Modélisation d'expertise scientifique pour la constitution de comités de programme

Hong Diep Tran

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *19 décembre 2017* **par :**

Hong Diep TRAN

Modélisation d'expertise scientifique pour la constitution de comités de programme

---

JURY

Sabine Loudcher	PR, Univ. Lyon 2	Rapporteuse
Christian Sallaberry	MCF HDR, Univ. Pau (UPPA)	Rapporteur
Mohand Boughanem	PR, Univ. Toulouse 3	Examinateur
Chérifa Boukacem-Zeghmouri	PR, Univ. Lyon 1	Examinatrice
Gilles Hubert	MCF HDR, Univ. Toulouse 3	Directeur
Guillaume Cabanac	MCF HDR, Univ. Toulouse 3	Co-directeur

École doctorale et spécialité :
MITT : Image, Information, Hypermédia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse UMR 5505
Directeurs de Thèse :
Gilles Hubert et *Guillaume Cabanac*
Rapporteurs :
Sabine Loudcher et *Christian Sallaberry*

Remerciements

Không thày, đố mày làm nên, comme le dit ce proverbe vietnamien très connu : « impossible de réussir dans la vie sans professeurs ».

Pour arriver jusqu'ici aujourd'hui, j'ai dû faire beaucoup d'efforts, avec l'aide et le soutien indispensable de ma famille, des amis et des collègues. Certes, je ne suis pas la seule à fournir ces efforts, il en est de même pour toutes celles et ceux réalisant un doctorat. Pourtant, malgré tous ces efforts, tout le monde n'y arrive pas. En ce qui me concerne, je pense que mes professeurs, Gilles Hubert et Guillaume Cabanac, ont vraiment fait beaucoup pour moi. Leur dire simplement « Je vous remercie beaucoup » est vraiment trop peu par rapport à ce qu'ils ont fait pour moi, et ce que je pense d'eux. Ils m'ont particulièrement accompagnée durant mon doctorat. Certains disent que faire des recherches est particulièrement difficile, d'autres disent que ce n'est pas si difficile, mais personne n'ose dire qu'il ne faut pas consacrer beaucoup de temps et d'efforts pour faire une thèse, et que l'accomplissement finit par arriver un jour. Pour moi, être doctorante, ce fut vraiment des efforts longs et intenses. Je suis arrivée en France avec de nombreuses lacunes, en termes de connaissances, d'expérience de travail. Il a également fallu surmonter les manques engendrés par une vie à l'étranger, sans mes proches, loin de ma maison. Premièrement, bien qu'ils soient mes professeurs, ils sont devenus également des proches. Avec chaque sourire, chaque parole ou geste d'encouragement qui semblaient anodins, ils m'ont donné de l'énergie, de la motivation et de la confiance dans mon travail. Chaque jour, j'ai beaucoup appris de mes professeurs, la connaissance, la méthode de travail et même l'attitude dans le travail. Cela a joué un rôle décisif dans l'accomplissement de ma thèse. Sans cela, il aurait été impossible d'y arriver, même en étant suffisamment déterminée.

Je ne suis pas une doctorante particulièrement forte. À certains moments je me suis sentie incapable de poursuivre mon chemin. Pourtant, à l'heure où j'écris ces mots, je suis en train de faire mes derniers pas pour finaliser cette thèse. Les vietnamiens disent souvent qu'une personne, même avec des compétences très limitées, en présence de bons enseignants, peut également réussir. Ainsi c'est mon cas, j'ai eu beaucoup de chance de rencontrer de bons professeurs.

Je souhaite remercier sincèrement mes professeurs, Gilles Hubert et Guillaume Cabanac pour le temps parfait que j'ai passé à travailler avec eux.

Je souhaite également remercier sincèrement les rapporteurs de ma thèse, Madame le professeur Sabine Loudcher (Université Lyon 2) et Monsieur le professeur Christian Sallaberry (Université de Pau et des pays de l'Adour) pour avoir accordé du temps à une lecture attentive et détaillée de mon manuscrit ainsi que pour leurs remarques encourageantes et constructives. Je tiens également à remercier les examinateurs, Monsieur le professeur Mohand Boughanem et Madame le professeur Chérifa Boukacem-Zeghmouri pour avoir accepté de participer à mon jury de thèse.

Je souhaite remercier les membres de l'équipe IRIS, et également ceux de l'équipe SIG pour m'avoir aidé à terminer mon travail.

Je tiens à remercier chaleureusement Thi Kieu Thu Raffi, Thibaut Thonet et mes amis de m'avoir accompagnée et soutenue durant ces quatre années loin de ma famille et de mon pays.

Je souhaite exprimer mes remerciements au professeur Tran Van Chu et à M. Dao Duy Phuong (*Vietnam National Forestry University* et *Foreign Education Department*) de m'avoir donnée l'occasion et de bonnes conditions pour accomplir mon travail de recherche en France.

Je souhaite également exprimer mes remerciements au professeur Tran Van Cong, et aux enseignants du Département de français – HANU et de l'Institut français de l'Ambassade de France au Vietnam pour m'avoir donnée des leçons précieuses de langue et de culture française.

Finalement et, d'une importance particulière à mes yeux, je voudrais dédier quelques mots à ma famille : mes parents et mes enfants. Pour que je puisse me concentrer sur mon travail, ils m'ont aidé dans toutes les tâches, petites comme grandes. Ils l'ont fait comme si c'était naturel, avec joie et affection. Les enfants ont dû surmonter les difficultés du quotidien sans leur mère. Ce fut ma grande motivation pour travailler. Leur joie suscitée par l'obtention de mon doctorat et mon retour à la maison sera peut-être plus grande que la mienne. Je ne pouvais pas rentrer auprès d'eux les mains vides, ce fut vraiment la grande motivation pour mon travail.

Résumé

La publication scientifique permet de communiquer les progrès en sciences auprès des chercheurs et du grand public. Les articles paraissent dans les revues spécialisées et les actes de conférences, usuellement après évaluation par les pairs. Les comités de rédaction et de programme sous-jacents représentent la clé de voûte du processus d'évaluation. Avec le développement des revues et le nombre croissant de conférences scientifiques organisées chaque année, rechercher des experts pour participer à ces comités mobilise les réseaux des scientifiques les plus actifs. C'est une activité chronophage mais critique pour maintenir la confiance que place la société en la science.

Cette thèse se focalise sur la tâche de suggestion de membres de comité de programme (CP) pour des conférences scientifiques. Elle comporte trois volets.

Premièrement, nous formulons l'hypothèse que l'invitation d'un chercheur à participer à un CP peut s'expliquer par différentes preuves d'expertise. Les activités de publication et le rayonnement scientifique d'un chercheur contribueraient à forger cette expertise tant recherchée. Nous proposons une modélisation de cette expertise scientifique multifacette basée sur un graphe hétérogène pondéré.

Deuxièmement, nous définissons des indicateurs scientométriques pour étudier la constitution de CP passés, nous permettant de distinguer quantitativement les critères prépondérant à l'inclusion des chercheurs au CP.

Troisièmement, nous concevons une approche de suggestion de CP pour une conférence donnée, en combinant les résultats des indicateurs scientométriques susmentionnés. Cette approche vise à aider au renouvellement des CP en suggérant des chercheurs actifs, proches de la thématique de la conférence donnée et dont l'impact des travaux sur la communauté est reconnu.

Notre approche de suggestion de CP est expérimentée pour une des conférences de premier plan de notre communauté de recherche : SIGIR (*Special Interest Group on Information Retrieval*), en considérant ses éditions de 1971 à 2015, ainsi que les conférences proches thématiquement telles que CIKM, ECIR, WSDM et WWW. Nous évaluons la pertinence de notre approche en séparant le jeu de données en deux ensembles délimités par une année frontière : observation sur les éditions jusqu'à cette année-là et test sur les éditions postérieures. Ceci permet de quantifier la part des membres de CP à la fois suggérés par notre approche et pertinents, car figurant dans les CP effectifs de la conférence considérée.

Abstract

Academic publishing enables to communicate scientific progress to researchers and, more generally, to the society. Articles appear in specialized journals and conference proceedings, usually after peer review. The underlying editorial and program committees represent the cornerstone of the evaluation process. With the development of journals and the growing number of scientific conferences held annually, searching for experts who would serve in these committees mobilizes the networks of the most active scientists. It is a time-consuming and yet critical activity that maintains the trust the society places in science.

This PhD thesis focuses on the task of suggesting program committee (PC) members for scientific conferences. It is organized into three parts.

First, we hypothesize that a researcher's invitation to participate in a PC can be explained by different proofs of expertise. The publication activities and the scientific influence of a researcher would contribute to build up this precious expertise. We propose a modelling of this multifaceted scientific expertise based on a weighted heterogeneous graph.

Second, we define scientometric indicators to study the composition of past PCs, enabling us to distinguish the leading criteria for the inclusion of researchers in PCs quantitatively.

Third, we design a PC suggestion approach for a given conference, combining the results of the aforementioned scientometric indicators. This approach aims to help for the renewal of PCs by suggesting active researchers who are close to the topics of the given conference and whose impact on the community is recognized.

Our approach of PC suggestion is experimented in the context of the leading conferences from our research community: SIGIR (Special Interest Group on Information Retrieval), considering its 1971–2015 editions, and topically close conferences such as CIKM, ECIR, WSDM, and WWW. We evaluated the relevance of our approach by splitting the dataset into two sets delimited by a boundary year. The observation phase considers those editions up to this specific year; the testing phase targets the later editions. This makes it possible to quantify the share of PC members both suggested by our approach and relevant, as they appear in the actual PCs of the considered conference.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

Sommaire

1.1 Évaluation et diffusion des savoirs

Les progrès de la recherche sont diffusés au travers de multiples canaux, tels que les ouvrages, les articles de revues et de conférences, les communications orales et la littérature grise. Les innovations sont, en règle générale, soumises à l'appréciation de la communauté scientifique via le processus d'évaluation par les pairs. Ce dernier repose sur l'expertise de chercheurs qui sont chargés d'évaluer l'originalité de la création ou de la découverte, la robustesse des méthodes employées et la capacité des résultats avancés à repousser le front de recherche. Ces experts siègent au sein des comités : chaque revue s'adosse à un comité de rédaction, chaque conférence s'adosse à un comité de programme.

La constitution et l'entretien d'un comité est une tâche critique : la qualité de la sélection réalisée — opérée sur les contributions à évaluer — est intimement liée à l'expertise des membres du comité, à leur complémentarité sur le plan des savoirs, et à l'adéquation de leurs domaines de compétence eu égard au thème de la revue ou de la conférence. Crowcroft, Keshav et McKeown (2009) mettent en garde contre les écueils auxquels les communautés doivent faire face à l'heure actuelle : accroissement

du nombre de soumissions d'articles à expertiser — de qualité en déclin —, superficialité des expertises de certains évaluateurs peu scrupuleux, ainsi que favoritisme bénéficiant indûment à certains scientifiques proches du pilotage des communautés.

1.2 Pilotage des communautés scientifiques

Les chercheurs mobilisés au niveau du pilotage des revues et conférences doivent en permanence connaître le positionnement de leur communauté scientifique sur l'atlas de la science (Börner, 2010). C'est en fonction de cette information stratégique qu'ils l'animent et la font évoluer. Par exemple, des événements scientifiques sont coorganisés par des communautés proches, tels que les conférences jointes (*joint conferences*) et les numéros spéciaux interdisciplinaires dans les revues.

Le pilotage des communautés scientifiques met à profit les études quantitatives des sciences et de l'innovation. Le domaine interdisciplinaire associé, nommé « scientométrie » (Nalimov & Mulchenko, 1969 ; Leydesdorff & Milojević, 2015), étend les méthodes de la bibliométrie (Pritchard, 1969) qui porte sur l'analyse des documents scientifiques. Les termes scientométrie et bibliométrie sont couramment employés de façon interchangeable (De Bellis, 2009, p. 5) ou utilisés comme synonymes (Larivière, 2015, p. 27). La scientométrie contribue à apporter un éclairage aux multiples questions que formulent les animateurs de communautés scientifiques, telles que :

- par étude des *contenus textuels* des articles, quels sont les sujets au cœur — respectivement à la périphérie — de notre communauté ? Quels nouveaux concepts émergent ?
- par étude des *co-signatures* d'articles, quelles sont les collaborations entre institutions, entre chercheurs, entre disciplines, entre nations ?
- par étude des *bibliographies*, quelle est la proximité conceptuelle entre les articles et, par extension, entre les chercheurs associés ?
- par étude des *citations reçues*, quels sont les travaux les plus plébiscités par et en dehors de la communauté ?

Ce besoin en matière de réflexivité apparaît, par exemple, au sein de l'association INFORSID [1] qui organise le congrès francophone éponyme depuis 1983. C'est à l'occasion du congrès de Lille en 2011 que le bureau exécutif d'INFORSID, en charge du pilotage scientifique de l'association, a animé un atelier de réflexion sur « la recherche en systèmes d'information et ses nouvelles frontières ». Le résultat de cette réflexion

1. Informatique des organisations et systèmes d'information et de décision, http://inforsid.fr

a « dressé un état des lieux des thèmes de recherche les plus actifs au niveau francophone » (Collectif INFORSID, 2012, p. 9).

L'association INFORSID a souhaité approfondir cette réflexion en finançant une étude scientométrique sur les systèmes d'information [2] dans le cadre d'une « action spécifique » portée par Guillaume Cabanac et Gilles Hubert en 2012–2013. Le résultat de cette étude, l'anthologie d'INFORSID en ligne [3], représente la communauté INFORSID au travers de différentes perspectives relatives aux :

1. *éditions* passées du congrès, représentées sur une carte de la France et des pays francophones ayant accueilli le congrès ;
2. *comités de programme* ayant pris en charge le volet scientifique lié à la sélection des articles via l'évaluation par les pairs, pour chaque édition du congrès. Chaque membre du comité de programme est identifié, son affiliation notée et son rôle au sein d'éventuels comités de rédaction de revues internationales souligné (Cabanac, 2012) ;
3. *articles* publiés dans les actes de chaque édition du congrès ;
4. *auteurs* co-signataires d'article, avec leur affiliation pour chaque article publié ;
5. *thématiques* des recherches publiées lors des trois décennies du congrès, extraites des titres d'articles et observées sur les périodes 1983–1993, 1994–2003 et 2004–2013 ;
6. *lieux de publication proches*, plébiscités par les auteurs d'INFORSID. Ce sont des conférences et revues publiant des travaux réalisés par les chercheurs dont au moins un article est publié dans les actes d'INFORSID.

1.3 Genèse de la suggestion de membres de comités

L'anthologie d'INFORSID exploite les données collectées manuellement à partir des actes pour mettre en œuvre un processus de suggestion de membres de comité de programme (CP). Plusieurs éléments ont motivé cette proposition originale. Premièrement, renouveler une part significative des membres pour chaque édition contribue à dynamiser la communauté scientifique. En effet, une des missions confiées aux

2. http://web.archive.org/web/2017/http://inforsid.org/?q=node/31
3. https://www.irit.fr/~Guillaume.Cabanac/inforsid

membres du CP consiste à diffuser l'appel à communications dans leur environnement scientifique, afin de faire connaître le congrès et d'attirer de nouveaux auteurs. Deuxièmement, les présidents de CP sont familiers de leurs thématiques, pour lesquelles ils peuvent mobiliser des experts qu'ils connaissent ; cependant, ils éprouvent des difficultés à identifier les experts nécessaires au bon déroulement de l'évaluation en ce qui concerne les thématiques desquelles ils sont plus distants intellectuellement. Troisièmement, à expertise et aptitude égales, certains collègues se rendent plus visibles que d'autres ; ils sont donc davantage sollicités, conséquence directe de l'effet Matthieu (Merton, 1968). Les collègues experts, quoique discrets, sont donc des cibles privilégiées pour intégrer le CP et contribuer à une évaluation par les pairs de qualité.

La suggestion de membres de CP pour une édition donnée opère ainsi : les auteurs les plus présents dans les actes des éditions passées, tout en étant les moins sollicités pour siéger au CP sont listés — certains n'ont d'ailleurs jamais été invités au CP. Ces suggestions, mises à jour chaque année, ont été exploitées par les président·e·s des CP successifs, de 2013 à 2017, pour solliciter des collègues. Les retours informels indiquent un intérêt pour cette approche et une pertinence de certaines suggestions (hormis la suggestion de collègues retraités ou décédés) qui ont permis de solliciter avec succès des collègues méconnus du pilotage du CP.

L'association EGC [4], communauté nationale dont les thématiques sont connexes à celles d'INFORSID, a lancé en 2016 le défi [5] « Communauté EGC : quelle histoire et quel avenir ? ». Nous avons saisi cette opportunité pour développer plus avant notre approche initiée précédemment. En particulier, nous avons rassemblé des experts spécialistes de la lexicométrie et des études de genre autour de ce travail (Cabanac, Hubert, Tran, Favre & Labbé, 2016). Le premier aspect est illustré par la figure 1.1 présentant les thèmes principaux couverts par les conférences EGC. Le second aspect est illustré par la tableau 1.1 listant les suggestions de membres de CP pour l'édition 2016 de la conférence. Notons que ces suggestions incluent une indication relative au sexe des personnes suggérées, avec pour objectif de promouvoir la diversité femme–homme dans le CP — ce dernier étant constitué de 70–75 % d'hommes jusqu'alors.

Cette thèse prolonge cette réflexion — initiée en 2013 dans le cadre d'INFORSID et étendue en 2016 dans le cadre d'EGC — au sujet de la suggestion de membres de CP pour aider au renouvellement des instances d'évaluation des conférences.

4. Extraction et gestion des connaissances, http://www.egc.asso.fr
5. http://www.egc.asso.fr/Manifestations_dEGC/71-FR-Defi_EGC

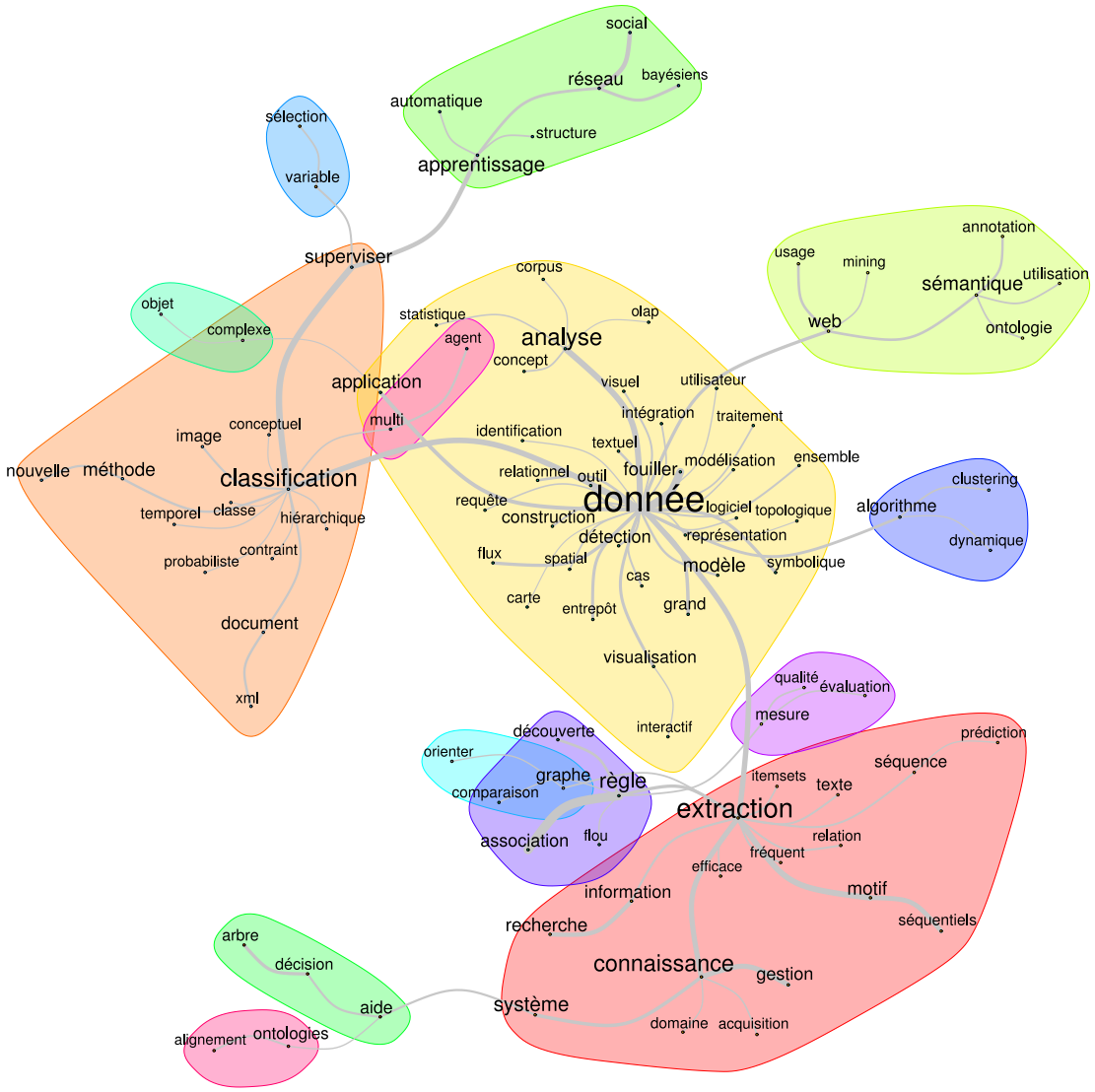

FIGURE 1.1 – Thèmes présents dans les 1 212 titres des articles publiés à EGC lors des éditions 2001–2016 et étudiés dans le cadre du défi EGC de 2016 (Cabanac, Hubert, Tran, Favre & Labbé, 2016). Ce visuel a été réalisé par analyse des mots reliés avec le logiciel de lexicométrie Iramuteq (Ratinaud, 2009). Il révèle les expressions les plus fréquentes, impliquant les mots donnée, extraction et classification, notamment.

TABLEAU 1.1 – Suggestion de chercheurs pour constituer un CP pour l'édition 2016 de la conférence EGC (Cabanac, Hubert, Tran, Favre & Labbé, 2016). Les auteurs réunissent les trois conditions suivantes : 1) plus de deux articles publiés à EGC et 2) dont le dernier article EGC est postérieur à 2008 et 3) n'ayant jamais siégé au CP d'EGC ou alors antérieurement à 2010. Les suggestions sont ordonnées selon trois clés de tri décroissantes : par dernier CP, par nombre d'articles publiés à EGC et par date du dernier article (cf. lignes en pointillés).

Sexe	Identité	# articles	dernier article EGC	# CP	dernier CP
♀	Cherif Chiraz Latiri	10	2015	0	
♀	Fadila Bentayeb	10	2014	0	
♂	Baghdad Atmani	9	2014	0	
♂	Arnaud Soulet	8	2015	0	
♂	Matthias Studer	8	2010	0	
…	…	…	…	…	…
♀	Marie-Odile Cordier	3	2009	0	
♂	Edwin Diday	16	2009	9	2009
♀	Annie Morin	8	2011	2	2009
♀	Sylvie Desprès	5	2012	3	2009
♂	David Auber	4	2011	1	2009
♀	Marie-Christine Rousset	3	2010	8	2009
♂	Henri Briand	19	2011	5	2006
♂	Yves Kodratoff	11	2009	4	2005
♂	Einoshin Suzuki	5	2013	2	2005
♀	Sylvie Szulman	3	2010	1	2005
♀	Marie-Luce Picard	4	2010	1	2004

1.4 Problématiques et contribution de la thèse

L'accueil positif des propositions décrites dans la section précédente et implémentées pour les deux associations a motivé le travail conduit dans le cadre de la présente thèse. Le processus de suggestion de membres de CP — introduit en 2012 — innovait de par l'exploitation des données d'implications dans les comités. Cependant, il ne tirait pas parti de données bibliométriques usuellement exploitées telles que les citations, les bibliographies complètes (publications dans les conférences proches thématiquement) et les caractéristiques des individus (ancienneté dans la recherche). Par ailleurs, la pertinence des résultats fournis par le processus de suggestion n'avait pas été évaluée, hormis par des retours informels enthousiastes (Cabanac et al., 2016).

La contribution de cette thèse, dont les résultats sont publiés dans les actes de la conférence nationale CORIA (Tran, Cabanac & Hubert, 2016) et dans les actes de la conférence internationale RCIS (Tran, Cabanac & Hubert, 2017) consiste en :

1. un processus de suggestion de membres de CP exploitant une variété de traces numériques considérées comme preuves d'expertise. Ce processus s'appuie sur :
   (a) la modélisation de la conférence étudiée et du domaine scientifique dans lequel elle s'inscrit,
   (b) la définition d'indicateurs scientométriques pour quantifier l'expertise des individus suivant différentes facettes ;
2. l'évaluation de la pertinence des résultats de ce processus selon un protocole reposant sur des mesures éprouvées du domaine de la recherche d'information.

1.5 Organisation du mémoire

Ce mémoire est organisé en deux parties, l'une consacrée à l'état de l'art et l'autre à la contribution. La partie I propose une synthèse de la littérature couvrant la recherche d'expert (chapitre 2), la recherche de contenus informationnels en fonction d'un domaine d'expertise (chapitre 3) et les travaux de scientométrie les plus proches de notre problématique (chapitre 4).

La partie II expose notre contribution en matière de modélisation de la sphère académique des conférences (chapitre 5), de mise au point d'indicateurs scientométriques d'expertise scientifique (chapitre 6) sous-tendant le processus de suggestion de scienti-

fiques concourant au renouvellement des CP de conférences (chapitre 7). Le chapitre 8 conclut le mémoire et suggère des pistes de recherche en perspective à cette thèse.

Première partie

État de l'art : de la recherche d'expert à la proposition de comités de programme

Résumé (État de l'art).

Faisant l'hypothèse qu'un comité de programme d'une conférence est constitué d'un panel d'experts du domaine couvert par la conférence, la problématique abordée dans cette thèse peut être rapprochée de la recherche d'expert qui a fait l'objet de différents travaux, ainsi que de travaux connexes. Les revues et conférences constituent également un objet d'étude en scientométrie, dont les contributions sont reprises dans le domaine informatique, et notamment en recherche d'information.

Cette partie consacrée à l'état de l'art lié à la problématique de constitution des comités de programme de conférences est donc divisée en trois chapitres :

- le chapitre 2 présente les travaux représentatifs de la recherche d'expert ;
- le chapitre 3 étend la problématique à la recherche de contenus informationnels en lien avec de l'expertise ;
- le chapitre 4 fait le lien avec les travaux de scientométrie et leur insertion dans les travaux d'informatique.

Chapitre 2

Recherche d'expert

Sommaire

2.1 Introduction

La recherche d'expert (*expert search*) est une thématique de recherche apparue il y a une quinzaine d'années dans le domaine de la recherche d'information (RI). Balog, Fang, de Rijke, Serdyukov et Si (2012) présentent un tour d'horizon des travaux sur la recherche d'expert et notamment le besoin, les tâches, les approches et les évaluations. Deux tâches essentielles sont associées à la recherche : profiler les experts et trouver les experts. Des travaux représentatifs relatifs à ces deux tâches sont présentés dans les sections suivantes. Le *profilage* d'expert (*expert profiling*) s'intéresse à la représentation de l'expert en termes de thématiques et d'indicateurs d'expertise à partir de différentes sources d'information. La *découverte* d'expert (*expert finding*) consiste à identifier les experts à partir d'un besoin exprimé sous forme d'une requête. Cette tâche a été principalement vue comme une tâche de recherche d'information dans laquelle les représentations des experts remplacent les représentations des documents habituellement manipulées. Les travaux sur la recherche d'expert s'appuient sur les modèles de RI populaires comme les modèles vectoriels, probabilistes et modèles de

langue. Ce dernier type de modèle est le plus présent dans la littérature, comme le reflète la section 2.3.1. Une majorité des travaux de la littérature s'intéresse à la tâche de découverte d'expert, à la faveur de collections de test (*benchmarks*) construites et rendues disponibles dans le cadre des différentes éditions de la tâche *TREC Enterprise Track* organisées durant 4 ans (Balog & de Rijke, 2008).

