CS-490W
Web Information Search and Management

Luo Si
Department of Computer Science
Purdue University

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)

---

**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
- ...................
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 e-mails, 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
Some core concepts of IR

Information Need

Query → Retrieval Model → Indexed Objects

Retrieved Objects

Representation

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

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

News Categories

SciTech

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 1,000 years, according to the National Academy of Sciences ...

National panel supports 90% global warming evidence Boston Globe

No More Crashing Global Temp Trend Crown Free Press
Guardian Unlimited - Scotsman Times - Reuters - Forbes - all 411 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

Tiny shell may be world’s oldest bauble MORE »

Researchers identify what may be oldest known jewelry Voice of America

BBC News - New York Times - People’s Daily Online - Telegraph.co.uk - all 79 related »
IR Applications: Text Categorization

Medical Subject Headings (Categories)

- Anatomy [A]
- Organizations [O]
- Disease [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques [E]
- Psychiatry and Psychology [F]
- Behavior and Behavior Mechanisms [BF]
- Psychological Phenomena and Processes [BP]
- Mental Disorders [MPD]
- Biological Sciences [B]

IR Applications: Document Clustering

Java

Top 265 results of at least $1.536.691$ retrieved for the query Java (Details)

- Sun offers Java beta release (Amelia already), Jan. 16, 2006
- Java errors fixed, Free Java version all fixed instantly: Free download available now (dhtmlx)
- Free JSP Editor - Eclipse (BEA Workshop Short for JSP, Eclipse JSP IDE, EJ63: Download now (www.bae.com

1. Java Technology [N]
   - Sun’s home for Java: Offers Windows, Solaris, and Linux Java Development Kits (JDK) product information
   - java.sun.com (Eclipse, J2SE, J2EE, J2ME, OpenJDK)

2. Java programming language [N]
   - Java is an object-oriented programming language developed initial colleagues at Sun Microsystems. Initially called OAK (named after office), it was intended to replace C++ although the feature set was
   - java.sun.com (Eclipse, J2SE, J2EE, J2ME, OpenJDK)
IR Applications: Content Based Filtering

Keyword Matching

Other Customers with similar tastes

IR Applications: Collaborative Filtering

Modern Information Retrieval (Paperback)
By Ricardo Baeza-Yates, Bernd K.T. Suster, Jon M. Kleinberg
Look Inside

Customers who bought this item also bought:

- Managing Gigabytes: Compressing and Indexing Documents and Images (The Morgan Kaufmann Series in Information Systems) by Jan M. Witten
- Mining the Web: Analysis of Hypertext and Semi-Structured Data (The Morgan Kaufmann Series in Data Management Systems) by Anand R. Das and Suresh C. Surana
- Foundations of Statistical Natural Language Processing by Christopher D. Manning
- Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series) by David A. Garnier
- Information Retrieval: Data Structures and Algorithms by William A. Frakes
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**

  ![Diagram](image)

IR Applications: Question Answering

Who is the president of united state in 1972

- **Richard M. Nixon**
  Biography of Richard M. Nixon, the thirty-seventh President of the United States (1969-1974).
  [www.whitehouse.gov/history/presidents/nixon.html](http://www.whitehouse.gov/history/presidents/nixon.html) - Cached - Save

- **Harry S. Truman**
  Biography of Harry S. Truman, the thirty-third President of the United States (1945-1953).
  [www.whitehouse.gov/history/presidents/truman.html](http://www.whitehouse.gov/history/presidents/truman.html) - Cached - Save

More Results from www.whitehouse.gov
IR Applications: Web Search

Crawled into a centralized database

IR Applications: Federated Search

Valuable → Searched by Federated Search
IR Applications: Expertise Search

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

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

CiteSeer: Only retrieving 1000 documents.


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 contexts found.


Latest Variables Models for Semantic Orientations of Phrases - Hanno Takamura Takamura (Correct)

IR Applications: Multimedia Retrieval

Query

Color Histogram Wavelet...

Feature Extraction

Retrieval Model

Picture

Feature Extraction
IR Applications: Information Visualization

Partial Structure of pages from a Web subset visualized by Mapuccino

Grading Policy:

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

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:

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:
- Introduction to Information Retrieval (Manning, C.; Raghavan, P.; Schütze, H.
  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
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%