دانشگاه بوعلی سینا

Information Retrieval

Amin Nazari

Spring 2025

Outline

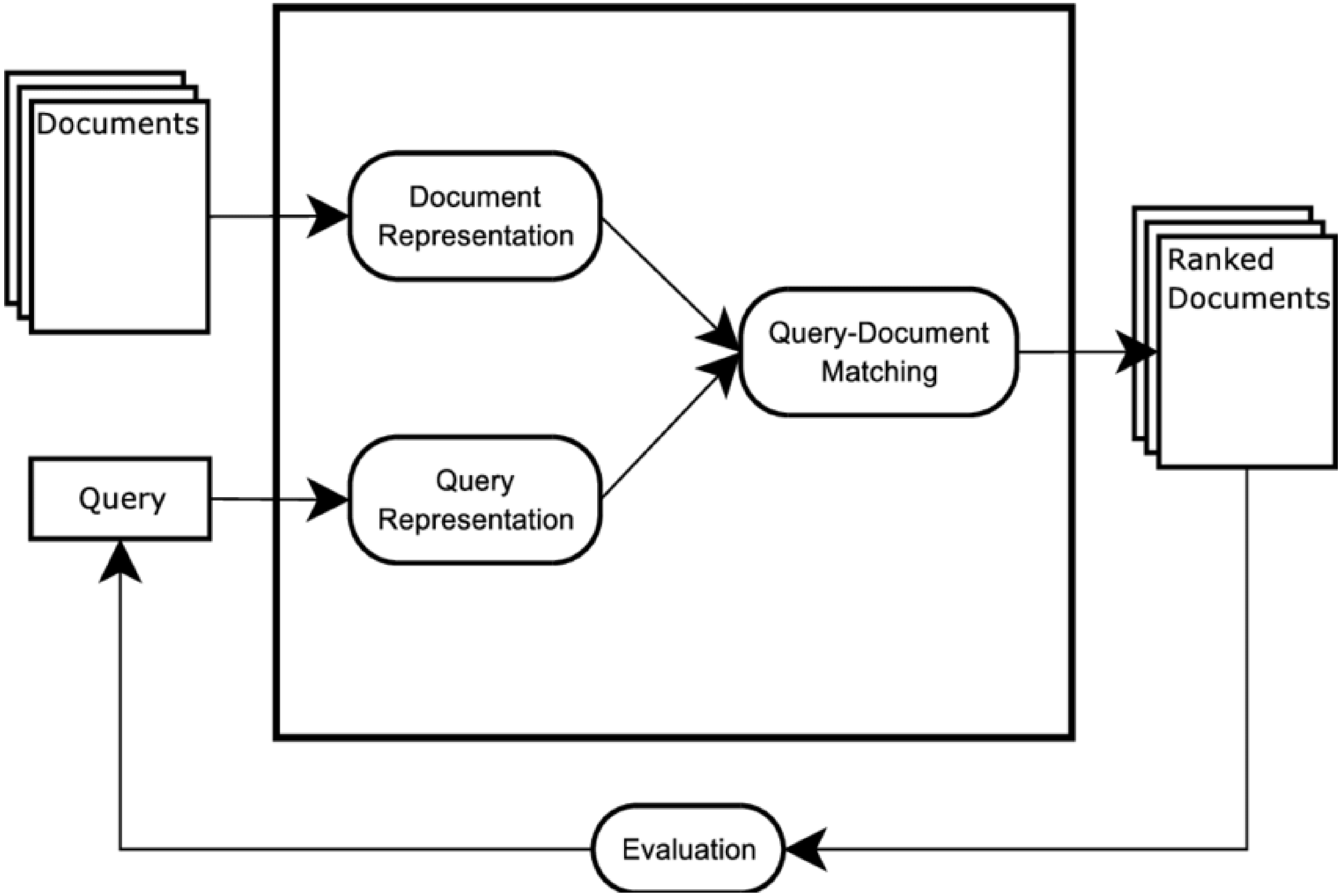
- What is Information Retrieval?
- IR applications
- IR vs Information Storage and Retrieval
- IR vs Data retrieval (DBMS)
- Why study it?
- Reference
- Outlines
- Grading

What is Information Retrieval?

- Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- Unstructured data types
  - Text
  - Audio
  - Image
  - Video
- Why text?