Craswell, de Vries et Soboroff (2005) présentent la première édition de la tâche *TREC Enterprise Track* et plus particulièrement la sous-tâche *Expert Search* relative à la découverte d'expert correspondant à une requête exprimant une compétence recherchée. Un jeu de 50 requêtes thématiques a été proposé avec les jugements de pertinence associés. L'objectif était de trouver les experts correspondant à la requête — au lieu des documents qui y sont relatifs et issus de l'intranet de l'entreprise. Cependant, les collections fournies ne permettent que d'évaluer les approches orientées vers l'objectif de cette tâche TREC. D'autres objectifs en lien avec la recherche d'expert, comme celui auquel s'intéresse la présente thèse, restent plus difficiles à évaluer comme le soulignent Jayasinghe, Karimi et Ayre (2016).

2.2 Approches de profilage d'expert

Le profilage d'expert repose, en général, sur la construction automatique de profils d'expert. Ces profils sont basés sur un ensemble de mots-clés, et construits à partir de sources d'information qui sont principalement des documents, tels que les documents d'entreprise (Serdyukov, Taylor, Vinay, Richardson & White, 2011), des curricula vitæ (Ribeiro, Santos, Gonçalves & Laender, 2015), ou encore des bibliographies académiques (Cabanac, 2011).

Les travaux suivants, centrés sur le profilage, ont proposé des modèles de profils plus complexes que de simples ensembles de mots-clés, ainsi que des approches pour la constitution automatique de profils.

Yimam et Kobsa (2000) proposent une architecture pour la modélisation de l'expertise, baptisée DEMOIR. Cette architecture combine modélisation d'expertise centralisée et distribuée. Elle s'appuie sur une représentation matricielle des relations entre experts et documents, et la fusion de matrices pour générer les modèles d'expertise.

Balog et de Rijke (2007) distinguent les notions de profil thématique et de profil social, dans le contexte d'intranet d'entreprise. Ils proposent des méthodes pour constituer le profil thématique, basées sur la RI et le filtrage. Les évaluations portent

principalement sur ce profil, avec les collections de *TREC Enterprise Track*. Les auteurs notent eux-mêmes que la solution de profil social introduite et basée sur un graphe de collaboration nécessite d'être développée plus avant.

Afzal et Maurer (2011) ont proposé une visualisation hyperbolique pour l'identification d'experts universitaires, basée sur des profils multifacette. Les facettes sont basées sur différentes mesures, telles que le nombre de publications et le nombre de citations reçues, y compris dans chaque thématique. Les profils sont construits à partir d'une base de fiches descriptives des experts et de données issues du Web. Des pondérations sont appliquées aux profils en fonction des publications, citations et fonctions éditoriales détenues par les experts. Les profils des chercheurs sont reliés à la classification hiérarchique ACM qui sert de base à la visualisation hyperbolique permettant d'identifier les experts associés à chaque entrée de la classification. L'approche proposée n'a cependant pas fait l'objet d'une évaluation.

Serdyukov et al. (2011) se sont intéressés à une nouvelle tâche correspondant à l'annotation (*tagging*) automatique pour définir les domaines d'expertise des employés d'une entreprise. L'approche proposée tire parti d'un algorithme d'apprentissage automatique basé sur différentes sources de preuves d'expertise pour suggérer les annotations (*tags*) pertinentes pour décrire un employé. Les sources concernaient les documents d'entreprise écrits ou liés aux employés, des documents issus du Web, les listes de discussions internes à l'entreprise et les clics issus des fichiers de consultation (*logs*) du moteur de recherche intranet de l'entreprise. Les expérimentations menées ont montré la possibilité de trouver un ensemble de *tags* décrivant partiellement un employé. Les résultats ont souligné l'intérêt d'utiliser les clics du moteur de recherche sur l'intranet. En revanche, les documents du Web ne sont pas pertinents pour cette tâche.

Ribeiro et al. (2015) ont étudié la recommandation de *tags* pour le profilage automatique d'expertise dans le domaine scientifique. Neuf systèmes de recommandation de *tags* orientés contenu ont été analysés. Ces systèmes exploitent trois sources de preuves d'expertise différentes : les titres de publications, les résumés et les mots-clés. Les analyses ont montré de bonnes performances de la plupart de ces systèmes de recommandation avec une amélioration potentielle via leur combinaison à l'aide de stratégies de *learning to rank*.

2.3 Approches de découverte d'expert

Deux types d'approches pour la découverte d'expert se distinguent dans la littérature : les approches textuelles et celles orientées graphe. La grande majorité des travaux se concentre sur une approche textuelle, avec de nombreux travaux basés sur des modèles de langue comme le montre la section suivante consacrée aux approches textuelles. La disponibilité des collections *TREC Enterprise Track* explique le développement des approches textuelles de découverte d'expertise à partir de documents liés aux experts, au détriment d'autres types de travaux. Quelques travaux orientés graphe ont, toutefois, été proposés, exploitant la structure de graphe pour évaluer le niveau d'expertise des individus au sein d'un réseau social.

2.3.1 Approches textuelles de découverte d'expert

Les profils d'experts en termes de thématiques ont été généralement représentés par des vecteurs de termes, permettant le calcul de mesures de similarité classiques, comme le cosinus entre les profils et les requêtes thématiques (Demartini, Gaugaz & Nejdl, 2009).

Dans le cadre de la participation à la première édition de la tâche *TREC Enterprise* en 2005, Cao, Liu, Bao et Li (2005) ont proposé un modèle de langue constitué de deux volets : un modèle de pertinence (appariement du document par rapport à la requête) et un modèle de fenêtre de co-occurrence (appariement de la requête par rapport à un expert). Différentes métadonnées et combinaisons de métadonnées, ainsi qu'une méthode de clustering, sont ensuite utilisées pour réordonner la liste d'experts obtenue.

Macdonald et Ounis (2006) considèrent la recherche d'expert comme un problème de vote. Chaque profil de candidat est constitué d'un ensemble de documents associés au candidat et représentant son expertise. Une recherche *adhoc* de documents, à partir d'une requête, est ensuite réalisée. Chaque document restitué est assimilé à un vote implicite pour le candidat associé. Les techniques de fusion de données, telles que CombSum ou RRF sont adaptées et appliquées pour déterminer le score final d'un candidat en réponse à une requête. Les expérimentations menées sur la collection *TREC Enterprise* 2005 ont montré que la qualité des documents en termes de structure est importante dans l'amélioration des résultats. Macdonald et Ounis (2009) poursuivirent les travaux précédents en étudiant l'intégration d'une normalisation de la taille des profils candidats ainsi que d'une expansion de requête basée sur les documents les mieux

classés. Les résultats de l'étude soulignent l'intérêt de la normalisation et un apport très nuancé de l'expansion de requête.

Petkova et Croft (2006) ont présenté une approche générale pour représenter les connaissances d'un potentiel expert comme un mélange de modèles de langue issus des documents qui lui sont associés. L'approche se base sur la collecte de preuves d'expertise à partir de plusieurs sources hétérogènes, suivant une approche de type modèle de langue pour trouver les associations entre documents et experts et leurs degrés d'association. Les expérimentations menées sur la collection *TREC Enterprise* 2005 ont étudié différents aspects, comme la prise en compte de la structure des documents ou encore la combinaison de différentes définitions de l'expert. Les travaux poursuivis dans (Petkova & Croft, 2007) considèrent les entités nommées pour la recherche d'expert. Une nouvelle représentation des documents est proposée qui met l'accent sur les termes proches des entités apparaissant dans un document. L'étude a notamment montré que l'extraction d'entités nommées n'était pas un facteur prépondérant de performance dans le cas de grandes collections.

Balog, Azzopardi et de Rijke (2006) ont étudié deux stratégies basées sur des modèles génératifs probabilistes pour la recherche d'expert dans une collection de documents. Ces stratégies sont basées sur l'estimation des associations entre documents et experts suivant les informations attribuables aux experts et trouvées dans les documents. La première stratégie est basée sur des modèles « orientés candidats » et la seconde sur des modèles « orientés documents ». Les expérimentations réalisées sur la collection *TREC Enterprise* ont montré de meilleures performances pour la seconde stratégie.

Serdyukov, Chernov et Nejdl (2007) procèdent en se basant sur un modèle de langue en deux étapes. La première étape, empruntée à Balog et al. (2006), utilise le modèle de langue de l'expert pour trouver les top-k experts potentiels pertinents au regard d'un requête. La deuxième étape utilise les top-k documents retournés par un modèle KL de RI sur la collection de documents pour raffiner le modèle de langue de la requête. Un principe d'entropie croisée (*cross-entropy*) est appliqué pour obtenir un score pour chaque expert et les classer. Les expérimentations sont réalisés sur la collection de *TREC Enterprise Track*.

H. Fang et Zhai (2007) présentent un cadre général probabiliste pour la recherche d'expert, s'appuyant sur les points de convergence entre la recherche d'expert et la recherche d'information *adhoc*. Deux familles de modèles génératifs sont proposées :

basées candidats et basées requêtes. Les estimations des modèles sont améliorées à l'aide (1) d'une modélisation des mentions aux candidats dans les documents permettant une pondération des représentations d'un expert candidat, (2) de la modélisation des relations entre documents pour l'expansion de requête, et (3) de priorités basées sur le nombre de mentions des adresses de courriel des candidats. Les expérimentations sont réalisées sur les collections *TREC Enterprise Track* des éditions 2005 et 2006.

Balog, Bogers, Azzopardi, de Rijke et van den Bosch (2007) introduisent trois types de modèles de langue pour la découverte d'expert : basé sur les modèles de candidats, basé sur les modèles de documents et basé sur les modèles thématiques. Le contexte visé est celui de l'intranet d'université *via* la collection *UvT Expert*. Les modèles sont raffinés en incluant des connaissances du domaine (autres formulation des requêtes), du contexte (organisation hiérarchique), et des aspects de nature multilingue.

Serdyukov et Hiemstra (2008) proposent une méthode de recherche d'expert basées sur une modélisation des documents comme des mélanges de modèles de langue individuels. La méthode considère les termes des documents comme étant générés par les personnes mentionnées dans les documents. Le classement final des experts-candidats combine les deux preuves d'expertise que sont la probabilité de générer la requête à partir du modèle de langue individuel du candidat et une probabilité a priori (*prior*) correspondant au niveau d'activité du candidat dans les discussions sur le sujet. Des expérimentations ont été menées sur les collections *TREC Expert Search* des éditions 2005 et 2006.

Balog et de Rijke (2008) ont complété leurs précédents travaux au travers de l'intégration d'une preuve globale d'expertise, obtenue en utilisant des informations non directement liées au nom de l'expert ou la page dans laquelle son nom apparaît. Ces informations sont liées à la fréquence de rejets de documents de certains candidats ou à la cohérence des documents dans lesquels le nom du candidat apparaît.

Zhu, Huang, Song et Rüger (2010) ont étudié différentes caractéristiques (*features*) liées aux documents pour la recherche d'expert. Le contexte d'utilisation est celui d'intranet d'entreprise. Le modèle probabiliste rassemble un modèle de co-occurrence entre document et expert, un modèle de RI entre document et requête ainsi qu'une estimation de priorités des documents basée sur des *features* (Craswell et al., 2005) telles que PageRank, *in-degree* (nombre de liens entrants) et taille d'URL. Les expérimentations ont porté sur les collections *TREC Expert Search* des éditions 2006 et 2007.

Y. Fang, Si et Mathur (2010) ont proposé un cadre d'apprentissage discriminant

pour dériver des modèles discriminants pour la recherche d'expert. La recherche d'expert est abordée comme un problème de classification qui considère les paires requête-expert pertinentes comme des données positives et les paires non-pertinentes comme des données négatives. Ce type de modèle intègre naturellement en un modèle unique différentes preuves liées aux documents et associant les experts. Les expérimentations menées sur deux collections de la tâche *TREC Enterprise* ont montré l'efficacité (*effectiveness*) et la robustesse de l'approche proposée.

Poursuivant leurs travaux vers les méthodes d'apprentissage, Macdonald et Ounis (2011) ont proposé une nouvelle approche utilisant les techniques de *learning to rank* pour apprendre le modèle d'ordonnancement d'agrégats. Une tâche d'ordonnancement d'agrégats est une tâche où les objets manipulés sont représentés par des ensembles de documents devant être ordonnés en réponse à une requête. La recherche d'expert est considérée comme un exemple de tâche d'ordonnancement d'agrégats. L'approche définit un ensemble de *features* basé sur les modèles de pondération TF·IDF et BM25, ainsi que les méthodes de fusion CombSum et CombMnz. L'approche est présentée comme plus simple que celle Y. Fang et al. (2010) car elle peut intégrer des techniques de *learning to rank* existantes. Des expérimentations menées sur deux collections de la tâche *TREC Enterprise* ont montré l'efficacité et la robustesse de l'approche proposée.

Dans la même idée, Moreira, Calado et Martins (2015) ont proposé plus récemment deux approches pour combiner un ensemble d'estimateurs d'expertise issus de trois sources de preuves. La première approche est basée sur les algorithmes de *learning to rank* et la seconde sur les algorithmes d'agrégation des rangs. Les trois sources de preuves qui correspondent aux documents liés aux experts, aux citations et aux profils des experts donnent lieu à trois catégories d'estimateurs : similarité textuelle (TF·IDF, BM25), information de profil (nombre de publications dans des conférences et nombre de publications dans des revues) et citations. Une contribution apportée concerne l'utilisation de différents indicateurs scientométriques, notamment dérivés du Hirsch index (Hirsch, 2005) tels que le *Contemporary Hirsch index* et le *Trend Hirsch index*. Les expérimentations menées à l'aide des bases bibliographiques DBLP (Ley, 2002) et ArnetMiner (Tang et al., 2008) ont montré l'intérêt de combiner les trois catégories d'estimateurs d'expertise.

Récemment dans ce contexte, S. Liang et de Rijke (2016) ont introduit une évolution de la tâche de recherche d'expert au travers d'une nouvelle tâche qui consiste à trouver des groupes d'individus ayant une expertise sur un sujet donné. Cinq modèles

de langues sont proposés : deux modèles d'agrégation, un modèle de document et deux modèles de requêtes. Pour les expérimentations, une collection a été construite à partir des collections *TREC Enterprise* des éditions 2005 et 2006, associées à trois vérités terrain : une binaire (le groupe est pertinent si au moins un de ses membres est expert), une graduelle (suivant la proportion de membres experts au sein du groupe) et une numérique (suivant le nombre d'experts dans le groupe). Les expérimentations ont montré que collecter les preuves d'expertise à partir des documents était la manière la plus efficace pour trouver un groupe ayant une expertise sur un sujet donné.

2.3.2 Approches orientées graphe pour la découverte d'expert

Jun Zhang, Ackerman et Adamic (2007) se sont intéressés aux communautés basées sur le Web pour la recherche et le partage d'expertise, en exploitant les différences en termes de topologie des réseaux formés par ces communautés. Il a présenté une caractérisation de réseau, notamment sur le réseau spécifique du Java Forum, et analysé l'impact sur différentes mesures visant à identifier le niveau d'expertise d'un utilisateur. Les mesures analysées comprenaient des mesures statistiques simples comme le nombre de personnes aidées par une personne donnée, la mesure ExpertiseRank dérivée de PageRank (Brin & Page, 1998) ou encore la mesure d'autorité de HITS (Kleinberg, 1999). Les expérimentations menées suggèrent que les caractéristiques structurelles des communautés influencent les performances des algorithmes étudiés.

Jing Zhang, Tang et Li (2007) se sont intéressés à la recherche d'expert dans un réseau social. Une approche de propagation a été proposée tenant compte des informations locales aux personnes et des relations entre personnes dans le réseau. Les informations locales sont utilisées pour calculer un score initial d'expertise pour chaque nœud (représentant une personne). Un coefficient de propagation d'expertise est associé à chaque lien interpersonnel et un principe de propagation de score est mis en œuvre jusqu'à convergence des scores.

2.4 Bilan

La plupart des travaux liés à la recherche d'expert se sont orientés vers la représentation des experts sous forme de profils thématiques à base de mots-clés et de modèles de RI pour évaluer la correspondance entre les profils d'experts et un besoin de compé-

tence exprimé sous forme de requête. De plus, les travaux sont, pour la plupart, liés à la tâche *TREC Enterprise* proposée de 2005 et 2008, fournissant des collections disponibles pour l'évaluation des approches proposées.

La présente thèse s'intéresse à la composition des comités de programme de conférences. Le besoin de compétences pour une conférence est complexe, multiple et plus vague (que dans le cas de la recherche d'expert) car reposant sur un ensemble de thématiques, souvent non exhaustif. Une recherche textuelle basée sur de telles requêtes risquerait d'engendrer des résultats fortement bruités. De plus, le contenu des documents (majoritairement des publications) associés aux experts potentiels est accessible librement de manière inégale (Khabsa & Giles, 2014), limitant ainsi la possibilité de recourir aux approches textuelles.

Les approches orientées graphe proposées pour la découverte d'expert se sont limitées à l'exploitant de réseaux sociaux basés sur les relations interpersonnelles. Pour l'analyse des compositions de comités de programme — auxquels s'intéresse la présente thèse — l'hypothèse est que de multiples relations impliquant les experts potentiels doivent être considérées. Une exploitation de réseaux plus complexes est proposée dans les chapitres suivants consacrés à la contribution apportée dans le cadre de cette thèse.

D'autres travaux, non axés directement sur la recherche d'expert, quoique liés à la notion d'expertise, de compétence ou de de notoriété sont présents dans la littérature. Cette thèse partage un certain nombre de principes avec ces travaux qui sont présentés dans le chapitre suivant.

Chapitre 3

Recherche d'items liés à la notion d'expertise

Sommaire

3.1 Introduction

La recherche d'expert présentée dans le chapitre précédent constitue une part importante des travaux liés à la notion d'expertise. L'objectif était d'identifier les experts correspondant à un besoin exprimé sous forme de requête thématique. D'autres travaux se sont intéressés à la notion d'expertise, de compétence ou de notoriété visant des objectifs différents ou cherchant à répondre à d'autres expressions de besoins. Cette fois encore, deux grands types d'approches se distinguent : les approches *textuelles* et les approches *orientées graphe*. Un nombre plus important d'approches orientées graphe figure dans la littérature dans ce cadre, contrairement aux approches de recherche d'expert. Il est à noter, cependant, que certaines approches orientées graphe intègrent la manipulation d'éléments textuels.

3.2 Approches textuelles

Les approches textuelles s'inscrivent dans des contextes spécifiques mobilisant des éléments de contenu qui permettent des traitements particuliers.

Nanba et Okumura (1999) ont proposé une approche pour aider un auteur à rédiger un état de l'art (*survey*) sur un sujet donné. L'approche exploite les citations contenues dans les documents. Plusieurs types de références sont distingués, traduisant par exemple le fait que les travaux présentés se basent sur les travaux cités ou qu'ils se comparent à eux. Une analyse du texte introduisant une référence permet de déterminer le type de référence.

McDonald et Ackerman (2000) ont proposé une architecture de système de recommandation pour l'identification d'experts. Le système de recommandation était basé sur un ensemble d'heuristiques établies après l'étude du domaine d'application, à savoir les logiciels médicaux. L'objectif était d'aider les utilisateurs à trouver les experts susceptibles de pouvoir apporter un support technique. Le système de recommandation de type collaboratif exploite des profils d'experts constitués d'items notés pour recommander des experts répondant à une tâche spécifique du domaine. Les profils sont constitués à partir des documents d'une base de support technique qui recense différentes informations comme les problèmes, les personnes affectées et le module logiciel concerné, décrits suivant un thésaurus du domaine. Les performances du système n'ont cependant pas été évaluées.

Liu, Curson et Dew (2005) se sont intéressés à l'utilisation d'un modèle RDF (*resource description framework*) pour l'appariement d'expertise. L'objectif était d'intégrer sémantiquement de multiples sources d'expertise issues de sources de données hétérogènes pour aider des utilisateurs à identifier les « bons » experts. L'étude a été menée dans le contexte académique pour aider les candidats à un doctorat à identifier les directeurs de thèse potentiels avant de postuler explicitement auprès d'une université. Les sources de données exploitées étaient les publications, les projets et les données de ressources humaines. Un modèle RDF combiné à l'utilisation d'une ontologie de domaine représentait les relations entre un chercheur et les concepts d'expertise.

Plus récemment et dans la même idée, Zhixing Huang et Qiu (2010) ont proposé de catégoriser les citations entre articles dans le but de définir des réseaux de citations sémantiques. Les catégories distinguées regroupent, par exemple, la comparaison de travaux, la similitude entre travaux ou encore l'utilisation de précédents travaux.

Bien que n'étant pas étroitement liées à l'approche présentée dans cette thèse, ces approches distinguent différents types de citations, ce qui pourrait constituer une extension future à cette thèse.

3.3 Approches orientées graphe

Les approches orientées graphe exploitent différents types de réseaux, tels que : les réseaux de citations entre publications, des réseaux de conversations par mail entre personnes ou encore des réseaux de collaborations académiques. Elles se fondent sur des mesures basées sur les structures des graphes, telles que les mesures de distance entre nœuds du graphe, des mesures d'autorité ou de centralité. Elles visent, en général, à identifier les nœuds proches ou importants liés à un nœud donné.

Bollacker, Lawrence et Giles (2000) présentent le système CiteSeer [1] visant à faciliter la découverte de publications académiques « utiles » sur le Web. CiteSeer utilise des profils utilisateurs créés explicitement par les utilisateurs puis adaptés en fonction des interactions de l'utilisateur avec le système. Les recommandations faites par le système se basent sur la correspondance entre le profil utilisateur et les publications. CiteSeer extrait notamment les citations contenues dans les publications pour créer un graphe de citations. Les recommandations se basent sur des contraintes exprimées par l'utilisateur sur ses centres d'intérêts, sur la correspondance aux niveaux des mots-clés et sur les liens entre documents basés sur leurs citations communes issues du graphe de citations. Une perspective concerne l'évolution de CiteSeer pour identifier et proposer des concepts ou axes de recherche émergents aux utilisateurs.

Zan Huang, Chung, Ong et Chen (2002) ont proposé une approche de système de recommandation hybride dans un contexte de bibliothèque numérique. Les recommandations sont faites sur la base des achats antérieurs de livres par les utilisateurs. L'approche se base sur une représentation sous forme de graphe à deux niveaux : le niveau *livre* avec les liens de similarité entre livres, le niveau *client* caractérisé par les liens de similarité entre clients, et enfin les liens représentant les achats entre les deux niveaux. Un algorithme de réseau de neurones d'Hopfield a été utilisé pour exploiter les associations fortes entre livres, entre clients et entre livres et clients. Les tests effectués dans le contexte d'une librairie en ligne n'ont cependant pas permis de constater

1. http://citeseerx.ist.psu.edu

une amélioration significative liée à cette approche.

Dom, Eiron, Cozzi et Zhang (2003) ont étudié un ensemble de mesures de graphes pour classer les courriels de personnes en fonction de leur niveau d'expertise sur des sujets. L'étude se base sur des graphes où les nœuds représentent les personnes et les liens représentent l'existence de courriels entre les personnes. Les liens sont orientés, indiquant la supériorité d'une personne par rapport à l'autre en termes de niveau d'expertise. Les mesures étudiées comprenaient notamment l'affinité, le PageRank, la fonction de pouvoir positionnel (PPF) ou encore la mesure d'autorité HITS. Les expérimentations menées sur des données synthétiques et sur des données réelles collectées auprès d'une organisation ont donné l'avantage aux mesures PageRank et PPF.

Liben-Nowell et Kleinberg (2003) se sont focalisés sur la prédiction de lien dans les réseaux sociaux. Ce type de question peut, par exemple, intervenir lors de l'exploitation d'un réseau social académique pour suggérer des collaborations à des chercheurs. Plusieurs approches de prédiction de lien basées sur différentes mesures de proximité entre les nœuds du graphe ont été étudiées, telles que les voisins communs et la distance de Katz (1953). Les résultats des expérimentations sur de grands réseaux de co-auteurs ont montré que les futures interactions pouvaient être extraites de la seule topologie du réseau.

Strohman, Croft et Jensen (2007) présentent un système de recommandation de citations à partir d'une requête-document fournie par l'utilisateur, exploitant un graphe de citations. L'approche se base sur une recherche de cent documents les plus similaires à la requête en termes de contenu. Cet ensemble est ensuite complété par les documents que ceux-ci citent. Enfin, différents facteurs, tels que les années de publication, la similarité textuelle, les co-citations, les auteurs communs, la distance de Katz (1953) ou le nombre de citations reçues, sont employés pour composer la liste des citations recommandées.

Rodriguez et Bollen (2008) ont introduit une approche pour automatiser l'attribution de relecteurs d'article dans le contexte de conférence avec évaluation par les pairs (*peer review*). La problématique attaquée consiste à trouver un équilibre entre identifier un relecteur qualifié tout en évitant les conflits d'intérêts entre auteurs et relecteurs. L'approche est basée sur une modélisation du réseau des co-auteurs et un algorithme de propagation par essaims particulaires (*particle swarm*). Les expérimentations menées sur les données de l'édition 2005 de la *ACM/IEEE Joint Conference on Digital Libraries (JCDL)* ont montré les limites de l'approche en termes de détection des conflits

d'intérêts, penchant de ce fait pour une application durable du processus traditionnel d'affectation des relecteurs.

Tang et al. (2008) présentent la plateforme ArnetMiner [2] pour l'extraction et la fouille de réseaux sociaux académiques. D'une part, des profils de chercheurs sont construits, composés d'informations issues de leurs pages Web personnelles et des publications dont ils sont signataires. D'autre part, un graphe représentant le réseau académique est construit à partir, entre autres, des relations de co-signature de publications et des citations entre publications. Trois modèles génératifs permettent de dégager les thématiques liées aux articles, aux auteurs et aux lieux de publication (c'est-à-dire conférences et revues). Ils ont été appliqués à la recherche d'expertise à partir d'une requête donnée pour trouver des articles, des auteurs ou des conférences. Une exploitation du réseau académique présenté était, par exemple, la recherche d'associations existant entre deux personnes.

Ekstrand et al. (2010) se sont intéressés à la recommandation de littérature scientifique pour tout nouveau chercheur désirant se familiariser avec son domaine de recherche. L'objectif était de proposer une liste d'articles à partir d'un ensemble initial d'article lus par un utilisateur. Pour cela, plusieurs méthodes ont été étudiées pour étendre les algorithmes de filtrage orienté contenu et de filtrage collaboratif à l'aide de mesures d'influence d'un article dans le graphe de citations. Les méthodes PageRank, HITS et SALSA (un algorithme stochastique similaire à HITS mais basé sur une marche aléatoire) sont comparées. Les expérimentations montrent les performances intéressantes des systèmes de recommandation par filtrage collaboratif.

Lao et Cohen (2010) ont proposé une méthode d'apprentissage des mesures de proximité entre nœuds d'un graphe orienté étiqueté représentant la littérature scientifique en biologie. La méthode se base sur la pondération des chemins entre nœuds du graphe et les principes de marche aléatoire (*random walk*). La méthode a été expérimentée sur deux jeux de données dérivés de deux bases génomiques, dans le cadre de quatre tâches : la recommandation de lieu de publication, la recommandation de références bibliographiques, la recherche d'expert et la recommandation de gènes. Les résultats des expérimentations révèlent de meilleures performances pour la méthode proposée comparée à l'algorithme RWR (*random walk with restart*) de la littérature. Cependant, la définition des requêtes utilisées ainsi que la vérité terrain utilisée pour le calcul des mesures d'évaluation ne sont pas clairement présentées.

2. http://arnetminer.org

Cabanac (2011) a proposé une approche de recommandation socio-thématique de chercheurs, exploitant des graphes de co-signature et de co-participation potentielle aux conférences par les chercheurs. La recommandation se fonde sur la définition de mesures de similarité sociale (proximité, connectivité) et thématique (basée sur l'agrégation des titres des articles publiés par un auteur et la mesure cosinus). Les expérimentations menées avec les données de DBLP [3] et avec un panel de 71 chercheurs ont souligné l'intérêt des indices liés aux liens sociaux, qui permettent d'améliorer de 11 % la pertinence des recommandations thématiques.

Dans le contexte de la littérature scientifique, Y. Liang, Li et Qian (2011) ont, quant à eux, proposé un système de recommandation d'articles à partir d'un réseau de citations. Ils ont défini une nouvelle métrique basée sur la dépendance de citation entre deux articles, ainsi qu'un modèle de recommandation basé sur la distance de Katz (1953). Une évaluation de l'approche a été réalisée en utilisant la collection *ACL Anthology Network* [4] (Radev, Muthukrishnan, Qazvinian & Abu-Jbara, 2013 ; Radev, Joseph, Gibson & Muthukrishnan, 2016) en se comparant avec plusieurs lignes de référence (*baselines*) : co-citation, co-coupling, CCIDF, HITS Vector-based et Katz. Les expérimentations ont montré l'intérêt de la métrique proposée.

