Sujet de stage de M2

17/12/2014

Laboratoire Informatique d'Avignon

Sujet : Extraction de réseaux sociaux implicites à partir de textes

Mots-clés : Fouille de texte, Identification d'entités nommées, Comparaison d'évènements

Encadrants : Vincent Labatut (MCF - LIA) & Guillaume Marrel (MCF - LBNC)

1. Objectif

Le but de ce projet est de développer une méthode d'extraction d'information permettant de traiter un corpus de textes afin d'identifier les interactions sociales qui y sont implicitement décrites, et de produire un graphe en donnant une représentation explicite.

En d'autres termes, on veut donner une structure (le graphe) à des données non-structurées (les textes). L'intérêt du graphe est qu'il peut ensuite être utilisé comme support pour de multiples analyses, telles que celles issues du domaine des réseaux complexes [1, 2] : identification de nœuds centraux, détection de communautés, etc.

L'outil a vocation à être utilisé dans un contexte de sciences humaines et sociales, puisqu'il cible l'étude d'interactions sociales contemporaines ou historiques. On se propose de l'appliquer dans un premier temps à des données ciblant des personnalités publiques.

2. Contexte

L'analyse de *réseaux complexes* est un domaine inter-disciplinaire relativement récent [3, 4] visant à étudier des systèmes complexes du monde réel en les modélisant sous forme de graphes. La plupart des travaux existants se concentrent sur des données structurées, généralement sous forme tabulaire (bases de données relationnelles).

Les données textuelles, qui sont considérées comme non-structurées, sont pour leur part relativement ignorées, car leur traitement est plus difficile. Or, elles constituent une part non-négligeable des données existantes et produites par le passé et aujourd'hui, ne serait-ce que parce que le Web est essentiellement textuel. À ce titre, leur exploitation constitue un enjeu important à la fois pour l'informatique et pour les sciences sociales. Le travail proposé ici a pour but de traiter ce problème, par l'extraction de données relationnelles à partir de textes biographiques ou d'échanges épistolaires.

3. Méthode

La méthode proposée consiste à considérer les textes biographiques ou épistolaires comme des séquences d'évènements. Ces évènements évoqués ou relatés sont décrits par certaines caractéristiques sélectionnés à l'avance, telles que : acteurs, lieu, date, objet, etc.

La première étape porte sur la détection de ce type d'évènement. Celle-ci peut être réalisée notamment via des outils de détection d'entités nommées. Pour simplifier le problème, on se concentrera dans un premier temps sur des textes issus de l'encyclopédie en ligne Wikipedia, car ils sont facilement accessibles et contiennent de nombreux hyperliens utilisables pour améliorer la détection d'entités. Cette partie du travail pourra bénéficier d'une plate-forme déjà développée lors d'un projet précédent [5]. On pourra ensuite étendre notre outil au traitement d'autres types de textes, voire d'autres langues.

Nous comptons identifier les interactions entre individus par recoupement des évènements identifiés, en partant de l'hypothèse que deux personnes qui ont participé au même évènement se connaissent probablement, et qu'une fréquence élevée de coparticipation augmente encore cette probabilité. La deuxième étape consiste donc à définir des méthodes de comparaison entre évènements, en tenant compte des différences de granularité, aussi bien spatiales que temporelles.

La dernière étape porte sur l'évaluation empirique de l'outil. La détection d'entités nommées est menée sur la base d'un corpus annoté, en comparant les entités estimée automatiquement par notre méthode à celles désignées manuellement par les annotateurs. La méthode d'évaluation de la deuxième tâche, en

revanche, reste à définir. Le niveau de pertinence du graphe d'interaction final ne peut être déterminé que par des experts du domaine étudié, d'où l'intérêt d'impliquer dans ce projet des historiens et des politistes.

4. Déroulement

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

1. Prendre en main l'outil de détection d'entités nommées.
2. Proposer et implémenter une méthode de comparaison d'évènements.
3. Concevoir et implémenter un module permettant d'explorer le corpus de textes.
4. Évaluer les performances de l'outil sur un corpus (Wikipedia dans un premier temps).

5. Production

Outil. Ce stage devra déboucher sur un outil opérationnel. Le but est ici d'implémenter une version basique de la chaîne de traitement complète, de manière à obtenir un programme fonctionnel à défaut d'être optimal. Une fois cet objectif atteint, l'outil pourra être amélioré point par point. L'outil livré devra être finalisé dans le sens où il sera complètement documenté (code source commenté, mode d'emploi), afin de pouvoir être utilisé par des non-spécialistes et servir de base à une poursuite du projet.

Application. Le bon fonctionnement de l'outil devra être montré en l'évaluant sur des données réelles. Nous projetons dans un premier temps de traiter des notices biographiques Wikipedia décrivant de personnalités publiques. Dans un second temps, un corpus de correspondances pourra également être traité.

6. 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.

7. Références

1. da Fontoura Costa, L., et al., *Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications.* Advances In Physics, 2011. **60**(3): p. 329-412.
2. da Fontoura Costa, L., et al., *Characterization of complex networks: A survey of measurements.* Advances in Physics, 2007. **56**(1): p. 167-242.
3. Watts, D. and S.H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature, 1998. **393**(6684): p. 409-410.
4. Barabási, A.-L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509.
5. Atdağ, S. and V. Labatut. *A Comparison of Named Entity Recognition Tools Applied to Biographical Texts*. in *2nd International Conference on Systems and Computer Science*. 2013.