Information Retrieval: Course Introduction

Animesh Mukherjee, Saptarshi Ghosh

CSE, IITKGP

Course Info

Course Website:

`https://sites.google.com/view/information-retrieval-2022/home`

Course Info

Course Website:

https://sites.google.com/view/information-retrieval-2022/home

Meeting Times

- Wednesday: 12:00 - 13:00 (NC141)
- Thursday: 11:00 - 12:00 (NC141)
- Friday: 09:00 - 10:00 (NC141)

Course Info

Teaching Assistants

- Paramita Das
- Shounak Paul
- Sayantan Adak
- Punyajoy Saha

Course Info

Pre-requisites

- Data structures and algorithms
- Probability and Statistics
- Basics of Machine Learning
- Basics of Natural Language Processing
- Basics of Graph algorithms
- Programming in Python/Java

Books and Materials

Text book:

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*, Cambridge university press.
- Available online: `https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Books and Materials

Text book:

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*, Cambridge university press.
- Available online: `https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Other materials:

- Lecture Slides
- Additional Readings to be given as necessary

Course Evaluation Plan: Tentative

- Mid-Sem : 20%
- End-Sem : 40%
- Term Project: 40%

What is Information Retrieval?

Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need (usually specified using a user query) from within large collections.

What is a document?

web pages, emails, books, news stories, scholarly papers, text messages, Powerpoint, PDF, forum postings, patents, tweets, question answer postings, etc.

Document 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

Document vs. Database Records

Example bank database query

- Find records with balance > $50,000 in branches located in Amherst, MA.
- Matches easily found by comparison with field values of records

Example search engine query

- *bank scandals in 2019 in India*
- This text must be compared to the text of entire news stories

So, what do we do in IR?

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

What is the "killer" app?

Searching for the pages on WWW

IR over text and other modes of data

- IR does not necessarily deal with text data
- Both the documents and the query can be in other modes as well, e.g., similar image search
- In this course, we will consider only textual IR

Typical IR tasks

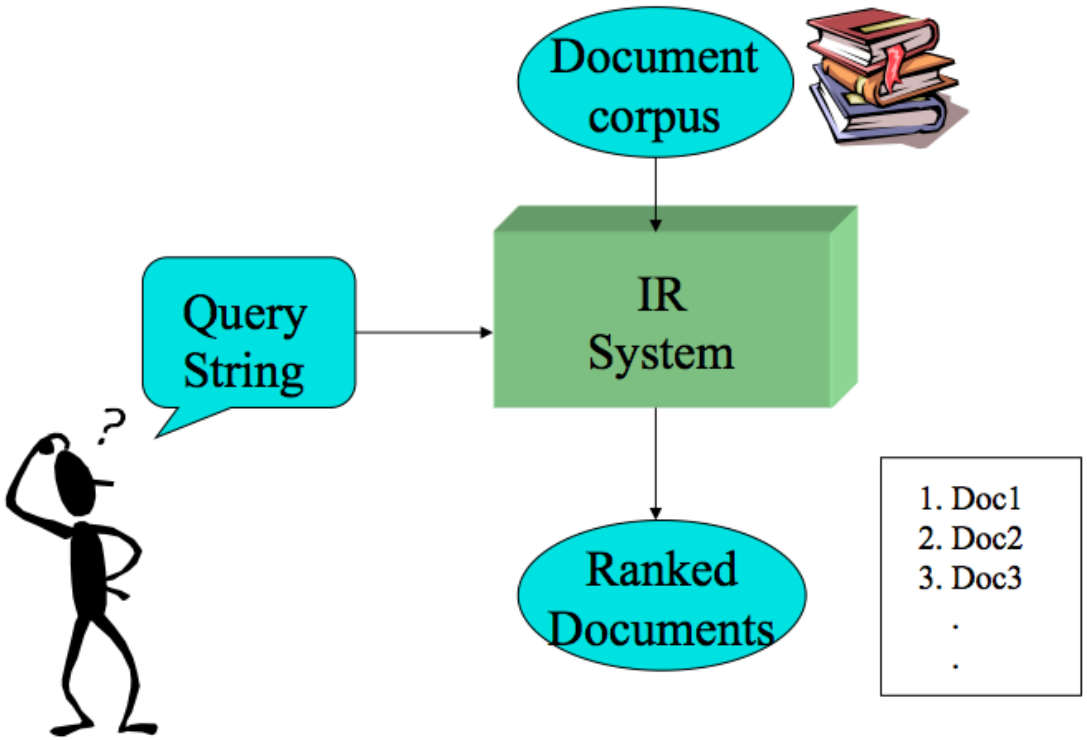
Given:

