CS172
INFORMATION RETRIEVAL

Prof. Mariam Salloum

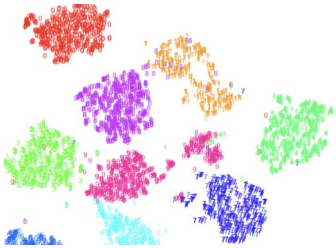
Introductions …

Prof. Mariam Salloum

Office: Bourns Hall A (Room 159B)

Email: msalloum@cs.ucr.edu

- Current research work on Data Integration and Big Data Visualization.
  - Given millions of relevant sources/sites to a particular query, how to identify relevant sources, query sources in optimized fashion then perform record linkage and data fusion.
  - How to visualize large datasets in high dimensions.



Outline

- What is this course about?
- Course Logistics
- Some IR applications – Overview

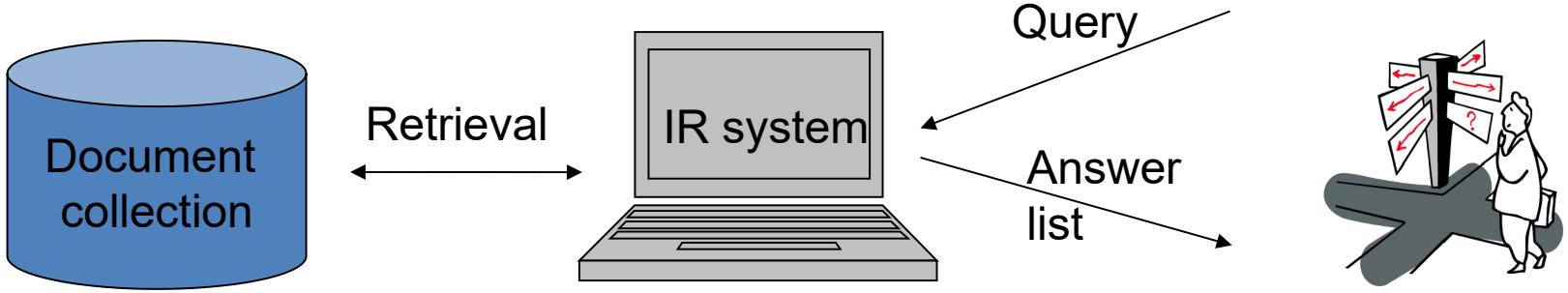
What is this course about?

- This course will cover core technologies for ***searching*** and ***organizing*** content.

- The following are some of the questions we will consider:
  - How do search engines work?
  - Why are some search engines better than others?
  - Can we think of the web as a big database/knowledge base and support efficient database style query processing?
  - How can we find useful pearls and patterns in the massive accessible data on the Internet?

