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 emails, 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
    relevant
- 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
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

Global warming has been a popular topic among scientists

World’s oldest bling: two tiny 100,000-year-old shells

News Categories

Top Stories
- World
- U.S.
- Business
- SciTech
- Sports
- Entertainment
- Health

Most Popular

© 2017 Christopher W. Clifton
IR Applications: Text Categorization

Medical Subject Headings (Categories):
- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques [E]
- Psychology and Psychotherapy [F]
- Psychological Phenomena and Mental Disorders [G]
- Behavioral Sciences and Behavior Alchemy [H]

IR Applications: Document Clustering

Clusty

Top 265 results of at least 94,536,061 retrieved for the query Java (Details)

1. Java Technology
   - Sun's home for Java. Offers Windows, Solaris, and Linux Java Development Kits (JDKs).
   - Product Information:
     - Java.sun.com
     - JavaSoft

2. Java Programming Language
   - Java is an object-oriented programming language developed initially by Sun Microsystems. Initially called Oak (named after office), it was intended to replace C++. Although the feature set bet

© 2017 Christopher W. Clifton
IR Applications: Content Based Filtering

Keyword Matching

IR Applications: Collaborative Filtering

Other Customers with similar tastes

Introduction to Information Retrieval [Hardcover]

Customers Who Bought This Item Also Bought

Foundations of Statistical Natural Language Processing (1st) by Christopher D. Manning $16.60

Algorithms of the Intelligent Web by Michael Hernandez $24.74

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition) by Trevor Hastie $59.00

Speech and Language Processing (2nd Edition) by Daniel Jurafsky $101.17
IR Applications: Information Extraction

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.

- Recognize Relationship between entities

What type of effect of gliclazide on this patient with diabetes

- Inference based on the relationship between entities

Inherited Disease → Gene → Chemical

Drug discovery

IR Applications: Question Answering

Web Search

- Richard M. Nixon
  Biography of Richard M. Nixon, the thirty-seventh President of the United States (1969-1974).
  www.whitehouse.gov/history/presidents/n37.html • Cached • Save

- Biography of Harry S. Truman
  Biography of Harry S. Truman, the thirty-third President of the United States (1945-1953).
  www.whitehouse.gov/history/presidents/t33.html • Cached • Save

More Results from www.whitehouse.gov

President Elect - 1972
This site makes use of cascading style sheets. If you're reading this you need to either upgrade your browser or make sure...

Home State, PR, NY, ...
www.presidentelect.org/e1972.html • Cached • Save
IR Applications: Web Search

Crawled into a centralized database

IR Applications: Federated Search

Valuable → Search by Federated Search
IR Applications: Expertise Search

INDURE: Indiana database of university research expertise
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

Citation/Link Analysis

- 24 citations found. Only retrieving 100 documents.

CiteSeer

This paper is cited in the following contexts:

- A Probabilistic Model of Gaze Imitation and Shared Attention - Matthew Hoffman Card
  - Next 50

IR Applications: Multimedia Retrieval

- Query
- Color Histogram
- Wavelet...
- Feature Extraction
- Retrieval Model
- Feature Extraction
- Pictures
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:
  – Linjie Li, li2477@purdue.edu, office hours: TBD
  – Hogun Park, hogun@purdue.edu, office hours: TBD
  – Rajkumar Pujari, rpujari@purdue.edu, office hours: TBD
• Webpage: http://www.cs.purdue.edu/~clifton/cs47300/
• Email list: fall-2017-cs-47300-le1@lists.purdue.edu
• Piazza signup: piazza.com/purdue/fall2017/cs47300
• Prerequisites: CS25100. Having had a Stat course (e.g., STAT 35000) will help.

Readings

Workload

• Homeworks
  – 4-5 written assignments
  – 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%