Suggestion d'experts pour renouveler le comité de programme d'une conférence

Hong Diep Tran, Guillaume Cabanac, Gilles Hubert

IRIT UMR 5505 CNRS, Université Paul Sabatier – Toulouse 3
{hong-diep.tran, guillaume.cabanac, gilles.hubert}@irit.fr

RÉSUMÉ. *Le processus d'évaluation par les pairs permet de valider les progrès scientifiques communiqués dans des articles de recherche. Cette grande responsabilité repose sur les comités éditoriaux des journaux, sur les comités de programme des conférences et sur chacun de leurs membres. De plus, avec un grand nombre de conférences scientifiques organisées chaque année, la recherche d'experts pour participer au comité de programme devient une tâche fréquente et coûteuse. Dans cet article, nous proposons une modélisation d'expert basée sur différentes preuves d'expertise, notamment sur les citations, pour émettre des suggestions d'experts dans le cas d'une recherche de membres pour renouveler le comité de programme d'une conférence.*

ABSTRACT. *The peer review process enables the scientific community to validate the new knowledge documented in research activites. It relies on the editorial boards of academic journals, on the program committees of conferences, and their members. In addition, with a large number of scientific conferences held each year, searching for experts that would be invited to join program committees is an increasingly hard task. In this paper, we propose an expert modeling method based on different features of expertise — in particular on citation networks — in order to issue suggestions for renewing the members of conference program committees.*

MOTS-CLÉS : *Recherche d'experts, graphe de citations, comité de programme.*

KEYWORDS: *Expert retrieval, citation graph, program committee.*

1. Introduction

La recherche d'experts est une problématique posée depuis une quinzaine d'années (Balog *et al.*, 2012). Des systèmes automatiques ont été proposés pour répondre à ce problème dans de nombreux cas spécifiques (Fang et Zhai, 2007 ; Rodriguez et Bollen, 2008 ; Afzal et Maurer, 2011), bien qu'ils ne semblent pas, en réalité, largement utilisés. En parallèle au fort développement de la science et de la technologie, un grand nombre de conférences sont organisées annuellement afin de diffuser les découvertes scientifiques et de les partager avec la communauté scientifique. Notre étude s'intéresse à l'une des tâches importantes de l'organisation d'une conférence : trouver des experts pour renouveler son comité de programme. La réalisation de cette tâche, basée habituellement sur les accointances entre chercheurs, est lourde et présente le risque d'oublier des experts qui pourraient être sollicités. Rodriguez et Bollen (2008) ont montré la nécessité d'automatiser cette tâche. De nombreuses études s'intéressent à la recherche d'experts pour des buts différents, par exemple trouver un relecteur pour un manuscrit soumis à un journal (Rodriguez et Bollen, 2008), suggérer des chercheurs correspondant à des intérêts scientifiques (Cabanac, 2011), trouver le directeur de thèse approprié pour un étudiant (Liu et Dew, 2004), ou encore trouver un professionnel de santé pour remplacer un absent en cas d'urgence (McDonald et Ackerman, 2000). Cependant, à notre connaissance, il n'existe pas d'approche relative à la recherche de membres de comité de programme.

Différents problèmes de recherche d'experts ont trouvé une réponse dans les systèmes de recherche d'information (Macdonald et Ounis, 2008 ; Serdyukov et Hiemstra, 2008), c'est-à-dire une recherche de profils thématiques d'experts à partir de la formulation d'un besoin d'expertise. Une difficulté qui influence la qualité de la recherche réside dans la détermination du besoin de l'utilisateur, davantage pour la recherche scientifique car ce besoin est souvent implicite (Balog *et al.*, 2012). Les profils des experts qui représentent le domaine, les compétences ou les intérêts scientifiques des individus sont déterminés à partir de diverses preuves d'expertise, telles que les activités professionnelles, les activités scientifiques ou encore les citations formulées par les chercheurs. Plusieurs études ont montré une bonne efficacité de systèmes de recherche basés sur les citations (Strohman *et al.*, 2007 ; De Bellis, 2009 ; Liang *et al.*, 2011) et l'apport de la combinaison avec d'autres traces d'activités scientifiques des chercheurs (Cabanac, 2011).

Les citations constituent un objet d'étude central en sociologie des sciences (Kessler, 1963 ; Small, 1973 ; Cronin, 1984). Cependant, saisir le sens exact de chaque citation de manière automatique reste un problème ouvert. À ce jour, peu de travaux ont utilisé les graphes de citations pour la recherche d'experts, la plupart se sont retreints aux tâches de recherche d'articles ou de documents scientifiques.

L'objectif de cet article vise la suggestion d'experts-candidats pour renouveler le comité de programme d'une conférence. Ce contexte diffère des précédents évoqués précédemment par le peu de contenu textuel décrivant une conférence. En effet, la description d'une conférence se limite en général à un ensemble de sujets dont les

grandes lignes sont évoquées par quelques mots-clés. Une recherche dans une base bibliographique à partir de ces mots-clés conduirait inévitablement à un résultat fortement « bruité ». Ainsi, l'approche proposée dans cet article se base sur des éléments autres que textuels. Notre contribution dans cet article est triple, en proposant :

1) une modélisation de l'espace de recherche se basant sur différents éléments (tels que la conférence, les publications, les experts-candidats) et les types de relations entre ces éléments (comme être auteur d'une publication, citer un article, participer à un comité de programme précédent),

2) une méthode de calcul de la proximité entre un expert-candidat et la conférence,

3) une série d'expérimentations pour valider notre approche.

