CS 115: COMPUTING FOR THE SOCIO-TECHNO WEB

SEARCH ENGINES & SEARCH ALGORITHMS
WHAT IS THE SHAPE OF THE WEB?

“Map of the Internet” (1998)
WHAT ARE YOU TRYING TO FIND?

Types of queries:

• **Informational** - want to learn about something
  - Peripheral neuropathy

• **Navigational** - want to go to that page
  - Wellesley College

• **Transactional** - want to do something (web-mediated)
  - Access a service
    - Wellesley weather
  - Downloads
    - Mars surface images
  - Shop
    - Nikon SLR camera

• **Gray areas**
  - Find a good hub (resource collection)
    - car rental Boston
  - Exploratory search “see what’s there”
    - morality of abortion
WHAT IS THE SHAPE OF THE WEB?

“Map of the Internet” (1998)
THE WEB IS A DIRECTED GRAPH

- Like a map of a country with cities and one-way roads

- Directed Graph of Nodes and Arcs (one-way connections)
  - Nodes = web pages
  - Arcs = hyperlinks from a page to another

- Why is this cool? Because...
  - it can be explored
  - it can be indexed
WHAT IS THE SHAPE OF THE WEB?

<table>
<thead>
<tr>
<th>Region</th>
<th>SCC</th>
<th>IN</th>
<th>OUT</th>
<th>TENDRILS</th>
<th>DISC.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>56,463,993</td>
<td>43,343,168</td>
<td>43,166,185</td>
<td>43,797,944</td>
<td>16,777,756</td>
<td>203,549,046</td>
</tr>
</tbody>
</table>

Bow-tie shape of the web

- Central core: 56 million pages
- IN: 44 million pages
- OUT: 44 million pages
- Tendrils and tubes: 44 million pages
- Disconnected components: 17 million pages
strongly connected component (SCC) in a directed graph is a subset of the nodes such that:
(i) every node in the subset has a path to every other; and
(ii) the subset is not part of some larger set with the property that every node can reach every other.
BOWTIE TERMINOLOGY: LARGEST SCC, CORE, IN, OUT, ISLANDS, TENDRILS
WHAT IS THE SHAPE OF THE WEB?

“Map of the Internet” (1998)
HOW BIG IS THE WEB?

Number of accessible web pages (the *visible* web)
- 1 trillion unique URLs are indexed

The *deep web* (or *hidden* or *invisible* web)
- “contains 400-550 times more information”

*Coverage* (i.e. the proportion of the web indexed) is crucial for search engines.
- Today, less than 15% pages are indexed!
Google: 79.98%
Bing: 9.78%
Yahoo!: 8.39%
AOL: 0.83%
DuckDuckGo: 0.4%
Other: 0.61%
3.5 million searches per minute!

Estimated Unique Monthly Visitors: 1.6 billion
HOW FAR DO YOU LOOK FOR RESULTS?

(Click-Through Rate)
Crawling & Indexing

The journey of a query starts before you ever type a search, with crawling and indexing the web of trillions of documents.

How Search Works

These processes lay the foundation — they're how we gather and organize information on the web so we can return the most useful results to you. Our index is well over 100,000,000 gigabytes, and we've spent over one million computing hours to build it. Learn more about the basics in this short video.
HOW SEARCH ENGINES WORK

User

Search

Web

Web spider

Indexer

Indexes

Ad indexes

The Web

A

B

C

D

E
1st Gen. Ranking = Content Similarity...

**Content Similarity Ranking:**
The more rare words two documents share, the more similar they are

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>157</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brutus</td>
<td>4</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>232</td>
<td>227</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>worser</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Documents are treated as “**bags of words**”
  (no effort to “understand” the contents)
- Term frequencies are computed
  (Above: a few rare words in W.S. books)
- Very common words (“a”, “of”, “the”, “in”, …) are ignored
**Content Similarity Ranking:**
The more *rare* words two documents share, the more similar they are

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>worser</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- **Above:** *Normalized* term frequencies.
- **Results are** *ranked* by sorting the sum of normalized frequencies of the query terms

**Query:** “In the mercy of Caesar”  
**Rank Results:** Antony & Cleo
1st Generation Spam: Keyword Stuffing

- “Keyword stuffing”: Add keywords, text, to increase content similarity
The reputation “PageRank” of a page \( P_i \) =
the sum
of a fraction of the reputations
of all pages \( P_j \) that point to \( P_i \)

Idea similar to academic reputation

Beautiful Math behind it

The PageRank of a page
- Can be calculated by
  a sequence of matrix multiplications
- It is equivalent to
  the probability of randomly surfing to a page

How To Spam?
WHAT IS THE SHAPE OF THE WEB?

“Map of the Internet” (1998)