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

Stella Zevio

Université Sorbonne Paris Nord

Thèse présentée pour obtenir le grade de docteur
Université Sorbonne Paris Nord

Laboratoire d'Informatique de Paris Nord - CNRS UMR 7030
École doctorale Galilée

Discipline : Informatique

N° attribué par la bibliothèque : |___|___|___|___|___|___|___|___|___|___|

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

par Stella Zevio

Sous la direction de Thierry Charnois, Professeur des Universités

Co-encadrée par Guillaume Santini **et** Haïfa Zargayouna, Maîtres de conférences

Membres du jury:

Examinateur : Christophe Fouqueré, Professeur des Universités, Université Sorbonne Paris Nord, Villetaneuse, France

Examinatrice : Christine Largeron, Professeur des Universités, Université Jean-Monnet, Saint-Etienne, France

Examinatrice : Sylvie Ranwez, Professeur, IMT Mines Alès, Alès, France

Rapportrice : Catherine Faron Zucker, Maître de conférences, Université de Nice Sophia Antipolis, Nice, France

Rapporteur : Christophe Rigotti, Maître de conférences, Institut National des Sciences Appliquées de Lyon, Lyon, France

Date de soutenance : 23 février 2021

Remerciements

Ces travaux de thèse ont été réalisés sur une période s'étendant de l'année 2017 à 2020, au sein du Laboratoire d'Informatique de Paris Nord (LIPN), unité mixte de recherche du CNRS rattachée à l'Université Paris 13. J'ai été dirigée par Thierry Charnois, Professeur des Universités et co-encadrée par Haïfa Zargayouna et Guillaume Santini, maîtres de conférences. Cette thèse est l'accomplissement d'années d'anticipation et d'efforts, a été préparée en amont dès mes premières années d'études supérieures et parachève un objectif que je chérissais de très longue date. Je tiens donc à exprimer ma gratitude envers l'ensemble des personnes ayant pu rendre cette concrétisation possible.

Le premier de mes remerciements s'adresse à mon directeur de thèse, Pr. Thierry Charnois, pour la grande confiance qu'il m'a accordée et la bienveillance dont il a su faire preuve tout au long de ces trois années durant lesquelles j'ai eu le plaisir d'être son étudiante. Je souhaite ensuite remercier mon encadrante, Dr. Haïfa Zargayouna, pour l'amitié dont elle a su faire preuve tout au long de son encadrement, effaçant les moments de doute qui ont parfois égrené cette aventure. Je souhaite également remercier le dernier mais non le moindre de mes encadrants, Dr. Guillaume Santini, pour sa patience et pour les nombreux efforts qu'il a déployés pour m'introduire à la fouille de graphes attribués qui m'était jusqu'alors inconnue. Je le remercie également de m'avoir épaulée tout au long de ces trois années de thèse et d'avoir fait preuve d'une si grande disponibilité. Je souhaite également remercier les membres de ce jury de thèse, les examinateurs, Pr. Christophe Fouqueré, Pr. Christine Largeron, Pr. Sylvie Ranwez ainsi que les rapporteurs, Dr Catherine Faron Zucker et Dr. Christophe Rigotti pour avoir accepté d'évaluer mes travaux durant cette période particulière.

Mes travaux de thèse ont été financés par le fonds ministériel FUI associé au projet PCU (Plateforme de Connaissances Unifiées). Ce projet avait pour objectif la conception d'une plateforme de valorisation des données industrielles et académique. Il rassemblait de nombreux partenaires académiques (LIPN, ESILV) et industriels (Smile, Proxem, Armadillo et Wallix). Je tiens à remercier l'ensemble des partenaires du projet avec qui j'ai pu interagir et tiens à souligner le cadre très enrichissant dans lequel j'ai pu valoriser mes travaux de thèse. Produire une plateforme sémantique dans le cadre du projet a été une expérience enrichissante et formatrice [1].

Je souhaite également remercier l'ensemble des membres de mon laboratoire pour leur accueil chaleureux. Les nombreuses discussions que nous avons pu entretenir autour de la machine à café (ou plutôt à chocolat chaud pour ma part) ont contribué à animer ces

1. PCU : `https://github.com/zevio/PCU`

trois années de thèse. Mes travaux se situant à l'interface de deux équipes de recherche du laboratoire, j'ai eu la chance d'interagir assidûment avec de nombreux chercheurs. Synthétiser deux visions d'une même problématique dans ces travaux de thèse a parfois été une tâche ardue mais toujours enthousiasmante. Au sein de l'équipe Apprentissage Artificiel et Applications (A3) du LIPN, mes remerciements vont particulièrement à Dr. Henry Soldano dont les remarques pertinentes, les explications détaillées et les importantes contributions ont permis de réaliser de grandes avancées dans mes travaux. Au sein de l'équipe Représentation des Connaissances et Langage Naturel (RCLN) du LIPN, mes remerciements les plus vifs vont vers mon co-bureau, Dr. Davide Buscaldi. Je le remercie d'avoir si souvent partagé son expérience et ses conseils sur le monde de la recherche académique en général. Je remercie également les chercheurs du domaine auprès de qui j'ai pu recueillir des informations cruciales et des conseils inestimables pour l'achèvement de ces travaux, Dr. Francesco Osborne, Dr. Richard Berendsen, M. Steffen Remus et M. Tim Fischer.

Je souhaite également remercier les directrices successives du LIPN durant la période pendant laquelle j'ai effectué ma thèse, Pr. Laure Petrucci et Pr. Frédérique Bassino, ainsi que mon tuteur, Pr. Christophe Fouqueré, et l'assistante de direction, Mme Brigitte Guéveneux, pour avoir supervisé le bon déroulement de la thèse au niveau administratif ainsi que pour leur bienveillance. Pour ces mêmes raisons, je tiens à remercier mon école doctorale, Galilée, en particulier le directeur, Pr. Dominique Ledoux et le directeur adjoint, Pr. Olivier Bodini. J'en profite également pour mentionner les membres de mon comité de suivi de thèse à mi-parcours, particulièrement Dr. Catherine Faron Zucker pour ses retours éclairants à un tournant de mes travaux et qui ont permis, je l'espère, de les améliorer dans ce manuscrit final. Je remercie chaleureusement les administrateurs système du laboratoire, en particulier M. Xavier Monnin. La pancarte dans son bureau ne mentait pas, les administrateurs système sont vraiment des héros. Merci d'avoir résolu des problèmes aussi inextricables et terre-à-terre que des lignes manquantes dans les bases de données du laboratoire ou les pannes du serveur Gaïa sur lequel j'hébergeais mes calculs. Sans son aide précieuse, mes expériences n'auraient pas pu être matérialisées en si grand nombre.

Je tiens à remercier l'ensemble de l'équipe enseignante de l'IUT de Villetaneuse pour leur accueil chaleureux et leur confiance renouvelée au cours de ces trois années de monitorat. La mission d'enseignement qui m'a été attribuée au sein de l'IUT a été une véritable parenthèse enchantée dans mon contrat doctoral. J'adresse un remerciement particulier à mes étudiants à qui j'espère avoir transmis mon enthousiasme ainsi qu'un peu de mes connaissances et de mon expérience. Un remerciement spécial est adressé à M. Rémi Duret et M. Elliott Falguerolle, que j'ai eu le plaisir d'encadrer durant un stage clôturant leur DUT. Dans ce cadre, ils ont réalisé de jolies contributions techniques à ce travail de thèse

en proposant respectivement un parseur NRI (format d'entrée du logiciel MinerLC [2] de l'équipe A^3 du LIPN) $\leftrightarrow$ RDF [3] et un visualiseur RDF [4].

Enfin, il m'est impossible de ne pas remercier les personnes qui ont contribué à l'accomplissement de cette thèse avant même qu'elle ne commence. Je remercie les professeurs de l'Université de Montpellier, particulièrement Dr. Mathieu Lafourcade pour ses encouragements et ses conseils délivrés tout au long de ma formation. Je tiens également à remercier les anciens doctorants du Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) ayant effectué leur thèse autour de l'année 2015, époque durant laquelle j'assistais assidûment aux séminaires doctoraux. À cette époque, j'essayais de comprendre leurs travaux et écoutais leurs récits avec un très fort intérêt avant de finir par présenter mes propres travaux en 2017. Je remercie également les différents professeurs qui m'ont accordé leur confiance et m'ont introduite au monde de la recherche scientifique dans le cadre des stages que j'ai eu la chance de pouvoir effectuer durant ma formation. Je remercie également Pr. Shigeo Sugimoto et Dr. Tetsuya Mihara, membres du laboratoire de la *Faculty of Library, Information and Media Science* de l'Université de Tsukuba au Japon, pour leur accueil chaleureux en 2017 et pour avoir maintenu notre collaboration en parallèle de mes travaux de thèse cette année là en dépit de la distance géographique.

Je remercie également mes camarades doctorants pour avoir partagé mon quotidien pendant ces trois dernières années avec humour et bonne humeur. Je tiens à exprimer ma gratitude envers mes amis, dont beaucoup ont été rencontrés sur les bancs de la fac et m'ont encouragée depuis les prémices de ma formation universitaire. Je remercie également Haru, Cloud et Casca pour m'avoir tenu compagnie durant la rédaction de ce manuscrit et y avoir parfois contribué en insérant quelques caractères indésirables. Enfin, je remercie mon époux pour son soutien inconditionnel, ses nombreux encouragements ainsi que son aide précieuse au quotidien. Mon remerciement le plus chaleureux s'adresse à mes parents pour avoir toujours été bienveillants, présents, impliqués et pour m'avoir toujours encouragée à poursuivre mes rêves.

Je souhaite enfin remercier l'ensemble des chercheurs qui ont relu ou porté attention à ce travail, qu'il s'agisse des relecteurs de comités de lecture dans les conférences ou de chercheurs croisés au hasard des manifestations scientifiques. Je remercie également par avance ceux qui, je l'espère, s'intéresseront à ces travaux dans le futur ou les trouveront dignes d'intérêt.

2. MinerLC : `https://lipn.univ-paris13.fr/MinerLC/`
3. Parseur NRI $\leftrightarrow$ RDF (réalisé par Rémi Duret) : `https://github.com/zevio/RDF_extraction_connaissances`
4. Visualiseur RDF (réalisé par Elliott Falguerolle) : `https://github.com/zevio/visu-RDF`

À mes parents.

Table des matières

Introduction

Dans le milieu académique, la recherche d'experts est une problématique récurrente. En effet, il est essentiel d'assigner des experts appropriés lors du montage de comités de programme de manifestations scientifiques, de projets de recherche ou de comités de recrutement par exemple. Au sein des entreprises, la recherche d'experts est également une thématique fondamentale puisqu'il est constamment nécessaire d'identifier les experts internes pour l'attribution des tâches et projets. De plus, il est parfois nécessaire de faire appel à un expert extérieur pour la résolution de problèmes ponctuels. En outre, il est pertinent d'identifier des individus compétents lors de phases de recrutement, en particulier lorsqu'il s'agit de domaines d'expertise récents, sur lesquelles le recruteur n'a parfois que peu de maîtrise par exemple.

L'identification des experts, de leurs expertises associées et d'éventuels conflits d'intérêt constituent des tâches essentielles pour la recherche d'experts. Une étape préalable consiste en la définition des profils d'experts considérés. Définir un profil d'expert réside en l'assignation d'expertises appropriées à un individu. Pour définir le profil d'expert d'un individu, nous devons donc identifier ses expertises. Une expertise a été définie dans l'état de l'art comme la connaissance, la compétence, l'aptitude ou le comportement d'un individu. Cependant, il est également nécessaire de prendre en compte la réputation d'un individu pour estimer son niveau d'expertise. En effet, l'attribution du statut d'expert à un individu n'est pas seulement conditionnée à la nature ou à la quantité d'expertises qu'il maîtrise mais également à la réputation qu'il a construit auprès de ses pairs. Par exemple, dans la communauté scientifique, la réputation d'un individu peut être estimée en évaluant la réception de ses travaux par les autres membres de la communauté. En effet, si un membre de la communauté scientifique produit des travaux fréquemment cités ou s'il entretient de nombreux liens de coauteurs sur une thématique récurrente, il s'agit probablement d'un expert pertinent sur cette même thématique. Du point de vue des entreprises de services du numérique, les revues de code peuvent par exemple représenter un processus de validation interne similaire à celui du comité de lecture dans le domaine de la recherche académique.

Les connaissances décrivant l'expertise ou la réputation d'un individu peuvent être dérivées du contenu des documents qu'il rédige, des relations qu'il entretient avec ses pairs ou encore de son activité sur les réseaux sociaux ou les communautés en ligne. Le texte constitue une source de connaissances privilégiée. Dans le milieu académique, les publications scientifiques recèlent des connaissances pertinentes pour construire les profils d'expert. En effet, elles véhiculent les expertises ainsi que les liens de collaboration

scientifique nécessaires à l'évaluation de la réputation des membres de la communauté scientifique. Pour les entreprises, ce sont les rapports d'activité qui sont considérés ainsi que les *curriculum vitæ* dans une moindre mesure, puisque basés sur l'auto-assignation des expertises.

Initialement, la recherche d'experts était basée sur l'auto-assignation des expertises par les individus. Cette dernière repose sur la capacité des experts à fournir une description détaillée et complète de leurs expertises et à la maintenir à jour au gré des évolutions professionnelles et de l'acquisition de nouvelles expertises. En raison du caractère chronophage de la description et de la maintenance des profils d'expert des individus, des systèmes de recherche d'experts automatiques ont vu le jour. La recherche d'experts étant un sujet de recherche important, de nombreux systèmes de recherche d'experts ont été proposés. Le plus grand nombre d'entre eux tire parti d'une extraction automatique des expertises à partir de textes. Leur approche est donc généralement basée sur de l'apprentissage automatique. Un expert est alors défini comme un individu dont la description des thématiques associées correspond au mieux à une requête considérée.

Parmi les systèmes de recherche d'experts, certains prennent en compte des indicateurs liés à la réputation des individus tels que le nombre de citations de leurs travaux. Des méthodes à base de graphe inspirées de l'analyse des réseaux sociaux ont été employées pour automatiser la définition des profils d'expert. Celles-ci font généralement abstraction du contenu des documents rédigés par les individus et se basent sur une analyse des relations sociales entretenues par les individus au sein d'une communauté. Les courriels ont notamment été explorés pour l'identification d'experts. Un individu répondant à de nombreuses questions envoyées par ses pairs peut être considéré comme un expert. Plus récemment, des méthodes hybrides ont vu le jour, combinant des méthodes à base d'apprentissage et des méthodes à base de graphe. Si dans ces systèmes, les méthodes à base de graphe permettent d'enrichir les profils d'experts construits automatiquement à l'aide de méthodes d'apprentissage, la prise en compte de la réputation des individus à l'aide d'une validation par les pairs reste anecdotique. En effet, dans les systèmes de recherche d'experts actuels, l'analyse de la proximité entre les profils d'experts et les requêtes est favorisée.

Dans le cadre du projet PCU (Plateforme de Connaissances Unifiées) visant à produire une plateforme *open source* industrielle de valorisation des données de l'entreprise, le cas d'application de la recherche d'experts a été considéré. Le projet et les travaux de thèse présentés dans ce manuscrit ont été financés par un Fonds Unique Interministériel (FUI). Des partenaires académiques (LIPN [5], ESILV [6]) et industriels (Smile [7], Proxem [8],

5. Laboratoire d'Informatique de Paris Nord (LIPN) : `https://lipn.univ-paris13.fr`
6. École supérieure d'ingénieurs Léonard-de-Vinci (ESILV) : `https://www.esilv.fr`
7. Smile : `https://www.smile.eu/fr`
8. Proxem : `https://www.proxem.com`

Armadillo [9], Wallix [10]) ont contribué au projet, soutenus par le pôle de compétitivité Systematic Paris-Région [11] – GTLL [12] ainsi que par la Bpifrance [13] et la Région Île-de-France [14].

Dans le cadre de ces travaux de thèse, nous avons posé la problématique suivante : comment détecter des experts et leurs expertises associées à partir de documents en prenant en compte une validation par les pairs ? Considérons un graphe d'expertise, c'est-à-dire un graphe représentant des experts potentiels reliés entre eux par une relation sociale, par exemple une relation de citation. Considérons que le graphe d'expertise est attribué, c'est-à-dire que les experts potentiels sont étiquetés par leurs expertises ainsi que d'autres caractéristiques telles que leurs années de publication, par exemple. Notre hypothèse est la suivante : en se focalisant sur les zones denses d'un graphe d'expertise, il serait possible de détecter des individus considérés comme experts sur un ensemble de thématiques ainsi que leurs caractéristiques communes maximales.

Nous proposons une approche d'extraction de connaissances à partir de textes combinant des méthodes de fouille de texte classiques et une méthode émergente de fouille de graphes attribués, l'abstraction de graphe. Nous proposons d'appliquer cette approche au cas d'usage de la recherche d'experts dans le milieu académique. À partir d'un corpus de textes qui recèle des connaissances sur les expertises d'une communauté, les expertises et liens de collaboration entre experts sont extraits. Nous obtenons alors un corpus annoté. Ce corpus annoté est représenté sous forme de graphes attribués, c'est-à-dire de structures de données représentant un ensemble de sommets reliés entre eux et étiquetés par un ensemble d'attributs.

L'ensemble des graphes attribués que nous construisons permet de décrire les experts, les expertises ou les documents et de représenter les différents liens existant entre eux. Par exemple, le graphe de coauteurs permet de représenter les chercheurs et de les relier s'ils ont au moins une publication commune. Un autre exemple est celui du graphe de citation entre auteurs qui nous permet de représenter les chercheurs et les liens dirigés de citation existant entre eux. Nous appliquons sur l'ensemble des graphes attribués une méthode de fouille de graphe attribué afin de nous focaliser sur les zones denses. Cette méthode nous permet d'identifier des communautés d'experts sur une thématique particulière ainsi que leur ensemble d'expertises communes.

Enfin, dans un souci d'interopérabilité, nous nous intéressons à la représentation de l'ensemble des graphes attribués sous forme de graphe de connaissances. Les graphes de connaissances sont des réseaux représentant des entités du monde réel reliées entre elles. Ils

9. Armadillo : `https://www.armadillo.fr`
10. Wallix : `https://www.wallix.com/fr`
11. Systematic Paris-Région : `https://systematic-paris-region.org`
12. Groupe Thématique Logiciel Libre (GTLL) du Pôle Systematic Paris-Région
13. Bpifrance `https://www.bpifrance.fr`
14. Région Île-de-France : `https://www.iledefrance.fr`

permettent d'organiser de façon efficiente les connaissances d'un domaine pour en faciliter l'exploitation ultérieure. Le formalisme de représentation d'un graphe de connaissances est un ensemble de triplets RDF, standard du web sémantique. Parmi les graphes de connaissances les plus connus peuvent être cités le Google Knowledge Graph, Wikipedia, ou encore WordNet par exemple. Les graphes de connaissances sont utilisés dans de nombreux domaines, notamment dans le domaine académique. En effet, des graphes de connaissances scientifiques ont récemment été construits afin d'organiser les connaissances concernant les publications scientifiques, pour des applications telles que la recherche d'experts ou la recommandation de collaborateurs. Il est donc essentiel de proposer une méthode nous permettant de traduire les graphes attribués en un graphe de connaissances scientifique, respectant le formalisme RDF.

Nous avons proposé d'expérimenter notre méthode sur un corpus de publications scientifiques : le corpus ACL Anthology. Nous utilisons également un corpus de taille modeste pour illustrer notre approche, le corpus des Ateliers Recherche d'Information SEmantique (RISE). Ce dernier est rédigé principalement en langue française sur les thématiques de recherche d'information, web sémantique, extraction de connaissances, traitement automatique des langues naturelles et multimédia. Le corpus ACL Anthology a une taille plus importante et est rédigé en langue anglaise. Il aborde les thématiques de la linguistique informatique et du traitement de la langue naturelle. Pour évaluer notre méthode, nous proposons un cadre d'évaluation original. En effet, si des jeux de données d'évaluation sont disponibles, ils sont plutôt adaptés au cadre de la recherche d'experts dans le milieu de l'entreprise. Le jeu de données TU, adapté à l'évaluation de méthodes de recherche d'experts à partir de publications scientifiques, ne semble malheureusement plus maintenu à ce jour. Nous proposons donc un jeu d'évaluation original, construit à partir de chapitres de livre décrivant l'état de l'art de différentes thématiques. Nous évaluons les résultats obtenus à partir du corpus ACL Anthology à l'aide de cette méthode d'évaluation. Nous considérons que les experts d'une thématique retrouvés par le système sont validés s'ils ont publié des travaux appartenant aux références du chapitre de livre associé à la même thématique. Trois thématiques sont évaluées : l'extraction d'information (*information extraction*), l'analyse syntaxique (*syntactic parsing*) et la désambiguïsation lexicale (*word sense disambiguation*). Les métriques associées à notre méthode d'évaluation sont la précision, le rappel et la f-mesure. Nous proposons également une évaluation de notre méthode et une comparaison aux principales *baselines* du domaine à l'aide de la plateforme d'évaluation LT ExpertFinder.

Dans le chapitre 1, nous décrivons l'état de l'art de la recherche d'experts. Dans le chapitre 2, nous abordons l'état de l'art de la représentation de connaissances sous forme de graphes, notamment de graphes de connaissances scientifiques et de graphes attribués pertinents pour la recherche d'experts. Dans le chapitre 3, nous décrivons plus en détails la méthode émergente de fouille de graphe que nous avons utilisée, l'abstraction de

graphe, nos apports à cette méthode ainsi que la manière dont nous l'avons appliquée à notre problématique. Puis, dans le chapitre 4, nous détaillons l'approche que nous proposons à la lumière de cet état de l'art. Nous y détaillons également le formalisme RDF que nous avons employé pour représenter les connaissances extraites à partir de publications scientifiques sous forme de graphe de connaissances scientifique, en utilisant les graphes attribués comme support de connaissances intermédiaire. Afin de faciliter la compréhension du lecteur, nous illustrons notre approche par des exemples sur le corpus des ateliers Recherche d'Information SEmantique (RISE). Dans le chapitre 5, nous proposons des expérimentations le corpus ACL Anthology et décrivons les résultats obtenus. Dans le chapitre 6, nous évaluons les résultats obtenus lors des expérimentations décrites dans le chapitre 5, discutons de la pertinence de notre approche et proposons des pistes de réflexion supplémentaires que nous synthétisons dans la conclusion de ces travaux.

Première partie

Proposition scientifique

Chapitre 1

Recherche d'experts

La recherche d'experts consiste en l'identification d'un ensemble d'individus que l'on considère comme experts d'une thématique particulière. Il s'agit d'une problématique historiquement liée à celle du profilage d'experts (Balog, De Rijke et al. 2007 ; S. Lin et al. 2017). Cette dernière implique d'identifier des expertises et de les assigner aux individus adéquats (Bordea 2013). Une expertise est définie comme « une compétence, connaissance, aptitude ou l'un des comportements d'un individu » (Draganidis et Mentzas 2006). La recherche d'experts est principalement motivée par l'identification d'une source d'information fiable (expert, document source) ou par l'identification d'un individu capable d'exercer une fonction précise (Yimam-Seid et Kobsa 2003 ; Balog, Azzopardi et De Rijke 2006), telle qu'une expertise de projet ou la résolution d'un problème posé par exemple. Il est souvent utile de solliciter un expert lors d'évaluations, pour compléter ou remplacer les informations précédemment obtenues à partir de documents ou de bases de données (Yimam-Seid et Kobsa 2003).

Dans le milieu académique en particulier, l'identification d'individus considérés comme experts sur un sujet de recherche spécifique est utile dans de nombreuses applications (Deng, King et Lyu 2008). Par exemple, une problématique récurrente consiste à identifier des relecteurs compétents lors de la constitution de comités de lecture (Rodriguez et Bollen 2008 ; Mimno et McCallum 2007). En effet, l'acceptation d'articles scientifiques au sein des conférences est très largement soumise à un processus de sélection consistant en la relecture des articles par un comité de lecture. Ce processus de sélection est basé sur un système de validation par les pairs. Une autre application importante de la recherche d'experts dans le milieu académique est l'identification d'experts pour le montage de comités d'évaluation de projets de recherche et l'attribution de financements (Hettich et Pazzani 2006). Dans le milieu de l'entreprise, la recherche d'experts est également une tâche essentielle (Mockus et Herbsleb 2002). Par exemple, des problématiques récurrentes consistent à identifier un expert interne pour l'assigner à un projet, ou encore à faire occasionnellement appel à un spécialiste externe pour résoudre un problème spécifique. Cependant, si l'identification des experts constitue une tâche essentielle, il s'agit également d'une tâche difficile. En effet, les connaissances concernant les experts et leurs expertises associées sont rares, coûteuses à acquérir, difficiles à estimer et en constante évolution (Maybury 2006). Dans ce chapitre, nous abordons l'état de l'art de la recherche d'experts.

1.1 Auto-assignation des expertises

Initialement, les systèmes de recherche d'experts étaient basés sur l'auto-assignation des expertises à partir d'une sélection de mots-clefs prédéfinis (Angelova, Boeva et Tsiporkova 2017). Cette méthode repose sur la capacité des experts à fournir une description détaillée et complète de leurs expertises mais également à maintenir cette description à jour au gré des évolutions professionnelles et de l'acquisition de nouvelles expertises (Yimam-Seid et Kobsa 2003). Cependant, les descriptions fournies par les experts sont rarement complètes et habituellement un peu trop générales tandis que les expertises recherchées lors de requêtes sont généralement riches, spécifiques et de granularité très fine (Yimam-Seid et Kobsa 2003). En effet, les requêtes peuvent aller de « Quels sont les experts en programmation logique ? » à « Quels sont les individus capables de modifier le microcode de gestion du vecteur d'interruption dans le module de redémarrage du processeur XZY999 ? » comme indiqué dans un exemple de l'état de l'art (Kautz, Selman, Milewski et al. 1996). L'exemple précédent illustre bien la variabilité de granularité au sein des requêtes dans le cadre de la recherche d'experts ainsi que l'impossibilité pour un individu de maintenir une description exhaustive de ses propres expertises. En outre, la manière de définir une même expertise ainsi qu'un même niveau de maîtrise varie également selon les individus (Mockus et Herbsleb 2002). De plus, les expertises des individus évoluent constamment (Mockus et Herbsleb 2002). L'auto-assignation des expertises est donc une méthode très coûteuse en temps humain pour les experts et en ressources pour les institutions, fournissant des données incomplètes et rapidement obsolètes si un effort de maintenance continu n'est pas fourni (Yimam-Seid et Kobsa 2003). D'autres méthodes préliminaires se basaient sur la construction manuelle de bases de données listant des experts par catégorie (Davenport, Prusak et al. 1998). Cette méthode est tout aussi coûteuse que l'auto-assignation des expertises car elle nécessite également un effort de construction important ainsi qu'un effort de maintenance continu.

1.2 Automatisation de la recherche d'experts

Afin d'automatiser la recherche d'experts et de s'affranchir d'une partie des problématiques de coût élevé en ressources et en temps posées par l'auto-assignation des expertises, d'autres sources de connaissances ont été explorées. Les e-mails ont d'abord été exploités (Campbell et al. 2003). Des approches basées sur le contenu permettant d'extraire les expertises au sein du texte ont été employées, parfois combinées à une analyse des modes de communication entre utilisateurs (Campbell et al. 2003). Par exemple, l'algorithme *Hyperlink-Induced Topic Search* (HITS) (Kleinberg 1999) a été largement utilisé. Ce dernier tire parti de la notion d'autorité dans un graphe. Dans le cas applicatif considéré, un utilisateur peut être considéré comme une autorité s'il répond régulièrement à des

questions de ses pairs par e-mail. Considérons un graphe représentant les utilisateurs et établissant un lien dirigé d'un utilisateur u_1 vers un utilisateur u_2 si u_1 pose une question par e-mail à u_2. Une autorité est un sommet possédant un grand nombre de liens entrants. Les documentations de logiciels et les données issues de systèmes *change management* en particulier (Mockus et Herbsleb 2002) ont également été utilisées dans le milieu de l'entreprise. Par exemple, l'expertise d'un individu peut être quantifiée par son expérience, c'est-à-dire par le nombre de modifications qu'il apporte au code source ou à la documentation. Le système P@NOPTIC (Craswell et al. 2001) a été l'un des premiers à permettre l'extraction automatique d'expertise à partir de données hétérogènes (Deng, King et Lyu 2008). Le système considère les employés les plus souvent mentionnés dans le contexte d'une technologie comme experts de cette technologie. L'originalité résidait dans sa capacité à analyser l'ensemble des documents d'un réseau intranet.

Si les premières méthodes de recherche d'experts étaient focalisées sur des domaines spécifiques et des formats de données précis, des efforts ont été produits pour s'orienter vers des méthodes automatiques de recherche d'experts à partir de documents hétérogènes (Balog, Azzopardi et De Rijke 2006). De nombreuses méthodes de recherche d'experts ont émergé suite aux tâches *Enterprise Track* de la *Text REtrieval Conference* (TREC) ayant eu lieu à partir de 2005 jusqu'en 2008 (Msr, Craswell et De Vries 2005 ; Soboroff, Vries et Craswell 2006 ; Bailey et al. 2007 ; Balog, Thomas et al. 2008). Dans le cadre de ces tâches, la recherche d'experts a été considérée comme une tâche de recherche d'information (Balog, Azzopardi et De Rijke 2006). Le principe était le suivant : à partir de données extraites du Web, d'une liste d'individus considérés comme des experts potentiels et d'un ensemble d'expertises, la tâche consiste à déterminer les experts associés à chaque expertise. Les tâches *Enterprise Track* des conférences TREC ont fourni une plateforme commune aux chercheurs (Balog, Azzopardi et De Rijke 2006), leur permettant d'évaluer les méthodes de recherche d'experts.

1.3 Évaluation des méthodes de recherche d'experts

Nous introduisons l'évaluation des méthodes de recherche d'experts afin de pouvoir fournir au lecteur une estimation compréhensible des performances des méthodes de recherche d'experts de l'état de l'art que nous présentons dans la suite de ce manuscrit. Nous présentons d'abord les jeux de données les plus utilisés par la communauté des chercheurs en recherche d'experts pour évaluer leurs méthodes, ainsi que les métriques associées.

1.3.1 Jeux de données d'évaluation

Pour évaluer les méthodes de recherche d'experts, plusieurs jeux de données sont communément employés. Les plus connus sont les corpus *Enterprise Track* de TREC (Msr,

Craswell et De Vries 2005 ; Soboroff, Vries et Craswell 2006), le jeu de données TU (Tilburg University) (Berendsen et al. 2013) qui est une version améliorée de UvT (Universiteit van Tilburg) (A. Bogers et Balog 2007) ainsi que le corpus ACL Anthology (Radev et al. 2013).

Les corpus *Enterprise Track* issus des conférences TREC (Text REtrieval Conference) ont été utilisés dans le cadre de la recherche d'experts dans le milieu de l'entreprise (Balog, Azzopardi et De Rijke 2006). Les éditions 2005 et 2006 sont publiquement accessibles [1]. Deux éditions supplémentaires ont également été tenues en 2007 et 2008 mais les jeux de données associés ne sont plus disponibles. Le jeu de données utilisé en 2006 est le même que celui employé en 2005. Les données extraites du web lors de la première édition de la tâche correspondent à un *crawl* des pages web du *World Wide Web Consortium* (W3) [2] datant de juin 2004 et comportant 331037 documents hétérogènes (Msr, Craswell et De Vries 2005). Parmi ces documents figurent des emails, du code, des pages wiki ainsi que des pages web personnelles, par exemple.

Quant à TU, il s'agit d'un jeu de données créé en 2008 et fondé sur une base de données regroupant des informations sur 1147 chercheurs et enseignants affiliés à l'Université de Tilbourg aux Pays-Bas. Pour certains d'entre eux, une description de leurs thématiques de recherche ainsi qu'une liste de publications scientifiques sont disponibles. De plus, les experts potentiels représentés par ce jeu de données ont auto-assigné leurs expertises parmi une liste de 2507 expertises. TU est un corpus multilingue, disponible en néerlandais et en anglais. Certaines expertises sont décrites en néerlandais mais pas en anglais. En revanche, toutes les descriptions anglaises ont un équivalent néerlandais. TU est accessible sur demande [3]. L'intérêt principal de ce jeu de données consiste en la liste d'experts associée à chaque thématique de recherche constituant un *gold standard* très utile pour évaluer les performances d'un système de recherche d'expert. De plus, lorsque cela est possible, un ensemble de thématiques similaires est associé à chaque thématique de publication. Il s'agit de l'un des seuls jeux de données pour la recherche d'experts dans le milieu académique associé à un ensemble d'experts validés par une annotation manuelle [4].

Enfin, ACL Anthology est composé de publications scientifiques sur les thématiques de la linguistique informatique et du traitement du langage naturel, publiées lors des conférences ACL (*Association for Computational Linguistics*) (Radev et al. 2013). Il s'agit du corpus par défaut de LT Expertfinder, qui est le plus récent cadre d'évaluation des méthodes de recherche d'experts à ce jour (Fischer, Remus et Biemann 2019). LT

1. Corpus Enterprise Track issus des conférences TREC : `https://trec.nist.gov/data/enterprise.html`

2. World Wide Web Consortium (W3) : `https://www.w3.org`

3. Jeu de données TU : `https://ilps.science.uva.nl/resources/tu-expert-collection/`

4. Nos demandes d'accès à ce jeu de données par le biais du laboratoire *Information and Language Processing Systems* (ILPS) de l'Université d'Amsterdam et du Dr Berendsen ont été infructueuses. Le jeu de données n'est probablement plus maintenu à ce jour.

ExpertFinder permet d'utiliser des jeux de données interchangeables mais est configuré par défaut avec le jeu de données ACL Anthology (Radev et al. 2013). Il existe plusieurs versions de ce jeu de données (Bird et al. 2008 ; Radev et al. 2013 ; Gábor, Haïfa Zargayouna et al. 2016). LT ExpertFinder permet de réaliser une comparaison des résultats obtenus avec de nombreuses méthodes issues de l'état de l'art. Cependant, aucun *gold standard* n'est associé au corpus ACL Anthology. Il est donc difficile d'estimer la qualité et l'exhaustivité des experts retrouvés par un système de recherche d'experts sur ce jeu de données. Nous ne pouvons obtenir qu'un ensemble de résultats ordonnés pour chacune de ces méthodes.

1.3.2 Métriques d'évaluation

Des métriques pour l'évaluation des méthodes de recherche d'experts ont été proposées lors des conférences TREC, comme mAP (*mean average precision*) qui est devenu un standard dans de nombreuses communautés (détection d'objet dans le domaine de la vision par ordinateur, recherche d'information). La métrique MRR (*mean reciprocal rank*) est également très utilisée.

mAP est basé sur la métrique AP (*average precision*), elle-même basée sur la précision et le rappel. Nous utiliserons un exemple pour illustrer les différentes métriques :

Exemple. Soit q, une requête de la forme « Quels sont les experts sur l'expertise X ? ». Considérons un ensemble d'experts E_{GS} validés par le gold standard sur la requête q, c'est-à-dire dont l'expertise est validée par le jeu d'évaluation. Soit $E_{GS} = \{x, y\}$. Considérons également un système à évaluer qui répond à la requête q en proposant un ensemble d'experts ordonnés, E_S, avec $E_S = \{w, x, y, z\}$. On note que parmi les experts retrouvés par le système, x et y appartiennent également aux experts suggérés par le gold standard sur la requête q.

Précision

La précision permet d'évaluer la qualité des résultats obtenus, c'est-à-dire la proportion d'experts retrouvés par le système qui sont réellement pertinents. Le calcul de la précision est le suivant :

$$\text{Précision} = \frac{\text{retrouvés} \cap \text{pertinents}}{\text{retrouvés}}$$

Dans l'exemple considéré, on a retrouvé w, x, y et z et seulement x et y peuvent être considérés comme pertinents. La précision est donc de 0.5. Cependant, la précision ne prend pas en compte l'ordre des réponses à la requête.

Pour prendre en compte l'ordre des réponses à la requête dans le calcul de la précision, la précision au rang k notée P@k a été introduite. Dans l'exemple considéré, on a P@1 = 0, puisque w n'appartient pas au *gold standard*. P@2 = 0.5, puisque w n'appartient pas

au *gold standard* mais x en fait effectivement partie. De même, on a P@3 = 0.66 et P@4 = 0.5.

Rappel

Le rappel permet d'évaluer l'exhaustivité des résultats obtenus, c'est-à-dire la proportion d'experts appartenant au *gold standard* effectivement retrouvés par le système. Le calcul du rappel est le suivant :

$$\text{Rappel} = \frac{\text{retrouvés} \cap \text{pertinents}}{\text{pertinents}}$$

Dans l'exemple considéré, on a retrouvé x et y qui peuvent être considérés comme pertinents. L'ensemble des experts pertinents d'après le *gold standard* est constitué par les experts x et y. Le rappel est donc de 1. Le système permet donc de retrouver l'ensemble des résultats pertinents sur la requête q.

F-mesure

La f-mesure est une mesure globale des performances d'un système, basée sur une combinaison de la précision et du rappel. Le calcul de la f-mesure est le suivant :

$$\text{f-mesure}(q) = 2 * \frac{\text{précision} * \text{rappel}}{\text{précision} + \text{rappel}} \tag{1.1}$$

AP (*average precision*)

La métrique AP (*average precision*) permet d'estimer la capacité d'un système à ordonner les résultats d'une requête. Soit GTP l'ensemble des résultats pertinents retrouvés d'après le *gold standard* et rel@k une fonction renvoyant 1 si le résultat au rang k est pertinent, 0 sinon. Le calcul de AP est le suivant :

$$AP = \frac{1}{|GTP|} \sum_{k=1}^{n} P@k * rel@k$$

Dans l'exemple considéré, on a retrouvé 2 experts qui peuvent être considérés comme pertinents, donc GTP = 2. On a $AP = \frac{1}{2}(0 + 0.5 + 0.66 + 0) = 0.83$. Si x et y avaient été positionnés aux deux premiers rangs de la liste de résultats, on aurait eu $AP = \frac{1}{2}(1 + 1 + 0 + 0) = 1$. Cette mesure permet donc de pénaliser les systèmes qui ne sont pas capables de fournir des résultats pertinents dès les premiers rangs.

mAP (*mean average precision*)

Pour un ensemble de requêtes Q, il est possible de calculer mAP (*mean average precision*) qui est la moyenne des métriques AP pour chacune des requêtes q appartenant à

Q. Cette métrique permet d'obtenir un résultat unique, donnant une idée générale sur les performances moyennes d'un système. Le calcul de mAP est le suivant :

$$mAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AP(i)$$

MRR (*mean reciprocal rank*)

MRR est une mesure permettant d'évaluer des systèmes proposant une liste ordonnée de réponses à une requête. Elle est basée sur le rang réciproque (*reciprocal rank*). Le rang réciproque correspondant à l'inverse de la position de la meilleure réponse (*rank*) dans la liste de réponses proposées par le système pour une unique requête q (de 1 à N, N étant le nombre de réponses) :

$$\frac{1}{rank}$$

Dans le cas où le système ne trouve aucune réponse à la requête ou si le système ne retrouve aucune réponse adéquate à la requête, le rang réciproque est égal à 0. MMR permet de calculer le rang réciproque moyen (*mean reciprocal rank*) pour un ensemble de requêtes Q :

$$\text{MRR} = \frac{1}{|}Q| \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

1.4 Approches centrées sur le profil d'expert ou sur les documents

Deux approches principales émergent quant à la représentation de l'expertise d'un individu (Balog, Y. Fang et al. 2012 ; Cifariello, Ferragina et Ponza 2019). La première est l'approche centrée sur le profil d'expert, la seconde l'approche centrée sur les documents source d'expertise. L'approche centrée sur le profil d'expert consiste à créer pour chaque individu son profil d'expert, c'est-à-dire à lui assigner l'ensemble de ses expertises à partir des documents qui lui sont associés (Balog, Azzopardi et De Rijke 2006 ; Van Gysel, Rijke et Worring 2016). Répondre à une requête consiste alors à ordonner les individus en fonction de leurs profils d'expert. Plus le contenu du profil d'expert correspond à la requête, mieux l'individu est classé. L'approche centrée sur le profil d'expert est indépendante des requêtes (Petkova et Croft 2008). L'approche centrée sur les documents consiste quant à elle à explorer les documents sources d'expertise qui correspondent au mieux à la requête puis à leur associer les individus qui en sont les auteurs (Cao et al. 2005 ; Macdonald et Ounis 2006b ; Balog, Y. Fang et al. 2012). Cette approche ne permet pas de modéliser directement les expertises d'un individu et est

donc dépendante des requêtes (Petkova et Croft 2008). Il est également nécessaire de combiner le score de pertinence des documents au degré d'implication des individus dans la rédaction de ceux-ci. En effet, tous les coauteurs d'un document ne peuvent pas être considérés comme uniformément experts de son contenu. Cependant, l'approche centrée sur les documents permet de mesurer l'impact d'un document particulier, ce que les approches centrées sur les profils d'expert ne permettent pas (Deng, King et Lyu 2008). Des approches hybrides ont donc été proposées (Serdyukov et Hiemstra 2008 ; Balog et De Rijke 2008 ; Van Gysel, Rijke et Worring 2016). Dans l'approche hybride, les individus sont uniformément considérés comme experts des documents dont ils sont les auteurs. La présence régulière d'un auteur dans les documents correspondant au mieux à une requête est un indicateur supplémentaire de son expertise (Serdyukov et Hiemstra 2008). Une approche hybride combinant approches centrées sur les documents et les profils d'experts a été proposée et testée sur les éditions 2007 et 2008 des tâches *Enterprise Track* de la conférence TREC (Balog et De Rijke 2008). Les performances de l'approche hybride ont été comparées à celles d'une approche centrée sur les documents ainsi qu'à celles d'une approche centrée sur les profils d'experts sur l'ensemble des métriques associées. Les métriques évaluées sont les suivantes : mAP (*mean average precision*) et MRR (*mean reciprocal rank*). Sur l'ensemble des métriques, l'approche hybride surpasse substantiellement les deux autres.

1.5 Indicateurs d'expertise

Une problématique essentielle de la recherche d'experts consiste à déterminer les indicateurs d'expertise (Macdonald et Ounis 2006b ; Balog, Azzopardi et De Rijke 2006 ; Balog, Azzopardi et Rijke 2009 ; Van Gysel, Rijke et Worring 2016). Les méthodes de recherche d'experts de l'état de l'art exploitent des indicateurs d'expertise variés. L'expertise d'un individu est très largement associée à sa mention dans des documents sources d'expertise (Petkova et Croft 2008 ; Balog, Y. Fang et al. 2012). Plus un individu est proche dans le texte de la mention d'une expertise, plus sa probabilité d'être associé à l'expertise est élevée (Balog, Azzopardi et Rijke 2009). Les auteurs d'un document sont généralement considérés comme experts de leur contenu (Balog, Y. Fang et al. 2012). Cependant, lorsque les documents sont longs et impliquent plusieurs auteurs dont certains n'ont rédigé qu'une portion du document, l'indicateur de paternité d'un texte n'est plus aussi pertinent (Balog, Y. Fang et al. 2012).

D'autres indicateurs peuvent alors être pris en compte. Par exemple, les modes de communication entre individus peuvent constituer un fort indicateur d'expertise dans le cadre de la recherche d'experts à partir de l'analyse d'e-mails (Campbell et al. 2003). Plus généralement, les interactions sociales entre individus peuvent constituer des indicateurs d'expertise fiables (Jun Zhang et Ackerman 2005 ; Jing Zhang, Tang et J. Li 2007 ;

Jun Zhang, Ackerman et Adamic 2007 ; Ehrlich, C.-Y. Lin et Griffiths-Fisher 2007 ; Fu et al. 2007b ; J. Li et al. 2007 ; Smirnova et Balog 2011 ; Tang et al. 2008 ; Balog, Y. Fang et al. 2012). Par exemple, les systèmes questions-réponses au sein des communautés en ligne, les graphes de citation, les réseaux sociaux scientifiques véhiculent des indicateurs d'expertise liés aux modes de communication et aux interactions sociales. Plus un individu produit de travaux cités, plus il répond de façon pertinente aux questions posées, plus il est probable qu'il s'agisse d'un expert.

Des indicateurs liés à la qualité du document peuvent également être pris en compte (par exemple, le nombre de citations du document). De même, il est rare que l'ensemble des coauteurs contribuent de façon équivalente à la rédaction d'un document. Par exemple, lorsqu'une publication scientifique est rédigée par un doctorant, ses encadrants et son directeur de thèse, il est probable que le directeur de thèse ait une vision plus large du domaine. Il est également probable que les autres coauteurs aient une meilleure expertise sur les aspects techniques de la proposition scientifique (Hashemi, Neshati et Beigy 2013). Des travaux ont donc suggéré de prendre en compte l'implication d'un auteur dans la rédaction d'un document. Pour cela, des caractéristiques structurelles, temporelles et basées sur l'activité des auteurs sont prises en compte (Hashemi, Neshati et Beigy 2013). Par exemple, pour l'auteur a_i de la publication scientifique soumise à la conférence c_i, le nombre de coauteurs de a_i dans c_i est un indicateur structurel permettant d'évaluer l'importance de a_i au sein de la liste de coauteurs. Le nombre de publications précédemment publiées ou le nombre d'années d'exercice constituent des indicateurs temporels. Enfin, le nombre de citations moyen de chaque publication associée à a_i constitue un exemple d'indicateur basé sur l'activité de l'auteur.

1.6 Sources d'expertise

Nous identifions trois sources principales de connaissances pour la recherche d'experts : le texte, les réseaux sociaux et les communautés en ligne.

1.6.1 Recherche d'experts à partir de texte

Le texte est devenu une source de connaissances majeure pour la recherche d'experts (Petkova et Croft 2008 ; Balog, Y. Fang et al. 2012 ; Al-Taie, Kadry et Obasa 2018). Le profilage d'experts à partir des *curriculum vitæ* (CV) ou des pages web personnelles au travers desquelles les individus fournissent des informations concernant leurs expertises soulève les mêmes problématiques que celles de la méthode d'auto-assignation précédemment décrite. En effet, les experts doivent maintenir leur CV ou leur page web à jour et fournir des descriptions riches et précises de leurs expertises, ce qui n'est généralement pas le cas. D'autres sources de connaissances textuelles sont donc considérées.

Les documents de travail sont principalement utilisés pour l'automatisation de la recherche d'experts. Pour les entreprises, il s'agit essentiellement des rapports d'activité. Dans le milieu académique, grâce à l'explosion du stockage de données en ligne et à la masse de données scientifiques ouvertes désormais disponible, les publications scientifiques constituent une source de données textuelle importante pour la recherche d'experts (Khan et al. 2017 ; Xia et al. 2017). D'autres sources de données textuelles peuvent également être explorées, comme les emails, messages en ligne ou articles de blog (Al-Taie, Kadry et Obasa 2018) par exemple.

L'acquisition d'emails ou de documents de travail décrivant les experts et expertises au sein d'une entreprise sont des tâches délicates pour des raisons de confidentialité. D'autre part, dans le milieu académique, les publications scientifiques ont l'avantage de constituer une masse de documents fiables, produits par l'ensemble de la communauté scientifique, porteurs d'expertises riches et précises et généralement faciles d'accès dans le cas où les données sont ouvertes. De plus, elles recèlent également les relations de collaboration scientifique entretenues au sein de la communauté scientifique. En effet, les références d'une publication matérialisent des relations de citation tandis que la liste des auteurs des publications matérialise les relations de coauteurs. Ces relations peuvent être analysées pour étudier les modes de collaboration des individus au sein de la communauté scientifique, à la manière de l'analyse des modes de communication entre utilisateurs au sein des emails (Campbell et al. 2003). Cependant, les méthodes de recherche d'experts à partir de texte sont plus généralement des approches basées sur le contenu du texte. Elles seront explicitées de manière détaillée lors de la présentation des différentes méthodes de recherche d'experts.

1.6.2 Recherche d'experts à partir de l'analyse des réseaux sociaux

Le texte n'est pas la seule source de connaissances disponible pour la recherche d'experts. L'analyse des réseaux sociaux constitue également une source de connaissances majeure pour la recherche d'experts (Balog, De Rijke et al. 2007).

Si l'investissement dans les réseaux sociaux peut sembler une activité frivole au premier abord pour un chercheur (Ovadia 2014), il s'agit d'un indicateur efficace de son niveau d'expertise (M.-C. Yu et al. 2016). Dans le cas du RG score par exemple [5] (M.-C. Yu et al. 2016), le niveau d'expertise est basé sur la réception des contributions scientifiques (publications, réponses pertinentes à des questions) par les pairs [6]. Plus les pairs ayant un RG score élevé entretiennent des relations de collaborations scientifiques avec un individu (publications communes, citations, évaluations positives), plus le niveau d'expertise de

5. Indicateur issu du réseau social scientifique ResearchGate

6. Comprendre le RG Score de ResearchGate : `https://explore.researchgate.net/display/support/RG+Score`

l'individu est estimé comme étant également élevé.

Prendre en compte les liens de collaboration scientifique d'un expert (copublication, citation, *etc.*) permet donc de prendre en compte une validation par les pairs, très répandue dans le milieu académique. À partir d'un réseau social académique, c'est-à-dire d'un réseau social dont la cible est la communauté scientifique, les profils d'expert ainsi que les relations entre les individus peuvent être pris en compte pour extraire les experts d'une thématique (Jing Zhang, Tang et J. Li 2007).

Les limites de l'analyse des réseaux sociaux tiennent à l'investissement des scientifiques dans ces derniers. En effet, l'ensemble des chercheurs ne sont pas inscrits ou actifs sur les réseaux sociaux. Malgré la pertinence des indicateurs d'expertise qui sont issus de ces derniers (M.-C. Yu et al. 2016), ils ne permettent pas de définir à eux seuls le niveau d'expertise de l'ensemble des individus de la communauté scientifique.

1.6.3 Recherche d'experts à partir de l'analyse des communautés en ligne

Les communautés en ligne forment des lieux virtuels sur Internet où les individus se rassemblent pour chercher ou partager des connaissances sur des sujets d'intérêt communs (Al-Taie, Kadry et Obasa 2018). Parmi les plateformes de questions-réponses les plus populaires nous pouvons citer Yahoo Answers ![7] ou Stack Overflow [8] par exemple. Ces plateformes sont enrichies par *crowdsourcing*, les utilisateurs posant et répondant à des questions d'autres utilisateurs. Elles ont suscité un fort intérêt parmi la communauté des chercheurs en recherche d'experts. Le contenu des messages échangés ainsi que l'analyse des modes de communications entre utilisateurs peuvent permettre d'identifier des requêtes (des questions) et des experts (des utilisateurs ayant fourni une réponse satisfaisante, ou des utilisateurs susceptibles de répondre à une question donnée).

Cependant, les connaissances véhiculées dans les communautés scientifiques en ligne ou les systèmes de questions-réponses ne sont pas toujours bien structurées et ont donc tendance à être de mauvaise qualité (G. A. Wang et al. 2013 ; Al-Taie, Kadry et Obasa 2018). La qualité de l'information disponible au sein des communautés en ligne dépend fortement de la taille de ces dernières. Plus la communauté est grande, moins la qualité de l'information disponible est élevée (Gu et al. 2007).

1.7 Méthodes de recherche d'experts

D'après la littérature, les méthodes de recherche d'experts principalement utilisées sont les méthodes basées sur l'apprentissage et les méthodes à base de graphes (Al-

7. Yahoo Answers ! : `https://answers.yahoo.com`
8. Stack Overflow : `https://stackoverflow.com`

Taie, Kadry et Obasa 2018). les méthodes de recherche d'experts dépendent fortement de l'approche d'automatisation de la recherche d'experts adoptée. Les méthodes basées sur l'apprentissage sont principalement utilisées dans le cadre de l'automatisation de la recherche d'experts à partir de documents textuels (Petkova et Croft 2008) tandis que les méthodes à base de graphes sont principalement employées dans le cadre de la recherche d'experts à partir de réseaux sociaux (Jing Zhang, Tang et J. Li 2007), ou plus généralement lorsque les indicateurs d'expertise exploitent des interactions sociales. Dans la figure 1.1, nous présentons la hiérarchie existant entre les méthodes de recherche d'experts issues de l'état de l'art que nous présentons dans la suite de ce manuscrit.

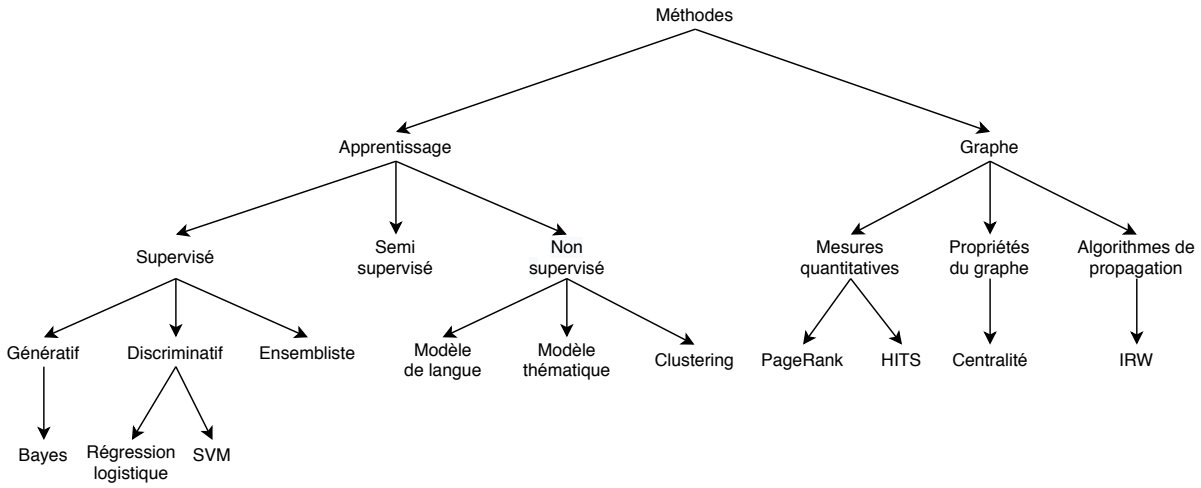


Figure 1.1 – Méthodes de recherche d'experts

Pour pouvoir comparer les performances de ces méthodes, nous nous basons sur les résultats obtenus sur les éditions 2005 à 2007 des tâches *Enterprise Track* de la conférence TREC. Dans la table 1.1, nous rappelons le tableau récapitulatif des performances des trois meilleurs systèmes officiels et de quelques systèmes notables supplémentaires en précisant la méthode utilisée. Ce tableau est issu de travaux existants (Balog, Azzopardi et Rijke 2009). Dans la suite de ce manuscrit, nous faisons référence à cette table lorsqu'une méthode de l'état de l'art que nous présentons a été testée sur les tâches *Enterprise Track*. Ainsi, nous pouvons estimer les performances d'une méthode et les comparer par rapport aux autres méthodes de l'état de l'art. Bien que les corpus des éditions 2005 et 2006 soient identiques, les performances de systèmes ayant concouru durant les deux éditions peuvent varier, l'outil d'évaluation ayant été modifié entre les deux éditions. De plus, les résultats présentés dans cette table correspondent aux meilleurs résultats obtenus par les systèmes, non à leurs performances moyennes.

Méthode	TREC 2005		TREC 2006		TREC 2007	
	mAP	MRR	mAP	MRR	mAP	MRR
TREC Enterprise 2005-2007 top 3						
2005 1^{er} Apprentissage ensembliste (Fu, W. Yu et al. 2005)	**.2749**	**.7268**				
2005 2^{eme} Modèle de langue et *clustering* (Cao et al. 2005)	.2688	.6344				
2005 3^{eme} *Clustering* (He et Zhifeng Yang 2005)	.2174	.6068				
2006 1^{er} Modèle de langue et *clustering* (Zhu, Song, Rüger et al. 2006)			**.6431**	**9609**		
2006 2^{eme} Modèle de langue (Bao et al. 2006)			.5947	.9358		
2006 3^{eme} Modèle de langue (You et al. 2006)			.5639	.9043		
2007 1^{er} Apprentissage ensembliste (même système que 1^{er} 2005) (Fu, Xue et al. 2007)					**.4632**	**.6333**
2007 2^{eme} PageRank (Duan et al. 2007)					.4427	.6131
2007 3^{eme} Modèle de langue et *clustering* (même système que 1^{er} 2006) (Zhu, Song et Rüger 2007)					.4337	.5802
Autres approches notables						
Modèle génératif (Model 1) (*run* officiel) (Balog, Azzopardi et De Rijke 2006)	.1883	.4692	.3206	.7264		
Modèle génératif (Model 1B) (Balog, Azzopardi et Rijke 2009)	**.2725**	**.6800**	.4291	.8912	**.4633**	**.6236**
Modèle génératif (Model 2) (Balog, Azzopardi et De Rijke 2006) (*run* officiel)	.2053	.6088	**.4660**	**.9354**		
Apprentissage ensembliste (Macdonald et Ounis 2008)	**.2917**		.5712			
Modèle de langue (Petkova et Croft 2008)	.2850	.6496				
Modèle log-linéaire (Van Gysel, Rijke et Worring 2016)	.248	.618	.484	.833	.344	.513

Table 1.1 – Performances des trois meilleurs systèmes de recherche d'experts et d'autres systèmes notables évalués lors des éditions 2005 à 2007 des tâches *Enterprise Track* de la conférence TREC (Balog, Azzopardi et Rijke 2009)

1.7.1 Méthodes à base d'apprentissage

L'apprentissage automatique est la détection automatique de motifs pertinents dans les données (Shalev-Shwartz et Ben-David 2014). Les méthodes de recherche d'experts

à base d'apprentissage peuvent être organisées en trois groupes en fonction du niveau de supervision de l'apprentissage automatique : les méthodes supervisées, les méthodes semi-supervisées et les méthodes non supervisées.

Méthodes d'apprentissage supervisé

Les méthodes de recherche d'experts à base d'apprentissage supervisé principalement employées consistent en des méthodes de fouille de texte utilisées pour extraire les expertises enfouies dans les documents textuels (Z. Wu et al. 2014 ; Angelova, Boeva et Tsiporkova 2017 ; Al-Taie, Kadry et Obasa 2018). Les méthodes supervisées sont basées sur un modèle entraîné sur un corpus déjà annoté. La qualité des résultats obtenus par le biais de telles méthodes dépend notamment de la qualité du vocabulaire acquis à partir du corpus d'entraînement. Les limites de ces méthodes tiennent donc à la qualité et à la disponibilité du corpus d'entraînement qui doit être le plus exhaustif possible dans son contenu et ses annotations. Les chercheurs ont donc privilégié les méthodes non supervisées, ne nécessitant pas de données d'entraînement annotées (Cifariello, Ferragina et Ponza 2019). Parmi les méthodes supervisées de recherche d'experts existent les modèles probabilistes génératifs et discriminatifs ainsi que les méthodes ensemblistes.

Modèles génératifs D'après les modèles génératifs, les experts potentiels sur une requête sont ordonnés en fonction de leur probabilité $P(e|q)$, avec e un expert, q une requête. Cette probabilité peut être estimée de deux manières différentes : soit directement (*candidate generation models*), soit par refactorisation à l'aide du théorème de Bayes (*topic generation models*) (Balog, Y. Fang et al. 2012 ; H. Fang et ChengXiang Zhai 2007). On obtient alors :

$$P(e|q) = \frac{P(q|e)P(e)}{P(q)}$$

Cette probabilité peut être simplifiée par $P(q|e)P(e)$. En effet, la probabilité $P(q)$ est constante selon une requête donnée. Cette variante du modèle génératif permet donc de dire que la probabilité qu'un individu donné soit un expert sur la requête q est équivalente à la probabilité $P(q|e)$ d'obtenir la requête q lorsque l'on a l'individu e, étant donné la probabilité $P(e)$ que l'individu e soit un expert *a priori* (Balog, Y. Fang et al. 2012). Parmi les modèles génératifs, nous pouvons citer *Model 1* et *Model 2* (Balog, Y. Fang et al. 2012). Le premier a une approche centrée sur les profils d'expert tandis que le second a une approche centrée sur les documents. Les approches basées sur le modèle probabiliste génératif sont largement fondées sur des méthodes de modélisation linguistique (S. Lin et al. 2017). Le principe consiste à représenter chaque document par un ensemble de termes à l'aide d'un modèle de langue et d'ordonner les documents par correspondance des termes qui les représentent avec une requête donnée (Balog, Azzopardi et De Rijke 2006 ;

S. Lin et al. 2017). Par exemple, les algorithmes TF/IDF et BM25 sont largement utilisés (Hu et al. 2006). *Model 2* en particulier est très populaire et est largement utilisé comme méthode de référence pour l'évaluation des méthodes de recherche d'experts (Fischer, Remus et Biemann 2019). En effet, *Model 2* a surpassé *Model 1* sur l'ensemble des métriques associées à l'évaluation des performances des systèmes de recherche d'experts sur le corpus W3C TREC Enterprise lors des éditions 2005 et 2006 de la tâche *Enterprise Track* de TREC (Balog, Azzopardi et De Rijke 2006 ; Msr, Craswell et De Vries 2005). Dans la table 1.1, nous décrivons les résultats obtenus par Model 1 et Model 2 sur les métriques mAP (respectivement 0.1883 en 2005, 0.3206 en 2006 pour Model 1 et 0.2053 en 2005, 0.4660 en 2006 pour Model 2) et MRR (respectivement 0.4692 en 2005, 0.7264 en 2006 pour Model 1 et 0.6088 en 2005, 0.9354 en 2006 pour Model 2). Les performances de Model 2 ont permis de le classer dans le top 5 des systèmes d'experts évalués lors de ces tâches (Balog, Azzopardi et De Rijke 2006). Dans Model 1, les auteurs considèrent que l'ensemble des indicateurs d'expertise que l'on peut extraire à partir d'un document décrivent l'expert potentiel auquel le document est associé. Pour alléger cette contrainte, Model 1B a été proposé. Ce modèle prend en considération la probabilité $P(t|d, e)$ qu'un terme t de la requête q soit associé à un expert potentiel e dans une fenêtre donnée (Balog, Azzopardi et Rijke 2009). D'après la table 1.1, les performances obtenues par Model 1B sur les métriques mAP (0.2725 en 2005, 0.4291 en 2006 et 0.4633 en 2007) et MRR (0.6800 en 2005, 0.8912 en 2006 et 0.6236 en 2007) sont supérieures à celles obtenues par Model 2 sur les éditions 2005 et 2007. Les auteurs démontrent qu'avec une étape de paramétrage de la taille de la fenêtre, Model 1B surpasse Model 2. Cependant, en raison de sa simplicité de mise en œuvre et de ses performances moyennes sur les métriques mAP et MRR, Model 2 reste le modèle phare des trois modèles présentés par les auteurs. Il s'agit également de l'une des méthodes principales du domaine de la recherche d'experts (Fischer, Remus et Biemann 2019).

Modèles discriminatifs Quant aux modèles discriminatifs, ils permettent d'évaluer directement la corrélation entre les experts potentiels e et la requête q par la probabilité $P(e, q)$. D'après le *probability ranking principle* (PRP) issu de la recherche d'information, les documents doivent être ordonnés par ordre décroissant de pertinence (Balog, Y. Fang et al. 2012). La pertinence est calculée selon le ratio suivant :

$$log \frac{P(r = 1|e, q)}{P(r = 0|e, q)}$$

avec r $\in \{1, 0\}$ une variable binaire permettant de mesurer la pertinence (1 si le document est pertinent, 0 sinon) (Balog, Y. Fang et al. 2012). La tâche de recherche d'experts a été transformée en un problème de classification binaire grâce à un modèle discriminatif (Y. Fang, Si et Mathur 2011). Les paires (requête, expert) pertinentes sont considérées

comme des instances positives tandis que les autres sont considérées comme des instances négatives (S. Lin et al. 2017). Lorsque cela est possible, « il est préférable de résoudre un problème [de classification] directement plutôt que de résoudre un problème plus général en tant qu'étape intermédiaire [comme la modélisation de $P(e|q)$] » (Vladimir Vapnik et Vlamimir Vapnik 1998 ; Ng et Jordan 2002). De ce fait, l'emploi de modèles discriminatifs pour résoudre un problème est généralement favorisé à l'emploi de modèles génératifs. Dans la pratique, les modèles discriminatifs sont préférés aux modèles génératifs lorsqu'il y a suffisamment de données d'entraînement disponibles (Ng et Jordan 2002 ; S. Lin et al. 2017). Une régression logistique (B. Li et King 2010) a notamment été employée pour identifier des utilisateurs capables de répondre à des questions dans des systèmes de questions-réponses. Ces utilisateurs sont considérés comme des experts. Les auteurs ont construit automatiquement une évaluation à partir d'utilisateurs extraits de Yahoo ! Answers [9], un système de questions-réponses populaire. Les auteurs ont au mieux obtenu une MRR de 0.0541. Des travaux ont également été produits sur cette même problématique, en utilisant des machines à vecteurs de support, sur un *crawl* différent de Yahoo ! Answers (Zhou, Lyu et King 2012). En sélectionnant 60 % de données d'entraînement pour 40 % de données de test et en combinant un ensemble de caractéristiques globales et locales, les auteurs ont obtenu une f-mesure de 0.6374, une précision de 0.8414 et un rappel de 0.513. L'avantage des machines à vecteurs de support réside dans leur robustesse face à des données bruitées.

Méthodes d'apprentissage ensemblistes D'autres approches d'apprentissage ont été employées, comme les méthodes d'apprentissage ensemblistes basées sur un processus de vote. Les modèles basés sur le vote sont inspirées de techniques de fusion de données (Macdonald et Ounis 2006b). Elles permettent d'associer des experts potentiels à une requête à travers la fusion ou combinaison d'informations (Balog, Y. Fang et al. 2012). Les documents sont ordonnés par ordre de pertinence par rapport à une requête donnée. Chaque document vote pour les individus qu'il mentionne selon des poids déterminés par une stratégie d'agrégation. Différentes stratégies d'agrégation ont été employées pour résoudre le problème de la recherche d'experts (Macdonald et Ounis 2006b). Les méthodes d'apprentissage ensemblistes permettent donc d'agréger un ensemble de votes réalisés par des documents de manière isolée pour obtenir une stratégie de sélection globale plus forte. Il a été démontré que l'utilisation de modèles diversifiés dans le cadre d'une méthode d'apprentissage ensembliste permettait d'obtenir des résultats plus précis (Kuncheva et Whitaker 2003). En effet, les modèles employés sont considérés comme des boîtes noires et seuls les résultats sont agrégés. Aussi, il est possible de combiner des modèles génératifs tels que Model 2 avec des modèles discriminatifs dans le cadre d'une méthode de recherche d'experts ensembliste (Balog, Y. Fang et al. 2012). Les

9. Yahoo ! Answers : `https://answers.yahoo.com`

méthodes d'apprentissage ensemblistes ont presque exclusivement été employées par les meilleurs systèmes des *KDD Cup* des compétitions *Data Mining and Knowledge Discovery* organisées par le *ACM Special Interest Group on Knowledge Discovery and Data Mining* (Balog, Y. Fang et al. 2012). Le meilleur système évalué lors des éditions 2005 et 2007 des taches *Enterprise Track* des conférences TREC est également basé sur une méthode d'apprentissage ensembliste (Fu, W. Yu et al. 2005 ; Fu, Xue et al. 2007). Nous rappelons les performances obtenues par ce système sur les métriques mAP (0.2749 en 2005, 0.4632 en 2007) et MRR (0.7268 en 2005, 0.6333 en 2007) dans la table 1.1. Il est à noter que le système n'a pas concouru en 2006. Il a cependant été comparé au meilleur système de 2006 en 2007 et a obtenu de meilleures performances sur l'ensemble des métriques. Un autre système notable basé sur une méthode d'apprentissage ensembliste a été testé sur la métrique mAP et sur les corpus des éditions 2005 et 2006 (respectivement 0.2917 et 0.5712 d'après la table 1.1). Dans leurs travaux, les auteurs comparent 12 techniques de vote différentes et déterminent l'efficacité du paradigme de vote, en dehors de toute considération du modèle de pondération employé (Macdonald et Ounis 2008). Ils concluent cependant qu'une meilleure représentation initiale du document améliore les performances obtenues à l'aide d'une méthode d'apprentissage ensembliste.

