Vers une cartographie automatique des thématiques et profils d'experts associés à une conférence scientifique : 9 ans d'ateliers Recherche d'Information SEmantique (RISE)

Stella Zevio, Haïfa Zargayouna, Guillaume Santini, Thierry Charnois

Vers une cartographie automatique des thématiques et profils d'experts associés à une conférence scientifique : 9 ans d'ateliers Recherche d'Information SEmantique (RISE)

Stella Zevio* Haïfa Zargayouna* Guillaume Santini* Thierry Charnois*
(*) LIPN - CNRS UMR 7030 - Université Paris XIII
99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
stella.zevio@lipn.univ-paris13.fr
haifa.zargayouna@lipn.univ-paris13.fr
guillaume.santini@lipn.univ-paris13.fr
thierry.charnois@lipn.univ-paris13.fr

Cet article a été publié dans les actes de l'atelier RISE 2018, conjoint à TALN et CORIA 2018

RÉSUMÉ

La recherche d'experts est une problématique essentielle pour les entreprises comme dans le milieu académique. La construction d'un profil d'expert, c'est-à-dire l'assignation d'expertises à l'individu qui les maîtrise, est une étape nécessaire à la cartographie des experts d'une entreprise ou d'un laboratoire. Les documents de travail contiennent les connaissances requises pour la construction de profils d'experts. Les publications scientifiques en particulier recèlent de connaissances cruciales pour la construction d'un profil d'expert de chercheur. Nous proposons de cartographier automatiquement les thématiques et profils d'experts associés aux ateliers Recherche d'Information SEmantique en appliquant des méthodes de fouille de texte et de fouille de graphe sur les actes de ces neuf dernières années, principalement rédigés en français. Nous obtenons des graphes de connaissances représentant les différentes communautés de chercheurs ayant participé aux ateliers.

ABSTRACT

Towards an automatic mapping of thematics and expert profiles associated to a scientific conference : 9 years of Recherche d'Information SEmantique (RISE) workshops

Expert finding is a key issue for both companies and academia. Building an expert profile (i.e., assignment of expertises to the individual who owns them) is a necessary step in the mapping of a company or a laboratory experts. Working papers contain required knowledge to build experts profiles. In particular, scientific publications contain crucial knowledge for building a researcher's expert profile. We propose to automatically map thematics and expert profiles associated with the Recherche d'Information SEmantique workshops, by using text mining and graph mining methods on the past nine years acts mainly written in French. We obtain knowledge graphs representing the different communities of researchers who participated in the workshops.

MOTS-CLÉS : recherche d'experts, recherche d'information, fouille de texte, fouille de graphe, abstraction de graphe.

KEYWORDS: expert mining, information retrieval, text mining, graph mining, graph abstraction.

1 Introduction

La recherche d'experts sur un sujet constitue un problème important pour les entreprises comme dans le milieu académique. En effet, les entreprises doivent continuellement assigner un expert interne à une tâche ou un projet. Dans le milieu académique se pose également la question de l'assignation d'un chercheur à un comité de programme, de recrutement, au montage d'un projet de recherche ou à une expertise de projet. Une étape nécessaire à la cartographie des experts d'une entreprise ou d'un laboratoire est la construction des profils d'experts des membres de la structure concernée. Construire un profil d'expert consiste en l'assignation d'expertises à l'individu qui les maîtrise. Les documents de travail (CV, rapports d'activités pour les entreprises, rapports d'équipes et publications scientifiques dans le milieu académique) contiennent les connaissances requises pour la construction de profils d'experts. Dans le milieu académique, ce sont les publications scientifiques en particulier qui recèlent de connaissances cruciales pour la construction d'un profil d'expert de chercheur.

