PROPOSITION DE SUJET DE THESE

Campagne 2017

Laboratoire L3i

Sujet de la thèse :

Structuration unifiée de contenus hétérogènes pour proposer une fouille interactive

Résumé du travail proposé :

De nombreuses organisations publiques et privées sont confrontées à l'éparpillement de leurs données, à la présence multiple de documents sous différentes formes et au besoin de retrouver des informations dans cette masse peu ou pas organisée. Notamment, 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 qui en ferait 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, l'objet de cette thèse est d'analyser des contenus hétérogènes pour en extraire une information enrichie sémantiquement, de les indexer et de proposer des outils de visualisation et de navigation interactifs hétérogènes.

Mots clés :

Big data / Structuration et indexation de contenus hétérogènes / Analyse de documents / Détection de communautés / Visualisation interactive des contenus

Informations complémentaires :

Encadrant(s) :

- Mickaël Coustaty (directeur de thèse)
- Accompagnement scientifique réalisé par : Jean-Loup Guillaume et Jean-Marc Ogier

Equipe :

- ☒ **Images et Contenus**
- ☐ **Dynamique des systèmes et adaptativité**
- ☒ **Modèle et Connaissance**

Domaine d'application stratégique :

- ☐ **E-éducation**
- ☐ **Environnement et développement durable**
- ☐ **E-culture**
- ☒ **Valorisation de contenus numériques**

Date de début du contrat : Dernier trimestre 2017
Durée du contrat : 3 ans

Contexte de l'étude:

Les données papiers et numériques produites par les grandes institutions publiques ou privées intègrent différents types de contenus très hétérogènes dont la cohérence globale est difficile à appréhender. Un exemple de telles données sont les contenus liés à l'activité de l'Université de La Rochelle : celles-ci représentent un potentiel d'information riche qui nécessiterait d'être extraites, structurées et agrégées pour renseigner ses partenaires (convention de stage avec les sujets, termes des contrats de recherche, site web de l'offre de formation et des laboratoires, noms des vacataires et enseignements assurés, entreprises d'origine, etc.). 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.

Dans cette thèse, nous souhaitons étudier la combinaison de mécanismes d'*information spotting* et d'*information retrieval* afin de proposer des solutions pour rechercher et visualiser de l'information de manière interactive. Le principe consiste, dans un premier temps, à extraire automatiquement de l'information à partir des contenus présents dans les systèmes d'information (scan de documents, informations structurées et non structurées), de l'organiser au sein d'une structure de données complexe (tel que des graphes ou des hypergraphes) qui représentera les différents types de liens qui peuvent exister entre des données (même type d'information, données concernant une même entité, etc.) et de calculer des clusters de données proches spatialement ou sémantiquement. Enfin, des outils de visualisation et de navigation interactifs seront testés afin d'aider l'utilisateur à interagir avec le système mais également de comprendre ces interactions afin de pouvoir proposer de nouvelles méthodes pour réorganiser l'espace de recherche.

Description du sujet :

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 l'interface de ces deux domaines qui est ciblée 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 communautés, de sous-graphes denses) et enrichir les méthodes développées en analyse de graphes 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 s'intéressent à l'*information spotting* qui consiste à retrouver des contenus similaires sans les reconnaître [2], et essayer de créer des liens entre des contenus textuels et des représentations images [1] en plongeant leurs descriptions dans un espace de représentation commun. Cela consiste donc à extraire des entités types à partir d'un corpus significatif de données (textes ou images) et trouver un espace de représentation hybride entre texte et image. La question majeure

qui n'est pour le moment que peu adressée consiste à proposer un espace de représentation commun, entre des éléments textuels et des éléments images, comme c'est le cas dans [3,4]. Cet espace doit 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 [8].

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 ensuite 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 dans des graphes multiplexes (graphes avec plusieurs couches de différents niveaux sémantiques), des approches de clustering consensuel, naturellement plus stables, 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 similarité entre des nœuds basée sur la propagation des labels, en temps réel et dans un contexte big data [6,7].

Références bibliographiques :

[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

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Contacts

Email(s) : mickael.coustaty@univ-lr.fr