

Un modèle de Recherche d'Information Sociale pour l'Accès aux Ressources Bibliographiques : Vers un réseau social pondéré

Lamjed Ben Jabeur*, Lynda Tamine* et Mohand Boughanem*

*IRIT, Université Paul Sabatier
118 Route de Narbonne
F-31062 Toulouse CEDEX 9 , FRANCE
{jabeur, tamine, bougha}@irit.fr

Résumé. Cet article propose une nouvelle approche, basée sur les réseaux sociaux, pour l'accès aux ressources bibliographiques. Nous introduisons un modèle d'information sociale dont les auteurs sont les principales entités et les relations sont extraites à partir des liens de coauteur et de citation. En effet, ces relations sont pondérées en tenant compte des interactions entre les auteurs et des annotations sociales produites par les utilisateurs. Dans ce modèle, la pertinence d'un document est estimée par combinaison de la pertinence thématique et de la pertinence sociale, qui est à son tour dérivée de l'importance sociale des auteurs associés. Nous évaluons la viabilité de notre modèle sur une collection d'articles scientifiques dont les annotations sociales sont extraites depuis le réseau social académique *CiteULike.org*. Les résultats obtenus montrent la supériorité des performances de notre modèle par rapport à la recherche d'information traditionnelle.

1 Introduction

Les moteurs de recherche académiques, tels que *GOOGLE SCHOLAR*¹ et *CITeseerX*², ont donné aux chercheurs la possibilité d'accéder à diverses sources d'information scientifique et de gérer leurs références bibliographiques. Depuis leur apparition, les systèmes de recherche d'information scientifique ont été confrontés à une problématique majeure qui consiste à évaluer l'importance des publications scientifiques. Les premiers travaux y ont apporté une solution en utilisant les indicateurs bibliométriques. Certains travaux successeurs ont modélisé les ressources bibliographiques avec des structures hypertextes dont les hyperliens représentent les citations. Dans une telle approche, l'importance des documents est calculée en appliquant l'algorithme de *PageRank* sur le graphe de citation.

¹<http://scholar.google.fr/>

²<http://citeseerx.ist.psu.edu/>

Avec l'apparition des réseaux sociaux académiques tels que CITEULIKE³ et ACADEMIA⁴, cette vision a été élargie en considérant que les articles scientifiques sont produits et consommés par des entités sociales et leur importance peut être estimée à partir du contexte de production et d'utilisation. Cette approche a été portée par le courant de la recherche d'information sociale où les acteurs sont représentés au moyen d'un réseau social et la pertinence du document est calculée en appliquant les mesures de centralité sociale introduites par l'analyse des réseaux sociaux Wasserman et al. (1994). En effet, la recherche d'information sociale suppose qu'un document pertinent est produit par un auteur important d'où l'importance scientifique des ressources bibliographiques peut être dérivée à partir de l'importance sociale des auteurs associés.

Dans ce contexte, inspirés des travaux de Mutschke (2001) et Kirsch et al. (2006) qui s'intéressent à l'accès aux ressources bibliographiques et à la représentation des auteurs par un réseau social, nous proposons un modèle générique de la recherche d'information sociale qui est déployé particulièrement pour l'accès aux ressources bibliographiques. Comparativement aux approches précédentes, ce modèle comprend des nouvelles entités sociales représentées par les annotateurs et les annotations sociales, des nouvelles relations sociales telles que la citation et l'annotation et des mesures de pondération attribuées aux différentes relations du réseau.

Le reste de cet article est organisé comme suit. Dans la section 2, nous présentons une synthèse des approches proposées pour l'accès aux ressources bibliographiques. La section 3 donne un aperçu de notre modèle générique de recherche d'information sociale. La section 4 décrit l'instanciation du modèle générique dans le cadre précis de l'accès aux ressources bibliographiques. La section 5 décrit les expérimentations menées pour valider notre modèle. Enfin, la section 6 conclut le papier et annonce les perspectives.

2 Recherche d'information dans les ressources bibliographiques

Les travaux précurseurs dans le domaine de la recherche d'information dans les ressources bibliographique ont considéré les liens de citation comme étant un indicateur de qualité et d'autorité des publications scientifiques. Nous citons dans ce contexte les indicateurs bibliométriques basés sur le nombre de citations tel que le facteur d'impact mesurant l'importance d'une publication ou d'une revue scientifique Garfield (2006). Cependant, l'indice de citation est insuffisant pour estimer la pertinence d'un document. Des nouvelles mesures qui tiennent compte de la croissance de citations reçues sont alors proposées pour ordonner les documents selon leurs âges et le nombre de citation prévues Hauff et Azzopardi (2005) Meij et de Rijke (2007). D'autre part, certaines approches représentent les citations par des hyperliens et modélisent les ressources bibliographiques sous forme d'un graphe. Dans ce cas, les documents sont classés selon leurs autorités calculées par les algorithmes de la recherche d'information hypertexte tels que *PageRank* et *HITS* Page et al. (1998) Langville et Meyer (2005).

