WWW: A Dynamic Corpus

- Pages are created on the fly – the "hidden Web".
- Many pages change regularly.

![Graph](image)

Figure 11: Change intervals for pages with the average change interval of 10 days.


Web Search Engines

Largest Indexes: [12/31/02, Overture owns 2 & 3]

Searches per day: [2/03]
1. Google (250M), 2. Overture (167M), 3. Inktomi (80M)

Most Popular Service: [Fall 02, % of US users who searched at least once during a month]
1. Google (29.5%), 2. Yahoo (28.9%), 3. MSN (27.6%)

Search Engines: Who Really Powers What?

- Overture/AlltheWeb/AltaVista: paid ads plus recently acquired general search
  - Yahoo, MSN, Lycos, HotBot
- Google
  - AOL, Yahoo, Netscape (recent), HotBot
- Inktomi
  - MSN, LookSmart, HotBot (backup)
- LookSmart
  - MSN

Search Engines Features

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Boolean</th>
<th>Proximity</th>
<th>Stem</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>-, or</td>
<td>phrase</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>AlltheWeb</td>
<td>and, or, and not, +, -, or, with()</td>
<td>phrase</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>AltaVista</td>
<td>+, -, and, or, and not</td>
<td>phrase, near</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Teoma</td>
<td>-, or</td>
<td>phrase</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Search Engines Types

Web search has moved from straightforward adaptation of static IR techniques to specialized ones.
Indexes: as with static IR, map query to URLs
Directories: organize info like Yellow Pages
Meta-Search: submit query to multiple search engines in parallel

Early Indexes Ranking: TFIDF Variants

- Most search engines use a modified Vector model.
  - Term factors: frequency, meta words, words in headers
  - Document factors: IDF, query frequency
- Some cache full text (e.g., Google, CiteSeer).

Recent Indexes Ranking: Link Analysis

- Exploit human judgment about what matters... if someone bothered to add a link to a page on their page, then it must be worthwhile.
- Pages that share links are likely to be about the same topics.
- Issue: how to use link information?

HITS (Hypertext Induced Topic Search) [Kleinberg]

Rank a small subgraph of Web based on:
- Authorities: have many pages pointing to them
- Hubs: have many outgoing links
Weigh links based on context or TFIDF for page.
**HITS: Subgraph generation**

1. Initialize S as all pages containing query term(s)
2. Add to S any page that:
   - points to a page in S
   - is pointed to by a page in S
   - Need to track connectivity

**HITS: Link Weights**

Tags: increase weights of terms in tags, titles, meta fields or near beginning of sections.

Anchor text: increase weight of terms immediately around links and add terms to vector for referenced page

---

**PageRank: Link Analysis for Google [Page]**

- Models a user browsing the Web as a ergodic Markov chain: nxn transition probability matrix for n pages.
- User may follow link or jump to a random other page.
- Ergodic means path exists between any two states and non-zero probability of being in any state.
- Estimate probabilities by iterating over actual Web links.

---

**PageRank Calculation**

PageRank:

\[
PR(a) = q + (1-q) \sum_{i=1}^{n} \frac{PR(p_i)}{C(p_i)}
\]

where q is probability of jumping to random page, 1-q is probability of following a random hyperlink on current page, C(a) is # of outgoing links of page a and p_i are pages pointing to a.

Principal Eigenvector of transition matrix for Web
Google Architecture
[Brin & Page 1998]

- Indexes by *hits*: word occurrence information (position, font, and capitalization/tags)
- Ranks by PageRank combined with dot product of weighted hits and query term proximity within document

Figure 1: High Level Google Architecture

Google

FAST: carefully constructed data structures which expedite distributed computation.
Precision: additional information (proximity and linkage) focuses search on best pages
Comprehensive: crawlers have indexed most pages on Web
Archival: stores compressed versions of all crawled pages.
Mining/Caching: tunes to frequent queries and caches results

Search

Motivation: General search engines each cover 14-55% of Web, and not exactly the same sites.
Key Idea: If at first you don't succeed, try another search engine; in fact, try several at the same time.
Issues: How many to search? Which to search? How to keep up with interface and search engine changes? How to combine results?

- Common in early Web search (e.g., Yahoo, NorthernLight)
- Hierarchical taxonomy of Web pages.
- Human generated had more meaning, but indexed less of Web.
- Usually combined with index based search engine
- Architecture circa 1998
- Learned which search engines to query based on which returned results and which links user followed.