- A corpus of textual natural-language documents.
- A user query in the form of a textual string.

Find:

- A ranked set of documents that are relevant to the query.

IR System



The system should be able to retrieve the relevant docs efficiently

So, what is relevance?

Relevant document contains the information that a person was looking for when they submitted the query. This may include:

- Being on the proper subject.
- Being timely (recent information).
- Being authoritative (from a trusted source).
- Satisfying the goals of the user and his/her intended use of the information (information need).

Simplest notion of Relevance from Retrieval Models' Perspective

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that (most of) the words in the query appear frequently in the document, in any order (*bag of words*).

Problems with Keywords Search

Term mismatch

May not retrieve relevant documents that include synonymous terms

- PRC vs. China
- car vs. automobile

Ambiguity

May retrieve irrelevant document that include ambiguous terms (due to polysemy)

- 'Apple' (company vs. fruit)
- 'Java' (programming language vs. Island)

An Intelligent IR system will

- Take into account the *meaning* of the words used.
- Adapt to the user based on *direct* or *indirect* feedback.
- Take into account the *importance* of the page.
- ...

Where do we find the latest happenings in the field?

Top Conferences in the field

- SIGIR
- WWW
- WSDM

Other Venues

- ECIR
- ACM Transactions on Information Systems
- Information Retrieval (Springer), Information Processing & Management (Elsevier), etc.

Active Areas of Research

Compiled based on some recent papers at SIGIR and related conferences, just indicative, not exhaustive

What to retrieve

- *Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval*. SIGIR 2015.
- *Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale.* SIGIR 2016.
- *On Application of Learning to Rank for E-Commerce Search*. SIGIR 2017.
- *Understanding and Modeling Success in Email Search.* SIGIR 2017.
- *ANNE: Improving Source Code Search using Entity Retrieval Approach.* WSDM 2017.
- *Exploiting Food Choice Biases for Healthier Recipe Recommendation.* SIGIR 2017.
- *Toward an Interactive Patent Retrieval Framework based on Distributed Representations.* SIGIR 2018.
- *A Test Collection for Evaluating Legal Case Law Search.* SIGIR 2018.
- *Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos.* SIGIR 2019.

Personalization, Privacy, Social, Bias, Fairness, etc.

- *Predicting Which Topics You Will Join in the Future on Social Media*. SIGIR 2017.
- *Why People Search for Images using Web Search Engines*. WSDM 2018.
- *The Utility and Privacy Effects of a Click*. SIGIR 2017.
- *How Do Biased Search Result Rankings Affect User Attitudes on Debated Topics?*. SIGIR 2021.
- *When Fair Ranking Meets Uncertain Inference*. SIGIR 2021.

What do we cover in this course

IR Basics

- Boolean retrieval
- Term vocabulary & postings lists
- Scoring, term weighting & the vector space model
- Dictionaries and tolerant retrieval
- Index construction and compression
- Evaluation in information retrieval
- Relevance feedback & query expansion
- Probabilistic information retrieval
- Language models for information retrieval

Course Contents

Web Search, Applications, Recent Advances

- Web crawling and Link analysis (HITS, PageRank)
- Summarization
- Domain-specific IR - case studies
- Fairness and Bias in IR