RÉSOPO

Extraction de réseaux sociaux implicites à partir de notices biographiques d'acteurs politiques d'hier et d'aujourd'hui

Réponse AAP Agor@nTic 2015

Porteur :	Vincent LABATUT	MCF d'informatique (LIA)
Équipe :	Frédéric MONIER	PR d'histoire (Hemoc-CNE)
	Guillaume MARREL	MCF de science politique (LBNC)
Axes :	B / C / E	

1. Présentation

Le but de ce projet est de développer une méthode automatique d'extraction d'informations relationnelles dans un corpus de textes biographiques retraçant l'existence ou l'activité de personnalités publiques. Il s'agit d'identifier les interactions sociales qui y sont explicitement ou implicitement décrites, et de produire un graphe donnant une représentation explicite du réseau social ainsi observé.

L'intérêt du graphe en tant qu'outil de modélisation est qu'il peut ensuite être utilisé comme support pour de multiples analyses issues du domaine des réseaux complexes (da Fontoura Costa et al., 2011; da Fontoura Costa, Rodrigues, Travieso, & Villas Boas, 2007) (identification de nœuds centraux, détection de communautés, etc.), ou de méthodes plus spécifiquement développées en sciences humaines et sociales pour la visualisation et l'étude de réseaux relationnels historiques ou contemporains (Beaurepaire & Taurisson, 2003; Mercklé, 2011).

La méthode proposée consiste à considérer les textes biographiques comme des mises en récit de séquences d'évènements, caractérisés par certains aspects, tels que la date, le lieu, l'objet, les acteurs impliqués, etc. L'identification d'évènements peut être réalisée via des outils de détection d'entités nommées. Lorsqu'elles ne sont pas explicites, les relations entre plusieurs individus peuvent être identifiées par la co-participation à un même événement.

L'objectif est de reconstituer des réseaux sociaux complets ou bien des réseaux personnels (ou égocentrés). Le projet vise l'expérimentation de cette méthode sur deux corpus distincts de textes biographiques d'acteurs du monde politique : historique, le premier concerne l'entourage parlementaire national du dirigeant socialiste Léon Blum, entre 1919 et 1940 ; le second, contemporain, porte sur le réseau de pouvoir de l'actuel député-Président du Conseil régional de la Région PACA, Michel Vauzelle.

Dans sa phase expérimentale en 2015, le projet consiste concrètement en l'encadrement pluridisciplinaire d'un stage financé de Master 2 d'informatique, avec un partenaire du monde économique, l'invitation de deux chercheurs dans le cadre d'un séminaire de recherche et la présentation des premiers résultats à l'occasion d'une conférence en informatique ou en sciences sociales.

2. Description

Contexte. L'analyse de *réseaux complexes* est un domaine interdisciplinaire relativement récent (Barabási & Albert, 1999; Watts & Strogatz, 1998) visant à étudier des systèmes complexes du monde réel en les modélisant sous forme de graphes[1]. Il est largement basé sur les nombreux travaux antérieurs dédiés à l'analyse de réseaux sociaux.

Un réseau social peut être défini comme un « système de relations sociales plus ou moins dense, informel et durable, associant des personnes ayant un intérêt réciproque à agir de façon solidaire, sans pour autant

[1] Le terme graphe est utilisé ici dans son sens mathématique, i.e. un ensemble de sommets et d'arrêtes.

appartenir aux mêmes organisations ou aux mêmes sphères sociales » (Nay, 2011, p. 512). Au-delà des mécanismes prévus par le droit, le concept permet en sciences sociales de mettre en lumière des systèmes de solidarité interpersonnels autres que ceux que dessinent les présentations officielles des organisations, des professions, des sphères de décision etc. Fondés sur des relations d'affinité, de connaissance, de coopération, de formation commune, le réseau social sert à comprendr[2]e les « univers sociaux », les logiques du recrutement des acteurs, les mécanismes décisionnels, l'échange des ressources, les jeux de concurrence et la consolidation des systèmes de pouvoir et de domination.

