Search Engine Indexing

• Link (structure) Indexing
  index the links between web pages

• Text Indexing
  traditional IR indexing
Link Indexing

• Neiborhood information:
  for each page, incoming and outgoing links are indexed.

• Two adjacent lists are used to store neiborhood information
  – incoming link adjacent list
  – outgoing link adjacent list
Text Indexing

- Challenges
  - time to build index
  - storage to store index
  - index rebuild
Distributed Text Indexing

- Index Partitioning
  - local inverted file
    each machine is responsible for a disjoint subset of pages in collection. A query is broadcasted to all machines, each of which returns disjoint lists of pages.
  - global inverted file
    each machine stores a inverted file for a subset of terms. Search for a term is distributed to the appropriate machine.
Additional Index Information

In addition to the location of the term, other information may be useful to store. Examples are:

• term fond style: bold or italic

• section headings

• anchor
Page Rank

- Relevance Ranking
  - similarity between the query and the document

- Importance Ranking
  - a page is important if it is pointed by many important pages

- A page is ranked according to its relevance and importance
Simple PageRank Algorithm

- N(i): the number of outgoing links from page i
- B(i): the set of pages pointing to page i
- r(i): importance rank of page i

\[ r(i) = \sum_{j \in B(i)} \frac{r(j)}{N(j)} \]
A Simple PageRank Algorithm

1. $s$ = a random vector
2. $r = A^T \ast s$
3. if $||r - s|| < \varepsilon$, $r$ is the PageRank vector
4. $s = r$, goto 2.

$r = (r_1, ..., r_m)$ and $r_i$ is the rank of page $i$

$A$ is an $m$ by $m$ matrix, in which $a_{ij} = 1/N(i)$ if page $i$ points to page $j$, and $a_{ij} = 0$ otherwise.
Need for Meta-Search Engine

- many different web search engines
- each search engine has different coverage and emphasis
- each employs different algorithms for collecting, indexing, and search links
- each returns a different set of results
- users may need to query several search engines for the best search results
MetaSearch Engine

• Meta-search engines automate the process of using several search engines by submitting a single query to multiple search engines

• Examples:
  – ProFusion (University of Kansas)
  – MetaCrawler (University of Washington)
  – SavvySearch (Colorado State University)
MetaSearch Engine Functions

• Collecting a query from the user
• Processing the query
  – determine search engines to be used
• Dispatching the query to multiple search engines
• Processing the results
  – eliminate duplicate links
  – rerank the links
• Presenting the results
Text Languages

- Metadata
- Text Formats
- Text Models
- Markup Languages
- Multimedia Formats
Metadata

- Data about data and help user find the documents

- Examples of metadata of documents
  - size of the documents
  - author of the documents
  - content rating
  - topics
  - key words
Text Representation

- Text Coding Scheme
  - ASCII
  - Unicode

- Text Format
  - word
  - RTF
  - postscript
  - PDF
Entropy

Entropy quantifies the amount of information a text carries. The larger the entropy, the more information

\[ E = - \sum_{i=1}^{n} p_i \log_2 p_i \]

p_i is the probability of ith symbols
n is the number of symbols in the alphabet
Zipf’s Law

- A few words occur very often
  - 2 most frequent words can account for 10% of occurrences
  - top 6 words are 20%, top 50 words are 50%
- Many words are infrequent
- Rank * Frequency = Constant
  - r*Pr = A
  - r is rank
  - Pr is the probability of the word of rank r
  - A is around 0.1
# Example of Frequent Words

<table>
<thead>
<tr>
<th>Frequent Word</th>
<th>Number of Occurrences</th>
<th>Percentage of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>7,398,934</td>
<td>5.9</td>
</tr>
<tr>
<td>of</td>
<td>3,893,790</td>
<td>3.1</td>
</tr>
<tr>
<td>to</td>
<td>3,364,653</td>
<td>2.7</td>
</tr>
<tr>
<td>and</td>
<td>3,320,687</td>
<td>2.6</td>
</tr>
<tr>
<td>in</td>
<td>2,311,785</td>
<td>1.8</td>
</tr>
<tr>
<td>is</td>
<td>1,559,147</td>
<td>1.2</td>
</tr>
<tr>
<td>for</td>
<td>1,313,561</td>
<td>1.0</td>
</tr>
<tr>
<td>The</td>
<td>1,144,860</td>
<td>0.9</td>
</tr>
<tr>
<td>that</td>
<td>1,066,503</td>
<td>0.8</td>
</tr>
<tr>
<td>said</td>
<td>1,027,713</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Frequencies from 336,310 documents in the 1 GB TREC Volume 3 Corpus 125,720,891 total word occurrences and 508,209 unique words
Zipf’s Law

Distribution of the word frequencies and the ith most frequent word appears \( f(i) \) times

\[
f(i) = n/(i^\theta H_v(\theta))
\]

\[
H_V(\theta) = \sum_{i=1}^{V} \frac{1}{i^\theta}
\]

\( n \) is the number of word occurrences in a document
\( V \) is the number of words in a vocabulary
\( \theta \) is often between 1 and 2.
Vocabulary Growth Heaps’ Law

New words occur less frequently as collection grows

Heaps’ Law:
The number of distinct words in a document
\[ V = Kn^\beta \]
\( n \) is the number of words in a document
\( K \) is normally between 10 and 100
\( \beta \) is often between 0 and 1.
Markup Language

- **SGML**
  - a standard of markup languages

- **HTML**
  - an instance of SGML

- **XML**
  - a subset of SGML