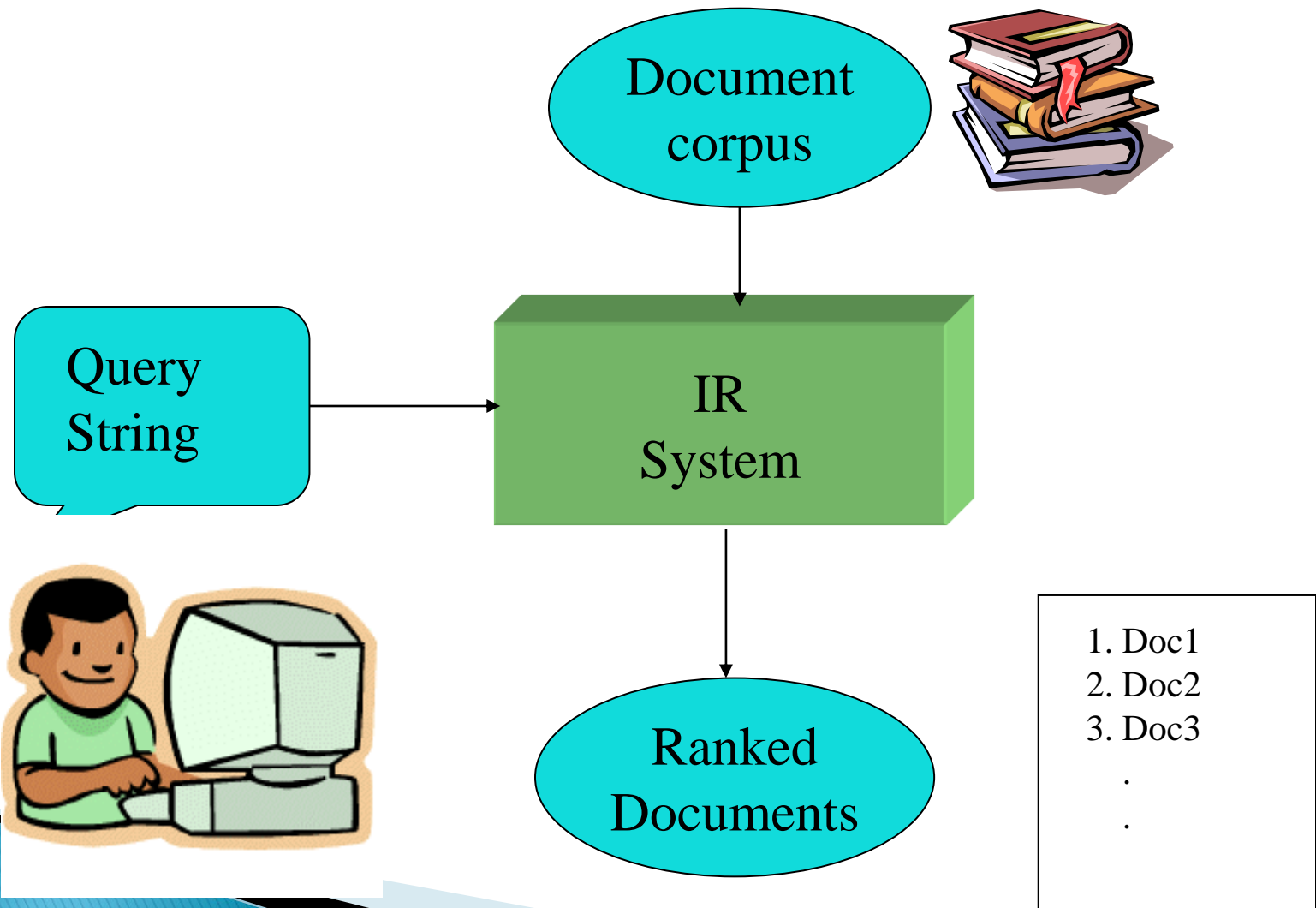
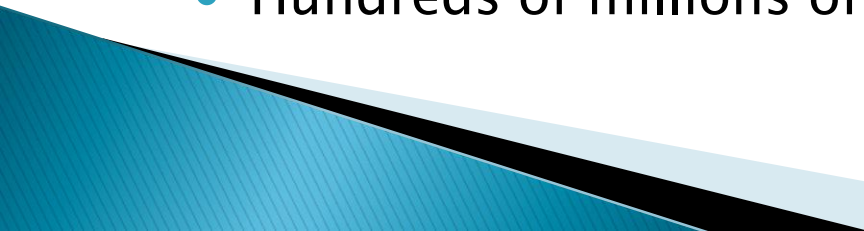


