PURDUE
UNIVERSITY

CS54701: Information Retrieval

Course Overview

12 January 2016

Prof. Chris Clifton

Indiana
Center for
Database
Systems
™

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 ~~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

Course Information

- Contact Information: Professor Clifton
  - – Office: LWSN 2142F, x4-6005.
  - – 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

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
  - Text: Introduction to Information Retrieval, by Manning, C.; Raghavan, P.; Schütze, H. Cambridge University Press, 2008, ISBN 0521865719
  - 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

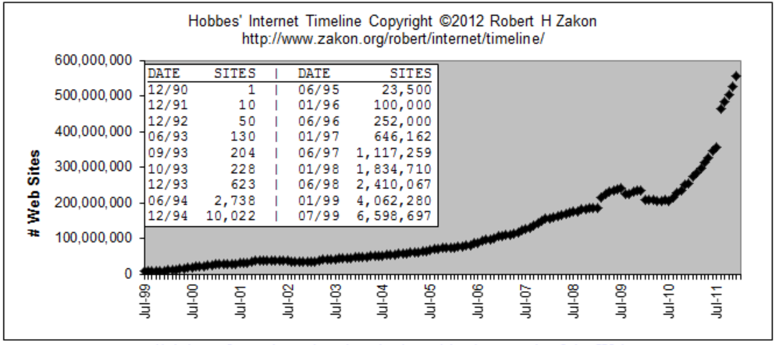
For more details see the course web page

Why Information Retrieval:

Information Overload:

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

Figure: WWW Growth



DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

Why Information Retrieval:

Information Retrieval (IR) mainly studies unstructured data:

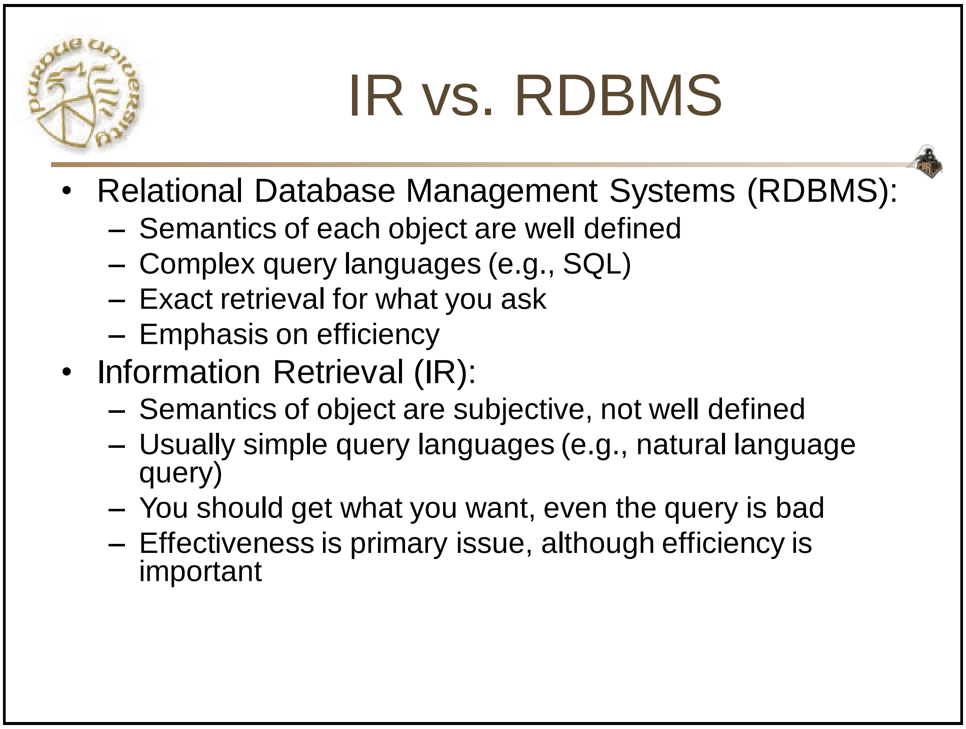
Text in Web pages or emails; image; audio; video; protein sequences..

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:

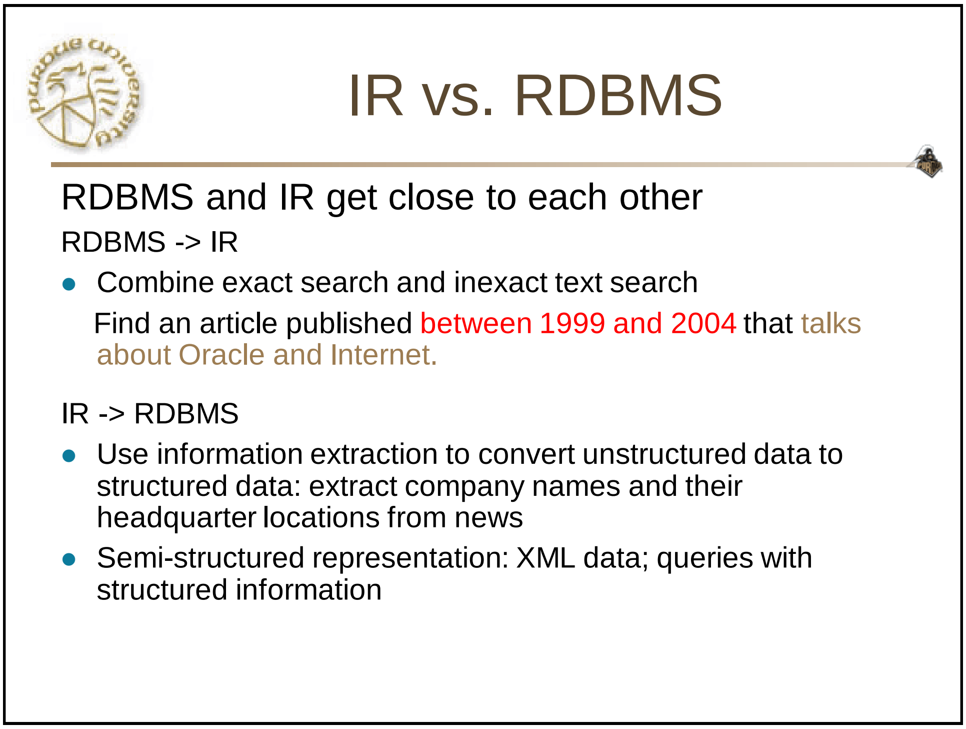
No structure: no primary key as in RDBMS