3.4 Bilan

Les travaux présentés dans ce chapitre, tout comme ceux introduits dans la présente thèse, sont liés aux notions d'expertise, de compétence ou de notoriété. Néanmoins, leurs objectifs diffèrent et, par conséquent, les données exploitées diffèrent. Aucune de ces recherches ne s'est intéressée aux comités de programme de conférences ou à une problématique de même type.

Comme justifié dans le chapitre précédent, la problématique de la présente thèse ne se prête *a priori* pas à une recherche basée sur du contenu textuel. Néanmoins, à l'image de certains des travaux décrits dans ce chapitre, l'intégration du contenu textuel peut être envisagée dès lors qu'il est disponible. Cependant, dans le contexte des conférences, peu de contenu textuel est généralement accessible librement. C'est pourquoi nous nous sommes orientés vers d'autres sources de preuves d'expertise, tout

3. Les données de la *DBLP computer science bibliography* sont diffusées en libre accès au format XML *via* http://dblp.org/xml/.

4. http://clair.eecs.umich.edu/aan

comme les travaux orientés graphe présentés dans ce chapitre.

De plus, les travaux exploitent, pour la plupart, une source de preuve d'expertise unique, des graphes composés d'un seul type de nœud et un seul type de relation entre ces nœuds. L'analyse des comités de programme fait intervenir de multiples éléments tels que les conférences, les chercheurs ou les publications et relations entre ces éléments. Cette thèse s'appuie la représentation sous forme de graphe, plus spécifiquement de graphe triparti : comprenant des nœuds de nature diverses — auteurs, publications, conférences, revues… — et formant un graphe hétérogène. À l'instar de plusieurs contributions identifiées dans l'état de l'art, notre approche se base sur l'exploitation de distances exploitant la structure de graphe, telles que la mesure de Katz (1953).

Cette thèse s'appuie également sur des concepts forgés dans le domaine de la scientométrie portant sur l'étude quantitative de la science et de l'innovation (Leydesdorff & Milojević, 2015). Plusieurs travaux se sont inspirés de la scientométrie en lien avec les notions d'expertise et de réputation dans le contexte des conférences. Parmi eux, deux travaux se sont directement intéressés à la composition des comités de programme des conférences, problématique sur laquelle se concentre cette thèse. Le chapitre suivant se focalise sur ces travaux.

Chapitre 4

Scientométrie et informatique pour caractériser l'expertise

Sommaire

4.1 Introduction

La scientométrie est l'étude quantitative de la science et de l'innovation (Leydesdorff & Milojević, 2015). Nombre d'ouvrages de vulgarisation emploient *scientométrie*, *bibliométrie* et *infométrie* de manière interchangeable (De Bellis, 2009, p. 5) et ces derniers sont souvent considérés comme synonymes (Larivière, 2015, p. 27). L'institutionnalisation de la scientométrie comme discipline scientifique remonte aux années 1950 (D. J. Price, 1951). La scientométrie mesure et analyse la science pour comprendre, notamment, comment celle-ci se structure et évolue. Elle s'appuie sur la définition et l'analyse d'indicateurs quantitatifs valorisés en observant les activités de la recherche.

Cette thèse exploite le même matériau que les méthodes scientométriques. C'est pourquoi nous la lions à la scientométrie, introduite dans la section 4.2. Différents travaux dans le domaine de l'informatique s'appuient sur les contributions issues de la scientométrie (Cabanac, 2016). Nous détaillons plus particulièrement les travaux en lien avec la notion d'expertise ou de réputation dans la section 4.3. Récemment, la

composition des comités de programme de conférences a suscité l'intérêt de la communauté informatique, comme illustré dans la section 4.4.

4.2 Point de vue scientométrique

La scientométrie s'est attachée à mesurer la science à différents niveaux tels que les auteurs, les revues, les institutions ou encore les pays. Les publications — et notamment les citations qu'elles véhiculent — constituent le matériau d'une grande partie des travaux (De Bellis, 2009).

De nombreux indicateurs quantitatifs issus des activités de la recherche ont été définis et analysés pour répondre à différentes questions concernant la structuration et l'évolution de la recherche (Todeschini & Baccini, 2016). Par exemple, Leydesdorff et Rafols (2011) se sont intéressés à la définition d'indicateurs pour mesurer le degré d'interdisciplinarité dans les revues.

Des indicateurs de notoriété célèbres ont été proposés, comme le h-index au niveau auteur (Hirsch, 2005) ou le facteur d'impact (*journal impact factor*) au niveau des revues (Garfield, 1955, 2006). Par ailleurs, Bouyssou et Marchant (2011) ont, par exemple, analysé différentes mesures de classement des auteurs et de leurs départements universitaires. Les résultats montrent la nécessité d'utiliser une mesure adaptée à chaque niveau. Marchant (2009), qui a analysé différentes mesures de classement d'auteurs de publications, suggère par exemple d'utiliser celle qui correspond le mieux au problème considéré. Wildgaard (2015) a comparé 17 indicateurs au niveau auteur et souligné que de nombreux facteurs influençaient les résultats, comme la couverture des travaux dans les bases bibliographiques. La visibilité dans la base peut être très différente de la visibilité au sein de la communauté.

L'utilisation d'indicateurs scientométriques en dehors de leur cadre de définition reste par conséquent délicate (Leydesdorff, Bornmann, Comins & Milojević, 2016). Dans ce contexte, l'utilisation de la scientométrie dans un but d'évaluation de la recherche suscite de nombreuses mises en garde (Billaut, Bouyssou & Vincke, 2010 ; Gevers, 2014 ; Gingras, 2014). La comparaison des chercheurs à l'aide d'indicateurs qui n'ont pas été définis pour cet objectif demeure très controversée.

Au-delà de la définition d'indicateurs, de nombreuses problématiques intéressent la scientométrie dans sa compréhension du fonctionnement de la recherche (Cabanac, 2015). Parmi ces problématiques, certains travaux se sont intéressés aux conférences

scientifiques :

- Bartneck et Hu (2009) ont présenté une étude scientométrique de la conférence CHI en informatique (*Computer-Human Interaction*). L'étude portait sur l'évolution d'un certain nombre d'indicateurs tels que le nombre d'articles, le nombre de pages par article, le nombre d'auteurs ou encore le h-index des organisations associées aux publications ;
- Sakr et Alomari (2012) se sont intéressés aux conférences prestigieuses dans le domaine des bases de données, soulignant au passage l'intérêt de ce type de conférences pour les chercheurs en informatique. L'étude portait sur l'évolution des comités de programme des quatre plus prestigieuses conférences du domaine en termes de taille des comités notamment en rapport avec le développement de la communauté, en termes de chevauchement des comités de programme ou de renouvellement de ces comités. Les résultats ont montré une forte dynamique de ces conférences durant la période étudiée ;
- Küngas et al. (2013) ont étudié des classements des conférences pour évaluer à quel point ils étaient fondés sur des critères objectifs, comme le taux d'acceptation. L'application de méthodes d'apprentissage suggère que le taux d'acceptation est un indicateur pertinent pour prédire le classement d'une conférence. Il est cependant préférable de le combiner à des indicateurs bibliométriques (comme le nombre de citations des articles de la conférence) pour identifier les conférences les mieux classées ;
- dans le cadre d'un volume sur les contributions séminales à l'ingénierie des systèmes d'information, Jarke, Pham et Klamma (2013) font montre de réflexivité en analysant l'évolution de la communauté CAiSE (*Conference on Advanced Information Systems Engineering*) à l'aune des thématiques et des réseaux de cosignataires ;
- Kergosien, Bessagnet, Sallaberry, Le Parc-Lacayrelle et Royer (2016) ont davantage orienté leur étude de l'évolution de la communauté nationale EGC suivant des aspects spatio-temporels liés aux affiliations des auteurs.

4.3 De la scientométrie à l'informatique

Un certain nombre de travaux avec des objectifs très différents s'appuient sur des préoccupations ou des contributions issues du domaine de la scientométrie.

Y. Chen, Wei, Wu et Hu (2006) se sont par exemple focalisés sur la recherche de documents similaires au sein d'une base bibliographique telle le *Web of Science*. Une mesure de similarité entre documents est proposée, prenant en compte à la fois le texte et les citations des notices bibliographiques. Les listes de références des documents sont considérées comme des listes d'items pondérés. La similarité entre deux documents prend en compte les références communes.

Zhuang, Elmacioglu, Lee et Giles (2007) indiquaient le rôle majeur des conférences dans le domaine de l'informatique (voir aussi J. Chen & Konstan, 2010 ; Freyne, Coyle, Smyth & Cunningham, 2010). Face à la prolifération des conférences, un constat était qu'il devenait de plus en plus difficile d'évaluer la qualité d'une conférence. L'article propose un ensemble d'heuristiques pour identifier automatiquement le niveau d'une conférence en collectant des caractéristiques relatives aux membres de comités de programme. Cette fouille se basait sur des indicateurs tels que le nombre moyen de publications ou de co-auteurs des membres de comités de programme, leur visibilité ou encore leur influence.

Wang, Tong et Zeng (2013) ont proposé une approche pour classer des articles suivant leur prestige estimé. L'approche se base sur un graphe hétérogène exploitant différents types d'information comme les citations, les auteurs, ainsi que les conférences et revues. Elle s'appuie sur une combinaison des algorithmes PageRank et HITS ainsi que deux stratégies de prise en compte du temps pour l'estimation du prestige futur des articles. Les expérimentations menées sur deux collections (*ArXiv KDD Cup* 2003 et CORA) ont montré des résultats encourageants.

Vasilescu, Serebrenik, Mens, van den Brand et Pek (2014) se sont intéressés aux conférences en génie logiciel, en proposant des indicateurs pour mesurer leur « bonne santé ». Il s'agissait de mesurer la stabilité de la communauté, l'ouverture aux nouveaux auteurs ou encore la représentativité du comité de programme. L'étude des conférences a montré des disparités entre les conférences sur certaines mesures.

Avin, Lotker, Peleg et Turkel (2015) ont étudié douze ACM/IEEE conférences pour vérifier s'il pouvait exister un biais dans la sélection des articles lié aux articles soumis par des collaborateurs passés des membres des comités de programme. L'approche s'est basée sur la construction d'un réseau social par édition de conférence, rassemblant les individus impliquées dans l'édition de la conférence (auteurs et membres de comité de programme). Les analyses effectuées pour les douze conférences ont montré globalement une certaine équité, même si des biais étaient ponctuellement constatés.

Moreira et al. (2015) ont proposé deux approches pour combiner un ensemble d'estimateurs d'expertise issus de trois sources de preuves. Une contribution soulignée était l'utilisation de différents indicateurs scientométriques, notamment dérivés du Hirsch (2005) index, tels que le *Contemporary Hirsch index* et le *Trend Hirsch index*.

Ribas et al. (2015) se sont focalisés sur l'estimation de la réputation d'entités, telles que les personnes ou les organisations. Ils proposent un modèle de marche aléatoire pour classer des entités à partir de sources de réputation appropriées. Les sources et cibles sont les nœuds d'un graphe hétérogène, connectés par des liens traduisant les transferts de réputation. Les expérimentations menées sur une collection spécifique construite autour de la base gouvernementale brésilienne CAPES [1] ont montré une meilleure efficacité (*effectiveness*) et robustesse de l'approche par rapport à des approches basées sur les citations pour identifier les lieux de publications et chercheurs réputés.

Loudcher, Jakawat, Morales et Favre (2015) présentent une revue de la littérature OLAP (*Online analytical processing*) à des fins d'analyse bibliométrique. Une approche basée sur des graphes enrichis par des cubes de données (liées à des entrepôts de données) permet de réaliser des analyses à différents niveaux de granularité : auteurs, affiliations, années, notamment (Jakawat, Favre & Loudcher, 2016).

Plus récemment, Pradhan, Paul, Maheswari, Nandi et Chakraborty (2017) ont défini une nouvelle mesure de performance au niveau auteur basée sur l'algorithme de PageRank. La mesure combine les effets des citations et des collaborations d'un auteur en utilisant un réseau pondéré multi-niveau pour classer les auteurs.

Une approche pour identifier les chercheurs influents est proposée dans (de La Robertie, Pitarch, Takasu & Teste, 2017). L'approche se base sur un graphe hétérogène regroupant chercheurs, articles et conférences et relations de signature, co-signature et citation. Le graphe est combiné à une ressource extérieure d'indicateurs bibliométriques, notamment sur le « prestige » des conférences. L'approche applique un principe de renforcement mutuel entre les différents types d'éléments du graphe par propagation de score. Les expérimentations menées sur la collection *Microsoft Academic Search* [2] (Sinha et al., 2015) ont montré une meilleure efficacité (*effectiveness*) de l'approche par rapport à différents variantes basées sur l'algorithme de HITS.

1. http://www.capes.gov.br
2. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

4.4 Suggestion de membre de comités de programme

L'intérêt d'aider à l'élaboration ou au renouvellement des comités de programme de conférences a récemment émergé. À notre connaissance, peu de travaux se sont jusqu'ici focalisés sur cette problématique.

Han, Jiang, Yue et He (2013) ont étudié la possibilité de recommander des membres de comités de programme en combinant une approche basée contenu pour la recherche d'expert et des indicateurs complémentaires liés à l'historique de publication des candidats, leur proximité sociale et leur notoriété (*authority*). Les résultats des expérimentations ont montré que l'historique de publication et la proximité sociale sont de bons indicateurs, contrairement à l'autorité. Ces travaux apportent un éclairage sur certains indicateurs appropriés pour la suggestion de membres de CP, dont s'inspire l'approche présentée dans cette thèse.

Sfyris, Fragkos et Doulkeridis (2016) se sont intéressés à la sélection d'experts répondant à un ensemble de critères, tels que la couverture d'un ensemble de compétences afin de pourvoir un comité d'une organisation ou couvrir les thématiques d'une conférence. L'approche proposée construit une collection de profils textuels de candidats et sélectionne un groupe d'experts suivant un processus de RI. Bien qu'expérimenté dans le contexte des comités de programme de conférences, les CP suggérés par l'approche proposée n'ont pas été confrontés aux CP réels. Les évaluations reposent sur des mesures de qualité des chercheurs tels que le h-index, mais dont la pertinence est contestée (Bornmann & Marx, 2011 ; Waltman & van Eck, 2012). De plus, cette approche est uniquement basée sur des descriptions textuelles.

4.5 Bilan

La scientométrie s'intéresse depuis longtemps au fonctionnement de la science, aux activités des chercheurs et par conséquent à leurs publications et aux lieux de publications. De nombreux indicateurs ont été définis pour répondre à différentes questions. Certains indicateurs ont été utilisés dans le but d'établir des classements de chercheurs, d'institutions ou de lieux de publications. L'utilisation de ces indicateurs dans ce contexte reste cependant très délicate. La problématique à laquelle s'intéresse la présente thèse est différente puisqu'elle ne vise pas le classement de chercheurs, de publications ou de conférences et revues. L'objectif est d'identifier les critères qui peuvent

expliquer la composition de comités de programme et, potentiellement, aider à leur renouvellement.

À notre connaissance peu de travaux se sont jusqu'ici focalisés sur cette problématique. Aider à l'élaboration ou au renouvellement des comités de programme de conférences n'a suscité l'intérêt de la communauté scientifique que récemment. Ces travaux proposent des approches basées principalement sur l'analyse d'informations textuelles.

Différents travaux de la littérature, notamment en lien avec la scientométrie ont montré l'intérêt d'exploiter les données bibliographiques liées aux activités des chercheurs. C'est pour cette raison que l'approche présentée dans les chapitres suivants vise à exploiter ce type d'information pour aider à la composition et au renouvellement des comités de programme des conférences.

Deuxième partie

Contribution à la suggestion de membres de comités de programme de conférences

Résumé (Contribution).

Cette thèse se focalise sur la tâche de suggestion de membres de comité de programme (CP) pour des conférences scientifiques. Cette partie II fait l'objet de trois chapitres :

- le chapitre 5 propose une modélisation basée sur un graphe hétérogène pondéré de l'expertise scientifique multifacette liée aux activités de publication et au rayonnement scientifique d'un chercheur ;
- le chapitre 6 définit des indicateurs scientométriques pour étudier la constitution de CP passés et quantifier l'influence de chaque critère dans la participation de chercheurs au CP ;
- le chapitre 7 propose une approche de suggestion de CP pour une conférence donnée, en combinant les résultats des indicateurs scientométriques définis. Cette approche vise notamment à aider au renouvellement des CP.

Des expérimentations ont été menées pour une des conférences de premier plan de notre communauté de recherche : SIGIR (*Special Interest Group on Information Retrieval*), en considérant ses éditions de 1971 à 2015, ainsi que les conférences proches thématiquement, telles que CIKM, ECIR, WSDM et WWW.

Chapitre 5

Modélisation de la sphère académique des conférences

Sommaire

5.1 Motivation et problématique

La présente thèse s'intéresse à la constitution des comités de programme des conférences. Les conférences constituent une cible de publication privilégiée par les chercheurs de différentes disciplines. Différents travaux ont souligné l'importance de certaines conférences phares pour le domaine de l'informatique (Bar-Ilan, 2010 ; J. Chen & Konstan, 2010 ; Freyne et al., 2010).

L'hypothèse faite dans cette thèse est que chaque membre d'un comité de programme d'une conférence est lié à cette conférence.

L'activité principale d'un chercheur se traduit par la publication d'articles. Publier un article dans une conférence constitue une preuve évidente d'un lien direct entre cette conférence et les co-signataires de cet article. La participation à l'organisation d'une conférence est une autre forme de preuve directe d'un lien existant entre cette conférence et le chercheur impliqué dans son organisation.

Au-delà des liens directs qui peuvent exister entre un chercheur et une conférence, d'autres preuves de liens, plus indirectes, peuvent être identifiées. Cela apparaît, par exemple, lorsqu'un chercheur fait référence à un article de la conférence dans une de ses publications.

Une première tâche consiste donc à recenser les preuves de liens existant entre les chercheurs et les conférences. Cette tâche peut s'appuyer sur les données issues des conférences elles-mêmes — mais celles-ci ne sont souvent pas accessibles librement. Un certain nombre d'informations sont toutefois disponibles via des bases bibliographiques qui peuvent exister, telles que l'*ACM Digital Library*[1] en informatique. Combiner différentes sources complémentaires permet d'enrichir les preuves de liens entre chercheurs et conférences.

L'analyse des comités de programme de conférences au travers des liens existant entre leurs membres et les conférences nécessite une modélisation appropriée. La modélisation sous forme de graphe est une représentation qui facilite l'exploitation des relations existant entre des objets. Différents travaux présentés dans les chapitres précédents se sont basés sur ce type de modélisation. Cependant, la plupart de ces travaux considère un seul type de relation entre un seul type d'objet : les liens de citations entre articles, par exemple. L'étude sur laquelle se focalise cette thèse requiert un graphe plus riche, modélisant plusieurs types de relations. Une première contribution réside donc dans la proposition d'une modélisation des relations entre conférences et chercheurs sous forme de graphe multi-parti. Cette modélisation est réalisée à deux niveaux : au niveau d'une conférence, puis au niveau du domaine de la conférence intégrant les autres conférences du domaine.

5.2 Sources de preuves de liens entre conférences et chercheurs

Les activités des chercheurs en lien avec les conférences sont visibles au travers de différentes sources d'information.

Les informations directement produites par les conférences — comme leurs pages Web et leurs actes (*proceedings*) — regroupent des preuves « officielles » de liens entre chercheurs et conférences. Les actes d'une conférence listent notamment les noms des

1. http://dl.acm.org

personnes membres du comité de programme et du comité d'organisation, en complément des différents articles publiés avec leur contenu textuel et des informations sur les auteurs et leurs affiliations. Cependant, notons que les actes de conférences ne sont généralement pas accessibles librement, ce qui empêche l'exploitation du contenu textuel des articles publiés.

Les pages Web personnelles ou institutionnelles des chercheurs peuvent également révéler un certain nombre de liens entre les chercheurs et les conférences. Cependant, l'existence de ces pages est inégale et les informations fournies sont très souvent de nature hétérogène. Il est ainsi difficile d'obtenir les mêmes informations pour l'ensemble des chercheurs par ce biais.

Les bases bibliographiques constituent des sources d'information communément exploitées par les travaux de la littérature relatifs aux activités des chercheurs. Certaines bases sont multi-disciplinaire comme par exemple le *Web of Science*[2], d'autres étant focalisées sur une discipline particulière comme la base *DBLP*[3] relative à l'informatique. L'intérêt de ces bases est qu'elles proposent un format homogène, différentes informations et une bonne couverture d'une discipline. Cependant, certaines bases ne sont pas accessibles librement, ce qui peut constituer un frein à l'exploitation de certaines informations. La combinaison de différentes bases bibliographiques permet également d'obtenir davantage d'informations consolidées. Cette thèse s'appuie sur une hypothèse raisonnable d'accès à un certain nombre de bases bibliographiques accessibles librement et fournissant des informations complémentaires. Cette hypothèse a été vérifiée dans le contexte de la discipline informatique. Les informations exploitées dans cette thèse sont ainsi :

— les éditions des conférences,
— les articles qu'elles ont publiés,
— les citations entre articles,
— les auteurs des articles et
— les compositions de leurs comités de programme.

La modélisation issue de ces informations et décrite dans les sections suivantes demeure extensible pour intégrer des informations supplémentaires.

2. http://isiknowledge.com/wos
3. http://dblp.org

5.3 Modélisation

Cette thèse se focalise sur les conférences et leurs comités de programme, sur lesquels repose la sélection des articles qu'elles publient. Elle vise à connaître les relations entre une conférence et un chercheur qui peuvent expliquer sa participation à un comité de programme.

Les relations entre un chercheur et une conférence peuvent être multiples et diverses. Une autre activité des chercheurs vis-à-vis des conférences concerne l'organisation de celles-ci. Le comité de programme joue un rôle essentiel dans le déroulement d'une conférence. Participer à un comité de programme est donc une relation importante entre un chercheur et une conférence.

Comme dans de nombreux travaux de la littérature, l'activité de publication est considérée comme une des activités principales des chercheurs. Divers « lieux de publication » sont recensés, comme les revues, les conférences ou encore les livres. L'importance accordée à ces lieux de publication dépend de la discipline, du domaine au sein de la discipline, voire même du chercheur lui-même. L'informatique accorde, par exemple, une grande importance aux conférences contrairement à d'autres disciplines qui privilégient les revues. Publier un article dans une conférence révèle une relation directe entre un chercheur et cette conférence via l'article publié. D'autres relations peuvent exister entre un chercheur et une conférence au travers des publications d'articles. Par exemple, un chercheur peut écrire un article publié dans une revue et faire référence dans cet article à un article publié dans une conférence. Ce principe indique alors une relation d'un autre type entre le chercheur et la conférence, via la notion de citation entre articles.

Les objets manipulés pour représenter les relations entre chercheurs et conférences sont par conséquent de trois types :

1. les conférences, et plus précisément les éditions des conférences,
2. les articles publiés,
3. les chercheurs qui ont signé ces articles et ceux qui ont participé à un comité de programme des conférences.

Pour répondre à la problématique de représentation de trois types d'objets et des relations existant entre ces objets, nous avons opté pour une modélisation des données sous forme de graphe qui représente naturellement des relations entre des objets.

De nombreux travaux de la littérature se basent sur une modélisation des données sous forme de « graphe simple », c'est-à-dire de graphe composé d'un seul type de nœud et d'un seul type d'arête, tels que les graphes de citations entre articles (Y. Liang et al., 2011). Certains travaux, en revanche, exploitent des graphes bipartis, c'est-à-dire composés de deux types de nœuds et d'un type d'arête. Serdyukov, Rode et Hiemstra (2008) exploitent, par exemple, des graphes représentant des liens de signature entre des nœuds articles et des nœuds auteurs pour la découverte d'expert.

Dans la même idée et afin de prendre en compte les trois types d'objets manipulés dans cette thèse, le type de graphe choisi est un graphe triparti. Ces trois types d'objets correspondent aux trois types de nœuds existant dans un graphe.

Comme mentionné précédemment, plusieurs relations entre chercheurs et conférences peuvent être considérées. Compte tenu des trois types d'objets modélisés, ces relations se basent sur quatre types de liens entre ces objets, comme illustré dans la figure 5.1.

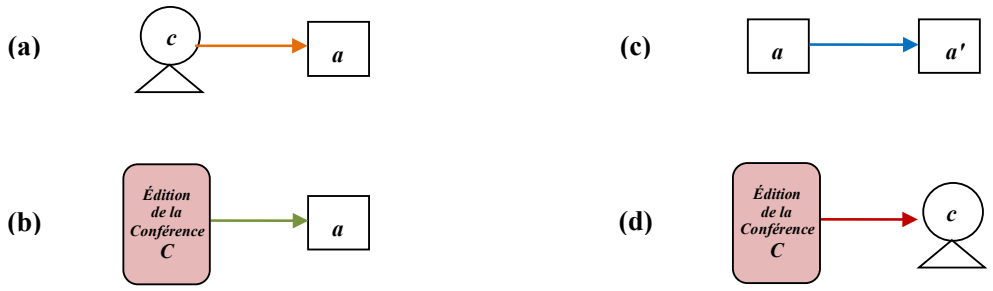

FIGURE 5.1 – Les quatre types de liens considérés entre les trois objets modélisés. (a) Lien de signature entre un chercheur c et l'article a dont il est auteur. (b) Lien de publication entre l'édition C d'une conférence et l'article a dans laquelle il est publié. (c) Lien de citation entre un article a qui cite un article a'. (d) Lien de participation au comité de programme entre le chercheur c et l'édition C d'une conférence.

Le lien de signature entre un chercheur et un article représente le fait que le chercheur apparaît dans la liste des auteurs de l'article. De multiples liens entre différents chercheurs permettent de déduire des relations de co-signature entre chercheurs.

Le lien de publication entre une édition de conférence et un article représente le fait que l'article a été publié dans l'édition donnée de la conférence.

Le lien de citation entre deux articles représente le fait qu'un article cite l'autre. Ces liens permettent des relations de citations entre articles et permettent d'inférer une

proximité conceptuelle entre les articles via l'identification de co-citations et le couplage bibliographique (Kessler, 1963 ; Small, 1973 ; Lawrence, Giles & Bollacker, 1999).

Le lien de participation au comité de programme représente le fait qu'un chercheur a été membre du comité de programme de l'édition de la conférence. Ce type de lien est à notre connaissance inexploité dans la littérature à ce jour. Il est cependant *a priori* indispensable pour la problématique de cette thèse, relative à la composition du comité de programme d'une édition de la conférence.

Les différentes relations existant entre un chercheur donné et une édition donnée d'une conférence sont représentées par les chemins reliant le nœud correspondant au chercheur et le nœud correspondant à l'édition de la conférence. Chaque chemin est composé de différents types de liens reliant différents types de nœuds intermédiaires suivant la relation qu'il traduit.

Cette modélisation peut être appliquée à différents niveaux. Nous l'avons appliquée, dans cette thèse, au niveau d'une conférence et au niveau du domaine dans lequel s'inscrit une conférence.

5.3.1 Modèle de conférence

Le modèle d'une conférence est basé sur trois types de nœuds :

1. les éditions de la conférence,
2. les articles liés aux éditions de la conférence,
3. les chercheurs liés à la conférence, soit en tant qu'auteur, soit en tant que membre du comité de programme.

Les relations considérées entre les chercheurs et une conférence donnée, illustrées dans la figure 5.2 sont :

1. *membre interne* lorsque le chercheur a été membre du comité de programme d'une édition de la conférence,
2. *auteur interne* lorsque le chercheur est auteur d'un article qui a été publié dans les actes d'une édition de la conférence,
3. *auteur externe citant* lorsque le chercheur est auteur d'un article publié en dehors de la conférence qui cite un article publié dans une édition de la conférence,

4. *auteur externe cité* lorsque le chercheur est auteur d'un article publié en dehors de la conférence, lui-même cité par un article publié dans une édition de la conférence.

