Cours "Recherche d'information–Fouille de données : principes de base"

MR. Amini et M. Clausel

October 2, 2015

Qu'est ce qu'est la recherche d'information?
La recherche d'information (RI) consiste en la recherche de données non structurées, généralement des documents contenant des informations spécifiques. Un bon exemple est la recherche d'information sur internet, où l'on peut rechercher une information sans avoir exactement d'idée précise de ce que l'on va trouver (si on cherche par exemple une destination pour un voyage). L'approche en RI va consister à ordonner les données (par exemple les pages web que l'on explore) par rapportà une mesure de similarité par rapport à la requête qui est généralement un ensemble de mots clés (penser à une requête internet). Le point central en RI est donc de construire un système qui respectera au mieux l'ordre qu'il faut attribuer aux données.

Objectif
L'objectif de cette formation est de vous faire découvrir les premiers principes et quelques algorithmes de base en RI.

La formation est proposée sous une double étiquette Collège doctoral/MaiMoSiNE (Maison de la Modélisation et de la Simulation) avec une priorité d'accès aux étudiants du collège doctoral de Grenoble.

Prérequis :

- Connaissances de base en C/C^{++} avec notion de programmation objet. Construction d'un executable et utilisation d'une bibliothèque
- Algèbre linéaire matricielle de base (rang d'une matrice, valeurs propres vecteurs propres)
- Bases de probabilités

Contenu de la formation :
La gestion de grandes bases de données nécessite des modèles et des représentations pertinentes des données afin de pouvoir aller rechercher de manière efficace l'information.

Ce module est destiné à des scientifiques voulant acquérir des méthodes de bases en recherche d'information. Les différentes manières de représenter un document seront abordées ainsi que les différents modèles utilisés en RI.

L'enseignement comprend des exposés théoriques d'une à deux heures suivis de travaux pratiques en C basés sur les notions présentées.

Planning :
Chaque séance sera constituée d'un cours puis d'un TP.

1. Séances1,2 et 3 : 5/04/2016, 8/04/2016 et 12/04/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Représentation et indexation d'un document
   - Prétraitements linguistiques : segmentation, normalisation, filtrage par un anti-dictionnaire.
   - Les deux lois de base en recherche d'information : loi de Zipf et de Heaps
   - Représentation vectorielle d'un document, pondération des termes.
2. Séances 4, 5 et 6 : 15/04/2016, 26/04/2016 et 29/04/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Clustering
   - Algorithmes EM et Classification EM.
   - Mesures d'évaluation et expériences sur une base de données textuelle
3. Séances 7 et 8 : 3/05/2016 et 6/05/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Recherche de thèmes latents
   - Modèles : Latent Semantic Indexing (LSI) & Probabilistic LSI
   - Application à une collection de données
4. Séances 9, et10 : 10/05/2016 et 13/05/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Minimisation du risque empirique/Optimisation
   - Descente de gradient : le cas déterministe
   - Un exemple d'algorithme stochastique : le gradient stochastique
5. Séances 11 et 12 : 17/05/2016 et 20/05/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Classification
   - Perceptron / Adaline
   - Application à une collection de données