Information Retrieval

Prof. Alberto Sillitti

1. Introduction

About the instructor

- PhD in Computer Engineering

- Former full professor with research focus on Software Engineering
  - Software quality
  - Data analysis/ML/AI
  - Agile software development
  - Open source

- Co-funder, CEO, and Chief Scientist
  - Consulting companies for improving the quality of their software
  - Consulting companies in AI and ML

CENTRE FOR APPLIED SOFTWARE ENGINEERING

CASE

Innopolis faculty teaching this course

- Lectures: Alberto Sillitti
- Labs:
  - Kamil Sabbagh
  - Mahmoud Mousatat
  - Kelvin Asu Ekuri
  - Ahmad Taha

Office hours on demand

Grading criteria

- Assignments: 30%
- Midterm: 35%
- Final: 35%
- Participation: extra 5%

Letter - grade

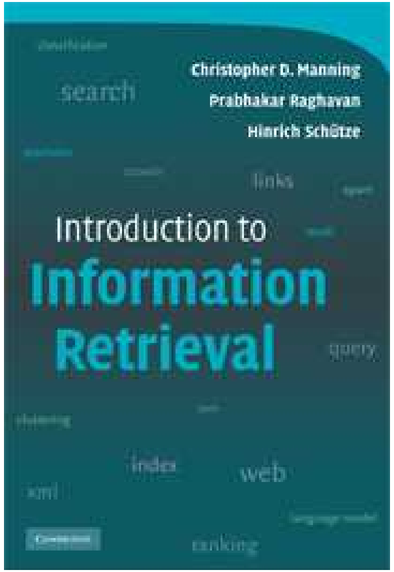
- A from 90%
- B from 75%
- C from 60%

Recommended literature

The main **book** is “An Introduction to Information Retrieval” by Manning, Raghavan, Schütze (2009 edition) (https://www-nlp.stanford.edu/IR-book/)

Slides (derived from the ones of the previous years by Stanislav Protasov and Leonard Johard)

Other materials will be published in Moodle.



Topics

- Introduction to IR
- Basics
  - Web crawling
  - Quality assessment
- Text processing
  - Indexes
  - Text management
  - Search
  - Language modelling
- Vector modelling
- Media processing

What Is IR?

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).

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

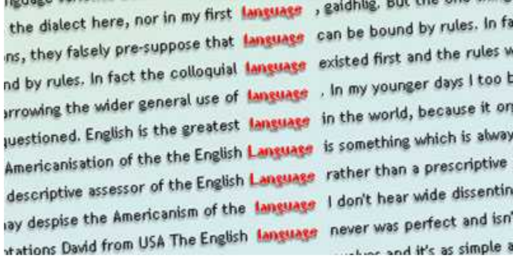
Scales of IR systems (1/2)

- 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

Scales of IR systems (2/2)

- Till **AI**
  - LLM (Large Language Models)
  - SLM (Small Language Models)
  - Fine-tuning
  - RAG (Retrieval Augmented Generation)
  - Agents