Cet article est organisé comme suit. La section 2 présente un état de l'art relatif à la recherche d'expertise et à la recommandation d'experts. La section 3 introduit notre contribution concernant la modélisation de l'espace de recherche d'experts et la proximité entre expert-candidat et conférence. La section 4 décrit les expérimentations réalisées pour valider notre approche. Enfin, la section 5 conclut l'article et présente les perspectives à ce travail en lien avec la diversification des suggestions selon les caractéristiques des individus (genre, thème, localisation, ancienneté).

2. État de l'art

2.1. *Recherche d'expertise*

Selon Balog *et al.* (2012), la recherche d'experts comprend deux tâches principales : *profiler l'expert* et *trouver l'expert*. Nous les détaillons dans les sections suivantes.

2.1.1. *Profiler l'expert*

La recherche d'experts est motivée par une première question : *quelle est l'expertise d'un chercheur ?* Une réponse à cette question consiste à profiler les experts, c'est-à-dire extraire les données appropriées indiquant le domaine et la thématique de la personne afin de les analyser et de les comparer. Il existe différents types de profils d'experts liés au type de preuve d'expertise utilisé. Par exemple, un profil thématique se basera sur le contenu textuel des publications alors qu'un profil social se basera sur les activités relationnelles en situation professionnelle de l'expert.

Le profil d'un expert est très souvent thématique (Balog *et al.*, 2012) et représenté sous forme de vecteur de termes représentant ses connaissances. Elles sont analysées suivant des éléments de connaissance élémentaires. Pour la recherche d'experts, il faut déterminer l'ensemble de ces connaissances élémentaires et mesurer ensuite le niveau d'expertise sur chaque connaissance élémentaire. Le nombre de connaissances utilisées influence la précision et l'efficacité du système. Trop peu de connaissances peut conduite à ne représenter que trop partiellement l'expertise de la personne. Trop de connaissances conduit à un nombre et une complexité de calculs trop importants. La

détermination de l'espace de recherche est également essentielle et repose généralement sur les bases bibliographiques. Cependant, ces données sont souvent hétérogènes et ne sont pas toujours disponibles dans tous les domaines scientifiques. Par exemple, Liang *et al.* (2011) mentionnent l'*ACL Anthology Network* comme seul jeu de données exploitable pour leur modèle de recommandation de publications.

Une fois la zone de recherche déterminée, la modélisation des experts et l'estimation du niveau d'expertise sur chaque aspect de connaissance constituent d'autres problématiques. Ces travaux dépendent du type de preuve d'expertise utilisé. Balog *et al.* (2012) présentent un panorama des différentes approches, chaque approche utilisant un ensemble spécifique de preuves et y appliquant une modélisation ainsi qu'une mesure de similarité ad hoc. De même, Cabanac (2011) utilise les titres d'articles comme preuve et modélise les experts sous forme de vecteurs de termes, le niveau d'expertise de l'expert sur chaque terme étant établi sur la base de la valeur *tf.idf*.

2.1.2. *Trouver l'expert*

Une seconde question liée à la recherche d'experts est : *quels sont les experts dans ce domaine ?* Pour une thématique donnée, il faut estimer le niveau d'expertise des experts-candidats afin de les classer. Il existe un grand nombre de méthodes pour mesurer la force du lien entre une thématique et un expert-candidat. En particulier, on peut estimer le lien thématique par les interactions entre l'objet qui porte le thème donné et l'expert-candidat. Par exemple, si un chercheur a publié de nombreux articles dans une conférence, il est évident que ce chercheur travaille dans le domaine de cette conférence ; si un chercheur a participé à de nombreuses reprises à la conférence, c'est sûrement un expert du domaine de cette conférence.

Une première problématique concerne l'identification de preuves d'expertise de l'expert-candidat. Les preuves sont principalement liées aux publications scientifiques. Ce sont les articles dont l'expert est auteur, les articles qu'il a téléchargés, ceux qu'il a lus ou ceux qu'il a cités. Les activités professionnelles de l'expert sont d'autres preuves montrant efficacement le lien thématique entre experts. Il existe de nombreuses activités et liens professionnels qui peuvent être utilisés comme preuves, par exemple, les liens de collaboration entre collègues, les co-signatures d'articles, les participations à une même conférence ou une même manifestation scientifique, les liens de citation (un article cite l'autre), les liens de co-citation (deux articles citent un même troisième) et les liens de couplage bibliographique (deux articles sont cités par un même troisième), etc.

Une seconde problématique consiste à modéliser et mesurer la similarité thématique entre un thème donné et les experts-candidats. En fonction du type de preuve d'expertise, les approches appliquent une modélisation et une mesure appropriées. Une méthode largement utilisée pour la modélisation thématique repose sur le modèle vectoriel et les mesures de similarité associées. Plus récemment, des approches se sont intéressées aux modèles de graphes, où les objets de recherche et l'espace de recherche sont modélisés sous la forme d'un graphe. Déterminer la proximité thématique dans le graphe se ramène alors un problème de parcours de graphe. Strohman

et al. (2007) ainsi que Liang *et al.* (2011) définissent, par exemple, des modèles basés sur les graphes de citations. Cabanac (2011) propose un modèle basé sur le graphe des co-auteurs et le graphe des participants aux conférences.

2.2. *Utilisation du graphe des citations pour la recherche d'expertise*

En réalisant une citation, un auteur exprime un lien conceptuel entre l'auteur, son article et l'article cité. Les citations peuvent donc être considérées comme une preuve d'expertise intéressante. Plusieurs approches ont ainsi cherché à analyser les relations de citation et à créer des bases de données de citations (Garfield, 1955 ; Nanba et Manabu, 1999 ; Tang *et al.*, 2008 ; Tang *et al.*, 2009 ; Huang et Qiu, 2010). Cependant, un problème est que les citations véhiculent des sens différents, et chaque sens de citation correspond à un niveau de lien entre l'article qui cite et l'article cité. Par exemple, Nanba et Manabu (1999), Tang *et al.* (2009) ainsi que Liang *et al.* (2011) extraient le contexte des citations pour classer les citations, alors que Huang et Qiu (2010) et Liang *et al.* (2011) s'intéressent à la sémantique des citations. En réalité, il existe un grand nombre de raisons de citer (Cronin, 1984). Extraire les citations dans le contenu de l'article et déterminer précisément leur raison d'être est très délicat.