Nous proposons d'établir une méthode de cartographie automatique des thématiques et profils d'experts associés à une conférence scientifique. Pour cela, nous nous intéressons à l'identification automatique de thématiques au sein des actes scientifiques, à la génération d'un graphe de connaissances reliant les experts (auteurs des publications) entre eux et aux thématiques sur lesquelles ils ont publié, ainsi qu'à l'accès à des connaissances implicites au sein de ce graphe, à l'aide des méthodes d'abstractions de graphe explicitées dans la section 5. Pour permettre l'identification automatique de thématiques au sein des actes scientifiques, nous utilisons des méthodes de traitement du language naturel et de fouille de textes sur des corpus de documents multilingues (en langues française et anglaise). Pour célébrer la dixième édition des ateliers Recherche d'Information SEmantique (RISE), nous proposons d'appliquer cette méthode aux actes de l'atelier sur la dernière décennie. Les actes sont principalement rédigés en français. Nous cherchons à identifier les différentes communautés de chercheurs ayant participé aux ateliers.

Cet article est structuré de la manière suivante : nous présentons le travail de thèse et l'état de l'art dans la section 2. Nous abordons la méthode de fouille de texte appliquée sur les actes de RISE dans la section 3. Puis, nous décrivons le graphe de connaissances issu des données obtenues à partir de la fouille de texte dans la section 4, nous abordons la méthode de fouille de graphe appliquée sur le graphe de connaissances dans la section 5 et enfin, nous présentons les conclusions et perspectives de ce travail dans la section 6.

2 Découverte et enrichissement de connaissances à partir de textes pour la recherche d'experts

Ce travail s'inscrit dans le cadre du projet PCU (Plateforme de Connaissance Unifiées) visant à produire une plateforme open-source industrielle de valorisation de données d'entreprise, et plus particulièrement dans le cadre d'une thèse FUI, dont le titre est celui de la section. Le cas d'application du projet et de la thèse est la recherche d'experts sur un sujet.

Une expertise est définie comme une «compétence, connaissance, aptitude ou le comportement d'un individu» (Draganidis & Mentzas, 2006). Ici, nous définirons une expertise comme une thématique de publication. Construire un profil d'expert consiste à associer à un individu ses expertises (Bordea, 2013), ici à associer à un chercheur l'ensemble des thématiques sur lesquelles il a publié. Dans notre

approche, les chercheurs sont reliés entre eux par des relations de co-publication (co-auteurs) ou de co-citation, par exemple. Les chercheurs sont décrits par leur profil d'expert.

Une méthode de recherche d'expert communément utilisée dans l'état de l'art consiste en la construction de profils d'experts à partir de textes par le biais d'extraction de thématiques ou de phrases-clefs dans les textes (Bordea, 2013; Sateli *et al.*, 2017), et de leur assignation à des individus, en liant parfois les thématiques extraites avec des ontologies pour respecter l'interopérabilité sémantique. Les méthodes de fouille de texte que nous utilisons sont basées sur ces principes et sont décrites plus en détails dans la section 3. Contrairement aux approches de recherche d'experts basées sur l'application de méthodes de fouille de texte, notre approche combine application de méthodes de fouille de texte et d'une méthode de fouille de graphes particulière, l'abstraction de graphes. L'abstraction de graphes est une extension du cadre de l'analyse de concepts formels. Elle est décrite plus en détails à la section 5. L'abstraction de graphe est appliquée sur un graphe de connaissances obtenu à la suite de l'application des méthodes de fouille de textes.

