Course overview and introduction

CE-324: Modern Information Retrieval

Sharif University of Technology

M. Soleymani

Fall 2015

Some slides have been adapted from: Profs. Manning, Nayak & Raghavan lectures (CS-276, Stanford)

Text books

- Main:
  - **Introduction to Information Retrieval**, C.D. Manning, P. Raghavan and H. Schuetze, Cambridge University Press, 2008.
    - Free online version is available at: http://informationretrieval.org/

- Recommended:
  - Modern Information Retrieval, R. Baeza-Yates and B. Ribeiro-Neto, Addison Wesley, Second Edition, 2011.
  - Managing Gigabytes: Compressing and Indexing Documents and Images, I.H. Witten, A. Moffat, and T.C. Bell, Second Edition, Morgan Kaufmann Publishing,1999.
  - Information Retrieval: Implementing and Evaluating Search Engines, S. Büttcher, C.L.A. Clarke and G.V. Cormack, MIT Press, 2010.

Information Retrieval (IR)

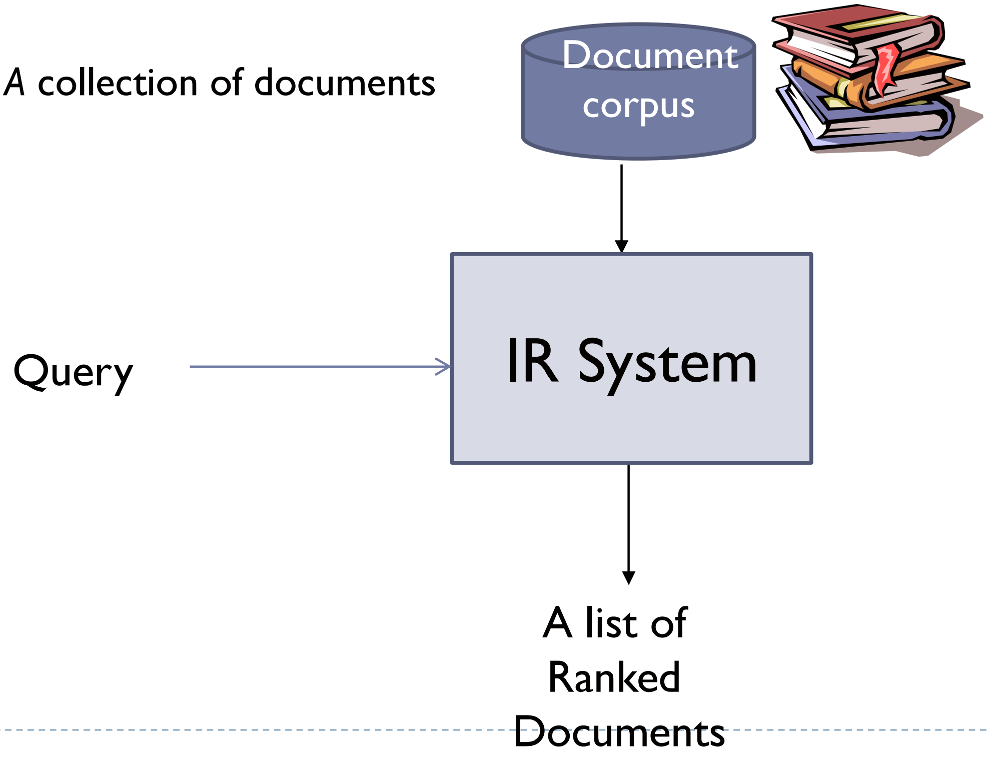
- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections [IIR Book].

- Retrieving *relevant* documents to a query (while retrieving as few non-relevant documents as possible)
  - especially from *large* sets of documents *efficiently*.
  - Deals with the representation, storage, organization of information items, and access to them

Typical IR system

- Given: corpus & user query
- Find: A ranked set of docs relevant to the query.

Corpus: *A* collection of documents



Document corpus

IR System

Query

A list of Ranked Documents

Basic Definitions

- **Document**: a unit decided to build a retrieval system over
  - textual: a sequence of words, punctuation, etc that express ideas about some topic in a natural language.

