Information Retrieval and Recommender Systems

Guide to AI and Data Courses · Part IV: Data

Indexing | Query and ranking | Collaborative filtering | Evaluation | Dense vector search | Neighboring courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides:
https://sercankulcu.github.io/files/ai_courses/slides/Chapter_28_Information_Retrieval_and_Recommender_Systems.html

INTRODUCTION

What Question Does It Ask?

> **How are the most relevant documents or products picked from millions of candidates?**

Information retrieval studies systems that find the documents most relevant to a query among millions and **rank** them. Recommender systems do the same job without a query: they predict what a user will like from past behavior. Both return a ranked list, not an answer.

At a glance

APPROACH	INPUT	OUTPUT
matching and ranking	query, documents, user history	ranked result list

SUCCESS MEASURE	MATH LOAD	DATA NEEDS
precision, recall, NDCG	medium: vector spaces	lots of interaction data

A Short History

1972
Spärck Jones proposes inverse document frequency (IDF): a rare term says more.

1975
Salton and colleagues publish the vector space model: documents and queries are term vectors.

1992
Xerox PARC's Tapestry system coins the term "collaborative filtering".

1998
Brin and Page introduce the Google search engine, built on PageRank.

2006
The Netflix Prize begins; matrix factorization stands out in the winning solutions.

2020
Dense vector retrieval and retrieval-augmented generation (RAG) begin to spread.

THE COURSE

Core Concepts

Inverted index
For each term, a list of the documents containing it.

TF-IDF and BM25
Scoring documents by term frequency and rarity.

Vector space model
Measuring similarity by the cosine of the angle.

Precision, recall
Share of results that are relevant; share of relevant ones found.

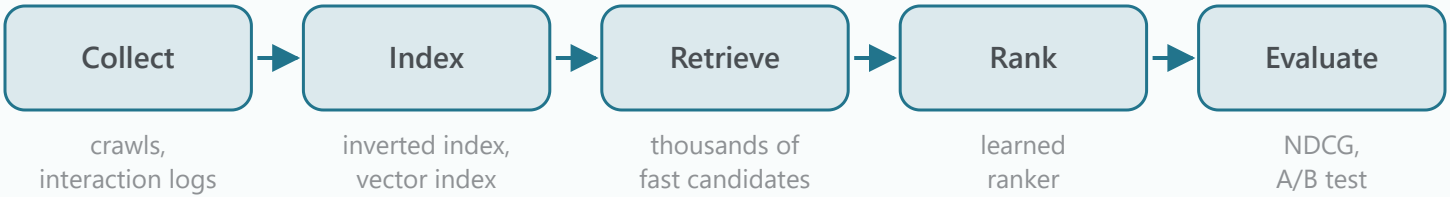
Ranking metrics
MAP, MRR, NDCG: position in the list matters.

Collaborative filtering
Recommending from what similar users liked.

Matrix factorization
Splitting the user-item table into latent factors.

Cold start
New users or items with no history.

A Typical Workflow



Example: A video platform collects watch logs, indexes videos as vectors, retrieves a few hundred candidates from millions, ranks them by probability of being watched, and compares the new ranker with the old one in an A/B test.

THE COURSE

What Does the Course Cover?

1. Boolean retrieval and the inverted index
2. Index construction, compression and dictionaries
3. Text preprocessing: tokenization, stemming
4. Scoring: TF-IDF and the vector space model
5. Probabilistic retrieval and BM25
6. Evaluating retrieval: precision, MAP, NDCG
7. Link analysis: PageRank and HITS
8. Learning to rank
9. Dense retrieval and approximate nearest neighbours
10. Content-based recommendation
11. Collaborative filtering and matrix factorization
12. Evaluating recommenders: offline metrics, A/B tests

Skills you will gain

building a search index in Elasticsearch · BM25 experiments with Pyserini · vector search with FAISS · training recommenders with Surprise · measuring rankings with NDCG and MAP

A typical semester; the order and coverage vary between universities.

TRY IT

Searching a Help Center

A shopping site's help center has nine articles, and customers type questions in their own words. Pick a sample question or type your own; articles are ranked by TF-IDF weights and cosine similarity.

