La Rochelle Université

D R P I
Direction Recherche
Partenariats Innovation

SUJET

SUJET	
Laboratoire L3i	**École doctorale** S2iM

Sujet de thèse

Intitulé scientifique

Fouille, structuration et navigation interactive dans des corpus hétérogènes pour aider la mise en relation de données métiers

Intitulé vulgarisé *(explicite pour un non spécialiste)*

Rassembler, corréler et visualiser l'ensemble des données d'un domaine ou d'une personne hébergées dans les différents systèmes d'informations d'administrations

Direction de la thèse *(identité du·de la·des directeur·trice·s, grade, HDR)*

Mickaël Coustaty

Priorité scientifique :

Préciser la transition dans laquelle le projet s'inscrit (explication argumentée) ou sa dimension interdisciplinaire

Le sujet de thèse présenté s'intègre dans la priorité scientifique « Transition Numérique » et des liens seront réalisés avec le CEJEP sur tout ce qui a trait à la gestion et la protection des données utilisateurs.

Descriptif du sujet

Éléments d'explication du sujet (enjeux scientifiques, applicatifs, sociétaux…).

De nombreuses organisations publiques ou privées sont confrontées à l'éparpillement de leurs données (papiers et numériques), la présence multiple d'un document sous différentes formes et le besoin de retrouver l'information sous toutes ces formes. En effet, ces différents types de contenus très hétérogènes ne permettent pas d'obtenir une cohérence globale et d'uniformiser la manière de les représenter et de rechercher de l'information. Un exemple de telles données sont les contenus liés à l'activité de l'Université de La Rochelle (convention de stage avec les sujets, termes des contrats de recherche, site web de l'offre de formation et des laboratoires, noms des vacataires, entreprises d'origine, enseignements assurés par les vacataires). Ces données représentent un potentiel d'information riche qui nécessiterait d'être extraites, structurées et agrégées pour renseigner ses partenaires académiques et socio-économiques du territoire. Un autre exemple pourrait consister à analyser l'ensemble des contenus manipulés par une mairie afin de lui permettre de retrouver les informations concernant une personne ou une entité, et proposer des services innovant de recherche et de visualisation de ces contenus agrégés. Ce deuxième exemple rentre en résonance avec le règlement (UE) 2016/679 du Parlement européen relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, obligera à partir de mai 2018 tout organisme public à fournir la liste des données détenues sur un tiers par celui-ci s'il en fait la demande.

Bien que de nombreux outils existent pour retrouver une information exacte et structurée, peu sont capables de traiter des contenus hétérogènes (documents papiers, numériques) issus de sources variées (bases de données métiers, relationnelles, NoSQL…). Ainsi, dans cette thèse, nous souhaitons étudier la combinaison de mécanismes d'information spotting et d'information retrieval afin de proposer un mécanisme interactif pour rechercher et visualiser de l'information. L'idée sera alors d'extraire automatiquement de l'information à partir des contenus présents dans les systèmes d'information (scan de documents, informations structurées), de l'organiser au sein d'une structure de données complexe (graphe ou hypergraphe) qui représentera les différents types de liens qui peuvent exister entre des données (même type d'information, données concernant une même entité) et de calculer des clusters de données proches spatialement ou sémantiquement. Enfin, des outils de visualisation interactifs seront

testés afin de comprendre comment l'utilisateur interagit avec le système et ainsi proposer de nouvelles méthodes pour réorganiser l'espace de recherche.

Le propre de ce sujet repose dans le fait qu'il se situe à l'interface de deux domaines de recherche : la reconnaissance, l'interprétation, et l'indexation de contenus numériques d'une part, et l'étude des graphes de terrain, c'est-à-dire de réseaux réels modélisables par des graphes d'autre part. C'est donc le travail à l'interface de ces deux domaines qui est visé avec la volonté de proposer de nouvelles méthodes de structuration et d'indexation des contenus à partir des méthodes utilisées sur des grands graphes (détection de communauté, de sous-graphes denses) et enrichir les méthodes développées en analyse de graphe afin de les enrichir avec les informations et les caractéristiques usuelles utilisées en analyse de documents et de contenus numériques. Les verrous scientifiques se situent donc dans chacun de ces domaines et à l'interface en mélangeant ces approches.