Certains travaux récents réutilisent les mesures de centralité introduites par le domaine d'analyses des réseaux sociaux telles que les mesures de *Betweenness* et de *Closeness* afin

³<http://www.citeulike.org/>

⁴<http://www.academia.edu/>

d'identifier les ressources centrales dans le graphe des documents Bollen et al. (2005) ou pour déduire la pertinence d'un document à travers la centralité de ses auteurs appliquant ainsi ces mesures sur le graphe des auteurs Mutschke (2001) Kirchhoff et al. (2008). Pour représenter le réseau social, la plupart des travaux se limitent à des simples liens de coauteur entre les nœuds du graphe Yan et Ding (2009) Mutschke (2001). Toutefois, les modèles présentés par Newman (2000) et Liu et al. (2005) proposent de pondérer les liens entre les auteurs selon la fréquence et l'exclusivité de leurs associations de coauteur.

D'autres modèles de recherche d'information sociale intègrent les documents comme des nœuds du réseau social Korfiatis et al. (2006). Dans ce cas, les relations entre les documents et les auteurs sont extraites à partir des interactions de collaboration, de publication et de citation. Des telles approches interprètent la pertinence par le degré de confiance, d'autorité et de popularité des documents dans le réseau social Kazai et Milic-Frayling (2008). Ces facteurs de pertinence sociale sont modélisés soit par des probabilités de transition sur le graphe social (approche intégrée) Amer-Yahia et al. (2007), soit par un score combiné (approche modulaire) Kirchhoff et al. (2008) Kirsch et al. (2006).

Inscrit dans ce cadre, nous proposons un modèle de recherche d'information sociale qui associe la pertinence des ressources bibliographiques à l'importance sociale de leurs auteurs. Comparativement aux travaux proches du domaine et de notre précédente contribution Tamine et al. (2009), notre proposition présentée dans ce papier s'en distingue par les points clés suivants :

1. nous proposons une formalisation d'un modèle générique de recherche d'information sociale,
2. en plus des liens de coauteur, nous exploitons deux autres types de relations sociales : la citation et l'annotation sociale,
3. nous attribuons à ces relations des poids qui tiennent compte de la position des acteurs dans le réseau social et de leurs mutuelles collaborations.

3 Le modèle générique de recherche d'information sociale

Un modèle de recherche d'information offre un support théorique pour représenter des documents et des requêtes et mesure leur degré de similitude assimilée à la pertinence. Formellement, et en se basant sur la représentation proposée par Baeza-Yates et Ribeiro-Neto (1999), nous décrivons le modèle générique de recherche d'information sociale par un quintuplet $[D, Q, G, F, R(q_i, d_j, G)]$ où D est l'ensemble des documents, Q est l'ensemble des requêtes, G est le réseau d'information sociale, F est la fonction d'appariement des documents et des requêtes et $R(q_i, d_j, G)$ est une fonction de classement qui intègre divers facteurs de la pertinence sociale et qui tient compte de la topologie du réseau. Cette fonction peut être définie par la combinaison de sous-ensemble des facteurs de la pertinence sociale suivants : la pertinence thématique, l'importance sociale des acteurs, la distance sociale, la popularité, la fraîcheur de l'information et le nombre de marque-pages reçus Amer-Yahia et al. (2007).

En ce qui concerne le réseau d'information sociale G , il représente les entités sociales qui interagissent au voisinage du document. Nous proposons donc d'y inclure tous les acteurs et les données qui permettent d'évaluer sa pertinence sociale comme illustré dans la figure 1. Les acteurs y représentent les producteurs et les consommateurs d'information (respectivement les

auteurs et les utilisateurs) tandis que les données comprennent les documents et les annotations sociales (les *tags*, les votes, et les avis). Dans le cadre de leurs collaborations et interactions sociales, les acteurs participent à produire de l'information et à enrichir les documents par les annotations sociales.

