Argumentaire sur la demande 2014 de poste MC 27/61
UTLN, UMR CNRS LSIS, équipe Dynamiques de l'Information (DYNI)

« SCIENCE DES MASSES DE DONNEES » (3 pages)

1. Introduction

L'équipe DYNI dont la dernière publication de poste, malgré ses demandes récurrentes, remonte à la campagne 2010-11, a été évaluée A+ par l'Aeres qui en soutient fortement le développement, comme le montre cet extrait de son rapport : *« Le projet (DYNI) est pertinent et ambitieux mais nécessiterait sans doute plus de forces dans l'équipe dont la taille est petite en nombre de permanents (2PR 4MC) tous par ailleurs enseignants-chercheurs. Avis global: Equipe jeune et très dynamique, avec une approche originale et ambitieuse qui s'appuie sur et se nourrit de compétences multi-disciplinaires pour aborder de nombreux domaines d'applications à fort impact sociétal - Points forts & opportunités : Les thématiques de recherche développées sont tout à fait d'actualité et tendent à lever des verrous scientifiques et technologiques importants. L'équipe est pluridisciplinaire et possède les compétences nécessaires à ses ambitions. Elle a développé plusieurs modèles originaux de recherche d'information multimodale qui ont été très bien évalués dans le cadre de campagnes d'évaluation internationales. Les applications de ses recherches rencontrent des besoins industriels et sociétaux croissants en termes de surveillance et de détection d'événements. »*

Les recherches de DYNI évoluent vers le traitement des Masses de Données (MD) (ou *big data*). La production des MD numériques double tous les 3 ans depuis 1980, et dépasse les 10e8 octets par jour, au cœur des industries et services (Google..., finances, télécommunications, bio-informatique, physique, neurosciences, sciences sociales...). Ces enjeux ont motivé le lancement d'une «Big Data Initiative» aux Etats-Unis, mais aussi en France avec le programme CNRS MD (MASTODONS) qui finance depuis 3 ans le projet « Scaled Acoustic BioDiversity » (http://sabiod.org) piloté par DYNI fédérant 5 UMRs et des équipes internationales.

Les verrous des MD consistent à organiser / modéliser des bases de MD hétérogènes de grand volume, à distribuer et manipuler les calculs délicats en grandes dimensions (GD). Cela nécessite la représentation de données de grande dimension souvent peu étiquetées et d'une variabilité considérable. Réduire cette variabilité passe par la construction de représentations invariantes. Une approche cascade des opérateurs contractants, afin d'obtenir des structures de plus en plus spécialisées et invariantes, comme par transformations en ondelettes récurrentes, ou réseaux de neurones profonds. La régulation par parcimonie joue un rôle important dans ces algorithmes.

D'autre part, la prise de décision à partir des MD doit par nature être principalement non-supervisée : les MD n'impliquent pas de profiter de plus d'informations. Il est en effet très difficile, voire impossible, d'annoter précisement des bases à grande échelle. De ce fait, le recours à des méthodes d'apprentissage non-supervisées est une bonne direction pour raffiner les décisions en tirant profit de l'énorme quantité d'information. Se placent alors naturellement les méthodes génératives, à très bonnes performances par exemple en analyse de réseaux sociaux (mélanges de blocs latents), ou en classification à l'échelle de documents (topic modeling) à travers des modèles bayésiens non-paramétriques (analyse de Dirichlet latente). D'ailleurs l'approche en plein essor est générative par réseaux de neurones profonds.