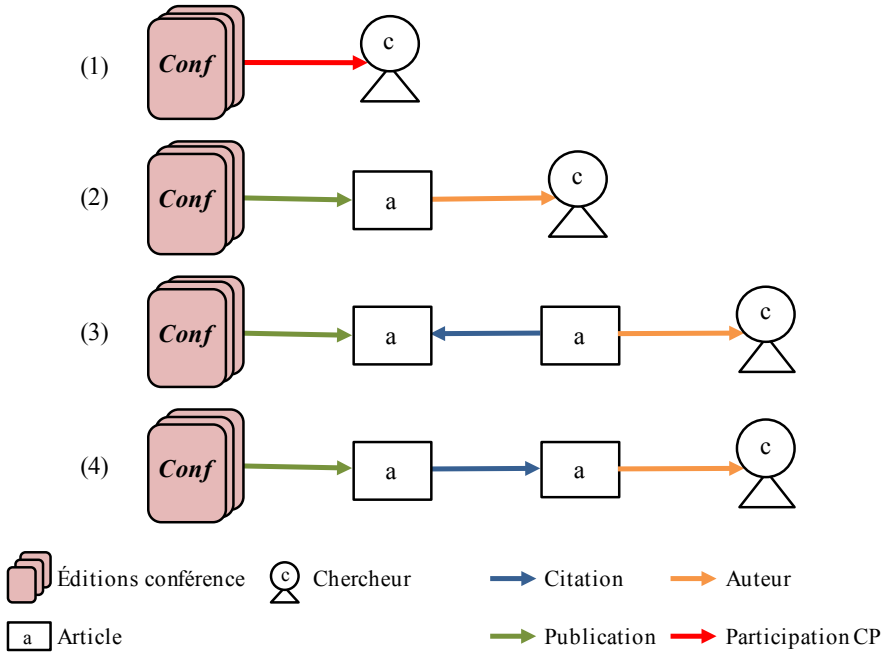

FIGURE 5.2 – Relations considérées entre les chercheurs et une conférence.

Comme indiqué précédemment, le principe de modélisation peut s'appliquer à différents niveaux. Les objets et relations pris en compte dépendent donc du niveau de modélisation considéré. Pour construire le modèle d'une conférence, les éléments pris en compte sont :

1. les éditions de la conférence,
2. les chercheurs ayant participé à au moins un comité de programme de la conférence,
3. les articles publiés dans les éditions de la conférence,
4. les chercheurs ayant signé les articles publiés dans la conférence,
5. les articles cités par les articles publiés dans les éditions de la conférence et les articles citant les articles publiés dans les éditions de la conférence,
6. les chercheurs ayant signé les articles cités ou citant.

La figure 5.3 décrit la progression de la prise en compte des objets constituant le graphe d'une conférence. La figure 5.4, quant à elle, montre un exemple de graphe représentant une conférence.

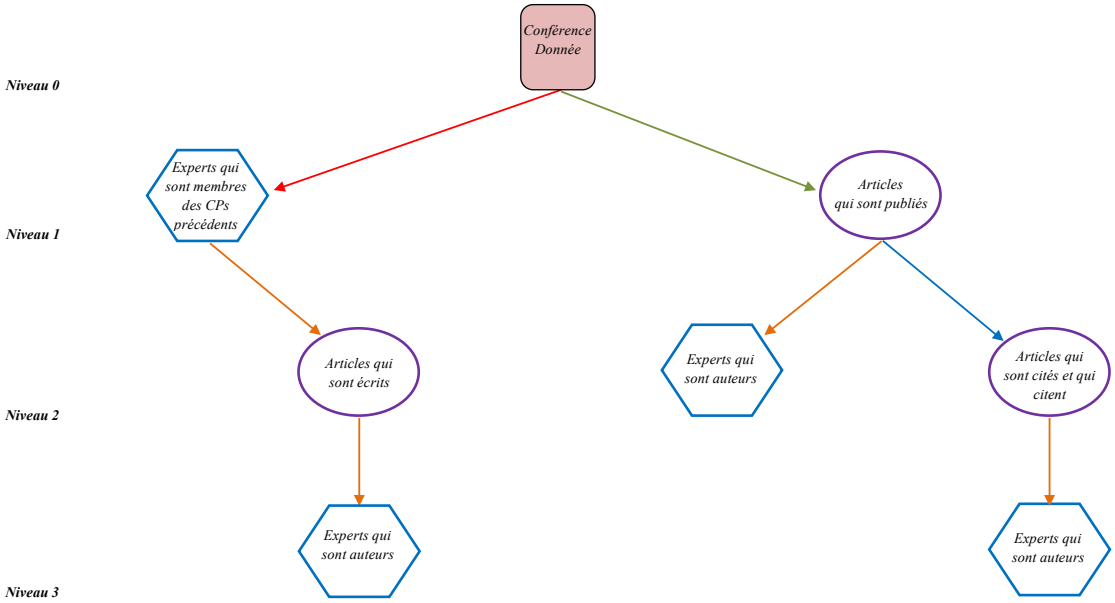

FIGURE 5.3 – Information considérées pour la construction d'un modèle de conférence.

5.3.2 Modèle de domaine scientifique

Chaque conférence est liée à un ou plusieurs domaines scientifiques. Un domaine scientifique rassemble donc plusieurs conférences qui partagent un certain nombre de thématiques de recherche. Les chercheurs qui s'intéressent à un domaine scientifique particulier soumettent des articles dans les différentes conférences qui sont rattachées à ce domaine. Ils sont donc, *a priori* pour nombre d'entre eux, les auteurs des articles publiés dans ces conférences.

Nous appelons « conférences proches » celles d'un domaine scientifique partageant certaines thématiques relatives à ce domaine-là. Ce partage implique par conséquent un lien entre conférences proches. Les chercheurs en relation avec les conférences proches d'une conférence donnée sont donc en relation également avec cette conférence donnée.

Selon la modélisation proposée dans ce chapitre, chacune des conférences proches

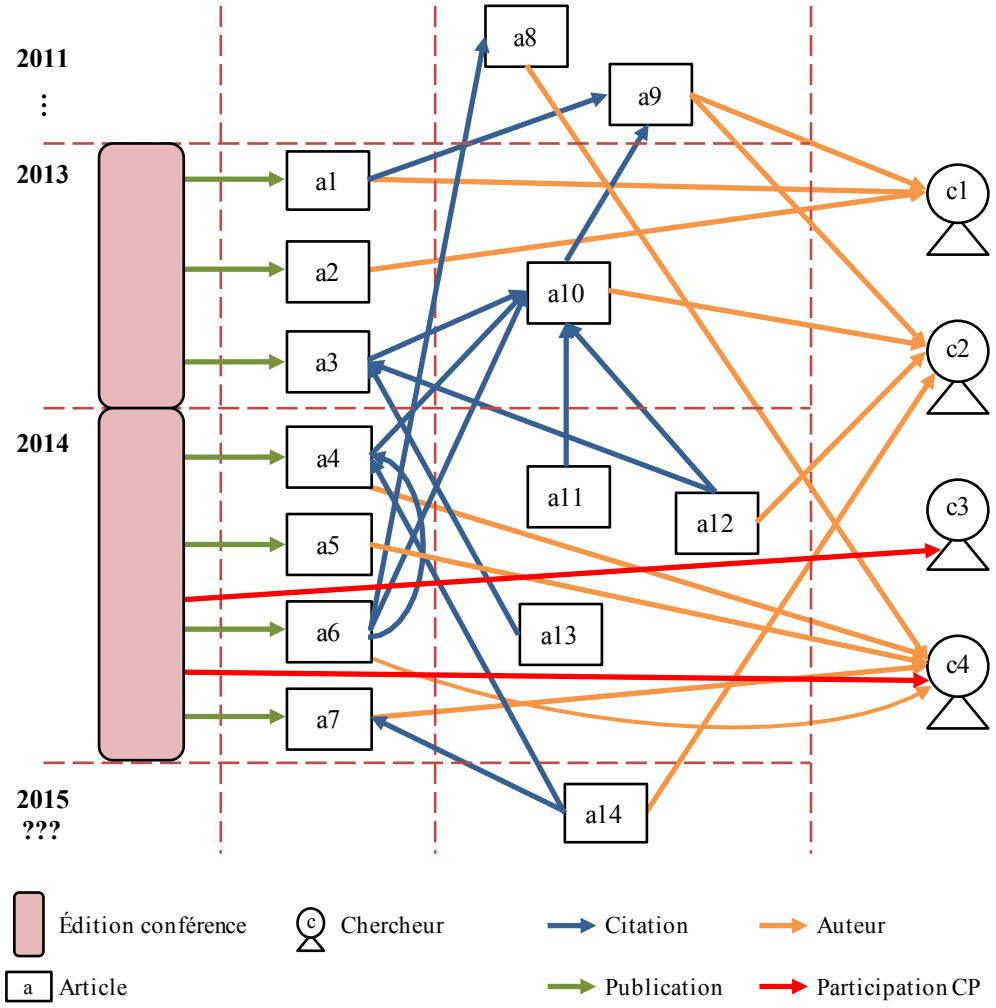

FIGURE 5.4 – Exemple de modèle de conférence sous forme de graphe triparti.

peut être modélisée par un graphe triparti. Ces graphes peuvent exhiber des chercheurs ou des articles en commun. Les chercheurs en commun sont, par exemple, ceux qui ont signé des articles dans les deux conférences ou qui ont participé à des comités de programme des éditions des deux conférences. Comme illustré en figure 5.5, les conférences proches sont ainsi connectées au travers de trois types de relations :

1. un article d'une conférence donnée cite un article de la conférence proche ou bien il est cité par un article de la conférence proche,
2. un chercheur auteur d'un article d'une des deux conférences proches a égale-

ment participé au comité de programme de l'autre conférence,

3. un chercheur a été membre de comités de programme des deux conférences proches.

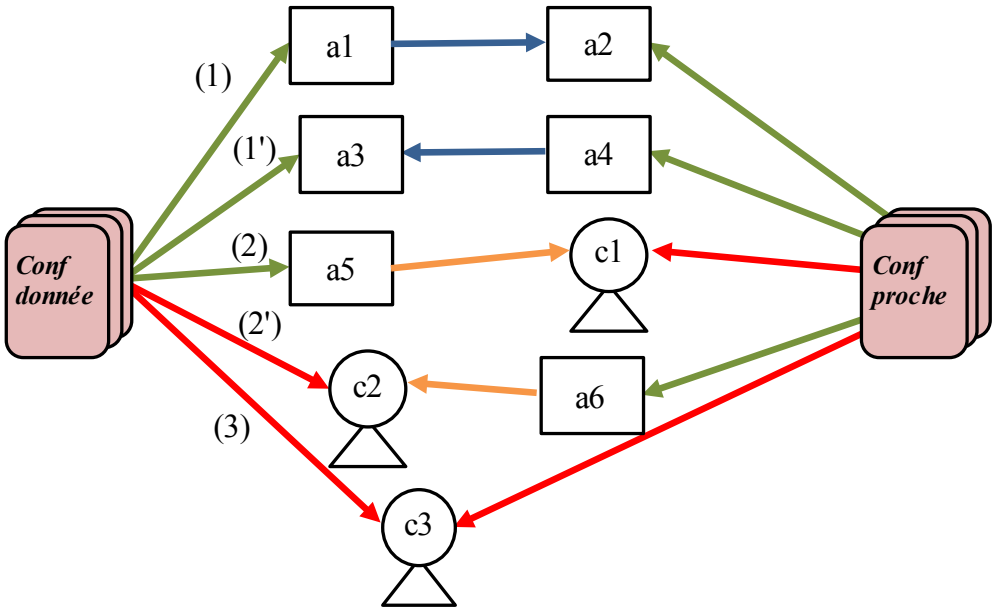

FIGURE 5.5 – Types de relations entre conférences proches.

La fusion des différents graphes modélisant les conférences proches forme ainsi un graphe modélisant la sphère académique de ces conférences. Dans un tel graphe, toutes les conférences sont équivalentes. La présente thèse se focalise sur l'étude d'une conférence au sein d'un domaine scientifique. Elle privilégie donc la partie du graphe de domaine centrée sur cette conférence.

Pour réduire la complexité introduite par les différents types de relations manipulées et, plus particulièrement, les relations entre les graphes de conférences proches, nous avons choisi de fusionner ces dernières dans une seule relation, appelée relation inter-conférence, comme illustrée par le figure 5.6.

De plus, le graphe d'une conférence proche est réduit pour ne prendre en compte que les chercheurs directement en relation avec celle-ci, comme illustré par la figure 5.7. Les deux types de relations considérés sont alors : 1) le chercheur est auteur d'un article publié dans la conférence proche et 2) le chercheur a été membre de l'un des comités de programme de la conférence proche.

Le graphe réduit représentant un domaine ainsi obtenu est illustré par la figure 5.8.

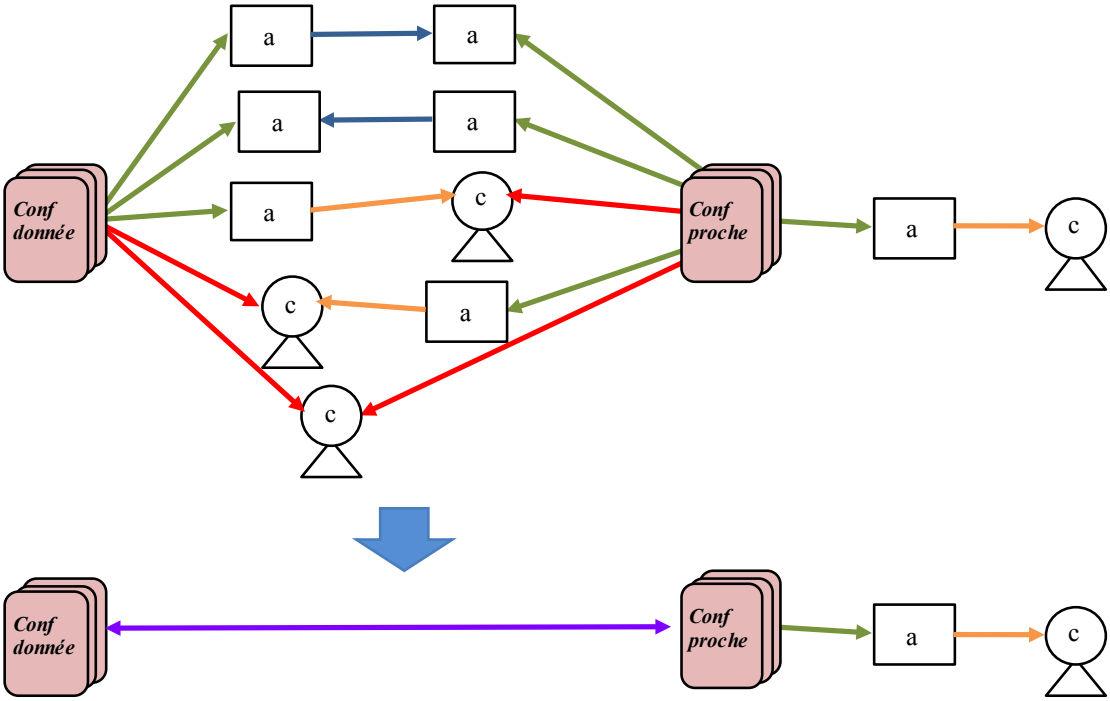

FIGURE 5.6 – Fusion des relations entre conférences proches en une seule relation inter-conférence.

5.4 Bilan

L'étude de la composition des comités de programme — à laquelle nous nous intéressons dans la présente thèse — sous-tend une modélisation adaptée des données manipulées. Cette étude est basée sur les relations entre les chercheurs et les conférences. Ces relations correspondent aux activités de publication des chercheurs et aux activités de participation aux comités de programme. Les objets manipulés sont ainsi de trois types : édition de conférence, article et chercheur.

Une modélisation sous forme de graphe nous est apparue particulièrement adaptée, compte tenu de l'exploitation de relations entre objets sous-jacente. De nombreux travaux de la littérature s'appuient sur une modélisation sous forme de graphe. Cependant, un grand nombre de ces travaux exploite des graphes « simples », composés d'un seul type de nœud et d'un seul type de lien. Certains exploitent des graphes biparti représentant les relations entre documents et auteurs. Dans notre contexte, nous avons opté pour une modélisation via des graphes tripartis, compte tenu des trois types d'objets manipulés.

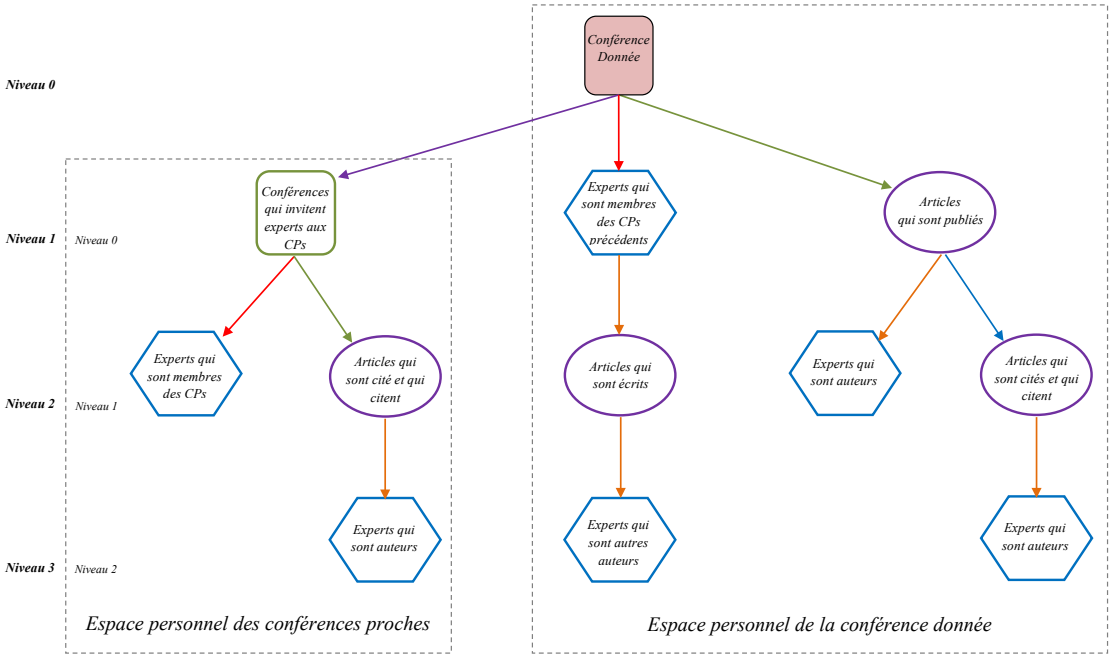

FIGURE 5.7 – Prise en compte des objets pour la construction d'un modèle de domaine.

La modélisation est applicable à différents niveaux suivant l'étendue des relations considérées entre les chercheurs et les conférences. La modélisation a été appliquée au niveau d'une conférence et au niveau d'un domaine regroupant plusieurs conférences proches, c'est-à-dire partageant des thématiques du domaine. Pour limiter la complexité du modèle de domaine, une méthode de réduction du modèle de domaine a été conçue et présentée.

Ce modèle « réduit » de domaine sert de base aux contributions présentées dans la partie suivante, à savoir, une analyse des relations expliquant la composition de comités de programme, *via* la définition d'indicateurs scientométriques, puis une méthode de suggestion de membre de comité de programme.

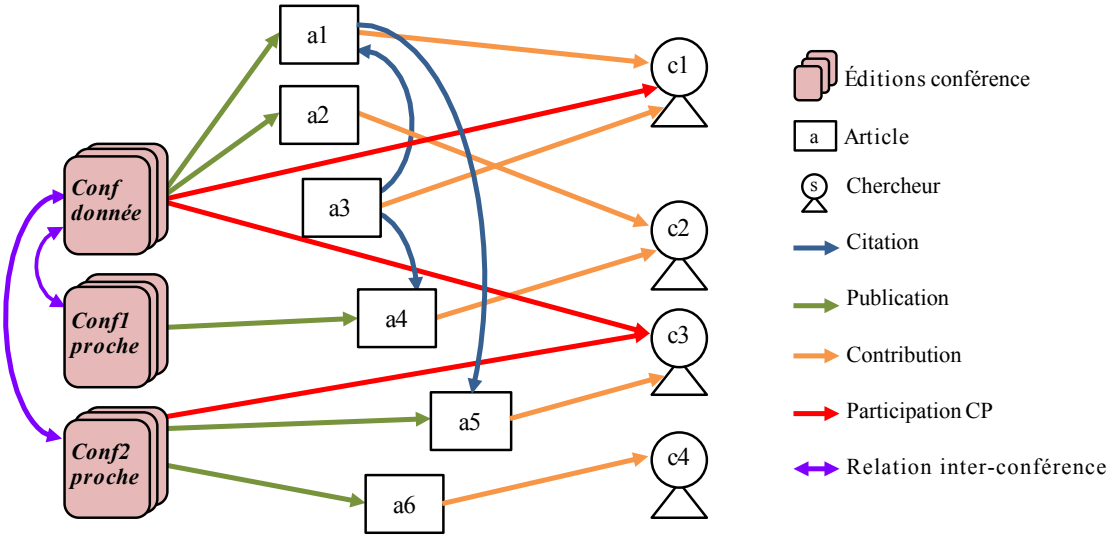

FIGURE 5.8 – Exemple de modèle réduit de domaine scientifique.

Chapitre 6

Étude de la composition des comités de programme

Sommaire

6.1 Motivation et problématique

La sélection des articles pour constituer le programme d'une conférence repose généralement sur les résultats du processus d'évaluation par les pairs, qui est géré par les membres du comité de programme de la conférence. Les compositions des comités de rédaction des revues, qui constituent une cible de publication importante pour les chercheurs, sont plutôt stables dans le temps. En revanche, les comités de programme de conférence (CP) sont habituellement renouvelés chaque année par les président·e·s des CP selon des critères idiosyncrasiques et non divulgués. Ici, le défi consiste à identifier les chercheurs actifs dans les multiples sujets abordés par la conférence pour les inviter à rejoindre le (nouveau) CP de l'édition à venir de la conférence. C'est d'autant plus difficile pour les grandes conférences, telles que CIKM, SIGIR ou WWW en

informatique, par exemple. Celles-ci attirent des centaines de soumissions d'articles et impliquent ainsi des centaines d'évaluateurs (Cabanac & Preuss, 2013). L'automatisation des tâches liées à l'organisation de la conférence est encore un problème de recherche ouvert (S. Price & Flach, 2017).

Les travaux décrits dans ce chapitre visent à identifier un ensemble d'indicateurs d'expertise susceptibles d'être considérés dans la décision d'inviter un chercheur à siéger dans un comité de programme d'une conférence. Contrairement aux contextes applicatifs habituels liés à la découverte d'expertise, un appariement basé sur le contenu des bases de données bibliographiques et les descriptions des thématiques de la conférence (par exemple, l'appel à communications) paraît inenvisageable en raison de l'étendue et du manque de spécificité de ces descriptions. C'est pourquoi les indicateurs conçus dans cette thèse ne sont pas basés sur des contenus textuels, mais sur des réseaux d'information (par exemple, auteurs, données de citation, comité de programme) qui traduisent les relations directes ou indirectes des chercheurs avec une conférence.

La contribution de ce chapitre est la définition de six indicateurs d'expertise basés sur la modélisation du domaine d'une conférence présentée dans le chapitre précédent. Ces indicateurs se basent sur les différents rôles que jouent les chercheurs vis-à-vis d'une conférence pour quantifier chaque type de relation que peut avoir un chercheur avec une conférence. Nous avons mené des expérimentations dans le cadre des éditions de la conférence SIGIR pour révéler les indicateurs habituellement impliqués dans la composition des CP et ceux jusqu'ici délaissés mais qui pourraient servir à suggérer de nouveaux membres de CP à solliciter.

6.2 Indicateurs scientométriques de rôles d'un chercheur vis-à-vis d'une conférence

En se basant sur la modélisation proposée au chapitre précédent, chaque rôle joué par un chercheur est représenté par un chemin reliant un chercheur et la conférence. Pour chacun de ces rôles, nous définissons un indicateur en se basant sur la notion de chemin dans le graphe hétérogène. Préalablement à la définition d'indicateurs associés aux différents rôles d'un chercheur nous définissons une pondération pour les différents types de liens constituant le graphe modélisant le domaine d'une conférence.

6.2.1 Pondération des liens du graphe de domaine

Le modèle de domaine proposé au chapitre précédent est basé sur une représentation sous forme de graphe composé de trois types de nœuds (éditions de conférences, articles et chercheurs) et de quatre types de liens (publication d'un article dans une édition d'une conférence, citation entre articles, signature d'un article par un chercheur et participation d'un chercheur au comité de programme d'une édition d'une conférence).

Dans l'objectif de distinguer les implications des chercheurs vis-à-vis d'une conférence au travers de la définition d'un ensemble d'indicateurs scientométriques, il paraît pertinent de pondérer les liens du graphe de domaine. Cette pondération se base sur différentes propriétés des nœuds reliés.

Nous définissons donc quatre pondérations correspondant aux quatre types de liens inhérents au graphe et modélisant les données relatives aux conférences :

1. lien de publication d'un article par une conférence,
2. lien de citation entre articles,
3. lien de signature d'un article par un chercheur,
4. lien de participation d'un chercheur à un comité de programme.

Ces pondérations, spécifiques à notre étude, sont fondées sur différentes hypothèses faites sur l'importance accordée aux objets impliqués dans les relations entre chercheurs et conférences.

De plus, comme introduit dans le chapitre précédent, nous considérons le graphe de domaine simplifié, dans lequel les relations entre une conférence donnée et une conférence proche sont traduites au travers d'un seul lien inter-conférence. Ce lien est également pondéré en fonction des liens existant entre les deux conférences.

6.2.1.1 Poids du lien de *publication* entre un article et l'édition de la conférence qui le publie

Lors de l'année $Année_x$, le poids du lien de publication d'un article a publié lors de l'édition de l'année $Année_a$ de la conférence C, dénoté $Impact_{a,\, C_{Année_x}}$, traduit l'importance de l'article vis-à-vis de la conférence. Bien que tous les articles publiés dans une conférence puissent être considérés comme de même importance, différents travaux en scientométrie distinguent l'impact des articles pour une communauté. Les articles

les plus cités sont généralement considérés comme ayant davantage d'impact auprès de la communauté. Une première hypothèse faite dans cette thèse est que ces articles contribuent davantage à la notoriété et au prestige de la conférence. De plus, les articles récemment publiés sont porteurs, a priori, des thématiques actuelles d'une conférence. Une seconde hypothèse est donc que les articles récents doivent être privilégiés pour modéliser les conférences et les chercheurs.

Le poids d'un lien de publication dépend ainsi de deux autres facteurs :

1. du nombre d'articles citant l'article a, noté $Citations_a$. Parmi les articles qui sont publiés par la conférence, ceux qui sont les plus cités constituent un point d'entrée plus important vers la conférence. Ils contribuent à la visibilité et à l'intérêt pour la conférence ;
2. de l'écart de temps entre la publication de l'article a et l'édition de la conférence $Année_x$ pour laquelle on étudie la composition du comité de programme. Les articles les plus proches de l'édition $Année_x$ de la conférence traitent des thèmes les plus actuels, et par conséquent plus importants pour constituer le comité de programme à venir.

Ainsi, l'impact d'un article a publié lors de l'édition de l'année $Année_a$ de la conférence C pour la communauté lors de l'édition de l'année $Année_x$ est estimé par

$$Impact_{a,\, C_{Année_x}} = \frac{Citations_a}{Citations_{max}} \cdot e^{-\frac{Année_x - Année_a}{\Delta A_{max}}} \tag{6.1}$$

où $Citations_{max}$ est le nombre de citations reçues par l'article le plus cité et ΔA_{max} est l'écart de temps maximum en années entre l'article le plus ancien et l'édition de conférence considérée.

Par conséquent, plus la publication sera récente et citée, plus le poids du lien de publication sera élevé, comme l'illustre la Figure 6.1 a).

6.2.1.2 Poids du lien de *citation* entre un article citant et un article cité

Le poids du lien de citation entre deux articles $a_{cité}$ et a_{citant}, dénoté $Citation_{a_{citant}, a_{cité}}$, traduit l'importance de la citation entre les deux articles. Une première hypothèse est qu'un article est d'autant plus incontournable qu'il reçoit un grand nombre de citations. Une seconde hypothèse que le lien entre deux articles est d'autant plus fort qu'un laps

de temps long s'est écoulé entre les deux publications. Cette situation suggère que l'article cité, quoiqu'ancien, est toujours d'utilisé pour des recherches actuelles.

