CADT

Cambodia Academy of Digital Technology

Introduction to

“Information Web Retrival Analytics”

Lecturer

Mr. Natt KORAT

Lecturer-Researcher,

Institute of Digital Research & Innovation (IDRI),

Cambodia CADT Academy of Digital Technology (CADT)

About me

- **Academic background**
  - Master of computer science specialize in AI and data science (2026), CADT
  - Bachelor of computer science (2024), AGA
  - Bachelor of Psychology (2022), RUPP
- **Work experience**
  - Lecturer-Researcher, IDRI, CADT, 2026 – Present
  - Assistant Lecturer-Researcher, IDRI, CADT, 2022-2026
  - Data engineeer, Z1 data, 2021-2022

Same query, two engines

- **In pairs**: pick a query, search it on 2 different search services, compare top 3 results.

Python	Coffee near me	Phnom Penh to Siem Reap

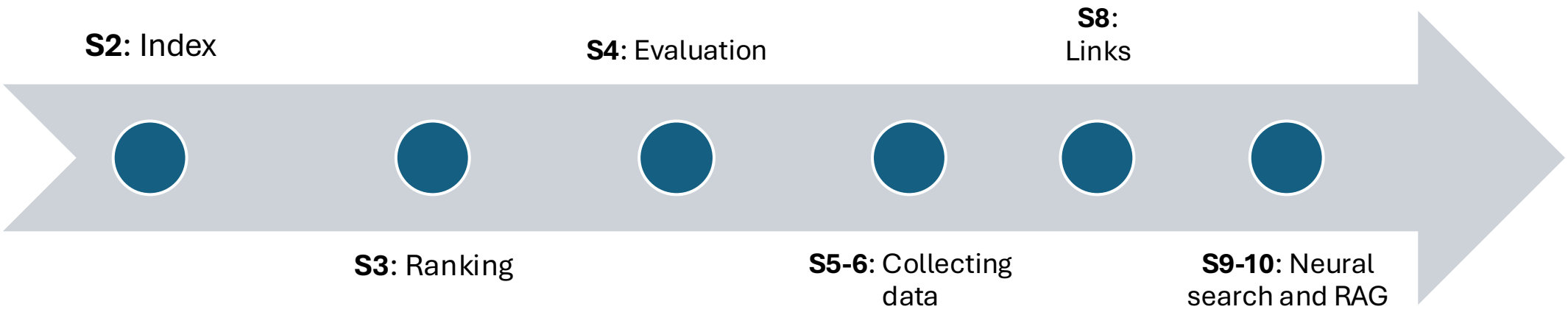
- What’s your observation?

Search is everywhere

- Web, email, phone, e-commerce, code, legal and medical papers, company documents.
- Chatbots that “read your documents” start with retrieval RAG

What we will build

- A small search engine over a Web corpus, in teams



How the course run

- **Each week**: 1.5 h lecture + 3 h lab, short quiz every beginning of the lab session
- Attendance 10%, quiz & homework 15%, project & participation 35%, midterm 20%, and final 20%
- Reading guide and project milestones

What is information retrieval (IR)?

- Finding material (usually documents) of an unstructured nature such as text, images, audios, etc. that satisfies an information need, from large collections.
- Three key words:
  - Unstructured
  - Information need
  - Large collections



Structured vs. unstructured data

ID	NAME	CLASS	MARK	GENDER
1	John Deo	Four	75	female
2	Max Ruin	Three	85	male
3	Arnold	Three	55	male
4	Krish Star	Four	60	female
5	John Mike	Four	60	female
6	Alex John	Four	55	male
7	My John Rob	Five	78	male
8	Asruid	Five	85	male
9	Tes Qry	Six	78	male
10	Big John	Four	55	female



Khmer language

114 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

"Cambodian language" redirects here. For other languages spoken in Cambodia, see Demographics of Cambodia § Languages.

Khmer (/kəˈmɛər/ *kə-MAIR*;[3] ខ្មែរ, UNGEGN: *Khmêr* [kʰmae]) is an Austroasiatic language spoken natively by the Khmer people and is an official language and national language of Cambodia. The language is also widely spoken by Khmer people in Eastern Thailand and Isan, Thailand, as well as in the Southeastern and Mekong regions of Vietnam.

Khmer has been influenced considerably by Sanskrit and Pali especially in the royal and religious registers, through Hinduism and Buddhism,[4] due to Old Khmer being the language of the historical empires of Chenla and Angkor.

The vast majority of Khmer speakers speak *Central Khmer*, the dialect of the central plain where the Khmer are most heavily concentrated. Within Cambodia, regional accents exist in remote areas but these are regarded as varieties of Central Khmer. Two exceptions are the speech of the capital, Phnom Penh, and that of the Khmer Khe in Stung Treng province, both of which differ sufficiently enough from Central Khmer to be considered separate dialects of Khmer.