- **Information need**: information required by the user about some topics

- **Query**: formulation of the information need

Minimize search overhead

- Search overhead: Time spent in all steps leading to the reading of items containing the needed information
  - Steps: query generation, query execution, scanning results, reading non-relevant items, etc.

- The amount of online data has grown at least as quickly as the speed of computers

Data retrieval vs. information retrieval

- Data retrieval
  - which items contain a set of keywords? Or satisfy the given (e.g., regular expression like) user query?
  - well defined structure and semantics
  - a single erroneous object implies failure!

- Information retrieval
  - information about a subject
  - semantics is frequently loose (natural language is not well structured and may be ambiguous)
  - small errors are tolerated

Heuristic nature of IR

- Problem: Semantic gap between query and docs
  - A doc is relevant if the user perceives that this doc contains his information need
  - How to extract information from docs and how to use it to decide relevance

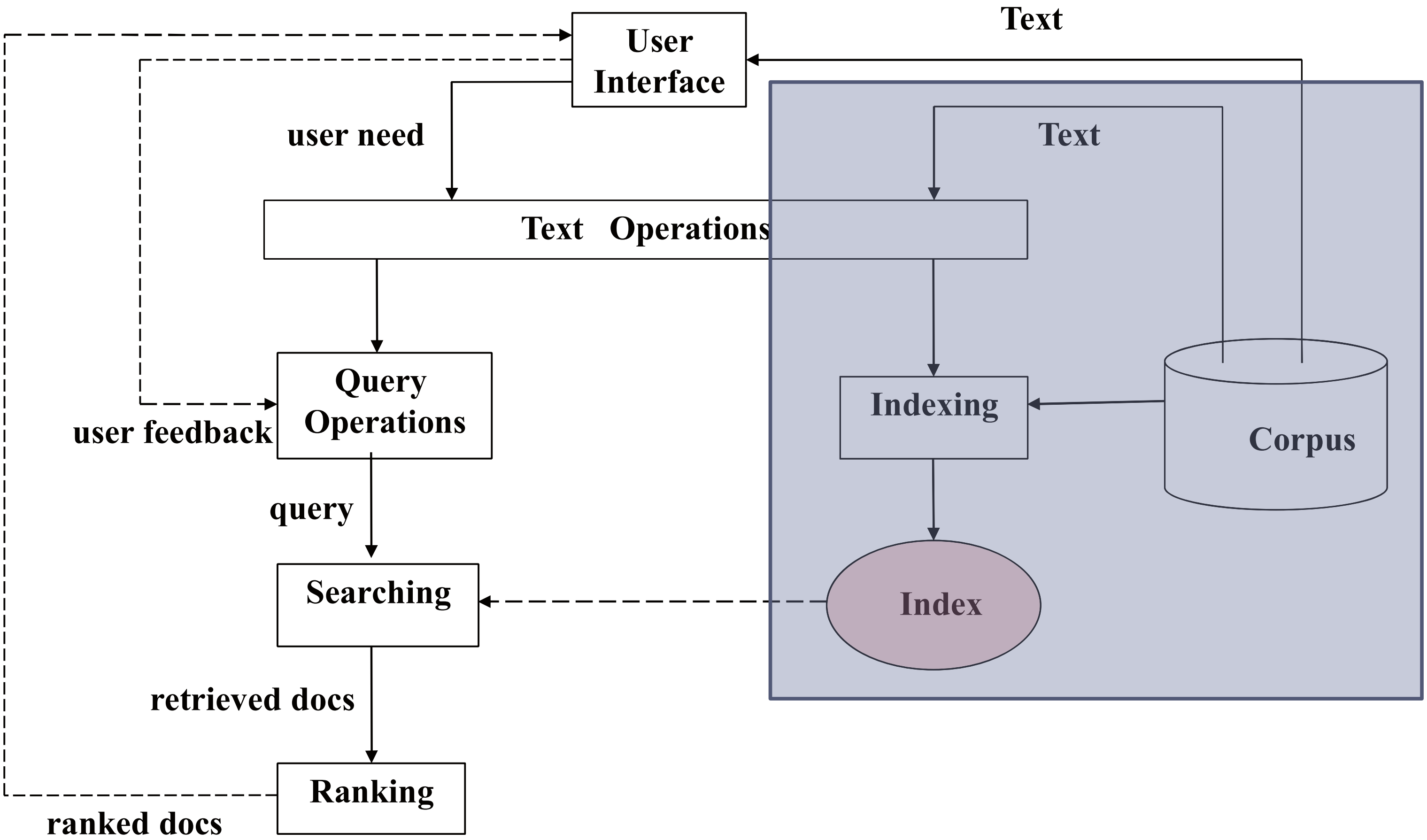
- Solution: IR system must interpret and rank docs according to the amount of relevance to the user's query.
  - "The notion of relevance is at the center of IR."

