Fouille de texte et fouille de graphe appliquées à la recherche d'experts

Stella Zevio, Guillaume Santini, Haïfa Zargayouna, Thierry Charnois

Fouille de texte et fouille de graphe appliquées à la recherche d'experts

Stella Zevio, Guillaume Santini, Haïfa Zargayouna, Thierry Charnois

LIPN CNRS UMR 7030, Villetaneuse, France
zevio@lipn.univ-paris13.fr

La recherche d'experts est une problématique récurrente dans le milieu académique. En effet, les chercheurs doivent continuellement être assignés à des comités de programme, de recrutement, à des projets de recherche ou à des expertises de projet. Cette problématique est historiquement liée à celle du profilage d'experts (Lin *et al.*, 2017). Elle implique d'identifier des expertises (définies par Draganidis & Mentzas (2006) comme des compétences, connaissances, aptitudes ou comportements) et de les assigner aux individus adéquats (Bordea, 2013). Identifier les relations qu'un chercheur entretient au sein de la communauté scientifique est également essentiel pour définir son niveau d'expertise en prenant en compte une validation par les pairs. Notre objectif est donc d'identifier des ensembles d'experts (chercheurs) validés par leurs pairs et leur expertises (thématiques de publication) associées.

Afin d'automatiser cette tâche, les expertises et relations entre experts sont extraits automatiquement à partir de texte. Dans le milieu académique, les publications scientifiques recèlent de connaissances pertinentes pour construire les profils d'expert. Nous utilisons des méthodes classiques de fouille de texte (Khan *et al.*, 2017) pour extraire les thématiques de publication et les liens de collaboration scientifique entre chercheurs (citation, co-auteurs, *etc.*) à partir des publications scientifiques. Les connaissances extraites à partir du texte sont communément représentées sous forme de graphe attribué. L'originalité de notre approche consiste en l'adjonction d'une méthode de fouille de motifs exploitant les caractéristiques topologiques d'un graphe afin d'extraire de nouvelles connaissances.

La méthode de fouille de graphe que nous utilisons est l'abstraction de graphe. Elle est inspirée de l'analyse des réseaux sociaux et a déjà été utilisé pour la détection de k-communautés fréquentes (Soldano *et al.*, 2015). Partant du principe que les chercheurs et leurs expertises associées sont représentées sous la forme d'un graphe attribué, nous détectons des k-communautés fortement connectées dans le graphe. Les relations connues liant les auteurs ou les publications sont utilisées pour modéliser l'insertion d'un expert et de ses expertises dans une communauté scientifique et comme critère garantissant une certaine validation par les pairs. Ne seront considérées que les associations experts/expertises supportées par des structures connexes dans le réseau de relations scientifiques. Les structures connexes que nous cherchons à identifier s'appellent des cœurs.

L'identification des cœurs (*core*) d'un graphe est une approche classique d'exploration de la structure des graphes complexes. Le cœur d'un graphe est un sous-graphe maximal dont l'ensemble des sommets vérifient une propriété topologique. La première définition de cœur est celle du cœur *k-core* (Seidman, 1983) pour laquelle l'ensemble des sommets vérifient la propriété d'avoir un degré supérieur ou égal à k. Cette notion a été généralisée (Batagelj & Zaversnik, 2011) et permet la définition de nouveaux cœurs garantissant d'autres propriétés topologiques. Dans notre étude, nous nous intéressons à quatre propriétés topologiques.

Notre postulat est le suivant : si les experts sont fortement liés entre eux (par un réseau dense de relations de citation ou de co-publication par exemple), ils partageraient alors un ensemble d'expertises communes plus grand. En se focalisant sur les k-communautés fréquentes du graphe, notre hypothèse est que nous pourrions à la fois détecter des communautés d'experts, mais également les expertises communes maximales qu'ils partagent.

