Information Retrieval on the World Wide Web

Dr. Bulu Maharana
bulumaharana@gmail.com

Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.suniv.ac.in
Email: bulumaharana@gmail.com
Why Web IR an Issue?
# World Internet Usage and Population Statistics

**June 30, 2014 - Mid-Year Update**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Africa</strong></td>
<td>1,125,721,038</td>
<td>4,514,400</td>
<td>297,885,898</td>
<td>26.5 %</td>
<td>6,498.6 %</td>
<td>9.8 %</td>
</tr>
<tr>
<td><strong>Asia</strong></td>
<td>3,996,408,007</td>
<td>114,304,000</td>
<td>1,386,188,112</td>
<td>34.7 %</td>
<td>1,112.7 %</td>
<td>45.7 %</td>
</tr>
<tr>
<td><strong>Europe</strong></td>
<td>825,824,883</td>
<td>105,096,093</td>
<td>582,441,059</td>
<td>70.5 %</td>
<td>454.2 %</td>
<td>19.2 %</td>
</tr>
<tr>
<td><strong>Middle East</strong></td>
<td>231,588,580</td>
<td>3,284,800</td>
<td>111,809,510</td>
<td>48.3 %</td>
<td>3,303.8 %</td>
<td>3.7 %</td>
</tr>
<tr>
<td><strong>North America</strong></td>
<td>353,860,227</td>
<td>108,096,800</td>
<td>310,322,257</td>
<td>87.7 %</td>
<td>187.1 %</td>
<td>10.2 %</td>
</tr>
<tr>
<td><strong>Latin America / Caribbean</strong></td>
<td>612,279,181</td>
<td>18,068,919</td>
<td>320,312,562</td>
<td>52.3 %</td>
<td>1,672.7 %</td>
<td>10.5 %</td>
</tr>
<tr>
<td><strong>Oceania / Australia</strong></td>
<td>36,724,649</td>
<td>7,620,480</td>
<td>26,789,942</td>
<td>72.9 %</td>
<td>251.6 %</td>
<td>0.9 %</td>
</tr>
<tr>
<td><strong>WORLD TOTAL</strong></td>
<td>7,182,406,565</td>
<td>360,985,492</td>
<td>3,035,749,340</td>
<td>42.3 %</td>
<td>741.0 %</td>
<td>100.0 %</td>
</tr>
</tbody>
</table>

Information Retrieval

- Representation, storage, organisation, and access to information items
- (Usually) keyword-based representation

Primary goal of an IR system

"Retrieve all the documents which are relevant to a user query, while retrieving as few non-relevant documents as possible."
What is different about Web?

(1) Pages:
1. Bulk ..........................
1. Lack of stability ..........
1. Heterogeneity
   - Type of documents .. Text, pictures, audio, scripts, ...
   - Quality .................. From dreck to ICDE papers ...
   - Language ............... 100+
1. Duplication
   - Syntactic ............... 30% (near) duplicates
   - Semantic ................ ??
1. Non-running text .......... many home pages, bookmarks,
1. High linkage .............. ≥ 8 links/page in the average
The Big Challenge

Meet the user needs given the heterogeneity of Web pages
What is the difference about the Web?

(2) **Users:**

1. **Make poor queries**
   - Short (2.35 terms avg)
   - Imprecise terms
   - Sub-optimal syntax (80% queries without operator)
   - Low effort

1. **Wide variance in**
   - Needs

1. **Specify behavior**
   - 85% look over one result screen only
   - 78% of queries are not modified
   - Follow links
   - See various user studies in CHI, Hypertext, SIGIR, etc.
The Bigger Challenge

Meet the user needs given the heterogeneity of Web pages and the poorly made queries.
Why don’t the users get what they want from Web?

- User need
- User request (verbalized)
- Query to IR system
- Results

Translation problems

Polysemy Synonymy

Example

I need to get rid of mice in the basement

What’s the best way to trap mice alive?

mouse trap

Software, toy cars, inventive products, etc
User Tasks

**Pull technology**
- User requests information in an interactive manner
- 3 retrieval tasks
  - Browsing (hypertext)
  - Retrieval (classical IR systems)
  - Browsing and retrieval (modern digital libraries and web systems)

**Push technology**
- automatic and permanent pushing of information to user
- software agents
- example: news service
- filtering (retrieval task) relevant information for later inspection by user
IR Architecture
Information Retrieval Models

• An IR model governs how a document and a query are represented and how the relevance of a document to a user query is defined.