L'exploitation de la relation de citation pour la recherche d'expertise passe généralement par la création d'un graphe de citations. Sur le graphe des citations, la proximité entre deux articles n'est pas seulement attestée par la citation directe entre ces deux articles, mais aussi par la citation indirecte. Clairement, deux articles proches sur un thème ne se citent pas forcément l'un l'autre (en particulier, un article antérieur à un autre ne peut mécaniquement pas le citer). Pour estimer la proximité entre les nœuds dans le graphe des citations, deux types d'approches se distinguent :

– *celles basées sur la proximité entre les nœuds* : c'est-à-dire basées sur la citation directe. Small (1973) examine la relation de co-citation : deux articles sont proches s'ils citent un même troisième article. De même, Kessler (1963) définit la relation de « couplage bibliographique » lorsque deux articles sont cités par un même troisième article. Cependant, ces deux cas exhibent une faiblesse illustrée dans l'exemple dans la Figure 1. Quand trois articles *d1, d2* et *d3* citent un même quatrième article, l'article *d1* est considéré en relation avec les articles *d2* et *d3*, mais on ne peut pas estimer si *d1* est plus proche de *d2* ou *d3* ;

– *celles basées sur la structure du graphe*, c'est-à-dire basées sur la force et le nombre de chemins entre les nœuds du graphe de citations. Liben-Nowell et Kleinberg (2003) ainsi que Liang *et al.* (2011) ont utilisé la distance de Katz (1953) pour estimer la pertinence entre deux nœuds. Dans le cas de Liben-Nowell et Kleinberg (2003), toutes les citations directes sont traitées de la même façon : toute citation est considérée de poids identique. Liang *et al.* (2011) considèrent, par contre, que les citations ont différentes forces. Ils déterminent une force de citation basée sur la classification des citations ; puis la pertinence entre deux nœuds est estimée sur la base de cette différence des citations. Cependant, comme évoqué précédemment, l'utilisation du contexte de citation pour estimer la force du lien nécessite des données qui sont

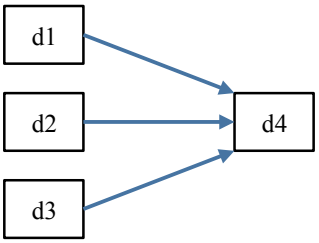

Figure 1. *Proximité des nœuds d1, d2 et d3 basée sur la relation de co-citation.*

difficiles à collecter à large échelle. Un autre problème relatif à la relation de citation entre deux articles concerne l'orientation des arcs : un article ne peut qu'en citer un autre qui a été publié avant. Elle est donc représentée par un arc (c.-à-d., orienté) dans le graphe. Selon Katz (1953), il existe une relation entre deux nœuds quand il y a au moins un chemin reliant l'un à l'autre. Des articles conceptuellement proches (par co-citation ou couplage bibliographique) ne sont cependant pas forcément reliés par un chemin (cf. Figure 2) et il serait préjudiciable d'ignorer cette proximité.

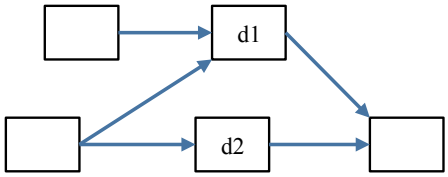

Figure 2. *Exemple de documents proches d1 et d2 pour lesquels il n'existe pourtant pas de chemin entre les co-citations et les couplages bibliographiques.*

2.3. *Système de recommandation d'experts*

Les systèmes de recommandation visent à suggérer des éléments pouvant intéresser un utilisateur sans qu'il ait à expliciter un besoin. Une problématique majeure concerne la détermination des intérêts de l'utilisateur. Concrètement, il s'agit de modéliser les éléments liés à l'utilisateur pour ensuite analyser ces éléments et construire un profil complet représentant les centres d'intérêt de l'utilisateur. Il existe une grande diversité d'approches de systèmes de recommandation. Cependant, deux approches sont principalement utilisées (Bobadilla *et al.*, 2013) : le *filtrage basé sur le contenu* et le *filtrage collaboratif*.

Le filtrage basé sur le contenu utilise les informations (généralement des termes) décrivant les éléments manipulés afin de calculer des similarités entre éléments et recommander des éléments candidats similaires à un élément initial. Par exemple, pour

recommander des articles scientifiques au regard du domaine d'expertise d'un chercheur donné, on cherchera la preuve d'expertise dans le contenu textuel des articles dont il est l'auteur (mais aussi éventuellement les articles qu'il a téléchargés, qu'il a lus, etc.) pour construire un profil et calculer ensuite une similarité entre le contenu d'autres d'articles et le profil du chercheur.

Différentes études ont porté sur la recherche des preuves d'expertise concernant les articles, comme les termes apparaissant dans ces articles. Ces termes sont extraits, par exemple, du contenu des articles, de leur titre (Lin et Wilbur, 2007 ; Lao et Cohen, 2010 ; Cabanac, 2011), de leur résumé (Ekstrand *et al.*, 2010), de leur introduction ou de leurs mots-clés (Huang *et al.*, 2002).

