Web

- Web opened the door for many important applications

- Information Retrieval
  - Web Search
  - Information Recommendation by content or by collaborative information

- Web Services
- Semantic Web
- Web 2.0
- XML
- Social Network

………………………..
Why Information Retrieval:

Information Overload:

“... The world produces between 1 and 2 exabytes ($10^{18}$ bytes) of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. ...” (Lyman & Hal 03)

Why this course?

• Managing Data is one of the primary uses of computers
• Most of this data is NOT contained in structured databases
  – Merrill Lynch estimates that more than 85 percent of all business information exists as unstructured data - commonly appearing in e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, ... and Web pages.
  – Text in Web pages or emails; image; audio; video; protein sequences..
  – Therefore, no carefully structured queries
• How do we find this?
  Information Retrieval
Information Retrieval: Challenges

- Data is unstructured
  - Need to guess what is important
- Query is unstructured
  - Need to guess user intent
- But computers don’t guess!

*Inferring relevance and intent from data, query is the science of Information Retrieval*

IR vs. RDBMS

- Relational Database Management Systems (RDBMS):
  - Semantics of each object are well defined
  - Complex query languages (e.g., SQL)
  - Exact retrieval for what you ask
  - Emphasis on efficiency
- Information Retrieval (IR):
  - Semantics of object are subjective, not well defined
  - Usually simple query languages (e.g., natural language query)
  - You should get what you want, even the query is bad
  - Effectiveness is primary issue, although efficiency is important
IR and other disciplines

Some core concepts of IR

- Information Need
- Representation
- Query
- Retrieval Model
- Indexed Objects
- Retrieved Objects
- Returned Results
- Evaluation/Feedback
Some core concepts of IR

Query Representation:
- Bridge lexical gap: system and systems; create and creating (stemmer)
- Bridge semantic gap: car and automobile (feedback)

Document Representation:
- Internal representation of document contents: a list of documents that contain specific word (inverted document list)
- Representation of document structure: different fields (e.g., title, body)

Retrieval Model:
- Algorithms that best match meaning of user query and available documents. (e.g., vector space model and statistical language modeling)
IR Applications

Information Retrieval: a gold mine of applications

- Web Search
- Information Organization: text categorization; document clustering
- Information Recommendation by content or by collaborative information
- Information Extraction: deep analysis of the surface text data
- Question-Answering: find the answer directly
- Federated Search: explore hidden Web
- Multimedia Information Retrieval: image, video
- Information Visualization: Let user understand the results in the best way

IR Applications: Text Categorization

Google News

Global warming has been a popular topic among scientists
DailyTech - 3 hours ago
The Earth's average temperature over the past quarter century has been the hottest in four centuries -- and part of the world has been warmer during the past 25 years than any period in the past 1,200 years, according to the National Academy of Sciences ...

No More Outrageous Global Warming Claims
Detroit Free Press

Utopian Unlimited - Seattle Times - Reuters - Forbes - all 441 related »

World's oldest bling: two tiny 100,000-year-old shells
Guardian Unlimited - 5 hours ago
They may not compare with today's diamond-encrusted bling, but in their own way, they are of far greater value. Two tiny shells have been confirmed as the world's oldest known items of jewellery, probably used on a necklace about 100,000 years ago.

Sea shells may be world's oldest beads MIND BLOWING
Researchers Uncover What May Be Oldest Known Jawels Voice of America

USA Today - New York Times - People's Diary Online - Telegraph.co.uk - all 79 related »
IR Applications: Text Categorization

Medical Subject Headings (Categories)

- Anatomy
- Organism
- Disease
- Chemicals and Drugs
- Analytical, Diagnostic and Therapeutic Techniques and Equipment
- Psychology and Psychopathology
- Behavior and Behavior Mechanisms
- Psychological Phenomena and Processes
- Mental Disorders
- Behavioral Sciences

IR Applications: Document Clustering

Clusty

Top 250 results of at least 94,536,661 retrieved for the query "Java".

1. Java Technology
2. Java Programming Language
IR Applications: Content Based Filtering

Keyword Matching

IR Applications: Collaborative Filtering

Other Customers with similar tastes
Bring structure and semantic meaning to text:

- **Entity detection**
  
  An 80-year-old woman with diabetes mellitus was treated with gliclazide. Prior to the gliclazide administration, her urinary excretion of albumin, serum urea nitrogen and serum creatinine were normal. After the medication, oliguria, edema and azotemia developed. On the twenty-fourth day when the edema was severe and generalized, gliclazide administration was terminated.

  *Diabetes: entity of disease*  
  *gliclazide: entity of drug*

- **Recognize Relationship between entities**

  What type of effect of gliclazide on this patient with diabetes

- **Inference based on the relationship between entities**

  ![Diagram of relationships between entities: Inherited Disease -> Gene -> Chemical (Drug discovery)]
IR Applications: Web Search

Web

Crawled into a centralized database

IR Applications: Federated Search

Valuable → Searched by Federated Search
IR Applications: Expertise Search

**INDURE:** Indiana database of university research expertise

[www.indure.org](http://www.indure.org)

IR Applications: Citation/Link Analysis

- **Linear Collider Accelerator in Japan**
- **U.S. Government Lab**
- **Nobel Prize Organization**
IR Applications: Citation/Link Analysis


IR Applications: Multimedia Retrieval

- Query
- Color Histogram
- Wavelet...
- Feature Extraction
- Retrieval Model
- Feature Extraction
Course Goals

- Learn the techniques behind Web search engines, E-commerce recommendation systems, etc.
- Get hands on project experience by developing real-world applications, such as building a small-scale Web search engine, a Web page management system, or a movie recommendation system.
- Learn tools and techniques to do research in the area of information retrieval or text mining.
- Lead to the amazing job opportunities in Search Technology and E-commerce companies such as Google, Microsoft, Yahoo! and Amazon.
Logistics

- Time and location: MWF 10:30-11:20, ME 1130
- Instructor: Chris Clifton, clifton@cs.purdue.edu
  LWSN 2142F, office hours: TBD
- Teaching assistants:
  - Rajkumar Pujari, rpujari@purdue.edu, office hours: TBD
  - TBD
- Email list: fall-2018-CS-47300-le1@lists.purdue.edu
- Piazza signup: piazza.com/purdue/fall2018/cs47300
- Prerequisites: CS25100. Having had a Stat course (e.g., STAT 35000) will help.

Readings

Workload

- Homeworks
  - 4-5 written assignments
  - 2-3 more substantial programming projects
  - Late policy: 15% off per day late, maximum of 5 days
  - Five extension days to be used at your discretion
    - Must state explicitly in heading of work being turned in that you are using late days
    - No fractional days
    - May not be used to extend submission past last day of class.

- Exams
  - Midterms (2) and final exam

Grading

- Participation: 26%
- Homework: 40%
- Midterms: 24%
- Final: 10%
Classroom Interaction Preferences

- **Uses**
  - Feedback to instructor: Do you understand what I’m trying to present, or do I need to go over it again?
  - Participation: Did you respond?
  - Knowledge and Understanding: Did you get it right?
    
    *Only a few questions will be scored, in many cases I may ask questions to get you thinking and don’t expect you to know the answer.*

- **Methods**
  - iClicker
  - HotSeat
  - TopHat