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

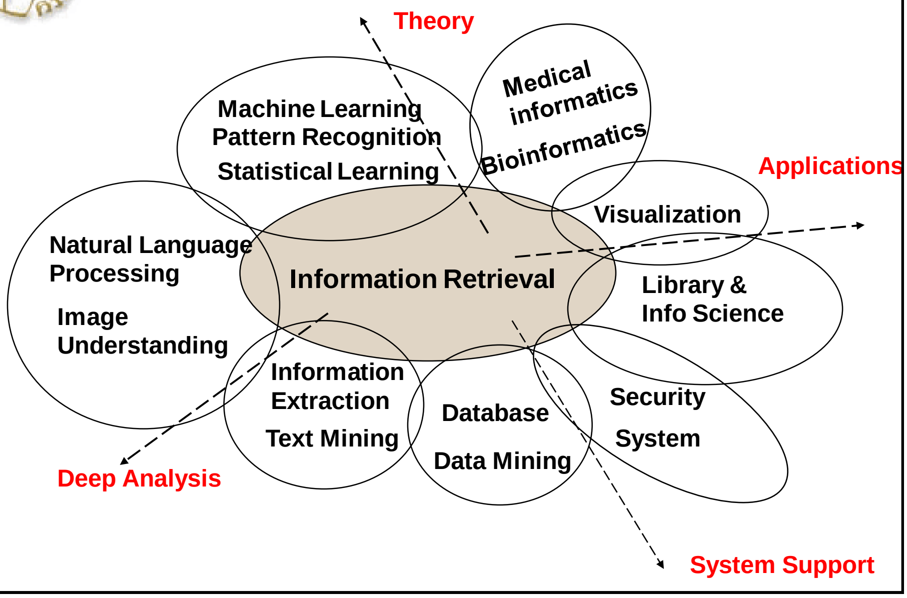
- Combine exact search and inexact text search

  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



Theory

Machine Learning
Pattern Recognition
Statistical Learning

Medical informatics
Bioinformatics

Applications

Visualization

Natural Language Processing

Image Understanding

Information Retrieval

Library & Info Science

Information Extraction
Text Mining

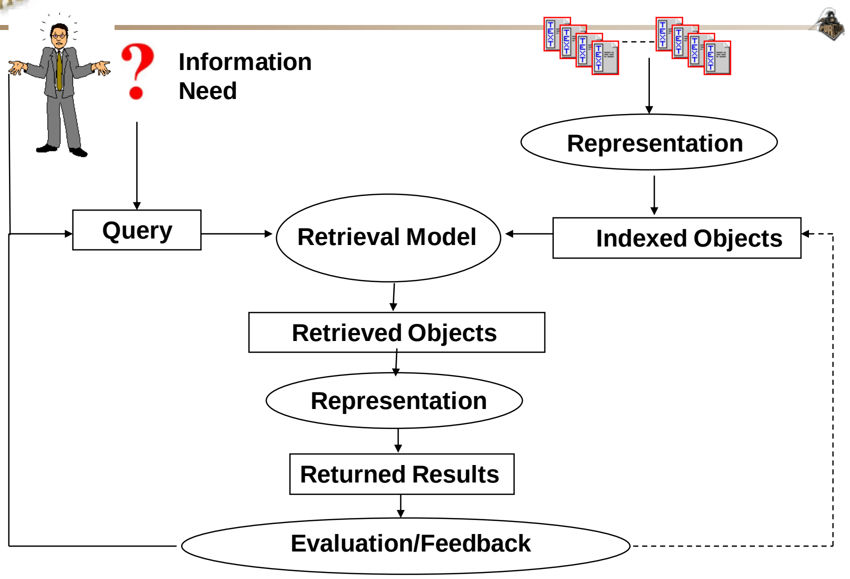
Database
Data Mining

Security
System

Deep Analysis

System Support

Some core concepts of IR



Information Need

Representation

Query

Retrieval Model

Indexed Objects

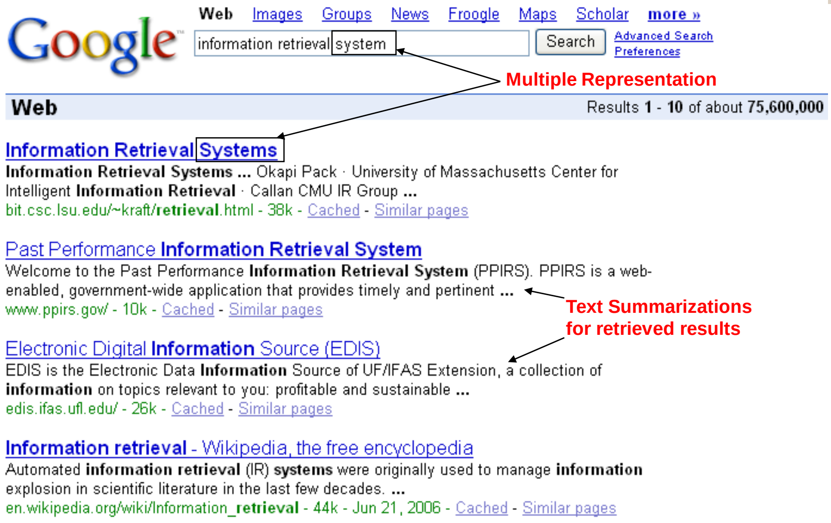
Retrieved Objects

Representation

Returned Results

Evaluation/Feedback

Some core concepts of IR



Web Images Groups News Froogle Maps Scholar more »

information retrieval system

Search

Advanced Search
Preferences

Multiple Representation

Web Results **1** - **10** of about **75,600,000**

Information Retrieval Systems
Information Retrieval Systems ... Okapi Pack · University of Massachusetts Center for Intelligent **Information Retrieval** · Callan CMU IR Group **...**
bit.csc.lsu.edu/~kraft/**retrieval**.html - 38k - Cached - Similar pages

