Fouille de données sur le Web

Proposition d'option d'enseignement pour le corps Mines/Télécoms

1 Contexte général

Le World Wide Web est aujourd'hui incontournable à la fois comme source d'informations et comme moyen de communication. Le cours présenté ici permettra aux élèves du corps des Mines et Télécommunications de comprendre comment le Web fonctionne, comment les données textuelles et multimédia y sont organisées et comment les moteurs de recherche présents et à venir fouillent, indexent et mettent à disposition ces données. Ces aspects sont cruciaux pour des fonctionnaires des grands corps techniques de l'état qui pourront être amené à utiliser le Web pour de la veille technologique, à coordonner la mise en place de sites Web d'informations et de services pour des administrations ou grandes entreprises publiques, ou encore à proposer des solutions pour une meilleure exploitation des données d'un système d'information interne comportant un Intranet.

2 Contenu du cours

Les grandes parties du cours sont détaillées ci-dessous. Les parties a), b), c) (données textuelles) sont sous la responsabilité du département INFRES de TELECOM ParisTech, tandis que les parties d), e) et f) (données multimédia) seront organisées par le département TSI.

a) Le World Wide Web. Ce cours donne des bases sur le Web, ses protocoles et langages, et décrit la manière dont on peut le parcourir pour l'archiver ou l'indexer.

- Métriques, Internet et le World Wide Web, Protocoles du Web : HTTP, HTTPS
- Introduction aux langages et formats du Web, côté client et côté serveur
- Conception d'un crawleur du Web

b) Recherche d'informations textuelles. Ce cours décrit les méthodes classiques d'indexation et de recherche dans des grands corpus textuels, qui sont à la base de tous les moteurs de recherche.

- Prétraitement textuel (découpage en lexèmes, lemmatisation, etc.)
- Construction d'un index inversé, pondération par tf-idf ou modèles statistiques
- Requêtes renvoyant les k premiers résultats (algorithmes de Fagin, etc.)
- Classification dans l'espace vectoriel des documents

c) Le Web vu comme un graphe. Considérer le Web comme un graphe a permis de nettes améliorations dans les résultats fournis par les moteurs de recherche.

- Modèle du graphe du Web, typologie du graphe
- Algorithmes HITS (Kleinberg) et PageRank (GOOGLE)
- Contrer le *spamdexing*
- Extensions aux réseaux sociaux, publicité ciblée, systèmes de recommandation

d) Base de représentation des signaux et des images. Ce module présente les bases de l'analyse et du traitement des signaux audio et des images. L'objectif est d'introduire des outils de bases (échantillonnage, filtrage, transformées), donner un aperçu de quelques méthodes classiques de traitement (restauration, débruitage, segmentation).

e) Représentations des contenus multimédias. Le stockage et le transfert de grandes masses de données (vidéo, musique, images) sur les réseaux fixes ou mobiles se sont considérablement développés ; il est maintenant possible de réaliser ces transferts en des temps raisonnables, même en présence de contraintes sur les bandes passantes disponibles, grâce à l'existence d'algorithmes de compression efficaces.

f) Apprentissage et fouille de données multimédia. La fouille de données multimédias nécessite de mettre en œuvre des techniques performantes permettant de représenter les données pour être en mesure de les mesurer leur proximité, les classer (de façon supervisée ou non) des contenus, les hiérarchiser… Dans ce module, on présentera les fondements de l'apprentissage et les problèmes qu'il permet d'aborder. Un accent tout particulier sera mis sur les méthodes de classification de données en grande dimension, l'intérêt de ces concepts et techniques étant illustré au travers d'applications à la fouille de données multimédias (images, signaux audio).

g) Défis pour le Web d'aujourd'hui et de demain. Ce dernier cours, proposé par un intervenant du moteur de recherche Exalead, décrit des aspects plus prospectifs de la recherche sur le Web : Web profond, Web 2.0, Web sémantique, applications pratiques de la recherche multimédia, calcul massivement distribué, MapReduce …

3 Modes pédagogiques

Le contenu détaillé ci-dessus est prévu pour un volume d'une soixantaine d'heures de cours et travaux dirigés. Le format précis des modules d'enseignement n'étant pas encore défini, le format du cours pourra être adapté (et éventuellement allégé ou développé) le cas échéant. Les séances d'enseignement pourront associer cours et travaux dirigés, ces derniers permettant d'appliquer sur des exemples les algorithmes et systèmes présentés, et d'approfondir la discussion de certains sujets. Un équilibre sera recherché entre théorie (algorithmes fondamentaux), pratique (langages et systèmes concrets, applications) et ouverture sur les problèmes auxquels sont confrontés les moteurs de recherche et outils de fouille actuels et futurs.

En fonction du temps disponible, il pourra être fait appel à des conférenciers extérieurs pour aborder des sujets connexes. Nous avons obtenu l'accord de principe des conférenciers suivants :

— Serge ABITEBOUL (DR INRIA Saclay) : données et calcul distribué sur le Web
— Julien MASANÈS (directeur European Archive) : archivage du Web

D'autres intervenants pourront donner un éclairage « métier » (sur les aspects liés à la régulation, aux nouveaux usages, à la protection des droits).

Afin de mettre en application les algorithmes et techniques vues en cours, il sera demandé aux étudiants de réaliser un projet individuel ou en groupe, à choisir parmi une liste.

Les seuls prérequis nécessaires sont une maîtrise d'outils de bases de statistiques et de mathématiques appliquées (enseignés dans les écoles dorigine) ; une connaissance de la programmation est souhaitée pour la réalisation des projets, mais certains sujets viseront les élèves n'ayant pas d'expérience dans ce domaine.

4 Intervenants

TELECOM ParisTech : Éric MOULINES (TSI), Pierre SENELLART (INFRES) ; d'autres enseignants de ces deux départements pourront également intervenir.

Exalead : Florian DOUETTEAU (ingénieur R&D)

Conférenciers extérieurs : Serge ABITEBOUL (DR INRIA Saclay), Julien MASANÈS (directeur European Archive), autres intervenants non encore connus