CS6120 Intelligent Media Systems

Lecture 3: Web Search

A. O'Riordan, 2008

Parts based on notes by B. Smyth

Overview

- Information Retrieval & Web Search
  - The Anatomy of a Search Engine
  - Search Engine Evaluation
  - Meta-Search
- Search Using Links and Connectivity
  - Hubs & Authorities, PageRank
- Search Engine Market
- Contextualizing Search
  - Custom Search
  - Explicit v. Implicit Context

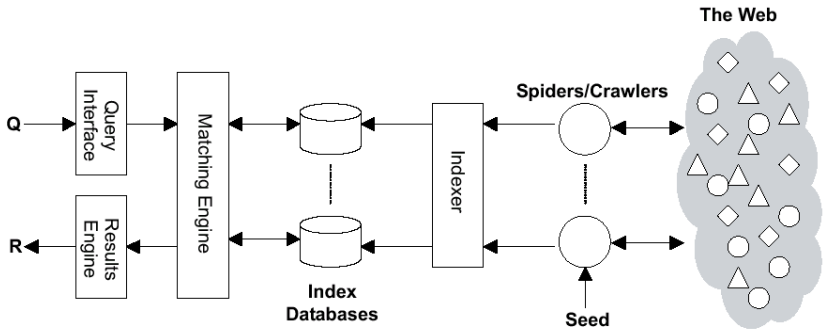
Information Retrieval

- Traditional libraries are storehouses of books, and the task of cataloguing and indexing is carried out manually by librarians
- The concept of automated information retrieval (IR) was developed in the 1950s, became an active research area in the 1960s, and achieved mass usage in the 1990s with the Web.
- Systems deal primarily with textual information

The Retrieval Process

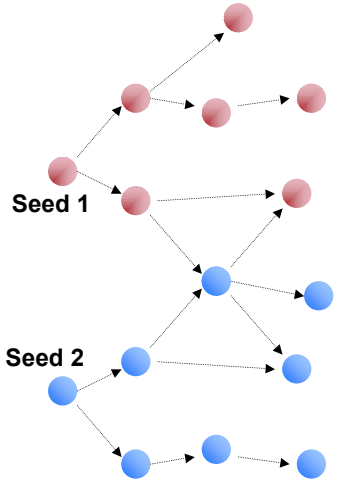
- In order to obtain information you formulate an information need in the form of a query that is issued to a document collection.
- Relevant documents (pages), possibly ranked, are then returned to the user

The Anatomy of a Search Engine



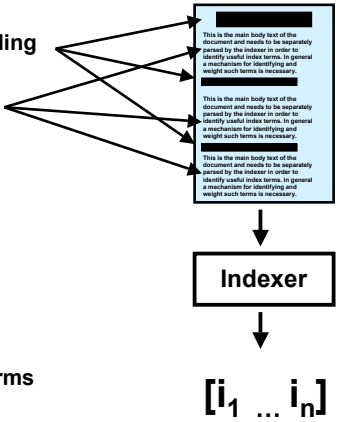
The Spider/Crawler

- Web spiders harvest Web page content by traversing the Web graph starting at some seed page
- The graph is traversed by following outward links
- Issues
  - Order of Traversal (Breath-First v. Depth-First)
  - Security (eg. robots.txt)
  - Topic-specific spiders
  - Centralized v. Distributed crawling



The Indexer

- Sources of Index Terms
  - Meta-Tags, Title and Heading Terms
  - Content Terms, Link Text
- Representation
  - Bag of Words
  - Vector Space Model
- Automatic Indexing
  - Selection of index terms
  - Relative importance of terms



Representation

- Bag of Words
  - Documents and queries represented as unordered lists of words … grammatical information lost.
  - Stop-word removal (e.g. “a”, “the”, “and”, etc.)
  - Stemming (e.g. “compelling” ⇒ “compel”)
  - Query-Document similarity ⇒ term overlap metric

> Scientists have found compelling new evidence of possible ancient microscopic life on Mars, derived from magnetic crystals in a meteorite that fell to Earth from the red planet, NASA announced on Monday.

