Information retrieval

Stanislav Protasov

2022

Course team live in 463

Stanislav — s.protasov@innopolis.ru

Anastasiia — a.puzankova@innopolis.ru

Marina — m.lisnichenko@innopolis.university

Patrik — p.kenfack@innopolis.university

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

Major statements

Course consists of 15 weeks including **15 lectures and 15 labs.**

Course ends **in the end of April.**

No exam.

Course materials are in **moodle, github** and telegram.

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** (4) will cost you up to **60** points in total (15 points each)
- **Quizzes** (4): 4 short quizzes, up to **40** in total (10 point each)
- **Contests** (3-4) can bring you up to 5 additional points each.
  - +2 points for each successful completion of the task
    OR
  - +5 points for each of top-10 solutions.

Grades distribution:

- **A = 84+**
- B = 72-83 (rounded to integer)
- C = 60-71
- **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]

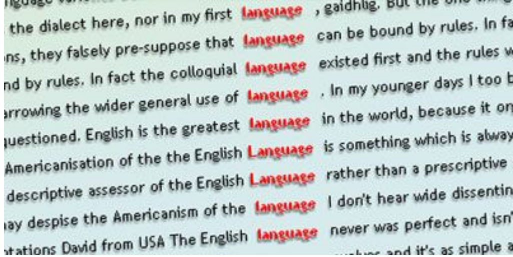
Let’s speculate on the definition

1. Where are borders among **algorithms, IR, and DB**?
    a. How these disciplines answer the question “**How old is John Doe**”?
    b. What is the difference in terms of software?
2. Is IR a static area?
3. Name some IR systems

Scales of IR systems

- From **personal information retrieval**
  - Indexing vs `find -r /`
  - Classification (e.g. photo collection) 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)



```
 the dialect here, nor in my first  language  , gaidhlig.
ns, they falsely pre-suppose that  language  can be bound by rules. In fa
nd by rules. In fact the colloquial  language  existed first and the rules w
arrowing the wider general use of  language  . In my younger days I too b
uestioned. English is the greatest  language  in the world, because it on
Americanisation of the the English  Language  is something which is alway
descriptive assessor of the English  Language  rather than a prescriptive
ay despise the Americanism of the  language  I don't hear wide dissentin
tations David from USA The English  language  never was perfect and isn'
```

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*

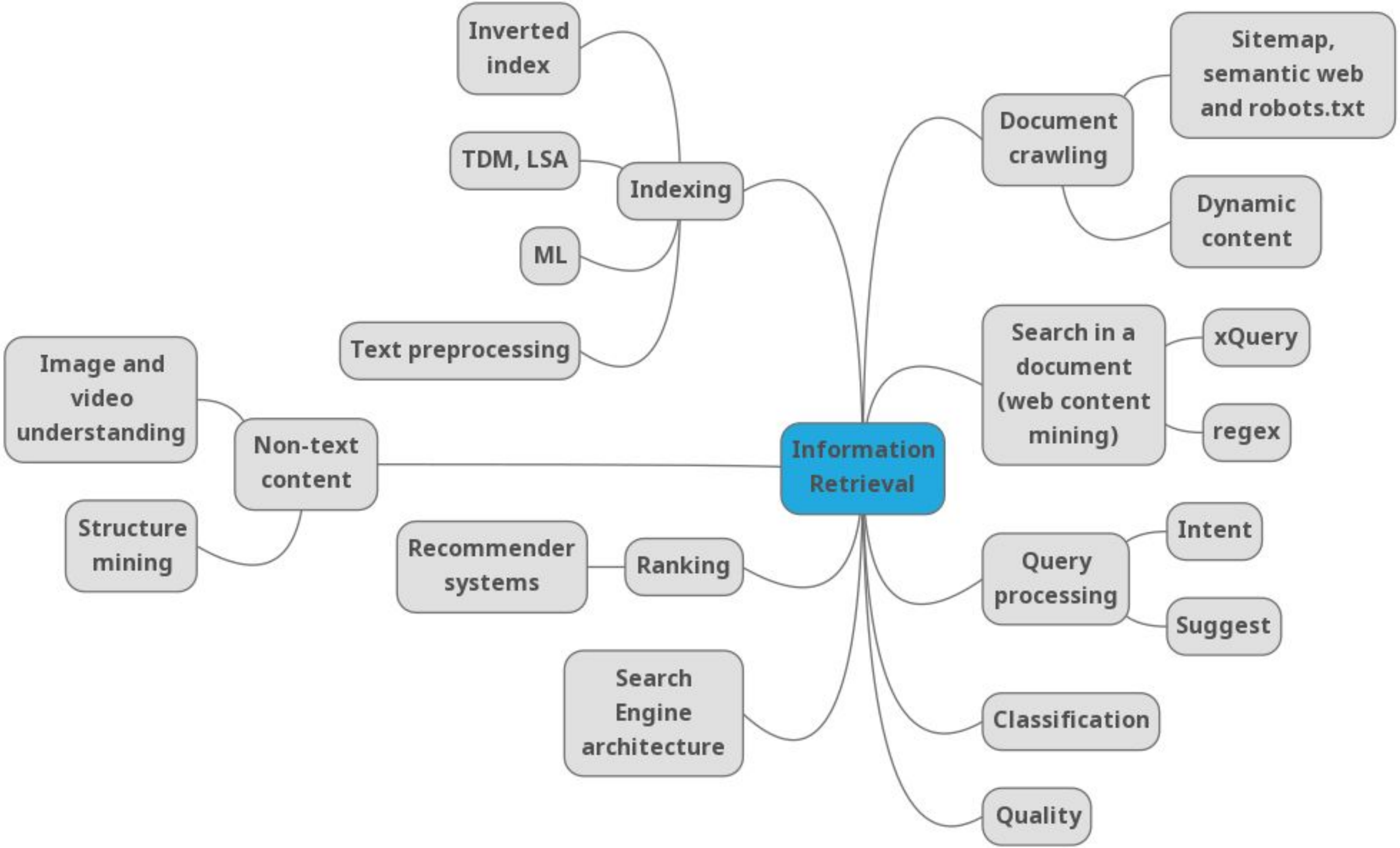
What does IR care about?