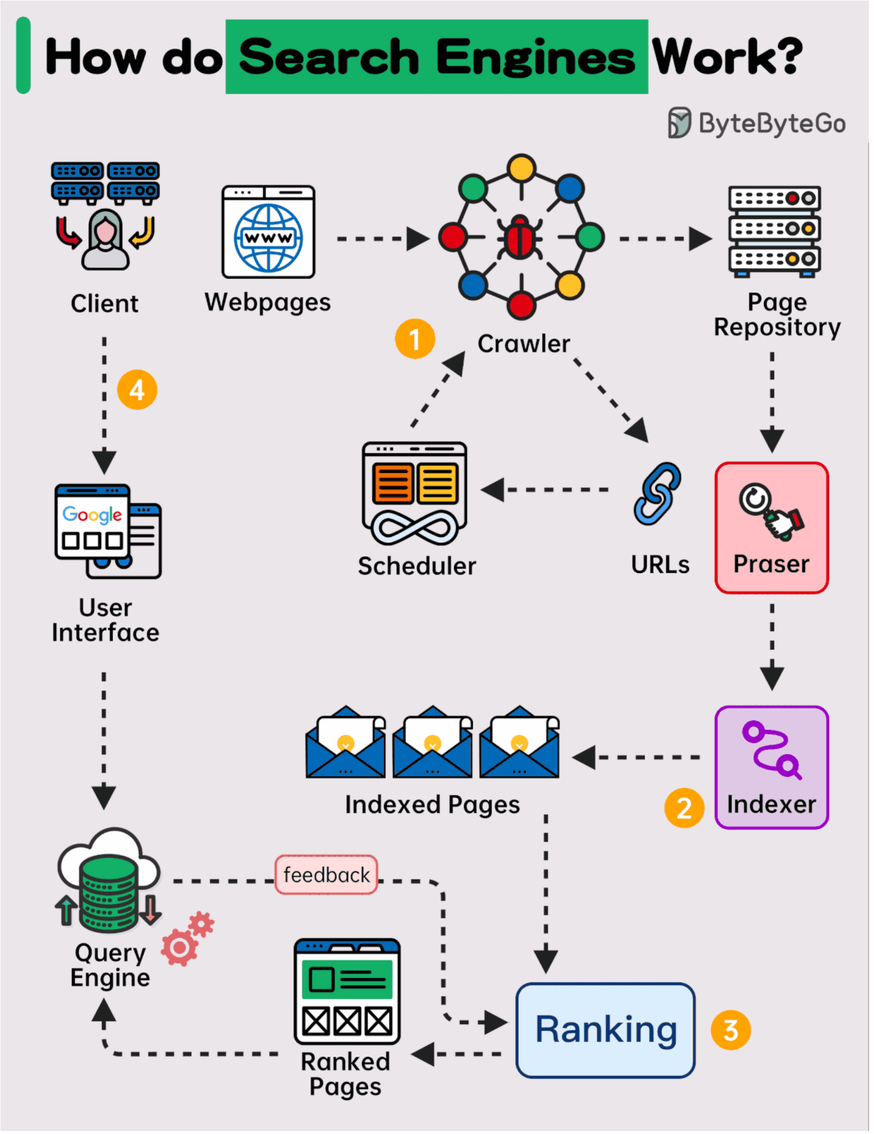
IR schema


Documents
Document Representation
Query-Document Matching
Ranked Documents
Query
Query Representation
Evaluation