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 **...**
www.ppirs.gov/ - 10k - Cached - Similar pages

Text Summarizations for retrieved results

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 **...**
edis.ifas.ufl.edu/ - 26k - Cached - Similar pages

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. **...**
en.wikipedia.org/wiki/Information_**retrieval** - 44k - Jun 21, 2006 - Cached - Similar pages

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)

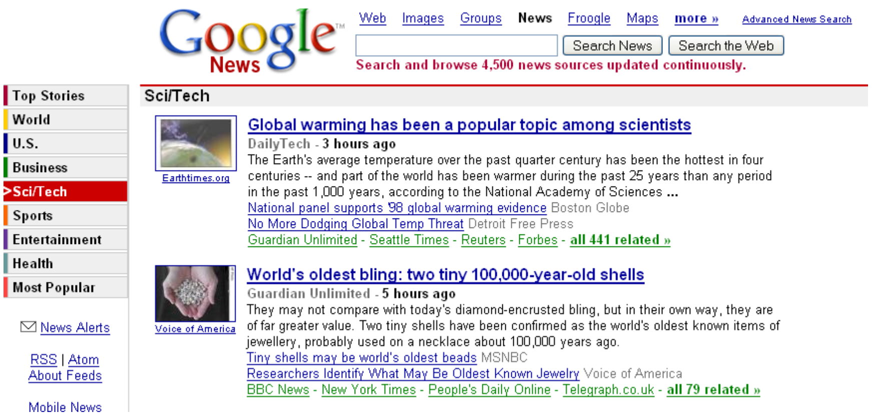
IR Applications

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

News Categories



Google News

Web Images Groups **News** Froogle Maps **more »** Advanced News Search

Search News | Search the Web

Search and browse 4,500 news sources updated continuously.

- Top Stories
- World
- U.S.
- Business
- >Sci/Tech
- Sports
- Entertainment
- Health
- Most Popular

News Alerts

RSS | Atom
About Feeds

Mobile News

Sci/Tech

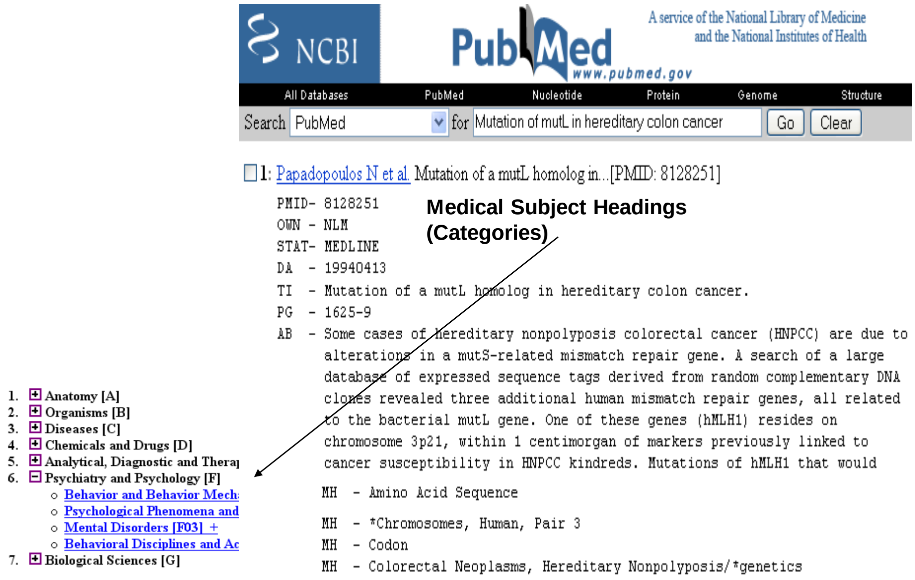
Earthtimes.org

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 »**

Voice of America

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 »**

IR Applications: Text Categorization



Mutation of mutL in hereditary colon cancer

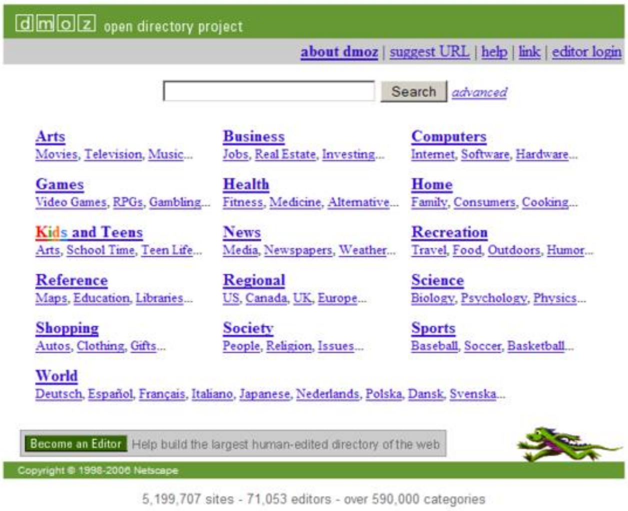
□ 1: Papadopoulos N et al. Mutation of a mutL homolog in...[PMID: 8128251]

Medical Subject Headings (Categories)

```
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
```