Search and Information Retrieval

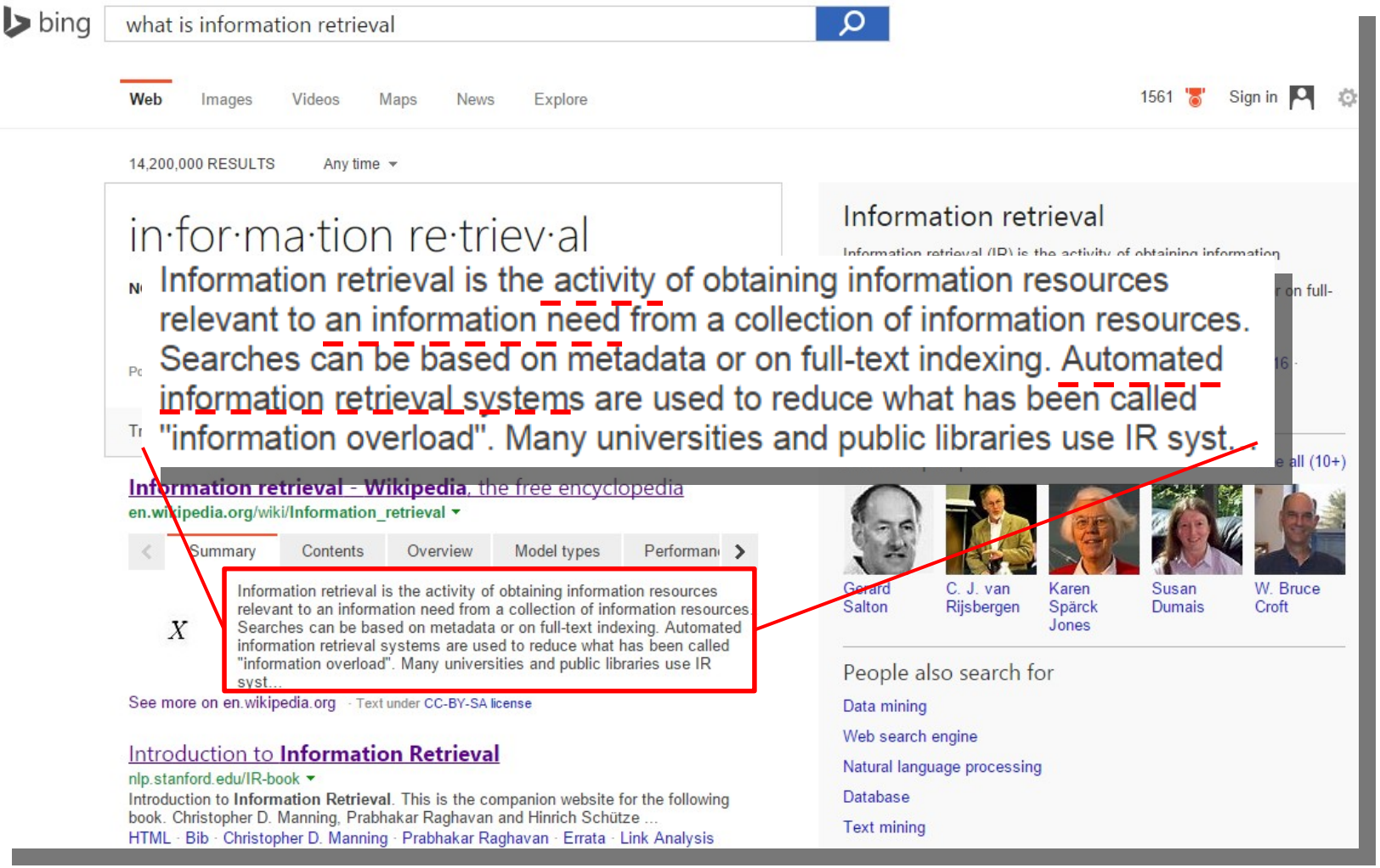
- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).



Course Topics

- As part of the course, we will cover:
  - Introduction to Information Retrieval (IR) principles including:
    - Efficient text indexing,
    - Searching document collections,
    - Crawling, and
    - Link analysis
  - Advanced topics like search in social networks.
  - Technologies for Big Data management and processing, with an introduction to the Map-Reduce paradigm, and NoSQL databases.
  - Brief topics on document clustering and classification.

What is information retrieval?



Why information retrieval

- Information overload
  - "*It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information.*" – wiki

- Handling unstructured data
  - Structured data: database system is a good choice
  - Unstructured data is more dominant
    - Text in Web documents or emails, image, audio, video…