Cependant, il existe d'autres preuves d'expertise, sans contenu textuel, qui ne sont donc pas exploitables par ce type d'approche. Le filtrage basé sur le contenu est souvent combiné à un filtrage collaboratif pour améliorer le système de recommandation. Le filtrage collaboratif a été popularisé par Goldberg *et al.* (1992), et depuis, de nombreux modèles utilisent ce principe. Il s'agit de déterminer les éléments similaires à un élément initial par l'intermédiaire d'éléments tiers. Les preuves d'expertise utilisées par ce type d'approche sont alors les activités professionnelles, les activités scientifiques des éléments initiaux que sont les individus, comme les activités de citation (Liang *et al.*, 2011), de co-signature et de co-participation à des conférences (Cabanac, 2011).

3. Suggestion de membres pour constituer un comité de programme

Suggérer des experts pour renouveler un comité de programme requiert de modéliser les preuves d'expertise des scientifiques de la communauté. Il existe par exemple des preuves spécifiques, comme la participation de l'expert aux comités de programme passés. Nous proposons un modèle de suggestion d'experts basé sur diverses preuves. Les différences entre notre modèle et les travaux existants sont les suivantes :

– au niveau de la modélisation, le graphe de citation est intégré dans un graphe plus riche, où non seulement les articles sont modélisés en tant que nœuds, mais aussi les experts-candidats et la conférence donnée. De même, les liens entre les nœuds représentent différents types de relations ;

– différentes définitions sont proposées pour estimer la force des relations variées dans le graphe (différents arcs). La relation entre une paire d'articles est notamment calculée sans considérer le contexte de citation, pour éviter de dépendre de données « plein texte » souvent indisponibles ;

– le graphe initial est transformé en graphe non-orienté pour déterminer les chemins entre la conférence et un expert candidat. Cela évite de perdre des relations de co-citation et de couplage bibliographique, comme évoqué précédemment.

3.1. *Modélisation de l'espace de recherche*

Les notations utilisées pour la modélisation de l'espace de recherche sont listées dans le Tableau 1.

Notation	Description
Co	Conférence
d	Document correspondant à un article scientifique
d_{citant}	Article d_{citant} qui cite un autre article
$d_{cité}$	Article $d_{cité}$ qui est cité par un autre article
c	Chercheur expert-candidat
l_d	Force du lien de publication entre la conférence et l'article d qui y est publié
$v_{d_{citant},d_{cité}}$	Force du lien de citation de d_{citant} vers $d_{cité}$
$w_{d,c}$	Force du lien d'auteur entre un expert-candidat c et son article d
u_c	Force du lien entre la conférence et l'expert c qui a été membre du comité de programme
M_a	Ensemble des membres du comité de programme de l'édition de l'année a
Q_d	Ensemble des articles qui citent l'article d
R_d	Ensemble des auteurs de l'article d
t_d	Année de publication de l'article d
t_{p_c}	Année de participation de l'expert c au comité de programme p
t_x	Année de nouvelle édition de conférence pour laquelle suggérer des membres de comité

Tableau 1. *Notations utilisées pour la modélisation de l'espace de recherche.*

Nous construisons un graphe orienté avec :

– trois types de **nœuds** :

1) la conférence *Co* donnée pour laquelle on renouvelle le comité de programme (par exemple, SIGIR). Chaque graphe possède un nœud unique de ce type ;

2) les articles publiés à la conférence *Co*, ainsi que tous les articles, extérieurs à la conférence, qu'ils citent et qui les citent ;

3) les auteurs de ces articles et les membres de comités de programme passés.

– et quatre types de **liens** entre les nœuds suivants :

1) d'un article vers un autre pour modéliser une citation ;

2) d'un article publié dans une édition de la conférence *Co* donnée ;

3) d'un auteur vers l'article qu'il co-signe ;

4) d'un membre de comité de programme vers la conférence *Co* donnée.

Pour deux documents d_{citant} et $d_{cité}$, l'auteur de d_{citant} connaît le contenu scientifique de $d_{cité}$ puisqu'il y fait référence dans son article. Il indique ainsi une proximité conceptuelle entre les deux articles. Certaines approches, comme (Liang *et al.*, 2011), parcourent le graphe orienté, d'un document vers d'autres. Dans notre approche, se restreindre au graphe *orienté* ne permet pas d'exploiter toute la richesse des relations de proximité conceptuelle entre l'ensemble des éléments portés par le graphe.

En revanche, notre proposition considère les relations conférence-article, conférence-expert et article-auteur comme non-orientées pour déterminer la proximité

entre la conférence et les nœuds experts-candidats. Un exemple de graphe est présenté dans la Figure 3.

