CSC 101
Fluency with Information Technology and Computing

Chapter 5
Locating Information on the WWW

Brian McBride
Department of Computer Science, Engineering and Physics (CSEP)
University of Michigan - Flint

Email: brmcbrid@umflint.edu

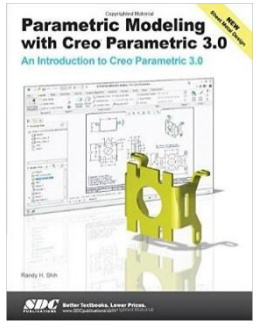
How do you locate information on the World Wide Web?

Use a Web browser

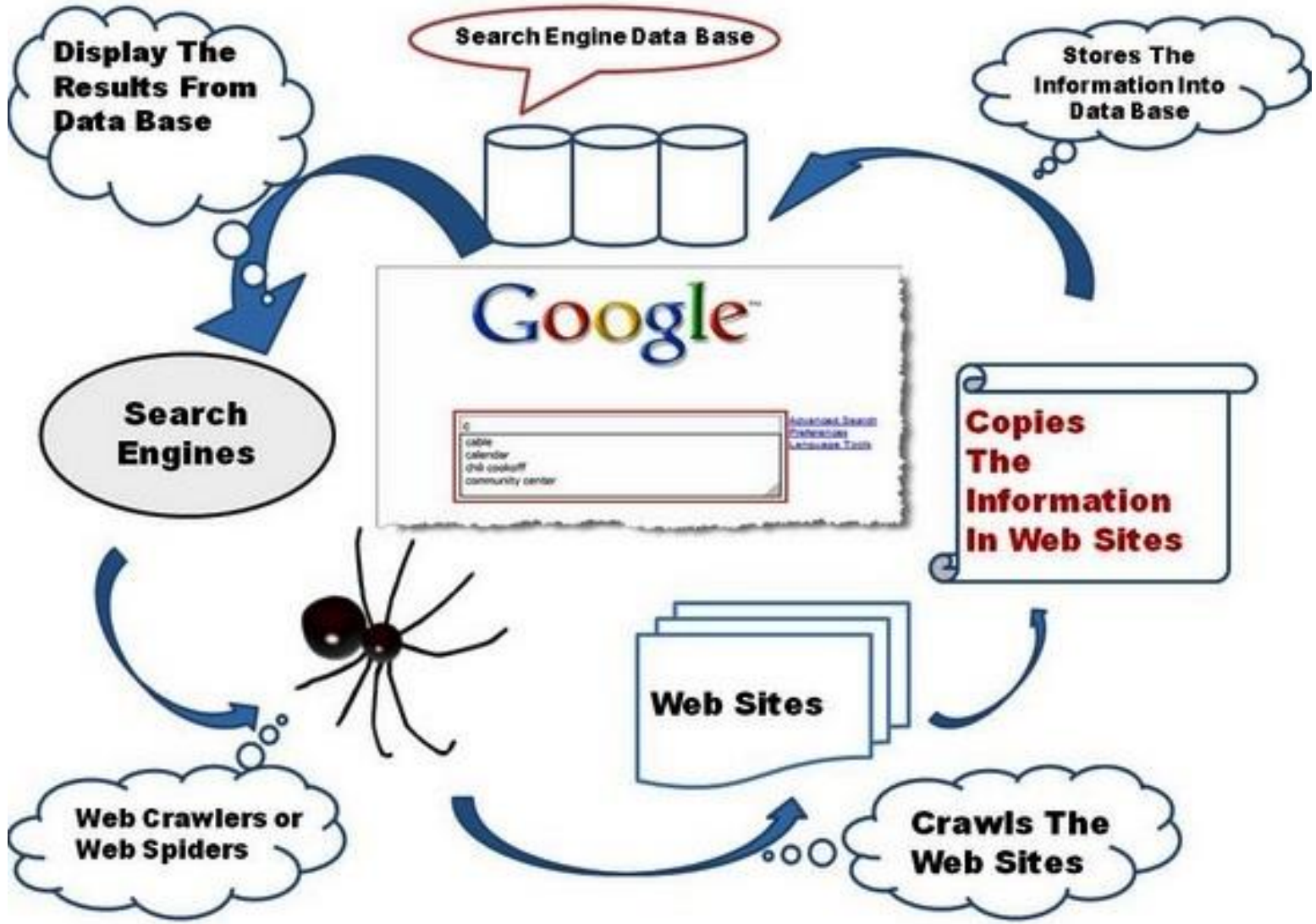
- Type in a URL to navigate to a Web page
  - Click links on page to navigate to other Web pages
- Go to Google, Bing, or Yahoo to conduct a search
  - Creates a list of “Hits”
  - Read page title and snippet to determine relevance

