

Web Content

Extracted from <http://www.sc.edu/beaufort/library/pages/bones/bones.shtml>

Search Engines

- Search engines are huge databases of web files that have been assembled automatically.

- Search engines compile their databases by employing software agents, called "spiders" or "robots" ("bots"), to crawl through web space, from link to link, identifying and perusing pages.
- These agents typically index most of the words on the publicly available pages at the site.

Searching

- In using a search engine, you're asking the engine to scan its index of sites and match your keywords and phrases with those in the texts of documents within the engine's database.

Currency

- **Remember: when you are using a search engine, you are NOT searching the entire web as it exists at that moment. You are actually searching a portion of the web, captured in the index created at an earlier date.**
- Spiders regularly return to the web pages they index to look for changes. When changes occur, the index is updated to reflect the new information.

Note

- Web site administrators can prevent spiders from navigating their site. This implies an index can only access those sites who have permitted that access.

Metasearch Engines

- Metasearch engines do not compile their own searchable databases. Instead, they search the databases of individual search engines simultaneously.
- Metasearchers provide a quick way of finding out which engines are retrieving the best results for you in your search.

Subject Directories

- Subject directories, unlike search engines, are created and maintained by human editors, not electronic spiders or robots. The editors review and select sites for inclusion in their directories on the basis of previously determined selection criteria. The resources they list are usually annotated. Directories tend to be smaller than search engine databases, typically indexing only the home page or top level pages of a site.

Examples

- Beaucoup
- LookSmart
- Open Directory Project

Gateways

- There are two kinds of gateways: library gateways and portals.
 - Library gateways are collections of databases and informational sites, arranged by subject, that have been assembled, reviewed and recommended by specialists, usually librarians.
 - Portals are directories that have been created or taken over by commercial interests reconfigured to act as gateways to the web. These portal sites also offer additional services such as email, current news, stock quotes, travel information and maps.
 - Vortals, or vertical portals, are subject-specific directories, as opposed to the broader, more generalized smorgasbord of subjects and other links commonly found in portals.

Examples

- Gateways and Vortals

Evaluating Web Content

- Where is it hosted?
- Who is responsible for content?
- Who is the author?
- Is the material current?
- Who sponsors the page?
- Is the content reviewed or refereed?

FYI

Citing Web Content

- APA
- MLA
- Turabian and Chicago Style Guides

Rationale

Keeping Up

- Webopedia
 - Terms and definitions
- Wikipedia
 - Free Encyclopedia

Perspective

- Searching currently relies on the extraction of linguistic information, words, from files. Crawlers work their way through documents extracting words.
- Textual information about images can be easily searched using existing technology, but requires humans to personally describe every image in the database.

Content-Based Search

- "Content-based" means that the search will analyze the actual contents of the file.
- Images, video, audio, 3-D Models

Properties

- The term 'content' refers to information that can be derived from the object itself. Without the ability to examine content, searches must rely on metadata such as captions or keywords, which may be laborious or expensive to produce.

Feature Retrieval

- Current content-based image retrieval systems make use of lower-level features like texture, color, and shape, although some systems take advantage of very common higher-level features like faces (see facial recognition system).

Semantic Retrieval

- The ideal system from a user perspective would involve what is referred to as *semantic* retrieval, where the user makes a request like "find pictures of dogs" or even "find pictures of Abraham Lincoln".
- This type of open-ended task is very difficult for computers to perform
 - pictures of Chihuahuas and Great Danes look very different, and Lincoln may not always be facing the camera or in the same pose.

Current Efforts

- Much effort in image retrieval issues.
- Video and audio are harder problems.