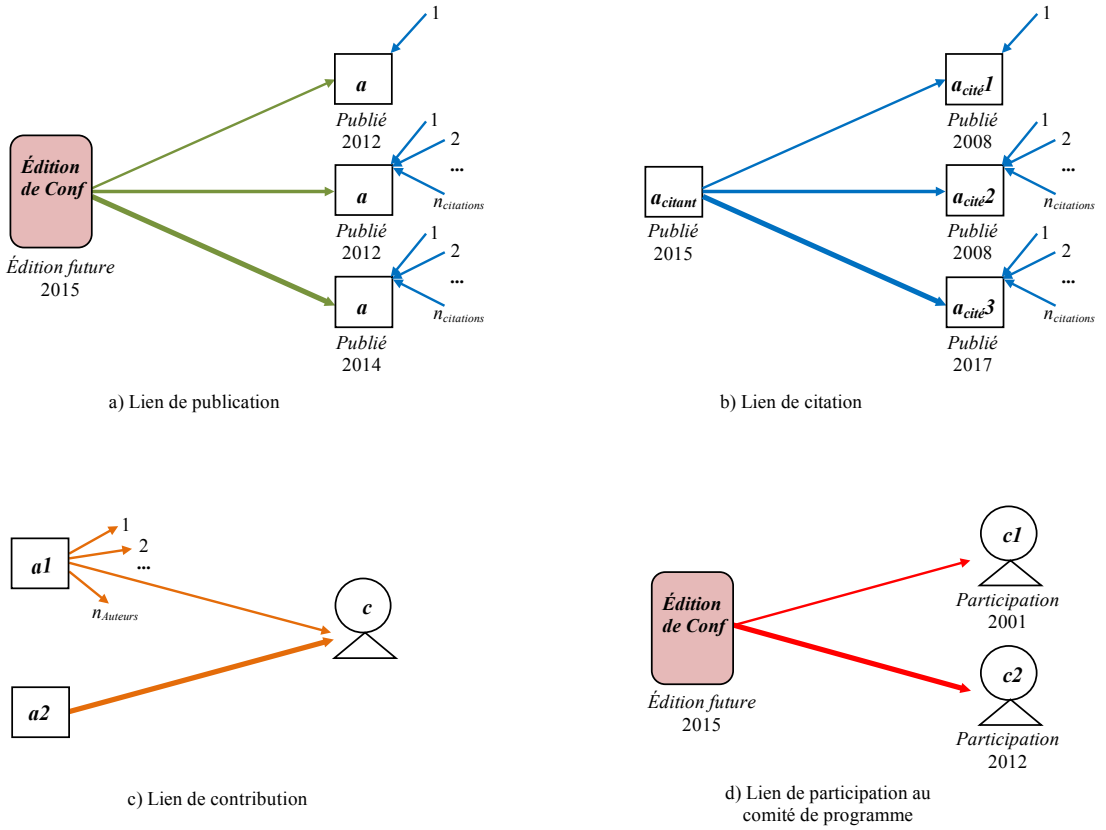

FIGURE 6.1 – Preuves d'expertise pour estimer la force des principaux liens.

Le poids d'un lien de citation dépend ainsi de deux facteurs :

1. de l'écart de temps entre la publication de l'article $a_{cité}$ et celle de l'article a_{citant}. Une citation vers un article publié il y a longtemps suggère que cet article est encore d'actualité. Ainsi, la force de la citation croît en fonction de l'écart entre les années de publication ;

2. du nombre d'articles qui citent $a_{cité}$. Outre la valorisation habituelle des articles très cités, c'est-à-dire qui reflètent un fort impact auprès de la communauté scientifique, il s'agit également de valoriser les articles citant ces articles très cités.

Le poids du lien de *Citation* entre l'article a_{citant} et l'article $a_{cité}$ est donc défini par

$$Citation_{a_{citant},\, a_{cité}} = \frac{Citations_{a_{cité}}}{Citations_{max}} \cdot e^{\frac{Année_{a_{citant}} - Année_{a_{cité}}}{\Delta A'_{max}}} \tag{6.2}$$

où $Citations_{max}$ est inchangé et $\Delta A'_{max}$ est le plus long écart de temps en années entre deux articles liés par une relation de citation. L'utilisation de l'exponentielle pour le facteur relatif à l'écart de temps permet d'accentuer les écarts entre les valeurs.

Par conséquent, plus l'article cité sera influant et ancien par rapport à l'article qui le cite, plus le poids du lien de citation sera élevé, comme l'illustre la figure 6.1 b).

6.2.1.3 Poids de la *Contribution* d'un chercheur à un article

Le poids du lien de contribution entre un article a et un chercheur c (qui fait partie de ses auteurs) traduit l'importance de la contribution de ce chercheur à cet article. Différentes hypothèses existent en scientométrie, comme par exemple, considérer une répartition égale des contributions des différents auteurs, ou bien privilégier la contribution du premier auteur (Kim & Kim, 2015). Dans cette thèse, compte tenu qu'il est impossible de quantifier la participation de chaque co-signataire d'un article à la lumière des seules méta-données d'une notice bibliographique (Kosmulski, 2012), l'hypothèse d'une répartition uniforme des contributions a été retenue.

Le poids de la contribution entre un article a et un chercheur c est défini suivant un seul facteur comme suit :

$$Contribution_{c,\, a} = \frac{1}{Auteurs_a} \tag{6.3}$$

où $Auteurs_a$ est le nombre de co-signataires de l'article (Van Hooydonk, 1997).

Par conséquent, moins il y aura de signataires sur un article, plus le poids du lien entre un article et un chercheur sera élevé, comme l'illustre la figure 6.1 c).

6.2.1.4 Poids du lien de *Participation* au comité de programme entre un chercheur et une conférence

Une participation au comité de programme d'une conférence indique qu'un chercheur est reconnu en lien avec des thématiques de la conférence. L'hypothèse faite dans cette thèse est que l'importance de cette reconnaissance s'érode avec le temps.

Pour l'édition $Année_x$ de la conférence C étudiée, le poids d'une participation au comité de programme de l'année $Année_p$ de C dépend ainsi de l'écart de temps entre les années :

$$ParticipationCP_{c,\, C_{Année_p},\, C_{Année_x}} = e^{-\frac{Année_x - Année_p}{\Delta A''_{max}}} \tag{6.4}$$

où $\Delta A''_{max}$ est l'écart de temps maximum en années entre l'édition de conférence considérée et le plus ancien comité de programme.

Plus le comité de programme auquel a participé le chercheur est récent, plus le poids du lien de participation sera élevé, comme l'illustre la figure 6.1 d).

6.2.1.5 Poids du lien *inter-conférence* entre deux conférences proches

Le poids du lien entre deux conférences traduit à quel point les deux conférences sont proches. Comme introduit dans le chapitre précédent, ce lien fusionne différentes relations qui existent entre les deux conférences, comme illustré par la figure 6.2 :

1. l'article a' publié par la conférence proche C' cite l'article a publié par la conférence étudiée C ;
2. l'article a' publié par la conférence proche C' est cité par l'article a publié par la conférence étudiée C ;
3. un chercheur c qui est un auteur d'un article de la conférence étudiée C a également participé à un comité de programme de la conférence proche C' ;
4. un chercheur c qui est un auteur d'un article de la conférence proche C' a également participé à un comité de programme de la conférence étudiée C ;
5. un chercheur c a participé à un comité de programme de la conférence étudiée C et à un comité de la conférence proche C'.

Le poids du lien inter-conférence agrège donc différents facteurs correspondant aux relations existant entre les deux conférences. Nous basons la similarité entre la conférence étudiée et la conférence proche sur le poids des cinq types de chemins représentant les relations considérées. Le poids de chaque chemin [1] est défini suivant les poids des liens composant ce chemin :

1. L'étoile ($\star$) dénote une normalisation par la valeur maximale du facteur pour obtenir l'ensemble des valeurs dans l'intervalle $[0, 1]$.

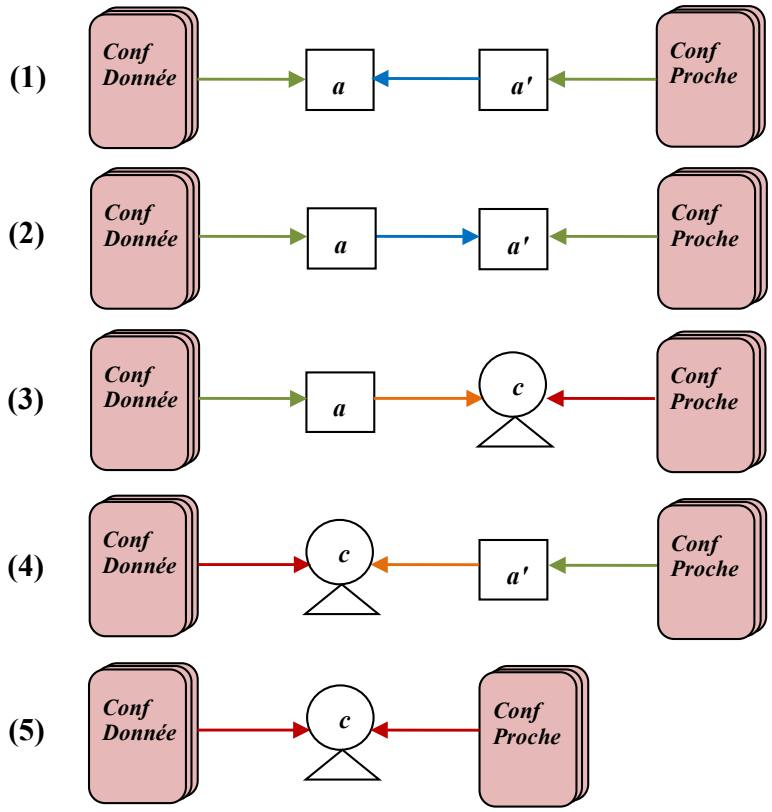

FIGURE 6.2 – Les types de relations fusionnées dans le lien inter-conférence.

1. étant donné un article a de l'édition x de la conférence C cité par l'article $a^{'}$ de l'édition $y \geqslant x$ de la conférence $C^{'}$, le chemin entre la conférence étudiée C et la conférence proche $C^{'}$ a pour poids :

$$Chemin1_{C_{Année_x}, C^{'}_{Année_y}, a, a^{'}} = Impact^{\star}_{a, C_{Année_x}} + Citation^{\star}_{a^{'}, a} + Impact^{\star}_{a^{'}, C^{'}_{Année_y}}. \quad (6.5)$$

2. étant donné un article a de l'édition x de la conférence C citant l'article $a^{'}$ de l'édition $y \leqslant x$ de la conférence $C^{'}$, le chemin entre la conférence étudiée C et la conférence proche $C^{'}$ a pour poids :

$$Chemin2_{C_{Année_x}, C^{'}_{Année_y}} = Impact^{\star}_{a, C_{Année_x}, a, a^{'}} + Citation^{\star}_{a, a^{'}} + Impact^{\star}_{a^{'}, C^{'}_{Année_y}}. \quad (6.6)$$

3. étant donné un article a de l'édition x de la conférence C et un de ses co-signataires c, le chemin entre la conférence étudiée C et la conférence proche

C' a pour poids :

$$\begin{aligned} Chemin3_{C_{Année_x}, C'_{Année_y}, a, c, p} &= Impact^{\star}_{a, C_{Année_x}} \\ &+ Contribution^{\star}_{c,a} \\ &+ ParticipationCP^{\star}_{c, C'_{Année_p}, C'_{Année_y}} \end{aligned} \tag{6.7}$$

4. étant donné un article un chercheur c, membre du CP de l'édition p de la conférence C et co-signataire d'un article a' publié par la conférence C', le chemin entre la conférence étudiée C et la conférence proche C' a pour poids :

$$\begin{aligned} Chemin4_{C_{Année_x}, C'_{Année_y}, a, c, p} &= ParticipationCP^{\star}_{c, C_{Année_p}, C_{Année_x}} \\ &+ Contribution^{\star}_{c,a} \\ &+ Impact^{\star}_{a', C'_{Année_y}} \end{aligned} \tag{6.8}$$

5. étant donné un chercheur c, membre du CP de la conférence C pour l'édition p et du CP de la conférence C' pour l'édition p', le chemin entre la conférence étudiée C et la conférence proche C' a pour poids :

$$\begin{aligned} Chemin5_{C_{Année_x}, C'_{Année_y}, c, p, p'} &= ParticipationCP^{\star}_{c, C_{Année_p}, C_{Année_x}} \\ &+ ParticipationCP^{\star}_{c, C'_{Année_{p'}}, C'_{Année_y}} \end{aligned} \tag{6.9}$$

Enfin, la similarité entre ces deux conférences C et C', agrège les poids des différents chemins existant entre les modèles des deux conférences, comme suit :

$$\begin{aligned} Similarité_{C_{Année_x}, C'_{Année_y}} &= \sum_{\forall (a,a') | C \to a \leftarrow a' \leftarrow C'} Chemin1^{\star}_{C_{Année_x}, C'_{Année_y}, a, a'} \\ &+ \sum_{\forall (a,a') | C \to a \to a' \leftarrow C'} Chemin2^{\star}_{C_{Année_x}, C'_{Année_y}, a, a'} \\ &+ \sum_{\forall (a,c,p) | C \to a \to c \leftarrow C'_p} Chemin3^{\star}_{C_{Année_x}, C'_{Année_y}, a, c, p} \\ &+ \sum_{\forall (a,c,p) | C \to c \leftarrow a \leftarrow C'_p} Chemin4^{\star}_{C_{Année_x}, C'_{Année_y}, a, c, p} \\ &+ \sum_{\forall (c,p,p') | C_p \to c \leftarrow C'_{p'}} Chemin5^{\star}_{C_{Année_x}, C'_{Année_y}, c, p, p'} \end{aligned} \tag{6.10}$$

avec la notation « $x \rightarrow y$ » indiquant la présence d'un lien dirigé dans le graphe de domaine scientifique réduit (figure 5.8 en page 55).

Les poids des liens ainsi calculés sont utilisés pour définir les indicateurs présentés dans la section suivante, mesurant l'implication d'un chercheur dans une conférence au travers des différents rôles qu'il peut jouer : auteur d'un article de la conférence, auteur d'un article d'une conférence proche, membre d'un précédent comité de programme de la conférence, etc.

6.2.2 Définition des indicateurs d'influence

Chaque rôle joué par un chercheur vis-à-vis d'une conférence peut contribuer à la notoriété du chercheur, et par conséquent, à son éventuelle influence au sein de la conférence.

Dans le but d'étudier dans quelle mesure un rôle joué par un chercheur par le passé peut expliquer sa participation à un comité de programme, nous définissons un indicateur pour chacun des rôles suivant qu'un chercheur peut jouer vis-à-vis d'une conférence :

1. Rôle 1 – Membre de comité de programme interne : le chercheur a été membre de comités de programme de la conférence ;
2. Rôle 2 – Auteur interne : le chercheur est auteur d'articles publiés dans la conférence ;
3. Rôle 3 – Auteur externe citant : le chercheur a été auteur d'articles citant des articles publiés dans la conférence ;
4. Rôle 4 – Auteur externe cité : le chercheur est auteur d'articles cités par des articles publiés dans la conférence ;
5. Rôle 5 – Auteur externe proche : le chercheur est auteur d'articles publiés dans une conférence proche ;
6. Rôle 6 – Membre de comité de programme proche : le chercheur a été membre de comités de programme d'une conférence proche.

Dans notre modèle, présenté au chapitre précédent, chaque rôle est représenté pour un chemin reliant le chercheur à la conférence. C'est sur cette base que nous définissons nos indicateurs en s'inspirant de la mesure de Katz (1953), basée sur la notion de chemin dans un graphe. La définition de chaque indicateur est ainsi basée sur les poids des

liens qui composent le chemin traduisant le rôle joué par le chercheur vis-à-vis de la conférence.

6.2.2.1 Rôle 1 – Membre de comité de programme interne

Cet indicateur se base sur les chemins reliant le chercheur c à la conférence C pour l'année x *via* des liens de participation à un comité de programme de C. Ce type de chemin n'est composé que d'un lien de participation entre c et C. Le poids d'un chemin est défini par

$$Poids\_Rôle1_{c,\,C_{Année_x},p} = ParticipationCP_{c,\,C_{Année_p},\,C_{Année_x}} \tag{6.11}$$

où $ParticipationCP_{c,\,C_{Année_p},\,C_{Année_x}}$ est le poids du lien de participation du chercheur c au comité de programme de l'édition $Année_p$ de la conférence C.

L'indicateur est ensuite défini comme l'agrégation des poids des chemins correspondant au rôle de membre d'un comité de programme de la conférence étudiée :

$$Indicateur\_Rôle1_x = \sum_{\forall (c,p) | C_p \rightarrow c} Poids\_Rôle1^{\star}_{c,\,C_{Année_x},p}. \tag{6.12}$$

Pour rappel, l'étoile ($\star$) dénote une normalisation par la valeur maximale du facteur pour obtenir l'ensemble des valeurs dans l'intervalle $[0, 1]$.

6.2.2.2 Rôle 2 – Auteur interne

Cet indicateur se base sur les chemins représentant les différentes signatures d'articles publiés dans la conférence C par le chercheur c. Ce type de chemin est composé d'un lien de signature entre le chercheur et un article ainsi que d'un lien de publication entre la conférence et l'article. Le poids de chaque chemin dépend ainsi de deux facteurs, la contribution du chercheur dans l'article et l'impact de l'article pour la conférence :

$$Poids\_Rôle2_{c,\,C_{Année_x},a} = Contribution^{\star}_{c,\,a} + Impact^{\star}_{a,\,C_{Année_x},a} \tag{6.13}$$

L'indicateur est ensuite défini comme l'agrégation des poids des chemins corres-

pondant au rôle d'auteur d'article de la conférence étudiée :

$$Indicateur\_Rôle2_x = \sum_{\forall (a,c) | C \to a \to c} Poids\_Rôle2^{\star}_{c,\, C_{Année_x},\, a}. \tag{6.14}$$

6.2.2.3 Rôle 3 – Auteur externe citant

Le rôle d'auteur externe citant pour un chercheur c intervient lorsque celui-ci signe un article a'' extérieur à la conférence C dans lequel il cite un article a publié dans la conférence C. Chaque chemin représentant ce rôle est composé de trois liens : un lien de signature entre le chercheur et l'article citant, un lien de citation entre l'article extérieur citant et l'article de la conférence, et un lien de publication entre l'article cité et la conférence.

Le poids de chaque chemin de ce type dépend ainsi de trois facteurs :

$$Poids\_Rôle3_{c,\, C_{Année_x},\, a,\, a''} = Contribution^{\star}_{c,\, a''} + Citation^{\star}_{a'',\, a} + Impact^{\star}_{a,\, C_{Année_x}} \tag{6.15}$$

L'indicateur associé à ce rôle est ensuite défini comme l'agrégation de tous les rôles de ce type joués par le chercheur :

$$Indicateur\_Rôle3_x = \sum_{\substack{\forall (a,a'',c) | C \to a \leftarrow a'' \to c \\ a'' \not\leftarrow \Omega}} Poids\_Rôle3^{\star}_{c,\, C_{Année_x},\, a,\, a''}. \tag{6.16}$$

où Ω regroupe la conférence étudiée et les conférences proches et $a'' \not\leftarrow \Omega$ signifie que l'article a'' est publié par une conférence qui n'est ni la conférence étudiée, ni une conférence proche.

6.2.2.4 Rôle 4 – Auteur externe cité

À l'inverse du rôle précédent, le rôle d'auteur externe cité intervient lorsqu'un chercheur c signe un article a'' extérieur à la conférence C qui est cité par un article a de la conférence C. Chaque chemin associé à ce rôle est composé de trois liens : un lien de signature entre le chercheur et l'article cité, un lien de citation entre l'article extérieur cité et l'article de la conférence qui le cite, et un lien de publication entre l'article citant et la conférence.

Le poids de chaque chemin de ce type dépend donc de trois facteurs :

$$Poids\_Rôle4_{c,\,C_{Année_x},a,a''} = Contribution^{\star}_{c,\,a''} + Citation^{\star}_{a,\,a''} + Impact^{\star}_{a,\,C_{Année_x}} \tag{6.17}$$

L'indicateur associé au rôle d'auteur externe cité est ensuite défini comme l'agrégation de tous les poids des chemins de ce type entre par le chercheur et la conférence :

$$Indicateur\_Rôle4_x = \sum_{\substack{\forall(a,a'',c)|C\to a\to a''\to c \\ \nwarrow\,\Omega}} Poids\_Rôle4^{\star}_{c,\,C_{Année_x},a,a''}. \tag{6.18}$$

6.2.2.5 Rôle 5 – Membre de comité de programme proche

Ce rôle est analogue au rôle 1 mais pour une conférence proche C'. Le chemin associé à ce rôle intègre, en plus du lien de participation entre le chercheur et la conférence proche, le lien inter-conférence entre la conférence et la conférence proche.

Le poids de ce type de chemin est défini suivant deux facteurs :

$$Poids\_Rôle5_{c,\,C_{Année_x},C'_{Année_y},p} = Proximité^{\star}_{C_{Année_x},\,C'_{Année_y}} + ParticipationCP_{c,\,C'_{Année_p},\,C'_{Année_y}} \tag{6.19}$$

L'indicateur associé, agrège les poids des chemins correspondant à ce rôle :

$$Indicateur\_Rôle5_x = \sum_{\forall(c,p)|c\leftarrow C'_p\leftrightsquigarrow C} Poids\_Rôle5^{\star}_{c,C_{Année_x},\,C'_{Année_y},p} \tag{6.20}$$

où « $C' \leftrightsquigarrow C$ » indique que C' est une conférence proche de C.

6.2.2.6 Rôle 6 – Auteur externe proche

Ce rôle est le pendant du rôle 2 pour une conférence proche. Le chemin traduisant ce rôle intègre le lien inter-conférence en plus des liens entre la conférence proche C' et l'article a' publié dans celle-ci ainsi qu'entre l'article a' et l'auteur c.

Le poids de ce type de chemin dépend de trois facteurs :

$$Poids\_Rôle6_{c,\,C_{Année_x},C'_{Année_y},a'} = Similarité^{\star}_{C_{Année_x},\,C'_{Année_y}} + Contribution^{\star}_{c,\,a'} + Impact^{\star}_{a',\,C'_{Année_y}} \tag{6.21}$$

Le sixième et dernier indicateur correspondant au rôle d'auteur d'article d'une conférence proche, qui agrège tous les chemins de ce type entre un chercheur et la conférence proche est défini par :

$$Indicateur\_Rôle6_x = \sum_{\forall (a',c) | c \leftarrow a' \leftarrow C' \leftrightsquigarrow C} Poids\_Rôle6^{\star}_{c, C_{Année_x}, C'_{Année_y}, a'}. \tag{6.22}$$

Ces indicateurs ont été expérimentés dans le contexte d'une conférence phare en recherche d'information, la conférence SIGIR. Ces expérimentations sont décrites dans la section suivante.

6.3 Expérimentations

Les expérimentations ont eu pour objectif d'utiliser ces indicateurs pour étudier la composition de comités de programme d'une conférence. Il s'agissait de mesurer à quel point chaque indicateur permet d'expliquer une participation à un comité de programme donné. Pour ces expérimentations, nous avons tout d'abord collecté les données relatives à la conférence SIGIR représentative du domaine de la recherche d'information pour construire ensuite le modèle de domaine de cette conférence.

6.3.1 Cadre expérimental

Pour rassembler les données relatives à une conférence dans le but de construire le modèle de domaine de cette conférence, nous appliquons le processus suivant, illustré par la figure 6.3 :

1. Collecte des données concernant les articles publiés par la conférence étudiée – *Article interne*,
2. Collecte des données concernant les comités de programme de la conférence étudiée – *Participation interne*,
3. Collecte des données concernant les articles qui citent et qui sont cités par les articles internes – *Article externe*,
4. Extraction des données concernant les auteurs des articles internes – *Auteur interne – AI*,

5. Extraction des données concernant les membres des comités de programme de la conférence donnée – *Membre du comité de programme interne – CPI*,
6. Extraction des données concernant les articles cités par les articles internes – *Article externe*,
7. Extraction des données concernant les auteurs des articles externes – *Auteur externe – AE*,
8. Collecte des données concernant les articles publiés dans les conférences proches – *Article publié dans une conférence proche*,
9. Collecte des données concernant les participations aux comités de programme des conférences proches – *Participation externe*,
10. Extraction des données concernant les auteurs des articles publiés par les conférences proches – *Auteur dans une conférence proche – AIE*,
11. Extraction des données concernant les membres des comités de programme des conférences proches – *Membre du comité de programme d'une conférence proche – CPE*.

Ce processus a été appliqué pour construire un jeu de données exploitable pour valider nos propositions. Il a été appliqué pour rassembler les données relatives aux différentes éditions de la conférence SIGIR (*Special Interest Group on Information Retrieval*). L'intérêt de cette conférence réside dans le fait qu'elle soit l'une des principales conférences du domaine de la recherche d'information, qu'elle existe depuis 1971 et qu'elle possède ainsi de nombreuses éditions et de nombreux comités de programme. De plus, elle totalise un grand nombre de publications, de chercheurs qui en sont auteurs ou qui ont participé à ses comités de programme. Un atout supplémentaire est que la plupart des données nécessaires à nos expérimentations sont disponibles dans les bases bibliographiques ACM et DBLP (depuis l'édition de 1978) permettant d'avoir ainsi des données homogènes.

Le premier travail consistant en la collecte de l'ensemble des articles publiés dans les éditions de la conférence a permis d'obtenir 3 554 articles pour les 40 éditions de la conférence SIGIR de 1971 à 2015. Ceci constitue l'ensemble initial d'articles.

Le second travail de collecte d'articles concernant les articles qui citent les articles de la conférence SIGIR et également ceux cités par les articles de la conférence SIGIR a ajouté 29 907 articles.

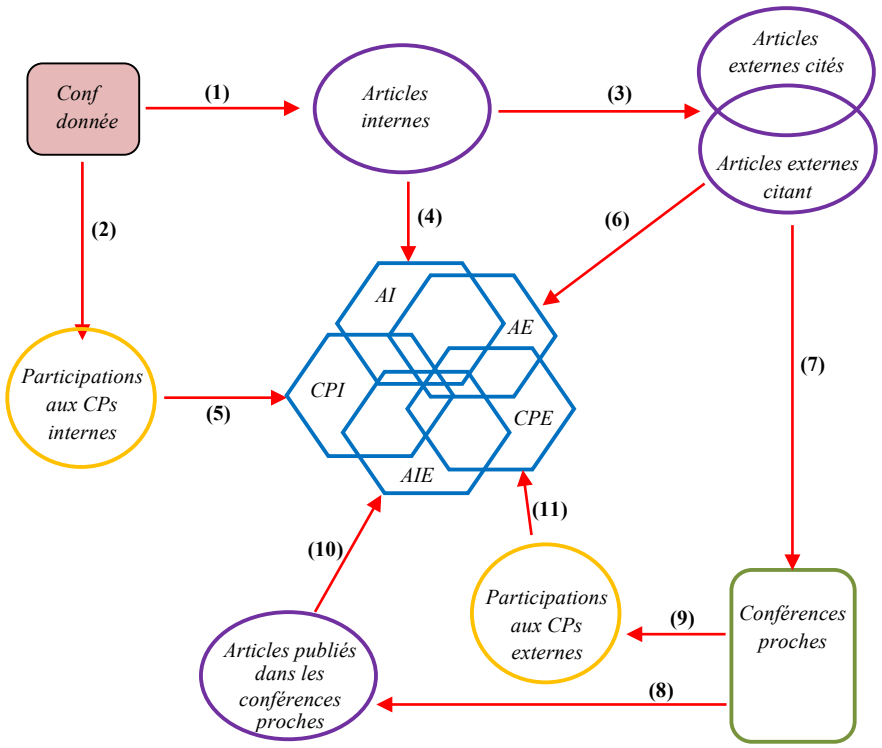

FIGURE 6.3 – Processus de collecte des données liées à la conférence étudiée.

Pour les conférences proches de SIGIR, nous avons considéré huit conférences majeures qui partagent des thématiques avec SIGIR : ACL, CIKM, ECIR, IJCAI, KDD, RecSys, WSDM et WWW. Le travail de collecte des articles publiés par ces conférences a conduit à l'ajout de 35 181 articles.

Enfin, toutes les participations des experts pour les différents comités de programme des éditions disponibles ont été recensées, depuis 2004 pour CIKM, ECIR, KDD, SIGIR et WWW, depuis 2007 pour RecSys, depuis 2003 pour IJCAI, depuis 2005 pour ACL, et depuis 2010 pour WSDM. Le nombre d'éditions utilisées pour chaque conférence dépend de leur disponibilité. Ce travail a permis de recenser 50 930 participations.

L'ensemble des chercheurs recensés dans nos données — qui concerne tous les auteurs des articles collectés et tous les membres des comités de programme collectés — représente un total de 52 071 chercheurs.

