# CartoLabe : Cartographie et identification d'individus influents guidées par l'expert

Equipe : TAO
Encadrant : Philippe Caillou - Directrice : Michèle Sebag
Philippe.caillou@inria.fr

## Problématique Générale

L'analyse des données massives (« Big Data ») est devenue un enjeu majeur non seulement au niveau de la recherche, mais également au niveau économique (c'est un des sept axes du plan Innovation 2030 défini sous l'égide du ministère du développement productif[1]). Une des possibilités offerte par la masse de données disponibles concerne l'identification d'individus influents à partir de leur activité en ligne[2][3]. Cette catégorie regroupe aussi bien l'identification de cyber-activistes (hackers), de faiseurs d'opinions conduisant à des activités terroristes, mais aussi celle des scientifiques ou journalistes les plus influents pour diffuser les thèses climato-sceptiques.

Ce problème déjà complexe est rendu encore plus ardu par le fait que l'objectif même est fortement dépendant de l'utilisateur-expert : ce qui est important pour identifier un meneur terroriste sera différent de ce qui est important pour un hacker militant. Sur le même thème, deux experts peuvent même chercher des réponses très différentes alors même que leur demande initiale est identique : par exemple, un spécialiste du renseignement peut chercher à détecter les individus les plus à même de transmettre largement une fausse information à caractère terroriste, alors qu'un autre souhaitera identifier les individus les plus actifs dans le débat islamiste. La requête pourrait être dans les deux cas : individus influents avec le mot clef « islamisme », mais les deux expert n'ont ni la même notion d'influence, ni obligatoirement la même notion de ce qui définit la sphère islamiste.

Une solution pourrait être de passer un temps important à spécifier le problème avec l'expert, afin qu'il précise son objectif et détaille sa requête. Cette démarche nécessite toutefois une forte interaction expert-informaticien, et que l'expert soit capable d'expliciter son objectif et sa vision du domaine. Une solution alternative que nous souhaitons appliquer ici est de fournir un outil générique avec lequel l'expert va interagir et qui va progressivement s'adapter pour correspondre à l'objectif et à la vision de l'expert.

A partir d'un ou plusieurs mots clefs et/ou d'un ou plusieurs individus-exemples connus, l'objectif est d'obtenir un classement d'individus influents et une cartographie des thèmes et des groupes/individus identifiés. A partir de ces éléments, l'expert pourra alors corriger les résultats en déplaçant/supprimant des individus, ce qui permettra à l'outil d'adapter les distances sous-jacentes utilisées pour calculer les similarités et les fonctions de score. Le classement et la cartographie pourront alors s'adapter à l'objectif et à la vision de l'expert.

Ce type d'approche a l'avantage à la fois de fournir un outil qui ne nécessite plus l'intervention de l'informaticien, et d'être générique et donc de pouvoir s'adapter à d'autres sujets. De nombreuses extensions sont par ailleurs possibles, comme la visualisation et

---

[1] http://www.innovation2030.org/fr/
[2] Freimut Bodendorf, Carolin Kaiser: Detecting opinion leaders and trends in online social networks. CIKM-SWSM 2009: 65-68
[3] Francesco Bonchi: Influence Propagation in Social Networks: A Data Mining Perspective. IEEE Intelligent Informatics Bulletin 12(1): 8-16 (2011)

l'utilisation des trajectoires temporelles des individus et/ou des groupes d'individus identifiés. Il serait alors possible d'identifier des personnes avant qu'elles ne deviennent influentes, c'est-à-dire dont la trajectoire logique les conduiraient à devenir une personne influente sur le réseau.

## Programme de Thèse

Ce projet fait suite à une étude réalisée dans le cadre d'un contrat avec la DATAR sur la cartographie des pôles de compétitivités[4]. Pour cela, nous avons rapproché chaque pôle des thèmes scientifiques définis par la nomenclature de l'INIST. Chacun des 39 thèmes de brevet de l'INIST est décrit par un texte en langage naturel ; chaque pôle est décrit par trois types de textes: sa page Web publique; l'ensemble des projets ANR labellisés par le pôle (résumés); l'ensemble des projets FUI du pôle (résumés). Cette caractérisation a permis par la suite de construire aussi bien une cartographie des Pôles dans l'espace des thématiques que des indicateurs d'adéquation thématique des Pôles à leur région, en exploitant les « dires d'expert » pour tenir compte des attendus des experts et de leur connaissance du domaine (voir un exemple Figure 1).