# *Web Search Engines*

# IR System Overview



# Search engine characteristics

- Unedited – anyone can enter
    - Quality issues
    - Spam
  - Varied information types
    - Phone book, brochures, catalogs, dissertations, news reports, weather, all in one place!
  - Different kinds of users
    - Online catalogs
      - scholars searching scholarly literature
    - Web
      - Every type of person with every type of goal
  - Scale
    - Hundreds of millions of searches/day; billions of docs
- 

# Web Search Queries

- Web search queries are SHORT
  - ~2.4 words on average (Aug 2000)
  - Has increased, was 1.7 (~1997)
- User Expectations
  - Many say “the first item shown should be what I want to see”!
  - This works if the user has the most popular/common notion in mind

# Directories vs. Search Engines

## ▶ Directories

- Hand-selected sites
- Search over the contents of the *descriptions* of the pages
- Organized in advance into categories

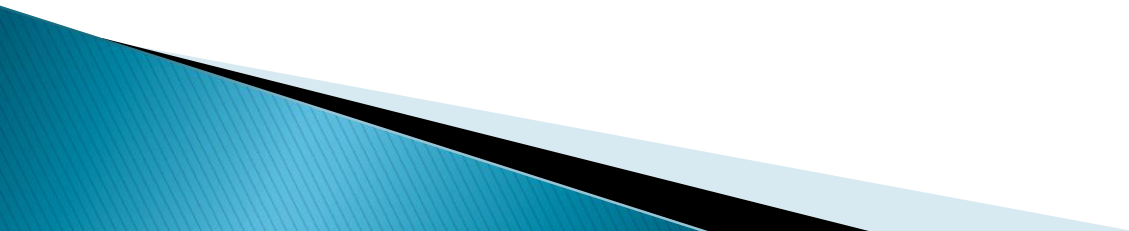
## ▶ Search Engines

- All pages in all sites
- Search over the contents of the *pages themselves*
- Organized after the query by relevance rankings or other scores

