Information retrieval

Stanislav Protasov

2021

Course team

Stanislav Protasov, room 463
s.protasov@innopolis.ru

Anastasiia Puzankova, room 463
a.puzankova@innopolis.ru

Course news telegram channel

Agenda

1. How the course in taught and organized
    a. Lectures and labs
    b. Grading
    c. Exam
2. What is “information retrieval” (IR)
    a. Definitions
    b. Topic overview

How the course is taught and organized

Lectures, labs and hometasks

Course consists of 8 weeks, including 15 lectures and 15 labs

1. On Wednesday 9:10AM there is a 3-hours lecture by Stanislav
2. On Wednesday 2:30PM there is a *lab related to homework* by Anastasiia
3. On Thursday 2:30PM second *hands-on lab* by Anastasiia
4. **Hometasks** are extensions of labs with **deadline on Monday 6:00 PM**

Course materials are in **moodle, github**

Main **book** is "An Introduction to Information Retrieval" by Manning, Raghavan, Schütze; other materials will be published in Moodle or referred in github

Grading and exam

- **Hometasks** (7) will cost you up to **70** points in total (10 points each)
- **Contest labs** (7) can bring you up to 5 points each. Work in **teams up to 2**.
  - +2 points for each successful completion
  - +3 points for each submission from top 3 results

Grades distribution:

- **A = 84+**
- B = 70-83
- C = 60-69
- **Fail = 0-59**

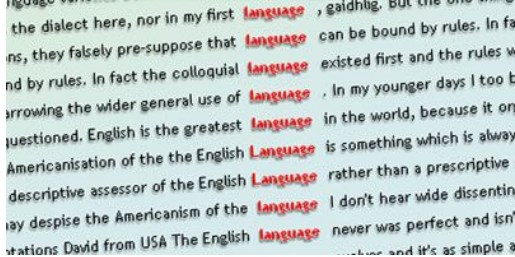
Information retrieval

Definition

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).
[The Book]

Scales of IR systems

- From **personal information retrieval**
  - Indexing vs `find -r /`
  - Classification and Filters
  - Background monitoring
- Via **enterprise and domain-specific search**
  - Specific domain information (law, chemistry, math)
  - Enterprise network (machine access)
- To **Web search**
  - Large scale
  - Commercial interest (SEO, exploits, advertisements)
  - Very heterogeneous data



Major research milestones (1)

Early days (late 1950s to 1960s): foundation of the field

- Luhn's work on automatic indexing (KWIC)
- Cleverdon's Cranfield evaluation methodology and index experiments
- Salton's early work on SMART system and experiments

1970s-1980s: a large number of retrieval models

- Vector space model
- Probabilistic models

Major research milestones (2)

1990s: further development of retrieval models and new tasks

- Language models
- TREC evaluation
- Web search

2000s-present: more applications, especially Web search and interactions with other fields

- Learning to rank
- Scalability (e.g., MapReduce)
- Real-time search

Highlights about today's IR

- Process **quickly** (no grep)
- **Flexible** match (consider language, typos, …)
- Ranked retrieval (closer to query, to intent, to user, ...)
  - ***Relevance*** *(relevant) - the user perceives as containing information of value with respect to their personal information need*