Article	cosine similarity
1. Damaged item arrived	0.45
2. Where is my order?	0.31
3. How do I cancel my order?	0.17
4. What is the shipping fee?	0.05
5. Can I change the billing address?	0.04
6. When is my money refunded?	0.03
7. I forgot my password	0.03
8. How do I return an item?	0.00
9. Paying in instalments	0.00

my order arrived damaged

arrived damaged | shipping fee | refund | password

TF-IDF | Counts only (TF)

Term stems and IDF: my **0.81**, orde **0.81**, arri **1.50**, dama **2.20**
Matching articles: **7** / 9 · gap between 1st and 2nd: **0.13**

Top result: Photograph the **damaged** item and write to us within 48 hours; we will exchange or refund it.

IDF down-weights words such as “my” and “order” that appear in many articles, so the distinctive “damaged” stands out. Compare with the TF button. Cutting each word to its first four letters is a crude stemmer: damage, damaged → dama.

CONTEXT

Prerequisites

Courses worth knowing before this one, and why they are needed.

Course	Why it is needed
Linear Algebra	vector space model, matrix factorization
Data Structures	hash tables, trees and priority queues: the index and the top k results
Algorithms	sorting, merging and complexity: speed on large indexes

CONTEXT

Are You Ready for This Course?

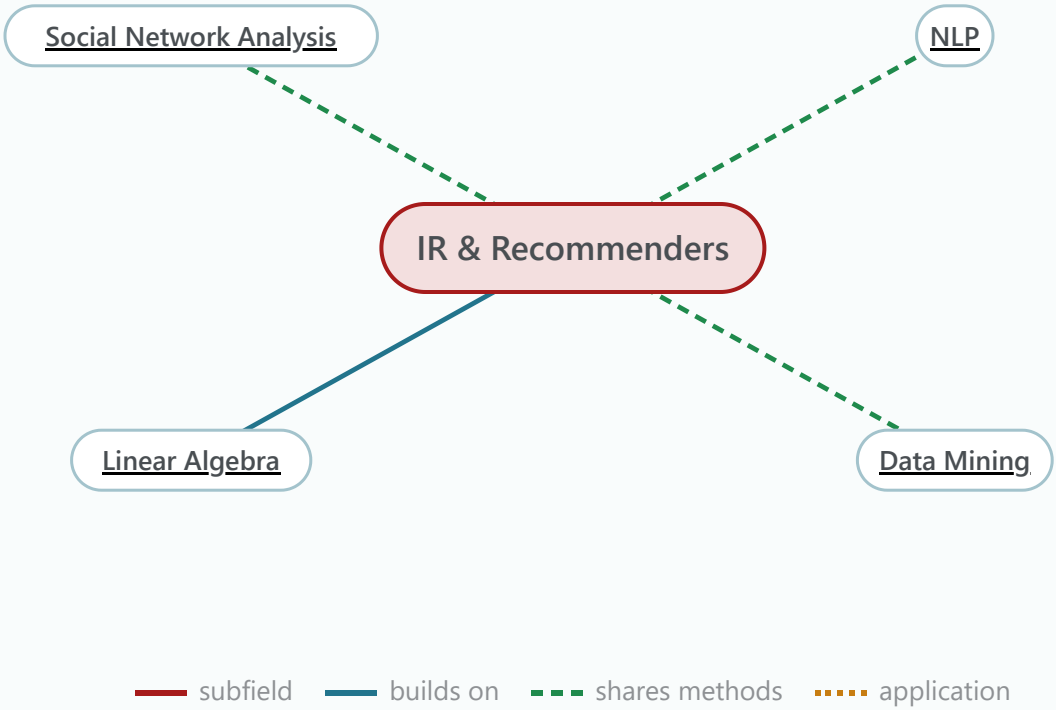
Ask yourself; for every "no", the right-hand column shows where to start.

Ask yourself	If not, start with
Can you compute and interpret the cosine similarity of two vectors?	Linear Algebra
Can you explain why lookup in a hash table takes constant time?	A data structures course: hash tables and trees
Can you compute precision and recall from a confusion matrix?	Machine Learning
Can you split a text into words and count word frequencies?	Natural Language Processing and Text Mining

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

Neighbouring Courses



Overlapping Courses

Relation	Course	Shared ground
builds on	This course builds on Linear Algebra	vector space model, matrix factorization
shares methods	Shares methods with Natural Language Processing and Text Mining	text representation, TF-IDF, semantic search
shares methods	Shares methods with Data Mining	association, collaborative filtering
shares methods	Shares methods with Social Network Analysis	PageRank, link analysis