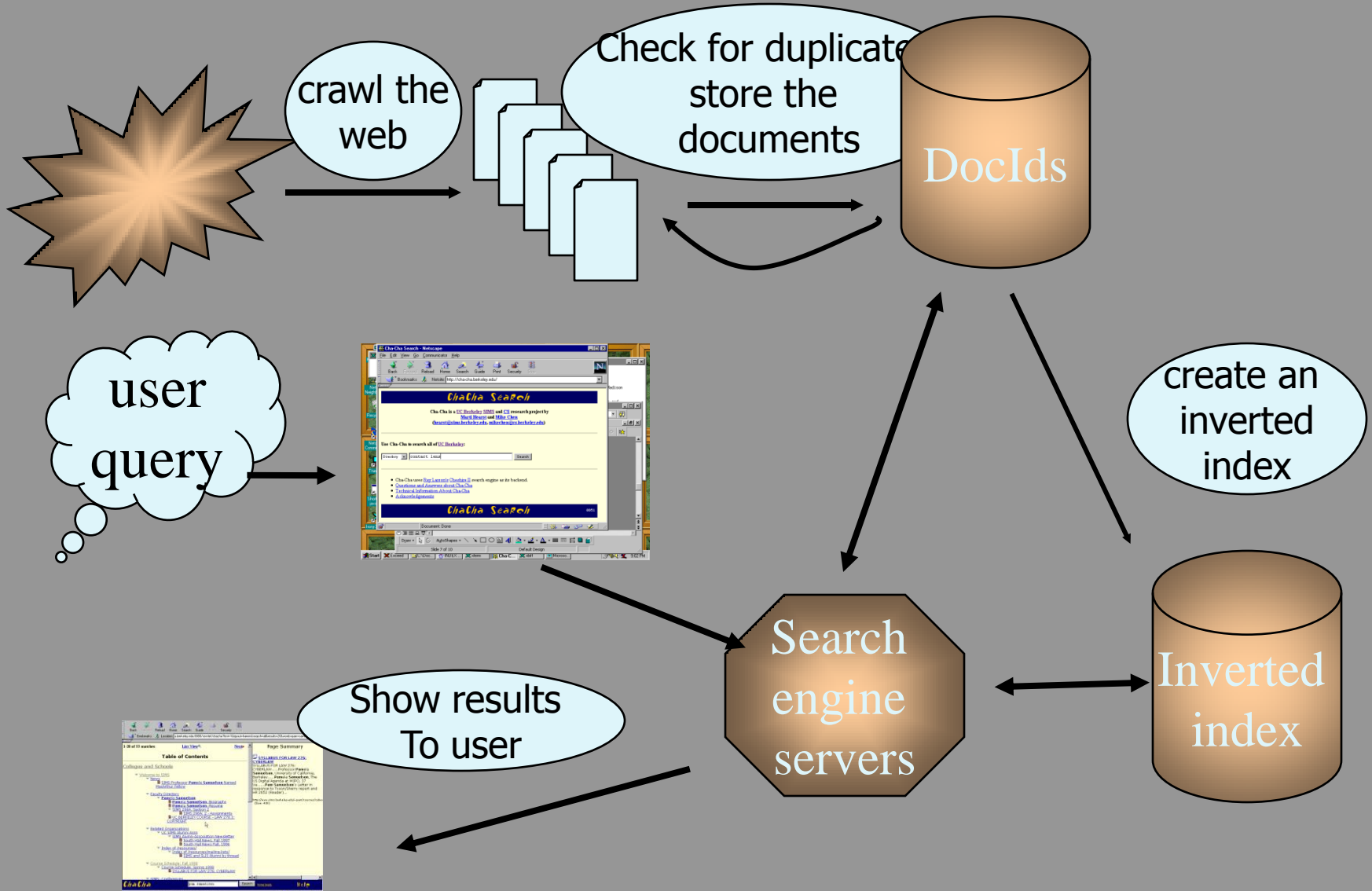
# Web Spam

- ▶ What are the types of Web spam?
  - Add extra terms to get a higher ranking
    - Repeat “cars” thousands of times
  - Add irrelevant terms to get more hits
    - Put a dictionary in the comments field
    - Put extra terms in the same color as the background of the web page
  - Add irrelevant terms to get different types of hits
    - Put “Madonna” in the title field in sites that are selling cars
  - Add irrelevant links to boost your link analysis ranking
- ▶ There is a constant “arms race” between web search companies and spammers

