

INFORMATION RETRIEVAL: COURSE INTRODUCTION CS60092

Somak Aditya

Assistant Professor

Department of CSE, IIT Kharagpur

January 3rd , 2023



Brief Introduction



Prof. Somak Aditya

***Tr*²AIL Lab**

**Building *Tr*ansparent & *Tr*usted *AI* systems
using *L*ogic**

Office: CSE 305

<https://cse.iitkgp.ac.in/~saditya/>

Short Bio

Theses (Masters and PhD)

-  Ph.D., CS, Arizona State University. "*Knowledge and Reasoning for Image Understanding*".
-  M.E., CS, Indian Institute of Science. "*Generic Incremental K-Means Clustering*"

Post-PhD Experience

-  Postdoc Researcher, Microsoft Research
-  Researcher, Adobe Research

Organizational Activities

- › [CVPR 2022]: Open-Domain Retrieval Under a Multi-Modal Setting, (<https://asu-apg.github.io/odrum/>) *IIT KGP, Arizona State University, FAIR, DeepMind, Microsoft Azure, IDIAP*
- › [CIKM 2021]: "Knowledge Injection in Neural Networks", (<https://sites.google.com/view/kinn2021/>) *Intel Labs, Arizona State University, MSRI, Univ. College of London*
- › [IJCAI 2021]: "Is Neuro-Symbolic SOTA still a myth for Natural Language Inference?", (<http://nsnli.github.io/>) *MSRI, UT Austin, IBM Research, KU Leuven*
- › [KR 2018]: Integrating learning of Representations and models with deductive Reasoning that leverages Knowledge, *Arizona State University, IBM Research, Verisk Analytics AI*

Course Website

- <https://adityasomak.github.io/courses/irspring23/>
- Course Timings
 - Mon 12:00-12:55 pm,
 - Tue 10-11:55 am
- My Office: CSE 305
- Teaching Assistants
 - Bishal Santra, Sachin Vashishtha, Vivek Karde, Deepak Chaudhary

Books and Materials

- Reference Book
 - Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval, Cambridge university press.
- Lecture Materials
 - Lecture Slides
 - Course Notes
 - Slides/lectures by Prof. Subbarao Kambhampati (Ex-AAAI President, Professor ASU <http://rakaposhi.eas.asu.edu/cse494/>)

Course Evaluation Plan (Tentative)

- Mid-Sem: 20%
- Final Exam: 40%
- Class Performance/Viva: 5-10%
- Term Project: 30-35% (extremely important)

Term Project Dates (Tentative)

- Distribute Project Topics ~ Jan 13
- Form groups of 4/5. Propose 2-3 choices ~ Jan 20
- Assign projects ~ Jan 27

- April 1-7 (Tentative)
 - Submit short 4 page project reports. Submit running code (Google Collab/Jupyter Notebook).
 - Guest Judges (Other Faculties/Industry Researchers)
 - Short Presentations (with demos)