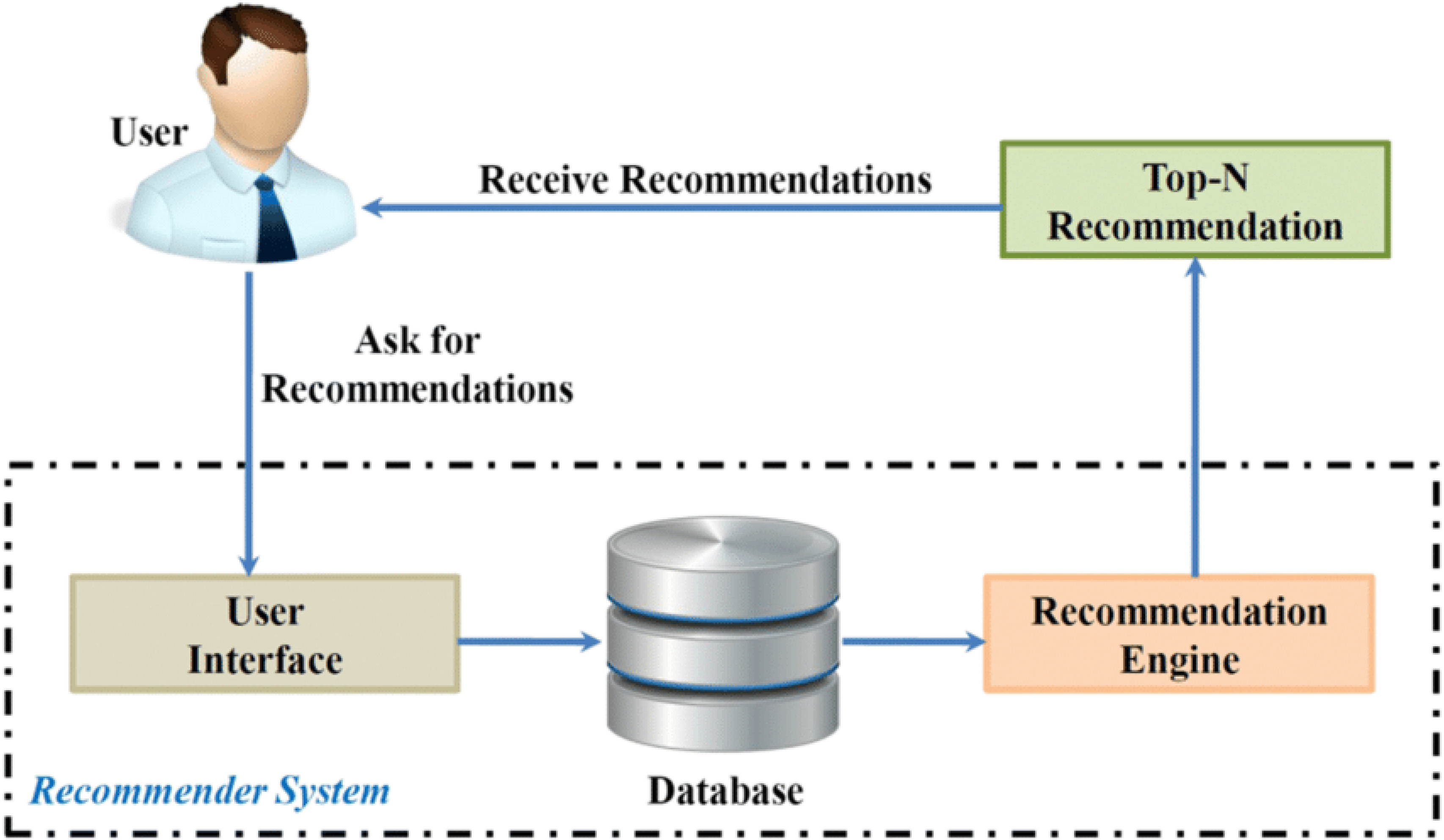
Applications of IR

- Search engine
- Recommender systems (News, Movies, Posts, Usres, Books, Tour, Music, Drug, …)
- Documentation management (Digital Library)
- Question Answering Systems
- Plagiarism Detection