> ~~a~~, ancient, announced, compelling, crystals, derived, earth, evidence, fell, found, from (2×), have, ~~in~~, life, magnetic, mars, meteorite, microscopic, monday, nasa, new, ~~of~~, ~~on~~ (2×), planet, possible, red, scientists, ~~that~~, ~~the~~, ~~to~~

Representation

- Vector Space Model
  - Documents, terms and queries treated as vectors.
  - SMART Project (1960s).
- Documents, Terms and Queries
  - Documents are vectors of terms
  - Terms are vectors of documents
  - Queries are vectors of terms

	t_1	t_2	t_3	t_4	...
doc_1	1	0	0	1	...
doc_2	0	1	0	1	...
doc_3	0	0	1	1	...
doc_4	1	1	1	0	...
⋮	⋮	⋮	⋮	⋮	

	t_1	t_2	t_3	t_4	t_5
q	1	1	0	0	1

Automatic Indexing

- TFIDF Weighting = Term frequency x Inverse document frequency
- Terms that occur *frequently* in a given document are likely to be important index terms.
- Terms that occur in the target document but that are *rare* within the collection of documents, are likely to be important.
- The TFIDF weight of a term i in document j is a combination of these factors.

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Matching Engine

- The matching engine is concerned with determining the similarity between a query and a document …
- Similarity (vector space model)
  - Similarity between query and document calculated as the inner product of their vector representations.

$$q \bullet d = \sum_{i=1}^{n} (q_i \cdot d_i)$$

- Eg …

	t_1	t_2	t_3	t_4	…
doc_1	1	0	0	1	…
doc_2	0	1	0	1	…
doc_3	0	0	1	1	…
doc_4	1	1	1	0	…
⋮	⋮	⋮	⋮	⋮	⋮

	t_1	t_2	t_3	t_4	t_5
q	1	1	0	0	1

	$q \bullet d$
doc_1	2
doc_2	2
doc_3	0
doc_4	3

Matching Engine

- **Similarity Problem**
  - Similarity is biased towards longer documents and queries so need to adjust the similarity score to factor in document length and query length by dividing by norms
  - Cosine similarity of query q and document d is

$$\text{sim}(q,d) = \frac{\sum_{i=1}^{n} (q_i \cdot d_i)}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

The Query Engine

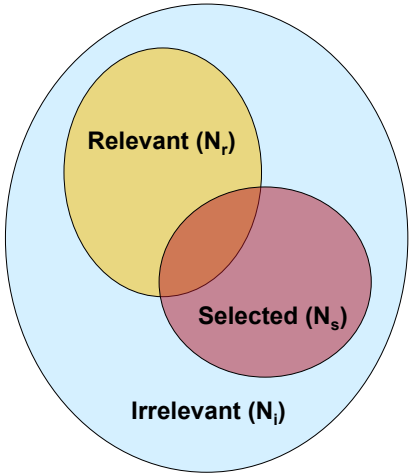
- **Query Representation**
  - Query terms plus logical operators (AND, OR, NOT) plus search-specific operators (e.g. NEAR)
  - Different search engines often utilise a range of different query features – non-standard support
  - Example: In Google Search the OR is implicit
- **Query Transformation**
  - Given that Web queries are often vague …
  - … active research focused on the elaboration and transformation of queries, especially vague queries, into richer query formats.

The Results Engine

- **Formatting**
  - Responsible for the formatting and presentation of the results
  - Result splitting, duplicate removal etc
- **Ranking**
  - Ranking is a critical function in order to present results in order of their likely relevance
  - Ranking metrics (content-based, connectivity, etc.)

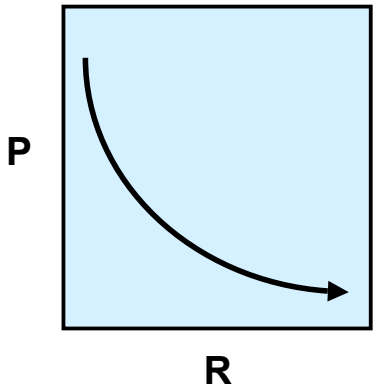
Search Engine Evaluation

- How doe we evaluate the performance of a search engine?
- How do we compare the performance of two systems?
- Precision (P) - Percentage of selected (retrieved) documents that are relevant
- $P = |N_r \cap N_s| / |N_s|$
- Recall (R) - Percentage of relevant documents that have been selected
- $R = |N_r \cap N_s| / |N_r|$



Precision v. Recall

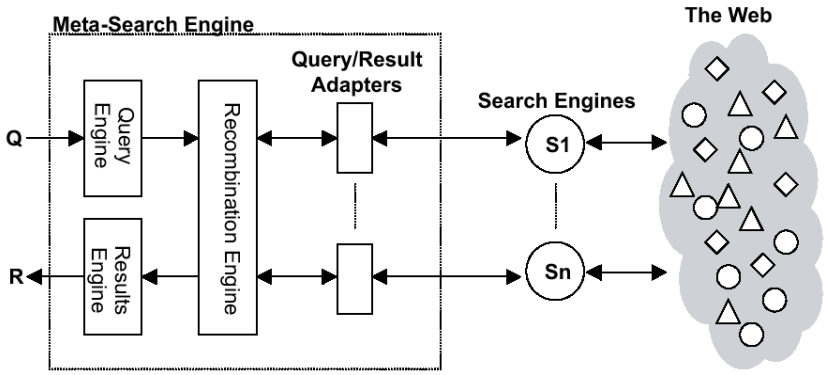
- What happens if we increase the number of selected documents?
- Inverse relationship between precision and recall
  - Increase precision ⇒ decrease recall
- Precision v. Recall – which is more important?
  - Depends on application …
  - Web search v. legal search?