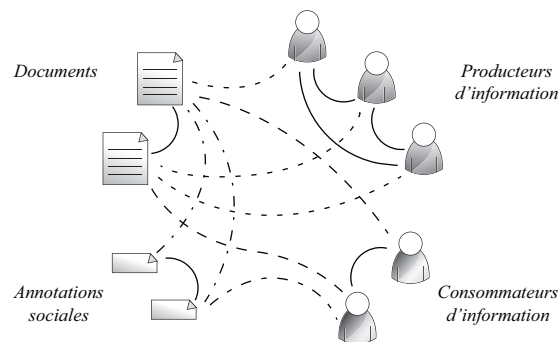


FIG. 1 – Le réseau d'information sociale

Le réseau d'information sociale peut être représenté par un graphe $G = (V, E)$ où l'ensemble des nœuds $V = A \cup U \cup D \cup T$ représente les entités sociales avec A , U , D et T correspondant respectivement aux auteurs, aux utilisateurs, aux documents et aux annotations sociales. L'ensemble des arcs $E \subseteq (V \times V)$ représente les relations sociales reliant les différents types des nœuds (publier, co-auteur, amitié, citation, annotation...etc.).

Le réseau d'information sociale est appréhendé différemment. Du point de vue du producteur d'information, le réseau social regroupe les nœuds documents et auteurs et met en évidence le contexte social de la production des ressources. De même, la vue consommateurs d'information représente le contexte d'utilisation sociale des documents et l'interaction entre les utilisateurs. Cette vue intègre 3 types des nœuds à savoir : les documents, les annotations sociales et les utilisateurs.

Dans la suite, nousinstancions ce modèle générique au cadre de la production et de la consommation de ressources bibliographiques.

4 Le réseau d'information sociale des ressources bibliographiques

Les travaux du domaine modélisent essentiellement le réseau social des ressources bibliographiques en se basant uniquement sur le point de vue producteurs d'information. Cependant, l'introduction des réseaux sociaux académiques sur le web (par exemple CITEULIKE et ACADMEICA) a permis aux utilisateurs de participer et de fournir par la suite des descripteurs sociaux aux ressources bibliographiques. Contrairement aux réseaux des amis tels que FACEBOOK⁵ et MYSPACE⁶ et dont les relations entre les individus expriment principalement un lien

⁵<http://www.facebook.com/>

⁶<http://www.myspace.com/>

d'amitié, les réseaux sociaux académiques incluent des relations sociales spécifiques entre les entités d'information.

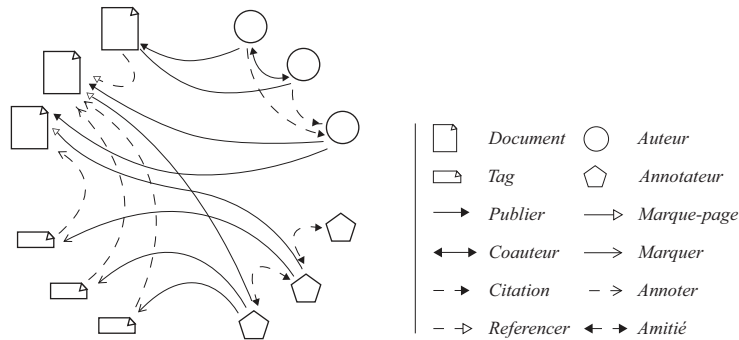


FIG. 2 – Le réseau d'information sociale pour les ressources bibliographiques

D'un point de vue des producteurs d'information, nous identifions les relations sociales suivantes qui impliquent les documents et les auteurs :

- Publier** : une relation dirigée relie chaque auteur $a_i \in A$ avec sa publication $d_j \in D$.
- Référencer** : une relation dirigée qui associe un document $d_i \in D$ avec ses références bibliographiques,
- Coauteur** : une relation entre deux auteurs $a_i, a_j \in A$ ayant collaboré pour produire un document en commun.
- Citation** : une relation sociale dirigée entre un auteur $a_i \in A$ et un second auteur $a_j \in A$ avec a_j est cité dans l'un des documents écrit par a_i .

Du point de vue des consommateurs d'information, nous identifions les relations sociales suivantes qui intègrent les documents, les tags et les annotateurs :

- Marque-page** : en attribuant un tag à un document, l'annotateur $u_i \in U$ et le document $d_j \in D$ sont associés avec une relation sociale de marque-page.
- Annoter** : relie un tag $t_j \in T$ avec un document $d_i \in D$ assigné au moins une fois pour décrire son contenu.
- Marquer** : relie un annotateur $u_i \in U$ et un tag $t_j \in T$ utilisé au moins une fois pour marquer un document.
- Amitié** : relie deux annotateurs $u_i, u_j \in U$. Cette relation peut être décrite explicitement par un lien d'amitié ou implicitement par le fait d'appartenir à un même groupe.

