COMP0084 Information Retrieval and Data Mining (IRDM)

Lecture 1: Introduction

Ingemar J. Cox

Computer Science, UCL

Lecturers

Prof. Ingemar J. Cox i.cox@ucl.ac.uk

Dr. Vasileios (Bill) Lampos v.lampos@ucl.ac.uk

Prof. Emine Yilmaz e.yilmaz@cs.ucl.ac.uk

Teaching assistants

Aarzoo Dhiman (aarzoo.dhiman@ucl.ac.uk)

Yue Feng (yue.feng.20@ucl.ac.uk)

Michael Morris (michael.morris.19@ucl.ac.uk)

Fanghua Ye (fanghua.ye.19@ucl.ac.uk)

Office hours

- Time:
  - 9AM-10AM Tuesdays
  - Office hour: Tuesdays 9-10AM. Please use MS Teams and call Dr Vasileios Lampos. Please avoid group calls.
  - **Please do not discuss Coursework 1 in the forum or any other public medium**.
    Please ask directly during the office hours or any time via an email. The tutor will respond either via email or via a public announcement to all students.

Questions

- There a two types of questions:
  - Course specific
  - Other
- All course-specific questions should be sent via Moodle
- Any other (personal) questions should be sent directly to the appropriate person

Overwhelmed with Information and Data

Information Retrieval and Data Mining

- Making optimal decision heavily relies on gathered information and knowledge

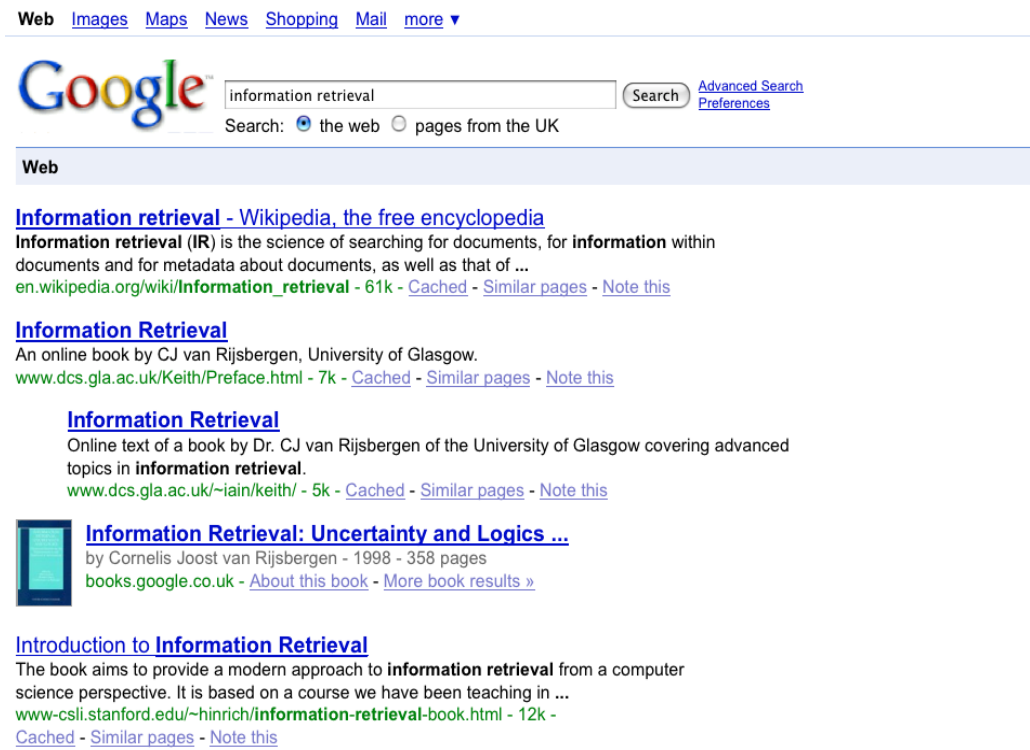
- Need <u>Information Retrieval</u> (IR):
  - obtain relevant information by searching large-scale unstructured data (data -> relevant information)
- Need <u>Data Mining</u> (DM):
  - extract knowledge, insights and useful information by mining the obtained information and data (data -> patterns and knowledge)

Outline

- Course administration
  - Outline of the syllabus
  - Learning outcomes
  - Project assignment
  - Grading system
  - Supporting
- Introduction of IR and DM

What will you expect from this course?

