Information Retrieval 1

Summary

Ilya Markov
i.markov@uva.nl

University of Amsterdam

IR1 2021

1. One of the largest IR1 courses so far
2. Fully online

Outline

1. IR1 in numbers
2. Course organization
3. Intermezzo: Information Retrieval
4. Course content

Outline

1. IR1 in numbers
2. Course organization
3. Intermezzo: Information Retrieval
4. Course content

Participants

- 205 registered intially
- 190 registered now
- 175 submitted assignment 1
- 172 submitted assignment 2
- 30–40 attended Q&A and flipped classroom sessions

Team

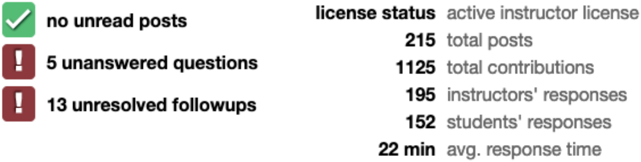
- 3 lecturers (+ Harrie)
- 1 senior TA
- 2 TAs working on assignments
- 8 TAs communicating with you

Hours per week

- 40 – coordination, lectures, Q&A, flipped classroom
- 30 – senior TAing
- 10 – TAing

170 hours per week spent by the teaching team

Piazza



- ✓ **no unread posts**
- ! **5 unanswered questions**
- ! **13 unresolved followups**

license status	active instructor license
215	total posts
1125	total contributions
195	instructors' responses
152	students' responses
22 min	avg. response time

Outline

1. IR1 in numbers
2. Course organization
3. Intermezzo: Information Retrieval
4. Course content

Course as collaboration

A course is a collaboration
between teachers and students

Your contribution

- Followed course guidelines and instructions
- Attended Q&A and flipped classroom sessions
- Submitted assignments on time
- Spotted mistakes in lectures, Q&As and assignments
- Provided feedback
- Responded to questions on Piazza
- Reported issues

Thank you!

Our contribution

- Prepared the course (slides, videos, assignments, flipped classrooms, timetable, Piazza, exam, etc., etc., etc.)
- Worked with you during LCs and Q&As
- Responded to questions on Piazza
- Responded to your feedback and requests
    - Some implemented (YouTube, flipped classroom, performance lower bounds, etc.)
    - Some will be considered for IR1 2022 (benchmarking, difficulties in assignment 2, LCs vs. Piazza, etc.)
    - Some could not be implemented (change of timetable, extension of deadlines, etc.)
- Experimented (flipped classrooms, +0.5 to the grade for answers on Piazza, etc.)
- Graded assignments and exam (some grading still TBD)
- Had weekly team meetings to make sure the course ran smoothly

We all made mistakes

- It is fine to make mistakes
- When we made mistakes, we all did our best to acknowledge them and to apologize

But in the end

- I am grateful to you
- I am proud of the team

We all made IR1 2021 a success!

Please give us feedback using online form

1. What did you like?
2. What can be improved (actionable items)?
   - I would like to have an additional lecture on . . . , because . . .
   - 15 mins meetings with my TA was too short for me, so I would like to have 30 mins instead
   - I would like to have weekly assignments instead of one every three weeks, because . . .
   - Etc.

Become a TA for IR1 2022

- Help shaping the course
- Meet students and help them with assignments
- Reply on Piazza
- Grade assignments
- Grade 1–2 exam questions

Outline

1. IR1 in numbers
2. Course organization
3. Intermezzo: Information Retrieval
4. Course content

What is IR?

Information Retrieval is about technology to connect people to information

Why studying IR?

Nowadays, IR problems are everywhere

- Text processing and analysis
- Various forms of ranking
- Ranked offline/online evaluation
- Learning from user interactions
- Etc.

What is so special about IR?

1. Relevance
   - "No one ever saw me but everyone knows I exist"
   - No precise definition
   - Highly subjective
   - Different in different scenarios
2. Ranking
   - Depends on relevance
   - Dependencies between ranked items

IR and AI

- IR uses AI
- IR learns from users (and, thus, contributes to AI)
- IR + NLP = set of techniques to work with text

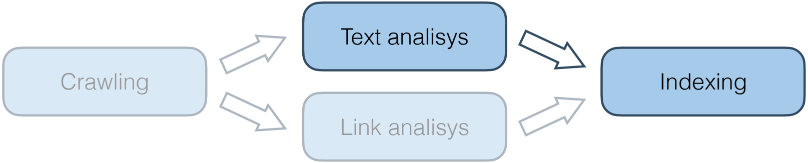
Outline

1. IR1 in numbers
2. Course organization
3. Intermezzo: Information Retrieval
4. Course content

IR1 2021

1. Basic techniques (IR0 recap)
2. Four pillars of IR
   - Evaluation
   - Document representation and matching
   - Learning to rank
   - IR-user interaction
3. IR scenarios
4. Current developments

