INFORMATION RETRIEVAL

L01. INTRODUCTION

SUZAN VERBERNE 2022

Universiteit Leiden

COURSE INFORMATION

- Brightspace page: https://brightspace.universiteitleiden.nl/d2l/home/91761
- Lectures:
  - Tuesday, 9.15-11.00
  - Sitterzaal, Huygens/Oort & online on Mediasite: https://weblectures.leidenuniv.nl/Mediasite/Channel/informationretrieval

COURSE INFORMATION

- Prof. dr. ir. Wessel Kraaij
  - https://www.universiteitleiden.nl/en/staffmembers/wessel-kraaij#tab-1
- Dr. Suzan Verberne
  - http://liacs.leidenuniv.nl/~verbernes/
- Teaching assistants:
  - Arian Askari (PhD student)
  - Juan Bascur Cifuentes (LIACS TA)
  - Yi Lin (LIACS TA)
- Contact: ircourse@liacs.leidenuniv.nl

WHO ARE YOU?

Quick round (raise hands): what is your master programme?

- Computer Science
- Artificial Intelligence
- Data Science
- Bioinformatics
- Media Technology
- ICT in Business and the Public Sector
- Other

WHO ARE YOU?

Quick round (raise hands)

- Who has taken Introduction to Deep Learning?
- Who has taken Text Mining?
- Who can program in Python?
- Who is familiar with the Big-O notation?
- Who knows what the vector space model is?
- Who is familiar with probabilistic models?

TODAY'S LECTURE

- Course goals
- Introduction to information retrieval
  - Including a small exercise
- Structure of this course
- Boolean retrieval

COURSE GOALS

COURSE GOALS

- https://studiegids.universiteitleiden.nl/courses/105168/information-retrieval
- You will learn about:
  - fundamentals of models
  - data, experimentation, evaluation
  - challenges and limitations
  - practical applications

COURSE LITERATURE

- We use the following reading materials:
  - Chapters from the textbook:
    Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, **Introduction to Information Retrieval**, 2008
    https://nlp.stanford.edu/IR-book/
  - Bhaskar Mitra and Nick Craswell,
    An Introduction to Neural Information Retrieval, 2018
    https://www.microsoft.com/en-us/research/uploads/prod/2017/06/fntir2018-neuralir-mitra.pdf
  - Additional readings

- All literature will be distributed on Brightspace, as are the slides

INTRODUCTION TO IR

Universiteit Leiden

Suzan Verberne 2022

WHAT IS INFORMATION RETRIEVAL?

- The book (Manning et al.):
  - *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).*

- There is a collection of unstructured documents
- A user is seeking information
  - searching for relevant documents
  - creating queries

WHAT IS INFORMATION RETRIEVAL?

- IR systems support the search process by
  - Analyzing queries and documents
  - Retrieving documents by computing a relevance score for each document given a query
  - Ranking the documents by that relevance score

- Evaluation based on:
  - Effectiveness: quality of the result
  - Efficiency: speed of finding the result

CHALLENGES

- The real user need is hidden
- User queries are short and ambiguous
- Natural language text is
  - Unstructured
  - Noisy
  - Redundant
  - Infinite
  - Sparse
  - Multilingual

QUERIES ARE SHORT AND AMBIGUOUS

leiden

- What does the query refer to?
  - The city?
  - The university?
  - The Dutch verb ‘leiden’?
- What should be the mode of the answer?
  - Photos?
  - A map?
  - A home page (www.leiden.nl)?
  - A Wikipedia article?
- What type of information is the user interested in?
  - Touristic information?
  - Historic information?
  - News events?
- With only the keyword query, the IR system cannot know

EXERCISE

albert einstein leiden

- How can a search engine determine the relevance of a document given a query?
- List a few factors of relevance that you would use if you were implementing a web search engine
- Discuss for a few minutes with your neighbour