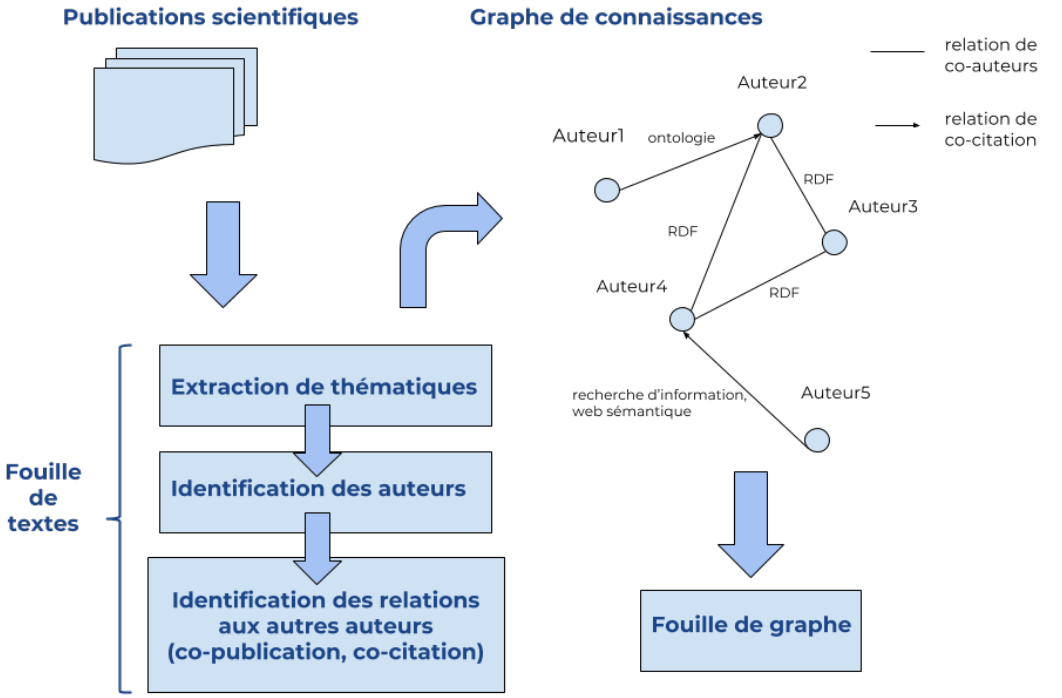

FIGURE 1 – Notre méthode de recherche d'experts sous la forme d'un flux de traitement

Un graphe de connaissances est une représentation graphique de sujets (les sommets du graphe) reliés entre eux par des relations sémantiques (les arêtes du graphe). Ici, les sommets sont les chercheurs, et les relations sémantiques les liens de co-publication ou de co-citation. Le graphe est orienté, c'est-à-dire que les relations vont d'un sommet vers un autre et sont représentées par des arcs. Ce graphe de connaissances est étiqueté, c'est-à-dire qu'il s'agit d'un graphe orienté dont les arcs sont porteurs d'étiquettes (les thématiques issues de la publication concernée par la relation de co-publication ou de co-citation). Notre méthode de recherche d'experts est représentée dans la figure 1.

3 De la fouille de texte au graphe de connaissances

Les actes des ateliers RISE contiennent 48 publications rédigées par 87 auteurs différents depuis 2009 jusqu'à 2017. Les thématiques officielles de l'atelier sont la recherche d'information, le web sémantique, l'extraction de connaissances, le traitement automatique des langues naturelles et le multimédia. Les actes sont disponibles au format PDF. Nous avons utilisé Apache Tika pour convertir les actes au format texte.

Les auteurs des publications sont extraits manuellement pour le besoin de cette expérience. Une automatisation de cette tâche doit être envisagée à l'avenir. Pour extraire les thématiques associées à un auteur, nous cherchons à associer à un auteur les thématiques sur lesquelles il a publié. Nous formulons l'hypothèse selon laquelle un auteur est expert du contenu qu'il publie. Pour cette tâche, nous utilisons habituellement un algorithme d'extraction de phrases-clefs (Hernandez *et al.*, 2017) entraîné sur l'anglais et donc dépendant de la langue. Les publications de RISE étant principalement écrites en langue française, nous avons donc extrait manuellement les mots-clefs présents dans les publications pour le besoin de l'expérience. En effet, les mots-clefs sont un sous-ensemble des thématiques. Nous devons envisager une automatisation de cette tâche à l'avenir, et un entraînement de l'algorithme sur le français. Pour extraire automatiquement les relations sémantiques nouvelles et connues entre les thématiques, nous utilisons un algorithme (Gábor *et al.*, 2016) éprouvé sur un corpus ACL (Bird *et al.*, 2008). Dans cette expérience, les liens sémantiques entre thématiques ne sont pas traités.