Méthodes d'apprentissage semi-supervisé

Les méthodes d'apprentissage semi-supervisé permettent de réduire le nombre d'exemples étiquetés utilisés lors de l'entraînement du modèle d'apprentissage. Une méthode d'apprentissage semi supervisé pour la recherche d'experts a été proposée, utilisant le renforcement mutuel dans les systèmes de questions-réponses (Bian et al. 2009). Les systèmes de questions-réponses sont des communautés en ligne dans lesquelles les utilisateurs peuvent poser des questions ou répondre à des questions posées par d'autres utilisateurs. Au sein des systèmes de questions-réponses, la qualité du contenu proposé par les utilisateurs varie considérablement et la réputation ou l'expertise d'un contributeur peut permettre de sélectionner des contenus pertinents. Les méthodes précédemment proposées pour identifier des experts au sein des systèmes questions-réponses nécessitaient une forte supervision (Agichtein et al. 2008) ou se basaient sur des méthodes inspirées de l'analyse des réseaux sociaux (Jun Zhang, Ackerman et Adamic 2007). Ces dernières ne permettent pas de considérer la pertinence du contenu des contributions mais se basent sur une analyse des interactions sociales des individus au sein du système. Cependant, dans les systèmes de questions-réponses, il arrive qu'un utilisateur ayant une bonne réputation fournisse une réponse incorrecte et qu'un nouvel utilisateur ou un utilisateur contribuant peu fournisse une excellente réponse. La méthode exploitant le renforcement mutuel permet d'alléger la supervision en combinant à la fois une analyse de la réputation du contributeur et une analyse du contenu de ses contributions (Bian et al. 2009). Elle se base sur différentes caractéristiques permettant d'évaluer la pertinence d'une réponse (par exemple,

le nombre total de réponses fournies par l'utilisateur et sélectionnées comme « meilleure réponse », le nombre de questions posées par l'utilisateur ou encore son score d'autorité). Les auteurs évaluent leur méthode sur un *crawl* de Yahoo ! Answers et obtiennent 0.483 sur la métrique mAP et 0.865 sur la métrique MRR. Cependant, les limites de ce modèle tiennent toujours à la nécessité d'acquérir un nombre d'exemples étiquetés pour entraîner le modèle, même si le nombre d'exemples requis est limité par rapport aux méthodes supervisées.

Méthodes d'apprentissage non supervisé

Les méthodes de recherche d'experts non supervisées ont l'avantage de ne pas nécessiter de données d'entraînement préalablement annotées. Elles sont donc largement répandues dans la littérature (Cao et al. 2005 ; Macdonald et Ounis 2006b ; Balog, Azzopardi et De Rijke 2006 ; Balog, Azzopardi et Rijke 2009 ; Van Gysel, Rijke et Worring 2016). Nous présentons les modèles de langue, les modèles thématiques et les méthodes à base d'analyse de *clusters* de documents ou d'utilisateurs.

Modèles de langue Les modèles de langue sont des modèles statistiques permettant d'estimer une distribution de probabilité sur des séquences de mots. Ces probabilités sont apprises à l'aide d'un large corpus d'entraînement non annoté. Il existe principalement deux catégories de modèles de langue : les modèles de langue statistiques et les modèles de langue basés sur des réseaux de neurones artificiels. Les premiers utilisent des méthodes statistiques traditionnelles et sont constitués par les n-grammes. Des règles de syntaxe ou de sémantique peuvent également être explicitées et appliquées pour estimer la probabilité d'apparition de séquences de mots. Les modèles n-grammes permettent d'estimer la probabilité d'apparition d'un mot à partir d'une séquence de mots donnée. Le modèle permet donc de prédire l'apparition d'un mot w en considérant n mots précédents.On note w_i le mot à la position i parmi la séquence de n mots. On a $P(w_1^n)$, la probabilité d'apparition d'une séquence de n mot :

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

Des n-grammes particuliers peuvent être considérés, comme les bigrammes ou les trigrammes par exemple. Ces derniers permettent d'estimer la probabilité d'apparition d'un mot à partir du mot ou des deux mots précédents respectivement. Si le modèle n-gramme permet d'estimer la probabilité d'apparition des mots, il faut estimer la probabilité d'apparition des termes d'une requête au sein d'un document. La probabilité $P(q|d)$, c'est-à-dire la probabilité d'apparition de l'ensemble des termes (ou éléments) de la requête q dans le document d, est déterminée par le produit des probabilités individuelles associées aux

termes (Balog, Y. Fang et al. 2012) :

$$P(q|d) = \prod_{t \in q} P(t|d)^{n(t,q)}$$

On a $n(t, q)$ le nombre d'apparitions du terme t dans la requête q, $P(t|d)$ la probabilité que le terme t apparaisse dans le document d estimée à l'aide du modèle de langue statistique.

Une problématique émerge, liée à l'absence potentielle d'un seul terme de la requête dans le document. En effet, dès lors qu'un seul terme de la requête est absent du document, l'ensemble de la probabilité $P(q|d)$ est égale à 0. Cette problématique est particulièrement vérifiée lors de l'application d'un modèle de langue au domaine de la recherche d'experts au sein des communautés en ligne. En effet, la taille des questions posées par les utilisateurs au sein des communautés en ligne sont souvent courtes et ne permettent pas d'établir suffisamment de chevauchement sémantique entre les documents et les requêtes (L. Chen et Nayak 2008 ; Al-Taie, Kadry et Obasa 2018). Pour résoudre ce problème, le lissage $P(t|\theta d)$ permet de s'assurer que toutes les probabilités $P(t|d)$ soient supérieures à 0. De ce fait, il est possible d'améliorer le modèle de langue en lui permettant de retrouver des documents pertinents auparavant écartés à cause de l'absence d'un unique terme de q dans le document (Chengxiang Zhai et Lafferty 2017).

Les modèles de langue sont utilisés pour représenter les documents dans la plupart des systèmes les mieux classés lors des éditions *Enterprise Track* des conférences TREC, comme l'indique la table 1.1. Ils ont parfois été utilisés en combinaison avec d'autres méthodes de recherche d'experts, notamment avec des algorithmes de *clustering* (Cao et al. 2005 ; Zhu, Song, Rüger et al. 2006 ; Zhu, Song et Rüger 2007). Les trois systèmes les mieux classés lors de l'édition 2006 utilisent tous un modèle de langue (Zhu, Song, Rüger et al. 2006 ; Bao et al. 2006 ; You et al. 2006). Le plus performant est utilisé en combinaison avec une méthode de *clustering* tandis que le moins performant est utilisé seul. En considérant le système le plus performant et le système le moins performant des trois, on obtient respectivement 0.6431 et 0.5639 sur la métrique mAP ainsi que 0.9609 et 0.9043 sur la métrique MRR. Un autre système notable exploitant un modèle de langue a également été évalué sur le corpus de l'édition 2005 de la tâche *Enterprise Track* des conférences TREC et a obtenu 0.2850 sur la métrique mAP ainsi que 0.6496 sur la métrique MRR (Petkova et Croft 2008), soit des performances supérieures à celles obtenues par le deuxième meilleur système officiel de l'édition.

Cependant, une autre problématique émerge, celle du fléau de la dimension (Bellman 1966 ; Al-Taie, Kadry et Obasa 2018). Dans le cas du modèle de langue, le fléau de la dimension est lié à l'estimation de la probabilité d'apparition d'une séquence de mots de taille n, $P(w_1^N)$, par un modèle entraîné sur un vocabulaire V. Il existe $|V|^n - 1$ paramètres libres que le modèle de langue doit estimer lors du calcul de $P(w_1^N)$, ce qui constitue une explosion combinatoire. Par exemple, pour $n = 10$ et pour un vocabulaire de 100000 mots

acquis par le système, le nombre de paramètres libres est de $100000^{10} - 1$.

Pour réduire l'impact du fléau de la dimension, il est possible de représenter les mots comme des vecteurs de nombres réels. Cette technique de représentation s'appelle un plongement lexical (ou *word embedding*). Elle permet une représentation compacte de la distribution de probabilités sur les séquences de mots. L'hypothèse est la suivante : des mots qui apparaissent dans un contexte similaire ont probablement une signification apparentée. Le vecteur de mot intègre des caractéristiques liées à la sémantique ou à la syntaxe (nombre, genre) du mot et représente l'ensemble de ces caractéristiques en un seul nombre réel. Ainsi, les mots « chat » et « chien » peuvent par exemple être représentés par des vecteurs peu distants dans l'espace vectoriel. Considérons que la phrase « le chat marche dans la chambre » fait partie du corpus d'entraînement (Bengio et al. 2003). À l'aide d'une représentation des mots sous forme de vecteurs de nombres réels, le modèle de langue devrait pouvoir généraliser cette phrase d'entraînement et estimer la probabilité associée à la phrase « le chien court dans le salon » à l'aide de la similarité des caractéristiques syntaxiques et sémantiques liées aux mots (Bengio et al. 2003). Les limitations des modèles de langue basés sur une représentation des mots sous forme de vecteurs de nombres réels sont liées à celles du modèle vectoriel. Les homonymies et les polysémies d'un mot ne sont pas prises en compte, puisque chaque mot du modèle vectoriel est représenté par un unique vecteur (Al-Taie, Kadry et Obasa 2018).

Un algorithme d'apprentissage profond basé sur un réseau de neurones artificiel permet de découvrir automatiquement les caractéristiques syntaxiques et sémantiques liées aux mots lors de la phase d'apprentissage. Pour cette raison et grâce à leurs performances, les modèles de langue basés sur des réseaux de neurones artificiels ont supplanté les modèles de langage statistiques. Un neurone formel est une représentation inspirée du neurone biologique consistant en une fonction mathématique à plusieurs paramètres ajustables. Un réseau de neurones est constitué d'un ensemble de neurones formels organisés au sein d'une architecture. Cette dernière détermine la manière dont les neurones sont connectés au sein du réseau. Inspirés par les réseaux de neurones et la représentation des mots et des experts potentiels par des vecteurs de nombres réels, des travaux ont abouti à un modèle log-linéaire pour la recherche d'experts (Van Gysel, Rijke et Worring 2016). Dans ce modèle, les profils d'experts sont construits à l'aide d'une méthode non supervisée basée sur une représentation des mots par des vecteurs de nombres réels. Les publications scientifiques sont utilisées pour décrire le profil d'expert des chercheurs. Les requêtes sont également représentées par des vecteurs de nombres réels et les experts potentiels sont ordonnés par valeur décroissante du produit scalaire entre le plongement lexical représentant une requête et celui représentant un profil d'expert. Les auteurs ont notamment évalué leur méthode sur les éditions 2005 à 2007 des tâches *Enterprise Track* des conférences TREC. Les performances sont décrites dans la table 1.1. Ils obtiennent 0.248 sur la métrique mAP et 0.618 sur la métrique MRR pour l'édition 2005, 0.484 sur la métrique

mAP et 0.833 sur la métrique MRR pour l'édition 2006 ainsi que 0.344 sur la métrique mAP et 0.513 sur la métrique MRR pour l'édition 2007. La limitation de ce modèle réside dans le manque d'interprétabilité de la méthode utilisés. En effet, les concepts exploités pour ordonner les profils d'experts sont latents et ne peuvent être employés pour expliquer la correspondance entre un profil d'expert et une requête (Cifariello, Ferragina et Ponza 2019).

Modèles thématiques Dans le cadre de la recherche d'experts, un modèle thématique associe les requêtes aux individus correspondants avec une représentation des individus par un ensemble de variables latentes équivalent à des thématiques ou expertises (Balog, Y. Fang et al. 2012). Chaque variable latente ou thématique correspond à un ensemble de termes extraits du document. Une thématique regroupe donc plusieurs termes proches du point de vue de la sémantique. Chaque individu est représenté par la somme pondérées des thématiques qu'il aborde (Deng, King et Lyu 2008). Les principales méthodes basées sur un modèle thématique sont LSI ou LSA (analyse sémantique latente ou indexation sémantique latente) (Deerwester et al. 1990), pLSI ou pLSA (analyse sémantique latente probabiliste ou indexation sémantique latente probabiliste) (Hofmann 1999 ; Hofmann 2017) et LDA (allocation de Dirichlet latente) (Blei, Ng et Jordan 2003). LSA (aussi appelée LSI) exploite une matrice document-terme pour produire des matrices document-thématique et thématique-terme à l'aide d'un procédé issu de l'algèbre linéaire, la décomposition en valeurs singulières (SVD). Dans la matrice document-terme, chaque ligne représente un document et chaque colonne un mot. Une entrée de la matrice $w_{i,j}$ correspond au score TF-IDF du terme j dans le document i :

$$w_{i,j} = tf_{i,j} * log\frac{N}{df_j}$$

avec $tf_{i,j}$ le nombre d'occurrences du terme j dans le document i, df_j le nombre de documents contenant le terme j. Un terme rare dans le corpus et fréquent dans le document considéré aura un score TF-IDF élevé. À partir de la matrice document-terme, les thématiques peuvent être identifiées en capturant les relations existant entre les termes et les documents. La décomposition en valeurs singulières permet de réduire les dimensions de la matrice en la factorisant. En sélectionnant les t plus grandes valeurs singulières de M et en ne conservant que les t premières colonnes de U et V, avec t un hyperparamètre correspondant au nombre de thématiques que l'on souhaite faire émerger, on obtient :

$$M = U_t * S_t * V_t$$

avec $M \in \mathbb{R}^{(m*n)}$ la matrice document-terme, S la matrice diagonale des valeurs singulières de M, $U \in \mathbb{R}^{(m*t)}$ la matrice document-thématique et $V \in \mathbb{R}^{(n*t)}$ la matrice

terme-thématique. Plutôt qu'une décomposition en valeurs singulières, pLSA (aussi appelée pLSI) emploie un modèle probabiliste pour identifier les thématiques à partir de la matrice document-terme. L'idée générale de pLSA consiste à obtenir un modèle probabiliste et un ensemble de variables latentes permettant de générer les données observées dans la matrice document-terme. Soit d un document, z une thématique, w un terme. Considérons $P(z|d)$, la probabilité que la thématique z soit présente dans le document d et $P(w|z)$ la probabilité que le terme w appartienne à la thématique z. On obtient $P(D, W)$ la probabilité d'obtenir le terme w dans le document d en considérant le document :

$$P(D, W) = P(D) \sum_{Z} P(Z|D)P(W|Z)$$

La probabilité $P(D)$ est obtenue directement à partir du corpus. Les probabilités $P(Z|D)$ et $P(W|Z)$ sont estimées à l'aide d'un algorithme espérance-maximisation (EM). Il s'agit d'un algorithme permettant de trouver les paramètres du maximum de vraisemblance d'un modèle probabiliste dépendant de variables latentes. Il existe une autre manière d'obtenir $P(D, W)$, en considérant la thématique :

$$P(D, W) = \sum_{Z} P(D|Z)P(Z)P(W|Z)$$

Dans ce cas, il est possible d'établir un parallèle avec LSA. On a $P(D|Z)$ correspondant à U_t, $P(Z)$ correspondant à S_t et $P(W|Z)$ correspondant à V_t. LDA correspond à une version bayésienne de pLSA. À partir d'un nombre de thématiques t fixées, une thématique est attribuée à chaque terme du document considéré à l'aide d'une distribution de Dirichlet. Une extension de LDA est la modélisation ATM (*author-topic model*) qui permet de modéliser simultanément l'expertise des individus auteurs de documents ainsi que les thématiques abordées dans les documents à l'aide d'hyperparamètres (Rosen-Zvi et al. 2012). Une autre extension, la modélisation APT (*author-persona-topic*) ajoute les *persona*, c'est-à-dire des identités diverses sous lesquelles un auteur peut écrire (Mimno et McCallum 2007). Pour la recherche d'experts dans le milieu académique, le modèle ACT (*author-conference-topic*) a été introduit et permet de modéliser une variable latente représentant les conférences dans lesquelles les articles de recherche ont été publiés (Tang et al. 2008). Enfin, le modèle CAT (*citation-author-topic*) permet de représenter les auteurs cités par un document (Tu et al. 2010). Chacune des méthodes de modélisation thématique a ses propres limitations. LSA ne permet pas de prendre en compte les synonymes ou la polysémie (Al-Taie, Kadry et Obasa 2018). Des problématiques de surapprentissage sont soulevées par pLSA, puisque le nombre de paramètres augmente avec le nombre de documents (Aggarwal 2014 ; Al-Taie, Kadry et Obasa 2018). De plus, il est difficile de généraliser le modèle lors de l'introduction de nouveaux documents, puisque la probabilité $P(D)$ n'est pas paramétrée. Enfin, le temps d'estimation

des thématiques latentes peut être très long mais des travaux ont permis de suggérer une modélisation thématique à partir d'un ensemble de thématiques prédéfinies à l'avance (Deng, King et Lyu 2008). Cependant, la sélection de l'ensemble des thématiques reste une problématique essentielle (Balog, Y. Fang et al. 2012).

Analyse de *clusters* de documents ou d'utilisateurs Les méthodes de recherche d'experts à base de *clustering* sont des méthodes qui permettent de regrouper les documents ou individus similaires en *clusters* (Xiaoyong Liu et Croft 2004). Par exemple, les potentiels experts ont été regroupés à l'aide d'un algorithme des k-moyennes après application d'un modèle de langue et étiquetage par leurs collaborateurs et thématiques abordées (Cao et al. 2005). L'algorithme des k-moyennes consiste à regrouper les données en k *clusters* en minimisant la distance de chaque élément à un centroïde (au centre d'un *cluster*). L'inconvénient de cette méthode consiste au fait de fixer le nombre de *clusters* à l'avance, sans garantie qu'il s'agit du nombre de *clusters* optimal. Le deuxième meilleur système de l'édition 2005 des tâches *Enterprise Track* des conférences TREC a employé une méthode de *clustering*. Les performances du système sur les métriques mAP (0.2688) et MRR (0.6344) sont décrites dans la table 1.1. D'autres travaux se sont focalisés sur le regroupement des documents associés aux profils d'experts à l'aide d'un algorithme de *clustering* afin de déterminer les principaux domaines de recherche des individus (Macdonald, Hannah et Ounis 2008). Le troisième système le plus performant de l'édition 2005 de la tâche *Enterprise Track* des conférences TREC emploie un *clustering* pour générer des requêtes à partir de *clusters* d'utilisateurs ayant envoyé, reçu ou répondu à des emails (He et Zhifeng Yang 2005). Une requête est construite à partir des thématiques associées à chaque *cluster*. Les auteurs obtiennent respectivement des performances de 0.2174 sur la métrique mAP et 0.6068 sur la métrique MRR. Un algorithme basé sur le *clustering* des trajectoires des chercheurs, *Temporal Semantic Topic-Based Clustering* (TST), a également été utilisé pour identifier des communautés de chercheurs partageant une trajectoire scientifique similaire (Osborne, Scavo et Motta 2014). Appliquer un *clustering* sur les trajectoires des chercheurs consiste à identifier des communautés de chercheurs travaillant sur les mêmes thématiques durant les mêmes périodes. Les limites de cet algorithme sont liées au fait que l'on regroupe dans les *clusters* uniquement des experts ayant des trajectoires similaires. En effet, l'analyse de *clusters* d'utilisateurs ne permet d'obtenir que des regroupements d'experts ayant le même niveau d'expertise (Al-Taie, Kadry et Obasa 2018). Cette approche est plus indiquée pour identifier des communautés de chercheurs sans identifier de hiérarchie entre les membres de la communauté.

1.7.2 Méthodes à base de graphes

Des méthodes de recherche d'experts à base de graphes ont également été explorées à la suite de l'introduction des graphes d'expertise (Serdyukov et Hiemstra 2008). Ces derniers représentent des experts et des documents, reliés par une relation de paternité. En effet, un individu est considéré comme expert du contenu qu'il rédige. Les graphes d'expertise ont ensuite été redéfinis comme des graphes dont les sommets sont des individus (experts ou non) reliés entre eux par un type de relation prédéfini (par exemple, liens de citation, de coauteurs, *etc.*) (Pal et Konstan 2010). Formellement, G est un graphe tel que $G = (V, E)$ où V est l'ensemble des sommets du graphe G, E est l'ensemble des relations existant entre les paires de nœuds du graphe.

Afin d'identifier les experts au sein des graphes d'expertise, diverses méthodes peuvent être appliquées : les mesures quantitatives (à l'aide des algorithmes PageRank (Page et al. 1999) et HITS (Kleinberg 1999)), les méthodes exploitant les propriétés du graphe telles que la centralité (Al-Taie, Kadry et Obasa 2018) ainsi que les algorithmes de propagation (S. Lin et al. 2017). Des statistiques simples telles que le h-index ou le nombre de citations peuvent également être employées (Fischer, Remus et Biemann 2019). Les limites des méthodes à base de graphes sont liées à la complexité exponentielle des calculs et au passage à l'échelle sur de plus grands jeux de données (Pal et Counts 2011 ; Al-Taie, Kadry et Obasa 2018). De manière générale, ces méthodes ne permettent pas de prendre en compte les expertises associées aux individus, mais ne se basent que sur une évaluation de leur autorité. Des problématiques peuvent émerger dans le cas de requêtes complexes ou de domaines d'expertise émergents.

Méthodes exploitant des mesures quantitatives

Initialement, HITS et PageRank ont été créé pour classer les pages web par ordre décroissant de pertinence. PageRank ordonne les pages web en fonction d'un score d'importance. Si une page web i a une importance r et d liens sortants, chaque lien j reçoit la fraction suivante de l'importance de la page i : $\frac{r}{d}$. Une page obtient donc un score d'importance correspondant à la somme des scores d'importance des pages possédant un lien entrant vers elle (aussi appelées *backlinks*). On a :

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

avec r_i le score d'importance de la page i, d_i le degré sortant de la page i. PageRank considère donc le nombre et la qualité des liens vers une page pour déterminer son importance. L'hypothèse sur laquelle se base l'algorithme est la suivante : plus une page web est importante, plus il est probable que d'autres pages renvoient vers cette dernière. Il s'agit d'un algorithme très populaire puisqu'il est à la base du moteur de recherche Google.

Initialement, PageRank était basé sur un modèle de surfeur aléatoire. Ce modèle suppose qu'un utilisateur clique au hasard sur les liens d'une page. On estime la probabilité que l'utilisateur clique sur une page considérée en simulant le comportement d'un surfeur aléatoire à l'aide d'un algorithme de marche aléatoire dans le graphe. PageRank correspond donc à la probabilité qu'une page soit visitée par un tel utilisateur sous ce modèle. Plus récemment, PageRank a évolué pour se baser sur un modèle de surfeur raisonnable, prenant en compte un ensemble de caractéristiques susceptibles de mener un utilisateur à cliquer sur une page (par exemple, la position du lien dans une page web). Dans le cadre de la recherche d'experts, PageRank est basé sur sur l'analyse des citations et des liens de coauteurs d'un individu (Page et al. 1999). Cet algorithme a été utilisé pour ordonner les auteurs de publications scientifiques à l'aide de leur nombre de citations et des liens de coauteurs qu'ils entretiennent (Fischer, Remus et Biemann 2019). Un auteur cité par d'autres auteurs considérés comme importants sera considéré comme un expert.

Le problème de la recherche d'experts au sein de communautés en ligne ou de réseaux sociaux a également été ramené à l'identification d'individus faisant figures d'autorité au sein de réseaux sociaux sur des thématiques particulières et à l'aide de méthodes à base de graphes. L'algorithme *Hyperlink-Induced Topic Search* (HITS) (Kleinberg 1999) a été utilisé pour identifier des experts lors de l'analyses d'e-mails, en considérant qu'un expert est une autorité (Campbell et al. 2003). Dans ces travaux, une autorité correspond à un individu répondant régulièrement à des questions de ses pairs par e-mail. Plus précisément, une autorité correspond à un sommet du graphe possédant un degré entrant élevé depuis des noeuds appelés *hubs*.

Les performances obtenues sur l'identification d'experts à partir d'emails à l'aide d'une approche basée sur le contenu et d'un algorithme HITS ont été comparées (Campbell et al. 2003). L'approche basée sur le contenu consistait en un simple comptage du nombre d'emails envoyé à propos d'une thématique particulière par chaque individu. Un ensemble de mots-clefs était associé à chaque thématique possible. Les individus ont été ordonnés en fonction du nombre d'emails envoyés sur une thématique considérée. Les emails constituaient deux corpus distincts : *OrgA*, un corpus d'emails issus d'une institution académique et *OrgB*, un corpus d'emails issus d'une entreprise. Sur le corpus *OrgA*, les auteurs obtiennent une précision de 0.52 pour HITS, 0.38 pour l'approche basée sur le contenu. Sur le corpus *OrgB*, les auteurs obtiennent une précision de 0.67 pour HITS, 0.5 pour l'approche basée sur le contenu. L'approche basée sur l'algorithme HITS a donc surpassé l'approche basée sur l'analyse du contenu des emails dans cette étude.

HITS et PageRank ont été comparés dans le cadre de travaux sur l'identification d'experts au sein d'e-mails (Dom et al. 2003 ; H. Chen et al. 2006). Les deux algorithme ont notamment été comparés sur la métrique mAP à partir d'un corpus d'emails issu de l'édition 2005 des tâches *Enterprise Track* des conférences TREC (H. Chen et al. 2006). PageRank a obtenu une mAP de 0.2005 tandis que HITS a obtenu une mAP de

0.4219. Un système de recherche d'experts basé sur PageRank a été classé second lors de l'évaluation des systèmes de l'édition 2007 des tâches *Enterprise Track* des conférences TREC. Ce système obtient 0.4427 sur la métrique mAP et 0.6131 sur la métrique MRR (Duan et al. 2007). La différence entre PageRank et HITS réside dans la considération des liens sortants. En effet, HITS permet de considérer les liens entrants (pour identifier des autorités) et sortants (pour identifier des *hubs*) tandis que PageRank ne considère que les liens entrants (Chien, Hoong et Ho 2014).

Méthodes exploitant des propriétés du graphe

Traditionnellement, l'autorité d'un individu au sein d'un réseau est mesurée par l'exploitation de propriétés du graphe qui représente ce réseau. La propriété principalement exploitée est la centralité (Balog, Y. Fang et al. 2012). Plusieurs mesures permettent d'évaluer la centralité dans un graphe : la centralité de degré, la centralité de proximité, la centralité d'intermédiarité et la centralité de vecteur propre.

La centralité de degré repose sur le calcul du degré d'un individu dans le graphe. Le degré d'un individu dans le graphe correspond au nombre de ses voisins directs. Plus un noeud est connecté aux autres dans le graphe, plus il peut être considéré comme une autorité car possédant de nombreux contacts directs.

La centralité de proximité (*closeness*) permet d'évaluer la proximité d'un individu avec l'ensemble des autres individus du réseau. La distance entre des nœuds dans un graphe est évaluée à l'aide de la longueur du plus court chemin qui les sépare.

Quant à la centralité d'intermédiarité (*betweenness*), elle permet d'identifier les individus que l'on peut considérer comme des intermédiaires pour accéder à d'autres individus du réseau. Les individus intermédiaires correspondent à des individus occupant une position centrale dans le réseau. Cette mesure diffère des autres car elle n'est basée ni sur l'analyse du nombre de liens directs d'un nœud ni sur sa distance aux autres nœuds. Elle réside dans le comptage du nombre de fois qu'un nœud du graphe agit comme un point de passage le long du plus court chemin entre deux autres nœuds.

Enfin, la centralité de vecteur propre (*eigenvector*) s'inspire de PageRank et de HITS et propose une version du calcul de degré prenant en compte l'importance des liens directs qu'un individu entretient avec les autres membres du réseau (Cunningham, Everton et Murphy 2016). Les relations entretenues avec des individus centraux ont un poids plus fort tandis que les relations entretenues avec des individus en périphérie de réseau ou entretenant eux-mêmes peu de liens directs ont un poids plus faible.

Algorithmes de propagation

Quant aux algorithmes de propagation, ils permettent de propager le score d'expertise d'un individu aux individus du réseau avec lesquels ils partagent une relation sociale

(Jing Zhang, Tang et J. Li 2007). Ce « score d'expertise » peut être calculé pour chaque individu à partir de son profil d'expert. Au sein des graphes d'expertise, les algorithmes de propagation permettent de propager le score d'expertise d'un individu à ses coauteurs. Ils se basent sur une marche aléatoire dans lesquels les experts sont ordonnés par le nombre de visites d'un marcheur aléatoire. Différents algorithmes de propagation ont été proposés, selon la technique de marche aléatoire employée.

Les algorithmes de propagation *k-step*, *infinite step* et la marche aléatoire avec absorption sont fondés sur une propagation multiple des scores d'expertise, permettant de consulter plusieurs documents ou plusieurs profils d'experts (S. Lin et al. 2017). Ces méthodes ont permis d'obtenir de meilleurs résultats qu'avec une propagation en une étape (Serdyukov et Hiemstra 2008). L'approche *k-step* permet de considérer qu'un utilisateur réalisera k étapes finies avant d'interrompre le processus de recherche d'expert et de se satisfaire d'un expert potentiel ou d'un document. La probabilité calculée est celle de sélectionner un expert potentiel donné au bout de k étapes par une consultation des profils d'experts ou des documents initialement ordonnés. L'approche *infinite step* considère que l'utilisateur réalisera un nombre infini d'étapes. Un processus de décision markovien permet de déterminer les experts potentiels ou documents les plus consultés, donc les plus pertinents pour un utilisateur (Serdyukov et Hiemstra 2008). Enfin, l'approche basée sur une marche aléatoire avec absorption permet de calculer la probabilité de sélectionner un expert potentiel considéré comme le seul expert valable, au bout d'un nombre suffisant d'étapes de marche aléatoire dans le graphe.

Statistiques simples

Des statistiques simples telles que le h-index ou le nombre de citations peuvent également être utilisées pour identifier des experts au sein de la communauté scientifique (Fischer, Remus et Biemann 2019). Le h-index (Hirsch 2005) est un indicateur mesurant l'impact des travaux d'un chercheur. Un chercheur a un h-index h si « h de ses travaux ont au moins h citations ». Le calcul du h-index permet de prendre en considération l'impact des publications. Cependant, le nombre de publications joue un rôle fondamental. En effet, le nombre maximum de publications d'un auteur correspond à son h-index maximal. Ainsi, les chercheurs privilégiant un petit nombre de publications soumises à des conférences sélectives peuvent obtenir un h-index bien moins élevé qu'un chercheur plus prolifique mais dont les publications ont moins d'impact (Costas et Bordons 2007).

L'analyse des graphes de citation (An, Janssen et Milios 2004) est également très répandue. Les sommets d'un graphe de citation sont généralement constitués par des publications scientifiques. Il existe une relation dirigée d'une publication p_1 vers une publication p_2 si p_1 cite p_2. Le graphe est dit orienté puisque les relations entre les paires de sommets sont dirigées. L'analyse des graphes de citation permet de prendre en compte

les relations de citation existant entre publications scientifiques (et par projection, entre leurs auteurs), pour déterminer les publications scientifiques les plus fortement citées donc phares. En effet, si une publication est fortement citée par les autres membres de la communauté scientifique, elle revêt un intérêt particulier. Elle constitue probablement les fondements d'un domaine de recherche ou propose une méthode permettant d'obtenir des résultats probants. Par projection, les auteurs de publications phares peuvent être considérés comme des experts de leur domaine, puisqu'ils sont capables de produire des publications phares. Cependant, les statistiques simples telles que le h-index ou le nombre de citations ne permettent pas d'identifier des experts à elles seules et nécessitent d'être complétées par d'autres indicateurs dont nous avons discuté dans la section 1.5 (Martin 1996 ; Hirsch 2005 ; Van Raan 2006).

1.7.3 Méthodes hybrides

Des méthodes hybrides de recherche d'experts récemment été proposées, combinant des méthodes à base d'apprentissage et des méthodes à base de graphes. Les méthodes à base de graphes permettent généralement d'obtenir de meilleures performances que les méthodes basées uniquement sur le contenu (Al-Taie, Kadry et Obasa 2018) comme nous l'avons rappelé dans la section 1.7.2. Cependant, les limites des méthodes à base de graphes résident dans leur absence de prise en compte du contenu des documents. En effet, les méthodes à base de graphes ne permettent pas de considérer les expertises enfouies dans les documents mais ne se basent que sur la réputation des individus, ce qui est problématique dans le cas de thématiques émergentes par exemple.

L'état de l'art suggère que les problèmes liés à l'identification et la classification des experts peuvent être évités en combinant l'identification d'expertises avec l'identification des relations entre experts (Angelova, Boeva et Tsiporkova 2017). Des travaux récents indiquent que la combinaison des méthodes de fouille de texte et de fouille de graphe aboutit à de meilleures performances que sur l'utilisation de l'une des deux méthodes seule, bien que leur combinaison reste anecdotique (Ganguly et Pudi 2017).

Parmi les méthodes hybrides, une extension de PageRank pour l'analyse de l'influence des utilisateurs sur des thématiques particulières au sein du réseau social Twitter, TwitterRank, a été proposée (Weng et al. 2010). Celle-ci permet de prendre en compte les interactions sociales entre utilisateurs via l'analyse de la structure du réseau social mais également les similarités entre utilisateurs concernant les thématiques abordées. L'étude a démontré que TwitterRank surpasse PageRank (Weng et al. 2010).

Wiser (Cifariello, Ferragina et Ponza 2019) est un système de recherche d'experts combinant des approches centrées sur les documents exploitant une méthode d'apprentissage non supervisé et une extraction de connaissances issues du *Wikipedia Knowledge Graph*. Plus précisément, le contenu des documents est annoté sémantiquement.

L'annotation sémantique consiste en l'identification d'entités issues d'une base de connaissances (ici *Wikipedia Knowledge Graph*) et mentionnées dans un texte ou document. Chaque chercheur indexé dans le système Wiser est donc associé à un graphe extrait de Wikipedia représentant les thématiques abordées au sein de ses travaux. Ces thématiques, entités ou encore expertises sont reliées entre elles dans le graphe si elles sont similaires, c'est-à-dire si les vecteurs qui les représentent dans un modèle de langue sont similaires. Contrairement au modèle log-linéaire (Van Gysel, Rijke et Worring 2016), Wiser est un système interprétable, puisque les entités qui permettent de réaliser des correspondances entre requêtes et documents sont explicitées lors de l'annotation sémantique. Wiser a été évalué sur le jeu de données TU décrit dans la section 1.3.1 et comparé à Model 2 (Balog, Y. Fang et al. 2012) et au modèle log-linéaire (Van Gysel, Rijke et Worring 2016) sur les métriques mAP et MRR. Model 2 a obtenu respectivement 0.253 et 0.302 sur les métriques mAP et MRR, tandis que l'algorithme log-linéaire a obtenu respectivement 0.287 et 0.363 sur ces mêmes métriques. Wiser a surpassé ces deux systèmes, obtenant 0.385 sur la métrique mAP et 0.459 sur la métrique MRR à l'aide de la meilleure configuration du système.

1.8 Systèmes état de l'art

Des plateformes de données scientifiques ont déjà été développées. Au sein de ces plateformes, la représentation de connaissances extraites à partir de textes par un graphe est répandue. ArnetMiner (Tang et al. 2008) permet d'extraire des profils d'experts de chercheurs à partir du web et a pour objectif de représenter ces profils sous forme d'un réseau social académique. Il fouille et indexe automatiquement les publications scientifiques disponibles en ligne. Aussi appelé AMiner, il est basé sur une modélisation thématique, plus précisément une allocation de Dirichlet latente. CareerMap (K. Wu et al. 2018) est un composant de ArnetMiner permettant de visualiser la trajectoire d'un chercheur. Cette trajectoire est extraite de la base de données de publications scientifiques de ArnetMiner. Elle matérialise les changements d'affiliation des chercheurs au cours de leur carrière, c'est-à-dire de structures (universités, laboratoires, lieux géographiques) auxquelles les chercheurs sont rattachés. Cependant, l'implémentation du système n'est pas *open source*. Il s'agit d'un service web à la manière de Google Scholar, Microsoft Academic Search, PubMed, arXiv, ACM Digital Library ou encore DBLP. De plus, seul le contenu des publications scientifiques est analysé, les indicateurs liés aux relations de collaboration scientifique n'étant pas considérés.

Saffron est un système *open source* configurable basé sur des méthodes de traitement du langage naturel telles que l'extraction de termes, l'annotation sémantique et l'extraction de relations sémantiques (Monaghan et al. 2010). L'une des fonctionnalités phare du système est la construction de taxonomies à partir des relations sémantiques extraites.

Construire une taxonomie permet de représenter la hiérarchie existant entre les thématiques et d'identifier des relations d'hyperonymie, d'hyponymie ou de synonymie entre certains termes, par exemple. Dans Saffron, un modèle de langue probabiliste est utilisé pour extraire les termes à partir de texte. Le web est également fouillé pour enrichir les profils d'experts. Une méthode d'apprentissage supervisé basée sur un modèle discriminatif, plus précisément sur des machines à vecteurs de support, est employée pour construire une taxonomie en organisant les termes dans une hiérarchie. En raison de ses fonctionnalités riches et configurables et de son caractère open source, Saffron est l'une des plateformes phare de la recherche d'experts. Si les liens de coauteurs entretenus par les chercheurs peuvent être pris en compte dans Saffron pour caractériser le profil d'expert des individus, la plateforme semble davantage tournée vers l'analyse du contenu des publications et ne semble pas privilégier l'estimation de la réputation des individus afin d'identifier des experts au sein des publications scientifiques.

CSSeer (H.-H. Chen et al. 2013) est un système de recherche d'expert basé sur Wikipedia et CiteSeer^X (Caragea et al. 2014), une librairie digitale qui stocke et indexe les publications scientifiques du domaine de l'informatique. CSSeer extrait des phrases-clefs à partir du titre et du résumé d'une publication et utilise cette information pour inférer l'expertise des auteurs qui lui sont associés (H.-H. Chen et al. 2013). CSSeer a été comparé à ArnetMiner et à Microsoft Academic Search sur vingt requêtes de recherche d'expert simples (H.-H. Chen et al. 2013). Les résultats obtenus étant disparates, les auteurs de l'étude comparative ont suggéré d'utiliser des systèmes de recherche d'experts variés pour obtenir une liste d'experts la plus exhaustive possible.

L'une des plateformes phares, Rexplore (Osborne, Motta et Mulholland 2013), propose un réseau sémantique de thématiques de publication de granularité fine, liées entre elles par des relations sémantiques. Cette plateforme permet de suivre les tendances de thématiques de recherche d'un domaine et est plutôt orientée sur leur analyse que sur des problématiques de recherche d'experts, bien qu'elle permette également de répondre à des requêtes avancées sur ce domaine. Elle propose également des visualisations avancées pour l'exploration des données. Au sein de Rexplore, l'algorithme utilisé est un algorithme d'apprentissage non supervisé basé sur le *clustering* des trajectoires de chercheurs. Nous renvoyons le lecteur à notre présentation de l'algorithme dans la section adéquate pour plus de détails. Rexplore est à l'origine de Smart Topic Miner (Osborne, Salatino et al. 2016), une application sémantique d'aide à l'annotation à destination du comité de rédaction de la revue scientifique Springer Nature Computer Science. Les limites de Rexplore sont liées au fait que l'on ne regroupe dans les *clusters* que des experts ayant des trajectoires similaires. Cette approche semble plus appropriée pour une analyse des tendances et pour identifier des communautés de chercheurs sans identifier de hiérarchie entre les membres de la communauté. La prise en compte des relations entretenues par les experts au sein de la communauté permettrait de définir le niveau d'expertises des

membres d'une communauté scientifique.

CL Scholar (M. Singh et al. 2018) exploite des connaissances enfouies dans le texte des publications scientifiques du corpus ACL Anthology (Radev et al. 2013) pour construire un graphe de connaissances représentant le domaine de la linguistique informatique. Ce graphe de connaissances représente les publications scientifiques, les auteurs, les conférences et les thématiques de publication. Les publications sont notamment décrites par les années de publication, le titre, le résumé. De même, les auteurs sont décrits par un ensemble de propriétés telles que leur nom ou leur affiliation, par exemple. Les informations nécessaires à la construction du graphe de connaissances sont extraites à partir des publications scientifiques au format PDF à l'aide d'un algorithme de reconnaissance optique de caractères basé sur un réseau de neurones récurrent (Mathew, A. K. Singh et Jawahar 2016). Les auteurs utilisent également une méthode à base de graphe pour trier par ordre de pertinence les documents indexés par le système. Des statistiques simples comme le nombre de citations sont employées pour ordonner les documents. Le système est équipé de fonctionnalités de requêtage. Les requêtes peuvent être soumises en langage naturel et permettent de répondre à des questions pertinentes telles que « Quels sont les publications scientifiques sur le domaine de publication X acceptées dans la conférence Y ? » ou encore « Est-ce que l'auteur A a publié sur le domaine de publication X ? » (M. Singh et al. 2018).

Enfin, le système de recherche d'experts Wiser (Cifariello, Ferragina et Ponza 2019) est basé sur une approche centrée sur les documents exploitant une méthode d'apprentissage non supervisé combinée à de l'annotation sémantique réalisée à partir du Wikipedia Knowledge Graph. Nous avons présenté la méthode de recherche d'experts employée par Wiser lors de la présentation des méthodes de recherche d'experts hybrides. Son approche combinant méthodes à base d'apprentissage et méthodes à base de graphe permet d'obtenir des résultats très prometteurs. Cependant, la méthode de recherche d'experts sur laquelle Wiser repose ne permet pas de prendre en compte une validation par les pairs naturellement présente dans le milieu académique. Elle est plutôt basée sur la proximité sémantique entre profils d'experts et requêtes.

1.9 Conclusion

La recherche d'experts est une problématique essentielle. De ce fait, elle a suscité l'intérêt des chercheurs qui ont mis au point de nombreuses méthodes d'automatisation de la recherche d'experts et produit de nombreuses plateformes et systèmes de recherche d'experts. Les systèmes état de l'art actuels sont des systèmes hybrides, combinant les deux principales méthodes historiques de recherche d'experts que sont les méthodes d'apprentissage et les méthodes à base de graphe.

Si les connaissances initialement extraites des profils d'experts par des systèmes comme

Saffron, CSSeer ou Wiser sont enrichies par la fouille du web, les indicateurs d'expertise basés sur la validation par les pairs ne sont pas ou peu considérés dans ces systèmes. Les indicateurs d'expertise sont plutôt basés sur la proximité sémantique entre les profils d'experts et les requêtes. Des indicateurs simples de validation par les pairs tels que le nombre de citations ou le h-index sont parfois inclus aux méthodes traditionnelles d'extraction de connaissances à partir de texte. Cependant, comme cela a été suggéré par l'état de l'art, ces statistiques simples ne sont pas suffisantes pour caractériser à elles seules le niveau d'expertise de ses auteurs. En effet, l'état de l'art suggère d'exploiter des indicateurs d'expertise variés, tels que la réputation de l'expert potentiel ou le contenu des documents dans lequel celui-ci est mentionné. Il manque donc un système basé sur une méthode hybride combinant des indicateurs d'expertise basés sur le contenu des documents ainsi que sur la réputation de leurs auteurs. L'identification des limitations du h-index a mis en lumière la difficulté d'estimer la réputation d'un expert potentiel. Ainsi, la réputation d'un individu devrait être estimée à l'aide de multiples indicateurs. De plus, la réputation d'un individu doit être matérialisée par une validation par ses pairs.

Dans ce cadre, la fouille de graphe est une méthode intéressante pour la recherche d'experts, puisqu'elle permet de prendre en compte la réputation d'un individu en se basant sur une analyse des interactions sociales des individus au sein d'un réseau. À la lumière de l'état de l'art, la combinaison de méthodes de fouille de texte employées pour construire une représentation de connaissances sous forme de graphe et de méthodes de fouille de graphe semble particulièrement indiquée. En effet, le graphe est une structure adaptée à la représentation d'experts potentiels sous forme de sommets reliés en fonction des interactions sociales existant entre eux, ce qui permet de prendre en compte une validation par les pairs.

Dans le chapitre 2, nous nous intéressons à la représentation de connaissances sous forme de graphe dans le cadre de la recherche d'experts. Nous émettons des hypothèses d'expertise, c'est-à-dire des suppositions concernant la caractérisation de l'expertise d'un individu. Ces hypothèses sont émises en fonction de l'éclairage porté sur les connaissances disponibles. Nous nous intéressons à deux structures de représentation des connaissances en particulier, les graphes de connaissances et les graphes attribués. Nous proposons d'aborder l'utilisation de ces représentations dans le cadre de la recherche d'experts ainsi que d'identifier les indicateurs d'expertise pertinents dans ce cadre, en considérant une validation de l'expertise par les pairs. Le graphe de connaissances est utilisé dans un objectif de partage et de réutilisation des connaissances. Les graphes attribués sont utilisés comme représentation intermédiaire au graphe de connaissances. L'application d'abstractions de graphe sur ces graphes attribués permet de valider les hypothèses d'expertise que nous suggérons dans le chapitre 2. L'abstraction de graphe présentée lors du chapitre 3 permet d'identifier les experts et leurs expertises associées à partir de la fouille des graphes attribués. Les résultats obtenus à la suite de l'abstraction de graphe sont utilisés pour

enrichir le graphe de connaissances et présenter nos résultats de manière interprétable et interopérable.

Chapitre 2

Représentation de connaissances dans le cadre de la recherche d'experts

Les données sont des éléments bruts, obtenus à partir de mesures ou d'observations. Si l'information est la mise en forme d'une donnée et est située à un niveau d'abstraction supérieur, c'est la connaissance qui permet d'interpréter l'information et d'en donner un sens. Une connaissance se partage, se stocke et s'interprète en accord avec un référentiel commun. Dans le cadre de la recherche d'experts, l'expertise a été décrite comme une connaissance tacite (Reuber, Dyke et Fisher 1990). Elle n'est pas explicitement élicitée (Bradley, Paul et Seeman 2006) et nécessite donc d'être extraite, ce qui constitue un processus coûteux et chronophage. Il est donc nécessaire lorsque cela est possible de représenter les connaissances extraites afin d'en faciliter la réutilisation dans d'autres applications et la compréhension par une machine. Pour rendre la recherche d'experts interprétable et intéropérable, c'est-à-dire pour en faciliter la compréhension, le partage et la réutilisation, il est nécessaire de représenter les connaissances facilitant la recherche d'experts dans une structure de données respectant un formalisme standardisé. L'information ne doit plus seulement être compréhensible pour un être humain, mais également être aisément interprétable par une machine.

Le graphe est une structure largement répandue dans l'état de l'art pour représenter les connaissances. Comme le suggère l'état de l'art de la recherche d'experts présenté lors du chapitre 1, il s'agit d'une structure de données essentielle dans le cadre de la recherche d'experts.

Avec l'avènement du Web sémantique dont les objectifs principaux sont le partage et l'enrichissement des données disponibles sur le Web, le graphe de connaissances est devenu une représentation de connaissances incontournable lorsque l'on cherche à publier et à partager du contenu réutilisable et sur lequel les machines peuvent raisonner. Dans cet objectif, nous nous intéressons aux graphes de connaissances lors de ce chapitre.

Le graphe permet également de représenter les experts potentiels et les interactions sociales existant entre eux, ce qui permet de prendre en compte une validation de l'expertise par les pairs. Les experts et expertises associées peuvent être identifiés à l'aide d'une méthode de fouille de graphe, présentée lors du chapitre 3. Cette méthode explore

les graphes attribués, formalisme utilisé comme représentation intermédiaire au graphe de connaissances dans ces travaux de thèse. Les graphes attribués sont des graphes dont les sommets sont étiquetés à l'aide d'un langage de description.

Un graphe de connaissances est une représentation des connaissances sous forme d'entités reliées entre elles. Les entités forment un réseau sémantique structuré représentant une base de connaissances. Dans ce réseau sémantique, les entités et les relations peuvent être formellement interprétées dans le monde réel. Les graphes de connaissances sont largement répandus et utilisés dans de nombreux domaines tels que le traitement du langage naturel ou le Web sémantique par exemple. En effet, ils permettent d'organiser et fouiller les connaissances (X. Chen, Jia et Xiang 2020). Le plus connu d'entre eux est le Google Knowledge Graph (Singhal 2012). Les graphes de connaissances respectent un formalisme strict au niveau syntaxique et sémantique. Ils sont généralement représentés au format RDF, sous forme de triplets <sujet, prédicat, objet>. Le format RDF représente un standard du Web sémantique. Ce formalisme permet de représenter et diffuser des connaissances dans un objectif d'interprétabilité et d'interopérabilité (Silvello et al. 2017).

Dans ce chapitre, nous nous intéressons à la représentation de connaissances sous forme de graphes, plus particulièrement sous forme de graphes de connaissances, avec un intérêt particulier porté sur le cadre de la recherche d'experts. Nous abordons succinctement l'état de l'art de la représentation des connaissances sous forme de graphes, puis plus précisément sous forme de graphes de connaissances. Nous abordons brièvement les méthodes de génération d'un graphe de connaissances ainsi que les mécanismes de raisonnement à partir de graphes de connaissances. Nous abordons également le cas particulier des graphes de connaissances scientifiques, qui nous intéressent notamment dans le cadre de la recherche d'experts dans le milieu académique. Nous nous intéressons aux connaissances pertinentes dans ce cadre, afin d'identifier les indicateurs d'expertise permettant de générer des graphes attribués facilitant la recherche d'experts, en considérant une validation de l'expertise par les pairs. Ces graphes attribués sont explorés à l'aide d'une méthode de fouille de graphe présentée lors du chapitre 3. Enfin, les graphes attribués construits servent de représentation intermédiaire à un graphe de connaissances scientifique, généré dans un objectif de partage et de réutilisation des connaissances. L'approche permettant de générer le graphe de connaissances scientifique à partir des graphes attribués est présentée dans le chapitre 4.

2.1 Représentation de connaissances sous forme de graphes

Nous nous intéressons à la représentation de connaissances sous forme de graphes.

Afin de rendre les connaissances compréhensibles pour les machines et leur permettre de raisonner, les chercheurs ont suggéré de nombreuses structures de représentation des connaissances à travers l'état de l'art. Le graphe est une structure de données privilégiée. Il représente un ensemble de sommets reliés entre eux. Différents formalismes de représentation de connaissances sous forme de graphes ont été proposés.

Nous présentons des exemples de graphes de connaissances, le processus de construction de ces graphes et un survol des raisonnements possibles sur ces graphes.

2.1.1 Graphes de connaissances

Le Web sémantique a popularisé la représentation de connaissances sous forme de graphes de connaissances. Le Web sémantique est le modèle du Web des données ouvertes et liées proposé par le W3C (Word Wide Web Consortium) (Berners-Lee, Hendler et Lassila 2001). Il s'agit d'une extension du Web standardisée, dans laquelle l'information est structurée de manière à être aisément compréhensible par les humains et les machines, réutilisable et facile à partager. Le Web sémantique s'appuie sur la famille de langages OWL pour décrire les ontologies. Une ontologie est une spécification explicite d'une conceptualisation d'un domaine (Gruber 1993). Elle représente le consensus d'un groupe d'individus sur la spécification explicite de la conceptualisation d'un domaine particulier. OWL se base sur le formalisme RDF et y ajoute des fonctionnalités de raisonnement inspirées des logiques de description.

Les graphes de connaissances sont une représentation des connaissances sous forme d'entités du monde réel reliées entre elles. Les graphes de connaissances respectent le formalisme RDF. Dans un triplet RDF, le sujet représente une ressource à décrire, le prédicat une propriété appliquée sur le sujet, l'objet la valeur associée au prédicat appliqué sur le sujet. Par exemple, l'information « John Smith est l'auteur du document dont l'URL est `http://www.bar.com/some.doc` » peut être représentée par le triplet <`http://www.bar.com/some.doc`, *bib :author*, John Smith> [1]. Le prédicat est toujours représenté sous la forme d'un URI (*Unique Resource Identifier*).

Parmi les exemples de graphes de connaissances les plus connus se trouvent le Google Knowledge Graph (Singhal 2012), Wikidata [2], DBPedia (Lehmann et al. 2015), WordNet (Miller 1995), Facebook Social Graph [3], LinkedIn Knowledge Graph [4], Satori (Microsoft Knowledge Graph) [5], Freebase (Bollacker et al. 2008) ou encore YAGO (Suchanek, Kasneci et Weikum 2008).

1. Exemple de triplet RDF fourni par W3 : `https://www.w3.org/TR/WD-rdf-syntax-971002/#model`
2. Wikidata : `https://www.wikidata.org/`
3. Facebook Social Graph : `https://developers.facebook.com/docs/graph-api/`
4. LinkedIn Knowledge Graph : `https://engineering.linkedin.com/blog/2016/10/building-the-linkedin-knowledge-graph`
5. Microsoft Knowledge Graph : Satori : `https://searchengineland.com/library/bing/bing-satori`

Google a introduit son Knowledge Graph en 2012, permettant d'améliorer la pertinence de ses résultats de recherche en interprétant des entités du monde réel et les liens existant entre elles (Singhal 2012). À la suite de cette contribution majeure ainsi que de l'émergence des données liées, de nombreux graphes de connaissances ont été proposés dans l'état de l'art. Microsoft a également proposé son propre graphe de connaissances avec Satori, afin d'améliorer les performances de son moteur de recherche, Bing. Wikidata est un réseau sémantique structuré et multilingue basé sur Wikipedia. Comme Wikipedia, Wikidata est alimenté par les utilisateurs. DBPedia (Lehmann et al. 2015) est également un réseau sémantique basé sur Wikipedia, à la différence qu'il extrait automatiquement les informations structurées présentes dans Wikipedia et les représente au format RDF à travers une ontologie. WordNet est une base de données lexicale pour la langue anglaise créée en 1985 (Miller 1995). Facebook Social Graph et LinkedIn Knowledge Graph représentent un ensemble de profils d'utilisateurs et les liens existant entre eux. D'autres entités sont également représentées dans ces graphes de connaissances. Concernant Facebook, il s'agit de photos, lieux géographiques et centres d'intérêt. Pour LinkedIn, il s'agit d'emplois, de compétences, de lieux géographiques, d'entreprises ou d'institutions par exemple. Freebase (Bollacker et al. 2008) était une base de connaissances collaborative libre agglomérant des connaissances issues de sources diverses telles que Wikipedia, NNDB, Fashion Model Directory et MusicBrainz. Elle a depuis été rachetée par Google et son contenu a été transféré dans Wikidata en 2015. Quant à YAGO (Suchanek, Kasneci et Weikum 2008), il s'agit d'une base de connaissances utilisée dans l'intelligence artificielle Watson conçue par IBM, représentant des connaissances extraites de Wikipedia, WordNet et GeoNames. Elle a la particularité d'attacher des informations temporelles et spatiales à un grand nombre des entités qu'elle représente (X. Chen, Jia et Xiang 2020).

2.1.2 Construction d'un graphe de connaissances

Initialement, les graphes de connaissances étaient construits manuellement par des experts (Lenat 1995), ce qui représentait une tâche considérablement coûteuse et chronophage. De nos jours, ils sont plutôt alimentés par les utilisateurs ou construits automatiquement à partir de la fouille du Web sémantique, à l'aide de chaînes de traitement basées sur du traitement du langage naturel et de l'extraction d'information. Wikidata ou encore Freebase sont des exemples de graphes de connaissances alimentés par les utilisateurs (Bollacker et al. 2008). Quant à DBPedia ou YAGO (Suchanek, Kasneci et Weikum 2008), ils sont construits automatiquement à partir de la fouille du Web sémantique (Paulheim 2017). Les graphes alimentés par des experts sont généralement de taille restreinte mais contiennent des connaissances de haute qualité, tandis que les méthodes de génération automatique de graphes de connaissances ou d'alimentation par les utilisa-

teurs permettent de traiter de larges volumes de données mais produisent inévitablement des données bruitées.

Puisqu'un graphe de connaissances représente des concepts du monde réel, il est matériellement impossible qu'il puisse être exhaustif (Paulheim 2017), c'est-à-dire qu'il puisse couvrir l'ensemble des concepts existant. De ce fait, la couverture d'un graphe de connaissances est nécessairement limitée. D'autre part, alimenter le graphe de connaissances avec de nouveaux concepts augmente le risque d'introduire des concepts erronés. La problématique réside donc dans l'obtention d'un graphe de connaissances pour lequel un compromis entre couverture et justesse est attendu (Paulheim 2017).

2.1.3 Mécanismes de raisonnement à partir de graphes de connaissances

Les mécanismes de raisonnement à partir de graphes de connaissances consistent en l'ajout de connaissances manquantes à partir de connaissances existantes ou de sources externes et en l'identification et la suppression des erreurs précédemment introduites (Paulheim 2017 ; X. Chen, Jia et Xiang 2020). L'inférence est le processus qui permet de découvrir de nouvelles connaissances à partir de connaissances existantes. Différents mécanismes de raisonnement à partir de graphes de connaissances existent. Il existe des mécanismes de raisonnement à partir de règles, à base de représentations distribuées ou de réseaux neuronaux (X. Chen, Jia et Xiang 2020). Pour plus de détails sur les différents mécanismes de raisonnement à partir de graphes de connaissances, nous renvoyons le lecteur aux différents états de l'art du domaine (X. Chen, Jia et Xiang 2020 ; Paulheim 2017). Nous précisons que les méthodes à base de règles sont facilement interprétables tandis que les méthodes à base de réseaux neuronaux sont les plus performantes bien que leur complexité soit grande.

Dans le cas où l'application requiert des résultats aisément interprétables, par exemple dans les systèmes d'aide à la décision médicale, les méthodes d'inférence à base de règles sont préférées pour le peuplement de graphes de connaissances (L. Chen et Nayak 2008). En effet, les applications employant des graphes de connaissances sont nombreuses. Les graphes de connaissances et les mécanismes de raisonnement à base de graphes de connaissances peuvent aider à récolter des données médicales ou diagnostiquer des pathologies par exemple (X. Chen, Jia et Xiang 2020 ; Yuan et al. 2017).

Afin de faciliter l'utilisation des graphes de connaissances dans le cadre d'applications dans des domaines de spécialité, des graphes de connaissances adaptés à des domaines particuliers ont donc été proposés. Dans le domaine biomédical par exemple, les graphes de connaissances sont généralement construits à l'aide d'une combinaison d'ontologies, de données ouvertes et liées et de données expérimentales. Les graphes de connaissances sont également utilisés pour organiser de manière efficiente les connaissances concernant

le milieu académique. Ils sont alors appelés graphes de connaissances scientifiques. Des applications telles que la recherche d'experts dans le milieu académique ou la recommandation de collaborateurs tirent parti de tels graphes.

2.2 Représentation de connaissances adaptée à la recherche d'experts

Nous nous intéressons à la représentation de connaissances adaptée à la recherche d'experts. Nous abordons le cas particulier des graphes de connaissances scientifiques, utilisés pour organiser les connaissances décrivant le milieu académique. Des graphes attribués pertinents pour la recherche d'experts peuvent être employés comme structure de représentation de connaissances intermédiaire au graphe de connaissances scientifiques. Ces graphes sont notamment construits pour l'application d'une méthode de fouille de graphe permettant d'identifier les experts et expertises associées présentée lors du chapitre 3. Ils sont également utilisés pour représenter les différentes relations existant entre les entités du monde académique.

2.2.1 Graphes de connaissances scientifiques

Les graphes de connaissances scientifiques permettent de représenter les thématiques de recherche, les publications scientifiques ainsi que leurs auteurs (Sinha et al. 2015 ; Tang et al. 2008 ; Nuzzolese et al. 2016 ; Jaradeh et al. 2019 ; Fenner et Aryani 2019 ; Shotton 2013 ; Manghi et al. 2010 ; R. Wang et al. 2018). Récemment, des graphes de connaissances scientifiques ont été générés automatiquement afin d'organiser et fouiller les connaissances concernant le milieu académique de façon efficace (Dessi et al. 2020 ; Angioni et al. 2020).

Parmi les graphes de connaissances scientifiques les plus connus peuvent être cités Open Academic Graph [6] (unifiant Microsoft Academic Graph (Sinha et al. 2015) et AMiner (Tang et al. 2008)), scholarlydata.org (Nuzzolese et al. 2016), PID Graph (Fenner et Aryani 2019), Open Research Knowledge Graph (Jaradeh et al. 2019), OpenCitations (Shotton 2013) ainsi que OpenAIRE (Manghi et al. 2010) ou AceKG (Academic Knowledge Graph) (R. Wang et al. 2018).

Open Academic Graph est un large graphe de connaissances scientifique unifiant Microsoft Academic Graph et AMiner. Microsoft Academic Graph (Sinha et al. 2015) contient près de 300 millions de publications scientifiques et y associe des métadonnées telles que les citations, les auteurs, les institutions, les revues, conférences et thématiques de publication. AMiner (Tang et al. 2008) représente près de 200 millions de publications

6. Open Academic Graph : `https://www.openacademic.ai/oag/`

obtenues à l'aide d'un système d'extraction de profils d'experts de chercheurs basé sur la fouille du Web. Open Research Knowledge Graph (Jaradeh et al. 2019), OpenCitations (Shotton 2013) et OpenAIRE (Manghi et al. 2010) sont des graphes de connaissances scientifiques libres. Open Research Knowledge Graph représente environ 10 millions de publications, OpenCitations 55 millions et OpenAIRE 110 millions. PID Graph (Fenner et Aryani 2019) est composé de publications scientifiques, d'auteurs et de jeux de données représentés par des PID, aussi appelés identifiants pérennes. Quant à scholarlydata.org (Nuzzolese et al. 2016), il s'agit d'un projet à l'origine de la conference-ontology, une ontologie représentant le domaine des conférences scientifiques. AceKG (R. Wang et al. 2018) est un réseau sémantique académique basé sur une ontologie, librement disponible en ligne et décrivant 3,13 milliards de triplets représentant des faits académiques.

La tâche *SemEval 2018 Task Semantic Relation Extraction and Classification in Scientific Papers Challenge* (Gábor, Buscaldi et al. 2018) a fourni un cadre d'évaluation pour l'identification des relations sémantiques au sein des résumés de publications scientifiques. Cette tâche avait pour objectif l'identification automatique des relations sémantiques existant entre des entités spécifiques au sein d'un corpus de publications scientifiques. Les entités correspondaient à des thématiques de publication. Les applications soulevées par de cette tâche sont l'identification de publications scientifiques traitant de problématiques similaires ainsi que le suivi de la progression de l'état de l'art sur une problématique particulière. Les travaux concernant la construction automatique de graphes de connaissances scientifiques ont donc connu un regain d'intérêt suite à cette tâche.

Inspirée par les réflexions résultantes de cette tâche, une combinaison d'apprentissage profond et de méthodes issues du traitement du langage naturel (Buscaldi et al. 2019) a notamment permis de générer un graphe de connaissances scientifique décrivant plus de 10000 entités et 25000 relations sémantiques extraites d'environ 12000 publications scientifiques dans le domaine du Web sémantique. Ces travaux ont permis par la suite à donner naissance à AI-KG, un graphe de connaissances scientifique généré automatiquement (Dessi et al. 2020) et décrivant plus de 14 millions de triplets RDF concernant 300000 publications scientifiques. Un autre graphes de connaissances scientifique généré automatiquement, AIDA (Angioni et al. 2020), permet de prendre en compte la dualité entre milieu académique et milieu de l'entreprise dans la représentation des thématiques de recherche.

De manière générale, les initiatives de publication et d'ouverture des données et de génération de graphes de connaissances scientifiques sont largement encouragées dans l'état de l'art (Dessi et al. 2020), qu'il s'agisse de favoriser le développement des ontologies décrivant le milieu académique[7] ou certains domaines de recherche (Salatino et al. 2018b) ou encore d'encourager la publication sur le Web sémantique (Shotton 2009) par le biais du paradigme des données ouvertes liées.

7. The Bibliographic Ontology (BIBO) : `http://bibliontology.com`

2.2.2 Graphes attribués pertinents pour la recherche d'experts

Une problématique essentielle consiste à identifier les connaissances requises pour la génération automatique d'un graphe de connaissances scientifique adapté à la recherche d'experts. Les connaissances acquises peuvent être dérivées du contenu des publications scientifiques et représentées sous la forme de graphes attribués. Les graphes attribués pertinents pour la recherche d'experts constituent alors une représentation des connaissances intermédiaire, préalable à la construction d'un graphe de connaissances scientifique.

Afin de construire des graphes pertinents pour la recherche d'experts, il est nécessaire d'identifier ce qui caractérise l'expertise d'un individu. Les suppositions concernant la caractérisation de l'expertise d'un individu s'appellent des hypothèses d'expertise (Stankovic et al. 2010) et sont modélisées à l'aide de règles de la forme suivante : « *Si (condition) alors l'individu A est probablement un expert du domaine X* ». Les hypothèse d'expertise concernant un individu peuvent être obtenues à partir du contenu textuel publié en ligne (par exemple, à partir de publications scientifiques) ainsi qu'à partir de sa réputation. Dans le chapitre 1, nous avons suggéré de combiner ces deux indicateurs et de les extraire à partir des publications scientifiques. La réputation d'un individu peut être évaluée par la densité des collaborations scientifiques qu'il entretient avec les membres de la communauté scientifique. Les graphes d'expertise permettent de représenter les liens de collaboration scientifique entretenus par les membres de la communauté scientifique. Puisque l'ensemble des graphes présentés dans ce chapitre représentent des individus du monde réel, les documents qu'ils rédigent, les expertises extraites de ces documents et les liens existant entre eux, ils peuvent donc tous être considérés comme des graphes de connaissances isolément ou combinés.

Graphes d'expertise

Dans l'état de l'art, des graphes pertinents pour la recherche d'experts appelés graphes d'expertise ont été introduit. Dans le chapitre 1, nous avons rappelé qu'un graphe d'expertise (Pal et Konstan 2010) est un graphe dont les sommets sont des individus (experts ou non) reliés entre eux par un type de relation prédéfini. Nous identifions les graphes de coauteurs et les graphes de citation entre auteurs (que nous appelons graphes de citation A) comme graphes d'expertise pertinents pour la recherche d'experts.

Graphes de coauteurs Les graphe de coauteurs représentent des auteurs de publications scientifiques reliés entre eux s'ils ont au moins une publication commune. Dans la figure 2.1, nous présentons un graphe de coauteurs attribué. Les auteurs ou sommets du graphe (A_i) sont étiquetés par des thématiques de publication (t_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les auteurs.

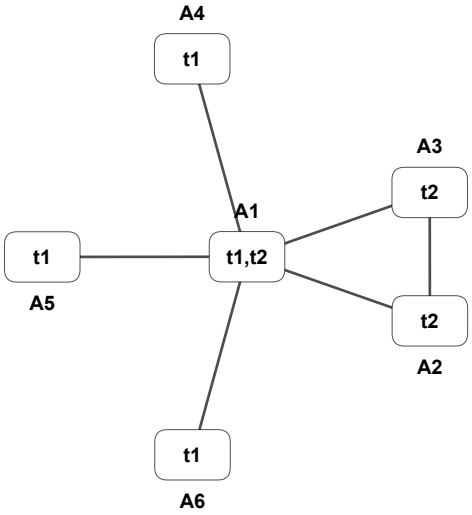


FIGURE 2.1 – Graphe de coauteurs

Identifier des experts au sein de graphes de coauteurs permet d'identifier des membres de la communauté scientifique ayant collaboré avec un grand nombre d'individus différents. Notre hypothèse est la suivante : *si un individu rédige des publications scientifiques avec des membres variés de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine*. En effet, en partageant un lien de coauteur avec un membre de la communauté scientifiques, les individus valident implicitement sa condition d'expert. Il s'agirait également d'un indicateur de la prolificité d'un individu, puisqu'un individu entretenant un grand nombre de liens de coauteurs produit au moins autant de publications scientifiques. Nous supposons que s'intéresser aux liens de coauteurs entretenus au sein de la communauté scientifique permettrait également de détecter des conflits d'intérêt lors de l'assignation d'un chercheur à un comité, par exemple. Nous pourrions également proposer de nouveaux collaborateurs potentiels à un individu.

Graphes de citation entre auteurs Les graphes de citation entre auteurs (graphes de citation A) représentent des auteurs de publications scientifiques. Dans un graphe de citation entre auteurs, la relation entre deux auteurs est dirigée. Il existe un lien d'un auteur A_i vers un auteur A_j si A_i cite au moins l'un des travaux de A_j. Dans la figure 2.2, nous présentons un graphe de citation A attribué. Les auteurs ou sommets du graphe (A_i) sont étiquetés par des thématiques de publication (t_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les auteurs.

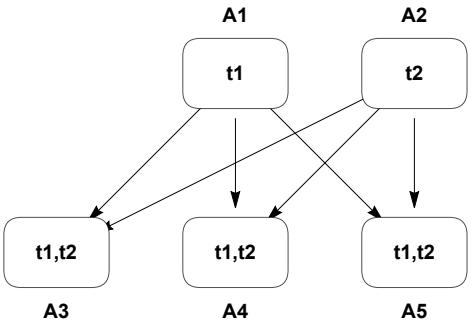


FIGURE 2.2 – Graphe de citation entre auteurs

Identifier des experts au sein de graphes de citation entre auteurs permet d'identifier des individus entretenant suffisamment de relations de citation. Les individus ayant été cités à de nombreuses reprises peuvent naturellement être considérés comme des experts. L'hypothèse d'expertise associée est la suivante : *si un individu est cité par un grand nombre de membres de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine.*

Graphes décrivant les documents sources d'expertise

D'autres graphes peuvent également se révéler utiles pour caractériser l'expertise d'individus. En effet, si la recherche d'experts consiste principalement en l'identification d'un individu capable d'exercer une fonction ou d'être une source d'information fiable, les documents sources d'expertise peuvent également être considérés comme des sources d'information fiables. Dans le milieu académique, l'identification des publications phares d'un domaine est une problématique importante. Les auteurs d'une publication phare d'un domaine de recherche peuvent être considérés comme experts de ce domaine. Nous identifions les graphes de copublication et les graphes de citation entre documents sources comme pertinents pour la recherche d'experts.

Graphes de copublication Les graphes de copublication représentent des documents sources d'expertise reliés entre eux s'ils ont au moins un auteur en commun. Dans la figure 2.3, nous présentons un graphe de copublication attribué. Les documents sources d'expertise ou sommets du graphe (D_i) sont étiquetés par des thématiques de publication (t_i) ainsi que par des auteurs (a_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les documents.