Meta-Search

- There are lots of search engines out there …
- … so why not leverage the work of multiple search engines at a time?
- Meta-Search Engines
  - Query submitted to many underlying search engines
  - Results combined and (usually) reformatted into a single result list
  - Examples: mamma, ixquick, dogpile
- Technical Issues
  - Query Transformation
  - Result Combination & Ranking
  - Search Engine Selection

Meta-Search

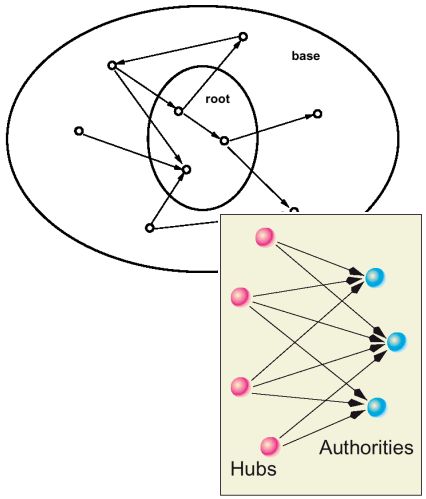


Beyond Content-Based Search

- **The Content Pitfalls**
  - **Poor quality Web pages with the correct terms**
  - **High-quality pages without the correct terms**
  - **Content-based techniques are easy to corrupt – repeated use of key search terms**
- **What about using connectivity information?**
  - **Consider a page P that is linked to by lots of other pages …**

Hubs and Authorities

- **3-Phase Search**
  - **1 – Select documents that are deemed relevant to a given query (root set)**
  - **2 - Expand root set into base set by following in and out links**
  - **3 – Compute Hubs and Authorities**
- **Hub & Authorities**
  - **A good Authority is linked to by many Hubs and vice versa**



Calculating Hubs & Authorities

- **In summary …**
  - **Each page is associated with an authority weight x<p> and a hub weight y <p>**
  - **An iterative algorithm updates the weights of each page in the base set by the use of two update operations.**
  - **One updates the authority weights and other operation updates the hub weights**

$$x^{\langle p \rangle} \leftarrow \sum_{q:(q,p)\in E} y^{\langle q \rangle} \qquad y^{\langle p \rangle} \leftarrow \sum_{q:(p,q)\in E} x^{\langle q \rangle}$$

  - **The algorithm terminates when an equilibrium is reached**

PageRank & Google

- Developed at Google in late 1990s by Brin and Page
- The PageRank of a page u, R(u), depends on
  - B_u = set of pages that point to u (in-links of u)
  - F_u = set of pages that u points to (out-links of u)
- Specifically, R(u) is proportional to the normalised sum of the PageRanks of u's in-links
  - $N_v = |F_u|$

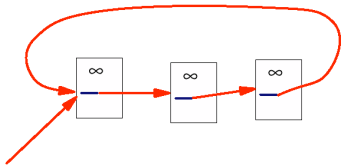
Iterative computation starting from random PageRanks until convergence occurs

Simplified PageRank

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Rank Sinks & Random Surfers

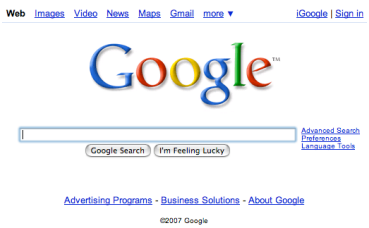
- The above is the naïve version of PageRank and it brings with it a number of problems and issues.
  - Rank sinks (loops of pages that accumulate rank without distributing it)
  - Modified PageRank, R'(u) and the Random Surfer Model – E(u) is the probability of a random jump to another Web page



$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

Major Search Engines

- Google Search (1998), Google Inc.
- Yahoo! Search (2004), Yahoo! Inc.
- Live Search (2006), Microsoft Corp.
- Ask.com (2006), IAC Search & Media
- Baidu 百度 (2000), Baidu.com, Inc.