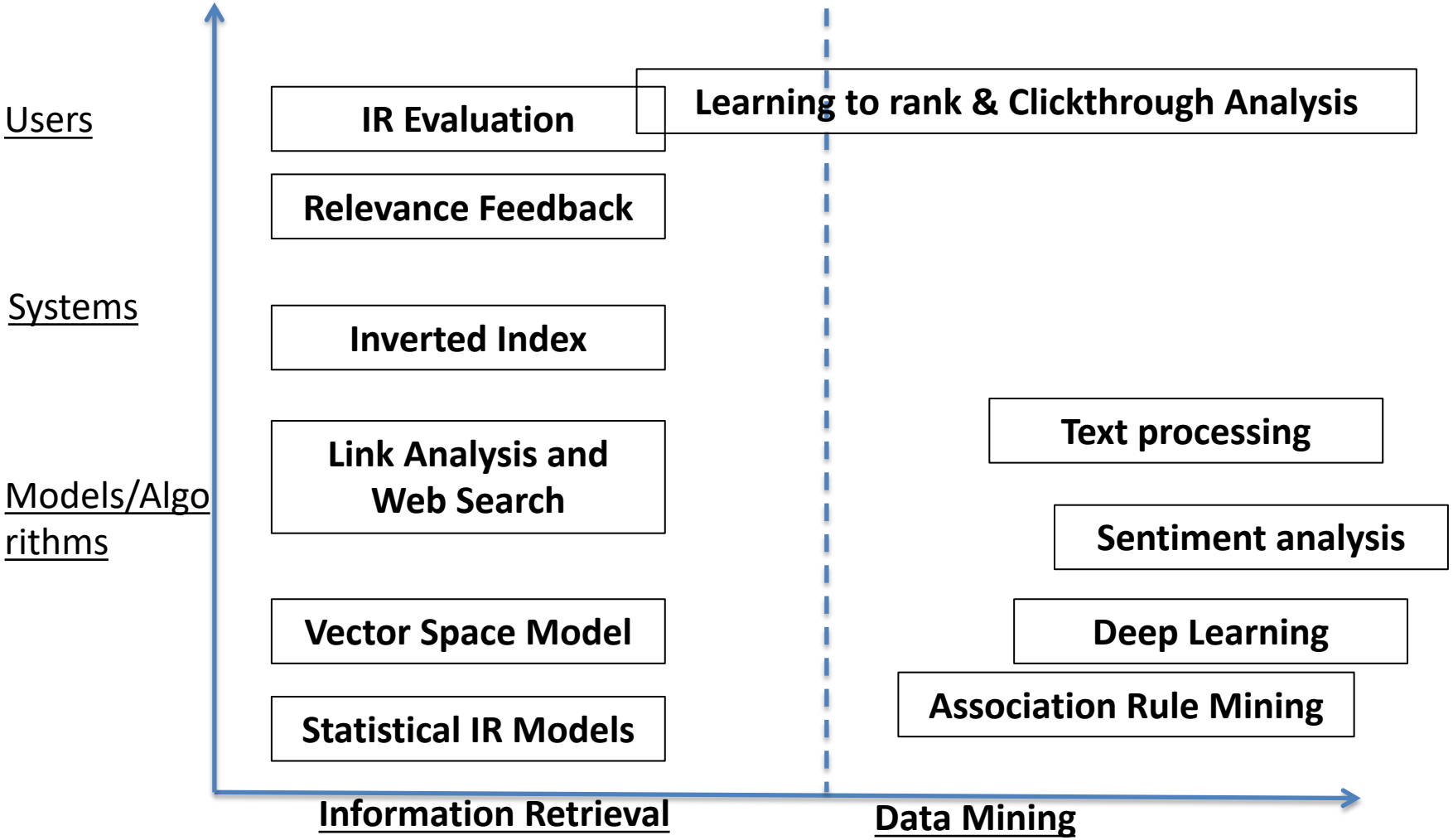
- Learn how a *search engine* (e.g. Google) works