Google Searches

- Advanced Search
- Search by Image
  - e.g.



- Site Search
  - (e.g. Final Exam Schedule:www.umflint.edu)
- Related Sites
  - (e.g. related:www.umflint.edu)


Display The Results From Data Base
Search Engine Data Base
Stores The Information Into Data Base
Google
Search Engines
Copies The Information In Web Sites
Web Sites
Web Crawlers or Web Spiders
Crawls The Web Sites

How a Search Engine Works

- The first step, ***crawling***, visits every Web page that it can find
- How are the pages found?
  - The crawler has a **todo** list that is loaded with a set of pages to start
  - When a URL is found while crawling a page, it adds that URL to the **todo** list
- The main work of the crawler is to build an index

How a Search Engine Works

- The index is a list of tokens (or words) that are associated with the page
- The token might be part of the page’s title
- There are other ways for a token to be associated with a page
- For each token, the crawler creates a list of the URLs associated with that token

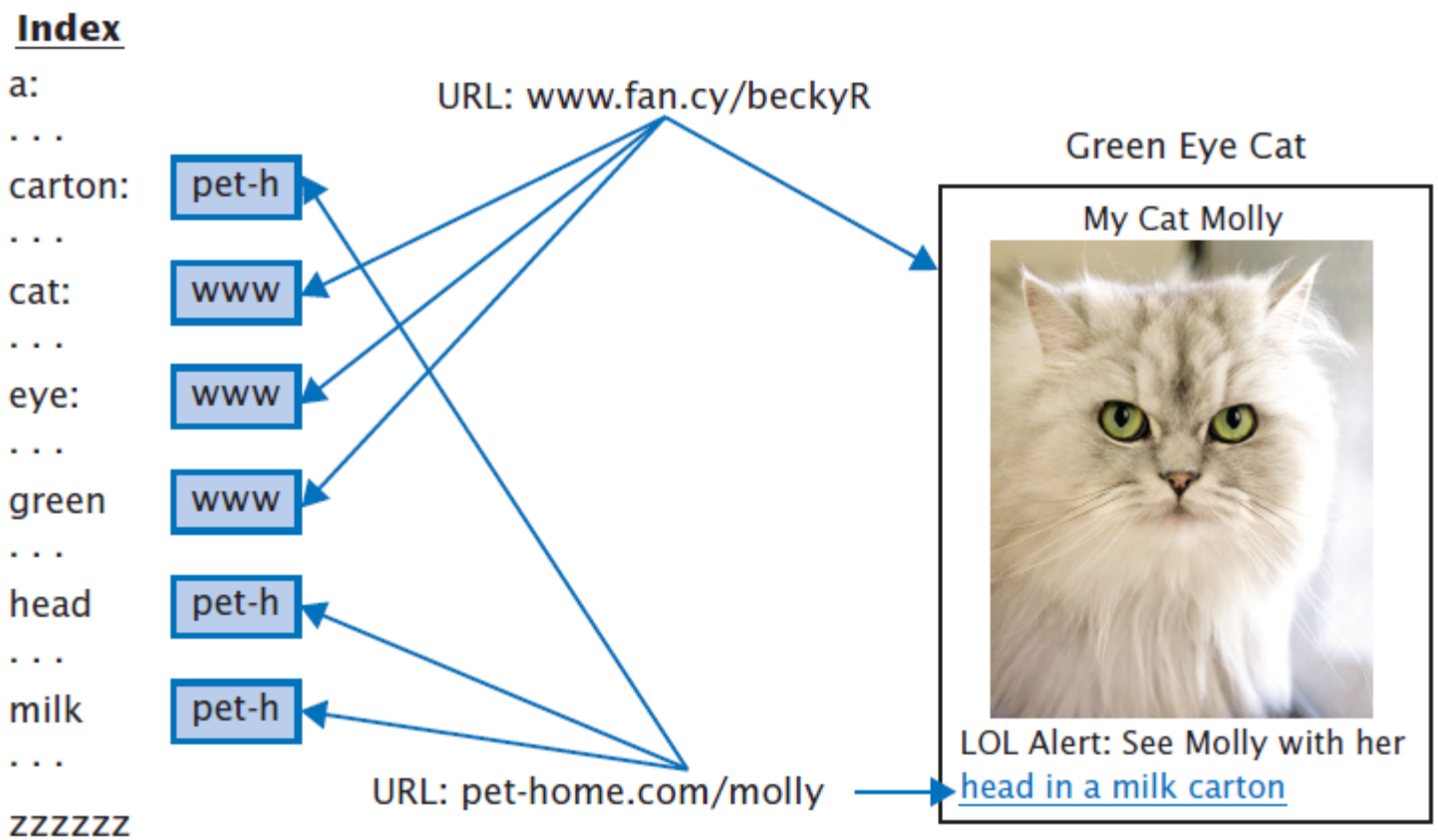


Figure 5.1 Crawling over the Green Eye Cat page: The crawler adds the page's URL to the lists for each word in its title; for words in the anchor text, the link URL is added to their lists.

How a Search Engine Works

- The second step is ***query processing***
- The user presents tokens (aka search terms) to the query processor
- The search engine then looks up the word in the index and returns a ***hit*** list
- By creating the index ahead of time, search engines are able to answer user queries very quickly

Multiword Searches

- With a multiple-word query, the pages returned should be appropriate for *all* of the queried words
- **AND-query**
  - Each page returned *should* be associated with *all* the words
  - There is no index entry corresponding to a *set*
  - There is only a list for the individual words

Intersecting Queries

human **powered** **flight**

Rules for Intersecting Alphabetized Lists

- To intersect several alphabetized lists:
  1. Put a marker (arrow) at the start of each token’s index list
  2. If all markers point to the same URL, save it, because all tokens are associated with the page
  3. Move the marker(s) to the next position for whichever URL is earliest in the alphabet
  4. Repeat Steps 2–3 until some marker reaches the end of the list