Les thématiques et relations extraites peuvent ensuite être liées à des concepts d'ontologies (FOAF, *conference ontology*, etc.) pour une meilleure interopérabilité sémantique. Le graphe de connaissances que nous avons généré relie les sommets (auteurs de publications) entre eux par des relations de co-auteurs (si un auteur a co-publié au moins une fois avec un autre, alors ils sont reliés dans le graphe).

4 Description du graphe de connaissances

Le graphe de connaissances construit à partir des données issues de la fouille de textes contient 87 sommets (auteurs de publications) et 185 arêtes (liens de co-auteurs). Le degré moyen (nombre moyen de co-auteurs) dans le graphe est de 4. La distribution du degré dans le graphe est représentée dans la figure 2.

Nous avons obtenu un classement des participants ayant le plus fort degré dans le graphe. Le classement, limité au 5 degrés les plus élevés est disponible dans la table 1. Didier Schwab a le plus fort degré dans le graphe. Il a participé à un article avec les membres de la clique de 12 co-auteurs qui sont en deuxième position du classement, et il forme un lien de co-auteur avec une autre communauté. Ce résultat peut être visualisé dans le graphe de connaissances présenté dans les figures 3, 4 ou 5, sans descripteurs visibles. Le graphe obtenu n'est pas connexe. Certains auteurs ont publié seuls, d'autres ont publié en triades n'ayant pas d'autres liens de co-auteur avec les autres auteurs dans RISE. Cependant, ces auteurs ont pu co-publier avec d'autres auteurs dans le graphe dans le cadre d'autres conférences.

Les descripteurs, binaires, sont les thématiques, les dates de publication et les localisations de laboratoires. Les dates en particulier sont représentées par une binarisation des intervalles des dates

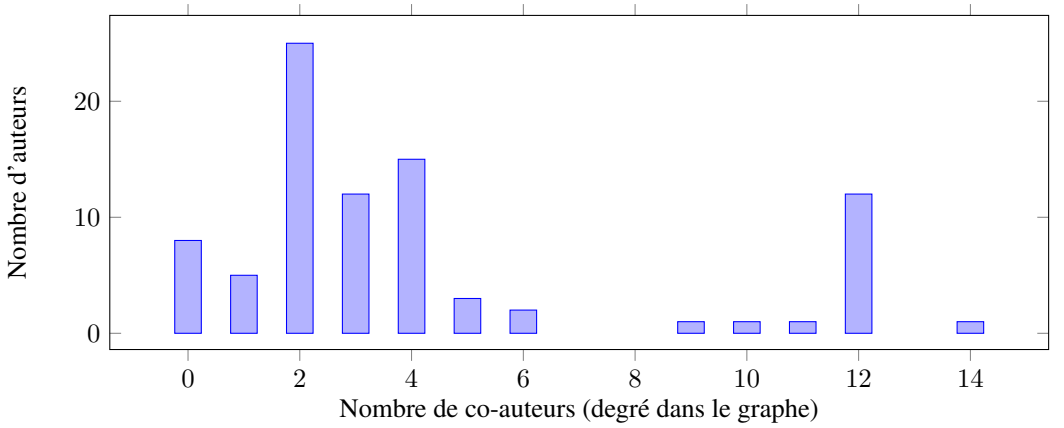

FIGURE 2 – Distribution du nombre de co-auteurs (degré) dans le graphe

(de 2009 à 2017, nous avons les descripteurs 2009_2010_2011, 2010_2011_2012, 2011_2012_2013, *etc.*). Nous avons choisi des intervalles de trois dates pour pouvoir relier des publications à celles des années les ayant immédiatement précédé ou suivi.