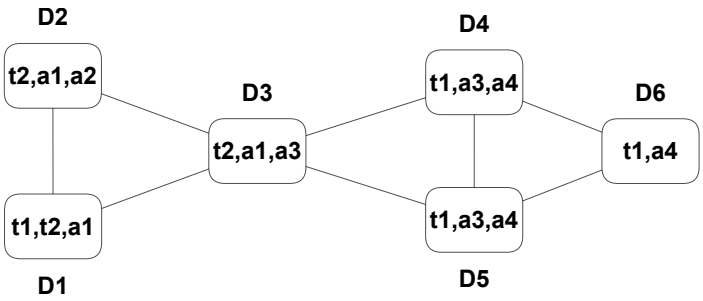


FIGURE 2.3 – Graphe de copublication

Identifier des documents sources d'expertise ayant suffisamment d'auteurs en commun permet, par projection, de suivre l'activité d'un individu travaillant sur des thématiques similaires ou produisant des travaux connexes. En utilisant les années ou périodes de publication dans le langage d'étiquetage, il est possible d'identifier des individus travaillant sur des thématiques récurrentes sur une longue période. Nous pouvons définir l'hypothèse d'expertise suivante : *si un individu rédige un grand nombre de publications scientifiques sur une thématique récurrente, alors il est probable qu'il s'agisse d'un expert de son domaine*. Il peut s'agir d'un indicateur de la prolificité d'un chercheur.

Enfin, les coauteurs d'un auteur prolifique peuvent éventuellement être considérés comme des experts, puisqu'un expert a collaboré avec eux. L'hypothèse d'expertise associée est la suivante : *si un membre de la communauté scientifique est coauteur d'un individu qui rédige un grand nombre de publications scientifiques sur une thématique récurrente, alors il est probable qu'il s'agisse également d'un expert de son domaine*. Nous supposons également que s'intéresser aux documents produits par un ensemble d'auteurs ayant collaboré ensemble permettrait de proposer des documents connexes lorsqu'une requête aboutirait à la proposition d'un document comme source d'expertise pertinente.

Graphes de citation entre documents source d'expertise Les graphes de citation entre documents source d'expertise (graphes de citation D) représentent des documents source d'expertise. Il existe une relation dirigée d'un sommet d_1 vers un sommet d_2 si le document d_1 cite le document d_2. Dans la figure 2.4, nous présentons un graphe de citation D attribué. Les documents sources d'expertise ou sommets du graphe (D_i) sont étiquetés par des thématiques de publication (t_i) et des auteurs (a_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les documents.

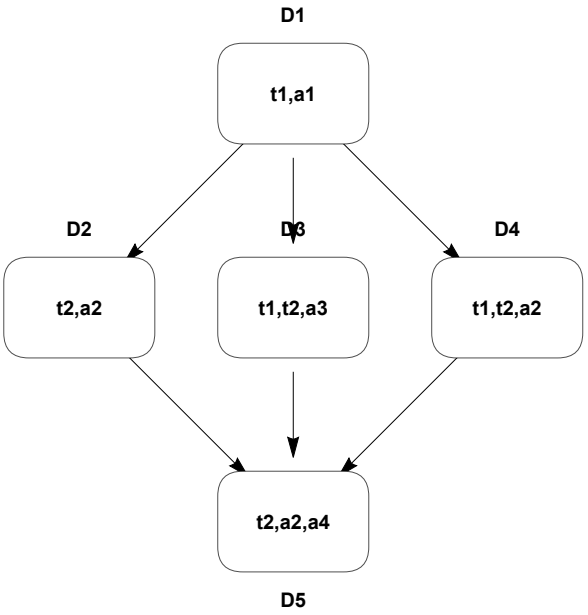


FIGURE 2.4 – Graphe de citation entre documents sources d'expertise

Identifier des documents sources d'expertise impliqués dans suffisamment de relations de citation permet d'identifier des publications phares, c'est-à-dire des publications scientifiques ayant été abondamment citées. L'hypothèse d'expertise associée est la suivante : *si un individu est auteur d'une publication phare, alors il est probable qu'il s'agisse d'un expert de son domaine.*

Nous supposons qu'identifier des documents sources d'expertise impliqués dans suffisamment de relations de citation permettrait également identifier les articles de revue, c'est-à-dire les documents citant un grand nombre de publications phares dans un domaine de publication précis. Les articles de revue constituent une source d'expertise extrêmement pertinente et sont un bon point d'entrée à l'apprentissage d'un domaine de recherche. L'hypothèse d'expertise associée est la suivante : *si un individu est auteur d'un article de revue, alors il est probable qu'il s'agisse d'un expert de son domaine.*

Nous supposons également que lorsque des articles de revue pour le domaine considéré et sur la période temporelle considérée n'existent pas, il serait possible de suggérer un ensemble de publications phares du domaine à citer pour la rédaction d'un article de revue. La rédaction d'une revue de la littérature est une tâche fastidieuse et chronophage, utile pour la communauté scientifique et consacrant l'acquisition d'un certain niveau d'expertise sur les thématiques qui y sont abordées, puisqu'elle exige une compréhension globale d'une thématique de recherche. Une revue de la littérature nécessitant généralement de réaliser un grand nombre de citations différentes sur une thématique récurrente, proposer un ensemble de publications phares à citer permettrait de faciliter le travail de compilation des chercheurs.

Enfin, nous proposons une dernière hypothèse d'expertise associée au graphe de citation D : *si un individu est auteur d'un article citant un grand nombre de publications*

phares, alors il est probable qu'il s'agisse d'un expert de son domaine. En effet, cette hypothèse se base sur l'indicateur d'expertise suivant : si un auteur est capable d'identifier les références appropriées lorsqu'il rédige une publication scientifique sur une thématique particulière, alors il est probable qu'il s'agisse d'un individu possédant une vision précise du domaine, donc un expert.

Graphes décrivant les expertises

D'autres graphes peuvent également se révéler utiles pour acquérir des connaissances cruciales sur les thématiques d'un domaine de recherche et les liens existant entre elles. En effet, il peut être utile de connaître les liens entre thématiques de recherche afin de hiérarchiser les thématiques associées à un domaine de recherche inconnu pour un utilisateur, étendre une requête ou découvrir des tendances de thématiques de recherche et leur évolution au cours du temps. Nous identifions les graphes de co-occurrences et les graphes de citation entre expertises comme pertinents pour la recherche d'experts.

Graphes de co-occurrences Les graphes de co-occurrences ont pour sommets des expertises reliées si elles co-occurrent dans au moins une publication commune. Dans la figure 2.5, nous présentons un graphe de co-occurrences attribué. Les expertises ou sommets du graphe (E_i) sont étiquetés par des thématiques de publication (t_i) et des auteurs (a_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les expertises.

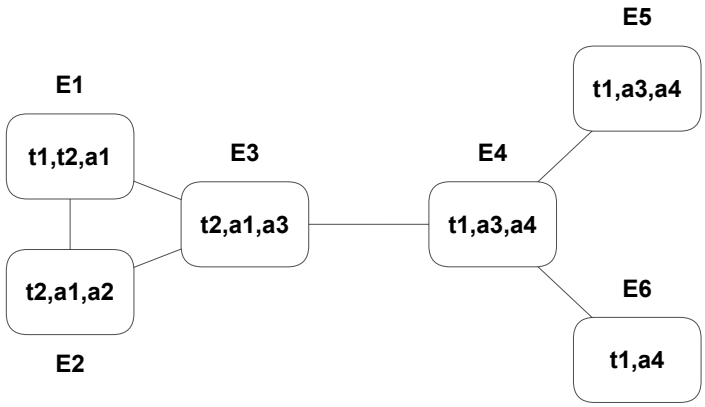


FIGURE 2.5 – Graphe de co-occurrences

Nous supposons qu'identifier des expertises apparaissant simultanément dans un nombre suffisamment grand de publications permettrait d'identifier les liens existant entre les différentes thématiques de recherche ainsi que les tendances de recherche et similarités sémantiques entre thématiques. Nous supposons également que l'acquisition de ces connaissances permettrait à terme d'automatiser la description d'un domaine de recherche, d'en identifier les tendances sans avoir besoin de consulter un expert du domaine et d'étendre une

requête. Par exemple, pour la requête « Quels sont les experts en reconnaissance d'entités nommées ? », il est possible de proposer l'extension : « Quels sont les experts en traitement du langage naturel ? », puisque le traitement du langage naturel et la reconnaissance d'entités nommées sont des thématiques de recherche similaires.

Graphes de citation entre expertises Les graphes de citation entre expertises (graphes de citation E) sont des graphes orientés, dont le sommets sont des expertises. Il existe une relation dirigée d'un sommet e_1 vers un sommet e_2 si l'expertise e_1 apparaît dans le document source d_1, l'expertise e_2 apparaît dans le document source d_2 et le document d_1 cite le document d_2. Dans la figure 2.6, nous présentons un graphe de citation E attribué. Les expertises ou sommets du graphe (E_i) sont étiquetés par des thématiques de publication (t_i) et des auteurs (a_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les expertises.

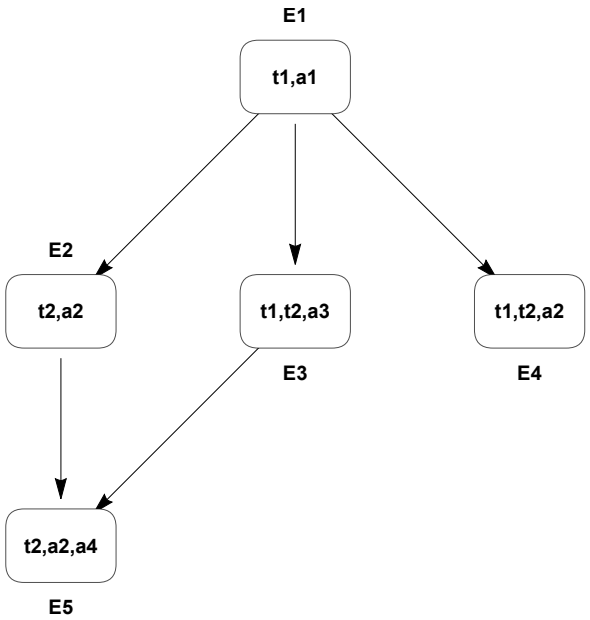


FIGURE 2.6 – Graphe de citation entre expertises

Considérons un ensemble d'expertises apparaissant dans un ensemble de documents cités, ainsi qu'un ensemble d'expertises apparaissant dans les documents les citant. Nous supposons qu'identifier des expertises apparaissant dans des documents impliqués dans suffisamment de relations de citation permettrait d'identifier l'émergence de thématiques de recherche à partir de domaines de recherche qui les ont inspirées. Nous supposons également que cela permettrait d'identifier des liens de similarité entre expertises.

Graphes bipartis

Jusqu'à présent, nous avions considéré des graphes n'ayant qu'un seul type de sommets. Or, il existe des graphes bipartis composés de deux ensembles de sommets ayant

chacun un rôle défini. Par exemple, un graphe biparti peut être composé d'un ensemble de sommets correspondant à des publications et à leurs auteurs. Nous identifions trois graphes bipartis pertinents pour la recherche d'experts dans le milieu académique : le graphe biparti publication-auteurs, auteurs → publications citées et publications → auteurs cités.

Graphe biparti publications-auteurs Les graphes publications-auteurs sont des graphes bipartis dont les sommets sont composés de publications et de leurs auteurs. Il existe une relation entre un sommet publication et un sommet auteur si l'auteur a rédigé la publication. Dans la figure 2.7, nous présentons un graphe biparti publications-auteurs attribué. Les sommets du graphe correspondant à des documents sources d'expertise (D_i) sont étiquetés par des thématiques de publication (t_i) et des auteurs (a_i). Les sommets du graphe correspondant à des auteurs (A_i) sont étiquetés par des thématiques de publication (t_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les documents et les auteurs.

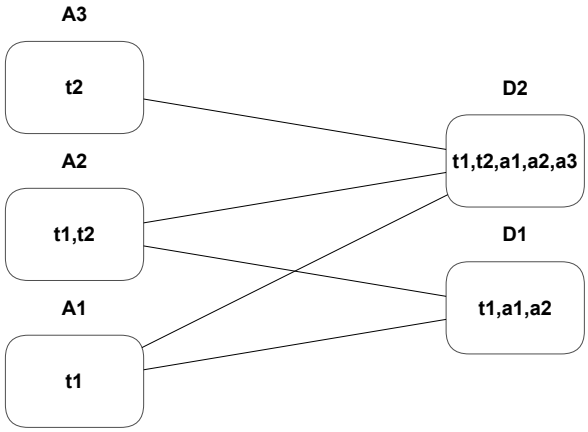


FIGURE 2.7 – Graphe biparti publications-auteurs

Le graphe de coauteurs est une projection du graphe biparti publications-auteurs. Il peut être utile de manipuler le graphe biparti plutôt que le graphe de coauteurs afin de conserver l'information concernant les relations de paternité existant entre auteurs et documents. L'intérêt du graphe biparti publications-auteurs est donc similaire à celui du graphe de coauteurs en ajoutant l'information concernant les liens de paternité vers les documents.

Graphe biparti auteurs → publications citées Les graphes auteurs → publications citées sont des graphes bipartis orientés dont les sommets sont composés d'auteurs et des publications qu'ils citent. Il existe une relation dirigée d'un sommet auteur vers un sommet publication si l'auteur cite la publication. Dans la figure 2.8, nous présentons un graphe biparti auteurs → publications citées attribué. Les sommets du graphe correspondant à

des documents sources d'expertise (D_i) sont étiquetés par des thématiques de publication (t_i) et des auteurs (a_i). Les sommets du graphe correspondant à des auteurs (A_i) sont étiquetés par des thématiques de publication (t_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les documents et les auteurs.

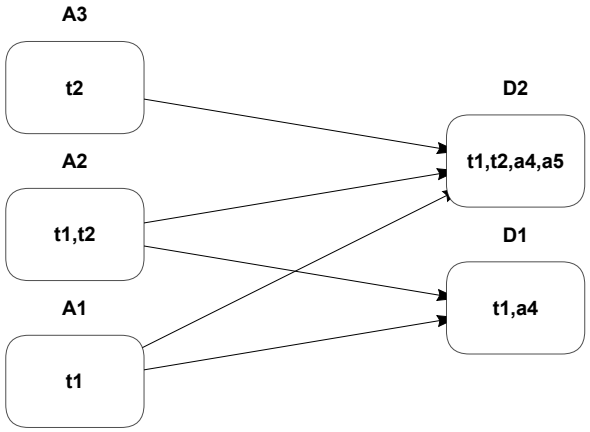


FIGURE 2.8 – Graphe biparti auteurs → publications citées

Identifier des experts au sein de graphes auteurs → publications citées permet notamment d'identifier des auteurs de publications phares, c'est-à-dire des auteurs de documents fortement cités par les membres de la communauté scientifique. Les auteurs de publications phares peuvent naturellement être considérés comme des experts. L'hypothèse d'expertise associée est la suivante : *si un individu est auteur d'une publication phare, alors il est probable qu'il s'agisse d'un expert de son domaine.*

De plus, les individus qui citent des documents de façon appropriée peuvent éventuellement être considérés comme des experts. L'hypothèse d'expertise associée est la suivante : *si un individu cite des publications phares d'un domaine, alors il est probable qu'il s'agisse d'un expert de ce domaine.*

Graphe biparti publications → auteurs cités Les graphes publications → auteurs cités sont des graphes bipartis orientés dont les sommets sont composés de publications et des auteurs qu'elles citent. Il existe une relation dirigée d'un sommet publication vers un sommet auteur si la publication cite l'auteur. Dans la figure 2.9, nous présentons un graphe biparti publications → auteurs cités attribué. Les sommets du graphe correspondant à des documents sources d'expertise (D_i) sont étiquetés par des thématiques de publication (t_i) et des auteurs (a_i). Les sommets du graphe correspondant à des auteurs (A_i) sont étiquetés par des thématiques de publication (t_i). Il est également possible d'utiliser les années ou périodes de publication pour étiqueter les documents et les auteurs.

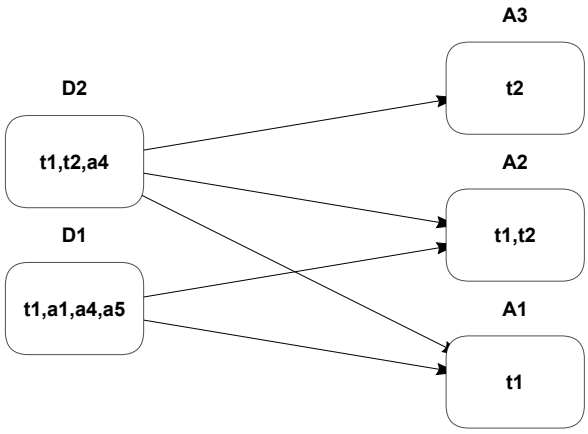


FIGURE 2.9 – Graphe biparti publications → auteurs cités

Identifier des experts au sein de graphes auteurs → publications citées permet d'identifier les individus fortement cités par la littérature. Les auteurs fortement cités peuvent naturellement être considérés comme des experts, puisqu'ils ont produit au moins une contribution notable dans le domaine. L'hypothèse d'expertise associée est la suivante : *si un individu est fortement cité par les membres de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine.*

De plus, les publications qui citent des auteurs de façon appropriée peuvent éventuellement être considérés comme des articles de revue ou des publications pertinentes dans le domaine abordé. L'hypothèse d'expertise associée est la suivante : *si un individu cite des auteurs appropriés du domaine dans l'une de ses publications, alors il est probable qu'il s'agisse d'un expert de ce domaine.*

2.3 Conclusion

De nombreuses initiatives de publication de données scientifiques sur le Web sémantique ont fleuri au sein de l'état de l'art et donné naissance à des graphes de connaissances scientifiques. Ces graphes permettent de représenter et d'organiser efficacement les connaissances concernant le milieu académique et respectent un formalisme strict. Il s'agit de structures de représentation des connaissances réutilisables et facilement interprétables par une machine.

Les initiatives de publication et d'ouverture des données dans un objectif d'interprétabilité et d'interopérabilité sont favorisées dans l'état de l'art. Notre objectif est de produire des connaissances utiles pour la communauté scientifiques, réutilisables par les membres de la communauté scientifique et connectées aux productions déjà existantes. Dans le cadre de cet objectif d'interprétabilité et d'interopérabilité, nous suggérons d'identifier les expertises au sein des publications scientifiques en utilisant des ontologies décrivant les domaines de recherche. Nous proposons d'utiliser la Computer Science Ontology (SALATINO

et al. 2018b) décrivant le domaine de l'informatique.

Nous avons identifié un ensemble de graphes attribués pertinents pour la recherche d'experts que nous utilisons comme étape intermédiaire de représentation des connaissances avant la génération du graphe de connaissances scientifique. Ces graphes permettent de représenter de manière la plus exhaustive possible les interactions existant entre les entités du monde académique. Dans le chapitre 3, nous explorons ces graphes attribués à l'aide d'une méthode de fouille de graphe afin d'identifier les experts et leurs expertises associées à l'aide des hypothèses d'expertise que nous avons émises durant ce chapitre. La méthode de fouille de graphe que nous employons, l'abstraction de graphe, permet de se focaliser sur des zones denses du graphe, les cœurs de graphe. Nous supposons que les cœurs de graphe matérialisent des communautés permettant d'identifier des experts et leurs expertises associées. Dans le chapitre 4, nous proposons un modèle basé sur trois schémas RDF existants (Sinha et al. 2015 ; Angioni et al. 2020 ; Silvello et al. 2017) permettant de générer un graphe de connaissances scientifique à partir des graphes attribués pertinents pour la recherche d'experts ainsi que des experts et expertises associées identifiés à l'aide de l'abstraction de graphe.

Chapitre 3

Abstraction de graphe

Les réseaux sociaux ou les grands graphes constituent des structures largement répandues qui suscitent un intérêt grandissant auprès des chercheurs en intelligence artificielle. L'une des problématiques les plus répandues consiste à simplifier ces structures complexes, à les *abstraire* en proposant de les réduire à des structures plus petites représentant une version condensée du réseau original, c'est-à-dire véhiculant les informations les plus pertinentes au sein de ce réseau. La détection de communautés en particulier consiste en l'identification de sous-structures dont les sommets sont d'une part densément connectés aux nœuds faisant partie de la communauté et d'autre part faiblement connectés au reste du réseau. En considérant un réseau de collaboration scientifique dans lequel les experts seraient fortement connectés entre eux et partageraient un ensemble d'expertises communes, il est possible de modéliser la recherche d'experts comme un problème de détection de communautés dans des réseaux de collaboration scientifique. Les communautés au sein des réseaux peuvent être vues comme des structures cœur/périphérie, le cœur étant la partie interne densément connectée et la périphérie l'ensemble des sommets extérieurs entretenant peu de relations avec le cœur (Soldano, Santini, Bouthinon et Lazega 2017).

Le cœur d'un graphe est un sous-graphe maximal dans lequel l'ensemble des sommets respectent une contraire topologique en son sein, c'est-à-dire une contrainte de connectivité. Identifier le cœur d'un graphe est une approche classique d'exploration de la structure des graphes complexes. Dans le domaine de la fouille de graphes attribués, des travaux récents ont combiné la fouille de motifs au sein des graphes attribués avec la considération de contraintes de connectivité. La méthode résultant de cette combinaison s'appelle l'abstraction de graphe. L'originalité de l'abstraction de graphe consiste en l'énumération de motifs clos abstraits, c'est-à-dire de motifs clos au sein de cœurs de graphes attribués.

Nous proposons de modéliser le problème de l'identification des experts et de leurs expertises associées comme un problème de découverte de connaissances à l'aide de fouille de motifs au sein de cœurs de graphes attribués. Un graphe attribué est un graphe dont les sommets sont étiquetés à l'aide d'un langage de description. Des exemples de graphes attribués pertinents pour la représentation de connaissances dans le cadre de la recherche d'experts dans le milieu académique sont les graphes d'expertise, comprenant les graphes de citation et les graphes de coauteurs. Les graphes de citation sont des graphes représentant des chercheurs liés entre eux par une relation dirigée si l'un cite l'autre. Les graphes

de coauteurs représentent les chercheurs liés entre eux s'ils ont déjà publié ensemble. Dans ces deux graphes, les chercheurs peuvent être étiquetés par leurs thématiques de publication. De tels graphes permettent de matérialiser une validation par les pairs des individus. Plus un individu entretient de liens de collaboration avec les autres membres de la communauté scientifique, plus sa réputation est considérée comme élevée. Les cœurs de graphes d'expertise représentent donc des communautés d'experts et les motifs clos abstraits, c'est-à-dire l'ensemble des expertises communes maximales qu'ils partagent au sein du cœur correspondent à leurs expertises communes maximales associées.

La fouille de motifs clos au sein des cœurs de graphe est pertinente pour l'identification de caractéristiques communes maximales partagées par des ensembles de sommets formant des cœurs. Adaptée à notre problématique de recherche d'experts, notre hypothèse est que cette méthode peut permettre d'explorer les graphes d'expertise afin d'énumérer les expertises communes maximales partagées par des ensembles d'experts validés par leurs pairs. Nous supposons que la validation par les pairs est matérialisée par les propriétés topologiques associées aux cœurs des graphes. Ces hypothèses sont suggérées par les récents travaux utilisant l'abstraction de graphe pour explorer les graphes de citation, à l'aide de la propriété topologique associée au cœur *hub*-autorité (Soldano, Santini, Bouthinon et Lazega 2017), inspirée par l'algorithme HITS (Kleinberg 1999).

Dans ce chapitre, nous abordons l'état de l'art de l'abstraction de graphe et définissons les propriétés topologiques associées aux cœurs que nous utilisons dans le cadre de l'adaptation de l'abstraction de graphe à notre problématique de recherche d'experts à partir de textes. L'abstraction de graphe étant une méthode exploratoire générant de nombreux résultats à interpréter, nous nous intéressons également à une méthode de sélection des motifs clos abstraits produits. Enfin, nous proposons une méthode de représentation étendue des motifs clos abstraits afin d'en assouplir la définition.

3.1 Travaux connexes

Les graphes sont des structures de représentation des connaissances largement utilisées dans l'état de l'art. Les graphes de terrain (aussi appelés grands graphes ou graphe complexes), sont des graphes représentant des réseaux construits à partir de données réelles. Les chercheurs ont proposé de nombreuses méthodes d'exploration de tels graphes. Elles peuvent être séparées en deux catégories : les méthodes exploitant les propriétés des graphes (par exemple le degré, le diamètre, le nombre et la taille des composantes connexes) et les méthodes de découverte de motifs locaux permettant d'identifier des sous-graphes particuliers (Mougel, Plantevit et al. 2011).

Parmi les méthodes fondées sur la découverte de motifs locaux dans l'objectif d'identifier des sous-graphes particuliers, nous pouvons citer les travaux de Mougel *et al.*, qui se sont intéressés à l'extraction d'ensembles de cliques dites homogènes, à l'aide

de contraintes (Mougel, Plantevit et al. 2011 ; Mougel, Rigotti et Gandrillon 2012). Le principe de leur méthode est le suivant : dans un graphe attribué, les auteurs extraient des ensembles maximaux de cliques tels que les sommets impliqués respectent une conjonction de contraintes. Cette conjonction de contraintes est paramétrable et peut être constituée par des contraintes s'appliquant sur le nombre de cliques séparées, la taille des cliques ainsi que le nombre d'étiquettes partagées (Mougel, Plantevit et al. 2011). Les ensembles maximaux de cliques identifiés à l'aide de cette méthode sont dits homogènes. Les auteurs ont notamment appliqué leur méthode à un graphe de coauteurs attribué issu de DBLP [1]. Dans ce graphe, le langage de description associe aux auteurs de publications scientifiques les conférences dans lesquelles ceux-ci publient. Les auteurs proposent une analyse quantitative des motifs obtenus à l'aide de l'extraction d'ensembles maximaux de cliques homogènes. Ils présentent également plus en détails quelques motifs qui mettent en évidence des interactions non triviales entre les cliques. Ces résultats semblent valider l'approche des auteurs dans le cadre de la détection de k-communautés fréquentes dans un graphe. Cependant, à notre connaissance, cette méthode n'a pas été évaluée dans le cadre de la recherche d'experts et n'a pas fait l'objet d'une analyse qualitative des motifs extraits à l'aide d'un protocole d'évaluation.

L'abstraction de graphe permet d'énumérer les motifs clos abstraits, c'est-à-dire des ensembles maximaux d'attributs partagés par des sommets impliqués dans des sous-graphes particuliers, les cœurs de graphe. Il s'agit donc également d'une méthode de découverte de motifs locaux permettant d'identifier des sous-graphes particuliers.

3.2 Cœurs de graphe

L'abstraction de graphe tire parti de la notion de cœur. Considérons un graphe G_V induit par un ensemble de sommets V. Le cœur du graphe G_V est un sous-graphe G_C induit par un ensemble maximal de sommets C (ou d'arêtes) vérifiant une propriété topologique $P(v, C)$ au sein du cœur, c'est-à-dire une contrainte de connectivité respectée par l'ensemble des sommets du cœur en son sein, pour tout $v \in C$. Des définitions variées de cœurs sont obtenues en fonction de la propriété topologique associée (Batagelj et Zaversnik 2011). Le cœur de $G_V(V, E_V)$ appelé $G_C(C, E_C)$ est obtenu en recherchant un point fixe par itérations successives en retirant du graphe G_V tout noeud (ou toute arête) ne respectant pas la propriété topologique P.

La première définition de cœur est celle du k-*core* (Seidman 1983) pour laquelle l'ensemble des sommets du cœur vérifient la propriété $P_{k-core}(v, C)$ suivante : pour tout sommet v de C, v a un degré supérieur ou égal à k au sein du cœur G_C. Si tout sommet du cœur G_C a un degré supérieur ou égal à k, cela signifie que chacun des sommets est impliqué dans au moins k relations différentes. La notion de cœur a été généralisée

1. Digital Bibliography & Library Project (DBLP) `https://dblp.org`

(Batagelj et Zaversnik 2011) et permet la définition de nouveaux cœurs garantissant d'autres propriétés topologiques. Pour les graphes non dirigés, il s'agit des cœurs k-dense et k-*nearstar*. Le cœur k-dense (Saito, Yamada et Kazama 2009) tire avantage d'une relaxation de la notion de clique pour identifier des zones fortement connexes du graphe. Une clique est un sous-ensemble de sommets induisant un sous-graphe complet de G, c'est-à-dire un graphe dont tous les sommets sont reliés deux à deux. Quant au cœur k-*nearstar* (Soldano et Santini 2014), il introduit une extension du k-*core* à ses premiers voisins. Le *hub-authority core* (Soldano, Santini, Bouthinon et Lazega 2017) a été introduit pour la recherche de collaborations dans les graphes sociaux dirigés. Tout cœur de graphe est obtenu en recherchant un point fixe à l'aide d'itérations successives consistant au retrait de tout nœud (ou arête) ne respectant pas la propriété topologique P au sein du graphe.

3.2.1 Cœur k-*core*

Le cœur k-*core* d'un graphe $G_V = G(V, E_V)$ est le sous-graphe maximal $G_C = G(C, E_C)$ tel que $C \subseteq V$ et $\forall v \in C, P_{k-core}(v, C)$ est vraie, c'est-à-dire que tous les sommets de G_C vérifient la propriété P_{k-core}. Cette propriété garantit que tous les sommets du k-*core* G_C ont un degré supérieur ou égal à k (Seidman 1983). Cette abstraction s'applique aux graphes non orientés. Une version orientée, le k-l *D-core* a également été définie dans l'état de l'art. Dans le k-l *D-core*, les degrés entrants et sortants des sommets appartenant au k-l *D-core* doivent avoir un degré supérieur ou égal à k et l respectivement au sein du k-l *D-core* (Giatsidis, Thilikos et Vazirgiannis 2013). La définition de la propriété P_{k-core} est la suivante :

Définition de la propriété P_{k-core}

Soient X un ensemble de sommets d'un graphe G, G_X le sous graphe de G induit par X et v un sommet. $P_{k\_core}(v, X)$ est vraie si et seulement si le sommet v a un degré supérieur ou égal à k dans le graphe G_X

La figure 3.1 illustre les étapes de l'obtention d'un 2-*core* $G_C = G(C, E_C)$ (en bas et en bleu) à partir d'un graphe $G_V = G(V, E_V)$ (en haut). Dans le graphe G_V, le sommet 5 du graphe ne respecte pas la propriété topologique d'un 2-*core*, puisque son degré est strictement inférieur à 2. Le sommet 5 est donc éliminé. Après élimination du sommet 5, le sommet 4 ne respecte plus la propriété topologique au sein du 2-*core* à son tour, bien qu'il soit de degré 2 dans le graphe originel. Il est donc également supprimé du 2-*core*. Ainsi, le 2-*core* G_C du graphe G_V est obtenu en atteignant un point fixe constitué par les sommets 1, 2 et 3 respectant tous la propriété topologique P_{k-core}. Tout sommet du

k-*core* est lié à au moins k autres sommets. Dans le cas du 2-*core*, tout sommet est lié à 2 autres sommets.

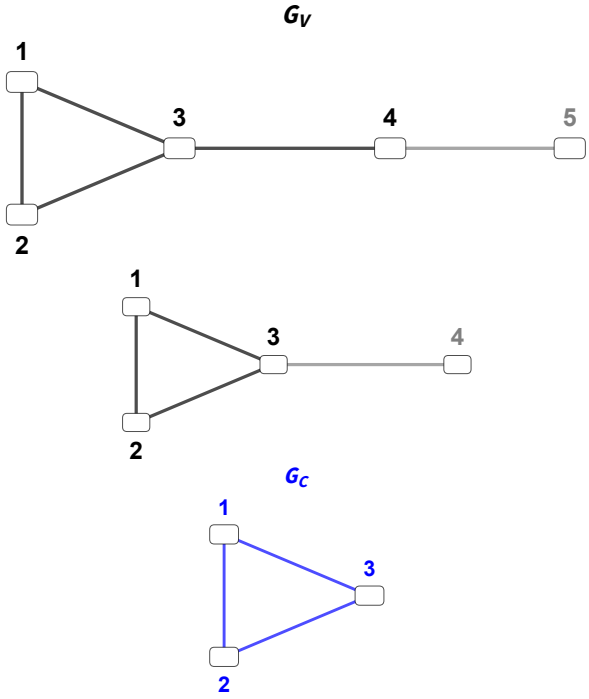


FIGURE 3.1 – Abstraction k-*core* : exemple d'un 2-*core*

3.2.2 Cœur k-*nearstar*

Le cœur k-*nearstar* d'un graphe $G_V = G(V, E_V)$ est le sous-graphe maximal $G_C = G(C, E_C)$ tel que $C \subseteq V$ et $\forall v \in C, P_{k-nearstar}$ est vraie, c'est-à-dire que tous les sommets de G_C vérifient la propriété $P_{k-nearstar}$. Cette propriété garantit que tous les sommets appartenant au k-*nearstar* ont un degré supérieur ou égal à k (étoiles) ou un voisin de degré supérieur ou égal à k au sein du k-*nearstar* (satellites) (SOLDANO et SANTINI 2014). Un sommet peut être à la fois satellite et étoile. Un satellite peut être associé à plusieurs étoiles. Cette abstraction s'applique aux graphes non orientés. Il s'agit d'une relaxation du k-*core* à ses voisins les plus proches. La définition de la propriété $P_{k-nearstar}$ est la suivante :

Définition de la propriété $P_{k-nearstar}$

Soient X un ensemble de sommets d'un graphe G, G_X le sous graphe de G induit par X et v un sommet. $P_{k\_nearstar}(v, X)$ est vraie si et seulement si le sommet v a un degré supérieur ou égal à k dans le graphe G_X ou si il est lié par une arête de G_X à un sommet $v' \in x$ ayant un degré supérieur ou égal à k dans G_X.

La figure 3.2 illustre les étapes de l'obtention d'un 5-*nearstar* $G_C = G(C, E_C)$ (en bas et en bleu) à partir d'un graphe $G_V = G(V, E_V)$ (en haut). Dans le graphe G_V, les sommets 7 et 8 ne respectent pas la propriété topologique. En effet, ils ne sont pas de degré supérieur ou égal à 5 et ne possèdent pas non plus de voisin de degré supérieur ou égal à 5. Les sommets 7 et 8 sont donc éliminés. Le 5-*nearstar* G_C du graphe G_V est obtenu en atteignant un point fixe constitué par les sommets 1, 2, 3, 4, 5 et 6. En effet, le sommet 1 est de degré 5 au sein du 5-*nearstar* tandis que les sommets 2, 3, 4, 5 et 6 ont tous un voisin de degré 5 dans le k-*nearstar* : le sommet 1.

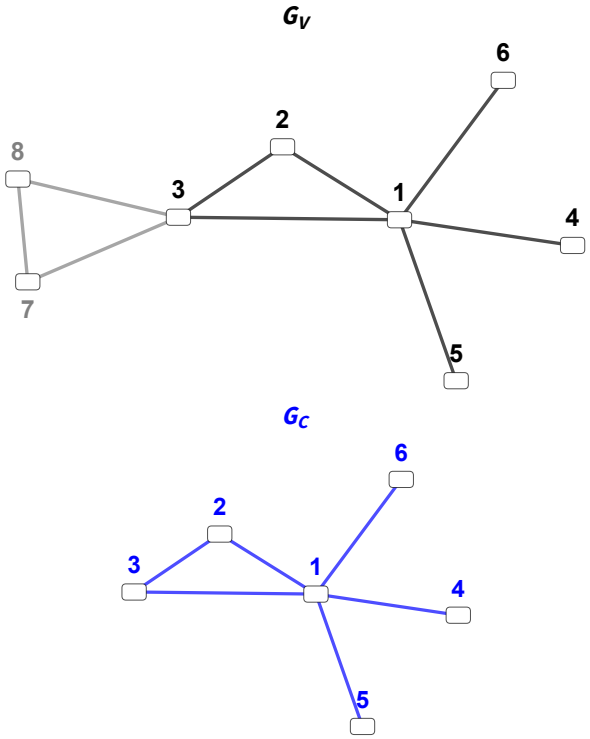


FIGURE 3.2 – Abstraction k-*nearstar* : exemple d'un 5-*nearstar*

3.2.3 Cœur k-dense

Le cœur k-dense d'un graphe $G = G(V, E)$ est le sous-graphe $G_C(V_C, E_C)$ maximal induit par un sous-ensemble d'arêtes E_C, tel que toutes les arêtes de G_C vérifient la propriété $P_{k-dense}$.

Cette propriété garantit que chacune des arêtes du k-dense G_C impliquent des sommets v_i et v_j ayant au moins $k-2$ voisins communs dans G_C (SAITO, YAMADA et KAZAMA 2009).

Cette abstraction s'applique aux graphes non orientés. Contrairement aux abstractions précédemment décrites, il s'agit d'une abstraction d'arêtes et non de sommets. La définition de la propriété $P_{k-dense}$ est la suivante :

Définition de la propriété $P_{k-dense}$

Soient E_X un sous-ensemble d'arêtes d'un graphe G, G_X le sous-graphe induit par E_X et (v_1, v_2) une arête. $P_{k-dense}((v_1, v_2), E_X)$ est vraie si et seulement si les sommets v_1 et v_2 partagent au moins $k-2$ voisins communs dans G_X.

La figure 3.3 illustre l'obtention d'un 4-dense $G_C = G(C, E_C)$ (en bas et en bleu) à partir d'un graphe $G_V = G(V, E_V)$ (en haut). Dans le graphe G_V, les arêtes $3 \leftrightarrow 5$ et $4 \leftrightarrow 5$ ne satisfont pas la propriété du 4-dense, puisque 3 et 5 (respectivement 4 et 5) n'ont pas au moins 2 voisins communs. Les arêtes (3,5) et (4,5) sont supprimées et par conséquent le sommet 5 disparaît. Les arêtes constituant le cœur 4-dense G_C du graphe G_V sont (1,2), (1,3), (1,4), (2,3), (2,4) et (3,4). En effet, toutes ces arêtes impliquent des sommets ayant au moins 2 voisins en commun dans le k-dense.

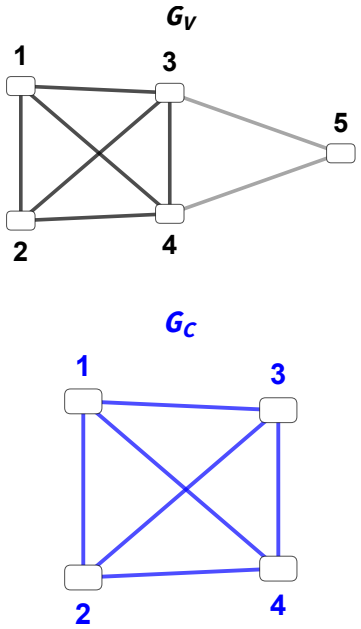


FIGURE 3.3 – Abstraction k-dense : exemple d'un 4-dense

3.2.4 Cœur h-a-*hub*-autorité

Le cœur h-a-*hub*-autorité d'un graphe orienté $G_V = G(V, E_V)$ est le sous-graphe maximal $G_C = G(C, E_C)$ tel que $C \subseteq V$ et $\forall v \in C, P_{h-a-hub-autorité}$ est vraie, c'est-à-dire que tous les sommets de G_C vérifient la propriété $P_{h-a-hub-autorité}$. Cette propriété garantit que tous les sommets appartenant au h-a-*hub*-autorité sont un *hub*, une autorité ou les deux rôles à la fois (SOLDANO, SANTINI, BOUTHINON et LAZEGA 2017) tel que défini ci-après.

En effet, cette abstraction de graphe permet de prendre en compte les différents rôles que peut remplir un sommet vis à vis des arcs qui le lient au reste du graphe. Dans un graphe dirigé, un sommet peut être la source d'une arête sortante ou la cible d'une arête entrante. Un noeud est défini comme un hub, h-*hub* en abrégé, (respectivement une autorité, a-aut en abrégé) s'il présente h arêtes sortantes (respectivement a arêtes entrantes) vers des nœuds autorités (respectivement depuis des nœuds *hubs*). Les définitions de *hubs* et autorités sont donc liées. Les noeuds peuvent remplir les deux rôles et ainsi être à la fois *hubs* et autorités. Cette abstraction s'applique uniquement aux graphes orientés.

Soit $H \subseteq C$ et $A \subseteq C$ les noeuds *hubs* et autorités du cœur G_C h-a-*hub*-autorité de G_V. Soit $d^H(v)$ le degré sortant du sommet v vers les noeuds autorités A. Soit $d^A(v)$ le degré entrant du sommet v depuis les noeuds *hubs* H. La définition de la propriété $P_{h-a-hub-autorité}$ est la suivante :

Définition de la propriété $P_{h-a-hub-autorité}$

Soient X un ensemble de sommets d'un graphe G, G_X le sous graphe de G induit par X, $H \subseteq X$ et $A \subseteq X$ deux sous-ensemble de sommets correspondant respectivement aux sommets *hubs* et autorités tels que $A \cup H = X$.

$P_{hub-autorité}(v, X)$ est vraie si et seulement si :

— si $v \in H$, $d^H(v) \geq h$

— si $v \in A$, $d^A(v) \geq a$

La figure 3.4 illustre l'obtention du 2-2-*hub*-autorité G_C (en bas et en bleu) à partir d'un graphe G_V (en haut). Dans le graphe G_V, les sommets 1 et 8 ne satisfont pas la propriété du 2-2-*hub*-autorité. En effet, les degrés entrants et sortants de 1 et 8 sont inférieurs à 2, ils ne peuvent alors n'être considérés ni comme des *hubs* ni comme des autorités. En raison de la suppression des sommets précédents, 4 ne peut plus être considéré comme un *hub* potentiel, puisque son degré sortant est inférieur à 2. De même, 5 ne peut plus être considéré comme une autorité potentielle, puisque son degré entrant est inférieur à 2. Les sommets 4 et 5 sont donc supprimés. Les *hubs* du 2-2-*hub*-autorité du graphe sont 2 et 3 (de degrés sortants 2) et les autorités sont 5 et 6 (de degrés entrants 2). Le sommet 7 est de degré entrant 2 (et non pas 3) car le sommet 4 ne peut être considéré comme une autorité et a été supprimé.

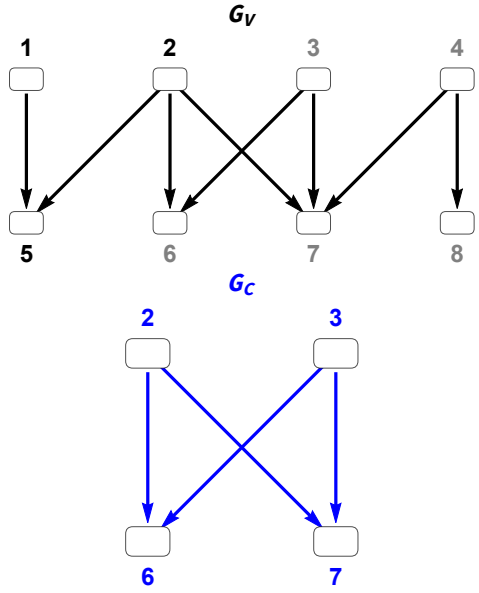


FIGURE 3.4 – Abstraction h-a-*hub*-autorité : exemple du 2-2-*hub*-autorité

3.3 Fouille de motifs clos abstraits

L'abstraction de graphe consiste en la fouille de motifs au sein des cœurs de graphes attribués. Elle étend la fouille de motifs clos et le cadre de l'analyse formelle de concept (SOLDANO, SANTINI, BOUTHINON et LAZEGA 2017) au cas où l'ensemble support des motifs est constitué par les noeuds d'un graphe attribué $G_V(V, E_V, I, L)$, I étant l'ensemble d'étiquettes (aussi appelés *items*) permettant de décrire les sommets et L le langage de description des sommets tel que $L = 2^I$. L est l'ensemble de sous-ensembles que l'on peut former à partir de I.

Soit $desc(v)$ la description d'un objet v par un ensemble d'étiquettes issues de I. Un objet v est reconnu par un motif q du langage L si sa description $desc(v)$ est plus spécifique que q (q est inclus dans $desc(v)$). On a $desc(v) \subseteq I$ et $desc(v) \in L$.

L'extension d'un motif $ext(q)$ est le sous-ensemble des noeuds du graphe qui sont reconnus par le motif q. Cette extension définit l'ensemble support du motif. On a : $ext(q) = \{v_i \in V, desc(v_i) \supseteq q\}$. À chaque motif q correspond donc un sous graphe induit par son extension.

Dans la figure 3.5, représente un graphe G(V,E,I) avec $I = \{a, b, c\}$, et $desc(1) = \{b, c\}$, $desc(2) = \{a, b, c\}$, $desc(3) = \{b, c\}$, $desc(4) = \{a, c\}$, $desc(5) = \{a, c\}$, $desc(6) = \{a, b\}$, $desc(7) = \{a, b\}$, et $desc(8) = \{a, b\}$.

Si l'on considère le motif a on a :

— ext(a) = {2,4,5,6,7,8}.

— Le sous-graphe induit par $ext(a)$ est donc $G_{ext(a)}$ = (ext(a), E_a) avec

— E_a = {(2,4), (4,5), (5,6), (6,7), (6,8), (7,8)}

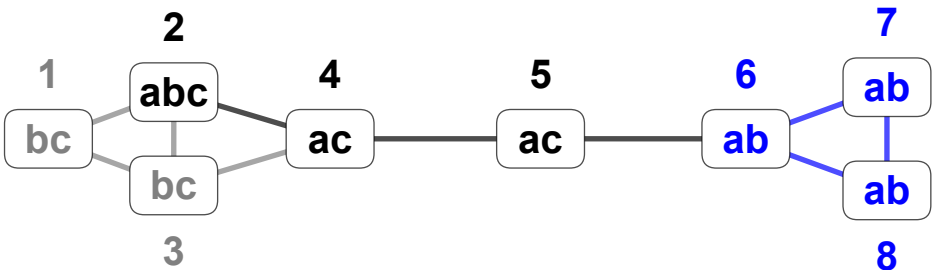


FIGURE 3.5 – Motif a et son extension dans un graphe attribué G

Inspirée par la fouille de motifs clos

Un motif clos c est défini comme le motif maximalement spécifique supporté par un ensemble d'objets X. Quand les objets décrits sont les sommets d'un graphe G, l'extension du motif clos c peut donc être associé à un sous graphe $G_X(X, E_X)$ induit de G. Dans la fouille de motifs clos abstraits, le sous-graphe G_X est réduit à un cœur de graphe G_C respectant une propriété topologique P particulière. Le motif clos abstrait c' est défini comme le motif maximalement spécifique supporté par les sommets de C. On a donc :

— $c \in L$ un motif clos,

— $ext(c) = X$, l'extension du motif clos

— $p \circ ext(c) = e$ l'extension abstraite (réduite au cœur)

— $int \circ p \circ ext(c) = c'$ le motif clos abstrait

avec

$$int(c) = \bigwedge_{v \in c} desc(v)$$

Cet opérateur est un opérateur de clôture permettant d'obtenir le motif clos abstrait en réalisant l'intersection des descriptions associées à l'extension abstraite du motif c'. À chaque motif c' et pour une propriété topologique P donnée, on définit un motif clos abstrait $int \circ p \circ ext(c) = c'$ ainsi que le cœur induit par son extension abstraite $G_{X'} = (X' = p(ext(c)), E_{X'})$. L'opérateur p permet d'obtenir un cœur de graphe et donc le motif clos abstrait de cœur avec $c' = f(c) = int \circ p \circ ext(c')$ (SOLDANO, SANTINI et BOUTHINON 2015). Si p est un opérateur intérieur monotone, alors les motifs clos abstraits adoptent une structure de treillis et il est possible de les énumérer au moyen des algorithmes standards de fouille de motifs.

Nous rappelons la définition d'opérateur intérieur : la propriété P(x) est monotone si lorsqu'elle vérifiée pour le sommet x dans un graphe $G(V, E)$ alors elle est toujours vrai pour le sommet x dans tout graphe $G(V' \supseteq V, E' \supseteq E)$.

Ainsi, il est possible d'énumérer l'ensemble des motifs clos abstraits au sein des graphes

attribués [2]. Les motifs clos s'organisent en treillis et, à la condition que p soit un opérateur intérieur, les motifs clos abstrait s'organisent également en treillis.

Dans la figure 3.5, pour obtenir le motif clos abstrait $f(a)$ en considérant le motif a, ext(a) est sélectionnée. Elle induit le sous-graphe G_a. Considérons la propriété topologique P_{k-core}. Le 2-*core* G_C du sous-graphe G_a est obtenu en appliquant P_{k-core} sur l'extension du motif a, $P_{k-core}(ext(a))$. G_C est constitué par les sommets 6, 7 et 8 et est indiqué en bleu dans la figure. Le motif clos abstrait $f(a)$ associé au cœur G_C du sous-graphe G_a est obtenu en appliquant l'opérateur d'intersection *int*, $f(a) = int(p(ext(a))) = ab$. L'extension abstraite de $f(a)$ est constituée par les sommets 6, 7 et 8. D'autres motifs clos abstraits supportés par des cœurs 2-*core* peuvent être énumérés au sein du graphe attribué G. Les motifs clos abstraits bc et c sont respectivement supportés par le cœur composé des sommets 1, 2 et 3 ainsi que par le cœur composé des sommets 1, 2, 3 et 4.

Inspirée par l'analyse formelle de concepts

Dans l'analyse formelle de concepts, un ensemble d'objets dotés de propriétés (ou décrits par un ensemble d'attributs) a une structure de treillis de concepts, aussi appelé treillis de Galois (Ganter, Stumme et Wille 2005). Un concept dans le treillis de Galois correspond à un ensemble d'objets (l'extension du concept) partageant un sous-ensemble d'attributs communs maximal (l'intension du concept) (Crampes et Plantié 2012). Dans un treillis de Galois, une hiérarchie existe entre les concepts sur lesquels un ordre partiel est appliqué. Le treillis de Galois représente donc l'ordre partiel sur les extensions ou de façon équivalente, sur les motifs clos. Les classes d'équivalence sont formées par les ensembles de concepts ayant une même extension. Dans le cadre de la fouille de motifs clos et lorsque le langage de description utilisé pour décrire les motifs est un treillis de Galois, un motif clos est unique et est le représentant de sa classe d'équivalence (Soldano, Santini et Bouthinon 2015). Le treillis de Galois représente l'ensemble des paires (q, X) où $q = int(X)$ et $X = ext(q)$.

Dans la figure 3.6, le treillis de Galois associé au graphe attribué introduit dans la figure 3.5 est représenté. Ce treillis correspond à l'ordre partiel sur les extensions et les motifs clos du graphe attribué G. Par exemple, le motif clos abc a pour extension le sommet 2. De même, le motif clos a a pour extension les sommets 2, 4, 5, 6, 7 et 8.

2. Ceci pourra être réalisé notamment avec l'outil MinerLC, logiciel *open source* d'énumération des motifs clos abstraits : `https://lipn.univ-paris13.fr/MinerLC/`

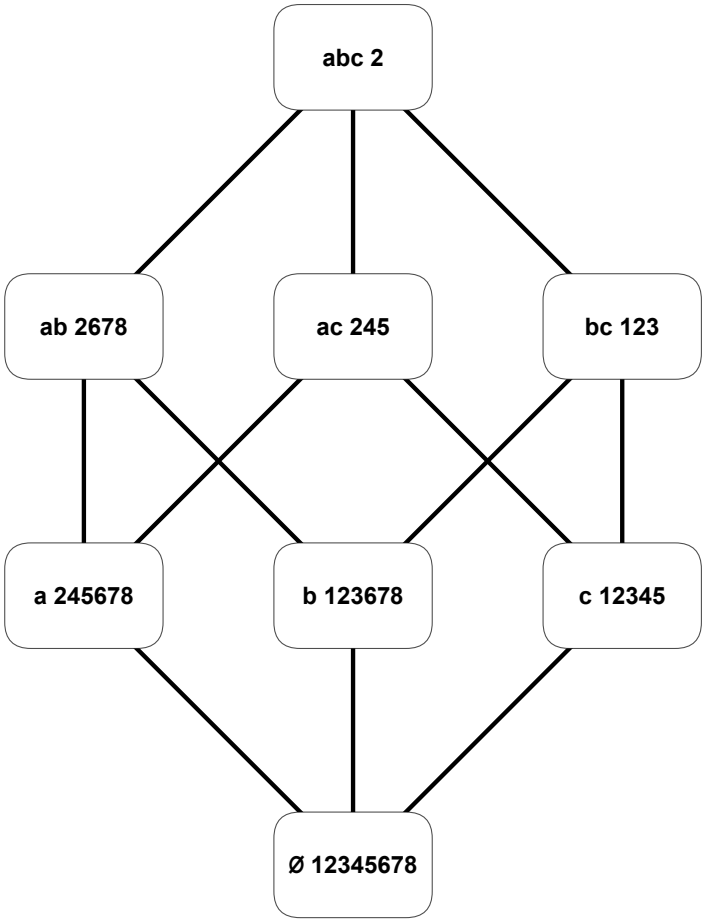


FIGURE 3.6 – Treillis de Galois associé au graphe attribué G

Quant au treillis de Galois abstrait, il représente l'ordre partiel sur les extensions abstraite ou de façon équivalente, sur les motifs clos abstraits. Les classes d'équivalence sont formées par les ensembles de concepts ayant une même extension abstraite. Un motif clos abstrait est unique et est le représentant de sa classe d'équivalence. Le treillis de Galois abstrait représente l'ensemble des paires (q, X) où $q = int(X)$ et $X = p(ext(q))$.

Dans la figure 3.7, le treillis de Galois abstrait associé au graphe attribué introduit dans la figure 3.5 est représenté. Ce treillis correspond à l'ordre partiel sur les extensions abstraites et les motifs clos abstraits du graphe attribué G. Une manière brutale mais simple d'obtenir le treillis de Galois abstrait à partir du treillis de Galois consiste à calculer pour chaque motif clos du treillis de Galois le motif clos abstrait correspondant, d'éliminer les doublons, puis d'ordonner.

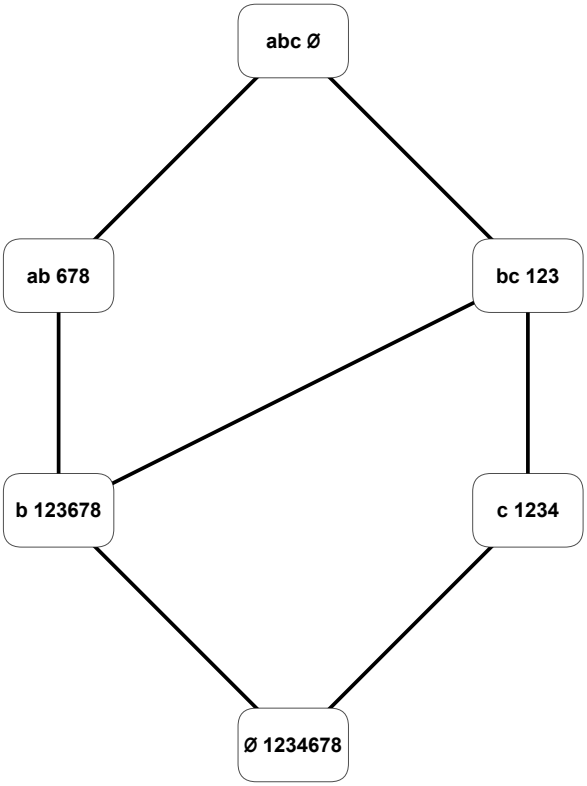


FIGURE 3.7 – Treillis de Galois abstrait associé au graphe attribué G

Pour obtenir les classes d'équivalence représentées dans le treillis de Galois abstrait de la figure 3.7 à partir du treillis de Galois de la figure 3.6, les calculs suivant ont été réalisés :

- f(abc) = int(p(2)) = int($\emptyset$) = $\emptyset$
- f(ab) = int(p(2678)) = int(678) = ab
- f(bc) = int(p(123)) = int(123) = bc
- f(ac) =int(p(245)) = int($\emptyset$) = $\emptyset$, fusion avec la classe d'équivalence dont le représentant est abc
- f(a) = int(p(245678) = int(678) = ab, fusion avec la classe d'équivalence dont le représentant est ab
- f(b) = int(p(123678) = int(123678) = b
- f(c) = int(p(12345) = int(1234) = c
- f($\emptyset$) = int(p(12345678) = int(1234678) = $\emptyset$

avec p, l'opérateur qui réduit l'ensemble de sommets aux sommets qui vérifient la propriété P_{2-core}.

On obtient donc six classes d'équivalence correspondant à des extensions abstraites distinctes (ainsi qu'à des motifs clos abstraits distincts) dans le treillis de Galois abstrait représenté dans la figure 3.7.

L'abstraction de graphe s'appuie sur la réduction de l'ensemble de noeuds support $ext(c)$ du motif au cœur du sous-graphe induit par ces noeuds que l'on appelle ensemble support abstrait (ou extension abstraite) $p(ext(c))$. Cette réduction présente un double intérêt. D'une part le nombre de motifs clos abstraits à énumérer est moindre par rapport au nombre de motifs clos puisque la réduction des ensembles support conduit à la fusion de certaines classes d'équivalence : deux motifs clos peuvent avoir la même extension abstraite et donc avoir le même motif clos abstrait associé. D'autre part, l'abstraction de graphe garantit que chaque motif clos abstrait énuméré est supporté par un sous-graphe vérifiant les propriétés topologiques du cœur. Ces propriétés expriment des contraintes de connexité entre éléments de l'extension abstraite du motif dans le cœur du graphe.

3.4 Fouille de bimotifs clos abstraits

Précédemment, nous avons présenté l'état de l'art de la fouille de motifs clos abstraits, c'est-à-dire de motifs clos au sein des cœurs de graphes. Cependant, seuls des graphes représentant un seul type de sommets ont été considérés. Un graphe biparti représente un ensemble de sommets partitionné en deux sous-ensembles de nature distincte (par exemple des individus et les documents qui leur sont associés). Dans un graphe biparti, les arêtes relient uniquement des sommets appartenant à des sous-ensembles différents. Un graphe biparti peut être orienté ou non orienté.

Chaque ensemble de sommets étant étiqueté à l'aide d'un langage de description distinct au sein d'un graphe attribué biparti, il est possible de définir des bimotifs (Soldano, Santini, Bouthinon, Bary et al. 2018). Un bimotif clos abstrait associé à un graphe biparti $G_V(V_1, V_2, E_V, I_1, I_2, L_1, L_2)$ est une paire de motifs de la forme $(f(q_1), f(q_2))$ dans lequel $f(q_1)$ et $f(q_2)$ sont les plus grands motifs partagés par C_1 et C_2 respectivement dans le cœur $G_C(C_1, C_2, V)$ du sous-graphe G_V.

La définition de cœur dans les graphes bipartis évolue également. Les bicœurs de graphes bipartis, induits par deux sous-ensembles de sommets (C_1, C_2) reposent sur une paire de propriétés topologiques (P_1, P_2). Chaque propriété topologique P_i est définie de la manière suivante : $P_i(v_i, C_1, C_2)$ avec v_i un sommet issu de C_i. Par exemple, le bicœur h-a-*hub*-autorité d'un graphe biparti orienté peut-être obtenu en considérant que l'ensemble des sommets de C_1 constituent les *hubs* tandis que l'ensemble des sommets de C_2 constituent les autorités. En d'autres termes, les sommets de C_1 ont tous un degré sortant supérieur ou égal à h vers des sommets autorités. De la même manière, les sommets de C_2 ont tous un degré entrant supérieur ou égal à a à partir de sommets *hubs*. Plus généralement, dans un bicœur $C = (C_1, C_2)$, l'ensemble des sommets $v_1 \in C_1$ doivent respecter la propriété topologique $P_1(v_1, C_1, C_2)$ et l'ensemble des sommets $v_2 \in C_2$ doivent respecter la propriété topologique $P_2(v_2, C_1, C_2)$.

Dans la figure 3.8, l'abstraction bi-h-a-*hub*-autorité est appliquée sur le graphe attribué

biparti $G_V(V_1, V_2, E_V, I_1, I_2, L_1, L_2)$ avec V_1 = {1,2,3}, V_2 = {4,5,6}, I_1 = {a,b,c,d} et I_2 = {w,x,y,z}. Le bicœur $G_C(C_1, C_2, V)$ de G_V est obtenu après abstraction de bicœur 2-2-*hub*-autorité. On a l'ensemble de sommets *hubs* C_1 = {1,2} en bleu dans la figure et l'ensemble de sommets autorités C_2 = {4,5,6} en rouge dans la figure. On obtient l'ensemble de bimotifs clos abstraits $(f(q_1), f(q_2))$ = (ab, wx).

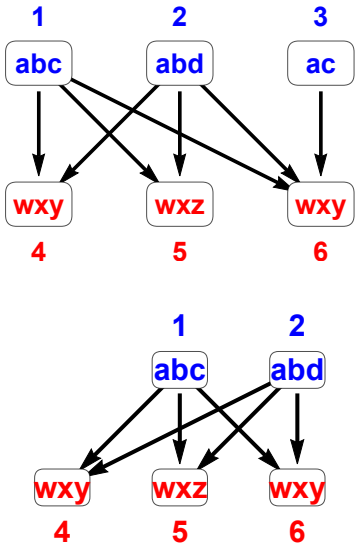


FIGURE 3.8 – Abstraction bi-h-a-*hub*-autorité : exemple du bi-2-2-*hub*-autorité

Dans l'exemple décrit dans la figure 3.8, les sous-ensembles I_1 et I_2 des langages de description L_1 et L_2 utilisés pour décrire les deux sous-ensembles de sommets V_1 et V_2 ne possèdent aucune intersection.

3.5 Fouille de bimotifs clos abstraits restreinte

Dans la section 3.4, nous avons présenté la fouille de bimotifs clos abstraits précédemment introduite dans l'état de l'art. La fouille de bimotifs clos abstraits restreinte consiste à définir un cadre dans le cas où les sous-ensembles I_1 et I_2 des langages de description L_1 et L_2 de graphes bipartis ont une intersection non vide. On a $I_1 \cap I_2 \neq \emptyset$. Dans le cas où les sous-ensembles I_1 et I_2 des langages de description L_1 et L_2 utilisés pour décrire les deux sous-ensembles de sommets dans un graphe biparti possèdent une intersection non vide, il est possible de ramener la fouille de bimotifs clos abstraits à un problème de fouille de motifs clos abstraits et donc d'unifier bimotifs clos abstraits et motifs clos abstraits. Nous considérons des bimotifs clos abstraits (q_1, q_2) dans lesquels q_1 et q_2 possèdent une intersection F. Deux cas extrêmes émergent :

- $F = \emptyset$. Dans ce cas, il s'agit d'un problème classique de fouille de bimotifs clos abstraits.

- $F = I_1 = I_2 = I$. Dans ce cas, il s'agit d'un problème classique de fouille de motifs clos abstraits.

Dans le cas où il existe un sous-ensemble non vide d'attributs communs $F \subseteq I_1 \cap I_2$ partagés par les deux ensembles de sommets V_1 et V_2 avec $I_1 \neq I_2$, nous définissons la fouille de bimotifs clos abstraits restreinte à la partie commune F de la manière suivante : pour tout bimotif (q_1, q_2) restreint à F et pour tout descripteur $i \in F$, i appartient à q_1 et q_2 à la fois ou i n'appartient ni à q_1 ni à q_2.

Pour tout bimotif (q_1, q_2) restreint à F, il existe un unique bimotif clos abstrait restreint à F. Son extension abstraite est identique à l'extension du bimotif (q_1, q_2). Nous pouvons donc définir un opérateur de clôture permettant d'obtenir le bimotif clos abstrait restreint à F.

Dans la figure 3.9, nous considérons le bicœur $G_C(C_1, C_2, V)$ de G_V obtenu après abstraction de bicœur 1-*core*. L'abstraction de bicœur 1-*core* garantit qu'il n'existe pas de sommets isolés dans le bicœur. Il existe deux sous-ensembles de sommets répartis en deux rôles : C_1 = {1,2,3} en bleu dans la figure et C_2 = {4,5,6} en rouge dans la figure. Le bimotif clos abstrait obtenu est $(f(q_1), f(q_2)) = (abx, acy)$. On a I_1 = {a,b,c,x,y} et I_2 = {a,c,w,y}. Il existe une intersection $I_1 \cap I_2$ non vide. $I_1 \cap I_2$ = {a,c,y}. Définissons une partie commune restreinte $F \subset I_1 \cap I_2$ avec F = {a,c}. La partie commune à $f(q_1), f(q_2)$ et F est : $f(q_1) \cap f(q_2) \cap F = a$. On a $f(q_1) - F = abx$ - $ac = bx$. On a $f(q_2) - F = acy$ - $ac = y$. On obtient alors le motif clos abstrait $f(q_1)$ restreint à F, $f(q_1)_F = a \cup bx = abx$ et le motif clos abstrait $f(q_2)$ restreint à F, $f(q_2)_F = a \cup y = ay$. On obtient alors le bimotif clos abstrait restreint à F : $(f(q_1)_F, f(q_2)_F) = (abx, ay)$.

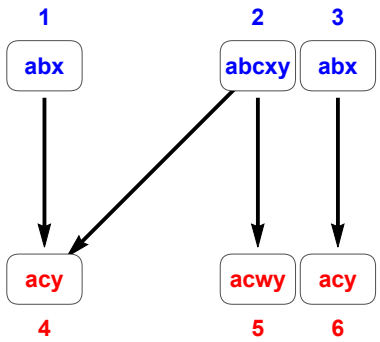


FIGURE 3.9 – Bicœur illustrant la fouille de bimotifs clos abstraits restreinte

3.6 Sélection de motifs

Une problématique récurrente de la fouille de motifs est la suivante : comment sélectionner un nombre restreint de motifs pertinents limitant les redondances ? L'algorithme $g\beta$, un schéma général de sélection d'un sous ensemble de motifs S maximal à partir d'un ensemble de motifs P en post-traitement a été proposé pour répondre à cette problématique-

tique (Soldano, Santini et Bouthinon 2019). Ce schéma permet de sélectionner des motifs en fonction des critères suivants :

- Dans S, les distances entre toutes paires de motifs doivent excéder un seuil β. On a $\forall(x, y) \in S^2$, $d(x, y) > \beta$
- S maximise la somme des mesures d'intérêt individuels de chacun de ses motifs

Ces critères supposent qu'il existe une définition de la distance entre deux motifs $d(x, y)$ ainsi qu'une définition de la mesure d'intérêt d'un motif, g.

Dans la figure 3.10, nous représentons un ensemble de trois motifs q_1, q_2 et q_3 avec $g(q_1) > g(q_2) > g(q_3)$, g étant la mesure d'intérêt d'un motif. Le premier motif sélectionné est q_1, puisqu'il s'agit du motif présentant la mesure d'intérêt la plus forte. Le second motif présentant la plus haute mesure d'intérêt est q_2. Il se situe à une distance supérieure au seuil β de l'ensemble des motifs préalablement sélectionnés. Il est donc également sélectionné. Quant au motif q_3, il se situe à une distance inférieure au seuil β d'au moins l'un des motifs préalablement sélectionnés, il est donc écarté de la sélection.

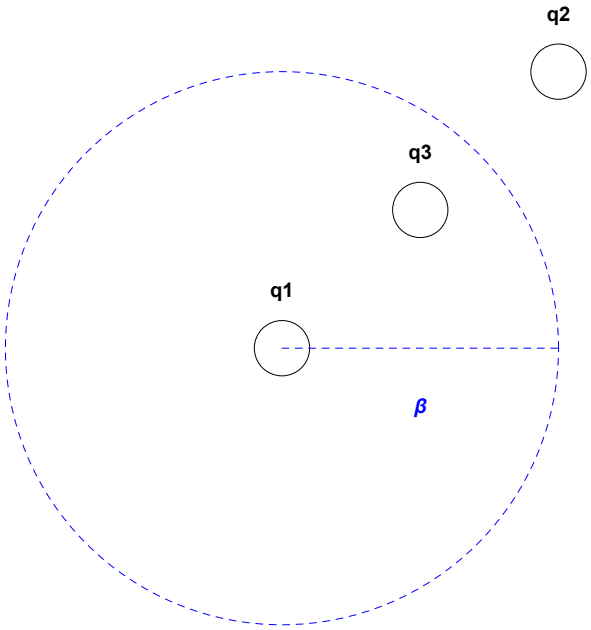


Figure 3.10 – Sélection de motifs à l'aide de l'algorithme gβ

Algorithme glouton gβ

L'algorithme $g\beta$ est un algorithme glouton garantissant la contrainte de distance entre motifs et retournant un sous-ensemble de motifs S qui constitue une solution approchée au problème de sélection. Dans le pire des cas, la complexité de g_β est $O(|P||S|)$ à la fois en nombre de comparaisons et en nombre de distances à calculer entre paires de motifs. En conséquence, $g\beta$ est très efficace lorsqu'une contrainte β forte est appliquée.

gβ consiste en :

- Une étape d'initialisation dans laquelle une liste vide S est définie et dans laquelle les motifs composant P sont triés par ordre décroissant d'intérêt g

- une étape de recherche dans laquelle chaque motif de P est soit rejeté lorsque sa distance à un motif de la liste S courante est plus petite ou égale à β, soit ajouté à S

Nous rappelons l'algorithme g_β (Soldano, Santini et Bouthinon 2019) :

Données : P, un ensemble de motifs
Résultat : S, un sous-ensemble de P pertinent et non redondant
P est ordonné par ordre décroissant de la mesure d'intérêt g;
$F \leftarrow P$;
$S \leftarrow \emptyset$;
tant que $F \neq \emptyset$ **faire**
  Trouver le premier élément $x \in F$ tel que $\forall y \in S$, $d(x, y) > \beta$;
  si *x a été trouvé* **alors**
    Supprimer x de F ainsi que ses prédecesseurs dans F;
    S $\leftarrow$ S $\cup \{x\}$;
  sinon
    $F \leftarrow \emptyset$;
  fin
fin
retourner S

Algorithme 1 : Algorithme g_β de sélection des motifs

Le sous-ensemble de motifs sélectionné S doit être maximal, c'est-à-dire qu'il n'existe pas de sous-ensemble S' de motifs plus grand que S tel que $\forall (x, y) \in S'$, $d(x, y) > \beta$. En considérant un graphe dont les sommets sont des éléments de P et dont les arêtes relient des sommets x, y tels que $d(x, y) \leq \beta$, S est un stable maximum. Un stable maximum est un sous-ensemble maximal de sommets V' d'un graphe $G(V, E)$ tel que $\forall x, y \in V'$, $(x, y) \notin E$, c'est-à-dire que les éléments du sous-ensemble de sommets V' ne sont pas voisins dans le graphe G. La version pondérée du problème du stable maximum consiste en l'identification du stable de poids maximal, en considérant une fonction de pondération associant un poids à chaque sommet. Le problème de sélection des motifs g_β est donc équivalent au problème du stable maximum en considérant la mesure d'intérêt g comme une fonction de pondération. En généralisant ce problème, il est possible de conclure que g_β est NP-complet (Brandstädt 2001).