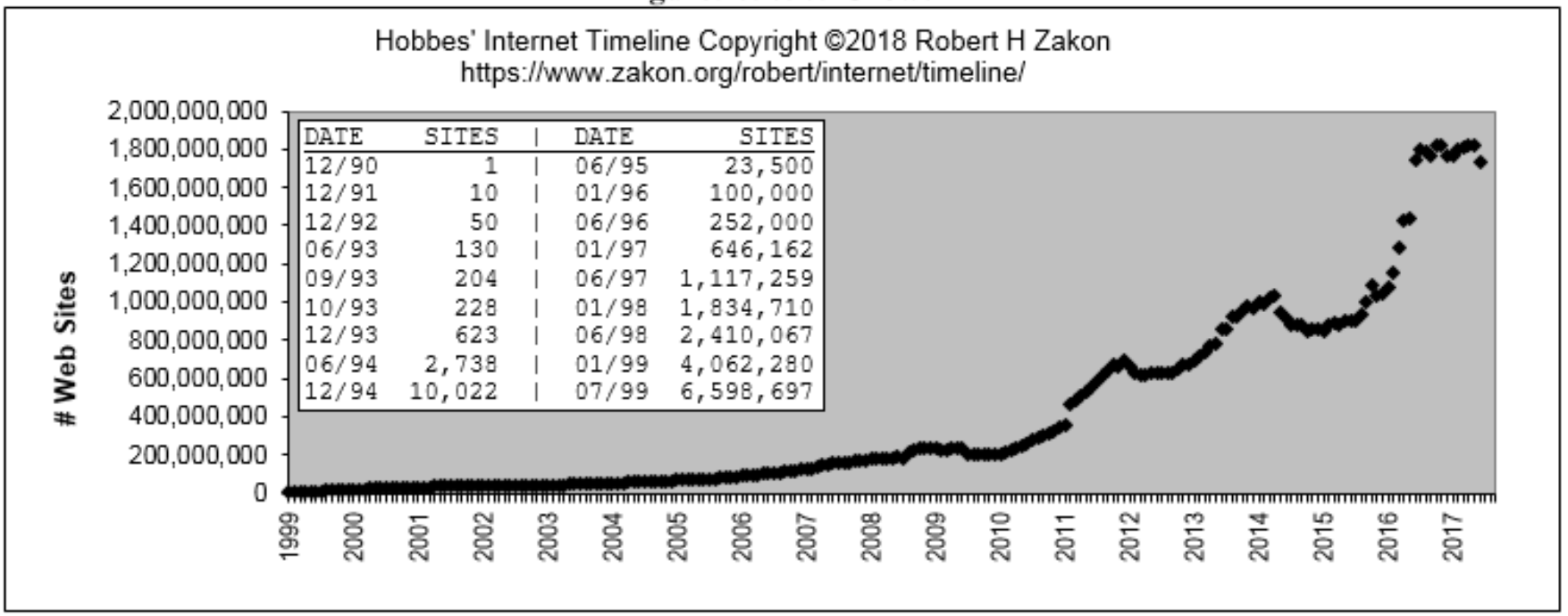
Why information retrieval

- Information overload

Figure: WWW Growth

Hobbes' Internet Timeline Copyright ©2018 Robert H Zakon
https://www.zakon.org/robert/internet/timeline/

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



Course Logistics

- **Lectures / Labs**
  - **Lecture:** TH: 2:00 – 3:20 PM
  - **Lab:** F 10:00 - 10:50 AM

- **My Info:**
  - **Office:** Bourns A (Room 159B)
  - **Email:** msalloum@cs.ucr.edu
  - **Website:** www.cs.ucr.edu/~msalloum

Course Logistics

- **iLearn :**
  - Will be used to turn-in assignments, and post grades.
- **Google Drive**
  - Lectures slides will be placed on a shared Google Drive
    LINK:
    https://drive.google.com/open?id=11yRclP5N4M5AyfLRlfbnpDxtv9fmu8kF
- **Campus Wire**
  - CampusWire will be used for discussions- announcements.
  - Questions relating to lecture or assignment should be posted to discussion board, not emailed to teachers, so any teacher/student can respond and fellow students benefit from answers.
  - **LINK:** https://campuswire.com/c/G16C76071
  - **CODE: 3522**

Textbook & Reading List

- **Required Book**

  Search Engines: Information Retrieval in Practice

  Bruce Croft, Donald Metzler, Trevor Strohman

  Addison Wesley; 1 edition (February 16, 2009)

