Information Retrieval and Web Search

Cornelia Caragea

Computer Science

University of Illinois at Chicago

Credits for slides: Hofmann, Mooney, Schutze

Logistics and Intro to Information Retrieval

Today

- Logistics
- What is this course about?
- What is Information Retrieval?

Logistics

- Instructor:
  - Dr. Cornelia Caragea
  - Email: cornelia@uic.edu
  - Office: 3-190E Daley Library
- Course time: Monday 3pm - 5:30pm, Room: TBH 180G
- Office hours: Monday 1pm - 2pm or by appointment in 3-190E Daley Library
- Prerequisites: Basic knowledge on probability and statistics, data structures and algorithms
- Class material: Will be made available on Blackboard.

Textbook

- Required: **Introduction to Information Retrieval**
  by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze
  Online version available at: http://nlp.stanford.edu/IR-book/
- Recommended: **Readings in Information Retrieval**
  by K.Sparck Jones and P. Willett
- Recommended: **Modern Information Retrieval**
  Ricardo Baeza-Yates and Berthier Ribeiro-Neto
- Recommended: **Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining**
  ChengXiang Zhai and Sean Massung

Topics

- The term vocabulary and postings lists
- Index construction
- Scoring, term weighting and the vector space model
- Computing scores in a complete search system
- Evaluation in information retrieval
- Relevance feedback and query expansion
- Language models for information retrieval

Topics II

- Web search basics
- Web crawling and indexes
- Link analysis - PageRank and its variants
- Text classification and mining
- Applications
  - Information Extraction
  - Knowledge Base Construction
  - Question Answering
  - Sentiment analysis and emotion detection on the Web

Software Resources

- Machine Learning in Python - scikit-learn
  http://scikit-learn.org/stable/
- Mallet: MAchine Learning for LanguagE Toolkit
  http://mallet.cs.umass.edu/
- Text Analyser: Text Content Analysis Tool
  http://www.usingenglish.com/resources/text-statistics.php
- Gephi: The Open Graph Viz Platform
  https://gephi.org/
- SNAP: Stanford Network Analysis Platform
  http://snap.stanford.edu/snap/
- PyTorch: https://pytorch.org/
- TensorFlow: https://www.tensorflow.org/

Course Grading

Section	Weight
Homework	20%
Reading assignments	10%
Exam 1 (October 11)	25%
Exam 2 (November 29)	25%
Class Project (December 9)	20%

Class Participation

- "We learn:
  10% of what we hear
  30% of what we see
  50% of what we see and hear
  70% of what we discuss
  80% of what we experience
  95% of what we teach others."
- "Teach me and I will forget;
  show me and I may remember;
  involve me and I will understand."
  Chinese Proverb.

Submission Policies

- Assignments will be submitted online through Blackboard.
    - Assignments are due by 11:59pm on the due date.
- Late submissions are not encouraged. I will accept late submissions, but there will be grading penalties.
    - Assignments may be turned in up to 3 days late, with a penalty of 5% for each day late. No credit will be given after 3 days.

Collaboration Policies

- You are encouraged to discuss the course material, concepts, and assignments, but you must write your answers independently.
- For each assignment, you are required to list students with whom you have discussed the assignment.
- Your submission should reflect your own knowledge and you should be able to reproduce the material you turn in at any time.
- Sharing answers will not be tolerated.
- Plagiarism will not be tolerated either.
- Appropriate citations for any external sources used in your work are mandatory. Never use sentences or phrases taken directly from a paper you are reviewing.

Who Am I?

- Cornelia Caragea, Associate Professor in CS@UIC <cornelia@uic.edu>
- Artificial Intelligence, Information Retrieval, Natural Language Processing, Machine Learning
  - Information Extraction
  - Text Summarization / Text Simplification / Text Generation / Text Classification
  - Scientific Document Mining
    - Building Scholarly Knowledge Graphs
    - Scientific Document Classification
    - Natural Language Inference
  - Mining Social Media
    - Domain Adaptation approaches for Disaster Management
    - Emotion Detection from Health Related Posts
  - Privacy Detection

What Is This Course About?

- Processing
- Indexing
- Retrieving

  · · · textual data

Need for IR?

