Nom du Syllabus :	Web Mining
Niveau :	M2
Année :	2
Trimestre :	2
Enseignant (e):	Yoann PITARCH

OBJECTIFS :

L'objectif de ce module "web mining" est double : il s'agira dans un premier temps de dresser un panorama assez large de ce qu'est le "web mining" et des techniques mises en œuvre pour fouiller ces données. Dans un second temps, les étudiants se confronteront eux-mêmes à cette problématique d'extraction de connaissances à travers la mise en place d'un TP/projet. D'un point de vue pratique, le début du module s'articulera sous la forme de cours formels et représentera 30% environ du volume horaire total. La seconde partie se déroulera sous la forme de séances de TPs.

Le "web mining" peut être vu comme l'application de techniques issues de la fouille de données ("data mining" en anglais) sur des données provenant du web. Le cours s'orientera vers la description du "web mining" à proprement parlé. Il est d'usage de diviser ce domaine en 3 sous-domaines : la fouille de contenu web (web content mining), la fouille de structures (web structure mining) et la fouille d'usage web (web usage mining). Pour chacune de ces catégories, nous décrirons les motivations, les applications, quelques travaux de recherche majeurs et conclurons en mentionnant quelques défis à relever.

Concernant la deuxième partie du module, l'objectif est de permettre aux étudiants d'expérimenter l'ensemble du processus d'extraction de connaissances : de la récolte des données jusqu'à l'analyse et l'interprétation des résultats d'analyse. Quelques bases de programmation seront alors nécessaires (essentiellement pour récupérer les données et les pré-traiter). Le langage de

programmation Perl sera alors abordé et quelques TPs permettront de consolider les connaissances apprises lors du cours. Nous utiliserons ensuite le logiciel R pour lancer quelques analyses et interpréter les résultats de cette analyse.

PLAN DU COURS :

Partie 1 : Web Mining

1/ Introduction. Nous aborderons l'histoire du Web, rappellerons quelques concepts clés du Data Mining et définirons les particularités des données issues du Web. Une première définition du Web Mining sera donnée et mettra en évidence les 3 catégories d'approches qui seront développées dans le reste du cours.

2/ Web Structure Mining. Nous borderons ici les approches qui visent à analyser la structure du Web. Comme nous le verrons, ces approches ont de nombreuses similarités avec l'analyse de réseaux sociaux que nous introduirons. Nous développerons ensuite l'algorithme PageRank, aborderons l'analyse de réseaux d'auteurs dans les publications scientifique et terminerons par quelques méthodes qui permettent de découvrir des communautés dans un réseau.

3/ Web Content Mining. Ici il s'agit non plus d'analyser les relations entre les entités mais plutôt le contenu des pages Web. Nous nous focaliserons sur un pan important des efforts menés dans ce domaine : l'extraction automatique d'opinion (a.k.a. Opinion Mining). Nous aborderons la classification automatique de documents en fonction des sentiments exprimés, verrons que souvent, le niveau de granularité du document est trop élevé pour capturer réellement ce que pense un utilisateur d'un produit (ou autre) et aborderons enfin la détection automatique de faux avis.

Université Toulouse 1 Capitole
Ecole d'économie de Toulouse –TSE
Manufacture des Tabacs
21, allée de Brienne
31042 Toulouse Cedex

4/ Web Usage Mining. L'analyse de l'utilisation du Web par les utilisateurs équivaut grossièrement à l'analyse des logs générés par les serveurs Web. Nous verrons que dans ce contexte, ce n'est pas l'analyse qui pose des difficultés (de nombreux algorithmes standards du Data Mining peuvent être appliqués) mais plutôt le pré-traitement de ces données pour permettre une extraction de connaissance pertinente. Parmi les difficultés associées à ces données on peut cite entre autres l'identification automatique de sessions utilisateurs, la complétion de chemin d'accès ou encore l'intégration de données. Après avoir décrits et données des éléments de réponses pour pallier ces difficultés, une modellisation des données est proposée et nous aborderons la découverte de motifs sur ces données.

Partie 2 : Perl

Ce cours permettra de donner les bases du langage de programmation Perl. Nous aborderons les aspects syntaxiques du langage et réaliserons quelques programmes simples. Ensuite, nous évoquerons l'utilisation d'expressions régulières, qui sont un outil très puissant pour le traitement de données textuelles. Ce savoir permettra de réaliser efficacement des pré-traitement pour les données sur lesquelles seront ensuite appliqués des approches de Web Mining à partir du logiciel R.

PRE-REQUIS :

- Module de Data Mining

BIBLIOGRAPHIE : Aucune lecture n'est nécessaire pour débuter ce module. Le support de cours fournira ensuite une bibliographie assez complète des points abordés.

Université Toulouse 1 Capitole
Ecole d'économie de Toulouse –TSE
Manufacture des Tabacs
21, allée de Brienne
31042 Toulouse Cedex