COS 435: Information Retrieval, Discovery, & Delivery

Questions about how we find, organize, evaluate and deliver information

Concept of Information in Digital Age

• What is information?
• Where do we find it?
• How do we extract it?

Some numbers from Web (no guarantees)

• From July 25, 2008 Google blog
  – trillion unique URLs crawled
• From IDC market analysis co in 2013
  – 1.9 zettabytes info created since Jan 1, 2011
• From factshunt.com, as of Dec. 31, 2013
  – 14.3 trillion live Webpages
  – 48 billion Webpages indexed by Google.Inc.
  – 14 billion Webpages indexed by Bing.
  – >1 yottabyte total data stored on Internet

Concept of Information in Digital Age

• What is information?
• How is it different from data?
• How is it different from knowledge?

Retrieval

Have
• Collection of “information objects”
  – “information object” is unit of information
  – think “document” or “image”
• Users who have information needs

Retrieval

Want
• Model to represent information objects
  – precise enough for retrieval
  – Efficient
• Query language for asking for info want
  – able to capture user’s information need
• Retrieval system to find relevant info
  – return “info objects” best satisfy query
  – experiment to get right query
  – “Know it when see it” correctness
<table>
<thead>
<tr>
<th>Unstructured information objects</th>
<th>Compare</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Information retrieval usually refers to <strong>unstructured</strong> objects:</td>
<td>• <strong>Structured information</strong>: database system</td>
</tr>
<tr>
<td>– Text</td>
<td>– tagged, typed</td>
</tr>
<tr>
<td>– Graphics: 2D, 3D</td>
<td>– well-defined semantic interpretation</td>
</tr>
<tr>
<td>– Music</td>
<td>– precise queries</td>
</tr>
<tr>
<td>– Video</td>
<td>• database query languages like SQL</td>
</tr>
<tr>
<td>– any help with semantic interpretation?</td>
<td>– precise response</td>
</tr>
<tr>
<td></td>
<td>• data matches query or not</td>
</tr>
<tr>
<td></td>
<td>• <strong>Semi-structured objects</strong>: tagged</td>
</tr>
<tr>
<td></td>
<td>– XML, HTML?</td>
</tr>
<tr>
<td></td>
<td>– some help with semantic interpretation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discovery</th>
<th>Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <strong>Content discovery</strong></td>
<td>• <strong>Information discovery</strong></td>
</tr>
<tr>
<td>What are the information objects?</td>
<td>– combinations</td>
</tr>
<tr>
<td>– constructed collections: digital libraries</td>
<td>– content analysis: data mining</td>
</tr>
<tr>
<td>• all in one (conceptually) place</td>
<td>• clustering</td>
</tr>
<tr>
<td>• curated?</td>
<td>– prediction</td>
</tr>
<tr>
<td>– harvested collections</td>
<td>– relationship analysis</td>
</tr>
<tr>
<td>• Web crawling</td>
<td>• network analysis</td>
</tr>
<tr>
<td>– databases behind Web pages</td>
<td>– metadata</td>
</tr>
<tr>
<td>• &quot;deep Web&quot;</td>
<td></td>
</tr>
<tr>
<td>– temporal issues</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Delivery</th>
<th>Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <strong>Content delivery</strong></td>
<td>• <strong>Information delivery</strong> - broadly construed:</td>
</tr>
<tr>
<td>– search tool and content repository over one <strong>umbrella organization</strong></td>
<td>– mode of interaction?</td>
</tr>
<tr>
<td>• e.g. Facebook, Library of Congress</td>
<td>• compare handheld, desktop</td>
</tr>
<tr>
<td>– Web search engines: <strong>actual Web pages not provided by search engines</strong></td>
<td>– user interfaces</td>
</tr>
<tr>
<td>• freshness issue</td>
<td>– visualization</td>
</tr>
<tr>
<td>• can get cached copy sometimes</td>
<td>• Analysis</td>
</tr>
<tr>
<td>– <strong>where content stored affects delivery</strong></td>
<td>– other?</td>
</tr>
<tr>
<td>• Storage Management</td>
<td></td>
</tr>
<tr>
<td>• Bandwidth management</td>
<td></td>
</tr>
</tbody>
</table>
What are efficiency issues?

- Large amounts data
  - build indexes
  - disks I/O or not?
  - distributed data
- Large volume of queries
  - distributed computing
- Expensive analysis
  - algorithm design
  - distributed computing

Search Engine

A system that implements information retrieval methods for a collection

- May create the collection
  - discovery of content
- Has a query language and retrieval model
- Has methods for presenting query results
  - system architecture + algorithms + implementation

Topics

- Information retrieval models for text documents
- Indexing and inverted files
- Ranking documents
- Using linking structure for Web content analysis
- User behavior-based relevance criteria
- Evaluating retrieval systems
- Social networks as sources of meta-info
- Social networks as sources of information
- Recommender systems

Topics cont.

- Privacy issues
- Web crawling
- System design of search engines: distributed storage and computing
- Document similarity
- Clustering
- Non-text media search
- Searching dynamic information sources

Course logistics

- **TA:** Yinda Zhang
- **Web site:**
  - COS home page -> courses -> schedule -> COS 435
    - General Information
    - Schedule and Assignments
    - Project description
- **Communication:** using Piazza
  - announcements
  - Q&A
- **Text:** *Introduction to Information Retrieval*
  - available online
  - 2 other online texts – see general info

Course Work

- Tests – two, take-home
- Homework, 6
- Project – pairs
  - your choosing with approval