Proposition de sujet de stage de M2

17/11/2014

Sujet : Extraction d'évènements à partir du Web : application à l'analyse d'agendas électroniques de personnalités publiques

Encadrants : Vincent Labatut & Guillaume Marrel

Objectif

Le but de ce stage est d'étudier dans quelle mesure il est possible de confronter les évènements inscrits dans l'agenda personnel d'une personnalité publique (personnalités médiatiques et/ou politiques…) avec ce qui en est dit sur le Web. À cette fin, nous proposons dans un premier temps de considérer à la fois l'agenda et le contenu disponible en ligne sous la forme de séquences d'évènements spatio-temporels, qui seront extraits au moyen de méthodes issues de la fouille de texte. La confrontation entre l'information planifiée dans l'agenda et celle que restitue le Web se fera alors par comparaison de ces évènements.

Contexte

L'activité d'un élu, par exemple, est plus ou moins commentée en ligne par les professionnels de la politique et de l'information, mais aussi de plus en plus par des citoyens ordinaires , sur les pages institutionnelles, les blogs, les murs Facebook, Twitter et autres réseaux sociaux. Le Web peut ainsi être considéré comme un miroir déformant de l'activité politique.

La maîtrise de cette déformation médiatique devient un enjeu stratégique de premier ordre avec l'accroissement de la dimension conversationnelle du Web : ceci change le rapport des leaders aux instruments de contrôle de leur image, à tel point que l'e-réputation fait aujourd'hui l'objet de conseils spécialisés. L'objectif de la recherche est de parvenir à objectiver ces déformations pour ensuite les interpréter et, à terme, dégager les variables de l'écho web-médiatique du travail politique dans les années 2010.

Données

Des tests seront menés sur un corpus fourni par Guillaume Marrel dans le cadre du projet Tr@nsPolo. Il s'agit d'un fichier au format iCalendar (`.ics`) issu d'un logiciel de gestion du temps (Outlook, Google calendar...) provenant d'un élu local volontaire. La période étudiée devra être la plus récente possible. Les tests pourront être effectués sur quelques semaines d'emploi du temps.

Ce type de fichier exporte une base de données comportant pour chaque événement planifié (rendez-vous, réunion, cérémonie, déplacement, etc...), un objet, une date, des horaires de début et de fin, un lieu et éventuellement des compléments d'information sur les participants à l'interaction. Plusieurs types d'événements pourront être pris en compte : déplacements, visites, rencontres, etc.

Méthode

On propose de traiter les agendas comme des séquences d'évènements, chacun étant représenté par un nombre limité d'attributs considérés comme pertinents parmi ceux disponibles dans les données fournies. On pourra par exemple se concentrer dans un premier temps sur les participants (qui ?), le lieu (où ?), le temps (quand ?) et l'objet (quoi ?). Le problème d'identifier un évènement dans un texte peut alors se ramener à une détection d'entités nommées (qui, où, quoi ?) et numériques (quand ?). Une partie du travail consiste donc à mettre en place un outil capable de réaliser ce traitement, ce qui nécessite un travail préliminaire de revue (de nombreux outils de ce type existent déjà). À noter qu'une plate-forme déjà définie pour un autre projet (Atdağ & Labatut 2014) peut être utilisée comme base de travail.

Une fois les entités détectées dans l'agenda de référence et dans les textes issus du Web, il est nécessaire d'identifier quels textes sont pertinents, et quels évènements de référence ils relatent. Un travail bibliographique permettra de déterminer si une mesure appropriée à cet usage a déjà été décrite dans la littérature (et peut donc être réutilisée) ou bien si une nouvelle mesure doit être conçue.

L'obtention des textes issus du Web constitue aussi un problème en soi. Plusieurs approches peuvent être adoptées ici. Dans un premier temps, on peut envisager de se concentrer sur certains sites jugés

pertinents (par exemples des journaux locaux) et pour lesquels des méthodes d'extraction du texte ont été définies manuellement. Cependant, il serait plus intéressant, du point de vue applicatif, de pouvoir accéder à l'ensemble du Web, et donc de définir une méthode automatique d'extraction générale. Une approche à base de règle semble possible, qui consisterait à identifier différentes structures de page types, ou bien une approche heuristique (par exemple, se concentrer sur l'élément HTML contenant le plus de texte dans la page), voire une approche reposant sur l'apprentissage automatique.

Le dernier point méthodologique concerne l'évaluation des résultats obtenus. En s'inspirant des travaux existant en recherche d'information, il faudra définir une approche adaptée au cas présent et consistante avec les besoins des chercheurs de sciences politiques. Ceci passe notamment pas l'affichage des résultats sous forme lisible et pratique, par exemple via la génération de rapports prenant la forme de pages HTML (en complément de documents plus classiques tels que des tables de mesures au format CSV).

Déroulement

Les différentes étapes du travail pourraient être les suivantes (cette organisation évoluera éventuellement en fonction du déroulement du projet) :

1. Analyser les besoins des chercheurs en sciences politiques, s'imprégner des données.
2. Développer le module capable d'extraire les entités de l'agenda et d'un texte quelconque.
3. Développer le module capable d'extraire le texte contenu dans des pages Web.
4. Définir et implémenter la mesure permettant de comparer les évènements.
5. Développer le module générant des rapports présentant les résultats obtenus.
6. Vérifier l'adéquation de l'outil produit avec les besoins des chercheurs en sciences politiques.

Production

Ce stage devra déboucher sur la réalisation d'un outil opérationnel, stable et réutilisable dans la perspective d'une poursuite de ce projet de recherche. Cela signifie en particulier que le code source rendu devra être correctement commenté, et qu'un document technique devra être rédigé pour expliquer comment installer et utiliser le logiciel. À noter qu'une partie de ce texte apparaîtra vraisemblablement aussi dans le mémoire lui-même.

Aspects pratiques

Durée : 6 mois (printemps-été 2015).

Indemnité : environ 436,05€/mois.

Lieu : Laboratoire Informatique d'Avignon (LIA), dépendant de l'UAPV, et situé sur le technopole Agroparc.

Formation requise : étudiant en informatique, en M2 ou dernière année de formation d'ingénieur.

Compétences demandées : programmation Java, notions de fouille de texte et/ou de fouille Web, de recherche d'information, volonté de travailler sur un projet pluridisciplinaire

Modalités de candidature : envoyez les pièces suivantes (format PDF) aux deux encadrants :

- CV ;
- Lettre de motivation relative au sujet ;
- Lettre(s) de recommandation ;
- Relevés de notes des deux dernières années (incluant les enseignements suivis) ;
- Tout autre document jugé utile.

Références

Samet Atdağ & Vincent Labatut, "A Comparison of Named Entity Recognition Tools Applied to Biographical Texts", 2nd International Conference on Systems and Computer Science, IEEE Press, p.228-233, 2013.

Guillaume Marrel & Laurent Godmer, "Que font vraiment les professionnels de la politique ? L'agenda électronique et l'emploi du temps d'une élue régionale", in Alice Mazeau (dir), *La représentation politique en pratiques*, Rennes, PUR, 2014.

Guillaume Marrel & Laurent Godmer, "La production de l'agenda. Comment se fabrique l'emploi du temps d'une vice-présidente de conseil régional", in Demazière Didier, Le Lidec Patrick (dir.), *Les mondes du travail politique. Les élus et leurs entourages*, Rennes, PUR, p.37-52, 2014.