Ce processus de construction du jeu de données a représenté un travail fastidieux mêlant traitement automatique de la base bibliographique DBLP (Ley, 2002) et traitement manuel de collecte des comités de programme.

6.3.2 Résultats

Les expérimentations visaient à valider les indicateurs proposés et répondre à la question : quels rôles peuvent expliquer la participation d'un chercheur à un comité de programme ?

Cette question concerne la possibilité de trouver des indicateurs qui caractérisent les membres qui composent les comités de programme. Pour répondre à cette question, nous avons comparé les listes de chercheurs ordonnancées selon les indicateurs définis avec les comités de programme officiels. La série d'expérimentations menée a consisté à calculer les 100 premiers chercheurs selon chacun des six indicateurs. Nous avons réalisé ces calculs pour les cinq éditions de SIGIR les plus récentes de notre jeu de données ($a \in [2011, 2015]$). Pour une année donnée a, nous avons donc comparé chaque liste obtenue S_a avec le comité de programme officiel de la même année O_a pour déterminer leur intersection ($S_a \cap O_a$). Le tableau 6.1 synthétise les précisions à 20 et 100 des listes obtenues suivant chaque indicateur.

Année (a)	$\lvert O_a \rvert$	**Rôle 1**		**Rôle 2**		**Rôle 3**		**Rôle 4**		**Rôle 5**		**Rôle 6**	
		P	P@20	P	P@20	P	P@20	P	P@20	P	P@20	P	P@20
2011	426	0,80	0,90	0,57	0,60	0,43	0,40	0,43	0,55	0,48	0,65	0,05	0,05
2012	486	0,84	1,00	0,64	0,70	0,45	0,40	0,47	0,55	0,51	0,65	0,05	0,05
2013	431	0,86	0,95	0,61	0,75	0,44	0,45	0,42	0,55	0,43	0,55	0,07	0,10
2014	448	0,86	0,90	0,59	0,75	0,43	0,40	0,45	0,25	0,42	0,60	0,09	0,20
2015	432	0,85	1,00	0,60	0,80	0,42	0,45	0,45	0,25	0,44	0,70	0,07	0,10

TABLEAU 6.1 – Précision des listes de 100 et 20 meilleurs chercheurs selon six indicateurs par rapport au comité de programme officiel correspondant.

Ces résultats soulignent la prédominance du Rôle 1 pour expliquer la participation à un comité de programme d'une édition e de la conférence donnée C. Avoir déjà participé à un comité de programme précédent de la même conférence constitue le rôle prépondérant pour un chercheur pour participer au prochain comité. Les rôles 2 et 5 sont également de bons indicateurs pour caractériser les chercheurs qui constituent un comité de programme. Les auteurs des articles publiés par la conférence C ou les membres de comités de programme d'une conférence proche sont susceptibles d'être invités comme un membre du comité de programme pour l'édition e de la conférence C. Les rôles 3 et 4 sont des indicateurs moins pertinents. Les auteurs des articles citant

Rôle	2011	2012	2013	2014	2015
1	42,9 %	46,6 %	47,8 %	47,1 %	50,0 %
2	15,9 %	20,7 %	20,9 %	17,6 %	15,6 %
3	6,3 %	6,9 %	4,5 %	8,8 %	9,4 %
4	3,2 %	5,2 %	9,0 %	10,3 %	9,4 %
5	27,0 %	19,0%	13,4%	10,3 %	9,4 %
6	4,8 %	1,7 %	4,5 %	5,9 %	6,3 %

Tableau 6.2 – Proportion par rôle de chercheurs issus d'un seul indicateur et présent dans le comité de programme.

ou cités par certains articles de la conférence C ont peu d'invitations pour rejoindre le comité de programme de l'édition e de la conférence C. Enfin, le rôle 6 n'est pas un bon indicateur pour caractériser des membres du comité de programme. Publier des articles dans les conférences proches ne se traduit généralement pas, pour les auteurs, par une invitation à participer au comité de programme de l'édition de la conférence e.

Une analyse des listes produites par les six indicateurs révèle qu'une majorité (entre 20 % et 24 %) des chercheurs des comités de programme qui apparaissent dans les listes produites par les indicateurs, n'apparaissent dans les 100 premiers que pour un seul indicateur et qu'aucun chercheur n'apparaît simultanément dans les six listes. Parmi, les chercheurs présents dans des comités de programme et qui ne font partie que d'une seule liste (et qui sont donc parmi les 100 premiers d'un seul indicateur), la répartition entre les six rôles, comme indiqué dans le tableau 6.2, met en lumière une évolution dans la considération des rôles. La part prépondérante des rôles 1 et 2 est restée plutôt stables. La part des rôles 3, 4 et 6 a augmenté tandis que la part du rôle 5 a fortement diminué.

Les six indicateurs définis sont donc tous révélateurs de membres de comités de programme. De plus, aucun des indicateurs n'englobe totalement l'un des autres. Les expérimentations menées valident donc l'intérêt des indicateurs proposés.

6.3.3 Bilan

Les comités de programme des conférences sont généralement constitués par les président·e·s de comité en se basant sur des critères qui leur sont propres et qui ne

sont pas dévoilés. Ce chapitre a été consacré à notre contribution à l'étude des comités de programme dans le but d'éclairer la manière dont ils sont constitués. Les bases bibliographiques disponibles actuellement permettent de rassembler des données qui renseignent sur les relations existant entre les chercheurs et les conférences comme, par exemple, la signature d'un article, la participation à un comité de programme ou encore la citation d'un article.

En se basant sur la modélisation des données bibliographiques d'une conférence et du domaine d'une conférence proposée au chapitre précédent, nous avons proposé six indicateurs associés aux rôles que peuvent jouer les chercheurs vis-à-vis d'une conférence étudiée. En utilisant ces indicateurs, nous avons cherché à comprendre la composition des comités de programme de la conférence.

Nous avons mené des expérimentations sur les données de la conférence SIGIR, bien connue dans le domaine de la recherche d'information. Nos résultats mettent en évidence les principaux indicateurs expliquant la participation d'un chercheur à un comité de programme : le chercheur a participé aux comités précédents de la conférence, il a publié des articles dans la conférence ou bien ses articles ont été cités par des articles de la conférence. À l'inverse, les résultats mettent l'accent sur le fait que citer des articles de la conférence ou publier des articles dans des conférences proches ne représentent pas des indicateurs expliquant une participation aux comités de programme d'une conférence.

Les travaux présentés dans ce chapitre constituent une première contribution. Des expérimentations complémentaires, notamment sur des données d'autres conférences, permettraient de conforter les indicateurs proposés mais aussi de les faire évoluer. La proposition et l'analyse d'autres indicateurs constitue également une piste de travaux futurs. La définition d'indicateurs présentés dans ce chapitre nous a également conduit vers la possibilité de suggérer des membres dans le but d'aider au renouvellement des comités de programme des conférences suivant des critères qui puissent être transparents pour une communauté de chercheurs. Ces travaux font l'objet du chapitre suivant.

Chapitre 7

Suggestion de membres de comités de programme pour une conférence

Sommaire

7.1 Motivation et problématique

Comme introduit précédemment, de nombreux travaux ont montré l'intérêt d'automatiser différentes tâches relatives à l'organisation des conférences, notamment l'attribution des soumissions aux évaluateurs (Rodriguez & Bollen, 2008 ; S. Price & Flach, 2017). À notre connaissance, très peu de travaux se sont intéressés à la suggestion de membres de comités de programme (Han et al., 2013 ; Sfyris et al., 2016). Ces approches, présentées dans le chapitre dédié à l'état de l'art, se sont basées sur des descriptions textuelles.

Sur la base des résultats de l'étude présentée dans le chapitre 6, nous proposons dans ce chapitre une approche pour la suggestion de membres de CP. L'approche se base sur la modélisation des données du domaine sous forme de graphe triparti et sur les six indicateurs précédemment définis. Nous définissons une mesure de similarité

entre un chercheur et une conférence pour une année donnée. Plusieurs variantes de la mesure peuvent être spécifiées. L'application de la mesure sur un ensemble de chercheurs permet ensuite de suggérer ceux obtenant les meilleurs scores comme membres potentiels d'un comité de programme.

L'approche a été appliquée dans le contexte de la conférence SIGIR. Une comparaison des suggestions de membres de CP avec les CP officiels a été réalisée sur plusieurs années, ainsi qu'une investigation plus approfondie sur le renouvellement des CP. Les résultats de ces expérimentations sont synthétisés dans ce chapitre.

7.2 Définition de la similarité entre une conférence et un chercheur

Pour répondre à la problématique de suggestion de chercheurs susceptibles d'être invités, nous avons opté pour la définition d'une mesure de similarité entre un chercheur et une conférence pour une année donnée. L'idée est ensuite de pouvoir suggérer les chercheurs les plus « proches » de la conférence à un instant donné.

La mesure de similarité est définie à partir des six indicateurs proposés dans le chapitre précédent (section 6.2.2, page 66). L'idée sous-jacente est que chacun de ces indicateurs représente une des facettes (ou rôles) au travers desquelles un chercheur est proche d'une conférence. La mesure de similarité entre un chercheur c et l'édition de l'année x d'une conférence C est définie comme une combinaison des six indicateurs normalisés associés à ces six rôles :

$$\mathit{Similarité}_{C,c,x} = \sum_{i=1}^{6} \eta_i \cdot \mathit{Indicateur\_Rôle}\, i_x^{\star} \tag{7.1}$$

où η_i est un facteur modulant l'influence de l'indicateur associé dans le calcul de la similarité. $\mathit{Indicateur\_Rôle}\, i_x^{\star}$ est la normalisation par le maximum de l'indicateur.

7.3 Expérimentations

Les expérimentations ont été menées dans le contexte de SIGIR en utilisant le même jeu de données que pour l'étude présentée dans le précédent chapitre. Le jeu de données concerne les éditions de la conférence SIGIR ainsi que les données des conférences

proches CIKM, WWW, ECIR, RecSys, IJCAI, KDD, ACL et WSDM. Le jeu de données concerne 52 071 chercheurs.

Les objectifs des expérimentations menées ont été de répondre à deux questions :

— **Q1**. Dans quelle mesure notre approche réussit-elle à suggérer les membres des CP officiels ?

— **Q2**. Dans quelle mesure notre approche réussit-elle à suggérer de nouveaux membres pour les prochains CP officiels ?

Q1 concerne la capacité de notre approche à identifier les membres de CP connus. Pour répondre à Q1 nous avons comparé des CP suggérés par notre approche avec les CP officiels.

Q2 concerne la capacité de notre approche à identifier de nouveaux membres potentiels. Étant donné qu'il n'existe aucune liste disponible de chercheurs identifiés comme nouveaux membres potentiels de nouveaux CP SIGIR (pouvant servir de vérité terrain), nous avons comparé les suggestions de nouveaux membres avec les nouveaux membres des CP officiels. Puisque notre proposition vise à aider les président·e·s de CP à renouveler les CP, répondre à Q2 est décisif.

Les suggestions de membres de CP ont été évaluées suivant des mesures traditionnelles de RI : la précision (ratio du nombre de suggestions correctes sur le nombre de suggestions) et le rappel (ratio du nombre de suggestions correctes sur le nombre de membres à suggérer).

Pour décrire les expérimentations et leurs résultats nous utiliserons les notations listées dans le tableau 7.1.

Nous avons testé sept configurations de notre approche, listées dans le tableau 7.2, correspondant à différentes combinaisons de valeurs des paramètres impliqués dans la définition de notre mesure de similarité (cf. section 7.2).

7.3.1 Suggestions de CP comparées aux CP officiels

Cette première série d'expérimentations a consisté à suggérer des comités de programme pour les cinq éditions de 2011 à 2015 de la conférence SIGIR. Pour chaque édition considérée, les données des éditions de SIGIR des années précédentes et des conférences proches ont été utilisées pour construire le modèle de domaine, c'est-à-dire les données bibliographiques des années 2004 à 2013 ont été utilisées pour réaliser les suggestions de CP de l'année 2014.

Tableau 7.1 – Notations utilisées dans les descriptions des expérimentations.

Notation	Description
O_a	Ensemble des membres du comité de programme officiel de l'année a.
O_a^-	Ensemble des membres de tous les comités de programme officiels antérieurs à l'année a.
O_a^{+i}	Ensemble des membres du comité de programme officiel de la i^{e} année après l'année a.
S_a	Ensemble des membres suggérés pour l'année a.
N_a	Ensemble des *nouveaux* membres du comité de programme officiel de l'année a, c'est-à-dire $N_a = O_a \setminus O_a^-$.
N_{S_a}	Ensemble des*nouveaux* membres parmi ceux suggérés pour l'année a, c'est-à-dire $N_{S_a} = S_a \setminus O_a^-$.
Ci	Intersection entre S_a et une combinaison de CP officiels (ou sous-ensembles de CP) O.

Tableau 7.2 – Les sept configurations testées avec les valeurs des paramètres.

	η_1	η_2	η_3	η_4	η_5	η_6
Configuration 1	1,0	1,0	1,0	1,0	1,0	1,0
Configuration 2	1,0	0,5	0,5	0,5	0,5	0,5
Configuration 3	0,5	1,0	0,5	0,5	0,5	0,5
Configuration 4	0,5	0,5	1,0	0,5	0,5	0,5
Configuration 5	0,5	0,5	0,5	1,0	0,5	0,5
Configuration 6	0,5	0,5	0,5	0,5	1,0	0,5
Configuration 7	0,5	0,5	0,5	0,5	0,5	1,0

Nous avons réalisé deux types de comparaisons pour les listes de membres suggérés S_a pour l'année a :

1. comparaison avec le CP officiel (O_a) de la même année a pour déterminer le nombre de suggestions correctes, c'est-à-dire, $C1 = S_a \cap O_a$;
2. comparaison avec l'ensemble des membres de CP qui ont participé à au moins un des CP antérieurs à l'année a (O_a^-), c'est-à-dire, $C2 = S_a \cap O_a^-$.

Pour chaque expérimentation, nous avons produit trois listes différentes :

1. liste de suggestions comprenant le même nombre de membres que le CP officiel

correspondant (*Nombre de membres suggérés* = $|O_a|$) ;

2. liste de suggestions comprenant 50 % de membres supplémentaires que le CP officiel correspondant (*Nombre de membres suggérés = 150 %* $|O_a|$) ;
3. liste de suggestions comprenant 100 % de membres supplémentaires que le CP officiel correspondant (*Nombre de membres suggérés = 200 %* $|O_a|$).

Le tableau 7.3 présente les résultats obtenus en appliquant les sept configurations testées de notre approche, comparés avec le CP officiel de l'année correspondante a (c'est-à-dire $C1$).

Les résultats montrent que notre approche suggère un CP qui partage entre 40 % et 53 % de membres avec le CP officiel. Il est possible de suggérer jusqu'à plus de 67 % des membres en doublant le nombre de suggestions, mais ceci au détriment de la précision. Les meilleurs résultats en termes de précision ont été atteints par la configuration2 de notre approche, qui privilégie les participations des chercheurs aux CP passés de la conférence. Les résultats les plus faibles ont été obtenus avec les configurations 5 et 6, qui privilégient les participations aux CP des conférences proches et les publications dans les conférences proches, respectivement.

Les résultats présentés dans le tableau 7.4 comparent les suggestions de membres avec l'ensemble des CP passés (O_a^-). Ils montrent qu'un rappel supérieur à 96 % peut être atteint avec la configuration2 en doublant la taille de la liste suggérée par rapport à la taille de l'ensemble des CP passés (cf. Tableau 7.4, colonnes $S_a = 200\,\%|O_a^-|$). Les configurations 5 et 6 sont celles qui suggèrent le moins de membres des CP existants. Elles constituent donc des sources susceptibles de suggérer davantage de candidats qui pourraient s'avérer pertinents pour siéger comme nouveaux membres de CP. Cette hypothèse a fait l'objet de la seconde série d'expérimentation dont les résultats sont présentés dans la section suivante.

7.3.2 Nouveaux membres suggérés comparés aux nouveaux membres des CP officiels

Chaque année, de nouveaux membres rejoignent le CP d'une conférence pour la première fois. Le tableau 7.5 montre la part non négligeable de nouveaux membres rejoignant chaque année les CP des conférences liées à notre cas d'étude. Cette observation renforce l'intérêt de notre approche pour aider les président·e·s de CP à identifier des chercheurs à inviter à rejoindre le prochain comité de programme.

TABLEAU 7.3 – Comparaison des CP suggérés *versus* les CP officiels.

Année	$\|O_a\|$	$\|S_a\| = \|O_a\|$			$\|S_a\| = 150\,\%\|O_a\|$			$\|S_a\| = 200\,\%\|O_a\|$		
		$C1$	R	P	$C1$	R	P	$C1$	R	P
$Configuration1 : \eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta_5 = \eta_6 = 1$										
2011	426	197	46,24 %	46,24 %	240	56,34 %	37,56 %	261	61,27 %	30,63 %
2012	486	218	44,86 %	44,86 %	255	52,47 %	34,98 %	281	57,82 %	28,91 %
2013	431	201	46,64 %	46,64 %	237	54,99 %	36,69 %	266	61,72 %	30,86 %
2014	448	200	44,64 %	44,64 %	236	52,68 %	35,12 %	266	59,38 %	29,69 %
2015	432	192	44,44 %	44,44 %	233	53,94 %	35,96 %	260	60,19 %	30,09 %
$Configuration2 : \eta_1 = 1; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	426	226	53,05 %	53,05 %	266	62,44 %	41,63 %	288	67,61 %	33,80 %
2012	486	243	50,00 %	50,00 %	273	56,17 %	37,45 %	300	61,73 %	30,86 %
2013	431	223	51,74 %	51,74 %	259	60,09 %	40,06 %	288	66,82 %	33,41 %
2014	448	217	48,44 %	48,44 %	260	58,04 %	38,69 %	300	66,96 %	33,48 %
2015	432	214	49,54 %	49,54 %	253	58,56 %	39,04 %	278	64,35 %	32,18 %
$Configuration3 : \eta_1 = 0,5; \eta_2 = 1; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	426	202	47,42 %	47,42 %	242	56,81 %	37,87 %	268	62,91 %	31,46 %
2012	486	222	45,68 %	45,68 %	259	53,29 %	35,53 %	286	58,85 %	29,42 %
2013	431	202	46,87 %	46,87 %	244	56,61 %	37,77 %	269	62,41 %	31,21 %
2014	448	203	45,31 %	45,31 %	240	53,57 %	35,71 %	270	60,27 %	30,13 %
2015	432	195	45,14 %	45,14 %	237	54,86 %	36,57 %	265	61,34 %	30,67 %
$Configuration4 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 1; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	426	196	46,01 %	46,01 %	239	56,10 %	37,40 %	262	61,50 %	30,75 %
2012	486	218	44,86 %	44,86 %	255	52,47 %	34,98 %	281	57,82 %	28,91 %
2013	431	201	46,64 %	46,64 %	239	55,45 %	37,00 %	266	61,72 %	30,86 %
2014	448	200	44,64 %	44,64 %	235	52,46 %	34,97 %	264	58,93 %	29,46 %
2015	432	191	44,21 %	44,21 %	232	53,70 %	35,80 %	259	59,95 %	29,98 %
$Configuration5 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 1; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	426	197	46,24 %	46,24 %	238	55,87 %	37,25 %	262	61,50 %	30,75 %
2012	486	219	45,06 %	45,06 %	256	52,67 %	35,12 %	283	58,23 %	29,12 %
2013	431	204	47,33 %	47,33 %	239	55,45 %	37,00 %	264	61,25 %	30,63 %
2014	448	199	44,42 %	44,42%	235	52,46 %	34,97 %	265	59,15 %	29,58 %
2015	432	192	44,44 %	44,44 %	232	53,70 %	35,80 %	258	59,72 %	29,86 %
$Configuration6 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 1; \eta_6 = 0,5$										
2011	426	182	42,72 %	42,72 %	218	51,17 %	34,12 %	240	56,34 %	28,17 %
2012	486	195	40,12 %	40,12 %	236	48,56 %	32,37 %	258	53,09 %	26,54 %
2013	431	186	43,16 %	43,16 %	221	51,28 %	34,21 %	246	57,08 %	28,54 %
2014	448	189	42,19 %	42,19 %	219	48,88 %	32,59 %	247	55,13 %	27,57 %
2015	432	183	42,36 %	42,36 %	218	50,46 %	33,64 %	244	56,48 %	28,24 %
$Configuration7 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 1$										
2011	426	173	40,61 %	40,61 %	215	50,47 %	33,65 %	243	57,04 %	28,52 %
2012	486	194	39,92 %	39,92 %	236	48,56 %	32,37 %	262	53,91 %	26,95 %
2013	431	187	43,39 %	43,39 %	221	51,28 %	34,21 %	243	56,38 %	28,19 %
2014	448	182	40,63 %	40,63 %	216	48,21v%	32,14 %	247	55,13 %	27,57 %
2015	432	182	42,13 %	42,13 %	216	50,00 %	33,33 %	243	56,25 %	28,13 %

TABLEAU 7.4 – Comparaison des CP suggérés par notre approche *versus* les CP passés.

Année	$\lvert O_a^- \rvert$	$\lvert S_a \rvert = \lvert O_a^- \rvert$			$\lvert S_a \rvert = 150\,\% \lvert O_a^- \rvert$			$\lvert S_a \rvert = 200\,\% \lvert O_a^- \rvert$		
		$C2$	R	P	$C2$	R	P	$C2$	R	P
$Configuration1 : \eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta_5 = \eta_6 = 1$										
2011	1386	683	49,28 %	49,28 %	832	60,03 %	40,02 %	1077	77,71 %	25,90 %
2012	1482	720	48,58 %	48,58 %	879	59,31 %	39,54 %	1143	77,13 %	25,71 %
2013	1537	785	51,07 %	51,07 %	948	61,68 %	41,13 %	1238	80,55 %	26,85 %
2014	1624	806	49,63 %	49,63 %	982	60,47 %	40,31 %	1263	77,77 %	25,92 %
2015	1706	855	50,12 %	50,12 %	1038	60,84 %	40,56 %	1340	78,55 %	26,18 %
$Configuration2 : \eta_1 = 1; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	1386	812	58,59 %	58,59 %	992	71,57 %	47,72 %	1294	93,36 %	31,12 %
2012	1482	851	57,42 %	57,42 %	1052	70,99 %	47,32 %	1386	93,52 %	31,17 %
2013	1537	927	60,31 %	60,31 %	1134	73,78 %	49,20 %	1479	96,23 %	32,08 %
2014	1624	962	59,24 %	59,24 %	1154	71,06 %	47,37 %	1465	90,21 %	30,07 %
2015	1706	1010	59,20 %	59,20 %	1220	71,51 %	47,67 %	1569	91,97 %	30,66 %
$Configuration3 : \eta_1 = 0,5; \eta_2 = 1; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	1386	694	50,07 %	50,07 %	835	60,25 %	40,16 %	1078	77,78 %	25,93 %
2012	1482	733	49,46 %	49,46 %	893	60,26 %	40,17 %	1158	78,14 %	26,05 %
2013	1537	801	52,11 %	52,11 %	964	62,72 %	41,82 %	1245	81,00 %	27,00 %
2014	1624	830	51,11 %	51,11 %	997	61,39 %	40,93 %	1266	77,96 %	25,99 %
2015	1706	878	51,47 %	51,47 %	1051	61,61 %	41,07 %	1350	79,13 %	26,38 %
$Configuration4 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 1; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	1386	684	49,35 %	49,35 %	832	60,03 %	40,02 %	1072	77,34 %	25,78 %
2012	1482	724	48,85 %	48,85 %	877	59,18 %	39,45 %	1145	77,26 %	25,75 %
2013	1537	790	51,40 %	51,40 %	948	61,68 %	41,13 %	1236	80,42 %	26,81 %
2014	1624	815	50,18 %	50,18 %	985	60,65 %	40,44 %	1256	77,34 %	25,78 %
2015	1706	858	50,29 %	50,29 %	1039	60,90 %	40,60 %	1338	78,43 %	26,14 %
$Configuration5 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 1; \eta_5 = 0,5; \eta_6 = 0,5$										
2011	1386	681	49,13 %	49,13 %	833	60,10 %	40,07 %	1074	77,49 %	25,83 %
2012	1482	723	48,79 %	48,79 %	882	59,51 %	39,68 %	1145	77,26 %	25,75 %
2013	1537	792	51,53 %	51,53 %	949	61,74 %	41,17 %	1236	80,42 %	26,81 %
2014	1624	812	50,00 %	50,00 %	984	60,59 %	40,39 %	1261	77,65 %	25,88 %
2015	1706	858	50,29 %	50,29 %	1040	60,96 %	40,64 %	1339	78,49 %	26,16 %
$Configuration6 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 1; \eta_6 = 0,5$										
2011	1386	611	44,08 %	44,08 %	749	54,04 %	36,03 %	983	70,92 %	23,64 %
2012	1482	646	43,59 %	43,59 %	795	53,64 %	35,76 %	1041	70,24 %	23,41 %
2013	1537	714	46,45 %	46,45 %	850	55,30 %	36,88 %	1123	73,06 %	24,35 %
2014	1624	726	44,70 %	44,70 %	882	54,31 %	36,21 %	1141	70,26 %	23,42 %
2015	1706	780	45,72 %	45,72 %	927	54,34 %	36,23 %	1221	71,57 %	23,86 %
$Configuration7 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 1$										
2011	1386	602	43,43 %	43,43 %	741	53,46 %	35,64 %	977	70,49 %	23,50 %
2012	1482	638	43,05 %	43,05 %	787	53,10 %	35,40 %	1057	71,32 %	23,77 %
2013	1537	699	45,48 %	45,48 %	863	56,15 %	37,44 %	1130	73,52 %	24,51 %
2014	1624	735	45,26 %	45,26 %	896	55,17 %	36,78 %	1154	71,06 %	23,69 %
2015	1706	782	45,84 %	45,84 %	946	55,45 %	36,97 %	1234	72,33 %	24,11 %

TABLEAU 7.5 – Part de membres qui ont participé à un CP pour la première fois.

Année	**Conf.**	$\|O_a\|$	$\|N_a\|$	**Ratio**	**Conf.**	$\|O_a\|$	$\|N_a\|$	**Ratio**
2011		426	92	21,60 %		570	320	56,14 %
2012		486	96	19,75 %		547	241	44,06 %
2013	***SIGIR***	431	55	12,76 %	***WWW***	955	521	54,55 %
2014		448	87	19,42 %		941	455	48,35 %
2015		432	82	18,98 %		502	173	34,46 %
2011		697	295	42,32 %		418	117	27,99 %
2012		716	407	56,84 %		677	341	50,37 %
2013	***CIKM***	515	131	25,44 %	***KDD***	696	306	43,97 %
2014		629	353	56,12 %		784	281	35,84 %
2015		700	226	32,29 %		794	274	34,51 %
2011		229	68	29,69 %		789	291	36,88 %
2012		250	94	37,60 %		649	154	23,73 %
2013	***ECIR***	233	22	9,44 %	***ACL***	748	232	31,02 %
2014		268	57	21,27 %		874	269	30,78 %
2015		311	88	28,30 %		885	228	25,76 %

Pour évaluer la capacité de notre approche à suggérer de nouveaux membres de CP, une possibilité consistait à comparer les nouveaux chercheurs suggérés avec un ensemble de candidats pertinents identifiés pour une année donnée, pouvant servir de vérité terrain. Malheureusement, il n'existe pas de telle liste de membres de CP potentiels dans le contexte de SIGIR, ni dans un autre contexte non plus. Une autre possibilité aurait été de soumettre nos suggestions à des assesseurs experts ; ce principe est cependant délicat à mettre en place.

En l'absence des ces possibilités, nous avons considéré comme vérité terrain les nouveaux membres des CP officiels de SIGIR. Pour répondre à la question Q2, nous avons donc comparé les nouveaux membres suggérés pour une année donnée a avec les nouveaux membres du CP officiel de la même année mais également des années suivantes. Les mesures d'évaluation de précision et de rappel ont également été calculées. Les expérimentations menées ont cependant quelques limites puisque les nouveaux membres de CP SIGIR apparaissant dans les éditions postérieures ne représentent qu'une vérité terrain incomplète. Il est nécessaire de construire une vérité terrain plus complète, qui devrait être indépendante des pratiques courantes utilisées pour la constitution de CP. Cependant, trouver des experts pour constituer cette vérité

terrain est une tâche ardue. Ceci constitue une perspective importante aux travaux de la présente thèse.