Le réseau d'information sociale pour les ressources bibliographiques peut être représenté en utilisant une notation graphique illustrée dans la figure 2.

4.1 Pondération des relations sociales

Les arcs reliant les nœuds sociaux expriment divers types de relations sociales et permettent d'optimiser d'une façon significative le processus d'exploration du réseau social. Si nous ex-

plorons le voisinage social d'un nœud⁷, les poids sur les arcs nous permettent de sélectionner le nœud suivant à chaque saut. Dans cet article, nous nous intéressons essentiellement au réseau social des publications scientifiques, pour cela nous définissons un modèle de pondération pour les associations auteur-auteur $e(a_i, a_j) \in (A \times A)$ et les associations auteur-document $e(a_i, d_j) \in (A \times D)$.

a. La relation de coauteur : représentée par un arc dirigé, cette relation connecte deux coauteurs ayant collaboré pour produire un document. Les coauteurs ont souvent des contacts personnels directs néanmoins la multiplicité de leurs collaborations exprime la similarité et le partage d'intérêt entre eux. En effet, les auteurs des publications scientifiques ont tendance à s'échanger les connaissances et diversifier leurs collaborations. Pour cette raison et afin de quantifier la similarité entre les coauteurs, nous proposons de tenir compte de la totalité des collaborations. Nous proposons d'assigner des poids asymétriques aux relations de coauteur comme suit :

$$Co(i, j) = \frac{A(i, j)}{A(i)} \quad (1)$$

Avec $A(i, j)$ est le nombre de documents co-écrits par les auteurs a_i et a_j . $A(i)$ représente le nombre des documents publiés par l'auteur a_i .

b. La relation de citation : représentée par un arc dirigé, les liens de citation expriment le transfert de connaissances entre les auteurs des publications scientifiques. Par conséquent, plus un auteur cite les publications d'un second auteur, plus il est influencé par ses idées d'une manière que tout les deux partagent des sujets similaires. Cette relation est asymétrique et son importance est souvent proportionnelle au nombre des publications. Afin de mesurer l'importance de cette association, nous tenons compte du nombre des citations entre les auteurs ainsi que le nombre total des citations énoncées par l'auteur source de la relation. Les relations de citation sont alors pondérées comme suit :

$$Ci(i, j) = \frac{C(i, j)}{C(j)} \quad (2)$$

Avec $C(i, j)$ est le nombre de fois que l'auteur a_i a cité l'auteur a_j et $C(i)$ représente le nombre de citations énoncées par l'auteur a_i .

c. La relation de publication : un auteur sera plus affilié à un sujet S s'il l'a fréquemment abordé dans ses publications. Ainsi, un coauteur sera davantage associé à son document d que ses coauteurs s'il a publié plusieurs documents sur le même sujet de d . Pour estimer les connaissances d'un coauteur a_k sur le sujet de son document d , nous proposons de comparer la quantité d'information importée via ses autres publications. Du point de vue des consommateurs, cela peut être estimé par la distribution des *tags* affectés au sous-ensemble des publications de chaque coauteur représenté par \mathcal{A}_k avec $a_k \in A$ est le coauteur que nous souhaitons mesurer l'affiliation au sujet de son document d . Nous calculons ainsi une distribution

⁷Voisinage social : l'ensemble des nœuds de proximité accessibles directement ou indirectement à partir d'un nœud du réseau.

de probabilité de l'ensemble des *tags* T assignés au document d dans la sous-collection des documents publiés par les coauteurs du document d .

$$w(a_k, d) = \sum_{t_i \in T} \frac{tf(t_i, \mathcal{A}_k)}{tf(t_i, \mathcal{A})} \quad (3)$$

Avec T l'ensemble des tags assignés au document d . $\mathcal{A} = \bigcup_{k=1}^m \mathcal{A}_k$ représente la sous-collection des documents publiée par les m coauteurs du document d . $tf(t_i, \mathcal{A}_k)$ est la fréquence de tag t_i dans le sous-ensemble des documents \mathcal{A}_k publiés par l'auteur a_k . $tf(t_i, \mathcal{A})$ représente la fréquence de tag dans la sous-collection des documents publiés par les coauteurs du document d .