Une phase exploratoire est en cours sur un échantillon du corpus ACL Anthology (Bird *et al.*, 2008; Gábor *et al.*, 2016) constitué de 13322 publications scientifiques, publiées entre 1985 et 2008 sur les thématiques de la linguistique informatique et du traitement de la langue

naturel. De très nombreux résultats émergent, et la problématique consiste à identifier les paramètres et résultats les plus intéressants. La mise en place d'une méthode d'évaluation adaptée, automatique et rigoureuse nous paraît nécessaire à ce stade. Pour évaluer les résultats obtenus, nous ne disposons d'aucun *gold standard*. La dernière publication issue de notre échantillon du corpus ACL datant de 2008. Nous voulons donc en construire un en se basant sur le chapitre d'un livre de référence datant de 2010 et intitulé Information Extraction (Hobbs & Riloff, 2010) pour valider nos résultats. Pour faire cela, nous identifions les publications citées par le chapitre et référencées dans le jeu de données ACL. Pour chacune de ces publications, nous nous reportons à la partie du chapitre pour laquelle la publication est citée et nous étiquetons la publication avec les mots clefs trouvés dans le texte. Cet étiquetage des publications nous donne une référence à laquelle comparer les motifs énumérés par l'abstraction de graphe. Des 92 références que nous avons automatiquement extraites de ce chapitre, nous retrouvons automatiquement 37 publications issues de l'échantillon du corpus ACL analysé avec des traitements simples.

L'analyse préliminaire des résultats que nous avons obtenu montre que notre méthode d'extraction de concepts sémantiques à partir du texte pourrait également être améliorée. Certains des concepts sémantiques extraits ne sont pas très pertinent (par exemple *paper*, *method* ou *based*). À l'instar du système proposé par Osborne *et al.* (2013) les thématiques et relations extraites pourraient être liées à des concepts d'ontologies pour une meilleure interopérabilité sémantique et pour pouvoir généraliser les concepts identifiés. En perspective, nous souhaitons lier les concepts extraits automatiquement dans les résumés des publications à des concepts d'ontologie plus ou moins spécialisés, afin d'enrichir les motifs clos abstraits.

Références

BATAGELJ V. & ZAVERSNIK M. (2011). Fast algorithms for determining (generalized) core groups in social networks. *Adv. Data Analysis and Classification*, **5**(2), 129–145.

BIRD S., DALE R., DORR B. J., GIBSON B., JOSEPH M. T., KAN M.-Y., LEE D., POWLEY B., RADEV D. R. & TAN Y. F. (2008). The ACL Anthology Reference Corpus : a Reference Dataset for Bibliographic Research in Computational Linguistics.

BORDEA G. (2013). *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. PhD thesis.

DRAGANIDIS F. & MENTZAS G. (2006). Competency Based Management : a Review of Systems and Approaches. *Information Management & Computer Security*, **14**(1), 51–64.

GÁBOR K., ZARGAYOUNA H., BUSCALDI D., TELLIER I. & CHARNOIS T. (2016). Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature. In *LREC 2016*, Proceedings of the LREC 2016 Conference.

HOBBS J. R. & RILOFF E. (2010). Information Extraction. *Handbook of Natural Language Processing*, **2**.

KHAN S., LIU X., SHAKIL K. A. & ALAM M. (2017). A Survey on Scholarly Data : From Big Data Perspective. *Information Processing & Management*, **53**(4), 923–944.

LIN S., HONG W., WANG D. & LI T. (2017). A Survey on Expert Finding Techniques. *Journal of Intelligent Information Systems*, **49**(2), 255–279.

OSBORNE F., MOTTA E. & MULHOLLAND P. (2013). Exploring Scholarly Data With Rexplore. In *International Semantic Web Conference*, p. 460–477 : Springer.

SEIDMAN S. B. (1983). Network structure and minimum degree. *Social Networks*, **5**, 269–287.

SOLDANO H., SANTINI G. & BOUTHINON D. (2015). Local Knowledge Discovery in Attributed Graphs. In *International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 250–257.