Large Digital Information Repositories

- World Wide Web ($> 10^{12}$ links)
- Scientific Literature Libraries (e.g., CiteSeer, ArnetMiner, Microsoft Academic Search, Google Scholar, Semantic Scholar)
- Digital Libraries
- Web Archiving Repositories
- Medical Information Portals (e.g., Medline)
- Patent Databases (e.g., US Patent Office)
- Online Encyclopedias (e.g., Wikipedia)

Various Needs for Information

- Search for documents that fall in a given topic
- Search for specific information
- Search an answer to a question
- Search for information in a different language
- . . .
- Search for images
- Search for music
- Search for a (candidate) friend

Definition of 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).

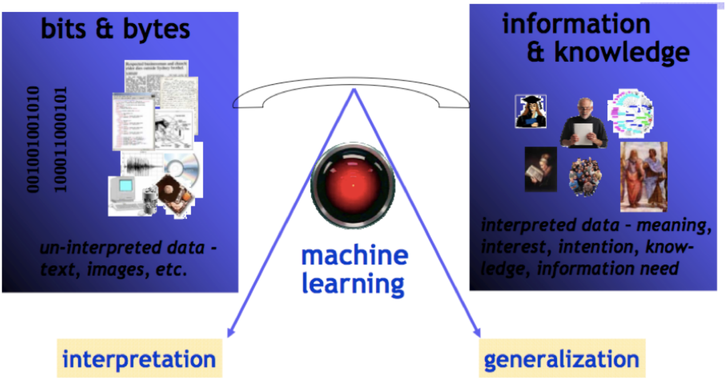
Other Definitions of IR

- **Maron & Kuhns (1960):** "IR deals with adequately identifying the information content of documentary data."
- **Salton (1989):** "Information-retrieval systems process files of records and requests for information, and identify and retrieve from the files certain records in response to the information requests. The retrieval of particular records depends on the similarity between the records and the queries, which in turn is measured by comparing the values of certain attributes to records and information requests."

Search

- We live in a search society - belief that (almost) everything is known, we just have to find the information.
- We search for everything - the right book, movie, car, house, vacation trip, bargain, search engine, etc.

Machine Learning in IR



bits & bytes

un-interpreted data - text, images, etc.

information & knowledge

interpreted data - meaning, interest, intention, knowledge, information need

machine learning

interpretation

generalization

unsupervised learning & data mining:
discover hidden regularities, generate semantically meaningful representations, predictive modeling, statistics

supervised learning:
generalize from given examples, classification & recognition, emulate human experts

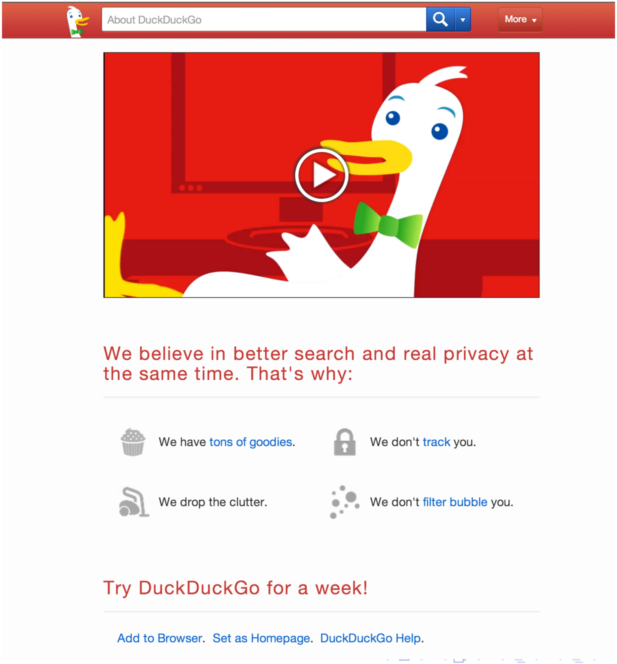
Machine Learning: Some IR Directions

- Text Clustering
  - Clustering of IR query results.
  - Automatic formation of hierarchies.
- Text Categorization
  - Automatic hierarchical classification.
  - Adaptive filtering/recommendation.
  - Automated spam filtering.
- Learning for Information Extraction
- Text Mining

Examples of IR Systems

- Conventional (library catalog)
    - Search by keyword, title, author, etc.
- Text-based (Google, Bing, DuckDuckGo)
    - Search by keywords.
- Question answering systems (START, Ask)
    - Search in (restricted) natural language
- Other:
    - Cross language information retrieval, music retrieval

DuckDuckGo



We believe in better search and real privacy at the same time. That's why:

We have tons of goodies.

We don't track you.