What you will expect from this course?

- Learn how *search engine* (Google) works
- Learn how to employ *Data Mining* to extract knowledge

Overview of the syllabus



Data -> Relevance Information -> Knowledge -> Decision

Learning outcomes

- You will:

  - *understand* the *theory* and *practical algorithms* underlying IR/DM systems

    retrieval, filtering, probabilistic ranking, etc

  - know how to *build* an IR/DM system

    practical algorithms, data structures, etc

  - know how to *evaluate* IR/DM systems

    does the new algorithm perform better?

Prerequisite

- A basic understanding of probability and statistics
  - Chain rule, Bayes' rule, Maximum likelihood estimation, Gaussian distribution, Multinomial distribution etc
- Basic machine learning knowledge (not essential)
  - Classification, clustering, regression, generative models etc
- For those who are not familiar with Probability and Statistics, please self-study some online materials
- You need to know Python programming for the course work. Java is also possible, but Python is strongly recommended.

Assessment Criteria

Coursework: (100%) (details will follow)

Coursework	Weight	Submission deadline
1	50%	Friday 03 Mar 2023 at 16:00 (UK time)
2	50%	Tuesday 28 Mar 2023 at 16:00 (UK time)

Project report

- Assessment is based on quality, not quantity
  - Write succinctly – clear, but brief
  - Style should be as a research article
    - NOT a press release
  - Provide ALL details needed to replicate the experiments
  - Speculation should only be present in the Conclusion

Cheating and Plagiarism

- You CANNOT:
  - –Use someone else's work (text,codes) and ideas as your own (copy/paste, recycle, etc)
  - –Employ a professional or anyone else (including systems such as ChatGPT) to produce work for you
- You CAN:
  - –Quote, paraphrase but you must mention the source
  - –use public data but interpretation and conclusions derived from that data i.e. the ‘write-up’ must be your own.

System support: Moodle

We use Moodle

Login/access
http://moodle.ucl.ac.uk using your UCL's account

Course name:
COMP0084: Information Retrieval and Data Mining

Support System: Moodle

- You will find:
  - Syllabus,
  - Reading assignments,
  - Projects,
  - Lecture notes,
  - Copies of papers
  - Software packages/Data files

Recommended textbooks

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction of information retrieval, Cambridge, 2008

  http://nlp.stanford.edu/IR-book/

- Pang-Ning Tan, Michael Steinbach,Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2006

  http://www-users.cs.umn.edu/~kumar/dmbook/index.php

COURSE OUTLINE

Week 1: Ingemar J. Cox

General introduction to information retrieval

Week 2: Bill Lampos & Emine Yilmaz

Text processing and indexing

Week 3 – Emine Yilmaz

IR evaluation

Week 4: Ingemar J. Cox

Content independent ranking

Relevance feedback

Week 5: Bill Lampos

Introduction to machine learning and data mining

Week 6: Emine Yilmaz

Deep learning for IR

Lecture 7: Bill Lampos

Topic models and word embeddings

Week 8: Ingemar J. Cox

Compression algorithms

Week 9:

Guest lectures

Week 10: Ingemar J. Cox

And guest lectures

Guest lectures

Elad Yom-Tov, Microsoft Research

Adam Tsakalidis, Queen Mary University

Peter Wijeratne, UCL

Rishabh Mehrotra, Spotify