Nous comptons utiliser cette expérience acquise dans le traitement de données textuelles et l'utilisation de l'expert pour la définition d'un domaine pour la généraliser dans une approche générique et l'appliquer également à la fonction de score pour l'identification d'individus influents. Nous disposons déjà, en plus des données publiques accessibles sur le web, de bases de données avec plus de 30M de tweets/blogs/articles, et nous comptons étendre cette base, notamment par la participation aux programmes tels que ceux d'accès complet proposés par Twitter[5]. A partir de ces données et de requêtes directes au web et aux réseaux sociaux, le doctorant devra donc mettre en place une méthodologie et développer un outil qui, à partir de mots clés/exemples d'individus, fourni une liste d'individus influents et une cartographie du domaine cible.

Afin de réaliser l'analyse, il se fondera sur deux types d'indicateurs :

- L'analyse de textes (tweets, blogs, …) disponibles en ligne ou issus de bases de données, à l'aide de méthodes de décomposition en valeurs singulières (SVD) pour déterminer une distance entre documents décrits comme des sacs de mots[6].
- L'analyse de la position dans les réseaux sociaux en tant que tels, en utilisant des métriques pouvant être corrélées à l'influence dans un graphe[7] (comme PageRank), d'abord sur le réseau statique puis sur sa dynamique (en particulier celle des liens Twitter).

---

[4] Philippe Caillou, Émilie-Pauline Gallié, Valérie Mérindol et Thierry Weil, Typologie des pôles de compétitivité basée sur leurs caractéristiques « héritées », 2012, http://www.datar.gouv.fr/sites/default/files/travaux_en_l_13_21032012.pdf

[5] https://blog.twitter.com/2014/introducing-twitter-data-grants

[6] Scott C. Deerwester and Susan T. Dumais and Thomas K. Landauer and George W.Furnas and Richard A. Harshman. Indexing by Latent Semantic Analysis Journal of theAmerican Society of Information Science, vol. 41 (6), pp 391-407, 1990.

[7] Henry Franks, Nathan Griffiths, Sarabjot Singh Anand: Learning influence in complex social networks. AAMAS 2013: 447-454