Pour cette seconde série d'expérimentations, comme pour la première, nous avons suggéré des CP pour chaque édition de 2011 à 2015. Pour chaque année a nous avons considéré les nouveaux membres suggérés N_{S_a}, c'est-à-dire, les membres qui n'étaient pas présents dans les CP officiels antérieurs à l'année a. Nous avons distingué trois cas :

1. l'ensemble des nouveaux membres suggérés qui sont effectivement des nouveaux membres du CP officiel correspondant, c'est-à-dire

$$C3 = N_{S_a} \cap (O_a \setminus O_a^-).$$

   Par exemple, un chercheur qui est suggéré pour 2011 participe effectivement au CP officiel de 2011 comme nouveau membre.

2. l'ensemble des nouveaux membres suggérés qui sont effectivement des nouveaux membres du CP officiel correspondant ou du suivant, c'est-à-dire

$$C4 = N_{S_a} \cap ((O_a^{+1} \cup O_a) \setminus O_a^-).$$

   Par exemple, un chercheur qui est suggéré pour 2011 participe effectivement comme nouveau membre au CP officiel de 2011 ou de 2012.

3. l'ensemble des nouveaux membres suggérés qui sont effectivement des nouveaux membres du CP officiel correspondant ou des deux suivants, c'est-à-dire

$$C5 = N_{S_a} \cap ((O_a^{+1} \cup O_a^{+2} \cup O_a) \setminus O_a^-).$$

   Par exemple, un chercheur qui est suggéré pour 2011 participe effectivement comme nouveau membre au CP officiel de 2011, de 2012 ou de 2013.

Les résultats pour le premier cas, c'est-à-dire la comparaison avec les nouveaux membres du CP correspondant ($C3$), reportés dans le tableau 7.6, montrent qu'une faible proportion des membres suggérés correspond aux nouveaux membres officiels de l'année correspondante. Les résultats montrent que la configuration de notre approche qui promeut les participations aux CP des conférences proches (Configuration 6) obtient la meilleure performance en moyenne par rapport aux autres configu-

rations. Doubler le nombre de suggestions permet d'obtenir un meilleur rappel, au prix d'une précision plus faible.

Les cas 2) et 3) avaient pour objectif de vérifier si un délai pouvait exister pour un chercheur avant d'être invité à rejoindre le CP d'une conférence. Les résultats reportés dans le tableau 7.7 pour le cas 2) et dans le tableau 7.8 pour le cas 3) montrent que davantage de nouveaux chercheurs suggérés deviennent pertinents car ils rejoignent les CP officiels des années suivantes. Ces résultats confirment l'influence des participations aux CP des conférences proches pour expliquer l'invitation de nouveaux membres. Cependant, ces résultats n'informent pas sur la pertinence des autres nouveaux membres suggérés et incitent à mener des expérimentations axées sur la validation de telles suggestions par des experts de SIGIR pour enrichir les possibilités d'évaluation. Cela fait partie des travaux futurs prévus.

7.4 Bilan

La constitution des comités de programme des conférences est généralement du ressort des président·e·s de CP. Une partie non négligeable des CP est renouvelée d'une année sur l'autre. Identifier des chercheurs au fait des thématiques de la conférence et correspondant aux critères fixés par le président·e du CP est une tâche difficile, notamment pour les conférences phares qui reposent sur des CP imposants (plus d'une centaine de membres).

Ce chapitre a couvert notre contribution à l'automatisation des tâches relatives à l'organisation des conférences, et plus particulièrement la suggestion de membres de comités de programme qui a récemment suscité l'intérêt de la communauté (Han et al., 2013 ; Sfyris et al., 2016). Contrairement à ces deux approches qui se basent sur des descriptions textuelles des conférences et des chercheurs, nous avons proposé une approche orientée graphe, basée sur nos contributions présentées dans les précédents chapitres et relatives à la modélisation de données bibliographiques et de définition d'indicateurs. Une mesure de similarité entre un chercheur et une conférence pour une année donnée a été proposée et utilisée pour suggérer des membres de comités de programme. Cette mesure, paramétrable, permet de mettre en œuvre plusieurs systèmes de suggestion de membre de CP.

Des expérimentations, dans le contexte de la conférence SIGIR, ont permis de répondre aux questions relatives à la capacité de notre approche à identifier des cher-

cheurs pertinents pour faire partie des CP de la conférence, y compris pour suggérer de nouveaux membres. Néanmoins, certaines limitations concernant les expérimentations ont été soulignées, liées à l'absence de réelle liste de nouveaux membres de CP potentiels pouvant servir de vérité terrain.

La contribution présentée dans ce chapitre constitue néanmoins une première étape ouvrant un certain nombre de perspectives, telles que :

- enrichir le protocole d'évaluation en faisant appel à des experts SIGIR, comme d'anciens président·e·s de CP, pour établir une liste plus complète de nouveaux membre de CP potentiels,
- affiner les configurations testées et les tester dans le contexte d'autres conférences,
- compléter notre approche pour en faire une approche multicritère pour la construction de comités de programme qui applique la configuration la plus appropriée pour répondre à un ensemble de critère spécifiés.

Tableau 7.6 – Nouveaux membres suggérés qui sont nouveaux membre du CP de l'année correspondante.

Année	$\lvert N_a\rvert$	$\lvert S_a\rvert = \lvert O_a\rvert$				$\lvert S_a\rvert = 150\,\%\lvert O_a\rvert$				$\lvert S_a\rvert = 200\,\%\lvert O_a\rvert$			
		$C3$	R	N_{S_a}	P	$C3$	R	N_{S_a}	P	$C3$	R	N_{S_a}	P
$Configuration1 : \eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta_5 = \eta_6 = 1$													
2011	92	0	0,00 %	102	0,00 %	3	3,26 %	199	1,51 %	6	6,52 %	337	1,78 %
2012	96	5	5,21 %	114	4,39 %	7	7,29 %	239	2,93 %	10	10,42 %	391	2,56 %
2013	55	2	3,64 %	82	2,44 %	2	3,64 %	182	1,10 %	3	5,45 %	292	1,03 %
2014	87	0	0,00 %	88	0,00 %	2	2,30 %	198	1,01 %	4	4,60 %	312	1,28 %
2015	82	0	0,00 %	75	0,00 %	1	1,22 %	165	0,61 %	3	3,66 %	274	1,09 %
$Configuration2 : \eta_1 = 1; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	92	0	0,00 %	49	0,00 %	1	1,09 %	122	0,82 %	3	3,26 %	233	1,29 %
2012	96	3	3,13 %	64	4,69 %	5	5,21 %	156	3,21 %	8	8,33 %	286	2,80 %
2013	55	0	0,00 %	33	0,00 %	2	3,64 %	107	1,87 %	2	3,64 %	202	0,99 %
2014	87	0	0,00 %	37	0,00 %	0	0,00 %	113	0,00 %	2	2,30 %	218	0,92 %
2015	82	0	0,00 %	28	0,00 %	0	0,00 %	90	0,00 %	2	2,44 %	183	1,09 %
$Configuration3 : \eta_1 = 0,5; \eta_2 = 1; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	92	1	1,09 %	98	1,02 %	4	4,35 %	194	2,06 %	6	6,52 %	321	1,87 %
2012	96	6	6,25 %	113	5,31 %	8	8,33 %	234	3,42 %	10	10,42 %	378	2,65 %
2013	55	2	3,64 %	78	2,56 %	3	5,45 %	171	1,75 %	3	5,45 %	280	1,07 %
2014	87	0	0,00 %	82	0,00 %	2	2,30 %	184	1,09 %	5	5,75 %	305	1,64 %
2015	82	0	0,00 %	70	0,00 %	1	1,22 %	155	0,65 %	5	6,10 %	264	1,89 %
$Configuration4 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 1; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	92	0	0,00 %	103	0,00 %	3	3,26 %	200	1,50 %	6	6,52 %	333	1,80 %
2012	96	5	5,21 %	117	4,27 %	7	7,29 %	239	2,93 %	10	10,42 %	385	2,60 %
2013	55	2	3,64 %	82	2,44 %	2	3,64 %	181	1,10 %	3	5,45 %	290	1,03 %
2014	87	0	0,00 %	87	0,00 %	2	2,30 %	195	1,03 %	4	4,60 %	308	1,30 %
2015	82	0	0,00 %	74	0,00 %	1	1,22 %	164	0,61 %	3	3,66 %	271	1,11 %
$Configuration5 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 1; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	92	1	1,09 %	103	0,97 %	3	3,26 %	199	1,51 %	6	6,52 %	331	1,81 %
2012	96	5	5,21 %	116	4,31 %	8	8,33 %	236	3,39 %	10	10,42 %	387	2,58 %
2013	55	2	3,64 %	79	2,53 %	2	3,64 %	179	1,12 %	3	5,45 %	291	1,03 %
2014	87	0	0,00 %	86	0,00 %	2	2,30 %	194	1,03 %	4	4,60 %	310	1,29 %
2015	82	0	0,00 %	72	0,00 %	1	1,22 %	164	0,61 %	3	3,66 %	272	1,10 %
$Configuration6 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 1; \eta_6 = 0,5$													
2011	92	3	3,26 %	133	2,26 %	4	4,35 %	252	1,59 %	5	5,43 %	399	1,25 %
2012	96	6	6,25 %	147	4,08 %	9	9,38 %	291	3,09 %	10	10,42 %	467	2,14 %
2013	55	2	3,64 %	111	1,80 %	3	5,45 %	223	1,35 %	4	7,27 %	355	1,13 %
2014	87	1	1,15 %	109	0,92 %	3	3,45 %	230	1,30 %	3	3,45 %	367	0,82 %
2015	82	1	1,22 %	96	1,04 %	2	2,44 %	198	1,01 %	5	6,10 %	322	1,55 %
$Configuration7 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 1$													
2011	92	0	0,00 %	146	0,00 %	2	2,17 %	255	0,78 %	5	5,43 %	393	1,27 %
2012	96	5	5,21 %	158	3,16 %	7	7,29 %	299	2,34 %	8	8,33 %	454	1,76 %
2013	55	2	3,64 %	115	1,74 %	2	3,64 %	234	0,85 %	2	3,64 %	360	0,56 %
2014	87	1	1,15 %	124	0,81 %	2	2,30 %	247	0,81 %	4	4,60 %	380	1,05 %
2015	82	0	0,00 %	105	0,00 %	2	2,44 %	217	0,92 %	3	3,66 %	331	0,91 %

TABLEAU 7.7 – Nouveaux membres suggérés qui sont nouveaux membres du CP de l'année correspondante ou de la suivante.

Année	$\lvert N_a \cup N_a^{+1}\rvert$	$\lvert S_a\rvert = \lvert O_a \cup O_a^+1\rvert$				$\lvert S_a\rvert = 150\%\lvert O_a \cup O_a^+1\rvert$				$\lvert S_a\rvert = 200\%\lvert O_a \cup O_a^+1\rvert$			
		$C4$	R	N_{S_a}	P	$C4$	R	N_{S_a}	P	$C4$	R	N_{S_a}	P
$Configuration1 : \eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta_5 = \eta_6 = 1$													
2011	188	8	4,26 %	176	4,55 %	13	6,91 %	363	3,58 %	17	9,04 %	566	3,00 %
2012	151	8	5,30 %	164	4,88 %	12	7,95 %	335	3,58 %	17	11,26 %	542	3,14 %
2013	142	2	1,41 %	152	1,32 %	6	4,23 %	298	2,01 %	10	7,04 %	499	2,00 %
2014	169	3	1,78 %	146	2,05 %	7	4,14 %	296	2,36 %	11	6,51 %	467	2,36 %
$Configuration2 : \eta_1 = 1; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	188	4	2,13 %	109	3,67 %	12	6,38 %	257	4,67 %	15	7,98 %	439	3,42 %
2012	151	6	3,97 %	99	6,06 %	9	5,96 %	236	3,81 %	14	9,27 %	420	3,33 %
2013	142	2	1,41 %	82	2,44 %	3	2,11 %	209	1,44 %	9	6,34 %	368	2,45 %
2014	169	0	0,00 %	80	0,00 %	4	2,37 %	200	2,00 %	8	4,73 %	346	2,31 %
$Configuration3 : \eta_1 = 0,5; \eta_2 = 1; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	188	9	4,79 %	171	5,26 %	14	7,45 %	350	4,00 %	19	10,11 %	550	3,45 %
2012	151	8	5,30 %	159	5,03 %	13	8,61 %	317	4,10 %	19	12,58 %	527	3,61 %
2013	142	3	2,11 %	140	2,14 %	8	5,63 %	288	2,78 %	13	9,15 %	485	2,68 %
2014	169	1	0,59 %	133	0,75 %	8	4,73 %	285	2,81 %	12	7,10 %	455	2,64 %
$Configuration4 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 1; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	188	8	4,26 %	177	4,52 %	13	6,91 %	360	3,61 %	17	9,04 %	562	3,02 %
2012	151	8	5,30 %	162	4,94 %	12	7,95 %	335	3,58 %	17	11,26 %	536	3,17 %
2013	142	2	1,41 %	149	1,34 %	7	4,93 %	299	2,34 %	11	7,75 %	497	2,21 %
2014	169	3	1,78 %	141	2,13 %	7	4,14 %	294	2,38 %	11	6,51 %	464	2,37 %
$Configuration5 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 1; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	188	8	4,26 %	178	4,49 %	14	7,45 %	361	3,88 %	17	9,04 %	563	3,02 %
2012	151	7	4,64 %	165	4,24 %	12	7,95 %	333	3,60 %	17	11,26 %	537	3,17 %
2013	142	2	1,41 %	150	1,33 %	7	4,93 %	297	2,36 %	10	7,04 %	499	2,00 %
2014	169	3	1,78 %	143	2,10 %	7	4,14 %	296	2,36 %	11	6,51 %	464	2,37 %
$Configuration6 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 1; \eta_6 = 0,5$													
2011	188	10	5,32 %	222	4,50 %	14	7,45 %	431	3,25 %	19	10,11 %	643	2,95 %
2012	151	11	7,28 %	204	5,39 %	14	9,27 %	399	3,51 %	18	11,92 %	609	2,96 %
2013	142	3	2,11 %	179	1,68 %	8	5,63 %	364	2,20 %	11	7,75 %	576	1,91 %
2014	169	3	1,78 %	175	1,71 %	8	4,73 %	346	2,31 %	13	7,69 %	538	2,42 %
$Configuration7 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 1$													
2011	188	6	3,19 %	231	2,60 %	13	6,91 %	419	3,10 %	16	8,51 %	630	2,54 %
2012	151	8	5,30 %	213	3,76 %	10	6,62 %	393	2,54 %	14	9,27 %	601	2,33 %
2013	142	3	2,11 %	202	1,49 %	4	2,82 %	368	1,09 %	9	6,34 %	571	1,58 %
2014	169	3	1,78 %	199	1,51 %	7	4,14 %	360	1,94 %	8	4,73 %	548	1,46 %

TABLEAU 7.8 – Nouveaux membres suggérés qui sont nouveaux membre du CP de l'année correspondante ou les deux suivantes.

Année	$\lvert N_a \cup N_a^{+1} \cup N_a^{+2} \rvert$	$\lvert S_a \rvert = \lvert O_a \cup O_a^{+1} \cup O_a^{+2} \rvert$				$\lvert S_a \rvert = 150\% \lvert O_a \cup O_a^{+1} \cup O_a^{+2} \rvert$				$\lvert S_a \rvert = 200\% \lvert O_a \cup O_a^{+1} \cup O_a^{+2} \rvert$			
		$C5$	R	N_{S_a}	P	$C5$	R	N_{S_a}	P	$C5$	R	N_{S_a}	P
$Configuration1 : \eta_1 = \eta_2 = \eta_3 = \eta_4 = \eta_5 = \eta_6 = 1$													
2011	243	11	4,53 %	220	5,00 %	19	7,82 %	442	4,30 %	23	9,47 %	684	3,36 %
2012	238	9	3,78 %	229	3,93 %	19	7,98 %	460	4,13 %	28	11,76 %	716	3,91 %
2013	224	5	2,23 %	201	2,49 %	13	5,80 %	403	3,23 %	18	8,04 %	636	2,83 %
$Configuration2 : \eta_1 = 1; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	243	7	2,88 %	141	4,96 %	15	6,17 %	330	4,55 %	19	7,82 %	551	3,45 %
2012	238	7	2,94 %	148	4,73 %	15	6,30 %	348	4,31 %	24	10,08 %	585	4,10 %
2013	224	3	1,34 %	126	2,38 %	7	3,13 %	284	2,46 %	15	6,70 %	501	2,99 %
$Configuration3 : \eta_1 = 0,5; \eta_2 = 1; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	243	13	5,35 %	210	6,19 %	17	7,00 %	424	4,01 %	25	10,29 %	671	3,73 %
2012	238	11	4,62 %	223	4,93 %	23	9,66 %	446	5,16 %	32	13,45 %	703	4,55 %
2013	224	6	2,68 %	191	3,14 %	17	7,59 %	389	4,37 %	21	9,38 %	618	3,40 %
$Configuration4 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 1; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	243	11	4,53 %	218	5,05 %	18	7,41 %	437	4,12 %	23	9,47 %	682	3,37 %
2012	238	9	3,78 %	230	3,91 %	20	8,40 %	459	4,36 %	29	12,18 %	712	4,07 %
2013	224	5	2,23 %	201	2,49 %	12	5,36 %	394	3,05 %	19	8,48 %	631	3,01 %
$Configuration5 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 1; \eta_5 = 0,5; \eta_6 = 0,5$													
2011	243	12	4,94 %	220	5,45 %	18	7,41 %	440	4,09 %	23	9,47 %	683	3,37 %
2012	238	9	3,78 %	226	3,98 %	20	8,40 %	457	4,38 %	29	12,18 %	712	4,07 %
2013	224	5	2,23 %	199	2,51 %	12	5,36 %	395	3,04 %	19	8,48 %	631	3,01 %
$Configuration6 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 1; \eta_6 = 0,5$													
2011	243	16	6,58 %	283	5,65 %	19	7,82 %	510	3,73 %	25	10,29 %	758	3,30 %
2012	238	15	6,30 %	280	5,36 %	22	9,24 %	537	4,10 %	29	12,18 %	795	3,65 %
2013	224	7	3,13 %	244	2,87 %	15	6,70 %	475	3,16 %	18	8,04 %	703	2,56 %
$Configuration7 : \eta_1 = 0,5; \eta_2 = 0,5; \eta_3 = 0,5; \eta_4 = 0,5; \eta_5 = 0,5; \eta_6 = 1$													
2011	243	10	4,12 %	278	3,60 %	17	7,00 %	508	3,35 %	20	8,23 %	759	2,64 %
2012	238	8	3,36 %	288	2,78 %	17	7,14 %	525	3,24 %	24	10,08 %	793	3,03 %
2013	224	5	2,23 %	258	1,94 %	10	4,46 %	471	2,12 %	16	7,14 %	713	2,24 %

Chapitre 8

Conclusion générale et perspectives

Conclusion

La publication scientifique permet de communiquer les progrès en sciences auprès des chercheurs et du grand public. Les articles paraissent dans les revues spécialisées et les actes de conférences, usuellement après évaluation par les pairs. Les comités de rédaction et de programme sous-jacents représentent la clé de voûte du processus d'évaluation (Zuckerman & Merton, 1971). Avec le développement des revues et le nombre croissant de conférences scientifiques organisées chaque année, rechercher des experts pour participer à ces comités mobilise les réseaux des scientifiques les plus actifs. C'est une activité chronophage mais critique pour maintenir la confiance que place la société dans la science (Benos et al., 2007 ; Crowcroft et al., 2009).

Cette thèse s'est focalisée sur la tâche de suggestion de membres de comité de programme pour des conférences scientifiques. Ce travail doctoral s'inscrit dans la continuité des études scientométriques entreprises pour révéler le paysage international du pilotage de la recherche en systèmes d'information (Cabanac, 2012), ainsi que l'histoire et les perspectives des communautés scientifiques nationales INFORSID et EGC (Cabanac et al., 2016) inscrites dans ce même domaine scientifique.

Les travaux les plus proches en recherche d'information exploitent très majoritairement le texte intégral des articles scientifiques pour recommander des experts en réponse à une requête portant sur un thème de recherche (chapitres 2 et 3). Au regard de notre problématique, l'accès massif au texte intégral des articles scientifiques à des fins de fouille de texte n'est pas envisageable concrètement.

Pour pallier cet écueil, nous avons tourné notre regard vers les travaux de scientométrie (chapitre 4) qui s'appuient sur les données bibliographiques pour en extraire des réseaux : de citations, de co-signataires, d'institutions, de pays, de mots employés

dans les titres d'articles, etc. Ces éléments liés aux activités de publication et au rayonnement scientifique contribuent, à nos yeux, à forger l'expertise scientifique des chercheurs, tant recherchée à des fins de pilotage de communautés scientifiques. Partant de l'hypothèse que l'invitation d'un chercheur à participer à un comité de programme peut s'expliquer par différentes preuves d'expertise, nous avons unifié ces éléments en les modélisant au sein d'un graphe hétérogène pondéré (Tran et al., 2016) dans le chapitre 5.

Cette modélisation a servi de socle pour définir des indicateurs scientométriques pour étudier la constitution de comités de programme passés (chapitre 6). La conférence de premier plan SIGIR (*Special Interest Group on Information Retrieval*), dont nous avons considéré les éditions de 1971 à 2015, a été retenue pour instancier ce modèle, reflet de la sphère académique SIGIR. Nous avons collecté l'ensemble des membres de comité de programme et des notices d'articles publiés dans cette même conférence. Nous avons alors distingué quantitativement les critères prépondérants pour l'inclusion des chercheurs au comité de programme de SIGIR dont, sans surprise, la participation aux comités de programme de SIGIR antérieurs ainsi que la publication dans les actes de SIGIR. Plus étonnant, les experts des comités de conférences proches thématiquement se retrouvent peu dans le comité SIGIR.

Enfin, nous avons conçu une approche de suggestion de comités de programme pour une conférence donnée, en combinant les résultats des indicateurs scientométriques susmentionnés (chapitre 7). Cette approche vise à aider au renouvellement des comités en suggérant des chercheurs actifs, proches de la thématique de la conférence donnée et dont l'impact des travaux sur la communauté est reconnu (Tran et al., 2017).

Notre approche de suggestion de membre de comité de programme a été expérimentée pour la conférence SIGIR, en considérant les données afférentes ainsi que celles relatives à ses conférences proches thématiquement telles que CIKM, ECIR, WSDM et WWW. Nous avons évalué la pertinence de notre approche en coupant le jeu de données en deux ensembles délimités par une année frontière : apprentissage sur les éditions jusqu'à cette année frontière-là et test sur les éditions postérieures. Ce procédé nous a permis de quantifier la part des membres de comité à la fois suggérés par notre approche et pertinents, car figurant dans les comités effectifs de la conférence considérée.

Perspectives

L'approche développée dans ce mémoire s'inscrit dans un programme de recherche plus ambitieux. Concernant la modélisation de la sphère académique d'une conférence, nous avons identifié de nombreuses variables dans un premier temps, sans avoir pu les collecter et les étudier ensuite. Par exemple, nous avons ambitionné de porter sur le graphe hétérogène le genre, l'ancienneté et les affiliations (institution et pays) des scientifiques, ainsi qu'une estimation *a priori* de leur réputation académique (Nicholas et al., 2015). Ces éléments nous auraient permis de questionner la participation des individus aux comités de programme au regard d'arbitrages opérant sur ces variables, afin de tendre vers un équilibre de représentation des femmes–hommes, des institutions, des pays impliqués, une rotation des personnes sollicitées issues d'un environnement commun, etc. Des échanges informels avec des président·e·s de comités de programme nous incitent à creuser cette piste, qui pourrait expliquer certains faux-négatifs dans nos listes de suggestion : ce sont des suggestions pertinentes mais, pour des raisons de stratégie du pilotage, ces personnes n'ont pas été sollicitées.

Ces réflexions incitent à questionner davantage les acteurs modélisés, dans une démarche qualitative complémentaire aux solutions proposées, reposant sur des indicateurs quantitatifs. Il nous paraît crucial d'interroger les président·e·s de comités de programme qui envoient les invitations, ainsi que les invités, à plusieurs titres. D'une part, nous nous interrogeons quant aux éléments guidant les président·e·s durant leur sélection des scientifiques à inviter. D'autre part, il est évident que certain·e·s invité·e·s refusent de participer au comité de programme, et ce, pour une foultitude de raisons : surcharge de travail, adéquation partielle à la thématique de la conférence, règles tacites de non-concurrence entre communautés scientifiques, etc. Ces personnes ont été pensées légitimes pour intégrer le comité de programme, elles sont cependant absentes des listes mentionnées dans les actes, sur lesquelles nous basons nos suggestions et leur validation.

Nous avons observé que certains membres participent au comité sans pour autant paraître posséder un tel capital scientifique au vu des indicateurs étudiés. Il conviendrait d'analyser finement ces cas, certes marginaux, mais révélateurs de pratiques inattendues. Par exemple, certains membres ne sont pas actifs dans la communauté universitaire mais occupent une position de responsable scientifique dans des divisions de recherche et développement du secteur industriel (pensons aux cas d'entreprises telles

que Google, Facebook, Twitter). Un autre exemple concerne le cas de *subreviewers* : des individus auxquels un membre du comité confie l'évaluation d'une ou plusieurs soumissions en se portant garant de cette personne. Ces relecteurs additionnels sont vraisemblablement invités à rejoindre le comité en priorité, les années suivantes, lorsqu'ils ont démontré leur capacité à réaliser des évaluations ponctuelles et de qualité (cf. les *Outstanding Reviewer Awards* [1] de SIGIR décernés à des relecteurs additionnels). L'analyse des réseaux de co-signature (*coauthorships*) ou de collaboration au sens large (projets de recherche, journées d'études, etc.) couplée à des entretiens permettrait de révéler et documenter ces stratégies auxquelles recourent certains scientifiques établis pour « faire entrer » leurs collaborateurs dans les comités de programme.

Décrypter les rouages gouvernant la constitution des comités de programme — ou tout autre collectif scientifique constitué par invitation et cooptation — au niveau des disciplines constitue, nous semble-t-il un axe de recherche ambitieux. Il s'agira de concevoir une méthode mixte mêlant les approches quantitative et qualitative pour sonder cette question et contraster les pratiques des communautés scientifiques d'horizons divers.

1. https://web.archive.org/web/201708/http://sigir.org/sigir2017/program/awards/

Bibliographie

Afzal, M. T. & Maurer, H. A. (2011). Expertise Recommender System for Scientific Community. *Journal of Universal Computer Science*, *17*(11), 1529-1549. doi :10.3217/jucs-017-11-1529. (cité p. 15)

Avin, C., Lotker, Z., Peleg, D. & Turkel, I. (2015). Social Network Analysis of Program Committees and Paper Acceptance Fairness. In *ASONAM'15 : Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (p. 488-495). New York, NY, USA : ACM. doi :10.1145/2808797.2809305. (cité p. 34)

Balog, K., Azzopardi, L. & de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *SIGIR'06 : Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 43-50). New York, NY, USA : ACM. doi :10.1145/1148170.1148181. (cité p. 17)

Balog, K., Bogers, T., Azzopardi, L., de Rijke, M. & van den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. In *SIGIR'07 : Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 551-558). New York, NY, USA : ACM. doi :10.1145/1277741.1277836. (cité p. 18)

Balog, K. & de Rijke, M. (2007). Determining Expert Profiles (With an Application to Expert Finding). In *IJCAI'07, Proceedings of the 20th International Joint Conference on Artificial Intelligence* (p. 2657-2662). Récupérée *via* http://ijcai.org/Proceedings/07/Papers/427.pdf. (cité p. 14)

Balog, K. & de Rijke, M. (2008). Non-local evidence for expert finding. In *CIKM'08 : Proceedings of the 17th ACM Conference on Information and Knowledge Management* (p. 489-498). New York, NY, USA : ACM. doi :10.1145/1458082.1458148. (cité p. 14, 18)

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P. & Si, L. (2012). Expertise Retrieval. *Foundations and Trends in Information Retrieval*, *6*(2-3), 127-256. doi :10.1561/1500000024. (cité p. 13)

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes : The case of computer science. *Scientometrics*, *83*(3), 809-824. doi :10.1007/s11192-009-0145-4. (cité p. 43)

Bartneck, C. & Hu, J. (2009). Scientometric Analysis of the CHI Proceedings. In *CHI'09 : Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 699-708). New York, NY, USA : ACM. doi :10.1145/1518701.1518810. (cité p. 33)

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., … Zotov, A. (2007). The ups and downs of peer review. *Advances in Physiology Education*, *31*(2), 145-152. doi :10.1152/advan.00104.2006. (cité p. 91)

Billaut, J.-C., Bouyssou, D. & Vincke, P. (2010). Should you believe in the Shanghai ranking ? – An MCDM view. *Scientometrics*, *84*(1), 237-263. doi :10.1007/s11192-009-0115-x. (cité p. 32)

Bollacker, K. D., Lawrence, S. & Giles, C. L. (2000). Discovering Relevant Scientific Literature on the Web. *IEEE Intelligent Systems*, *15*(2), 42-47. doi :10.1109/5254.850826. (cité p. 25)

Börner, K. (2010). *Atlas of Science : Visualizing What We Know*. Cambridge, MA : MIT Press. (cité p. 2).