Major research milestones (1/2)



```
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
questioned. 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 (2/2)

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*

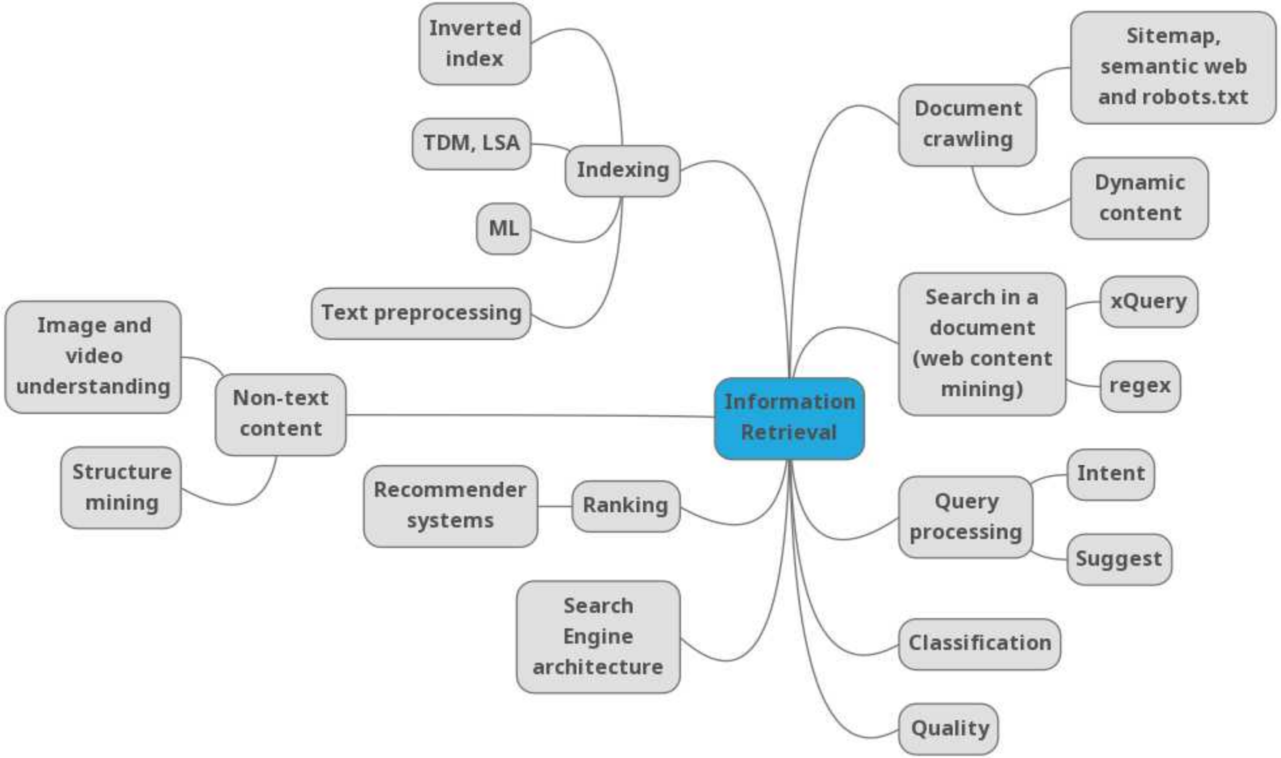
What does IR care about?

- **Query representation**
  - Lexical gap: no such word
  - Semantic gap: ranking model (system assumes),
    retrieval method (system encodes), human language
- **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 ...

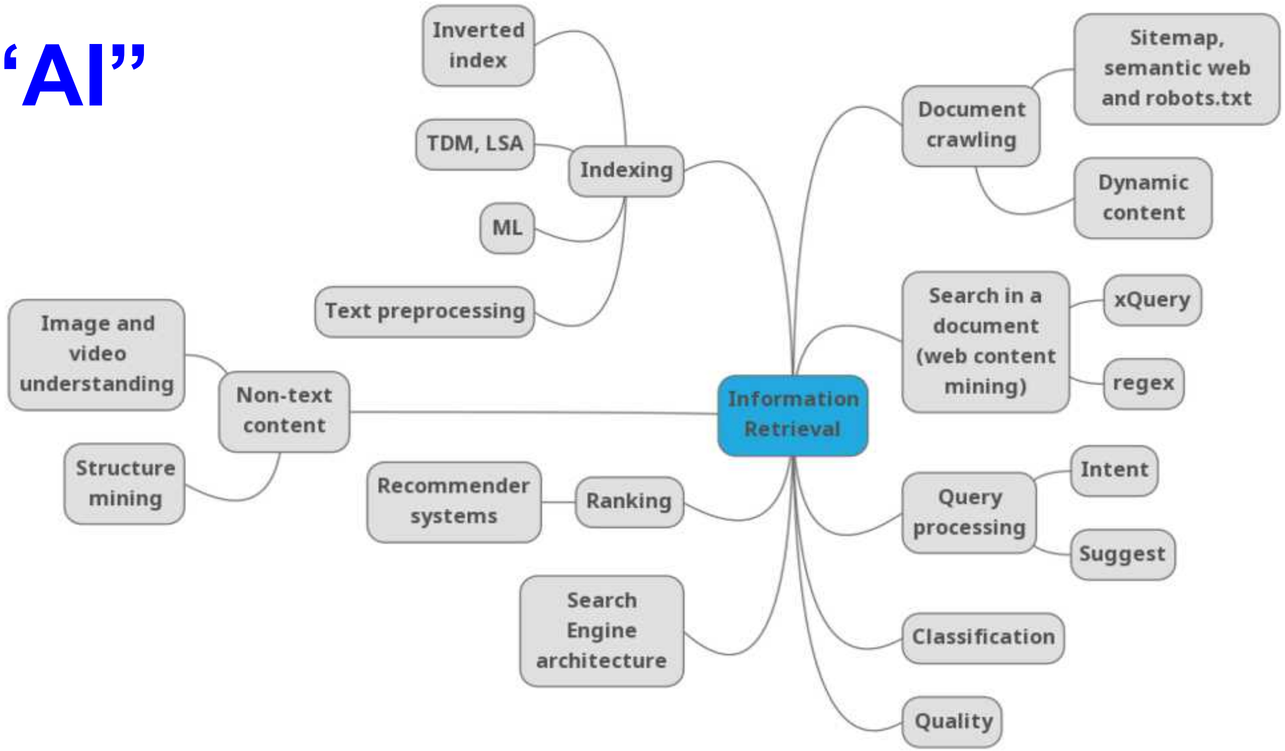
- Search (obviously)
- Recommendations
- Question answering
- Text mining
- Online ads
- Audio, images, video understanding
- ...

Topic overview (by 2020)



Topic overview (now)

“AI”



How search works: introduction (1/2)

- Watch this video
  - https://www.youtube.com/watch?v=0eKVizvYSUQ
- Answer the questions:
  - Did you understand how Google search works?
  - What is an **index**?
  - What is **scam** site?
  - Name or propose some **factors**
  - What is **side by side** and how is it used?

How search works: introduction (2/2)

- At home, read:
  - https://www.google.com/search/howsearchworks/
  - https://searchengineland.com/google-search-document-leak-ranking-442617
  - https://searchengineland.com/yandex-search-ranking-factors-leak-392323