Search Engines



RecSys



User

Receive Recommendations

Top-N Recommendation

Ask for Recommendations

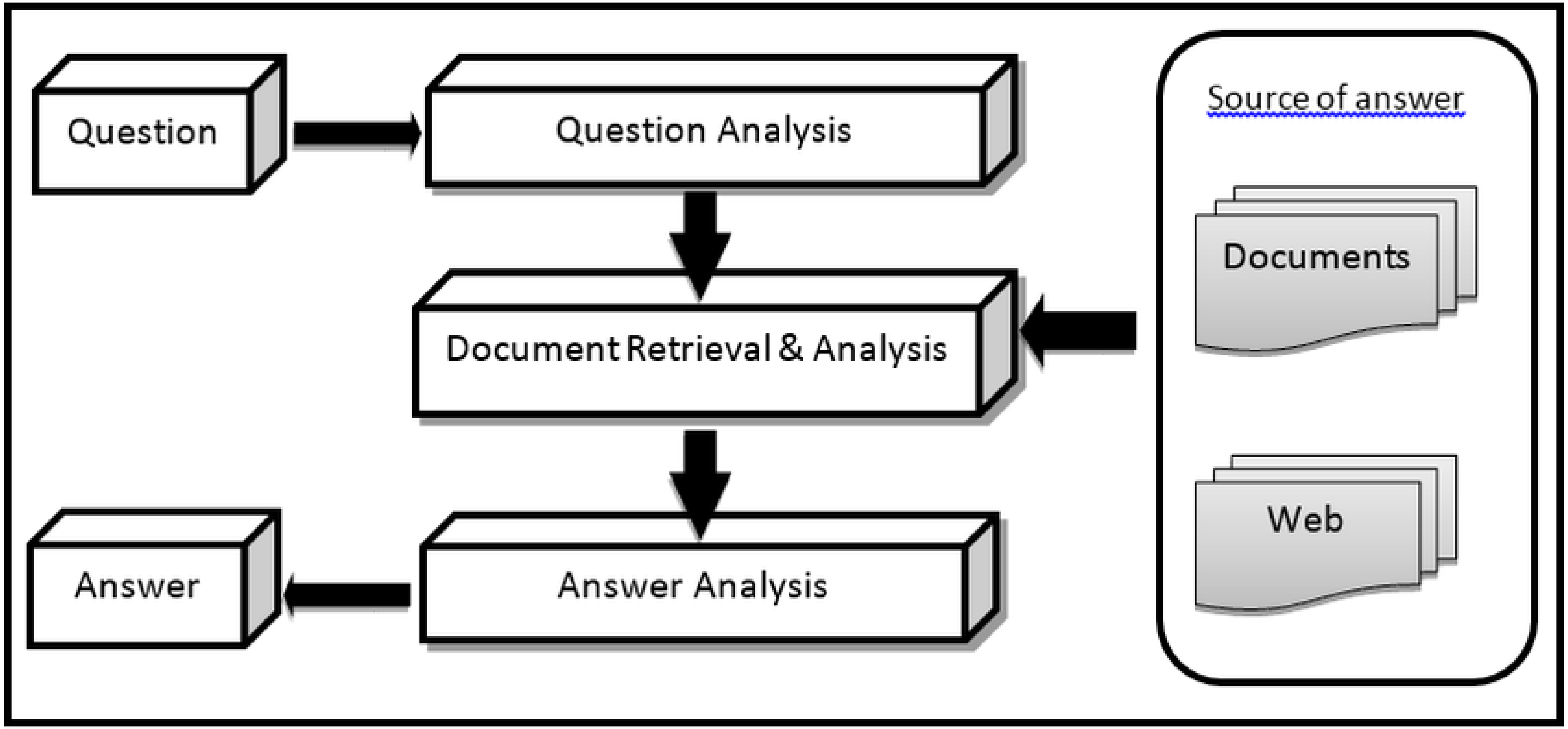
User Interface

Database

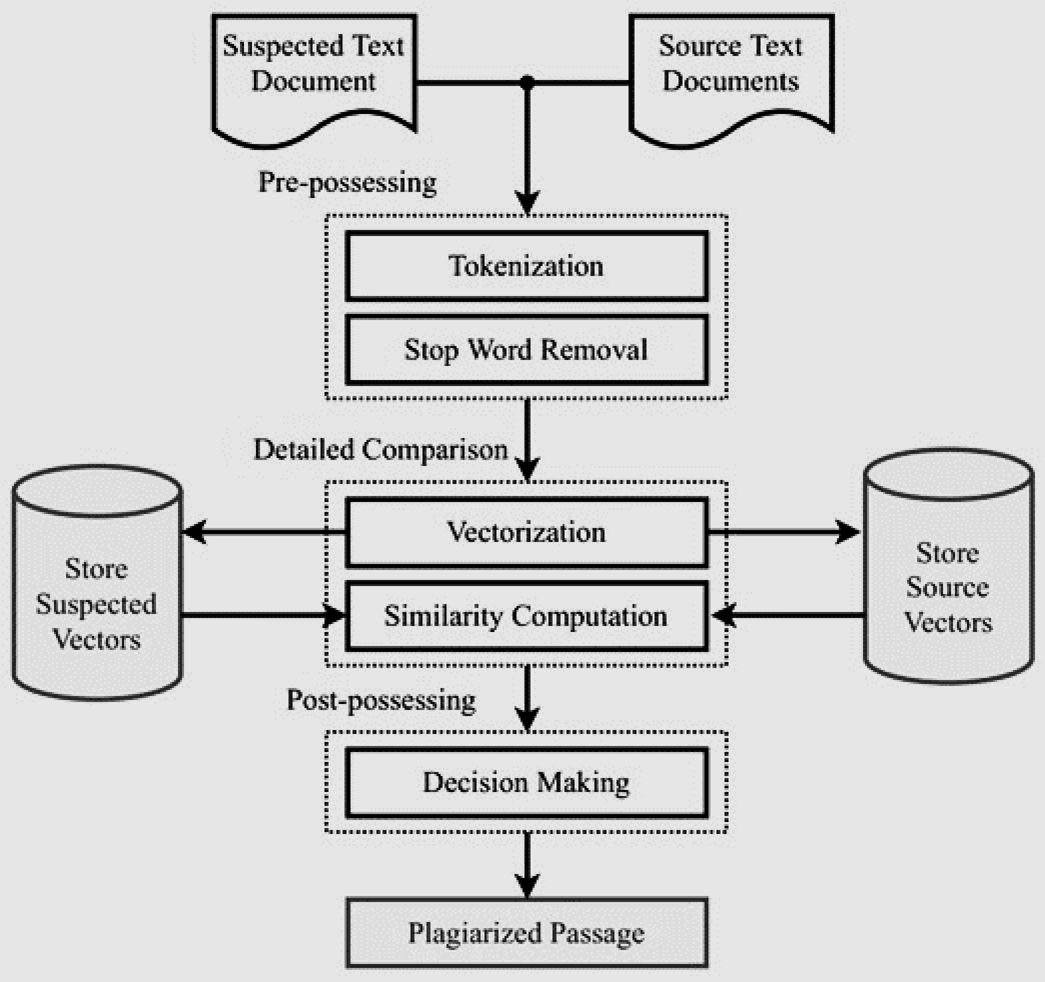
Recommendation Engine

Recommender System

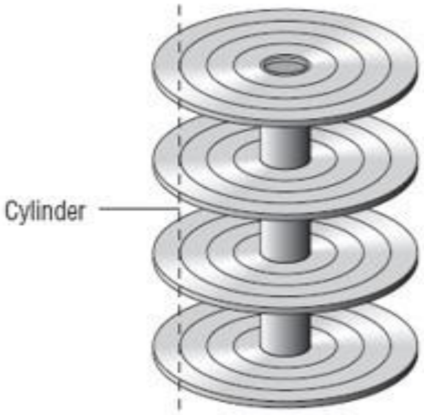
Question Answering Systems

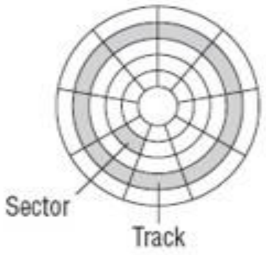


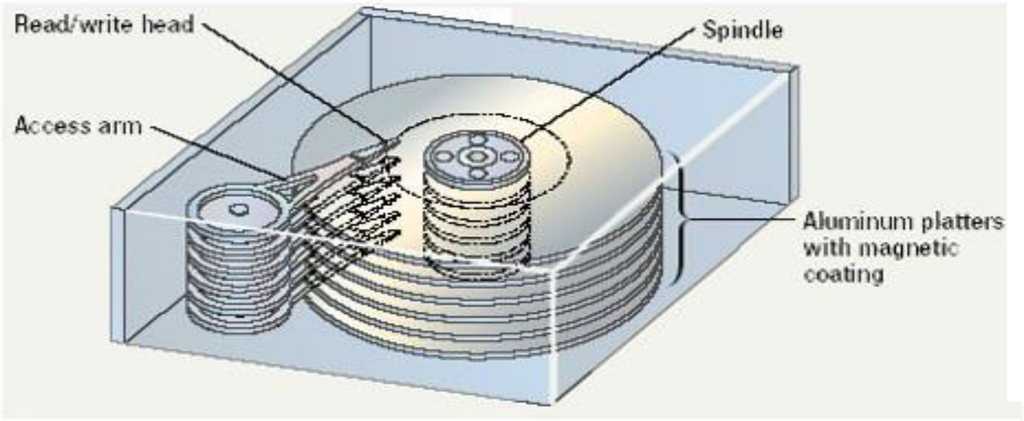
Plagiarism Detection

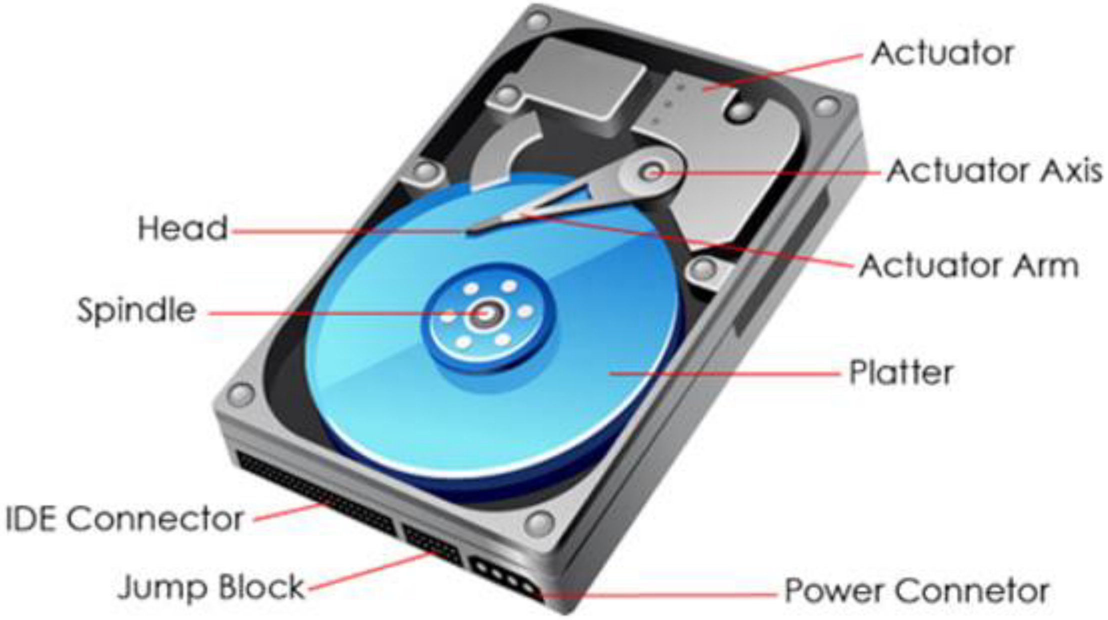


Information Storage and Retrieval