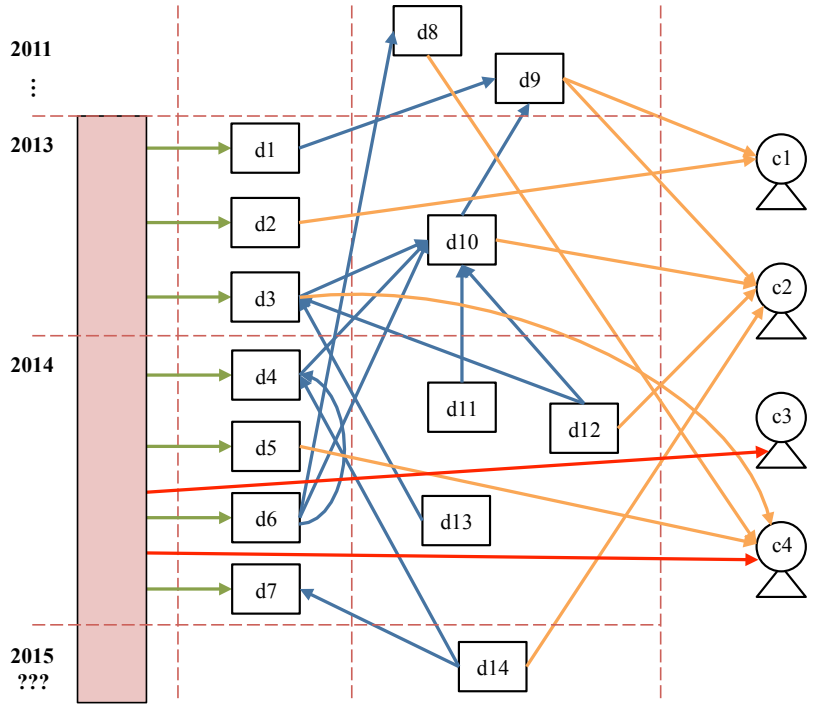

Figure 3. *Exemple de graphe modélisant l'espace de recherche.*

3.2. *Estimation de la force des liens dans le graphe*

Dans cette section, nous proposons une méthode pour mesurer la dépendance entre les objets modélisés qui correspondent aux quatre types de lien ci-dessus.

3.2.1. *Estimation de la force du lien de citation entre articles*

La force du lien de citation est estimée par deux facteurs :

1) l'écart de temps entre la publication de l'article $d_{cité}$ et celle de l'article d_{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) le nombre d'articles qui citent $d_{cité}$. Il s'agit de valoriser les articles citant des articles eux-mêmes très cités, qui témoignent d'un fort impact auprès de la communauté scientifique.

La force du lien de citation quand d_{citant} cite $d_{cité}$ est donc définie par

$$v_{d_{citant},d_{cité}} = \frac{|Q_{d_{cité}}|}{Q_{max}} \cdot \mathrm{e}^{\left(\frac{t_{d_{citant}} - t_{d_{cité}}}{\Delta t_{max}}\right)} \quad [1]$$

où Q_{max} est le nombre maximum d'articles qui citent un article et Δt_{max} est le plus long écart de temps, en valeur absolue, 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.

3.2.2. *Estimation de la force du lien entre une conférence et un article publié*

La relation entre la conférence et ses articles atteste de leur similarité thématique. Cette relation est a priori identique entre la conférence et tous les articles qui y sont publiés. Cependant, les articles les plus cités contribuent davantage à la notoriété de la conférence. De plus, les articles récemment publiés sont porteurs, a priori, des thématiques actuelles. Cette relation dépend ainsi de deux facteurs :

1) du nombre d'articles citant l'article d. 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 d et l'édition de la conférence t_x pour laquelle on souhaite renouveler le comité de programme. Les articles les plus récents traitent des thèmes les plus actuels, et par conséquent plus importants dans une optique de renouvellement de comité de programme.

Ainsi, la force du lien entre une conférence et un article publié lors de l'édition de l'année t_d est estimée par

$$l_d = \frac{|Q_d|}{Q_{max}} \cdot \mathrm{e}^{\left(\frac{t_d - t_x}{\Delta t_{max}}\right)} \quad [2]$$

où Δt_{max} est l'écart de temps maximum, en valeur absolue, entre l'article le plus ancien et l'édition de conférence considérée.

3.2.3. *Estimation de la force du lien entre article et auteur*

Étant donné qu'il est difficile de quantifier la participation relative de chaque co-signataire à un article à la lumière des seules méta-données d'une notice bibliographique, nous considérons un taux participation identique pour chaque co-signataire. La force de la relation entre un article d et un auteur c est donc définie par

$$w_{d,c} = \frac{1}{|R_d|}. \quad [3]$$

3.2.4. *Estimation de la force du lien entre conférence et membre de comité*

Une participation au comité de programme d'une conférence indique que l'expertise d'un individu est reconnue sur les thématiques de la conférence. La dépendance

est définie en fonction de l'écart de temps entre l'année de participation de l'expert au comité de programme t_{p_c} de cette conférence et l'édition de la conférence t_x pour laquelle on souhaite renouveler le comité de programme. Un expert ayant participé au comité de programme d'une édition récente est considéré plus pertinent qu'un expert ayant participé à un comité plus ancien.

Étant donné un expert c ayant une participation au comité de programme p, la force de la relation entre c et la conférence est définie par la formule

$$u_c = \mathrm{e}^{\left(\frac{t_{p_c} - t_x}{\Delta t_{max}}\right)} \quad [4]$$

où Δt_{max} est l'écart de temps maximum, en valeur absolue, entre l'édition de conférence considérée et le plus ancien comité de programme. L'utilisation de l'exponentielle permet d'accentuer les écarts de temps.

3.3. *Émettre une suggestion d'experts pour un comité de programme*

Pour émettre une suggestion, la proximité entre la conférence donnée et l'expert-candidat dépend de deux facteurs. Premièrement, elle dépend du nombre de chemins reliant le nœud de la conférence et le nœud de l'expert-candidat. La proximité augmente avec ce nombre. Deuxièmement, elle dépend de la force de chaque chemin reliant le nœud de la conférence et le nœud de l'expert-candidat. La force d'un chemin est déterminée par la somme des forces (normalisées) des liens qui constituent le chemin. Plus il y aura de chemins avec des forces élevées, plus la proximité entre la conférence et l'expert-candidat sera élevée.