- **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
- **Speed and space**
- **…**

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

Non-text content

Image and video understanding

Structure mining

Ranking

Recommender systems

Search Engine architecture

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

How search works

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

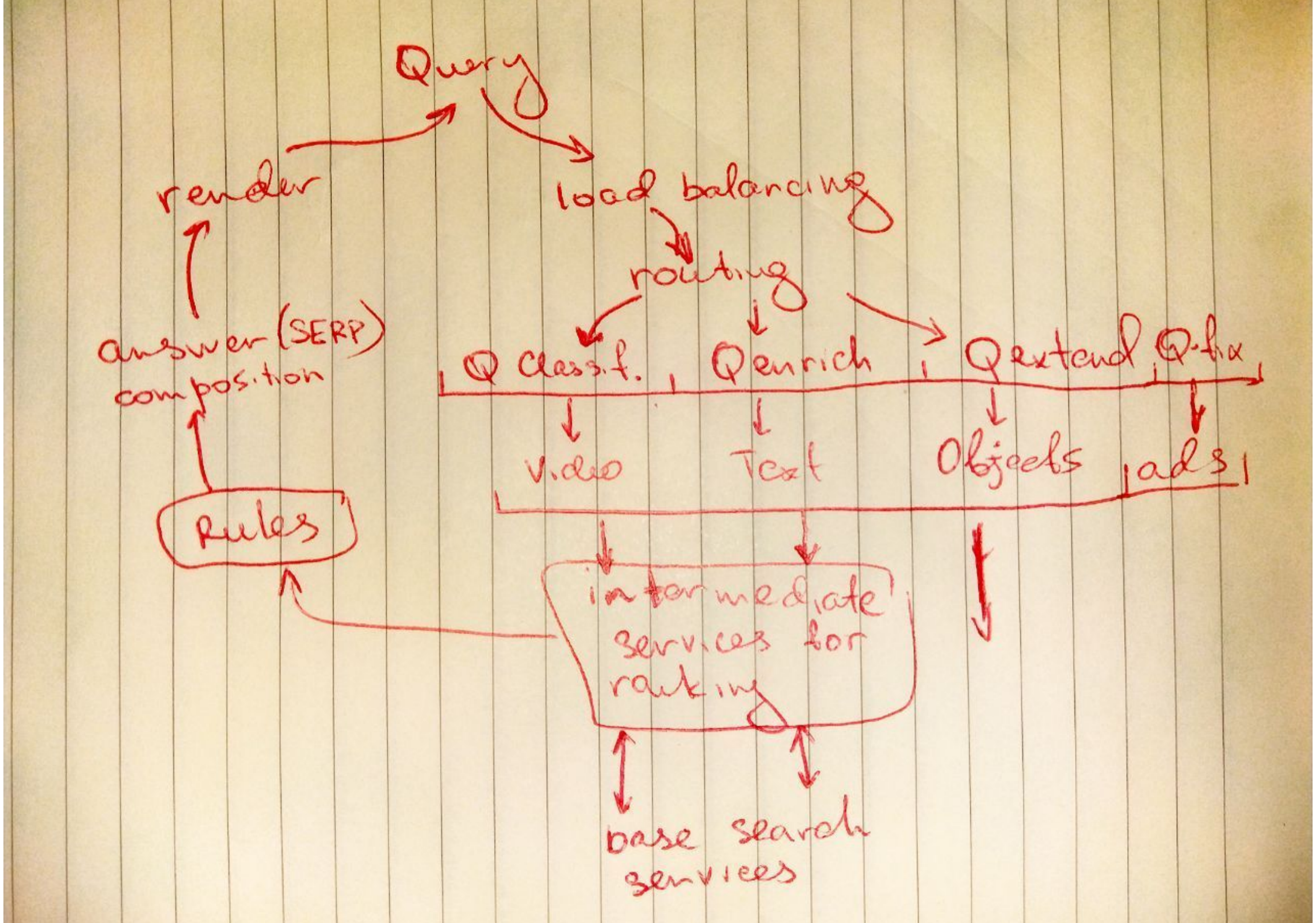
Answer the questions:

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