DIFFERENCES

Where It Differs (1/2)

Criterion	IR & Recommenders	Data Mining	NLP
Core question	How are the most relevant documents or products picked from millions of candidates?	How are hidden, useful patterns found in large piles of data?	How does a computer analyze, understand and generate human language?
Approach	matching and ranking	discovery-driven analysis	linguistics, statistics and deep learning
Input	query, documents, user history	large tables or transaction data	text
Output	ranked result list	rules, clusters, outliers	label, translation, summary, answer
Success measure	precision, recall, NDCG	interestingness, support, confidence	accuracy, BLEU, human evaluation
Math load	medium: vector spaces	medium	medium
Data needs	lots of interaction data	large	large text corpora

- **Data Mining:** Data mining discovers patterns in data; this course turns discovery into a ranked list for one user's current need.
- **NLP:** NLP analyzes the structure and meaning of language; this course represents text only to score and rank relevance.

DIFFERENCES

Where It Differs (2/2)

Criterion	IR & Recommenders	Machine Learning	Social Network Analysis
Core question	How are the most relevant documents or products picked from millions of candidates?	How does a program generalize from examples?	How are networks of people and their relationships analyzed?
Approach	matching and ranking	statistical learning	graph theory and statistics
Input	query, documents, user history	labeled or unlabeled examples	nodes and links
Output	ranked result list	a model that makes predictions	centrality, communities, diffusion
Success measure	precision, recall, NDCG	error on unseen data	modularity, interpretability
Math load	medium: vector spaces	high: linear algebra, probability, optimization	medium: graphs, matrices
Data needs	lots of interaction data	moderate to large	network data

- **Machine Learning:** ML tries to predict individual examples correctly; this course cares about list order, above all the top few results.
- **Social Network Analysis:** Social network analysis studies the structure of ties; this course uses that structure (e.g. PageRank) only as a ranking signal.

PRACTICE

Applications

Web search
Ranking the pages most relevant to a query.

E-commerce
Customers who bought this also bought...

Video and music
Personal feeds and playlists.

Enterprise search
Searching company documents and e-mail.

News feeds
Ranking posts by relevance.

RAG
Fetching the right documents for a language model.

PRACTICE

Risks and Ethics

Filter bubbles

Always showing more of what a user liked narrows the range of views; polarization from news and video feeds is debated.

Popularity bias

Popular items get recommended more, so they get clicked more; new and small producers stay invisible.

Query logs give people away

From search logs that AOL anonymized and released in 2006, journalists identified some of the users.

Ranking manipulation

Link farms, fake reviews and keyword stuffing distort rankings; the system must be designed against adversaries.

In depth: Explainable AI and AI Ethics.

PRACTICE

Tools and Datasets

Tools and libraries

Apache Lucene | Elasticsearch | FAISS | Pyserini | Surprise | scikit-learn

Well-known datasets

MovieLens | Netflix Prize | MS MARCO | TREC collections | Cranfield collection | Amazon product reviews

PRACTICE

Common Misconceptions

Common belief	In fact
A search engine just finds the words in the query.	The real work is **ranking** thousands of matches; users rarely look past the first page.
Recommending the top-rated items is enough.	The same list for everyone is not personalization; diversity and novelty are measured too.
High accuracy means good recommendations.	A model with low rating error can still rank badly; rank metrics such as NDCG are needed.
Dense vectors have made classic search obsolete.	BM25 is still a strong baseline; many systems combine the two.

NEXT STEPS

After This Course

Courses that build on or extend this one

This course is mostly an end point: it builds on the others.

Related careers

Search engineer | Recommender systems engineer | Data scientist | Ranking researcher

After this course you can move on to semantic search with natural language processing, link-based ranking with social network analysis, and indexing billions of documents with big data systems. Retrieval-augmented generation connects this course directly to large language models.

NEXT STEPS

Further Reading

- Manning, Raghavan and Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- Baeza-Yates and Ribeiro-Neto, *Modern Information Retrieval*, 2nd ed., Addison-Wesley, 2011.
- Aggarwal, *Recommender Systems: The Textbook*, Springer, 2016.
- Koren, Bell and Volinsky, "Matrix Factorization Techniques for Recommender Systems", *IEEE Computer* 42(8), 2009.

The End

Chapter 28: Information Retrieval and Recommender Systems

← Previous chapter · Map of the field · Next chapter →