• Main models:
  – Boolean model
  – Vector space model
  – Statistical language model
Boolean model

• Each document or query is treated as a “bag of words” or terms. Word sequence is not considered.

• Given a collection of documents $D$, let $V = \{t_1, t_2, \ldots, t_{|V|}\}$ be the set of distinctive words/terms in the collection. $V$ is called the vocabulary.

• A weight $w_{ij} > 0$ is associated with each term $t_i$ of a document $d_j \in D$. For a term that does not appear in document $d_j$, $w_{ij} = 0$.

\[ d_j = (w_{1j}, w_{2j}, \ldots, w_{|V|j}), \]
Boolean model (contd)

• Query terms are combined logically using the Boolean operators **AND**, **OR**, and **NOT**.
  – E.g., ((data AND mining) AND (NOT text))

• Retrieval
  – Given a Boolean query, the system retrieves every document that makes the query logically true.
  – Called **exact match**.

• The retrieval results are usually quite poor because term frequency is not considered.
Vector Space model

• Documents are also treated as a “bag” of words or terms.
• Each document is represented as a vector.
• However, the term weights are no longer 0 or 1. Each term weight is computed based on some variations of TF or TF-IDF scheme.
• **Term Frequency (TF) Scheme:** The weight of a term $t_i$ in document $d_j$ is the number of times that $t_i$ appears in $d_j$, denoted by $f_{ij}$. Normalization may also be applied.
TF-IDF term weighting scheme

• The most well known weighting scheme
  – TF: still term frequency
  – IDF: inverse document frequency.

\[ N: \text{total number of docs} \]
\[ df_i: \text{the number of docs that } t_i \text{ appears.} \]

• The final TF-IDF term weight is:

\[
\begin{align*}
    tf_{ij} &= \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|V|j}\}} \\
    idf_i &= \log \frac{N}{df_i} \\
    w_{ij} &= tf_{ij} \times idf_i
\end{align*}
\]
Retrieval in vector space model

• Query \( q \) is represented in the same way or slightly differently.

• **Relevance of \( d_i \) to \( q \):** Compare the similarity of query \( q \) and document \( d_i \).

• Cosine similarity (the cosine of the angle between the two vectors)

\[
\text{cosine}(d_j, q) = \frac{\langle d_j \cdot q \rangle}{\|d_j\| \times \|q\|} = \frac{\sum_{i=1}^{\|V\|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{\|V\|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{\|V\|} w_{iq}^2}}
\]

• Cosine is also commonly used in text clustering
An Example

• A document space is defined by three terms:
  – hardware, software, users
  – the vocabulary

• A set of documents are defined as:
  – A1=(1, 0, 0), A2=(0, 1, 0), A3=(0, 0, 1)
  – A4=(1, 1, 0), A5=(1, 0, 1), A6=(0, 1, 1)
  – A7=(1, 1, 1) A8=(1, 0, 1), A9=(0, 1, 1)

• If the Query is “hardware and software”
• what documents should be retrieved?
An Example (cont.)

• In Boolean query matching:
  – document A4, A7 will be retrieved ("AND")
  – retrieved: A1, A2, A4, A5, A6, A7, A8, A9 ("OR")

• In similarity matching (cosine):
  – q=(1, 1, 0)
  – S(q, A1)=0.71, S(q, A2)=0.71, S(q, A3)=0
  – S(q, A4)=1, S(q, A5)=0.5, S(q, A6)=0.5
  – S(q, A7)=0.82, S(q, A8)=0.5, S(q, A9)=0.5
  – Document retrieved set (with ranking)=
    • \{A4, A7, A1, A2, A5, A6, A8, A9\}
The bright side: Web advantages vs. classic IR

**User**
- Many tools available
- Personalization
- Interactivity (refine the query if needed)

**Collection/tools**
- Redundancy
- Hyperlinks
- Statistics
  - Easy to gather
  - Large sample sizes
- Interactivity (make the users explain what they want)
Web IR tools

1. General-purpose search engines:
   - direct: AltaVista, Excite, Google, Infoseek, Lycos, …
   - Indirect (Meta-search): MetaCrawler, DogPile, AskJeeves, InvisibleWeb, …

1. Hierarchical directories: Yahoo!, all portals.

1. Specialized search engines:
   - Home page finder: Ahoy
   - Shopping robots: Jango, Junglee, …
Web IR tools (cont...)