3.6.1 Distance entre motifs

Définir une distance entre motifs est une étape cruciale pour la sélection de motifs non redondants. Ce processus consiste à éliminer les motifs proches de ceux préalablement sélectionnés. Nous proposons d'utiliser la distance de Jaccard comme mesure de distance entre les extensions abstraites de paires de motifs clos abstraits. Il s'agit initialement d'une mesure de dissimilarité entre deux ensembles X et X', entre 0 et 1. La distance de Jaccard est définie de la manière suivante :

$$dJ(X, X') = 1 - \frac{|X \cap X'|}{|X \cup X'|}$$

Ainsi, la distance entre deux motifs q et q' d'extensions abstraites respectives C et C' est définie de la manière suivante :

$$d(q, q') = dJ(C, C')$$

Dans le cas de bimotifs q et q', d'extensions abstraites respectives (C_1, C_2) et (C'_1, C'_2), nous définissons la distance entre q et q' de la manière suivante :

$$d(q, q') = max(dJ(C_1, C'_1), dJ(C_2, C'_2))$$

Ainsi, si le bimotif q est sélectionné, q' est éliminé si et seulement si la distance de Jaccard entre C_1 et C'_1 ainsi que la distance de Jaccard entre C_2 et C'_2 sont toutes deux inférieures à β.

3.6.2 Mesure d'intérêt d'un motif

Définir une mesure d'intérêt sur les motifs permet d'identifier un ensemble de motifs pertinents. Dans l'étape d'initialisation de l'algorithme de sélection des motifs g_β, les motifs sont préalablement triés par ordre décroissant d'intérêt. Ce tri préalable est essentiel et peut avoir un impact important sur le résultat de la sélection. En effet, le motif ayant la mesure d'intérêt la plus élevée est toujours conservé. Il s'agit du premier motif dans la liste de motifs triés par ordre décroissant d'intérêt. Les motifs présentant une distance inférieure au seuil β avec l'un des motifs conservés sont ensuite systématiquement éliminés de la sélection. L'étape préalable de tri des motifs par ordre décroissant d'intérêt est donc cruciale.

La mesure d'intérêt que nous proposons de considérer est la modularité (Newman et Girvan 2004). Il s'agit d'une mesure qui est élevée lorsque le nombre de liens intra-groupes dans un sous-graphe est grand et le nombre de liens inter-groupes dans ce même sous-graphe est faible (Crampes et Plantié 2012). La définition formelle de la modularité est la suivante : $\sum_i (e_{ii} - a_i^2)$ avec e_{ii} le nombre d'arêtes intra-groupe au sein de la communauté

i et $a_i = \sum_j e_{ij}$ le nombre d'arêtes inter-groupes d'une communauté j vers la communauté i. Il s'agit d'une méthode simple et répandue pour la détection de communautés au sein de réseaux sociaux. De nombreuses heuristiques ont été proposé pour approcher l'optimum, notamment des algorithmes gloutons (Noack et Rotta 2009) comme celui que nous proposons. L'avantage des algorithmes gloutons repose sur leur simplicité.

3.7 Représentation étendue des motifs

Dans la section 3.6, nous avons présenté la méthode de sélection des motifs précédemment introduite (Soldano, Santini et Bouthinon 2019). Cette méthode permet d'éliminer les motifs redondants ou non pertinents. Nous proposons une représentation étendue des motifs, permettant d'étendre leur définition à l'aide des motifs qui ont été éliminés lors de l'étape de sélection.

Reprenons l'exemple illustré par la figure 3.10, dans lequel nous avons représenté un ensemble de trois motifs q_1, q_2 et q_3 avec $g(q_1) > g(q_2) > g(q_3)$, g étant la mesure d'intérêt d'un motif. On a $\forall (x, y) \in S$, $d(x, y) > \beta$ avec S l'ensemble des motifs sélectionnés à l'aide de l'algorithme g_β, $d(x, y)$ la distance entre paires de motifs, β un seuil. Pour rappel, l'algorithme g_β maximise la somme des intérêts individuels des motifs sélectionnés. Ainsi, dans la figure 3.10, q_1 est préalablement sélectionné, puisque il est le motif ayant la mesure d'intérêt g la plus élevée. Le motif q_2 est également sélectionné, puisqu'il se trouve à une distance respective des motifs préalablement sélectionnés toujours supérieure au seuil β. En effet, le seul motif préalablement sélectionné est q_1, et $d(q_1, q_2) > \beta$. Quant au motif q_3, il est éliminé de la sélection, puisqu'il se trouve à une distance d'un motif préalablement sélectionné inférieure au seuil β.

Un motif est éliminé en raison de sa proximité avec un motif préalablement sélectionné, ceci afin d'éviter les redondances. Le motif associé à la mesure d'intérêt maximale est conservé. Cependant, les motifs éliminés peuvent comporter des éléments du langage de description utiles pour étendre la définition des motifs sélectionnés.

Soit L le langage de description des motifs, qi un motif sélectionné dont on souhaite obtenir la représentation étendue, avec $qi \subseteq L$. Soit $ext(q_i) \subseteq V$, l'extension du motif. $Q = \{q_1, q_2, ..., q_i, ..., q_n\}$ est l'ensemble des motifs énumérés, $q_\beta \subseteq Q$ l'ensemble des motifs sélectionnés pour un seuil β à l'aide de la mesure de distance entre motifs d. On a $d(q_i, q_j)$ la distance entre deux motifs définie précédemment.

On définit $N_i = \{q_j \in Q \mid d(q_j, q_i) \leq \beta\}$, l'ensemble des motifs proches du motif sélectionné q_i. À l'aide de l'ensemble des motifs proches de q_i, $q_j \in N_i$, on calcule une représentation étendue t_i du motif sélectionné q_i.

Nous proposons deux méthodes différentes pour obtenir la représentation étendue d'un motif q_i. L'une est basée sur l'intension des motifs proches de q_i, l'autre sur leur extension. Une variante est également proposée, basée sur l'extension de q_i seulement.

3.7.1 Représentation étendue d'un motif basée sur l'intension

Nous proposons une représentation étendue t_i du motif q_i obtenue à l'aide de l'analyse de l'intension de l'ensemble des motifs proches de q_i, $q_j \in N_i$. Il s'agit d'une représentation étendue d'un motif q_i obtenue à partir des fréquences d'apparition des éléments du langage des sommets dans les motifs proches de q_i.

L'extension de t_i correspond à l'union des extensions des éléments $q_j \in N_i$:

$$ext(t_i) = \bigcup_{q_j \in N_i} ext(q_j)$$

L'intension de t_i correspond à l'ensemble des vecteurs de fréquence d'apparition obtenus pour chaque élément du langage dans les motifs proches de q_i :

$$int(t_i) = \{\ (l, f)\ \mid\ l \in I, f = \frac{\mid \{q_j \mid q_j \in N_i \text{ et } l \in q_j\} \mid}{\mid N_i \mid}\ \}$$

La représentation étendue t_i du motif q_i, $rep_{int}(q_i)$ est définie par son extension $ext(t_i)$ et son intention $int(t_i)$.

3.7.2 Représentation étendue d'un motif basée sur l'extension

Nous proposons une représentation étendue t_i du motif q_i obtenue à l'aide de l'analyse de l'extension du motif q_i. Il s'agit d'une représentation étendue d'un motif q_i obtenue à partir des fréquences d'apparition des éléments du langage des sommets de l'extension du motif q_i.

L'extension de t_i correspond à l'extension de q_i :

$$ext(t_i) = ext(q_i)$$

L'intension de t_i correspond à la proportion p de sommets de $ext(q_i)$ dans lesquels apparaît l'élément l du langage :

$$int(t_i) = \{(l, p)\ \mid l \in I \text{ et } p = \frac{\mid \{v \mid v \in ext(q_i) \text{ et }\ l \in desc(v)\} \mid}{\mid ext(q_i) \mid}\}$$

La représentation étendue t_i du motif q_i, $rep_{ext1}(q_i)$, est définie par son extension $ext(t_i)$ et son intention $int(t_i)$.

Nous proposons également une représentation étendue t_i du motif q_i obtenue à l'aide de l'analyse de l'extension de l'ensemble des motifs proches de q_i, $q_j \in N_i$. Il s'agit d'une représentation étendue d'un motif q_i obtenue à partir des fréquences d'apparition des éléments du langage des extensions des sommets proches du motif q_i, N_i.

L'extension de t_i correspond à l'union des extensions des éléments $q_j \in N_i$:

$$ext(t_i) = \bigcup_{q_j \in N_i} ext(q_j)$$

L'intension de t_i correspond à la proportion p de sommets de $q_j \in N_i$ dans lesquels apparaît l'élément l du langage :

$$int(t_i) = \{ \ (l,p) \ \mid l \in I \ et \ p = \frac{\mid \{v \mid v \in \bigcup_{q_j \in N_i} ext(q_j) \text{ et } \ l \in desc(v)\} \mid}{\mid \bigcup_{q_j \in N_i} ext(q_j) \mid} \ \}$$

La représentation étendue t_i du motif q_i, $rep_{ext2}(q_i)$, est définie par son extension $ext(t_i)$ et son intention $int(t_i)$.

3.8 Conclusion

L'abstraction de graphe est une méthode inspirée de l'analyse des réseaux sociaux permettant de simplifier des structures complexes et de les réduire à des sous-ensembles densément connectés. Dans le cadre de la recherche d'experts, l'utilisation de cette méthode nous semble prometteuse. En effet, combiner la fouille de motifs au sein des graphes attribués avec la considération de contraintes de connectivité supportées par des cœurs de graphe nous paraît utile pour identifier des ensemble d'experts et leurs expertises associées au sein des graphes d'expertise. Notre hypothèse est la suivante : nous supposons qu'en se focalisant sur les zones denses de graphes d'expertise obtenus à partir d'un corpus de publications scientifiques, il serait possible de détecter des individus considérés comme experts sur un ensemble de thématiques ainsi que leurs caractéristiques communes maximales.

Dans ce chapitre, un ensemble de propriétés topologiques associées à des cœurs de graphe ont été définis. La fouille de motifs clos abstraits au sein des graphes attribués a également été introduite par les chercheurs du domaine. Il s'agit d'une extension de la fouille de motifs clos et de l'analyse formelle de concepts. Elle permet d'énumérer des motifs clos partagés par des sommets impliqués dans des cœurs de graphe. La fouille de bimotifs clos abstraits au sein de graphes attribués bipartis a également été préalablement définie dans ce chapitre. Enfin, les motifs clos abstraits étant représentés par un treillis de Galois, il existe des motifs clos abstraits très similaires dans l'ensemble des motifs clos abstraits énumérés à partir d'un graphe attribué et d'une propriété topologique associée à un cœur. Une méthode a donc été définie afin d'opérer une sélection parmi les motifs clos abstraits dont la proximité est telle qu'ils sont considérés comme redondants et non pertinents.

À la lumière de cet état de l'art, nos contributions au domaine sont les suivantes : nous introduisons la fouille de bimotifs clos abstraits restreinte à un langage F. Le langage F

est commun aux deux sous-ensembles de sommets du graphe biparti. Il s'agit donc d'un sous-ensemble compris dans l'intersection des langages d'étiquetage associés aux deux sous-ensembles de sommets du graphe biparti. Cette restriction de la fouille de bimotifs clos abstraits nous permet de ramener le problème de la fouille de bimotifs clos abstraits à un problème de fouille de motifs clos abstraits. Ainsi, nous proposons d'unifier bimotifs clos abstraits et motifs clos abstraits. De plus, nous proposons une méthode de représentation étendue d'un motif, basé sur les motifs évictés lors du processus de sélection. En effet, lors du processus de sélection des motifs, les motifs redondants sont éliminés. Cependant, ces motifs sont proches de motifs préalablement conservés, et peuvent aider à étendre leur définition si nécessaire.

Si l'abstraction de graphes attribués nous semble pertinente pour la détection d'individus considérés comme experts sur un ensemble de thématiques ainsi que leurs caractéristiques communes maximales, il reste à tester la validité de cette approche dans le cadre de la recherche d'experts à partir de publications scientifiques. Dans le chapitre 4, nous présentons notre approche. Cette dernière est basée sur les éléments introduits dans les chapitres 1, 2 et 3.

Chapitre 4

Approche proposée

À la lumière des trois premiers chapitres, nous proposons une méthode originale pour la recherche d'experts. Son originalité repose dans la prise en compte d'une validation par les pairs à l'aide d'une combinaison de méthodes de fouille de texte et d'une méthode de fouille de graphe, l'abstraction de graphe. Nous appliquons cette méthode au cas d'usage de la recherche d'experts dans le milieu académique.

Nous considérons un corpus de publications scientifiques à partir duquel nous souhaitons identifier des experts et leurs expertises associées. Des connaissances pertinentes pour la recherche d'experts sont extraites du corpus à l'aide de méthodes d'apprentissage automatique et de méthodes symboliques tirant parti d'une ontologie. Ces connaissances permettent de construire des graphes attribués pertinents pour la recherche d'experts. Les graphes attribués que nous construisons décrivent les experts (les auteurs de publications scientifiques), les expertises (les thématiques de recherche) et les documents sources d'expertise (les publications scientifiques) présents dans le corpus ainsi que les liens pertinents existant entre eux. L'abstraction de graphe permet de se focaliser sur les zones denses des graphes attribués et de découvrir des experts et leurs expertises associées à l'aide de propriétés topologiques matérialisant des contraintes de connectivité. Ces propriétés topologiques permettent de prendre en compte une validation par les pairs, matérialisée par la densité des relations de collaboration scientifique que les individus entretiennent entre eux. Plus précisément, l'énumération de motifs clos abstraits permet de découvrir des connaissances à l'aide d'une fouille de motifs clos au sein de cœurs de graphes attribués.

Dans un souci d'interopérabilité, nous nous intéressons à la représentation de l'ensemble des graphes attribués sous forme de graphe de connaissances. Les graphes de connaissances représentent un standard de la représentation de connaissances au sein du Web sémantique. Ils sont utilisés dans de nombreux domaines, notamment dans le domaine académique. Ils sont alors appelés graphes de connaissances scientifiques. Les graphes de connaissances scientifiques permettent d'organiser les connaissances obtenues à partir de la littérature scientifique. Ils sont utilisés dans de nombreuses applications, parmi lesquelles la recherche d'experts ou la recommandation de collaborateurs. Afin de favoriser l'interprétabilité et l'interopérabilité de nos résultats, nous souhaitons respecter le standard imposé par les graphes de connaissances scientifiques. Nous proposons donc de traduire les graphes attribués en un graphe de connaissances scientifique, respectant le formalisme RDF.

Dans ce chapitre, nous présentons en détail l'approche que nous proposons. Il s'agit d'une méthode hybride, combinant une méthode à base d'apprentissage et une méthode à base de graphe. Cette combinaison nous permet de prendre en compte des indicateurs d'expertise liés au contenu publié par les individus, mais également liés à leur réputation. Cette réputation est estimée à l'aide d'une validation par les pairs, matérialisée par la densité des liens de collaboration scientifique entretenus par les individus au sein de la communauté scientifique.

4.1 Une méthode hybride prenant en compte des indicateurs d'expertise liés au contenu publié ainsi qu'une validation par les pairs

Les publications scientifiques constituent la source d'expertise que nous considérons pour la recherche d'experts dans le milieu académique. Il s'agit de documents universels au sein de la communauté scientifique. Les indicateurs d'expertise qui peuvent être extraits des publications sont multiples. Pour caractériser l'expertise des membres de la communauté scientifique, nous pouvons exploiter les liens citations ou les liens de coauteurs présents dans leurs travaux. Les publications recèlent à la fois d'expertises et de liens de collaboration scientifique (liens de coauteurs, citations) et sont donc une source de connaissances appropriée pour la recherche d'experts dans le milieu académique.

L'état de l'art de la recherche d'experts, rappelé dans le chapitre 1, a permis d'établir que le nombre de citations seul ne suffit pas à indiquer l'expertise d'un individu ou la pertinence d'une publication scientifique. Il suggère de combiner un ensemble d'indicateurs variés pour déterminer l'expertise des membres de la communauté scientifique. L'état de l'art suggère également d'exploiter les deux approches principales pour la recherche d'experts simultanément, les approches basées sur les documents ainsi que les approches basées sur le profil d'expert. De même, les méthodes hybrides combinant des méthodes de fouille de texte à base d'apprentissage et des méthodes de fouille de graphe inspirées de l'analyse des réseaux sociaux permettent d'obtenir des systèmes plus performants. En effet, cette combinaison permet de capturer les expertises présentes dans les documents et de prendre simultanément en compte la réputation des individus.

Les récentes contributions dans le domaine ont permis d'aborder le problème de la recherche d'expert sous des angles nouveaux. Inspirés par ces contributions, nous proposons de ramener la recherche d'experts à un problème de fouille de graphes attribués, notamment à partir de graphes d'expertise, ce qui constitue une contribution originale. Un graphe attribué est un graphe dont les sommets sont étiquetés à l'aide d'un langage de description. Ainsi, nous proposons notamment de représenter un corpus de publications scientifiques sous forme de graphes d'expertise (dont les sommets sont des auteurs

reliés entre eux par des relations de collaboration scientifique, par exemple par des liens de citation ou de coauteurs) étiquetés par les expertises (thématiques ou domaines de publication).

De récents travaux ont démontré qu'il était pertinent d'utiliser la fouille de motifs au sein des cœurs de graphe pour énumérer des ensembles maximaux d'attributs communs au sein des cœurs de graphe ainsi que les ensembles de sommets qui les supportent (Soldano, Santini, Bouthinon et Lazega 2017). Adaptée à notre problématique de recherche d'experts, nous supposons que cette méthode permettrait d'explorer les graphes d'expertise et d'énumérer les ensembles maximaux d'expertises communes ainsi que les experts qui leur sont associés, en prenant en compte une validation par les pairs matérialisée par les contraintes de connexité. Cette méthode a précédemment été utilisée dans l'exploration de graphes de citation, à l'aide d'une propriété topologique inspirée de HITS (Soldano, Santini, Bouthinon et Lazega 2017). Notre hypothèse est la suivante :

Hypothèse

En se focalisant sur les zones denses de graphes d'expertise obtenus à partir d'un corpus de publications scientifiques, il serait possible de détecter des individus considérés comme experts sur un ensemble de thématiques ainsi que leurs caractéristiques communes maximales.

À la lumière de l'état de l'art, nous proposons donc une approche originale pour la recherche d'experts basée sur la combinaison de méthodes classiques de fouille de texte à partir de publications scientifiques, d'une représentation des connaissances extraites à partir de texte sous forme de graphes attribués et d'une méthode de fouille de graphe attribués, l'abstraction de graphe.

Plus précisément, nous utilisons une annotation sémantique des résumés des publications à l'aide d'une ontologie pour identifier les expertises présentes dans les publications scientifiques. À partir des graphes attribués, nous utilisons une approche de fouille de motifs au sein des cœurs de graphe. Le cœur d'un graphe est une zone fortement dense dans laquelle l'ensemble des sommets respecte une propriété topologique vraie au sein d'un graphe (Soldano et Santini 2014). Par exemple, la propriété topologique associée à un cœur *2-core* est la suivante : « tout sommet du *2-core* a un degré supérieur ou égal à k au sein du *2-core* ». D'autres propriétés topologiques ont été définies dans l'état de l'art. Ces dernières matérialisent une validation par les pairs, c'est-à-dire l'existence d'un réseau de collaboration scientifique dense supportant les experts potentiels au sein de la communauté scientifique.

Ainsi, l'expertise d'un individu est validée par la combinaison d'indicateurs considérant le contenu des documents (correspondant aux attributs des graphes attribués) et par

des indicateurs considérant la réputation de l'individu (correspondant aux contraintes de connectivité au sein des graphes attribués). Notre approche est illustrée et résumée par la figure 4.1. Elle est détaillée dans la suite de ce chapitre, qui se fonde sur les chapitres 1, 2 et 3.

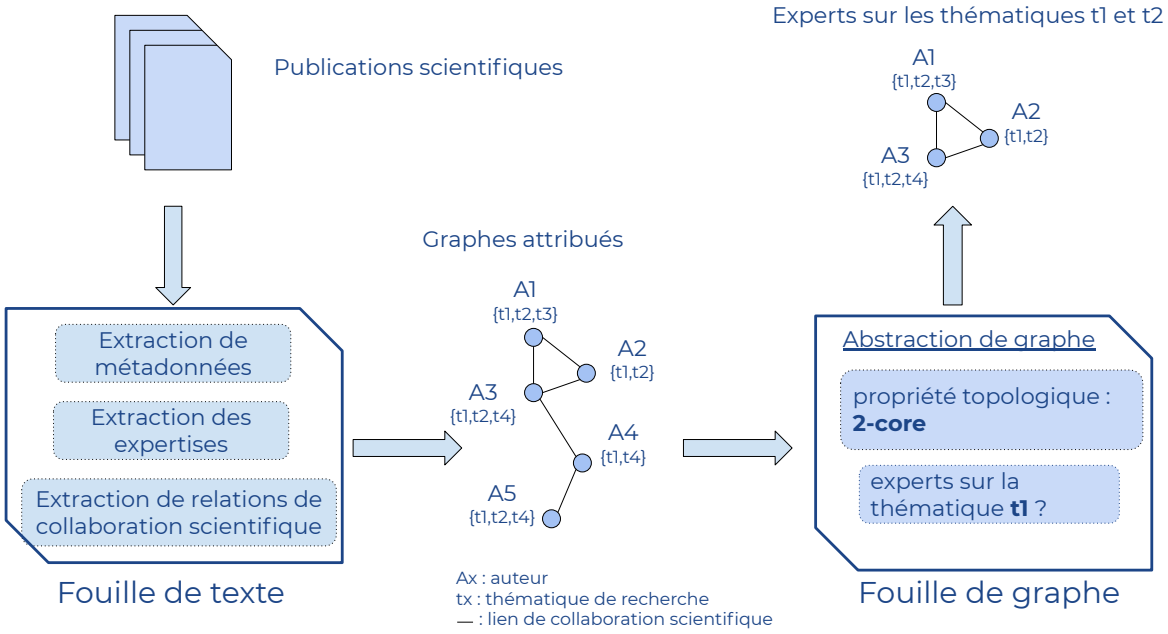


FIGURE 4.1 – Approche proposée. Dans la partie centrale, un graphe d'expertise attribué est représenté, plus précisément un graphe de coauteurs. Les auteurs ($A1, A2, A3, A4$ et $A5$) sont reliés s'ils ont au moins une publication commune. Ils sont étiquetés par les thématiques de recherche qu'ils ont abordé ($t1, t2, t3, t4$)

4.2 Recherche d'experts à partir de publications scientifiques

Notre approche permet d'identifier des experts et leurs expertises associées à partir d'un corpus de publications scientifiques, comme l'indique la figure 4.1.

Dans ce chapitre, nous proposons d'illustrer notre approche en considérant un petit corpus de publications scientifiques rédigé en français que nous avons annoté manuellement. Il s'agit du corpus de publications scientifiques issu des ateliers Recherche d'Information SEmantique (RISE)[1]. Ce corpus, de taille modeste, permet d'illustrer notre approche à l'aide d'exemples simples. Il est constitué des actes de 2009 à 2017. Ces derniers sont composés de 48 publications scientifiques (ZEVIO, Haïfa ZARGAYOUNA et al. 2018). Les thématiques officielles de ces ateliers sont la recherche d'information, le Web sémantique, l'extraction de connaissances, le traitement automatique des langues natu-

1. Ateliers Recherche d'Information SEmantique (RISE) :`https://sites.google.com/site/frenchsemanticir/`

relles et le multimédia. Le corpus est majoritairement rédigé en langue française, par 87 auteurs différents. Le nombre d'auteurs étant restreint, le corpus décrit les travaux d'une communauté scientifique de petite taille, dont on peut s'attendre à ce que ses membres partagent un grand nombre d'expertises communes. La densité des relations de collaboration scientifique entretenues au sein de cette communauté est faible.

4.3 Extraction de connaissances à partir de publications scientifiques

À partir d'un corpus de publications scientifiques, nous proposons d'extraire des métadonnées, d'identifier des expertises ainsi que d'extraire les relations de collaboration scientifique existantes, comme illustré par la partie gauche de la figure 4.1.

La figure 4.2 illustre les connaissances pertinentes que l'on peut extraire à partir de publications scientifiques. Il est possible d'extraire des métadonnées, par exemple le titre, la liste des auteurs, le résumé, la date de publication, la liste des références bibliographiques, ainsi que l'affiliation des auteurs, par exemple (Khan et al. 2017). Dans le milieu académique, les expertises correspondent à des thématiques de recherche, comme par exemple « extraction d'information » ou encore « apprentissage automatique ».

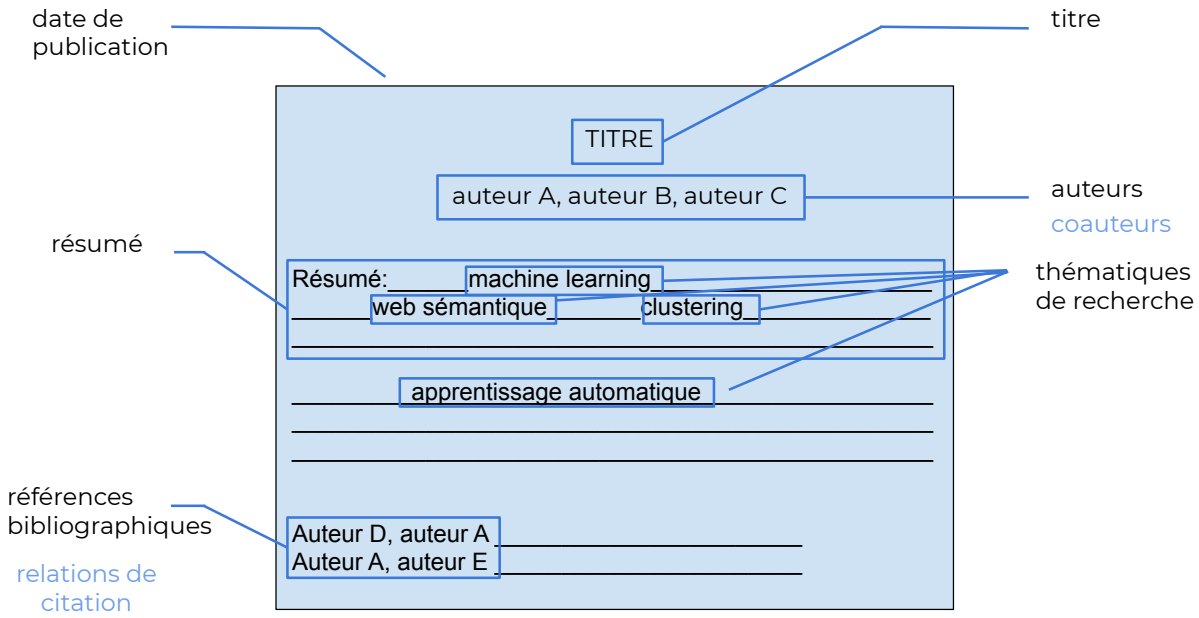


Figure 4.2 – Connaissances pertinentes dans les publications scientifiques

4.3.1 Extraction des thématiques de recherche

Les thématiques de recherche peuvent correspondre à des mots-clefs fournis par les auteurs ou être extraites automatiquement à partir du texte ou du résumé de la publication

scientifique, comme l'indique la figure 4.2.

Thématiques de recherche décrites par les auteurs

Le format des articles de RISE permet aux auteurs de fournir une liste de mots-clefs pour chaque publication. Ils correspondent à une description spécifique et condensée des thématiques de recherche abordées par les auteurs dans leur publication. La liste des mots-clefs associés à une publication se situe généralement à la suite de son résumé. Pour mieux définir notre approche, nous nous sommes posés la question suivante : les mots-clefs fournis par les auteurs sont-ils suffisants pour décrire les thématiques de recherche abordées par les publications scientifiques ?

Pour déterminer s'il est nécessaire d'employer des méthodes d'apprentissage automatique pour extraire les thématiques de recherche à partir du contenu des publications scientifiques ou si les mots-clefs fournis par les utilisateurs suffisent pour décrire les thématiques abordées par les publications scientifiques, nous proposons une expérience. Nous avons annoté manuellement les publications du corpus des ateliers Recherche d'Information SEmantique à l'aide des mots-clefs fournis par les auteurs et nous proposons d'analyser les résultats obtenus à l'aide de cette annotation. Le corpus étant de taille modeste, une annotation manuelle est possible dans un temps raisonnable.

Au sein du corpus des ateliers Recherche d'Information SEmantique, nous avons extrait 107 mots-clefs distincts fournis par les auteurs des publications. Chaque publication est décrite par 3 mots-clefs en moyenne avec une déviation standard de 2,5, un nombre minimum et maximum de 0 et 9 mots-clefs par publication respectivement. Dans la table 4.1, nous décrivons les mots-clefs les plus utilisés dans le corpus. Il s'agit de mots-clefs comportant au moins 3 occurrences, c'est-à-dire retrouvés dans au moins 3 publications différentes.

Numéro	Mot-clef	Nombre d'occurrences
1	recherche d'information	9
	ontologie	9
2	recherche d'information sémantique	7
3	annotation sémantique	4
4	XML	3
5	ressource sémantique	3
	modèle de langue	3
	indexation sémantique	3

TABLE 4.1 – Mots-clefs les plus utilisés pour décrire le corpus des ateliers Recherche d'Information SEmantique (RISE)

Parmi les mots-clefs décrits dans la table 4.1, nous pouvons constater que l'on ne trouve que des mots-clefs correspondant à des thématiques très générales. Par ailleurs,

un pourcentage non négligeable de publications ne comportent aucun mot-clef associé. En effet, 15 publications sur les 48 décrites par le corpus ne comportent aucun mot-clef associé, soit plus de 31 %. De plus le nombre de mots-clefs utilisés en moyenne reste relativement faible.

D'après les constatations réalisées sur le corpus des ateliers Recherche d'Information SEmantique, il semble nécessaire d'employer des méthodes de traitement automatique du langage naturel basées sur de l'apprentissage automatique pour identifier les thématiques de recherche à partir du contenu des publications scientifiques car de nombreuses publications scientifiques ne contiennent pas de mots-clefs, particulièrement lorsqu'aucun processus d'élicitation n'incite les auteurs à en fournir lors de la soumission. De plus, les mots-clefs, généralement peu nombreux, ne sont pas suffisamment spécifiques pour représenter l'ensemble des thématiques abordées dans la publication.

Thématiques de recherche présentes dans le texte

Pour identifier les thématiques de recherche présentes dans le texte, nous proposons d'employer des méthodes d'apprentissage supervisé combinées à des méthodes symboliques. Les méthodes symboliques sont principalement basées sur l'utilisation de ressources lexicales, terminologiques et ontologiques. Contrairement aux méthodes à base d'apprentissage, elles ne nécessitent pas de données d'entraînement préalablement annotées. Elles consistent à reconnaître des mots d'un lexique ou des concepts d'ontologies et terminologies dans un texte. À la différence de la simple reconnaissance d'un mot au sens lexical du terme, reconnaître un concept d'ontologie permet également d'accéder aux concepts plus généraux qui lui sont associés.

Nous proposons de comparer trois méthodes différentes. Nous proposons d'assimiler une thématique de recherche à un terme, à une phrase-clef ou à un concept d'ontologie. Nous avons sélectionné ces trois méthodes car elles sont répandues dans l'état de l'art. Le principe de l'identification d'une thématique reste identique quelle que soit la méthode employée : il réside dans l'identification d'un ensemble de mots importants dans le texte, correspondant à une thématique de recherche.

Identification des termes Un terme correspond à un ensemble de mots du texte revêtant un intérêt particulier suggéré par son étiquette grammaticale et son nombre d'occurrences dans le texte. Une étiquette grammaticale attache à un mot sa catégorie grammaticale (par exemple « déterminant », ou « verbe »). Les phrases nominales constituent généralement des termes pertinents. Une combinaison d'analyse morphosyntaxique et de calcul de la fréquence des mots dans le texte permet d'identifier les termes dans le texte. Nous proposons d'utiliser un algorithme d'extraction de termes populaire, TermSuite (Rocheteau et Daille 2011).

Identification des phrases-clefs Une phrase-clef se situe à un niveau d'abstraction supérieur au terme. Il s'agit d'un terme qui décrit au mieux le sujet du document. Une phrase-clef est comparable à un mot-clef. Cependant, contrairement au mot-clef, la phrase-clef n'est pas fournie par les auteurs mais doit être identifiée à partir du texte. Pour identifier des phrases-clefs à partir de termes candidats, des méthodes d'apprentissage supervisé sont généralement employées. À partir d'un ensemble de données d'entraînement et d'une méthode à base d'apprentissage, les termes candidats sont filtrés. Nous proposons d'utiliser un algorithme d'extraction de phrases-clefs (Hernandez, Buscaldi et Charnois 2017) pour identifier les phrases-clefs contenues dans les publications scientifiques.

Identification des concepts d'ontologie Une ontologie représente le consensus d'un groupe d'individus sur la spécification explicite de la conceptualisation d'un domaine particulier. Un concept d'ontologie correspond donc à un ensemble de mots spécifiques à un domaine. Considérons une ontologie de même domaine que celui abordé dans un corpus de publications scientifiques. Reconnaître les concepts de l'ontologie dans la publication semble une méthode naturellement indiquée pour en identifier les thématiques. Nous proposons d'utiliser la récente Computer Science Ontology (Salatino et al. 2018b) pour reconnaître les concepts sémantiques présents dans les résumés des publications scientifiques du domaine de l'informatique. Cette ontologie est adaptée aux textes rédigés en langue anglaise.

Dans le chapitre 5, nous proposons des expériences permettant de valider notre approche. Nous comparons la reconnaissance des expertises présentes dans le texte à l'aide de l'algorithme d'extraction de phrases-clefs (Hernandez, Buscaldi et Charnois 2017) et de l'algorithme d'extraction de termes (Cram et Daille 2016) avec la reconnaissance de concepts sémantiques isssus de la Computer Science Ontology (Salatino et al. 2018b) à partir d'un corpus de publications scientifiques rédigé en langue anglaise. Notre hypothèse, assez largement vérifiée dans l'état de l'art, est que la reconnaissance de concepts issus d'une ontologie permet d'identifier plus précisément les connaissances au sein de textes plutôt qu'à l'aide d'algorithmes d'extraction de termes ou de phrases-clefs. Nous proposons cependant de réaliser une expérience permettant de confirmer ce résultat dans la section 5.2.1 du chapitre 5.

4.3.2 Extraction des métadonnées

Les métadonnées peuvent être extraites automatiquement des publications scientifiques. De nombreux outils ont été proposés à cette fin dans l'état de l'art. Le plus performant d'entre eux, CERMINE, dispose d'une performance annoncée de 77,5 % et est de surcroît *open source* (Tkaczyk, Szostek et al. 2015). Il permet d'identifier automatiquement les métadonnées associées à une publication scientifiques telles que le titre, les

auteurs et leurs affiliations, la revue ou la conférence dans laquelle l'article a été publié, les références, les mots-clefs et le résumé. Il s'agit d'une chaîne de traitement modulaire, basée sur des méthodes d'apprentissage supervisé et non supervisé. Nous avons voulu estimer le bruit engendré par l'utilisation de CERMINE lors de la phase d'extraction des métadonnées, afin de faciliter la reproductibilité de ce travail et sa généricité. Pour cela, nous proposons une expérience permettant d'évaluer la qualité des auteurs extraits à l'aide de CERMINE sur le corpus des ateliers Recherche d'Information SEmantique. La liste des auteurs extraits automatiquement de ce corpus a été comparée à la liste construite manuellement.

Pour améliorer les résultats de la phase d'extraction des métadonnées, nous proposons une étape de filtrage permettant de réduire le bruit introduit. Pour cela, nous avons utilisé la liste des auteurs représentés par le corpus. Si un auteur extrait par CERMINE ne correspondait pas à l'un des auteurs figurant sur la liste, nous supprimions ce résultat. Après filtrage, nous avons obtenu une f-mesure de 66 %. Dans 16,66 % des cas, aucun auteur n'était retrouvé par le système. Les auteurs de CERMINE suggèrent de ré-entraîner le système afin d'obtenir de meilleurs résultats (Tkaczyk, Collins et al. 2018). Cependant, nous estimons que le bruit introduit reste élevé, considérant de surcroît une étape de filtrage qui n'est pas réalisable lorsque la liste des auteurs d'un corpus n'est pas connue au préalable. Lorsque cela est possible, nous suggérons donc d'utiliser des corpus de publications scientifiques préalablement annotés.

4.4 Des publications scientifiques aux graphes attribués

À la suite de l'étape d'extraction de connaissances à partir du corpus de publications scientifiques, nous obtenons un corpus annoté. Pour chacune des publications scientifiques du corpus, le corpus annoté fournit des informations cruciales pour la recherche d'experts (auteurs, auteurs cités, thématiques de recherche abordées). Selon l'éclairage porté sur ces informations, différentes perspectives peuvent être mises en lumière. Nous proposons de représenter ces connaissances sous forme de graphes attribués, notamment de graphes d'expertise attribués. Ces graphes représentent les connaissances extraites à partir du corpus de publications scientifique à l'aide de méthodes de fouille de texte, comme indiqué dans la partie centrale de la figure 4.1.

L'analyse des graphes de citation par exemple, permet d'identifier les auteurs les plus cités. L'analyse des graphes de coauteurs (dans lesquels il existe une relation entre deux auteurs s'ils ont au moins une publication commune) permet de détecter les conflits d'intérêt ou les experts qui collaborent le plus avec les autres membres de la communauté scientifique. Les différentes représentations sous forme de graphes permettent de struc-

turer différemment les informations disponibles. En comparant la qualité des experts et expertises associées identifiés à partir de ces différentes représentations, nous supposons qu'il est possible de déterminer les représentations les plus pertinentes pour la recherche d'experts dans le milieu académique.

Dans le chapitre 2, nous avons détaillé les graphes attribués que nous avons identifiés comme pertinents pour la recherche d'experts ainsi que les hypothèses d'expertise que nous avons suggérées à partir de ces graphes. Dans un graphe attribué, les sommets sont étiquetés à l'aide d'un langage de description (ou d'étiquetage). Par exemple, dans le cas du graphe de coauteurs, les sommets du graphe sont des auteurs étiquetés par les thématiques de recherche qu'ils abordent et par les années de publications durant lesquels ils ont publié. Ce formalisme permet de limiter la perte d'information lors de la représentation sous forme de graphes des connaissances extraites à partir de textes.

Le langage d'étiquetage des sommets des graphes attribués est composé des thématiques de recherche, des années de publication et éventuellement des auteurs. Si les sommets du graphe sont des publications, une publication est étiquetée par les thématiques de recherche qui lui sont associées, par son année de publication et par ses auteurs comme décrit dans le tableau 4.2. Si les sommets du graphe sont des auteurs, ceux-ci sont étiquetés par les thématiques de recherche de l'ensemble des publications qu'ils ont écrites ainsi que par les années durant lesquelles ils ont publié, comme décrit dans le tableau 4.2. Si les sommets du graphe sont des thématiques de recherche, elles sont étiquetées par les auteurs qui les ont abordées dans leurs publications ainsi que par les années durant lesquelles elles ont été abordées, comme décrit dans le tableau 4.2.

Nature des sommets du graphe	Langage d'étiquetage des sommets
Publications	Thématiques de recherche Années de publication Auteurs
Auteurs	Thématiques de recherche Années de publication
Thématiques de recherche	Auteurs Années de publication

TABLE 4.2 – Langage d'étiquetage des sommets dans nos graphes attribués en fonction de la nature des sommets

Nous avons identifié neuf graphes attribués pertinents pour la recherche d'experts, introduits dans le chapitre 2. Les graphes d'expertise, décrivant les auteurs, sont constitués des graphes de coauteurs et des graphes de citation entre auteurs (aussi appelés graphes de citation A). Les graphes décrivant les thématiques de recherche correspondent aux graphes de co-occurrences et aux graphes de citation entre expertises (aussi appelés graphes de citation E). Enfin, les graphes décrivant les publications scientifiques sont

composés des graphes de copublication et des graphes de citation entre documents (aussi appelés graphes de citation D). Nous avons également identifié des graphes bipartis, dont les sommets peuvent être de deux types différents. Les graphes bipartis publications-auteurs présentent des sommets qui décrivent soit des publications scientifiques, soit leurs auteurs. Les graphes bipartis auteurs → publications citées présentent des sommets qui décrivent soit les auteurs, soit les publications qu'ils citent. Les graphes bipartis publications → auteurs cités présentent des sommets qui décrivent soit les publications, soit les auteurs qu'elles citent.

Dans le cas du corpus de publications scientifiques issu des ateliers RISE, les relations de citation entre auteurs sont inconnues. Dans ce cas, il est naturellement impossible de construire le graphe de citation A. Dans la table A.1, nous présentons les graphes construits à partir du corpus RISE. Nous présentons le nombre de sommets, d'arêtes (ou d'arcs dans le cas de graphes orientés), le degré moyen des sommets et le nombre de descripteurs pour chaque graphe construit dans le cadre de cette expérience. Nous constatons que la densité des relations de collaboration scientifique entretenues au sein de la communauté scientifique représentée à partir du corpus RISE est relativement élevée pour la taille du corpus.

Graphe	Sommets	Arêtes/Arcs	Degré moyen	Descripteurs
1 - Graphe de coauteurs	87	191	4,4	110
2 - Graphe de copublication	48	87	3,6	197
3 - Graphe de co-occurrences	104	273	5,25	93
4 - Graphe biparti publications - auteurs	135	138	2	110

TABLE 4.3 – Graphes construits dans le cadre de l'expérimentation sur le corpus des ateliers Recherche d'Information SEmantique (RISE)

4.5 Identification des experts et de leurs expertises associées à partir des graphes attribués

En supposant que les individus et leurs expertises associées sont représentés sous la forme d'un graphe d'expertise attribué, se focaliser sur les individus fortement connectés dans le graphe est une piste possible pour la recherche d'experts. L'hypothèse de départ peut être la suivante : les experts fortement liés entre eux (par un réseau dense de relations de citation ou de co-publication par exemple) doivent probablement partager un ensemble d'expertises communes plus grand. Un expert n'est alors pas seulement défini par la seule énumération de ses compétences mais également par la densité des relations qu'il entretient avec les autres experts d'un réseau de collaboration scientifique. Le réseau dense de relations avec les autres experts est alors représenté par les zones denses dans le graphe.

Pour valider notre hypothèse et détecter ces réseaux d'experts, d'expertises et de documents sources, nous utilisons l'abstraction de graphe qui s'appuie sur l'identification de zones fortement denses dans un graphe. Elle introduit une étape de réduction aux sous-graphe coeurs en appliquant une contrainte topologique garantissant une forte connexité. Les experts faisant partie de ces coeurs sont plus fortement connectés à leurs pairs et donc constituent des candidats plus probables au statut d'expert.

L'abstraction de graphe est appliquée sur les graphes attribués représentant les connaissances extraites à partir de publications scientifiques, comme illustré dans la figure 4.1. Elle permet de réduire le graphe à un sous-graphe particulier appelé le cœur, respectant une contrainte topologique. Le cœur matérialise une zone dense du graphe. Identifier des cœurs permet d'identifier des experts sur des thématiques particulières au sein de la communauté scientifique.

Considérons l'exemple illustré par la figure 4.1. Dans la partie centrale de la figure, un graphe de coauteurs est représenté. Il s'agit plus précisément d'un sous-graphe induit par l'ensemble de tous les sommets présentant l'étiquette $t1$ dans un graphe plus grand. Dans cet exemple, le sous-graphe est induit par les sommets $A1$, $A2$, $A3$, $A4$ et $A5$. À titre d'exemple, une abstraction de cœur 2-*core* est appliquée sur le graphe de coauteurs décrit dans la partie centrale de la figure. On obtient un sous-graphe abstrait, le 2-*core* du graphe, composé par les sommets $A1, A2$ et $A3$ et présenté dans la partie droite de la figure. Ces sommets ont en commun un motif clos abstrait constitué par les thématiques $t1$ et $t2$. L'abstraction a permis d'apprendre que tous les sommets qui présentent $t1$ et vérifient la contrainte du 2-*core* présentent également $t2$. Si $A4$ et $A5$ sont également décrits par $t1$ et si $A5$ est également décrit par $t2$, seuls $A1$, $A2$ et $A3$ peuvent être considérés comme experts des thématiques $t1$ et $t2$ d'après notre approche, puisque validés par leurs pairs. La validation par les pairs est matérialisée par l'existence d'un réseau dense de collaboration scientifique. Dans l'exemple illustré par la figure 4.1, le graphe considéré est très petit, le cœur de graphe considéré simple et le paramètre d'abstraction faible. Sur des graphes complexes, en se focalisant sur des cœurs de graphes plus évolués à l'aide de paramètres d'abstraction plus contraints, nous supposons que les experts et expertises associées identifiés puissent être pertinents.

À notre connaissance, l'abstraction de graphe n'a jamais été utilisée dans le cadre de la recherche d'experts. Cette méthode a pourtant récemment été employée avec succès pour découvrir des k-communautés fréquentes dans un graphe (Soldano, Santini et Bouthinon 2015) dans d'autres domaines. L'appliquer à la problématique de recherche d'experts et valider l'hypothèse que nous avons formulée constitue l'objectif de ce travail et permet de développer les pistes de recherche suggérées par les plus récents états de l'art (Angelova, Boeva et Tsiporkova 2017 ; Al-Taie, Kadry et Obasa 2018), à savoir que cette approche permet de prendre en compte efficacement les relations entre pairs pour identifier les experts et leurs expertises associées.

4.5.1 Cœurs de graphe

Les abstractions de graphe pertinentes décrites dans la section 3.2 du chapitre 3 sont appliquées sur les graphes attribués construits. Pour rappel, nous utilisons des abstractions de cœurs k-*core*, k-dense et k-*nearstar* pour les graphes non orientés. Pour les graphes orientés, nous utilisons des abstractions de cœur h-a-*hub*-autorité.

Certaines combinaisons d'abstractions de cœurs appliquées sur des graphes particuliers expriment une approche naïve de la recherche d'experts, par exemple l'abstraction de degré (aussi appelée abstraction de cœur k-*core*) appliquée sur un graphe simple tel que le graphe de coauteurs. D'autres approches sont plus évoluées, notamment l'application de l'abstraction de cœur h-a-*hub*-autorité sur un graphe biparti associant aux auteurs les publications qu'ils citent.

4.5.2 Paramètres d'abstraction

Le choix de paramètres d'abstraction de graphe pertinents est fortement conditionné par la densité du graphe étudié : plus un graphe est dense, plus le paramètre d'abstraction doit être contraint pour obtenir un ensemble de motifs clos abstraits pertinent dont le nombre reste aisément interprétable. Ainsi, le paramètre choisi doit permettre d'identifier un ensemble de motifs clos abstraits suffisamment grand pour offrir une couverture pertinente des experts et de leurs expertises mais également suffisamment restreint pour que l'ensemble des résultats puisse être accessible et lisible pour un utilisateur.

Afin d'identifier un paramètre d'abstraction pertinent, les valeurs les plus contraintes sont d'abord appliquées puis progressivement relâchées jusqu'à l'obtention d'un compromis entre temps d'exécution, interprétabilité des résultats et nombre de motifs clos abstraits produits. Ce processus est exploratoire et nécessite d'appliquer plusieurs paramètres avant d'obtenir un compromis satisfaisant.

4.5.3 Motifs clos abstraits

Pour illustrer les experts et expertises associées obtenus, nous explorons le graphe de coauteurs, le graphe de copublication, le graphe de co-occurrences et le graphe biparti publications-auteurs cités que nous avons construit à partir du corpus des ateliers Recherche d'Information SEmantique. Des tableaux récapitulatifs décrivant l'ensemble des expériences réalisées sur le corpus des ateliers RISE sont disponibles dans la section A.1 de l'annexe.

La table 4.4 indique le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe de coauteurs obtenu à partir du corpus RISE. Les temps d'exécution de toutes ces expériences sont inférieurs à la seconde. Lorsque l'on relâche le paramètre d'abstraction de graphe appliqué, le nombre de motifs clos abs-

traits obtenus augmente. Ce phénomène est généralisable à toute abstraction appliquée sur n'importe quel graphe. Dans le cas des graphes construits à partir du corpus des ateliers Recherche d'Information SEmantique, le nombre de motifs clos abstraits obtenus est faible, même en relâchant la contrainte. Cela est dû à la taille restreinte des graphes et du langage de description.

Abstraction	Paramètre	Motifs
k-*core*	10	1
	8	1
	6	1
	4	7
	3	23
	2	67
k-dense	10	1
	8	1
	6	1
	4	23
	3	67
k-*nearstar*	10	3
	8	5
	6	13
	4	32
	3	56

TABLE 4.4 – Paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus RISE et nombre de motifs clos abstraits obtenus

Considérons le graphe de coauteurs obtenu à partir du corpus RISE. À l'aide de l'abstraction de cœur 2-*core*, nous obtenons 67 motifs clos abstraits. Prenons l'exemple du motif q suivant parmi ces 67 motifs : q = {ontologie}. Dans le graphe de coauteurs, l'extension de ce motif décrit un ensemble d'experts potentiels ayant comme caractéristiques communes maximales le motif q. Il s'agirait donc d'experts en ontologie selon nos hypothèses d'expertise. En effet, l'hypothèse d'expertise associée au graphe de coauteurs est la suivante : *si un individu rédige des publications scientifiques avec des membres variés de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine.* Nous obtenons un ensemble de 22 experts représentés dans la figure 4.3. Parmi l'ensemble des auteurs étiquetés par le concept sémantique « *ontologie* », seul Rami Harrathi ne fait pas partie du 2-*core* du graphe de coauteurs. En effet, il s'agit du seul auteur étiqueté par la thématique « *ontologie* » qui n'entretient un lien de coauteur qu'avec un seul autre individu au sein du corpus, en l'occurrence Sylvie Calabretto. En effet, Rami Harrathi a été le doctorant de Sylvie Calabretto. Une autre doctorante, Inès Bannour, a été identifiée comme experte en ontologie car elle appartient au 2-*core*, c'est-à-dire qu'elle a publié avec

au moins deux auteurs appartenant au 2-*core*. Nous supposons qu'identifier des experts au sein du graphe de coauteurs peut introduire du bruit. En effet, les doctorants ou jeunes chercheurs ne peuvent pas être identifiés comme des experts confirmés de leur domaine.

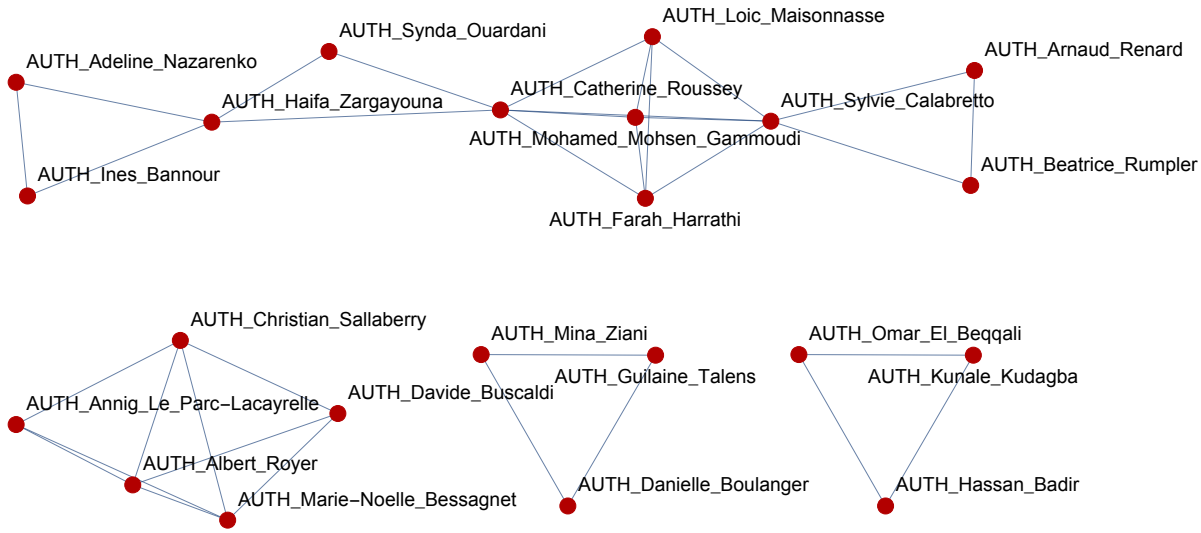
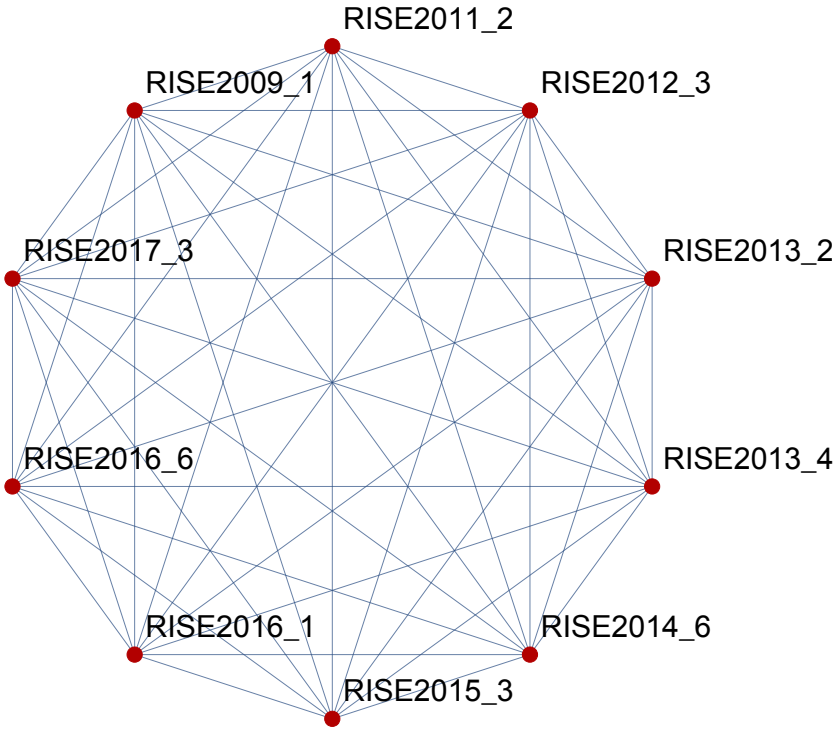


FIGURE 4.3 – Experts potentiels obtenus sur le motif {ontologie} à l'aide d'une abstraction de cœur 2-*core* sur le graphe de coauteurs construit à partir du corpus RISE

Considérons le graphe de copublication obtenu à partir du corpus RISE. À l'aide de l'abstraction de cœur 2-*core*, nous obtenons 41 motifs clos abstraits. Prenons l'exemple du motif q suivant parmi ces 41 motifs : q = {Jean-Pierre Chevallet}. Dans le graphe de copublication, l'extension de ce motif décrit un ensemble de documents ayant comme caractéristiques communes maximales le motif q. Il s'agit donc de documents ayant tous été rédigés par Jean-Pierre Chevallet. L'une des hypothèses d'expertise associées au graphe de copublication est la suivante : *si un individu rédige un grand nombre de publications scientifiques sur une thématique récurrente, alors il est probable qu'il s'agisse d'un expert de son domaine*. Sur le motif q, nous obtenons un ensemble de 10 documents représentés dans la figure 4.4. Dans cette figure, nous fournissons également un tableau associant à chaque identifiant de publication la liste de ses auteurs. Le motif ne décrit pas de thématique de recherche commune aux 10 documents. Il est intéressant de constater que dans ce cas, s'il n'est pas possible de caractériser la nature de l'expertise associée à Jean-Pierre Chevallet, il est cependant probable qu'il s'agisse néanmoins d'un acteur important de la communauté scientifique. Ses coauteurs peuvent également probablement être considérés comme des membres importants de la communauté scientifique. En effet, la seconde hypothèse d'expertise associée au graphe de copublication est la suivante : *si un membre de la communauté scientifique est coauteur d'un individu qui rédige un grand nombre de publications scientifiques sur une thématique récurrente, alors il est probable qu'il s'agisse également d'un expert de son domaine*. Cependant, les doctorants ou jeunes chercheurs peuvent entretenir des liens de coauteurs avec des membres éminents de la communauté

scientifique sans pour autant être considérés comme des experts confirmés. Néanmoins, il peut être pertinent de les identifier comme de futurs experts potentiels, ou de jeunes talents émergents.

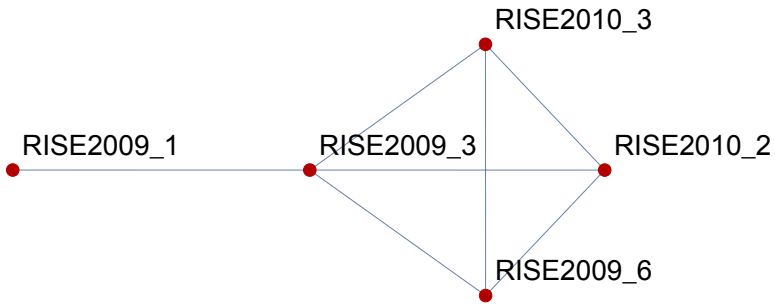


RISE2009_1	{Loic Maisonnasse, Eric Gaussier, Jean–Pierre Chevallet}
RISE2011_2	{Karam Abdulahhad, Jean–Pierre Chevallet, Catherine Berrut}
RISE2012_3	{Karam Abdulahhad, Jean–Pierre Chevallet, Catherine Berrut}
RISE2013_2	{Kian–Lam Tan, Jean–Pierre Chevallet, Philippe Mulhem}
RISE2013_4	{Mohannad Almasri, Jean–Pierre Chevallet}
RISE2014_6	{Mohannad Almasri, Kian–Lam Tan, Jean–Pierre Chevallet, Philippe Mulhem, Catherine Berrut}
RISE2015_3	{Mohannad Almasri, Jean–Pierre Chevallet, Catherine Berrut}
RISE2016_1	{Jean–Pierre Chevallet}
RISE2016_6	{Mohannad Almasri, Catherine Berrut, Jean–Pierre Chevallet}
RISE2017_3	{Jibril Frej, Jean–Pierre Chevallet, Didier Schwab}

FIGURE 4.4 – Experts potentiels obtenus sur le motif {Jean-Pierre Chevallet} à l'aide d'une abstraction de cœur 2-*core* sur le graphe de copublication construit à partir du corpus RISE

Considérons une autre abstraction de graphe appliquée sur le graphe de copublication obtenu à partir du corpus RISE. À l'aide de l'abstraction de cœur 4-*nearstar*, nous obtenons 15 motifs clos abstraits. Prenons l'exemple du motif q suivant parmi ces 15 motifs : q = {publications datant d'avant 2012}. Dans le graphe de copublication, l'extension de ce motif décrit un ensemble de documents ayant comme caractéristiques communes maximales le motif q. Il s'agit donc de documents ayant tous été publié avant 2012. Nous

obtenons un ensemble de 5 documents représentés dans la figure 4.5. Dans cette figure, nous fournissons également un tableau associant à chaque identifiant de publication la liste de ses auteurs. Comme pour le motif et son extension représentée dans la figure 4.4, ce motif ne décrit pas de thématique de recherche. Cependant, les 5 documents correspondant à l'extension du motif peuvent être considérés comme des publications phares, et leurs auteurs des experts potentiels.



RISE2009_1	{Loic Maisonnasse, Eric Gaussier, Jean–Pierre Chevallet}
RISE2009_3	{Farah Harrathi, Catherine Roussey, Sylvie Calabretto, Loic Maisonnasse, Mohamed Mohsen Gammoudi}
RISE2009_6	{Arnaud Renard, Sylvie Calabretto, Beatrice Rumpler}
RISE2010_2	{Rami Harrathi, Sylvie Calabretto}
RISE2010_3	{Samuel Gesche, Elod Egyed–Zsigmond, Sylvie Calabretto, Guy Caplat, Jean Beney}

FIGURE 4.5 – Experts potentiels obtenus sur le motif {publications datant d'avant 2012} à l'aide d'abstraction de cœur 4-*nearstar* sur le graphe de copublication construit à partir du corpus RISE

Considérons le graphe de co-occurrences obtenu à partir du corpus RISE. À l'aide de l'abstraction de cœur 2-*core*, nous obtenons 82 motifs clos abstraits. Prenons l'exemple du motif q suivant parmi ces 82 motifs : q = {publications datant d'avant 2012}. Dans le graphe de co-occurrences, l'extension de ce motif décrit un ensemble d'expertises ayant comme caractéristiques communes maximales le motif q. Il s'agit donc de thématiques de recherche ayant été abordées avant 2012. Nous obtenons un ensemble de 50 thématiques représentées dans la figure 4.6. Pour rappel, le graphe de co-occurrences ne permet pas de caractériser l'expertise d'un individu particulier mais permet d'acquérir des connaissances sur les thématiques d'un domaine de recherche et les liens existant entre elles. Pour le corpus RISE, ce graphe peut se révéler particulièrement utile puisque nous ne disposons pas de connaissances externes concernant les liens existant entre les thématiques de recherche. Dans la figure 4.6, nous identifions 3 cliques d'expertises. En prenant l'exemple de la clique de 5 expertises, nous pouvons supposer l'existence d'une similarité sémantique entre les thématiques multimedia, multilingue, masse de données, images et classification sémantique.

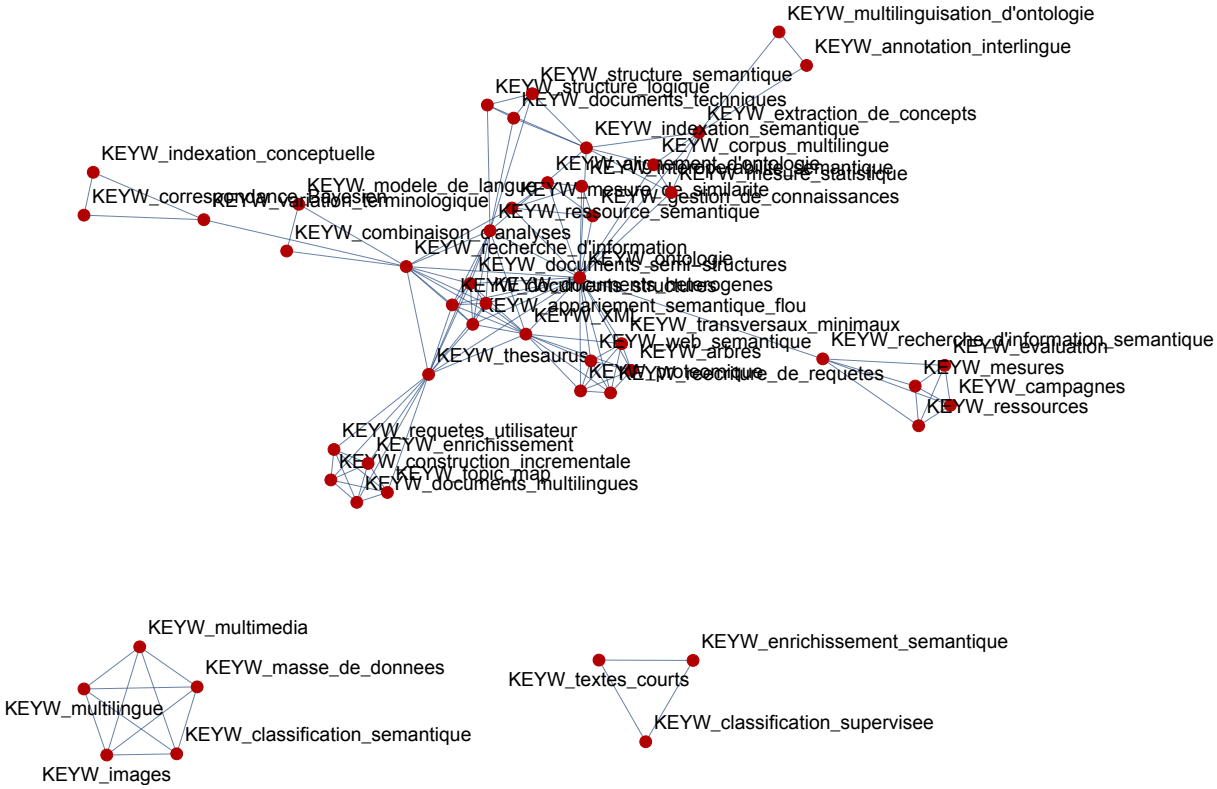


FIGURE 4.6 – Experts potentiels obtenus sur le motif q à l'aide d'abstraction de cœur 2-*core* sur le graphe de co-occurrences construit à partir du corpus RISE

Prenons l'exemple du motif q suivant : q = {Clément Jonquet, publications datant de 2012 à 2015}. Dans le graphe de co-occurrences, l'extension de ce motif décrit un ensemble d'expertises ayant comme caractéristiques communes maximales le motif q. Il s'agit donc de thématiques de recherche ayant été abordées par Clément Jonquet entre 2012 et 2015. Nous obtenons une clique de 7 expertises représentées dans la figure 4.7. À partir du graphe de co-occurrences, il est possible d'acquérir des connaissances concernant les liens existant entre les différentes thématiques de recherche. Les expertises abordées par un chercheur durant une période de temps possèdent probablement un lien de similarité sémantique. Par exemple, les expertises abordées par Clément Jonquet et représentées dans la figure 4.7 entretiennent probablement des liens de similarité sémantique.

Considérons le graphe biparti publications-auteurs obtenu à partir du corpus RISE. À l'aide de l'abstraction de cœur 3-*nearstar*, nous obtenons 66 bimotifs clos abstraits. Prenons l'exemple du bimotif q suivant parmi ces 66 motifs, avec q = $\{q_1, q_2\}$ et $q_1 = q_2 =$ {*recherche d'information*, publication avant 2014}. Le motif q_1 décrit les caractéristiques communes maximales partagées par les publications, q_2 décrit les caractéristiques communes maximales partagées par les auteurs. Dans le graphe biparti publications-auteurs, l'extension de ce bimotif décrit un ensemble d'experts et de documents ayant comme caractéristiques communes maximales le bimotif q. Il s'agit donc d'auteurs et de documents avec lesquels ils entretiennent un lien de paternité, les documents ayant été publiés avant

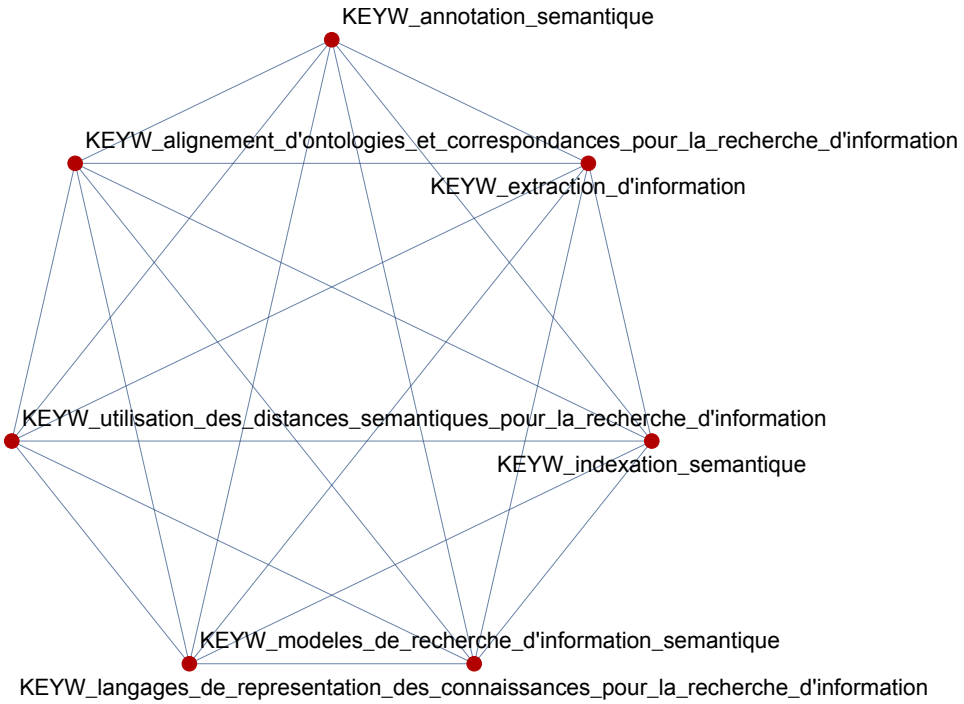


FIGURE 4.7 – Expertises obtenues sur le motif q à l'aide d'abstraction de cœur 2-*core* sur le graphe de co-occurrences construit à partir du corpus RISE

2014 sur la thématique de recherche d'information. Nous obtenons un ensemble de 12 experts et de 4 documents sources d'expertise représentés dans la figure 4.8.