1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊞ Diseases [C]
4. ⊞ Chemicals and Drugs [D]
5. ⊞ Analytical, Diagnostic and Thera
6. ⊟ Psychiatry and Psychology [F]
   - ○ Behavior and Behavior Mech
   - ○ Psychological Phenomena and
   - ○ Mental Disorders [F03] +
   - ○ Behavioral Disciplines and Ac
7. ⊞ Biological Sciences [G]

IR Applications: Text Categorization


dmoz open directory project
about dmoz | suggest URL | help | link | editor login
Search
advanced
Become an Editor

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Autos, Clothing, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

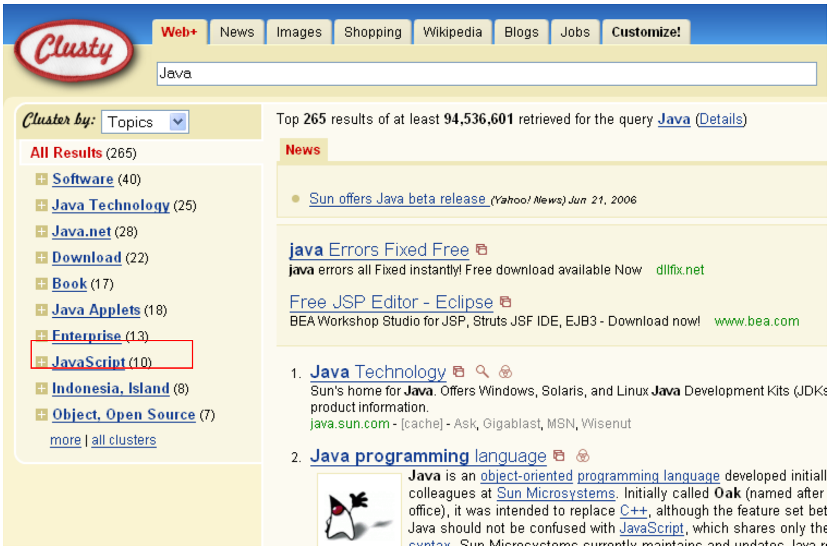
World
Deutsch, Español, Français, Italiano, Japanese, Nederlands, Polska, Dansk, Svenska...

Help build the largest human-edited directory of the web

Copyright © 1998-2006 Netscape

5,199,707 sites - 71,053 editors - over 590,000 categories

IR Applications: Document Clustering



Web+ | News | Images | Shopping | Wikipedia | Blogs | Jobs | Customize!

Java

Cluster by: Topics

All Results (265)
- Software (40)
- Java Technology (25)
- Java.net (28)
- Download (22)
- Book (17)
- Java Applets (18)
- Enterprise (13)
- JavaScript (10)
- Indonesia, Island (8)
- Object, Open Source (7)

more | all clusters

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

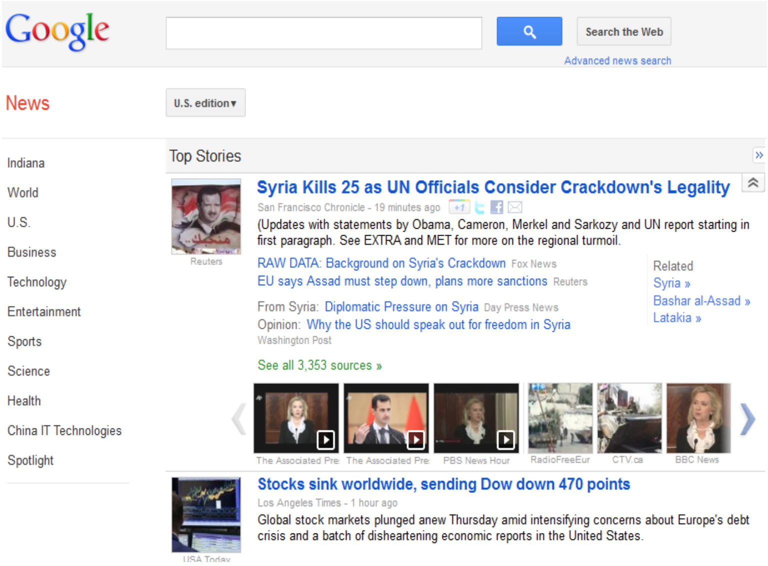
java Errors Fixed Free
java errors all Fixed instantly! Free download available Now dllfix.net

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 initiall
colleagues at Sun Microsystems. Initially called Oak (named after
office), it was intended to replace C++, although the feature set bet
Java should not be confused with JavaScript, which shares only the

IR Applications: Content Based Filtering



Search the Web

Advanced news search

News

U.S. edition▾

- Indiana
- World
- U.S.
- Business
- Technology
- Entertainment
- Sports
- Science
- Health
- China IT Technologies
- Spotlight

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.

Reuters

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

Related
Syria »
Bashar al-Assad »
Latakia »

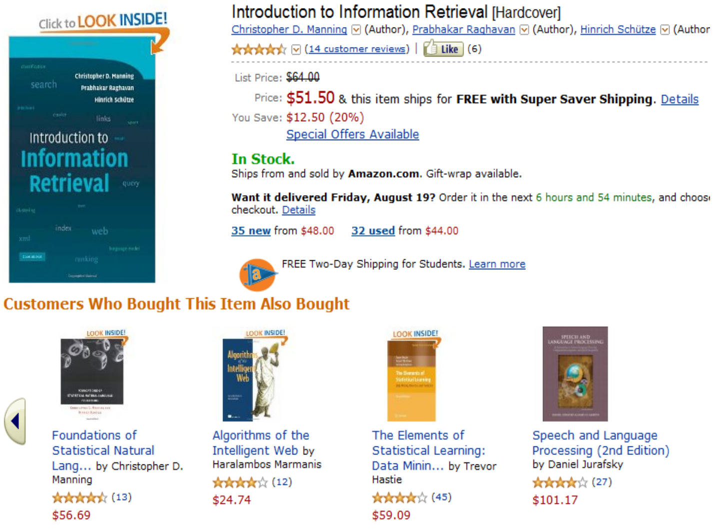
See all 3,353 sources »

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.

