CS54701: Information Retrieval

CS-54701
Information Retrieval

Luo Si
Department of Computer Science
Purdue University
Overview of Information Retrieval
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 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
IR vs. RDBMS

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
IR and other disciplines

- Information Retrieval
- Machine Learning
- Pattern Recognition
- Statistical Learning
- Natural Language Processing
- Image Understanding
- Information Extraction
- Text Mining
- Database
- Data Mining
- Medical informatics
- Bioinformatics
- Visualization
- Library & Info Science
- Security System
- System Support
- Theory
- Deep Analysis
Some core concepts of IR

Information Need

Query → Retrieval Model → Retrieved Objects → Representation → Returned Results → Evaluation/Feedback

Indexed Objects
Some core concepts of IR

**Multiple Representation**

Information Retrieval Systems

Information Retrieval Systems ... Okapi Pack · University of Massachusetts Center for Intelligent Information Retrieval · Callan CMU IR Group ...

Past Performance Information Retrieval System

Welcome to the Past Performance Information Retrieval System (PPIRS). PPIRS is a web-enabled, government-wide application that provides timely and pertinent ...

Electronic Digital Information Source (EDIS)

EDIS is the Electronic Data Information Source of UF/IFAS Extension, a collection of information on topics relevant to you: profitable and sustainable ...

Information retrieval - Wikipedia, the free encyclopedia

Automated information retrieval (IR) systems were originally used to manage information explosion in scientific literature in the last few decades. ...
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
- ..................................
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 98 global warming evidence Boston Globe

No More Dodging Global Temp Threat Detroit Free Press

Guardian 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.

Tiny shells may be world’s oldest beads MSNBC

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 »
1. Papadopoulos N et al. Mutation of a mutL homolog in... [PMID: 8128251]

PMID- 8128251
OWN - NLM
STAT- MEDLINE
DA - 19940413

TI - Mutation of a mutL homolog in hereditary colon cancer.
PG - 1625-9

AB - Some cases of hereditary nonpolyposis colorectal cancer (HNPCC) are due to alterations in a mutS-related mismatch repair gene. A search of a large database of expressed sequence tags derived from random complementary DNA clones revealed three additional human mismatch repair genes, all related to the bacterial mutL gene. One of these genes (hMLH1) resides on chromosome 3p21, within 1 centimorgan of markers previously linked to cancer susceptibility in HNPCC kindreds. Mutations of hMLH1 that would

MH - Amino Acid Sequence
MH - *Chromosomes, Human, Pair 3
MH - Codon
MH - Colorectal Neoplasms, Hereditary Nonpolyposis/*genetics
IR Applications: Document Clustering

Top 265 results of at least 94,536,601 retrieved for the query Java (Details)

News

- Sun offers Java beta release (Yahoo! News) Jun 21, 2006

java Errors Fixed Free
java errors all Fixed instantly! Free download available Now dlffix.net

Free JSP Editor - Eclipse
BEA Workshop Studio for JSP, Struts JSF IDE, EJB3 - Download now! www.bea.com

1. Java Technology
Sun's home for Java. Offers Windows, Solaris, and Linux Java Development Kits (JDKs) product information.
java.sun.com - [cache] - Ask, Gigablast, MSN, Wisenut

2. Java programming language
Java is an object-oriented programming language developed initially by colleagues at Sun Microsystems. Initially called Oak (named after the office), it was intended to replace C++, although the feature set between Java should not be confused with JavaScript, which shares only the syntax. Sun Microsystems currently maintains and updates Java.
IR Applications: Content Based Filtering

Keyword Matching

Google

News

United States edition

Top Stories

Syria Kills 25 as UN Officials Consider Crackdown's Legality

San Francisco Chronicle - 19 minutes ago

(Updates with statements by Obama, Cameron, Merkel and Sarkozy and UN report starting in first paragraph. See EXTRA and MET for more on the regional turmoil.

RAW DATA: Background on Syria's Crackdown

Fox News

EU says Assad must step down, plans more sanctions

Reuters

From Syria: Diplomatic Pressure on Syria

Day Press News

Opinion: Why the US should speak out for freedom in Syria

Washington Post

See all 3,353 sources »

Stocks sink worldwide, sending Dow down 470 points

Los Angeles Times - 1 hour ago

Global stock markets plunged anew Thursday amid intensifying concerns about Europe's debt crisis and a batch of disheartening economic reports in the United States.
IR Applications: Collaborative Filtering

Other Customers with similar tastes

- **Foundations of Statistical Natural Lang...** by Christopher D. Manning
  - Rating: ★★★★★ (13)
  - Price: $56.69

- **Algorithms of the Intelligent Web** by Haralambos Marmanis
  - Rating: ★★★★★ (12)
  - Price: $24.74

- **The Elements of Statistical Learning: Data Minin...** by Trevor Hastie
  - Rating: ★★★★★ (45)
  - Price: $59.09

- **Speech and Language Processing (2nd Edition)** by Daniel Jurafsky
  - Rating: ★★★★★ (27)
  - Price: $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.

  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
IR Applications: Question Answering

IBM DeepQA!!
IR Applications: Web Search

Google

Department of Computer Science, Purdue University
Dr. William Gorman, Purdue Computer Science’s Assistant to the Department Head, was honored with a service award from the Purdue Co-Op office. ...
www.cs.purdue.edu/ - 13k - Cached - Similar pages

CS Topic Generator
Computer Science is facing a major roadblock to further research. The problem is most evident with students, but afflicts many researchers as well: people ...
www.cs.purdue.edu/homes/dec/essay.topic.generator.html - 6k - Cached - Similar pages
[ More results from www.cs.purdue.edu ]

Department of Computer and Information Science :: IUPUI
School of Science · IUPUI Main Page · Purdue University · Indiana University · Certificate of Applied Computer Science for the Liberal Arts Major ...
www.cs.iupui.edu/ - 9k - Cached - Similar pages

ETCS - The School of Engineering Technology and Computer Science
Computer Labs · Contact ETCS · The College of ETCS wins Berger Award · IPFW Engineering Commercial - Quicktime Format ...
www.etcs.ipfw.edu/ - 19k - Cached - Similar pages
IR Applications: Federated Search
INDURE: Indiana database of university research database

www.indure.org

Welcome to INDURE
Indiana Database of University Research Expertise

Version 1.0 August 15th, 2008
IR Applications: Citation/Link Analysis

Linear Collider Accelerator In Japan

U.S. Government Lab

Nobel Prize Organization
2457 citations found. 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 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
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 (email spam detection 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. Cambridge University Press
- Other readings on the course web site

Office hour:
- TA office hour: TBD; Lecturer office hour: Tue 10:30-11:30
- Dzung Hong and Ravi Kiran Bukka
Course Description:

The Goal

- Introduce core concepts of information retrieval (what is behind search engine 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
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%