1SC4810 – Processing and analysis of massive unstructured data - the case of web data

Instructors: Myriam Tami, Wassila Ouerdane, Céline Hudelot
Department: DÉPARTEMENT INFORMATIQUE
Language of instruction: FRANCAIS
Campus: CAMPUS DE PARIS - SACLAY
Workload (HEE): 60
On-site hours (HPE): 34,50

Description

How can we automatically and quickly find relevant information to a particular need, based on a large amount of information ? This is typically what search engines such as Google or Baidu do effectively when they respond to the 4 million queries they each receive every second. The objective of this course is to describe the foundations and techniques of Information Retrieval (IR) on which these search engines are based. The course will also address some current challenges in the field such as the contributions of automatic and deep learning to IR or personnalisation and recommendation (collaborative filtering).

Quarter number
ST4

Prerequisites (in terms of CS courses)
Information systems and programming (ISP) - Algorithms design and Complexity.

Syllabus
Breakdown of the course into chapters:

- Introduction to Information Retrieval - Basics: indexing and inverted index.
- Boolean and vectorial search models.
- Probabilistic and language models.
- Evaluation of Information Retrieval Systems.
- Web search: crawling - case of large collections - distributed indexing (MAP REDUCE - Hadoop)
- Web search : link analysis
- Personalization: e.g. Recommendation systems

- Machine Learning for Information retrieval : categorization
- Machine Learning for Information retrieval : Learning to Rank

Contents of the Laboratories (practical work on machine and TDs)

- Indexing - Building of an inverted index
- Search Models : boolean and vectorial search models
- Search Models : probabilistic search models
- Evaluation of IR Systems
- Web search : MapReduce
- Web search : Link analysis -PageRank
- Recommandation systems
- Learning to Rank : point-wise approaches
- Learning to Rank : pair-wise approaches

Class components (lecture, labs, etc.)

9 sessions of 1h30 of classes, 13 sessions of 1h30 of Laboratories

Grading

Final control in the form of a written exam (1,30 hours) with calculator, handouts and course notes of the student authorized

Course support, bibliography

1. Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze (Cambridge University Press, 2008)
2. Information Retrieval: Implementing and Evaluating Search Engines, by S. Büttcher, C. Clarke, and G. Cormack.
3. Search Engines : Information Retrieval in Practice, by B. Croft, D. Metzler, and T. Strohman.
4. Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto.
5. Recherche d'information - Applications, modèles et algorithmes - Data mining, décisionnel et big data - Messa-Aminih Reza et Gaussier Eric - Eyrolles

Resources

o Teaching team : Céline Hudelot, Wassila Ouerdane, Myriam Tami, Bich-Liên Doan.
o Size of the Lab groups: 40 (3 or 4 groups)
o Software tools:

- Programming language: python
- Development tools : jupyter notebooks

Learning outcomes covered on the course

At the end of this course the student will have acquired a good understanding of the basic concepts of Information Retrieval and will be able to apply these concepts in practice. More precisely, it will be able to:

- Understand the problems of modeling, indexing and information processing related to Information Retrieval (IR).
- Understand how statistical text models and machine learning can be used to solve IR problems.
- Understand and make recommendations on the importance of data structures for effective access to information in large corpora.
- Apply the main concepts of Information Retrieval to the design and implementation of real applications of ad-hoc information retrieval.
- Analyze and evaluate the performance of information retrieval systems using test collections.
- Understand the current challenges of Information Retrieval such as large scale data processing or personnalisation.

Description of the skills acquired at the end of the course

- Understand the problems of modeling, indexing and information processing related to Information Retrieval (IR), Apply the main concepts of Information Retrieval for the design and implementation of real ad-hoc information system applications and Understand the current challenges of Information Retrieval such as large-scale date processing or personnalisation are part of **C1 "Analyze, design and build complex systems with scientific, technological, human and economic components"** and **C2 "Develop an in-depth competence in an engineering field and in a family of professions".**

- Understanding how statistical text models and learning can be used to solve IR problems and Understanding and making recommendations on the importance of data structures to enable effective access to information in large corpora are part of **C6 : "Be operational, responsible and innovative in the digital world"**

- Analyzing and evaluating the performance of information retrieval systems using test collections is part of **C3.3 "To concretely implement innovative ideas and commit to their decisions, to evaluate the solutions, to move to industrialization to deliver tangible results".**