USA Today

Keyword Matching

IR Applications: Collaborative Filtering



Click to LOOK INSIDE!

Introduction to Information Retrieval [Hardcover]

Christopher D. Manning (Author), Prabhakar Raghavan (Author), Hinrich Schütze (Author

(14 customer reviews) | Like (6)

List Price: ~~$64.00~~

Price: $51.50 & this item ships for **FREE with Super Saver Shipping**. Details

You Save: $12.50 (20%)

Special Offers Available

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Friday, August 19? Order it in the next 6 hours and 54 minutes, and choos checkout. Details

35 new from $48.00   32 used from $44.00

FREE Two-Day Shipping for Students. Learn more

Customers Who Bought This Item Also Bought

Foundations of Statistical Natural Lang... by Christopher D. Manning (13) $56.69

Algorithms of the Intelligent Web by Haralambos Marmanis (12) $24.74

The Elements of Statistical Learning: Data Minin... by Trevor Hastie (45) $59.09

Speech and Language Processing (2nd Edition) by Daniel Jurafsky (27) $101.17

Other Customers with similar tastes

IR Applications: Information Extraction

Bring structure and semantic meaning to text:

- Entity detection

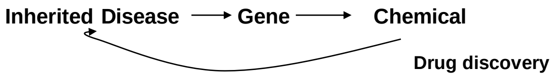
An 80-year-old woman with diabetes mellitus was treated with gliclazide. Prior to the gliclazide administration, her urinary excretion of albumin, serum urea nitrogen and serum creatinine were normal. After the medication, oliguria, edema and azotemia developed. On the twenty-fourth day when the edema was severe and generalized, gliclazide administration was terminated.

Diabetes: entity of disease   gliclazide: entity of drug

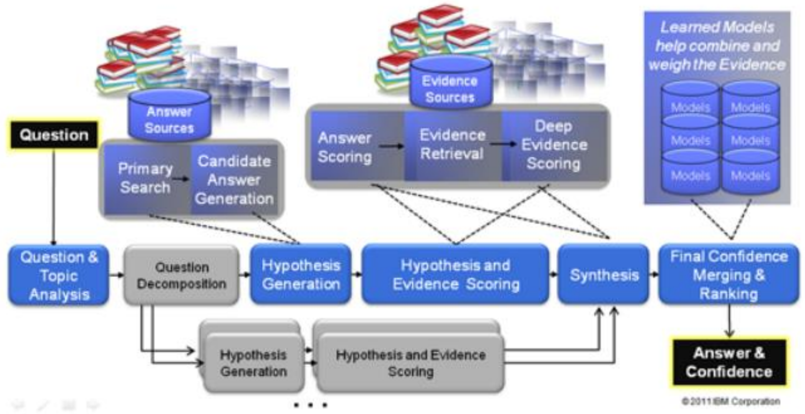
- Recognize Relationship between entities

What type of effect of gliclazide on this patient with diabetes

- Inference based on the relationship between entities

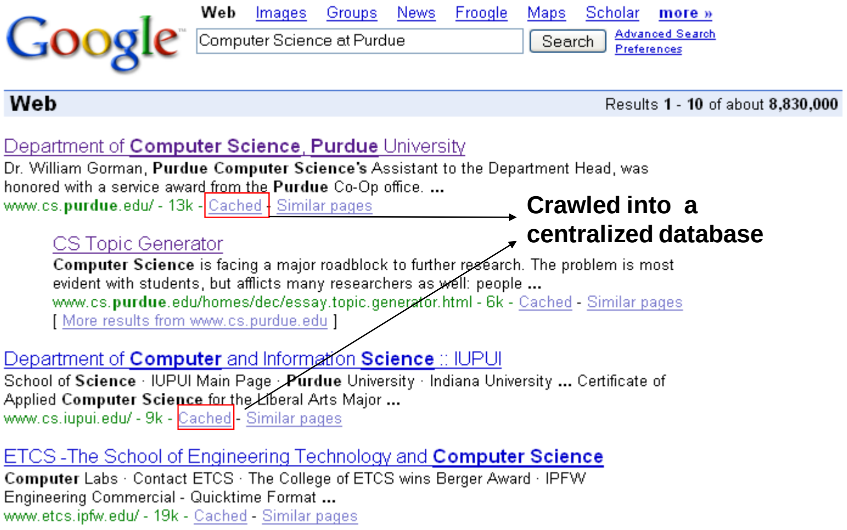


IR Applications: Question Answering



- IBM DeepQA!!