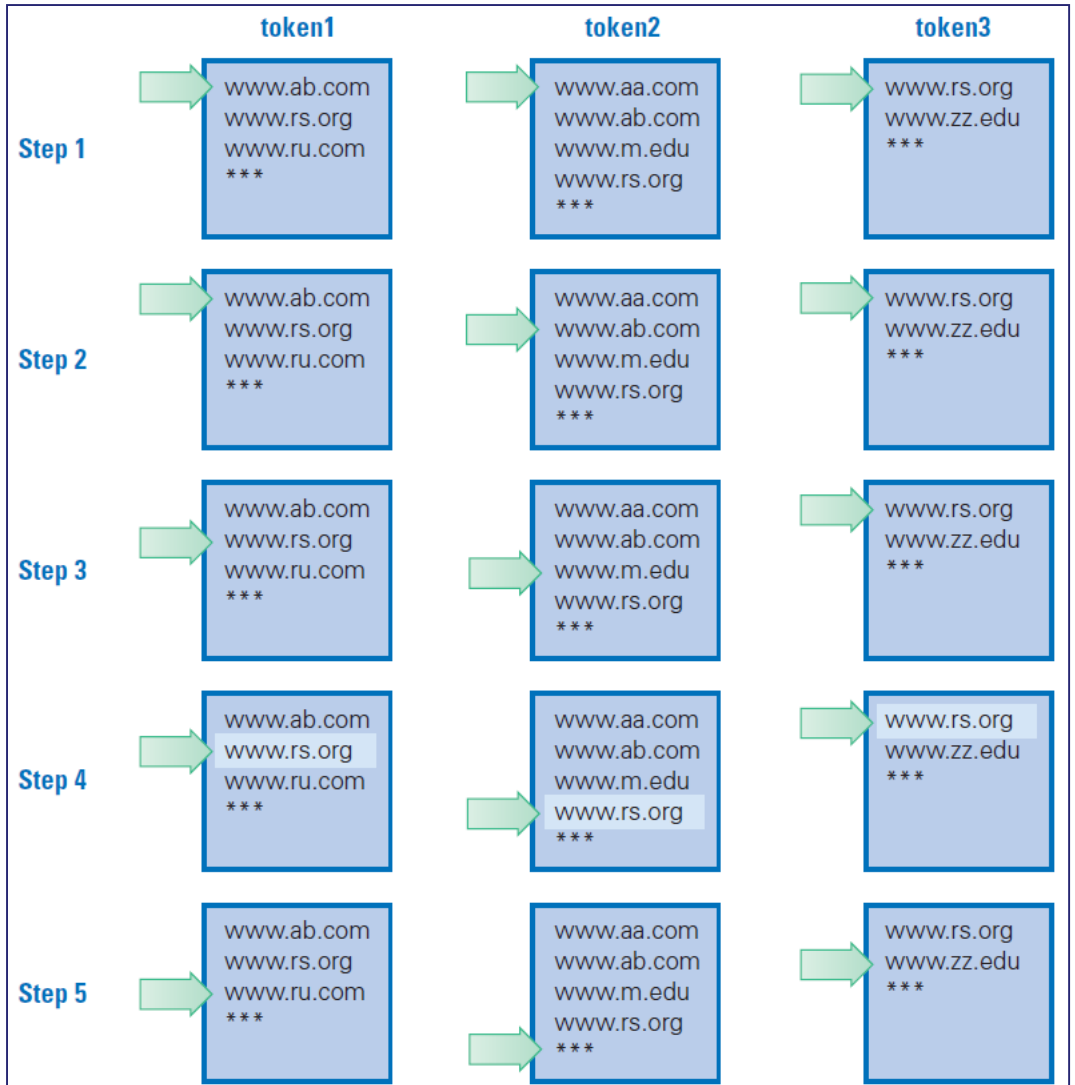


Figure 5.2 Illustrating the Intersecting Alphabetized Lists rules: In each step (row of boxes) one or more arrows advance; notice Step 3 where the arrow has advanced in two lists in the same step because the earliest URL is on both lists.

Power of an Indexed Search

- The computer:
  - takes the time to crawl the data (Web pages)
  - build an index first
  - find the index entries for each word
  - intersect the lists to find the information for an AND-query
- Search engines can look at billions of Web pages and return an answer in a quarter of a second

Descriptive Terms

- "***Hits***" on a page means the search term is "*associated*" with the page
- This does not mean the word is "*on*" the page
- Web page structure helps a lot to identify descriptive text
  - ***Title***—The <title> encloses a short phrase describing the whole page
  - ***Anchor text***—The highlighted link text, inside <a . . . > tags, describes the page it links to
  - ***Meta***—A <meta . . . > tag in the head section can hold a description of the page
  - ***Alt*** attributes—The <img . . . > tag has an alt attribute that gives a textual description
  - ***h1***—Text of top-level headers