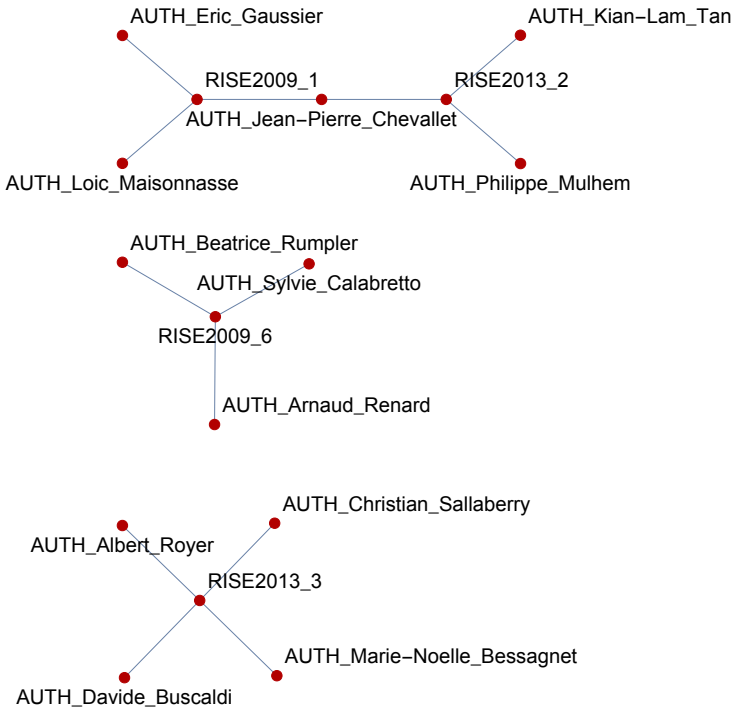


FIGURE 4.8 – Experts potentiels obtenus sur le motif q à l'aide d'abstraction de cœur sur le graphe biparti publications-auteurs construit à partir du corpus RISE

4.6 Construction d'un graphe de connaissances scientifique

Afin de favoriser l'interprétabilité et l'interopérabilité de notre méthode de recherche d'experts, nous proposons de générer un graphe de connaissances scientifiques à partir de la représentation sous forme de graphes attribués des connaissances extraites à partir du corpus de publications scientifiques étudié.

Nous avons identifié un ensemble de graphes pertinents pour la représentation de connaissances dans le domaine de la recherche d'experts dans la section 2.2.2. Isolément, ces graphes peuvent être considérés comme des graphes de connaissances s'ils en respectent le formalisme, puisqu'ils représentent des entités du monde réel formellement interprétables ainsi que les relations existant entre elles. Fusionnés, ils peuvent permettre de véhiculer des connaissances plus exhaustives sur le milieu académique et générer un graphe de connaissances scientifique pertinent pour la recherche d'experts.

Un graphe de connaissances scientifique respecte le formalisme RDF. Les graphes attribués sont sérialisés au format NRI, un format d'entrée compatible avec le logiciel MinerLC [2]. MinerLC permet d'énumérer les motifs clos abstraits d'un graphe attribué, c'est-à-dire appliquer des abstractions de graphe. Pour fusionner les graphes attribués en un graphe de connaissances ainsi que pour transformer un graphe de connaissances existant en graphes attribués pertinents pour la recherche d'experts, nous proposons un parseur NRI $\leftrightarrow$ RDF [3].

Un formalisme RDF adapté à la recherche d'experts a été suggéré dans l'état de l'art. Nous proposons une simplification de ce modèle ainsi qu'une réutilisation du schéma de Microsoft Academic Graph (Sinha et al. 2015) ainsi que certaines propriétés de Academia Industry Dynamics Knowledge Graph (Angioni et al. 2020) [4] (AIDA). En effet, AIDA fournit un schéma permettant de décrire les concepts issus de la *Computer Science Ontology* identifiés dans les publications scientifiques. De plus, AIDA réutilise des propriétés du schéma RDF de *Microsoft Academic Graph*, un graphe de connaissances largement utilisé dans l'état de l'art et décrivant plus de huit milliards de triplets représentant le milieu académique. Pour plus de détails nous renvoyons le lecteur au schéma RDF de *Microsoft Academic Graph* [5].

Les propriétés qui nous intéressent dans le Microsoft Academic Graph sont les suivantes :

2. MinerLC : `https://lipn.univ-paris13.fr/MinerLC/`
3. Parseur NRI $\leftrightarrow$ RDF réalisé par Rémi Duret, étudiant stagiaire en $2^{ème}$ année de DUT Informatique : `https://github.com/zevio/RDF_extraction_connaissances`
4. Academia Industry Dynamics Knowledge Graph : `http://aida.kmi.open.ac.uk`
5. Schéma RDF de Microsoft Academic Graph : `http://ma-graph.org/schema-linked-dataset-descriptions/`

Paper dcterms:title xsd:string (4.1)

Paper dcterms:abstract xsd:string (4.2)

Paper prism:publicationDate xsd:date (4.3)

Paper fabio:hasDiscipline FieldOfStudy (4.4)

Paper dcterms:creator Author (4.5)

Paper cito:cites Paper (4.6)

Author foaf:name xsd:string (4.7)

FieldOfStudy foaf:name xsd:string (4.8)

avec Paper, FieldOfStudy et Author des classes RDF correspondant aux publications scientifiques, expertises et auteurs. Nous ne considérons pas les classes Conference, Journal et Affiliation, bien que l'affiliation d'un auteur ou la conférence ou la revue dans laquelle un document est publié puissent constituer des indicateurs d'expertise pertinents. En effet, les informations concernant l'affiliation d'un auteur ne sont pas toujours disponible.

Nous suggérons d'utiliser une ontologie pour identifier les thématiques de recherche abordées dans les publications scientifiques. Les expertises correspondent donc à des concepts sémantiques. Dans le domaine de l'informatique, la *Computer Science Ontology* a récemment été publiée. Le schéma RDF de AIDA (Academia Industry Dynamics Knowledge Graph) propose une propriété permettant d'associer à une publication scientifique les concepts sémantiques issus de la *Computer Science Ontology* :

Paper aidaschema:hasTopic csos:Topic (4.9)

Dans l'état de l'art, un modèle de représentation des connaissances au format RDF a été proposé pour le cas d'usage de la recherche d'experts (Silvello et al. 2017). Parmi les propriétés RDF du schéma adapté au cas d'usage de la recherche d'experts (Silvello et al. 2017), les propriétés suivantes permettent de caractériser l'expertise associée à un individu :

Link ims:relation is-expert-in (4.10)

Link ims:has-source UserY (4.11)

Link ims:has-target ExpertiseA (4.12)

avec Link un nœud vide, UserY un auteur de publication scientifique, ExpertiseA une

thématique.

Ces propriétés tirent parti d'un nœud vide. Nous proposons une simplification permettant d'établir un lien direct entre individu et expertise :

$$\textit{AuthorA scho:is-expert-in ExpertiseA} \tag{4.13}$$

$$\tag{4.14}$$

Nous définissons l'espace de nom *scho*, « *ScholarMap* », le nom de notre système de recherche d'experts. Cette nouvelle propriété RDF nous permet d'enrichir le graphe de connaissances scientifique à l'aide de connaissances concernant l'expertise des individus, découvertes à l'aide de l'abstraction de graphe.

4.7 Synthèse de l'approche

Notre approche permet de découvrir et enrichir des connaissances à partir de textes pour la recherche d'experts. Elle consiste en quatre étapes : une étape de fouille de texte, une étape de représentation des connaissances sous forme de graphes attribués, une étape de fouille de graphe et une étape de génération d'un graphe de connaissances scientifique. Pour une meilleure interprétabilité et interopérabilité de nos résultats, nous fusionnons les graphes attribués considérés en un graphe de connaissances scientifique et enrichissons ce graphe des connaissances concernant les experts et leurs expertises associées découvertes durant la phase d'abstraction des graphes attribués. L'originalité de notre méthode réside dans la considération d'une validation des experts par leurs pairs, matérialisée par les propriétés topologiques considérées lors de la phase d'abstraction de graphe.

Dans la suite de cette thèse, nous validons notre proposition scientifique. Dans le chapitre 5, nous expérimentons notre méthode sur le jeu de données ACL Anthology, détaillons le protocole expérimental et le choix du paramétrage de la méthode. En effet, notre méthode étant exploratoire, elle nécessite une phase de paramétrage qu'il est nécessaire d'expliciter. Puis, dans le chapitre 6, nous évaluons les résultats obtenus sur ce même jeu de données à l'aide d'un protocole d'évaluation original.

Deuxième partie

Validation de la proposition scientifique

Chapitre 5

Expériences

Afin de tester la validité de notre approche présentée durant le chapitre 4, nous proposons de réaliser différentes expériences. Dans ce chapitre, nous présentons le jeu de données sur lequel nous avons expérimenté notre approche, décrivons le protocole expérimental et illustrons les résultats obtenus à l'aide d'exemples que nous analysons. Le protocole expérimental que nous proposons consiste en différentes étapes : une étape d'extraction des expertises à partir des résumés des publications scientifiques, une étape de construction de tables intermédiaires, une représentation des connaissances extraites sous forme de graphes attribués pertinents pour la recherche d'experts ainsi qu'une énumération des motifs clos abstraits afin d'identifier des experts et leurs expertises associées. Dans ce chapitre, nous présentons également les graphes que nous avons construits à partir du jeu de données, les paramètres d'abstraction de graphe que nous avons testés sur ces graphes ainsi que le nombre de motifs clos abstraits que nous avons obtenus à l'aide de ces paramètres. Nous présentons également le paramétrage de la sélection des motifs clos abstraits.

5.1 Jeu de données

Afin d'évaluer la validité de notre approche dans le milieu académique, nous considérons un jeu de données constitué par un corpus de publications scientifiques. Nous avons sélectionné le jeu de données ACL Anthology puisqu'il s'agit d'un corpus de publications scientifiques utilisé dans la plateforme d'évaluation LT ExpertFinder. En effet, la comparaison des résultats obtenus avec notre méthode par rapport aux résultats obtenus à l'aide des *baselines* disponibles dans LT ExpertFinder est particulièrement pertinente. Nous disposions d'une version du corpus ACL annotée (Gábor, Haïfa Zargayouna et al. 2016). Cette version est antérieure à celle hébergée sur LT ExpertFinder (Radev et al. 2013). Cela signifie que certaines des publications décrites dans la version du corpus ACL Anthology hébergée par LT ExpertFinder ne sont pas représentées dans notre version du corpus.

L'échantillon du corpus ACL Anthology (Gábor, Haïfa Zargayouna et al. 2016) [1] considéré est constitué de 13322 publications scientifiques, publiées entre 1985 et 2008 sur

1. Corpus ACL Anthology (échantillon annoté) : `https://github.com/zevio/ACL`

les thématiques de la linguistique informatique et du traitement de la langue naturel. Le corpus ACL Anthology original (Bird et al. 2008) est constitué de 48104 publications. L'échantillon considéré couvre donc environ 28 % du corpus original.

Pour chaque publication, nous disposons de 13 descripteurs :

1. Un identifiant
2. La liste des auteurs
3. Le titre
4. La revue scientifique dont la publication est issue
5. L'année de publication
6. La description de la publication sous forme de référence
7. Le résumé
8. Les références de la publication sous forme de chaîne de caractères
9. Les auteurs cités
10. Les titres des publications citées
11. Les revues scientifiques dont sont issues les publications citées
12. Les années de publication des publications citées
13. Les termes extraits automatiquement à partir du résumé à l'aide de TermSuite (Cram et Daille 2016)

5.2 Protocole expérimental

Le protocole expérimental que nous proposons est générique et s'adapte à tout type de corpus de publications scientifiques préalablement annoté. Pour chaque publication, nous devons connaître au préalable la liste des auteurs *a minima*, puisqu'ils constituent les experts potentiels. Dans ce protocole expérimental, nous décrivons l'étape d'extraction des expertises à partir des résumés des publications scientifiques, la construction de tables de connaissances, la construction de graphes attribués pertinents pour la recherche d'experts, l'abstraction appliquée sur ces graphes et enfin l'identification d'experts et de leurs expertises associées.

Dans les expériences proposées tout au long de ce chapitre, nous nous intéressons à la pertinence des experts découverts à la suite des énumérations de motifs clos abstraits réalisées sur le corpus de publications scientifiques. Dans un souci d'éviction des biais lors de l'évaluation des performances de notre système, nous avons choisi de réaliser nos expériences sur un corpus préalablement annoté. En effet, nous ne souhaitions pas introduire de biais dans l'évaluation des performances de notre système liés à la qualité des métadonnées extraites. L'évaluation du bruit engendré par l'utilisation d'un outil d'extraction des métadonnées a été discuté dans le chapitre 4.

5.2.1 Extraction des expertises

Nous avons extrait automatiquement les termes, phrases-clefs et concepts issus de la *Computer Science Ontology* à partir des résumés des publications scientifiques du corpus ACL Anthology. Nous comparons les résultats obtenues à partir de l'annotation par les termes, phrases-clefs et concepts extraits afin d'identifier la méthode la plus appropriée pour l'identification automatique des thématiques de recherche à partir de publications scientifiques. Pour réaliser cette comparaison, nous nous basons sur l'analyse des 25 termes, phrases-clefs et concepts les plus utilisés pour représenter le corpus.

Pour identifier les thématiques d'une publication scientifique, nous pouvons exploiter le résumé de la publication ou le texte entier. Dans nos expériences, nous avons considéré les résumés plutôt que le texte entier. La raison principale est la suivante : nous ne disposions pas du texte entier dans le corpus annoté que nous considérions. Considérer le résumé plutôt que le texte entier comporte des avantages. Le résumé constitue généralement un échantillon représentatif et détaillé du contenu des publications et présente le meilleur ratio de phrases-clefs par total de mots (Shah et al. 2003). Si le texte entier contient également des phrases-clefs pertinentes, toutes les sections d'une publication scientifique ne présentent pas un ratio identique de phrases-clefs par total de mots. De surcroît, le stockage des textes entiers des publications ainsi que le traitement nécessaire à l'identification de phrases-clefs pertinentes dans le texte entier peuvent augmenter significativement le temps de calcul (Shah et al. 2003). Il est également à noter que certaines phrases-clefs employées dans le texte peuvent s'appliquer à l'état de l'art et non pas aux réelles contributions apportées par les auteurs, donc véhiculer du bruit. Cependant, lorsque cela est possible, nous suggérons d'utiliser les textes entiers pour identifier les expertises à partir de publications scientifiques. En effet, des travaux sur le domaine de la pharmacogénomique ont notamment démontré que le texte entier véhicule des informations cruciales pour l'identification de concepts du domaine à partir de publications scientifiques (Garten et Altman 2009). Les auteurs ajoutent que la reconnaissance de concepts à partir du résumé n'est généralement pas suffisante pour identifier de tels concepts.

Reconnaissance de concepts issus de la Computer Science Ontology

Nous disposons des titres et des résumés des articles du corpus ACL Anthology. Nous avons utilisé l'ontologie *Computer Science Ontology* pour extraire les concepts correspondant à des thématiques de recherche dans les résumés. Cette ontologie est adaptée aux textes rédigés en langue anglaise et traite du même domaine que celui du corpus, c'est-à-dire l'informatique. Le corpus aborde plus précisément le domaine de la linguistique informatique et du traitement du langage naturel.

Nous avons réutilisé un outil de reconnaissance de concepts d'ontologie à partir de

texte[2]. Le titre, le résumé et un ensemble de termes associés sont nécessaires pour reconnaître les concepts enfouis dans les publications scientifiques à l'aide de l'ontologie. Les concepts sont reconnus à l'aide de deux méthodes. L'outil que nous réutilisons dispose d'un module de reconnaissance syntaxique ainsi que d'un module de reconnaissance sémantique (Salatino et al. 2018a). Le module de reconnaissance syntaxique identifie les concepts issus de l'ontologie dont les étiquettes sont explicitement reconnues dans le texte. Quant au module de reconnaissance sémantique, il infère la présence de concepts issus de l'ontologie à l'aide d'un étiquetage morphosyntaxique des termes et de plongements lexicaux. La sortie de ces deux modules se nomme *« union »*. La sortie *« enhanced »* ajoute aux concepts identifiés dans le texte leurs généralisations dans la hiérarchie de l'ontologie.

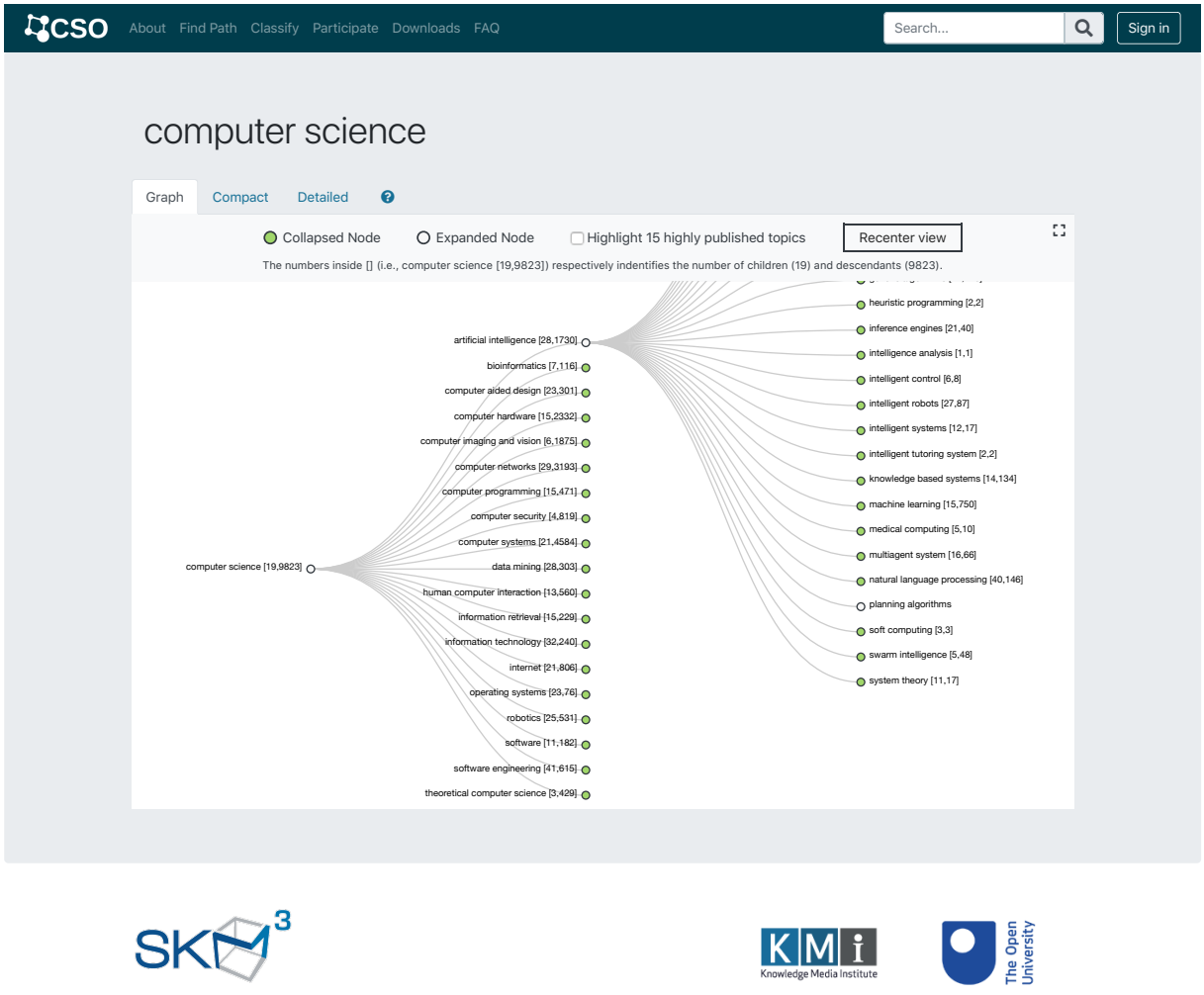


Figure 5.1 – Exemple de concept issu de la Computer Science Ontology

La figure 5.1 décrit l'insertion du concept « *machine learning* » dans la hiérarchie de concept de la Computer Science Ontology. Il a un parent direct, « *artificial intelligence* ». Les parents directs d'un concept dans la hiérarchie correspondent à ses premiers

2. CSO Classifier : `https://github.com/angelosalatino/cso-classifier`

concepts les plus généraux. Le concept « *artificial intelligence* » a lui-même un parent, « *computer science* ». L'ensemble des concepts les plus généraux obtenu à partir de *« machine learning »* est donc composé des concepts « *artificial intelligence* » et « *computer science* ».

```
        {
    "syntactic": [
        "natural language interfaces",
        "data management",
        "data-base management systems",
        "database management"
    ],
    "semantic": [
        "languages",
        "data management",
        "database management",
        "database",
        "data-base management systems"
    ],
    "union": [
        "natural language interfaces",
        "database management",
        "database",
        "data management",
        "languages",
        "data-base management systems"
    ],
    "enhanced": [
        "natural languages",
        "management information systems",
        "database systems",
        "information management",
        "object oriented programming",
        "computer programming languages",
        "query languages",
        "linguistics",
        "information storage and retrieval"
    ]
}
```

FIGURE 5.2 – Annotation du résumé d'une publication scientifique du corpus ACL Anthology à l'aide de la Computer Science Ontology

Un exemple d'annotation du résumé d'une publication dans ACL Anthology (ayant pour identifiant A83-1002, pour titre « Problems In Natural-Language Interface To DSMS With Examples From EUFID » et pour auteurs Marjorie Templeton et John D. Burger) est donné dans la figure 5.2. En examinant les premiers concepts généraux récupérés à l'aide de la sortie *« enhanced »* pour un petit échantillon de publications scientifiques issues

du corpus ACL, nous avons remarqué que les concepts étaient parfois trop généralistes et n'apportaient pas beaucoup de sémantique supplémentaire. Par exemple, dans l'exemple donné dans la figure 5.2, les concepts *« computer programming languages »* et *« object oriented programming »* sont trop généraux et ne nous semblent pas pertinents pour décrire les publications scientifiques dans le cadre de la recherche d'experts. Nous avons donc choisi de ne pas récupérer les premiers concepts les plus généraux et de récupérer la sortie *« union »*.

Numéro	concept	Nombre d'occurrences
1	languages	5161
2	semantics	574
3	syntactics	2431
4	syntactic structure	2261
5	natural languages	2113
6	linguistics	2072
7	parsing	1629
8	part of speech	1597
9	learning	1482
10	correlation analysis	1471
11	natural language processing	1426
12	machine translations	1201
13	semantic information	1054
14	dialogue	884
15	component	822
16	speech signals	811
17	word sense disambiguation	765
18	natural language understanding	729
19	computational linguistics	729
20	database	620
21	named entity recognition	619
22	speech recognition	604
23	user information	583
24	automatic speech recognition	546
25	information extraction	537

TABLE 5.1 – Les 25 concepts issus de la *Computer Science Ontology* les plus utilisés pour décrire le corpus ACL Anthology

Au sein de l'échantillon du corpus ACL Anthology, nous avons extrait 2714 concepts distincts. Chaque publication est décrite par 6,8 concepts en moyenne avec une déviation standard de 3,63, un nombre minimum et maximum de 0 et 55 concepts par publication respectivement. Dans la table 5.1, nous décrivons les 25 concepts les plus utilisés dans le corpus ACL Anthology. Une occurrence de concept correspond à l'utilisation du concept pour décrire une publication. Les multiples occurrences au sein d'une même publication

ne sont pas comptabilisées. Parmi les concepts décrits dans la table 5.1 se trouvent des concepts assez généraux (par exemple *« languages »*, *« learning »*, *« component »*) et d'autres plus spécifiques (par exemple *« natural language processing », « word sense disambiguation » ou « information extraction »*).

Extraction des termes

Nous avons également extrait automatiquement les termes présents dans les résumés des publications du corpus ACL Anthology à l'aide d'un algorithme d'extraction de termes, TermSuite (Rocheteau et Daille 2011). TermSuite est un outil d'extraction terminologique et d'alignment multilingue de termes. Il fonctionne selon le principe suivant, découpé en quatre phases distinctes :

- Une phase de traitements préliminaires tels que l'identification et la conversion des encodages de caractères et la détection de la langue
- Une phase d'analyse linguistique consistant au découpage du texte en mots, à une analyse morphosyntaxique et à une lemmatisation
- Une phase d'extraction terminologique monolingue consistant en la détection d'occurrences de termes simples et complexes, en la normalisation et au regroupement des termes en fonction de leurs variations et au filtrage statistique
- Une phase d'alignement terminologique bilingue consistant en un alignement contextuel par paires de langues

Au sein de l'échantillon du corpus ACL Anthology, nous avons extrait 6766 termes distincts à l'aide de TermSuite. Chaque publication est décrite par 20,2 termes en moyenne avec une déviation standard de 15, un nombre minimum et maximum de 0 et 130 termes par publication respectivement. Dans la table 5.2, nous décrivons les 25 termes les plus utilisés dans le corpus. Une occurrence de terme correspond à l'utilisation du terme pour décrire une publication. Les multiples occurrences au sein d'une même publication ne sont pas comptabilisées. Parmi les termes décrits dans la table 5.2 ne se trouvent que des termes vides, n'apportant aucune sémantique supplémentaire (par exemple « *paper* », « *based* » ou « *result* ») ou des termes généraux (par exemple *« parsing »* ou *« semantic »*).

Numéro	Terme	Nombre d'occurrences
1	paper	5681
2	based	3225
3	system	2916
4	results	2549
5	approach	2252
6	text	1945
7	method	1906
8	data	1899
9	information	1798
10	corpus	1724
11	language	1629
12	model	1585
13	task	1490
14	english	1473
15	performance	1454
16	word	1413
17	systems	1392
18	words	1380
19	automatic	1262
20	semantic	1260
21	problem	1230
22	evaluation	1206
23	parsing	1203
24	algorithm	1200
25	methods	1138

TABLE 5.2 – Les 25 termes les plus utilisés pour décrire le corpus ACL Anthology

Extraction des phrases-clefs

Enfin, nous avons également extrait les phrases-clefs présentes dans les résumés des publications du corpus ACL Anthology à l'aide d'un algorithme d'extraction de phrases-clefs (HERNANDEZ, BUSCALDI et CHARNOIS 2017). Cet algorithme se base sur un modèle CRF (*Conditional Random Fields)*. Le modèle CRF permet d'étiqueter de potentielles phrases-clefs candidates puis de les filtrer par leur étiquetage morphosyntaxique.

Nous avons identifié 23879 phrases-clefs distinctes au sein du corpus ACL Anthology. Chaque publication est décrite par 3,3 phrases-clefs en moyenne avec une déviation standard de 2,9, un nombre minimum et maximum de 0 et 40 phrases-clefs par publication respectivement. Dans la figure 5.3, nous décrivons les 25 phrases-clefs les plus utilisées dans le corpus. Une occurrence de phrase-clef correspond à l'utilisation d'une phrase-clef pour décrire une publication. Les multiples occurrences au sein d'une même publication ne sont pas comptabilisées. Parmi les phrases-clefs décrites dans la table 5.3 se trouvent des thématiques de recherche pertinentes (par exemple *« natural language processing »* ou

« word sense disambiguation »). Parfois, certaines thématiques sont disponibles sous une orthographe différente (par exemple *« natural language processing »* et *« Natural Language Processing »*). Aucun regroupement sémantique n'est employé dans l'algorithme d'extraction de phrases-clefs que nous avons considéré. De plus, nous retrouvons un grand nombre d'abréviations (par exemple *« NLP »* ou *« MT »*). Les abréviations ne sont pas pertinentes pour la recherche d'experts car une abréviation est soumise à l'ambiguïté (par exemple, dans le milieu médical, IVG peut signifier « insuffisance ventriculaire gauche » ou « interruption volontaire de grossesse »).

Numéro	Phrases-clef	Nombre d'occurrences
1	natural language processing	239
2	training data	238
3	machine translation	214
4	NLP	192
5	MT	186
6	natural language	169
7	information extraction	103
8	statistical machine translation	99
9	WSD	94
10	POS	93
11	NER	89
12	word sense disambiguation	76
13	SMT	76
14	knowledge base	69
15	classification	69
16	Natural Language Processing	64
17	HPSG	63
18	SVMs	62
19	HMM	62
20	IR	60
21	IE	60
22	QA	57
23	NE	56
24	machine learning	55
25	SVM	51

TABLE 5.3 – Les 25 phrases-clefs les plus utilisées pour décrire le corpus ACL Anthology

Comparaison des méthodes d'extraction de thématiques de recherche dans les résumés des publications

Nous avons employé trois méthodes d'extraction automatique de thématiques de recherche dans les résumés des publications scientifiques à lors de nos expériences réalisées sur le corpus ACL Anthology. Nous réutilisons un outil de reconnaissance de concepts issus

d'une ontologie, un algorithme d'extraction de termes ainsi qu'un algorithme d'extraction de phrases-clefs. Sur le corpus ACL Anthology, nous avons obtenu 2714 concepts, 6766 termes, 5078 termes racinisés et 23879 phrases-clefs uniques. Ces thématiques potentielles décrivent les résumés de 13322 publications. Nous sélectionnons les 25 concepts, termes et phrases-clefs les plus utilisés dans le corpus et comparons ces thématiques potentielles pour identifier la méthode d'identification des thématiques la plus appropriée. Nous utilisons également des exemples de thématiques potentielles peu représentées dans le corpus pour déterminer la méthode la plus appropriée.

Si la reconnaissance de concepts d'ontologie nous permet d'obtenir le plus faible nombre de thématiques de recherche différentes, elle nous permet cependant d'obtenir des thématiques pertinentes et suffisamment spécifiques au domaine considéré. Parmi les concepts peu représentés dans le corpus que nous avons pu analyser, nous avons trouvé des concepts qui ne correspondent pas à une thématique pertinente, comme « university » par exemple. Cependant, les concepts d'ontologie correspondent généralement à des thématiques de recherche pertinentes. Parfois, certains des concepts peuvent correspondent à des thématiques un peu trop générales, comme « languages », « parsing » ou « learning » par exemple. Les concepts multitermes tels que « syntactic structure » ou « natural language processing » semblent plus spécifiques.

Quant à la méthode d'extraction de termes à l'aide de TermSuite, elle ne peut pas être considérée comme une méthode d'extraction de thématiques de recherche valide. En effet, en considérant les termes les plus représentés dans le corpus ACL Anthology, cette méthode ne nous permet pas d'obtenir des thématiques pertinentes car les termes ne sont pas suffisamment spécifiques. Par exemple, « paper », le terme le plus représenté dans le corpus, ne constitue pas une thématique de recherche. Parmi les 25 termes les plus représentés dans le corpus, seuls « language », « semantic » ou « parsing » peuvent éventuellement être considérés comme des thématiques pertinentes, bien qu'encore un peu trop générales. Parmi les termes peu représentés dans le corpus sont également retrouvées des formes singuliers et pluriels (par exemple, « *categorial grammar* » et « *categorial grammars* ») ou des formes non structurées telles que « *annota-tor* » ou « *analysis-* » qui doivent être nettoyées à l'aide d'un post-traitement.

Si l'algorithme d'extraction de phrases-clefs permet d'obtenir le plus grand nombre de thématiques potentielles différentes, il ne s'agit pas de la méthode permettant d'obtenir les thématiques les plus fiables. Parmi les 25 phrases-clefs les plus employées dans le corpus, nous obtenons de nombreuses abréviations ainsi que des doublons sous une orthographe différente. Parmi les phrases-clefs peu représentées dans le corpus, nous obtenons par exemple « => » ou « *∼10% error reduction* » qui ne sont pas pertinentes en tant que thématiques de recherche potentielles. De plus, nous obtenons également des formes singuliers et pluriels tels que « *3D scene* » et « *3D scenes* », nécessitant un post-traitement. Nous avons donc sélectionné la méthode de reconnaissance de concepts dans les résumés

des publications comme méthode automatique d'extraction des thématiques. Cette méthode nous permet d'obtenir une meilleure proportion de thématiques valides, c'est-à-dire de thématiques suffisamment spécifiques au domaine considéré lorsque l'on considère les thématiques les plus représentées dans le corpus.

Cependant, lorsqu'aucune ontologie du domaine n'est disponible, l'algorithme d'extraction de phrases-clefs peut permettre d'obtenir des thématiques relativement satisfaisantes pour peu qu'un regroupement sémantique soit employé et que le modèle soit éventuellement ré-entraîné. En effet, à l'instar de l'algorithme d'extraction de termes, l'algorithme d'extraction de phrases-clefs ne nécessite pas de connaissance préalable du domaine. Cependant, contrairement à l'algorithme d'extraction de termes, un effort d'entraînement doit néanmoins être réalisé.

5.2.2 Représentation des années de publication sous forme d'intervalles

Pour chaque publication du corpus ACL Anthology annoté, nous disposons d'une métadonnée correspondant à l'année de publication. Il est naturel de considérer que deux articles ayant été publié durant la même année partagent une caractéristique commune. Par extension, nous pouvons considérer que deux publications partagent également une caractéristique commune lorsqu'elles ont été publié durant une même période temporelle, en considérant une période temporelle plus étendue. Selon le jeu de données considéré et les thématiques abordées, cette période peut être plus ou moins longue. Par exemple, il peut être pertinent de considérer des périodes temporelles correspondant à des décennies entières ou à quelques années seulement selon le domaine de recherche et la communauté scientifique associée. En effet, certaines communautés scientifiques sont plus prolixes que d'autres.

Nous proposons de représenter les dates de publication décrites dans un corpus de publications scientifiques sous forme d'intervalles d'années. Pour que la définition des intervalles d'années ne soit pas arbitraire, nous proposons d'utiliser des fractiles pour diviser les publications en intervalles d'années de publication contenant un nombre de publications identique.

Concernant le corpus ACL Anthology, nous avons choisi d'utiliser des quintiles, c'est-à-dire de créer cinq intervalles d'années de publication. Le choix du fractile est fortement lié à la taille du corpus. Si le corpus décrit peu de publications, créer un grand nombre d'intervalles n'est pas un choix judicieux.

Dans la figure 5.3, nous présentons la répartition du nombre de publications par intervalle d'années dans le corpus ACL Anthology. À l'aide des quantiles, nous avons déterminé les intervalles suivants : avant 1993, entre 1993 et 1999, entre 2000 et 2004, entre 2005 et 2007 ainsi qu'après 2007. Ces intervalles nous permettent d'obtenir des effectifs les plus

homogènes possibles.

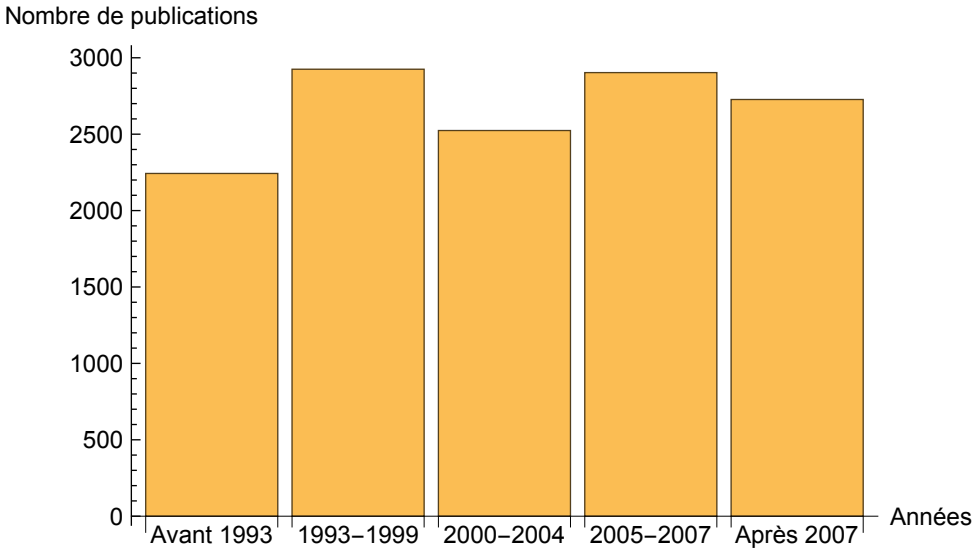


FIGURE 5.3 – Répartition du nombre de publications par intervalle d'années dans le corpus ACL Anthology

Pour décrire les publications du corpus ACL Anthology à l'aide de leur date de publication, nous proposons les descripteurs suivants : {<=1992, >1992, <=1999, >1999, <=2004, >2004, <=2007, >2007}. Par exemple, une publication $p1$ datant de 1997 sera associée aux descripteurs suivants : {>1992, <=1999, <=2004, <=2007}. Une publication $p2$ datant de 2003 sera associée aux descripteurs suivants : {>1992, >1999, <=2004, <=2007}. Les publications $p1$ et $p2$ auront les descripteurs suivants en commun : {>1992, <=2004, <=2007}.

De la même manière, les auteurs peuvent être étiquetés par les descripteurs correspondant aux années durant lesquelles ils ont publié. Par exemple, considérons l'auteur a_1 ayant rédigé les publications p_1 et p_2 précédemment définies. L'auteur a_1 sera associé aux descripteurs suivants : {>1992, <=1999, >1999, <=2004, <=2007}. De même, les thématiques de recherche peuvent être étiquetées par les descripteurs correspondant aux années durant lesquelles elles ont été abordées.

5.2.3 Construction des graphes attribués

Les graphes attribués identifiés lors du chapitre 2 sont ensuite construits à partir des tables de connaissances intermédiaires. Pour réaliser nos expériences sur les corpus annotés, nous utilisons l'environnement de développement Wolfram Mathematica [3]. Nous avons développé des *notebook* nous permettant de décrire chacune de nos expériences en langage naturel, d'exécuter le code et d'en afficher les visualisations. Ces notebooks sont disponibles en ligne [4]. Selon le corpus considéré et les connaissances disponibles, tout ou

3. Wolfram Mathematica : `http://www.wolfram.com/mathematica/`
4. ScholarMap : `https://depot.lipn.univ-paris13.fr/zevio/scholarmap`

une partie des graphes sont construits.

Dans le chapitre 2, nous avons identifié 9 graphes pertinents pour la recherche d'experts que nous rappelons dans la table 5.4. Ces graphes sont les graphes de coauteurs, de citation A, de copublication, de citation D, de co-occurrences, de citation E, bipartis publications-auteurs, bipartis auteurs → publications citées et bipartis publications → auteurs cités.

Graphe	Orienté ?	Sommets	Arêtes/Arcs	Descripteurs
1 - Graphe de coauteurs	Non	Auteurs	Coauteur	Thématiques Années
2 - Graphe de citation A	Oui	Auteurs	Citation	Thématiques Années
3 - Graphe de copublication	Non	Publications	Auteur commun	Auteurs Thématiques Années
4 - Graphe de citation D	Oui	Publications	Citation	Auteurs Thématiques Années
5 - Graphe de co-occurrences	Non	Thématiques	Co-occurrence	Auteurs Années
6 - Graphe de citation E	Oui	Thématiques	Citation	Auteurs Années
7 - Graphe biparti publications - auteurs	Non	Publications	Auteurs	Auteurs Thématiques Années
		Auteurs		Thématiques Années
8 - Graphe biparti auteurs → publications citées	Oui	Publications	Citation	Auteurs Thématiques Années
		Auteurs		Thématiques Années
9 - Graphe biparti publications → auteurs cités	Oui	Publications		Auteurs Thématiques Années
		Auteurs		Thématiques Années

TABLE 5.4 – Graphes pertinents pour la recherche d'experts

Pour rappel, les graphes de coauteurs et de citation A représentent les auteurs décrits par leurs thématiques de recherche et les intervalles d'années durant lesquelles ils ont publié. Le graphe de coauteurs représente les liens de coauteurs entre auteurs. Il n'est donc pas orienté. Le graphe de citation A représente les liens de citation entre auteurs. Il est donc orienté. Quant aux graphes de copublication et de citation D, ils représentent les publications décrites par les auteurs qui les ont rédigées, les thématiques qu'elles abordent

et les intervalles d'années correspondant à leur date de publication. Le graphe de copublication lie deux publications si elles ont un auteur commun. Il n'est donc pas orienté. Le graphe de citation D représente les liens de citation entre publications. Il est donc orienté. Concernant les graphes de co-occurrences et de citation E, ils représentent les thématiques de recherche décrites par les auteurs qui les abordent et les intervalles d'années durant lesquelles elles ont été abordées. Le graphe de co-occurrences lie deux thématiques si elles apparaissent dans une même publication. Il n'est donc pas orienté. Le graphe de citation E représente les liens existant entre deux thématiques e_1 et e_2 lorsque e_1 apparaît dans une publication p_1 citant une publication p_2 dans laquelle la thématique t_2 est abordée. Le graphe de citation E est donc orienté. Enfin, les graphes bipartis représentent à la fois les publications et les auteurs. Les publications sont décrites par les auteurs qui les ont rédigées, les thématiques qu'elles abordent et les intervalles d'années correspondant à leur date de publication. Les auteurs sont décrits par les thématiques qu'ils abordent et les intervalles d'années durant lesquelles ils ont publié. Le graphe biparti publication-auteurs représente les publications et leurs auteurs. Il n'est pas orienté. Le graphe biparti auteurs $\rightarrow$ publications citées représente les auteurs et les publications qu'ils citent. Le graphe biparti publications $\rightarrow$ auteurs cités représente les publications et les auteurs qu'elles citent. Ces deux derniers graphes sont orientés.

Selon les connaissances extraites disponibles, les graphes précédents sont construits. Par exemple, dans le cas d'un corpus de publications scientifiques annoté uniquement par une liste d'auteurs, il est impossible de construire le graphe de citation A, décrivant les liens de citation existant entre les auteurs. Dans la table A.6, nous présentons les graphes construits dans le cadre des expériences sur le corpus ACL Anthology. Nous présentons le nombre de sommets, d'arêtes (ou d'arcs dans le cas de graphes orientés), le degré moyen des sommets et le nombre de descripteurs pour chaque graphe construit dans le cadre des expériences.

Le degré moyen des sommets des graphes représentant les thématiques de recherche du corpus ACL Anthology est très élevé. Dans le graphe de co-occurrences, le degré moyen est de 76,4 et dans le graphe de citation E, le degré moyen est de 223,9. Il s'agit donc de deux graphes extrêmement denses. Cela signifie que les thématiques de recherche abordées dans le corpus ACL Anthology sont sémantiquement très proches.

5.2.4 Énumération des motifs clos abstraits

Afin d'identifier des experts et leurs expertises associées, nous nous focalisons sur les cœurs de graphes pertinents pour la recherche d'experts à l'aide d'abstractions de graphe. Cette méthode est décrite dans le chapitre 3. Nous présentons les abstractions de cœurs que nous avons utilisées, les paramètres d'abstraction testés, le nombre de motifs clos abstraits obtenus en fonction des paramètres ainsi que le nombre de motifs clos abstraits

Graphe	Sommets	Arêtes/Arcs	Degré moyen	Descripteurs
1 - Graphe de coauteurs	10724	30698	5,7	2722
2 - Graphe de citation A	10724	195277	36,4	2722
3 - Graphe de copublication	13322	151904	22,8	13446
4 - Graphe de citation D	13322	54949	8,2	13446
5 - Graphe de co-occurrences	2714	103578	76,4	10732
6 - Graphe de citation E	2714	303541	223,9	10732
7 - Graphe biparti publications - auteurs	24046	32412	2,7	2722
8 - Graphe biparti auteurs → publications citées	24046	103950	8,6	2722
9 - Graphe biparti publications → auteurs cités	24046	111982	9,3	2722

TABLE 5.5 – Graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

sélectionnés. Des tableaux décrivant l'ensemble de ces résultats sont fournis en annexe, dans la section A.2.

Abstraction

Les abstractions de graphe pertinentes décrites dans la section 3.2 du chapitre 3 sont appliquées sur les graphes attribués construits. Pour rappel, nous utilisons des abstractions de cœurs k-*core*, k-dense et k-*nearstar* pour les graphes non orientés. Pour les graphes orientés, nous utilisons des abstractions de cœur h-a-*hub*-autorité. Certaines abstractions de cœurs sont basées sur des approches naïves, par exemple l'abstraction de degré (aussi appelée abstraction de cœur k-*core*). D'autres sont plus évoluées, notamment l'abstraction de cœur h-a-*hub*-autorité. Pour plus de détails sur les abstractions de graphe, nous renvoyons le lecteur au chapitre 3.

Paramètres d'abstraction

Le choix des paramètres d'abstraction de graphe dépend fortement de la densité du graphe étudié : plus un graphe est dense, plus le paramètre d'abstraction doit être élevé, ceci pour augmenter la contrainte. L'objectif du paramétrage est d'obtenir un ensemble de motifs clos abstraits suffisamment grand pour permettre d'identifier des experts pertinents sur des ensembles d'expertises variés mais également suffisamment restreint pour être toujours aisément interprétable.

Des paramètres d'abstraction les plus grands possible sont d'abord appliqués, ce qui augmente la contrainte relative à l'abstraction de graphe. Puis, cette contrainte est progressivement relâchée par le biais de paramètres d'abstraction plus faibles, jusqu'à l'obtention d'un compromis entre temps d'exécution, interprétabilité des résultats et nombre

de motifs clos abstraits produits. Ce processus est exploratoire et nécessite d'appliquer plusieurs paramètres avant d'obtenir un bon compromis. Pour chacune des expériences réalisées, nous précisons le temps d'exécution en secondes de chacune des expériences. La machine sur laquelle nous avons réalisé nos expériences a pour processeur un modèle Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz et dispose de 32 threads (2 CPU 8 cœurs avec hyperthreading).

Le temps d'exécution est une information fournie directement dans les fichiers de résultats de MinerLC. Si le temps d'exécution est égal à 0, cela signifie que le processus s'est exécuté en moins d'une seconde.

Nombre de motifs clos abstraits

Dans un souci de lisibilité pour le lecteur, nous ne détaillons pas l'ensemble des résultats obtenus dans la suite de ce chapitre mais les présentons de manière synthétique. Pour plus de détails, nous renvoyons le lecteur à l'annexe, où des tableaux récapitulatifs décrivant l'ensemble des expériences réalisées et des résultats obtenus sont disponibles dans la section A.2.

Après application d'une abstraction de graphe, plusieurs cas de figure se présentent. À l'aide de certains paramètres, le temps d'exécution est quasiment instantané mais le nombre de motifs clos abstraits obtenus est faible. Cela signifie que le paramètre d'abstraction est trop contraint et qu'il faut le relâcher.

D'autres paramètres impliquent une explosion du nombre de motifs clos obtenus. Parallèlement, le temps d'exécution augmente. Si le temps d'exécution est trop élevé (supérieur à quelques heures d'exécution), nous considérons que le compromis entre nombre de motifs clos abstraits obtenus et temps d'exécution n'est pas optimal.

Parfois, le temps d'exécution est si élevé que nous ne sommes pas en mesure de laisser le processus s'achever. Dans ce cas, le paramètre d'abstraction testé n'est pas considéré puisque notre méthode ne permet pas de passer à l'échelle. Il n'est donc évidemment pas non plus possible de relâcher davantage le paramètre d'abstraction de graphe.

Idéalement, nous considérons que les meilleurs paramètres d'abstraction de graphe sont ceux qui nous permettent d'obtenir un nombre maximal de motifs clos abstraits dans un temps d'exécution raisonnable (que l'on estime à quelques heures au maximum sur la machine sur laquelle nous avons lancé les calculs). Cependant, le nombre de motifs clos abstraits obtenus n'est pas un indicateur suffisant de la pertinence du paramètre d'abstraction de graphe. Les motifs clos abstraits obtenus et leurs extensions doivent nous permettre d'identifier des experts pertinents et leurs expertises associées. Dans le chapitre 6, nous évaluons les résultats obtenus sur le corpus ACL Anthology.

Dans la suite de ce chapitre, nous détaillons les résultats obtenus sur un échantillon de graphe afin d'illustrer nos conclusions. Nous nous focalisons sur un graphe simple, le graphe de coauteurs ainsi que deux graphes élaborés, les graphes bipartis auteurs →

publications citées et publications → auteurs cités. Les profils des courbes de variation du temps d'exécution ou du nombre de clos abstrait en fonction de la contrainte que l'on peut observer lors de ces trois expériences sont similaires à ceux observés durant les autres expériences.

Le graphe de coauteurs étant un graphe non orienté, des abstractions de graphe simples ont été appliquées : les abstractions de cœurs k-*core*, k-dense et k-*nearstar*. Dans la table 5.6, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe de coauteurs obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution. Ce tableau est également disponible en annexe, dans la section A.2.

Abstraction	Paramètre	Motifs	Temps
k-*core*	15	13	0
	12	74	4
	10	544	26
	8	2981	141
	6	17875	640
k-dense	15	20	1
	12	221	13
	10	1263	64
	8	6833	230
	6	47535	924
k-*nearstar*	50	43	46
	45	90	67
	40	232	120
	30	1365	402
	20	14530	2258

TABLE 5.6 – Paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus et temps d'exécution (en secondes)

Dans la figure 5.4, nous présentons le nombre de motifs clos abstraits obtenus à l'aide des différents paramètres d'abstraction appliqués sur le graphe de coauteurs sous forme de graphique.

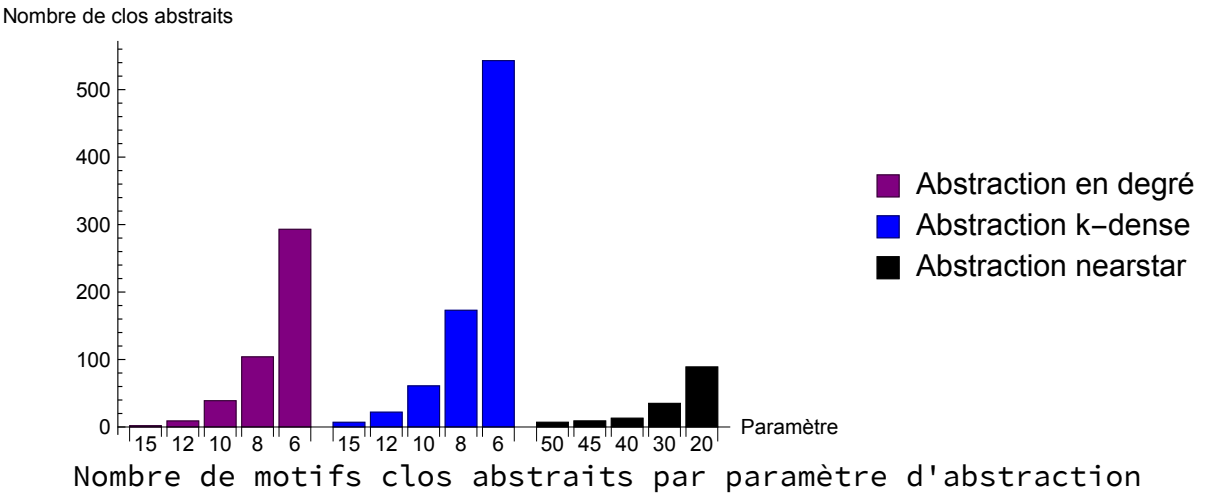


FIGURE 5.4 – Nombre de motifs clos abstraits obtenus à l'aide des différents paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Pour l'abstraction de cœur k-*core*, nous avons d'abord paramétré k par 15, puis nous avons progressivement relâché la contrainte en faisant varier k de 12 à 6 par pas de 2. En relâchant la contrainte, nous obtenons un nombre plus élevé de motifs clos abstraits. Parallèlement, le temps nécessaire pour que l'expérience se termine croît. Pour le paramètre 15-*core*, le temps d'exécution est instantané (inférieur à une seconde). Parallèlement, le nombre de motifs clos abstraits obtenus est faible, puisque nous n'obtenons que 13 motifs clos abstraits. Pour le paramètre le plus relâché, le 6-*core*, nous obtenons 17875 motifs clos abstraits en un peu plus de 10 minutes. Ce paramètre permet d'obtenir un bon compromis entre le nombre de motifs clos abstraits obtenus et le temps d'exécution. Cependant, lorsque l'on relâche le paramètre d'abstraction, nous relâchons également la contrainte de validation par les pairs. Dans le chapitre 5, nous évaluons la qualité des experts obtenus et de leurs expertises associées à l'aide des différents paramètres d'abstraction de graphe.

Sur le graphe de coauteurs, le temps d'exécution maximal des expériences que nous avons réalisé est d'environ 38 minutes. Ce temps d'exécution est obtenu à l'aide de l'abstraction 20-*nearstar*. À l'aide de cette abstraction, nous obtenons 14530 motifs clos abstraits. Cette abstraction étant un peu plus complexe que l'abstraction en degré, pour un nombre de motifs clos abstraits obtenus similaire, le temps d'exécution est plus important. Selon la complexité de l'abstraction de cœur employée, le temps d'exécution est également impacté.

Concernant les graphes bipartis, plus élaborés, une abstraction de graphe plus complexe a été appliquée, l'abstraction h-a-*hub*-autorité. Dans la table 5.7, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe biparti auteurs → publications citées obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution.

Abstraction	Paramètre	Motifs	Temps
h-a-*hub*-autorité	20 20	53	320
	15 15	147	2365
	12 12	408	5662
	10 10	851	8742
	8 8	2039	13923
	6 6	5483	21026
	5 5	10090	26510
	4 4	22131	32933
	3 3	56934	40175

TABLE 5.7 – Paramètres d'abstraction appliqués sur le graphe biparti auteurs → publications citées entre expertises construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus et temps d'exécution (en secondes)

Nous avons appliqué des abstractions de cœur *hub*-autorité avec les paramètres h et a suivants : 20 20, 15 15, 12 12, 10 10, 8 8, 6 6, 5 5, 4 4 et 3 3. Avec l'abstraction 20-20-*hub*-autorité, nous obtenons 53 motifs clos abstraits en un peu plus de 5 minutes. À l'aide de l'abstraction 3-3-*hub*-autorité, nous obtenons 56934 motifs clos abstraits en un peu plus de 11 heures. Le nombre de motifs clos abstraits obtenus a explosé, et le temps d'exécution s'est fortement allongé. Un bon compromis entre nombre de motifs clos abstraits et temps d'exécution est obtenu à l'aide de l'abstraction 8-8-*hub*-autorité. Avec ce paramètre, nous obtenons 2039 motifs clos abstraits en quasiment 4 heures.

Dans la table 5.8, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe biparti publications → auteurs cités obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution.

Abstraction	Paramètre	Motifs	Temps
h-a-*hub*-autorité	20 20	2	1
	15 15	170	883
	12 12	842	3499
	10 10	1941	6613
	8 8	5046	12243
	6 6	14376	21079
	5 5	25967	27019
	4 4	55150	35236

TABLE 5.8 – Paramètres d'abstraction appliqués sur le graphe biparti publications → auteurs cités entre expertises construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus et temps d'exécution (en secondes)

Nous avons appliqué des abstractions de cœur *hub*-autorité avec les paramètres h et a suivants : 20 20, 15 15, 12 12, 10 10, 8 8, 6 6, 5 5 et 4 4. Avec l'abstraction 20-20-*hub*-autorité, nous obtenons 2 motifs clos abstraits en une seconde. À l'aide de l'abstraction 4-

4-*hub*-autorité, nous obtenons 55150 motifs clos abstraits en presque 10 heures. Le nombre de motifs clos abstraits obtenus a explosé, et le temps d'exécution s'est fortement allongé. Un meilleur compromis entre nombre de motifs clos abstraits et temps d'exécution est obtenu à l'aide de l'abstraction 10-10-*hub*-autorité. Avec ce paramètre, nous obtenons 1941 motifs clos abstraits en moins de 2 heures.

Dans le chapitre 6, nous évaluons la qualité des experts et des expertises associées identifiés à l'aide des différents paramètres d'abstraction de graphe sur les graphes construits dans le cadre des expériences menées sur le corpus ACL Anthology.

Nombre de motifs clos abstraits sélectionnés

Les motifs clos sont sélectionnés d'après la méthode décrite dans la section 3.6 du chapitre 3. La sélection des motifs clos abstraits permet de conserver un nombre restreint de motifs pertinents limitant les redondances. Dans ce chapitre, nous avons rappelé que la sélection des motifs clos abstraits se fonde sur une mesure d'intérêt portée sur les motifs ainsi que sur le calcul de la distance entre deux motifs. Lors de nos expériences, nous avons utilisé la modularité comme mesure d'intérêt et la distance de Jaccard comme mesure de distance entre motifs. La distance de Jaccard est paramétrée par un seuil β. Dans l'ensemble des motifs clos abstraits sélectionnés, les distances entre toutes paires de motifs excèdent le seuil β et la somme des mesures d'intérêt individuels de chacun des motifs sélectionnés est maximisée.

Pour rappel, le seuil β peut être compris entre 0 et 1. Plus le seuil β est élevé, plus la distance entre paires de motifs sélectionnés est élevée et par conséquent, plus la sélection est drastique. Les motifs clos abstraits sélectionnés sont donc très différents mais également moins nombreux. Pour paramétrer le seuil β, plusieurs variables doivent être prises en compte. La nature du graphe considéré ainsi que le paramètre d'abstraction de graphe influent sur la sélection des motifs clos abstraits. Au sein de graphes de terrains comme ceux que nous considérons lors de nos expériences, il existe généralement une zone plus dense que les autres. En augmentant le paramètre d'abstraction de graphe, seuls les sous-graphes appartenant à cette zone très dense vérifient la contrainte topologique. Tous les motifs clos abstraits obtenus sont donc centrés sur une même zone du graphe et ont alors tendance à se ressembler. La sélection permet d'écarter un grand nombre de motifs clos abstraits jugés redondants. La proportion de motifs sélectionnés est alors faible et le nombre de motifs clos sélectionnés obtenus peut être négligeable. En considérant une contrainte topologique plus faible, les zones du graphe vérifiant la contrainte topologique sont plus nombreuses, par conséquent les motifs clos abstraits obtenus sont également plus nombreux et variés. Ainsi, la sélection permet de conserver un plus grand nombre de motifs clos abstraits. Ce phénomène est illustré par les exemples fournis dans les tables 5.11, 5.12 et 5.13.

Dans la figure 5.5, nous représentons le nombre de motifs clos abstraits sélectionnés

obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe de coauteurs construit à partir du corpus ACL Anthology, pour un seuil β de 0.8. Cette figure illustre le phénomène d'augmentation du nombre de motifs clos abstraits sélectionnés obtenus lors du relâchement du paramètre d'abstraction de graphe, quelle que soit l'abstraction de cœur employée. Dans la section A.2 de l'annexe, nous fournissons l'ensemble des courbes décrivant le nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur les graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology. Les courbes ont toutes la même allure.

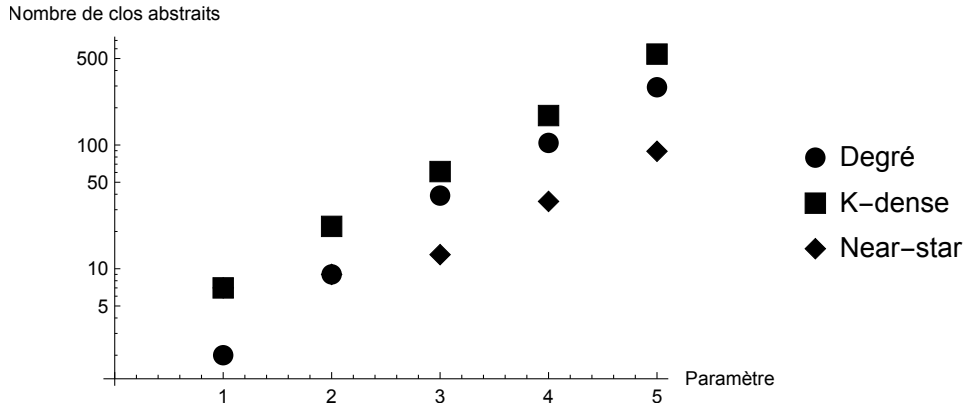


FIGURE 5.5 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, pour un seuil β de 0.8

Lors de nos expériences, nous avons comparé le nombre de motifs clos abstraits obtenus en fonction du paramètre β. Nous avons réalisé cette comparaison sur un graphe complexe, le graphe biparti auteurs $\rightarrow$ publications citées. Dans les tables 5.9 et 5.10, nous décrivons le nombre de motifs clos sélectionnés ainsi que le pourcentage de motifs clos abstraits sélectionnés en fonction du paramètre d'abstraction appliqué et du seuil β choisi.

β	20	15	12	10	8	6	5	4	3
0	0	13	89	266	816	2619	5710	14522	42425
0.2	0	7	39	101	325	987	2096	5088	14138
0.4	0	6	20	52	169	515	1069	2554	6716
0.6	0	4	10	30	87	241	492	1202	3053
0.8	0	3	5	12	26	79	164	393	962
1	0	1	1	1	1	1	1	1	1

TABLE 5.9 – Nombre de motifs clos abstraits sélectionnés en fonction du paramètre d'abstraction et du seuil β sur le graphe biparti auteurs $\rightarrow$ publications citées du corpus ACL Anthology

β	20	15	12	10	8	6	5	4	3
0	0	8,8	21,8	31,3	40	47,8	56,6	65,6	74,5
0.2	0	4,8	9,6	11,9	15,9	18	20,8	23	24,8
0.4	0	4,1	4,9	6,1	8,3	9,4	10,6	11,5	11,8
0.6	0	2,7	2,5	3,5	4,3	4,4	4,9	5,4	5,4
0.8	0	2	1,2	1,4	1,3	1,4	1,6	1,8	1,7
1	0	0,7	0,2	0,1	$< 0,1$	$< 0,1$	$< 0,1$	$< 0,1$	$< 0,1$

TABLE 5.10 – Pourcentage de motifs clos abstraits sélectionnés en fonction du paramètre d'abstraction et du seuil β sur le graphe biparti auteurs $\rightarrow$ publications citées du corpus ACL Anthology

Nous souhaitons limiter autant que possible les redondances tout en conservant un nombre de motifs clos abstraits sélectionnés suffisant pour décrire des experts potentiels variés. Nous constatons que pour la valeur 1 du seuil β, nous n'obtenons au mieux qu'un seul motif clos abstrait sélectionné. Ce seuil est donc trop contraint. Pour la valeur 0,8 du seuil β, nous sélectionnons de 1 à 2 % des motifs clos abstraits obtenus à l'aide des différents paramètres d'abstraction. Ce paramètre nous permet d'obtenir une sélection très forte. Nous avons fixé la valeur du seuil β à 0,8 sur l'ensemble des expériences réalisées sur le corpus ACL Anthology. Ainsi, nous obtenons une méthode de sélection homogène sur l'ensemble des expériences réalisée, bien que cette méthode de sélection aboutit à des résultats variés selon le graphe étudié.

Dans la table 5.11, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection pour chacun des paramètres d'abstraction testés sur le graphe de coauteurs obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution, pour un seuil β de 0.8. Ce tableau est également disponible en annexe, dans la section A.2.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
k-*core*	15	13	2	15,38	0
	12	74	9	12,16	4
	10	544	39	7,17	26
	8	2981	104	3,49	141
	6	17875	293	1,64	640
k-dense	15	20	7	35	1
	12	221	22	9,95	13
	10	1263	61	4,83	64
	8	6833	173	2,53	230
	6	47535	543	1,14	924
k-*nearstar*	50	43	7	16,28	46
	45	90	9	10	67
	40	232	13	5,6	120
	30	1365	35	2,56	402
	20	14530	89	0,61	2258

TABLE 5.11 – Paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection, pourcentage de motifs clos abstraits sélectionnés et temps d'exécution (en secondes), pour un seuil β de 0.8

Dans la table 5.12, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe biparti auteurs → publications citées obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution, pour un seuil β de 0.8.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
h-a-*hub*-autorité	20 20	53	1	1,89	320
	15 15	147	4	2,72	2365
	12 12	408	7	1,72	5662
	10 10	851	15	1,76	8742
	8 8	2039	33	1,62	13923
	6 6	5483	92	1,68	21026
	5 5	10090	178	1,76	26510
	4 4	22131	408	1,84	32933
	3 3	56934	1007	1,77	40175

TABLE 5.12 – Paramètres d'abstraction appliqués sur le graphe biparti auteurs → publications citées entre expertises construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection, pourcentage de motifs clos abstraits sélectionnés et temps d'exécution (en secondes), pour un seuil β de 0.8

Dans la table 5.13, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe biparti publications → auteurs

cités obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution, pour un seuil β de 0.8.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
h-a-*hub*-autorité	20 20	2	1	50	1
	15 15	170	6	3,53	883
	12 12	842	17	2,02	3499
	10 10	1941	34	1,75	6613
	8 8	5046	90	1,78	12243
	6 6	14376	227	1,58	21079
	5 5	25967	398	1,53	27019
	4 4	55150	734	1,33	35236

TABLE 5.13 – Paramètres d'abstraction appliqués sur le graphe biparti publications → auteurs cités entre expertises construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes), pour un seuil β de 0.8

Comme les tables 5.11, 5.12 et 5.13 l'illustrent, la sélection des motifs paramétrée au seuil $\beta = 0.8$ implique une sélection drastique des motifs clos abstraits obtenus sur l'ensemble des paramètres d'abstraction appliqués.

5.2.5 Identification des experts et de leurs expertises associées

Les motifs clos abstraits obtenus à l'aide d'abstractions de graphe ainsi que leurs extensions permettent d'identifier des experts et leurs expertises associées. Nous identifions les experts à l'aide des hypothèses d'expertise définies dans le chapitre 2, pour chacun des graphes que nous avons identifié comme pertinent pour la recherche d'experts. Nous illustrons les résultats que nous obtenons sur chaque graphe à l'aide d'exemples. L'évaluation des résultats est disponible dans le chapitre 6.

Nous avons utilisé un livre, Handbook of Natural Language Processing (INDURKHYA et DAMERAU 2010), afin de nous baser sur ses références pour discuter de l'expertise des individus identifiés par notre méthode en tant qu'experts potentiels. Ce livre réalise une revue de la littérature. Il est constitué de plusieurs chapitres abordant des thématiques de recherche variées, situées dans le domaine de la linguistique informatique. De surcroît, il a été publié en 2010, soit deux ans après la dernière publication du corpus ACL Anthology. Notre supposition est la suivante : si un auteur identifié comme expert sur un motif décrivant une thématique particulière est présent dans les références du chapitre de livre associé à cette même thématique, alors on peut confirmer son statut d'expert. Certains des chapitres du livre abordent les thématiques d'extraction d'information (*information extraction*), d'analyse syntaxique (*syntactic parsing*) ou de désambiguïsation lexicale (*word sense disambiguation*). Nous utilisons donc des exemples de motifs décrivant l'une de ces

thématiques.

Graphe de coauteurs

Considérons le graphe de coauteurs obtenu à partir du corpus ACL Anthology. À l'aide de l'abstraction de cœur 8-*core*, nous obtenons 2981 motifs clos abstraits dont 104 sont conservés après l'étape de sélection des motifs pour un seuil β paramétré à 0.8. Prenons l'exemple du motif q suivant parmi ces 104 motifs : q = {*information extraction*, *named entity recognition*, *natural language processing*, *text mining*, auteurs ayant publié entre 1993 et 2007}. Dans le graphe de coauteurs, l'extension de ce motif décrit un ensemble d'experts ayant comme caractéristiques communes maximales le motif q. Selon nos hypothèses, il s'agit donc d'experts sur les thématiques *information extraction*, *named entity recognition*, *natural language processing* et *text mining* ayant publié entre 1993 et 2007. Nous obtenons un ensemble de 18 experts représentés dans la figure 5.6. Les 18 experts obtenus peuvent être regroupés en deux cliques de coauteurs de 9 auteurs respectivement.

Les auteurs appartenant à une même clique sont les auteurs d'une même publication. Les auteurs appartenant à la clique située en haut de la figute 5.6 ont rédigé la publication intitulée Information Extraction Research And Applications : Current Progress And Future Directions publiée en 1998 et ayant pour identifiant X98-1013. Quant aux auteurs appartenant à la clique située en bas de la figute 5.6, ils ont rédigé la publication intitulée Maytag : A Multi-Staged Approach To Identifying Complex Events In Textual Data publiée en 2006 et ayant pour identifiant E06-2012. Ces deux publications abordent bien les thématiques d'extraction d'information, de reconnaissance d'entités nommées, de traitement du langage naturel et de fouille de texte.

Parmi les 18 experts potentiels identifiés, nous pouvons aisément identifier Jerry R. Hobbs, qui est un chercheur éminemment reconnu dans le domaine de l'intelligence artificielle et de la linguistique informatique. Quelques-uns de ses collègues du SRI International sont également présents dans la clique à laquelle il appartient, par exemple Megumi Kameyama ou Douglas Appelt. Cependant, il est difficile de confirmer le statut d'expert des individus identifiés à l'aide de nos seules connaissances.