Condensing the data (indexing)

- Indexing the corpus to speed up the searching task
  - Using the index instead of linearly scanning the docs that is computationally expensive for large collections
  - Indexing depends on the query language and IR model

- **Term** (index unit): A word, phrase, and other groups of symbols used for retrieval
  - Index terms are useful for remembering the document themes

Typical IR system architecture



IR system components

- **Text Operations** forms index terms
  - Tokenization, stop word removal, stemming, …

- **Indexing** constructs an index for a corpus of docs.

- **Query Operations** transform the query to improve retrieval:
  - Query expansion using a thesaurus or query transformation using relevance feedback

- **Searching** retrieves docs that are related to the query.

IR system components (continued)

- **Ranking** scores retrieved documents according to their relevance.

- **User Interface** manages interaction with the user:
  - Query input and visualization of results
  - Relevance feedback

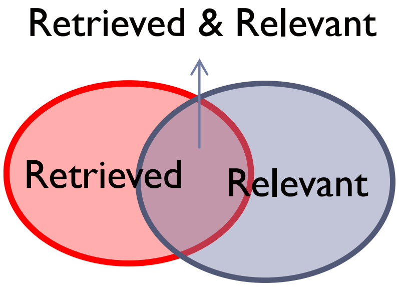
Evaluation of results

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

    $= |Retrieved \cap Relevant| \;/\; |Retrieved|$

- **Recall**: Fraction of relevant docs that are retrieved
  - *Recall = relevant retrieved / relevant exist*

    $= |Retrieved \cap Relevant| \;/\; |Relevant|$



Structured vs. unstructured docs

- Unstructured text (free text): a continuous sequence of tokens

- Structured text (fielded text): text is broken into fields that are distinguished by tags or other markup