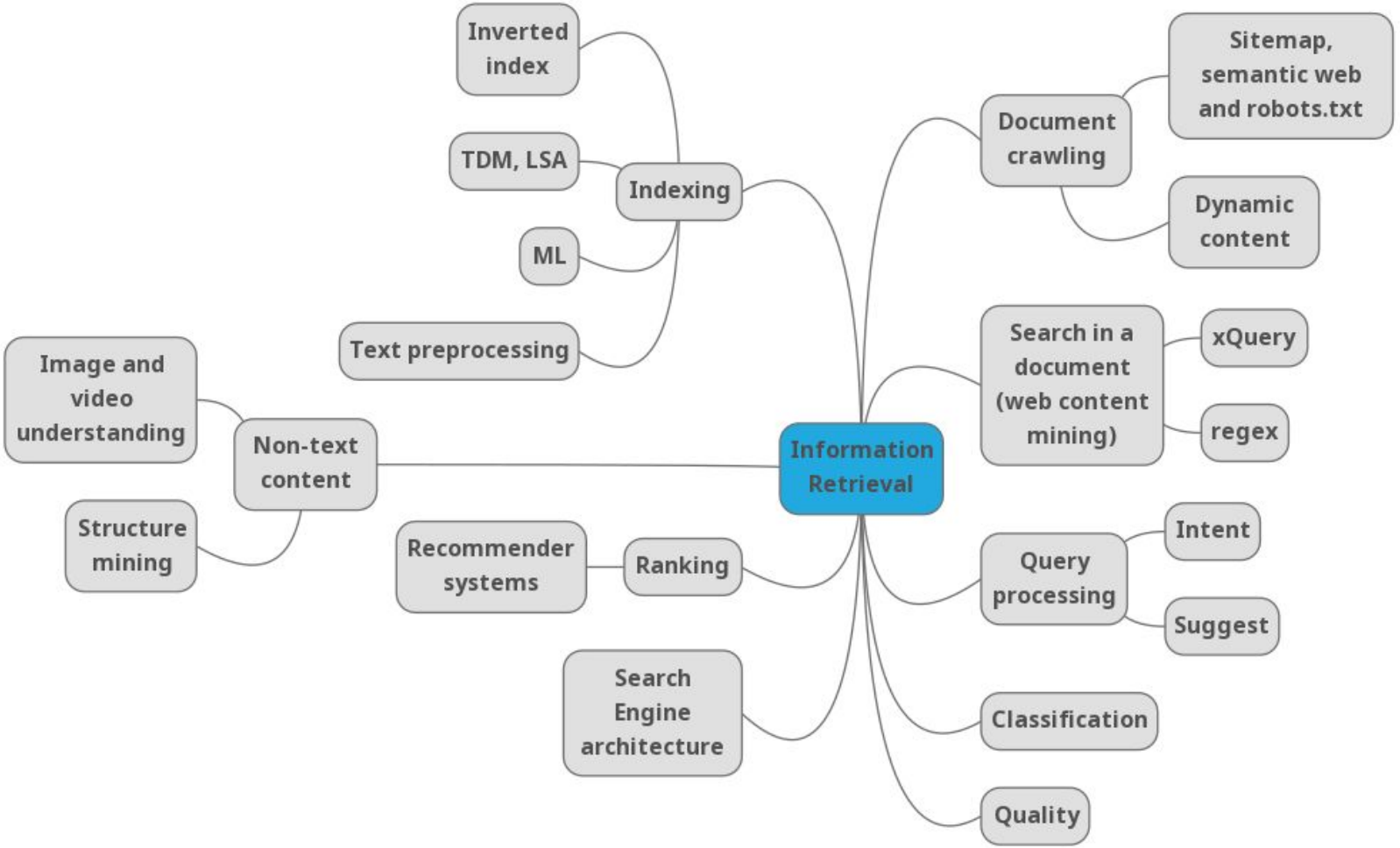
What does IR care about?

- **Query representation**
  - Lexical gap
  - Semantic gap: ranking model vs. retrieval method
- **Document representation**
  - Specific data structure for efficient access
  - Lexical gap and semantic gap
- **Retrieval model**
  - Algorithms that find the most relevant documents for the given information need

IR covers ...

- Search (obviously)
- Recommendations
- Question answering
- Text mining
- Online ads
- Audio, images, video understanding
- ...

Topic overview (by 2020)


Information Retrieval
Indexing
Inverted index
TDM, LSA
ML
Text preprocessing
Document crawling
Sitemap, semantic web and robots.txt
Dynamic content
Search in a document (web content mining)
xQuery
regex
Query processing
Intent
Suggest
Classification
Quality
Ranking
Recommender systems
Search Engine architecture
Non-text content
Image and video understanding
Structure mining