Dans les références du chapitre Information Extraction du livre Handbook of Natural Language Processing, nous avons retrouvé 7 des 18 auteurs considérés comme experts potentiels sur ce motif. Nous avons retrouvé Jerry R. Hobbs, ainsi que ses coauteurs Douglas Appelt, Megumi Kameyane, Andrew Kehler, John Bear et David J. Israel. Parmi la seconde clique, nous avons retrouvé Benjamin Wellner. Il est à noter que le chapitre Information Extraction du livre a été écrit par Jerry R. Hobbs et Elen Riloff. L'étude des références d'articles de revue semble effectivement aider à déterminer l'expertise des individus. Nous constatons que chaque clique supportant ce motif contient au moins un expert confirmé d'après ce référentiel.

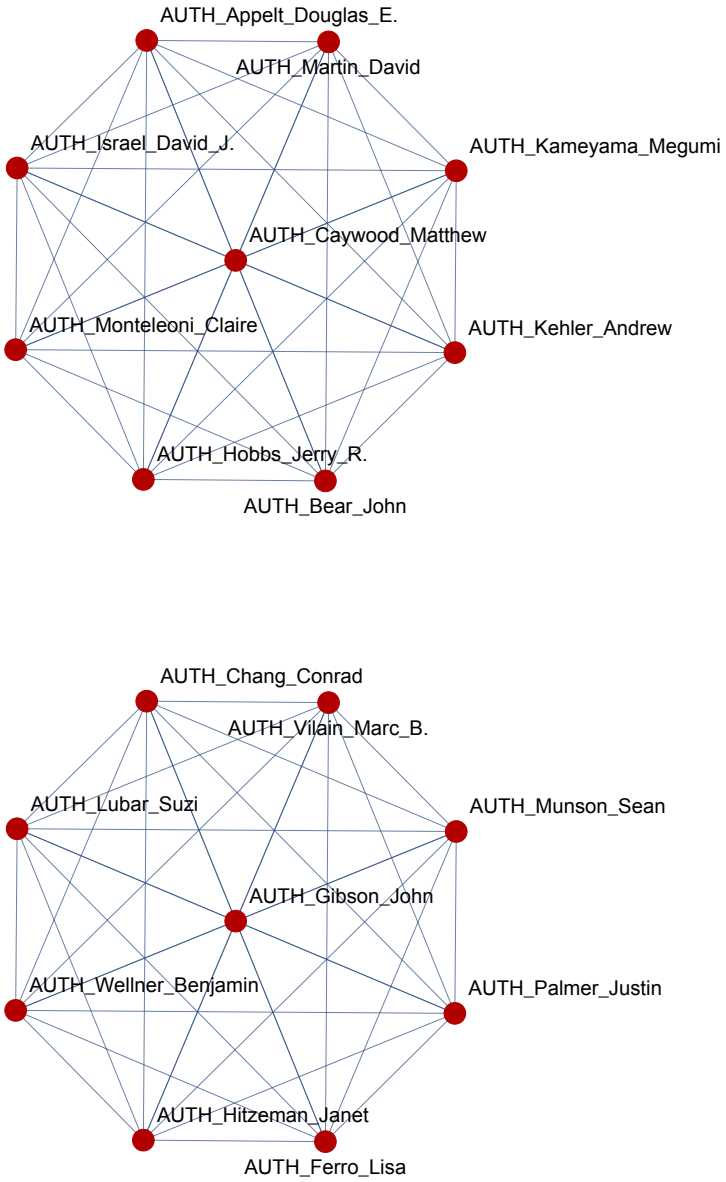


FIGURE 5.6 – Experts obtenus sur le motif q à l'aide d'une abstraction de cœur 8-*core* sur le graphe de coauteurs construit à partir du corpus ACL Anthology

Graphe de citation A

Considérons le graphe de citation A obtenu à partir du corpus ACL Anthology. À l'aide de l'abstraction de cœur 35-35-*hub*-autorité, nous obtenons 2528 motifs clos abstraits dont 2 sont conservés après sélection des motifs pour un seuil β paramétré à 0.8. Sur ce graphe, le nombre de motifs clos abstraits obtenus est très faible, car il est difficile d'appliquer une contrainte relâchée sans que le temps d'exécution n'explose, ce en raison de la densité du graphe. Prenons l'exemple du motif q suivant parmi ces 2 motifs : q = { *semantic information*, auteurs ayant publié entre 1993 et 2007}. Dans le graphe de citation A, l'extension de ce motif décrit un ensemble d'experts ayant comme caractéristiques communes maximales le motif q. Il s'agit donc d'experts sur la thématique *semantic information* ayant publié entre 1993 et 2007 selon nos hypothèses d'expertise. En effet, l'hypothèse

d'expertise associée au graphe de citation est la suivante : *si un individu est fortement cité par des membres variés de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine.* Nous obtenons un ensemble de 317 experts, ce qui semble un nombre très élevé, malgré la forte contrainte appliquée. Le nombre d'experts obtenus est trop grand pour être représenté dans une figure. Le graphe de citation A obtenu à partir du corpus ACL Anthology étant très dense, aucun des motifs clos abstraits obtenu après abstraction n'est supporté par un ensemble d'experts suffisamment petit pour être représenté dans une figure.

Pour ce motif, nous obtenons 225 auteurs, dont 176 font partie des auteurs citant, 141 des auteurs cités. Le livre Handbook of Natural Language Processing ne contient pas de chapitre décrivant la thématique *semantic information*. Cependant, nous pouvons utiliser un chapitre décrivant une thématique connexe. Nous avons sélectionné le chapitre Information Extraction. Sur l'ensemble des experts potentiels identifiés sur ce motif, 13 % sont identifiés comme des experts confirmés grâce au chapitre Information Extraction. Sur les experts potentiels citant, 15 % ont pu voir leur expertise confirmée à l'aide du chapitre. Sur les experts potentiels cités, 17 % sont considérés comme des experts confirmés.

Graphe de copublication

Considérons le graphe de copublication obtenu à partir du corpus ACL Anthology. À l'aide de l'abstraction de cœur 4-*nearstar*, nous obtenons 21777 motifs clos abstraits dont 2185 sont conservés après sélection des motifs pour un seuil β paramétré à 0.8. Prenons l'exemple du motif q suivant parmi ces 2185 motifs : q = {*syntactic parsing*, publié après 1992}. Dans le graphe de copublication, l'extension de ce motif décrit un ensemble de documents ayant comme caractéristiques communes maximales le motif q. Il s'agit donc de documents ayant tous été publié après 1992 sur la thématique *syntactic parsing*. Nous obtenons un ensemble de 18 documents représentés dans la figure 5.7. Les auteurs de ces 18 documents sont considérés comme experts de la thématique *syntactic parsing* après 1992.

Dans la figure 5.8, nous représentons les auteurs associés à chaque publication décrite par le motif.

Nous rappelons la principale hypothèse d'expertise associée au graphe de copublication : *si un individu rédige un grand nombre de publications scientifiques sur une thématique récurrente, alors il est probable qu'il s'agisse d'un expert de son domaine.*

Sur le motif considéré, nous avons identifié 30 auteurs différents. Pour confirmer l'expertise de ces experts potentiels, nous utilisons le chapitre Syntactic Parsing du livre Handbook of Natural Language Processing, puisque le motif décrit la thématique *syntactic parsing*. Parmi les 30 experts potentiels, 3 ont été retrouvés dans les références du chapitre : Lauri Karttunen, Joakim Nivre et Yorick Wilks.

Cependant, certains auteurs éminents ne sont pas retrouvés dans les références du

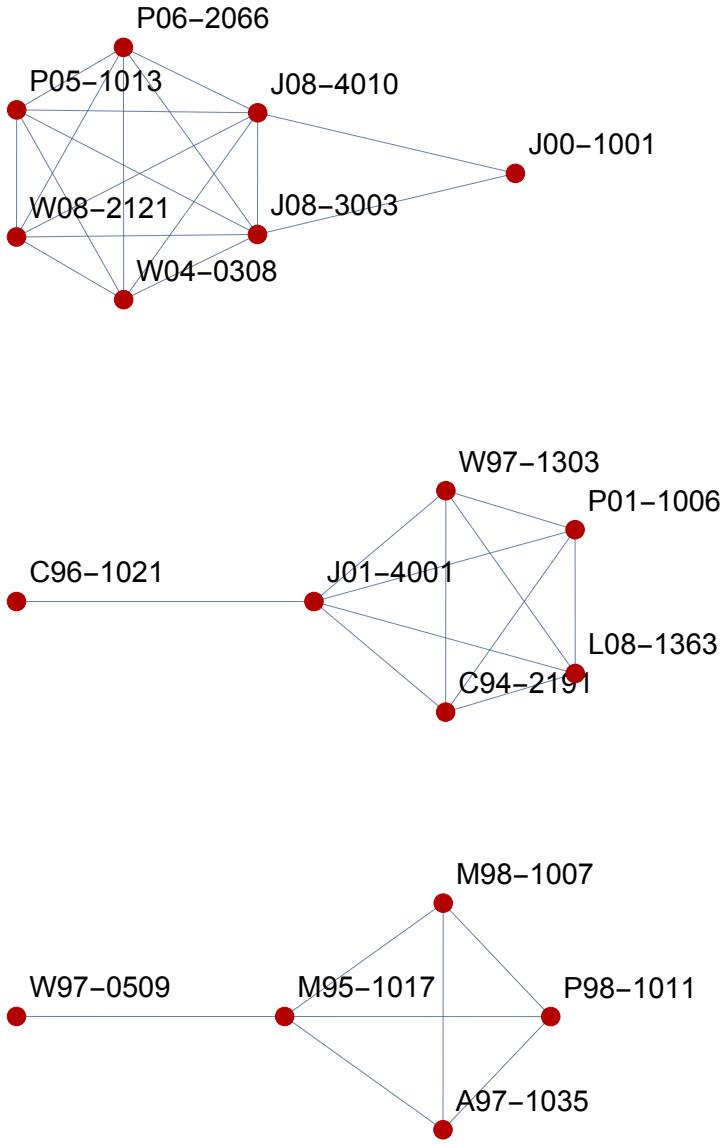


FIGURE 5.7 – Documents sources d'expertise obtenus sur le motif {*syntactic parsing*, publié après 1992} à l'aide d'abstraction de cœur 4-*nearstar* sur le graphe de copublication construit à partir du corpus ACL Anthology

chapitre, alors que nous pouvons aisément les reconnaître. Par exemple, Ruslan Mitkov est un chercheur phare du domaine. Cet auteur est notamment retrouvé dans la clique centrale de la figure 5.7. Il est l'auteur de cinq publications décrites par le motif. Ces publications possèdent les identifiants suivants : C94-2191, J01-4001, L08-1363, P01-1006 et W97-1303.

Graphe de citation D

Considérons le graphe de citation D obtenu à partir du corpus ACL Anthology. À l'aide de l'abstraction de cœur 3-3-*hub*-autorité, nous obtenons 774 motifs clos abstraits dont 145 sont conservés après sélection des motifs pour un seuil β paramétré à 0.8. Pre-

A97–1035	{AUTH_Cunningham_Hamish, AUTH_Humphreys_Kevin, AUTH_Gaizauskas_Robert_J., AUTH_Wilks_Yorick}
C94–2191	{AUTH_Mitkov_Ruslan}
C96–1021	{AUTH_Kennedy_Christopher, AUTH_Boguraev_Branimir_K.}
J00–1001	{AUTH_Karttunen_Lauri, AUTH_Oflazer_Kemal}
J01–4001	{AUTH_Mitkov_Ruslan, AUTH_Lappin_Shalom, AUTH_Boguraev_Branimir_K.}
J08–3003	{AUTH_Eryiğit_Gülşen, AUTH_Nivre_Joakim, AUTH_Oflazer_Kemal}
J08–4010	{AUTH_Eryiğit_Gülşen, AUTH_Nivre_Joakim, AUTH_Oflazer_Kemal}
L08–1363	{AUTH_Orăsan_Constantin, AUTH_Cristea_Dan, AUTH_Mitkov_Ruslan, AUTH_Branco_António_H.}
M95–1017	{AUTH_Gaizauskas_Robert_J., AUTH_Wakao_Takahiro, AUTH_Humphreys_Kevin, AUTH_Cunningham_Hamish, AUTH_Wilks_Yorick}
M98–1007	{AUTH_Humphreys_Kevin, AUTH_Gaizauskas_Robert_J., AUTH_Azzam_Saliha, AUTH_Huyck_Christian, AUTH_Mitchell_B., AUTH_Cunningham_Hamish, AUTH_Wilks_Yorick}
P01–1006	{AUTH_Barbu_Catalina, AUTH_Mitkov_Ruslan}
P05–1013	{AUTH_Nivre_Joakim, AUTH_Nilsson_Jens}
P06–2066	{AUTH_Kuhlmann_Marco, AUTH_Nivre_Joakim}
P98–1011	{AUTH_Azzam_Saliha, AUTH_Humphreys_Kevin, AUTH_Gaizauskas_Robert_J.}
W04–0308	{AUTH_Nivre_Joakim}
W08–2121	{AUTH_Surdeanu_Mihai, AUTH_Johansson_Richard, AUTH_Meyers_Adam, AUTH_Màrquez_Lluís, AUTH_Nivre_Joakim}
W97–0509	{AUTH_Wakao_Takahiro, AUTH_Ehara_Terumasa, AUTH_Sawamura_Eiji, AUTH_Abe_Yoshiharu, AUTH_Shirai_Katsuhiko}
W97–1303	{AUTH_Mitkov_Ruslan}

FIGURE 5.8 – Auteurs associés aux documents sources d'expertise obtenus sur le motif {*syntactic parsing*, publié après 1992} à l'aide d'abstraction de cœur 4-*nearstar* sur le graphe de copublication construit à partir du corpus ACL Anthology

nons l'exemple du motif q suivant parmi ces 145 motifs : q = {*information extraction*, publié entre 2000 et 2007}. Dans le graphe de citation D, l'extension de ce motif décrit un ensemble de documents ayant comme caractéristiques communes maximales le motif q. Il s'agit donc de documents ayant été publié entre 2000 et 2007 sur la thématique « *information extraction* ». Nous obtenons un ensemble de 10 documents représentés dans la figure 5.9. Les auteurs de ces 10 documents sont considérés comme experts de la thématique *information extraction* entre 2000 et 2007. Dans la figure, les sommets bleus sont des *hubs* et les sommets rouges des autorités. Les sommets étant à la fois *hubs* et autorités sont colorés en rouge et en bleu. Selon le rôle (*hub* ou autorité) associé à un document, l'expertise de son auteur est caractérisée à l'aide d'un indicateur différent. Trois hypothèses d'expertise ont été suggérées à partir du graphe de citation. La première est la suivante : *si un individu est auteur d'une publication phare, alors il est probable qu'il s'agisse d'un expert de son domaine*. Dans le graphe de citation, les publications phares correspondent aux autorités. La seconde hypothèse d'expertise est la suivante : *si un individu est auteur d'un article de revue, alors il est probable qu'il s'agisse d'un expert de son domaine*. Enfin, la troisième hypothèse d'expertise est la suivante : *si un individu est auteur d'un article citant un grand nombre de publications phares, alors il est probable qu'il s'agisse d'un expert de son domaine*. Dans le graphe de citation, les *hubs* peuvent être considérés comme des articles de revue ou des articles ayant pour références des publications phares du domaine.

Dans la figure 5.10, nous représentons les auteurs associés à chaque publication décrite par le motif.

Sur le motif considéré, nous avons identifié 16 auteurs différents. Pour confirmer l'expertise de ces experts potentiels, nous utilisons le chapitre Information Extraction du livre Handbook of Natural Language Processing, puisque le motif décrit la thématique

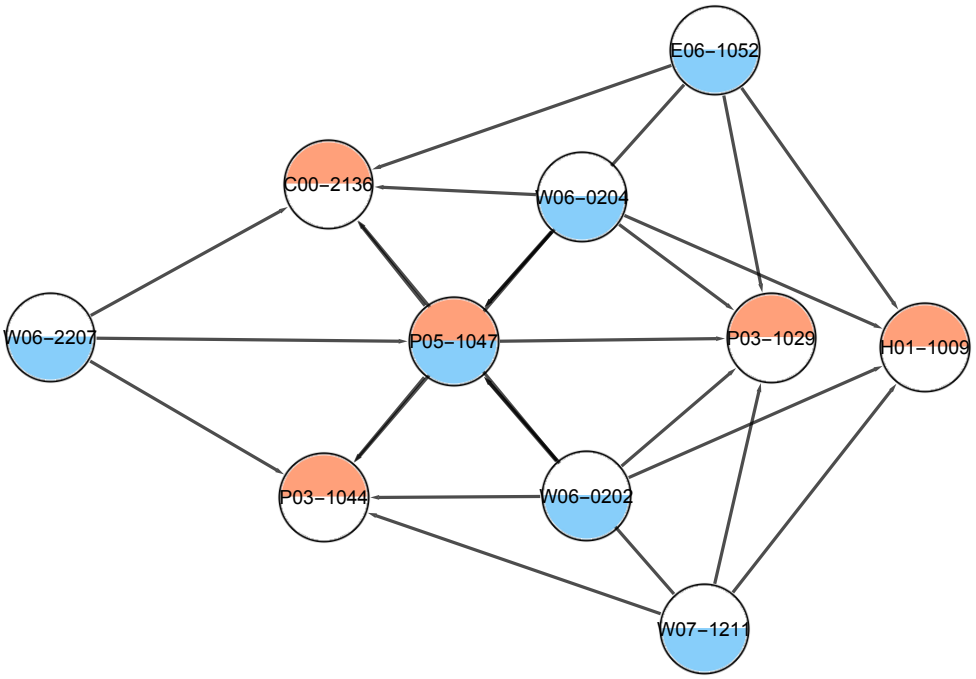


FIGURE 5.9 – Experts obtenus sur le motif q à l'aide d'abstraction de cœur sur le graphe de citation D construit à partir du corpus ACL Anthology

information extraction. Parmi les 16 experts potentiels, 10 ont été retrouvés dans les références du chapitre. 4 font partie des auteurs de publications citantes, 8 des auteurs de publications citées. Sur l'ensemble des experts potentiels identifiés sur ce motif, 63 % sont identifiés comme des experts confirmés à l'aide du chapitre Information Extraction. Sur les experts potentiels identifiés comme auteurs de publications citantes, 40 % ont pu voir leur expertise confirmée à l'aide du chapitre. Sur les experts potentiels identifiés comme auteurs de publications citées, 100 % sont considérés comme des experts confirmés.

Graphe biparti publications-auteurs

Considérons le graphe biparti publications-auteurs obtenu à partir du corpus ACL Anthology. À l'aide de l'abstraction de cœur 14-*nearstar*, nous obtenons 907 bimotifs clos abstraits dont 65 sont conservés après sélection des motifs pour un seuil β paramétré à 0.8. Prenons l'exemple du bimotif q suivant parmi ces 65 bimotifs, avec q = $\{q_1, q_2\}$. Le motif q_1 décrit les caractéristiques communes maximales partagées par les publications, q_2 décrit les caractéristiques communes maximales partagées par les auteurs. On a q_1 = {*information extraction*} , q_2 = {*information extraction*, auteurs ayant publié après 2007}. Dans le graphe biparti publications-auteurs, l'extension de ce bimotif décrit un ensemble d'experts et de documents ayant comme caractéristiques communes maximales le bimotif q. Il s'agit donc d'auteurs ayant publié après 2007 et de documents avec lesquels ils entretiennent un lien de paternité. Nous obtenons un ensemble de 16 experts et 27 documents représentés dans la figure 5.11. Nous obtenons deux cliques. Dans la première

C00–2136	{AUTH_Yangarber_Roman, AUTH_Grishman_Ralph, AUTH_Tapanainen_Pasi, AUTH_Huttunen_Silja}
E06–1052	{AUTH_Romano_Lorenza, AUTH_Kouylekov_Milen, AUTH_Szpektor_Idan, AUTH_Dagan_Ido, AUTH_Lavelli_Alberto}
H01–1009	{AUTH_Sudo_Kiyoshi, AUTH_Sekine_Satoshi, AUTH_Grishman_Ralph}
P03–1029	{AUTH_Sudo_Kiyoshi, AUTH_Sekine_Satoshi, AUTH_Grishman_Ralph}
P03–1044	{AUTH_Yangarber_Roman}
P05–1047	{AUTH_Stevenson_Mark, AUTH_Greenwood_Mark_A.}
W06–0202	{AUTH_Stevenson_Mark, AUTH_Greenwood_Mark_A.}
W06–0204	{AUTH_Greenwood_Mark_A., AUTH_Stevenson_Mark}
W06–2207	{AUTH_Surdeanu_Mihai, AUTH_Turmo_Jordi, AUTH_Ageno_Alicia}
W07–1211	{AUTH_Greenwood_Mark_A., AUTH_Stevenson_Mark}

FIGURE 5.10 – Auteurs associés aux documents sources d'expertise obtenus sur le motif {*information extraction*, publié entre 2000 et 2007} à l'aide d'abstraction de cœur 3-3-*hub*-autorité sur le graphe de citation D construit à partir du corpus ACL Anthology

clique, l'étoile est un auteur, les satellites ses publications associées. Dans la seconde clique, l'étoile est une publication scientifique, les satellites des auteurs.

Ralph Grishman, l'étoile de la première clique, est un chercheur éminent. Quant aux chercheurs constituant les satellites de la seconde clique, si certains sont reconnaissables, comme German Rigau ou Eneko Agirre, aucun d'entre eux n'est retrouvé dans les références du chapitre Information Extraction du livre Handbook of Natural Language Processing.

Graphe biparti auteurs → publications citées

Considérons le graphe biparti auteurs → publications citées obtenu à partir du corpus ACL Anthology. À l'aide de l'abstraction de cœur 4-4-*hub*-autorité, nous obtenons 22131 motifs clos abstraits dont 408 sont conservés après sélection des motifs pour un seuil β paramétré à 0.8. Prenons l'exemple du bimotif q suivant parmi ces 408 motifs, avec q = $\{q_1, q_2\}$. Le motif q_1 décrit les caractéristiques communes maximales partagées par les auteurs, le motif q_2 les caractéristiques communes maximales partagées par les documents cités. On a q_1 = { *syntactic parsing*, auteurs ayant publié entre 2000 et 2004 ainsi qu'après 2004 }, q_2 = { *syntactic parsing*, publications datant d'avant 2000}. Dans le graphe biparti auteurs → publications citées, l'extension de ce bimotif décrit un ensemble d'experts et de documents cités ayant comme caractéristiques communes maximales le bimotif q. Il s'agit donc d'experts sur la thématique *syntactic parsing* ayant publié entre 2000 et 2004 ainsi qu'après 2004, citant des publications scientifiques sur cette même thématique publiés avant 2000. Nous obtenons un ensemble de 9 experts et 8 documents représentés dans la figure 5.12. Dans la figure, les sommets bleus sont des *hubs* (des auteurs) et les sommets rouges des autorités (des publications). Deux hypothèses d'expertise ont été suggérées à partir du graphe biparti auteurs → publications citées. La première est la suivante : *si un individu est auteur d'une publication phare, alors il est probable qu'il s'agisse d'un expert*

de son domaine. Dans le graphe biparti auteurs → publications citées, les publications phares correspondent aux autorités. La seconde hypothèse d'expertise est la suivante : *si un individu cite des publications phares d'un domaine, alors il est probable qu'il s'agisse d'un expert de ce domaine*. Dans le graphe biparti auteurs → publications citées, les *hubs* peuvent être considérés comme des auteurs citant des publications phares du domaine.

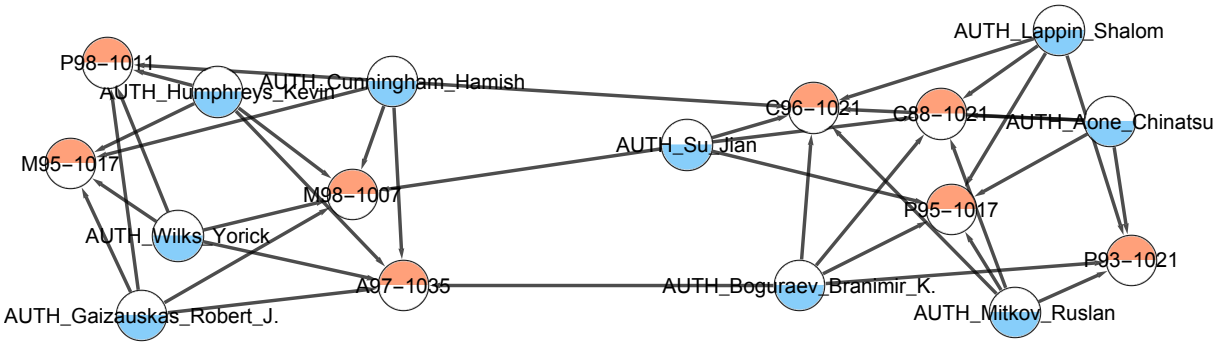


FIGURE 5.12 – Experts obtenus sur le bimotif q à l'aide d'abstraction de cœur 4-4-*hub*-autorité sur le graphe biparti auteurs → publications citées construit à partir du corpus ACL Anthology

Dans la figure 5.13, nous représentons les auteurs associés à chaque publication décrite par le bimotif.

A97–1035	{AUTH_Cunningham_Hamish, AUTH_Humphreys_Kevin, AUTH_Gaizauskas_Robert_J., AUTH_Wilks_Yorick}
C88–1021	{AUTH_Carbonnell_Jaime_G., AUTH_Brown_Ralf_D.}
C96–1021	{AUTH_Kennedy_Christopher, AUTH_Boguraev_Branimir_K.}
M95–1017	{AUTH_Gaizauskas_Robert_J., AUTH_Wakao_Takahiro, AUTH_Humphreys_Kevin, AUTH_Cunningham_Hamish, AUTH_Wilks_Yorick}
M98–1007	{AUTH_Humphreys_Kevin, AUTH_Gaizauskas_Robert_J., AUTH_Azzam_Saliha, AUTH_Huyck_Christian, AUTH_Mitchell_B., AUTH_Cunningham_Hamish, AUTH_Wilks_Yorick}
P93–1021	{AUTH_Aone_Chinatsu, AUTH_McKee_Douglas}
P95–1017	{AUTH_Aone_Chinatsu, AUTH_William_Scott}
P98–1011	{AUTH_Azzam_Saliha, AUTH_Humphreys_Kevin, AUTH_Gaizauskas_Robert_J.}

FIGURE 5.13 – Auteurs associés aux documents sources d'expertise obtenus sur le bimotif {{ *syntactic parsing*, auteurs ayant publié entre 2000 et 2004 ainsi qu'après 2004 }, { *syntactic parsing*, publications datant d'avant 2000}} à l'aide d'abstraction de cœur 4-4-*hub*-autorité sur le graphe biparti auteurs → publications citées construit à partir du corpus ACL Anthology

Sur le motif considéré, nous avons identifié 18 auteurs différents. Puisque le motif décrit la thématique *syntactic parsing*, nous avons utilisé le chapitre Syntactic Parsing du livre Handbook of Natural Language Processing pour confirmer l'expertise des experts potentiels identifiés. Parmi les 18 experts potentiels, 9 font partie des auteurs citants, 15 des auteurs de publications citées. Aucun n'a été retrouvé à l'aide du chapitre. Certains des experts potentiels sont cependant aisément reconnaissables, comme Yorick Wilks, ou Ruslan Mitkov.

Graphe biparti publications → auteurs cités

Considérons le graphe biparti publications → auteurs cités obtenu à partir du corpus ACL Anthology. À l'aide de l'abstraction de cœur 8-8-*hub*-autorité, nous obtenons 5046 motifs clos abstraits dont 90 sont conservés après sélection des motifs pour un seuil β paramétré à 0.8. Prenons l'exemple du bimotif q suivant parmi ces 90 motifs, avec q = $\{q_1, q_2\}$. Le motif q_1 décrit les caractéristiques communes maximales partagées par les publications, le motif q_2 les caractéristiques communes maximales partagées par les auteurs cités. On a q_1 = {*information extraction*, publication entre 1993 et 2007}, q_2 = {*information extraction*, auteurs ayant publié entre 1993 et 1999}. Dans le graphe biparti auteurs → publications citées, l'extension de ce bimotif décrit un ensemble d'experts et de documents cités ayant comme caractéristiques communes maximales le bimotif q. Il s'agit donc de publications sur la thématique *information extraction* publiées entre 1993 et 1999, citant des auteurs ayant publié sur cette même thématique entre 1993 et 1999. Nous obtenons un ensemble de 21 experts et 22 documents représentés dans la figure 5.14. Dans la figure, les sommets bleus sont des *hubs* (des publications) et les sommets rouges des autorités (des auteurs). Deux hypothèses d'expertise ont été suggérées à partir du graphe biparti publications → auteurs cités. La première est la suivante : *si un individu est fortement cité par les membres de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine*. Dans le graphe biparti publications → auteurs cités, les auteurs cités correspondent aux autorités. La seconde hypothèse d'expertise est la suivante : *si un individu cite des auteurs appropriés du domaine dans l'une de ses publications, alors il est probable qu'il s'agisse d'un expert de ce domaine*. Dans le graphe biparti publications → auteurs cités, les *hubs* peuvent être considérés comme des publications citant des auteurs appropriés du domaine. Donc, les auteurs des pubications *hubs* peuvent être considérés comme des experts.

Dans la figure 5.15, nous représentons les auteurs associés à chaque publication décrite par le bimotif.

Sur le bimotif considéré, nous avons identifié 59 auteurs différents. Pour confirmer l'expertise de ces experts potentiels, nous utilisons le chapitre Information Extraction du livre Handbook of Natural Language Processing, puisque le bimotif décrit la thématique *information extraction*. Parmi les 59 experts potentiels, 19 ont été retrouvés dans les références du chapitre. 15 font partie des auteurs de publications citantes, 7 des auteurs cités. Sur l'ensemble des experts potentiels identifiés sur ce motif, 32 % sont identifiés comme des experts confirmés à l'aide du chapitre Information Extraction. Sur les experts potentiels identifiés comme auteurs de publications citantes, 31 % ont pu voir leur expertise confirmée à l'aide du chapitre. Sur les experts potentiels identifiés comme auteurs de publications citées, 33 % sont considérés comme des experts confirmés.

5.2.6 Représentation étendue de motifs clos abstraits

Dans la section 3.7 du chapitre 3, nous avons défini plusieurs méthodes permettant d'obtenir une représentation étendue d'un motif clos abstrait. Les motifs clos abstraits éliminés lors du processus de sélection peuvent comporter des éléments du langage de description utiles pour étendre la définition des motifs sélectionnés. Ainsi, il est possible de suggérer une requête de recherche d'experts similaire à celle soumise par un utilisateur.

Prenons l'exemple d'un motif clos abstrait décrivant l'une des thématiques associées à notre *gold standard*. Nous avons choisi l'exemple du bimotif q suivant, décrivant la thématique *information extraction* :

q = {{*information extraction*, auteurs ayant publié après 1992}{*information extraction*, publications datant d'avant 2008}}

Ce bimotif a été obtenu à l'aide d'une abstraction 4-4-*hub*-autorité sur le graphe biparti auteurs → publications citées. Nous utilisons le chapitre Information Extraction du livre Handbook of Natural Language Processing pour évaluer les performances de notre méthode sur ce bimotif. Les experts potentiels identifiés à l'aide de notre méthode sont validés lorsqu'ils apparaissent dans les références du chapitre de livre. Sur les experts potentiels identifiés sur ce bimotif, nous obtenons une précision de 0.25, un rappel de 0.68 et une f-mesure de 0.36.

Nous proposons trois représentations de ce bimotif : une représentation étendue du bimotif basée sur l'intension, $rep_{int}(q)$, ainsi que deux représentations étendues du bimotif basée sur l'extension, $rep_{ext1}(q)$ et $rep_{ext2}(q)$. La représentation étendue d'un bimotif basée sur l'intention utilise les vecteurs de fréquence d'apparition de chaque élément du langage de description dans les bimotifs proches du bimotif considéré. La première des représentations étendues d'un bimotif basées sur l'extension est obtenue à partir de l'analyse de l'extension du bimotif q tandis que la seconde est obtenue à l'aide des extensions de l'ensemble des bimotifs proches de q que l'on nomme N_j.

Nous avons obtenu 18 bimotifs proches de q, $Nj = \{qi \in Q | d(q_j, d_i) \leq \beta\}$ avec β paramétré à 0.3. Ces bimotifs sont les suivants :

1. q
2. {{*information extraction*, auteurs ayant publié après 2004}, {*information extraction*, publication datant d'avant 2008}}
3. {{*information extraction*, auteurs ayant publié après 1999 et avant 2005 ainsi qu'après 2004 et avant 2008}, {*information extraction*, publications datant d'avant 2008}}
4. {{*information extraction*, auteurs ayant publié après 1999 et avant 2005 ainsi qu'après 2004 et avant 2008}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}
5. {{*information extraction*, auteurs ayant publié après 2004} {*information extraction*, auteurs ayant publié après 1992 et avant 2008}}

6. {{*information extraction, auteurs ayant publié après 2004 et avant 2008*}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}

7. {{*information extraction, auteurs ayant publié après 2004 et avant 2008*}, {*information extraction*, publications datant d'avant 2008}}

8. {{*information extraction, auteurs ayant publié après 1992 et avant 2005*}, {*information extraction*, publications datant d'avant 2008}}

9. {{*information extraction, auteurs ayant publié après 1999 et avant 2005*}, {*information extraction*, publications datant d'avant 2008}}

10. {{*information extraction, auteurs ayant publié après 1999 et avant 2005*}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}

11. {{*information extraction, auteurs ayant publié après 1992 et avant 2005*}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}

12. {{*information extraction, auteurs ayant publié après 1999*}, {*information extraction*, publications datant d'avant 2008}}

13. {{*information extraction, auteurs ayant publié après 1999*}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}

14. {{*information extraction, auteurs ayant publié après 1999 et avant 2008*}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}

15. {{*information extraction, auteurs ayant publié après 1999 et avant 2008*}, {*information extraction*, publications datant d'avant 2008}}

16. {{*information extraction, auteurs ayant publié après 1992*}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}

17. {{*information extraction, auteurs ayant publié après 1992 et avant 2008*}, {*information extraction*, publications datant d'après 1992 et d'avant 2008}}

18. {{*information extraction, auteurs ayant publié après 1992 et avant 2008*}, {*information extraction*, publications datant d'avant 2008}}

Les performances obtenues à l'aide de notre méthode sur les bimotifs proches de q sont indiquées dans la table 5.14. Sur 12 des 17 bimotifs qui ne correspondent pas à q, nous obtenons une meilleure f-mesure que sur le motif q.

Bimotif	P	R	F
1	.25	.68	.36
2	.25	.66	.36
3	**.27**	.66	**.39**
4	**.29**	.66	**.41**
5	**.26**	.66	**.37**
6	**.26**	.66	**.38**
7	.25	.66	.36
8	**.27**	.68	**.39**
9	**.27**	.68	**.39**
10	**.29**	.67	**.40**
11	**.28**	.67	**.40**
12	.25	.68	.36
13	**.26**	.67	**.37**
14	**.26**	.67	**.38**
15	.25	.68	.36
16	.25	.67	**.37**
17	**.26**	.67	**.37**
18	.25	.68	.36

TABLE 5.14 – Performances de notre méthode sur les bimotifs Nj (avec q le premier bimotif)

$rep_{int}(q)$ = {{*information extraction*, auteurs ayant publié après 1999 et avant 2008},{ *information extraction*, publications datant d'avant 2008}}

Sur ce bimotif, nous obtenons une précision de 0.25, un rappel de 0.68 et une f-mesure de 0.36, tout comme pour le bimotif q. En réalité, la f-mesure obtenue sur la représentation étendue basée sur l'intension est légèrement meilleure, mais cette amélioration n'est visible qu'à partir des centièmes, donc nous la considérons comme négligeable. Les fréquences associées aux éléments du langage appartenant à la représentation étendue sont indiquées dans la table 5.15.

Élément du langage	Fréquence d'apparition
information extraction	1
Publications datant d'avant 2008	1
Auteurs ayant publié après 1992	1
Auteurs ayant publié avant 2008	0.66
Auteurs ayant publié après 1999	0.66

TABLE 5.15 – Fréquences associées aux éléments du langage appartenant à la représentation étendue du bimotif q basée sur l'intention, $rep_{int}(q)$

Le calcul des représentations étendues basées sur l'extension du bimotif q ou sur les extensions de tous les bimotifs proches de q génère de nombreux candidats potentiels parmi

les éléments du langage de description pour l'extension d'un bimotif. La première représentation étendue basée sur l'extension, $rep_{ext1}(q)$, analyse la proportion de sommets de l'extension du bimotif q dans lesquelles apparaît tout élément du langage de description. La seconde représentation étendue basée sur l'extension, $rep_{ext2}(q)$, analyse la proportion de sommets de toutes les extensions des bimotifs proches du motif q dans lesquelles apparaît tout élément du langage de description. Pour restreindre la représentation étendue du bimotif, il est possible de définir un seuil α tel que $p > \alpha$, avec p la proportion de sommets de l'extension du bimotif q (de toutes les extensions des bimotifs proches du bimotif q respectivement) dans lesquelles apparaît tout élément du langage de description. Par exemple, définissons $\alpha = 0.8$, ce qui signifie que tout élément du langage appartenant à la représentation étendue du bimotif apparaît dans au moins 80 % des sommets de l'extension du bimotif q (des extensions des bimotifs proches de q respectivement). On a :

$rep_{ext1}(q) = \{\{$*information extraction*, auteurs ayant publié après 1999 et avant 2005, après 2004 et avant 2008 ainsi qu'après 2007$\}$,$\{$*information extraction*, publications datant d'après 1999 et avant 2005$\}\}$

Ce bimotif n'est pas obtenu à l'aide de l'abstraction appliquée qui nous a permis d'obtenir le motif q sur le graphe considéré.

Tout élément du langage de description apparaissant dans au moins 80 % des sommets de l'extension du bimotif q appartient à la représentation étendue du bimotif q basée sur l'extension, $rep_{ext1}(q)$. Dans la table 5.16, nous présentons ces éléments du langage ainsi que la proportion de sommets de l'extension du bimotif q dans lesquels ils apparaissent.

Élément du langage	Proportion de sommets de $ext(q)$
information extraction	1
Publications datant d'avant 2007	1
Auteurs ayant publié après 1992	1
Publications datant d'après 1992	0.99
Auteurs ayant publié avant 2008	0.99
Auteurs ayant publié après 1999	0.98
Auteurs ayant publié après 2004	0.93
Auteurs ayant publié avant 2005	0.93
Publications datant d'avant 2005	0.89
Publications datant d'après 1999	0.85
Auteurs ayant publié après 2007	0.80

TABLE 5.16 – Proportions de sommets dans lesquels les éléments du langage appartenant à la représentation étendue du bimotif q basée sur l'extension, $rep_{ext1}(q)$ apparaissent

On a $rep_{ext2}(q) = rep_{ext1}(q)$.

Tout élément du langage de description apparaissant dans au moins 80 % des sommets de toutes les extensions des bimotifs proches du bimotif q appartient à la représentation étendue du bimotif q basée sur l'extension, $rep_{ext2}(q)$. Ces éléments du langage ainsi que

la proportion de sommets des extensions des bimotifs proches du bimotif q dans lesquels ils apparaissent sont les mêmes que ceux indiqués dans la table 5.16.

Les représentations étendues $rep_{int}(q)$, $rep_{ext1}(q)$ et $rep_{ext2}(q)$ correspondent à des bimotifs qui ne sont pas retrouvés dans les bimotifs clos abstraits obtenus à l'aide du paramètre d'abstraction appliqué ayant permis d'obtenir le bimotif q sur le graphe considéré. Nous constatons que la méthode basée sur l'intension nous permet d'obtenir une représentation étendue très différente de celles basées sur l'extension, qui elles nous permettent d'obtenir une représentation identique sur cet exemple. Ces représentations permettent de mettre en lumière des connaissances intéressantes.

À l'aide de la représentation basée sur l'intention $rep_{int}(q)$, nous avons pu déterminer que les deux tiers des motifs proches de q décrivent des auteurs ayant publié après 1999. De la même manière, deux tiers des motifs proches de q décrivent des auteurs ayant publié avant 2008. Soit l l'élément du langage suivant : « auteurs ayant publié après 1999 ». Nous constatons que la proportion de sommets de l'extension du motif q (ou de toutes les extensions des motifs proches de q) dans lesquels apparaissent l atteint cette fois-ci 98 %. La représentation étendue d'un motif, particulièrement les représentations étendues basées sur l'extension permettent d'obtenir des suggestions d'extension de requêtes de recherche d'expert qui, si elles ne permettent pas toujours d'augmenter les performances obtenues, permettent de mettre en lumière des connaissances pertinentes.

5.3 Conclusion

Nous avons testé la validité de notre approche en proposant des expériences sur le corpus ACL Anthology. Nous proposons un protocole expérimental générique consistant en l'extraction des thématiques de recherche à partir des résumés des publications scientifiques, la construction de tables de connaissances intermédiaires, la construction de graphes attribués pertinents pour la recherche d'experts puis, enfin, l'abstraction de ces graphes.

Nous avons également proposé un comparatif des différentes méthodes d'extraction d'expertises à partir des publications scientifiques. Nous comparons une méthode basée sur l'extraction de concepts à l'aide d'une ontologie, une seconde basée sur une extraction de termes et une dernière sur une extraction de phrases-clefs. Nous démontrons que si l'extraction de concepts ne permet pas d'obtenir le plus grand nombre d'expertises extraites, elle permet d'obtenir des expertises plus pertinentes.

Enfin, nous illustrons les experts et expertises associées obtenus lors de l'abstraction de graphe à l'aide d'exemples, pour chacun des graphes construits à partir du corpus ACL Anthology. Pour chaque exemple, nous proposons une analyse qualitative des experts et expertises associées identifiés.

Dans le chapitre 6, nous proposons une évaluation de l'ensemble des résultats obtenus

sur le corpus ACL Anthology. Cette évaluation est basée sur un jeu de données d'évaluation construit à partir de chapitres d'une revue de la littérature. Nous proposons également une comparaison des résultats obtenus sur le corpus ACL Anthology avec les résultats obtenus sur les principales méthodes de recherche d'experts issues de l'état de l'art à l'aide de la plateforme d'évaluation LT ExpertFinder.

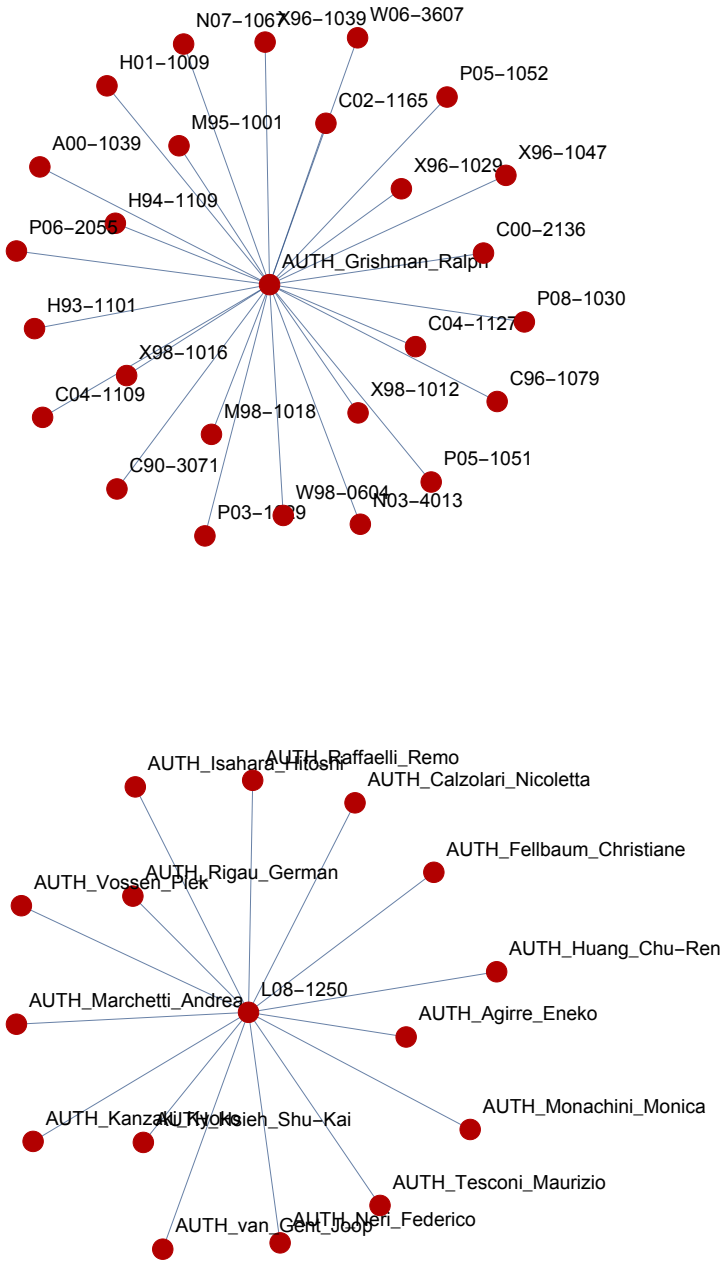


FIGURE 5.11 – Experts obtenus sur le motif q à l'aide d'abstraction de cœur 4-*near-star* sur le graphe publications-auteurs construit à partir du corpus ACL Anthology

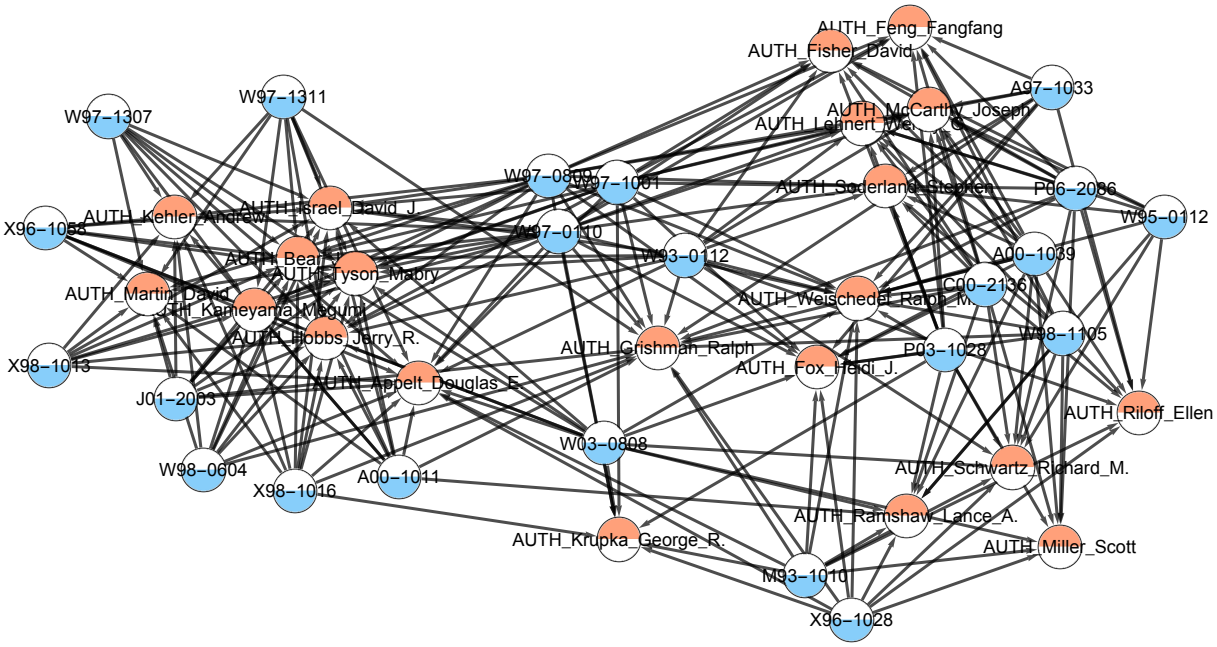


FIGURE 5.14 – Experts obtenus sur le bimotif q à l'aide d'abstraction de cœur 8-8-*hub*-autorité sur le graphe biparti publications → auteurs cités construit à partir du corpus ACL Anthology

A00–1011	{AUTH_Aone_Chinatsu, AUTH_Ramos-Santacruz_Mila}
A00–1039	{AUTH_Yangarber_Roman, AUTH_Grishman_Ralph, AUTH_Tapanainen_Pasi, AUTH_Huttunen_Silja}
A97–1033	{AUTH_Radev_Dragomir_R., AUTH_McKeown_Kathleen_R.}
C00–2136	{AUTH_Yangarber_Roman, AUTH_Grishman_Ralph, AUTH_Tapanainen_Pasi, AUTH_Huttunen_Silja}
J01–2003	{AUTH_Kehler_Andrew, AUTH_Appelt_Douglas_E., AUTH_Bear_John}
M93–1010	{AUTH_The_PLUM_Research_Group}
P03–1028	{AUTH_Chieu_Hai_Leong, AUTH_Ng_Hwee_Tou, AUTH_Lee_Yoong_Keok}
P06–2086	{AUTH_Rosenfeld_Benjamin, AUTH_Feldman_Ronen}
W03–0808	{AUTH_Srihari_Rohini_K., AUTH_Li_Wei, AUTH_Niu_Cheng, AUTH_Cornell_Thomas_L.}
W93–0112	{AUTH_Waterman_Scott}
W95–0112	{AUTH_Riloff_Ellen, AUTH_Shoen_Jay}
W97–0110	{AUTH_Chai_Joyce, AUTH_Biermann_Alan_W.}
W97–0809	{AUTH_Chai_Joyce, AUTH_Biermann_Alan_W.}
W97–1001	{AUTH_Bagga_Amit, AUTH_Chai_Joyce}
W97–1307	{AUTH_Kameyama_Megumi}
W97–1311	{AUTH_Humphreys_Kevin, AUTH_Gaizauskas_Robert_J., AUTH_Azzam_Saliha}
W98–0604	{AUTH_Meyers_Adam, AUTH_Macleod_Catherine, AUTH_Yangarber_Roman, AUTH_Grishman_Ralph, AUTH_Barrett_Leslie, AUTH_Reeves_Ruth}
W98–1105	{AUTH_Collins_Michael_John, AUTH_Miller_Scott}
X96–1028	{AUTH_Weischedel_Ralph_M., AUTH_Boisen_Sean, AUTH_Bikel_Daniel_M., AUTH_Bobrow_Robert_J., AUTH_Crystal_Michael, AUTH_Ferguson_William, AUTH_Wechsler_Allan, AUTH_The_PLUM_Research_Group}
X96–1058	{AUTH_Kameyama_Megumi}
X98–1013	{ …9 }
X98–1016	{AUTH_Yangarber_Roman, AUTH_Grishman_Ralph}

FIGURE 5.15 – Auteurs associés aux documents sources d'expertise obtenus sur le bimotif {{*information extraction*, publication entre 1993 et 2007}, {*information extraction*, auteurs ayant publié entre 1993 et 1999}} à l'aide d'abstraction de cœur 8-8-*hub*-autorité sur le graphe biparti publications → auteurs cités construit à partir du corpus ACL Anthology

Chapitre 6

Évaluation

Afin de valider les résultats obtenus lors du chapitre 5, nous avons réalisé une évaluation des résultats obtenus grâce aux expériences réalisées. Dans ce chapitre, nous présentons les outils et le protocole d'évaluation. À l'aide de ces derniers, nous discutons la validité des résultats obtenus lors des expériences présentées durant le chapitre 5. Nous proposons une évaluation des performances de notre méthode à l'aide d'un jeu de données d'évaluation construit à partir d'articles de revue. Nous proposons également une comparaison des performances de notre méthode avec les performances obtenues à l'aide de méthodes issues de l'état de l'art grâce au cadre d'évaluation pour les méthodes de recherche d'experts, LT ExpertFinder.

6.1 Les outils d'évaluation

Afin d'évaluer les résultats obtenus lors des expériences réalisées sur notre échantillon du corpus ACL Anthology (Gábor, Haïfa Zargayouna et al. 2016), nous utilisons différents outils d'évaluation. Parmi ceux-ci, nous utilisons LT ExpertFinder (Fischer, Remus et Biemann 2019), une cadre d'évaluation pour les méthodes de recherche d'experts. LT ExpertFinder permet de comparer les résultats obtenus à l'aide de différentes méthodes de recherche d'experts issues de l'état de l'art sur le corpus ACL Anthology. Cependant, ce cadre ne propose pas d'évaluation quantifiée des performances des différentes méthodes à ce jour. En d'autres termes, s'il est possible d'obtenir la liste des experts retrouvés à l'aide d'une méthode pour une requête donnée à l'aide de ce cadre d'évaluation, il n'est pas possible d'obtenir une quelconque valeur associée à une métrique d'évaluation (par exemple précision, rappel, f-mesure, mAP, *etc.*) pour cette même méthode. À ce jour, il n'existe pas de *gold standard* sur les experts présents dans le corpus ACL Anthology. Les résultats ne peuvent être évalués manuellement puisque le jeu de données couvre un nombre d'experts potentiels bien trop important (10724 individus au total).

Afin de remédier à cette difficulté et de proposer une évaluation des performances de notre méthode et des différentes méthodes issues de l'état de l'art sur ce corpus, nous avons exploré des méthodes alternatives pour l'évaluation des méthodes de recherche d'experts. Il existe un *benchmark* proposé par AMiner (Tang et al. 2008) qui suggère

1781 experts sur 13 thématiques [1]. Cependant, le nombre d'experts validés pour chacune des thématiques est parfois très faible et différent du nombre d'experts annoncé. Par exemple, 91 experts sont annoncés sur la thématique « information extraction », tandis que seulement 20 sont effectivement présents dans la liste à ce jour [2]. Pour cette raison, nous avons construit automatiquement des *gold standard* sur quelques-unes des thématiques de publications du corpus à partir d'articles de revue publiés après 2008, la dernière date de publication des articles de notre échantillon ACL Anthology.

6.1.1 Les *gold standards* créés à partir d'articles de revue

Pour évaluer les performances de notre méthode ainsi que celles des différentes méthodes issues de l'état de l'art sur l'échantillon du corpus ACL Anthology, nous avons construit automatiquement des *gold standards* créés à partir d'articles de revue sur les thématiques *information extraction* (présente dans 537 publications de l'échantillon du corpus, soit 4%), *word sense disambiguation* (présente dans 765 publications de l'échantillon du corpus, soit 5,7%) et *syntactic parsing* (présente dans 199 publications de l'échantillon du corpus, soit 1,5%).

Chacun des articles de revue présentés correspond à un chapitre du livre Handbook of Natural Language Processing (Indurkhya et Damerau 2010) publié en 2010 et compilant un ensemble d'articles de revue sur des thématiques du traitement automatique du langage naturel. Les articles que nous avons sélectionnés couvrent des thématiques abordées dans les publications de l'échantillon du corpus ACL Anthology. Les auteurs des publications présentes dans les références des articles de revue sont donc considérés comme des experts des thématiques couvertes par les articles de revue et forment un *gold standard* sur ces mêmes thématiques. Nous supposons que les publications présentes à la fois dans l'échantillon du corpus ACL Anthology et dans les références des articles de revue peuvent être considérées comme des publications phares et leurs auteurs comme des experts.

Dans les exemples que nous avons utilisés lors du chapitre 5 pour illustrer les experts et expertises obtenues à l'aide de notre méthode sur les différents graphes étudiés, nous avons constaté que l'utilisation d'articles de revue pouvait faciliter l'évaluation de l'expertise des individus. Cependant, cette méthode reste limitée. En effet, les articles de revue considérés sont rarement exhaustifs. Certains experts peuvent être retrouvés par le système mais ne pas être présents dans les références de l'article de revue. Nous notons qu'un biais peut être introduit lors de l'évaluation des performances. En effet, les experts obtenus à partir des *gold standard* sont des auteurs cités. Il est donc probable que les performances augmentent

1. Liste d'experts proposés par AMiner : `https://static.aminer.cn/lab-datasets/expertfinding/#new_expert_list`

2. Liste d'experts suggérée par AMiner sur « information extraction » : `https://static.aminer.cn/lab-datasets/expertfinding/datasets/Information-Extraction.txt`

lorsqu'on identifie des experts assimilés à des auteurs fortement cités.

« Information Extraction », Hobbs et Riloff, 2010

Pour évaluer la qualité des résultats obtenus sur la thématique *information extraction*, nous proposons d'utiliser l'article de revue Information Extraction de Hobbs et Riloff publié en 2010 (Hobbs et Riloff 2010). L'article de revue comporte 90 références bibliographiques, dont 47 font également partie de notre échantillon du corpus ACL Anthology d'après une identification automatique. Après vérification manuelle, nous retrouvons bien les mêmes 47 références bibliographiques à l'intersection de l'échantillon du corpus ACL Anthology et des références de l'article de revue. Nous retrouvons 97 auteurs du corpus dans les références du chapitre de livre.

« Word Sense Disambiguation », Yarowsky (2010)

Pour évaluer la qualité des résultats obtenus sur la thématique *word sense disambiguation*, nous proposons d'utiliser l'article de revue Word Sense Disambiguation de Yarowsky publié en 2010 (Yarowsky 2010). L'article de revue comporte 118 références bibliographiques, dont 46 font également partie de notre échantillon du corpus ACL Anthology d'après une identification automatique. Nous retrouvons 78 auteurs du corpus dans les références du chapitre de livre.

« Syntactic Parsing », Ljunglöf et Wirén (2010)

Pour évaluer la qualité des résultats obtenus sur la thématique *syntactic parsing*, nous proposons d'utiliser l'article de revue Syntactic Parsing de Ljunglöf et Wirén publié en 2010 (Ljunglöf et Wirén 2010). L'article de revue comporte 132 références bibliographiques, dont 36 font également partie de notre échantillon du corpus ACL Anthology d'après une identification automatique. Nous retrouvons 57 auteurs du corpus dans les références du chapitre de livre.

6.1.2 Le cadre d'évaluation comparative LT ExpertFinder

LT ExpertFinder est un cadre d'évaluation *open source*[3] des méthodes de recherche d'experts (Fischer, Remus et Biemann 2019). Il permet de comparer les résultats obtenus à l'aide de différentes méthodes issues de l'état de l'art sur la dernière version du corpus ACL Anthology[4] créée à ce jour, dont les publications s'étendent de 1965 à 2016. Dans cette version, le corpus décrit plus de 23000 publications scientifiques décrites par

3. LT ExpertFinder : `https://github.com/uhh-lt/lt-expertfinder`
4. Version de ACL Anthology employée au sein de LT ExpertFinder : `http://tangra.cs.yale.edu/newaan/`

plus de 18000 auteurs entretenant plus de 124000 relations de citation et 142000 relations de coauteurs distinctes.

LT ExpertFinder utilise également des données extraites à partir de Wikidata [5] ainsi que de Google Scholar [6] afin d'enrichir les profils d'experts. Des mots-clefs extraits des publications scientifiques sont également utilisés pour enrichir les profils d'experts à l'aide d'un outil d'extraction de mots-clefs [7].

Différentes méthodes issues de l'état de l'art sont utilisées au sein de LT Expert Finder : Model 2 (Balog, Y. Fang et al. 2012), divers algorithmes de propagation (Serdyukov et Hiemstra 2008), PageRank (Page et al. 1999) ainsi que quelques statistiques simples telles que le nombre de citations ou le h-index (Hirsch 2005). Nous rappelons succintement les définitions de ces modèles et renvoyons le lecteur au chapitre 1 pour plus de détails sur ces différentes méthodes.

Model 2 (Balog *et al.* (2012)

Model 2 (Balog, Y. Fang et al. 2012) est l'une des méthodes les plus utilisées comme méthode de référence dans le cadre de l'évaluation de méthodes de recherche d'experts (Fischer, Remus et Biemann 2019). Il s'agit d'un modèle d'apprentissage génératif centré sur les documents. Il permet d'ordonner les experts potentiels e sur la requête q en fonction de leur probabilité P(e|q). Cette dernière est estimée par refactorisation du théorème de Bayes P(e|q) = P(e|q)P(e)P(q) en P(e|q)P(e), la probabilité d'obtenir la requête q lorsque l'on a l'individu e, étant donné la probabilité P(e) que l'individu e soit un expert *a priori*.

Algorithmes de propagation

Les algorithmes de propagation sont des méthodes de recherche d'experts à base de graphe (Serdyukov et Hiemstra 2008). Au sein des graphes d'expertise, c'est-à-dire de graphes représentant des experts potentiels et des documents reliés entre eux par un lien de paternité (Serdyukov et Hiemstra 2008), ces algorithmes permettent de propager le score d'expertise d'un individu à ses coauteurs. Ils se basent sur une marche aléatoire au sein des graphes d'experts, dans lesquels les experts sont ordonnés par le nombre de visites d'un marcheur aléatoire. Différents algorithmes de propagation ont été proposé, selon la technique de marche aléatoire employée.

Dans LT ExpertFinder, les approches *k-step* et *infinite step* sont proposées. Dans l'approche *k-step*, l'algorithme réalise un nombre d'itérations k. Dans l'approche *infinite step*, le processus de marche aléatoire est infini. Cette méthode réalise des étapes de marche itératives jusqu'à ce que l'ordre de classement des experts converge. Des versions

5. Wikidata : `https://www.wikidata.org/`
6. Google Scholar : `https://scholar.google.de/`
7. LT KeyTerms : `https://github.com/uhh-lt/lt-keyterms`

pondérées de ces algorithmes ont été proposées au sein de LT ExpertFinder. Dans ces versions, des arêtes supplémentaires représentant les citations de documents ainsi que les liens de coauteurs ont été introduites. De plus, des poids sur les arêtes ont été inclus : les documents cités avec la date de publication la plus récente sont privilégiés. En effet, les auteurs supposent qu'un utilisateur préférera lire les publications les plus récentes (Fischer, Remus et Biemann 2019). De même, les liens de coauteurs sont pondérés par le nombre total de liens de coauteurs entretenus. Enfin, pour déterminer l'importance d'un lien de paternité entre un auteur et un document, les liens de paternité sont pondérés par le h-index de l'auteur.

PageRank

PageRank (Page et al. 1999) est une méthode initialement créée pour classer les pages Web par ordre décroissant de pertinence et qui a été utilisée dans le cadre de la recherche d'experts. Il s'agit d'une méthode à base de graphe qui a été utilisée pour ordonner les auteurs de publications scientifiques à l'aide de mesures quantitatives, c'est-à-dire de leur nombre de citations et du nombre de liens de coauteurs qu'ils entretiennent. Initialement, cet algorithme a été créé pour classer les pages Web par ordre décroissant de pertinence.

Statistiques simples

Des méthodes statistiques simples sont aussi implémentées dans LT ExpertFinder, telles que le nombre de citations ou le h-index (global ou local). Le h-index local correspond au h-index de l'auteur par rapport à l'ensemble des documents correspondant à une requête tandis que le h-index global de l'auteur correspond à son h-index par rapport à l'ensemble du corpus. Pour rappel, le h-index est un indice représentant à la fois le nombre de publications d'un auteur et son nombre de citations par publication.

6.2 Protocole d'évaluation

Nous proposons un protocole d'évaluation pour l'évaluation des résultats obtenus sur chacun des jeux de données étudiés. Les métriques utilisées sont la précision, le rappel et la f-mesure, largement employés dans l'état de l'art et présentés dans le chapitre 1. Les performances de notre méthode sont évaluées à l'aide des *gold standard* créés à partir des articles de revue. Enfin, une comparaison des performances de notre méthode avec les performances obtenues à l'aide des méthodes issues de l'état de l'art disponibles dans le cadre d'évaluation LT ExpertFinder est réalisée. Cette comparaison permet de situer les performances de notre méthode par rapport à celles obtenues avec les méthodes de l'état de l'art.

6.2.1 Évaluation des performances obtenues à partir du corpus ACL Anthology

Nous utilisons des *gold standard* construits automatiquement à partir d'articles de revue. Ces articles de revue sont utilisés pour évaluer les performances du système à partir de résultats obtenus sur les thématiques *information extraction*, *word sense disambiguation* et *syntactic parsing*.

Il est important de noter que nous avons évalué les performances de notre méthode sur tous les motif clos abstrait obtenus décrivant l'une des thématiques représentées par les *gold standard*, ce pour tout paramètre d'abstraction de graphe appliqué sur l'ensemble des graphes étudiés. Chaque motif clos abstrait décrivant *information extraction*, *syntactic parsing* ou *word sense disambiguation* est donc associé à une précision, un rappel et une f-mesure. Dans un souci de lisibilité, pour chaque paramètre d'abstraction appliqué, nous présentons les meilleures performances obtenues ainsi que le nombre de motifs clos abstraits évalués. La meilleure performance correspond à la meilleure f-mesure obtenue.

6.2.2 Comparaison aux méthodes issues de l'état de l'art à l'aide de LT ExpertFinder

Afin de comparer notre système à l'état de l'art, nous utilisons la plateforme LT ExpertFinder pour obtenir les experts associés à une requête, ce pour différentes méthodes de recherche d'experts.

Les experts obtenus à l'aide des méthodes de l'état de l'art permettent d'obtenir une liste d'experts triée, contrairement aux experts obtenus à l'aide de notre méthode. Pour nous permettre de comparer nos performances à celles obtenues à l'aide de méthodes ordonnées, nous nous inspirons de la méthode d'évaluation de résultats ordonnés suggérée par le *Stanford NLP Group* affilié à l'Université de Stanford [8].

Nous considérons une méthode permettant d'obtenir une liste d'experts triée par ordre décroissant de pertinence. Pour i de 1 à k, avec k le nombre d'experts obtenus avec la méthode de recherche d'experts considérée pour la requête (ou le motif) q, nous obtenons un couple (rappel, précision). Nous traçons la courbe rappel-précision associée à la méthode de recherche d'experts. Soit p_q et r_q la précision et le rappel respectivement obtenus pour la requête q à l'aide de notre méthode de recherche d'experts. Pour comparer p_q et r_q avec les performances obtenues avec une méthode de recherche d'experts issue de l'état de l'art, nous comparons p_q et r_q à p_i et r_i avec : p_i la précision et r_i le rappel obtenus sur les j premiers experts, avec $j \leq k$ et j égal au nombre d'experts obtenus à l'aide de notre méthode.

8. Évaluation de résultats ordonnés (Stanford) : `https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html`

Un autre paramétrage est nécessaire pour comparer nos résultats à ceux obtenus avec LT ExpertFinder. En effet, nous avons utilisé un échantillon du corpus ACL Anthology s'étendant de 1980 à 2008, tandis que la plateforme LT ExpertFinder exploite une version plus récente et plus riche, s'étendant de 1965 à 2016. Lors de l'écriture de requêtes sur la plateforme LT ExpertFinder, nous avons donc réduit les périodes de publication pour obtenir des experts décrits . Par exemple, si la requête est la suivante : {publication avant 2005}, alors la période temporelle que nous paramétrons dans LT ExpertFinder est : 1980 à 2004. De même, si la requête est la suivante : {publication après 1992}, alors la période temporelle paramétrée est : 1993 à 2008.

6.3 Évaluation des performances de notre méthode

Afin d'évaluer les performances de notre méthode de recherche d'experts pour chaque graphe construit dans le cadre des expériences sur le corpus ACL Anthology, nous évaluons les performances de notre méthode de recherche d'experts à l'aide des métriques décrites dans la section 6.2 et à l'aide des *gold standard* décrits dans la section 6.1.1.

Pour des raisons de lisibilité, il est difficile de présenter l'ensemble des résultats évalués. Dans la section A.3 de l'annexe, nous fournissons des tableaux décrivant le nombre de motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque graphe construit à partir du corpus ACL Anthology. De plus, nous fournissons des tableaux décrivant les meilleures performances obtenues sur les motifs décrivant au moins l'une des thématiques associées aux *gold standard* pour chaque paramètre d'abstraction testé sur les graphes construits à partir du corpus ACL Anthology.

Dans les tables 6.1, 6.2 et 6.3, nous présentons les meilleures performances obtenues sur les motifs clos contenant *information extraction*, *word sense disambiguation* et *syntactic parsing*, pour chaque paramètre d'abstraction appliqué sur tout graphe étudié. Nous précisons le paramètre d'abstraction de graphe appliqué qui a permis d'obtenir ces performances, ainsi que le nombre de motifs sélectionnés et obtenus à l'aide de ce paramètre. La meilleure performance correspond à la meilleure f-mesure obtenue. Pour ce résultat, nous décrivons également la précision et le rappel.

