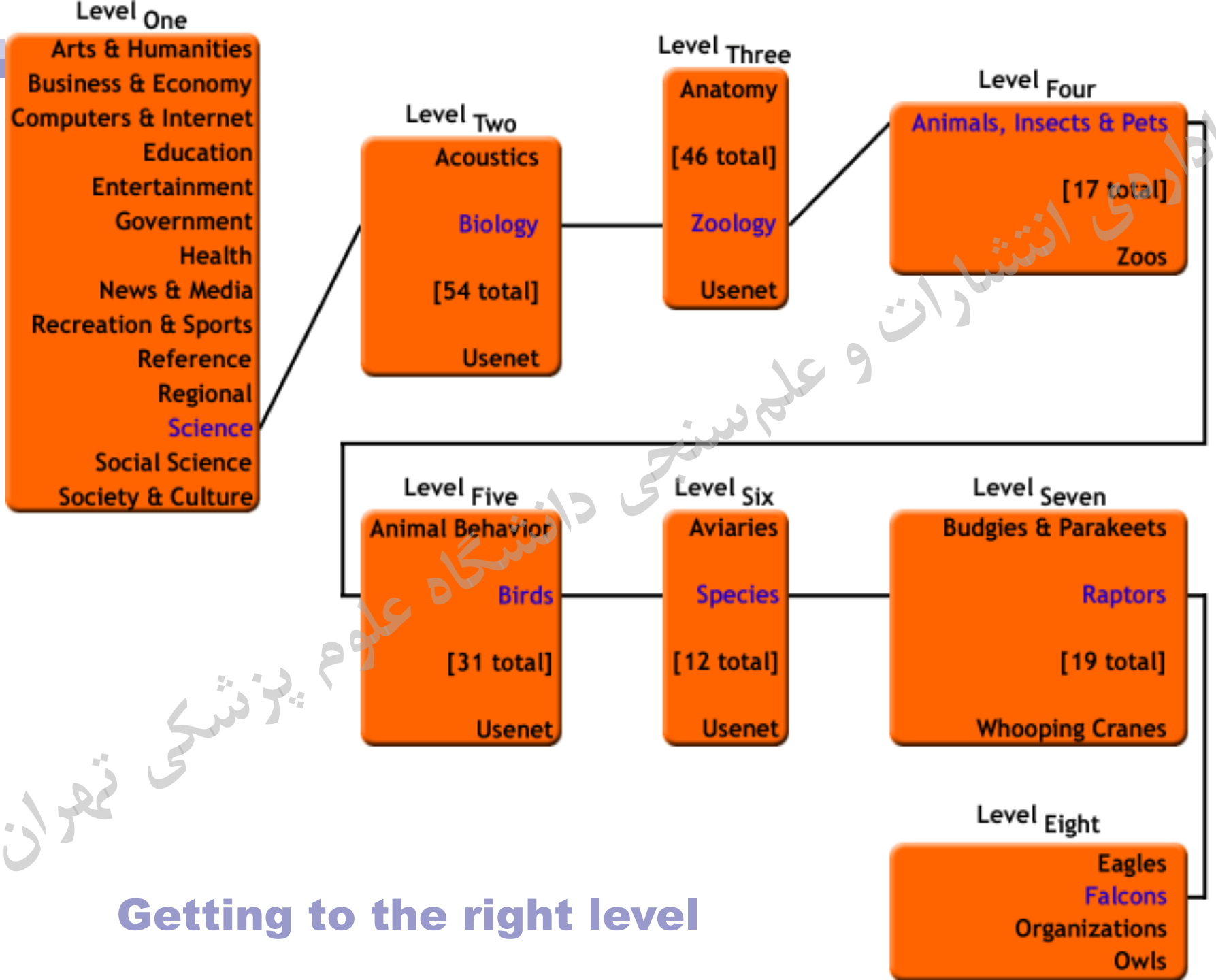
- There are two ways:

1- Using Search Engines

2- Using Directories

Directories

- 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.



Getting to the right level

Directories

- For searches that are easily classified, 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 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.

Examples of subject directories

Yahoo	http://dir.yahoo.com
Open Directory	http://dmoz.org
LookSmart	http://www.looksmart.com
Librarian Index	http://lii.org
Infomine	http://infomine.ucr.edu
Academic Info	http://www.academicinfo.net
About.com	http://www.about.com

Search engines vs. Directories

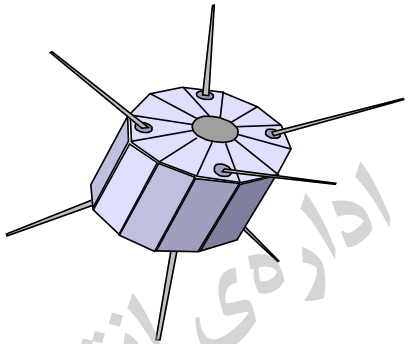
- *Search engines* indexes words or terms in internet documents.

  They are machine-based.
- *Directories* classifies web documents or locations into an arbitrary subject classification scheme or taxonomy.