IR Applications: Web Search



Computer Science at Purdue

Web Results **1** - **10** of about **8,830,000**

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

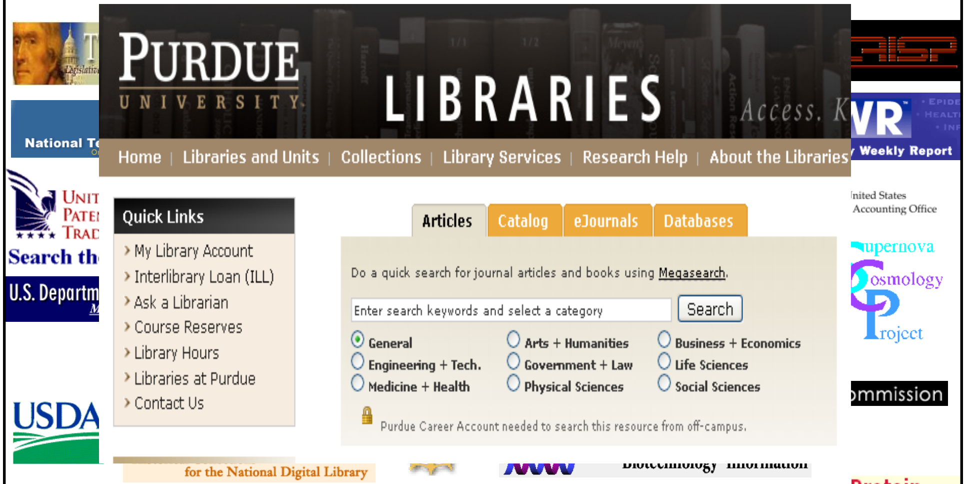
CS Topic Generator
Computer Science is facing a major roadblock to further research. The problem is most evident with students, but afflicts many researchers as well: people **...**
www.cs.**purdue**.edu/homes/dec/essay.topic.generator.html - 6k - Cached - Similar pages
[More results from www.cs.purdue.edu]

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

Crawled into a centralized database

IR Applications: Federated Search



National T
UNIT PATE TRAD
Search th
U.S. Departm
USDA
Weekly Report
nited States Accounting Office
upernova Cosmology Project
mmission
for the National Digital Library
Biotechnology Information
Protein Interaction Search:

PURDUE UNIVERSITY LIBRARIES Access. K

Home | Libraries and Units | Collections | Library Services | Research Help | About the Libraries

Quick Links

- My Library Account
- Interlibrary Loan (ILL)
- Ask a Librarian
- Course Reserves
- Library Hours
- Libraries at Purdue
- Contact Us

Articles | Catalog | eJournals | Databases

Do a quick search for journal articles and books using Megasearch.

Enter search keywords and select a category [Search]

● General ○ Arts + Humanities ○ Business + Economics
○ Engineering + Tech. ○ Government + Law ○ Life Sciences
○ Medicine + Health ○ Physical Sciences ○ Social Sciences

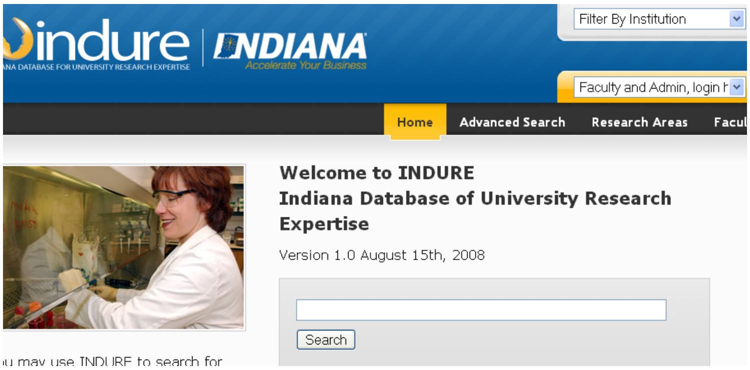
Purdue Career Account needed to search this resource from off-campus.

Valuable → **Searched by *Federated Search***

IR Applications: Expertise Search

INDURE: Indiana database of university research expertise

www.indure.org



indure | INDIANA Accelerate Your Business
INDIANA DATABASE FOR UNIVERSITY RESEARCH EXPERTISE

Filter By Institution
Faculty and Admin, login h

Home | Advanced Search | Research Areas | Facul

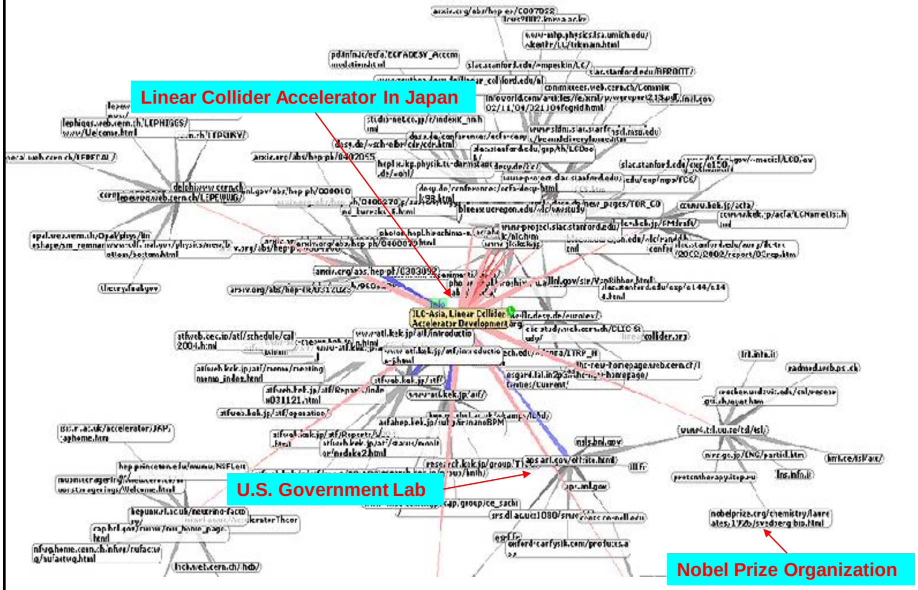
Welcome to INDURE
Indiana Database of University Research Expertise