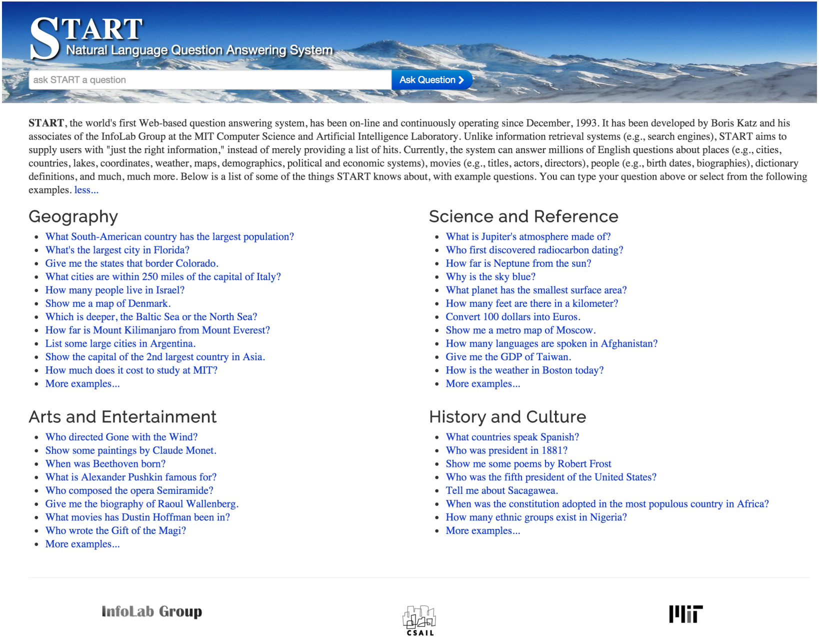
We drop the clutter.

We don't filter bubble you.

Try DuckDuckGo for a week!

Add to Browser. Set as Homepage. DuckDuckGo Help.

START



START, the world's first Web-based question answering system, has been on-line and continuously operating since December, 1993. It has been developed by Boris Katz and his associates of the InfoLab Group at the MIT Computer Science and Artificial Intelligence Laboratory. Unlike information retrieval systems (e.g., search engines), START aims to supply users with "just the right information," instead of merely providing a list of hits. Currently, the system can answer millions of English questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more. Below is a list of some of the things START knows about, with example questions. You can type your question above or select from the following examples. less...

Geography

- What South-American country has the largest population?
- What's the largest city in Florida?
- Give me the states that border Colorado.
- What cities are within 250 miles of the capital of Italy?
- How many people live in Israel?
- Show me a map of Denmark.
- Which is deeper, the Baltic Sea or the North Sea?
- How far is Mount Kilimanjaro from Mount Everest?
- List some large cities in Argentina.
- Show the capital of the 2nd largest country in Asia.
- How much does it cost to study at MIT?
- More examples...

Arts and Entertainment

- Who directed Gone with the Wind?
- Show some paintings by Claude Monet.
- When was Beethoven born?
- What is Alexander Pushkin famous for?
- Who composed the opera Semiramide?
- Give me the biography of Raoul Wallenberg.
- What movies has Dustin Hoffman been in?
- Who wrote the Gift of the Magi?
- More examples...

Science and Reference

- What is Jupiter's atmosphere made of?
- Who first discovered radiocarbon dating?
- How far is Neptune from the sun?
- Why is the sky blue?
- What planet has the smallest surface area?
- How many feet are there in a kilometer?
- Convert 100 dollars into Euros.
- Show me a metro map of Moscow.
- How many languages are spoken in Afghanistan?
- Give me the GDP of Taiwan.
- How is the weather in Boston today?
- More examples...

History and Culture

- What countries speak Spanish?
- Who was president in 1881?
- Show me some poems by Robert Frost
- Who was the fifth president of the United States?
- Tell me about Sacagawea.
- When was the constitution adopted in the most populous country in Africa?
- How many ethnic groups exist in Nigeria?
- More examples...

IR systems links

- Search for Web pages
  http://www.google.com
- Search for images
  http://www.picsearch.com
- Search for image content
  http://wang.ist.psu.edu/IMAGE/
- Search for answers to questions
  http://www.ask.com
  http://start.csail.mit.edu/
- Music retrieval
  http://www.rotorbrain.com/foote/musicr/

Information Retrieval

- The processing, indexing and retrieval of textual documents.
- Concerned firstly with retrieving relevant documents to a query.
- Concerned secondly with retrieving from large sets of documents efficiently.