Co-auteurs	Auteur
14	Didier Schwab
12	Valérie Bellynck, Hervé Blanchon, Christian Boitet, Liming Chen Gabriela Csurka, Emmanuel Dellandréa, Achille Falaise, Ningning Liu Luca Marchesotti, David Rouquet, Alexandre Saidi, Sandra Skaff
11	Sylvie Calabretto
10	Catherine Roussey
9	Jean-Pierre Chevallet

TABLE 1 – Classement des participants ayant le plus grand nombre de co-auteurs (5 plus grand nombres de co-auteurs)

5 Abstraction de graphe sur le graphe de connaissances

D'autres connaissances implicites peuvent être obtenues à l'aide d'abstractions du graphe. L'abstraction de graphes est une extension de la fouille de motifs clos et du cadre de l'analyse de concepts formels aux graphes attribués (Soldano *et al.*, 2017). Un graphe attribué est un graphe G=(V,E), où V est un ensemble de sommets, E un ensemble d'arcs et des étiquettes sont portées sur les sommets et/ou les arêtes. Ici nous considérons que les sommets sont étiquetés par les thématiques, dates de publications et localisations de laboratoires habituellement portées par les publications pour plus de simplicité. L'abstraction de graphe consiste à réduire l'ensemble des sommets dans le graphe à un *core*, c'est-à-dire à un sous-graphe dense, dont les noeuds satisfont une contrainte topologique à l'intérieur du sous-graphe induit. Les sommets appartenant à l'ensemble sont porteurs d'un motif

clos, c'est-à-dire d'un motif le plus spécifique possible partagé par les sommets. Un motif est un sous-ensemble d'attributs. Nous appliquons un algorithme (Soldano *et al.*, 2015) énumérant l'ensemble des couples de motifs clos (le motif le plus spécifique partagé par un ensemble d'objets) et d'extensions abstraites (l'ensemble de sommets présentant un motif clos et vérifiant une contrainte topologique dans le graphe).

Soit t, la contrainte topologique suivante : tout sommet du graphe doit être de degré supérieur ou égal à 2. Soit la thématique «ontologie» utilisée comme motif clos. L'application de l'abstraction du graphe énumérant l'ensemble des couples de motif clos «ontologie» et d'extensions abstraites vérifiant la contrainte topologique t permet d'obtenir le graphe représenté dans la figure 3. Nous pouvons remarquer que Catherine Roussey et Haïfa Zargayouna, deux co-organisatrices des ateliers, font partie de la communauté «ontologie» des chercheurs. L'application de l'abstraction du graphe énumérant l'ensemble des couples de motif clos «recherche d'information» et d'extensions abstraites vérifiant la même contrainte topologique t permet d'obtenir le graphe représenté dans la figure 4. Nous pouvons remarquer que Jean-Pierre Chevallet, dernier co-organisateur des ateliers fait partie de la communauté «recherche d'information» des chercheurs. Enfin, l'intersection des deux motifs clos «ontologie» et «ri» permet d'obtenir le graphe représenté dans la figure 5. Il est intéressant de remarquer qu'un chercheur comme Loïc Maisonnasse ayant bien publié sur les deux thématiques n'apparaît pas dans cette abstraction de graphe. En effet, il ne satisfaisait pas les contraintes topologiques posées.

