UNIVERSITY of DELAWARE

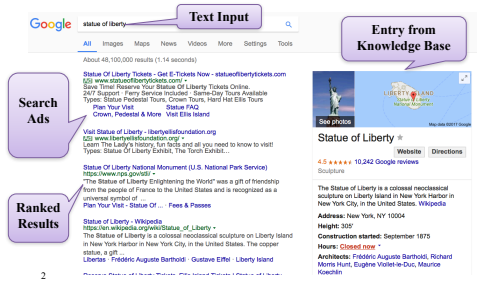
# Overview of Search

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE

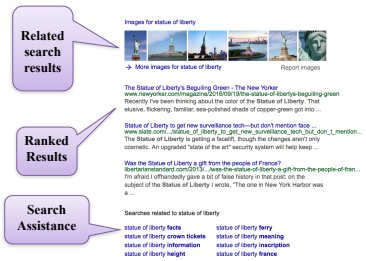
## Layout of Search Results Pages






UNIVERSITY of DELAWARE

## Layout of Search Results Pages (Cont.)






UNIVERSITY of DELAWARE

## How well do you know about Web search engines?

- How many queries are submitted per day to a major search engine?
- How much data is in the index of a major search engine?

  A petabyte is 10 million gigabytes
- How many computers are used for a Web search engine?

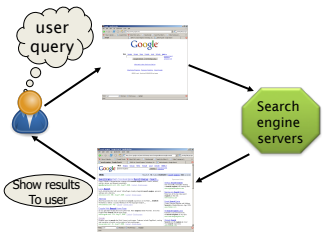
UNIVERSITY of DELAWARE

## What is Text Retrieval (Search)?

- There exists a collection of text documents
- User gives a query to express the information need
- A retrieval system returns relevant documents to users

UNIVERSITY of DELAWARE

## Web Search Engine Architecture

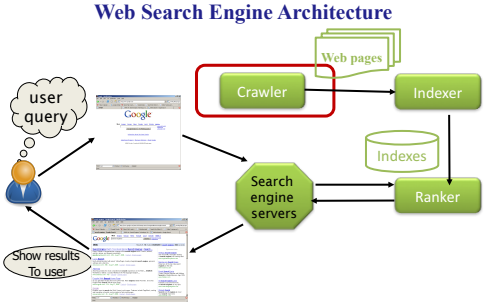


## How Web Search Engines Work

- **Crawler:** Download web pages from the Internet
- **Indexer:** Organize the contents of the pages in a way that allows efficient retrieval
- **Ranker:** Take a query as the input, and identify relevant web pages and return the results.



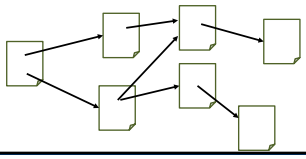
## Web Search Engine Architecture






## Crawler - Basics

- Start with a list of seed pages
- Keep a list of URLs that have been visited, and those to be visited.
- When downloading a page, look at the hyperlinks on the page and add the links in that page to the list to be crawled.



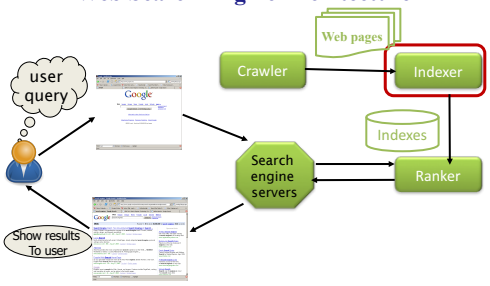


## Crawler - Challenges

- Servers can be down or slow
- Hyperlinks can get the crawler into cycles
- Some websites have junk in the web pages
- Now many pages have dynamic content
  - So called "hidden" or "deep" Web
  - E.g., course search web site
- Freshness: need to keep checking changes
- The web is HUGE



## Web Search Engine Architecture






## Indexing

- Goal:
  - store the information for efficient retrieval
- Basic steps:
  - Make a dictionary of all the words
  - For each word, list all the documents it occurs in

## Inverted Index Example

Doc 1
... news about

Doc 2
... news about organic food campaign ...