L'enjeu réside principalement dans la construction de modèles, notamment statistiques, avec des garanties d'optimalité. En effet, face à la complexité et au volume des données, il est fondamental de décliner un fondement théorique solide pour ces méthodes à fin de tirer totalement parti des MD

2. Synthèse du profil Recherche du MC sollicité par DYNI

Cet(te) MC complétera dans DYNI la recherche d'algorithmes d'analyse de masse de données multimodales complexes dans un contexte qui peut être dynamique. Il se décline en deux sous thèmes, le/la MC en développera l'un des deux:

2.a) Traitement de MD issues du Web, stockage et manipulation optimisés dans un univers "en nuages" de MD hétérogènes : Le/la MC proposera des modèles de représentation adaptés à des MD hétérogènes dans leur forme (textes, audiovisuel, vidéos, sons, etc), et dans leurs structures, annotées, dynamiques (évoluant au cours du temps), pouvant être distribuées et associées aux connaissances des domaines auxquelles elles

sont rattachées. La compétence recherchée viendra en appui pour travailler sur des architectures, modèles et algorithmes pour l'intégration de nos modèles et langages dans des environnements distribués de MD. En particulier, le changement majeur dans le domaine des bases de données que constitue l'émergence d'un univers 'en nuages' et les systèmes associés 'NoSQL' renouvelle les problématiques de l'indexation des données et d'optimisation des requêtes.

2.b) Théorie des modèles statistiques sur MD, classification et indexation : *Le/la MC* proposera des modèles théoriques d'apprentissage statistique, de représentations parcimonieuses de MD pour leur classification. Les validations porteront notamment sur les MD hétérogènes d'observatoires environnementaux, ou MD audiovisuelles, textuelles ou encore fonctionnelles. La compétence recherchée portera sur les modèles probabilistes pour l'apprentissage à l'échelle et/ou les descripteurs issus de modèles basés sur des estimés locaux de densités, comme les approches hiérarchiques par réseaux profonds, offrant une représentation multi-échelle des données. L'apprentissage joint à des données de plus haut niveau, ou de modèles à variables latente pour des données fonctionnelles, serait un plus. Un intérêt particulier devra être donné au développement théoriques des modèles, et un effort particulier devra être réalisé pour assurer sur chaque approche le passage à l'échelle des algorithmes d'apprentissage.

Mots clés : Bases de données, modèles et langages, optimisation de requêtes, ontologie, informatique en nuages (cloud), masses de données, apprentissage de modèles statistiques à variable latente, classification non/semi/supervisée, analyse de données fonctionnelles, réseaux de neurones profonds, ondelettes, données multimodales, fossé sémantique, masse de donnée, audiovisuel, bioacoustique, web.

3. Activité de qualité avérée de l'équipe dans le domaine de recrutement

Les recherches de DYNI dans ce thème sont reconnues au niveau national (évaluation Aeres A+), et soutenues par l'IUF, la FRIIAM, des projets DGA/ONERA/INRIA, et plusieurs ANRs (ROSE, OTIM, ANCL, COGNILEGO, LEMONS (pré-acceptée en 04.2014), des PEPS…), ainsi que par les pôles MER et OPTITEC, et par les Axes transversaux Mer et Information de l'UTLN. Les publications de DYNI dans le domaine sont de rang A+ avec les revues IEEE TSP, ASA, Neurocomputing,... et les conférences du thème (ICML, NIPS, IJCAI,...), (voir lsis.org).

En 2014, DYNI a initié et supporte sur ce thème de recherche des projets de maturations avec la SATT (brevet USA), des projets DGA (OPTIM 1 et 2). Une collaboration est en cours (CIFRE) en modèles de recherche / indexation d'information (CIFRE PME Coexel), en analyse de scène (coll. avec PME Prolexia IFREMER DCNS, projets FUI SYCIE et RAPID PHRASE en cours 2013-2017). Le MC sera également acteur dans le projet 2012-2017 de la Mission Interdisciplinaire du CNRS inter-instituts MASSE de DONNEES MASTODONS «Scaled Acoustic Biodiversity» (http://sabiod.univ-tln.fr - PI DYNI), qui fédère à ce jour une trentaine de chercheurs sur les méthodes de traitement de MD bioacoustiques et qui structure le paysage national sur le sujet (suivant la politique de la DS INS2I CNRS). De plus ces activités s'inscrivent et sont soutenues par le programme de l'Institut Universitaire de France 'Complex Big Data Scene Analysis' que H. Glotin pilote.