Bornmann, L. & Marx, W. (2011). The h index as a research performance indicator. *European Science Editing*, *37*(3), 77-80. (cité p. 36).

Bouyssou, D. & Marchant, T. (2011). Ranking scientists and departments in a consistent manner. *Journal of the American Society for Information Science and Technology*, *62*(9), 1761-1769. doi :10.1002/asi.21544. (cité p. 32)

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107-117. doi :10.1016/S0169-7552(98)00110-X. (cité p. 20)

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, *87*(3), 597-620. doi :10.1007/s11192-011-0358-1. (cité p. 14, 28)

Cabanac, G. (2012). Shaping the landscape of research in Information Systems from the perspective of editorial boards : A scientometric study of 77 leading journals. *Journal of the American Society for Information Science and Technology*, *63*(5), 977-996. doi :10.1002/asi.22609. (cité p. 3, 91)

Cabanac, G. (2015). In Praise of Interdisciplinary Research through Scientometrics. In P. Mayr, I. Frommholz & P. Mutschke (Éd.), *BIR'15 : Proceedings of the Second Workshop on Bibliometric-enhanced Information Retrieval co-located with the 37th*

European Conference on Information Retrieval (ECIR 2015) (T. 1344, p. 5-13). CEUR Workshop Proceedings. CEUR-WS. (cité p. 32).

Cabanac, G. (2016). *Interroger le texte scientifique* (Habilitation à diriger des recherches, Université Toulouse 3 – Paul Sabatier, Toulouse). (cité p. 31).

Cabanac, G., Hubert, G., Tran, H. D., Favre, C. & Labbé, C. (2016). Un regard lexico-scientométrique sur le défi EGC 2016. In *EGC'16 : Actes des 16e journées Extraction et Gestion des Connaissances* (p. 419-424). RNTI. Paris : Hermann. (cité p. 4-7, 91).

Cabanac, G. & Preuss, T. (2013). Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the American Society for Information Science and Technology*, *64*(2), 405-415. (cité p. 58).

Cao, Y., Liu, J., Bao, S. & Li, H. (2005). Research on Expert Search at Enterprise Track of TREC 2005. In *TREC'05 : Proceedings of the 14th Text REtrieval Conference*. NIST. Gaithersburg, MD, USA. Récupérée *via* http://trec.nist.gov/pubs/trec14/papers/microsoft-asia.ent.pdf. (cité p. 16)

Chen, J. & Konstan, J. A. (2010). Conference paper selectivity and impact. *Communications of the ACM*, *53*(6), 79-83. doi :10.1145/1743546.1743569. (cité p. 34, 43)

Chen, Y., Wei, J., Wu, S. & Hu, Y. (2006). A similarity-based method for retrieving documents from the SCI/SSCI database. *Journal of Information Science*, *32*(5), 449-464. doi :10.1177/0165551506065814. (cité p. 34)

Collectif INFORSID. (2012). La recherche en systèmes d'information et ses nouvelles frontières. *Ingénierie des Systèmes d'Information*, *17*(3), 9-68. doi :10.3166/isi.17.3.9-68. (cité p. 3)

Craswell, N., de Vries, A. P. & Soboroff, I. (2005). Overview of the TREC 2005 Enterprise Track. In *TREC'05 : Proceedings of the Fourteenth Text REtrieval Conference*. NIST. Gaithersburg, MD, USA. Récupérée *via* http://trec.nist.gov/pubs/trec14/papers/ENTERPRISE.OVERVIEW.pdf. (cité p. 14, 18)

Crowcroft, J., Keshav, S. & McKeown, N. (2009). Scaling the academic publication process to Internet scale [Viewpoint]. *Communications of the ACM*, *52*(1), 27-30. doi :10.1145/1435417.1435430. (cité p. 1, 91)

De Bellis, N. (2009). *Bibliometrics and Citation Analysis : From the Science Citation Index to Cybermetrics*. Lanham, MD : Scarecrow Press. (cité p. 2, 31, 32).

de La Robertie, B., Pitarch, Y., Takasu, A. & Teste, O. (2017). Identifying Authoritative Researchers in Digital Libraries Using External a Priori Knowledge. In *SAC'17 :*

Proceedings of the Symposium on Applied Computing (p. 1017-1022). New York, NY, USA : ACM. doi :10.1145/3019612.3019809. (cité p. 35)

Demartini, G., Gaugaz, J. & Nejdl, W. (2009). A Vector Space Model for Ranking Entities and Its Application to Expert Search. In *ECIR'09 : Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (p. 189-201). Springer-Verlag. doi :10.1007/978-3-642-00958-7_19. (cité p. 16)

Dom, B., Eiron, I., Cozzi, A. & Zhang, Y. (2003). Graph-based ranking algorithms for e-mail expertise analysis. In *DMKD@SIGMOD'03 : Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery* (p. 42-48). doi :10.1145/882082.882093. (cité p. 26)

Ekstrand, M. D., Kannan, P., Stemper, J. A., Butler, J. T., Konstan, J. A. & Riedl, J. (2010). Automatically building research reading lists. In *RecSys'10 : Proceedings of the 2010 ACM Conference on Recommender Systems* (p. 159-166). doi :10.1145/1864708.1864740. (cité p. 27)

Fang, H. & Zhai, C. (2007). Probabilistic Models for Expert Finding. In *ECIR'07 : Proceedings of the 29th European Conference on IR Research on Advances in Information Retrieval* (p. 418-430). doi :10.1007/978-3-540-71496-5_38. (cité p. 17)

Fang, Y., Si, L. & Mathur, A. P. (2010). Discriminative models of integrating document evidence and document-candidate associations for expert search. In *SIGIR'10 : Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 683-690). New York, NY, USA : ACM. doi :10.1145/1835449.1835563. (cité p. 18, 19)

Freyne, J., Coyle, L., Smyth, B. & Cunningham, P. (2010). Relative status of journal and conference publications in Computer Science. *Communications of the ACM*, *53*(11), 124-132. doi :10.1145/1839676.1839701. (cité p. 34, 43)

Garfield, E. (1955). Citation Indexes for Science : A New Dimension in Documentation through Association of Ideas. *Science*, *122*(3159), 108-111. doi :10.1126/science.122.3159.108. (cité p. 32)

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. *Journal of the American Medical Association*, *295*(1), 90-93. doi :10.1001/jama.295.1.90. (cité p. 32)

Gevers, M. (2014). Scientific performance indicators : A critical appraisal and a country-by-country analysis. In W. Blockmans, L. Engwall & D. Weaire (Éd.), *Bibliometrics : Use and abuse in the review of research performance* (Chap. 5, T. 87, p. 43-53).

Wenner-Gren International Series. London : Portland Press. Récupérée *via* http://www.portlandpress.com/pp/books/online/wg87/087/0043/0870043.pdf. (cité p. 32)

Gingras, Y. (2014). Les dérives de l'évaluation de la recherche : du bon usage de la bibliométrie. Paris : Raisons d'agir. (cité p. 32).

Han, S., Jiang, J., Yue, Z. & He, D. (2013). Recommending program committee candidates for academic conferences. In *CompSci@CIKM 2013 : Proceedings of the workshop on Computational scientometrics : theory & applications* (p. 1-6). doi :10.1145/2508497.2508498. (cité p. 36, 77, 86)

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569-16572. doi :10.1073/pnas.0507655102. (cité p. 19, 32, 35)

Huang, Z. [Zan], Chung, W., Ong, T. & Chen, H. (2002). A graph-based recommender system for digital library. In *JCDL'02 : Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (p. 65-73). doi :10.1145/544220.544231. (cité p. 25)

Huang, Z. [Zhixing] & Qiu, Y. (2010). A multiple-perspective approach to constructing and aggregating Citation Semantic Link Network. *Future Generation Computer Systems*, *26*(3), 400-407. doi :10.1016/j.future.2009.07.006. (cité p. 24)

Jakawat, W., Favre, C. & Loudcher, S. (2016). Graphs enriched by cubes for OLAP on bibliographic networks. *International Journal of Business Intelligence and Data Mining*, *11*(1), 85-107. doi :10.1504/ijbidm.2016.076435. (cité p. 35)

Jarke, M., Pham, M. C. & Klamma, R. (2013). Evolution of the CAiSE Author Community : A Social Network Analysis. In J. Bubenko, J. Krogstie, O. Pastor, B. Pernici, C. Rolland & A. Sølvberg (Éd.), *Seminal Contributions to Information Systems Engineering* (p. 15-33). Berlin : Springer. doi :10.1007/978-3-642-36926-1_2. (cité p. 33)

Jayasinghe, G. K., Karimi, S. & Ayre, M. (2016). Evaluation of Retrieval Algorithms for Expertise Search. In *ADCS'16 : Proceedings of the 21st Australasian Document Computing Symposium* (p. 85-88). New York, NY, USA : ACM. doi :10.1145/3015022.3015035. (cité p. 14)

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39-43. (cité p. 26, 28, 29, 66).

Kergosien, É., Bessagnet, M.-N., Sallaberry, C., Le Parc-Lacayrelle, A. & Royer, A. (2016). Un regard lexico-scientométrique sur le défi EGC 2016. In *EGC'16 : Actes des*

16e journées Extraction et Gestion des Connaissances (p. 371-382). RNTI. Paris : Hermann. (cité p. 33).

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, *14*(1), 10-25. doi :10.1002/asi.5090140103. (cité p. 48)

Khabsa, M. & Giles, C. L. (2014). The Number of Scholarly Documents on the Public Web. *PLoS ONE*, *9*(5), e93949. doi :10.1371/journal.pone.0093949. (cité p. 21)

Kim, J. [Jinseok] & Kim, J. [Jinmo]. (2015). Rethinking the comparison of coauthorship credit allocation schemes. *Journal of Informetrics*, *9*(3), 667-673. doi :10.1016/j.joi.2015.07.005. (cité p. 62)

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604-632. doi :10.1145/324133.324140. (cité p. 20)

Kosmulski, M. (2012). The order in the lists of authors in multi-author papers revisited. *Journal of Informetrics*, *6*(4), 639-644. doi :10.1016/j.joi.2012.06.006. (cité p. 62)

Küngas, P., Karus, S., Vakulenko, S., Dumas, M., Parra, C. & Casati, F. (2013). Reverse-engineering Conference Rankings : What Does It Take to Make a Reputable Conference ? *Scientometrics*, *96*(2), 651-665. doi :10.1007/s11192-012-0938-8. (cité p. 33)

Lao, N. & Cohen, W. W. (2010). Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, *81*(1), 53-67. doi :10.1007/s10994-010-5205-8. (cité p. 27)

Larivière, V. (2015). Bibliométrie. In J. Prud'homme, P. Doray & F. Bouchard (Éd.), *Sciences, technologies et sociétés de A à Z* (p. 26-29). Libre accès. Montréal : Presses Universitaires de Montréal. (cité p. 2, 31).

Lawrence, S., Giles, C. L. & Bollacker, K. D. (1999). Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, *32*(6), 67-71. doi :10.1109/2.769447. (cité p. 48)

Ley, M. (2002). The DBLP Computer Science Bibliography : Evolution, Research Issues, Perspectives. In A. H. F. Laender & A. L. Oliveira (Éd.), *SPIRE'02 : Proceedings of the 9th international conference on String Processing and Information Retrieval* (T. 2476, p. 1-10). LNCS. Springer. doi :10.1007/3-540-45735-6_1. (cité p. 19, 72)

Leydesdorff, L., Bornmann, L., Comins, J. A. & Milojević, S. (2016). Citations : Indicators of Quality ? The Impact Fallacy. *Frontiers in Research Metrics and Analytics*, *1*, 1. doi :10.3389/frma.2016.00001. (cité p. 32)

Leydesdorff, L. & Milojević, S. (2015). Scientometrics. In J. D. Wright (Éd.), *International Encyclopedia of the Social & Behavioral Sciences* (2e éd., T. 21, p. 322-327). Am-

sterdam : Elsevier. doi :10.1016/b978-0-08-097086-8.85030-8. (cité p. 2, 29, 31)

Leydesdorff, L. & Rafols, I. (2011). Indicators of the interdisciplinarity of journals : Diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87-100. doi :10.1016/j.joi.2010.09.002. (cité p. 32)

Liang, S. & de Rijke, M. (2016). Formal Language Models for Finding Groups of Experts. *Information Processing & Management*, *52*(4), 529-549. doi :10.1016/j.ipm.2015.11.005. (cité p. 19)

Liang, Y., Li, Q. & Qian, T. (2011). Finding Relevant Papers Based on Citation Relations. In *WAIM'11 : Proceedings of the 12th International Conference on Web-Age Information Management* (p. 403-414). doi :10.1007/978-3-642-23535-1_35. (cité p. 28, 47)

Liben-Nowell, D. & Kleinberg, J. M. (2003). The link prediction problem for social networks. In *CIKM'03 : Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management* (p. 556-559). New York, NY, USA : ACM. doi :10.1145/956863.956972. (cité p. 26)

Liu, P., Curson, J. & Dew, P. (2005). Use of RDF for expertise matching within academia. *Knowledge and Information Systems*, *8*(1), 103-130. doi :10.1007/s10115-004-0152-y. (cité p. 24)

Loudcher, S., Jakawat, W., Morales, E. P. S. & Favre, C. (2015). Combining OLAP and information networks for bibliographic data analysis : A survey. *Scientometrics*, *103*(2), 471-487. doi :10.1007/s11192-015-1539-0. (cité p. 35)

Macdonald, C. & Ounis, I. (2006). Voting for candidates : adapting data fusion techniques for an expert search task. In *CIKM'06 : Proceedings of the ACM CIKM International Conference on Information and Knowledge Management* (p. 387-396). New York, NY, USA : ACM. doi :10.1145/1183614.1183671. (cité p. 16)

Macdonald, C. & Ounis, I. (2009). Searching for Expertise : Experiments with the Voting Model. *The Computer Journal*, *52*(7), 729-748. doi :10.1093/comjnl/bxm112. (cité p. 16)

Macdonald, C. & Ounis, I. (2011). Learning Models for Ranking Aggregates. In *ECIR'11 : Proceedings of the 33rd European Conference on Research Advances in Information Retrieval* (p. 517-529). doi :10.1007/978-3-642-20161-5_52. (cité p. 19)

Marchant, T. (2009). Score-based bibliometric rankings of authors. *Journal of the American Association for Information Science and Technology*, *60*(6), 1132-1137. doi :10.1002/asi.21059. (cité p. 32)

McDonald, D. W. & Ackerman, M. S. (2000). Expertise recommender : a flexible recommendation system and architecture. In *CSCW'00 : Proceeding on the ACM Conference on Computer Supported Cooperative Work* (p. 231-240). doi :10.1145/358916.358994. (cité p. 24)

Merton, R. K. (1968). The Matthew Effect in Science : The reward and communication systems of science are considered. *Science*, *159*(3810), 56-63. Réédité dans (Merton, 1973, Chap. 20). doi :10.1126/science.159.3810.56. (cité p. 4)

Merton, R. K. (Éd.). (1973). *The Sociology of Science : Theoretical and Empirical Investigations*. Chicago, IL : The University of Chicago Press. (cité p. 102).

Moreira, C., Calado, P. & Martins, B. (2015). Learning to Rank Academic Experts in the DBLP Dataset. *Expert Systems*, *32*(4), 477-493. doi :10.1111/exsy.12062. (cité p. 19, 35)

Nalimov, V. V. & Mulchenko, Z. M. (1969). *Naukometriya. Izuchenie Razvitiya Nauki kak Informatsionnogo Protsessa [Scientometrics. Study of the Development of Science as an Information Process]*. [English translation : 1971. Washington, D.C. : Foreign Technology Division. U.S. Air Force Systems Command, Wright-Patterson AFB, Ohio. (NTIS Report No. AD735-634)]. Moscou : Nauka. (cité p. 2).

Nanba, H. & Okumura, M. (1999). Towards Multi-paper Summarization Using Reference Information. In *IJCAI'99 : Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (p. 926-931). Récupérée *via* http://ijcai.org/Proceedings/99-2/Papers/038.pdf. (cité p. 24)

Nicholas, D., Herman, E., Jamali, H. R., Rodríguez-Bravo, B., Boukacem-Zeghmouri, C., Dobrowolski, T. & Pouchot, S. (2015). New ways of building, showcasing, and measuring scholarly reputation. *Learned Publishing*, *28*(3), 169-183. doi :10.1087/20150303. (cité p. 93)

Petkova, D. & Croft, W. (2006). Hierarchical Language Models for Expert Finding in Enterprise Corpora. In *ICTAI'06 : Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence* (p. 599-608). IEEE. doi :10.1109/ictai.2006.63. (cité p. 17)

Petkova, D. & Croft, W. B. (2007). Proximity-based document representation for named entity retrieval. In *CIKM'07 : Proceedings of the Sixteenth ACM Conference*

on Information and Knowledge Management (p. 731-740). New York, NY, USA : ACM. doi :10.1145/1321440.1321542. (cité p. 17)

Pradhan, D., Paul, P. S., Maheswari, U., Nandi, S. & Chakraborty, T. (2017). C^3-index : A PageRank Based Multi-faceted Metric for Authors' Performance Measurement. *Scientometrics*, *110*(1), 253-273. doi :10.1007/s11192-016-2168-y. (cité p. 35)

Price, D. J. (1951). Quantitative Measures of the Development of Science. *Archives Internationales d'Histoire des Sciences*, *4*(14), 86-93. Récupérée *via* http://garfield.library.upenn.edu/price/pricequantitativemeasures1951.pdf. (cité p. 31)

Price, S. & Flach, P. A. (2017). Computational Support for Academic Peer Review : A Perspective from Artificial Intelligence. *Communications of the ACM*, *60*(3), 70-79. doi :10.1145/2979672. (cité p. 58, 77)

Pritchard, A. (1969). Statistical bibliography or bibliometrics ? [Documentation notes]. *Journal of Documentation*, *25*(4), 348-349. doi :10.1108/eb026482. (cité p. 2)

Radev, D. R., Joseph, M. T., Gibson, B. & Muthukrishnan, P. (2016). A bibliometric and network analysis of the field of computational linguistics. *Journal of the Association for Information Science and Technology*, *67*(3), 683-706. doi :10.1002/asi.23394. (cité p. 28)

Radev, D. R., Muthukrishnan, P., Qazvinian, V. & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation*, *47*(4), 919-944. doi :10.1007/s10579-012-9211-2. (cité p. 28)

Ratinaud, P. (2009). *IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. Récupérée *via* http://www.iramuteq.org. (cité p. 5)

Ribas, S., Ribeiro-Neto, B., Santos, R. L., de Souza e Silva, E., Ueda, A. & Ziviani, N. (2015). Random Walks on the Reputation Graph. In *ICTIR'15 : Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (p. 181-190). New York, NY, USA : ACM. doi :10.1145/2808194.2809462. (cité p. 35)

Ribeiro, I. S., Santos, R. L. T., Gonçalves, M. A. & Laender, A. H. F. (2015). On Tag Recommendation for Expertise Profiling : A Case Study in the Scientific Domain. In *WSDM'15 : Proceedings of the 8th ACM International Conference on Web Search and Data Mining* (p. 189-198). doi :10.1145/2684822.2685320. (cité p. 14, 15)

Rodriguez, M. A. & Bollen, J. (2008). An algorithm to determine peer-reviewers. In *CIKM'08 : Proceedings of the 17th ACM Conference on Information and Knowledge Management* (p. 319-328). New York, NY, USA : ACM. doi :10.1145/1458082.1458127. (cité p. 26, 77)

Sakr, S. & Alomari, M. (2012). A decade of database conferences : A look inside the program committees. *Scientometrics*, *91*(1), 173-184. doi :10.1007/s11192-011-0530-7. (cité p. 33)

Serdyukov, P., Chernov, S. & Nejdl, W. (2007). Enhancing Expert Search Through Query Modeling. In *ECIR'07 : Proceedings of the 29th European Conference on Research Advances in Information Retrieval* (p. 737-740). doi :10.1007/978-3-540-71496-5_81. (cité p. 17)

Serdyukov, P. & Hiemstra, D. (2008). Modeling Documents as Mixtures of Persons for Expert Finding. In *ECIR'08 : Proceedings of the 30th European Conference on Research Advances in Information Retrieval* (p. 309-320). doi :10.1007/978-3-540-78646-7_29. (cité p. 18)

Serdyukov, P., Rode, H. & Hiemstra, D. (2008). Modeling multi-step relevance propagation for expert finding. In *CIKM'08 : Proceedings of the 17th ACM Conference on Information and Knowledge Management* (p. 1133-1142). New York, NY, USA : ACM. doi :10.1145/1458082.1458232. (cité p. 47)

Serdyukov, P., Taylor, M., Vinay, V., Richardson, M. & White, R. W. (2011). Automatic People Tagging for Expertise Profiling in the Enterprise. In *ECIR'11 : Proceedings of the 33rd European Conference on Research Advances in Information Retrieval* (p. 399-410). doi :10.1007/978-3-642-20161-5_40. (cité p. 14, 15)

Sfyris, G. A., Fragkos, N. & Doulkeridis, C. (2016). Profile-Based Selection of Expert Groups. In *TPDL'16 : Proceedings of the Conference on Theory and Practice of Digital Libraries* (p. 81-93). doi :10.1007/978-3-319-43997-6_7. (cité p. 36, 77, 86)

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *WWW'15 : Proceedings of the 24th International Conference on World Wide Web Companion* (p. 243-246). Republic & Canton of Geneva, Switzerland : International World Wide Web Conferences Steering Committee. doi :10.1145/2740908.2742839. (cité p. 35)

Small, H. (1973). Co-citation in the scientific literature : A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265-269. doi :10.1002/asi.4630240406. (cité p. 48)

Strohman, T., Croft, W. B. & Jensen, D. D. (2007). Recommending citations for academic papers. In *SIGIR'07 : Proceedings of the 30th Annual International ACM Conference*

on Research and Development in Information Retrieval (p. 705-706). New York, NY, USA : ACM. doi :10.1145/1277741.1277868. (cité p. 26)

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. & Su, Z. (2008). ArnetMiner : Extraction and mining of academic social networks. In *SIGKDD'08 Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining* (p. 990-998). doi :10.1145/1401890.1402008. (cité p. 19, 27)

Todeschini, R. & Baccini, A. (2016). *Handbook of bibliometric indicators : Quantitative tools for studying and evaluating research*. Weinheim : Wiley-VCH. doi :10.1002/9783527681969. (cité p. 32)

Tran, H. D., Cabanac, G. & Hubert, G. (2016). Suggestion d'experts pour renouveler le comité de programme d'une conférence. In *CORIA'16 : Actes de la 13^{e} conférence en recherche d'information et applications* (p. 105-120). doi :10.24348/sdnri.2016.20. (cité p. 7, 92)

Tran, H. D., Cabanac, G. & Hubert, G. (2017). Expert suggestion for conference program committees. In S. Assar, Ó. Pastor & H. Mouratidis (Éd.), *RCIS'17 : Proceedings of the 11th International Conference on Research Challenges in Information Science* (p. 221-232). IEEE. doi :10.1109/rcis.2017.7956540. (cité p. 7, 92)

Van Hooydonk, G. (1997). Fractional counting of multiauthored publications : Consequences for the impact of authors. *Journal of the American Society for Information Science*, *48*(10), 944-945. doi :fs6xvz. (cité p. 62)

Vasilescu, B., Serebrenik, A., Mens, T., van den Brand, M. G. & Pek, E. (2014). How healthy are software engineering conferences ? *Science of Computer Programming*, *89*(100), 251-272. doi :10.1016/j.scico.2014.01.016. (cité p. 34)

Waltman, L. & van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, *63*(2), 406-415. doi :10.1002/asi.21678. (cité p. 36)

Wang, Y., Tong, Y. & Zeng, M. (2013). Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information. In *AAAI'13 : Proceedings of the 27th Conference on Artificial Intelligence* (p. 933-939). AAAI Press. Récupérée *via* https://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6363/7304. (cité p. 34)

Wildgaard, L. (2015). A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health

in Web of Science and Google Scholar. *Scientometrics*, *104*(3), 873-906. doi :10.1007/s11192-015-1608-4. (cité p. 32)

Yimam, D. & Kobsa, A. (2000). DEMOIR : A Hybrid Architecture for Expertise Modeling and Recommender Systems. In *WETICE : Proceedings of the 9th IEEE International Workshops on Enabling Technologies : Infrastructure for Collaborative Enterprises* (p. 67-74). doi :10.1109/enabl.2000.883706. (cité p. 14)

Zhang, J. [Jing], Tang, J. & Li, J. (2007). Expert Finding in a Social Network. In *DASFAA'07 : Proceedings of the 12th International Conference on Database Systems for Advanced Applications* (p. 1066-1069). doi :10.1007/978-3-540-71703-4_106. (cité p. 20)

Zhang, J. [Jun], Ackerman, M. S. & Adamic, L. A. (2007). Expertise networks in online communities : structure and algorithms. In *WWW'07 : Proceedings of the 16th International Conference on World Wide Web* (p. 221-230). doi :10.1145/1242572.1242603. (cité p. 20)

Zhu, J., Huang, X., Song, D. & Rüger, S. M. (2010). Integrating multiple document features in language models for expert finding. *Knowledge and Information Systems*, *23*(1), 29-54. doi :10.1007/s10115-009-0202-6. (cité p. 18)

Zhuang, Z., Elmacioglu, E., Lee, D. & Giles, C. L. (2007). Measuring conference quality by mining program committee characteristics. In *JCDL'07 : Proceedings of the 2007 conference on Digital libraries* (p. 225-234). New York, NY : ACM. doi :10.1145/1255175.1255220. (cité p. 34)

Zuckerman, H. & Merton, R. K. (1971). Patterns of Evaluation in Science : Institutionalisation, Structure and Functions of the Referee System. *Minerva*, *9*(1), 66-100. doi :10.1007/BF01553188. (cité p. 91)

Résumé

La publication scientifique dans les revues spécialisées et les actes de conférences permet de communiquer les progrès en sciences. Les comités de rédaction et de programme sous-jacents représentent la clé de voûte du processus d'évaluation. Avec le développement des revues et le nombre croissant de conférences scientifiques organisées chaque année, rechercher des experts pour participer à ces comités est une activité chronophage mais critique. Cette thèse se focalise sur la tâche de suggestion de membres de comité de programme (CP) pour des conférences scientifiques. Elle comporte trois volets. Premièrement, nous proposons une modélisation basée sur un graphe hétérogène pondéré de l'expertise scientifique multifacette des chercheurs. Deuxièmement, nous définissons des indicateurs scientométriques pour quantifier les critères impliqués dans la constitution de CP. Troisièmement, nous concevons une approche de suggestion de membres de CP pour une conférence donnée, en combinant les résultats des indicateurs scientométriques susmentionnés. Notre approche est expérimentée pour une des conférences de premier plan de notre communauté de recherche : SIGIR, en considérant ses éditions de 1971 à 2015, ainsi que les conférences proches thématiquement.

Abstract

Academic publishing in specialized journals and conference proceedings is the main way to communicate progress in science. The underlying editorial and program committees represent the cornerstone of the evaluation process. With the development of journals and the increasing number of scientific conferences held annually, searching for experts who would serve in these committees is a time-consuming and yet critical activity. This PhD thesis focuses on the task of suggesting program committee (PC) members for scientific conferences. It is organized into three parts. First, we propose a modelling of the multifaceted scientific expertise of researchers based on a weighted heterogeneous graph. Second, we define scientometric indicators to quantify the criteria involved in the composition of CPs. Third, we design a CP member suggestion approach for a given conference, combining the results of the aforementioned scientometric indicators. Our approach is experimented in the context of leading conferences of our research community : SIGIR, considering its editions from 1971 to 2015, and topically close conferences.