FICHE DE POSTE INGENIEUR de RECHERCHE/Post-Doct

Domaine de recherche :	*Projet de recherche :*
Apprentissage par machine, analyse de données dynamique	ISTEX-R

Poste INGENIEUR de RECHERCHE / Post-Doctoral

Nom du responsable direct :	Date de début du contrat et durée :
Jean-Charles Lamirel	A partir de Janvier 2015 – 12 Mois

A propos de l'INRIA et du poste (contexte)

Le projet ISTEX-R est un projet de recherche appliquée qui vise à intégrer et à mettre à disposition sur la plateforme ISTEX des outils d'accès au contenu opérant sur des textes intégraux et permettant de construire et capitaliser des connaissances sur un domaine scientifique ou technique. La plateforme ISTEX est alimentée par l'achat ou la mise à disposition par des éditeurs scientifiques ou techniques d'un très grand volume de textes intégraux (http://www.istex.fr/).

ISTEX-R part des acquis pour aller vers une analyse plus fine du contenu : comment caractériser l'évolution des recherches et des connaissances dans le temps ? L'évolution des connaissances dans un domaine se fait souvent par des glissements subtils d'une problématique vers une autre et par un enrichissement progressif des connaissances. La construction de cartes diachroniques vise à outiller l'expert d'un domaine pour lui permettre d'observer ce type d'évolution.

L'accès au contenu des textes sur un domaine passe également par la conceptualisation de ce domaine, et par celle du contenu des textes proprement dits, qui permettent de capitaliser les connaissances exprimées au travers des textes.

Mission

La mission du stage sera celle d'implanter, d'adapter et de valider l'exploitation de différentes méthodes d'analyse dynamique de l'information textuelle (**analyse diachronique**, **clustering incrémental**, **synthèse automatique de contenu**, **visualisation**) dans le cadre de la plate-forme ISTEX. Seront privilégiées les méthodes d'analyse dynamique développées au LORIA dans le cadre d'une collaboration antérieure avec l'INIST qui a fait l'objet de l'action CPER TALC MCFIID (Multiclustering de flux d'informations incrémental et distribué). Ces méthodes devront pouvoir être lancées en temps réel sur des résultats de recherche, ou en temps différé sur des corpus, ceci dans le cadre d'une plate-forme interactive destinée aux utilisateurs finaux.

Descriptif du poste (activités)

Le développement de **méthodes d'analyse dynamique de l'information**, comme le clustering incrémental et les méthodes de détection de nouveauté, devient une préoccupation centrale dans un grand nombre d'applications dont le but principal est de traiter de larges volumes d'information textuelles variant au cours du temps. Jusqu'à présent les techniques statistiques existantes s'avèrent inefficaces lorsqu'elles sont appliquées au texte du fait des contraintes intrinsèques de la représentation des données textuelles en machine, qui tendent à produire des espaces de description fortement multidimensionnels et épars.

Afin de repérer et analyser les émergences, ou de détecter des changements dans les caractéristiques des données, nous proposons de suivre deux approches différentes et complémentaires :

1- réaliser des classifications statiques à différentes périodes de temps et analyser les évolutions entre les différentes périodes (approche par pas de temps ou analyse diachronique),

2- développer des méthodes de classification qui permettent de suivre directement ou de pister les évolutions : les méthodes de classification incrémentales proprement dites (clustering incrémental) et les méthodes de détection de nouveauté (classification supervisée incrémentale).

L'**analyse diachronique** par pas de temps est fondée sur l'application d'une méthode de classification automatique sur des données associées à deux, ou plus, périodes de temps successives, et sur l'étude de l'évolution des résultats de classification obtenus. Nous développons dans ce cadre des solutions inédites basées sur l'exploitation de nouvelles mesures d'évaluation de la qualité du clustering, indépendantes de la méthode de clustering utilisée, que nous combinons à une approche multi-vues exploitant le raisonnement bayésien non supervisé. Cette combinaison de techniques originales nous permet à la fois d'optimiser et d'automatiser le processus de comparaison entre périodes.

En ce qui concerne le **clustering incrémental**, nous proposons une nouvelle approche incrémentale sans paramètres, fondée sur une adaptation générique des méthodes de clustering neuronales à topologie libre, à base de gaz de neurones croissants, et sur l'exploitation conjointe de distances ad-hoc, spécifiques au traitement des données textuelles, comme les distances de maximisation de traits, en lieu et place des distances usuelles. Nous comptons également expérimenter ce nouveau type de distances sur une large gamme de méthodes de clustering statiques et incrémentales, appliquées au texte, et mettre au point de nouvelles méthodes incrémentales, telles que des méthodes prometteuses fondées sur les composantes connexes des propriétés associées aux données des classes.

La **synthèse du contenu** des documents a pour but de répondre aux problèmes d'indisponibilité des métadonnées. Il s'agit également dans ce cas d'isoler les informations centrales véhiculées par un document indépendamment de sa source de production. Nous avons récemment proposé dans ce cadre une approche originale qui exploite la compétition entre les blocs (page, paragraphe ou structure logique si elle est présente) coordonnée par les distances de maximisation de traits. Elle présente l'avantage d'être indépendante de la langue et de ne nécessite que des prétraitements linguistiques très élémentaires des textes, voire aucun. Elle est polyvalente puisque qu'elle permet de créer des résumés sur forme de graphes d'interaction entre les mots ou sous forme de liste de phrases à fort pouvoir d'information, ou encore de créer des index. Elle nécessite cependant d'être testée au passage à l'échelle.

La **visualisation** de résultats de classification incrémentale reste également un problème important. Sans cette étape, des tableaux de nombres et de mots sont les seules sorties que l'expert peut analyser, avec toutes les difficultés que l'on imagine. Il est probable qu'après avoir exploré diverses pistes la solution idéale ne réside pas dans un seul type de visualisation, mais plutôt dans une combinaison d'approches.

Profil recherché (compétences attendues)

Doctorat (préférentiellement) ou diplôme d'ingénieur dans le domaine du traitement numérique, analyse statistique, classification automatique, avec de fortes compétences en développement informatique. Connaissances de bon niveau en analyse de données et en mathématiques appliquées. Une connaissance des problèmes de classification ou de clustering sur des données textuelles serait un plus très apprécié, de même que des connaissances de base en traitement automatique des langues. Une connaissance supplémentaire concernant le développement d'interfaces et d'applications Web serait également très appréciée.

Compétences en programmation requises : C et C++, Python, Java ; Anglais : lu, écrit.

Avantages (conditions de travail)
(salaire brut et net, restauration collective, éventuels déplacements,...)

Le poste est basé au LORIA à Vandœuvre les Nancy.

Salaire selon les diplômes et suivant la grille de la fonction publique.

Cet emploi donnera lieu à des publications et peut être considéré comme une formation postdoctorale pour les titulaires de doctorat.

Informations complémentaires (diplômes requis, horaires,...)

Doctorat (préférentiellement) ou diplôme d'ingénieur.

Coordonnées de la personne à contacter :

Jean-Charles Lamirel – Habilité à Diriger des Recherches - Equipe Synalp – LORIA

Email : lamirel@loria.fr

GSM : 06 24 36 54 91