Information Retrieval (5LN712)

Introduction and Outline

Ali Basirat

Department of Linguistics and Philology
Uppsala University

March 30, 2020

Table of contents

Information Retrieval (IR)

Definition

Information retrieval is finding material of an unstructured nature that satisfies an information need from within large collections [1].

- Material is usually documents
- Unstructured data refers to data which does not have, clear, semantically overt, easy-for-a-computer structure, like text

Major Use Cases

- Semi-structured searches such as finding documents with certain properties
- Browsing and filtering document collections

Example

Which play of Shakespeare contain the words `Brutus` AND `Caesar` AND NOT `Calpurnia`

Searching Approaches

- Grepping: linear search through documents one-by-one and marking those that matches our query, e.g., find the documents that contain `Brutus` and `Caesar`, and exclude those that have `Calpurnia`
- Indexing: index the document in advance, e.g., each term is represented by a binary vector whose elements corresponding to the documents indicate the presence of the term in a document. Depending on the query, a small number of Boolean (logical) operations are performed on the vectors to find the documents that satisfy our needs.

Searching Approaches

Example (Indexing)

Which play of Shakespeare contain the words `Brutus` AND `Caesar` AND NOT `Calpurnia`

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

`Brutus` AND `Caesar` AND NOT `Calpurnia`=100100

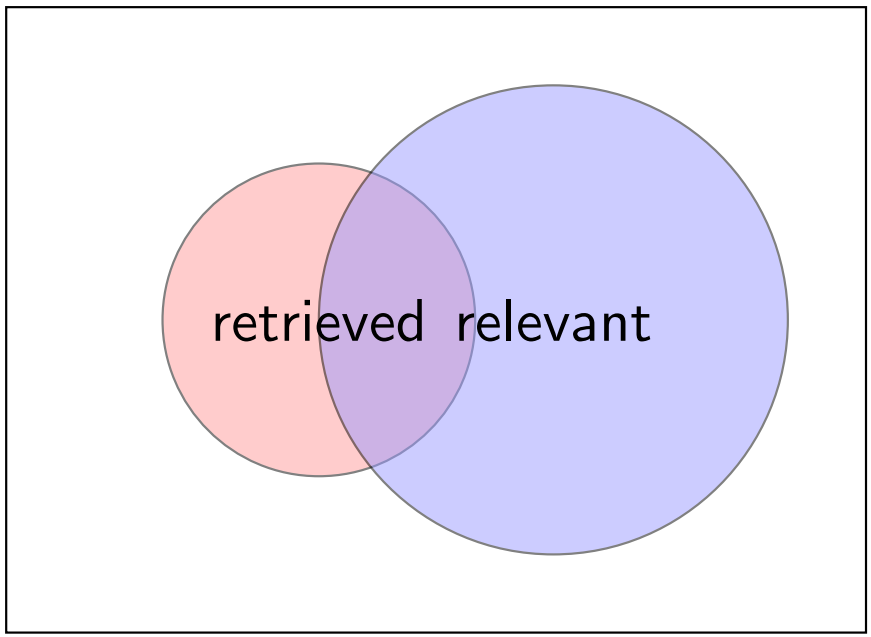
Searching Approaches

Advantages of Indexing:

- Process documents quickly
- Allow more flexible matching
- Allow ranked retrieval

Evaluation

- Precision: what fraction of the retrieved documents are relevant to the information need?
- Recall: what fraction of the relevant documents in the collection are retrieved by the system.



IR, a bird-eye view

- A collection of documents are indexed based on their terms
- A query made by an end user is translated into a Boolean expression
- The expression is evaluated on the indexed documents
- A set of most relevant documents is retrieved
- We want to know how good are the retrieved documents

Table of contents

Intended Learning Outcomes

At the end of the course, you should be able to:

- explain in detail the most common techniques of text indexing, text classification, and information extraction
- explain various types of information retrieval models
- evaluate an information retrieval system
- analyze and critically review scientific publications in the field of information retrieval
- apply basic tools for indexing and information retrieval
- implement some of the basic tools of information retrieval
- formulate and critically discuss the methodological assumptions made by the approaches mentioned in the course
- present results in a professional way

Teaching

- 9 lectures
- 3 lab sessions
- Supervision available on demand:
  - email
  - IR Forum in the student portal
  - knock on the office door - canceled due to the corona crisis
  - book a meeting - depends on the situation
- Do not expect the slides to be self contained.

Content

- Boolean retrieval
- Scoring and term weighting
- Evaluation of information retrieval systems
- Probabilistic information retrieval
- Language models for information retrieval
- Relevance feedback and query expansion
- Text classification and näive Bayes
- Vector space model and vector space classification
- Matrix decomposition and latent semantic indexing
- Information extraction

Examination

- Exercises and assignments - only for VG score
- 3 lab reports with G and VG scores
- Seminar presentation on selected chapters. You will be divided into groups of two or three. Each group should have a 20-minute-presentation on some selected topics. (Depending on how our online communication tools function, we may or may not cancel the seminars)
- Literature review - to critically review and analyze two selected scientific publication. The reviews are prepared by groups of two or three members.
- Individual/group projects - to *write a proposal* and *report results* in a scientific way. You are allowed to work on your own topic as far as it is confirmed by the course instructor. A few project topics are also available on the course web page.

Examination

- to pass (G): three Gs for the lap reports, a seminar presentation, a literature review, a project proposal, and a project report
- to pass with distinction (VG): in addition to the requirements for G, you should do at least half of the exercises and assignments, two VGs for the labs, and prepare high quality reports for the literature review and the project.

References I

- Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich.
  Introduction to Information Retrieval.
  Cambridge University Press, 2008.
- Dan Jurafsky and James H. Martin.
  Speech and Language Processing.
  3rd edition, 2019.
- Online freely available contents in Wikipedia and Mathworld Wolfram

Boolean Retrieval

- Boolean Queries
- Indexing mechanism
- Efficient indexing through inverted indices
- How to process arbitrary Boolean queries

Scoring, Term Weighting, and the Vector Space Model

- Boolean retrieval returns a large set of relevant documents
- We want to know which documents are more relevant
- Ranked retrieval augments Boolean retrieval with relevance rankings
- Documents as vectors

Evaluation in Information Retrieval

- How to evaluate a set of (un?)ranked retrieved documents
- Precision/recall/F-score

Probabilistic and Language Models for IR

- How to rank documents based on their relevance probability
- How to estimate the relevance probabilities
- The formulation of document relevance based on language models

Document Classification

- Bayesian approach
- Documents as vectors
- Vector space classification
- Latent semantic indexing

Information Extraction

- To extract structured information from textual data
- Named entity recognition
- Relation extraction

Registration

The course registration is done online.