Recap: Binary Representation

<table>
<thead>
<tr>
<th>Powers of 2</th>
<th>2^0</th>
<th>2^1</th>
<th>2^2</th>
<th>2^3</th>
<th>2^4</th>
<th>2^5</th>
<th>2^6</th>
<th>2^7</th>
<th>2^8</th>
<th>2^9</th>
<th>2^10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
<td>1024</td>
</tr>
</tbody>
</table>

$2^{10} = 1024 \approx 10^3$

**Fact:** Every integer can be *uniquely* represented as a sum of powers of 2.

**Ex:** $25 = 16 + 8 + 1$

$= 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$

$[25]_2 = 11001$
Misconceptions about Computers

Just a calculator on steroids

Just maintains large amount of data

Just does what the programmer tells it

Weather Forecast

Airline Reservation System

Yes, but …
Various meanings of search

- Look up “Shirley Tilghman” in online phonebook.
- In consumer database, find “credit-worthy” consumers.
- Find web pages relevant to “computer music.”
- Among all cell phone conversations originating in Country X, identify suspicious ones.
- Search all religion and philosophy books of the world for meaning of life.
These are major scientific problems with many components

- Engineering
- Algorithms
- Linguistics
- Statistical Modeling
- Ethics, Policy, Society
Electronic Phonebook

- **ASCII**: Agreed-upon convention for representing letters with numbers

- **Example**:

```plaintext
<table>
<thead>
<tr>
<th>T</th>
<th>i</th>
<th>l</th>
<th>g</th>
<th>h</th>
<th>m</th>
<th>a</th>
<th>n</th>
<th>,</th>
<th>2</th>
<th>5</th>
<th>8</th>
<th>-</th>
<th>6</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>105</td>
<td>108</td>
<td>103</td>
<td>104</td>
<td>109</td>
<td>97</td>
<td>110</td>
<td>44</td>
<td>50</td>
<td>53</td>
<td>56</td>
<td>45</td>
<td>54</td>
<td>49</td>
<td>48</td>
<td>48</td>
</tr>
</tbody>
</table>
```

- Sorted Phonebook
  = sorted array of numbers

- Use binary search
Rest of the lecture: Web Search
World Wide Web (simplified view)

URL: Unique address for each document

Browser

Web Page

Hyperlink
Future lecture:
Physical infrastructure of the Web

Routers, gateways, DNS, ...
Logical Structure of the Web

Page A ➔ Page B ➔ Page C ➔ Page D

“Directed graph”
“edges” = link from one node to another

Important: This logical structure is created by independent actions of 100s of millions of users
1st step for search engines: create snapshot of the web

- **Webcrawler**: “browser on autopilot”
  - Maintains array of web pages it has seen
  - 2 types of pages: “visited”, “fully explored”
  - Do forever
    
    ```
    \{
        Pick any webpage marked “visited” from array.
        Mark it “fully explored.”
        Open all its linked pages in browser.
        Save them in array and mark them “visited.”
    \}
    ```

  Better: just the pages not “fully explored” yet.
The WebCrawler Index is now available for searching! The index is broad: it contains information from as many different servers as possible. It's a great tool for locating several different starting points for exploring by hand. The current index is based on the contents of documents located on nearly 4000 servers, world-wide.

Check it out at:

http://www.biotech.washington.edu/WebCrawler/WebQuery.html

Other information is available from there, including a description of the WebCrawler (the robot itself), and a list of the 25 most frequently referenced sites on the Web.

Brian Pinkerton
Dept of Computer Science and Engineering
University of Washington

[http://thinkpink.com/bp/WebCrawler/History.html]
Still Feasible Today?

- About 15 billion web pages today.
- Say 10 kb (10,000 bytes) of data per page
- $15 \times 10^{13}$ bytes to store the web
- $\approx 150,000$ Gb
- $\approx 500$ hard disks
- $\approx$ $50,000$ in ‘07
Princeton Shape Search Engine

[http://shape.cs.princeton.edu/search.html]
Finding Forrester

How does Google find Forrester Cole…?

Forrester Cole

foole@cs.<this school>.edu

Department of Computer Science
35 Olden St.
Princeton NJ 08544

I am a third year Ph.D. candidate in the computer graphics group at Princeton. My advisor is Adam Finkelstein.

Prior to coming to Princeton I was a programmer with Pandemic Studios in Los Angeles, where I worked on Mercenaries.

Teaching

I am a teaching assistant for COS116: The Computational Universe for spring 2007.

Lab Hours: TBA

Office Hours (CS413): TBA

Research

My current research investigates how artists select lines for line
Searching for “computer music”

Ideas?

- Identify all pages that contain “computer music”.
- Sort according to number of occurrences of “computer music” in the page.
- Human staff computes answers to all possible questions.
Some pitfalls

- “Spamming” by unscrupulous websites
- Synonymy (car, auto, vehicle …)
- Polysemy (jaguar: car or cat?)
Solution

IBM’s CLEVER – 1996

Google’s PAGERANK – 1997

Take advantage of the link structure of the web

Web link confers “approval”
Authorities: Sites that are viewed “with respect” by many
- New York Times
- International Computer Music Association

Hubs: Clearinghouses of information
- “My favorite computer music links”

Typically Authorities point to hubs and hubs point to authorities

Circular Definition?
Breaking Circularity

- Iterative algorithm
- Start with

   Pages containing “Computer music”

   All pages they point to

- At every step each page has:

  - “Hub Score”
  - “Authority Score”

Initially all 1
Score Calculation

- Do forever
  
  { 
  
  Next Hub Score for page ← Sum of current Authority Scores of pages that link to it.

  Next Authority Score for page ← Sum of current Hub Scores of pages that link to it.

  }

Fact The scores converge.

(Proof uses Linear Algebra, Eigenvalues)
- By Product – Algorithm reveals clusters

Example:

“Abortion”

- Pro-Choice
- Pro-Life

- Data Mining – Process of finding answers that are not in the data and must be inferred.

Example: “How is a person who shops at Whole Foods & REI likely to vote?”
Concerns

From **users**:  
- Privacy  
- Privacy  
- Privacy

From **Computer scientists**:  
- Formalize privacy  
- How to safeguard privacy while allowing legitimate computations
Next Time...

Digital Audio / Music (Perry Cook)