Pour déterminer les chemins entre le nœud de la conférence et le nœud d'un expert-candidat, le graphe est considéré comme non-orienté. Trois types de chemin sont considérés :

– le premier type appelé *auteur externe* (AE) considère l'expert-candidat c en tant qu'auteur d'un article d' qui est cité ou qui cite l'article d de cette conférence. Dans ce cas, le i^{e} chemin possède trois segments dont les forces sont calculées par les fonctions l_d, $v_{d_{citant},d_{cité}}$ et $w_{d,c}$ selon

$$P_c^{(i)}(AE) = \begin{cases} l_d^{(i)} + v_{d,d'}^{(i)} + w_{d',c}^{(i)} & \textbf{si} \text{ l'article } d \text{ publié dans la conférence cite l'article } d', \\ l_d^{(i)} + v_{d',d}^{(i)} + w_{d',c}^{(i)} & \textbf{si} \text{ l'article } d \text{ publié dans la conférence est cité par l'article } d'. \end{cases} \quad [5]$$

– le deuxième type appelé *auteur interne* (AI) considère l'expert-candidat c en tant qu'auteur de l'article d publié par la conférence donnée. Dans ce cas, le i^{e} chemin possède deux segments dont les forces sont calculées par les fonctions l_d et $w_{d,c}$ selon

$$P_c^{(i)}(AI) = l_d^{(i)} + w_{d,c}^{(i)}. \quad [6]$$

– le troisième type appelé *comité de programme* (CP) considère l'expert-candidat c en tant que membre d'un précédent comité de programme. Ici, le chemin possède un seul segment dont la force est calculée par la fonction u_c :

$$P_c^{(i)}(CP) = u_c^{(i)} \quad [7]$$

Pour chaque type de chemin $X \in \{AE, AI, CP\}$ est ensuite calculée la somme des forces de tous les chemins menant du nœud de la conférence à celui de l'expert-candidat selon

$$P_c^{(\star)}(X) = \sum_i P_c^{(i)}(X). \quad [8]$$

Enfin, la proximité entre la conférence et chaque expert-candidat c est déterminée par la somme de tous les chemins entre les deux nœuds :

$$\mathit{Prox}_c = \frac{a \cdot P_c^{(\star)}(AE)}{\max\limits_{c' \in R_\bullet \cup M_\bullet} \left(P_{c'}^{(\star)}(AE)\right)} + \frac{b \cdot P_c^{(\star)}(AI)}{\max\limits_{c' \in R_\bullet \cup M_\bullet} \left(P_{c'}^{(\star)}(AI)\right)} + \frac{c \cdot P_c^{(\star)}(CP)}{\max\limits_{c' \in R_\bullet \cup M_\bullet} \left(P_{c'}^{(\star)}(CP)\right)} \quad [9]$$

où a, b et c sont les paramètres pour ajuster l'importance de chaque facteur motivant l'expertise et $R_\bullet \cup M_\bullet$ désigne l'ensemble des auteurs et membres de comités de programme recensés.

4. Expérimentations

4.1. *Collection de données*

Pour valider notre proposition, nous l'avons expérimentée sur les données des différentes éditions de la conférence SIGIR (*Special Interest Group on Information Retrieval*). L'intérêt de cette conférence est qu'elle fait partie des principales conférences du domaine de la recherche d'information, qu'elle existe depuis 1971 et possède donc de nombreuses éditions et de nombreux comités de programmes. De plus, elle totalise un grand nombre de publications, de chercheurs qui ont participé aux comités de programme ou comme auteurs. Un autre atout 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.

Un premier travail a consisté en la collecte de l'ensemble des articles publiés dans les éditions de la conférence : 3 761 articles des 38 éditions de la conférence SIGIR depuis 1978. Ces articles forment un premier ensemble noté « B1 ». Un second travail de collecte a porté sur les articles cités par les articles de la conférence SIGIR (ensemble noté « B2 »). Nous avons également collecté les articles qui citent les articles de la conférence SIGIR (ensemble « B3 »). Enfin, les participations aux différents comités de programme de la conférence SIGIR ont également été collectées manuellement depuis l'édition 2005, en consultant les actes du congrès.

4.2. *Résultats expérimentaux*

Les expérimentations ont consisté à produire des suggestions pour les CP des cinq éditions de 2011 à 2015. Pour chaque édition, les articles des ensembles B2 et B3 ont été restreints aux éditions précédentes de SIGIR de un à dix ans auparavant. Par exemple, les données de 2001 à 2010 ont été utilisées pour les suggestions du comité de 2011. Nous avons testé quatre configurations de paramètres pour a, b et c indiquées dans les tableaux de résultats présentés.

Les suggestions de membres de CP sont évaluées par rapport à une vérité terrain exploitant l'historique de SIGIR. Ainsi, pour une année donnée, les suggestions sont comparées par intersection avec le CP de cette année là (Tableau 2), puis avec l'ensemble des membres des CP antérieurs (Tableau 3).

Le Tableau 2 présente les résultats (cardinal de l'intersection et taux de recouvrement du CP par les suggestions) pour une suggestion du même nombre (respectivement de 50% en plus et de 100 % en plus) d'experts que le comité correspondant.