Version 1.0 August 15th, 2008

[Search]

u may use INDURE to search for

IR Applications: Citation/Link Analysis

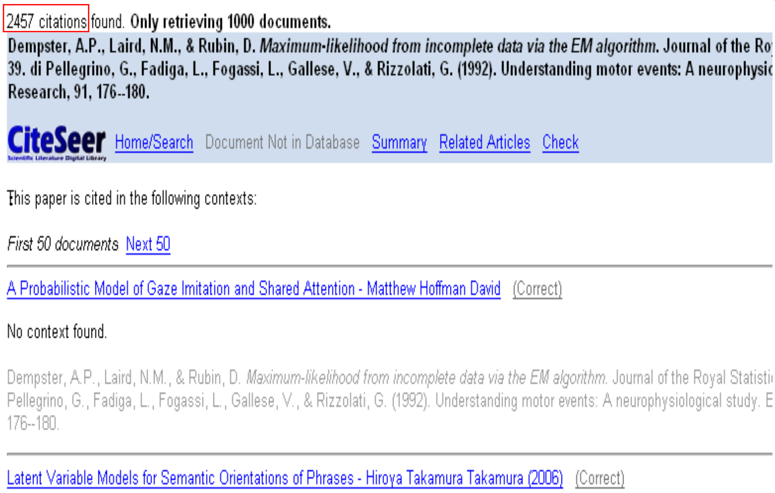


Linear Collider Accelerator In Japan

U.S. Government Lab

Nobel Prize Organization

IR Applications: Citation/Link Analysis



2457 citations found. Only retrieving 1000 documents.

Dempster, A.P., Laird, N.M., & Rubin, D. *Maximum-likelihood from incomplete data via the EM algorithm.* Journal of the Ro
39. di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolati, G. (1992). Understanding motor events: A neurophysic
Research, 91, 176--180.

CiteSeer Home/Search Document Not in Database Summary Related Articles Check

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)

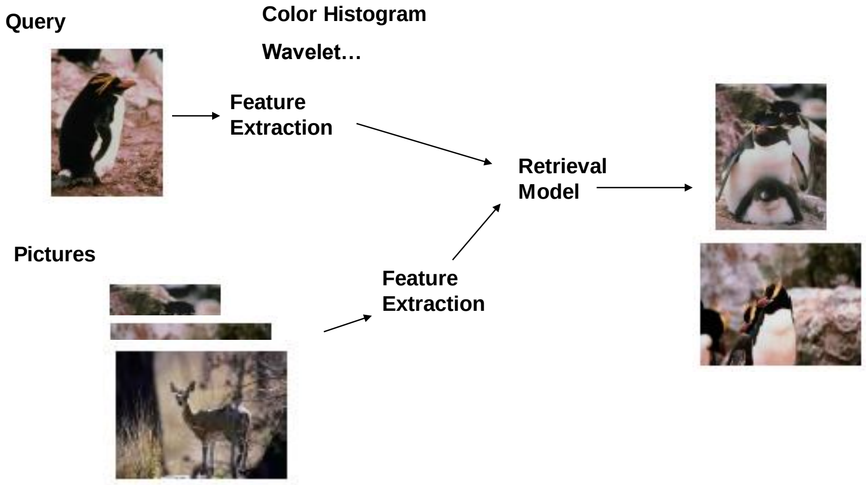
No context found.

Dempster, A.P., Laird, N.M., & Rubin, D. *Maximum-likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statisti
Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolati, G. (1992). Understanding motor events: A neurophysiological study. E
176--180.

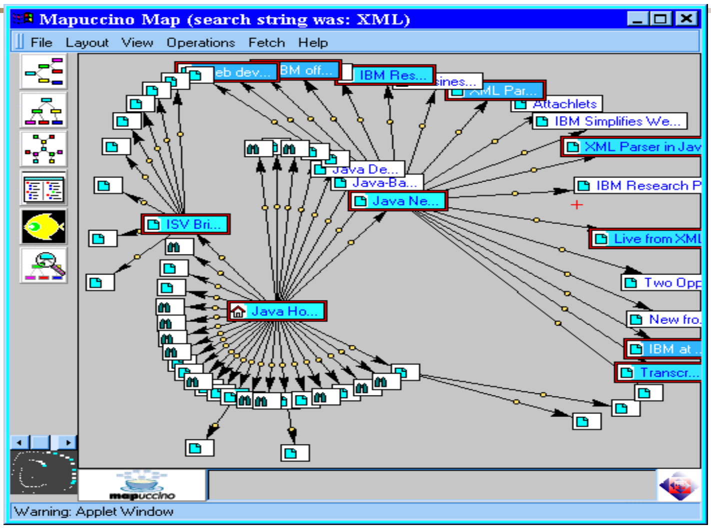
Latent Variable Models for Semantic Orientations of Phrases - Hiroya Takamura Takamura (2006) (Correct)