1 Search-by-example: Alexa’s “What’s related”, Excite’s “More like this”, Google’s “Googlescout”, etc.

1 Collaborative filtering: Firefly, GAB, ...

1 ...

1 Meta-information:
- Search Engine Comparisons
- Query log statistics
- ...

Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.sruv.ac.in
Email: bulumiharana@gmail.com
Algorithmic issues related to search engines

1. Collecting documents
   - Priority
   - Load balancing
     - Internal
     - External
   - Trap avoidance
   - ...

1. Processing and representing the data
   - Query-independent ranking
   - Graph representation
   - Index building
   - Duplicate elimination
   - Categorization

1. Processing queries
   - Query-dependent ranking
   - Duplicate elimination
   - Query refinement
   - Clustering
   - ...

Sambalpur University
Post Graduate Deptt. of Lib. & Info. Science
Jyoti Vihar-768019, Odisha (INDIA)
URL: http://www.smu.ac.in
Email: bulumaharana@gmail.com
Ranking of Web Pages

1. **Goal:** order the answers to a query in decreasing order of value
   - **Query-independent:** assign an intrinsic value to a document, regardless of the actual query.
   - **Query-dependent:** value is determined only wrt a particular query.
   - **Mixed:** combination of both valuations.

2. **Examples**
   - **Query-independent:** length, vocabulary, publication data, number of citations (indegree), etc.
   - **Query-dependent:** cosine measure
Considerations for Search Engines

- Scalability
- Content Freshness
- Speed of service
- Relevancy of Search results
Units of a Search Engine

1. Crawling
2. Indexing
3. Querying
4. Searching
5. Ranking
6. Browsing
Text pre-processing

• Word (term) extraction: easy
• Stopwords removal
• Stemming
• Frequency counts and computing TF-IDF term weights.
Stopwords removal

• Many of the most frequently used words in English are useless in IR and text mining – these words are called *stop words*. 
  – the, of, and, to, ....
  – Typically about 400 to 500 such words
  – For an application, an additional domain specific stopwords list may be constructed

• Why do we need to remove stopwords?
  – Reduce indexing (or data) file size
    • stopwords accounts 20-30% of total word counts.
  – Improve efficiency and effectiveness
    • stopwords are not useful for searching or text mining
    • they may also confuse the retrieval system.
Stemming

- Techniques used to find out the root/stem of a word. E.g.,
  - user     engineering
  - users    engineered
  - used     engineer
  - using

- stem: use     engineer

Usefulness:
- improving effectiveness of IR and text mining
  - matching similar words
  - Mainly improve recall
- reducing indexing size
  - combing words with same roots may reduce indexing size as much as 40-50%.
Basic stemming methods

Using a set of rules. E.g.,

• remove ending
  – if a word ends with a consonant other than s, followed by an s, then delete s.
  – if a word ends in es, drop the s.
  – if a word ends in ing, delete the ing unless the remaining word consists only of one letter or of th.
  – If a word ends with ed, preceded by a consonant, delete the ed unless this leaves only a single letter.
  – ......

• transform words
  – if a word ends with “ies” but not “eies” or “aies” then “ies --> y.”
Frequency counts + TF-IDF

• Counts the number of times a word occurred in a document.
  – Using occurrence frequencies to indicate relative importance of a word in a document.
    • if a word appears often in a document, the document likely “deals with” subjects related to the word.

• Counts the number of documents in the collection that contains each word

• TF-IDF can be computed.
Evaluation: Precision, Recall & E Measure

• Given a query:
  – Are all retrieved documents relevant?
  – Have all the relevant documents been retrieved?

• Measures for system performance:
  – The first question is about the precision of the search
  – The second is about the completeness (recall) of the search.
  – E-Measure: Normalization of Recall and Precision
Search Result Ranking

- Ranking is the process in which the closeness of a document to the user query is measured.
- Although there are many ranking techniques used by SEs, most of the ranking algorithms are not known.
Popular Ranking Techniques

1. **Boolean Spread**
   Number of Query terms found in the page and its neighborhood pages

2. **Vector Space**
   Term Frequency (TF) and Inverse Document Frequency (IDF)

3. **Most cited**
   Pages being pointed to in the answer set (authorities) and pages in the answer set which have outgoing links (hubs).....Chances of hyperlink spamming