Outside of Cambodia, three distinct dialects are spoken by ethnic Khmers native to areas that were historically part of the Khmer Empire. The Northern Khmer dialect is spoken by over a million Khmers in the southern regions of Northeast Thailand and is treated by some linguists as a separate language. Khmer Krom, or Southern Khmer, is the first language of the Khmer of Vietnam, while the Khmer living in the remote Cardamom Mountains speak a very conservative dialect that still displays features of the Middle Khmer language.

Khmer is primarily an analytic, isolating language. There are no inflections, conjugations or case endings. Instead, particles and auxiliary words are used to indicate grammatical relationships. General word order is subject–verb–object, and modifiers follow the word they modify. Classifiers appear after numbers when used to count nouns, though not always so consistently as in languages like Chinese. In spoken Khmer, topic-comment structure is common, and the perceived social relation between participants determines which sets of vocabulary, such as pronouns and honorifics, are proper.

Khmer
Cambodian
ភាសាខ្មែរ / *Phéasa Khmêr* ខេមរភាសា / *Khemôrôphéasa*
ភាសាខ្មែរ
Phéasa Khmêr ('Khmer language') written in Khmer script
Pronunciation [pʰiə.ˈsaː kʰmae] [kʰɛ.maʔ.raʔ.pʰiə.ˈsaː]
Native to Cambodia Thailand (East and Isan) Vietnam (Mekong Delta and Southeast)
Ethnicity Khmer
Speakers L1: 19 million (2019)[1] L2: 1 million (2024)[1] Total: 21 million (2019–2024)[1]
Language family Austroasiatic • **Khmer**
Early forms Proto-Khmeric • Old Khmer • Middle Khmer
Writing system Khmer script Khmer Braille
Official status
Official language in Cambodia
Recognised minority language in Thailand Vietnam
Regulated by Royal Academy of Cambodia, National Council of Khmer Language[2]

Core Vocabulary

- Document -> collection(corpus) -> information need -> query -> relevance -> ranked results
- Example need: “I need to get from Phnom Penh to Siem Reap tonight”
- Queries: “bus phnom penh siem reap” or “cheap transport siem reap” …

Relevance

- Depends on the user, the task, the context, and the time
- Topical relevance vs. usefulness
- Binary or graded judgments, made by people

Why is it hard? Vocabulary mismatch

- Car | automobile | vehicle
- Cheap | budget | affordable
- Ambiguity: Java, Python, Apple
- Short queries, long documents
- Mixed languages

IR vs. database queries

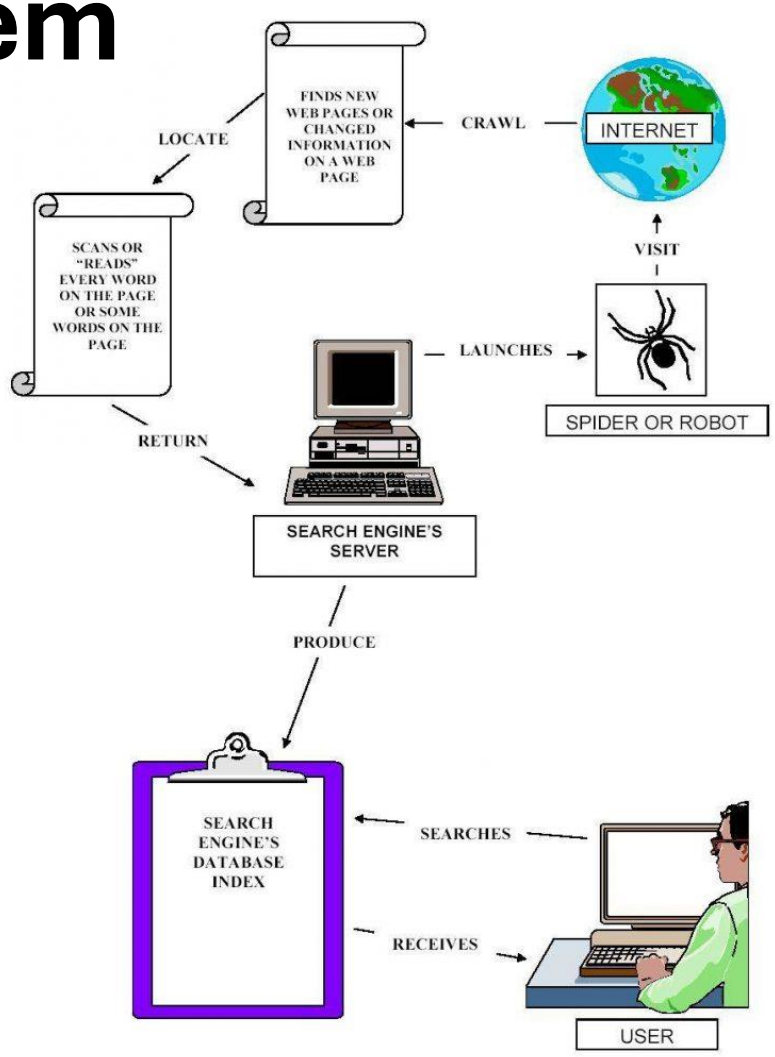
	IR	Database queries
Data	Unstructured text	Structured records
Query	Keywords or a question	SQL
Matching	Partial and approximate	exact
Result	Ranked list	A set of rows
Success	Degrees of relevance	Right or wrong