Graphe	Abstraction	Motifs	MP	MR	MF
Coauteurs	20-*nearstar*	3	.13	.38	.19
Citation A	-	-	-	-	-
Copublication	6-*core*	2	.19	.33	.24
	8-*nearstar*	4	.37	.18	.24
	4-*nearstar*	**12**	.27	.22	.24
Citation D	2-2-*hub*-autorité (a)	3	.40	.43	**.42**
Biparti publications-auteurs	3-*core*	3	.21	.21	.21
Biparti auteurs → publications citées	4-4-*hub*-autorité (h,a)	4	**.46**	.38	**.42**
Biparti publications → auteurs cités	5-5-*hub*-autorité (a)	6	.38	**.72**	**.42**

TABLE 6.1 – Meilleures performances obtenues sur les motifs contenant *information extraction* à partir des graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe	Abstraction	Motifs	MP	MR	MF
Coauteurs	20-*nearstar*	2	.15	.27	.19
Citation A	-	-	-	-	-
Copublication	4-*nearstar*	9	.49	.22	.30
Citation D	2-2-*hub*-autorité (h,a)	15	.32	.38	.35
Biparti publications-auteurs	4-*nearstar*	73	..60	.01	.19
Biparti auteurs → publications citées	5-5-*hub*-autorité (h)	6	.10	.29	.14
Biparti publications → auteurs cités	10-10-*hub*-autorité	1	**.84**	**.54**	**.66**

TABLE 6.2 – Meilleures performances obtenues sur les motifs contenant *word sense disambiguation* à partir des graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe	Abstraction	Motifs	MP	MR	MF
Coauteurs	6-dense	4	.14	.02	.03
Citation A	-	-	-	-	-
Copublication	4-*nearstar*	1	.03	.02	.02
Citation D	2-2-*hub*-autorité	1	0	0	0
Biparti publications-auteurs	4-*nearstar*	13	**.12**	**.07**	**.09**
Biparti auteurs → publications citées	3-3-*hub*-autorité	2	0	0	0
Biparti publications → auteurs cités	4-4-*hub*-autorité (h)	2	.07	**.07**	.07

TABLE 6.3 – Meilleures performances obtenues sur les motifs contenant *syntactic parsing* à partir des graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans les tables 6.1, 6.2 et 6.3, nous présentons les meilleures performances obtenues sur les motifs clos contenant *information extraction*, *word sense disambiguation* et *syntactic parsing*, pour chaque paramètre d'abstraction appliqué sur tout graphe étudié. Dans

les tables 6.4, 6.5 et 6.6, nous présentons les meilleures performances moyennes obtenues sur les motifs clos contenant *information extraction*, *word sense disambiguation* et *syntactic parsing*, pour chaque paramètre d'abstraction appliqué sur tout graphe étudié. Nous précisons le paramètre d'abstraction de graphe appliqué qui a permis d'obtenir ces performances, ainsi que le nombre de motifs sélectionnés et obtenus à l'aide de ce paramètre. La meilleure performance correspond à la meilleure f-mesure obtenue. Pour ce résultat, nous décrivons également la précision et le rappel.

Graphe	Abstraction	Motifs	P	R	F
Coauteurs	20-*nearstar*	3	.11	.15	.10
Citation A	-	-	-	-	-
Copublication	14-*nearstar*	2	.28	.20	.18
Citation D	2-2-*hub*-autorité (a)	3	.36	.18	.20
	2-2-*hub*-autorité (h,a)	3	.25	.23	.20
Biparti publications-auteurs	3-*core*	3	.22	.09	.11
Biparti auteurs → publications citées	4-4-*hub*-autorité (h,a)	4	**.55**	.30	.26
Biparti publications → auteurs cités	6-6-*hub*-autorité (h,a)	2	.24	**.35**	**.35**

TABLE 6.4 – Meilleures performances moyennes obtenues sur les motifs contenant *information extraction* à partir des graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe	Abstraction	Motifs	P	R	F
Coauteurs	30-*nearstar*	1	.21	.09	.13
Citation A	-	-	-	-	-
Copublication	14-*nearstar*	1	.26	.14	.18
	12-*nearstar*	2	.16	.28	.18
Citation D	3-3-*hub*-autorité (a)	1	.79	.19	.31
Biparti publications-auteurs	4-*core*	1	.50	.05	.09
Biparti auteurs → publications citées	6-6-*hub*-autorité (h)	2	.10	.13	.09
	6-6-*hub*-autorité (h,a)	2	.08	.15	.09
Biparti publications → auteurs cités	10-10-*hub*-autorité (a)	1	**.84**	**.54**	**.66**

TABLE 6.5 – Meilleures performances moyennes obtenues sur les motifs contenant *word sense disambiguation* à partir des graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe	Abstraction	Motifs	P	R	F
Coauteurs	6-dense	4	**.04**	.01	.02
Citation A	-	-	-	-	-
Copublication	4-*nearstar*	1	.03	.02	.02
Citation D	2-2-*hub*-autorité	1	0	0	0
Biparti publications-auteurs	4-*nearstar*	13	.02	.02	.02
Biparti auteurs → publications citées	3-3-*hub*-autorité	2	0	0	0
Biparti publications → auteurs cités	4-4-*hub*-autorité (h)	2	**.04**	**.04**	**.04**

TABLE 6.6 – Meilleures performances moyennes obtenues sur les motifs contenant *syntactic parsing* à partir des graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

Nous constatons que les meilleures performances (en considérant les performances dans le meilleur des cas comme les performances moyennes) sont obtenues à partir d'abstractions de graphe tirant parti de cœurs évolués, à l'aide de paramètres contraints, sur des graphes élaborés.

Sur la thématique *information extraction*, la meilleure f-mesure, 0.42, est obtenue à partir d'une abstraction de cœur ou 5-5-*hub*-autorité, 4-4-*hub*-autorité ou 2-2-*hub*-autorité en ne considérant que les autorités, sur les graphes bipartis publications → auteurs cités, auteurs → publications citées et citation D respectivement. La meilleure f-mesure moyenne, 0.35 sur 2 motifs, est obtenue à partir d'une abstraction de cœur 6-6-*hub*-autorité en considérant *hubs* et autorités sur le graphe biparti publications → auteurs cités.

Sur la thématique *word sense disambiguation*, la meilleure f-mesure, 0.66, est obtenue à partir d'une abstraction de cœur 10-10-*hub*-autorité en ne considérant que les autorités, sur le graphes bipartis publications → auteurs cités. Il s'agit de la meilleure performance que nous ayons obtenu. Il s'agit également de la meilleure f-mesure moyenne, obtenue sur un unique motif.

Nous constatons que les performances moyennes et dans le meilleur des cas sont similaires lorsque l'on considère les motifs décrivant les thématiques *information extraction* et *word sense disambiguation*. Sur les motifs décrivant la thématique *syntactic parsing*, les performances s'effondrent. En effet, sur la thématique *syntactic parsing*, la meilleure f-mesure n'est que de 0.09. Elle est obtenue à partir d'une abstraction de cœur 4-*nearstar*, sur le graphes biparti publications - auteurs. La meilleure f-mesure moyenne, 0.04 sur 2 motifs, est obtenue à partir d'une abstraction de cœur 4-4-*hub*-autorité en ne considérant que les *hubs* sur le graphe biparti publications → auteurs cités.

L'effondrement des performances de notre méthode sur cette thématique peut être expliquée par différents facteurs. Le *gold standard* peut ne pas être suffisamment exhaustif. Dans ce cas, la méthode pourrait nous permettre de retrouver des auteurs que l'on peut considérer comme experts sur la thématique *syntactic parsing* au sein du corpus ACL

Anthology mais qui n'appartiennent pas au *gold standard*. Cette hypothèse justifierait la baisse de précision, mais pas la baisse de rappel. La chute du rappel indique que notre méthode n'est pas efficace pour retrouver des experts appartenant au *gold standard* sur cette thématique. Il est peu probable que le *gold standard* indique des auteurs que l'on ne peut pas considérer comme des experts sur la thématique. Considérons le premier des 2 bimotifs que nous avons obtenu à l'aide d'une abstraction 3-3-*hub*-autorité sur le graphe biparti auteurs → publications citées et qui décrivent la thématique *syntactic parsing*. Sur cette expérience, nous avons obtenu une f-mesure de 0. Le premier bimotif est le suivant : {{*syntactic parsing*, auteurs ayant publié après 1992 et avant 2008},{*syntactic parsing*, publications datant d'avant 2005}}. Ce motif a une extension de 66 experts potentiels. Cependant, aucun des 66 experts potentiels n'est retrouvé dans les références du chapitre Syntactic Parsing du livre Handbook of Natural Language Processing. En l'absence d'autre *gold standard*, il est difficile d'estimer la qualité des experts potentiels identifiés sur ce motif, donc de réaliser une analyse permettant de comprendre cette chute de performances. Parmi les 66 experts potentiels retrouvés à l'aide de notre méthode, nous indiquons cependant que certains sont éminemment reconnus dans la communauté scientifique. Par exemple, nous retrouvons Hamish Cunningham, Christopher D. Manning, Yuji Matsumoto ou encore Ruslan Mitkov. La chute des performances semble donc plutôt liée à la qualité du *gold standard*.

Dans le meilleur des cas, les approches naïves impliquant des abstractions de graphe tirant parti de cœurs naïfs, à l'aide de paramètres relâchés, sur des graphes simples ne permettent d'obtenir que des performances moyennes. Sur des graphes simples, les meilleures performances sont obtenues à l'aide d'abstractions de graphes tirant parti de cœurs évolués, à l'aide de paramètres contraints. Cependant, les performances restent moyennes. La meilleure f-mesure moyenne associée à une abstraction du graphe de coauteurs est de 0.13. Elle est obtenue à l'aide d'une abstraction de cœur 30-*nearstar*, sur l'unique motif contenant la thématique *word sense disambiguation*. De même, la meilleure f-mesure moyenne associée à une abstraction du graphe de copublication est de 0.18. Elle est obtenue à l'aide de différents paramètres d'abstraction : par une abstraction de cœur 14-*nearstar* ainsi que par une abstraction de cœur 12-*nearstar*, toutes deux sur des motifs contenant la thématique *word sense disambiguation*.

Le graphe de citation entre auteurs est un cas particulier. Pour ce graphe, nous n'obtenons que 7 motifs clos abstraits à l'aide du paramètre d'abstraction le plus relâché que l'on peut appliquer dans un temps d'exécution raisonnable (soit presque 25 heures d'exécution sur la configuration matérielle décrite dans la section 5.2.4 du chapitre 5). Aucun des motifs clos abstraits obtenus à l'aide de ce paramètre d'abstraction ne décrit d'experts potentiels sur l'une des thématiques associée au *gold standard*. Il est donc difficile d'évaluer les performances obtenues sur ce graphe. Ces constatations tendent à démontrer que sur des graphes très denses, il est difficile d'appliquer des paramètres d'abstraction

très contraints en raison de la complexité des calculs. On obtient donc peu de motifs clos abstraits et ces motifs clos abstraits sont peu variés.

Nous constatons également que lorsque l'on applique un paramètre d'abstraction relâché, le rappel augmente. En effet, les motifs sont plus généraux. L'augmentation du rappel fait également augmenter la f-mesure, ce qui explique que la plupart des paramètres d'abstraction permettant d'obtenir les meilleures performances sont relativement peu contraints. Cependant, lorsque le paramètre d'abstraction est plus contraint, les motifs sont plus spécifiques, la validation par les pairs est plus stricte et la précision augmente.

Nous pouvons également tirer quelques conclusions concernant les hypothèses d'expertise suggérées lors du chapitre 2 grâce aux performances de notre méthode dans le meilleur des cas. Premièrement, les plus faibles performances sont obtenues à partir du graphe de coauteurs et du graphe biparti publications-auteurs. L'hypothèse d'expertise associée à ces graphes est la suivante : *si un individu rédige des publications scientifiques avec des membres variés de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine*. Il semble donc que la rédaction de publications avec des membres variés de la communauté scientifique ne soit pas un critère déterminant dans l'évaluation de l'expertise d'un individu, puisque nous n'obtenons, dans le meilleur des cas, qu'une f-mesure de 0.19 sur le graphe de coauteurs, 0.21 sur le graphe biparti publications-auteurs.

Nous obtenons des performances moyennes dans le meilleur des cas sur le graphe de copublication. L'hypothèse d'expertise associée à ce graphe est la suivante : *si un individu rédige un grand nombre de publications scientifiques sur une thématique récurrente, alors il est probable qu'il s'agisse d'un expert de son domaine*. Ce critère semble être un meilleur indicateur que le précédent, puisque nous obtenons une f-mesure de 0.30 dans le meilleur des cas.

Cependant, les graphes représentant des sommets reliés par une relation de citation sont ceux permettant d'obtenir les meilleures performances. Les meilleures performances dans le meilleur des cas sont obtenues à partir du graphe biparti publications → auteurs cités. Les hypothèses d'expertises associées au graphe biparti publications → auteurs cités sont les suivantes : *si un individu est fortement cité par les membres de la communauté scientifique, alors il est probable qu'il s'agisse d'un expert de son domaine* et *si un individu cite des auteurs appropriés du domaine dans l'une de ses publications, alors il est probable qu'il s'agisse d'un expert de ce domaine*. Dans le cas de l'application d'une abstraction de cœur *hub*-autorité, la première hypothèse d'expertise porte sur les autorités (les auteurs cités) tandis que la seconde porte sur les *hubs* (les auteurs de publications citantes).

Considérons le motif sur lequel nous obtenons la meilleure performance. Cette performance est obtenue à l'aide de l'abstraction de graphe 10-10-*hub*-autorité sur le graphe biparti publications → auteurs cités. Le bimotif associé est le suivant : q_m = {{*word sense disambiguation*},{*word sense disambiguation*, auteurs ayant publié avant 2005}}. La meilleure performance est obtenue en ne considérant que les autorités du bimotif q_m,

c'est-à-dire la partie du bimotif décrivant les auteurs cités. Il s'agit de la partie droite du bimotif. Sur la partie du bimotif concernant les autorités, nous obtenons une f-mesure de 0.66. Sur la partie du bimotif concernant les *hubs*, nous obtenons une f-mesure de 0.32, c'est-à-dire une f-mesure bien moins élevée. Il semble que la rédaction d'un article phare, c'est-à-dire d'un article fortement cité soit un critère prépondérant de l'expertise d'un individu. Cependant, la capacité à citer une littérature pertinente semble être un critère au moins aussi pertinent que la rédaction d'un grand nombre de publications sur une thématique de recherche récurrente pour déterminer l'expertise d'un individu.

6.4 Comparaison aux méthodes de l'état de l'art

Pour comparer notre méthode aux méthodes principales de l'état de l'art, nous utilisons la plateforme d'évaluation LT ExpertFinder (Fischer, Remus et Biemann 2019). Cette plateforme permet d'obtenir un ensemble d'experts pour une requête donnée et selon une méthode de l'état de l'art sélectionnée et est décrite dans la section 6.1.2.

Nous restreignons notre comparatif à un échantillon de motifs. Sur cet échantillon, nous comparons les performances obtenues à l'aide de notre méthode à celles obtenues à l'aide de Model 2, de l'algorithme de propagation *infinite-step* (*Infinite Random Walk*), de PageRank, du h-index et du nombre de citations.

Premièrement, nous sélectionnons le motif pour lequel nous obtenons les meilleures performances avec notre méthode. Il s'agit du bimotif q_m, obtenu à l'aide d'une abstraction de cœur 10-10-*hub*-autorité appliquée sur le graphe biparti publications $\rightarrow$ auteurs cités. On a q_m = {{*word sense disambiguation*},{*word sense disambiguation*, auteurs ayant publié avant 2005}}. La meilleure performance est obtenue en ne considérant que les autorités du bimotif q_m, c'est-à-dire la partie du bimotif décrivant les auteurs cités (à droite du bimotif). On note ce motif $q_m(a)$ = {*word sense disambiguation*, auteurs ayant publié avant 2005}.

Dans le protocole d'évaluation, nous avons suggéré de comparer les performances obtenues à l'aide de notre méthode et celles obtenues à l'aide des méthodes issues de l'état de l'art en utilisant les courbes rappel-précision obtenues. Pour rappel, les courbes rappel-précision sont construites pour rendre notre les principales méthodes issues de l'état de l'art comparables à notre méthode. En effet, les principales méthodes issues de l'état de l'art permettant d'obtenir une liste d'experts triée par ordre décroissant de pertinence, tandis que notre méthode ne trie pas les experts. Pour chaque méthode issue de l'état de l'art, nous considérons le couple (rappel, précision) au rang j, avec j le nombre d'experts obtenus à l'aide de notre méthode sur le motif considéré. Dans la figure 6.1, nous décrivons les courbes rappel-précision associées au motif $q_m(a)$ obtenues à l'aide de LT ExpertFinder pour chacune des méthodes issues de l'état de l'art.

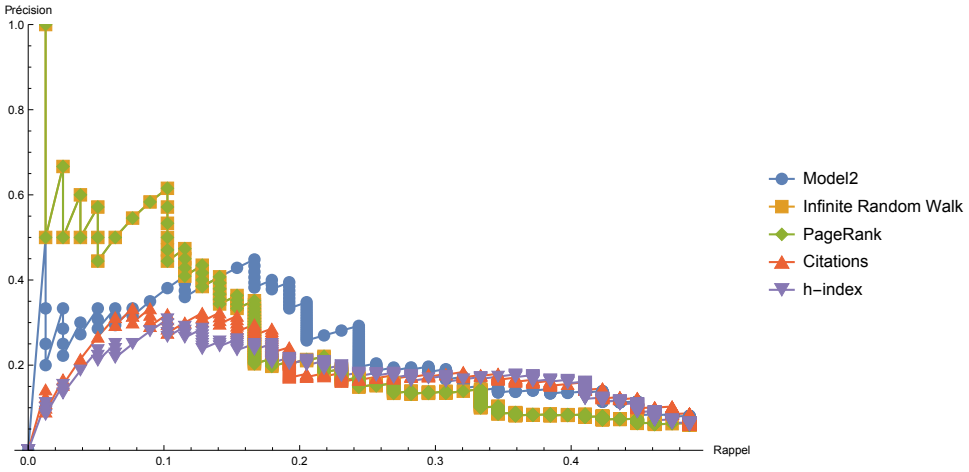


FIGURE 6.1 – Courbes rappel-précision associées au motif $q_m(a)$ obtenues à l'aide de LT ExpertFinder

Nous constatons qu'en ne considérant que les premiers experts ordonnés par ordre décroissant de pertinence, PageRank est le modèle le plus performant. Lorsque l'on considère un nombre plus important d'experts, Model2 surpasse alors PageRank.

Dans la table 6.7, nous comparons les performances obtenues sur la partie concernant les autorités du bimotif q_m, $q_m(a)$ à l'aide des différentes méthodes de recherche d'experts. Nous comparons les performances obtenues à l'aide des méthodes issues de l'état de l'art avec celles obtenues à l'aide de notre méthode. Nous considérons le couple (rappel, précision) obtenus au rang $j = |ext(q_m(a))|$. Cela signifie que l'on ne considère que les j premiers experts obtenus sur chacune des méthodes issues de l'état de l'art, avec j égal au nombre d'experts obtenus sur le motif $q_m(a)$ à l'aide de notre méthode. À l'aide de notre méthode, nous obtenons 50 experts sur ce motif. Le cadre d'évaluation LT ExpertFinder permet quant à lui d'ordonner 642 experts potentiels par ordre décroissant de niveau d'expertise sur ce motif. Chaque méthode issue de l'état de l'art permet d'obtenir un ordre différent, mais le nombre d'experts potentiels à trier est identique. Sur ce motif, nous obtenons des performances bien supérieures à celles obtenues avec les méthodes de l'état de l'art. En effet, notre précision, rappel et f-mesure sont plus de deux fois supérieures à celles obtenues avec Model 2, la plus performante des méthodes de l'état de l'art.

Méthode	Précision	Rappel	F-mesure
ScholarMap	**.84**	**.54**	**.66**
Model 2	.32	.21	.25
Infinite Random Walk	.26	.17	.20
PageRank	.26	.17	.20
h-index	.24	.15	.19
Citations	.28	.18	.22

TABLE 6.7 – Comparaison des performances obtenues sur le motif $q_m(a)$ à l'aide des méthodes de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti publications → auteurs cités construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Pour étoffer le comparatif entre les performances que nous avons obtenues et celles obtenues avec les méthodes de l'état de l'art, nous proposons un comparatif sur un ensemble de motifs pour lesquels nous avons obtenu des performances plus représentatives des performances obtenues en moyenne avec notre méthode (c'est-à-dire pour une f-mesure entre 0.3 et 0.4). Nous avons sélectionné un échantillon de motifs. Nous avons sélectionné les bimotifs décrivant la thématique *information extraction* obtenus à l'aide d'une abstraction de cœur 3-3-*hub*-autorité sur le graphe biparti auteurs → publications citées. Avec ce paramètre, nous obtenons 13 bimotifs clos abstraits Q = $\{q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10}, q_{11}, q_{12}, q_{13}\}$.

Nous décrivons les motifs appartenant à l'ensemble de motifs Q étudié :

— q_1 = {{*information extraction*, auteurs ayant publié entre 2000 et 2004}, {*information extraction*, documents publiés avant 2008}}

— q_2 = {{*information extraction*, auteurs ayant publié entre 2005 et 2007}, {{*information extraction*, documents publiés après 2004}}

— q_3 = {{*information extraction*, *languages*, auteurs ayant publié entre 1993 et 2000}, {*information extraction*, *languages*, documents publiés entre 1993 et 2007}}

— q_4 = {{*information extraction*, auteurs ayant publié entre 1993 et 2004},{*information extraction*, documents publiés entre 2000 et 2004}}

— q_5 = {{*information extraction*, auteurs ayant publié avant 2000},{*information extraction*, documents publiés avant 2000}}

— q_6 = {{*information extraction*, *learning*, auteurs ayant publié entre 2000 et 2007},{*information extraction*, *learning*, documents publiés entre 1993 et 2007}}

— q_7 = {{*information extraction*, *user information*, auteurs ayant publié entre 1993 et 2007}, {{*information extraction*, *user information*, documents publiés entre 1993 et 2007}}

— q_8 = {{*information extraction*, *languages*, *natural language processing*, auteurs ayant publié entre 2000 et 2007}, {*information extraction*, *languages*, *natural language processing*, documents publiés entre 1993 et 2007}}

— q_9 = {{*information extraction*, *conditional random field*, auteurs ayant publié entre 2005 et 2007}, {*information extraction*, *conditional random field*, documents publiés entre 2000 et 2007}}

— q_{10} = {{*information extraction*, *syntactics*, auteurs ayant publié après 2004}, {*information extraction*, *syntactics*, documents publiés entre 2000 et 2007}}

— q_{11} = {{*information extraction*, *languages*, *named entity recognition*, auteurs ayant publié entre 1993 et 2004}, {*information extraction*, *languages*, *named entity recognition*, documents publiés entre 1993 et 1999}}

— q_{12} = {{*information extraction*, *semantics*, auteurs ayant publié entre 1993 et 1999},{*information extraction*, *semantics*, documents publiés entre 1993 et 1999}}

— q_{13} = {{*information extraction*, *named entity recognition*, auteurs ayant publié entre 2000 et 2004}, {{*information extraction*, *named entity recognition*, documents publiés entre 2000 et 2004}}

Pour chaque bimotif, nous proposons une évaluation des performances obtenues en ne considérant que la partie du bimotif décrivant les *hubs* (les autorités respectivement). Lorsque l'on considère les experts obtenus à partir des *hubs* ainsi que des autorités à l'aide de notre méthode, il suffit de réaliser une simple union sur les experts obtenus en considérant isolément les deux rôles. Or, les experts étant ordonnés dans LT ExpertFinder et la version actuelle de la plateforme ne permettant pas de combiner des motifs et d'obtenir une liste d'experts triée associée à deux motifs, nous ne proposons pas d'évaluation en considérant à la fois *hubs* et autorités.

Dans la table 6.8, nous présentons le top 3 des méthodes permettant d'obtenir les meilleures performances sur les bimotifs de l'ensemble Q.

Motif	1^{er}			2^{eme}			3^{eme}		
	P	R	F	P	R	F	P	R	F
q_1	**ScholarMap (h)**			ScholarMap (a)			Model2, IRW (a)		
	.28	**.57**	**.38**	.21	.62	.31	.21	.31	.25
q_2	ScholarMap (a)			ScholarMap (h)			IRW (h)		
	.46	.31	.37	.46	.20	.28	.27	.11	.16
q_3	ScholarMap (h)			ScholarMap (a)			h-index, citation (h)		
	.38	.15	.22	.22	.20	.21	.13	.05	.07
q_4	ScholarMap (a)			ScholarMap (h)			IRW (a)		
	.43	.30	.35	.53	.21	.30	.44	.12	.19
q_5	ScholarMap (h)			ScholarMap (a)			Citations (h,a)		
	.33	.08	.13	.17	.08	.11	.21/.20	.05	.08
q_6	ScholarMap (a)			ScholarMap, Model2, IRW (h)			Model2, IRW (a)		
	.54	.21	.30	.61	.11	.19	.5	.08	.14
q_7	ScholarMap (h)			ScholarMap (a)			-		
	.33	.03	.06	0	0	0	-	-	-
q_8	ScholarMap (a)			ScholarMap (h)			IRW (h)		
	.32	.11	.17	.44	.04	.08	.22	.02	.04
q_9	ScholarMap (a)			ScholarMap, IRW (h)			PageRank (h)		
	.58	.07	.13	.50	.04	.08	.38	.08	.06
q_{10}	**ScholarMap (h)**			-			-		
	.71	.05	.1	-	-	-	-	-	-
	ScholarMap (a)			-			-		
	.27	.06	.1	-	-	-	-	-	-
q_{11}	ScholarMap (h)			IRW (h)			-		
	.57	.04	.08	.43	.03	.06	-	-	-
q_{12}	ScholarMap (a)			-			-		
	.29	.02	.04	-	-	-	-	-	-
q_{13}	IRW (h)			ScholarMap (a)			-		
	.66	.02	.04	.25	.01	.02	-	-	-

TABLE 6.8 – Top 3 des méthodes de recherche d'experts permettant d'obtenir les meilleures performances sur l'ensemble de bimotifs Q obtenus à l'aide d'abstraction de cœur 3-3-*hub*-autorité sur le graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Nous constatons que notre méthode (ScholarMap) est la plus performante sur 12 des 13 bimotifs. Sur le motif q_{13}, *Infinite Random Walk* surpasse notre méthode avec une très faible f-mesure de 0.04. Sur ce même bimotif, nous obtenons une f-mesure de 0.02 à l'aide de notre méthode. Nos performances sont donc très proches. Cependant, la précision de *Infinite Random Walk* est bien supérieure à celle de notre méthode sur ce motif. Sur les autres motifs, notre méthode permet d'obtenir une f-mesure significativement plus importante. Par exemple, nous obtenons une f-mesure de 0.37 sur le bimotif q_2 tandis que l'on obtient une f-mesure de 0.16 avec la méthode la plus performante de l'état de l'art sur ce même bimotif (*Infinite Random Walk*). Sur les motifs de l'ensemble Q, *Infinite Random Walk*, puis Model 2 sont les méthodes de l'état de l'art les plus performantes.

Sur 3 bimotifs, seule notre méthode permet d'obtenir des résultats. Dans deux cas, il s'agit de bimotifs décrivant des thématiques de recherche qui ne sont pas décrites dans la plateforme LT ExpertFinder. Ces thématiques ont été inférées à l'aide de la *Computer Science Ontology*. Par exemple, le bimotif q_7 décrit la thématique *user information*. Cette thématique de recherche n'existe pas dans la plateforme LT ExpertFinder. En effet, il s'agit d'un concept sémantique inféré à l'aide de la *Computer Science Ontology*. De ce fait, notre méthode est la seule permettant d'obtenir un ensemble d'experts décrivant cette thématique. De même, une liste d'experts associée au bimotif q_{10} n'est disponible qu'à l'aide de notre méthode et grâce à l'inférence opérée à l'aide de la *Computer Science Ontology*. Enfin, le bimotif q_{12} illustre un autre cas de figure. Pour ce bimotif, aucune méthode de l'état de l'art ne permet de retrouver des experts validés par le *gold standard*.

Pour les bimotif q_{11} et $q13$, seules notre méthode ainsi que *Infinite Random Walk* permettent d'identifier des experts.

Sur quelques bimotifs, le rappel obtenu est faible tandis que la précision est élevée. Généralement, nous obtenons une précision satisfaisante et un rappel faible. Cela signifie que nous retrouvons rarement l'ensemble des experts présents dans les références des articles de revue appropriés mais que les experts potentiels identifiés sont généralement corrects.

Enfin, les meilleures performances sont obtenues en considérant les *hubs* sur 7 bimotifs sur 13. Il semblerait donc que citer la littérature appropriée puisse finalement être un indicateur d'expertise déterminant.

6.5 Conclusion

N'ayant pu avoir accès au jeu de données du domaine adapté à la recherche d'experts dans le milieu académique et disposant d'une vérité terrain [9], nous avons proposé un nouveau protocole d'évaluation. Nous avons créé des *gold standard* à partir d'articles de revue et utilisé la plateforme d'évaluation comparative LT ExpertFinder pour proposer une évaluation des résultats obtenus à l'aide de notre méthode de recherche d'experts.

À partir d'articles de revue publiés peu de temps après la dernière publication du corpus, nous pouvons identifier des experts sur une thématique de recherche particulière. Ces experts correspondent aux auteurs des références des articles de revue et constituent un *gold standard*. Nous avons proposé trois *gold standard* sur les thématiques *information extraction*, *word sense disambiguation* et *syntactic parsing*.

La plateforme d'évaluation LT ExpertFinder nous permet de comparer nos résultats à ceux obtenus à l'aide de méthodes variées de recherche d'experts de l'état de l'art telles que des méthodes d'apprentissage supervisé basées sur un modèle génératif (Model 2), des méthodes à baxe de graphes telles que des algorithmes de propagation (Infinite Random

9. Jeu de donnée TU (Université de Tilbourg)

Walk) ou des méthodes exploitant des mesures quantitatives (PageRank) ainsi que des statistiques simples (nombre de citations, h-index).

À l'aide de métriques classiques (précision, rappel et f-mesure), nous avons proposé une évaluation des résultats obtenus pour chaque graphe pertinent pour la recherche d'experts que nous avons construit à partir du corpus ACL Anthology. Dans le meilleur des cas, les performances les plus satisfaisantes sont obtenues à partir d'abstractions de graphe tirant parti de cœurs évolués, à l'aide de paramètres contraints, sur des graphes élaborés. Sur des graphes très denses, la méthode a cependant des difficultés à passer à l'échelle en raison de la complexité des calculs.

Ensuite, nous avons comparé les résultats obtenus à l'aide de notre méthode avec les méthodes issues de l'état de l'art du domaine. Nous avons proposé cette évaluation comparative sur un échantillon de motifs clos abstraits. Nous avons proposé une comparaison sur la requête pour laquelle nous avons obtenu notre meilleure performance ainsi que sur un sous-ensemble de 26 requêtes correspondant à 13 bimotifs clos abstraits pour lesquels nous avions obtenu des performances représentatives de nos performances moyennes. Dans la quasi-totalité des cas observés, notre méthode permet d'obtenir de meilleures performances qu'avec n'importe quelle autre méthode de l'état de l'art sur ce jeu d'évaluation. L'utilisation d'une ontologie permet également de décrire un corpus de publications scientifiques avec une richesse sémantique plus importante qu'à l'aide des autres méthodes. En effet, nous avons démontré que notre méthode permet d'identifier des thématiques de recherche qui ne sont pas explicitement retrouvées dans le texte. Cette richesse sémantique est obtenue à l'aide d'inférences ontologiques.

Concernant les hypothèses d'expertise, nous concluons que les indicateurs d'expertises les plus déterminants par ordre décroissant sont les suivants : être l'auteur de publications phares (fortement citées), citer une littérature scientifique appropriée, être l'auteur de nombreuses publications sur une thématique récurrente, entretenir des liens de coauteurs avec de nombreux chercheurs du domaine.

Conclusion générale

Nous proposons une méthode permettant de découvrir des connaissances à partir de textes à l'aide de méthodes de fouille de texte. Ces connaissances sont enrichies à l'aide d'une représentation de connaissances sous forme de graphes et de l'application d'une méthode de fouille de graphe, l'abstraction de graphe. L'originalité de notre approche réside dans la combinaison de ces méthodes. Nous avons appliqué notre approche au cas d'usage de la recherche d'experts à partir de publications scientifiques pour le milieu académique.

Nous avons combiné des méthodes de fouille de texte et des méthodes de fouille de graphe afin de prendre en compte des indicateurs d'expertise liés au contenu publié par les individus ainsi qu'à leur réputation. Les publications scientifiques décrivent les thématiques de recherche abordées par les individus et recèlent également les relations de collaboration scientifique (relations de coauteurs et de citations) permettant d'évaluer la réputation des individus. Les connaissances extraites à partir des publications scientifiques à l'aide des méthodes de fouille de texte sont représentées sous forme de graphes. Ces graphes représentent les experts (les auteurs de publications scientifiques), leurs expertises associées (les thématiques de recherche qu'ils abordent) ainsi que les documents sources d'expertise (les publications scientifiques) et les liens existant entre eux. Ils correspondent à des graphes de terrain, aussi appelés grands graphes ou graphes complexes, et constituent des réseaux construits à partir de données réelles. L'abstraction de graphe permet d'explorer ces graphes. En se focalisant sur les zones denses des graphes, nous proposons de découvrir des motifs locaux permettant d'identifier des sous-graphes particuliers, les cœurs de graphe. Les sommets impliqués dans les cœurs de graphe respectent une contrainte topologique, c'est-à-dire une contrainte de connexité. Dans un graphe d'expertise, les cœurs de graphes matérialisent des experts validés par leurs pairs. Nous avons émis l'hypothèse suivante : en se focalisant sur les zones denses de graphes d'expertise obtenus à partir d'un corpus de publications scientifiques, il serait possible de détecter des individus considérés comme experts sur un ensemble de thématiques ainsi que leurs caractéristiques communes maximales.

Nous avons appliqué notre méthode sur le corpus de publications scientifiques ACL Anthology. Nous avons évalué les performances de notre méthode à l'aide d'un protocole d'évaluation original. En effet, le seul jeu de données du domaine, adapté au milieu académique et associé à une vérité terrain, TU (Université de Tilbourg), ne semble plus disponible. Ne disposant pas de *gold standard*, nous avons créé un jeu d'évaluation original. Notre protocole consiste en l'utilisation d'articles de revue en tant que *gold standard*.

En supposant que les experts d'un domaine sont présents dans les références d'articles de revue du même domaine, il est possible d'évaluer l'expertise des individus. Nous avons constaté que notre méthode permettait effectivement d'identifier des experts et leurs expertises associées. Les meilleures performances sont obtenues à partir d'abstractions de graphe tirant parti de cœurs évolués, à l'aide de paramètres contraints, sur des graphes élaborés. Concernant les hypothèses d'expertise, nous avons conclu que les indicateurs d'expertises les plus déterminants dans le milieu académique par ordre décroissant sont les suivants : être l'auteur de publications phares (fortement citées), citer une littérature scientifique appropriée, être l'auteur de nombreuses publications sur une thématique récurrente, entretenir des liens de coauteurs avec de nombreux chercheurs du domaine. Nous avons construit un graphe de connaissances fusionnant les graphes attribués représentant le corpus ACL Anthology. Pour cela, nous avons proposé une méthode permettant d'obtenir un graphe de connaissances à partir de graphes attribués, mais également d'obtenir des graphes attribués pertinents pour la recherche d'experts à partir d'un graphe de connaissances scientifique.

Nous avons comparé les performances de notre méthode à celles obtenues à l'aide des principales méthodes de l'état de l'art, à l'aide de la plateforme d'évaluation LT Expert-Finder, qui exploite également le corpus ACL Anthology. La plateforme ne permet pas d'obtenir directement les performances des méthodes mais permet d'obtenir des listes d'experts triés sur une requête, pour chaque méthode. Nous réutilisons donc notre protocole d'évaluation. Nous avons comparé nos performances à celles obtenues à l'aide de Model 2, *Infinite Random Walk*, PageRank, du h-index et du nombre de citations. Nous avons réalisé une évaluation comparative sur un échantillon de motifs clos abstraits sur lesquels nous avons obtenu des performances moyennes, ainsi que sur le motif clos abstrait sur lequel nous avons obtenu les meilleures performances. Dans la quasi-totalité des cas, notre méthode permet d'obtenir la meilleure f-mesure. Dans certains cas, notre méthode est même la seule à permettre d'obtenir des experts, puisque les motifs décrivent des thématiques de recherche inférées à l'aide de la *Computer Science Ontology*.

Les limites de notre méthode résident dans la complexité des calculs engendrée par la nature exploratoire de la méthode. En effet, contrairement aux autres méthodes de l'état de l'art qui permettent d'obtenir des experts à partir de requêtes fournies par l'utilisateur, l'énumération des motifs clos abstraits nous permet de découvrir automatiquement des ensembles de caractéristiques communes partagées par des experts. Sur des graphes très denses, la méthode a cependant des difficultés à passer à l'échelle en raison de la complexité des calculs. Il est alors difficile d'appliquer un paramètre d'abstraction de graphe contraint. Par conséquent, les motifs clos abstraits obtenus sont peu nombreux et peu variés.

Nous suggérons d'explorer de nouvelles pistes pour améliorer ces travaux. Nous avons réalisé nos expériences sur un corpus de publications scientifiques concernant les conférences ACL (*Association for Computational Linguistics*). Il est pertinent de prendre en

compte l'impact des conférences ou des revues dans lesquelles un expert potentiel publie lorsque l'on étudie un corpus de publications scientifiques issues de conférences variées. Les conférences ou les revues peuvent donc être considérées comme des attributs dans le langage de description. De même, l'affiliation d'un expert potentiel, c'est-à-dire l'université ou la structure à laquelle il est rattaché peut être un critère pertinent lors de l'évaluation de son expertise. Une stratégie d'agrégation pourrait également être mise en place. Les experts potentiels sont retrouvés en étudiant plusieurs graphes. Cependant, aucune stratégie de pondération n'est pour l'instant appliquée sur les experts, qui peuvent être retrouvés sur plusieurs graphes. Selon l'importance des indicateurs d'expertise que nous avons pu identifier, il pourrait également être utile d'appliquer une pondération. Il serait également pertinent d'explorer l'apprentissage d'ontologies à partir des graphes de co-occurrences et de citation entre expertises. Enfin, il serait pertinent de réaliser une comparaison qualitative des différentes méthodes de fouille de graphe. Par exemple, il serait pertinent de comparer l'abstraction de graphe à l'extraction d'ensembles de cliques homogènes à l'aide de contraintes (Mougel, Plantevit et al. 2011 ; Mougel, Rigotti et Gandrillon 2012) en utilisant le cas applicatif de la recherche d'experts ainsi que notre protocole d'évaluation pour estimer la qualité des communautés identifiées.

Annexe A

Annexes

A.1 Résultats obtenus sur le corpus Recherche d'Information SEmantique

Dans la table A.1, nous présentons les graphes construits dans le cadre des expériences sur le corpus Recherche d'Information SEmantique. Nous présentons le nombre de sommets, d'arêtes, le degré moyen des sommets et le nombre de descripteurs pour chaque graphe construit dans le cadre des expériences.

Graphe	Sommets	Arêtes	Degré moyen	Descripteurs
1 - Graphe de coauteurs	87	191	4,4	110
2 - Graphe de copublication	48	87	3,6	197
3 - Graphe de co-occurrences	104	273	5,25	93
4 - Graphe biparti publications - auteurs	135	138	2	110

Table A.1 – Graphes construits dans le cadre de l'expérimentation sur le corpus des ateliers Recherche d'Information SEmantique (RISE)

Graphe de coauteurs

Dans la table A.2, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe de coauteurs obtenu à partir du corpus RISE ainsi que le temps d'exécution. Les expériences réalisées sur ce graphe se sont toutes exécutées en moins d'une seconde.

Abstraction	Paramètre	Motifs
k-*core*	10	1
	8	1
	6	1
	4	7
	3	23
	2	67
k-dense	10	1
	8	1
	6	1
	4	23
	3	67
k-*nearstar*	10	3
	8	5
	6	13
	4	32
	3	56

TABLE A.2 – Paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus RISE et nombre de motifs clos abstraits obtenus

Graphe de copublication

Dans la table A.3, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe de copublication obtenu à partir du corpus RISE. Les expériences réalisées sur ce graphe se sont toutes exécutées en moins d'une seconde.

Abstraction	Paramètre	Motifs
k-*core*	10	0
	8	1
	6	3
	4	11
	3	23
	2	41
k-dense	10	1
	8	2
	6	5
	4	23
	3	41
k-*nearstar*	10	1
	8	2
	6	6
	4	15
	3	27

TABLE A.3 – Paramètres d'abstraction appliqués sur le graphe de copublication construit dans le cadre de l'expérimentation sur le corpus RISE et nombre de motifs clos abstraits obtenus

Graphe de co-occurrences

Dans la table A.4, nous indiquons le nombre de motifs clos abstraits obtenus pour chacun des paramètres d'abstraction testés sur le graphe de co-occurrences obtenu à partir du corpus RISE. Les expériences réalisées sur ce graphe se sont toutes exécutées en moins d'une seconde.

Abstraction	Paramètre	Motifs
k-*core*	10	0
	8	1
	6	7
	4	31
	3	46
	2	82
k-dense	10	0
	8	3
	6	9
	4	46
	3	82
k-*nearstar*	10	11
	8	17
	6	32
	4	64
	3	81

TABLE A.4 – Paramètres d'abstraction appliqués sur le graphe de co-occurrences construit dans le cadre de l'expérimentation sur le corpus RISE et nombre de motifs clos abstraits obtenus

Graphe biparti publications - auteurs

Dans la table A.5, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection pour chacun des paramètres d'abstraction testés sur le graphe biparti publications-auteurs obtenu à partir du corpus RISE. Les expériences réalisées sur ce graphe se sont toutes exécutées en moins d'une seconde.

Abstraction	Paramètre	Motifs
k-*core*	4	0
	3	1
	2	12
k-dense	3	0
k-*nearstar*	10	2
	8	3
	6	4
	4	22
	3	66

TABLE A.5 – Paramètres d'abstraction appliqués sur le graphe de biparti publications-auteurs construit dans le cadre de l'expérimentation sur le corpus RISE, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

A.2 Résultats obtenus sur le corpus ACL Anthology

Dans la table A.6, nous présentons les graphes construits dans le cadre des expériences sur le corpus ACL Anthology. Nous présentons le nombre de sommets, d'arêtes ou d'arc, le degré moyen des sommets et le nombre de descripteurs pour chaque graphe construit dans le cadre des expériences.

Graphe	Sommets	Arêtes/Arcs	Degré moyen	Descripteurs
1 - Graphe de coauteurs	10724	30698	5,7	2722
2 - Graphe de citation A	10724	195277	36,4	2722
3 - Graphe de copublication	13322	151904	22,8	13446
4 - Graphe de citation D	13322	54949	8,2	13446
5 - Graphe de co-occurrences	2714	103578	76,4	10732
6 - Graphe de citation E	2714	303541	223,9	10732
7 - Graphe biparti publications - auteurs	24046	32412	2,7	2722
8 - Graphe biparti auteurs → publications citées	24046	103950	8,6	2722
9 - Graphe biparti publications → auteurs cités	24046	111982	9,3	2722

TABLE A.6 – Graphes construits dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe de coauteurs

Dans la table A.7, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection paramétrée par un seuil $\beta = 0.8$, pour chacun des paramètres d'abstraction testés sur le graphe de coauteurs obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
k-*core*	15	13	2	15,38	0
	12	74	9	12,16	4
	10	544	39	7,17	26
	8	2981	104	3,49	141
	6	17875	293	1,64	640
k-dense	15	20	7	35	1
	12	221	22	9,95	13
	10	1263	61	4,83	64
	8	6833	173	2,53	230
	6	47535	543	1,14	924
k-*nearstar*	50	43	7	16,28	46
	45	90	9	10	67
	40	232	13	5,60	120
	30	1365	35	2,56	402
	20	14530	89	0,61	2258

TABLE A.7 – Paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

Dans la figure A.1, nous représentons le nombre de motifs clos abstraits sélectionnés obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe de coauteurs construit à partir du corpus ACL Anthology, pour un seuil $\beta = 0.8$. Les paramètres sont numérotés. Pour chaque abstraction de cœur, les paramètres sont numérotés par ordre d'application. Pour rappel, les paramètres les plus contraints (dont la valeur du paramètre d'abstraction est la plus élevée) sont appliqués en premier. Les valeurs des paramètres sont consultables dans la table A.7.

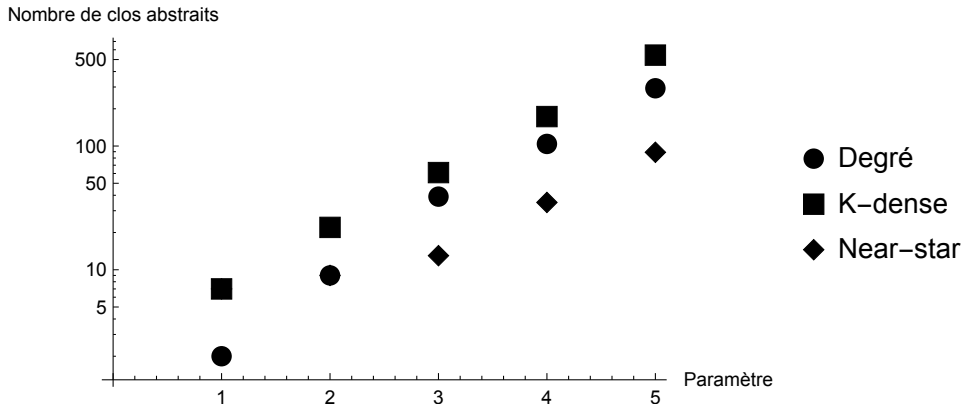


FIGURE A.1 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe de citation A

Dans la table A.8, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection paramétrée par un seuil $\beta = 0.8$, pour chacun des paramètres d'abstraction testés sur le graphe de citation A obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution. Le graphe de citation A construit à partir du corpus ACL Anthology étant très dense, des paramètres fortement contraints ont été appliqués. Le paramètre d'abstraction 25-25-*hub*-autorité s'est exécuté en 25 heures et nous a permis d'obtenir 7 motifs clos abstraits après sélection. Pour des raisons liées à la complexité des calculs, nous n'avons pas pu relâcher la contrainte au-delà de ce paramètre.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
h-a-*hub*-autorité	45 45	18	1	5,56	316
	45 30	252	1	0,4	5215
	40 30	704	1	0,14	10328
	35 40	435	1	0,23	7472
	35 30	1879	2	0,1	18320
	30 40	1333	2	0,15	15910
	30 35	2528	2	0,08	18320
	35 35	893	2	0,22	11992
	30 30	4837	3	0,06	32488
	25 25	23087	7	0,03	89373

TABLE A.8 – Paramètres d'abstraction appliqués sur le graphe de citation entre auteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

Dans la figure A.2, nous représentons le nombre de motifs clos abstraits sélectionnés obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe de citation A construit à partir du corpus ACL Anthology, pour un seuil $\beta = 0.8$. Les paramètres sont numérotés. Pour chaque abstraction de cœur, les paramètres sont numérotés par ordre d'application. Pour rappel, les paramètres les plus contraints (dont la valeur du paramètre d'abstraction est la plus élevée) sont appliqués en premier. Les valeurs des paramètres sont consultables dans la table A.8.

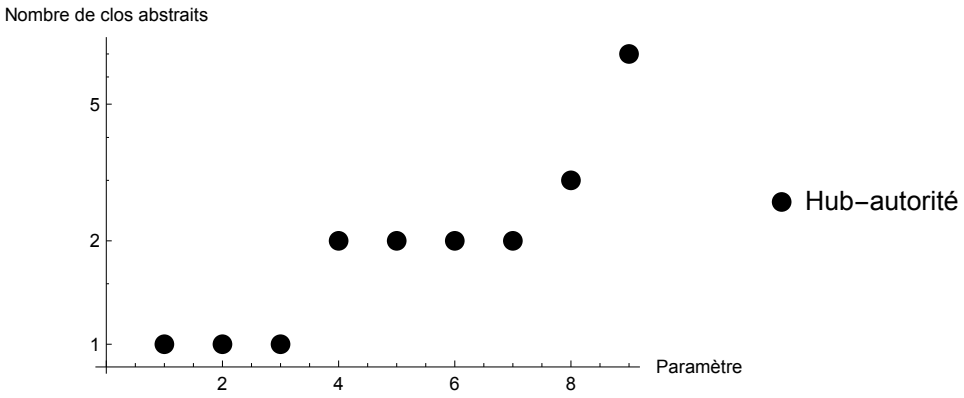


FIGURE A.2 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe de citation A construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe de copublication

Dans la table A.9, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection paramétrée par un seuil $\beta = 0.8$, pour chacun des paramètres d'abstraction testés sur le graphe de copublication obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
k-*core*	15	1720	287	16,69	2483
	12	2774	409	14,74	2918
	10	4059	541	13,32	3946
	8	6270	781	12,46	4643
	6	10393	1178	11,33	5440
k-dense	20	1010	187	18,51	1735
	18	1303	229	17,57	2232
	16	1720	287	16,69	2377
	14	2345	364	15,52	3183
	12	3336	480	14,39	3125
	10	5023	650	12,94	3790
	8	7941	926	11,66	4303
	6	13852	1520	10,97	5091
k-*nearstar*	20	1223	209	17,09	5651
	18	1525	254	16,66	5943
	16	1959	317	16,18	6144
	14	2568	386	15,03	6501
	12	3465	495	14,29	6151
	10	4962	651	13,12	6439
	8	7440	907	12,19	6248
	6	12055	1364	11,31	6471
	4	21777	2185	10,03	6222

TABLE A.9 – Paramètres d'abstraction appliqués sur le graphe de copublication construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

Dans la figure A.3, nous représentons le nombre de motifs clos abstraits sélectionnés obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe de copublication construit à partir du corpus ACL Anthology, pour un seuil $\beta = 0.8$. Les paramètres sont numérotés. Pour chaque abstraction de cœur, les paramètres sont numérotés par ordre d'application. Pour rappel, les paramètres les plus contraints (dont la valeur du paramètre d'abstraction est la plus élevée) sont appliqués en premier. Les valeurs des paramètres sont consultables dans la table A.9.

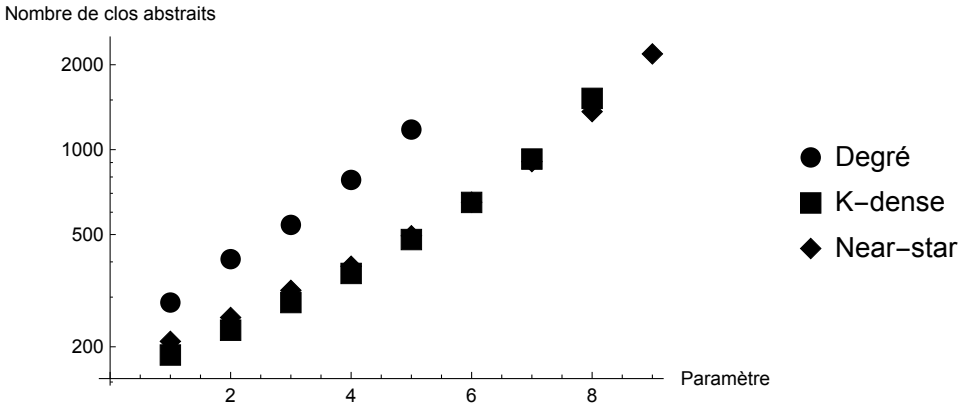


FIGURE A.3 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe de copublication construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe de citation D

Dans la table A.10, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection paramétrée par un seuil $\beta = 0.8$, pour chacun des paramètres d'abstraction testés sur le graphe de citation D obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
h-a-*hub*-autorité	8 8	7	3	42,86	8
	6 6	44	10	22,73	98
	4 4	235	48	20,43	485
	3 3	774	145	18,73	829
	2 2	3807	678	17,81	1252

TABLE A.10 – Paramètres d'abstraction appliqués sur le graphe de citation entre documents construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

Dans la figure A.4, nous représentons le nombre de motifs clos abstraits sélectionnés obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe de citation D construit à partir du corpus ACL Anthology, pour un seuil $\beta = 0.8$. Les paramètres sont numérotés. Pour chaque abstraction de cœur, les paramètres sont numérotés par ordre d'application. Pour rappel, les paramètres les plus contraints (dont la valeur du paramètre d'abstraction est la plus élevée) sont appliqués en premier. Les valeurs des paramètres sont consultables dans la table A.10.

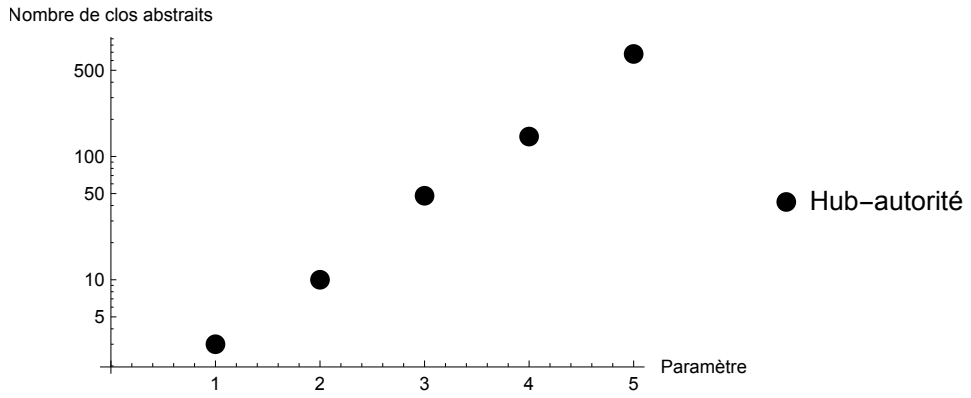


FIGURE A.4 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe de citation D construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe biparti publications - auteurs

Dans la table A.11, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection paramétrée par un seuil $\beta = 0.8$, pour chacun des paramètres d'abstraction testés sur le graphe biparti publications-auteurs obtenu à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
k-*core*	8	0	0	0	0
	6	4	1	25	0
	4	224	24	10,71	13
	3	2244	156	6,95	307
k-dense	3	0	0	0	0
k-*nearstar*	20	233	25	10,73	310
	18	351	34	9,69	528
	16	462	47	10,17	780
	14	907	65	7,17	1210
	12	1751	104	5,94	1579
	10	4465	181	4,05	2014
	8	10890	390	3,58	2593
	6	31766	937	2,95	3394
	4	150816	3414	2,26	5003

TABLE A.11 – Paramètres d'abstraction appliqués sur le graphe de biparti publications-auteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

Dans la figure A.5, nous représentons le nombre de motifs clos abstraits sélectionnés obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe biparti publications-auteurs construit à partir du corpus ACL Anthology, pour un seuil $\beta = 0.8$. Les paramètres sont numérotés. Pour chaque abstraction de cœur, les paramètres sont numérotés par ordre d'application. Pour rappel, les paramètres les plus contraints (dont

la valeur du paramètre d'abstraction est la plus élevée) sont appliqués en premier. Les valeurs des paramètres sont consultables dans la table A.11.

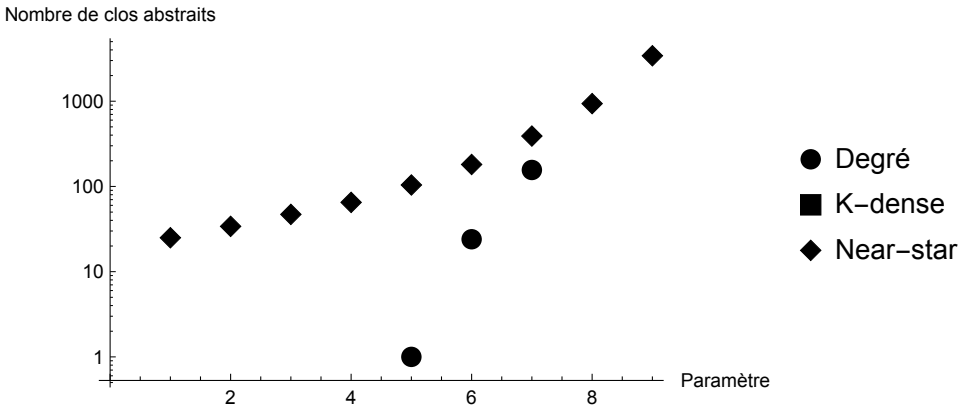


FIGURE A.5 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications-auteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe biparti auteurs → publications citées

Dans la table A.12, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection paramétrée par un seuil $\beta = 0.8$, pour chacun des paramètres d'abstraction testés sur le graphe biparti auteurs → publications citées obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
h-a-*hub*-autorité	20 20	53	1	1,89	320
	15 15	147	4	2,72	2365
	12 12	408	7	1,72	5662
	10 10	851	15	1,76	8742
	8 8	2039	33	1,62	13923
	6 6	5483	92	1,68	21026
	5 5	10090	178	1,76	26510
	4 4	22131	408	1,84	32933
	3 3	56934	1007	1,77	40175

TABLE A.12 – Paramètres d'abstraction appliqués sur le graphe biparti auteurs → publications citées entre expertises construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

Dans la figure A.6, nous représentons le nombre de motifs clos abstraits sélectionnés obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe biparti auteurs → publications citées construit à partir du corpus ACL Anthology, pour un seuil $\beta = 0.8$. Les paramètres sont numérotés. Pour chaque abstraction de cœur, les paramètres sont

numérotés par ordre d'application. Pour rappel, les paramètres les plus contraints (dont la valeur du paramètre d'abstraction est la plus élevée) sont appliqués en premier. Les valeurs des paramètres sont consultables dans la table A.12.

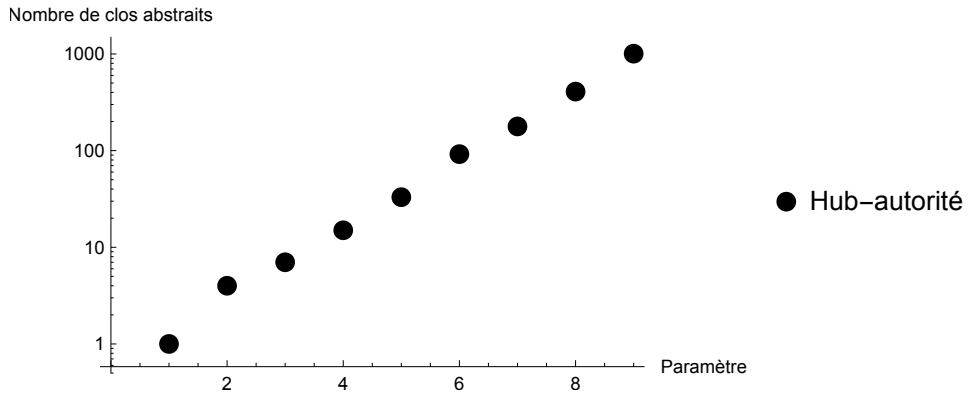


FIGURE A.6 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe biparti publications → auteurs cités

Dans la table A.13, nous indiquons le nombre de motifs clos abstraits obtenus avant et après sélection paramétrée par un seuil $\beta = 0.8$, pour chacun des paramètres d'abstraction testés sur le graphe biparti publications → auteurs cités obtenu à partir du corpus ACL Anthology ainsi que le temps d'exécution.

Abstraction	Paramètre	Motifs	Sélection	Pourcentage	Temps
h-a-*hub*-autorité	20 20	2	1	50	1
	15 15	170	6	3,53	883
	12 12	842	17	2,02	3499
	10 10	1941	34	1,75	6613
	8 8	5046	90	1,78	12243
	6 6	14376	227	1,58	21079
	5 5	25967	398	1,53	27019
	4 4	55150	734	1,33	35236

TABLE A.13 – Paramètres d'abstraction appliqués sur le graphe biparti publications → auteurs cités entre expertises construit dans le cadre de l'expérimentation sur le corpus ACL Anthology, nombre de motifs clos abstraits obtenus avant et après sélection et temps d'exécution (en secondes)

Dans la figure A.7, nous représentons le nombre de motifs clos abstraits sélectionnés obtenus pour chacun des paramètre d'abstraction appliqués sur le graphe biparti publications → auteurs cités construit à partir du corpus ACL Anthology, pour un seuil $\beta = 0.8$. Les paramètres sont numérotés. Pour chaque abstraction de cœur, les paramètres sont

numérotés par ordre d'application. Pour rappel, les paramètres les plus contraints (dont la valeur du paramètre d'abstraction est la plus élevée) sont appliqués en premier. Les valeurs des paramètres sont consultables dans la table A.13.

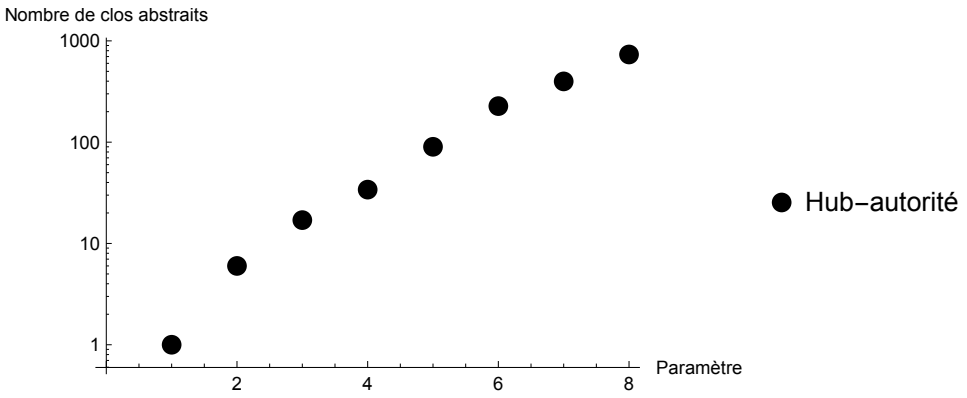


FIGURE A.7 – Nombre de motifs clos abstraits sélectionnés obtenus pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications → auteurs cités construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

A.3 Évaluation des résultats obtenus sur le corpus ACL Anthology

Nous évaluons les performances de notre méthode sur les résultats obtenus à partir du corpus ACL Anthology.

A.3.1 Graphes d'expertise

Nous évaluons les performances obtenues à partir des graphes d'expertise, c'est-à-dire du graphe de coauteurs et du graphe de citation A.

Graphe de coauteurs

Nous présentons le nombre de motifs clos abstraits obtenus à partir du graphe de coauteurs ainsi que les performances de notre méthode en moyenne et dans le meilleur des cas, pour chaque paramètre d'abstraction appliqué et pour chaque thématique associée au *gold standard*.

Nombre de motifs clos abstraits Dans la table A.14, nous décrivons le nombre de motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque paramètre d'abstraction appliqué sur le graphe de coauteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	IE	WSD	SP
k-*core*	15	2	0	0	0
	12	9	0	0	0
	10	39	2	1	0
	8	104	2	1	0
	6	293	10	6	3
k-dense	15	7	1	0	0
	12	22	1	0	0
	10	61	2	0	0
	8	173	6	3	0
	6	543	15	10	4
k*nearstar*	50	7	0	0	0
	45	9	0	0	0
	40	13	0	0	0
	30	35	2	1	0
	20	89	3	2	0

TABLE A.14 – Motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* à partir d'abstractions du graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Performances en moyenne et dans le meilleur des cas Dans la table A.15, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe de coauteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	15	0	-	-	-			
	12	0	-	-	-	-	-	-
	10	2	0	0	0	0	0	0
	8	2	.03	.0.1	.0.1	.06	.01	.02
	6	10	.07	.04	.04	.08	.21	.11
k-dense	15	1	0	0	0	0	0	0
	12	1	0	0	0	0	0	0
	10	2	.02	.02	.02	.04	.03	.04
	8	6	.04	.02	.02	.04	.08	.06
	6	15	**.14**	.05	.05	.09	.31	.14
k*nearstar*	50	0	-	-	-	-	-	-
	45	0	-	-	-	-	-	-
	40	0	-	-	-	-	-	-
	30	2	.08	.04	.05	.06	.05	.06
	20	3	.11	**.15**	**.10**	**.13**	**.38**	**.19**

TABLE A.15 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *information extraction* pour chacun des paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.16, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *word sense disambiguation*, pour chacun des paramètres d'abstraction appliqués sur le graphe de coauteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	15	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	1	0	0	0	0	0	0
	8	1	0	0	0	0	0	0
	6	6	.09	.04	.05	**.27**	.08	.12
k-dense	15	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	3	.09	.03	.05	.15	.05	.08
	6	10	.02	.02	.02	.05	.13	.07
k*nearstar*	50	0	-	-	-	-	-	-
	45	0	-	-	-	-	-	-
	40	0	-	-	-	-	-	-
	30	1	**.21**	.09	**.13**	.21	.09	.13
	20	2	.10	**.14**	.11	.15	**.27**	**.19**

TABLE A.16 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *word sense disambiguation* pour chacun des paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.17, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *syntactic parsing*, pour chacun des paramètres d'abstraction appliqués sur le graphe de coauteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	15	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	0	-	-	-	-	-	-
	6	3	.01	**.01**	.01	.02	**.02**	.02
k-dense	15	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	0	-	-	-	-	-	-
	6	4	**.04**	**.01**	**.02**	**.14**	**.02**	**.03**
k*nearstar*	50	0	-	-	-	-	-	-
	45	0	-	-	-	-	-	-
	40	0	-	-	-	-	-	-
	30	0	-	-	-	-	-	-
	20	0	-	-	-	-	-	-

TABLE A.17 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *syntactic parsing* pour chacun des paramètres d'abstraction appliqués sur le graphe de coauteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe de citation A

Dans la table A.18, nous décrivons le nombre de motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque paramètre d'abstraction appliqué sur le graphe de citation A construit à partir du corpus ACL Anthology. Nous n'avons obtenu aucun motif décrivant l'une des thématiques associées aux *gold standard*. Toute évaluation ultérieure sur ce graphe en particulier est donc compromise.

Abstraction	Paramètre	Motifs	IE	WSD	SP
h-a-*hub*-autorité	35 35	893	0	0	0
	30 30	4837	0	0	0
	25 25	23087	0	0	0

TABLE A.18 – Motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* à partir d'abstractions du graphe de citation A construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

A.3.2 Graphes décrivant les documents sources d'expertise

Nous évaluons les performances obtenues à partir des graphes décrivant les documents sources d'expertise, c'est-à-dire du graphe de copublication et du graphe de citation D.

Graphe de copublication

Nous présentons le nombre de motifs clos abstraits obtenus à partir du graphe de copublication ainsi que les performances de notre méthode en moyenne et dans le meilleur des cas, pour chaque paramètre d'abstraction appliqué et pour chaque thématique associée au *gold standard*.

Nombre de motifs clos abstraits Dans la table A.19, nous décrivons le nombre de motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque paramètre d'abstraction appliqué sur le graphe de copublication construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	IE	WSD	SP
k-*core*	15	287	0	0	0
	12	409	0	0	0
	10	541	1	0	0
	8	781	2	3	0
	6	1178	2	9	0
k-dense	20	187	0	0	0
	18	229	0	0	0
	16	287	0	0	0
	14	364	0	0	0
	12	480	1	0	0
	10	650	1	2	0
	8	926	2	3	0
	6	1520	3	9	0
k-*nearstar*	20	209	1	0	0
	18	254	1	1	0
	16	317	1	1	0
	14	386	2	1	0
	12	495	2	2	0
	10	651	3	3	0
	8	907	4	4	0
	6	1364	4	8	0
	4	2185	12	9	1

TABLE A.19 – Motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* à partir d'abstractions du graphe de copublication construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Performances en moyenne et dans le meilleur des cas Dans la table A.20, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe de copublication construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	15	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	1	.20	.02	.04	.20	.02	.04
	8	2	.19	.13	.12	.18	.25	.21
	6	2	.20	.18	.14	.19	.33	**.24**
k-dense	20	0	-	-	-	-	-	-
	18	0	-	-	-	-	-	-
	16	0	-	-	-	-	-	-
	14	0	-	-	-	-	-	-
	12	1	.20	.02	.04	.20	.02	.04
	10	1	.20	.02	.04	.20	.02	.04
	8	2	.19	.16	.13	.19	.30	.23
	6	3	.27	.16	.13	.16	.38	.22
k-*nearstar*	20	1	.16	.14	.15	.16	.14	.15
	18	1	.16	.14	.15	.16	.14	.15
	16	1	.11	.19	.14	.11	.19	.14
	14	2	**.28**	**.20**	**.18**	.15	.31	.20
	12	2	.27	**.20**	.17	.14	.32	.19
	10	3	.24	.15	.12	.12	.33	.18
	8	4	.19	.15	.13	**.37**	.18	**.24**
	6	4	.22	.19	.14	.12	**.47**	.20
	4	12	.20	.11	.10	.27	.22	**.24**

TABLE A.20 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *information extraction* pour chacun des paramètres d'abstractions appliqués sur le graphe de copublication construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.21, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *word sense disambiguation*, pour chacun des paramètres d'abstraction appliqués sur le graphe de copublication construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	15	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	3	.25	.17	.15	.21	.40	.28
	6	9	**.29**	.12	.13	.21	.49	.29
k-dense	20	0	-	-	-	-	-	-
	18	0	-	-	-	-	-	-
	16	0	-	-	-	-	-	-
	14	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	2	.20	.20	.17	.22	.36	.27
	8	3	.24	.17	.15	.20	.41	.27
	6	9	.26	.11	.11	.19	.54	.28
k-*nearstar*	20	0	-	-	-	-	-	-
	18	1	.24	.10	.14	.24	.10	.14
	16	1	.24	.10	.14	.24	.10	.14
	14	1	.26	.14	**.18**	.26	.14	.18
	12	2	.16	**.28**	**.18**	.19	.47	.27
	10	3	.17	.24	.16	.18	.55	.27
	8	4	.23	.19	.14	.17	.58	.26
	6	8	.25	.12	.09	.14	**.60**	.23
	4	9	.26	.14	.12	**.49**	.22	**.30**

TABLE A.21 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *word sense disambiguation* pour chacun des paramètres d'abstractions appliqués sur le graphe de copublication construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.22, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *syntactic parsing*, pour chacun des paramètres d'abstraction appliqués sur le graphe de copublication construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	15	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	0	-	-	-	-	-	-
	6	0	-	-	-	-	-	-
k-dense	20	0	-	-	-	-	-	-
	18	0	-	-	-	-	-	-
	16	0	-	-	-	-	-	-
	14	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	0	-	-	-	-	-	-
	6	0	-	-	-	-	-	-
k-*nearstar*	20	0	-	-	-	-	-	-
	18	0	-	-	-	-	-	-
	16	0	-	-	-	-	-	-
	14	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	0	-	-	-	-	-	-
	6	0	-	-	-	-	-	-
	4	1	**.03**	**.02**	**.02**	**.03**	**.02**	**.02**

TABLE A.22 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *syntactic parsing* pour chacun des paramètres d'abstractions appliqués sur le graphe de copublication construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe de citation D

Nous présentons le nombre de motifs clos abstraits obtenus à partir du graphe de citation D ainsi que les performances de notre méthode en moyenne et dans le meilleur des cas, pour chaque paramètre d'abstraction appliqué et pour chaque thématique associée au *gold standard*.