Guest Lecture Announcement



Dr. Aniruddha (Ani) Kembhavi

Director of Computer Vision

Allen Institute of AI (AI2), Seattle, US

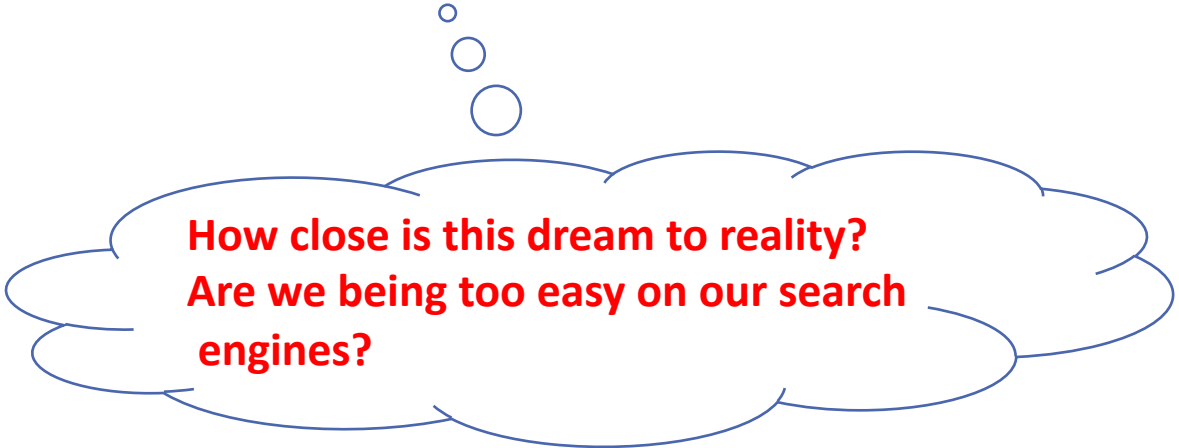
<https://anikem.github.io/>

Date: March 13 (Tentative/TBA)

Time: 9 am Online (Tentative/TBA)

Information Retrieval (informally)

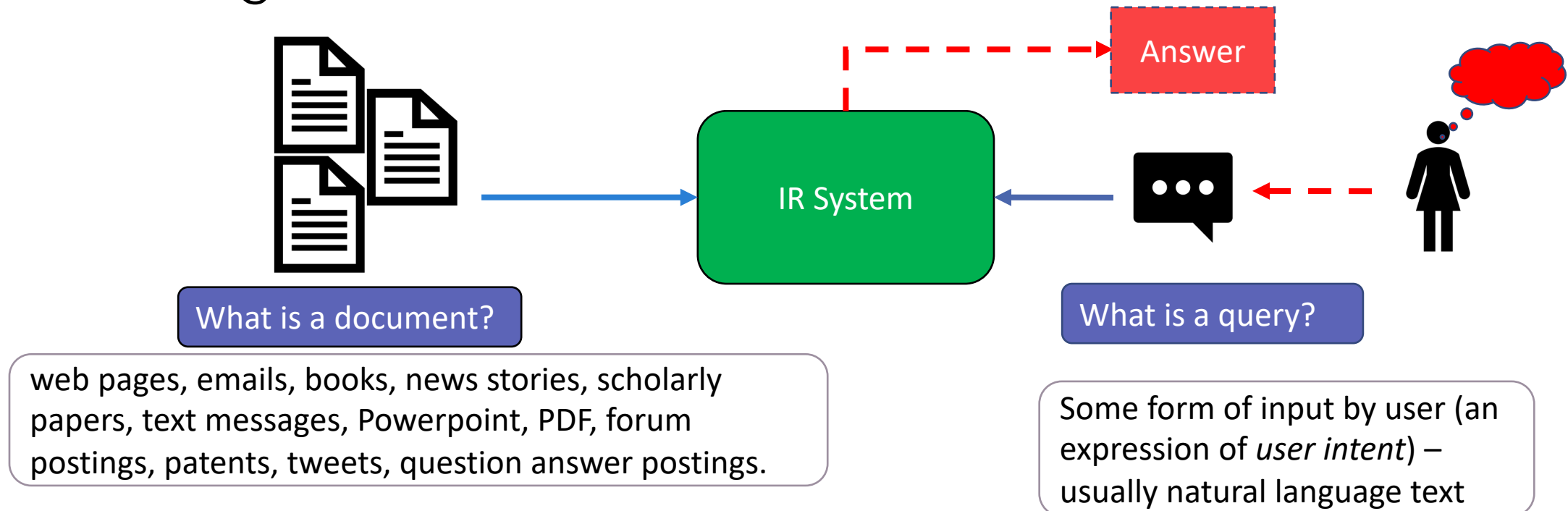
- ❖ Read all the web & remember what information is where
- ❖ Be able to reason about connections between information
- ❖ *Read my mind and answer questions (or better yet) satisfy my needs, even before I articulate them 😊*



**How close is this dream to reality?
Are we being too easy on our search engines?**

Information Retrieval (formally)

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.



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

!!!Some say entire AI (conceptually)
is an extension of database
systems!!!

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.

A thought bubble with a blue outline and a scalloped edge, containing the text "Efficiency in terms of ..?".