Certains algorithmes de centralité sociale ne supportent pas la multiplicité des arcs de même sens entre deux nœuds. Nous proposons donc de combiner les poids des relations de coauteur et de citation comme suit :

$$w(a_i, a_j) = \frac{1}{4} * (1 + Co(i, j)) * (1 + Ci(i, j)) \quad (4)$$

4.2 Estimation de la pertinence sociale des documents

La sémantique de l'importance sociale des documents dépend de la nature de l'application. Par exemple, l'importance sociale d'un article de blog est estimé à travers la popularité et le nombre des *tags* reçus. La mesure de *Degree* est alors la plus appropriée pour calculer son importance. Pour notre part, l'objectif est de sélectionner la mesure d'importance sociale qui met en évidence les ressources bibliographiques de qualité. Ainsi, nous calculons pour chaque auteur a_i un score d'importance sociale $C_G(a_i)$ en utilisant une des mesures d'importance suivantes : *Betweenness*, *Closeness*, *PageRank*, le score "Authority" de *HITS* et le score "Hub" de *HITS*. Ces mesures sont appliquées seulement sur le sous-graphe d'auteurs $G_a = (A, E_a)$ avec $E_a \subseteq (A \times A)$. Les arcs sont pondérés comme décrit précédemment et désignent soit une relation de coauteur ou un lien de citation. Ensuite, un score d'importance sociale est propagé aux documents par une somme pondérée des scores sociaux des ses auteurs :

$$Imp_G(d) = \sum_{i=1}^m w(a_i, d) C_G(a_i) \quad (5)$$

Nous combinons par la suite le score $Imp_G(d)$ avec une métrique de la recherche d'information traditionnelle. L'idée consiste à estimer la pertinence du document dans le graphe social et de présenter une réponse plus précise à l'utilisateur, en combinant la pertinence thématique et l'importance sociale. Intuitivement, un utilisateur est susceptible d'évaluer un document comme pertinent s'il couvre le sujet de la requête et si les auteurs correspondants sont socialement importants. Sur cette base, nous définissons la fonction de classement R par la combinaison linéaire de deux scores normalisés de la pertinence comme suit :

$$R(q, d, G) = \alpha RSV(q, d) + (1 - \alpha) Imp_G(d) \quad (6)$$

Avec $\alpha \in [0, 1]$ est un paramètre de pondération, $RSV(q, d)$ est une mesure de similarité thématique entre la requête q et un document d et $Imp_G(d)$ est l'importance sociale du document d dans le réseau social G .

5 Évaluation expérimentale

Dans le but d'évaluer l'impact de notre modèle étendu sur l'efficacité de la recherche, nous avons mené une série d'expérimentations sur une collection d'articles scientifiques. Les objectifs de cette évaluation sont de :

- mesurer l'impact de la pondération des relations sociales sur l'estimation de la pertinence sociale des documents comme proposé précédemment,
- comparer les différentes mesures d'importance sociale afin de déterminer la mesure permettant de mieux exprimer l'importance des ressources bibliographiques,
- mener une évaluation comparative de notre modèle relativement à un modèle recherche d'information classique.

5.1 Cadre d'évaluation

Les campagnes d'évaluations tel que TREC proposent un cadre standard pour évaluer et comparer les systèmes de recherche d'information. Cependant, les collections disponibles ne sont pas adaptées pour évaluer les modèles de recherche d'informations sociale en l'absence des données indispensables à la construction du réseau social. Afin de valider notre proposition, nous avons construit un corpus des documents scientifiques issus de la conférence ACM SIGIR de 1978 à 2008. Nous décrivons dans la suite les caractéristiques de la collection des documents ainsi que les requêtes et les mesures d'évaluation utilisées.

– *Corpus des ressources bibliographiques*

La collection SIGIR comprend 2871 auteurs avec une moyenne de 2 relations de coauteur et 16 liens de citation par auteur. Comme indiqué au tableau 1, les relations de citation dominent le réseau social avec 9 fois plus des liens que les associations de coauteur. En intégrant les relations de citation dans le réseau social, les communautés dispersées et de petite taille se restructure en plus larges composantes connexes. Par conséquent, la composante "géante" reliant la majorité des nœuds est élargie pour inclure 84% des auteurs comme le montre la figure 3.

