COS 435: Information Retrieval, Discovery, & Delivery

Andrea LaPaugh

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

Concept of Information in Digital Age

- What is information?
- How is it different from data?

Concept of Information in Digital Age

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

#### 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

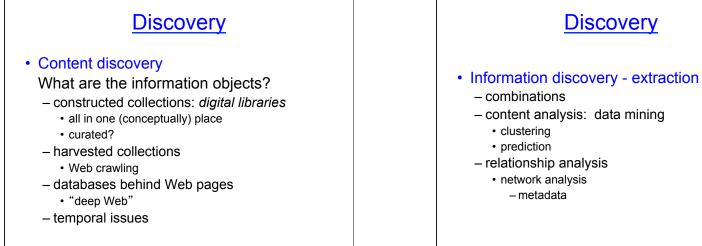
   precise enough for retrieval
  - Efficient
- Query language for asking for info want – able to capture user's information need
- Retrieval system to find relevant info
  - return "info objects" best satisfy query
  - experiment to get right query
  - "Know it when see it" correctness

# Unstructured information objects

- · Information retrieval usually refers to unstructured objects:
  - Text
  - Graphics: 2D, 3D
  - Music
  - Video
  - any help with semantic interpretation?

# Compare

- Structured information: database system - tagged, typed
  - well-defined semantic interpretation
  - precise queries
    - database guery languages like SQL
  - precise response
    - data matches query or not
- Semi-structured objects: tagged
  - XML, HTML?
  - some help with semantic interpretation



### **Delivery**

#### Content delivery

- search tool and content repository over one umbrella organization
  - e.g. Facebook, Library of Congress
- Web search engines: actual Web pages not provided by search engines
  - freshness issue
  - · can get cached copy sometimes
- where content stored affects delivery
  - Storage Management
  - Bandwidth management

### **Delivery**

- Information delivery broadly construed:
  - mode of interaction?
    - compare handheld, desktop
  - user interfaces
  - visualization
  - Analysis
  - other ?

#### 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

# 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: Mayank Mahajan
- Web site:

COS home page -> courses -> schedule -> COS 435

- General Information
- Schedule and Assignments (today)
- Project description (this week)
- · 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 single or pairs
  - your choosing with approval
  - Examples from 2016
    - Generating Playlists Using Spotify Data
    - Personalized Reddit Search App
    - Trending Tweet Prediction and Recommendation