Information Retrieval (5LN712)
Introduction and Outline

Ali Basirat

Department of Computer and Information Science (IDA)
Linköping University
ali.basirat@liu.se

Department of Linguistics and Philology
Uppsala University
ali.basirat@lingfil.uu.se

March 29, 2021
# Table of contents

1. **Introduction to IR**

2. **Information Extraction**

3. **Course Information**

4. **Overview**

5. **Registration**
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.
**Example (Indexing)**

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

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Brutus</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>worser</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Brutus AND Caesar AND NOT Calpurnia = 100100
Advantages of Indexing:

- Process documents quickly
- Allow more flexible matching
- Allow ranked retrieval
- **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’s-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

1. Introduction to IR
2. Information Extraction
3. Course Information
4. Overview
5. Registration
Information Extraction (IE) I

Definition
Information extraction is the process of turning unstructured information embedded in a text into structured data.²

Use cases
• Template filling: to extract tabular data from text
• Knowledge base population: to extract facts from text
• Language Analysis: to study the semantics of words and sentences
Example

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

- SpokesmanFor(Time Wagner, American Airlines)
- UnitOf(United, UAL Corp.)
- UnitOf(American, AMR)
Two important tasks:

- Named entity recognition (NER)
- Relation extraction (RE)
Named Entity Recognition (NER)

Example

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].
Relation extraction: to find and classify relations hold between named entities

$$\text{RE} : E \times E \ldots E \rightarrow R$$

E is an entity space, and R is a relation space

Example

- (Time Wagner, American Airlines) → SpokesmanFor
- (United, UAL Corp.) → UnitOf
- (American, AMR Corp.) → UnitOf
• Information retrieval finds relevant documents for an information need
• Information extraction converts the unstructured information in relevant documents to a structured format
• The structured information can be processed in a more systematic way for various tasks

---

2 Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd ed. draft)
1 Introduction to IR

2 Information Extraction

3 Course Information

4 Overview

5 Registration
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
• explain various types of information extraction models
• explain the common techniques of document representation and document classification
• evaluate an information retrieval and an information extraction system
• analyze and critically review scientific publications in the field of information retrieval/extraction
• 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
• 11 lectures
• 3 lab sessions
• Supervision available on demand:
  • email
  • Online meetings
• Do not expect the slides to be self contained (i.e., read the books and other related references).
Teaching assistants

- Harm Lameris: harm.lameris.2403 at student.uu.se
- Xingran Zhu: xingran.zhu.2781 at student.uu.se
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 naïve Bayes
• Vector space model and vector space classification
• Matrix decomposition and latent semantic indexing
• Neural information retrieval
• Information extraction
References

- Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd ed. draft) - Ch. 18, 25
- Bhaskar Mitra and Nick Craswell (2018), An Introduction to Neural Information Retrieval
Pages

- Course homepage: https://cl.lingfil.uu.se/~abasirat/ir/ir
- The student portal (studentportalen)
Group divisions

- Depending on the number of students, we may or may not have group divisions
- Each group includes only two students
- You will have both individual and group assignments
- The groups (or individuals) will be distributed into two major groups supervised by the two assistants
Examination items

- Exercises and assignments - only for VG score (individual)
- Lab: 3 lab reports with G and VG scores (group)
- Seminar: 20-minute-presentation on some selected topics (group)
- Literature review: to critically review and analyze two selected scientific publication (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. (group)
Grading

- to pass (G): three Gs for the lab 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, get two VGs for the labs, and prepare high quality reports for the literature review and the project.
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
Neural Information Retrieval

- Term and document embeddings
- Standard neural network architectures for IR
Information Extraction

- To extract structured information from textual data
- Named entity recognition
- Relation extraction
The course registration is done online through: https://forms.gle/1WNVaLigd9eV7Ts7A