Nombre de documents	2053
Nombre d'auteurs	2871
Relations de coauteur	5047
Relations de citation	45880
Relations de coauteur et/ou citation	52516
Composant le plus large	2430 (84%)

Tableau. 1 – *Caractéristiques statistiques de corpus SIGIR*

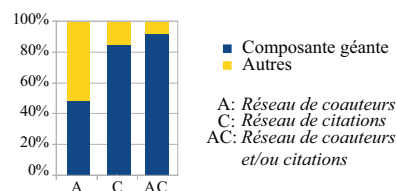


Figure. 3 – *Composante géante du réseau social SIGIR*

En plus du contenu textuel des publications scientifiques, le corpus SIGIR inclut d'autres informations sur les auteurs et les citations. Un auteur fait partie du réseau social s'il a publié au moins un article dans le cadre de la conférence ACM SIGIR. Deux auteurs sont en relation sociale à condition :

- Qu'ils aient publié un article en commun dans la conférence ACM SIGIR.
- Que dans son article SIGIR, le premier auteur cite une publication du deuxième auteur. L'article cité ne doit pas forcément appartenir au corpus SIGIR .

Nous avons enrichi ce corpus avec les données collectées depuis le réseau social académique CITEULIKE. Nous regroupons ainsi tous les *tags* utilisés pour annoter les documents SIGIR ainsi que les identifiants des utilisateurs correspondants. Pour indexer cette collection, nous avons utilisé le moteur de recherche TERRIER⁸.

– *Collection de requêtes test et jugements de pertinence associés*

Étant donné que les *tags* sont des mots clés générés par les utilisateurs dans le but d'annoter les documents et que les requêtes sont une représentation des besoins des utilisateurs d'un besoin en information, les *tags* peuvent alors constituer les requêtes dans notre cadre d'évaluation. Nous considérons que les *tags* les plus populaires sont de haute importance sociale et nous sélectionnons comme des requêtes les 25 *tags* les plus utilisés pour annoter les documents SIGIR.

Pour constituer la collection des documents pertinents, nous supposons qu'un document est pertinent s'il est annoté au moins une seule fois par le *tag* (requête) et que ce dernier est parmi les 3 *tags* les plus affectés aux documents. La collection finale contient 25 requêtes et 223 documents pertinents avec une moyenne de 8.9 documents par requête.

– *Mesures d'évaluation*

Afin d'étudier les performances de notre modèle et de comparer les mesures d'importance sociale nous étudions les précisions aux 5^{ème} et 10^{ème} documents résultats que nous notons respectivement $p@5$ et $p@10$. Ces mesures évaluent la capacité du système à retourner des résultats pertinents parmi les premiers documents retournés.

5.2 Comparaison des mesures d'importance sociale

Les différentes mesures d'importance sociale mettent en évidence les entités clés d'un réseau social. Ces mesures ont une sémantique qui varie d'une application sociale à une autre. Dans le contexte des publications scientifiques, la mesure de *Betweenness* est considérée comme un indicateur d'interdisciplinarité et met en évidence les auteurs connectant plusieurs partitions dispersées de la communauté scientifique. La mesure de *Closeness*, basée sur les chemins les plus courts, reflète la proximité et l'indépendance d'un auteur à son voisinage social. Les mesures de *PageRank* et le score d' *Authority* de *HITS* distinguent les sources d'autorité dans le réseau social. En revanche, le score de *Hub* de *HITS* identifie les auteurs ayant une importante activité sociale tout en se basant sur des sources d'autorité, appelés les auteurs "Centraux".

Nous avons appliqué les mesures d'importance sociale, à savoir : *Betweenness*, *Closeness*, *PageRank*, *Authority* (*HITS*), et *Hub* (*HITS*) sur un modèle binaire et pondéré du réseau social. Nous notons l'application de ces mesures sur le modèle pondéré du réseau respectivement

⁸<http://www.terrier.org>

par *W-Betweenness*, *W-Closeness*, *W-PageRank*, *W-Authority*, et *W-Hub*. Les performances de recherche sont présentées dans le tableau 2.

	p@5	p@10		p@5	p@10
Closness	0,0211	0,0526	W-Closness	0,0316	0,0579
Betweenness	0,0421	0,0526	W-Betweenness	0,0316	0,0316
PageRank	0,0211	0,0421	W-PageRank	0,0316	0,0421
Authority	0,0316	0,0368	W-Authority	0,0316	0,0368
Hub	0,0316	0,0579	W-Hub	0,0421	0,0632

TAB. 2 – Comparaison des mesures d'importance sociale

En comparant les précisions $p@5$ et $p@10$, nous constatons que la mesure de *Hub* permet de mieux classer les ressources bibliographiques retournées initialement par un modèle de recherche d'information classique. Nous concluons donc que l'importance des publications scientifiques peut être estimée par la *Centralité* de leurs auteurs. En effet, la pertinence sociale dans le contexte de ressources bibliographiques est interprétée par l'intense activité de l'auteur, proportionnelle au nombre de publications et de collaborations, tout en s'appuyant sur des ressources d'autorité.