Nombre de motifs clos abstraits Dans la table A.23, nous décrivons le nombre de motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque paramètre d'abstraction appliqué sur le graphe de citation D construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	IE	WSD	SP
h-a-*hub*-autorité	8 8	3	0	0	0
	6 6	10	0	0	0
	4 4	48	0	0	0
	3 3	145	1	1	0
	2 2	678	3	15	1

TABLE A.23 – Motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* à partir d'abstractions du graphe de citation D construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Performances en moyenne et dans le meilleur des cas Dans la table A.24, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe de citation D construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
Hubs								
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	1	.40	.04	.07	.40	.04	.07
	2 2	3	.18	.16	.15	.27	.40	.32
Autorités								
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	1	1	.08	.15	1	.08	.15
	2 2	3	.36	.18	**.20**	.40	.43	**.42**
Hubs et autorités								
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	1	**.63**	.10	.18	**.63**	.10	.18
	2 2	3	.25	**.23**	**.20**	.26	**.54**	.35

TABLE A.24 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *information extraction* pour chacun des paramètres d'abstraction appliqués sur le graphe de citation D construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.25, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *word sense disambiguation*, pour chacun des paramètres d'abstraction appliqués sur le graphe de citation D construit à partir du

corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
		Hubs						
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	1	.26	.14	.18	.26	.14	.18
	2 2	15	.37	.13	.13	.36	.31	.33
		Autorités						
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	1	**.79**	.19	**.31**	**.79**	.19	.31
	2 2	15	.57	.15	.18	.23	**.68**	.34
		Hubs et autorités						
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	1	.34	**.23**	.27	.34	.23	.27
	2 2	15	.40	.18	.18	.32	.38	**.35**

TABLE A.25 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *word sense disambiguation* pour chacun des paramètres d'abstraction appliqués sur le graphe de citation D construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.26, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *syntactic parsing*, pour chacun des paramètres d'abstraction appliqués sur le graphe de citation D construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
Hubs								
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	0	-	-	-	-	-	-
	2 2	1	.	.	.			
Autorités								
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	0	-	-	-	-	-	-
	2 2	1	.	.	.			
Hubs et autorités								
h-a-*hub*-autorité	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	4 4	0	-	-	-	-	-	-
	3 3	0	-	-	-	-	-	-
	2 2	1	0	0	0	0	0	0

TABLE A.26 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *syntactic parsing* pour chacun des paramètres d'abstraction appliqués sur le graphe de citation D construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

A.3.3 Graphes bipartis

Nous évaluons les performances obtenues à partir des graphes bipartis, c'est-à-dire du graphe biparti publications-auteurs, du graphe biparti auteurs → publications citées et du graphe biparti publications → auteurs cités.

Graphe biparti publications-auteurs

Nous présentons le nombre de bimotifs clos abstraits obtenus à partir du graphe biparti publications-auteurs ainsi que les performances de notre méthode en moyenne et dans le meilleur des cas, pour chaque paramètre d'abstraction appliqué et pour chaque thématique associée au *gold standard*.

Nombre de bimotifs clos abstraits Dans la table A.27, nous décrivons le nombre de bimotifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications-auteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Biotifs	IE	WSD	SP
k-*core*	8	0	0	0	0
	6	0	0	0	0
	4	24	0	1	0
	3	156	3	3	0
k-dense	3	0	0	0	0
k-*nearstar*	20	25	1	0	0
	18	34	1	0	0
	16	47	1	0	0
	14	65	1	0	0
	12	104	3	1	0
	10	181	4	2	0
	8	390	12	4	0
	6	937	28	15	2
	4	3414	86	73	13

TABLE A.27 – Biotifs clos abstraits obtenus sur les thématiques associées aux *gold standard* à partir d'abstractions du graphe biparti publications-auteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Performances en moyenne et dans le meilleur des cas Dans la table A.36, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les bimotifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe biparti publications-auteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Biotifs	P	R	F	MP	MR	MF
k-*core*	8	0	-	-	-	-	-	-
	6	0	-	-	-	-	-	-
	4	0	-	-	-	-	-	-
	3	3	**.22**	**.09**	**.11**	.21	**.21**	**.21**
k-dense	3	0	-	-	-	-	-	-
k-*nearstar*	20	1	.04	.01	.02	.04	.01	.02
	18	1	.04	.01	.02	.04	.01	.02
	16	1	.04	.01	.02	.04	.01	.02
	14	1	.02	.01	.01	.02	.01	.01
	12	3	.03	.07	1	.07	.01	.02
	10	4	.04	.02	.02	.04	.06	.05
	8	12	.04	.02	.02	.04	.12	.06
	6	28	.08	.03	.04	.21	.06	.10
	4	86	.11	.05	.05	**.30**	.13	.19

TABLE A.28 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *information extraction* pour chacun des paramètres d'abstraction appliqués sur le graphe biparti publications-auteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.37, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les bimotifs contenant *word sense disambiguation*, pour chacun des paramètres d'abstraction appliqués sur le graphe biparti publications-auteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	8	0	-	-	-	-	-	-
	6	0	-	-	-	-	-	-
	4	1	**.50**	.05	**.09**	.50	.05	.09
	3	3	.18	.03	.04	.21	.05	.08
k-dense	3	0	-	-	-	-	-	-
k-*nearstar*	20	0	-	-	-	-	-	-
	18	0	-	-	-	-	-	-
	16	0	-	-	-	-	-	-
	14	0	-	-	-	-	-	-
	12	1	.06	**.09**	.07	.06	.09	.07
	10	2	.07	.06	.05	.05	.12	.07
	8	4	.06	.06	.04	.05	.17	.08
	6	15	.07	.04	.04	.05	**.29**	.09
	4	73	.11	.05	.05	**.60**	.01	**.19**

TABLE A.29 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *word sense disambiguation* pour chacun des paramètres d'abstraction appliqués sur le graphe biparti publications-auteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.38, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les bimotifs contenant *syntactic parsing*, pour chacun des paramètres d'abstraction appliqués sur le graphe biparti publications-auteurs construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
k-*core*	8	0	-	-	-	-	-	-
	6	0	-	-	-	-	-	-
	4	0	-	-	-	-	-	-
	3	0	-	-	-	-	-	-
k-dense	3	0	-	-	-	-	-	-
k-*nearstar*	20	0	-	-	-	-	-	-
	18	0	-	-	-	-	-	-
	16	0	-	-	-	-	-	-
	14	0	-	-	-	-	-	-
	12	0	-	-	-	-	-	-
	10	0	-	-	-	-	-	-
	8	0	-	-	-	-	-	-
	6	2	.01	**.02**	.01	.02	.02	.03
	4	13	**.02**	**.02**	**.02**	**.12**	**.07**	**.09**

TABLE A.30 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *syntactic parsing* pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications-auteurs construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe bipartis auteurs → publications citées

Nous présentons le nombre de bimotifs clos abstraits obtenus à partir du graphe biparti auteurs → publications citées ainsi que les performances de notre méthode en moyenne et dans le meilleur des cas, pour chaque paramètre d'abstraction appliqué et pour chaque thématique associée au *gold standard*.

Nombre de motifs clos abstraits Dans la table A.31, nous décrivons le nombre de bimotifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque paramètre d'abstraction appliqué sur le graphe biparti auteurs → publications citées construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	IE	WSD	SP
h-a-*hub*-autorité	20 20	1	0	0	0
	15 15	4	0	0	0
	12 12	7	0	0	0
	10 10	15	0	0	0
	8 8	33	0	2	0
	6 6	92	0	2	0
	5 5	178	3	6	0
	4 4	408	4	13	1
	3 3	1007	13	35	2

TABLE A.31 – Biotifs clos abstraits obtenus sur les thématiques associées aux *gold standard* à partir d'abstractions du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Performances en moyenne et dans le meilleur des cas Dans la table A.32, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe biparti auteurs → publications citées construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
		Hubs						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	3	.44	.01	.02	.39	.04	.06
	4 4	4	**.66**	.19	.20	.33	.52	.40
	3 3	13	.40	.12	.14	.28	**.57**	.38
		Autorités						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	3	.34	.06	.10	.37	.12	.18
	4 4	4	**.66**	.25	.23	**.56**	.31	.37
	3 3	13	.32	.16	.17	.46	.31	.37
		Hubs et autorités						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	3	.33	.06	.10	.33	.12	.18
	4 4	4	.55	**.30**	**.26**	.46	.38	**.42**
	3 3	13	.34	.20	.20	.40	.38	.39

TABLE A.32 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *information extraction* pour chaque paramètre d'abstraction appliqué sur le graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.33, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe biparti auteurs → publications citées construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
			Hubs					
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	2	.11	.05	.07	.07	.07	.07
	6 6	2	.10	.13	**.09**	.09	.23	.13
	5 5	6	.11	.07	.06	.10	.29	**.14**
	4 4	13	.09	.05	.04	.08	.30	.12
	3 3	35	.09	.05	.04	.08	.30	.12
			Autorités					
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	2	.16	.06	.07	.07	.08	.08
	6 6	2	.09	.10	.08	.07	.15	.10
	5 5	6	**.12**	.06	.06	.07	.20	.10
	4 4	13	**.12**	.05	.04	.08	.28	.12
	3 3	35	**.12**	.05	.04	.08	.28	.12
			Hubs et autorités					
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	2	**.12**	.08	.08	**.17**	.05	.08
	6 6	2	.08	**.15**	**.09**	.08	.26	.12
	5 5	6	.11	.10	.08	.08	.32	.12
	4 4	13	.09	.06	.05	.11	**.12**	.12
	3 3	35	.09	.06	.05	.11	.12	.12

TABLE A.33 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *word sense disambiguation* pour chaque paramètre d'abstraction appliqué sur le graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.34, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe biparti auteurs → publications citées construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
Hubs								
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	0	-	-	-	-	-	-
	4 4	4	0	0	0	0	0	0
	3 3	2	0	0	0	0	0	0
Autorités								
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	0	-	-	-	-	-	-
	4 4	4	0	0	0	0	0	0
	3 3	2	0	0	0	0	0	0
Hubs et autorités								
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	0	-	-	-	-	-	-
	4 4	4	0	0	0	0	0	0
	3 3	2	0	0	0	0	0	0

TABLE A.34 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *syntactic parsing* pour chaque paramètre d'abstraction appliqué sur le graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Graphe biparti publications → auteurs cités

Nous présentons le nombre de motifs clos abstraits obtenus à partir du graphe biparti publications → auteurs cités ainsi que les performances de notre méthode en moyenne et dans le meilleur des cas, pour chaque paramètre d'abstraction appliqué et pour chaque thématique associée au *gold standard*.

Nombre de motifs clos abstraits Dans la table A.35, nous décrivons le nombre de motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications → auteurs cités construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	IE	WSD	SP
h-a-*hub*-autorité	20 20	1	0	0	0
	15 15	6	0	0	0
	12 12	17	0	0	0
	10 10	34	0	1	0
	8 8	90	1	3	0
	6 6	227	2	7	0
	5 5	398	7	15	0
	4 4	734	11	20	2

TABLE A.35 – Motifs clos abstraits obtenus sur les thématiques associées aux *gold standard* à partir d'abstractions du graphe biparti publications → auteurs cités construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Performances en moyenne et dans le meilleur des cas Dans la table A.36, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe de biparti publications → auteurs cités construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
		Hubs						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	1	.31	.15	.21	31	.15	.21
	6 6	2	.24	**.35**	.24	.22	.58	.32
	5 5	7	.41	.23	.21	.21	**.71**	.32
	4 4	11	.27	.19	.18	.30	.35	.32
		Autorités						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	1	.33	.07	.12	.33	.07	.12
	6 6	2	.36	.21	.23	.40	.36	.38
	5 5	7	**.49**	.14	.18	.38	**.71**	**.42**
	4 4	11	.45	.19	.16	**.63**	.23	.33
		Hubs et autorités						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	1	.32	.20	.24	.32	.20	.24
	6 6	2	.24	**.35**	**.25**	.22	.61	.32
	5 5	7	.41	.26	.23	.45	.26	.33
	4 4	11	.28	.22	.20	.33	.44	.38

TABLE A.36 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *information extraction* pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications → auteurs cités construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.36, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe de biparti publications → auteurs cités construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
			Hubs					
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	1	.29	.36	.32	.29	.36	.32
	8 8	3	.40	.34	.28	.42	.29	.35
	6 6	7	.34	.25	.21	.29	.36	.32
	5 5	15	.24	.17	.15	.23	.36	.28
	4 4	20	.27	.17	.16	.44	.23	.30
			Autorités					
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	1	**.84**	.54	**.66**	**.84**	.54	**.66**
	8 8	3	.72	.33	.36	.47	**.67**	.55
	6 6	7	.77	.26	.31	.60	.35	.44
	5 5	15	.68	.18	.24	.56	40.	.47
	4 4	20	.64	.18	.22	.45	.32	.42
			Hubs et autorités					
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	1	.42	**.64**	.51	.42	.64	.51
	8 8	3	.44	.46	.35	.44	.40	.42
	6 6	7	.42	.36	.30	.40	.41	.40
	5 5	15	.34	.25	.23	.41	.36	.38
	4 4	20	.35	.24	.22	.49	.32	.39

TABLE A.37 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *word sense disambiguation* pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications → auteurs cités construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Dans la table A.36, nous décrivons les performances moyennes ainsi que les meilleures performances obtenues sur les motifs contenant *information extraction*, pour chacun des paramètres d'abstraction appliqués sur le graphe de biparti publications → auteurs cités construit à partir du corpus ACL Anthology.

Abstraction	Paramètre	Motifs	P	R	F	MP	MR	MF
		Hubs						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	0	-	-	-	-	-	-
	4 4	2	**.04**	**.04**	**.04**	**.07**	**.07**	**.07**
		Autorités						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	0	-	-	-	-	-	-
	4 4	2	.06	.04	.05	.13	.07	.09
		Hubs et autorités						
h-a-*hub*-autorité	20 20	0	-	-	-	-	-	-
	15 15	0	-	-	-	-	-	-
	12 12	0	-	-	-	-	-	-
	10 10	0	-	-	-	-	-	-
	8 8	0	-	-	-	-	-	-
	6 6	0	-	-	-	-	-	-
	5 5	0	-	-	-	-	-	-
	4 4	2	.03	.04	.03	.06	.07	.07

TABLE A.38 – Performances moyennes (f-mesure F, couple précision-rappel P-R associé) et meilleures performances (meilleure f-mesure MF, couple précision-rappel MP-MR associé) obtenues sur les motifs contenant *syntactic parsing* pour chaque paramètre d'abstraction appliqué sur le graphe biparti publications → auteurs cités construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

A.4 Comparaison des résultats obtenus sur le corpus ACL Anthology avec ceux obtenus à l'aide des méthodes de l'état de l'art

Pour étoffer le comparatif des résultats obtenus à l'aide de notre méthode avec ceux obtenus à l'aide des méthodes de l'état de l'art, nous avons sélectionné les bimotifs décrivant la thématique *information extraction* obtenus à l'aide d'une abstraction de cœur 3-3-*hub*-autorité sur le graphe biparti auteurs → publications citées. Avec ce paramètre,

nous obtenons 13 bimotifs clos abstraits Q = {q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10}, q_{11}, q_{12}, q_{13}}.

On a q_1 = {{*information extraction*, auteurs ayant publié entre 2000 et 2004}, {*information extraction*, documents publiés avant 2008}}. Dans la table A.39, nous décrivons les performances obtenues sur le bimotif q_1 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	.28	**.57**	**.38**
Model 2	**.47**	.23	.15
Infinite Random Walk	.14	.28	.19
PageRank	.10	.20	.13
h-index	.13	.26	.17
Citations	.12	.24	.16
	Autorités		
ScholarMap	**.21**	**.62**	**.31**
Model 2	**.21**	.31	.25
Infinite Random Walk	**.21**	.31	.25
PageRank	.10	.14	.12
h-index	.14	.21	.17
Citations	.14	.20	.16

TABLE A.39 – Comparaison des performances obtenues sur le motif q_1 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_2 = {{*information extraction*, auteurs ayant publié entre 2005 et 2007}, {{*information extraction*, documents publiés après 2004}}. Dans la table A.40, nous décrivons les performances obtenues sur le bimotif q_2 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**.46**	**.2**	**.28**
Model 2	.34	.14	.2
Infinite Random Walk	.27	.11	.16
PageRank	.22	.09	.13
h-index	.22	.09	.13
Citations	.2	.08	.12
	Autorités		
ScholarMap	**.46**	**.31**	**.37**
Model 2	.38	.09	.15
Infinite Random Walk	.38	.09	.15
PageRank	.25	.06	.1
h-index	.21	.05	.08
Citations	.17	.04	.06

TABLE A.40 – Comparaison des performances obtenues sur le motif q_2 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs $\rightarrow$ publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a $q_3 = \{\{$*information extraction*, *languages*, auteurs ayant publié entre 1993 et 2000$\}$, $\{$*information extraction*, *languages*, documents publiés entre 1993 et 2007$\}\}$. Dans la table A.41, nous décrivons les performances obtenues sur le motif q_3 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**.38**	**.15**	**.22**
Model 2	.05	.02	.03
Infinite Random Walk	.05	.02	.03
PageRank	.1	.04	.06
h-index	.13	.05	.07
Citations	.13	.05	.07
	Autorités		
ScholarMap	**.22**	**.2**	**.21**
Model 2	.05	.02	.03
Infinite Random Walk	.07	.02	.03
PageRank	.1	.03	.05
h-index	.13	.04	.06
Citations	.1	.03	.05

TABLE A.41 – Comparaison des performances obtenues sur le motif q_3 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_4 = {{*information extraction*, auteurs ayant publié entre 1993 et 2004},{*information extraction*, documents publiés entre 2000 et 2004}}. Dans la table A.42, nous décrivons les performances obtenues sur le bimotif q_4 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**.53**	**.21**	**.30**
Model 2	.24	.09	.13
Infinite Random Walk	.29	.11	.16
PageRank	.24	.09	.13
h-index	.24	.09	.13
Citations	.18	.07	.1
	Autorités		
ScholarMap	.43	**.3**	**.35**
Model 2	.3	.08	.13
Infinite Random Walk	**.44**	.12	.19
PageRank	.19	.05	.08
h-index	.26	.07	.11
Citations	.22	.06	.1

TABLE A.42 – Comparaison des performances obtenues sur le motif q_4 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs $\rightarrow$ publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a $q_5 = \{\{$*information extraction*, auteurs ayant publié avant 2000$\}$,$\{$*information extraction*, documents publiés avant 2000$\}\}$. Dans la table A.43, nous décrivons les performances obtenues sur le bimotif q_5 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**.33**	**.08**	**.13**
Model 2	.17	.04	.06
Infinite Random Walk	.25	.06	.1
PageRank	.25	.06	.1
h-index	.25	.06	.1
Citations	.21	.05	.08
	Autorités		
ScholarMap	.17	**.08**	**.11**
Model 2	.16	.04	.07
Infinite Random Walk	**.24**	.06	.1
PageRank	**.24**	.06	.1
h-index	**.24**	.06	.1
Citations	.2	.05	.08

TABLE A.43 – Comparaison des performances obtenues sur le motif q_5 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_6 = {{*information extraction*, *learning*, auteurs ayant publié entre 2000 et 2007},{*information extraction*, *learning*, documents publiés entre 1993 et 2007}}. Dans la table A.44, nous décrivons les performances obtenues sur le bimotif q_6 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**.61**	**.11**	**.19**
Model 2	**.61**	**.11**	**.19**
Infinite Random Walk	**.61**	**.11**	**.19**
PageRank	.22	.04	.07
h-index	.17	.03	.05
Citations	.16	.03	.05
	Autorités		
ScholarMap	**.54**	**.21**	**.3**
Model 2	.5	.08	.14
Infinite Random Walk	.5	.08	.14
PageRank	.38	.06	.11
h-index	.13	.02	.04
Citations	.19	.03	.05

TABLE A.44 – Comparaison des performances obtenues sur le motif q_6 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_7 = {{*information extraction*, *user information*, auteurs ayant publié entre 1993 et 2007}, {{*information extraction*, *user information*, documents publiés entre 1993 et 2007}}. Dans la table **??**, nous décrivons les performances obtenues sur le bimotif q_7 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode. Le concept sémantique *user information* appartenant au bimotif q_7 n'existe pas dans la plateforme LT ExpertFinder. En effet, il s'agit d'un concept inféré à l'aide de la *Computer Science Ontology*. De ce fait, notre méthode est la seule permettant d'obtenir un ensemble d'experts décrivant ce concept.

Méthode	Précision	Rappel	F-mesure
Hubs			
ScholarMap	**.33**	**.03**	**.06**
Model 2	-	-	-
Infinite Random Walk	-	-	-
PageRank	-	-	-
h-index	-	-	-
Citations	-	-	-
Autorités			
ScholarMap	**0**	**0**	**0**
Model 2	-	-	-
Infinite Random Walk	-	-	-
PageRank	-	-	-
h-index	-	-	-
Citations	-	-	-

TABLE A.45 – Comparaison des performances obtenues sur le motif q_7 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_8 = {{*information extraction*, *languages*, *natural language processing*, auteurs ayant publié entre 2000 et 2007}, {*information extraction*, *languages*, *natural language processing*, documents publiés entre 1993 et 2007}}. Dans la table A.46, nous décrivons les performances obtenues sur le bimotif q_8 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**.44**	**.04**	**.08**
Model 2	.11	.01	.02
Infinite Random Walk	.22	.02	.04
PageRank	.11	.01	.02
h-index	.11	.01	.02
Citations	0	0	0
	Autorités		
ScholarMap	**.32**	**.11**	**.17**
Model 2	.13	.01	.02
Infinite Random Walk	.13	.01	.02
PageRank	.13	.01	.02
h-index	.13	.01	.02
Citations	0	0	0

TABLE A.46 – Comparaison des performances obtenues sur le motif q_8 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs $\rightarrow$ publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a $q_9 = \{\{$*information extraction*, *conditional random field*, auteurs ayant publié entre 2005 et 2007$\}$, $\{$*information extraction*, *conditional random field*, documents publiés entre 2000 et 2007$\}\}$. Dans la table A.47, nous décrivons les performances obtenues sur le bimotif q_8 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	.5	**.04**	**.08**
Model 2	**.63**	.05	.1
Infinite Random Walk	.5	.04	.08
PageRank	.38	.08	.06
h-index	.13	.01	.02
Citations	.25	.02	.04
	Autorités		
ScholarMap	.58	.07	**.13**
Model 2	**.83**	.05	.1
Infinite Random Walk	.33	.02	.04
PageRank	.17	.01	.02
h-index	.17	.01	.02
Citations	.33	.02	.04

TABLE A.47 – Comparaison des performances obtenues sur le motif q_9 à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_{10} = {{*information extraction*, *syntactics*, auteurs ayant publié après 2004}, {*information extraction*, *syntactics*, documents publiés entre 2000 et 2007}}. Dans la table A.48, nous décrivons les performances obtenues sur le bimotif q_{10} à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode. Une liste d'experts associée au bimotif q_{10} n'est disponible qu'à l'aide de notre méthode et grâce à l'inférence opérée à l'aide de la *Computer Science Ontology*. Nous obtenons une bonne précision et un très faible rappel, donc une très faible f-mesure car le nombre d'experts étiqueté est faible (mais précis).

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**.71**	**.05**	**.1**
Model 2	-	-	-
Infinite Random Walk	-	-	-
PageRank	-	-	-
h-index	-	-	-
Citations	-	-	-
	Autorités		
ScholarMap	**.27**	**.06**	**.1**
Model 2	-	-	-
Infinite Random Walk	-	-	-
PageRank	-	-	-
h-index	-	-	-
Citations	-	-	-

TABLE A.48 – Comparaison des performances obtenues sur le motif q_{10} à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_{11} = {{*information extraction*, *languages*, *named entity recognition*, auteurs ayant publié entre 1993 et 2004}, {*information extraction*, *languages*, *named entity recognition*, documents publiés entre 1993 et 1999}}. Dans la table A.49, nous décrivons les performances obtenues sur le bimotif q_{11} à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
Hubs			
ScholarMap	**.57**	**.04**	**.08**
Model 2	0	0	0
Infinite Random Walk	.43	.03	.06
PageRank	0	0	0
h-index	0	0	0
Citations	0	0	0
Autorités			
ScholarMap	**0**	**0**	**0**
Model 2	0	0	0
Infinite Random Walk	0	0	0
PageRank	0	0	0
h-index	0	0	0
Citations	0	0	0

TABLE A.49 – Comparaison des performances obtenues sur le motif q_{11} à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_{12} = {{*information extraction*, *semantics*, auteurs ayant publié entre 1993 et 1999},{*information extraction*, *semantics*, documents publiés entre 1993 et 1999}}. Dans la table A.50, nous décrivons les performances obtenues sur le bimotif q_{12} à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	**0**	**0**	**0**
Model 2	0	0	0
Infinite Random Walk	0	0	0
PageRank	0	0	0
h-index	0	0	0
Citations	0	0	0
	Autorités		
ScholarMap	**.29**	**.02**	**.04**
Model 2	0	0	0
Infinite Random Walk	0	0	0
PageRank	0	0	0
h-index	0	0	0
Citations	0	0	0

TABLE A.50 – Comparaison des performances obtenues sur le motif q_{12} à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

On a q_{13} = {{*information extraction*, *named entity recognition*, auteurs ayant publié entre 2000 et 2004}, {{*information extraction*, *named entity recognition*, documents publiés entre 2000 et 2004}}. Dans la table A.51, nous décrivons les performances obtenues sur le bimotif q_{13} à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode.

Méthode	Précision	Rappel	F-mesure
	Hubs		
ScholarMap	0	0	0
Model 2	0	0	0
Infinite Random Walk	**.66**	**.02**	**.04**
PageRank	0	0	0
h-index	0	0	0
Citations	0	0	0
	Autorités		
ScholarMap	**.25**	**.01**	**.02**
Model 2	0	0	0
Infinite Random Walk	0	0	0
PageRank	0	0	0
h-index	0	0	0
Citations	0	0	0

TABLE A.51 – Comparaison des performances obtenues sur le motif $q_1 3$ à l'aide des méthodes issues de l'état de l'art ainsi qu'à l'aide de notre méthode (ScholarMap) à partir du graphe biparti auteurs → publications citées construit dans le cadre de l'expérimentation sur le corpus ACL Anthology

Bibliographie

Publications liées à la thèse

[Zev+18] Stella Zevio, Haïfa Zargayouna et al., « Vers une cartographie automatique des thématiques et profils d'experts associés à une conférence scientifique : 9 ans d'ateliers Recherche d'Information SEmantique (RISE) », *in* : *Actes de la dixième édition de l'atelier Recherche d'Information SEmantique (RISE)* (2018), p. 6–13.

[Zev+19] Stella Zevio, Guillaume Santini, Haıfa Zargayouna et al., « Fouille de texte et fouille de graphe appliquées à la recherche d'experts », *in* : *IC 2019 : 30es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 30th French Knowledge Engineering Conference)* (2019), p. 222–223.

[Zev+20] Stella Zevio, Guillaume Santini, Henry Soldano et al., « A Combination of Semantic Annotation and Graph Mining for Expert Finding in Scholarly Data », *in* : *GEM 2020 : ECML PKDD 2020 Workshop on Graph Embedding and Mining* (2020).

[Zev18] Stella Zevio, « Knowledge Discovery and Enrichment from Scholarly Data for Expert Finding », *in* : *The 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW) Doctoral Consortium* (2018).

Références

[ABT17] Milena Angelova, Veselka Boeva et Elena Tsiporkova, « Advanced Data-driven Techniques for Mining Expertise », *in* : *30th Annual Workshop of the Swedish Artificial Intelligence Society (SAIS) 137* (2017), p. 45–52.

[Agg14] Charu C Aggarwal, « Data Clustering », *in* : *Algorithms and Application* (2014).

[Agi+08] Eugene Agichtein et al., « Finding High-quality Content in Social Media », *in* : *Proceedings of the 2008 International Conference on Web Search and Data Mining* (2008), p. 183–194.

[AJM04] Yuan An, Jeannette Janssen et Evangelos E Milios, « Characterizing and Mining the Citation Graph of the Computer Science Literature », *in* : *Knowledge and Information Systems* 6.*6* (2004), p. 664–678.

[AKO18] Mohammed Zuhair Al-Taie, Seifedine Kadry et Adekunle Isiaka Obasa, « Understanding Expert Finding Systems : Domains and Techniques », *in* : *Social Network Analysis and Mining* 8.*1* (2018), p. 1–9.

[Ang+20] Simone Angioni et al., « AIDA : a Knowledge Graph about Research Dynamics in Academia and Industry », *in* : *International Semantic Web Conference* (2020).

[B+07] Krisztian Balog, Maarten De Rijke et al., « Determining Expert Profiles (With an Application to Expert Finding) », *in* : *International Joint Conference on Artificial Intelligence* 7 (2007), p. 2657–2662.

[BAD06] Krisztian Balog, Leif Azzopardi et Maarten De Rijke, « Formal Models for Expert Finding in Enterprise Corpora », *in* : *Proceedings of the 29th Annual International Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval* (2006), p. 43–50.

[Bai+07] Peter Bailey et al., « Overview of the TREC 2007 Enterprise Track », *in* : *Text REtrieval Conference (TREC)* (2007).

[Bal+07] Krisztian Balog, Toine Bogers et al., « Broad Expertise Retrieval in Sparse Data Environments », *in* : *Proceedings of the 30th Annual International Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval* (2007), p. 551–558.

[Bal+08] Krisztian Balog, Paul Thomas et al., *Overview of the TREC 2008 Enterprise Track*, rapp. tech., 2008.

[Bal+12] Krisztian Balog, Yi Fang et al., « Expertise Retrieval », *in* : *Foundations and Trends in Information Retrieval* 6.*2–3* (2012), p. 127–256.

[Bao+06] Shenghua Bao et al., « Research on Expert Search at Enterprise Track of TREC 2006 », *in* : *Text REtrieval Conference (TREC)* (2006).

[BAR09] Krisztian Balog, Leif Azzopardi et Maarten de Rijke, « A Language Modeling Framework for Expert Finding », *in* : *Information Processing & Management* 45.*1* (2009), p. 1–19.

[BB07] AM Bogers et Krisztian Balog, « UvT Expert Collection Documentation », *in* : *ILK Research Group Technical Report* 7 (2007).

[BD08] Krisztian Balog et Maarten De Rijke, *Combining Candidate and Document Models for Expert Search*, rapp. tech., Amsterdam University (Netherlands), 2008.

[Bel66] Richard Bellman, « Dynamic Programming », *in* : *Science* 153.*3731* (1966), p. 34–37.

[Ben+03] Yoshua Bengio et al., « A Neural Probabilistic Language Model », *in* : *Journal of Machine Learning Research* 3 (2003), p. 1137–1155.

[Ber+13] Richard Berendsen et al., « On the Assessment of Expertise Profiles », *in* : *Journal of the American Society for Information Science and Technology* 64.*10* (2013), p. 2024–2044.

[BG94] Michael Buckland et Fredric Gey, « The Relationship Between Recall and Precision », *in* : *Journal of the American Society for Information Science* 45.*1* (1994), p. 12–19.

[BHL01] Tim Berners-Lee, James Hendler et Ora Lassila, « The Semantic Web », *in* : *Scientific American* 284.*5* (2001), p. 34–43.

[Bia+09] Jiang Bian et al., « Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement », *in* : *Proceedings of the 18th International Conference on World Wide Web* (2009), p. 51–60.

[Bir+08] Steven Bird et al., « The ACL Anthology Reference Corpus : a Reference Dataset for Bibliographic Research in Computational Linguistics », *in* : *The Sixth International Conference on Language Resources and Evaluation* (2008).

[Blo+08] Vincent D Blondel et al., « Fast Unfolding of Communities in Large Networks », *in* : *Journal of Statistical Mechanics : Theory and Experiment* 2008.*10* (2008), p. 1–12.

[BNJ03] David M Blei, Andrew Y Ng et Michael I Jordan, « Latent Dirichlet Allocation », *in* : *Journal of Machine Learning Research* 3.*Jan* (2003), p. 993–1022.

[Bol+08] Kurt Bollacker et al., « Freebase : A Collaboratively Created Graph Database for Structuring Human Knowledge », *in* : *Proceedings of the 2008 Association for Computing Machinery (ACM) Special Interest Group on Management of Data (SIGMOD) International Conference on Management of Data* (2008), p. 1247–1250.

[Bor10] Georgeta Bordea, « Concept extraction applied to the task of expert finding », *in* : *Extended Semantic Web Conference* (2010), p. 451–456.

[Bor13] Georgeta Bordea, « Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining », thèse de doct., 2013.

[BPS06] John H Bradley, Ravi Paul et Elaine Seeman, « Analyzing the Structure of Expert Knowledge », *in* : *Information & Management* 43.*1* (2006), p. 77–91.

[Bra01] Andreas Brandstädt, « On Robust Algorithms for the Maximum Weight Stable Set Problem », *in* : *International Symposium on Fundamentals of Computation Theory* (2001), p. 445–458.

[Bus+19] Davide Buscaldi et al., « Mining Scholarly Data for Fine-Grained Knowledge Graph Construction », *in* : *CEUR Workshop Proceedings* 2377 (2019), p. 21–30.

[BZ11] Vladimir Batagelj et Matjaz Zaversnik, « Fast Algorithms for Determining (Generalized) Core Groups in Social Networks », *in* : *Advances in Data Analysis and Classification* 5.*2* (2011), p. 129–145.

[Cam+03] Christopher S Campbell et al., « Expertise Identification Using Email Communications », *in* : *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (2003), p. 528–531.

[Cao+05] Yunbo Cao et al., « Research on Expert Search at Enterprise Track of TREC 2005 », *in* : *Text REtrieval Conference (TREC)* (2005).

[Car+14] Cornelia Caragea et al., « Citeseer X : A Scholarly Big Dataset », *in* : *European Conference on Information Retrieval* (2014), p. 311–322.

[CB07] Rodrigo Costas et Maria Bordons, « The H-index : Advantages, Limitations and its Relation With Other Bibliometric Indicators at the Micro Level », *in* : *Journal of Informetrics* 1 (2007), p. 193–203.

[CD16] Damien Cram et Béatrice Daille, « Terminology Extraction with Term Variant Detection », *in* : *Proceedings of ACL-2016 System Demonstrations* (2016), p. 13–18.

[CEM16] Daniel Cunningham, Sean Everton et Philip Murphy, *Understanding Dark Networks : A Strategic Framework for the Use of Social Network Analysis*, 2016.

[CFP19] Paolo Cifariello, Paolo Ferragina et Marco Ponza, « WISER : A Semantic Approach for Expert Finding in Academia Based on Entity Linking », *in* : *Information Systems* 82 (2019), p. 1–16.

[Che+06] Haiqiang Chen et al., « Social Network Structure Behind the Mailing Lists : ICT-IIIS at TREC 2006 Expert Finding Track », *in* : *Text REtrieval Conference (TREC)* (2006).

[Che+13] Hung-Hsuan Chen et al., « CSSeer : An Expert Recommendation System Based on CiteseerX », *in* : *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (2013), p. 381–382.

[CHH14] Ong Kok Chien, Poo Kuan Hoong et Chiung Ching Ho, « A Comparative Study of HITS vs PageRank Algorithms for Twitter Users Analysis », *in* : *2014 International Conference on Computational Science and Technology (ICCST)* (2014), p. 1–6.

[CJX20] Xiaojun Chen, Shengbin Jia et Yang Xiang, « A Review : Knowledge Reasoning Over Knowledge Graph », *in* : *Expert Systems with Applications* 141 (2020), p. 112948.

[CN08] Lin Chen et Richi Nayak, « Expertise Analysis in a Question Answer Portal for Author Ranking », *in* : *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (2008), p. 134–140.

[CP12] Michel Crampes et Michel Plantié, « Détection de communautés dans les graphes bipartis », *in* : *MARAMI : Conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques* (2012).

[Cra+01] Nick Craswell et al., « P@NOPTIC Expert : Searching for Experts Not Just for Documents », *in* : *Ausweb Poster Proceedings, Queensland, Australia* 15 (2001), p. 17.

[D+98] Thomas H Davenport, Laurence Prusak et al., *Working Knowledge : How Organizations Manage What They Know*, 1998.

[Dee+90] Scott Deerwester et al., « Indexing by Latent Semantic Analysis », *in* : *Journal of the American Society for Information Science* 41.*6* (1990), p. 391–407.

[Des+20] Danilo Dessi et al., « AI-KG : An Automatically Generated Knowledge Graph of Artificial Intelligence », *in* : *International Semantic Web Conference* (2020), p. 127–143.

[DKL08] Hongbo Deng, Irwin King et Michael R Lyu, « Formal Models for Expert Finding on DBLP Bibliography Data », *in* : *2008 Eighth IEEE International Conference on Data Mining* (2008), p. 163–172.

[DM06] Fotis Draganidis et Gregoris Mentzas, « Competency Based Management : a Review of Systems and Approaches », *in* : *Information Management & Computer Security* 14.*1* (2006), p. 51–64.

[Dom+03] Byron Dom et al., « Graph-based Ranking Algorithms for E-mail Expertise Analysis », *in* : *Proceedings of the 8th Association for Computing Machinery (ACM) Special Interest Group on Management of Data (SIGMOD) Workshop on Research Issues in Data Mining and Knowledge Discovery* (2003), p. 42–48.

[Dua+07] Huizhong Duan et al., « Research on Enterprise Track of TREC », *in* : *Text REtrieval Conference (TREC)* (2007).

[ELG07] Kate Ehrlich, Ching-Yung Lin et Vicky Griffiths-Fisher, « Searching for Experts in the Enterprise : Combining Text and Social Network Analysis », *in* : *Proceedings of the 2007 International Association for Computing Machinery (ACM) Conference on Supporting Group Work* (2007), p. 117–126.

[FA19] M Fenner et A Aryani, *Introducing the PID Graph*, 2019.

[FRB19] Tim Fischer, Steffen Remus et Chris Biemann, « LT ExpertFinder : An Evaluation Framework for Expert Finding Methods », *in* : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (2019), p. 98–104.

[FSM11] Yi Fang, Luo Si et Aditya P Mathur, « Discriminative Probabilistic Models for Expert Search in Heterogeneous Information Sources », *in* : *Information Retrieval* 14.*2* (2011), p. 158–177.

[Fu+05] Yupeng Fu, Wei Yu et al., « THUIR at TREC 2005 : Enterprise Track », *in* : *Proceedings of the Fourteenth Text REtrieval Conference (TREC)* 500-266 (2005).

[Fu+07a] Yupeng Fu, Yufei Xue et al., « THUIR at TREC 2007 : Enterprise Track », *in* : *Proceedings of The Sixteenth Text REtrieval Conference (TREC))* 500-274 (2007).

[Fu+07b] Yupeng Fu et al., « A CDD-based Formal Model for Expert Finding », *in* : *Proceedings of the Sixteenth Association for Computing Machinery (ACM) Conference on Conference on Information and Knowledge Management* (2007), p. 881–884.

[Fu+07c] Yupeng Fu et al., « Finding Experts using Social Network Analysis », *in* : *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)* (2007), p. 77–80.

[FZ07] Hui Fang et ChengXiang Zhai, « Probabilistic Models for Expert Finding », *in* : *European Conference on Information Retrieval* (2007), p. 418–430.

[GA09] Yael Garten et Russ B Altman, « Pharmspresso : a Text Mining Tool for Extraction of Pharmacogenomic Concepts and Relationships from Full Text », *in* : *BMC Bioinformatics* 10.*S2* (2009), S6.

[Gáb+16] Kata Gábor, Haïfa Zargayouna et al., « Semantic Annotation of the Anthology Corpus for the Automatic Analysis of Scientific Literature (ACL) », *in* : *Language Resources and Evaluation Conference (LREC)*, Proceedings of the LREC 2016 Conference (2016).

[Gáb+18] Kata Gábor, Davide Buscaldi et al., « SemEval-2018 Task 7 : Semantic Relation Extraction and Classification in Scientific Papers », *in* : *Proceedings of The 12th International Workshop on Semantic Evaluation* (2018), p. 679–688.

[GP17] Soumyajit Ganguly et Vikram Pudi, « Paper2vec : Combining Graph and Text Information for Scientific Paper Representation », *in* : *European Conference on Information Retrieval* (2017), p. 383–395.

[Gru93] Thomas R Gruber, « A Translation Approach to Portable Ontology Specifications », *in* : *Knowledge Acquisition* 5.*2* (1993), p. 199–220.

[GSW05] Bernhard Ganter, Gerd Stumme et Rudolf Wille, *Formal Concept Analysis : Foundations and Applications*, t. 3626, 2005.

[GTV13] Christos Giatsidis, Dimitrios M Thilikos et Michalis Vazirgiannis, « D-cores : Measuring Collaboration of Directed Graphs Based on Degeneracy », *in* : *Knowledge and Information Systems* 35.*2* (2013), p. 311–343.

[Gu+07] Bin Gu et al., « Competition Among Virtual Communities and User Valuation : The Case of Investing-related Communities », *in* : *Information Systems Research* 18.*1* (2007), p. 68–85.

[HBC17] Simon David Hernandez, Davide Buscaldi et Thierry Charnois, « 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* (2017), p. 995–999.

[Hir05] Jorge E Hirsch, « An Index to Quantify an Individual's Scientific Research Output », *in* : *Proceedings of the National Academy of Sciences* 102.*46* (2005), p. 16569–16572.

[HNB13] Seyyed Hadi Hashemi, Mahmood Neshati et Hamid Beigy, « Expertise Retrieval in Bibliographic Network : a Topic Dominance Learning Approach », *in* : *Proceedings of the 22nd Association for Computing Machinery (ACM) International Conference on Information & Knowledge Management* (2013), p. 1117–1126.

[Hof17] Thomas Hofmann, « Probabilistic Latent Semantic Indexing », *in* : *Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Forum* 51.*2* (2017), p. 211–218.

[Hof99] Thomas Hofmann, « Probabilistic Latent Semantic Analysis », *in* : *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (1999), p. 289–296.

[HP06] Seth Hettich et Michael J Pazzani, « Mining for Proposal Reviewers : Lessons Learned at the National Science Foundation », *in* : *Proceedings of the 12th Association for Computing Machinery (ACM) SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006), p. 862–871.

[HR10] Jerry R Hobbs et Ellen Riloff, « Handbook of Natural Language Processing », *in* : t. 2, 2010, chap. Information Extraction, p. 511–532.

[Hu+06] Guoping Hu et al., « A Supervised Learning Approach to Entity Search », *in* : *Asia Information Retrieval Symposium* (2006), p. 54–66.

[HY05] Conglei Yao Bo Peng Jing He et Zhifeng Yang, « CNDS Expert Finding System for Text REtrieval Conference (TREC) », *in* : *Proceedings of the Fourteenth Text REtrieval Conference (TREC)* (2005).

[ID10] Nitin Indurkhya et Fred J Damerau, *Handbook of Natural Language Processing*, t. 2, 2010.

[Jar+19] Mohamad Yaser Jaradeh et al., « Open Research Knowledge Graph : Next Generation Infrastructure for Semantic Scholarly Knowledge », *in* : *Proceedings of the 10th International Conference on Knowledge Capture* (2019), p. 243–246.

[K+96] Henry Kautz, Bart Selman, Al Milewski et al., « Agent Amplified Communication », *in* : *Association for the Advancement of Artificial Intelligence (AAAI)/Innovation Applications of Artificial Intelligence (IAAI)* 1 (1996), p. 3–9.

[Kap+12] Stelios Kapetanakis et al., « Monitoring Financial Transaction Fraud with the Use of Case-based Reasoning », *in* : *17th UK Case-Based Reasoning Workshop* (2012).

[Kar72] Richard M Karp, « Reducibility Among Combinatorial Problems », *in* : *Complexity of Computer Computations*, 1972, p. 85–103.

[Kha+17] Samiya Khan et al., « A Survey on Scholarly Data : From Big Data Perspective », *in* : *Information Processing & Management* 53.*4* (2017), p. 923–944.

[Kle99] Jon M Kleinberg, « Authoritative Sources in a Hyperlinked Environment », *in* : *Journal of the Association for Computing Machinery (ACM)* 46.*5* (1999), p. 604–632.

[KW03] Ludmila I Kuncheva et Christopher J Whitaker, « Measures of Diversity in Classifier Ensembles and Their Relationship With the Ensemble Accuracy », *in* : *Machine learning* 51.*2* (2003), p. 181–207.

[LC04] Xiaoyong Liu et W Bruce Croft, « Cluster-based Retrieval Using Language Models », *in* : *Proceedings of the 27th Annual International Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval* (2004), p. 186–193.

[Leh+15] Jens Lehmann et al., « DBpedia–A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia », *in* : *Semantic Web* 6.*2* (2015), p. 167–195.

[Len95] Douglas B Lenat, « CYC : A Large-scale Investment in Knowledge Infrastructure », *in* : *Communications of the Association for Computing Machinery (ACM)* 38.*11* (1995), p. 33–38.

[Li+07] Juanzi Li et al., « EOS : Expertise Oriented Search Using Social Networks », *in* : *Proceedings of the 16th International Conference on World Wide Web* (2007), p. 1271–1272.

[Lin+17] Shuyi Lin et al., « A Survey on Expert Finding Techniques », *in* : *Journal of Intelligent Information Systems* 49.*2* (2017), p. 255–279.

[Liu+05] Xiaoming Liu et al., « Co-authorship Networks in the Digital Library Research Community », *in* : *Information Processing & Management* 41.*6* (2005), p. 1462–1480.

[Liu07] Bing Liu, *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*, 2007.

[LK10] Baichuan Li et Irwin King, « Routing Questions to Appropriate Answerers in Community Question Answering Services », *in* : *Proceedings of the 19th Association for Computing Machinery (ACM) International Conference on Information and Knowledge Management* (2010), p. 1585–1588.

[LLT11] Theodoros Lappas, Kun Liu et Evimaria Terzi, « A Survey of Algorithms and Systems for Expert Location in Social Networks », *in* : *Social Network Data Analytics*, 2011, p. 215–241.

[LW10] Peter Ljunglöf et Mats Wirén, « Handbook of Natural Language Processing », *in* : t. 2, 2010, chap. Syntactic Parsing, p. 59–91.

[MA00] David W McDonald et Mark S Ackerman, « Expertise Recommender : a Flexible Recommendation System and Architecture », *in* : *Proceedings of the 2000 Association for Computing Machinery (ACM) Conference on Computer Supported Cooperative Work* (2000), p. 231–240.

[Man+10] Paolo Manghi et al., « An Infrastructure for Managing EC Funded Research Output-The OpenAIRE Project », *in* : *The Grey Journal : An International Journal on Grey Literature* 6.*1* (2010).

[Mar96] Ben Martin, « The Use of Multiple Indicators in the Assessment of Basic Research », *in* : *Scientometrics* 36.*3* (1996), p. 343–362.

[May06] Mark T Maybury, *Expert Finding Systems*, rapp. tech., Technical Report MTR06B000040, MITRE Corporation, 2006.

[MC96] Marie-Laure Mugnier et Michel Chein, « Représenter des connaissances et raisonner avec des graphes », *in* : *Revue d'intelligence artificielle* 10.*1* (1996), p. 7–56.

[MCD05] Nick Craswell Msr, Nick Craswell et Arjen P De Vries, « Overview of the TREC-2005 Enterprise Track », *in* : 2005.

[MH02] Audris Mockus et James D Herbsleb, « Expertise Browser : A Quantitative Approach to Identifying Expertise », *in* : *Proceedings of the 24th International Conference on Software Engineering.* (2002), p. 503–512.

[MHO08] Craig Macdonald, David Hannah et Iadh Ounis, « High Quality Expertise Evidence for Expert Search », *in* : *European Conference on Information Retrieval* (2008), p. 283–295.

[Mil95] George A Miller, « WordNet : a Lexical Database for English », *in* : *Communications of the Association for Computing Machinery (ACM)* 38.*11* (1995), p. 39–41.

[MM07] David Mimno et Andrew McCallum, « Expertise Modeling for Matching Papers with Reviewers », *in* : *Proceedings of the 13th Association for Computing Machinery (ACM) SIGKDD International Conference on Knowledge Discovery and Data Mining* (2007), p. 500–509.

[MO06a] Craig Macdonald et Iadh Ounis, « Searching for Expertise Using the Terrier Platform », *in* : *Proceedings of the 29th Annual International Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval* (2006), p. 732–732.

[MO06b] Craig Macdonald et Iadh Ounis, « Voting for Candidates : Adapting Data Fusion Techniques for an Expert Search Task », *in* : *Proceedings of the 15th Association for Computing Machinery (ACM) International Conference on Information and Knowledge Management* (2006), p. 387–396.

[MO07] Craig Macdonald et Iadh Ounis, « A Belief Network Model for Expert Search », *in* : *Proceedings of 1st Conference on Theory of Information Retrieval (ICTIR)* (2007).

[MO08] Craig Macdonald et Iadh Ounis, « Voting Techniques for Expert Search », *in* : *Knowledge and Information Systems* 16.*3* (2008), p. 259–280.

[Mon+10] Fergal Monaghan et al., « Exploring your Research : Sprinkling some Saffron on Semantic Web Dog Food », *in* : *Semantic Web Challenge at the International Semantic Web Conference* 117 (2010), p. 420–435.

[Mou+11] Pierre-Nicolas Mougel, Marc Plantevit et al., « Extraction sous contraintes d'ensembles de cliques homogènes. », *in* : *Extraction et Gestion des Connaissances (EGC)* (2011), p. 443–454.

[Mou+14] Pierre-Nicolas Mougel, Christophe Rigotti, Marc Plantevit et al., « Finding Maximal Homogeneous Clique Sets », *in* : *Knowledge and Information Systems* 39.*3* (2014), p. 579–608.

[MRG12] Pierre-Nicolas Mougel, Christophe Rigotti et Olivier Gandrillon, « Finding Collections of k-clique Percolated Components in Attributed Graphs », *in* : *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2012), p. 181–192.

[MSJ16] Minesh Mathew, Ajeet Kumar Singh et CV Jawahar, « Multilingual OCR for Indic Scripts », *in* : *12th IAPR Workshop on Document Analysis Systems (DAS)* (2016), p. 186–191.

[NG04] Mark EJ Newman et Michelle Girvan, « Finding and Evaluating Community Structure in Networks », *in* : *Physical Review E* 69.*2* (2004), p. 026113.

[NJ02] Andrew Y Ng et Michael I Jordan, « On Discriminative vs. Generative Classifiers : A Comparison of Logistic Regression and Naive Bayes », *in* : *Advances in Neural Information Processing Systems* (2002), p. 841–848.

[NR09] Andreas Noack et Randolf Rotta, « Multi-level Algorithms for Modularity Clustering », *in* : *International Symposium on Experimental Algorithms* (2009), p. 257–268.

[Nuz+16] Andrea Giovanni Nuzzolese et al., « Conference Linked Data : The scholarlydata Project », *in* : *International Semantic Web Conference* (2016), p. 150–158.

[OMM13] Francesco Osborne, Enrico Motta et Paul Mulholland, « Exploring Scholarly Data With Rexplore », *in* : *International Semantic Web Conference* (2013), p. 460–477.

[Osb+16] Francesco Osborne, Angelo Salatino et al., « Automatic Classification of Springer Nature Proceedings with Smart Topic Miner », *in* : *International Semantic Web Conference* (2016), p. 383–399.

[OSM14] Francesco Osborne, Giuseppe Scavo et Enrico Motta, « Identifying Diachronic Topic-based Research Communities by Clustering Shared Research Trajectories », *in* : *European Semantic Web Conference* (2014), p. 114–129.

[Ova14] Steven Ovadia, « ResearchGate and Academia. edu : Academic Social Networks », *in* : *Behavioral & social sciences librarian* 33.*3* (2014), p. 165–169.

[Pag+99] Lawrence Page et al., *The PageRank Citation Ranking : Bringing Order to the Web*. Rapp. tech., Stanford InfoLab, 1999.

[Pau17] Heiko Paulheim, « Knowledge Graph Refinement : A Survey of Approaches and Evaluation Methods », *in* : *Semantic Web* 8.*3* (2017), p. 489–508.

[PC07] Desislava Petkova et W Bruce Croft, « Proximity-based Document Representation for Named Entity Retrieval », *in* : *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (2007), p. 731–740.

[PC08] Desislava Petkova et W Bruce Croft, « Hierarchical Language Models for Expert Finding in Enterprise Corpora », *in* : *International Journal on Artificial Intelligence Tools* 17.*01* (2008), p. 5–18.

[PC11] Aditya Pal et Scott Counts, « Identifying Topical Authorities in Microblogs », *in* : *Proceedings of the Fourth Association for Computing Machinery (ACM) International Conference on Web Search and Data Mining* (2011), p. 45–54.

[PK10] Aditya Pal et Joseph A Konstan, « Expert Identification in Community Question Answering : Exploring Question Selection Bias », *in* : *Proceedings of the 19th Association for Computing Machinery (ACM) International Conference on Information and Knowledge Management* (2010), p. 1505–1508.

[Pro+16] Thiago B Procaci et al., « Finding Topical Experts in Question & Answer Communities », *in* : *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)* (2016), p. 407–411.

[Rad+13] Dragomir R Radev et al., « The ACL Anthology Network Corpus », *in* : *Language Resources and Evaluation* 47.*4* (2013), p. 919–944.

[RB08] Marko A Rodriguez et Johan Bollen, « An Algorithm to Determine Peer-reviewers », *in* : *Proceedings of the 17th Association for Computing Machinery (ACM) Conference on Information and Knowledge Management* (2008), p. 319–328.

[RD11] Jérôme Rocheteau et Béatrice Daille, « TTC TermSuite : A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora », *in* : *18ème Conférence francophone sur le Traitement Automatique des Langues Naturelles Conference (TALN 2011)* (2011).

[RDF90] A Rebecca Reuber, Lorraine S Dyke et Eileen M Fisher, « Using a Tacit Knowledge Methodology to Define Expertise », *in* : *Proceedings of the 1990 Association for Computing Machinery (ACM) SIGBDP Conference on Trends and Directions in Expert Systems* (1990), p. 262–274.

[Ria+12] Fatemeh Riahi et al., « Finding Expert Users in Community Question Answering », *in* : *Proceedings of the 21st International Conference on World Wide Web* (2012), p. 791–798.

[Ros+12] Michal Rosen-Zvi et al., *The Author-Topic Model for Authors and Documents*, 2012.

[Sal+18a] Angelo Salatino et al., « Classifying Research Papers with the Computer Science Ontology », *in* : *ISWC 2018 Posters & Demonstrations and Industry Tracks* (oct. 2018), sous la dir. de Marieke van Erp.

[Sal+18b] Angelo Salatino et al., « The Computer Science Ontology : A Large-Scale Taxonomy of Research Areas », *in* : (2018).

[SB11] Elena Smirnova et Krisztian Balog, « A User-oriented Model for Expert Finding », *in* : *European Conference on Information Retrieval* (2011), p. 580–592.

[SB14] Shai Shalev-Shwartz et Shai Ben-David, *Understanding Machine Learning : From Theory to Algorithms*, 2014.

[Sei83] Stephen B. Seidman, « Network Structure and Minimum Degree », *in* : *Social Networks* 5 (1983), p. 269–287.

[SH08] Pavel Serdyukov et Djoerd Hiemstra, « Modeling Documents as Mixtures of Persons for Expert Finding », *in* : *European Conference on Information Retrieval* (2008), p. 309–320.

[Sha+03] Parantu K Shah et al., « Information Extraction from Full Text Scientific Articles : Where are the Keywords ? », *in* : *BMC Bioinformatics* 4.*1* (2003), p. 20.

[Sho09] David Shotton, « Semantic Publishing : the Coming Revolution in Scientific Journal Publishing », *in* : *Learned Publishing* 22.*2* (2009), p. 85–94.

[Sho13] David Shotton, « Publishing : Open Citations », *in* : *Nature News* 502.*7471* (2013), p. 295.

[Sil+17] Gianmaria Silvello et al., « Semantic Representation and Enrichment of Information Retrieval Experimental Data », *in* : *International Journal on Digital Libraries* 18.*2* (2017), p. 145–172.

[Sin+15] Arnab Sinha et al., « An Overview of Microsoft Academic Service (MAS) and Applications », *in* : 2015, p. 243–246.

[Sin+18] Mayank Singh et al., « CL Scholar : The ACL Anthology Knowledge Graph Miner », *in* : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations* (2018), p. 16–20.

[Sin12] Amit Singhal, « Introducing the Knowledge Graph : Things, not Strings », *in* : *Official Google Blog* 5 (2012).

[SKW08] Fabian M Suchanek, Gjergji Kasneci et Gerhard Weikum, « YAGO : A Large Ontology from Wikipedia and WordNet », *in* : *Journal of Web Semantics* 6.*3* (2008), p. 203–217.

[Sol+17] Henry Soldano, Guillaume Santini, Dominique Bouthinon et Emmanuel Lazega, « Hub-Authority Cores and Attributed Directed Network Mining », *in* : *International Conference on Tools with Artificial Intelligence (ICTAI)* (2017), p. 1120–1127.

[Sol+18] Henry Soldano, Guillaume Santini, Dominique Bouthinon, Sophie Bary et al., « Bi-pattern Mining of Two Mode and Directed Networks », *in* : *Companion Proceedings of the Web Conference 2018* (2018), p. 1287–1294.

[Sow76] John F Sowa, « Conceptual Graphs for a Data Base Interface », *in* : *IBM Journal of Research and Development* 20.*4* (1976), p. 336–357.

[Sow92] John F Sowa, « Conceptual Graphs as a Universal Knowledge Representation », *in* : *Computers & Mathematics with Applications* 23.*2-5* (1992), p. 75–93.

[SPA07] Grimm Stephan, Hitzler Pascal et Abecker Andreas, « Knowledge Representation and Ontologies », *in* : *Semantic Web Services : Concepts, Technologies, and Applications* (2007), p. 51–105.

[SS14] Henry Soldano et Guillaume Santini, « Graph Abstraction For Closed Pattern Mining In Attributed Networks », *in* : *European Conference in Artificial Intelligence (ECAI)*, Frontiers in Artificial Intelligence and Applications 263 (2014), p. 849–854.

[SSB15] Henry Soldano, Guillaume Santini et Dominique Bouthinon, « Local Knowledge Discovery in Attributed Graphs », *in* : *International Conference on Tools with Artificial Intelligence (ICTAI)* (2015), p. 250–257.

[SSB19] Henry Soldano, Guillaume Santini et Dominique Bouthinon, « Attributed Graph Pattern Set Selection Under a Distance Constraint », *in* : *International Conference on Complex Networks and Their Applications* (2019), p. 228–241.

[Sta+10] Milan Stankovic et al., « Looking for Experts ? What can Linked Data do for you ? », *in* : *Linked Data on the Web (LDOW)* (2010).

[SVC06] Ian Soboroff, Arjen P de Vries et Nick Craswell, « Overview of the TREC 2006 Enterprise Track », *in* : t. 6, 2006, p. 1–20.

[SYK09] Kazumi Saito, Takeshi Yamada et Kazuhiro Kazama, « The k-Dense Method to Extract Communities from Complex Networks », *in* : *Mining Complex Data*, 2009, p. 243–257.

[Tan+08] Jie Tang et al., « Arnetminer : Extraction and Mining of Academic Social Networks », *in* : *Proceedings of the 14th Association for Computing Machinery (ACM) SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008), p. 990–998.

[Tka+15] Dominika Tkaczyk, Paweł Szostek et al., « CERMINE : Automatic Extraction of Structured Metadata from Scientific Literature », *in* : *International Journal on Document Analysis and Recognition (IJDAR)* 18.*4* (2015), p. 317–335.

[Tka+18] Dominika Tkaczyk, Andrew Collins et al., « Machine Learning vs. Rules and Out-of-the-box vs. Retrained : An Evaluation of Open-source Bibliographic Reference and Citation Parsers », *in* : *Proceedings of the 18th Association for Computing Machinery (ACM)/IEEE on Joint Conference on Digital Libraries* (2018), p. 99–108.

[Tru13] Richard J Trudeau, *Introduction to Graph Theory*, 2013.

[Tu+10] Yuancheng Tu et al., « Citation Author Topic Model in Expert Search », *in* : *Coling 2010 : Posters* (2010), p. 1265–1273.

[Van06] Anthony FJ Van Raan, « Comparison of the Hirsch-index With Standard Bibliometric Indicators and With Peer Judgment for 147 Chemistry Research Groups », *in* : *Scientometrics* 67.*3* (2006), p. 491–502.

[VRW16] Christophe Van Gysel, Maarten de Rijke et Marcel Worring, « Unsupervised, Efficient and Semantic Expertise Retrieval », *in* : *Proceedings of the 25th International Conference on World Wide Web* (2016), p. 1069–1079.

[VV98] Vladimir Vapnik et Vlamimir Vapnik, *Statistical Learning Theory Wiley*, 1998.

[Wan+13] G Alan Wang et al., « ExpertRank : A Topic-Aware Expert Finding Algorithm for Online Knowledge Communities », *in* : *Decision Support Systems* 54.*3* (2013), p. 1442–1451.

[Wan+18] Ruijie Wang et al., « AceKG : A Large-scale Knowledge Graph for Academic Data Mining », *in* : *Proceedings of the 27th Association for Computing Machinery (ACM) International Conference on Information and Knowledge Management* (2018), p. 1487–1490.

[Wen+10] Jianshu Weng et al., « TwitterRank : Finding Topic-sensitive Influential Twitterers », *in* : *Proceedings of the Third Association for Computing Machinery (ACM) International Conference on Web Search and Data Mining* (2010), p. 261–270.

[Wu+14] Zhaohui Wu et al., « Towards Building a Scholarly Big Data Platform : Challenges, Lessons and Opportunities », *in* : *Proceedings of the 14th Association for Computing Machinery (ACM)/IEEE-CS Joint Conference on Digital Libraries* (2014), p. 117–126.

[Wu+18] Kan Wu et al., « CareerMap : Visualizing Career Trajectory », *in* : *Science China Information Sciences* 61.*10* (2018), p. 1–3.

[Xia+17] F. Xia et al., « Big Scholarly Data : A Survey », *in* : *IEEE Transactions on Big Data* 3.*1* (2017), p. 18–35.

[Yan+09] Zi Yang et al., « Expert2Bólè : From Expert Finding to Bólè Search », *in* : *Proceedings of the Association for Computing Machinery (ACM) SIGKDD International Conference on Knowledge Discovery and Data Mining,(KDD'09)* (2009), p. 1–4.

[Yar10] David Yarowsky, « Handbook of Natural Language Processing », *in* : t. 2, 2010, chap. Word Sense Disambiguation, p. 315–338.

[YK03] Dawit Yimam-Seid et Alfred Kobsa, « Expert-Finding Systems for Organizations : Problem and Domain Analysis and the DEMOIR Approach », *in* : *Journal of Organizational Computing and Electronic Commerce* 13.*1* (2003), p. 1–24.

[You+06] Ganmei You et al., « Ricoh Research at Text REtrieval Conference (TREC) : Enterprise Track », *in* : *Proceedings of the Fifteenth Text REtrieval Conference (TREC)* (2006).

[Yu+16] Min-Chun Yu et al., « ResearchGate : An Effective Altmetric Indicator for Active Researchers ? », *in* : *Computers in Human Behavior* 55 (2016), p. 1001–1006.

[Yua+17] KQ Yuan et al., « Construction Techniques and Research Development of Medical Knowledge Graph », *in* : *Application Research of Computers* 35.*7* (2017), p. 1–12.

[ZA05] Jun Zhang et Mark S Ackerman, « Searching for Expertise in Social Networks : a Simulation of Potential Strategies », *in* : *Proceedings of the 2005 International Association for Computing Machinery (ACM) SIGGROUP conference on Supporting Group Work* (2005), p. 71–80.

[ZAA07] Jun Zhang, Mark S Ackerman et Lada Adamic, « Expertise Networks in Online Communities : Structure and Algorithms », *in* : *Proceedings of the 16th International Conference on World Wide Web* (2007), p. 221–230.

[Zhu+06] Jianhan Zhu, Dawei Song, Stefan Rüger et al., « The Open University at TREC 2006 Enterprise Track Expert Search Task », *in* : *Proceedings of The Fifteenth Text REtrieval Conference (TREC)* (2006).

[ZL17] Chengxiang Zhai et John Lafferty, « A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval », *in* : *Association for Computing Machinery (ACM) Special Interest Group on Information Retrieval (SIGIR) Forum* 51.*2* (2017), p. 268–276.

[ZLK12] Tom Chao Zhou, Michael R Lyu et Irwin King, « A Classification-based Approach to Question Routing in Community Question Answering », *in* : *Proceedings of the 21st International Conference on World Wide Web* (2012), p. 783–790.

[ZSR07] Jianhan Zhu, Dawei Song et Stefan Rüger, « The Open University at TREC 2007 Enterprise Track », *in* : *Proceedings of the Sixteenth Text REtrieval Conference (TREC)* (2007).

[ZTL07] Jing Zhang, Jie Tang et Juanzi Li, « Expert Finding in a Social Network », *in* : *International Conference on Database Systems for Advanced Applications* (2007), p. 1066–1069.

Résumé en français

La recherche d'experts consiste en l'identification d'un ensemble d'individus que l'on considère comme experts d'une thématique particulière. Il s'agit d'une problématique essentielle dans le milieu académique. En effet, il est constamment nécessaire d'identifier des chercheurs appropriés lors de la constitution de comités de lecture ou d'évaluation de projets de recherche, par exemple. De même, il est particulièrement utile d'identifier automatiquement les experts d'un domaine de recherche à partir de la littérature scientifique. Nous proposons une approche de découverte et d'enrichissement de connaissances basée sur une annotation sémantique des articles scientifiques, sur leur représentation sous forme de réseaux de collaboration scientifique et leur fouille à l'aide d'une méthode d'abstraction de graphe. Cette méthode permet de se focaliser sur les zones denses des réseaux et de découvrir des experts et leurs expertises associées à l'aide de contraintes de connectivité. Ces dernières permettent de prendre en compte une validation par les pairs, matérialisée par la densité des relations de collaboration scientifique que les individus entretiennent entre eux. Nous expérimentons notre approche sur un corpus de publications scientifiques, proposons une méthode d'évaluation originale de nos résultats et comparons nos performances aux méthodes de recherche d'experts implémentées dans le cadre d'évaluation LT ExpertFinder. Nous obtenons des performances supérieures à l'état de l'art et découvrons que les indicateurs d'expertise les plus déterminants sont la rédaction d'articles fortement cités mais également la capacité à citer la littérature scientifique appropriée.

Discipline

Informatique

Mots-clefs

recherche d'experts, annotation sémantique, fouille de données, fouille de texte, fouille de graphe, graphes attribués, abstraction de graphe, analyse de réseau

Intitulé et adresse du laboratoire

Laboratoire d'Informatique de Paris Nord (LIPN),
99 Avenue Jean Baptiste Clément, 93430 Villetaneuse

Title

Knowledge discovery and enrichment from texts for expert finding

Abstract

Expert finding consists in the identification of a set of individuals who are considered to be experts in a particular topic. This is an essential problem in the academic world. Indeed, it is constantly necessary to identify suitable researchers when setting up reading or evaluation committees for research projects, for example. Indeed, it is particularly useful to automatically identify experts on a specific field from the scientific literature. We suggest an approach for knowledge discovery and enrichment based on a semantic annotation of scientific articles, on their representation in the form of scientific collaboration networks and their exploration using a graph abstraction method. This method makes it possible to focus on dense areas of networks and to discover experts and their associated expertise using connectivity constraints. The latter make it possible to take into account a validation by peers, materialized by the density of scientific collaboration relations that individuals maintain with each other. We test our approach on a corpus of scientific publications, propose an original method for evaluating our results and compare our performance to expert research methods implemented in the LT ExpertFinder evaluation framework. We obtain better performance than the state of the art and discover that the most decisive indicators of expertise are the writing of highly cited articles but also the ability to cite the appropriate scientific literature.

Keywords

expert finding, semantic annotation, data mining, text mining, graph mining, attributed graphs, graph abstraction, network analysis