Why this course?

- Managing Data is one of the primary uses of computers
- Most of this data is NOT contained in structured databases
  - Therefore, no carefully structured queries
- How do we find this?
  
  Information Retrieval
Information Retrieval: Challenges

• Data is unstructured
  – Need to guess what is 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*

Course Information

• Contact Information: Professor Clifton
  – Office: LWSN 2142F, x4-6006.
  – clifton@cs.purdue.edu

• Teaching Assistant: Balamurugan Anandan
  – Office: LWSN 2149
  – banandan@cs.purdue.edu

• Office Hours: M 15:00-16:00, W 15:30-16:00, R 17:00-17:30
  – LWSN 2142F or LWSN 2149
  – WebEx

• Course Web Page: [http://www.cs.purdue.edu/homes/clifton/cs54701](http://www.cs.purdue.edu/homes/clifton/cs54701)
Course Methodology

• Lectures to present the concepts
  – Unless otherwise noted, all the material you are expected to know will be covered in class
  – Interaction (questions/discussion/thinking) encouraged

• Reading will fill in the details
  – Other readings (e.g., research literature) will be made available where appropriate

• Homework and Projects get you to understand what you’ve read and heard
  – 3-5 written homeworks
  – 3-4 programming projects, final project
  – Suggested exercises from the text (ungraded)

Evaluation and Grading

• Points earned as follows:
  – Midterm (15%)
  – Final Exam during Finals week (25%)
  – Homeworks / projects (55%)
    • Larger projects may be given higher weight (final project 25%)
  – Instructor’s evaluation (5%)
    • In-class discussions/participation
    • Out of class discussions, email
    • Overall perception of quality of your work in ways that may not be reflected in your scores

• Late work penalized 10% per day
• Qualifying Exam: One hour supplement to regular exam
  – Passing the qualifier requires both suitable performance in the course and on the supplemental exam

For more details see the course web page
Why Information Retrieval:

Information Overload:

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

![WWW Growth](image)

Why Information Retrieval:

Information Retrieval (IR) mainly studies unstructured data:

Text in Web pages or emails; image; audio; video; protein sequences.

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.

Unstructured data:

No structure: no primary key as in RDBMS

Semantic meaning unknown: natural language processing systems try to find the meaning in the unstructured text
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

RDBMS and IR get close to each other

RDBMS -> IR
- Combine exact search and inexact text search
  Find an article published between 1999 and 2004 that talks about Oracle and Internet.

IR -> RDBMS
- Use information extraction to convert unstructured data to structured data: extract company names and their headquarter locations from news
- Semi-structured representation: XML data; queries with structured information
Some core concepts of IR

- Information Need
- Representation
- 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
IR Applications: Text Categorization

Medical Subject Headings (Categories)

- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic [E]
- Psychiatry and Psychology [P]
  - Behavior and Behavior Mechanisms
  - Psychological Phenomena
  - Mental Disorders [R53]
  - Behavioral Disabilities [R65]
- Biological Sciences [G]

IR Applications: Text Categorization
IR Applications: Document Clustering

IR Applications: Content Based Filtering

Keyword Matching
IR Applications: Collaborative Filtering

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.

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

Inherited Disease → Gene → Chemical

Drug discovery
• IBM DeepQA!!
IR Applications: Federated Search

Valuable Searched by Federated Search

IR Applications: Expertise Search

INDURE: Indiana database of university research expertise

www.indure.org
IR Applications: Citation/Link Analysis

This paper is cited in the following contexts:

First 50 documents Next 50

A Probabilistic Model of Gaze Imitation and Shared Attention - Matthew Hoffman (David) (Correct)

No context found.

IR Applications: Multimedia Retrieval

- Query
- Color Histogram
- Wavelet...
- Feature Extraction
- Retrieval Model
- Feature Extraction
- Pictures

IR Applications: Information Visualization

Partial Structure of pages from a Web subset visualized by Mapuccino
Assignments (30%)

- Written assignments
- Algorithm design and implementation (about 3 assignments)
  - Implement and improve common retrieval algorithms
  - Create and compare algorithms for information retrieval applications (email spam detection and recommendation system)
- Late submission
  - 10% deduction per day (24 hours)
- Discussion encouraged
  - But work submitted should be your own
  - *If given a similar problem, would you be able to do it?*
  - If unsure, document discussion

Project (25%)

- Goal
  - Show your knowledge and creative ideas on real applications
  - Leading to research report/publication (optional)
- Topics
  - Suggested by the lecturer or any related topic proposed by you
- Project progress
  - Project proposal
  - Project final report and presentation
Course Description

• Introduce core concepts of information retrieval (what is behind search engines like Google)
• Wide coverage of many information retrieval applications (e.g., text categorization, filtering systems)
• Get hands on experience by developing practical systems/components (e.g., email spam detection)
• Prepare students for doing cutting-edge research in information retrieval and related fields
• Open the door to the amazing job opportunities in Search Technology and E-commerce companies