Le/la MC renforcera les collaborations nationales actuelles avec ENS Paris (Mallat), UPMC LIP6 et LAM, LIA, LIPADE, LIPN, IFSTTAR ; et internationales avec les universités de New-York / FaceBook (LeCun), Cornell, Heifi, Beijing, Pavia, Queensland (Pr. Geoff. McLachLan) amplifiant ainsi le rayonnement du LSIS et de l'UTLN. DYNI a notamment une activité croissante avec la Chine via un projet NSFC avec Heifi univ 2014-2017, et actuellement des discussions avec Hong-Kong univ (PIC soumis). DYNI porte plusieurs collaborations via SABIOD en modèles de flux MD de plusieurs observatoires acoustiques terrestres en Europe, ou sous-marins à l'international : Nemo en Sicile, ONC au Canada, Baobab à Madagascar, Héracles en Nouvelle Calédonie, ou nationale avec Antares sur Toulon et la bouée instrumentée LSIS-MIO au sud Port-Cros. Enfin le MC participera à la thématique d'apprentissage de modèles statistiques notamment développée dans la proposition ANR LEMONS pour 2015-2018 retenue aux sélections ANR.

DYNI (co)organise un ou deux congrès internationaux par an (IJCNN 2012, IJCNN 2013, BIO2S2013, ICML13 Atlanta http://sabiod.org/icml4b (rang A+), NIPS4B13 Nevada http://sabiod.org/nips4b (rang A+), ICML4B2014, 9^{e} édition école internationale ERMITES,... OCEAN IEEE 2019). Actuellement ces actions sont en flux tendu et doivent être soutenues par cet MC complémentaire.

4. Besoin de renforcement / complément / remplacement

DYNI n'est pas en difficulté en terme scientifique, mais a besoin de compléter ses compétences dans ce domaine des MD qui est en pleine expansion afin d'assurer des avancements théoriques nécessaires pour conserver sa visibilité internationale.

Un support envisageable pour ce poste est le départ prochain en retraite de Dr. Gontard, MC 27 qui a été en lien avec le LSIS et Imath.

5. Recrutement fédérateur entre équipes intra ou inter-laboratoires

Cet MC sera potentiellement en collaboration avec plusieurs équipes du LSIS travaillant sur des modèles MD, notamment avec ESCODI partie UTLN sur des modèles bioinspirés (cf Colloque BIO2S 2013 coorganisé), et avec DIMAG sur les modèles probabilistes de fouille de MD (possibilité de modèles conjoints). Cet MC sera aussi en relation sur l'UTLN avec Imath en calcul sur MD, et pourrait également collaborer sur des projets de traitement de données de capteurs environnementaux, physiques ou autres, comme avec le MIO (BOMBYX) ou Protée.

Des collaborations sont naturelles dans la FRIIAM, avec le LIF et l'équipe QUARMA (ex codirection de thèse connexe) et l 'équipe Base de Données du LIF, également participante à un projet MASTODON du CNRS. Des collaborations sont envisagées avec I3S à Nice (Dr. Precioso).

6. Cohérence avec la politique d'axes transverses UTLN

Le poste demandé s'intègre naturellement dans les perspectives de rayonnement des Axes transversaux Mer et Information, et de son école doctorale Mer & Sciences, de part les développements de modèles mathématiques et d'algorithmes pour l'analyse de MD en sciences de l'environnement, dont environnement maritime (acoustique notamment). Il s'insère en pointe du projet de calcul scientifique monté en coll. avec Imath et l'Axe Information. Des liens avec des équipes de l'IUT sont également établis sur le volet capteurs distribués et données en ligne.

7. Besoins particuliers d'encadrement en master ou doctorat

Outre son implication en licence SI, cet MC sera actif dans les modules du master informatique DAPM, notamment sur les problématiques traitement de données (Traitement données multimodales, Traitement de la parole,modèles de données), et sur des modules à prévoir en sciences des données: Base de données massives, Calcul en masse, Calcul distribué, etc. Des liens seraient possibles avec d'autres masters : Vision, Sciences Physiques / Environnement, master biologie, pour leurs aspects traitements de données scientifiques (audiovisuelles, acoustique, …).