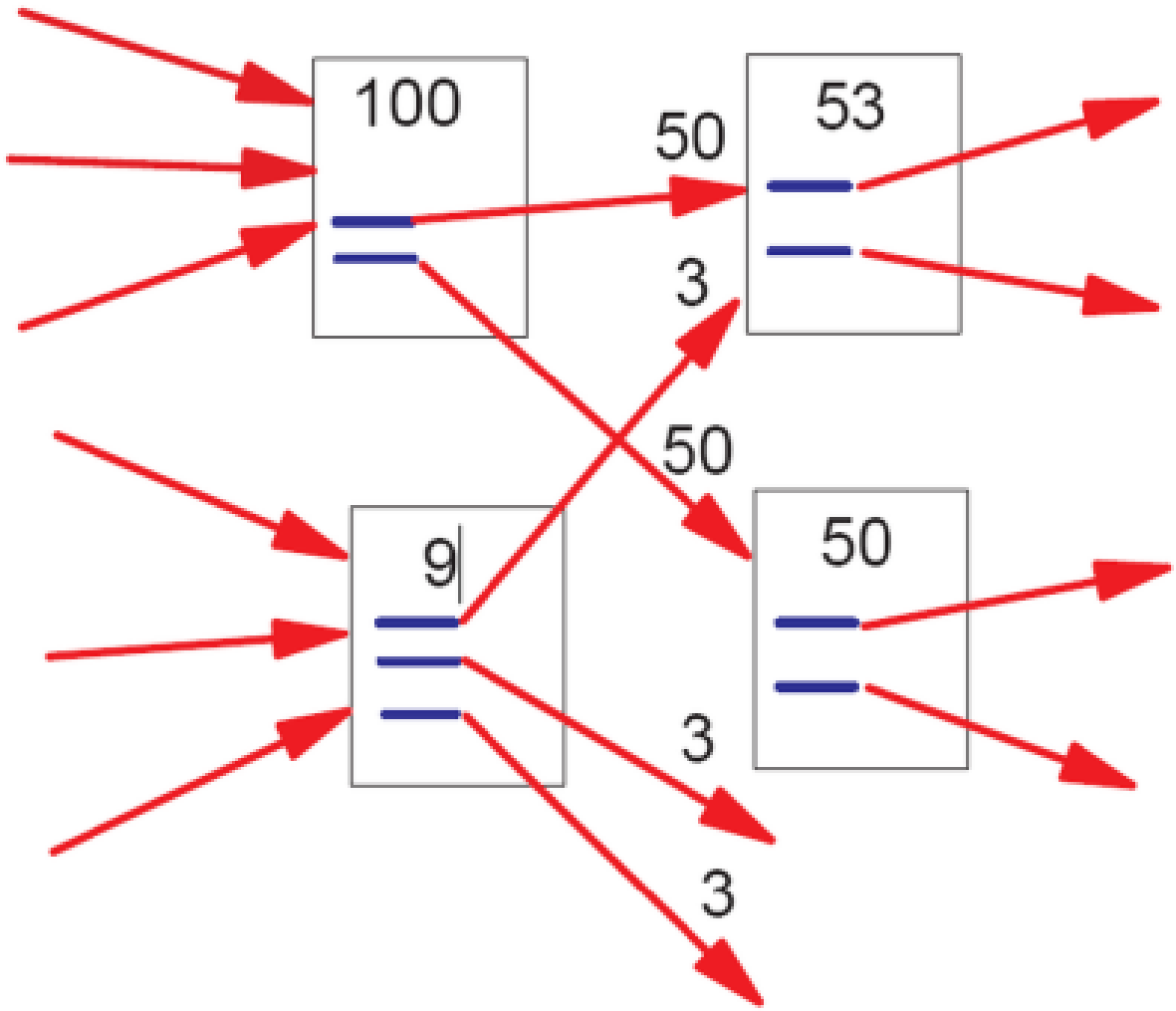
Page Rank

- Why, when the hit list is returned, the page you're looking for is often first on the hit list or in the top 10?
- The order in which hits are returned to a query is determined by a number called the ***PageRank***
- The higher the PageRank, the closer to the top of the list

Links to Other Pages

- Google pioneered page ranking as a way to determine which pages are likely to be most important
- PageRanking works like a voting system:
  - If page A links to page B, A's link adds to B's importance
- Pages that are linked-to by many pages have a higher page ranking and are assumed to be more important