# Web Search Architecture

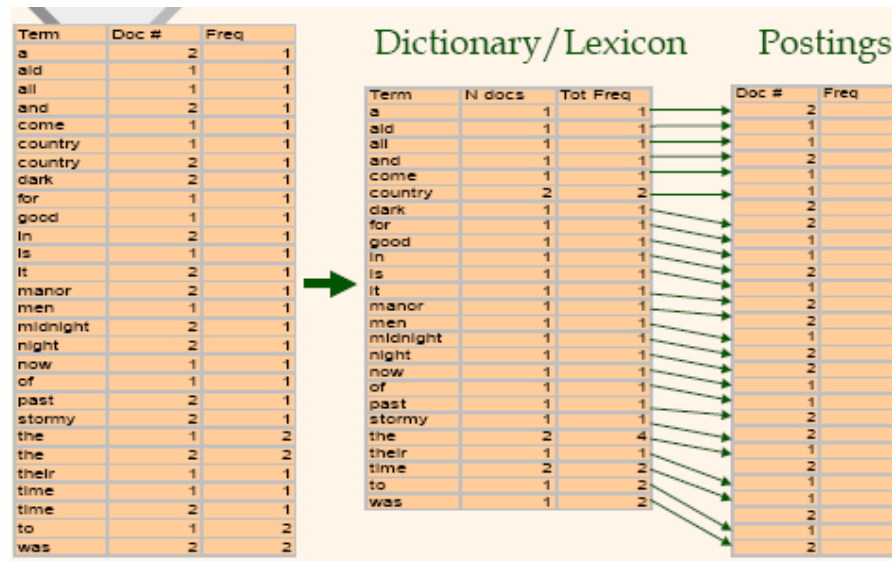


# Standard Web Search Engine Architecture





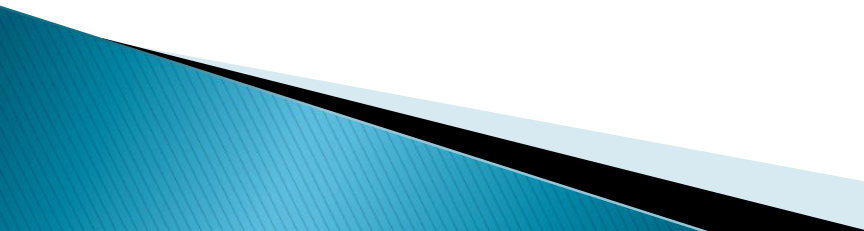
# How Inverted Files are Created?



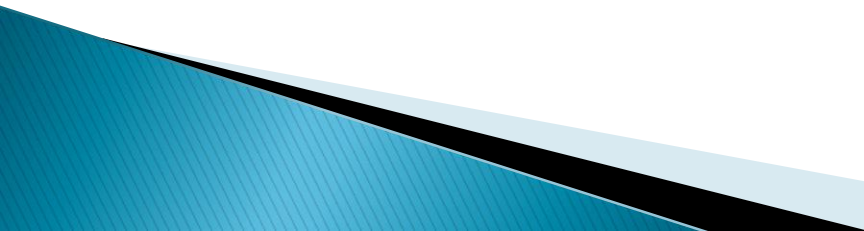
# *Inverted indexes*

- ▶ Permit fast search for individual terms
- ▶ For each term, you get a list consisting of:
  - document ID
  - frequency of term in doc (optional)
  - position of term in doc (optional)
- ▶ These lists can be used to solve Boolean queries:
  - country  $\rightarrow$  d1, d2
  - manor  $\rightarrow$  d2
  - country AND manor  $\rightarrow$  d2
- ▶ Also used for statistical ranking algorithms

## Inverted Indexes for Web Search Engines

- ▶ Inverted indexes are still used, even though the web is so huge
  - ▶ Some systems partition the indexes across different machines; each machine handles different parts of the data
  - ▶ Other systems duplicate the data across many machines; queries are distributed among the machines
  - ▶ Most do a combination of these
- 

# Web Crawlers

- ▶ How do the web search engines get all of the items they index?
  - ▶ Main idea:
    - Start with known sites
    - Record information for these sites
    - Follow the links from each site
    - Record information found at new sites
    - Repeat
- 

# Web Crawling Algorithm

- ▶ More precisely:
  - Put a set of known sites on a queue
  - Repeat the following until the queue is empty:
    - Take the first page off of the queue
    - If this page has not yet been processed:
      - Record the information found on this page
        - Positions of words, links going out, etc
      - Add each link on the current page to the queue
      - Record that this page has been processed
- ▶ Rule-of-thumb: 1 doc per minute per crawling server

# Web Crawling Issues

- ▶ Keep out signs
  - A file called `norobots.txt` tells the crawler which directories are off limits
- ▶ Freshness
  - Figure out which pages change often
  - Recrawl these often
- ▶ Duplicates, virtual hosts, etc
  - Convert page contents with a hash function
  - Compare new pages to the hash table
- ▶ Lots of problems
  - Server unavailable
  - Incorrect html
  - Missing links
  - Infinite loops

Web crawling is *difficult* to do robustly!

# Two Categories of Search Tools

- ▶ **Search Engines**
  - Individual search engine
  - Meta-search engine
- ▶ **Subject Directories**



# Individual Search Engines

The title "Individual Search Engines" is displayed in a large, grey, sans-serif font. Above the word "Search" are three logos: "YAHOO!" in purple, "altavista" in blue with a red swoosh, and "Google" in its multi-colored font.