How search works

Watch this video: https://youtu.be/0eKVizvYSUQ

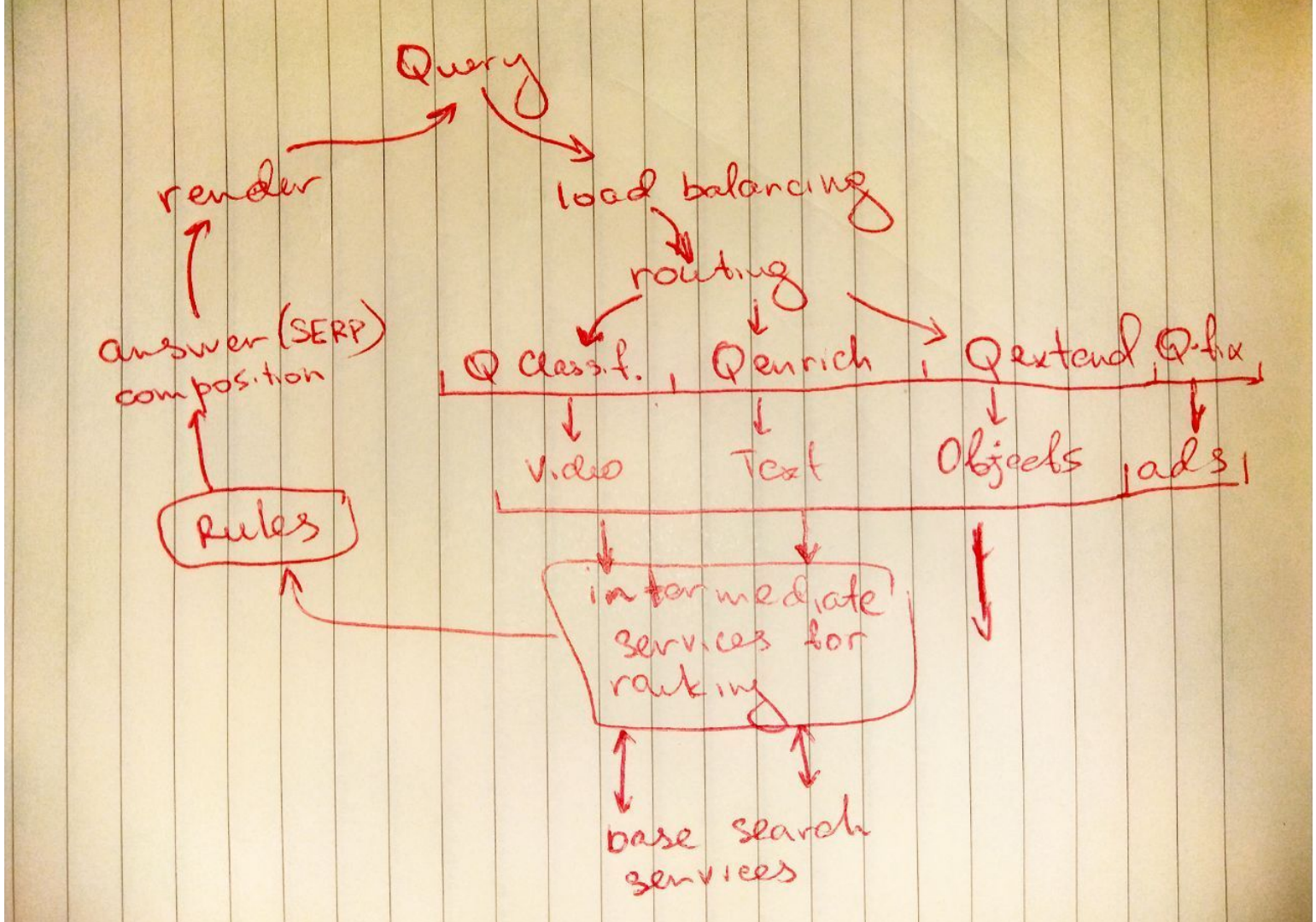
Answer the questions:

1. Did you understand how Google search works?
2. What is an **index**?
3. What is **scam** site?
4. Name or propose some **factors**
5. What is **side by side** and how is it used?

At home: read https://www.google.com/search/howsearchworks/

Whiteboard time!

Whiteboard


Query
render
load balancing
routing
answer (SERP) composition
Q class.f.
Q enrich
Q extend
Q-fix
video
Text
Objects
ads
Rules
intermediate services for ranking
base search services