Links to Other Pages

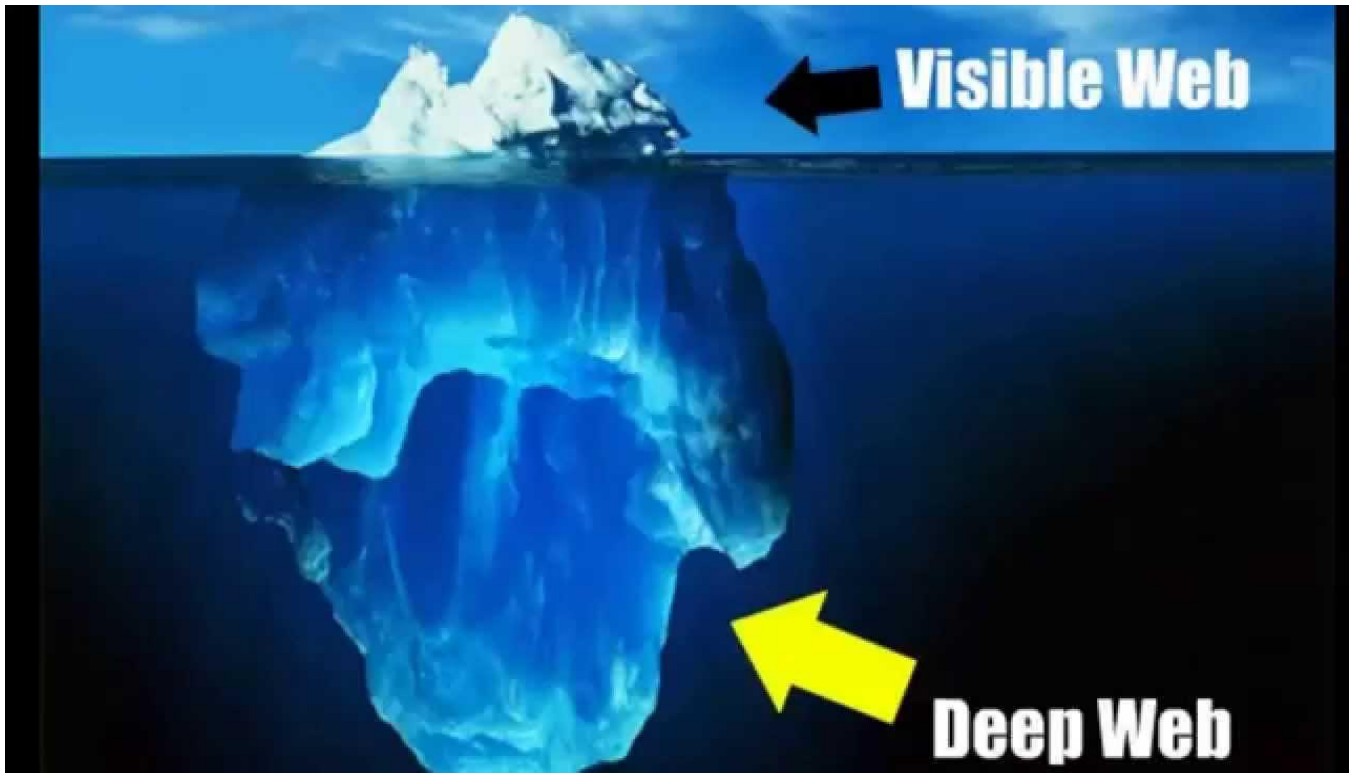


Links to Other Pages

- Links from pages with a high page ranking are also viewed as more important
- PageRank is computed by the crawler:
  - The crawler looks at page A
  - It notices the links to page B
  - It scores one for B
- Counting the number of links to a page is not sufficient