Jusqu'à récemment, les travaux issus des sciences sociales traitaient des graphes relativement petits, de l'ordre de quelques personnes (ou dizaines de personnes), et leur étude se faisait essentiellement de façon manuelle (Padgett & Ansell, 1993). Ceci était dû principalement à la difficulté de collecter des données, puis de les traiter. L'informatisation de la société a eu pour conséquence, d'une part, de faciliter la mise à disposition et la collecte de données, et d'autre part, d'accroître le pouvoir computationnel des chercheurs. Les graphes étudiés récemment sont beaucoup plus grands, de l'ordre de milliers, voire millions, de personnes (Ahn, Han, Kwak, Moon, & Jeong, 2007), et nécessitent des techniques spéciales d'analyse, qui ont forcément recours à l'outil informatique (Newman, 2003).

Mais avant de pouvoir étudier un graphe, il est nécessaire de construire cet objet mathématique, à partir des données empiriques disponibles. 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 beaucoup plus difficile. Or, elles constituent une part essentielle des données biographiques existantes et produites aujourd'hui, ne serait-ce que parce que le Web est essentiellement textuel. Et à ce titre, leur exploitation constitue un enjeu important.

But. Le but de ce projet est de définir, dans le cadre d'un stage de Master 2 d'informatique, une méthode d'extraction d'information permettant d'exploiter cette masse de données textuelles. Nous nous concentrons sur un type de texte en particulier : les *notices biographiques*. Nous voulons concevoir un outil capable de traiter une collection de biographies décrivant une population donnée, et d'y identifier des interactions implicites entre les individus constituant cette population. Le terme *implicite* signifie ici qu'une interaction n'est pas nécessairement mentionnée directement dans le texte, mais que son existence peut être inférée à partir de l'analyse de la collection.

Méthode. Nous proposons d'utiliser une *approche évènementielle* pour parvenir à nos fins. En effet, un texte biographique peut être considéré comme une séquence d'évènements, exprimés textuellement sous la forme de phrases du type "Monsieur X a réalisé la tâche Y le 1er janvier 2015 à Avignon". Dans le cas de phrases faisant mention de plusieurs personnes simultanément ("Messieurs X et Z ont étudié à l'UAPV"), l'interaction est explicitement formulée, ce qui permet d'inférer directement la présence d'arrêtes entre les sommets qui les représentent dans le graphe d'interaction que l'on veut extraire. Lorsque la phrase ne mentionne qu'une seule personne, le traitement est moins direct. On identifie d'abord tous les évènements décrits dans le corpus, puis on les recoupe de manière à identifier les personnes qui ont participé aux mêmes évènements. Notre méthode se base sur l'hypothèse que de tels co-participants se connaissent probablement, dans une mesure qui dépend de la nature et de la fréquence des évènements concernés.

Du point de vue technique, notre première tâche consistera à identifier les évènements dans les notices biographiques. Les outils de détection d'entités semblent particulièrement adaptés à cet objectif, car ils permettent de détecter les noms propres (lieux, personnes...) sous forme d'entités *nommées* et les dates sous forme d'entités *alphanumériques*. Ils ont aussi l'avantage d'être relativement indépendant de la langue traitée (Français, Anglais...). Deux démarches distinctes correspondent aux deux corpus de sources testés :

1. Corpus de textes biographiques historiques numérisés pour le réseau parlementaire de Léon Blum (1919-1940). Deux sources principales seront ici exploitées : 1) les notices biographiques des parlementaires de la Troisième République accessibles en ligne sur le site de l'assemblée nationale à partir de la base Sycomore[2] et du Sénat[3] ; 2) les notices biographiques compilées dans le *Dictionnaire*

[2] http://www.assemblee-nationale.fr/sycomore/
[3] http://www.senat.fr/senateurs-3eme-republique/index.html

biographique du mouvement ouvrier, Le Maitron, accessible en ligne avec abonnement[4] . Pour l'histoire politique, il s'agit là d'identifier par hypothèse le groupe des parlementaires socialistes de la période comme un réseau complet et d'y déceler les connexions explicites et implicites afin d'y situer la figure du dirigeant Léon Blum. L'objectif est de tester l'hypothèse de la construction de l'autorité du leader du Front Populaire au sein du groupe parlementaire socialiste (Hohl, 2007), alors même que le parti reste hostile à la personnalisation et peu favorable à l'émergence d'une figure charismatique.