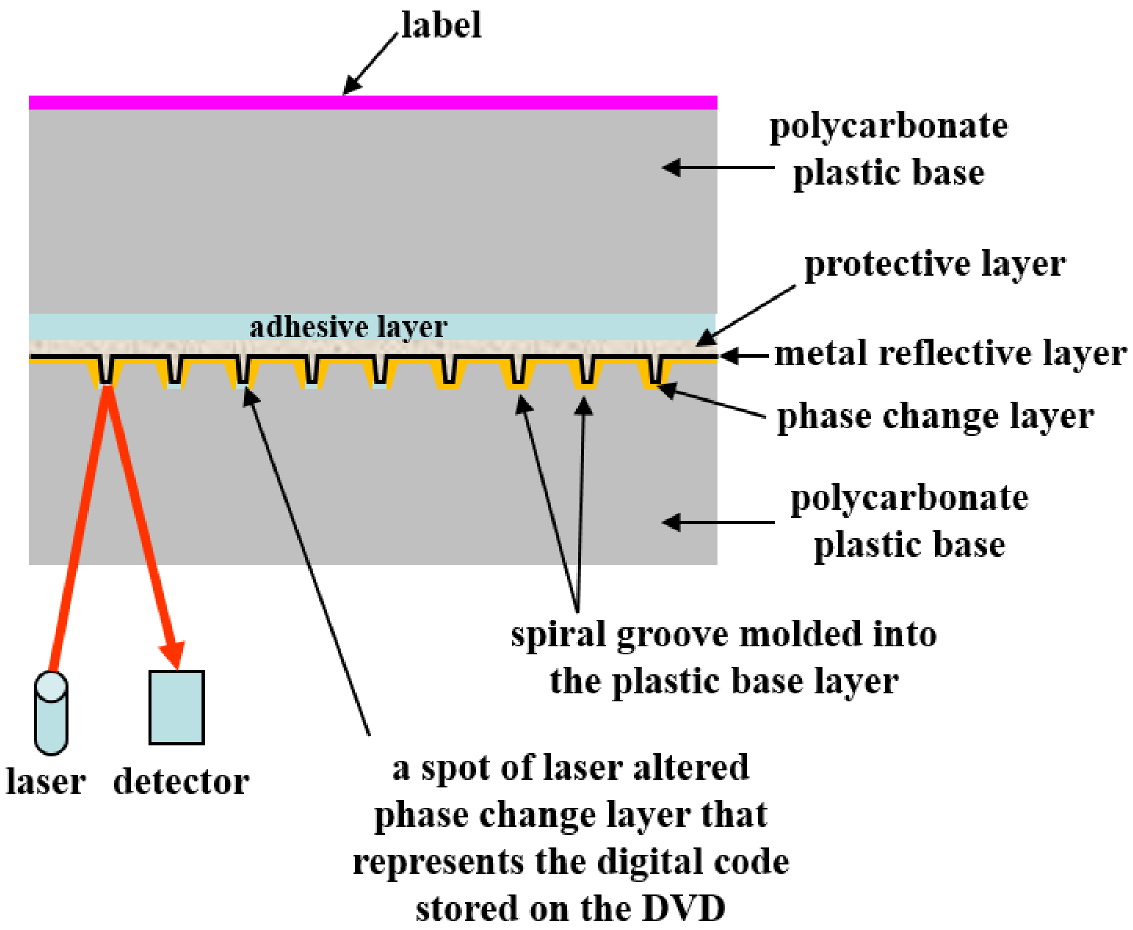


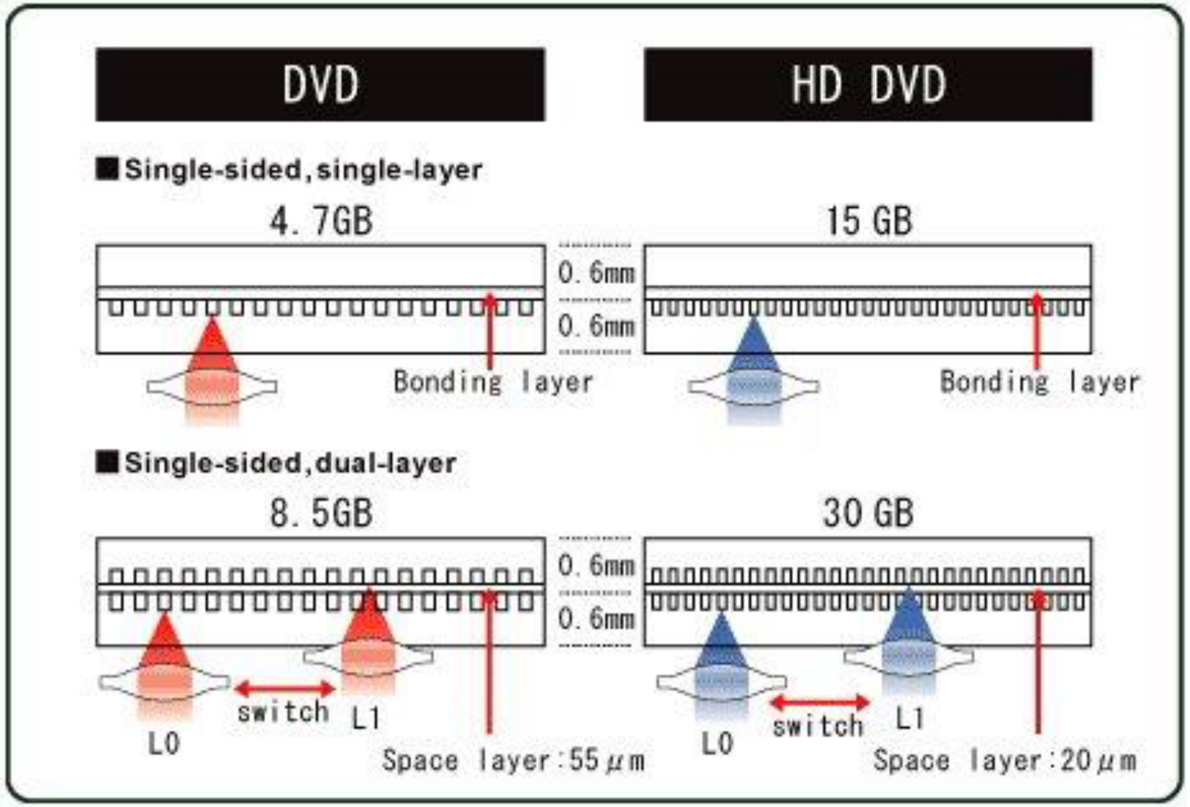




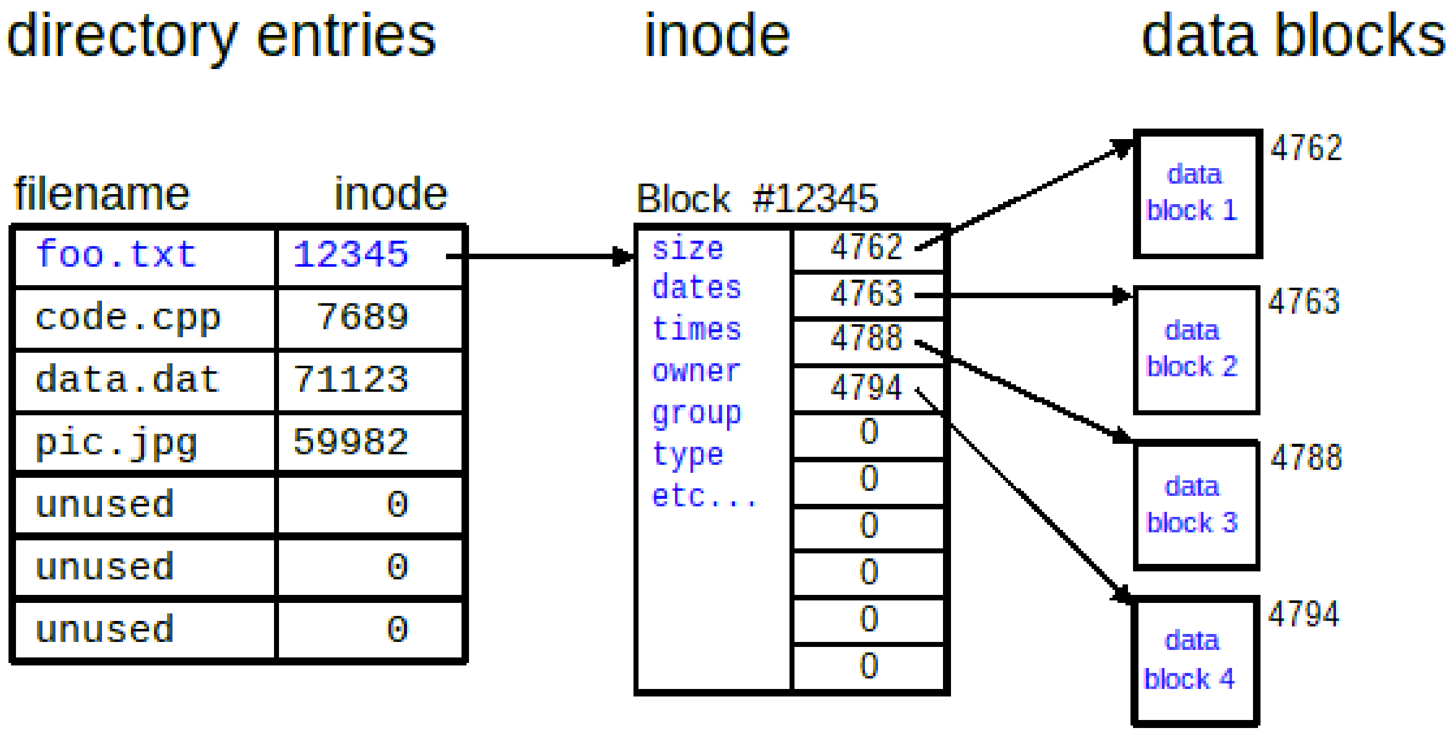


Information Storage and Retrieval


label
polycarbonate plastic base
protective layer
adhesive layer
metal reflective layer
phase change layer
polycarbonate plastic base
spiral groove molded into the plastic base layer
laser
detector
a spot of laser altered phase change layer that represents the digital code stored on the DVD



Information Storage and Retrieval



directory entries

filename	inode
foo.txt	12345