Where IR fits

- Web search
- Recommender systems
- Question answering
- RAG

Anatomy of a search system

- **Offline**: collect -> preprocess -> index
- **Online**: query -> process query -> retrieve -> rank -> present
- **Feedback loop**: logs, clicks, evaluation



Offline: build the index

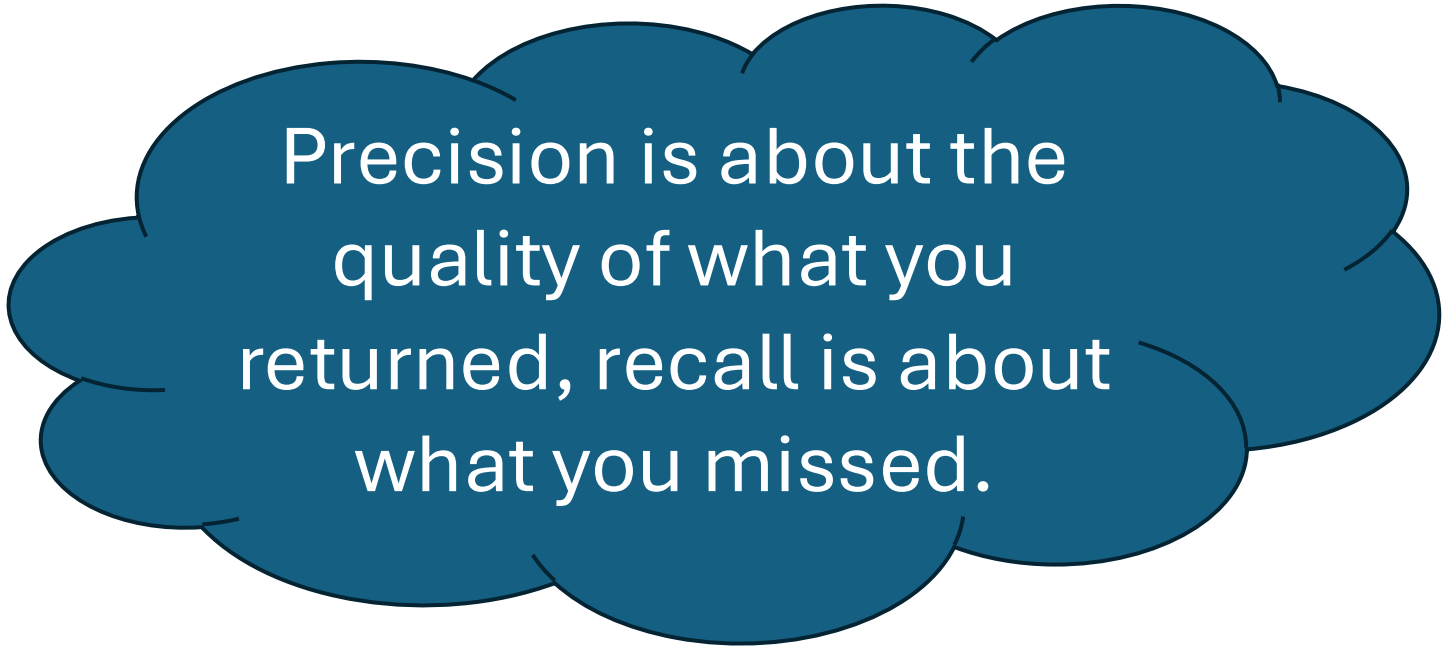
- Collect documents (crawl, feeds, files)
- Preprocess: clean, tokenize, normalize
- Index: structure that makes lookup fast
- Why offline: pay the cost once, answer man queries quickly

Online: answer the query

- Interpret the query
- Match against the index
- Score and rank
- Present snippets
- User expect answers in a fraction of a second

How do we know it works?

- Relevant documents in the collection 20; return 10; relevant among returned: 6
- Precision: 6/10
- Recall: 6/20
- Coming in session 4!



Recap

1. IR finds documents that satisfy an information need
2. A query is not the need, and relevant is judged by people
3. Databases and search engines have different goals
4. Offline indexing and online ranking
5. Language, scale and evaluation make it hard