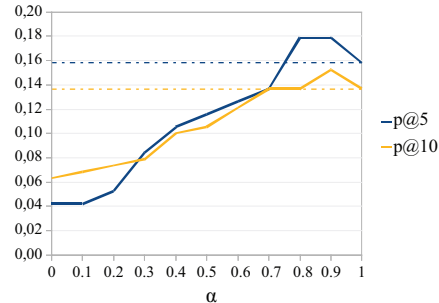
Pour la plupart des mesures d'importance, le modèle pondéré permet d'améliorer l'efficacité de la recherche. Cela est constaté avec les valeurs des précisions obtenues par les mesures de *W-Closeness*, de *W-PageRank* et de *W-Hub* dépassant leurs analogues appliquées sur un modèle binaire du réseau social. Nous concluons donc que les propriétés exprimées via la pondération des relations sociales, à savoir le partage des centres d'intérêt, l'influence et le transfert de connaissances, permettent de mieux identifier les auteurs *Centraux* et par la suite estimer la pertinence des ressources bibliographiques.

Pour évaluer l'efficacité de notre modèle, nous retenons la mesure de *W-Hub* comme étant la mesure qui permet de mieux exprimer l'importance sociale des ressources bibliographiques.

5.3 Evaluation de l'efficacité de notre modèle

Dans l'étape précédente, les résultats sont ordonnés uniquement selon les scores sociaux des documents. Dans ce cas, les mesures de précision $p@5$ et $p@10$ ne dépassent pas le seuil de 46% comparé à celles du système de recherche d'information classique basé sur le modèle *Okapi BM25* Jones et al. (2000) dont $p@5 = 0,158$ et $p@10 = 0,137$. Les mesures d'importance sociale ne sont donc pas capables de trier les résultats sans prendre en considération la similarité entre le document et la requête. Nous nous intéressons à présent à une combinaison linéaire de deux scores pour estimer la pertinence d'un document à la requête, comme indiqué dans la formule 6.

Nous avons étudié l'impact du paramètre α sur le processus de recherche d'information, et ceci pour la mesure de *W-Hub*. Lorsque $\alpha = 0$, seule la pertinence sociale est prise en considération. D'autre part, $\alpha = 1$ correspond au *baseline* BM25 puisque seule la pertinence thématique est prise en considération pour classer les documents. L'analyse des mesures $p@5$ et $p@10$ en fonction du paramètre α montre que les courbes, présentées sur la figure 4, présentent des pics dont les valeurs dépassent la valeur obtenue pour $\alpha = 1$, et cela lorsque la pertinence thématique est uniquement prise en considération. Donc, la combinaison des deux scores per-

FIG. 4 – Ajustement du paramètre α

met effectivement d'améliorer l'ordonnement final des documents. Les meilleures valeurs de paramètre α sont obtenues entre 0.8 et 0.9.

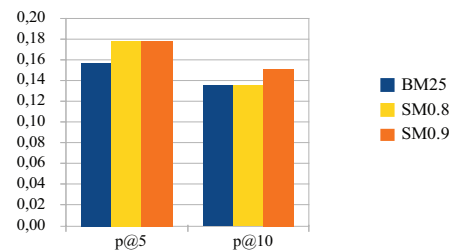


FIG. 5 – Évaluation l'efficacité de notre modèle

Nous avons comparé notre modèle avec un système de recherche d'information classique basé sur le modèle BM25 et utilisant l'algorithme de lemmatisation "SnowBall Stemmer". Nous utilisons le même modèle pour sélectionner les documents pertinents et pour calculer le score pertinence thématique $RSV(q, d)$. Comme décrit dans la figure 5, les meilleures valeurs du paramètre α permettent d'aboutir à une amélioration jusqu'à 13% par rapport la *baseline* BM25. Nous concluons donc que la combinaison de la pertinence thématique et l'importance sociale des documents permet d'accéder au ressources bibliographiques de qualité et d'améliorer l'efficacité de la recherche.