code.cpp	7689
data.dat	71123
pic.jpg	59982
unused	0
unused	0
unused	0

inode

Block #12345

size
dates
times
owner
group
type
etc...

4762
4763
4788
4794
0
0
0
0
0
0

data blocks

data block 1 4762
data block 2 4763
data block 3 4788
data block 4 4794

Files disadvantages

- Distinguished and Isolated Data
- Data Duplication / Data Redundancy
- Data Protection
- Issues with Transactions – ACID (Atomicity, Consistency, Isolation, and Durability)
- Concurrent issues

DR vs IR

Feature	Database System	Information Retrieval System
Data Type	Structured	Unstructured/Semi-structured
Query Language	SQL	Free-text/Natural Language
Result Matching	Exact	Relevance-based
Indexing	Structured (e.g., B-trees)	Inverted Index
Result Precision	Deterministic	Probabilistic
Examples	MySQL, PostgreSQL	Google, Elasticsearch

Why study IR?

- The most important problems in the domain of natural language processing (NLP)
- Hot topic research
- IR in LLM
- LLM in IR
- The roll of RecSys in e-commerce

Outlines

Theory	Practical
Text Preprocessing	Python
Boolean and vector-space retrieval models	Numpy
Evaluation and interface issues	Pandas
Document clustering and classification	MatplotLib
Traditional and machine learning-based ranking approaches	Nltk
Recommender System	Regex (Regular Expressions)
Web scrappy	NLTK (Natural Language Toolkit)
	TextBlob

Intro course

- Introduction to Information Retrieval (https://nlp.stanford.edu/IR-book/newslides.html)
- CS 276: Information Retrieval and Web Search (stanford.edu)
- Grading:
  - Assignments: 25
  - Midterm: 25
  - Presentation: 10
  - Final: 40