Overview

Web

Growth of the Web

"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 Information Retrieval:

Information Retrieval (IR) mainly studies unstructured data:

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

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

Web 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
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
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.
IR Applications: Web Search

Web

Crawled into a centralized database

IR Applications: Federated Search

Federated Search

IR Applications: Expertise Search

INDURE: Indiana database of university research database

IR Applications: Citation/Link Analysis

Citation/Link Analysis

IR Applications: Multimedia Retrieval

Multimedia Retrieval
IR Applications: Information Visualization

Partial Structure of pages from a Web subset visualized by Mapuccino

Grading Policy:

Project (30%):
- 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

Grading Policy:

Assignments: 30%
- Project: 30%
- Final exam: 30%
- Class attendance: 10%

Grading Policy:

Test(s) (30%):
- One or two tests? In class or not?
- Based on lecture contents (more) and required reading materials (less)
- Review session

Attendance (10%):
- Be interactive: the best way to learn is to ask questions
- Insightful questions/suggestion gives extra credit

Support System:

Course web page:
- Schedule, slides, reading materials, assignments, etc.

Textbook:
- Online free version
- Other recommended readings: on the course web page

Office hour:
- Wednesday 2:00-3:00 PM
- or reach me by: lsi@cs.purdue.edu

Grading Policy:

Assignments (30%):
- Algorithm design and implementation (about 3 assignments)
  - Implement and improve common retrieval algorithms
  - Create and compare algorithms for information retrieval applications (web page/email spam classification and recommendation system)
- Late submission
  - 90% credit for next two days, 50% afterwards
  - You may help each other by discussion (please indicate so in the submission), but copying/cheating may result in 0 credit
  - It is safe to start early…

Grading Policy:

Grading Policy:

Support System:
Course Description:

The Goal

- 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.

Lecture Review:

- Core concepts of information retrieval
  Query representation; document representation; retrieval model; evaluation

- Applications of information retrieval
  Web Search; Text Categorization; Document Clustering; Information Recommendation; Information Extraction; Question Answering....

Grade Policy
Assignments: 30%; Project: 30%; Final Exam: 30%; Class attendance: 10%