The continuum of computer “intelligence”

2/28/2006
COS 116
Instructor: Sanjeev Arora
Recap: Binary Representation

<table>
<thead>
<tr>
<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>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 programmer tells it

Yes, but …

Weather Forecast

Airline Reservation System
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
- Statistical Modeling
- Algorithms
- Ethics, Policy, Society
- Linguistics
Electronic Phonebook

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

<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

Web Results
Page 1 of 33,251 results containing shirley tilghman (0.38 seconds)

Google Press Center: Press Release
Google Appoints Shirley M. Tilghman, Ph.D., to its Board of Directors MOUNTAIN VIEW, Calif. - October 5, 2005 - Google inc. (NASDAQ
www.google.com/press/pressrel/tilghman_board.html Cached page

Princeton University Office of the President - President's Biography
Shirley M. Tilghman was elected Princeton University's 19th president on May 5, 2001, and assumed office on June 15, 2001. An exceptional teacher and a world-renowned scholar and leader in the field ...
www.princeton.edu/president/biography Cached page

Princeton - News - Shirley Tilghman named University's 19th President
Shirley Tilghman named Princeton University's 19th President Princeton, N.J. -- Shirley M. Caldwell Tilghman, a member of the Princeton University faculty since 1986, an ...
www.princeton.edu/pr/news/0162/0506-tilghman.htm Cached page
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, etc.
Logical Structure of the Web

**Important:** This logical structure is created by independent actions of 100s of millions of users

“Directed graph”

“edges” = link from one node to another
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

```plaintext
{  
  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.”
}
```
Feasibility Calculation

- About 15 billion web pages today.
- Say 10 Kilobytes (10,000 bytes) of data per page
- $15 \times 10^{13}$ bytes to store the web
- $\approx 150,000$ Gb
- $\approx 500$ Hard Disks (about $150,000$)
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
- Polysemy
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?**

Circular Definition – see Definition, Circular
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”
  \[\text{Initially all 1}\]
Score Calculation

- Do forever

\[
\{ \\
\text{Next Hub Score for page} & \quad \text{Sum of current Authority Scores of pages that link to it.} \\
\text{Next Authority Score for page} & \quad \text{Sum of current Hub Scores of pages that link to it.}
\}
\]

**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?”
Computer models and jurisprudence
Aug 25th 2005

[Fowler and Jeon, '05]
Concerns

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

From **Computer scientists**:
- Formalize privacy
- How to safeguard privacy while allowing legitimate computations
Qs for next time:

What is computation?

What can computers not do?

Also, 10-min discussion of readings for today’s lecture.