Année (a)	**Nb de membres CP** ($\lvert M_a \rvert$)	**Nb d'experts =** $\lvert M_a \rvert$		**Nb d'experts = 150 %** $\lvert M_a \rvert$		**Nb d'experts = 200 %** $\lvert M_a \rvert$	
		Nb d'experts dans M_a	*Taux recouvrement*	*Nb d'experts dans* M_a	*Taux recouvrement*	*Nb d'experts dans* M_a	*Taux recouvrement*
Configuration 1 : a=1 ; b=1 ; c=1							
2011	423	224	52,96 %	249	58,87 %	299	70,69 %
2012	478	215	44,98 %	251	52,51 %	312	65,27 %
2013	403	175	43,42 %	216	53,60 %	229	56,82 %
2014	421	178	42,28 %	235	55,82 %	251	59,62 %
2015	440	213	48,41 %	311	70,68 %	440	100,0 %
Configuration 2 : a=1 ; b=0,5 ; c=0,5							
2011	423	182	43,03 %	252	59,57 %	268	63,36 %
2012	478	186	38,91 %	251	52,51 %	284	59,41 %
2013	403	153	37,97 %	202	50,12 %	232	57,57 %
2014	421	184	43,71 %	225	53,44 %	266	63,18 %
2015	440	201	45,68 %	252	57,27 %	376	85,45 %
Configuration 3 : a=0,5 ; b=1 ; c=0,5							
2011	423	213	50,35 %	249	58,87 %	273	64,54 %
2012	478	218	45,61 %	250	52,30 %	278	58,16 %
2013	403	152	37,72 %	209	51,86 %	225	55,83 %
2014	421	184	43,71 %	235	55,82 %	258	61,28 %
2015	440	201	45,68 %	239	54,32 %	365	82,95 %
Configuration 4 : a=0,5 ; b=0,5 ; c=1							
2011	423	229	54,14 %	253	59,81 %	384	90,78 %
2012	478	248	51,88 %	314	65,69 %	354	74,06 %
2013	403	197	48,88 %	206	51,12 %	262	65,01 %
2014	421	206	48,93 %	224	53,21 %	290	68,88 %
2015	440	326	74,09 %	376	85,45 %	433	98,41 %

Tableau 2. *Comparaison des experts suggérés par notre système avec le CP original de l'année correspondante. Les résultats présentés correspondent à un nombre d'experts suggérés égal à* $\lvert M_a \rvert$*, puis 150 % de* $\lvert M_a \rvert$ *et enfin 200 % de* $\lvert M_a \rvert$*.*

Pour une taille de comité stable d'une année à l'autre, notre méthode arrive à suggérer approximativement la moitié de membres (effectifs) de CP. L'autre moitié comprend des experts suggérés via l'analyse des publications, des citations et des participations aux CP précédents. Ces personnes sont des candidats potentiels à considérer

pour renouveler le CP. Pour évaluer l'implication passée dans des CP de SIGIR de ces experts, le Tableau 3 montre la proportion des experts répondant à ce critère.

Année (a)	**Nb de membres CP** ($\|M_a\|$)	**Nb d'experts =** $\|M_a\|$		**Nb d'experts = 150%** $\|M_a\|$		**Nb d'experts = 200%** $\|M_a\|$	
		Nb d'experts dans $M_{a'}$ *avec* $a' < a$	*Taux recouvrement*	*Nb d'experts dans* $M_{a'}$ *avec* $a' < a$	*Taux recouvrement*	*Nb d'experts dans* $M_{a'}$ *avec* $a' < a$	*Taux recouvrement*
Configuration 1 : a=1 ; b=1 ; c=1							
2011	423	408	96,45 %	612	96,53 %	773	91,37 %
2012	478	456	95,40 %	682	95,12 %	869	90,90 %
2013	403	385	95,53 %	581	96,19 %	754	93,55 %
2014	421	404	95,96 %	600	95,09 %	795	94,42 %
2015	440	421	95,68 %	632	95,76 %	845	96,02 %
Configuration 2 : a=1 ; b=0,5 ; c=0,5							
2011	423	375	88,65 %	568	89,59 %	746	88,18 %
2012	478	414	86,61 %	632	88,15 %	819	85,67 %
2013	403	353	87,59 %	512	84,77 %	703	87,22 %
2014	421	366	86,94 %	533	84,47 %	713	84,68 %
2015	440	392	89,09 %	570	86,36 %	790	89,77 %
Configuration 3 : a=0,5 ; b=1 ; c=0,5							
2011	423	389	91,96 %	591	93,22 %	748	88,42 %
2012	478	419	87,66 %	667	93,03 %	822	85,98 %
2013	403	376	93,30 %	548	90,73 %	741	91,94 %
2014	421	385	91,45 %	564	89,38 %	759	90,14 %
2015	440	408	92,73 %	585	88,64 %	805	91,48 %
Configuration 4 : a=0,5 ; b=0,5 ; c=1							
2011	423	422	99,76 %	624	98,42 %	836	98,82 %
2012	478	474	99,16 %	698	97,35 %	937	98,01 %
2013	403	401	99,50 %	597	98,84 %	780	96,77 %
2014	421	418	99,29 %	624	98,89 %	815	96,79 %
2015	440	434	98,64 %	654	99,09 %	872	99,09 %

Tableau 3. *Comparaison des experts suggérés par notre système avec les CP originaux des années précédentes. Les résultats présentés correspondent à un nombre d'experts suggérés égal à* $|M_a|$*, puis 150 % de* $|M_a|$ *et enfin 200 % de* $|M_a|$*.*

Les résultats montrent l'effet des configurations sur l'ampleur du renouvellement suggéré pour les CP par des experts n'ayant jamais siégé. Par exemple, la configuration 4 favorise l'endogamie en promouvant les membres de CP précédents qui représentent quasiment 100 % des suggestions. Elle contraste avec la configuration 2 qui favorise l'exogamie en suggérant environ 15 % de scientifiques n'ayant pas siégé en CP quoique publiant dans le même espace intellectuel que la conférence ciblée. Ces deux exemples soulignent le potentiel d'adaptation des suggestions aux besoins exprimés par les présidents de CP souhaitant tantôt entretenir tantôt renouveler les pairs en charge de l'évaluation.

5. Conclusion et perspectives

Pour suggérer des experts afin de renouveler le CP d'une conférence, nous avons défini un modèle intégrant différents objets et relations. Ceux-ci constituent un graphe portant les preuves d'expertises observées dans diverses sources liées aux conférences, articles et participants à des CP.

