Introduction

CMPE 493
INTRODUCTION TO INFORMATION RETRIEVAL

Introduction

Arzucan Özgür

Department of Computer Engineering, Boğaziçi University
June 24, 2011
Instructor: Arzucan Ö zgür
Office: TBA
E-mail: ozgur@umich.edu

(Please include CMPE493 in your subject when sending e-mail.)
Ofis hours: Monday, Thursday, Friday after class, or by appointment
References


- Course Web Site: TBA
Grading

- Midterm Exam: 20%
- Final Exam: 30%
- Paper Presentation: 15%
- Project: 25%
- Class Participation: 10%
Final project format

- Research paper - using the SIGIR format. Students will be in charge of problem formulation, literature survey, hypothesis formulation, experimental design, implementation, and possibly submission to a conference like SIGIR or WWW.

- Software system - develop a working system or API. Students will be responsible for identifying a problem, implementing it and deploying it, either on the Web or as an open-source downloadable tool. The system can be either stand alone or an extension to an existing one.

- Survey paper - identify a topic of research in IR and summarize 10-15 recent papers on it, along with an introduction that compares and contrasts the papers involved.
Information Retrieval

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
Unstructured (text) vs. structured (database) data in 1996
Unstructured (text) vs. structured (database) data in 2009
Examples of search engines

- **Conventional (library catalog).**
  Search by keyword, title, author, etc.

- **Text-based (Google, Yahoo!, Bing).**
  Search by keywords. Limited search using queries in natural language.

- **Multimedia (QBIC, WebSeek)**
  Search by visual appearance (shapes, colors,… ).

- **Question answering systems (Ask, NSIR, Answerbus)**
  Search in (restricted) natural language

- **Other:**
  - cross language information retrieval, music retrieval
### IR systems on the Web

- Search for Web pages: [http://www.google.com](http://www.google.com)
- Search for images: [http://www.picsearch.com](http://www.picsearch.com)
- Search for image content: [http://wang14.ist.psu.edu/](http://wang14.ist.psu.edu/)
- Search for answers to questions: [http://www.askjeeves.com](http://www.askjeeves.com)
Information retrieval

Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of...

Journal of Information Retrieval - Springer link

Introduction to Information Retrieval

Information Retrieval

Information Retrieval - University of Glasgow; School of...

Google Directory - Computers > Software > Information Retrieval

ACM SIGIR Special Interest Group on Information Retrieval Home Page
EXTREME SOLIDARITY Far-Right Parties Form New Group in European...
Spiegel Online, Germany - 42 minutes ago
European Union expansion is a topic typically supported by those on the left of the continent’s political spectrum and opposed by those on the right. ...
Far-right EU lawmakers form coalition Olberlin
all 88 news articles »

Wild bird trade to be banned by European Union
EnjoyFrance.com, France - 1 hour ago
The European Union is going to ban the trade in wild birds starting in July, EU animal health officials have announced. Animal welfare campaigners are ...
Wild bird imports to end Green Consumer Guide
UN-Backed Body 'Disappointed' By Bird Trade Ban
Scoop.co.nz (press release)
EU To Ban Wild Birds Imports All Headline News
Earthtimes.org
all 8 news articles »

Russia, European Union A serious problem of trust
Monday Morning, Lebanon - 5 hours ago
 Merkel, in contrast, is wary of depending heavily on Russia for oil and gas and
The Capital of Turkey is Ankara.

Source: CIA World Factbook
See Also: Beb Profile, Encyclopedia
Search For: Flights, Geography, Government, People
Music Retrieval Demo

This is a small demonstration of some audio retrieval-by-similarity work I have recently been pursuing. The aim is to automatically find audio clips that sound "similar," in some sense, to an example clip. Here's a brief explanation of how the demo works, and some reasons why this use of other people's music doesn't constitute copyright infringement.

Below is a scrollable list of more than 250 sound clips, which are 7-second excerpts from longer musical recordings. Representative genres include jazz, pop, rock, rap, and techno, as well as Brazilian music, plainsong, solo piano, guitar, and "easy listening." Click 'Play' to play the selected clip or 'Search' to find music that sounds similar to your selection. The number to the left is a similarity score; the larger the number the closer the match. Clicking 'Reset' then 'Search' will give you an alphabetical listing of available artists/tracks.

This work is still preliminary, which hopefully excuses the occasional bizarre result. But even if you think Gregorian chant sounds nothing like Nat King Cole, do listen with an open ear: the similarities are often surprising.

Some things to search for:

- Piano music
- Grunge rock
- Acoustic guitar
- Reggae
- Jazz
- Medieval plainsong

Search for similar files | Play selected file | Reset
What does it take to build a search engine?

- Decide what to index
- Collect it
- Index it (efficiently)
- Keep the index up to date
- Provide user-friendly query facilities
What else?

- Understand the structure of the web for efficient crawling
- Understand user information needs
- Preprocess text and other unstructured data
- Cluster data
- Classify data
- Evaluate performance
Goals of the course

- Understand how search engines work
- Understand the limits of existing search technology
- Learn about the state of the art in IR research
- Learn to analyze textual and semi-structured data sets
- Learn to evaluate information retrieval
- Learn about standardized document collections
- Learn about text similarity measures
- Learn about semantic dimensionality reduction
- Learn about web crawling
- Learn to use existing software
- Understand the dynamics of the Web by building appropriate mathematical models
- Build working systems that assist users in finding useful information on the Web