2. Corpus de notices biographiques Wikipedia contemporaines pour le réseau notabiliaire de Michel Vauzelle. Il s'agit ici d'explorer le corpus de sources singulier, coopératif, presque illimité et sans cesse renouvelé, auquel correspond le projet lancé en 2001 de l'encyclopédie universelle et multilingue en ligne *Wikipedia.org,* pour l'extraction d'un réseau personnel de pouvoir plus ou moins implicite (Laurent, 2012). L'objectif est d'analyser la structure du réseau politique d'un leadership régional, sa dimension, les ressources échangées, leurs usages (Garrote, 2013). L'exploration progressera de la notice personnelle de Michel Vauzelle[5] vers les autres notices personnelles ou institutionnelles auxquelles les hyperliens de la première renvoient. Cette partie du travail pourra bénéficier d'une plateforme déjà développée lors d'un projet précédent (Atdağ & Labatut, 2013). On pourra ensuite étendre l'outil au traitement d'autres textes, voire d'autres langues.

La deuxième tâche consistera à analyser les évènements identifiés dans le corpus, afin de construire le graphe d'interaction, qui constitue notre résultat final. Il sera nécessaire de définir des méthodes de comparaison entre évènements, en tenant compte des différences de granularité, aussi bien spatiales que temporelles. Dans le graphe d'interaction, la création d'une arrête entre les sommets représentant deux personnes dépendra à la fois du nombre d'évènements auxquels elles ont co-participé, mais aussi de la nature de ces évènements, dans une mesure qui sera à déterminer.

La dernière étape renvoie à l'évaluation empirique de l'outil. La détection d'entités nommées est menée sur la base d'un corpus annoté, en comparant les entités estimées 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 les experts des domaines étudiés, le premier en histoire politique de la Troisième République, le second en sociologie des trajectoires et des carrières politiques locales.

3. Objectifs et résultats attendus

Ce projet doit être considéré comme une étude de faisabilité. Notre premier objectif est de développer un outil complet implémentant le traitement basique des différentes étapes décrites en section 2. Si les résultats sont concluants, un projet de plus grande envergure sera mis en place pour étendre et améliorer l'outil, point par point.

Si les méthodes implémentées seront dans un premier temps basiques, l'outil livré devra néanmoins 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.

4. Apports

En informatique, l'extraction de réseaux complexes à partir de données textuelles est un domaine très peu exploré. Les quelques travaux existants diffèrent sur le type de réseau traité et/ou sur la méthode utilisée pour l'extraire du texte. Ainsi, certains travaux s'intéressent aux réseaux sémantiques, qui visent à représenter les relations entre des concepts trouvés dans les textes (Carley, Columbus, & Landwehr, 2013; Kok & Domingos, 2008; Reinanda, Utama, Steijlen, & de Rijke, 2013), et non pas des interactions entre individus. D'autres travaux portent sur les réseaux d'interactions sociales, mais ce limitent aux relations explicitement décrites dans le texte (Agarwal, Kotalwar, & Rambow, 2013; Elson, Dames, & McKeown, 2010)

[4] http://maitron-en-ligne.univ-paris1.fr/

[5] http://fr.wikipedia.org/wiki/Michel_Vauzelle

ou dans sa structure (messages de forums, emails échangés, etc.) (Agarwal, Rambow, & Passonneau, 2010; Hassan, Abu-Jbara, & Radev, 2012).

L'apport de la méthode que nous proposons est d'extraire de l'information non seulement de chaque texte pris individuellement, via la détection d'entités nommées, mais aussi de la collection elle-même, à travers la comparaison d'évènements identifiés dans des textes distincts. Ceci nous permet de nous démarquer des approches existantes, qui reposent soit sur des relations décrites explicitement, soit sur des méthodes basées sur les co-occurrences.

