Sujet de thèse CIFRE : ERIC Lab - AMI Software

Contexte général de la collaboration

Cette proposition se place dans le cadre d'une collaboration entre la société AMI Software (Montpellier) et le laboratoire ERIC (Université de Lyon). Une première thèse CIFRE, entre les deux partenaires, portant sur l'analyse des opinions dans les médias sociaux vient d'être soutenue avec succès en juin 2015. Les deux partenaires collaborent par ailleurs dans le cadre du projet ANR ImagiWeb qui vise à analyser les opinions véhiculées sur Internet.

Contexte applicatif

Dans un cadre de veille sur Internet en général, et devant la masse importante de données que constitue aujourd'hui notre environnement numérique, il est primordial de se doter d'outils d'analyse performants permettant d'extraire une synthèse réduite de l'information clef, en particulier directement exploitable par les décideurs. Une telle synthèse se doit de mettre en évidence les tendances fortes ou les évènements marquants de l'environnement, mais aussi de déceler les phénomènes marginaux, éventuellement annonciateurs de changements à venir ou constituant les prémisses d'une crise potentielle. La thèse doit permettre d'avancer sur l'élaboration d'une telle synthèse, abordée sous l'angle des sujets abordés (par ex. des thématiques ou des événements) que le décideur désire suivre dans l'espace immense qu'est Internet. Pour cela, l'entreprise met à la disposition du doctorant son logiciel de veille qui lui permettra d'avoir un accès privilégié à un grand volume de données. Il est prévu que les algorithmes développés par celui-ci soient régulièrement intégrés à la solution « Lab » du logiciel distribué par AMI Software.

Contexte scientifique

Identifier, suivre et visualiser la manière dont l'information se distribue, voire circule, sur Internet est un verrou scientifique important qui nécessite de mobiliser de nombreux domaines de recherche : fouille de données, recherche d'information, apprentissage automatique, analyse des systèmes complexes, traitement automatique du langage naturel, visualisation. La tâche est rendue d'autant plus difficile qu'il s'agit de s'attaquer à des données complexes (données interconnectées, hétérogènes, évolutives) et volumineuses. Parmi les travaux de l'état de l'art, on peut citer le suivi des *mèmes* basé sur l'étude des citations réalisé par J. Leskovec [LES 2009] ou ses travaux plus récents sur les chemins de l'information [GOM 2013]. Des modèles de diffusion adaptés aux données issues du Web, tels que ceux inspirés du marketing viral ou de l'épidémiologie, ont également été proposés pour tenter de résoudre cette problématique [BAK 2012]. Dans le cadre de la thèse d'A. Lauf en collaboration avec l'Institut National des Langues et Civilisations Orientales, la société AMI a travaillé sur la modélisation et l'identification des buzz. Ils ont ainsi montré qu'il était possible de fournir une vision microscopique des divers thèmes abordés sur un ensemble de sites Internet ciblés [LAU 2014]. Du côté du laboratoire ERIC, de nouveaux modèles de diffusion adaptés aux réseaux sociaux ont été proposés [GUI 2012]. Enfin, la thèse de M. Dermouche, en commun entre AMI Software et ERIC, a permis de proposer un modèle original pour résumer l'évolution conjointe des thématiques et des opinions exprimées dans des corpus textuels [DER 2014]. Ces différents travaux entrepris par les partenaires serviront de base de départ à la future thèse.

Objectifs de la thèse

L'objectif de la thèse proposée est de travailler sur l'identification des trajectoires de l'information et de comprendre les mécanismes qui régissent ou influencent ces trajectoires. C'est également l'occasion de consolider les travaux déjà réalisés sur la problématique du « buzz » et de voir comment elle s'inscrit dans les trajectoires étudiées. La thèse est associée à un certain nombre de verrous scientifiques que nous souhaitons aborder ensemble, à savoir : a) prendre en compte un nombre important de sites Web variés, b) étudier la trajectoire d'un sujet (ex. : une thématique, un événement) à la fois dans le temps et dans l'espace, espace qu'il sera nécessaire de reconstruire, c) étudier l'influence que le type de sujet, l'opinion qui est exprimée, la présence d'acteurs au rôle particulier (ex. : influenceurs) peuvent avoir sur les trajectoires. Pour traiter le verrou (a), le doctorant pourra tirer profit de l'outil de veille performant mis à disposition par AMI Software qui lui donnera accès à un nombre important de sources de natures variées : sites Web, blogs, tweets, etc. Pour le verrou (b), une piste très prometteuse inspirée par [BOU 2014] consiste à reconstruire la topologie des données, nécessairement lacunaires, à partir des observations réalisées et des quelques informations topologiques à notre disposition (distance entre les sources d'information, similarité entre les auteurs de messages, proximités temporelles, etc.). Le verrou (c) pourra tirer partie des travaux réalisés dans le cadre des thèses de M. Dermouche et A. Lauf.