- ▶ Individual search engines use computer programs called “spiders” to match key search words with the web pages that contain them.
  - Returns a large volume of results
  - Information is not filtered for validity, authenticity, or adult content
  - Results are returned in the form of links to sites that match terms used in the search



# Take a look at search engines

- ▶ [www.Search.yahoo.com](http://www.Search.yahoo.com)
- ▶ [www.Ask.com](http://www.Ask.com)
- ▶ [www.Google.com](http://www.Google.com)

## Family Friendly Search

- ▶ Meta-search engines send requests for information to several search engines simultaneously and compile the results.
  - Duplicates are eliminated, thus yielding fewer results
- ▶ **Note:** Google Custom Search Engine allows the user to select which search engines will be used

# Take a look at meta-search engines

- ▶ <http://www.googlecustomsearch.com/>
- ▶ <http://www.mamma.com>
- ▶ <http://www.surfwax.com/>

- ▶ Developed and maintained by humans (instead of software robots) to search broad subject categories and their descriptions.
- ▶ More reliable than search engines
- ▶ Provide broad categories of information that allow users to drill down and narrow search results

- ▶ Often used in research by government agencies, medical industries, and educational institutions
- ▶ May be referred to as research database or searchable database
- ▶ Results may include non-HTML formats, such as PowerPoints, PDF documents, script, and photographs

# Take a look at subject directories

- ▶ <http://www.google.com/dirhp>
- ▶ [www.libraryresearch.com](http://www.libraryresearch.com)
- ▶ <http://www.eric.ed.gov/>
- ▶ <http://dir.yahoo.com/>
- ▶ <http://infomine.ucr.edu>
- ▶ <http://www.lii.org/>
- ▶ <http://www.about.com/>

# Compare

- ▶ Suppose you are planning a vacation camping trip in one of the NC State Parks. Compare the results of each search tool by searching for the words NC State Parks.
- ▶ [www.google.com](http://www.google.com) – 55+ pages of results
- ▶ [www.dogpile.com](http://www.dogpile.com) – 3 pages of results
- ▶ <http://www.lii.org/> – 4 results

# Two Categories of Search Tools

- ▶ **Search Engines**
  - Individual search engine
  - Meta-search engine
- ▶ **Subject Directories**





# Individual Search Engines

The title "Individual Search Engines" is displayed in a large, grey, sans-serif font. Above the word "Search" are three logos: "YAHOO!" in purple, "altavista" in blue with a red swoosh, and "Google" in its multi-colored font.

- ▶ Individual search engines use computer programs called “spiders” to match key search words with the web pages that contain them.
  - Returns a large volume of results
  - Information is not filtered for validity, authenticity, or adult content
  - Results are returned in the form of links to sites that match terms used in the search

# Take a look at search engines

- ▶ [www.Search.yahoo.com](http://www.Search.yahoo.com)
- ▶ [www.Ask.com](http://www.Ask.com)
- ▶ [www.Google.com](http://www.Google.com)

## Family Friendly Search

- ▶ Meta-search engines send requests for information to several search engines simultaneously and compile the results.
  - Duplicates are eliminated, thus yielding fewer results
- ▶ **Note:** Google Custom Search Engine allows the user to select which search engines will be used

# Take a look at meta-search engines

- ▶ <http://www.googlecustomsearch.com/>
- ▶ <http://www.mamma.com>
- ▶ <http://www.surfwax.com/>

- ▶ Developed and maintained by humans (instead of software robots) to search broad subject categories and their descriptions.
- ▶ More reliable than search engines
- ▶ Provide broad categories of information that allow users to drill down and narrow search results

- ▶ Often used in research by government agencies, medical industries, and educational institutions
- ▶ May be referred to as research database or searchable database
- ▶ Results may include non-HTML formats, such as PowerPoints, PDF documents, script, and photographs

# Take a look at subject directories

- ▶ <http://www.google.com/dirhp>
- ▶ [www.libraryresearch.com](http://www.libraryresearch.com)
- ▶ <http://www.eric.ed.gov/>
- ▶ <http://dir.yahoo.com/>
- ▶ <http://infomine.ucr.edu>
- ▶ <http://www.lii.org/>
- ▶ <http://www.about.com/>

# Compare

- ▶ Suppose you are planning a vacation camping trip in one of the NC State Parks. Compare the results of each search tool by searching for the words NC State Parks.
- ▶ [www.google.com](http://www.google.com) – 55+ pages of results
- ▶ [www.dogpile.com](http://www.dogpile.com) – 3 pages of results
- ▶ <http://www.lii.org/> – 4 results