CHAIRE MACHINE LEARNING FOR BIG DATA

Paris, le 25/11/16

Offre de stage

Sujet : Analyse des traces d'usage de Gallica

Encadrement

Florence D'Alché-Buc, Valérie Beaudouin, Christophe Prieur, François Roueff.

Lieu et dates du stage

Telecom ParisTech, 46 rue Barrault, 75013 Paris

Date de début du stage : à partir de février 2017

Équipe(s) d'accueil de la thèse

Equipe Statistique, Signal & Apprentissage du département TSI, département SES.

Mots clés

Analyse sémantique, apprentissage non-supervisé, modèles de Markov, processus ponctuels, elasticsearch, LDA, Word2Vec.

Sujet détaillé

Ce stage s'inscrit dans la continuité du partenariat entre la Bibliothèque nationale de France (BNF) et Télécom ParisTech, qui a donné naissance au *Laboratoire d'étude des usages du patrimoine numérique des bibliothèques* (Bibli-Lab). Il a pour but d'introduire des éléments de statistique décisionnelle permettant la compréhension des usages de Gallica, site de consultation de plus de trois millions de documents numérisés de la BNF. Une étude en cours a permis de mettre en place une architecture d'analyse des traces des logs du serveur Gallica et d'en produire une première analyse exploratoire. Elle a été motivée par des travaux de nature sociologique qui ont été conduites sur ce sujet ([1]). L'architecture de stockage utilise un serveur elasticsearch rapatrié sur la plateforme TeraLab [2], sur laquelle les algorithmes d'analyse sont aussi effectués (essentiellement en Python).

Les premiers éléments de l'étude ont montré que les logs de connexion permettent de distinguer des types d'usages de Gallica à partir de l'organisation temporelle et/ou séquentielle de chaque session.

Les travaux en cours, dans la continuité desquels s'inscrit ce stage s'intéressent :

1) d'une part à l'intégration des contenus des documents visités au cours de chaque session afin d'en extraire un comportement sémantique. Plusieurs approches sont envisagées : LDA, Word2Vect. [3,4].

2) d'autre part à l'étude de l'impact des campagnes de médiation menées par les équipes de communication de la BNF sur la consultation de Gallica.

Une originalité forte de ces études est d'avoir pour ambition de valider (ou du moins d'évaluer) sur l'ensemble de la base des logs de connexion la pertinence statistique d'usages typiques observés à partir d'enquête sociologiques, sous la forme de questionnaires en ligne ou d'interviews.

Les modèles envisagés sont typiquement des modèles de mélange de modèles classiques permettant de décrire une succession d'événements, tels que les modèles de Markov cachés (voir par exemple [5]). L'utilisation de modèles de mélange est nécessitée par l'approche non-supervisée d'apprentissage. Les modèles probabilistes intra-classes sont quant à eux dictés par les comportements à modéliser. Des modèles plus élaborés basés sur les processus ponctuels pourront être notamment explorés durant le stage, qui peut donc conduire aussi bien à des travaux de nature théorique qu'appliquée.

La Chaire Machine Learning for Big Data

Le traitement statistique des masses de données convoque à la fois mathématiques appliquées et informatique, à travers une discipline en plein essor : le Machine Learning ou apprentissage statistique.

Créée en septembre 2013 avec le soutien de la Fondation Télécom et financée à hauteur de près de 2 M€ par quatre entreprises partenaires : Criteo, PSA Peugeot Citroën, Safran et BNP Paribas, la Chaire Machine Learning for Big Data est portée par le mathématicien Stéphan Clémençon, Enseignant-Chercheur, Professeur au sein du Département du Traitement du Signal et des Images à Télécom ParisTech.

Proposant cinq axes de recherche méthodologiques, enrichis par des applications industrielles concrètes, cette Chaire a pour objectif d'animer, en interaction avec ses partenaires, une activité de recherche de pointe en Machine Learning, ainsi que de proposer des programmes de formation.

La variété des données aujourd'hui disponibles (nombres, images, textes, signaux), leur grande dimension et leur volumétrie rendent souvent inopérantes les méthodes statistiques traditionnelles reposant sur le prétraitement humain et un long travail de modélisation. Le Machine Learning vise donc à élaborer et étudier des algorithmes, à vocation prédictive le plus souvent, permettant à des machines d'apprendre automatiquement à partir des données et à effectuer des tâches de façon performante.

Les avancées technologiques, l'omniprésence des capteurs (systèmes embarqués, objets connectés, Internet…) et l'explosion des réseaux sociaux s'accompagnent d'un véritable déluge de données, propulsant les sciences de l'information au centre du processus de valorisation des masses de donnés. En plus de la collecte et du stockage, l'enjeu est de pouvoir analyser ces données afin d'optimiser les décisions et mettre au point de nouvelles applications.

Au-delà du buzz médiatique dont il fait l'objet, le Big Data est donc un sujet stratégique majeur, au cœur d'enjeux économiques et sociétaux considérables. Son impact est désormais perçu dans

presque tous les secteurs de l'activité humaine : de la recherche scientifique à la médecine en passant, entre autres, par la finance, le bâtiment, l'e-commerce, la défense ou les transports.

En savoir plus sur la Chaire, ses axes de recherche, ses activités, ses publications :

http://machinelearningforbigdata.telecom-paristech.fr

Profil du candidat

Etudiant titulaire d'un master 2 recherche

- Apprentissage statistique / reconnaissance des formes
- Traitement de la parole, traitement du langage naturel
- Bon niveau en programmation (Java, C/C++, Python)
- Bon niveau d'anglais

Candidatures

à envoyer à roueff@telecom-paristech.fr

- Curriculum Vitae
- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)
- Relevés de notes des années précédentes
- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Référence

[1] Valérie Beaudouin and Jérôme Denis. Observer et évaluer les usages de Gallica. Réflexion épistémologique et stratégique. Research report, BnF ; Telecom ParisTech, September 2014. URL https://halshs.archives-ouvertes.fr/halshs-01078530.

[2] Teralab, Plateforme *cloud computing* de stockage et de clacul. https://www.teralab-datascience.fr/fr/.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan) :993--1022, 2003.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111--3119, 2013.

[5] Lawrence R. Rabiner. Readings in speech recognition. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267--296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1-55860-124-4. URL http://dl.acm.org/citation.cfm?id=108235.108253.