- Early ones – Webcrawler (1994), Infoseek (1994), Lycos (2004)
- Many Takeovers/Acquisitions - Teoma acquired by IAC in 2001; Inktomi acquired by Yahoo! in 2002; goto/Overture acquired by Yahoo! in 2003

Search Engines Market Share

U.S. Core Search Engines by Search Share, September 2007

Core Search Entity	% Market
Google	57
Yahoo	23.7
Microsoft	10.3
Ask	4.7
Time Warner	4.3
Total core search	100.0

Notes:
1. Locations include home, work, and universities.
2. Data are based on the five major search engines, including partner searches and cross-channel searches.
Source: comScore, 2007

Towards Context in Search

- Search engines are poor at responding to vague/ambiguous queries
- Introducing context into Web search can increase relevancy/eliminate ambiguity
- Some traditional context-analysis techniques require too much user intervention

Query: "Michael Jordan"

Context? (Academic, Sport, Business?)


Professor


Basketball star


EDS chairman

Context in Web Search

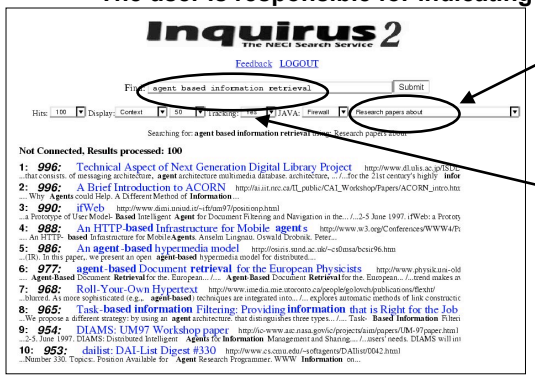
- Vague queries are commonplace in Web search
  - 2-3 query terms per search on average
- Context information is all too often absent
  - As a result search results tend to cover a variety of conflicting topics
  - Users may miss relevant results altogether
- Context Sensitive Search
  - Why not introduce context into search?
  - How? Explicit context v. Implicit Context

Custom Search

- uses knowledge of the user's current activities and goals to determine which information sources are best or direct search
- Idea around for a while
  - e.g. Towards Context-Based Search Engine Selection. David B. Leake and Ryan Scherle. In *Proc. of the 2001 Conf. on Intelligent User Interfaces (IUI2001)*
- Implementations
  - e.g. Google Custom Search Example
  - e.g. Yahoo!'s Y!C

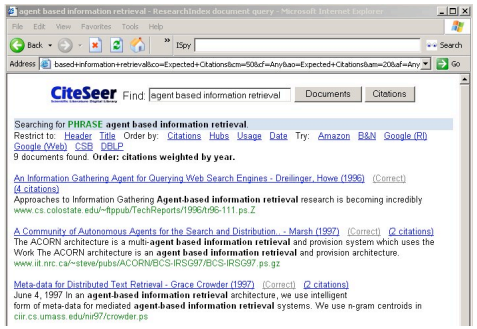
Explicit Context

- Inquirus (NEC) is a Web search engine that is designed to introduce explicit context terms into a search
  - The user is responsible for indicating the context terms



Explicit Context

- **Specialised Search Engines**
  - **Establish context by leveraging a specialised index**
  - **Context is explicitly indicated by the user's choice of specialised search engine**

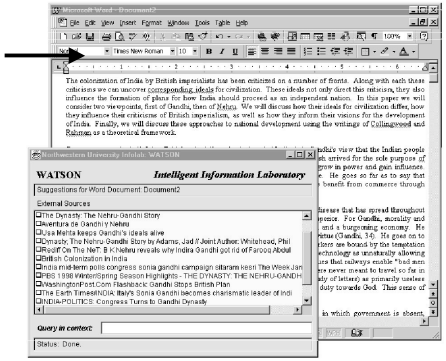


CiteSeer is a search engine devoted to academic research papers and as such all queries assume "research papers" to be the broad context.

Implicit Context – External Source

- **External context information taken from user activities surrounding search**
  - **E.g. Watson, INTELLEXT Inc., Topic-sensitive PageRank**

Watson monitors a user's word processing and uses context information from their recent text.



Implicit Context – Local Sources

- **Search results/past selections used to guide future search (e.g. SearchPad)**

By analysing the content of the top 20 search results it may be possible to identify additional query terms.

Note:
SearchPad was developed by Krishna Bharat then at Compaq, now Principal Scientist at Google and creator of Google News.

Further Reading

- **Information Retrieval**
  - **Numerous textbooks – a good one is *Modern Information Retrieval by Baeza-Yates and Ribeiro-Neto, 1999***
- ***Web Search***
  - ***Textbook - Google's PageRank and Beyond: The Science of Search Engine Rankings by Langville and Meyer, 2006***
  - ***Most up-to-date info on Web***
    - *e.g. SearchEngineWatch.com*
    - e.g. labs.google.com
    - Do a search!