- Semi-structured text
  - e.g. web page

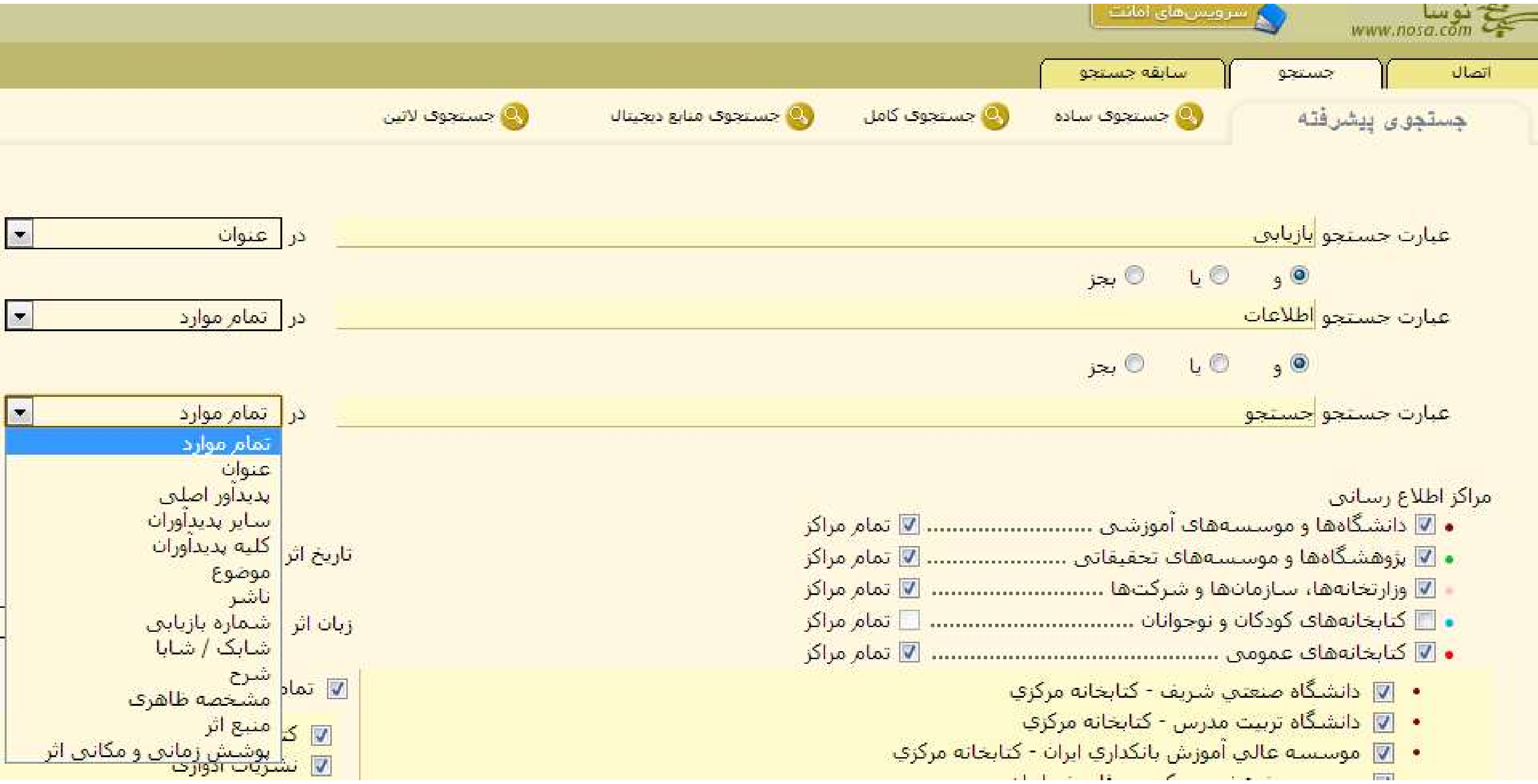
Databases vs. IR:
Structured vs. unstructured data

- Structured: data tends to refer to information in “tables”

Student Name	Student ID	Supervisor Name	GPA
Smith	20116671	Joes	12
Joes	20114190	Chang	14.1
Lee	20095900	Chang	19

Typically allows numerical range and exact match (for text) queries, e.g.,
GPA < 16 AND Supervisor = Chang.

Semi-structured data



عبارت جستجو بازیابی در عنوان

● و ○ یا ○ بجز

عبارت جستجو اطلاعات در تمام موارد

● و ○ یا ○ بجز

عبارت جستجو جستجو در تمام موارد

مراکز اطلاع رسانی

- ☑ دانشگاه‌ها و موسسه‌های آموزشی .... ☑ تمام مراکز
- ☑ پژوهشگاه‌ها و موسسه‌های تحقیقاتی .... ☑ تمام مراکز
- ☑ وزارتخانه‌ها، سازمان‌ها و شرکت‌ها .... ☑ تمام مراکز
- □ کتابخانه‌های کودکان و نوجوانان .... □ تمام مراکز
- ☑ کتابخانه‌های عمومی .... ☑ تمام مراکز

- ☑ دانشگاه صنعتي شريف - کتابخانه مرکزي
- ☑ دانشگاه تربيت مدرس - کتابخانه مرکزي
- ☑ موسسه عالي آموزش بانکداري ايران - کتابخانه مرکزي

Semi-structured data

- In fact almost no data is “unstructured”
  - E.g., this slide has distinctly identified zones such as the *Title* and *Bullets*

- Facilitates “semi-structured” search such as
  - *Title* contains <u>data</u> AND *Bullets* contain <u>search</u>

… to say nothing of linguistic structure

More sophisticated semi-structured search

- *Title* is about Object Oriented Programming AND *Author* something like stro*rup
  - * is the wild-card operator