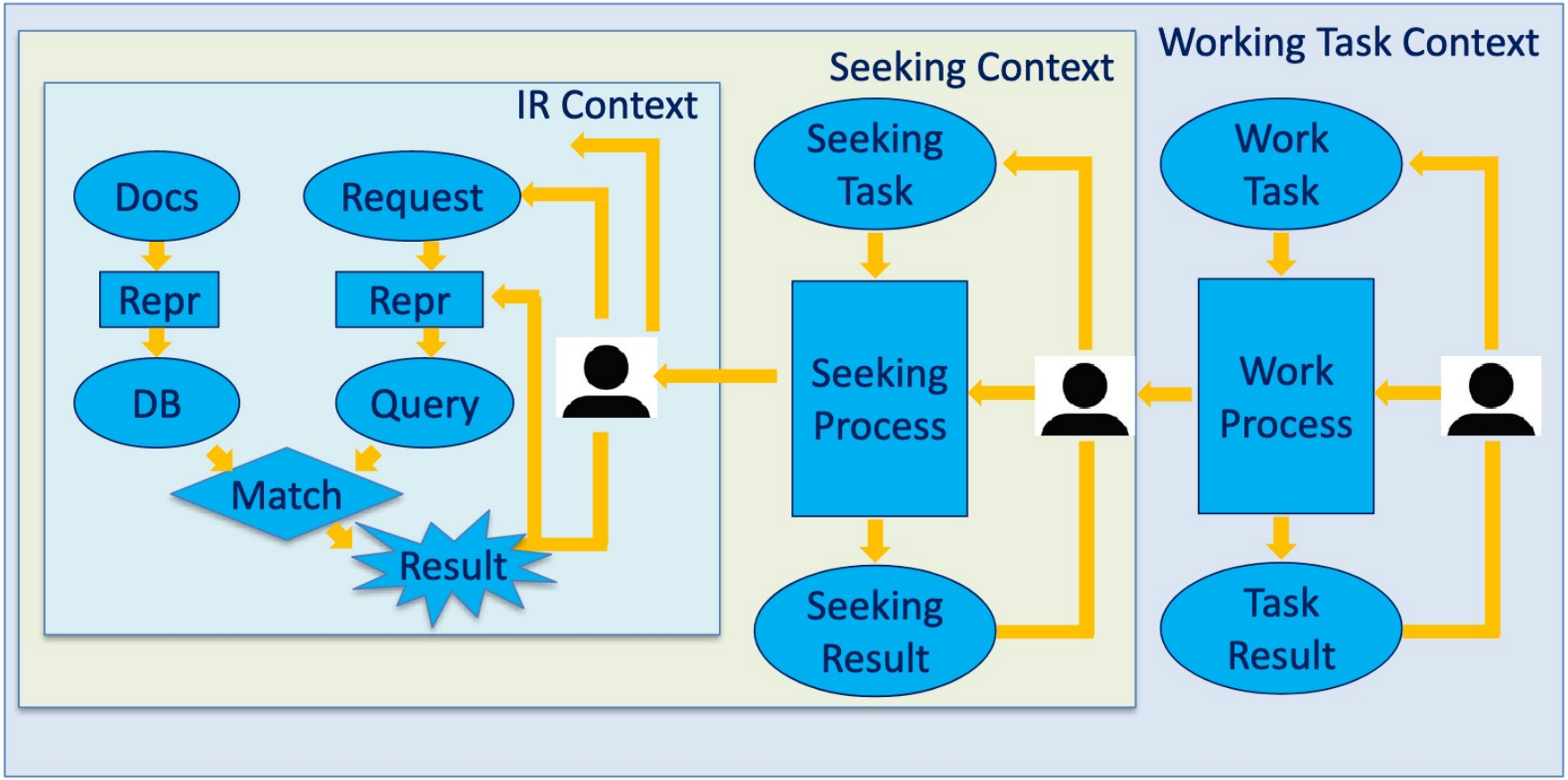
RELEVANCE FACTORS

- Term overlap: return documents that contain the query terms
  - Or related terms? Should ‘bicycle’ retrieve documents containing the term ‘bike’?
- Document importance: incoming hyperlinks
- Result popularity: clicks by previous users on the document
- The user history? (previous queries, previous activity)

FOUNDATIONS OF IR

- Database theory (scalability/storage)
- Natural Language Processing (term weighting, text representations)
- Machine Learning and Statistics (Learning to rank)
- Network analysis (Pagerank)
- Mathematics (Probability calculus, Logic)
- Library science (user information needs)
- Behavioural science (social media, user modeling)

IR IN CONTEXT



Recall, precision, efficiency,

Usability

Quality of end result

From Jaana Kekalainen

IR SETTINGS AND APPLICATIONS

- Web search
- Multimedia search
- Expert search
- Recommender systems
- Contextual search
- Exploratory search
- Conversational search
- Domain-specific search
- Personal collection search (think about email search)

COURSE STRUCTURE

COURSE OUTLINE

	Date	Teacher	Topic
1	8 Feb	WK + SV	Introduction and Boolean retrieval
2	15 Feb	SV	Evaluation and test collections
3	22 Feb	WK	Indexing and compression
	1 Mar		(No lecture)
4	8 Mar	WK	Vector Space Model
5	15 Mar	SV	Neural IR
6	22 Mar	WK	Probabilistic IR
7	29 Mar	WK	Language Modeling for IR
8	5 Apr	WK + SV	(Student presentations)
	12 Apr		(No lecture)
9	19 Apr	SV	Learning to rank
10	26 Apr	SV	Web search
11	3 May	SV	Query and session analysis
12	10 May	Cor Veenman	Guest lecture on Responsible IR
13	17 May	SV	Conversational search and domain specific IR

GENERAL STRUCTURE

- 13 lectures
- Homework:
  - Literature after the lecture
  - Homework exercise to be completed (individual)
- Two group projects:
  - A critical review about a recent IR paper (written and presentation)
  - Practical assignment (experimentation and report)
- Written individual exam

GRADING

- The course grade consists of the following components:
  - Weekly exercises (homework) – **10% (H)**
    At the end of the semester the homework grade is computed as: $\frac{n_{completed}}{n_{total}} * 10$
  - Critical review – **10% (R)**
    Written review and presentation (in teams)
  - Practical assignment – **20% (A)**
    Experimentation and report (in teams)
  - Final written exam (closed book, individual) – **60% (E)**
    The grade for the exam needs to be at least 5.5 to pass the course.

GRADING

- Weekly homework:
  - Submit through Brightspace (Assignments)
  - Since the purpose is exercising, you are allowed to make mistakes (you get your point for completing it)
  - The answers will be added to Brightspace after the deadline
- Passing the course:
  - The grade for the written exam should be 5.5 or higher in order to complete the course.
  - If the review or critical assignment is not submitted the grade for that task is 0.

BOOLEAN SEARCH

(BY WESSEL)

SUZAN VERBERNE 2022

Universiteit Leiden