# Web Technologies and Applications

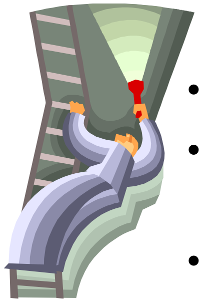
Winter 2001

## CMPUT 499: Search Engines

Dr. Osmar R. Zaïane

University of Alberta

## Everyday Activity



- We use search engines whenever we look for resources on the Internet
- How do these search engines work?
- How come they give different results and the results?
- The results are often very disappointing. Why aren't we satisfied?

## Course Content

- Introduction
- Internet and WWW
- Protocols
- HTML and beyond
- Animation & WWW
- Java Script
- Dynamic Pages
- Perl Intro.
- Java Applets
- Databases & WWW
- SGML / XML
- Managing servers
- **Search Engines**
- Web Mining
- CORBA
- Security Issues
- Selected Topics
- Projects

## Objectives of Lecture 13

### Search Engine

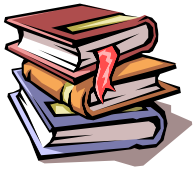
- Get a a general idea about the technologies behind search engines
- Get acquainted with inverted indexes
- Discuss ranking issues

# Outline of Lecture 13

- Inverted Indexes and Information Retrieval
- Anatomy of a Search Engine
- Web Crawler
- Ranking Results

# Information Retrieval

- Find resources (documents) that contain a certain list of keywords

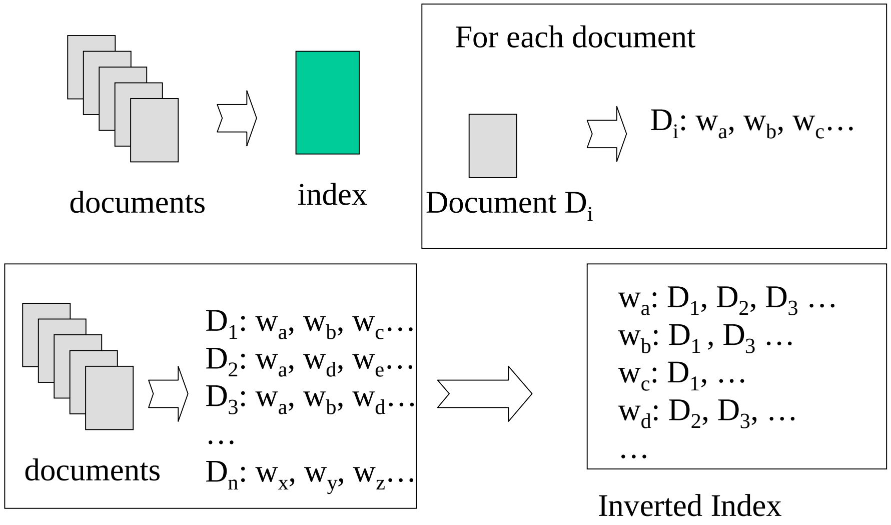


Find the pages where the phrase “alpha beta” occurs.

Searching sequentially is too expensive.

You would need an index to directly find the pages.

# Creating an Index



documents

index

For each document

$D_i$: $w_a$, $w_b$, $w_c$…

Document $D_i$

documents

$D_1$: $w_a$, $w_b$, $w_c$…
$D_2$: $w_a$, $w_d$, $w_e$…
$D_3$: $w_a$, $w_b$, $w_d$…
…
$D_n$: $w_x$, $w_y$, $w_z$…

$w_a$: $D_1$, $D_2$, $D_3$ …
$w_b$: $D_1$ , $D_3$ …
$w_c$: $D_1$, …
$w_d$: $D_2$, $D_3$, …
…