Doc 3
... news of presidential campaign ...
... presidential candidate ...

Dictionary (or lexicon)

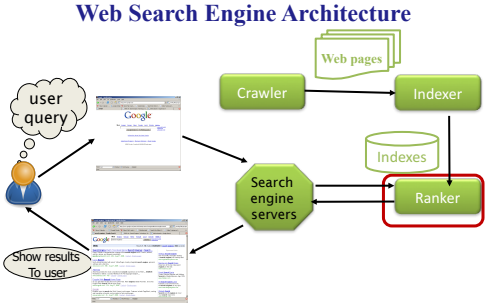
| Term | # docs | Total freq |
|---|---|---|
| news | 3 | 3 |
| campaign | 2 | 2 |
| presidential | 1 | 2 |
| food | 1 | 1 |
| ... | ... | ... |

Postings

| Doc ID | Freq | Position |
|---|---|---|
| 1 | 1 | p1 |
| 2 | 1 | p2 |
| 3 | 1 | p3 |
| 2 | 1 | p4 |
| 3 | 1 | p5 |
| 3 | 2 | p6, p7 |
| 2 | 1 | p8 |
| ... | ... | |
| ... | ... | |



## Web Search Engine Architecture






## Search - Formal Formulation

- *Vocabulary*: $V=\{w_1, w_2, \ldots, w_N\}$
- *Query*: $q = q_1,\ldots,q_m$, where $q_j \in V$
- *Document*: $d_i = d_{i1},\ldots,d_{il}$, where $d_{ij} \in V$
- *Collection*: $C= \{d_1, \ldots, d_k\}$
- *Set of relevant documents* $R(q) \subseteq C$
  - It is unknown and user dependent.
  - Query is a "hint" at which documents should be in R(q).
- *Task*: find $R'(q) \subseteq C$, which is an approximate of R(q)



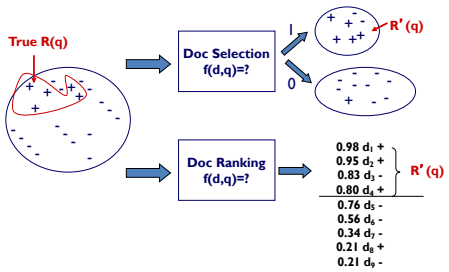
## Search - Computing R'(q)

- Strategy 1: Document selection
  - Implement a binary classifier to classify a document as either relevant or non-relevant with respect to a query
  - $R(q)=\{d \in C | f(d,q)=1\}$, where $f(d,q) \in \{0,1\}$ is an indicator function or classifier.
- Strategy 2: Document ranking
  - Implement a ranking function and rank all the documents in descending values of this ranking function.
  - $R(q) = \{d \in C | f(d,q) > \theta\}$, where $f(d,q) \in \Re$ is a relevance measure function; $\theta$ is a score threshold.

absolute relevance vs. relative relevance



## Search - Document Selection vs. Ranking






## Limitations of Document Selection

- The binary classifier is unlikely accurate
  - "Over-constrained" query: no relevant documents found
  - "Under-constrained" query: over delivery
  - It is extremely hard to find the right position between these two extremes.

UNIVERSITY of DELAWARE UNIVERSITY of DE

## Ranking is generally preferred.

- Relevance is a matter of degree
- A user can stop browsing anywhere, so the boundary is controlled by the user
  - High recall users would view more items
  - High precision users would view only a few
- Theoretical justification: Probability Ranking Principle
  - The strategy of ranking is optimal under the following assumptions:
    - The utility o fa document to a user is independent of the utility of any other documents
    - A user will browse the results sequentially.

**Relevance Modeling is an important topic in IR**