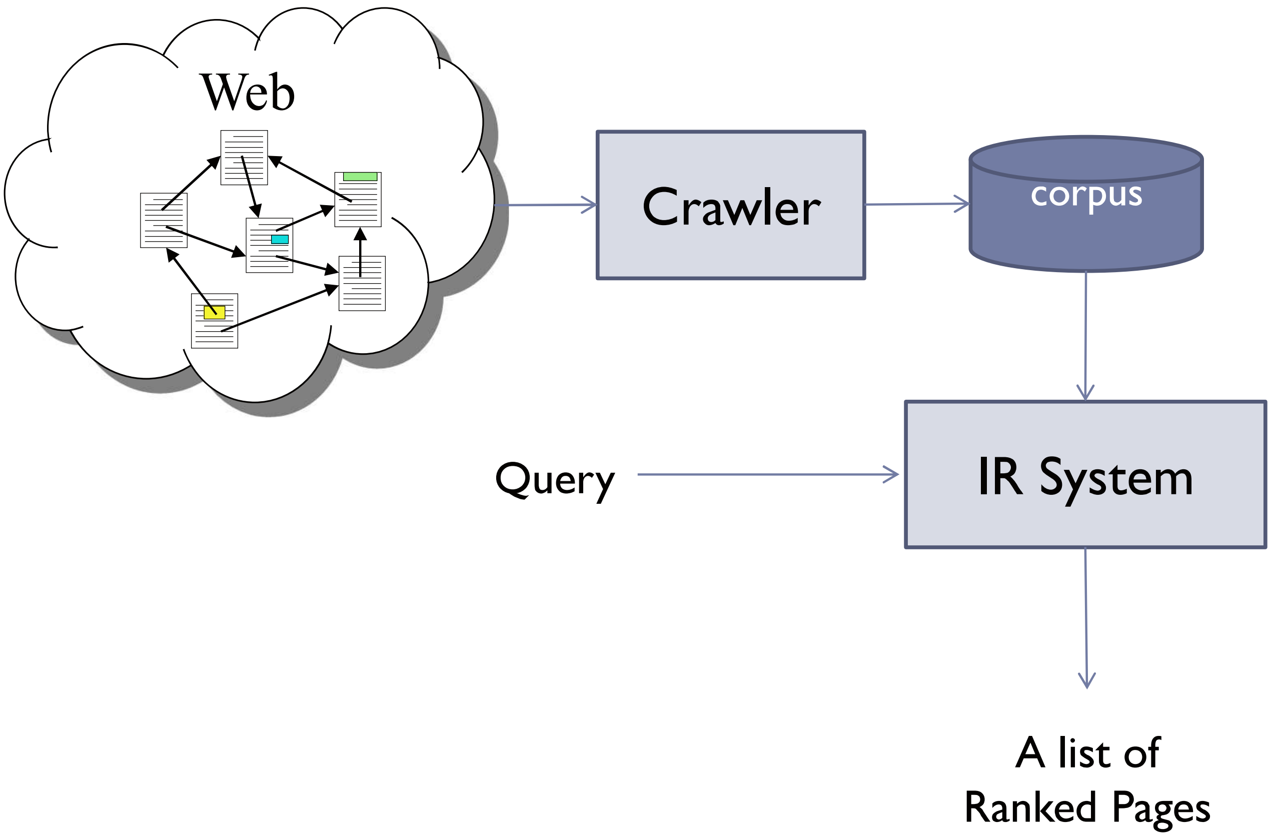
- Issues:
  - how do you process "about"?
  - how do you rank results?

Web Search

- Application of IR to (HTML) documents on the World Wide Web.

- Web IR
  - collect doc corpus by crawling the web
  - exploit the structural layout of docs
  - Beyond terms, exploit the link structure (ideas from social networks)
    - link analysis, clickstreams …

Web IR



The web and its challenges

- Web collection properties
  - Distributed nature of the web collection
  - Size of the collection and volume of the user queries
  - Web advertisement (web is a medium for business too)
  - Predicting relevance on the web
  - Docs change uncontrollably (dynamic and volatile data)
  - Unusual and diverse (heterogeneous) docs, users, and queries

Some main trends in IR models

- Boolean models: Exact matching
- Vector space model: Ranking docs by similarity to query
- PageRank: Ranking of matches by importance of documents
- Combinations of methods

Course main topics

- Introduction
- Indexing & text operations
- IR Models
  - Boolean, vector space, probabilistic
- Evaluation of IR systems
- Query operations
- Web IR
- Machine Learning in IR: Classification, clustering, and ranking
- Some advanced topics