En histoire, comme en science politique, le projet ouvre d'importantes perspectives empiriques et méthodologiques, en fournissant de manière expérimentale des instruments automatisés de fouille de textes et de données non-structurées, permettant de systématiser les sondes et d'élargir les corpus d'étude, en particulier pour l'approche sociologique des réseaux sociaux implicites, leur détection, leur visualisation et leur analyse (Barats, 2013).

5. Dimension interdisciplinaire

Notre outil a clairement vocation à être utilisé dans un contexte de sciences humaines et sociales, puisqu'il cible l'étude des interactions sociales dans le domaine du pouvoir politique, dans le passé et aujourd'hui.

La dimension interdisciplinaire apporte à l'informatique les enjeux socio-historiques, les éléments de contextualisation et d'évaluation pour tester les outils d'extraction d'information et de détection de réseaux sociaux. Elle fournit aux historiens du politique et aux politistes l'occasion d'équiper leurs recherches d'instruments d'une fouille de données textuelles systématisée et d'inventer alors de nouveaux corpus de données pour l'analyse des systèmes complexes que sont les réseaux d'acteurs du monde politique.

6. Positionnement dans Agorantic

Le projet implique des chercheurs appartenant à trois équipes membres d'Agorantic :

- Le **Laboratoire Informatique d'Avignon** (LIA) avec Vincent Labatut, pour tout ce qui porte sur le développement de l'outil lui-même.
- L'équipe **Histoire de l'Europe Moderne et Contemporaine** (HEMOC) du Centre Norbert Elias avec Frédéric Monier, pour l'application à des données historiques.
- Le **Laboratoire Biens, Normes et Contrats** (LBNC) avec Guillaume Marrel, pour les aspects applicatifs liés au monde politique contemporain.

Thématiquement, le projet s'intègre dans plusieurs axes de la FR :

- **Axe B** (Réseaux sociaux, structures, contenus et usages) : le but final est l'extraction de réseaux sociaux, donc notre projet est naturellement le plus proche de cet axe pour ce qui est de la méthodologie.
- **Axe C** (Patrimoine, territoire et politiques publiques) : mobilisant deux champs d'expérimentation en histoire politique et intellectuelle et en science politique, le projet concerne ici surtout le volet patrimonial et politique de la FR.
- **Axe E** (Accès au savoir, éthique, et méthodologies) : notre projet vise à produire un outil utilisable pour l'étude des objets thématiques traités dans les autres axes d'Agorantic.

7. Partenariats extérieurs

Le projet inclut un partenariat avec *Semaweb*[6], une société basée à Avignon et spécialisée en Web-marketing. Dans son activité de gestion de l'e-réputation de personnes privées ou morales, cette société utilise des outils de fouille de texte et est intéressée par le noyau applicatif développé ici. Notre outil d'extraction de réseau social implicite lui permettrait d'identifier à la fois les personnes relatives à un sujet donné (personnalité publique, marque, thème d'actualité...), mais aussi les connexions qui les relient, sous la forme d'un graphe d'interactions. Cependant, l'intérêt réel de ce graphe est qu'il peut faire l'objet d'analyses plus poussées, susceptibles de produire des résultats particulièrement pertinents. On peut ainsi identifier des individus, prédire l'évolution du réseau (apparition/disparition de liens), identifier des groupes

[6] http://www.semaweb.fr/

d'individus cohésifs, etc. Des outils issus de l'analyse de réseaux complexes permettent notamment de simuler la diffusion d'information et d'identifier les individus essentiels à ce processus, ce qui constitue un résultat particulièrement utile dans le contexte de l'activité de Semaweb.

L'entreprise est susceptible d'apporter une expertise et une connaissance du milieu politique et économique régional. Elle participera à l'encadrement du stage de Master2 planifié pour cette phase exploratoire, dans la perspective d'une participation significative dans le cadre du projet de recherche plus important qui pourrait découler de cette étude.