Efficiency in terms of ..?

A thought bubble with a blue outline and a scalloped edge, containing the text "What is relevance?".

What is relevance?

IR over text and other modes

- IR does not necessarily deal with text data.
 - Images, text, speech, what else?
- Both documents and queries can be in other modes.
- In this course, we will concentrate on textual IR.
 - Term project, image search might be included (optional).
 - Multi-lingual/cross-lingual search

Typical IR Tasks

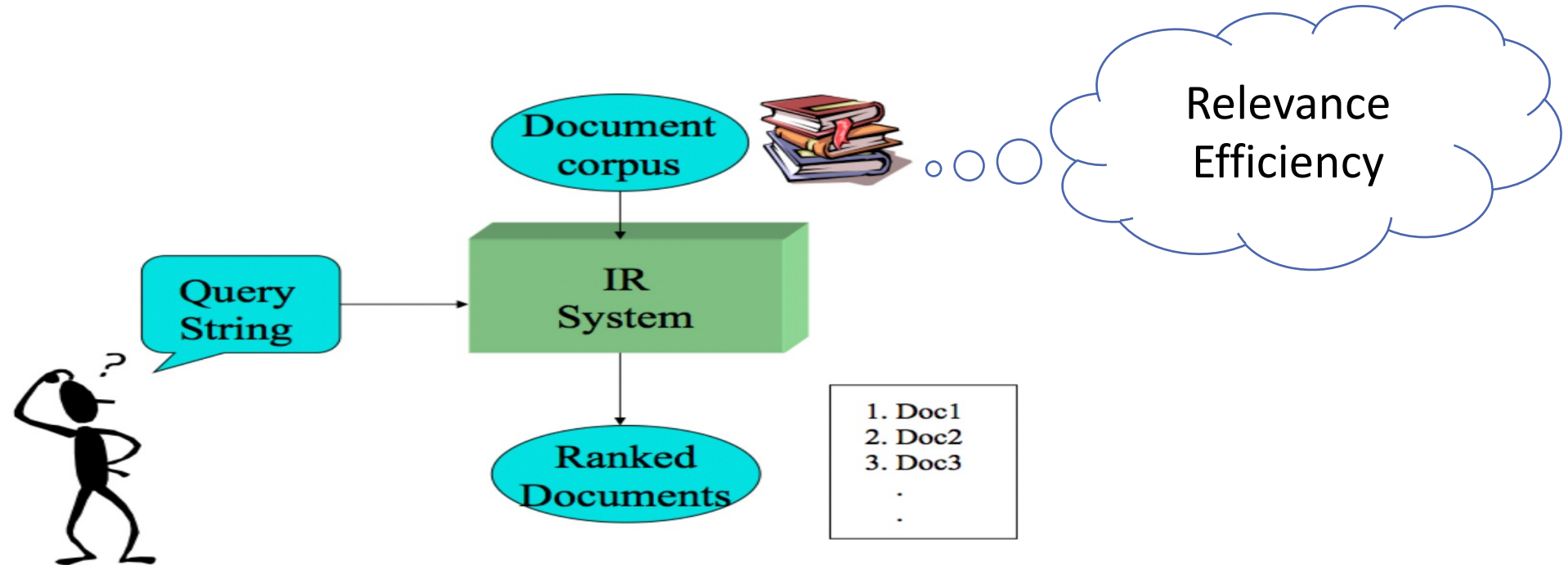
Given:

- 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

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).

Simpl(er) Notion of Relevance

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

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 vs. Coffee)
- 'Fall' (season/verb)

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.
- Estimate your "*thoughts*" (user intent)
- ...
- *Fair, ethical, transparent, privacy-preserving, secure ...*

What will you learn in this IR?

- ❖ (Some basic idea about) How search engines work
 - ❖ The Software/algorithm side.
 - ❖ Hardware side: http://videolectures.net/wsdm09_dean_cblirs/
 - ❖ How to make money out of it?
- ❖ Can web be seen as a collection of (semi)structured data/knowledge bases?
 - ❖ Unstructured → semi-structured
- ❖ Can we exploit the connectedness of the web pages? And How?
- ❖ (Will touch upon) Connections between NLP and IR.