Links to Other Pages

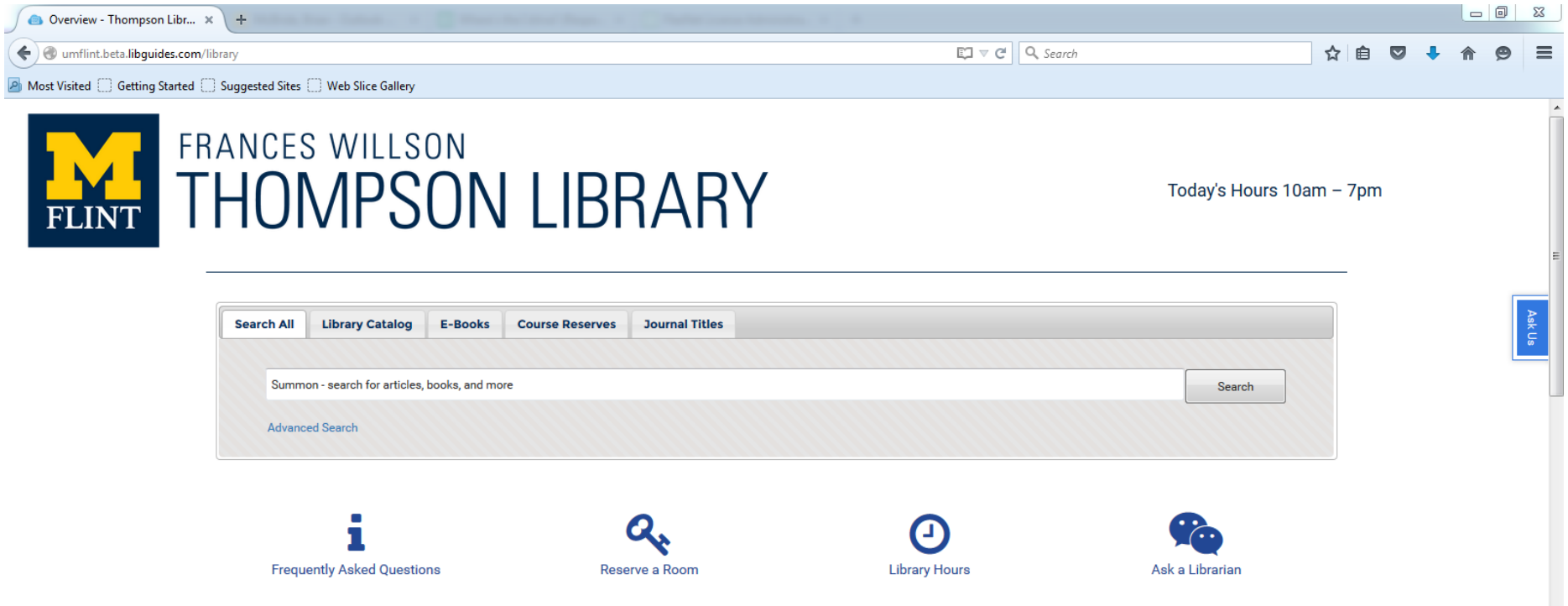
- After the crawling is completed, the PageRank computation is completed
- The query processor puts together the hit list
- The URLs are sorted by their page ranking, highest to lowest, and returned in that order


Visible Web
Deep Web


Visible Web
Web pages search engines can crawl and index
Invisible Web
Databases
Password-Protected Websites
Federal, State, County & City Public Records
Intranets
Message Boards
Website Archives
Forums
Classified Advertisements
Online Library Catalogs
Etc.

What does all this mean?

- Information is not free
- Not available through search engines
- Thompson Library


Overview - Thompson Libr...
umflint.beta.libguides.com/library
Search
Most Visited
Getting Started
Suggested Sites
Web Slice Gallery
FRANCES WILLSON
THOMPSON LIBRARY
FLINT
Today's Hours 10am – 7pm
Search All
Library Catalog
E-Books
Course Reserves
Journal Titles
Summon - search for articles, books, and more
Search
Advanced Search
Ask Us
Frequently Asked Questions
Reserve a Room
Library Hours
Ask a Librarian

•**Facets**: Most databases will allow you to filter your results using "facets." These are the options (normally located on the right or left hand side) that allow you to only display results that meet certain criteria such as peer-review, full-text, year of publication, etc. Using facets can really help to cut down the number of results you get from a search.

•**Bibliography Scanning:** When you find an article you like, look at the bibliography. There is a good chance that you'll find other articles that would be helpful to your research.

•**Find Alternate Keywords:** Often databases will list the keywords that are associated with the article you find. You can sometimes find this information in the abstract of the article as well.

•**Boolean Operators**:Use of Boolean operators (AND, OR, NOT) can sometimes be useful to help tie together or separate search terms. Use AND to only find articles that contain both of the keywords you're looking for, use OR to search for articles that use either one, and use NOT to eliminate a search term from your search.

•**Truncation and Wildcards:** Root words can have multiple endings Example: sun = suns, sunshine, sunny, sunlight. Likewise there are some words that are spelled differently, but mean the same thing. Example: color, colour