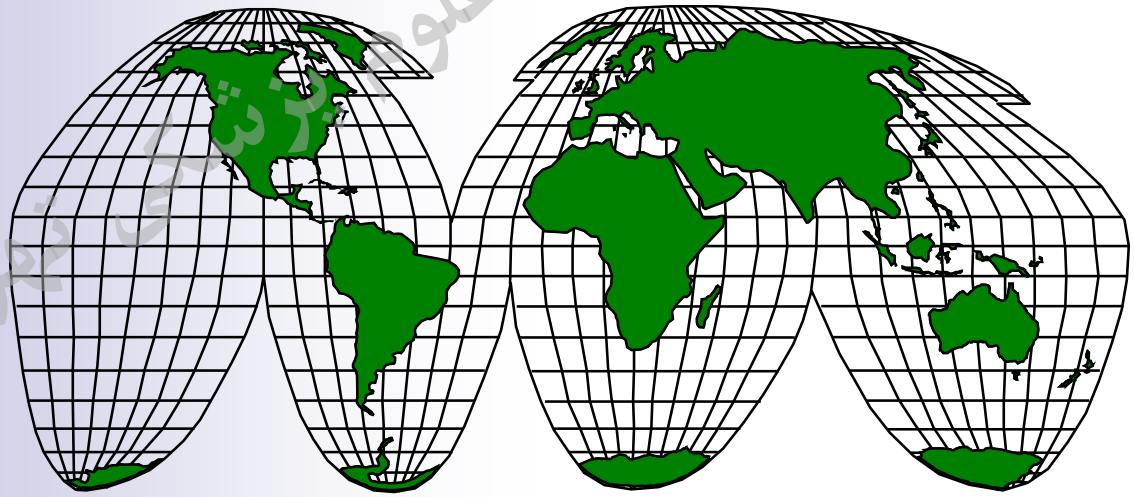
  They are human-based.

Problems with Directories

- If a given topic area has *not been specifically classified* by the search directories, finding any related information on that topic is made more difficult.
- *lack of coverage* because of the cost and time in individually assigning sites to categories.



Internet Search Strategies



Avoid Misspellings

- ***searching*** 269,000,000
- ***serching*** 207,000
- ***searchng*** 97,700
- ***seerching*** 3,860
- ***Sherching*** 5,670

!


A
B
C
Advanced Search
Preferences
Search Tips
Google
google
D
E
Google Search
I'm Feeling Lucky
F
G
Searched the web for google.
Results 1 - 10 of about 2,240,000. Search took 0.03 seconds.
H
Categories: Google Directory Computers > Internet > WWW > Searching the Web > Search Engines
I
Google
... Google Web Directory the web organized by topic Cool Jobs - Add Google to Your Site - Advertise with Us - Google in your Language - All About Google ...
J
Description: Lists the results in the order of popularity, determined by the number of links from other sites....
K
Category: Computers > Internet > WWW > Searching the Web > Search Engines
L
M
www.google.com/ - 3k - Cached - Similar pages
N
O
P
Q
Google Search:
Google, Search Tips. New! Use Your WAP Phone to Search
The Web with Google Google Web Directory ...
www.google.com/custom - 3k - Cached - Similar pages
[More results from www.google.com]
R

Internet Search Strategies

Search Recommendation 1

- *Recommendation:* Recognize and distinguish at least 2 to 3 concepts in query
- *Example:* “diabetes mellitus”, “sensory neuropathy”, biguanide*, treatment OR therapy
- *Why important:* triangulating on multiple query concepts, narrows and targets results, generally by more than 100 to 1000

Search Recommendation 2

- *Recommendation:* Put each concept in a parenthesis
- *Example:* (“diabetes mellitus”) (“sensory neuropathy”) (biguanide*) (treatment OR therapy)
- *Why important:* simple way to ensure the search engine evaluate your query the way you want, from left to right

Search Recommendation 3

- *Recommendation:* Use 6 to 8 words in query
- *Example:* Diabetes, mellitus, neuropathy, sensory, treatment, biguanide
- *Why important:* more keywords chosen at appropriate level, can reduce the universe of possible documents returned by 99%

Search Recommendation 4

- *Recommendation:* Use nouns or objects as query keywords
- *Example:* Diabetes
- *Why important:* actions (verbs), modifiers (adjectives, and adverbs), and conjunctions are either “thrown away” by search engines or too variable to be useful

Search Recommendation 5

- *Recommendation:* Try to pick up singular and pleural versions of the nouns
- *Example:* biguanide OR biguanides
- *Why important:* use asterisk wildcard. The wildcard tell the search engine to match all characters after it, preserving keyword slots and increasing coverage by 50% or more

Search Recommendation 6

- *Recommendation:* Use synonyms via the OR operator
- *Example:* treatment OR therapy
- *Why important:* cover the likely different ways a concept can be described. Generally avoid OR in other cases

Search Recommendation 7

- *Recommendation:* Combine keywords into phrases where possible
- *Example:* “diabetes mellitus”
- *Why important:* use quotes to denote phrases. Phrases restrict results to exact matches, narrows results by many time

Search Recommendation 8

- *Recommendation:* Link concepts with the AND operator
- *Example:* (“sensory neuropathy”) AND (“diabetes mellitus”) AND (biguanide*) AND (treatment OR therapy)
- *Why important:* AND glues the query together

Search Recommendation 9

- *Recommendation:* Order concepts with main subject first (Put Your Main Concept First)
- *Example:* (“sensory neuropathy”) (“diabetes mellitus”) (biguanide*) (treatment OR therapy)
- *Why important:* put main subject first. Engines tend to rank documents more highly that match first terms or phrases evaluated

Search Recommendation 10

- *Recommendation:* Refine your search if necessary
- *Why important:*
  - Many sites offer a "Refine search" option so you can modify your search term
  - Some have a "more like this" option
  - Or go BACK to the search box to change your query

Search Strategy we recommend:

1. Formulate the search question and its scope
2. Identify the important concepts within the question
3. Identify search terms to describe those concepts
4. Consider synonyms and variations of those terms
5. Prepare your search logic

Medical *Meta-Sites* examples

- Martindale Center
  - http://www.martindalecenter.com/
- Hardin Website
  - http://www.lib.uiowa.edu/hardin/md/

اگر میل داشتید Email بزنید !

kabiri@tums.ac.ir