Par ailleurs, un verrou supplémentaire à considérer, lié notamment au cadre applicatif de la thèse, concerne l'élaboration d'approches et d'outils de visualisation permettant de tirer profit des algorithmes développés. Il s'agit de rechercher et valider des approches innovantes et efficaces, issues du domaine émergent de la visualisation d'information. Cet aspect constitue un point critique conditionnant l'adoption et l'acceptation des nouveaux outils avancés mis à disposition des utilisateurs.

Méthodologie

Il est prévu d'articuler la thèse en trois parties :

a) La première partie consiste, au travers de plusieurs cas d'étude sur divers types de données réelles obtenues à l'aide de la plateforme d'AMI, à visualiser la trajectoire observée d'un nombre déterminé de sujets et/ou événements. En particulier, le doctorant pourra partir des travaux de Leskovec sur les citations et voir comment l'étendre avec d'autres types de sujets qui se répètent sur Internet, qu'il s'agisse de mots clefs, de motifs plus évolués (expressions, phrases) ou de thématiques. Cet aspect exploratoire pourra tirer profit des outils existants développés par AMI pour extraire et localiser des motifs récurrents extraits des documents (tels qu'un arbre de suffixe). Cette partie appliquée sera réalisée conjointement à un état de l'art qui se focalisera sur les trajectoires (spatiales et temporelles) que peuvent emprunter les sujets traités sur Internet (du moins la partie accessible par les moteurs de recherche analysés par la solution d'AMI Software), notamment lorsque l'information qui concerne la diffusion observée est incomplète.

b) La deuxième partie consiste à décrire puis à caractériser les différentes trajectoires observées. Il s'agit d'inférer le chemin de propagation de l'information le plus vraisemblable au vu des observations effectuées. L'un des problèmes auxquels sera confronté le doctorant est le fait que, malgré l'utilisation d'un outil de veille puissant, il ne sera pas toujours possible de connaître l'ensemble des sites Web qui composent la trajectoire. Il sera donc nécessaire de reconstruire cette trajectoire dans un contexte d'information lacunaire. Une piste possible consiste à analyser les déformations

constatées sur l'information et à en déduire des proximités entre sources, favorisant ainsi le rapprochement *a posteriori* de ces sources qui expliquent l'émergence de certaines trajectoires qui peuvent sembler inattendues [GOM 2013][BOU 2014].

c) La troisième partie consiste à exploiter la typologie des trajectoires et la caractérisation produite à l'étape précédente pour voir s'il est possible d'inférer sur de nouvelles observations. Cela peut permettre par exemple de réaliser des prédictions à court terme ou de retrouver l'origine la plus probable d'un « buzz ». Dans cette dernière partie, le doctorant devra mettre au point un outil de visualisation adaptée à la problématique. Cet outil permettra par exemple de mettre en valeur les sources les plus critiques à surveiller ou de naviguer dans la carte représentant la trajectoire de propagation d'un sujet donné.

Références

[BAK 2012] The role of social networks in information diffusion. E. Bakshy, I. Rosenn, C. Marlow and L.A. Adamic. Proceedings of WWW 2012, pp. 519-528.

[BOU 2014] Learning Social Network Embeddings for Predicting Information Diffusion. S. Bourigault, C. Lagnier, S. Lamprier, L. Denoyer, P. Gallinari. Proceedings of WSDM, 2014, ACM.

[DER 2014] A Joint Model for Topic-Sentiment Evolution over Time. Mohamed Dermouche, Julien Velcin, Leila Khouas, Sabine Loudcher. Proceedings of ICDM 2014, pp. 773-778.

[GUI 2012] A predictive model for the temporal dynamics of information diffusion in online social networks. A. Guille, H. Hacid. WWW 2012, Workshop on Mining Social Networks Dynamics, pp. 1145-1152.

[GOM 2013] Structure and Dynamics of Information Pathways in Online Media. M. Gomez-Rodriguez, J. Leskovec, B. Schoelkopf, Proceedings of WSDM, 2013, ACM.

[LAU 2014] Propagation du buzz sur le Web–identification, analyse, modélisation et représentation dans un contexte de veille. Aurélien LAUF. Thèse dirigée par Monique Slodzian et Mathieu Valette préparée à l'INALCO, en collaboration avec l'entreprise AMI Software, thèse soutenue en 2014.

[LES 2009] Meme-tracking and the dynamics of the news cycle. J. Leskovec, L. Backstrom, J. Kleinberg. Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining, 2009.

[SNO 2011] Refining causality: who copied from whom? T. Snowsill, N. Fyson, T. De Bie, and N. Cristianini. Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining, 2011.