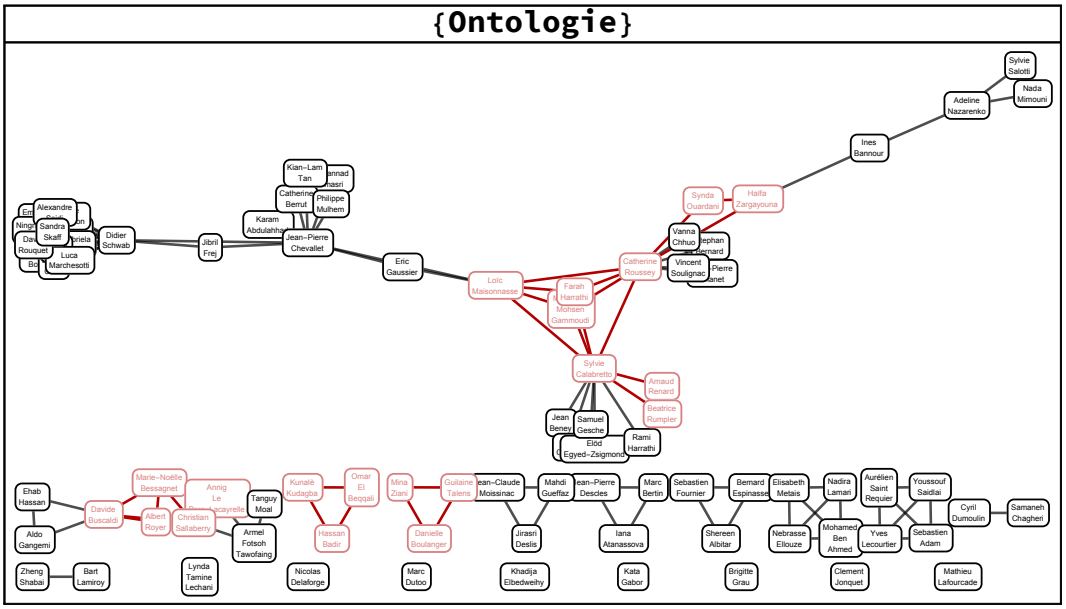

FIGURE 3 – Abstraction du graphe sur le motif «Ontologie», contrainte topologique de degré 2

L'intérêt est de découvrir des connaissances à partir de données structurées et d'accéder à des connaissances implicites, auxquelles nous n'aurions pas eu accès sans l'application d'une abstraction sur le graphe. La complexité de l'extraction de connaissances n'est plus liée à la nature non structurée des données (comme pour le texte) mais à la complexité combinatoire des connexions entre les sommets.

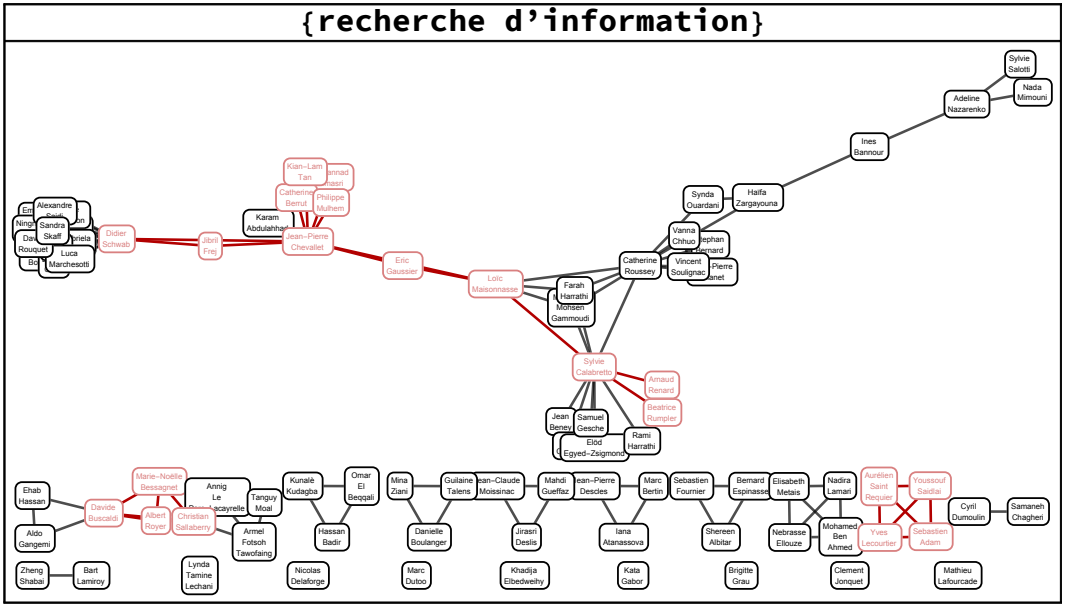

FIGURE 4 – Abstraction du graphe sur le motif «Recherche d'information», contrainte topologique de degré 2