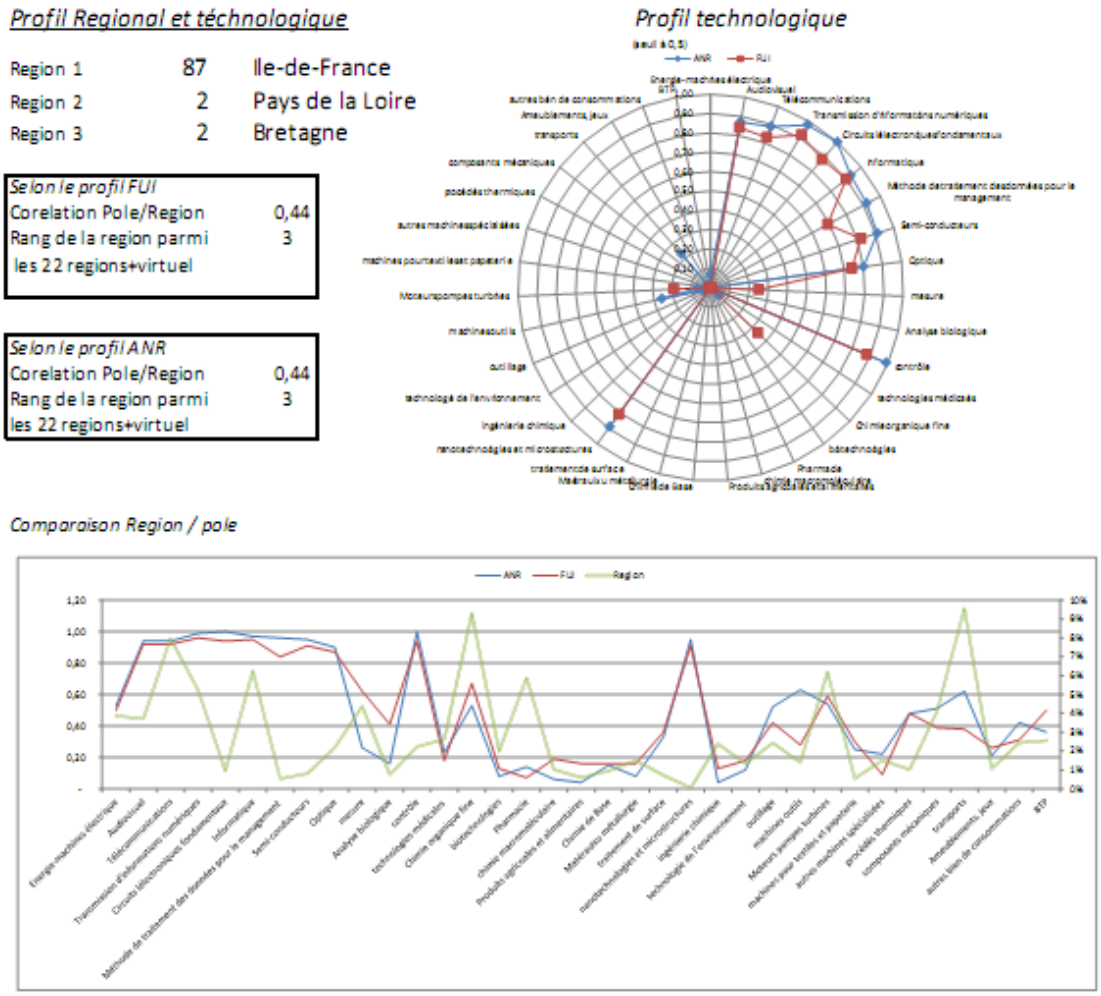


**Figure 1: caractérisation automatique d'un pôle de compétitivité à partir de sa page web et de ses projets de recherche**

Ces deux sources permettent de fournir des indicateurs d'influence reliés à la thématique, et peuvent être visualisés sur une carte à l'aide de techniques de mise à l'échelle multi-dimensionnelle (pour projeter en dimension 2 un ensemble de points dont la matrice de distance est connue, en préservant autant que possible ces distances)[8]. A l'aide de l'interface, l'utilisateur final pourra alors, soit en déplaçant des individus sur la carte, soit en sélectionnant certains d'entre eux pour dire qu'ils sont pertinents/proche d'autres individus, fournir des informations permettant d'améliorer la distance sous-jacente utilisée pour associer les individus à un thème et calculer le score d'influence[9]. Cet apprentissage interactif de distance devrait améliorer le résultat produit tout en apprenant la vision de l'utilisateur sur le domaine qu'il analyse. Les outils à utiliser se baseront sur les outils libres déjà bien développés et personnalisables tels que Apache Solr et Mahout, permettant un traitement de données massives.

Pour le doctorant, après une première phase d'état de l'art sur l'apprentissage de distances, le traitement de données textuelles et les indicateurs d'influence dans les réseaux

[8] Nonlinear Dimensionality Reduction by Locally Linear Embedding, Sam T. Roweis and Lawrence K. Saul, Science, 2000.

[9] Riad Akrour, Marc Schoenauer, Michèle Sebag: APRIL: Active Preference Learning-Based Reinforcement Learning. ECML/PKDD (2) 2012: 116-131

Kilian Q. Weinberger, Lawrence K. Saul: Distance Metric Learning for Large Margin Nearest Neighbor Classication. Journal of Machine Learning Research 10: 207-244 (2009)

sociaux, l'étape suivante sera donc de définir une démarche globale et de développer un premier prototype en utilisant un problème jouet pour lequel nous serons experts, tel que l'identification de scientifiques et journalistes influents sur les thèses climatosceptiques. Une fois une première version en place, de nombreuses voies de recherche sont disponibles. En particulier, l'analyse peut être étendue à l'étude de trajectoires afin d'analyser l'évolution des individus et de parvenir à identifier tôt de futur activistes/individus influents (dans ce cas, la SVD devra probablement porter non plus sur des matrices mais sur des tenseurs décrivant l'historique des documents correspondant à chaque individu).

# 5 Références de l'équipe

Riad Akrour, Marc Schoenauer, Michèle Sebag: APRIL: Active Preference Learning-Based Reinforcement Learning. ECML/PKDD (2) 2012: 116-131
Riad Akrour, Marc Schoenauer, Michèle Sebag: Preference-Based Policy Learning. ECML/PKDD (1) 2011: 12-27
Philippe Caillou, Émilie-Pauline Gallié, Valérie Mérindol et Thierry Weil, Typologie des pôles de compétitivité basée sur leurs caractéristiques « héritées », 2012,
How detailed should social networks be for labor market's models ?, Zach Lewkovicz; Samuel Thiriot; Philippe Caillou, SNAMAS@AISB 2011
Les pôles de compétitivité français en fonction de leur contexte initial d'émergence : essai de caractérisation, Emilie-Pauline Gallié; Valérie Mérindol; Philippe Caillou, ; Thierry Weil EvoReg 2011
Philippe Caillou; Javier Gil-Quijano, SimAnalyzer: automated description of groups dynamics in agent-based simulations, AAMAS 2012, 1353-1354