Inverted Index

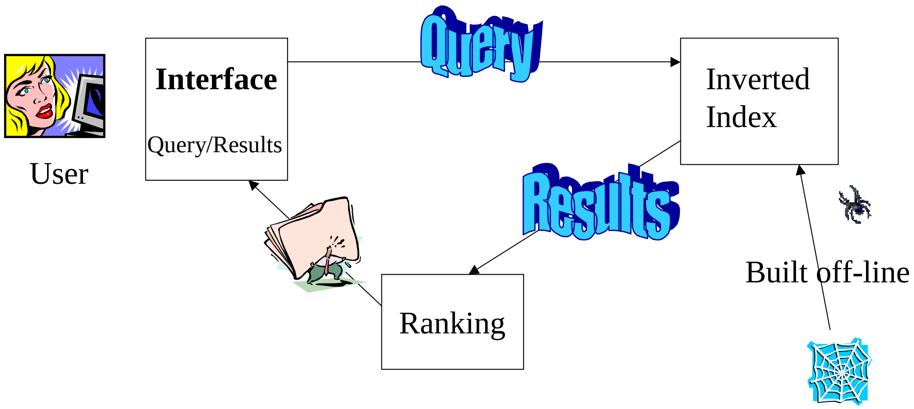
# Outline of Lecture 13

- Inverted Indexes and Information Retrieval
- Anatomy of a Search Engine
- Web Crawler
- Ranking Results

# Search Engine Components

- A Search Engine has an interface to enter queries
- A search engine has access to an inverted index already built
- A search engine ranks the results found in the index

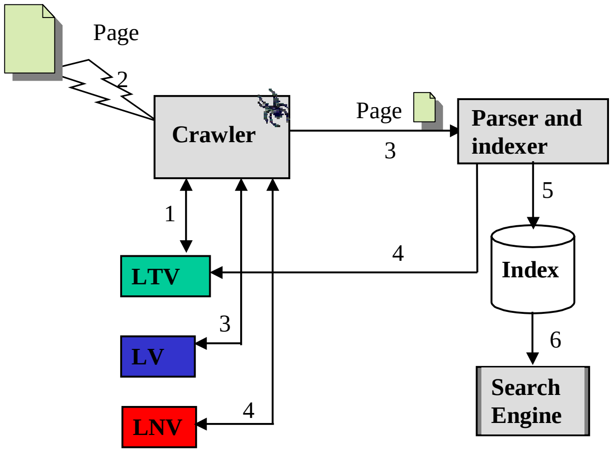
# A Search Engine Blocs



# Outline of Lecture 13

- Inverted Indexes and Information Retrieval
- Anatomy of a Search Engine
- Web Crawler
- Ranking Results

# Search Engine General Architecture



# Search Engines are not Enough

- Most of the knowledge in the World-Wide Web is buried inside documents.
- Search engines (and crawlers) barely scratch the surface of this knowledge by extracting keywords from web pages.
- There is text mining, text summarization, natural language statistical analysis, etc., but not the scope of this course.

# Outline of Lecture 13

- Inverted Indexes and Information Retrieval
- Anatomy of a Search Engine
- Web Crawler
- Ranking Results

# Relevancy Ranking

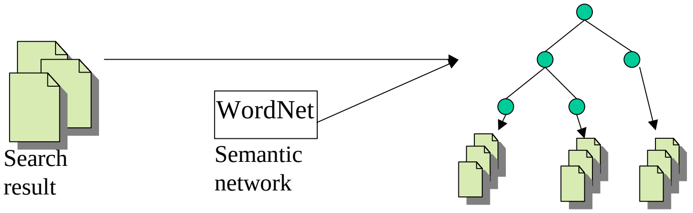
- Some search engine claim to have indexed about one billion documents
- Each search can yield a very large list of “supposedly relevant” documents
- Sifting through thousands of results is tedious and not necessary
- It is extremely important to rank the results since most users will look mainly at the 10 to 20 first documents.

# How do we Rank?

- Each Search Engine uses a different ranking function. Usually these ranking functions are not disclosed
- Parameters used in ranking:
  - Frequency of words
  - Location of words
  - Entirety of query
  - Size of document
  - Age of document
  - Existence in directory
  - Inward and outward Links
  - Metadata
  - Domain
  - And $$$$

# Ontology for Search Results

- There are still too many results in typical search engine responses.
- Reorganize results using a semantic hierarchy (Zaïane et al. 2001).



Search result

WordNet

Semantic network