Citation/Link importance

IR Applications: Multimedia Retrieval

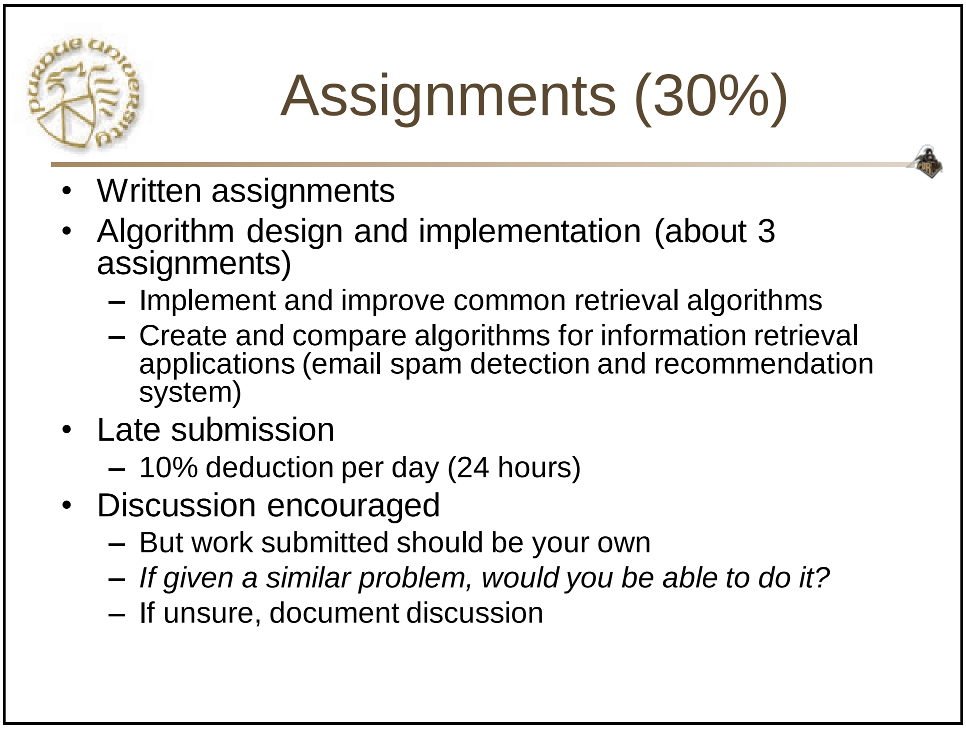


IR Applications: Information Visualization


Mapuccino Map (search string was: XML)

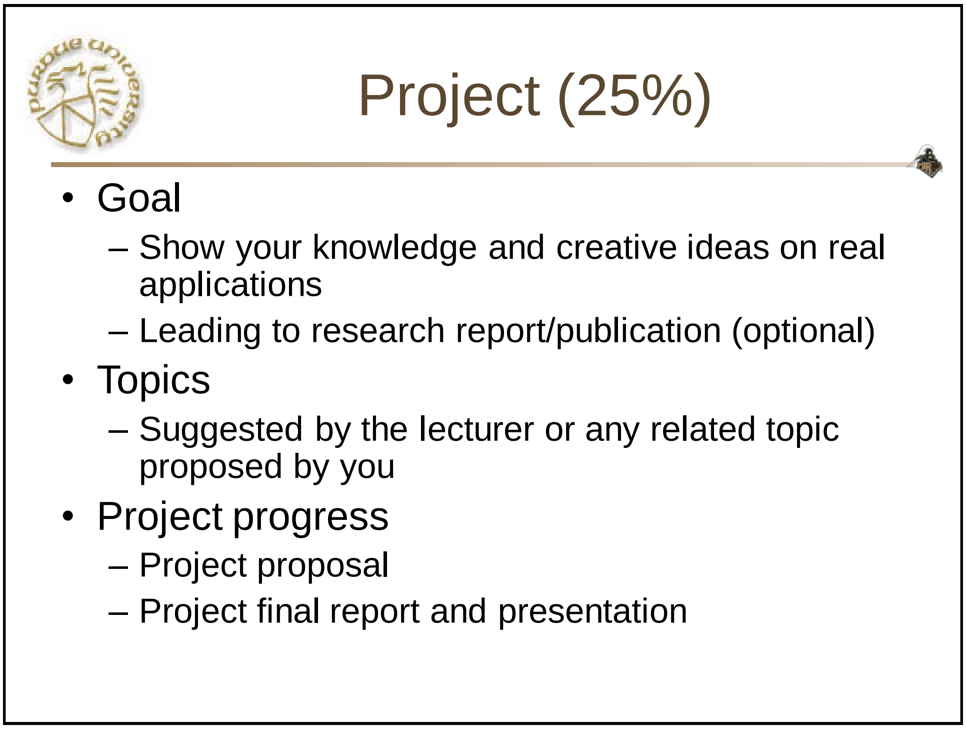
Warning: Applet Window

Partial Structure of pages from a Web subset visualized by Mapuccino



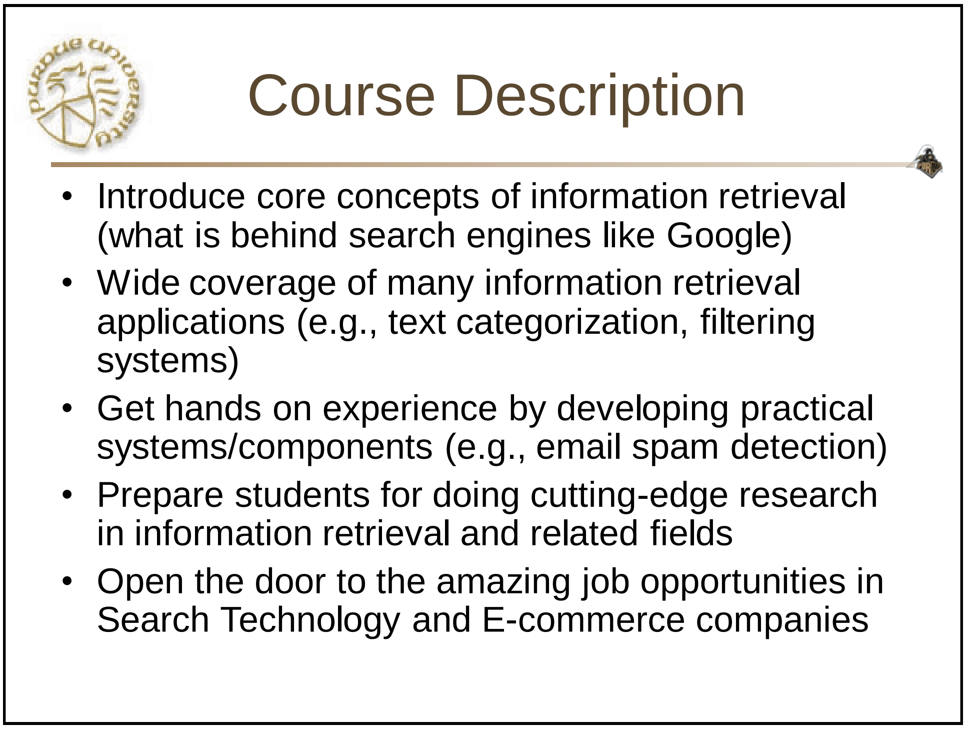
Assignments (30%)

- 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