Cours "Recherche d'information–Fouille de données" : applications à des données complexes

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). Ce cours se veut un prolongement au cours "Recherche d'information–Fouille de données : principes de base". L'objectif est d'offrir un panorama des différentes applications allant de la fouille de données dans un grand corpus de tweets jusqu'à l'exploration d'une base de données d'images ou de signaux.

Objectif
L'objectif de cette formation est de vous faire découvrir comment appliquer les principes et les principaux algorithmes en RI à des données comples (signaux, images, tweets)

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.

Pré-requis :

- 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 adapter les méthodes classiques de RI à des données complexes : images, signaux. Les différentes manières de représenter un document seront abordées ainsi que l'adaptation des différents modèles utilisés en RI.

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

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

1. Séances 1, 2 et 3 : 24/05/2016, 27/05/2016 et 31/05/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Le modèle LDA et ses extensions

   - Présentation du modèle et de ses variantes. Application à une base de données textuelles
   - Inférence statistique: EM variationel, Gibbs sampling
   - Le cas des tweets.

2. Séances 4,5 et 6 : 3/06/2016, 7/06/2016 et 10/06/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Recherche d'information pour des bases de données de signaux

   - Extraction de caractéristiques d'un signal
   - Classification de signaux
   - Le modèle Gaussian–LDA

3. Séances 7,8 et 9 : 14/06/2016, 17/06/2016 et 21/06/2016 de 13h30 à 16h30 (1h30 Cours ; 1h30 TP)
   Recherche d'information pour des bases de données d'image

   - Extraction de caractéristiques d'une image
   - Classification d'images
   - Adaptation du modèle LDA au contexte image