  ISBN-10: 0136072240/ISBN-13: 978-0136072249

  Available here : http://ciir.cs.umass.edu/downloads/SEIRiP.pdf

- **Reference Books**
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, Introduction to Information Retrieval, Cambridge University Press. 2008.
  - Modern Information Retrieval the concepts and technology behind search
  - Hearst, M.A. Search User Interfaces, Cambridge University Press, September, 2009

- **Reading List**
  - We will cover the state-of-art technology from research papers in big conferences

Requirements & Grading

Item	Percentage	Notes
Assignments	20%	Programming assignments to reinforce theory and concepts covered in lecture
Midterms (x2)	40%	There will be two midterm exams.
Quizzes/ Participation	10%	Class attendance and participation is required. Short quizzes will also be given in class to assess covered material.
Project	30%	Team project (group of 3) to implement a text search engine.

Homework Assignments

- Programming assignments will require coding in Java or Python.
- We will use a dedicated Hadoop cluster for ONE of the assignments (link-analysis using Map Reduce).
- Late policy
  - Assignments should be submitted via iLearn.
  - Assignments can be submitted 2 days past the deadline.
  - There will be a deduction of 10% penalty for each day late.

Outline

- What is this course about?
- Course Logistics
- **Some IR applications – Overview**

History of IR

- 1950s
- 1960-70's:
  - Initial exploration of text retrieval systems for "small" corpora of scientific abstracts, and law and business documents.
  - Development of the basic Boolean and vector-space models of retrieval.
- 1980's:
  - Large document database systems, many run by companies:
    - Lexis-Nexis
    - Dialog
    - MEDLINE

History of IR (Cont.)

- 1990's:
  - Searching FTPable documents on the Internet
  - Searching the World Wide Web
    - Lycos
    - Yahoo
    - Altavista
  - Organized Competitions/Conferences
    - NIST TREC
  - Recommender Systems
  - Automated Text Categorization & Clustering

History of IR (Cont.)

- 2000's
  - Link analysis for Web Search
    - Google
  - Automated Information Extraction
  - Question Answering
    - TREC Q/A track
  - Multimedia IR
    - Image
    - Video
    - Audio and music
  - Cross-Language IR
  - Social Network Search
    - Facebook Graph Search

Recent IR History

- 2010's
  - Intelligent Personal Assistants
    - Siri
    - Cortana
    - Google Now
    - Alexa
  - Complex Question Answering
    - IBM Watson
  - Distributional Semantics
  - Deep Learning

IR Systems

- Information retrieval (IR) deals with the representation, storage, organization of, and access to information items.
- The user describes the information they would like to access using a query (usually keywords).
- Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user.

IR System

Document Corpus

Query String

IR System

1. Doc1
2. Doc2
3. Doc3
.
.

Ranked Documents

What is a Document?

- Examples:
  - web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, etc.
- Common properties
  - Significant text content
  - Some structure (e.g., title, author, date for papers; subject, sender, destination for email)

Documents vs. Database Records

- Database records (or *tuples* in relational databases) are typically made up of well-defined fields (or *attributes*)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is more difficult

Comparing Text

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval

- Exact matching of words is not enough
  - Many different ways to write the same thing in a "natural language" like English
  - e.g., does a news story containing the text *"bank director in Riverside steals funds"* match the query?
  - Some stories will be better matches than others, how do we order them?