4. **Citation Rank** (Google’s PageRank)
The size of each face is proportional to the total size of the other faces which are pointing to it.
# Web IR Challenges

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Efforts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution of Web Content</td>
<td>Network limitations</td>
</tr>
<tr>
<td></td>
<td>Platform incompatibility</td>
</tr>
<tr>
<td>High Data volatility</td>
<td>Millions of pages added and eliminated; Domain Name Changes</td>
</tr>
<tr>
<td>Heterogeneity and Size of Web data</td>
<td>Varies in language, File Formats, Media</td>
</tr>
<tr>
<td>Lack of structure and data redundancy</td>
<td>No structure because of HTML, Mirroring or Proxy Servers, 30% of Web Pages duplicated</td>
</tr>
<tr>
<td>Poor Content Quality</td>
<td>Any body can post, no editorial process</td>
</tr>
<tr>
<td>Web Traps</td>
<td>Anti-spam protocols, URL aliases, Content duplication</td>
</tr>
<tr>
<td>Modeling the Web</td>
<td>Vector Space exhausted</td>
</tr>
<tr>
<td>Querying</td>
<td>Embedding structure in search queries</td>
</tr>
<tr>
<td>Distributed Architecture</td>
<td>Indexing Mechanisms to be replaced with Effective Search Agents</td>
</tr>
<tr>
<td>Ranking</td>
<td>Integrating the User in Search process</td>
</tr>
<tr>
<td>Hidden Web</td>
<td>Advance Search Agents</td>
</tr>
</tbody>
</table>
Web IR for Librarians
Small Directories

- Built by information specialists
- Selected, evaluated, annotated
- Organized into subject categories

- Librarians’ Internet Index (lii.org)
  - By a group of California library professionals
- Infomine
  - By UC consortium of library professionals
- Academic Info
  - By a librarian in Arizona
Larger Directories

• Google Web directory
  • [http://directory.google.com](http://directory.google.com)
    – 5+ million pages - less than 0.04% of Google web

• About.com – a collection of specialized directories
  – search by subject

• Yahoo’s directory
  • [http://dir.yahoo.com](http://dir.yahoo.com)
    – 4 million UNevaluated pages - about 0.06% of Yahoo! search
Finding “expert pages” and searchable databases

- Look in all the directories just mentioned
  - Databases and “expert pages” scattered throughout

- In routine searching:
  - If a site calls itself a directory or database, you can search on it
    genome database
    “cell biology” directory

- Look for society’s pages with collections of links
  genome society
  Home Page of “International mammalian genome society”
CRITICAL EVALUATION

Why Evaluate What You Find on the Web?

• Anyone can put up a Web page
  – about anything
• Many pages not kept up-to-date
• No quality control
  – most sites not “peer-reviewed”
    • less trustworthy than scholarly publications
  – no selection guidelines for search engines
Web Evaluation Techniques

Before you click to view the page...

• **Look at the URL** - personal page or site?

• Domain name appropriate for the content?
  - edu, com, org, net, gov, ca.us, uk, etc.

• Published by an entity that makes sense?
  - News from its source?
  - Advice from valid agency?
Web Evaluation Techniques

Scan the perimeter of the page

• Can you tell who wrote it?
  • name of page author
  • organization, institution, agency you recognize
  • e-mail contact by itself not enough

• Credentials for the subject matter?
  – Look for links to:
    “About us” “Philosophy” “Background” “Biography”

• Is it recent or current enough?
  • Look for “last updated” date - usually at bottom

• If no links or other clues...
  • truncate back the URL
Web Evaluation Techniques

Indicators of quality

• Sources documented
  • links, footnotes, etc.
    – As detailed as you expect in print publications?
  • do the links work?

• Information retyped or forged
  • why not a link to published version instead?

• Links to other resources
  • biased, slanted?
Web Evaluation Techniques

What Do Others Say?

• Search the URL in alexa.com
  – Who links to the site? Who owns the domain?
  – Type or paste the URL into the basic search box
  – Traffic for top 100,000 sites

• See what links are in Google’s Similar pages

• Look up the page author in Google
Web Evaluation Techniques

STEP BACK & ASK: Does it all add up?

• Why was the page put on the Web?
  • inform with facts and data?
  • explain, persuade?
  • sell, attract?
  • share, disclose?
  • as a parody or satire?

• Is it appropriate for your purpose?
Thank you !!