Les suggestions produites par notre modèle ont été confrontées aux CP soustendant plusieurs éditions de SIGIR. Un point délicat de la validation de notre approche concerne la vérité terrain employée. Faute de pouvoir questionner les prési-

dents de CP à propos des listes que nous suggérons, nous avons fait l'hypothèse que la suggestion d'un membre de CP « effectif » (potentiellement d'un CP dans le passé) était pertinente. C'est cependant une approche biaisée, en défaveur des scientifiques proches de la communauté ciblée quoique n'ayant jamais siégé au CP.

Nous cherchons à relever un défi : suggérer des scientifiques experts pour une conférence sans exploiter le contenu textuel (au sens RI) des articles publiés. Cette contrainte vient notamment des barrières posées par la plupart des éditeurs bloquant l'accès au plein texte des publications. Par conséquent, il nous semble pertinent d'exploiter les relations de citation et de co-signature traduisant une proximité conceptuelle entre scientifiques et la communauté formée par une conférence. Afin de promouvoir certains critères dans les suggestions (tels que le genre, l'âge et la localisation géographique), nous envisageons d'intégrer ces caractéristiques des scientifiques dans de futurs travaux.

6. Bibliographie

Afzal M., Maurer H., « Expertise Recommender System for Scientific Community », *Journal of Universal Computer Science*, vol. 17, p. 1529-1549, 2011.

Balog K., Fang Y., de Rijke M., Serdyukov P., Si L., « Expertise Retrieval », *Foundations and Trends in Information Retrieval*, vol. 6, p. 127-256, 2012.

Bobadilla J., Ortega F., Hernando A., Gutiérrez A., « Recommender systems survey », *Knowledge-Based Systems*, vol. 46, p. 109 - 132, 2013.

Cabanac G., « Accuracy of inter-researcher similarity measures based on topical and social clues », *Scientometrics*, vol. 87, n° 3, p. 597-620, 2011.

Cronin B., *The Citation Process : The Role and Significance of Citations in Scientific Communication*, Taylor Graham, London, 1984.

De Bellis N., *Bibliometrics and Citation Analysis*, Scarecrow Press, Lanham, 2009.

Ekstrand M., Kannan P., Stemper J. A., Butler J. T., Konstan J. A., Riedl J., « Automatically building research reading lists », *Conference on Recommender Systems (RecSys 2010)*, p. 159-166, 2010.

Fang H., Zhai C., « Probabilistic Models for Expert Finding », *29th European conference on IR research (ECIR'07)*, p. 418-430, 2007.

Garfield E., « Citation Indexes for Science : A New Dimension in Documentation through Association of Ideas », *Science*, vol. 122, p. 108-111, 1955.

Goldberg D., Nichols D., Oki B. M., Terry D., « Using collaborative filtering to weave an information tapestry », *Communications of the ACM – Special issue on information filtering*, vol. 35, n° 12, p. 61-70, 1992.

Huang Z., Chung W., Ong T., Chen H., « A graph-based recommender system for digital library », *ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*, p. 65-73, 2002.

Huang Z., Qiu Y., « A multiple-perspective approach to constructing and aggregating Citation Semantic Link Network », *Future Generation Computer Systems*, vol. 26, n° 3, p. 400-407, 2010.

Katz L., « A new status index derived from sociometric analysis », *Psychometrika*, vol. 18, n° 1, p. 39-43, 1953.

Kessler M., « Bibliographic coupling between scientific papers », *American Documentation*, vol. 14, p. 10-25, 1963.

Lao N., Cohen W., « Relational retrieval using a combination of path-constrained random walks », *Machine Learning*, vol. 81, p. 53-67, 2010.

Liang Y., Li Q., Qian T., « Finding Relevant Papers Based on Citation Relations », *Web-Age Information Management*, vol. 6897, p. 403-414, 2011.

Liben-Nowell D., Kleinberg J., « The link-prediction problem for social networks », *Twelfth international conference on Information and knowledge management (CIKM'03)*, p. 556-559, 2003.

Lin J., Wilbur W. J., « PubMed related articles : a probabilistic topic-based model for content similarity », *BMC Bioinformatics*, 2007.

Liu P., Dew P., « Using Semantic Web Technologies to Improve Expertise Matching within Academia », *International conference I-KNOW'04*, p. 370-378, 2004.

Macdonald C., Ounis I., « Voting techniques for expert search », *Knowledge and Information Systems*, vol. 16, n° 3, p. 259-280, 2008.

McDonald D. W., Ackerman M. S., « Expertise Recommender : A Flexible Recommendation System and Architecture », *ACM conference on Computer supported cooperative work (CSCW'00)*, p. 231-240, 2000.

Nanba H., Manabu M., « Towards multi-paper summarization reference information », *16th international joint conference on Artificial intelligence (IJCAI'99)*, p. 926-931, 1999.

Rodriguez M., Bollen J., « An Algorithm to Determine Peer-Reviewers », *17th ACM conference on Information and knowledge management (CIKM'08)*, p. 319-328, 2008.

Serdyukov P., Hiemstra D., « Modeling Documents As Mixtures of Persons for Expert Finding », *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, Springer-Verlag, Berlin, Heidelberg, p. 309-320, 2008.

Small H., « Co-citation in the scientific literature : A new measure of the relationship between two documents », *Journal of the American Society for Information Science*, vol. 24, p. 265-269, 1973.

Strohman T., Croft W., Jensen D., « Recommending Citations for Academic Papers », *30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07)*, p. 705-706, 2007.

Tang J., Zhang J., Yao L., Li J., Zhang L., Su Z., « ArnetMiner : extraction and mining of academic social networks », *14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, p. 990-998, 2008.

Tang J., Zhang J., Yu J. X., Yang Z., Cai K., Ma R., Zhang L., Su Z., « Topic Distributions over Links on Web », *Ninth IEEE International Conference on Data Mining (ICDM'09)*, p. 1010-1015, 2009.