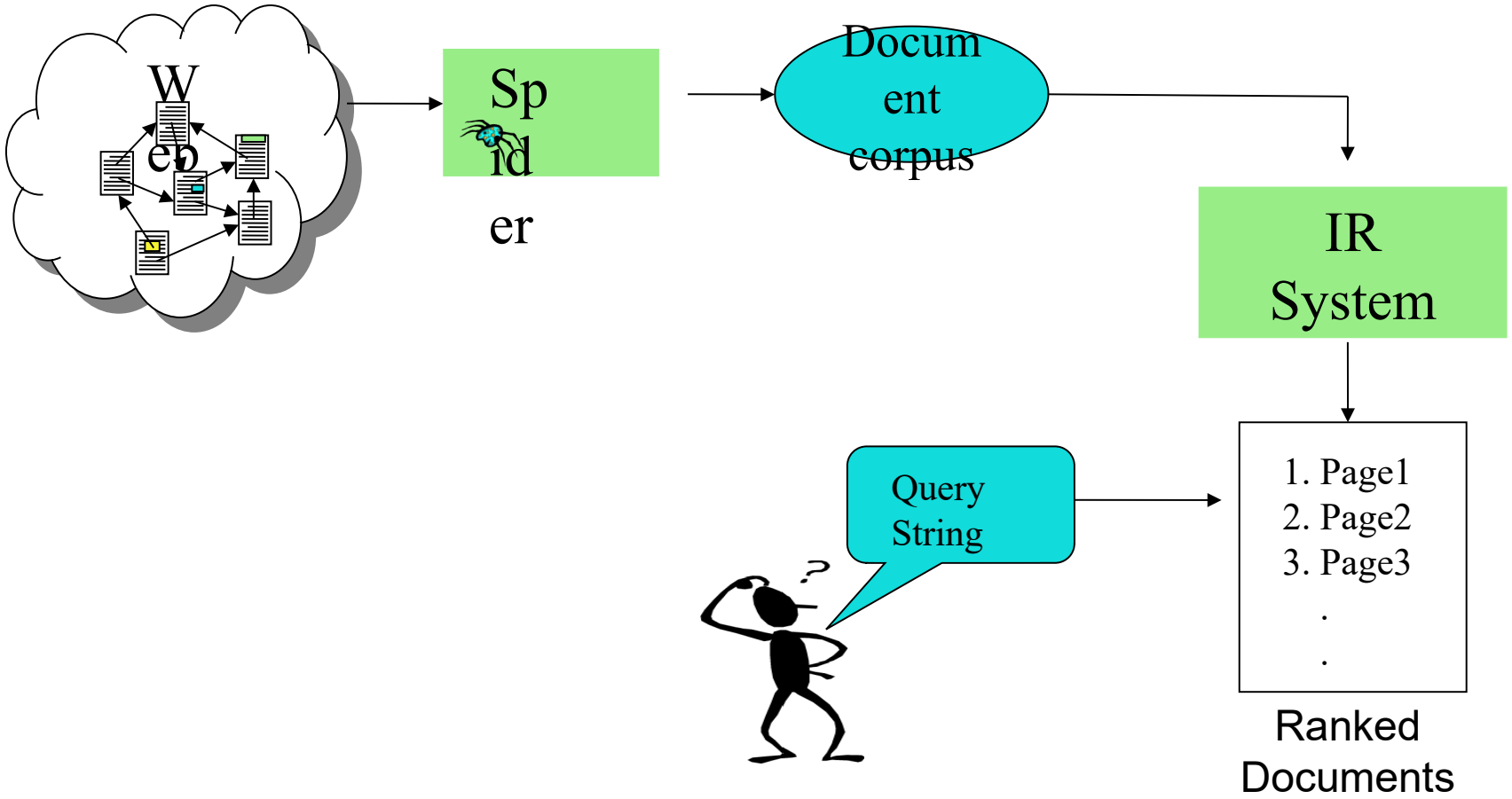
How good are the retrieved docs?

- *Precision* : Fraction of retrieved docs that are relevant to the user's information need

- *Recall* : Fraction of relevant docs in collection that are retrieved
  - More precise definitions and measurements to follow later

Dimensions of IR

- IR is more than just text, and more than just web search
  - although these are core fields of IR



Other IR-Related Tasks

- Automated document categorization
- Information filtering (spam filtering)
- Information routing
- Automated document clustering
- Recommending information or products
- Information extraction
- Information integration
- Question answering

Other Media

- New applications increasingly involve new media
  - e.g., video, photos, music, speech
- Like text, content is difficult to describe and compare
  - text may be used to represent them (e.g. tags)
- IR approaches to search and evaluation are appropriate

Related Areas

- Database Management
- Artificial Intelligence
- Natural Language Processing
- Machine Learning

Database Management

- Focused on structured data stored in relational tables rather than free-form text.
- Focused on efficient processing of well-defined queries in a formal language (SQL).
- Clearer semantics for both data and queries.
- Recent move towards semi-structured data (XML) brings it closer to IR.