Where to keep the tab on?

- Top Conferences in the field
 - SIGIR
 - WWW
 - ISDM
 - ECIR

- Language Conferences
 - EMNLP
 - ACL
 - CoNLL

Active Areas of Research (Workshop Titles)

- What to Retrieve
- Search Experience
- Personalization, Behavior, Conversation, Social, etc.
- *Cross-lingual/Multi-lingual search*
- *Multi-modal search*
- *Image Search*
- *Video Search*
- *Semantic Search*
- *ML/DL Efficiency for Web*
- *FATES*

[WWW 2021 Workshops \(a snapshot\)](#)

- Temporal Web-analytics
- Fairness, Accountability, Transparency, Ethics and Society on the Web (FATES 2021)
- Cross-lingual Event-centric Open Analytics
- Data-efficient Machine Learning for Web Applications (DeMaL)
- Scientific Knowledge Representation, Discovery, and Assessment (Sci-K)
- Natural Language Processing for Social Media (SocialNLP 2021)
- Deep Reinforcement Learning for Knowledge Discovery
- Knowledge Graphs for Online Discourse Analysis

What to Retrieve

- Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval. SIGIR 2015.
- On Application of Learning to Rank for E-Commerce Search. SIGIR 2017.
- Concept Embedded Convolutional Semantic Model for Question Retrieval. WSDM 2017.
- Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale. SIGIR 2016.
- Toward an Interactive Patent Retrieval Framework based on Distributed Representations. SIGIR 2018.
- ANNE: Improving Source Code Search using Entity Retrieval Approach. WSDM 2017.
- Exploiting Food Choice Biases for Healthier Recipe Recommendation. SIGIR 2017.
- Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos. SIGIR 2019.

Search Experience

- *Engaged or Frustrated? Disambiguating Emotional State in Search. SIGIR 2017.*
- *Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. SIGIR 2018.*
- *Understanding and Modeling Success in Email Search. SIGIR 2017.*
- *Using Information Scent to Understand Mobile and Desktop Web Search Behavior. SIGIR 2017.*

Personalization, Behavior, Conversation, Social, Bias, Fairness

- The Utility and Privacy Effects of a Click. SIGIR 2017.
- Predicting Which Topics You Will Join in Future on Social Media, SIGIR 2017
- Why People Search for Images using Web Search Engines. WSDM 2018.
- Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. SIGIR 2019.
- How do Biased Search Result Rankings Affect User Attitudes on Debated Topics?. SIGIR 2021
- *(Slightly Different – SM) Engagement Patterns of Peer-to-Peer Interactions on Mental Health Platforms, ICWSM 2020*

What will we cover?

- Boolean retrieval
- The term vocabulary & postings lists
- Skip Pointers, Phrase Queries and Positional Indexing
- Scoring, term weighting & the vector space model
- Dictionaries and Tolerant Retrieval
- Evaluation in information retrieval
- Index Construction and Compression
- Relevance feedback & query expansion
- *Probabilistic information retrieval*
- *Language models for information retrieval (+Current LM Primer)*

Course Contents (Tentative)

- *Link analysis – HITS, PageRank*
- *Word Vectors*
- ***Classification and Clustering with Vectors***
- *Learning to Rank*
- *Neural IR*

Tutorial: DL/NLP/PyTorch Primer

- ***Excluded (due to time)***
 - *Semantic Web, OWL, Image Retrieval, Cross-lingual/Cross-modal retrieval, Mathematical formula search*

Intelligent Logical Trusted Agents

CVIU '17, AAI '18, IJCAI ('15, '19)
UAI '18, WACV '19



See



Read

TaxiNLI, CoNLL 2020
TaxiXNLI, EMNLP MRL '21
*CheckList NLI**
*Multi-Hop NLI**

Ontology
Common-Sense

Knowledge

Learning

Reasoning

Logic

Machine Learning
Deep Learning

Semantic
Web/OWL

Embedding-
based IR