8. Budget prévisionnel et financement envisagés

Le budget prévisionnel du projet pour 2015 est évalué à 4500 € répartis comme suit :

Stage M2 (6 mois)	3 000€
Petit matériel, accès aux données, bibliographie	500€
Frais de mission entrants : invités (séminaire)	500€
Frais de mission sortants : communication (conférence)	500€
Total	**4 500€**
Montant sollicité auprès de la FR Agor@ntic	**3 000€**

9. Références

Agarwal, A., Kotalwar, A., & Rambow, O. (2013). *Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice In Wonderland.* Paper presented at the 6th International Joint Conference on Natural Language Processing.

Agarwal, A., Rambow, O., & Passonneau, R. J. (2010). *Annotation scheme for social network extraction from text.* Paper presented at the 4th Linguistic Annotation Workshop.

Ahn, Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007, May). *Analysis of topological characteristics of huge online social networking services.* Paper presented at the 16th International Conference on World Wide Web (WWW'07), Banff, Canada.

Atdağ, S., & Labatut, V. (2013, 2013). *A Comparison of Named Entity Recognition Tools Applied to Biographical Texts.* Paper presented at the 2nd International Conference on Systems and Computer Science.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509.

Barats, C. (Ed.). (2013). *Manuel d'analyse du web en sciences humaines et sociales.* Paris, FR: Armand Colin.

Beaurepaire, P.-Y., & Taurisson, D. (Eds.). (2003). *Les ego-documents à l'heure de l'électronique. Nouvelles approches des espaces et des réseaux relationnels.* Montpellier, FR: Publications de Montpellier III.

Carley, K. M., Columbus, D., & Landwehr, P. (2013). AutoMap User's Guide 2013: Carnegie Mellon University, School of Computer Science, Institute for Software Research.

da Fontoura Costa, L., Oliveira, O., Travieso, G., Aparecido Rodrigues, F., Villas Boas, P., Antiqueira, L., . . . Correa Rocha, L. (2011). Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. *Advances in Physics, 60*(3), 329-412.

da Fontoura Costa, L., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167-242.

Elson, D., Dames, N., & McKeown, K. (2010). *Extracting social networks from literary fiction.* Paper presented at the 48th Annual Meeting of the Association for Computational Linguistics.

Garrote, G. (2013). Entre franchissement et enfermement : pluralité et variabilité de configuration des réseaux notabiliaires territorialisés. *Collection HAL-SHS, Groupe f.m.r (flux, matrices, réseaux)*, 143-161.

Hassan, A., Abu-Jbara, A., & Radev, D. (2012). *Extracting Signed Social Networks From Text.*

Hohl, T. (2007). Divisions Parlementaires Socialistes Au Temps Du Cartel. *Parlement[s]. Revue d'histoire politique, 1*(7), 67-79.

Kok, S., & Domingos, P. (2008). *Extracting Semantic Networks from Text Via Relational Clustering.* Paper presented at the European Conference on Machine Learning and Knowledge Discovery in Databases.

Laurent, D. (2012). Wikipédia, Une Mine D'or Pour Les Chercheurs En TALN. Retrieved from http://blog.wikimedia.fr/wikipedia-une-mine-dor-pour-les-chercheurs-en-taln-4564

Mercklé, P. (2011). *La sociologie des réseaux sociaux.* Paris, FR: La Découverte.

Nay, O. (2011). *Lexique de science politique : Vie et institutions politiques* (2ème ed.). Paris, FR: Dalloz.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*, 167-256.

Padgett, J. F., & Ansell, C. K. (1993). Robust action and the rise of the Medici, 1400-1434. *Am. J. Sociol., 98*, 1259-1319.

Reinanda, R., Utama, M., Steijlen, F., & de Rijke, M. (2013). Entity Network Extraction based on Association Finding and Relation Extraction. *Lecture Notes in Computer Science, 8092*, 156-167.

Watts, D., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 409-410.