Artificial Intelligence

- Focused on the representation of knowledge, reasoning, and intelligent action.
- Formalisms for representing knowledge and queries:
  - First-order Predicate Logic
  - Bayesian Networks
- Recent work on web ontologies and intelligent information agents brings it closer to IR.

Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse.
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords.

- Relation to IR
- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*).
- Methods for identifying specific pieces of information in a document (*information extraction*).
- Methods for answering specific NL questions from document corpora or structured data like FreeBase or Google's Knowledge Graph.

Machine Learning

- Focused on the development of computational systems that improve their performance with experience.
- Automated classification and clustering of examples.

- Relation to IR
- Text Categorization
  - Adaptive filtering/routing/recommending.
  - Automated spam filtering.
- Text Clustering
  - Clustering of IR query results.
- Learning for Information Extraction
- Text Mining
- Learning to Rank

Main problems in IR

- Document and query indexing
  - How to best represent their contents?
- Query evaluation (or retrieval process)
  - To what extent does a document correspond to a query?
- System evaluation
  - How good is a system?
  - Are the retrieved documents relevant? (precision)
  - Are all the relevant documents retrieved? (recall)

Document indexing

- Goal = Find the important meanings and create an internal representation
- Factors to consider:
  - Accuracy to represent meanings (semantics)
  - Exhaustiveness (cover all the contents)
- What is the best representation of contents?
  - Char. string (char trigrams): not precise enough
  - Word: good coverage, not precise
  - Phrase: poor coverage, more precise
  - Concept: poor coverage, precise

Main problems in IR (Cont).

- Performance
  - Measuring and improving the efficiency of search
    - e.g., reducing *response time*, increasing *query throughput*, increasing *indexing speed*
  - *Indexes* are data structures designed to improve search efficiency
    - designing and implementing them are major issues for search engines

Main problems in IR (Cont).

- Dynamic data
  - The “collection” for most real applications is constantly changing in terms of updates, additions, deletions
    - e.g., web pages
  - Acquiring or “crawling” the documents is a major task
    - Typical measures are *coverage* (how much has been indexed) and *freshness* (how recently was it indexed)
  - Updating the indexes while processing queries is also a design issue

Main problems in IR (Cont).

- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithm, indexing strategy, interface for different applications

Spam

- For Web search, spam in all its forms is one of the major issues
- Affects the efficiency of search engines and, more seriously, the effectiveness of the results
- Many types of spam
  - e.g. spamdexing or term spam, link spam, “optimization”
- New subfield called *adversarial IR*, since spammers are “adversaries” with different goals

Social Networks

- Traditional Model
- Given ◦ a set of entities (humans) ◦ And their relations (network) ◦ Return ◦ Measures of centrality and importance ◦ Propagation of trust (Paths through networks) ◦ Many uses ◦ Spread of diseases ◦ Spread of rumours ◦ Popularity of people ◦ Friends circle of people Web-induced headaches ◦ Scale (billions of entities) ◦ Implicit vs. Explicit links ◦ Hypertext (inter-entity connections easier to track) ◦ Interest-based links & Simplifications ◦ Global view of social network possible… Consequently ◦ Ranking that takes link structure into account ◦ Authority/Hub ◦ Recommendations (collaborative filtering; trust propagation)

Course Goals

- To help you to understand search engines, evaluate and compare them, and modify them for specific applications
- Provide broad coverage of the important issues in information retrieval and search engines
  - includes underlying models and current research directions