Basic techniques



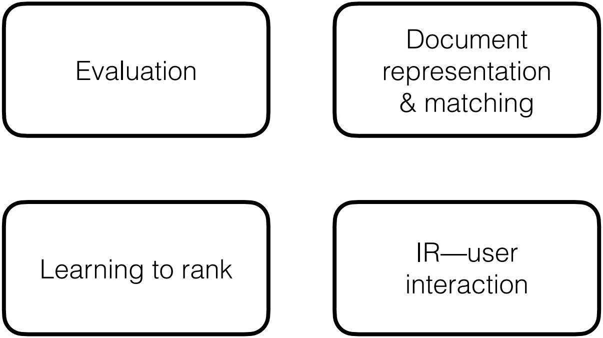
Text analysis

1. Statistical properties of written text
    - Zipf's law
    - Heaps' law
2. Text analysis pipeline
    - Stop-word removal
    - Stemming
    - Phrases

Indexing

1. Inverted index
   - Vocabulary
   - Inverted lists
2. Constructing an index
   - In-memory problem
   - Distributed indexing
3. Updating an index

Four pillars of IR



Evaluation

Document representation & matching

Learning to rank

IR—user interaction

(Offline) Evaluation

1. Offline evaluation metrics
    - Unranked: precision, recall
    - Ranked: AP, DCG
    - User-based: RBP, ERR
2. Test collections
    - Test documents
    - Test queries
    - Relevance judgements

Document representation and matching

1. Term-based retrieval
    - VSM+TF-IDF
    - QLM
    - BM25
2. Semantic retrieval
    - LSI
    - LDA
    - AWE
    - Doc2vec

avg word embedding

Document representation and matching

1. Vector-based
   - Documents and queries as vectors
   - Match using cosine similarity
   - Methods: VSM, LSI, AWE, Doc2vec
2. Distribution-based
   - Documents and queries as distributions
   - Match using QLM or Kullback-Leibler divergence
   - Methods: QLM, LDA

Learning to rank

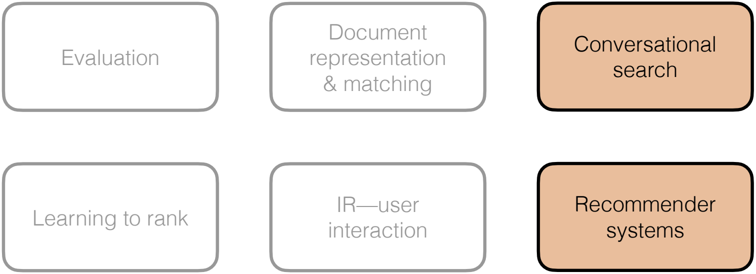
1. Point-wise (standard ML)
2. Pair-wise
   - Point-wise model $f(d_i)$, pair-wise loss $\mathcal{L}(d_i, d_j)$
   - Method: RankNet
3. List-wise
   - Replace cost with $|\Delta evaluation\_metric|$
   - Method: LambdaRank

IR-user interactions

6.6 online LTR will not be part of the explicit exam

1. Interactions and click models
   - Interactions are ambiguous and biased
   - Click models attempt to distinguish between bias and relevance
   - Methods: PBM, cascade model
2. Counterfactual ~~evaluation and~~ LTR
   - IPS
   - Propensity-weighted LTR
   - Estimation of position bias
3. Online evaluation ~~and LTR~~
   - A/B testing
   - Team draft/probabilistic/optimized interleaving
   - ~~Dueling bandit/pairwise differentiable gradient decent~~

IR Scenarios



Evaluation

Document representation & matching

Conversational search

Learning to rank

IR—user interaction

Recommender systems

Recommender systems

- Can be treated as a ranking problem with user profile instead of query
- All four pillars of IR are applicable directly
- Unique feature: explicit user ratings
- Collaborative filtering, e.g., matrix factorization

Conversational IR

- Very different from other IR scenarios
- Single vs. mixed initiative
- Standard IR evaluation can be adapted to some extent
- Document representation and matching can be reused, but...
- Initial question and conversation history are vital
- Not much research on LTR for conversational IR yet

Current developments

1. Neural models for passage matching and ranking
2. Query and document expansion
3. Weak supervision in LTR

Summary

Thanks everybody and good luck at the exam!