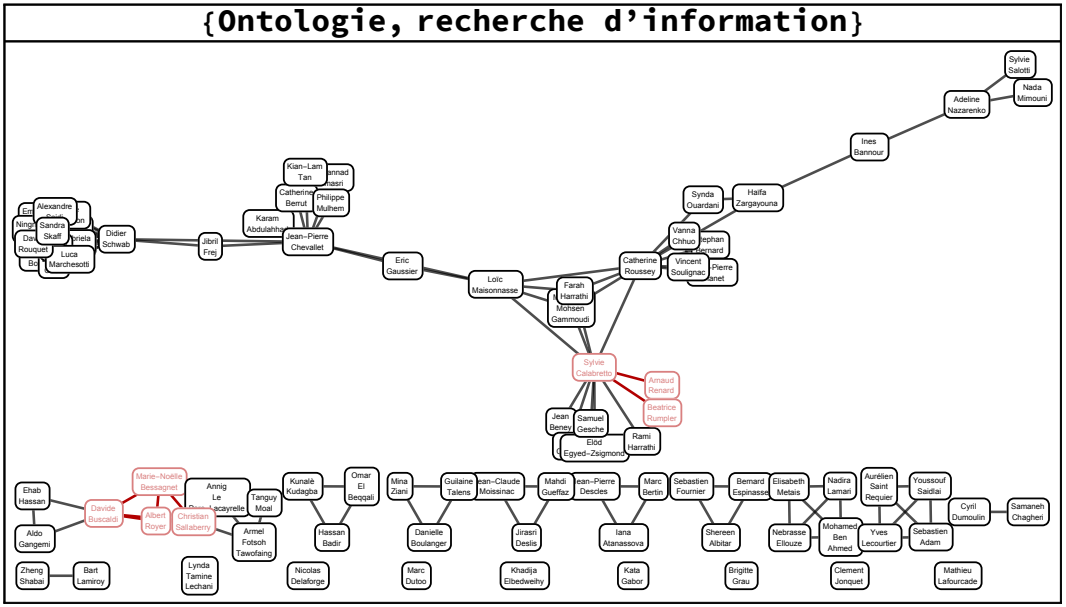

FIGURE 5 – Abstraction du graphe sur les motifs «Ontologie» et «Recherche d'information», contrainte topologique de degré 2

6 Conclusion

Nous avons mené une expérience sur les actes de l'atelier RISE nous permettant d'instancier les problématiques posées par la thèse et le projet PCU et d'identifier des résultats attendus, les méthodes et outils nécessaires à la mise en œuvre de la recherche d'experts au sein de publications scientifiques. Nous avons identifié l'intérêt de la fouille de textes combinée à de la fouille de graphe sur les publications scientifiques afin d'accéder à de nouvelles connaissances implicites (les mots-clefs ne suffisant pas à décrire une publication de façon exhaustive, et certaines connaissances n'étant accessibles qu'à un certain niveau d'abstraction).

Nous avons identifié un certain nombre de difficultés, notamment dans la réutilisation d'algorithmes d'apprentissage nécessitant un entraînement (nous n'avons pas encore pu adapter l'algorithme d'extraction de thématiques (Hernandez *et al.*, 2017) à la langue française). En effet, il n'est pas aisé de trouver des ressources pour le français de manière générale. À court terme nous souhaitons nous diriger vers plus d'automatisation, et proposer une cartographie entièrement automatisée des thématiques et profils d'experts associés à une conférence scientifique. Il serait également intéressant d'enrichir les connaissances obtenues avec des sources extérieures telles que DBLP par exemple. Les relations sémantiques pouvant lier des auteurs dans le graphe pourraient également être enrichies par des relations de co-citation, et non pas simplement des relations de co-auteurs. Des relations sémantiques pourraient également exister entre les thématiques.

Références

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.

HERNANDEZ S. D., BUSCALDI D. & CHARNOIS T. (2017). LIPN at SemEval-2017 Task 10 : Filtering Candidate Keyphrases from Scientific Publications with Part-of-Speech Tag Sequences to Train a Sequence Labeling Model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 995–999.

SATELI B., LÖFFLER F., KÖNIG-RIES B. & WITTE R. (2017). Scholarlens : Extracting Competences from Research Publications for the Automatic Generation of Semantic User Profiles. *PeerJ Computer Science*, **3**.

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.

SOLDANO H., SANTINI G., BOUTHINON D. & LAZEGA E. (2017). Hub-Authority Cores and Attributed Directed Network Mining. In *International Conference on Tools with Artificial Intelligence (ICTAI)*.