Analyse de contenus numériques : Les travaux les plus récents en analyse de documents visent à proposer des techniques d'information spotting qui consistent à retrouver des contenus similaires sans les reconnaître [2], et à venir créer des liens entre des contenus textuels et des représentations images [1]. Cela consiste donc à extraire des entités types à partir d'un corpus significatif de données (textes ou images). Une question importante qui n'est pour le moment que peu adressée pourrait consister à proposer un espace de représentation commun, entre des éléments textuels et des éléments images, comme proposé dans [3,4]. Cet espace devant permettre de rapprocher des contenus similaires issus de documents nativement numériques ou de contenus dématérialisés à l'aide de métriques usuelles. L'utilisation de méthodes à base de réseaux profonds pourra être également envisagée.
Analyse de réseaux d'information et de graphes : Une fois ces contenus résumés sous forme d'entités types et de leurs représentations vectorielles, des liens seront proposés entre les contenus les plus proches afin de construire un réseau d'information complexe. L'étude de ces réseaux consiste à extraire des informations complexes implicites (liens entre ces sources, détection de communauté ou de cluster dans des réseaux). Si les approches classiques de clustering de graphes ne sont pas utilisables directement pour calculer des communautés dynamiques, des approches de clustering consensuel peuvent être envisagées [5]. En particulier, des systèmes récents proposent de détecter des communautés (qui pourraient représenter des ensembles cohérents de données) à partir de recherche de similarité entre des nœuds basée sur la propagation des labels, en temps réel et dans un contexte big data [6,7].

[1] Jon Almazán, Albert Gordo, Alicia Fornés, Ernest Valveny: Word Spotting and Recognition with Embedded Attributes. IEEE Trans. Pattern Anal. Mach. Intell. 36(12): 2552-2566 (2014)
[2] Jon Almazán, Albert Gordo, Alicia Fornés, Ernest Valveny: Segmentation-free word spotting with exemplar SVMs. Pattern Recognition 47(12): 3967-3978 (2014)
[3] David Aldavert, Marçal Rusiñol, Ricardo Toledo, Josep Lladós: A study of Bag-of-Visual-Words representations for handwritten keyword spotting. IJDAR 18(3): 223-234 (2015)
[4] Nhu-Van Nguyen, Mickaël Coustaty, Jean-Marc Ogier: Multi-modal and Cross-Modal for Lecture Videos Retrieval. ICPR 2014: 2667-2672
[5] Stable community cores in complex networks. Massoud Seifi, Jean-Loup Guillaume, Ivan Junier, Jean-Baptiste Rouquier, Svilen Iskrov. 3rd Workshop on Complex Networks (CompleNet 2012), Floride
[6] Qi Song, Bo Li, Weiren Yu, Jianxin Li, Bin Shi: NSLPA: A Node Similarity Based Label Propagation Algorithm for Real-Time Community Detection. UCC 2014: 896-901
[7] Qi Song, Yinghui Wu, Xin Luna Dong: Mining Summaries for Knowledge Graph Search. ICDM 2016: 1215-1220

« Plus value » Etablissement :

Préciser l'ambition du sujet et ses retombées : impacts mesurables sur la dynamique du/des chercheurs, du laboratoire et de l'établissement

Ce sujet vise à développer un cadre générique d'analyse, de structuration et de visualisation de données hétérogènes. Ce cadre générique sera appliqué sur deux cas d'études qui apporteront une plus-value à l'établissement :

1. Construction d'objets visuels interactifs (graphes ou autres) permettant à une entreprise de pouvoir « fouiller » dans le patrimoine de l'ULR à partir de mot clefs, de concepts, ou de toute autre forme, dans un but de chercher une personne, une organisation, une formation, un laboratoire, qui pourrait répondre à une question posée par une entreprise
2. Détection et recherche de toutes les données associées à une personne dans le système d'informations hétérogènes de la ville de La Rochelle (Open Data ou non) afin de répondre aux besoins des collectivités locales en termes de législation sur la protection des données