En fait, les tags utilisés comme des requêtes dans notre évaluation expérimentale sont des termes utilisateur et ne figurent pas forcément dans le contenu du document. Par conséquent, seuls quelques documents pertinents sont sélectionnés ce qui explique ainsi la faible précision de la *baseline* qui affecte directement performances du modèle proposé.

En outre, les résultats sont comparables au modèle de recherche d'information classique basé sur la pertinence thématique. L'objectif principal des notre proposition est d'améliorer l'efficacité de la recherche en combinant l'importance sociale du document et sa pertinence thématique. Nous avons atteint cet objectif avec un taux d'amélioration de 13% par rapport à la *baseline* BM25.

6 Conclusion

Nous avons proposé un modèle de recherche d'information sociale générique puis nous l'avons instancié pour l'accès aux ressources bibliographiques. Ce modèle a la spécificité d'intégrer les relations sociales de citation, de production et d'annotation ainsi que la pondération des différentes relations sociales. Notre évaluation expérimentale sur la collection des documents scientifiques SIGIR montre que la mesure de *Hub* est la mesure qui permet de mieux évaluer l'importance sociale des documents scientifiques et prouve la supériorité de notre modèle comparativement à un modèle de recherche d'information classique.

En perspective, nous envisageons de comparer les performances de notre approche aux autres modèles qui intègrent la composante sociale. De plus, nous mènerons les expérimentations sur une collection de ressources bibliographiques de plus grande taille.

Références

- Amer-Yahia, S., M. Benedikt, et P. Bohannon (2007). Challenges in searching online communities. *IEEE Data Eng. Bull.* 30(2), 23–31.
- Baeza-Yates, R. et B. Ribeiro-Neto (1999). *Modern Information Retrieval*. New York : ACM Press.
- Bollen, J., H. V. de Sompel, J. A. Smith, et R. Luce (2005). Toward alternative metrics of journal impact : A comparison of download and citation data. *Information Processing & Management* 41(6), 419 – 1440. Special Issue on Infometrics.
- Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA* 295(1), 90–93.
- Hauff, C. et L. Azzopardi (2005). Age dependent document priors in link structure analysis. In *ECIR*, pp. 552–554.
- Jones, K. S., S. Walker, et S. E. Robertson (2000). A probabilistic model of information retrieval : development and comparative experiments. *Inf. Process. Manage.* 36(6), 779–808.
- Kazai, G. et N. Milic-Frayling (2008). Trust, authority and popularity in social information retrieval. In *CIKM '08*, New York, NY, USA, pp. 1503–1504. ACM.
- Kirchhoff, L., K. Stanoevska-Slabeva, T. Nicolai, et M. Fleck (2008). Using social network analysis to enhance information retrieval systems. In *in Applications of Social Network Analysis (ASNA) (Zurich)*, 12-9-2008.
- Kirsch, S. M., M. Gnasa, et A. B. Cremers (2006). Beyond the web : Retrieval in social information spaces. In *In Proceedings of the 28 th European Conference on Information Retrieval (ECIR 2006)*. Springer.
- Korfiatis, N. T., M. Poulos, et G. Bokus (2006). Evaluating authoritative sources using social networks : an insight from wikipedia. *Online Information Review* 30(3), 252–262.
- Langville, A. N. et C. D. Meyer (2005). A survey of eigenvector methods for web information retrieval. *SIAM Rev.* 47(1), 135–161.

- Liu, X., J. Bollen, M. L. Nelson, et H. Van de Sompel (2005). Co-authorship networks in the digital library research community. *Inf. Process. Manage.* 41(6), 1462–1480.
- Meij, E. et M. de Rijke (2007). Using prior information derived from citations in literature search. In *RIAO*.
- Mutschke, P. (2001). Enhancing information retrieval in federated bibliographic data sources using author network based stratagems. In *Reserach and Advanced Technology for Digital Libraries : 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001 ; Proceedings*.
- Newman, M. E. J. (2000). Who is the best connected scientist ? a study of scientific coauthorship networks. Working Papers 00-12-064, Santa Fe Institute.
- Page, L., S. Brin, R. Motwani, et T. Winograd (1998). The pagerank citation ranking : Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.
- Tamine, L., A. B. Jabeur, et W. Bahsoun (2009). An exploratory study on using social information networks for flexible literature access. In *FQAS*, pp. 88–98.
- Wasserman, S., K. Faust, et D. Iacobucci (1994). *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.
- Yan, E. et Y. Ding (2009). Applying centrality measures to impact analysis : A coauthorship network analysis. *J. Am. Soc. Inf. Sci. Technol.* 60(10), 2107–2118.